CHAPTER 2

Descriptive Inference

Social science research, whether quantitative or qualitative, involves the dual goals of describing and explaining. Some scholars set out to describe the world; others to explain. Each is essential. We cannot construct meaningful causal explanations without good description; description, in turn, loses most of its interest unless linked to some causal relationships. Description often comes first; it is hard to develop explanations before we know something about the world and what needs to be explained on the basis of what characteristics. But the relationship between description and explanation is interactive. Sometimes our explanations lead us to look for descriptions of different parts of the world; conversely, our descriptions may lead to new causal explanations.

Description and explanation both depend upon rules of scientific inference. In this chapter we focus on description and descriptive inference. Description is far from mechanical or unproblematic since it involves selection from the infinite number of facts that could be recorded. There are several fundamental aspects of scientific description. One is that it involves inference: part of the descriptive task is to infer information about unobserved facts from the facts we have observed. Another aspect involves distinguishing between that which is systematic about the observed facts and that which is nonsystematic.

As should be clear, we disagree with those who denigrate "mere" description. Even if explanation—connecting causes and effects—is the ultimate goal, description has a central role in all explanation, and it is fundamentally important in and of itself. It is not description versus explanation that distinguishes scientific research from other research; it is whether systematic inference is conducted according to valid procedures. Inference, whether descriptive or causal, quantitative or qualitative, is the ultimate goal of all good social science. Systematically collecting facts is a very important endeavor without which science would not be possible but which does not by itself constitute science. Good archival work or well-done summaries of historical facts may make good descriptive history, but neither are sufficient to constitute social science.

In this chapter, we distinguish description—the collection of facts—from descriptive inference. In section 2.1 we discuss the relationship between the seemingly contradictory goals of scholarship: discovering general knowledge and learning about particular facts. We are then able to explain in more detail the concept of inference in section 2.2. Our approach in the remainder of the book is to present ideas both verbally and through very simple algebraic models of research. In section 2.3 we consider the nature of these models. We then discuss models for data collection, for summarizing historical detail, and for descriptive inference in sections 2.4, 2.5, and 2.6, respectively. Finally, we provide some specific criteria for judging descriptive inferences in section 2.7.

2.1 General Knowledge and Particular Facts

The world that social scientists study is made up of particulars: individual voters, particular government agencies, specific cities, tribes, groups, states, provinces, and nations. Good social science attempts to go beyond these particulars to more general knowledge. Generalization, however, does not eliminate the importance of the particular. In fact, the very purpose of moving from the particular to the general is to improve our understanding of both. The specific entities of the social world—or, more precisely, specific facts about these entities—provide the basis on which generalizations must rest. In addition, we almost always learn more about a specific case by studying more general conclusions. If we wish to know why the foreign minister of Brazil resigned, it will help to learn why other ministers resigned in Brazil, why foreign ministers in other countries have resigned, or why people in general resign from government or even nongovernmental jobs. Each of these will help us understand different types of general facts and principles of human behavior, but they are very important even if our one and only goal is to understand why the most recent Brazilian foreign minister resigned. For example, by studying other ministers, we might learn that all the ministers in Brazil resigned to protest the actions of the president, something we might not have realized by examining only the actions of the foreign minister.

Some social science research tries to say something about a class of events or units without saying anything in particular about a specific event or unit. Studies of voting behavior using mass surveys explain the voting decisions of people in general, not the vote of any particular individual. Studies of congressional finance explain the effect of money on electoral outcomes across all congressional districts. Most such studies would not mention the Seventh Congressional District in Pennsylvania or any other district except, perhaps, in passing or as exceptions to a general rule. These studies follow the injunction of

Przeworski and Teune (1982): eliminate proper names. However, though these studies may not seek to understand any particular district, they should not ignore—as sometimes is unfortunately done in this tradition—the requirement that the facts about the various districts that go into the general analysis must be accurate.

Other research tries to tell us something about a particular instance. It focuses on the French Revolution or some other "important" event and attempts to provide an explanation of how or why that event came about. Research in this tradition would be unthinkable—certainly uninteresting to most of the usual readers of such research—without proper names. A political scientist may write effectively about patterns of relationships across the set of congressional campaigns without looking at specific districts or specific candidates but imagine Robert Caro's discussion (1982) of the 1948 Senate race in Texas without Lyndon Johnson and Coke Stevenson.[1] Particular events such as the French Revolution or the Democratic Senate primary in Texas in 1948 may indeed be of intrinsic interest: they pique our curiosity, and if they were preconditions for subsequent events (such as the Napoleonic Wars or Johnson's presidency), we may need to know about them to understand those later events. Moreover, knowledge about revolution, rebellion, or civil war in general will provide invaluable information for any more focused study of the causes of the French Revolution in particular.

We will consider these issues by discussing "interpretation," a claimed alternative to scientific inference (section 2.1.1); the concepts of uniqueness and complexity of the subject of study (section 2.1.2); and the general area of comparative case studies (section 2.1.3).

2.1.1 "Interpretation" and Inference

In the human sciences, some historical and anthropological researchers claim to seek only specific knowledge through what they call "interpretation." Interpretivists seek accurate summaries of historical detail. They also seek to place the events they describe in an intelligible context within which the meaning of actions becomes explicable. As Ferejohn (in Goldstein and Keohane 1993:228) has written, "We want

[1] Nor can we dismiss Caro as someone in another business: a journalist/biographer whose goal differs from that of the social scientist. His work addresses some of the same issues that a political scientist would. What leads to success or failure in an election campaign? What is the role of money and campaign finance in electoral success? What motivates campaign contributions? The discussion focuses on a particular candidacy in a particular district, but the subject matter and the puzzles posed overlap with standard political science.

social science theories to provide causal explanations of events . . . [and] to give an account of the reasons for or meanings of social action. We want to know not only what caused the agent to perform some act but also the agent's reasons for taking the action." Geertz (1973:17) also writes that "it is not in our interest to bleach human behavior of the very properties that interest us before we begin to examine it."

Scholars who emphasize "interpretation" seek to illuminate the intentional aspects of human behavior by employing *Verstehen* ("empathy: understanding the meaning of actions and interactions from the members' own points of view" [Eckstein 1975:83]). Interpretivists seek to explain the reasons for intentional action in relation to the whole set of concepts and practices in which it is embedded. They also employ standards of evaluation. "The most obvious standards are coherence and scope: an interpretative account should provide maximal coherence or intelligibility to a set of social practices, and an interpretative account of a particular set of practices should be consistent with other practices or traditions of the society" (Moon 1975: 173).

Perhaps the single most important operational recommendation of the interpretivists is that researchers should learn a great deal about a culture prior to formulating research questions. For only with a deep cultural immersion and understanding of a subject can a researcher ask the right questions and formulate useful hypotheses. For example, Duneier (1993) studied the collective life of working-class black and white men at one integrated cafeteria in Chicago. By immersing himself in this local culture for four years, he noticed several puzzles that had not previously occurred to him. For example, he observed that although these men were highly antagonistic to the Republican party, they articulated socially conservative positions on many issues.

Some scholars push the role of interpretation even further, going so far as to suggest that it is a wholly different paradigm of inquiry for the social sciences, "not an experimental science in search of law but an interpretive one in search of meaning" (Geertz 1973:5). In our view, however, science (as we have defined it in section 1.1.2) and interpretation are not fundamentally different endeavors aimed at divergent goals. Both rely on preparing careful descriptions, gaining deep understandings of the world, asking good questions, formulating falsifiable hypotheses on the basis of more general theories, and collecting the evidence needed to evaluate those hypotheses. The distinctive contribution of science is to present a set of procedures for discovering the answers to appropriately framed descriptive and causal questions.

Our emphasis on the methodology of inference is not intended to denigrate the significance of the process by which fruitful questions are formulated. On the contrary, we agree with the interpretivists that

it is crucial to understand a culture deeply before formulating hypotheses or designing a systematic research project to find an answer. We only wish to add that evaluating the veracity of claims based on methods such as participant observation can only be accomplished through the logic of scientific inference, which we describe. Finding the right answers to the wrong questions is a futile activity. Interpretation based on *Verstehen* is often a rich source of insightful hypotheses. For instance, Richard Fenno's close observations of Congress (Fenno 1978), made through what he calls "soaking and poking," have made major contributions to the study of that institution, particularly by helping to frame better questions for research. "Soaking and poking," says Putnam in a study of Italian regions (1993:12), "requires the researcher to marinate herself in the minutiae of an institution—to experience its customs and practices, its successes and its failings, as those who live it every day do. This immersion sharpens our intuitions and provides innumerable clues about how the institution fits together and how it adapts to its environment." Any definition of science that does not include room for ideas regarding the generation of hypotheses is as foolish as an interpretive account that does not care about discovering truth.

Yet once hypotheses have been formulated, demonstrating their correctness (with an estimate of uncertainty) requires valid scientific inferences. The procedures for inference followed by interpretivist social scientists, furthermore, must incorporate the same standards as those followed by other qualitative and quantitative researchers. That is, while agreeing that good social science requires insightful interpretation or other methods of generating good hypotheses, we also insist that science is essential for accurate interpretation. If we could understand human behavior only through *Verstehen*, we would never be able to falsify our descriptive hypotheses or provide evidence for them beyond our experience. Our conclusions would never go beyond the status of untested hypotheses, and our interpretations would remain personal rather than scientific.

One of the best and most famous examples in the interpretative tradition is Clifford Geertz's analysis of Gilbert Ryle's discussion of the difference between a twitch and a wink. Geertz (1973:6) writes

> Consider . . . two boys rapidly contracting the eyelids of their right eyes. In one, this is an involuntary twitch; in the other, a conspiratorial signal to a friend. The two movements are, as movements, identical; from an I-am-a-camera, "phenomenalistic" observation of them alone, one could not tell which was twitch and which was wink, or indeed whether both or either was twitch or wink. Yet the difference, however unphotographable, between a twitch and a wink is vast; as anyone unfortunate enough to have had the first taken for the second knows. The winker is communicating, and indeed communicating in a precise and special way: (1) deliberately, (2) to someone in particular, (3) to impart a particular message, (4) according to a socially established code, and (5) without cognizance of the rest of the company. As Ryle points out, the winker has done two things, contracted his eyelids and winked, while the twitcher has done only one, contracted his eyelids. Contracting your eyelids on purpose when there exists a public code in which doing so counts as a conspiratorial signal is winking.

Geertz is making an important conceptual point. Without the concept of "winking," given meaning by a theory of communication, the most precise quantitative study of "eyelid contracting by human beings" would be meaningless for students of social relations. In this example, the theory, which emerged from months of "soaking and poking" and detailed cultural study, is essential to the proper question of whether eyelid contraction even could be "twitches" or "winks." The magnificent importance of interpretation suggested by this example is clear: it provides new ways of looking at the world—new concepts to be considered and hypotheses to be evaluated. Without deep immersion in a situation, we might not even think of the right theories to evaluate. In the present example, if we did not think of the difference between twitches and winks, everything would be lost. If interpretation—or anything else—helps us arrive at new concepts or hypotheses, then it is unquestionably useful, and interpretation, and similar forms of detailed cultural understanding, have been proven again and again.

Having made a relevant theoretical distinction, such as that between a wink and a twitch, the researcher then needs to *evaluate* the hypothesis that winking is taking place. It is in such evaluation that the logic of scientific inference is unsurpassed. That is, the best way of determining the meaning of eyelid contractions is through the systematic methods described in this book. If distinguishing a twitch from wink were pivotal, we could easily design a research procedure to do so. If, for instance, we believe that particular eyelid contractions are winks imbued with political meaning, then other similar instances must also be observable, since a sophisticated signaling device such as this (a "public code"), once developed, is likely to be used again. Given this likelihood, we might record every instance in which this actor's eyelid contracts, observe whether the other key actor is looking at the right time, and whether he responds. We could even design a series of experiments to see if individuals in this culture are accustomed to communicating in this fashion. Understanding the culture, carefully de-

scribing the event, and having a deep familiarity with similar situations will all help us ask the right questions and even give us additional confidence in our conclusions. But only with the methods of scientific inference will we be able to evaluate the hypothesis and see whether it is correct.

Geertz's wink interpretation is best expressed as a causal hypothesis (which we define precisely in section 3.1): the hypothetical causal effect of the wink on the other political actor is the other actor's response given the eyelid contraction minus his response if there were no movement (and no other changes). If the eyelid contraction were a wink, the causal effect would be positive; if it were only a twitch, the causal effect would be zero. If we decided to estimate this causal effect (and thus find out whether it was a wink or a twitch), all the problems of inference discussed at length in the rest of this book would need to be understood if we were to arrive at the best inference with respect to the interpretation of the observed behavior.

If what we interpret as winks were actually involuntary twitches, our attempts to derive causal inferences about eyelid contraction on the basis of a theory of voluntary social interaction would be routinely unsuccessful: we would not be able to generalize and we would know it.[2]

Designing research to distinguish winks and twitches is not likely to be a major part of most political science research, but the same methodological issue arises in much of the subject area in which political scientists work. We are often called on to interpret the meaning of an act. Foreign policy decision makers send messages to each other. Is a particular message a threat, a negotiating point, a statement aimed at appealing to a domestic audience? Knowledge of cultural norms, of conventions in international communications, and of the history of particular actors, as well as close observation of ancillary features of the communication, will all help us make such an interpretation. Or consider the following puzzle in quantitative research: Voters in the United States seem to be sending a message by not turning out at the polls. But what does the low turnout mean? Does it reflect alienation with the political system? A calculation of the costs and benefits of voting with the costs being greater? Disappointment with recent candidates or recent campaigns? Could it be a consequence of a change in the minimum age of voting? Or a sign that nothing is sufficiently upsetting to get them to the polls? The decision of a citizen not to vote, like a wink or a diplomatic message, can mean many things. The sophisticated researcher should always work hard to ask the right questions and then carefully design scientific research to find out what the ambiguous act did in fact mean.

[2] For the sake of completeness, it is worth noting that we *could* imagine an altogether different theory in which an eyelid contraction was not a wink but still had a causal effect on other actors. For example, the twitch could have been misinterpreted. If we were also interested in whether the person with the eyelid contraction *intended* to wink, we would need to look for other observable consequences of this same theory.

We would also like to briefly address the extreme claims of a few proponents of interpretation who argue that the goal of some research ought to be feelings and musings with no observable consequences. This is hardly a fair characterization of all but a small minority of researchers in this tradition, but the claims are made sufficiently forcefully that they seem worth addressing explicitly. Like the over-enthusiastic claims of early positivists, who took the untenable position that unobservable concepts had no place in scientific research, these arguments turn out to be inappropriate for empirical research. For example, Psathas (1968:518) argues that

> any behavior by focusing only on that part which is overt and manifested in concrete, directly observable acts is naive, to say the least. The challenge to the social scientist who seeks to understand social reality, then, is to understand the meaning that the actor's act has for him.

Psathas may be correct that social scientists who focus on only overt, *observable*, behaviors are missing a lot, but how are we to know if we cannot see? For example, if two theories of self-conception have identical observable manifestations, then no observer will have sufficient information to distinguish the two. This is true no matter how clever or culturally sensitive the observer is, how skilled she is at interpretation, how well she "brackets" her own presuppositions, or how hard she tries. Interpretation, feeling, thick description, participant observation, nonparticipant observation, depth interviewing, empathy, quantification and statistical analysis, and all other procedures and methods are inadequate to the task of distinguishing two theories without differing observable consequences. On the other hand, if the two theories have some observable manifestations that differ, then the methods we describe in this book provide ways to distinguish between them.

In practice, ethnographers (and all other good social scientists) *do* look for observable behavior in order to distinguish among their theories. They may immerse themselves in the culture, but they all rely on various forms of observation. Any further "understanding" of the cultural context comes directly from these or other comparable observations. Identifying relevant observations is not always easy. On the contrary, finding the appropriate observations is perhaps the most difficult part of a research project, especially (and necessarily) for those areas of inquiry traditionally dominated by qualitative research.

2.1.2 "Uniqueness," Complexity, and Simplification

Some qualitatively oriented researchers would reject the position that general knowledge is either necessary or useful (perhaps even possible) as the basis for understanding a particular event. Their position is that the events or units they study are "unique." In one sense, they are right. There was only one French Revolution and there is only one Thailand. And no one who has read the biographical accounts or who lived through the 1960s can doubt the fact that there was only one Lyndon B. Johnson. But they go further. Explanation, according to their position, is limited to that unique event or unit: not why revolutions happen, but why the French Revolution happened; not why democratization sometimes seems to lag, but why it lags in Thailand; not why candidates win, but why LBJ won in 1948 or 1964. Researchers in this tradition believe that they would lose their ability to explain the specific if they attempted to deal with the general—with revolutions or democratization or senatorial primaries.

"Uniqueness," however, is a misleading term. The French Revolution and Thailand and LBJ are, indeed, unique. All phenomena, all events, are in some sense unique. The French Revolution certainly was; but so was the congressional election in the Seventh District of Pennsylvania in 1988 and so was the voting decision of every one of the millions of voters who voted in the presidential election that year. Viewed holistically, every aspect of social reality is infinitely complex and connected in some way to preceding natural and sociological events. Inherent uniqueness, therefore, is part of the human condition: it does not distinguish situations amenable to scientific generalizations from those about which generalizations are not possible. Indeed, as we showed in discussing theories of dinosaur extinction in chapter 1, even unique events can be studied scientifically by paying attention to the observable implications of theories developed to account for them.

The real question that the issue of uniqueness raises is the problem of complexity. The point is not whether events are inherently unique, but whether the key features of social reality that we want to understand can be abstracted from a mass of facts. One of the first and most difficult tasks of research in the social sciences is this act of *simplification*. It is a task that makes us vulnerable to the criticism of oversimplification and of omitting significant aspects of the situation. Nevertheless, such simplication is inevitable for all researchers. Simplification has been an integral part of every known scholarly work—quantitative and qualitative, anthropological and economic, in the social sciences and in the natural and physical sciences—and will probably always be. Even the most comprehensive description done by the best cultural interpreters with the most detailed contextual understanding will drastically simplify, reify, and reduce the reality that has been observed. Indeed, *the difference between the amount of complexity in the world and that in the thickest of descriptions is still vastly larger than the difference between this thickest of descriptions and the most abstract quantitative or formal analysis.* No description, no matter how thick, and no explanation, no matter how many explanatory factors go into it, comes close to capturing the full "blooming and buzzing" reality of the world. There is no choice but to simplify. Systematic simplification is a crucial step to useful knowledge. As an economic historian has put it, if emphasis on uniqueness "is carried to the extreme of ignoring all regularities, the very possibility of social science is denied and historians are reduced to the aimlessness of balladeers" (Jones 1981:160).

Where possible, analysts should simplify their descriptions only after they attain an understanding of the richness of history and culture. Social scientists may use only a few parts of the history of some set of events in making inferences. Nevertheless, rich, unstructured knowledge of the historical and cultural context of the phenomena with which they want to deal in a simplified and scientific way is usually a requisite for avoiding simplications that are simply wrong. Few of us would trust the generalizations of a social scientist about revolutions or senatorial elections if that investigator knew little and cared less about the French Revolution or the 1948 Texas election.

