6. The Promise and Perils of Statistics in International Relations

Bear F. Braumoeller and Anne E. Sartori

Students of international relations who are considering investing the time and effort necessary to learn statistics would be justified in first asking exactly what the statistical method is capable of doing. The answer can be summed up in a single sentence: it permits the researcher to draw inferences about reality based on the data at hand and the laws of probability. The ability to draw inferences is immensely helpful in assessing the extent to which the empirical expectations generated by theories are consistent with reality. It is also helpful in uncovering interesting questions or puzzles (e.g., Zinnes 1980) which occur when evidence is inconsistent with prior theoretical expectations.

In the sections that follow we attempt to highlight both the promise and the perils of the use of statistics in the pursuit of a better understanding of international political behavior. We do not aim to survey the vast literature in international relations that uses statistics; rather, we refer to particular works to illustrate our points. First, we discuss the advantages of the statistical method. These include the ability to aggregate information from large numbers of cases and to use the laws of probability to generalize well beyond those cases; the ability not just to describe associations among phenomena but to calculate the probabilities that such associations are the product of chance; and—as a direct result—the ability to gain a better understanding of the sources of human behavior in international affairs.

Despite our enthusiasm about applying statistical methods to international affairs in theory, we are cognizant of its shortcomings in practice. The shortcomings that concern us most are not the oft-stated worries of

many quantitative researchers—failures to satisfy regression assumptions, the need to ensure adequate levels of internal and external validity in our measures, and so on.[1] Such topics are covered at length in statistics and econometrics texts and need not be recited here. Rather, we are particularly concerned about a more fundamental problem: the widespread use of statistics with inadequate attention to the goal of *testing theories of international behavior.* In the following sections, we discuss two classes of shortcomings. The first pertains to the widespread neglect of the development of theory prior to the specification of a statistical model: statistical tests of theories usually have little worth unless the theories that they test are solid. The second concerns the process of deriving inferences from data, the finer points of which are too often neglected.

Advantages of the Statistical Method

One advantage of the statistical method is that it permits political scientists to aggregate information from a tremendous number of cases. This advantage is perhaps so obvious that its importance is often overlooked. To comprehend its magnitude we need only imagine trying to make sense of a thousand surveys of individual attitudes, beliefs, voting behavior, and so on, *without* the aid of statistics. The ability to extract even basic summary statistics from such a mass of data is immensely valuable: even something as unsophisticated as a sample mean—say, per capita GNP—conveys a wealth of information in compact and understandable form.

The ability to aggregate information is a potent stimulus for theorizing. Theory development often begins when a researcher uncovers an empirical puzzle that remains unexplained by prior theory. Such a puzzle leads to a search for an explanation, and eventually to new or better-developed theory. A puzzle can emerge from a single case, but the researcher often would like to know whether or not it indicates a widespread pattern of behavior. Only statistics can provide the answer to this question.[2] For example, statistical analyses indicate that a number of pairs of states (e.g., India and Pakistan) engage in a disproportionate number of wars (Goertz and Diehl 1992). The empirical discovery of this phenomenon, which the literature terms "enduring rivalry," has led to a number of attempts to explain the behavior of this set of dyads (e.g., Vasquez 1995; Bennett 1998; Diehl and Goertz 2000): what is it that makes states become rivals; why do rivals fight so often; and how do rivalries end?

The use of statistics also makes the terms of a given debate more explicit. Inference requires assumptions, whether implicit or explicit; statistics force scholars to be quite explicit about the nature of at least some assumptions. Transparency is valuable both because assumptions should be as clear as possible and because one can compensate for violated assumptions if they are understood.[3]

In addition to standards of inference, the use of statistics necessarily entails standards of evidence. Even the most scrupulous researcher can be hard-pressed to avoid selecting only the evidence that would support his or her theory. Here, too, standardization is an asset; the need for coding procedures forces the researcher to be explicit about criteria for measurement and mitigates the human tendency to notice only trends that are consistent with the theory under investigation. Quantification can be a considerable boon to both reliability and validity: in the former case, explicit tests of reliability can flag unacceptably noisy measures, while in the latter, details of the coding process make it clear what is and is not being measured.[4] For example, the Polity IV democracy index is an aid to scholars because the coding rules are specific and reliability can be calculated.

