Thinking Through Statistics

John Levi Martin

The University of Chicago Press Chicago and London

Contents

* 1 *
Introduction

Map: I'm going to start by identifying a meta-problem: most of what we learn in "statistics" class doesn't solve our actual problems, which have to do with the fact that we don't know what the true model is—not that we don't know how best to fit it. This book can help with that—but first, we need to understand how we can use statistics to learn about the social world. I will draw on pragmatism—and falsificationism—to sketch out what I think is the most plausible justification for statistical practice.

Statistics and the Social Sciences

What Is Wrong with Statistics

Most of statistics is irrelevant for us. What we need are methods to help us adjudicate between substantively different claims about the world. In a very few cases, refining the estimates from one model, or from one class of models, is relevant to that undertaking. In most cases, it isn't. Here's an analogy: there's a lot of criticism of medical science for using up a lot of resources (and a lot of monkeys and rabbits) trying to do something we know it can't do—make us live forever. Why do researchers concentrate their attention on this impossible project, when there are so many more substantively important ones? I don't deny that this might be where the money is, but still, there are all sorts of interesting biochemical questions in how you keep a ninety-nine-year-old millionaire spry. But if you look worldwide, and not only where the "effective demand" is, you note that the major medical problems, in contrast, are simple. They're things like nutrition, exercise, environmental hazards, things we've known about for years. But those things, simple though they are, are *difficult* to solve in practice. It's a lot more fun to concentrate on complex problems for which we can imagine a magic bullet.

So too with statistical work. Almost all of the discipline of statistics is about getting the absolutely best estimates of parameters from *true*

models (which I'll call "bestimates"). Statisticians will always admit that they consider their job only this—to figure out how to estimate parameters given that we already know the most important things about the world, namely the model we should be using. (Yes, there is also work on model selection that I'll get to later, and work on diagnostics for having the wrong model that I won't be able to discuss.) Unfortunately, usually, if we knew the right model, we wouldn't bother doing the statistics. The problem that we have isn't getting the bestimates of parameters from true models, it's about not having model results mislead us. Because what we need to do is to propose ideas about the social world, and then have the world be able to tell us that we're wrong . . . and having it do this more often when we *are* wrong than when we aren't.

How do we do this? At a few points in this book, I'll use a metaphor of carpentry. To get truth from data is a craft, and you need to learn your craft. And one part of this is knowing when *not* to get fancy. If you were writing a book on how to make a chair, you wouldn't tell someone to start right in after sawing up pieces of wood with 280 grit, extra fine, sandpaper. You'd tell them to first use a rasp, then 80 grit, then 120, then 180, then 220, and so on. But most of our statistics books are pushing you right to the 280. If you've got your piece in that kind of shape, be my guest. But if you're staring at a pile of lumber, read on.

Many readers will object that it simply isn't true that statisticians always assume that you have *the* right model. In fact, much of the excitement right now involves adopting methods for classes of models, some of which don't even require that the true model be in the set you are examining (Burnham and Anderson 2004: 276). These approaches can be used to select a best model from a set, or to come up with a better estimate of a parameter *across* models, or to get a better estimate of parameter uncertainty given our model uncertainty. In sociology, this is going to be associated with Bayesian statistics, although there are also related information-theoretic approaches. The Bayesian notion starts from the idea that we are thinking about a range of models, and attempting to compare a posteriori to a priori probability distributions—before and after we look at the data.

Like almost everyone else, I've been enthusiastic about this work (take a look at Raftery 1985; Western 1996). But we have to bear in mind that even with these criteria, we are only looking at a teeny fraction of all possible models. (There are some Bayesian statistics that don't require a set of models, but those don't solve the problem I'm discussing here.) When we do model selection or model averaging, we usually have a fixed set of possible variables (closer to the order of 10 than that of 100), and we usually don't even look at all possible combinations of variables. And we

usually restrict ourselves to a single family of specifications (link functions and error distributions, in the old GLM [General Linear Models] lingo).

