Designing Social Inquiry

SCIENTIFIC INFERENCE IN QUALITATIVE RESEARCH

Gary King
Robert O. Keohane
Sidney Verba

PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

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 (1983) 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 contributors? 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* ("emphathy: understanding the meaning of actions and interactions from the members' own points of view" [Eckstein 1975:81]). 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, be-

> tween 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 twiches 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 up-

[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.

setting 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.

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 meanings 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 (1968:510) 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 al-

ways 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 aimlesssness 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 indixes 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.[3]

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 1985:41). 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

[3] 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 "minitheory." 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. How-

ever, 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 clearer, 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 *sample 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 ex-

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

ample, 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.

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 nonsytematic 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_1 might be 23,562 people voting for Labor in district 1). 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.[10]

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 week 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 estimators 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:

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

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 commu-

[10] 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.

nity 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):

$$V(Y_i) = \sigma_i^2 \tag{2.3}$$

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.[11] 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.[12] However, for most purposes

[11] See King (1991b) for an elaboration of this distinction.

[12] 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" in an *estimator* from "substantive bias" in 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 $V(Y_{\text{typical}}) = \sigma^2$. 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 tradeoff 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, $\bar{Y} - 5/n$ is biased, but it is consistent because $5/n$ becomes zero as n approaches infinity.

some 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 an 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 d:

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

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:

$$\begin{aligned} E(d) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - 0.01\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - E(0.01) \\ &= \mu - 0.01 \end{aligned}$$

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

$$\begin{aligned} E(c) &= E\left(\frac{Y_1 + Y_2}{2}\right) \\ &= \frac{E(Y_1) + E(Y_2)}{2} \\ &= \frac{\mu + \mu}{2} \\ &= \mu \end{aligned}$$

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:

$$\begin{aligned} V(d) &= V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - 0.01\right) \\ &= V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - V(0.01) \\ &= \sigma^2/n \\ &= \sigma^2/650 \end{aligned}$$

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]

$$\begin{aligned} V(c) &= V\left(\frac{Y_1 + Y_2}{2}\right) \\ &= \frac{1}{4}[V(Y_1) + V(Y_2)] \\ &= \frac{1}{4}2\sigma^2 \\ &= \sigma^2/2 \end{aligned}$$

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.