Statistical techniques also permit us to assess the claim that observed associations among variables are due to chance. Such assessments are critical to the testing of theory, and they are often very difficult to make. The statistical method can make the task almost trivially easy. For example, the extent to which any given Third World country votes with the United States in the UN will naturally vary from year to year; as a result, it can be difficult to determine whether an increase or decrease following a change in domestic political regime is an indicator of realignment or simply the product of random fluctuation. Absent the ability to assess the odds that such fluctuations are due to chance, analysts could argue endlessly over their substantive significance.[5] Hagan (1989) addresses this question by testing to determine whether mean voting scores under a given regime differ significantly from mean voting scores under its successor; in about half of the eighty-seven cases he examines, he finds that random fluctuation is a highly improbable ($p < 0.05$) explanation for the difference in voting patterns across regimes. Although statistical testing does not answer the question with perfect certainty, it gives far more precise answers than could otherwise be obtained. In so doing it dramatically narrows potential areas of disagreement.

By answering the question of whether observed associations are the

plausible result of chance, the statistical method also permits us to draw causal inferences. Using statistics, one can investigate ancillary associations implied by a posited causal process and assess the probability that these associations are due to chance.[6] Because international relations scholars constantly seek to understand why actors behave as they do, this ability is perhaps the method's greatest contribution to the discipline. To continue the preceding example, one might wonder not just whether a given country's UN votes coincide to a greater or lesser degree with those of the United States but why. One obvious possibility would be that American foreign aid, to put it crudely, buys votes: American leaders use foreign assistance to induce cooperation. If this is the case, increases in American aid should be followed by an increased coincidence of votes in the UN on issues considered to be important by the United States. Wang (1999) tests this hypothesis by examining the voting records of sixty-five developing countries from 1984 to 1993 and finds that an increase in American foreign aid generally precedes an increase in voting alignment; moreover, the positive relationship between the two is very unlikely (again, $p < 0.05$) to be the result of chance. Absent statistical techniques, the effects of American aid could be debated one anecdote at a time without any conclusion in sight. Even the most meticulous case selection and comparison could never produce such precise results.

A final strength of the statistical method is the fact that it conveys the ability to test two explanations against one another with remarkable precision. For example, while tests of realist and of domestic-political explanations of conflict typically limit themselves to ruling out chance associations, Clarke (2001) tests realism against two domestic-political explanations and finds that realism "either does as well as the rival or better than the rival" theory (28).[7]

Potential Pitfalls

Despite the power of the statistical method, statistical evidence sometimes is far from persuasive. This failure typically stems from misunderstanding or ignorance of the underlying purpose of the method. It is critical for users of statistical techniques to realize that statistical models are models of human behavior and that, as a result, the assumptions that underlie them are substantively nontrivial. Common assumptions—such as simple additivity among variables—constitute theoretical assertions about how reality

works, and the prevalence of unreflective assumptions in statistical research has contributed to a widespread perception among formal modelers that statistical research is theoretically unsophisticated (see, e.g., Morton 1999, 3, 16–24 and passim). It need not be. In the following sections, we focus upon two sets of common errors, which we call errors of specification and errors of inference.

Errors of Specification

In order to convey useful information about the world, statistical tests must relate meaningfully to the causal mechanisms implied by the theories that they purport to evaluate. Failure to do so constitutes an error of specification. Three such errors are, in our view, of paramount importance. First, empirical researchers often spend too much effort calculating correlations with little or no attention to theory. Second, theory itself often is weak and difficult to test because it is too imprecise or too shallow. Finally, empirical researchers often impose a statistical model on the theory instead of crafting a model to test the theory. Under any of these circumstances, even the most sophisticated statistical techniques are futile.

The large literature on the democratic peace illustrates both the benefits of using statistics and the pitfalls of doing so with too little theory. Several studies demonstrated a relationship between democracy and peace and explained the relationship between the two by offering two theories, one based on liberal norms (e.g., Doyle 1986; Russett 1993) and the other based on the domestic political structure of democratic states (e.g., Rummel 1979; Morgan and Campbell 1991; Bueno de Mesquita and Lalman 1992).[8] Debate over whether or not there was, in Gertrude Stein's words, a "there there" ensued, with authors arguing both pro and con.[9] Researchers developed and tested additional hypotheses based on the generic notion of cooperation among democracies, yielding additional empirical insights.[10] Occasionally, they derived implications from the theories that would allow them to be tested against each other.[11] The result was an unusually comprehensive corpus of literature describing the behavior of democratic states.