In sum, we believe that, where possible, social science research should be both general and specific: it should tell us something about classes of events as well as about specific events at particular places. We want to be timeless and timebound at the same time. The emphasis on either goal may vary from research endeavor to research endeavor, but both are likely to be present. Furthermore, rather than the two goals being opposed to each other, they are mutually supportive. Indeed, *the best way to understand a particular event may be by using the methods of scientific inference also to study systematic patterns in similar parallel events.*

2.1.3 Comparative Case Studies

Much of what political scientists do is describe politically important events systematically. People care about the collapse of the Soviet Union, the reactions of the public in Arab countries to the UN-authorized war to drive Iraq from Kuwait, and the results of the latest congressional elections in the United States. And they rely on political sci-

entists for descriptions that reflect a more comprehensive awareness of the relationship between these and other relevant events—contemporary and historical—than is found in journalistic accounts. Our descriptions of events should be as precise and systematic as possible. This means that when we are able to find valid quantitative measures of what we want to know, we should use them: What proportion of Soviet newspapers criticize government policy? What do public opinion polls in Jordan and Egypt reveal about Jordanian and Egyptian attitudes toward the Gulf war? What percentage of congressional incumbents were reelected?

If quantification produces precision, it does not necessarily encourage accuracy, since inventing quantitative indices that do not relate closely to the concepts or events that we purport to measure can lead to serious measurement error and problems for causal inference (see section 5.1). Similarly, there are more and less precise ways to describe events that cannot be quantified. Disciplined qualitative researchers carefully try to analyze constitutions and laws rather than merely report what observers say about them. In doing case studies of government policy, researchers ask their informants trenchant, well-specified questions to which answers will be relatively unambiguous, and they systematically follow up on off-hand remarks made by an interviewee that suggest relevant hypotheses. Case studies are essential for description, and are, therefore, fundamental to social science. It is pointless to seek to explain what we have not described with a reasonable degree of precision.

To provide an insightful description of complex events is no trivial task. In fields such as comparative politics or international relations, descriptive work is particularly important because there is a great deal we still need to know, because our explanatory abilities are weak, and because good description depends in part on good explanation. Some of the sources of our need-to-know and explanatory weaknesses are the same: in world politics, for instance, patterns of power, alignments, and international interdependence have all been changing rapidly recently, both increasing the need for good description of new situations, and altering the systemic context within which observed interactions between states take place. Since states and other actors seek to anticipate and counter others' actions, causality is often difficult to establish, and expectations may play as important a part as observed actions in accounting for state behavior. A purported explanation of some aspect of world politics that assumes the absence of strategic interaction and anticipated reactions will be much less useful than a careful description that focuses on events that we have reason to believe are important and interconnected. Good description is better than bad explanation.

One of the often overlooked advantages of the in-depth case-study method is that the development of good causal hypotheses is complementary to good description rather than competitive with it. Framing a case study around an explanatory question may lead to more focused and relevant description, even if the study is ultimately thwarted in its attempt to provide even a single valid causal inference.

Comparative case studies can, we argue, yield valid causal inferences when the procedures described in the rest of this book are used, even though as currently practiced they often do not meet the standards for valid inference (which we explicate in chapter 3). Indeed, much of what is called "explanatory" work by historically-oriented or interpretative social scientists remains essentially descriptive because it does not meet these universally applicable standards. From this perspective, the advice of a number of scholars that comparative case studies must be be more systematic for description or explanation is fundamental.

For example, Alexander George recommends a method of "structured, focused comparison" that emphasizes discipline in the way one collects data (George and McKeown 1985; see also Verba 1967). George and his collaborators stress the need for a systematic collection of the same information—the same variables—across carefully selected units. And they stress the need for theoretical guidance—for asking carefully thought-out explanatory questions—in order to accomplish this systematic description, if causal inference is to be ultimately possible.[7]

The method of structured, focused comparison is a systematic way to employ what George and McKeown call the congruence procedure. Using this method, the investigator "defines and standardizes the data requirements of the case studies . . . by formulating theoretically relevant general questions to guide the examination of each case" (George and McKeown 1985:43). The point that George and McKeown (1985: 43) make is well taken: "Controlled comparison of a small n should follow a procedure of systematic data compilation." Such "structured-focused comparison" requires collecting data on the same variables across units. Thus, it is not a different method from the one that we emphasize here so much as it is a way of systematizing the information in descriptive case studies in such a way that it could conceivably

[7] The literature on comparative case studies is vast. Some of the best additional works are Eckstein (1975); Lijphart (1971); and Collier (1991).

be used for descriptive or causal inference. Much valuable advice about doing comparative case studies, such as this, is rudimentary but often ignored.

2.2 Inference: The Scientific Purpose of Data Collection

Inference is the process of using the facts we know to learn about facts we do not know. The facts we do not know are the subjects of our research questions, theories, and hypotheses. The facts we do know form our (quantitative or qualitative) data or observations.

In seeking general knowledge, for its own sake or to understand particular facts better, we must somehow avoid being overwhelmed by the massive cacophony of potential and actual observations about the world. Fortunately, the solution to that problem lies precisely in the search for general knowledge. That is, the best scientific way to organize facts is as observable implications of some theory or hypothesis. Scientific simplification involves the productive choice of a theory (or hypothesis) to evaluate; the theory then guides us to the selection of those facts that are implications of theory. Organizing facts in terms of observable implications of a specific theory produces several important and beneficial results in designing and conducting research. First, with this criterion for the selection of facts, we can quickly recognize that more observations of the implications of a theory will only help in evaluating the theory in question. Since more information of this sort cannot hurt, such data are never discarded, and the process of research improves.

Second, we need not have a complete theory before collecting data nor must our theory remain fixed throughout. Theory and data interact. As with the chicken and the egg, some theory is always necessary before data collection and some data are required before any theorizing. Textbooks on research tell us that we use our data to test our theories. But learning from the data may be as important a goal as evaluating prior theories and hypotheses. Such learning involves reorganizing our data into observable implications of the new theory. This reorganizing is very common early in many research processes, usually after some preliminary data have been collected; after the reorganization, data collection then continues in order to evaluate the new theory. We should always try to continue to collect data even after the reorganization in order to test the new theory and thus avoid using the same data to evaluate the theory that we used to develop it.[4]

[4] For example, Coombs (1964) demonstrated that virtually every useful data collection

Third, the emphasis on gathering facts as observable implications of a hypothesis makes the common ground between the quantitative and qualitative styles of research much clearer. In fact, once we get past thinking of cases or units or records in the usual very narrow or even naive sense, we realize that most qualitative studies potentially provide a very large number of observable implications for the theories being evaluated, yet many of these observations may be overlooked by the investigator. Organizing the data into a list of the specific observable implications of a theory thus helps reveal the essential scientific purpose of much qualitative research. In a sense, we are asking the scholar who is studying a particular event—a particular government decision, perhaps—to ask: "If my explanation is correct of why the decision came out the way it did, what else might I expect to observe in the real world?" These additional observable implications might be found in other decisions, but they might also be found in other aspects of the decision being studied: for instance, when it was made, how it was made, how it was justified. The crucial maxim to guide both theory creation and data gathering is: search for more observable implications of the theory.

Each time we develop a new theory or hypothesis, it is productive to list all implications of the theory that could, in principle, be observed. The list, which could then be limited to those items for which data have been or could easily be collected, then forms the basic operational guide for a research project. If collecting one additional datum will help provide one additional way to evaluate a theory, then (subject to the usual time, money, and effort constraints) it is worth doing. If an interview or other observation might be interesting but is not a potential observable implication of this (or some other relevant) theory, then it should be obvious that it will not help us evaluate our theory.

As part of the simplification process accomplished by organizing our data into observable implications of a theory, we need to systematize the data. We can think about converting the raw material of real-world phenomena into "classes" that are made up of "units" or "cases" which are, in turn, made up of "attributes" or "variables" or "parameters." The class might be "voters"; the units might be a sample of "voters" in several congressional districts; and the attributes or

---

task requires or implies some degree of theory, or "mini-theory." However, much quantitative data and qualitative history is collected with the explicit purpose of encouraging future researchers to use them for purposes previously unforeseen. Fifteen minutes with the *Statistical Abstract of the United States* will convince most people of this point. Data-collection efforts also differ in the degree to which researchers rigidly follow prior beliefs.

variables might be income, party identification, or anything that is an observable implication of the theory being evaluated. Or the class might be a particular kind of collectivity such as communities or countries, the units might be a selection of these, and the attributes or variables might be their size, the type of government, their economic circumstances, their ethnic composition, or whatever else is measureable and of interest to the researcher. These concepts, as well as various other constructs such as typologies, frameworks, and all manner of classifications, are useful as temporary devices when we are collecting data but have no clear hypothesis to be evaluated. However, in general, we encourage researchers not to organize their data in this way. Instead, we need only the organizing concept inherent in our theory. That is, our observations are either implications of our theory or irrelevant. If they are irrelevant or not observable, we should ignore them. If they are relevant, then we should use them. Our data need not all be at the same level of analysis. Disaggregated data, or observations from a different time period, or even from a different part of the world, may provide additional observable implications of a theory. We may not be interested at all in these subsidiary implications, but if they are consistent with the theory, as predicted, they will help us build confidence in the power and applicability of the theory. Our data also need not be "symmetric": we can have a detailed study of one province, a comparative study of two countries, personal interviews with government leaders from only one policy sector, and even a quantitative component—just so long as each is an observable consequence of our theory. In this process, we go beyond the particular to the general, since the characterization of particular units on the basis of common characteristics is a generalizing process. As a result, we learn a lot more about both general theories and particular facts.

In general, we wish to bring as much information to bear on our hypothesis as possible. This may mean doing additional case studies, but that is often too difficult, time consuming, or expensive. We obviously should not bring in irrelevant information. For example, treating the number of conservative-held seats in the British House of Commons as a monthly variable instead of one which changes at each national election, would increase the number of observations substantially but would make no sense since little new information would be added. On the other hand, disaggregating U.S. presidential election results to the state or even county level increases both the number of cases and the amount of information brought to bear on the problem.

Such disaggregated information may seem irrelevant since the goal is to learn about the causes of a particular candidate's victory in a race for the presidency—a fundamentally aggregate-level question. However, most explanations of the outcome of the presidential election have different observable implications for the disaggregated units. If, for instance, we predict the outcome of the presidential election on the basis of economic variables such as the unemployment rate, the use of the unemployment rates on a state-by-state basis provides many more observations of the implications of our theory than does the aggregate rate for the nation as a whole. By verifying that the theory holds in these other situations—even if these other situations are not of direct interest—we increase the confidence that the theory is correct and that it correctly explains the one observable consequence of the theory that is of interest.

2.3 Formal Models of Qualitative Research

A *model* is a simplification of, and approximation to, some aspect of the world. Models are never literally "true" or "false," although good models abstract only the "right" features of the reality they represent.

For example, consider a six-inch toy model of an airplane made of plastic and glue. This model is a small fraction of the size of the real airplane, has no moving parts, cannot fly, and has no contents. None of us would confuse this model with the real thing; asking whether any aspect of the model is true is like asking whether the model who sat for Leonardo DaVinci's *Mona Lisa* really had such a beguiling smile. Even if she did, we would not expect Leonardo's picture to be an exact representation of anyone, whether the actual model or the Virgin Mary, any more than we would expect an airplane model fully to reflect all features of an aircraft. However, we *would* like to know whether this model abstracts the correct features of an airplane for a particular problem. If we wish to communicate to a child what a real airplane is like, this model might be adequate. If built to scale, the model might also be useful to airplane designers for wind tunnel tests. The key feature of a real airplane that this model abstracts is its shape. For some purposes, this is certainly one of the right features. Of course, this model misses myriad details about an airplane, including size, color, the feeling of being on the plane, strength of its various parts, number of seats on board, power of its engines, fabric of the seat cushions, and electrical, air, plumbing, and numerous other critical systems. If we wished to understand these aspects of the plane, we would need an entirely different set of models.

Can we evaluate a model without knowing which features of the subject we wish to study? Clearly not. For example, we might think that a model that featured the amount of dirt on an airplane would not be of much use. Indeed, for the purposes of teaching children or wind

tunnel tests, it would be largely irrelevant. However, since even carpet dust can cause a plane to weigh more and thus use more expensive fuel, models of this sort are important to the airline industry and have been built (and saved millions of dollars).

All models range between restrictive and unrestrictive versions. Restrictive models are cleaner, more parsimonious, and more abstract, but they are also less realistic (unless the world really is parsimonious). Models which are unrestrictive are detailed, contextual, and more realistic, but they are also less clear and harder to estimate with precision (see King 1989: section 2.5). Where on this continuum we choose to construct a model depends on the purpose for which it is to be put and on the complexity of the problem we are studying.

Whereas some models are physical, others are pictorial, verbal, or algebraic. For example, the qualitative description of European judicial systems in a book about that subject is a model of that event. No matter how thick the description or talented the author, the book's account will always be an abstraction or simplification compared to the actual judicial system. Since understanding requires some abstraction, the sign of a good book is as much what is left out as what is included.

While qualitative researchers often use verbal models, we will use algebraic models in our discussion below to study and improve these verbal models. Just as with models of toy airplanes and book-long studies of the French Revolution, our algebraic models of qualitative research should not be confused with qualitative research itself. They are only meant to provide especially clear statements of problems to avoid and opportunities to exploit. In addition, we often find that they help us to discover ideas that we would not have thought of otherwise.

We assume that readers have had no previous experience with algebraic models, although those with exposure to statistical models will find some of the models that follow familiar. But the logic of inference in these models applies to both quantitative and qualitative research. Just because quantitative researchers are probably more familiar with our terminology does not mean that they are any better at applying the logic of scientific inference. Moreover, these models do *not* apply more closely to quantitative than to qualitative research; in both cases, the models *are* useful abstractions of the research to which they are applied. To ease their introduction, we introduce all algebraic models with verbal descriptions, followed by a box where we use standard algebraic notation. Although we discourage it, the boxes may be skipped without loss of continuity.

2.4 A Formal Model of Data Collection

Before formalizing our presentation of descriptive and causal inference—the two primary goals of social science research—we will develop a model for the data to be collected and for summarizing these data. This model is quite simple, but it is a powerful tool for analyzing problems of inference. Our algebraic model will not be as formal as that in statistics but nevertheless makes our ideas clearer and easier to convey. By *data collection*, we refer to a wide range of methods, including observation, participant observation, intensive interviews, large-scale sample surveys, history recorded from secondary sources, randomized experiments, ethnography, content analyses, and any other method of collecting reliable evidence. *The most important rule for all data collection is to report how the data were created and how we came to possess them.* Every piece of information that we gather should contribute to specifying observable implications of our theory. It may help us develop a new research question, but it will be of no use in answering the present question if it is not an observable implication of the question we seek to answer.

We model data with *variables*, *units*, and *observations*. One simple example is the annual income of each of four people. The data might be represented simply by four numbers: $9,000, $22,000, $21,000, and $54,292. In the more general case, we could label the income of four people (numbered 1, 2, 3, and 4) as y_1, y_2, y_3, and y_4. One variable coded for two unstructured interviews might take on the values "participatory," "cooperative," or "intransigent," and might be labeled y_1 and y_2. In these examples, the *variable* is y; the *units* are the individual people; and the *observations* are the values of the variables for each unit (income for dollars or degree of cooperation). The symbol y is called a variable because its values vary over the units, and in general, a variable can represent anything whose values change over a set of units. Since we can collect information over time or across sectional areas, units may be people, countries, organizations, years, elections, or decades, and often, some combination of these or other units. Observations can be numerical, verbal, visual, or any other type of empirical data.

For example, suppose we are interested in international organizations since 1945. Before we collect our data, we need to decide what outcomes we want to explain. We could seek to understand the size distribution of international organizational activity (by issue area or by organization) in 1990; changes in the aggregate size of international organizational activity since 1945; or changes in the size distribution of

international organizational activity since 1945. Variables measuring organizational activity could include the number of countries belonging to international organizations at a given time, the number of tasks performed by international organizations, or the sizes of budgets and staffs. In these examples, the units of analysis would include international organizations, issue areas, country memberships, and time periods such as years, five-year periods, or decades. At the data-collection stage, no formal rules apply as to what variables to collect, how many units there should be, whether the units must outnumber the variables, or how well variables should be measured. The only rule is our judgment as to what will prove to be important. When we have a clearer idea of how the data will be used, the rule becomes finding as many observable implications of a theory as possible. As we emphasized in chapter 1, empirical research can be used both to evaluate a priori hypotheses or to suggest hypotheses not previously considered; but if the latter approach is followed, new data must be collected to evaluate these hypotheses.

It should be very clear from our discussion that most works labeled "case studies" have numerous variables measured over many different types of units. Although case-study research rarely uses more than a handful of cases, the total number of observations is generally immense. It is therefore essential to distinguish between the number of cases and the number of observations. The former may be of some interest for some purposes, but only the latter is of importance in judging the amount of information a study brings to bear on a theoretical question. We therefore reserve the commonly used n to refer only to the number of observations and not to the number of cases. Only occasionally, such as when individual observations are partly dependent, will we distinguish between information and the number of observations. The terminology of the number of observations comes from survey sampling where n is the number of persons to be interviewed, but we apply it much more generally. Indeed, our definition of an "observation" coincides exactly with Harry Eckstein's (1975:85) definition of what he calls a "case." As Eckstein argues, "A study of six general elections in Britain may be, but need not be, an $n = 1$ study. It might also be an $n = 6$ study. It can also be an $n = 120{,}000{,}000$ study. It depends on whether the subject of study is electoral systems, elections, or voters." The "ambiguity about what constitutes an 'individual' (hence 'case') can only be dispelled by not looking at concrete entities but at the measures made of them. On this basis, a 'case' can be defined technically as a phenomenon for which we report and interpret only a single measure on any pertinent variable." The only difference in our usage is that since Eckstein's article, scholars have continued to use the word "case" to refer to a full case study, which still has a fairly imprecise definition. Therefore, wherever possible we use the word "case" as most writers do and reserve the word "observation" to refer to measures of one or more variables on exactly one unit.

We attempt in the rest of this chapter to show how concepts like variables and units can increase the clarity of our thinking about research design even when it may be inappropriate to rely on quantitative measures to summarize the information at our disposal. The question we pose is: How can we make descriptive inferences about "history as it really was" without getting lost in a sea of irrelevant detail? In other words, how can we sort out the essential from the ephemeral?

2.5 Summarizing Historical Detail

After data are collected, the first step in any analysis is to provide summaries of the data. Summaries describe what may be a large amount of data, but they are not directly related to inference. Since we are ultimately interested in generalization and explanation, a summary of the facts to be explained is usually a good place to start but is not a sufficient goal of social science scholarship.

Summarization is necessary. We can never tell "all we know" about any set of events; it would be meaningless to try to do so. Good historians understand which events were crucial, and therefore construct accounts that emphasize essentials rather than digressions. To understand European history during the first fifteen years of the nineteenth century, we may well need to understand the principles of military strategy as Napoleon understood them, or even to know what his army ate if it "traveled on its stomach," but it may be irrelevant to know the color of Napoleon's hair or whether he preferred fried to boiled eggs. Good historical writing includes, although it may not be limited to, a compressed verbal summary of a welter of historical detail.