The development of theory, however, proceeded at a much slower pace than the proliferation of statistical associations: with the exception of David Lake's (1992) article, which offered an explanation based on the relative rent-seeking behavior of democratic and nondemocratic states, vari-

ants of structural and normative theories dominated the study of democracies and peace for well over a decade. Recently, three additional contenders—the informational theory forwarded by Kenneth Schultz (1999), the institutional variant laid out by Bueno de Mesquita et al. (1999), and the evolutionary learning approach of Cederman (2001)—have rekindled interest in the democratic peace phenomenon. They have also raised an issue that may have widespread implications for the studies that preceded them: the question of what the independent variable should be. Although both the ability to generate audience costs and the existence of a broad constituency are correlated with democracy, for example, the correlations are not equal to one.[12] The development of new theory has brought to light the possibility that scores of books and articles have based their conclusions on measurements of the wrong causal variable.[13]

Unfortunately, simply paying attention to theory is not enough: many international relations theories are too imprecise or shallow to be subjected to tests against other theories. When a theory is imprecise, a wide range of relationships between independent and dependent variables is consistent with the theory. In the extreme, an imprecise theory may be entirely unfalsifiable. For example, as Lake and Powell (1999, 23) note, Waltzian neorealism suggests that states respond in one of two contradictory ways when confronted with a powerful adversary in a multipolar system: they either balance against an aggressive state or bandwagon with that state (Waltz 1979). If we see states balancing (or bandwagoning), is this behavior consistent with realism? Theoretically, the answer is yes, so that neither finding falsifies the theory. Similarly, the hypothesis that bipolarity is associated with the prevalence of peace is vague and untestable; only when polarity is carefully defined (see Wagner 1993) is this hypothesis falsifiable. In some cases, making a theory precise is merely a question of operationalizing variables. In others, as with polarity, lack of precision corresponds to inadequate definitions and is a weakness in the theory itself.

When a theory is shallow, it has testable implications, but only one or two. It may explain a broad range of phenomena, but it fail to explain even a few details of any one type of event. For example, scholars often use the Prisoners' Dilemma game to model international relations (see Snidal, chap. 10, fig. 2, this vol.). The Prisoners' Dilemma is a striking analogy, and it has been a useful step in theory development. The insights scholars glean from it are applicable to a broad range of problems.[14] Unfortunately, the trade-off in this case is depth.[15]

Because of its simplicity, a two-by-two game yields few implications about any specific substantive situation. Researchers usually derive implications from game-theoretic models by performing "comparative statics": they vary some feature of a model, usually the players' payoffs, and determine how the logical implications of the game change as a result. However, a twenty-two game has few elements that can be varied in this way; it portrays only one decision by each actor and four possible outcomes.

For example, a researcher might use the Prisoners' Dilemma game to investigate whether or not states' possession of nuclear weapons affects the probability of war. He or she might assume that states' payoffs differ depending upon whether or not both states have nuclear weapons. Perhaps, if both states possess nuclear weapons, mutual non-cooperation represents nuclear war and is each state's least-preferred outcome. (With these modified preferences, the two-by-two game is no longer a Prisoners' Dilemma.) If so, then one testable implication of the game is that the states are more likely to cooperate (less likely to go to war) if they have nuclear weapons than if they do not (Snidal, chap. 10).[16]

However, the model has few other implications since it has few components besides these payoffs to vary. Thus, the model cannot be tested against those alternative theories that also imply that jointly nuclear dyads are more peaceful. It can be tested only against the null hypothesis that the possession of nuclear weapons does not affect the probability of war. Shallow theory requires attention to theory development first. Statistical tests can do little to increase our understanding of the situation and must come later, when their empirical force can be brought to bear at the point of greatest theoretical leverage.

The final specification problem that we will discuss is inattention to the causal process or processes that generated the data. Correct specification of functional form requires close attention to theory, and widespread reliance on canned econometric techniques still tempts users to rely on statistical convention rather than theoretical logic. The form of a statistical test should be derived from the form of the theory, not vice versa. As a consequence, the ability to find a statistical test suitable for one's theory is crucial; the ability to design such a test when one does not exist would be ideal. Toward these ends we cannot overemphasize the importance of wide and deep familiarity with both mathematics and statistics. The old adage about hammers and nails is appropriate: when the only tool you have is regression, the world has a surprising tendency to look linear and additive.

Possession of a larger and more diverse methodological tool kit alleviates this problem to some degree, of course, but being up to date on the latest advances in maximum likelihood, Markov chain Monte Carlo, or Hilbert space methods will be of little use if the researcher gives insufficient attention to the choice of an estimator that is appropriate to the theory at hand and the causal process that generated the data. Another, equally obvious lesson is equally critical: *think about the theory.*

Attention to theory is not the only way to guard against misspecification, however. At times the data can suggest a markedly different functional form, perhaps one consistent with a different theory altogether, and an inattentive researcher can easily miss such a signal. As Anscombe (1973) pointed out, statistical models can be imposed on the data and can fit the data fairly well, even if their functional forms grossly misrepresent the relationship of interest. Table 1 and figure 1 demonstrate this point graphically: The regression results in table 1 suggest a significant linear relationship between *Y* and *X,* but they could have been generated by any one of the four data sets graphed in figure 1. In an era in which data sets can be obtained in moments and regressions run even more quickly, this underscores a fundamental lesson: *look at the data.*

The eyeball test is part of the intricate interplay between theory and data that occurs in skillful application of the scientific method. By thinking about the variables, the researcher often can anticipate the functional form that he or she sees in the data. For example, the relationship between the balance of forces and the probability that a state starts a war probably is not linear; moving from a 2-to-1 balance to a 3-to-1 balance probably has more of an effect than moving from a 100-to-1 to 101-to-1 balance. Thus, one might posit that the log of the military balance captures the hypothesized relationship better than the military balance itself. Never-

TABLE 1. Relationship between *Y* and *X*

	$n = 11$			$R^2 = 0.667$		
	$F(1,9) = 17.98$			Adj. $R^2 = 0.629$		
	Prob > $F = 0.002$			Root MSE = 1.237		
Y	**Coef.**	**S.E.**	***t***	**P > |*t*|**	**95% Conf. Interval**	
X	0.500	0.118	4.24	0.002	0.233	0.767
Constant	3.000	1.125	2.67	0.026	0.455	5.545

Source: From Anscombe (1973)

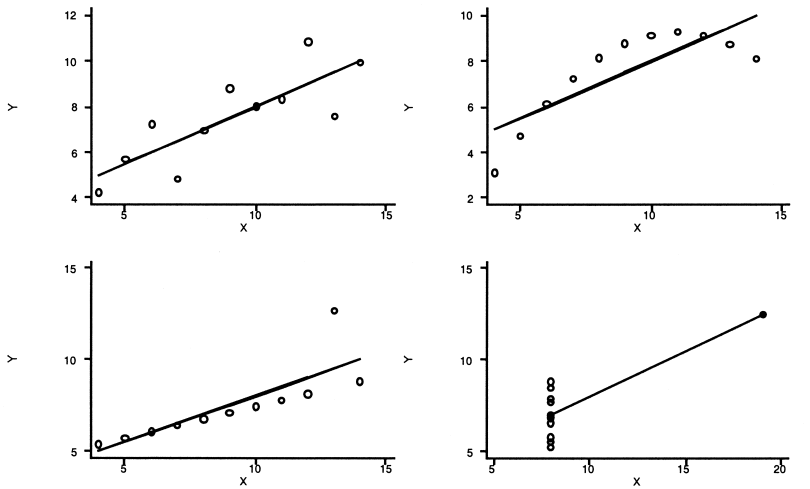

Fig. 1. Four data sets consistent with results in table 1

theless, in theorizing, one may miss important nonlinearities. A look at the data can provide a useful reminder that inadequate attention has been given to functional form.

The overall message is simple: statistical tests should correspond to theory that is well-developed. Toward this end, the use of formal theory can be especially useful in that it prompts careful thinking and forces the researcher to specify many important aspects of the situation under study. For example, a game-theoretic model requires the basic elements of theory: assumptions about which actors are important to the outcome being explained, what they care about and how strongly (the utility that they receive if various outcomes occur), the choices that are available to them, the order in which they can make choices, and the relationship of choices to outcomes. Game-theoretic models also must specify the information available to actors and their beliefs about any information about which they are uncertain. Varying any of these raw elements of the model produces implications about the relationships between the element (independent variable) and the action taken or outcomes (dependent variable).[17] Without any of the raw elements, the model cannot be solved. Thus, the researcher cannot deduce implications without specifying the required

assumptions. The statistical method does not force the user to provide, or even to think very hard about, any of these important elements of theory, nor does formal theory force the user to think about some of the intricacies of empirical testing or to say anything about the real world. Because each provides what the other lacks, the combination of the two methods constitutes a potent tool for inquiry.

Nevertheless, formalization is not a panacea for the problem of incomplete theory. The Prisoners' Dilemma model reveals three steps that the researcher must take in order to create testable (falsifiable) theory. First, the empirical researcher must specify precisely the real-world correspondents of the raw elements of the model (whether the model is formal or verbal). In the Prisoners' Dilemma example (Snidal, chap. 10, this vol.), the researcher must start by specifying what "cooperate" and "not cooperate" mean in the substantive problem at hand—possibly "no new arms" or "increase arms." He or she also must specify the real-world factors that constitute utility for a given actor. What factors determine how much the states benefit from a state of mutual disarmament? How much do they like or dislike the other outcomes? Like the Prisoners' Dilemma, many models can be used to explain several real-world situations. Nevertheless, research would progress more rapidly if game theorists in particular were more specific about some of the possible real-world referents of their models.

Second, while simplicity is a virtue, the model must be complex enough to capture an explanation of an interesting phenomenon. We emphatically agree with the edict often attributed to Einstein that "everything should be made as simple as possible, but not simpler."