Our model of the process of summarizing historical detail is a *statistic*. A statistic is an expression of data in abbreviated form. Its purpose is to display the appropriate characteristics of the data in a convenient format.[5] For example, one statistic is the *sample mean*, or average:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

[5] Formally, for a set of n units on which a variable y is measured $(y_1, \ldots, y_n)$, a statistic h is a real-valued function defined as follows: $h = h(y) = h(y_1, \ldots, y_n)$.

where $\sum_{i=1}^{n} y_i$ is a convenient way of writing $y_1 + y_2 + y_3 + \ldots + y_n$. Another statistic is the *simple maximum*, labeled y_{max}:

$$y_{max} = \text{Maximum}(y_1, y_2, \ldots, y_n) \tag{2.1}$$

The sample mean of the four incomes from the example in section 2.4 ($9,000, $22,000, $21,000, and $54,292) is $26,573. The sample maximum is $54,292. We can summarize the original data containing four numbers with these two numbers representing the sample mean and maximum. We can also calculate other sample characteristics, such as the minimum, median, mode, or variance.

Each summary in this model reduces all the data (four numbers in this simple example, or our knowledge of some aspect of European history in the other) to a single number. Communicating with summaries is often easier and more meaningful to a reader than using all the original data. Of course, if we had only four numbers in a data set, then it would make little sense to use five different summaries; presenting the four original numbers would be simpler. Interpreting a statistic is generally easier than understanding the entire data set, but we necessarily lose information by describing a large set of numbers with only a few.

What rules govern the summary of historical detail? The first rule is that *summaries should focus on the outcomes that we wish to describe or explain*. If we were interested in the growth of the average international organization, we would not be wise to focus on the United Nations, but if we were concerned about the size distribution of international organizations, from big to small, the United Nations would surely be one of the units on which we ought to concentrate. The United Nations is not a representative organization, but it is an important one. In statistical terms, to investigate the typical international organization, we would examine mean values (of budgets, tasks, memberships, etc.), but to understand the *range* of activity, we would want to examine the variance. A second, equally obvious precept is that *a summary must simplify the information at our disposal*. In quantitative terms, this rule means that we should always use fewer summary statistics than units in the original data, otherwise, we could as easily present all the original data without any summary at all.[6] Our summary should also be sufficiently simple that it can be understood by our audience. No phenomenon can be summarized perfectly, so standards of adequacy must depend on our purposes and on the audience. For example, a scientific paper on wars and alliances might include data involving 10,000 observations. In such a paper, summaries of the data using fifty numbers might be justified; however, even for an expert, fifty separate indicators might be incomprehensible without some further summary. For a lecture on the subject to an undergraduate class, three charts might be superior.

[6] This point is closely related to the concept of indeterminate research designs, which we discuss in section 4.1.

2.6 Descriptive Inference

Descriptive inference is the process of understanding an unobserved phenomenon on the basis of a set of observations. For example, we may be interested in understanding variations in the district vote for the Conservative, Labour, and Social Democratic parties in Britain in 1979. We presumably have some hypotheses to evaluate; however, what we actually observe is 650 district elections to the House of Commons in that year.

Naively, we might think that we were directly observing the electoral strength of the Conservatives by recording their share of the vote by district and their overall share of seats. But a certain degree of randomness or unpredictability is inherent in politics, as in all of social life and all of scientific inquiry.[7] Suppose that in a sudden fit of absentmindedness (or in deference to social science) the British Parliament had agreed to elections every week during 1979 and suppose (counterfactually) that these elections were independent of one another. Even if the underlying support for the Conservatives remained constant, each weekly replication would not produce the same number of votes for each party in each district. The weather might change, epidemics might break out, vacations might be taken—all these occurrences would affect voter turnout and electoral results. Additionally, fortuitous events might happen in the international environment, or scandals might reach the mass media; even if these had no long-term significance, they could affect the weekly results. Thus, numerous, transitory events could effect slightly different sets of election returns. Our observation of any one election would not be a perfect measure of Conservative strength after all.

As another example, suppose we are interested in the degree of conflict between Israelis (police and residents) and Palestinians in communities on the Israeli-occupied West Bank of the Jordan River. Official reports by both sides seem suspect or are censored, so we decide to conduct our own study. Perhaps we can ascertain the general level of conflict in different communities by intensive interviews or participa-

[7] See Popper (1982) for a book-length defense of indeterminism.

tion in family or group events. If we do this for a week in each community, our conclusions about the level of conflict in each one will be a function in part of whatever chance events occur the week we happen to visit. Even if we conduct the study over a year, we still will not perfectly know the true level of conflict, even though our uncertainty about it will drop.

In these examples, the variance in the Conservative vote across districts or the variance in conflict between West Bank communities can be conceptualized as arising from two separate factors: *systematic* and *nonsystematic* differences. Systematic differences in our voter example include fundamental and predictable characteristics of the districts, such as differences in ideology, in income, in campaign organization, or in traditional support for each of the parties. In hypothetical weekly replications of the same elections, systematic differences would persist, but the nonsystematic differences such as turnout variations due to the weather, would vary. In our West Bank example, systematic differences would include the deep cultural differences between Israelis and Palestinians, mutual knowledge of each other, and geographic patterns of residential housing segregation. If we could start our observation week a dozen different times, these systematic differences between communities would continue to affect the observed level of conflict. However, nonsystematic differences, such as terrorist incidents or instances of Israeli police brutality, would not be predictable and would only affect the week in which they happened to occur. With appropriate inferential techniques, we can usually learn about the nature of systematic differences even with the ambiguity that occurs in one set of real data due to nonsystematic, or random, differences.

Thus, *one of the fundamental goals of inference is to distinguish the systematic component from the nonsystematic component of the phenomena we study.* The systematic component is not more important than the nonsystematic component, and our attention should not be focused on one to the exclusion of the other. However, distinguishing between the two is an essential task of social science. One way to think about inference is to regard the data set we compile as only one of many possible data sets—just as the actual 1979 British election returns constitute only one of many possible sets of results for different hypothetical days on which elections could have been held, or just as our one week of observation in one small community is one of many possible weeks.

In descriptive inference, we seek to understand the degree to which our observations reflect either typical phenomena or outliers. Had the 1979 British elections occurred during a flu epidemic that swept through working-class houses but tended to spare the rich, our observations might be rather poor measures of underlying Conservative

strength, precisely because the nonsystematic, chance element in the data would tend to overwhelm or distort the systematic element. If our observation week had occurred immediately after the Israeli invasion of Southern Lebanon, we would similarly not expect results that are indicative of what usually happens on the West Bank.

The political world is theoretically capable of producing multiple data sets for every problem but does not always follow the needs of social scientists. We are usually only fortunate enough to observe one set of data. For purposes of a model, we will let this one set of data be represented by one variable y (say, the vote for Labor) measured over all $n = 650$ units (districts): $y_1, y_2, \ldots, y_n$ (for example, y_5 might be 23,562 people voting for Labor in district 5). The set of observations which we label y is a *realized variable*. Its values vary over the n units. In addition, we define Y as a *random variable* because it varies randomly across hypothetical replications of the same election. Thus, y_5 is the number of people voting for Labor in district 5, and Y_5 is the random variable representing the vote across many hypothetical elections that could have been held in district 5 under essentially the same conditions. The observed votes for the Labor party in the one sample we observe, $y_1, y_2, \ldots, y_n$, differ across constituencies because of systematic and random factors. That is, to distinguish the two forms of "variables," we often use the term *realized variable* to refer to y and *random variable* to refer to Y.

The same arrangement applies to our qualitative example. We would have no hope or desire of quantifying the level of tension between Israelis and Palestinians, in part because "conflict" is a complicated issue that involves the feelings of numerous individuals, organizational oppositions, ideological conflicts, and many other features. In this situation, y_5 is a realized variable which stands for the total conflict observed during our week in the fifth community, say El-Bireh.[8] The random variable Y_5 represents both what we observe in El-Bireh and what we could have observed; the randomness comes from the variation in chance events over the possible weeks we could have chosen to observe.[9]

One goal of inference is to learn about systematic features of the random variables $Y_1, \ldots, Y_n$. (Note the contradictory, but standard, terminology: although in general we wish to distinguish systematic from nonsystematic components in our data, in a specific case we wish to

[8] Obviously the same applies to all the other communities we might study.

[9] Note that the randomness is not exactly over different actual weeks, since both chance events and systematic differences might account for observed differences. We therefore create the more ideal situation in which we imagine running the world again with systematic features held constant and chance factors allowed to vary.

take a random variable and extract its systematic features.) For example, we might wish to know the expected value of the Labor vote in district 5 (the average Labor vote Y_5 across a large number of hypothetical elections in this district). Since this is a systematic feature of the underlying electoral system, the expected value is of considerable interest to social scientists. In contrast, the Labor vote in one observed election, y_5, is of considerably less long-term interest since it is a function of systematic features *and* random error.[11]

The expected value (one feature of the systematic component) in the fifth West Bank community, El-Bireh, is expressed formally as follows:

$$E(Y_5) = \mu_5$$

where $E(\cdot)$ is the expected value operation, producing the average across an infinite number of hypothetical replications of the work we observe in community 5, El-Bireh. The parameter μ_5 (the Greek letter mu with a subscript 5) represents the answer to the expected value calculation (a level of conflict between Palestinians and Israelis) for community 5. This parameter is part of our model for a systematic feature of the random variable Y_5. One might use the observed level of conflict, y_5, as an estimate of μ_5, but because y_5 contains many chance elements along with information about this systematic feature, better estimates usually exist (see section 2.7).

Another systematic feature of these random variables which we might wish to know is the level of conflict in the average West Bank community:

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu \qquad (2.2)$$

One estimator of μ might be the average of the observed levels of conflict across all the communities studied, $\bar{y}$, but other estimators for this systematic feature exist, too. (Note that the same summary of data in our discussion of summarizing historical detail from section 2.5 is used for the purpose of estimating a descriptive inference.) Other systematic features of the random variables include the variance and a variety of causal parameters introduced in section 3.1.

Still another systematic feature of these random variables that might be of interest is the variation in the level of conflict within a community even when the systematic features do not change: the extent to which observations over different weeks (different hypothetical realizations of the same random variable) produce divergent results. This is, in other words, the size of the nonsystematic component. Formally, this is calculated for a single community by using the variance (instead of the expectation):

$$V(Y_i) = \sigma_i^2 \qquad (2.3)$$

where σ^2 (the Greek letter sigma) denotes the result of applying the variance operator to the random variable Y_i. Living in a West Bank community with a high level of conflict between Israelis and Palestinians would not be pleasant, but living in a community with a high variance, and thus unpredictability, might be worse. In any event, both may be of considerable interest for scholarly researchers.

To understand these issues better, we distinguish two fundamental views of random variation.[12] These two perspectives are extremes on a continuum. Although significant numbers of scholars can be found who are comfortable with each extreme, most political scientists have views somewhere between the two.

Perspective 1: A Probabilistic World. Random variation exists in nature and the social and political worlds and can never be eliminated. Even if we measured all variables without error, collected a census (rather than only a sample) of data, and included every conceivable explanatory variable, our analyses would still never generate perfect predictions. A researcher can divide the world into apparently systematic and apparently nonsystematic components and often improve on predictions, but nothing a researcher does to analyze data can have any effect on reducing the *fundamental* amount of nonsystematic variation existing in various parts of the empirical world.

Perspective 2: A Deterministic World. Random variation is only that portion of the world for which we have no explanation. The division between systematic and stochastic variation is *imposed* by the analyst and depends on what explanatory variables are available and included in the analysis. Given the right explanatory variables, the world is entirely predictable.

These differing perspectives produce various ambiguities in the inferences in different fields of inquiry.[13] However, for most purposes

[11] Of course, y_5 may be of tremendous interest to the people in district 5 for that year, and thus both the random and systematic components of this event might be worth studying. Nevertheless, we should always try to distinguish the random from the systematic.

[12] See King (1995b) for an elaboration of this distinction.

[13] Economists tend to be closer to Perspective 1, whereas statisticians are closer to Perspective 2. Perspective 1 is also especially common in the field of engineering called "quality control." Physicists have even debated this distinction in the field of quantum mechanics. Early proponents of Perspective 2 subscribed to the "hidden variable theory"

these two perspectives can be regarded as observationally equivalent. This is especially true if we assume, under Perspective 2, that at least some explanatory variables remain unknown. Thus, observational equivalence occurs when these unknown explanatory variables in Perspective 2 become the interpretation for the random variation in Perspective 1. Because of the lack of any observable implications with which to distinguish between them, a choice between the two perspectives depends on faith or belief rather than on empirical verification.

As another example, with both perspectives, distinguishing whether a particular political or social event is the result of a systematic or nonsystematic process depends upon the choices of the researcher. From the point of view of Perspective 1, we may tentatively classify an effect as systematic or nonsystematic. But unless we can find another set of data (or even just another case) to check for the persistence of an effect or pattern, it is very difficult to make the right judgment.

From the extreme version of Perspective 2, we can do no more than describe the data—"incorrectly" judging an event as stochastic or systematic is impossible or irrelevant. A more realistic version of this perspective admits to Perspective 1's correct or incorrect attribution of a pattern as random or systematic, but it allows us some latitude in deciding what will be subject to examination in any particular study and what will remain unexplained. In this way, we begin any analysis with all observations being the result of "nonsystematic" forces. Our job is then to provide evidence that particular events or processes are the result of systematic forces. Whether an unexplained event or process is a truly random occurrence or just the result of as yet unidentified explanatory variables is left as a subject for future research.

This argument applies with equal force to qualitative and quantitative researchers. Qualitative research is often historical, but it is of most use as social science when it is also explicitly inferential. To conceptualize the random variables from which observations are generated and to attempt to estimate their systematic features—rather than merely summarizing the historical detail—does not require large-scale data collections. Indeed, one mark of a good historian is the ability to distinguish systematic aspects of the situation being described from idiosyncratic ones. This argument for descriptive inference, therefore, is certainly not a criticism of case studies or historical work. Instead,

of quantum mechanics. However, more modern work seems to provide a fundamental verification of Perspective 1: the physical world seems intrinsically probabilistic. We all await the resolution of the numerous remaining contradictions of this important theory and its implications for the nature of the physical world. However, this dispute in physics, although used to justify much of the philosophy of social science, is unlikely to affect the logic of inference or practice of research in the social sciences.

any kind of social science research should satisfy the basic principles of inference discussed in this book. Finding evidence of systematic features will be more difficult with some kinds of evidence, but it is no less important.

As an example of problems of descriptive inference in historical research, suppose that we are interested in the outcomes of U.S.–Soviet summit meetings between 1955 and 1990. Our ultimate purpose is to answer a causal question: under what conditions and to what extent did the summits lead to increased cooperation? Answering that question requires resolving a number of difficult issues of causal analysis, particularly those involving the direction of causality among a set of systematically related variables.[13] In this section, however, we restrict ourselves to problems of descriptive inference.

Let us suppose that we have devised a way of assessing—through historical analysis, surveying experts, counting "cooperative" and "conflictual" events or a combination of these measurement techniques—the extent to which summits were followed by increased superpower cooperation. And we have some hypotheses about the conditions for increased cooperation—conditions that concern shifts in power, electoral cycles in the United States, economic conditions in each country, and the extent to which previous expectations on both sides have been fulfilled. Suppose also that we hope to explain the underlying level of cooperation in each year, and to associate it somehow with the presence or absence of a summit meeting in the previous period, as well as with our other explanatory factors.

What we observe (even if our indices of cooperation are perfect) is only the degree of cooperation *actually* occurring in each year. If we observe high levels of cooperation in years following summit meetings, we do not know (without further study) whether the summits and subsequent cooperation are systematically related to one another. With a small number of observations, it could be that the association between summits and cooperation reflects randomness due to fundamental uncertainty (good or bad luck under Perspective 1) or to as yet unidentified explanatory variables (under Perspective 2). Examples of such unidentified explanatory variables include weather fluctuations leading to crop failures in the Soviet Union, shifts in the military balance, or leadership changes, all of which could account for changes in the extent of cooperation. If identified, these variables are alternative explanations—omitted variables that could be collected or examined

[13] In our language, as we will discuss in section 3.5 below, the issue is that of endogeneity. Anticipated cooperation could lead to the convening of summit meetings, in which case, instead of summit meetings explaining cooperation, anticipated cooperation would explain actual cooperation—hardly a startling finding if actors are rational!

to assess their influence on the summit outcome. If unidentified, these variables may be treated as nonsystematic events that could account for the observed high degree of superpower cooperation. To provide evidence against the possibility that random events (unidentified explanatory variables) account for the observed cooperation, we might look at many other years. Since random events and processes are by definition not persistent, they will be extremely unlikely to produce differential cooperation in years with and without superpower summits. Once again, we are led to the conclusion that only repeated tests in different contexts (years, in this case) enable us to decide whether to define a pattern as systematic or just due to the transient consequences of random processes.

Distinguishing systematic from nonsystematic processes is often difficult. From the perspective of social science, a flu epidemic that strikes working-class voters more heavily than middle-class ones is an unpredictable (nonsystematic) event that in one hypothetical replication of the 1979 election would decrease the Labor vote. But a persistent pattern of class differences in the incidence of a disabling illness would be a systematic effect lowering the average level of Labor voting across many replications.

The victory of one candidate over another in a U.S. election on the basis of the victor's personality or an accidental slip of the tongue during a televised debate might be a random factor that could have affected the likelihood of cooperation between the USSR and the United States during the Cold War. But if the most effective campaign appeal to voters had been the promise of reduced tensions with the USSR, consistent victories of conciliatory candidates would have constituted a systematic factor explaining the likelihood of cooperation.

Systematic factors are persistent and have consistent consequences when the factors take a particular value. Nonsystematic factors are transitory: we cannot predict their impact. But this does not mean that systematic factors represent constants. Campaign appeals may be a systematic factor in explaining voting behavior, but that fact does not mean that campaign appeals themselves do not change. It is the *effect* of campaign appeals on an election outcome that is constant—or, if it is variable, it is changing in a predictable way. When Soviet-American relations were good, promises of conciliatory policies may have won votes in U.S. elections; when relations were bad, the reverse may have been true. Similarly, the weather can be a random factor (if intermittent and unpredictable shocks have unpredictable consequences) or a systematic feature (if bad weather always leads to fewer votes for candidates favoring conciliatory policies).

In short, summarizing historical detail is an important intermediate

step in the process of using our data, but we must also make descriptive inferences distinguishing between random and systematic phenomena. Knowing what happened on a given occasion is not sufficient by itself. *If we make no effort to extract the systematic features of a subject, the lessons of history will be lost, and we will learn nothing about what aspects of our subject are likely to persist or to be relevant to future events or studies.*

2.7 Criteria for Judging Descriptive Inferences

In this final section, we introduce three explicit criteria that are commonly used in statistics for judging methods of making inferences—unbiasedness, efficiency, and consistency. Each relies on the random-variable framework introduced in section 2.6 but has direct and powerful implications for evaluating and improving qualitative research. To clarify these concepts, we provide only the simplest possible examples in this section, all from descriptive inference. A simple version of inference involves estimating parameters, including the expected value or variance of a random variable (μ or σ^2) for a descriptive inference. We also use these same criteria for judging causal inferences in the next chapter (see section 3.4). We save for later chapters specific advice about doing qualitative research that is implied by these criteria and focus on the concepts alone for the remainder of this section.

2.7.1 Unbiased Inferences

If we apply a method of inference again and again, we will get estimates that are sometimes too large and sometimes too small. Across a large number of applications, do we get the right answer on average? If yes, then this method, or "estimator," is said to be unbiased. This property of an estimator says nothing about how far removed from the average any one application of the method might be, but being correct on average is desirable.

Unbiased estimates occur when the variation from one replication of a measure to the next is nonsystematic and moves the estimate sometimes one way, sometimes the other. Bias occurs when there is a systematic error in the measure that shifts the estimate more in one direction than another over a set of replications. If in our study of conflict in West Bank communities, leaders had created conflict in order to influence the study's results (perhaps to further their political goals), then the level of conflict we observe in every community would be biased toward greater conflict, on average. If the replications of our

hypothetical 1979 elections were all done on a Sunday (when they could have been held on any day), there would be a bias in the estimates if that fact systematically helped one side and not the other (if, for instance, Conservatives were more reluctant to vote on Sunday for religious reasons). Or our replicated estimates might be based on reports from corrupt vote counters who favor one party over the other. If, however, the replicated elections were held on various days chosen in a manner unrelated to the variable we are interested in, any error in measurement would not produce biased results even though one day or another might favor one party. For example, if there were miscounts due to random sloppiness on the part of vote counters, the set of estimates would be unbiased.

If the British elections were always held by law on Sundays or if a vote-counting method that favored one party over another were built into the election system (through the use of a particular voting scheme or, perhaps, even persistent corruption), we would want an estimator that varied based on the mean vote that could be expected under the circumstances that included these systematic features. Thus, bias depends on the theory that is being investigated and does not just exist in the data alone. It makes little sense to say that a particular data set is biased, even though it may be filled with many individual errors.

In this example, we might wish to distinguish our definition of "statistical bias" as an *estimator* from "substantive bias" as an *electoral system*. An example of the latter are polling hours that make it harder for working people to vote—a not uncommon substantive bias of various electoral systems. As researchers, we may wish to estimate the mean vote of the actual electoral system (the one with the substantive bias), but we might also wish to estimate the mean of a hypothetical electoral system that doesn't have a substantive bias due to the hours the polls are open. This would enable us to estimate the amount of substantive bias in the system. Whichever mean we are estimating, we wish to have a statistically unbiased estimator.

Social science data are susceptible to one major source of bias of which we should be wary: people who provide the raw information that we use for descriptive inferences often have reasons for providing estimates that are systematically too high or low. Government officials may want to overestimate the effects of a new program in order to shore up their claims for more funding or underestimate the unemployment rate to demonstrate that they are doing a good job. We may need to dig deeply to find estimates that are less biased. A telling example is in Myron Weiner's qualitative study of education and child labor in India (1991). In trying to explain the low level of commitment to compulsory education in India compared to that in other countries, he had to first determine if the level of commitment was indeed low. In one state in India, he found official statistics that indicated that ninety-eight percent of school age children attend school. However, a closer look revealed that attendance was measured once, when children first entered school. They were then listed as attending for seven years, even if their only attendance was for one day! Closer scrutiny showed the actual attendance figure to be much lower.

A Formal Example of Unbiasedness. Suppose, for example, we wish to estimate μ in equation (2.2) and decide to use the average as an estimator, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. In a single set of data, $\bar{y}$ is the proportion of Labor voters averaged over all $n = 650$ constituencies (or the average level of conflict across West Bank communities). But considered across an infinite number of hypothetical replications of the election in each constituency, the sample mean becomes a function of 650 random variables, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Thus, the sample mean becomes a random variable, too. For some hypothetical replications, $\bar{Y}$ will produce election returns that are close to μ and other times they will be farther away. The question is whether $\bar{Y}$ will be right, that is, equal to μ, on average across these hypothetical replications. To determine the answer, we use the expected value operation again, which allows us to determine the average across the infinite number of hypothetical elections. The rules of expectations enable us to make the following calculations:

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \tag{2.4}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(Y_i)$$

$$= \frac{1}{n} n\mu$$

$$= \mu$$

Thus, $\bar{Y}$ is an unbiased estimator of μ. (This is a slightly less formal example than appears in formal statistics texts, but the key features are the same.)

2.7.2 Efficiency

We usually do not have an opportunity to apply our estimator to a large number of essentially identical applications. Indeed, except for some clever experiments, we only apply it once. In this case, unbiasedness is of interest, but we would like more confidence that the one estimate we get is close to the right one. Efficiency provides a way of distinguishing among unbiased estimators. Indeed, the efficiency criterion can also help distinguish among alternative estimators with a small amount of bias. (An estimator with a large bias should generally be ruled out even without evaluating its efficiency.)

Efficiency is a relative concept that is measured by calculating the variance of the estimator across hypothetical replications. For unbiased estimators, the smaller the variance, the more efficient (the better) the estimator. A small variance is better because our one estimate will probably be closer to the true parameter value. We are not interested in efficiency for an estimator with a large bias because low variance in this situation will make it unlikely that the estimate will be near the true value (because most of the estimates would be closely clustered around the wrong value). As we describe below, we are interested in efficiency in the case of a small amount of bias, and we may often be willing to incur a small amount of bias in exchange for a large gain in efficiency.

Suppose again we are interested in estimating the average level of conflict between Palestinians and Israelis in the West Bank and are evaluating two methods: a single observation of one community, chosen to be typical, and similar observations of, for example, twenty-five communities. It should be obvious that twenty-five observations are better than a single observation—so long as the same effort goes into collecting each of the twenty-five as into the single observation. We will demonstrate here precisely why this is the case. This result explains why we should observe as many implications of our theory as possible, but it also demonstrates the more general concept of statistical efficiency, which is also relevant whenever we are deciding the best way to evaluate different ways of combining gathered observations into an inference.

Efficiency enables us to compare the single-observation case study ($n = 1$) estimator of μ with the large-n estimator ($n = 25$), that is the average level of conflict found from twenty-five separate week-long studies in different communities on the West Bank. If applied appropriately, both estimators are unbiased. If the same model applies, the single-observation estimator has a variance of $V(Y_{typical}) = \sigma^2$. That is, we would have chosen what we thought was a "typical" district, which would, however, be affected by random variables. The variance of the large-n estimator is $V(\bar{Y}) = \sigma^2/25$, that is, the variance of the sample mean. Thus, the single-observation estimator is twenty-five times more variable (i.e., less efficient) than the estimate when $n = 25$. Hence, we have the obvious result that more observations are better.

More interesting are the conditions under which a more detailed study of our one community would yield as good or better results as our large-n study. That is, although we should always prefer studies with more observations (given the resources necessary to collect them), there are situations where a single case study (as always, containing many observations) is better than a study based on more observations, each one of which is not as detailed or certain.

All conditions being equal, our analysis shows that the more observations, the better, because variability (and thus inefficiency) drops. In fact, the property of *consistency* is such that as the number of observations gets very large, the variability decreases to zero, and the estimate equals the parameter we are trying to estimate.[14]

But often, not all conditions are equal. Suppose, for example, that any single measurement of the phenomenon we are studying is subject to factors that make the measure likely to be far from the true value (i.e., the estimator has high variance). And suppose that we have some understanding—from other studies, perhaps—of what these factors might be. Suppose further that our ability to observe and correct for these factors decreases substantially with the increase in the number of communities studied (if, for no other reason, than that we lack the time and knowledge to make corrections for such factors across a large number of observations). We are then faced with a trade-off between a case study that has additional observations internal to the case and twenty-five cases in which each contains only one observation.

If our single case study is composed of only one observation, then it is obviously inferior to our 25-observation study. But case-study researchers have significant advantages, which are easier to understand if formalized. For example, we could first select our community very carefully in order to make sure that it is especially representative of the rest of the country or that we understand the relationship of this community to the others. We might ask a few residents or look at newspaper reports to see whether it was an average community or whether

[14] Note that an estimator can be unbiased but inconsistent. For example, Y_1 is an unbiased estimator of μ, but it is inconsistent because as the number of units increase, this estimator does not improve (or indeed change at all). An estimator can also be consistent but biased. For example, $\hat{\sigma}^2 = S/n$ is biased, but it is consistent because S/n becomes zero as n approaches infinity.

same nonsystematic factor had caused the observation to be atypical, and then we might adjust the observed level of conflict to arrive at an estimate of the average level of West Bank conflict, μ. This would be the most difficult part of the case-study estimator, and we would need to be very careful that bias does not creep in. Once we are reasonably confident that bias is minimized, we could focus on increasing efficiency. To do this, we might spend many weeks in the community conducting numerous separate studies. We could interview community leaders, ordinary citizens, and school teachers. We could talk to children, read the newspapers, follow a family in the course of its everyday life, and use numerous other information-gathering techniques. Following these procedures, we could collect far more than twenty-five observations within this one community and generate a case study that is also not biased and more efficient than the twenty-five community study.

Consider another example. Suppose we are conducting a study of the international drug problem and want a measure of the percentage of agricultural land on which cocaine is being grown in a given region of the world. Suppose further that there is a choice of two methods: a case study of one country or a large-scale, statistical study of all the countries of the region. It would seem better to study the whole region. But let us say that to carry out such a study it is necessary (for practical reasons) to use data supplied to a UN agency from the region's governments. These numbers are known to have little relationship to actual patterns of cropping since they were prepared in the Foreign Office and based on considerations of public relations. Suppose, further, that we could, by visiting and closely observing one country, make the corrections to the government estimates that would bring that particular estimate much closer to a true figure. Which method would we choose? Perhaps we would decide to study only one country, or perhaps two or three. Or we might study one country intensively and use our results to reinterpret, and thereby improve, the government-supplied data from the other countries. Our choice should be guided by which data best answer our questions.

To take still another example, suppose we are studying the European Community and want to estimate the expected degree of regulation of an industry throughout the entire Community that will result from actions of the Commission and the Council of Ministers. We could gather data on a large number of rules formally adopted for the industrial sector in question, code these rules in terms of their stringency, and then estimate the average stringency of a rule. If we gather data on 100 rules with similar a priori stringency, the variance of our measure will be the variance of any given rule divided by 100 ($\sigma^2/100$), or less if the rules are related. Undoubtedly, this will be a better measure than using data on one rule as the estimator for regulatory stringency for the industry as a whole.

However, this procedure requires us to accept the formal rule as equivalent to the real regulatory activity in the sector under scrutiny. Further investigation of rule application, however, might reveal a large variation in the extent to which nominal rules are actually enforced. Hence, measures of formal rules might be systematically biased—for instance, in favor of overstating regulatory stringency. In such a case, we would face the bias-efficiency trade-off once again, and it might make sense to carry out three or four intensive case studies of rule implementation to investigate the relationship between formal rules and actual regulatory activity. One possibility would be to substitute as estimator based on these three or four cases—less biased and also less efficient—for the estimator based on 100 cases. However, it might be more creative, if feasible, to use the intensive case-study work for the three or four cases to correct the bias of our 100-case indicator, and then to use a corrected version of the 100-case indicator as our estimator. In this procedure, we would be combining the insights of our intensive case studies with large-n techniques, a practice that we think should be followed much more frequently than is the case in contemporary social science.

The argument for case studies made by those who know a particular part of the world well is often just the one implicit in the previous example. Large-scale studies may depend upon numbers that are not well understood by the naive researcher working on a data base (who may be unaware of the way in which election statistics are gathered in a particular locale and assumes, incorrectly, that they have some real relationship to the votes as cast). The researcher working closely with the materials and understanding their origin may be able to make the necessary corrections. In subsequent sections we will try to explicate how such choices might be made more systematically.

Our formal analysis of this problem in the box below shows precisely how to decide what the results of the trade-off are in the example of British electoral constituencies. The decision in any particular example will always be better when using logic like that shown in the formal analysis below. However, deciding this issue will almost always also require qualitative judgements, too.

Finally, it is worth thinking more specifically about the trade-offs that sometimes exist between bias and efficiency. The sample mean of the first two observations in any larger set of unbiased observations is

Formal Efficiency Comparisons. The variance of the sample mean $\bar{Y}$ is denoted as $V(\bar{Y})$, and the rules for calculating variances of random variables in the simple case of random sampling permit the following:

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

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i)$$

Furthermore, if we assume that the variance across hypothetical replication of each district election is the same as every other district and is denoted by σ^2, then the variance of the sample mean is

$$V(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) \tag{2.5}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= \frac{1}{n^2} n\sigma^2$$

$$= \sigma^2/n$$

In the example above, $n = 650$, so the large-n estimator has variance $\sigma^2/650$ and the case-study estimator has variance σ^2. Unless we can use qualitative, random-error corrections to reduce the variance of the case-study estimator by a factor of at least 650, the statistical estimate is to be preferred on the grounds of efficiency.

also unbiased, just as is the sample mean of all the observations. However, using only two observations discards substantial information; this does not change unbiasedness, but it does substantially reduce efficiency. If we did not also use the efficiency criterion, we would have no formal criteria for choosing one estimator over the other.

Suppose we are interested in whether the Democrats would win the next presidential election, and we ask twenty randomly selected American adults which party they plan to vote for. (In our simple version of random selection, we choose survey respondents from all adult Americans, each of which has an equal probability of selection.) Suppose that someone else also did a similar study with 1,000 citizens. Should we include these additional observations with ours to create a single estimate based on 1,020 respondents? If the new observations were randomly selected, just as the first twenty, it should be an easy decision to include the additional data with ours: with the new observations, the estimator is still unbiased and now much more efficient.

However, suppose that only 990 of the 1,000 new observations were randomly drawn from the U.S. population and the other ten were Democratic members of Congress who were inadvertently included in the data after the random sample had been drawn. Suppose further that we found out that these additional observations were included in our data but did not know which ones they were and thus could not remove them. We now know a priori that an estimator based on all 1,020 respondents would produce a slight overestimate of the likelihood that a Democrat would win the nationwide vote. Thus, including these 1,000 additional observations would slightly bias the overall estimate, but it would also substantially improve its efficiency. Whether we should include the observations therefore depends on whether the increase in bias is outweighed by the increase in statistical efficiency. Intuitively, it seems clear that the estimator based on the 1,020 observations will produce estimates fairly close to the right answer much more frequently than the estimator based on only twenty observations. The bias introduced would be small enough, so we would prefer the larger sample estimator even though in practice we would probably apply both. (In addition, we know the direction of the bias in this case and could even partially correct for it.)

If adequate quantitative data are available and we are able to formalize such problems as these, we can usually make a clear decision. However, even if the qualitative nature of the research makes evaluating this trade-off difficult or impossible, understanding it should help us make more reliable inferences.

Formal Comparisons of Bias and Efficiency. Consider two estimators, one a large-n study by someone with a preconception, who is therefore slightly biased, and the other a very small-n study that we believe is unbiased but relatively less efficient and is done by an impartial investigator. As a formal model of this example, suppose we wish to estimate μ and the large-n study produces estimator $\hat{\mu}$.

$$d = \left(\frac{1}{N}\sum_{i=1}^{N} Y_i - 0.01\right)$$

We model the small-n study with a different estimator of μ, c:

$$c = \left(\frac{Y_1 + Y_2}{2}\right)$$

where districts 1 and 2 are average constituencies, so that $E(Y_1) = \mu$ and $E(Y_2) = \mu$.

Which estimator should we prefer? Our first answer is that we would use neither and instead would prefer the sample mean $\bar{y}$; that is, a large-n study by an impartial investigator. However, the obvious or best estimator is not always applicable. To answer this question, we turn to an evaluation of bias and efficiency.

First, we will assess bias. We can show that the first estimator d is slightly biased according to the usual calculation:

$$E(d) = E\left(\frac{1}{N}\sum_{i=1}^{N} Y_i - 0.01\right)$$

$$= E\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) - E(0.01)$$

$$= \mu - 0.01$$

We can also show that the second estimator c is unbiased by a similar calculation:

$$E(c) = E\left(\frac{Y_1 + Y_2}{2}\right)$$

$$= \frac{E(Y_1) + E(Y_2)}{2}$$

$$= \frac{\mu + \mu}{2}$$

$$= \mu$$

By these calculations alone, we would choose estimator c, the result of the efforts of our impartial investigator's small-n study, since it is unbiased. On average, across an infinite number of hypothetical replications, for the investigator with a preconception, d would give the wrong answer, albeit only slightly so. Estimator c would give the right answer on average.

The efficiency criterion tells a different story. To begin, we calculate the variance of each estimator:

$$V(d) = V\left(\frac{1}{N}\sum_{i=1}^{N} Y_i - 0.01\right)$$

$$= V\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) - V(0.01)$$

$$= \sigma^2/n$$

$$= \sigma^2/650$$

This variance is the same as the variance of the sample mean because 0.01 does not change (has zero variance) across samples. Similarly, we calculate the variance of c as follows:[15]

$$V(c) = V\left(\frac{Y_1 + Y_2}{2}\right)$$

$$= \frac{1}{4}[V(Y_1) + V(Y_2)]$$

$$= \frac{1}{4}\,2\sigma^2$$

$$= \sigma^2/2$$

Thus, c is considerably less efficient than d because $V(c) = \sigma^2/2$ is 325 times larger than $V(d) = \sigma^2/650$. This should be intuitively clear as well, since c discards most of the information in the data set.

Which should we choose? Estimator d is biased but more efficient

[15] We assume the absence of spatial correlation across districts in the second line of the preceding and following calculations.

than c, whereas c is unbiased but less efficient. In this particular case, we would probably prefer estimator d. We would thus be willing to sacrifice unbiasedness, since the sacrifice is fairly small (0.01), in order to obtain a significantly more efficient estimator. At some point, however, more efficiency will not compensate for a little bias since we end up guaranteeing that estimates will be farther from the truth. The formal way to evaluate the bias-efficiency trade-off is to calculate the *mean square error* (MSE), which is a combination of bias and efficiency. If g is an estimator for some parameter γ (the Greek letter Gamma), MSE is defined as follows:

$$\text{MSE}(g) = V(g) + E(g - \gamma)^2 \tag{2.6}$$

$$= \text{variance} + \text{Squared bias}$$

Mean square error is thus the sum of the variance and the squared bias (see Johnston 1984:27–28). The idea is to choose the estimator with the minimum mean square error since it shows precisely how an estimator with some bias can be preferred if it has a smaller variance.

For our example, the two MSEs are as follows:

$$\text{MSE}(d) = \frac{\sigma^2}{650} + (0.01)^2 \tag{2.7}$$

$$= \frac{\sigma^2}{650} + 0.0001$$

and

$$\text{MSE}(c) = \frac{\sigma^2}{2} \tag{2.8}$$

Thus, for most values of σ^2, $\text{MSE}(d) < \text{MSE}(c)$ and we would prefer d as an estimator to c.

In theory, we should always prefer unbiased estimates that are as efficient (i.e., use as much information) as possible. However, in the real research situations we analyze in succeeding chapters, this trade-off between bias and efficiency is quite salient.

CHAPTER 3

Causality and Causal Inference

We have discussed two stages of social science research: summarizing historical detail (section 2.5) and making descriptive inferences by partitioning the world into systematic and nonsystematic components (section 2.6). Many students of social and political phenomena would stop at this point, eschewing causal statements and asking their selected and well-ordered facts to "speak for themselves."

Like historians, social scientists need to summarize historical detail and to make descriptive inferences. For some social scientific purposes, however, analysis is incomplete without causal inference. That is, just as causal inference is impossible without good descriptive inference, descriptive inference alone is often unsatisfying and incomplete. To say this, however, is not to claim that all social scientists must, in all of their work, seek to devise causal explanations of the phenomena they study. Sometimes causal inference is too difficult; in many other situations, descriptive inference is the ultimate goal of the research endeavor.