Third, the researcher often must resolve indeterminacy in a model before turning to empirical tests. Some game-theoretic models imply that a large number of outcomes are logically possible (corresponding to different equilibria). This indeterminacy does not make the models useless: they still narrow down the set of behaviors expected in a given situation. However, it does raise questions for empirical testing. For example, as Duncan Snidal discusses (chap. 10, this vol.), if the Prisoners' Dilemma is played repeatedly and players care sufficiently about the future, then many types of cooperative outcomes are possible (and mutual defection also is possible). Which outcome should the researcher expect to find in the real world? Game theory contains some tools for narrowing down the set of likely outcomes (called "equilibrium refinements"). However, multiple equilibria often remain, and some refinements seem worse than arbitrary.

Two equilibria of the same game can encompass very different substantive stories about the players' interactions. For example, some equilibria of repeated games specify that players forever punish those who decide not to act in a certain way. When a game-theoretic model leads to multiple equilibria, our preference is to consider each as its own explanation, with its own set of empirical implications.[18] If the results of statistical tests are inconsistent with the implications of an equilibrium, then that equilibrium is ruled out as an explanation for the outcome under investigation. Of course, researchers similarly can test different versions of the same, indeterminate verbal theory. For example, they can test a multiplicity of realisms. As with game-theoretic models, researchers should be up-front about the indeterminacy of the general model and about the specifics of the version that they are testing.

Game-theoretic modeling does not do away with the need to think about functional form and the nature of the error term. Researchers are increasingly considering how best to test the implications of game-theoretic models,[19] and international relations research is making progress on this front (see Signorino 1999a, 1999b; Smith 1999; Lewis and Schultz 2003; Sartori 2003). However, much work remains to be done. One thorny question is the extent to which factors outside the formal model (which is always quite simple), but thought to be theoretically important, should be considered in the statistical tests. For example, taken literally, a formal model may imply an unusual error structure (Signorino 1999a, 1999b). However, models are simplifications, and the error structure that comes literally from the model may not be the theorist's true best guess about the error in the underlying data-generating process. As the work on testing formal models progresses, it is our hope that researchers will continue to pay attention to the data as well as to theory. While the game-theoretic model may imply particular assumptions about the functional form and/or distribution of the error term, it is important to think about and look at the data before carrying these assumptions to the statistical model.

Errors of Inference

The two classes of problems that we have just discussed limit the extent to which statistical tests accurately assess the implications of a theory. A final set—not, we should emphasize, one that is unique to statistical methods—concerns the extent that tests of a given theory reveal information about

reality. This problem is a marked tendency to ignore some of the thornier problems involved in integrating data into larger-scale explanations. In particular, the complexity of the role that data play in the broader enterprise of theory testing is rarely appreciated. To put it more bluntly, statistics can take the place of thinking.

The first way in which statistics can do so is via the blind application of statistical significance to judge the importance of a variable. Although the notion of statistical significance is immensely useful, its abuse can lead to a multitude of sins. There is a persistent tendency to focus on statistical significance (the probability that an observed relationship between X and Y occurred by chance) without paying attention to substantive significance (the magnitude of the relationship between changes in X and changes in Y).

A data set with 50,000 observations, for example, permits us to uncover even the most minute relationships among variables and demonstrate that they were unlikely to have occurred by chance. Such relationships may, however, provide only very weak support for the theory under consideration. For example, a novice statistician who ran the analysis reported in table 2 might enthusiastically report very strong findings—a relationship between X and Y that is significant at the $p < 0.01$ level!—without ever realizing, as the data cloud in figure 2 makes clear, that the substantive relationship between the two is virtually nil.[20]

The relationship between the magnitude of a coefficient and substantive significance depends upon the problem at hand. There is no good quantitative rule for determining substantive significance. For example, assume that a researcher found that joint democracy decreased the probability of war from 0.03 to 0.001. One might be tempted to see this as an insubstantial decrease of 2.9 percentage points in the probability of the occurrence of war. However, given the extreme rarity of war, that seemingly minor decrease would imply that jointly democratic dyads experienced one-thirtieth as much war as other dyads. In our opinion, this particular result would be extremely substantively important, because of its implications for both theory and policy. Of course, what counts as substantively significant depends on the substance.

Political methodologists have succeeded in killing off widespread abuse of the R^2 coefficient (see Achen 1977) by distinguishing between degree of correlation and substantive significance, but this subtler form of confusion remains. The only good news is that, despite its tenacity, this tendency is at least decreasing. A survey of 211 articles on international relations from

the past decade of some of the field's top journals[21] revealed that, prior to 1996, only 16.4 percent of the quantitative articles discussed substantive significance, but after that point 38.8 percent contained such discussions.