Of course, we should always be explicit in clarifying whether the goal of a research project is description or explanation. Many social scientists are uncomfortable with causal inference. They are so wary of the warning that "correlation is not causation" that they will not state causal hypotheses or draw causal inferences, referring to their research as "studying association and not causation." Others make apparent causal statements with ease, labeling unevaluated hypotheses or speculations as "explanations" on the basis of indeterminate research designs.[1] We believe that each of these positions evades the problem of causal inference.

[1] In view of some social scientists' preference for explanation over "mere description," it is not surprising that students of complicated events seek to dress their work in the trappings of explanatory jargon; otherwise, they fear being regarded as doing inferior work. At its core, *real* explanation is always based on causal inferences. We regard arguments in the literature about "nonexistent explanation" as confusing terminology; in virtually all cases, these arguments are really about causal explanation or are internally inconsistent. If social scientists' failures to explain are not due to poor research or lack of imagination, but rather to the nature of the difficult but significant problems that they are examining, such feelings of inferiority are unjustified. Good description of important events is better than bad explanation of anything.

Avoiding causal language when causality is the real subject of investigation either renders the research irrelevant or permits it to remain undisciplined by the rules of scientific inference. Our uncertainty about causal inferences will never be eliminated. But this uncertainty should not suggest that we avoid attempts at causal inference. Rather we should draw causal inferences where they seem appropriate but also provide the reader with the best and most honest estimate of the uncertainty of that inference. It is appropriate to be bold in drawing causal inferences as long as we are cautious in detailing the uncertainty of the inference. It is important, further, that causal hypotheses be disciplined, approximating as closely as possible the rules of causal inference. Our purpose in much of chapters 4–6 is to explicate the circumstances under which causal inference is appropriate and to make it possible for qualitative researchers to increase the probability that their research will provide reliable evidence about their causal hypotheses.

In section 3.1 we provide a rigorous definition of causality appropriate for qualitative and quantitative research, then in section 3.2 we clarify several alternative notions of causality in the literature and demonstrate that they do not conflict with our more fundamental definition. In section 3.3 we discuss the precise assumptions about the world and the hypotheses required to make reliable causal inferences. We then consider in section 3.4 how to apply to causal inference the criteria we developed for judging descriptive inference. In section 3.5 we conclude this chapter with more general advice on how to construct causal explanations, theories, and hypotheses.

3.1 Defining Causality

In this section, we define causality as a *theoretical* concept independent of the data used to learn about it. Subsequently, we consider causal *inference* from our data. (For discussions of specific problems of causal inference, see chapters 4–6.) In section 3.1.1 we give our definition of causality in full detail, along with a simple quantitative example, and in section 3.1.2 we revisit our definition along with a more sophisticated qualitative example.

3.1.1 The Definition and a Quantitative Example

Our theoretical definition of causality applies most simply and clearly to a single unit.[2] As defined in section 2.4, a unit is one of the many elements to be observed in a study, such as a person, country, year, or political organization. For precision and clarity, we have chosen a single running example from quantitative research: the causal effect of incumbency status for a Democratic candidate for the U.S. House of Representatives on the proportion of votes this candidate receives. (Using only a Democratic candidate simplifies the example.) Let the dependent variable be the Democratic proportion of the two-party vote for the House. The key causal explanatory variable is then dichotomous, either the Democrat is an incumbent or not. (For simplicity throughout this section, we only consider districts where the Republican candidate lost the last election.)

Causal language can be confusing and our choice here is hardly unique. The "dependent variable" is sometimes called the "outcome variable." "Explanatory variables" are often referred to as "independent variables." We divide the explanatory variables into the "key causal variable" (also called the "cause" or the "treatment variable") and the "control variables." Finally, the key causal variable always takes on two or more values, which are often denoted by "treatment group" and "control group."

Now consider only the Fourth Congressional District in New York, and imagine an election in 1998 with a Democratic incumbent and one Republican (nonincumbent) challenger. Suppose the Democratic candidate received y_4^I fraction of the vote in this election (the subscript 4 denotes the Fourth District in New York and the superscript *I* refers to the fact that the Democrat is an *I*ncumbent). y_4^I is then a value of the dependent variable. To *define* the causal effect (a theoretical quantity), imagine that we go back in time to the start of the election campaign and everything remains the same, except that the Democratic incumbent decides not to run for re-election and the Democratic Party nominates another candidate (presumably the winner of the primary election). We denote the fraction of the vote that the Democratic (nonincumbent) candidate would receive by y_4^N (where *N* denotes a Democratic candidate who is a *N*on-incumbent).[3]

This *counterfactual* condition is the essence behind this definition of causality, and the difference between the actual vote (y_4^I) and the likely

[2] Our point of departure in this section is Holland's article (1986) on causality and what he calls "Rubin's Model." Holland bases his ideas on the work of numerous scholars. Donald Rubin's (1974, 1978) work on the subject was most immediately relevant, but he also cites Aristotle, Locke, Hume, Mill, Suppes, Granger, Fisher, Neyman, and others. We extend Holland's definition of a causal effect by using some ideas expressed clearly by Suppes (1970) and others concerning "probabilistic causality." We found this extension necessary since no existing approach alone is capable of defining causality with respect to a single unit and still allowing one to partition causal effects into systematic and nonsystematic components.

[3] See Coleman and King (1990) for details of this example. More generally, *I* and *N* can stand for the "treatment" and "control" group (or for any two treatments experimentally

vote in this counterfactual situation (y_i^0) is the causal effect, a concept we define more precisely below. We must be very careful in defining counterfactuals, although they are obviously counter to the facts, they must be reasonable and it should be possible for the counterfactual event to have occured under precisely stated circumstances. A key part of defining the appropriate counterfactual condition is clarifying precisely what we are holding constant while we are changing the value of the treatment variable. In the present example, the key causal (or treatment) variable is incumbency status, and it changes from "incumbent" to "non-incumbent." During this hypothetical change, we hold everything constant up to the moment of the Democratic Party's nomination decision—the relative strength of the Democrats and Republicans in past elections in this district, the nature of the nomination process, the characteristics of the congressional district, and the economic and political climate at the time, etc. We do not control for qualities of the candidates, such as name recognition, visibility, and knowledge of the workings of Congress, or anything else that follows the party nomination. The reason is that these are partly consequences of our treatment variable, incumbency. That is, the advantages of incumbency include name recognition, visibility, and so forth. If we did hold these constant, we would be controlling for and hence disregarding some of the most important effects of incumbency and as a result, would misinterpret its overall effect on the vote total. In fact, controlling for enough of the consequences of incumbency could make one incorrectly believe that incumbency had no effect at all.[6]

More formally, the causal effect of incumbency in the Fourth District in New York—the proportion of the vote received by the Democratic Party candidate that is attributable to incumbency status—would be the difference between these two vote fractions: $(y_i^1 - y_i^0)$. For reasons that will become clear shortly, we refer to this difference as the *realized*

administered in fact or in theory. Of course, the decision to call one value of an explanatory variable a treatment and the other a control is entirely arbitrary, if this language is used at all.

[6] Jon Elster (1983:34–36) has claimed "the meaning of causality can not be rendered by counterfactual statements" in many situations, such as those in which a third factor accounts for both the apparent explanatory and dependent variables. In our language, Elster is simply pointing to common problems of inference, which are always uncertain to some extent. However, these difficulties of inference do not invalidate a definition of causality in terms of counterfactuals. Despite his objections, Elster acknowledges that counterfactual statements "have an important role in causal analysis" (Elster 1983:36). Hence Elster's argument is more cogent, we think, as a set of valuable warnings against careless use of counterfactuals than as a critique of their fundamental definitional importance in causal reasoning.

causal effect and write it in more general notation for unit i instead of only district 4.[7]

$$\text{(Realized Causal Effect for unit } i) = y_i^1 - y_i^0 \tag{3.1}$$

Of course, this effect is defined only in theory since in any one real election we might observe either y_i^1 or y_i^0 or neither, but never both. Thus, this simple definition of causality demonstrates that we can never hope to know a causal effect for certain. Holland (1986) refers to this problem as *the fundamental problem of causal inference*, and it is indeed a fundamental problem since no matter how perfect the research design, no matter how much data we collect, no matter how perceptive the observers, no matter how diligent the research assistants, and no matter how much experimental control we have, we will never know a causal inference for certain. Indeed, most of the empirical issues of research designs that we discuss in this book involve this fundamental problem, and most of our suggestions constitute partial attempts to avoid it.

Our working definition of causality differs from Holland's, since in section 2.6 we have argued that social science always needs to partition the world into systematic and nonsystematic components, and Holland's definition does not make this distinction clearly.[8] To see the importance of this partitioning, think about what would happen if we could rerun the 1998 election campaign in the Fourth District in New York, with a Democratic incumbent and a Republican challenger. A slightly different total vote would result, due to nonsystematic features of election campaigns—aspects of politics that do not persist from one campaign to the next, even if the campaigns begin on identical footing. Some of these nonsystematic features might include a verbal gaffe, a surprisingly popular speech or position on an issue, an unexpectedly bad performance in a debate, bad weather during one candidate's rally or on election day, or the results of some investigative journalism. We can therefore imagine a variable that would express the values of the Democratic vote across hypothetical replications of this same election.

[7] We can specialize for district 4 by substituting "4" for "i" in the following equation.

[8] The reason for this is probably that Holland is a statistician who comes very close to an extreme version of "Perspective 2" random variation, which is described in section 2.6. In his description of the "statistical solution" to the problem of causal inference, he most closely approximates our definition of a causal effect, but this definition is mostly about using different units to solve the Fundamental Problem instead of retaining the definition of causality in just one. In particular, his expected value operator averages over units, whereas ours (described below) averages over hypothetical replications of the same experiment for just a single unit (see Holland 1986:947).

As noted above (see section 2.6), this variable is called a "random variable" since it has nonsystematic features: it is affected by explanatory variables not encompassed in our theoretical analysis or contains fundamentally unexplainable variability.[7] We define the random variable representing the proportion of votes received by the incumbent Democratic candidate as Y_4^I (note the capital Y) and the proportion of votes that would be received in hypothetical replications by a Democratic nonincumbent as Y_4^N.

We now define the *random causal effect* for district 4 as the difference between these two random variables. Since we wish to retain some generality, we again switch notation from district 4 to unit i:

$$\text{Random Causal Effect for unit } i = (Y_i^I - Y_i^N) \tag{3.2}$$

Just as in the definition of a random variable, a random causal effect is a causal effect that varies over hypothetical replications of the same experiment but also represents many interesting systematic features of elections.) If we could observe two separate vote proportions in district 4 at the same time, one from an election with and one without a Democratic incumbent running, then we could directly observe the realized causal effect in equation (3.1). Of course, because of the Fundamental Problem of Causal Inference, we cannot observe the realized causal effect. Thus, the realized causal effect in equation 3.1 is a single *unobserved* realization of the random causal effect in equation 3.2. In other words, across many hypothetical replications of the same election in district 4 with a Democratic incumbent, and across many hypothetical replications of the same election but with a Democratic nonincumbent, the (unobserved) realized causal effect becomes a random causal effect.

Describing causality as one of the systematic features of random variables may seem unduly complicated. But it has two virtues. First, it makes our definition of causality directly analogous to those systematic features (such as a mean or variance) of a phenomenon that serve

[7] As we explained in more detail in section 2.2, this phrasing can be confusing. A "random variable" contains some systematic component and thus is not always entirely unpredictable. Unfortunately, this language has a specific meaning in statistics and the concepts underlying it are important. The original reason for the terminology is that randomness does not mean "anything goes" or "anything could happen." Instead, it refers to *one of many possible very well-specified probabilistic processes*. For example, the random process governing which side of a coin lands upward when flipped in the air is a very different random process than the one governing the growth of the European Economic Community's bureaucracy or the uncertain political consequence of a change in Italy's electoral system. The key to our representation is that each of these "random" processes have systematic and probabilistic components.

as objects of descriptive inference: means and variances are also systematic features of random variables (as in section 2.2). Secondly, it enables us to partition a causal inference problem into systematic and nonsystematic components. Although many systematic features of a random variable might be of interest, the most relevant for our running example is the *mean causal effect* for unit i. To explain what we mean by this, we return to our New York election example.

Recall that the random variable refers to the vote fraction received by the Democrat (incumbent or nonincumbent) across a large number of hypothetical replications of the same election. We define the expected value of this random variable—the vote fraction averaged across these replications—for the nonincumbent as

$$E(Y_4^N) = \mu_4^N$$

and for the incumbent as

$$E(Y_4^I) = \mu_4^I$$

Then, the mean causal effect of incumbency in unit i is a systematic feature of the random causal effect and is defined as the difference between these two expected values (again generalized to unit i instead of to district 4):

$$\begin{aligned}
\text{Mean Causal Effect for unit } i = \beta_i \qquad & (3.3)\\
&= E(\text{Random Causal Effect for unit } i)\\
&= E(Y_i^I - Y_i^N)\\
&= E(Y_i^I) - E(Y_i^N)\\
&= \mu_i^I - \mu_i^N
\end{aligned}$$

where in the first line of this equation, β (beta) refers to this mean causal effect. In the second line, we indicate that the mean causal effect for unit i is just the mean (expected value) of the random causal effect, and in the third and fourth lines we show how to calculate the mean. The last line is another way of writing the difference in the means of the two sets of hypothetical elections. (The average of the difference between two random variables equals the difference of the averages.) To summarize in words: *the causal effect is the difference between the systematic component of observations made when the explanatory variable takes*

one value and the systematic component of comparable observations when the explanatory variable takes on another value.

The last line of equation 3.3 is similar to equation 3.1, and as such, the Fundamental Problem of Causal Inference still exists in this formulation. Indeed, the problem expressed this way is even more formidable because even if we could get around the Fundamental Problem for a realized causal effect, we would still have all the usual problems of inference, including the problem of separating out systematic and nonsystematic components of the random causal effect. From here on, we use Holland's phrase, the Fundamental Problem of Causal Inference, to refer to the problem that he identified as well as to these standard problems of inference, which we have added to his formulation. In the box on page 95, we provide a more general notation for causal effects, which will prove useful throughout the rest of this book.

Many other systematic features of these random causal effects might be of interest in various circumstances. For example, we might wish to know the variance in the possible (realized) causal effects of incumbency status on Democratic vote in unit i, just as with the variance in the vote itself that we described in equation 2.3 in section 2.6. To calculate the variance of the causal effect, we apply the variance operation

$$\text{variance of the causal effect in unit } i) = V(Y_i^1 - Y_i^0)$$

in which we avoid introducing a new symbol for the result of the variance calculation, $V(Y_i^1 - Y_i^0)$. Certainly new incumbents would wish to know the variation in the causal effect of incumbency so they can judge how closely their experience will be to that of previous incumbents and how much to rely on their estimated mean causal effect of incumbency from previous elections. It is especially important to understand that this variance in the causal effect is a fundamental part of the world and is not uncertainty due to estimation.

3.1.2 A Qualitative Example

We developed our precise definition of causality in section 3.1. Since some of the concepts in that section are subtle and quite sophisticated, we illustrated our points with a very simple running example from quantitative research. This example helped us communicate the concepts we wished to stress without also having to attend to the contextual detail and cultural sensitivity that characterize good qualitative research. In this section, we proceed through our definition of causality again, but this time via a qualitative example.

Political scientists would learn a lot if they could rerun history with everything constant save for one investigator-controlled explanatory variable. For example, one of the major questions that faces those involved with politics and government has to do with the consequences of a particular law or regulation. Congress passes a tax bill that is intended to have a particular consequence—lead to particular investments, increase revenue by a certain amount, and change consumption patterns. Does it have this effect? We can observe what happens after the tax is passed to see if the intended consequences appear; but even if they *do*, it is never certain that they result from the law. The change in investment policy might have happened anyway. If we could rerun history with and without the new regulation, then we would have much more leverage in estimating the causal effect of this law. Of course, we cannot do this. But the logic will help us design research to give us an approximate answer to our question.

Consider now the following extended example from comparative politics. In the wake of the collapse of the Soviet system, numerous governments in the ex-Soviet republics and in Eastern Europe have instituted new governmental forms. They are engaged—as they themselves realize—in a great political experiment: they are introducing new constitutions, constitutions that they hope will have the intended effect of creating stable democratic systems. One of the constitutional choices is between parliamentary and presidential forms of government. Which system is more likely to lead to a stable democracy is the subject of considerable debate among scholars in the field (Linz 1993; Horowitz 1993; Lijphart 1993). The debate is complex, not the least because of the numerous types of parliamentary and presidential systems and the variety of the other constitutional provisions that might accompany and interact with this choice (such as the nature of the electoral system). It is not our purpose to provide a thorough analysis of these choices but rather a greatly simplified version of the choice in order to define a causal effect in the context of this qualitative example. In so doing, we highlight the distinction between systematic and nonsystematic features of a causal effect.

The debate about presidential versus parliamentary systems involves varied features of the two systems. We will focus on two: the extent to which each system represents the varied interests of the citizenry and encourages strong and decisive leadership. The argument is that parliamentary systems do a better job of representing the full range of societal groups and interests in the government since there are many legislative seats to be filled, and they can be filled by representatives elected from various groups. In contrast, the all-or-nothing character of presidential systems means that some groups will feel left out of the government, be disaffected, and cause greater instability. On the other hand, parliamentary systems—especially if they adequately represent the full range of social groups and interests—are likely to be

deadlocked and ineffective in providing decisive government. These characteristics, too, can lead to disaffection and instability.[5]

The key purpose of this section is to formulate a precise definition of a causal effect. To do so, imagine that we could institute a parliamentary system and, periodically over the next decade or so, measure the degree of democratic stability (perhaps by actual survival or demise of democracy, attempted coups, or other indicators of instability), and in the same country and at the same time, institute a presidential system, also measuring its stability over the same period with the same measures. The *realized causal effect* would be the difference between the degree of stability observed under a presidential system and that under a parliamentary system. The impossibility of measuring this causal effect directly is another example of the fundamental problem of causal inference.

As part of this definition, we also need to distinguish between systematic and nonsystematic effects of the form of government. To do this, we imagine running this hypothetical experiment many times. We define the *mean causal effect* to be the average of the realized causal effects across replications of these experiments. Taking the average in this way causes the nonsystematic features of this problem to cancel out and leaves the mean causal effect to include only systematic features. Systematic features include indecisiveness in a parliamentary system or disaffection among minorities in a presidential one. Nonsystematic features might include the sudden illness of a president that throws the government into chaos. The latter event would not be a persistent feature of a presidential system; it would appear in one trial of the experiment but not in others.[6]

Another interesting feature of this example is the variance of the causal effect. Any country thinking of choosing one of these political systems would be interested in its mean causal effect on democratic stability; however, this one country gets only one chance—only one replication of this experiment. Given this situation, political leaders may be interested in more than the average causal effect. They may wish to understand what the maximum and minimum causal effects, or at least the *variance* of the causal effects, might be. For example, it may be that presidentialism reduces democratic stability on average

[5] These distinctions are themselves debated. Some argue that a presidential system can do a better representational job. And others argue that parliamentary systems can be more decisive.

[6] The distinction between a systematic and nonsystematic feature is by no means always clear-cut. The sudden illness of a president appears to be a nonsystematic feature of the presidential system. On the other hand, the general vulnerability of presidential systems to the vagaries of the health and personality of a single individual is a systematic effect that raises the likelihood that some nonsystematic feature will appear.

but that the variability of this effect is enormous—sometimes increasing stability a lot, sometimes decreasing it substantially. This variance translates into risk for a polity. In this circumstance, it may be that citizens and political leaders would prefer to choose an option that produces only slightly less stability on average but has a lower variance in causal effect and thus minimizes the chance of a disastrous outcome.

3.2 Clarifying Alternative Definitions of Causality

In section 3.1, we defined causality in terms of a causal effect: the mean causal effect is the difference between the systematic component of a dependent variable when the causal variable takes on two different values. In this section, we use our definition of causality to clarify several alternative proposals and apparently complicating ideas. We show that the important points made by other authors about "causal mechanisms" (section 3.2.1), "multiple" causality (section 3.2.2), and "symmetric" versus "asymmetric" causality (section 3.2.3) do not conflict with our more basic definition of causality.

3.2.1 "Causal Mechanisms"

Some scholars argue that the central idea of causality is that of a set of "causal mechanisms" posited to exist between cause and effect (see Little 1991:15). This view makes intuitive sense: any coherent account of causality needs to specify how the effects are exerted. For example, suppose a researcher is interested in the effect of a new bilateral tax treaty on reducing the United States's current account deficit with Japan. According to our definition of causality, the causal effect here is the reduction in the expected current account deficit with the tax treaty in effect as compared to the same situation (at the same time and for the same countries) with the exception that the treaty was not in effect. The causal mechanism operating here would include, in turn, the signing and ratification of the tax treaty, newspaper reports of the event, meetings of the relevant actors within major multinational companies, compensatory actions to reduce their total international tax burden (such as changing its transfer pricing rules or moving manufacturing plants between countries), further actions by other companies and workers to take advantage of the movements of capital and labor between countries, and so on, until we reach the final effect on the balance of payments between the United States and Japan.

From the standpoint of processes through which causality operates, an emphasis on causal mechanisms makes intuitive sense: any coher-

ent account of causality needs to specify how its effects are exerted. Identifying causal mechanisms is a popular way of doing empirical analyses. It has been called, in slightly different forms, "process tracing" (which we discuss in section 6.3.3), "historical analysis," and "detailed case studies." Many of the details of well-done case studies involve identifying these causal mechanisms.

However, identifying the causal mechanisms requires causal inference, using the methods discussed before. That is, to demonstrate the causal status of each potential linkage in such a posited mechanism, the investigator would have to define and then estimate the causal effect underlying it. To portray an internally consistent causal mechanism requires using our more fundamental definition of causality offered in section 3.1 for each link in the chain of causal events.

Hence our definition of causality is logically prior to the identification of causal mechanisms. Furthermore, there always exists in the social sciences an infinity of causal steps between any two links in the chain of causal mechanisms. If we posit that an explanatory variable causes a dependent variable, a "causal mechanisms" approach would require us to identify a list of causal links between the two variables. This definition would also require us to identify a series of causal linkages, to define causality for each pair of consecutive variables in the sequence, and to identify the linkages between any two of these variables and the connections between each pair of variables. This approach quickly leads to infinite regress, and at no time does it alone give a precise definition of causality for any one cause and one effect.

In our example of the effect of a presidential versus parliamentary system on democratic stability (section 3.1.2), the hypothesized causal mechanisms include greater minority disaffection under a presidential regime and lesser governmental decisiveness under a parliamentary regime. These intervening effects—caused by the constitutional system and, in turn, affecting political stability—can be directly observed. We could monitor the attitudes or behaviors of minorities to see how they differ under the two experimental conditions or study the decisiveness of the governments under each system. Yet even if the causal effect of presidential versus parliamentary systems could operate in different ways, our definition of the causal effect would remain valid. We can define a causal effect without understanding all the causal mechanisms involved, but we cannot identify causal mechanisms without defining the concept of causal effect.

In our view, identifying the mechanisms by which a cause has its effect often builds support for a theory and is a very useful operational procedure. Identifying causal mechanisms can sometimes give us more leverage over a theory by making observations at a different level of analysis into implications of the theory. The concept can also create new causal hypotheses to investigate. However, we should not confuse a definition of causality with the nondefinitional, albeit often useful, operational procedure of identifying causal mechanisms.

3.2.2 "Multiple Causality"

Charles Ragin, in a recent work (1987:34–52), argues for a methodology with many explanatory variables and few observations in order that one can take into account what he calls "multiple causation." That is, "The phenomenon under investigation has alternative determinants—what Mill (1843) referred to as the problem of 'plurality of causes.'" This is the problem referred to as "equifinality" in general systems theory (George 1982:11). In situations of multiple causation, these authors argue that the same outcome can be caused by combinations of different independent variables.[18]

Under conditions in which different explanatory variables can account for the same outcome as a dependent variable, according to Ragin, some statistical methods will falsely reject the hypothesis that these variables have causal status. Ragin is correct that some statistical models (or relevant qualitative research designs) could fail to alert an investigator to the existence of "multiple causality," but appropriate statistical models can easily handle situations like these (some of which Ragin discusses).

Moreover, the fundamental features of "multiple causality" are compatible with our definition of causality. They are also no different for quantitative than qualitative research. The idea contains no new features or theoretical requirements. For example, consider the hypothesis that a person's level of income depends both on high educational attainment *and* highly educated parents. Having one but not both is insufficient. In this case, we need to compare categories of our causal variable: respondents who have high educational attainment and highly educated parents, the two groups who have one but not the other, and the group with neither. Thus, the concept of "multiple causation" puts greater demands on our data since we now have four cat-

[18] This idea is often explained in terms of no explanatory variable being either necessary or sufficient for a particular value of a dependent variable to occur. However, this is misleading terminology because the distinction between necessary and sufficient conditions largely disappears when we allow for the possibility that causes are probabilistic. As Little (1991:27) explains, "Consider the claim that poor communication among superpowers during crises increases the likelihood of war. This is a probabilistic claim. It identifies a causal variable (poor communication) and asserts that this variable increases the probability of a given outcome (war). It cannot be translated into a claim about the necessary and sufficient conditions for war; however, it is irreducibly probabilistic."

egories of our causal variables, but it does not require a modification of our definition of causality. For our definition, we would need to measure the expected income for the same person, at the same time, experiencing each of the four conditions.

But what happens if different causal explanations generate the same values of the dependent variable? For example, suppose we consider whether or not one graduated from college as our (dichotomous) causal variable in a population of factory workers. In this situation, both groups could quite reasonably earn the same income (our dependent variable). One reason might be that this explanatory variable (college attendance) has no causal effect on income among factory workers, perhaps because a college education does not help one perform better. Alternatively, different explanations might lead to the same level of income for those educated and those not educated. College graduates might earn a particular level of income because of their education, whereas those who had no college education might earn the same level of income because of their four years of additional seniority on the job. In this situation wouldn't we be led to conclude that "college education" has no causal effect on income levels for those who will become factory workers?

Fortunately, our definition of causality requires that we more carefully specify the counterfactual condition. In the present example, the values of the key causal variable to be varied are (1) college education, as compared to (2) no college education but four additional years of job seniority. The dependent variable is starting annual income. Our causal effect is then defined as follows: we record the income of a person graduating from college who goes to work in a factory. Then, we go back in time four years, put this same person to work in the same factory instead of in college and, at the end of four years, measure his or her income "again." The expected difference between these two levels of income for this one individual is our definition of the mean causal effect. In the present situation, we have imagined that this causal effect is zero. But this does not mean that "college education has no effect on income," only that the average difference between treatment groups (1) and (2) is zero. In fact, there is no logically unique definition of "the causal effect of college education" since one cannot define a causal effect without at least two conditions. The conditions need not be the two listed here, but they must be very clearly identified.

An alternative pair of causal conditions is to compare a college graduate with someone without a college degree but with the same level of job seniority as the college graduate. In one sense, this is unrealistic, since the non-college graduate would have to do something for the four years while not attending college, but perhaps we would be willing to imagine that this person had a different, irrelevant job for those four years. Put differently, this alternative counterfactual is the effect of a college education compared to that of none, with job seniority held constant. Failure to hold seniority constant in the two causal conditions would cause any research design to yield estimates of our first counterfactual instead of this revised one. If the latter were the goal, but no controls were introduced, our empirical analysis would be flawed due to "omitted variable bias" (which we introduce in section 5.2).

Thus, the issues addressed under the label "multiple causation" do not confound our definition of causality although they may make greater demands in our subsequent analyses. The fact that some dependent variables, and perhaps all interesting social science–dependent variables, are influenced by many causal factors does not make our definition of causality problematic. The key to understanding these very common situations is to define the counterfactual conditions making up each causal effect very precisely. We demonstrate in chapter 5 that researchers need not identify "all" causal effects on a dependent variable to provide estimates of the one causal effect of interest (even if that were possible). A researcher can focus on only the one effect of interest, establish firm conclusions, and then move on to others that may be of interest (see sections 5.2 and 5.3).[11]

3.2.3 *"Symmetric" and "Asymmetric" Causality*

Stanley Lieberson (1985:63–64) distinguishes between what he refers to as "symmetrical" and "asymmetrical" forms of causality. He is interested in causal effects which differ when an explanatory variable is increased as compared to when it is decreased. In his words,

> In examining the causal influence of X_1 [an explanatory variable] on Y [a dependent variable], for example, one has also to consider whether shifts to a given value of X_1 from either direction have the same consequences for Y. . . . If the causal relationship between X_1 [an explanatory variable] and Y

[11] Our emphasis on distinguishing systematic from nonsystematic components of observations subject to causal inference reflects our general view that the world, at least as we know it, is probabilistic rather than deterministic. Hence, we also disagree with Ragin's premise (1987:15) that "explanations which result from applications of the comparative method are not conceived in probabilistic terms because every instance of a phenomenon is examined and accounted for if possible." Even if it were possible to collect a census of information on every instance of a phenomenon and every permutation and combination of values of the explanatory variables, the world still would have produced those data according to some probabilistic process (as defined in section 2.6). This

> [a dependent variable] is symmetrical or truly reversible, then the effect on *Y* of an increase in X_1 will disappear if X_1 shifts back to its earlier level (assuming that all other conditions are constant).

As an example of Lieberson's point, imagine that the Fourth Congressional District in New York had no incumbent in 1998 and that the Democratic candidate received 55 percent of the vote. Lieberson would define the causal effect of incumbency as the increase in the vote if the winning Democrat in 1998 runs as an incumbent in the next election in the year 2000. This effect would be "symmetric" if the absence of an incumbent in the subsequent election (in year 2002) caused the vote to return to 55 percent. The effect might be "asymmetric" if, for example, the incumbent Democrat raised money and improved the Democratic party's campaign organization; as a result, if no incumbent were running in 2002, the Democratic candidate might receive more than 55 percent of the vote.

Lieberson's argument is clever and very important. However, in our view, his argument does not constitute a *definition* of causality, but applies only to some causal *inferences*—the process of learning about a causal effect from existing observations. In section 3.1, we defined causality for a single unit. In the present example, a causal effect can be defined theoretically on the basis of hypothetical events occurring only in the 1998 election in the Fourth District in New York. Our definition is the difference in the systematic component of the vote in this district with an incumbent in this election and without an incumbent in the same election, time, and district.

In contrast, Lieberson's example involves no hypothetical quantities and therefore *cannot* be a causal definition. This example involves only what would actually occur if the explanatory variable changed in two real elections from nonincumbent to incumbent, versus incumbent to nonincumbent in two other elections. Any empirical analysis of this example would involve numerous problems of inference. We discuss many of these problems of causal inference in chapters 4–6. In the present example, we might ask whether the estimated effect seemed larger only because we failed to account for a large number of recently registered citizens in the Fourth District. Or, did the surge in support for the Democrat in the election in which she or he was an incumbent seem smaller than it should because we necessarily discarded districts where the Democrat lost the first election?

Thus, Lieberson's concepts of "symmetrical" and "asymmetrical" causality are important to consider in the context of causal inference. However, they should not be confused with a theoretical definition of causality, which we give in section 3.1.

seems to invalidate Ragin's "Boolean Algebra" approach as a general way of designing theoretical explanations or making inferences; to learn from data requires the same logic of scientific inference that we discuss in this book. However, his approach can still be valuable as a form of formal theory (see section 3.5.2): it enables the investigator to specify a theory and its implications in a way that might be much more difficult without it.

3.3 Assumptions Required for Estimating Causal Effects

How do we avoid the Fundamental Problem of Causal Inference and also the problem of separating systematic from nonsystematic components? The full answer to this question will consume chapters 4–6, but we provide an overview here of what is required in terms of the two possible assumptions that enable us to get around the fundamental problem. These are *unit homogeneity* (which we discuss in section 3.3.1) and *conditional independence* (section 3.3.2). These assumptions, like any other attempt to circumvent the Fundamental Problem of Causal Inference, always involve some untestable assumptions. It is the responsibility of all researchers to make the substantive implications of this weak spot in their research designs extremely clear and visible to readers. Causal inferences should not appear like magic. The assumptions can and should be justified with whatever side information or prior research can be mustered, but it always must be explicitly recognized.

3.3.1 *Unit Homogeneity*

If we cannot rerun history at the same time and the same place with different values of our explanatory variable each time—as a true solution to the Fundamental Problem of Causal Inference would require—we can attempt to make a second-best assumption: we can rerun our experiment in two different units that are "homogeneous." Two units *are homogeneous when the expected values of the dependent variables from each unit are the same when our explanatory variable takes on a particular value.* (That is, $\mu_1^1 = \mu_2^1$ and $\mu_1^0 = \mu_2^0$.) For example, if we observe $X = 1$ (an incumbent) in district 1 and $X = 0$ (no incumbent) in district 2, an assumption of unit homogeneity means that we can use the observed proportions of the vote in two separate districts for inference about the causal effect β, which we assume is the same in both districts. For a data set with n observations, unit homogeneity is the assumption that all units with the same value of the explanatory variables have the same expected value of the dependent variable. Of course, this is only an assumption and it can be wrong: the two districts might differ in

some unknown way that would bias our causal inference. Indeed, any two real districts will differ in some ways; application of this assumption requires that these districts must be the same on average over many hypothetical replications of the election campaign. For example, patterns of rain (which might inhibit voter turnout in some areas) would not differ across districts on average unless there were systematic climatic differences between the two areas.

In the following quotation, Holland (1986:947) provides a clear example of the unit homogeneity assumption (defined from his perspective of a realized causal effect instead of the mean causal effect). Since very little randomness exists in the experiment in the following example, his definition and ours are close. (Indeed, as we show in section 4.2, with a small number of units, the assumption of unit homogeneity is most useful when the amount of randomness is fairly low.)

> If [the unit] is a room in a house, *t* [for 'treatment'] means that I flick the light switch in that room, *c* [for 'control'] means that I do not, and [the dependent variable] indicates whether the light is on or not a short time after applying either *t* or *c*, then I might be inclined to believe that I can *know* the values of [the dependent variable for both *t* and *c*] by simply flicking the switch. It is clear, however, that it is only because of the plausibility of certain assumptions about the situation that this *belief* of mine can be shared by anyone else. If, for example, the light has been flicking off and on for no apparent reason while I am contemplating beginning this experiment, I might doubt that I would know the values of [the dependent variable for both *t* and *c*] after flicking the switch—at least until I was clever enough to figure out a new experiment!

In this example, the unit homogeneity assumption is that if we had flicked the switch (in Holland's notation, applied *t*) in both periods, the expected value (of whether the light will be on) would be the same. Unit homogeneity also assumes that if we had not flicked the switch (applied *c*) in both periods, the expected value would be the same, although not necessarily the same as when *t* is applied. Note that we would have to reset the switch to the off position after the first experiment to assure this, but we would also have to make the untestable assumption that flipping the switch on in the first period does not effect the two hypothetical expected values in the next period (such as if a fuse were blown after the first flip). In general, the unit homogeneity assumption is untestable for a single unit (although, in this case, we might be able to generate several new hypotheses about the causal mechanism by ripping the wall apart and inspecting the wiring).

A weaker, but also fully acceptable, version of unit homogeneity is the constant *effect* assumption. Instead of assuming that the expected value of the dependent variable is the same for different units with the same value of the explanatory variable, we need only to assume that the causal effect is constant. This is a weaker version of the unit homogeneity assumption, since the causal effect is only the difference between the two expected values. If the two expected values for units with the same value of the explanatory variable vary in the same way, the unit homogeneity assumption would be violated, but the constant effect assumption would still be valid. For example, two congressional districts could vary in the expected proportion of the vote for Democratic nonincumbents (say 45 percent vs. 65 percent), but incumbency could still add an additional ten percent to the vote of a Democratic candidate of either district.

The notion of unit homogeneity (or the less demanding assumption of constant causal effects) lies at the base of all scientific research. It is, for instance, the assumption underlying the method of comparative case studies. We compare several units that have varying values on our explanatory variables and observe the values of the dependent variables. We believe that the differences we observe in the values of the dependent variables are the result of the differences in the values of the explanatory variables that apply to the observations. What we have shown here is that our "belief" in this case necessarily relies upon an assumption of unit homogeneity or constant effects.

Note that we may seek homogeneous units across time or across space. We can compare the vote for the Democratic candidate when there is a Democratic incumbent running with the vote when there is no Democratic incumbent in the same district at different times or across different districts at the same time (or some combination of the two). Since a causal effect can only be estimated instead of known, we should not be surprised that the unit homogeneity assumption is generally untestable. But it is important that the nature of the assumption is made explicit. Across what range of units do we expect our assumption of a uniform incumbency effect to hold? All races for Congress? Congressional but not Senate races? Races in the North only? Races in the past two decades only?

Notice how the unit homogeneity assumption relates to our discussion in section 1.1.3 on complexity and "uniqueness." There we argued that social science generalization depends on our ability to simplify reality coherently. At the limit, simplifying reality for the purpose of making causal inferences implies meeting the standards for unit homogeneity: the observations being analyzed become, for the purposes of analysis, identical in relevant respects. Attaining unit homogeneity is often impossible: congressional elections, not to speak of revolutions, are hardly close analogies to light switches. But understanding

the degree of heterogeneity in our units of analysis will help us to estimate the degree of uncertainty or likely biases to be attributed to our inferences.

3.3.2 *Conditional Independence*

Conditional independence is the assumption that values are assigned to explanatory variables independently of the values taken by the dependent variables. (The term is sometimes used in statistics, but it does not have the same definition as it commonly does in probability theory.) That is, after taking into account the explanatory variables (or controlling for them), the process of assigning values to the explanatory variable is independent of both (or, in general two or more) dependent variables, Y_i^N and Y_i^E. We use the term "assigning values" to the explanatory variables to describe the process by which these variables obtain the particular values they have. In experimental work, the researcher actually *assigns* values to the explanatory variables; some subjects are assigned to the treatment group and others to the control group. In nonexperimental work, the values that explanatory variables take may be "assigned" by nature or the environment. What is crucial in these cases is that the values of the explanatory variables are not caused by the dependent variables. The problem of "endogeneity" that exists when the explanatory variables are caused, at least in part, by the dependent variables is described in section 5.4.

Large-*n* analyses that involve the procedures of random selection and assignment constitute the most reliable way to assure conditional independence and do not require the unit homogeneity assumption. Random selection and assignment help us to make causal inferences because they *automatically* satisfy three assumptions that underlie the concept of conditional independence: (1) that the process of assigning values to the explanatory variables is independent of the dependent variables (that is, there is no endogeneity problem); (2) that selection bias, which we discuss in section 4.3, is absent; and (3) that omitted variable bias (section 5.2) is also absent. Thus, if we are able to meet these conditions in any way, either through random selection and assignment (as discussed in section 4.2) or through some other procedure, we can avoid the Fundamental Problem of Causal Inference.