Moreover, much of the field seems to forget that the choice of a significance level for rejecting the null hypothesis is arbitrary. A better way to judge the certainty of one's results when the baseline is the null is simply to calculate the probability that an observed result is nonzero due to chance. Finding that this probability is 6 percent rather than 5 percent should decrease one's confidence in a study's finding only trivially. Unfortunately, researchers and journals often apply the "5 percent rule" and relegate such findings to the trash can.

Finally, levels of statistical significance are based on the assumption that a single test has been carried out on a single data set. Running multiple tests, or running the same set on different data sets, invalidates this assumption, and significance levels are therefore incorrect. Mock and Weisberg (1992) provide an amusing example of this point by examining the *Washington Post*'s assertion, based on data from the 1985–87 General Social Survey (GSS), that there is a relationship between an individual's partisanship and his or her zodiac sign. In fact, they demonstrate that such a relationship exists and is significant at the $p < 0.10$ level.[22] They then expand the sample to eleven separate years and demonstrate that there is only a significant relationship between sign and partisanship in one of them (1985). The probability of finding at least one significant relationship in eleven attempts, as they point out, is 0.69: far from being surprising, a result like the 1985 one is precisely what should be expected due to chance variation.

Few political scientists who utilize the statistical method would argue with the preceding example; even fewer, unfortunately, are mindful of its implications for the researcher who runs eleven regressions and finds one

TABLE 2. A Significant Regression Coefficient with 50,000 Observations

	$N = 50{,}000$ $F(1,9) = 17.98$ Prob > $F = 0.002$			$R^2 = 0.0002$ Adj. $R^2 = 0.0001$ Root MSE = 1.003		
Y	Coef.	S.E.	*t*	P > |*t*|	95% Conf. Interval	
X	0.013	0.004	2.85	0.004	0.004	0.022
Constant	1.994	0.004	444.6	0.000	1.985	2.003

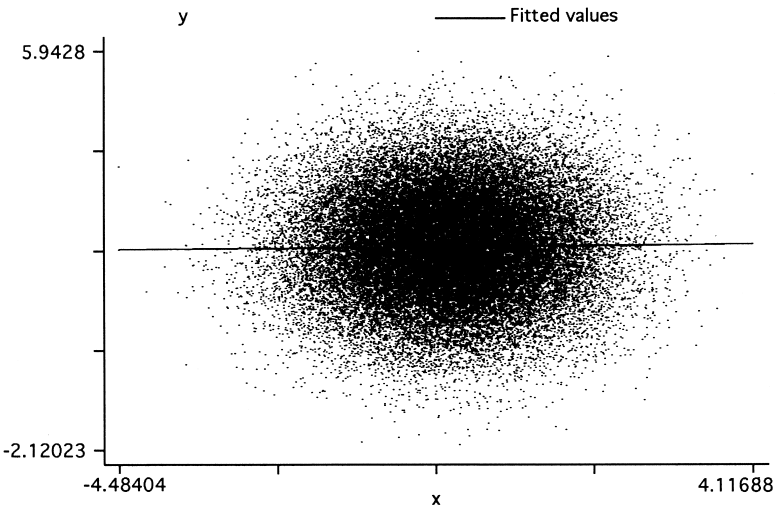

Fig. 2. Data summarized in table 2

significant relationship. Most researchers can point to one or two colleagues whom they suspect of mining data sets behind closed doors until significant results appear.[23] The variables rarely are zodiac signs, but the associations uncovered in this fashion are no less silly. Worse, publication bias is pervasive: nonresults typically do not make it to print (of the 211 articles in the previous survey, just 10, or 4.7 percent, reported only null results), and as a result insignificant relationships may be discarded or ignored until a significant one happens along.

The second way in which statistics take the place of thinking is that researchers simply accept or reject a theory based upon an assessment of how likely certain variables are to have nonzero effects.[24] The implications of a nonzero coefficient for the status of a theory are not always clear, and practitioners typically pay far too little attention to this rather critical issue. To those who believe along with Lakatos (1970) that theory A should be retained until superior theory B is found, simply accepting or rejecting theories seriatim based on whether or not variables have nonzero effects can constitute a sin of omission. Lakatos asserts that a theory should be retained despite empirical anomalies until a better theory can be found. If one is a Lakatosian, therefore, the ultimate way to assess a theory's performance is

to compare its success to that of another theory; this sometimes, but not always, can be accomplished by determining that particular variables that a theory points to have nonzero effects. To those who take a more Bayesian view of comparative theory testing, the degree to which various theories are believed to be true depends not on the results of a single statistical analysis but rather on the cumulation of results over time. Either way, it makes no sense simply to look at a parameter and its associated standard error and either accept or reject a theory based on their values.