Fortunately, random selection and assignment are *not* required to meet the conditional independence assumption. If the process by which the values of the explanatory variables are "assigned" is not independent of the dependent variables, we can still meet the conditional independence assumption if we learn about this process and

include a measure of it among our control variables. For example, suppose we are interested in estimating the effect of the degree of residential segregation on the extent of conflict between Israelis and Palestinians in communities on the Israeli-occupied West Bank. Our conditional independence assumption would be severely violated if we looked only at the association between these two variables to find the causal effect. The reason is that the Israelis and Palestinians who choose to live in segregated neighborhoods may do so out of an ideological belief about who ultimately has rights to the West Bank. Ideological extremism (on both sides) may therefore lead to conflict. A measure that we believe to be residential segregation might really be a surrogate for ideology. The difference between the two explanations may be quite important, since a new housing policy might help remedy the conflict if residential segregation were the real cause, whereas this policy would be ineffective or even counterproductive if ideology were really the driving force. We might correct for the problem here by also measuring the ideology of the residents explicitly and controlling for it. For example, we could learn how popular extremist political parties are among the Israelis and PLO affiliation is among the Palestinians. We could then control for the possibly confounding effects of ideology by comparing communities with the same level of ideological extremism but differing levels of residential segregation.

When random selection and assignment are infeasible and we cannot control for the process of assignment and selection, we have to resort to some version of the unit homogeneity assumption in order to make valid causal inferences. Since that assumption will be only imperfectly met in social science research, we will have to be especially careful to specify our degree of uncertainty about causal inferences. This assumption will be particularly apparent when we discuss the procedures used in "matching" observations in section 5.6.

Notation for a Formal Model of a Causal Effect. *We now generalize our notation for the convenience of later sections.* In general, we will have n realizations of a random variable Y_i. In our running quantitative example, n is the number of congressional districts (435), and the realization y_i of the random variable Y_i is the observed Democratic proportion of the two-party vote in district i (such as 0.56). The expected nonincumbent Democratic proportion of the two-party vote (the average over all hypothetical replications) in district i is μ_i^N. We define the explanatory variable as X_i, which is coded in the present example as zero when district i has no Democratic incum-

bent and as one when district i has a Democratic incumbent. Then, we can denote the mean causal effect in unit i as

$$\beta_i = E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = \mu_i^1 - \mu_i^N \tag{3.4}$$

and incorporate it into the following simple formal model:

$$\begin{aligned} E(Y_i) &= \mu_i^N + X_i(\mu_i^1 - \mu_i^N) \\ &= \mu_i^N + X_i\beta \end{aligned} \tag{3.5}$$

Thus, when district i has no incumbent, and $X_i = 0$, the expected value is determined by substituting zero into equation (3.5) for X_i, and the answer is as before:

$$\begin{aligned} E(Y_i|X_i = 0) &= \mu_i^N + 0\beta \\ &= \mu_i^N \end{aligned}$$

Similarly, when a Democratic incumbent is running in district i, the expected value is μ_i^1:

$$\begin{aligned} E(Y_i|X_i = 1) &= \mu_i^N + (1)\beta \\ &= \mu_i^N + \beta \\ &= \mu_i^N + (\mu_i^1 - \mu_i^N) \\ &= \mu_i^1 \end{aligned}$$

Thus, equation (3.5) provides a useful model of causal inference, and β—the difference between the two theoretical properties—is our causal effect. Finally, for future reference, we simplify equation (3.5) one last time. If we assume that Y_i has a zero mean (or is written as a deviation from its mean, which does not limit the applicability of the model in any way), then we can drop the intercept from this equation, and write it more simply as

$$E(Y_i) = X_i\beta \tag{3.6}$$

The parameter β is still the theoretical value of the mean causal effect, a systematic feature of the random variables, and one of our goals in causal inference. This model is a special case of "regression analysis," which is common in quantitative research, but regression coefficients are only sometimes coincident with estimates of causal effects.

3.4 Criteria for Judging Causal Inferences

Recall that by defining causality in terms of random variables, we were able to draw a strict analogy between it and other systematic features of phenomena, such as a mean or a variance, on which we focus in making descriptive inferences. This analogy enables us to use precisely the same criteria to judge causal inferences as we used to judge descriptive inferences in section 2.7: unbiasedness and efficiency. Hence, most of what we said on this subject in Chapter 2 applies equally well to the causal inference problems we deal with here. In this section, we briefly formalize the relatively few differences between these two situations.

In section 2.7 the object of our inference was a mean (the expected value of a random variable), which we designate as μ. We conceptualize μ as a fixed, but unknown, number. An estimator of μ is said to be unbiased if it equals μ on average over many hypothetical replications of the same experiment.

As above, we continue to conceptualize the expected value of a random causal effect, denoted as β, as a fixed, but unknown, number. The unbiasedness is then defined analogously: an estimator of β is unbiased if it equals β on average over many hypothetical replications of the same experiment. Efficiency is also defined analogously as the variability across these hypothetical replications. These are very important concepts that will serve as the basis for our studies of many of the problems of causal inference in chapters 4–6. The two boxes that follow provide formal definitions.

A Formal Analysis of Unbiasedness of Causal Estimates. In this box, we demonstrate the unbiasedness of the estimator of the causal effect parameter from section 3.1. The notation and logic of these ideas closely parallel those from the formal definition of unbiasedness in the context of descriptive inference in section 2.7. The simple linear model with one explanatory and one dependent variable is as follows:[12]

[12] In order to avoid using a constant term, we assume that all variables have zero means. This simplifies the presentation but does not limit our conclusions in any way.

$$E(Y_i) = \beta X_i$$

Our estimate of β is simply the least squares regression estimate:

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \tag{3.7}$$

To determine whether b is an unbiased estimator of β, we need to take the expected value, averaging over hypothetical replications:

$$\begin{aligned} E(b) &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \qquad (3.8)\\ &= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2} \\ &= \beta \end{aligned}$$

which proves that b is an unbiased estimator of β.

A Formal Analysis of Efficiency. Here, we assess the efficiency of the standard estimator of the causal effect parameter β from section 3.1. We proved in equation (3.8) that this estimator is unbiased and now calculate its variance:

$$\begin{aligned} V(b) &= V\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \qquad (3.9)\\ &= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \sum_{i=1}^{n} X_i^2 V(Y_i) \\ &= \frac{V(Y_i)}{\sum_{i=1}^{n} X_i^2} \end{aligned}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

Thus, the variance of this estimator is a function of two components. First, the more random each unit in our data (the larger is σ^2) is, the more variable will be our estimator b; this should be no surprise. In addition, the larger the observed variance in the explanatory variable ($\sum_{i=1}^{n} X_i^2$), the less variable will be our estimate of b. In the extreme case of no variability in X, nothing can help us estimate the effect of changes in the explanatory variable on the dependent variable, and the formula predicts an infinite variance (complete uncertainty) in this instance. More generally, this component indicates that efficiency is greatest when we have evidence from a larger range of values of the explanatory variable. In general, then, it is best to evaluate our causal hypotheses in as many diverse situations as possible. One way to think of this latter point is to think about drawing a line with a ruler, two dots on a page, and a shaky hand. If the two dots are very close together (small variance of X), errors in the placement of the ruler will be much larger than if the dots are farther apart (the situation of a large variance in X).

3.5 Rules for Constructing Causal Theories

Much sensible advice about improving qualitative research is precise, specific, and detailed; it involves a manageable and therefore narrow aspect of qualitative research. However, even in the midst of solving a host of individual problems, we must keep the big picture firmly in mind: each specific solution must help in solving whatever is the general causal inference problem one aims to solve. Thus far in this chapter, we have provided a precise theoretical definition of a causal effect and discussed some of the issues involved in making causal inferences. We take a step back now and provide a broader overview of some rules regarding theory construction. As we discuss (and have discussed in section 1.2), improving theory does not end when data collection begins.

Causal theories are designed to show the causes of a phenomenon or set of phenomena. Whether originally conceived as deductive or inductive, any theory includes an interrelated set of causal hypotheses. Each hypothesis specifies a posited relationship between variables that creates observable implications: if the specified explanatory variables

take on certain values, other specified values are predicted for the dependent variables. Testing or evaluating any causal hypothesis requires causal inference. The overall theory, of which the hypotheses are parts should be *internally consistent*, or else hypotheses can be generated that contradict one another.

Theories and hypotheses that fit these definitions have an enormous range. In this section, we provide five rules that will help in formulating good theories, and we provide a discussion of each with examples.

3.5.1 *Rule 1: Construct Falsifiable Theories*

By this first rule, we do not only mean that a "theory" incapable of being wrong is not a theory. We also mean that we should design theories so that they can be shown to be wrong as easily and quickly as possible. Obviously, we should not actually try to be wrong, but even an incorrect theory is better than a statement that is neither wrong nor right. The emphasis on falsifiable theories forces us to keep the right perspective on uncertainty and guarantees that we treat theories as tentative and not let them become dogma. We should always be prepared to reject theories in the face of sufficient scientific evidence against them. One question that should be asked about any theory (or of any hypothesis derived from the theory) is simply: what evidence would falsify it? The question should be asked of all theories and hypotheses but, above all, the researcher who poses the theory in the first place should ask it of his or her own.

Karl Popper is most closely identified with the idea of falsifiability (Popper 1968). In Popper's view, a fundamental asymmetry exists between confirming a theory (verification) and disconfirming it (falsification). The former is almost irrelevant, whereas the latter is the key to science. Popper believes that a theory once stated immediately becomes part of the body of accepted scientific knowledge. Since theories are general, and hypotheses specific, theories technically imply an infinite number of hypotheses. However, empirical tests can only be conducted on a finite number of hypotheses. In that sense, "theories are not verifiable" because we can never test all observable implications of a theory (Popper 1968:252). Each hypothesis tested may be shown to be consistent with the theory, but any number of consistent empirical results will not change our opinions since the theory remains accepted scientific knowledge. On the other hand, if even a single hypothesis is shown to be wrong, and thus inconsistent with the theory, the theory is falsified, and it is removed from our collection of scientific knowledge. "The passing of tests therefore makes not a jot of difference to the status of any hypothesis, though the failing of just one test may make a great deal of difference" (Miller 1988:22). Popper did not mean falsification to be a deterministic concept. He recognized that any empirical inference is to some extent uncertain (Popper 1982). In his discussion of disconfirmation, he wrote, "even if the asymmetry [between falsification and verification] is admitted, it is still impossible, for various reasons, that any theoretical system should ever be conclusively falsified" (Popper 1968:42).

In our view, Popper's ideas are fundamental for *formulating* theories. We should always design theories that are vulnerable to falsification. We should also learn from Popper's emphasis on the tentative nature of any theory. However, for *evaluating* existing social scientific theories, the asymmetry between verification and falsification is not as significant. Either one adds to our scientific knowledge. The question is less whether, in some general sense, a theory is false or not—virtually every interesting social science theory has at least one observable implication that appears wrong—than *how much of the world the theory can help us explain*. By Popper's rule, theories based on the assumption of rational choice would have been rejected long ago since they have been falsified in many specific instances. However, social scientists often choose to retain the assumption, suitably modified, because it provides considerable power in many kinds of research problems (see Cook and Levi 1990). The same point applies to virtually every other social science theory of interest. The process of trying to falsify theories in the social sciences is really one of searching for their bounds of applicability. If some observable implication indicates that the theory does not apply, we learn something; similarly, if the theory works, we learn something too.

For scientists (and especially for social scientists) evaluating properly formulated theories, Popper's fundamental asymmetry seems largely irrelevant. O'Hear (1989:43) made a similar point about the application of Popper's ideas to the physical sciences:

> Popper always tends to speak in terms of *explanations* of universal theories. But once again, we have to insist that proposing and testing universal theories is only part of the aim of science. There may be no true universal theories, owing to conditions differing markedly through time and space; this is a possibility we cannot overlook. But even if this were so, science could still fulfil [sic] many of its aims in giving us knowledge and true predictions about conditions in and around our spatio-temporal niche.

Surely this same point applies even more strongly to the social sciences.

Furthermore, Popper's evaluation of theories does not fundamentally distinguish between a newly formulated theory and one that has

withstood numerous empirical tests. When we are testing for the deterministic distinction between the truth or fiction of a universal theory (of which there exists no interesting examples), Popper's view is appropriate, but from our perspective of searching for the bounds of a theory's applicability, his view is less useful. As we have indicated many times in this book, we require all inferences about specific hypotheses to be made by stating a best guess (an estimate) and a measure of the uncertainty of this guess. Whether we discover that the inference is consistent with our theory or inconsistent, our conclusion will have as much effect on our belief in the theory. Both consistency and inconsistency provide information about the truth of the theory and should affect the certainty of our beliefs.[13]

Consider the hypothesis that Democratic and Republican campaign strategies during American presidential elections have a small net effect on the election outcome. Numerous more specific hypotheses are implied by this one, such as that television commercials, radio commercials, and debates all have little effect on voters. Any test of the theory must really be a test of one of these hypotheses. One test of the theory has shown that forecasts of the outcome can be made very accurately with variables available only at the time of the conventions—and thus before the campaign (Gelman and King 1993). This test is consistent with the theory (if we can predict the election before the campaign, the campaign can hardly be said to have much of an impact), but it does not absolutely verify it. Some aspect of the campaign could have some small effect that accounts for some of the forecasting errors (and few researchers doubt that this is true). Moreover, the prediction could have been luck, or the campaign could have not included any innovative (and hence unpredictable) tactics during the years for which data were collected.

We could conduct numerous other tests by including variables in the forecasting model that measure aspects of the campaign, such as relative amounts of TV and radio time, speaking ability of the candidates, and judgments as to the outcomes of the debates. If all of these hypotheses show no effect, then Popper would say that our opinion is not changed in any interesting way: the theory that presidential campaigns have no effect is still standing. Indeed, if we did a thousand similar tests and all were consistent with the theory, the theory could still be wrong since we have not tried every one of the infinite number of possible variables measuring the campaign. So even with a lot of results consistent with the theory, it still *might* be true that presidential campaigns influence voter behavior.

However, if a single campaign event—such as substantial accusations of immoral behavior—is shown to have some effect on voters, the theory would be falsified. According to Popper, even though this theory was not conclusively falsified (which he recognized as impossible), we learn more from it than the thousand tests consistent with the theory.

To us, this is not the way social science is or should be conducted. After a thousand tests in favor and one against, even if the negative test seemed valid with a high degree of certainty, we would not drop the theory that campaigns have no effect. Instead, we might modify it to say perhaps that normal campaigns have no effect except when there is considerable evidence of immoral behavior by one of the candidates—but since this modification would make our theory more restrictive, we would need to evaluate it with a new set of data before being confident of its validity. The theory would still be very powerful, and we would know somewhat more about the bounds to which the theory applies with each passing empirical evaluation. Each test of a theory affects both the estimate of its validity and the uncertainty of that estimate; and it may also affect to what extent we wish the theory to apply.

In the previous discussion, we suggested an important approach to theory, as well as issued a caution. The approach we recommended is one of sensitivity to the contingent nature of theories and hypotheses. Below, we argue for seeking broad application for our theories and hypotheses. This is a useful research strategy, but we ought always to remember that theories in the social sciences are unlikely to be universal in their applicability. Those theories that are put forward as applying to everything, everywhere—some versions of Marxism and rational choice theory are examples of theories that have been put forward with claims of such universality—are either presented in a tautological manner (in which case they are neither true nor false) or in a way that allows empirical disconfirmation (in which case we will find that they make incorrect predictions). Most useful social science theories are valid under particular conditions (in election campaigns without strong evidence of immoral behavior by a candidate) or in particular settings (in industrialized but not less industrialized nations, in House but not Senate campaigns). We should always try to specify the bounds of applicability of the theory or hypothesis. The next step is to

[13] Some might call us (or accuse us of being?) "justificationists" or even "probabilistic justificationists" (see Lakatos 1970), but if we must be labeled, we prefer the more coherent, philosophical Bayesian label (see Learner 1978, Zellner 1971, and Barnett 1982). In fact, our main difference with Popper is our goals. Given his precise goal, we agree with his procedure; given our goal, perhaps he might agree with ours. However, we believe that our goals are closer to those in use in the social sciences and are also closer to the ones likely to be successful.

raise the question: Why do these bounds exist? What is it about Senate races that invalidates generalizations that are true for House races? What is it about industrialization that changes the causal effects? What variable is missing from our analysis which could produce a more generally applicable theory? By asking such questions, we move beyond the boundaries of our theory or hypothesis to show what factors need to be considered to expand its scope.

But a note of caution must be added. We have suggested that the process of evaluating theories and hypotheses is a flexible one: particular empirical tests neither confirm nor disconfirm them once and for all. When an empirical test is inconsistent with our theoretically based expectations, we do not immediately throw out the theory. We may do various things: We may conclude that the evidence may have been poor due to chance alone; we may adjust what we consider to be the range of applicability of a theory or hypothesis even if it does not hold in a particular case and, through that adjustment, maintain our acceptance of the theory or hypothesis. Science proceeds by such adjustments; but they can be dangerous. If we take them too far we make our theories and hypotheses invulnerable to disconfirmation. The lesson is that we must be very careful in adapting theories to be consistent with new evidence. We must avoid stretching the theory beyond all plausibility by adding numerous exceptions and special cases.

If our study disconfirms some aspect of a theory, we may choose to retain the theory but add an exception. Such a procedure is acceptable as long as we recognize the fact that we are reducing the claims we make for the theory. The theory, though, is less valuable since it explains less; in our terminology, we have less *leverage* over the problem we seek to understand.[14] Furthermore, such an approach may yield a "theory" that is merely a useless hodgepodge of various exceptions and exclusions. At some point we must be willing to discard theories and hypotheses entirely. Too many exceptions, and the theory should be rejected. Thus, by itself, *parsimony, the normative preference for theories with fewer parts, is not generally applicable.* All we need is our more general notion of maximizing leverage, from which the idea of parsimony can be fully derived when it is useful. The idea that science is largely a process of explaining many phenomena with just a few makes clear that theories with fewer parts are not better or worse. To maximize leverage, we should attempt to formulate theories that explain as much as possible with as little as possible. Sometimes this formulation is achieved via parsimony, but sometimes not. We can conceive of examples by which a slightly more complicated theory will explain vastly more of the world. In such a situation, we would surely use the nonparsimonious theory, since it maximizes leverage more than the more parsimonious theory.[15]

[14] As always, when we do modify a theory to be consistent with evidence we have collected, then the theory (or that part of it on which our evidence bears) should be evaluated in a different context or new data set.

3.5.2 *Rule 2: Build Theories That Are Internally Consistent*

A theory which is internally inconsistent is not only falsifiable—it is false. Indeed, this is the only situation where the veracity of a theory is known without any empirical evidence: if two or more parts of a theory generate hypotheses that contradict one another, then no evidence from the empirical world can uphold the theory. Ensuring that theories are internally consistent should be entirely uncontroversial, but consistency is frequently difficult to achieve. One method of producing internally consistent theories is with formal, mathematical modeling. *Formal modeling* is a practice most developed in economics but increasingly common in sociology, psychology, political science, anthropology, and elsewhere (see Ordeshook 1986). In political science, scholars have built numerous substantive theories from mathematical models in rational choice, social choice, spatial models of elections, public economics, and game theory. This research has produced many important results, and large numbers of plausible hypotheses. One of the most important contributions of formal modeling is revealing the internal inconsistency in verbally stated theories.