However, investigating the match between a particular theory and data is often a useful exercise during what may be a long period of theory development. Most theories have multiple implications that can be taken as working hypotheses. Determining how likely variables are to have the signs that the theory implies provides useful information for refinement of the theory. In most cases, the data are consistent with some of the theory's implications and inconsistent with others. The researcher refines the theory, taking into account the results of the data analysis, and tests the new version by examining the new set of implications using a new data set. At the same time, empirical regularities uncovered during this period can give rise to alternative explanations that can also be developed and (ultimately) tested. While the researcher can compare the relative success of two or more theories during the early stages of theory development, such an exercise can also be counterproductive: it can distract the researcher from the many important issues involved in carefully testing the theory at hand.

When researchers do compare the relative usefulness of two or more theories, they often pay insufficient attention to how this should be done. Theories are generally assumed to be competing rather than potentially complementary parts of a larger theory, though there is no particular reason for this to be the case. Moreover, even if they are competing explanations, it is not at all clear that the way to compare them is to include variables representing each in an additive statistical equation. Doing so, though it comports with standard statistical practice, assumes that their effects cumulate in an additive fashion, which is probably not a reasonable representation of the either-or logic of evaluating competing theories.[25]

Finally, attempts to compare competing theories often result in a sin of commission—a decision to throw every plausible causal variable into the regression equation. Adding large numbers of variables often takes the place of careful investigation of the effect of the few variables truly relevant to the theory (Achen 2002). Moreover, if the variables that the competing

theory suggests are correlated in the sample with the variables of primary interest, then including these "control" variables can lead to incorrect conclusions about the primary theory being tested. In the absence of formal theory, Achen (2002) suggests "A Rule of Three" (ART): no more than three independent variables in a statistical specification. While informed opinion will surely differ regarding exactly how many independent variables should be permitted in a given equation, we agree that "garbage can" models—those with many independent variables and weak or absent microfoundations—represent a threat to inference that is currently underappreciated.

In short, it is often asserted or implied that theories have been proven correct by a successful rejection of the null hypothesis despite the inherent difficulty (some would say impossibility) of gauging precisely *how much* support for a theory is implied by support for a hypothesis that is consistent with it.[26] Here, we must confess, it is often far easier to criticize than to propose solutions, but the absence of solutions has become dangerously comfortable.

Even if researchers are meticulous in avoiding all of the pitfalls described earlier, they are typically unaware of a final threat to inference: simple computer error. In a series of articles, Bruce McCullough has endeavored to assess the reliability of commonly used econometric software,[27] and Micah Altman and Michael P. McDonald (2001) have extended these analyses to include the software packages most frequently used by political scientists. The results are the stuff of nightmares. One respected software package produced *t*-statistics that were half of the correct value when performing maximum likelihood analysis; another produced incorrect regression results when the names of the variables were too long. Few software packages were deemed entirely reliable for even fairly straightforward tasks. Therefore, when possible, it seems advisable to attempt to replicate findings using a different statistics package to avoid the possibility that important findings (or nonfindings) are simply artifacts of a bug in a particular statistics package.

So Why Bother?

We have stressed that statistical analyses are just one step in the scientific method of the study of international relations. While statistics can and should be used to generate stylized facts, the most common underlying

goal of research that uses statistics is to test and evaluate theories of international phenomena. Unfortunately, much research strays far from this goal in practice because the researcher fails to specify the theory carefully before testing it, because the statistical model conforms poorly to the theory, or because the researcher uses statistical "tests" without concern for their underlying meaning or relation to the theory.

Given the preceding discussion, students of international relations may wonder whether the expenditure in time and effort to learn statistical methods is worth the payoff. Our answer is an immediate yes. It is important not to make the best the enemy of the good: our critiques here are of ways in which international relations researchers often use the method rather than of the method itself. While the statistical method is of little value without theory, so, too, is theory insignificant without empirical tests. Absent empirical tests, we might work forever developing fundamentally incorrect theories.

The statistical method conveys tremendous advantages to the scholar wishing to test explanations of international events. It permits generalization, compels specificity, and conveys information with unparalleled precision. As recent issues of *Political Analysis* and the growing body of working papers amassed at the Society for Political Methodology website attest, increasingly sophisticated statistical methods are rapidly improving our ability to extract information from data, and the amount of data available to us continues to increase. In short, statistics provide a way of evaluating our understanding of the world that is simply unavailable via other means.

Recommended Readings

Statistical texts roughly in order of increasing difficulty

Gonick, L., and W. Smith. 1993. *The Cartoon Guide to Statistics.* New York: Harper Perennial. For students who find the prospect of mathematics horrifying, this book provides a remarkably gentle introduction up to the level of regression analysis.

Achen, C. H. 1982. *Interpreting and Using Regression.* Newbury Park, CA: Sage. This book provides invaluable advice to the student wishing to use regression in a thoughtful manner.

King, G. 1989. *Unifying Political Methodology.* Cambridge: Cambridge University Press. Reprint, Ann Arbor: University of Michigan Press, 1998. This book pro-

vides an introduction to maximum-likelihood estimation, which forms the basis of many current statistical models in political science.

Greene, W. H. 1993. *Econometric Analysis.* New York: Macmillan. This book covers many of the key topics of statistical analyses at an intermediate level.

Morton, R. B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science.* Cambridge: Cambridge University Press. A useful book for students wishing to pursue the idea of testing formal models.

Notes

1. This chapter assumes that the reader has at least an introductory knowledge of statistics. Those readers who do not are encouraged to see the recommended readings at the end of the chapter for definitions of terms. For definitions of external and internal validity, see Campbell and Stanley (1963).

2. In common terminology, statistical analyses can lead to the discovery of "empirical regularities" that could be explained by theory.

3. See, e.g., Kennedy (1998).

4. Reliability and validity assessment are often covered in passing in statistics books; for more specific treatments see Carmines and Zeller (1979) and Litwin (1995). Few international relations scholars assess reliability or validity, a fact that is quite striking given the manifest threats to both presented by their data.

5. We differ here from Clarke (2001), who argues that chance always is an uninteresting alternative explanation.

6. Lest we be misunderstood: correlation should never be equated with causation. Nevertheless, correlation provides valuable evidence in assessing claims of causation, as the following example demonstrates.

7. While Clarke's study is of nonnested models, researchers can compare nested models using simple, well-known techniques such as *F*-tests.

8. The literature on the democratic peace is vast, and we do not attempt to review all relevant works here. For further summary of the normative and structural theories, see Russett (1993).

9. See, e.g., Farber and Gowa (1995), Layne (1994), Spiro (1994), and Russett (1995). For a detailed case-by-case assessment of conflicts deemed dangerous to the finding, see Ray (1995).

10. Examples abound. See, for example, Dixon (1994) on democracy and the settlement of conflict, Simon and Gartzke (1996) on democracy and alliance, and Maoz and Russett (1993) on democracy and both involvement in and escalation of militarized interstate disputes (MIDs).

11. For an attempt to do precisely this, as well as an elaboration of one of the present authors' views on the subject, see Braumoeller (1997). The reader would

be justified in inferring that we can claim only limited impartiality on this point (limited, that is, to the other author).

12. For example, if relative size of constituency is the driving force behind the democratic peace, the Nineteenth Amendment produced a dramatic effect on the causal variable of interest. By most measures, however, the United States is not considered to have been half as democratic before women were allowed to vote as it was subsequently.

13. This possibility existed even in the case of structural and normative theories (Braumoeller 1997, fn. 7), but norms and structure are arguably more closely related to democracy. Nevertheless, Morgan and Campbell (1991) make this point with regard to the structural-constraints school and attempt to determine whether constrained states are more peaceful than others. Their results are discouraging for structuralists but quite encouraging for proponents of empirically informed theoretical progress.

14. See, for example, Oye (1986).

15. Some readers may argue that depth is about details, that taken to an extreme, our argument suggests that political scientists should examine the details of individual cases rather than develop theory. We are decidedly in favor of developing theory.

16. Technically, the game implies that mutual noncooperation never will occur in this situation. Usually, researchers translate the deterministic implications of game-theoretic models into probabilistic hypotheses about the world. In varying the payoffs so as to generate a testable implication of the Prisoners' Dilemma, one is comparing outcomes when the two-by-two game is a Prisoners' Dilemma to one in which the game has some other name; however, most of the structure of the situation remains the same.

17. We discussed the most common method of deriving implications, varying the payoffs, earlier in the chapter.

18. See Sartori (2004) for an example. Of course, the reasonableness of considering equilibria as competing explanations depends upon the model.

19. It is particularly encouraging that the National Science Foundation has a new program, "Empirical Implications of Theoretical Models," designed to encourage advances on this subject.

20. The data were simulated: $Y = Xb + e$, $X \sim \mathrm{N}(0,1)$, $e \sim \mathrm{N}(0,1)$, $b = 0.01$.

21. The *American Political Science Review, American Journal of Political Science, International Studies Quarterly, International Security, International Organization,* and *World Politics* were examined; we are grateful to Doug Stinnett for his careful research assistance.

22. Libras are most likely (30.1 percent) to be Republicans, while those born under the sign of Aquarius are most likely (49 percent) to be Democrats.

23. One of the authors was horrified when, at a recent conference, a speaker