However, as with other hypotheses, formal models do not constitute verified explanations without empirical evaluation of their predic-

[15] Another formulation of Popper's view is that "you can't prove a negative." You cannot, he argues, because a result consistent with the hypothesis might just mean that you did the wrong test. Those who try to prove the negative will always run into this problem. Indeed, their troubles will be not only theoretical but professional as well since journals are more likely to publish positive results rather than negative ones.

This has led to what is called *the file drawer problem*, which is chronic in the quantitative literature. Suppose no patterns exist in the world. Then five of every one hundred tests of any pattern will fall outside the 95 percent confidence interval and thus produce incorrect inferences. If we were to assume that journals publish positive rather than negative results, they will publish only those 5 percent that are "significant"; that is, they will publish only the papers that come to the wrong conclusions, and our file drawers will be filled with all the papers that come to the right conclusions! (See Iyengar and Greenhouse (1988) for a review of the statistical literature on this problem.) In fact, these incentives are well known by researchers, and it probably affects their behaviors as well. Even though the acceptance rate at many major social science journals is roughly 5 percent, the situation is not quite this bad, but it is still a serious problem. In our view, the file drawer problem could be solved if everyone adopted our alternative position. *A negative result is as useful as a positive one; both can provide just as much information about the world.* So long as we present our estimates and a measure of our uncertainty, we will be on safe ground.

tions. Formality does help us reason more clearly, and it certainly ensures that our ideas are internally consistent, but it does not resolve issues of empirical evaluation of social science theories. An assumption in a formal model in the social sciences is generally a convenience for mathematical simplicity or for ensuring that an equilibrium can be found. Few believe that the political world is mathematical in the same way that some physicists believe the physical world is. Thus, formal models are merely models—abstractions that should be distinguished from the world we study. Indeed, some formal theories make predictions that depend on assumptions that are vastly oversimplified, and these theories are sometimes not of much empirical value. They are only more precise in the abstract than are informal social science theories: they do not make more specific predictions about the real world, since the conditions they specify do not correspond, even approximately, to actual conditions.

Simplifications are essential in formal modeling, as they are in all research, but we need to be cautious about the inferences we can draw about reality from the models. For example, assuming that all omitted variables have no effect on the results can be very useful in modeling. In many of the formal models of qualitative research that we present throughout this book, we do precisely this. Assumptions like this are not usually justified as a feature of the world; they are only offered as a convenient feature of our model of the world. The results, then, apply exactly to the situation in which these omitted variables are irrelevant and may or may not be similar to results in the real world. We do not have to check the assumption to work out the model and its implications, but it is essential that we check the assumption during empirical evaluation. The assumption need not be correct for the formal model to be useful. But we cannot take untested or unjustified theoretical assumptions and use them in constructing empirical research designs. Instead, we must generally supplement a formal theory with additional features to make it useful for empirical study.

A good formal model should be abstract so that the key features of the problem can be apparent and mathematical reasoning can be easily applied. Consider, then, a formal model of the effect of proportional representation on political party systems, which implies that proportional representation fragments party systems. The key causal variable is the type of electoral system—whether it is a proportional representation system with seats allocated to parties on the basis of their proportion of the vote or a single-member district system in which a single winner is elected in each district. The dependent variable is the number of political parties, often referred to as the degree of party-system fragmentation. The leading hypothesis is that electoral systems based on proportional representation generate more political parties than do district-based electoral systems. For the sake of simplicity, such a model might well include only variables measuring some essential features of the electoral system and the degree of party-system fragmentation. Such a model would generate only a hypothesis, not a conclusion, about the relationship between proportional representation and party-system fragmentation in the real world. Such a hypothesis would have to be tested through the use of qualitative or quantitative empirical methods.

However, even though an implication of this model is that proportional representation fragments political parties, and even though no other variables were used in the model, using only two variables in an empirical analysis would be foolish. A study that indicates that countries with proportional representation have more fragmented party systems would ignore the problem of endogeneity (section 5.4), since countries which establish electoral systems based on a proportional allocation of seats to the parties may well have done so because of their already existent fragmented party systems. Omitted variable bias would also be a problem since countries with deep racial, ethnic, or religious divisions are probably also likely to have fragmented party systems, and countries with divisions of these kinds are more likely to have proportional representation.

Thus, both of the requirements for omitted variable bias (section 5.2) seem to be met: the omitted variable is correlated both with the explanatory and the dependent variable, and any analysis ignoring the variable of social division would therefore produce biased inferences.

The point should be clear: formal models are extremely useful for clarifying our thinking and developing internally consistent theories. For many theories, especially complex, verbally stated theories, it may be that only a formal model is capable of revealing and correcting internal inconsistencies. At the same time, formal models are unlikely to provide the correct empirical model for empirical testing. They certainly do not enable us to avoid any of the empirical problems of scientific inference.

3.5.3 *Rule 3: Select Dependent Variables Carefully*

Of course, we should do everything in research carefully, but choosing variables, especially dependent variables, is a particularly important decision. We offer the following three suggestions (based on mistakes that occur all too frequently in the quantitative and qualitative literatures):

First, dependent variables should be dependent. A very common mistake is to choose a dependent variable which in fact causes changes in our

explanatory variables. We analyze the specific consequences of endogeneity and some ways to circumvent the problem in section 5.4, but we emphasize it here because the easiest way to avoid it is to choose explanatory variables that are clearly exogenous and dependent variables that are endogenous.

Second, *do not select observations based on the dependent variable so that the dependent variable is constant*. This, too, may seem a bit obvious, but scholars often choose observations in which the dependent variable does not vary at all (such as in the example discussed in section 4.3.1). Even if we do not deliberately design research so that the dependent variable is constant, it may turn out that way. But, as long as we have not predetermined that fact by our selection criteria, there is no problem. For example, suppose we select observations in two categories of an explanatory variable, and the dependent variable turns out to be constant across the two groups. This is merely a case where the estimated causal effect is zero.

Finally we should ~~choose a dependent variable that adequately the variation we wish to explain~~. Although this point seems obvious, it is actually quite subtle, as illustrated by Stanley Lieberson (1985:103):

> A simple gravitational exhibit at the Ontario Science Centre in Toronto inspires a heuristic example. In the exhibit, a coin and a feather are both released from the top of a vacuum tube and reach the bottom at virtually the same time. Since the vacuum is not a total one, presumably the coin reaches the bottom slightly ahead of the feather. At any rate, suppose we visualize a study in which a variety of objects is dropped without the benefit of such a strong control as a vacuum—just as would occur in nonexperimental social research. If social researchers find that the objects differ in the time that they take to reach the ground, typically they will want to know what characteristics determine these differences. Probably such characteristics of the objects as their density and shape will affect speed of the fall in a nonvacuum situation. If the social researcher is fortunate, such factors together will fully account for all of the differences among the objects in the velocity of their fall. If so, the social researcher will be very happy because all of the variation between objects will be accounted for. The investigator, applying standard social research thinking will conclude that there is a complete understanding of the phenomenon *because all differences among the objects under study have been accounted for*. Surely there must be something faulty with our procedures if we can approach such a problem without even considering gravity itself.

The investigator's procedures in this example would be faulty only if the variable of interest were gravity. If gravity were the explanatory variable we cared about, our experiment does not vary it (since the experiment takes place in only one location) and therefore tells us nothing about it. However, the experiment Lieberson describes would be of great interest if we sought to understand variations in the time it will take for different types of objects to hit the ground when they are dropped from the same height under different conditions of air pressure. Indeed, even if we knew all about gravity, this experiment would still yield valuable information. But if, as Lieberson assumes, we were really interested in an inference about the causal effect of gravity, we would need a dependent variable which varied over observations with differing degrees of gravitational attraction. Likewise, in social science, we must be careful to ensure that we are really interested in understanding our dependent variable, rather than the background factors that our research design holds constant.

Thus, we need the entire range of variation in the dependent variable to be a possible outcome of the experiment in order to obtain an unbiased estimate of the impact of the explanatory variables. Artificial limits on the range or values of the dependent variable produce what we define (in section 4.3) as selection bias. For instance, if we are interested in the conditions under which armed conflict breaks out, we cannot choose as observations only those instances where the result is armed conflict. Such a study might tell us a great deal about variations among observations of armed conflict (as the gravity experiment tells us about variations in speed of fall of various objects) but will not enable us to explore the sources of armed conflict. A better design if we want to understand the sources of armed conflict would be one that selected observations according to our explanatory variables and allowed the dependent variable the possibility of covering the full range from there being little or no threat of a conflict through threat situations to actual conflict.

3.5.4 *Rule 4: Maximize Concreteness*

Our fourth rule, which follows from our emphasis on falsifiability, consistency, and variation in the dependent variable is to maximize concreteness. ~~We should choose observable, rather than unobservable~~ concepts wherever possible. Abstract, unobserved concepts such as utility, culture, interests, motivations, identification, intelligence, or the national interest are often used in social science theories. They can play a useful role in theory *formulation*, but they can be a hindrance to empirical *evaluation* of theories and hypotheses unless they can be defined in a way such that they, or at least their implications, can be observed and measured. Explanations involving concepts such as culture or national interest or utility or motivation are suspect unless we can

measure the concept independently of the dependent variable that we are explaining. When such terms are used in explanations, it is too easy to use them in ways that are tautological or have no differentiating, observable implications. An act of an individual or a nation may be explained as resulting from a desire to maximize utility, to fulfill intentions, or to achieve the national interest. But the evidence that the act maximized utility or fulfilled intentions or achieved the national interest is the fact that the actor or the nation engaged in it. It is incumbent upon the researcher formulating the theory to specify clearly and precisely what observable implications of the theory would indicate its veracity and distinguish it from logical alternatives.

In no way do we mean to imply by this rule that concepts like intentions and motivations are unimportant. We only wish to recognize that the standard for explanation in any *empirical* science like ours must be *empirical* verification or falsification. Attempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and less successful than for many imperfectly conceived specific and concrete concepts. The more abstract our concepts, the less clear will be the observable consequences and the less amenable the theory will be to falsification.

Researchers often use the following strategy. They begin with an abstract concept of the sort listed above. They agree that it cannot be measured directly; therefore, they suggest specific indicators of the abstract concept that can be measured and use them in their explanations. The choice of the specific indicator of the more abstract concept is justified on the grounds that it is observable. Sometimes it is the *only thing* that is observable (for instance, it is the only phenomenon for which data are available or the only type of historical event for which records have been kept). This is a perfectly respectable, indeed usually necessary, aspect of empirical investigation.

Sometimes, however, it has an unfortunate side. Often the specific indicator is far from the original concept and has only an indirect and uncertain relationship to it. It may not be a valid indicator of the abstract concept at all. But, after a quick apology for the gap between the abstract concept and the specific indicator, the researcher labels the indicator with the abstract concept and proceeds onward as if he were measuring that concept directly. Unfortunately, such reification is common in social science work, perhaps more frequently in quantitative than in qualitative research, but all too common in both. For example, the researcher has figures on mail, trade, tourism and student exchanges and uses these to compile an index of "societal integration" in Europe. Or the researcher asks some survey questions as to whether

respondents are more concerned with the environment or making money and labels different respondents as "materialists" and "post-materialists." Or the researcher observes that federal agencies differ in the average length of employment of their workers and converts this into a measure of the "institutionalization" of the agencies.

We should be clear about what we mean here. The gap between concept and indicator is inevitable in much social science work. And we use general terms rather than specific ones for good reasons: they allow us to expand our frame of reference and the applicability of our theories. Thus we may talk of legislatures rather than of more narrowly defined legislative categories such as parliaments or specific institutions such as the German Bundestag. Or we may talk of "decision-making bodies" rather than legislatures when we want our theory to apply to an even wider range of institutions. (In the next section we, in fact, recommend this.) Science depends on such abstract classifications—or else we revert to summarizing historical detail. But our abstract and general terms must be connected to specific measurable concepts at some point to allow empirical testing. The fact of that connection—and the distance that must be traversed to make it—must always be kept in mind and made explicit. Furthermore, the choice of a high level of abstraction must have a real justification in terms of the theoretical problem at hand. It must help make the connection between the specific research at hand—in which the particular indicator is the main actor—and the more general problem. And it puts a burden on us to see that additional research using other specific indicators is carried on to bolster the assumption that our specific indicators really relate to some broader concept. The abstract terms used in the examples above—"societal integration," "post-materialism," and "institutionalization"—may be measured reasonably by the specific indicators cited. We do not deny that the leap from specific indicator to general abstract concept must be made—we have to make such a leap to carry on social science research. The leap must, however, be made with care, with justification, and with a constant "memory" of where the leap began.

Thus, we do not argue against abstraction. But we do argue for a language of social research that is as concrete and precise as possible. If we have no alternative to using unobservable constructs, as is usually the case in the social sciences, then we should at least choose ideas with *observable consequences*. For example, "intelligence" has never been directly observed but it is nevertheless a very useful concept. We have numerous tests and other ways to evaluate the implications of intelligence. On the other hand, if we have the choice between "the institu-

tionalization of the presidency" and "size of the White House staff," it is usually better to choose the latter. We may argue that the size of the White House staff is related to the general concept of the institutionalization of the presidency, but we ought not to reify the narrower concept as identical to the broader. And, if size of staff means institutionalization, we should be able to find other measures of institutionalization that respond to the same explanatory variables as does size of staff. Below, we shall discuss "maximizing leverage" by expanding our dependent variables.

Our call for concreteness extends, in general, to the words we use to describe our theory. If a reader has to spend a lot of time extracting the precise meanings of the theory, the theory is of less use. There should be as little controversy as possible over what we mean when we describe a theory. To help in this goal of specificity, even if we are not conducting empirical research ourselves, we should spend time explicitly considering the observable implications of the theory and even possible research projects we could conduct. The vaguer our language, the less chance we will be wrong—but the less chance our work will be at all useful. It is better to be wrong than vague.

In our view, eloquent writing—a scarce commodity in social science—should be encouraged (and savored) in presenting the rationale for a research project, arguing for its significance, and providing rich descriptions of events. Tedium never advanced any science. However, as soon as the subject becomes causal or descriptive inference, where we are interested in observations and generalizations that are expected to persist, we require concreteness and specificity in language and thought.[16]

[16] The rules governing the best questions to ask in interviews are almost the same as those used in designing explanations: Be as concrete as possible. We should not ask conservative, white Americans, "Are you racist?", rather, "Would you mind if your daughter married a black man?" We should not ask someone if he or she is knowledgeable about politics; we should ask for the names of the Secretary of State and Speaker of the House. In general and wherever possible, we must not ask an interviewee to do our work for us. It is best not to ask for estimates of causal effects; we must ask for measures of the explanatory and dependent variables, and estimate the causal effect ourselves. We must not ask for motivations, but rather for facts.

This rule is not meant to imply that we should never ask people why they did something. Indeed, asking about motivations is often a productive means of generating hypotheses. Self-reported motivations may also be a useful set of observable implications. However, the answer given must be interpreted as the interviewee's response to the researcher's question, not necessarily as the correct answer. If questions such as these are to be of use, we should design research so that a particular answer given (with whatever justifications, embellishments, lies, or selective memories we may encounter) is an observable implication.

3.5.5 *Rule 5: State Theories in as Encompassing Ways as Feasible*

Within the constraints of guaranteeing that the theory will be falsifiable and that we maximize concreteness, the theory should be formulated so that it explains as much of the world as possible. We realize that there is some tension between this fifth rule and our earlier injunction to be concrete. We can only say that both goals are important, though in many cases they may conflict, and we need to be sensitive to both in order to draw a balance.

For example, we must not present our theory as if it only applies to the German Bundestag when there is reason to believe that it might apply to all independent legislatures. We need not provide evidence for all implications of the theory in order to state it, so long as we provide a reasonable estimate of uncertainty that goes along with it. It may be that we have provided strong evidence in favor of the theory in the German Bundestag. Although we have no evidence that it works elsewhere, we have no evidence against it either. The broader reference is useful if we remain aware of the need to evaluate its applicability. Indeed, expressing it as a hypothetically broader reference may force us to think about the structural features of the theory that would make it apply or not to other independent legislatures. For example, would it apply to the U.S. Senate, where terms are staggered, to the New Hampshire Assembly, which is much larger relative to the number of constituents, or to the British House of Commons, in which party voting is much stronger? An important exercise is stating what we think are systematic features of the theory that make it applicable in different areas. We may learn that we were wrong, but that is considerably better than not having stated the theory with sufficient precision in the first place.

This rule might seem to conflict with Robert Merton's ([1949] 1968) preference for "theories of the middle-range," but even a cursory reading of Merton should indicate that this is not so. Merton was reacting to a tradition in sociology where "theories" such as Parson's "theory of action" were stated so broadly that they could not be falsified. In political science, Easton's "systems theory" (1965) is in this same tradition (see Eckstein 1975:90). As one example of the sort of criticism he was fond of making, Merton ([1949] 1968: 43) wrote, "So far as one can tell, the theory of role-sets is not inconsistent with such broad theoretical orientations as Marxist theory, functional analysis, social behaviorism, Sorokin's integral sociology, or Parson's theory of action." Merton is not critical of the theory of role-sets, which he called a middle-range theory, rather he is arguing against those "broad theoretical orienta-

tions," with which almost any more specific theory or empirical observation is consistent. Merton favors "middle-range" theories but we believe he would agree that theories should be stated as broadly as possible as long as they remain falsifiable and concrete. Stating theories as broadly as possible is, to return to a notion raised earlier, a way of maximizing leverage. If the theory is testable—and the danger of very broad theories is, of course, that they may be phrased in ways that are not testable—then the broader the better; that is, the broader, the greater the leverage.

CHAPTER 4

Determining What to Observe

UP TO THIS POINT, we have presented our view of the standards of scientific inference as they apply to both qualitative and quantitative research (chapter 1), defined descriptive inference (chapter 2), and clarified our notion of causality and causal inference (chapter 3). We now proceed to consider specific practical problems of qualitative research design. In this and the next two chapters, we will use many examples, both drawn from the literature and constructed hypothetically, to illustrate our points. This chapter focuses on how we should select cases, or observations, for our analysis. Much turns on these decisions, since poor case selection can vitiate even the most ingenious attempts, at a later stage, to make valid causal inferences. In chapter 5, we identify some major sources of bias and inefficiency that should be avoided, or at least understood, so we can adjust our estimates. Then in chapter 6, we develop some ideas for increasing the number of observations available to us, often already available within data we have collected. We thus pursue a theme introduced in chapter 1: we should seek to derive as many observable implications of our theories as possible and to test as many of these as are feasible.

In section 3.3.2, we discussed "conditional independence": the assumption that observations are chosen and values assigned to explanatory variables independently of the values taken by the dependent variables. Such independence is violated, for instance, if explanatory variables are chosen by rules that are correlated with the dependent variables or if dependent variables cause the explanatory variables. Randomness in selection of units and in assigning values to explanatory variables is a common procedure used by some quantitative researchers working with large numbers of observations to ensure that the conditional independence assumption is met. Statistical methods are then used to mitigate the Fundamental Problem of Causal Inference. Unfortunately, random selection and assignment have serious limitations in small-*n* research. If random selection and assignment are not appropriate strategies, we can seek to achieve unit homogeneity through the use of intentional selection of observations (as discussed in section 3.3.1). In a sense, intentional selection of observations is our "last line of defense" to achieve conditions for valid causal inference.