Now I don't in any way mean to lessen the importance of this sort of work. And I think because of the ease of computerization, we're going to see more and more such exhaustive search through families of models. This should, I believe, increasingly be understood as "best practices," and it can be done outside of Bayesian framework to examine the robustness of our methods to other sorts of decisions. (For example, in an awesome paper recently, Frank et al. [2013] compared their preferred model to all possible permutations of all possible collapsings of certain variables to choose the best model.) But it doesn't solve our basic problem, which is not being able to be sure we're somewhere even *close* to the true model.

You might think that even if it doesn't solve our biggest problems, at least it can't *hurt* to have statisticians developing more rigorously defined estimates of model parameters. If we're lucky enough to be close to the true model, then our estimates will be way better, and if they aren't, no harm done. But in fact, it is often—though, happily, not invariably—the case that the approaches that are best for the *perfect* model can be *worse* for the wrong model.

When I was in graduate school, there was a lot of dumping on Ordinary Least Squares (OLS) regression. Almost never was it appropriate, we thought, and so it was, we concluded, the thing that thoughtless people would do and really, the smartest people wouldn't be caught dead within miles from a linear model anyway. We loved to list the assumptions of regression analysis, thereby (we thought) demonstrating how implausible it was to believe the results.

I once had two motorcycles. One was a truly drop dead gorgeous, 850 cc parallel twin Norton Commando, the last of the kick start only big British twins, with separate motor, gearbox, and primary chain, and a roar that was like music. The other was a Honda CB 400 T2—boring, straight ahead, what at the time was jokingly called a UJM—a "Universal Japanese Motorcycle." No character whatsoever.

I knew nearly every inch of that Commando—from stripping it down to replace parts, from poring over exploded parts diagrams to figure out what incredibly weird special wrench might be needed to get at some insignificant part. And my wife never worried about the danger of me having a somewhat antiquated motorcycle when we had young children. The worst that happened was that sometimes I'd scrape my knuckles on a particularly stuck nut. Because it basically stayed in the garage, sheltering a pool of oil on the floor, while I worked on it.

The Honda, on the other hand, was very boring. You just pressed a but-

ton, it started, you put it in gear, it went forward, until you got where you were going and turned it off.[1] If I needed to make an impression, I'd fire up the Norton. But if I needed to be somewhere "right now," I'd jump on the Honda. OLS regression is like that UJM. Easy to scorn, hard to appreciate—until you really need something to get done.

Proof by Anecdote

I find motorcycle metaphors pretty convincing. But if you don't, here's a simple example from some actual data, coming from the American National Election Study (ANES) from 1976. Let's say that you were interested in consciousness raising at around this time, and you're wondering whether the parties' different stances on women's issues made some sort of difference in voting behavior, so you look at congressional voting as a simple dichotomy, with 1 = Republican and 0 = Democrat. You're interested in the gender difference primarily, but with the idea that education may also make a difference. So you start out with a normal OLS regression. And we get what is in table 1.1 as model 1 (the R code is R1.1).

Gender isn't significant—there goes *that* theory—but education is. That's a finding! You write up a nice paper for submission, and show it to a statistician friend, very interested that those with more education are more likely to vote Republican. He smiles, and says that makes a lot of sense given that education would make people better able to understand the economic issues at hand (I think *he* is a Republican), but he tells you that you have made a major error. Your dependent variable is a dichotomy, and so you have run the wrong model. You need to instead use a logistic regression. He gives you the manual.

You go back, and re-run it as model 2. You know enough not to make the mistake of thinking that your coefficients from model 1 and model 2 are directly comparable. You note, to your satisfaction, that your basic findings are the same: the gender coefficient is around half its standard error, and the education coefficient around four times *its* standard error. So you add this to your paper, and go show it to an even more sophisticated statistician friend, and he says that your results make a lot of sense (I think he too is a Republican) but that you've made a methodological error. Actually, your cases are not statistically independent. ANES samples congressional districts,[2] and persons in the same congressional

1. In fact, it was so boring that I attached deer antlers to the front for a while, until a helpful highway patrolman told me that if I ever was in an accident, and an antler actually impaled a pedestrian, unlikely though that was, I would certainly get the electric chair, and then go straight to hell.

2. Actually in the 1976 ANES, the sampling units were not congressional districts but

Table 1.1. Proof by Anecdote

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Type of Model	OLS	LOGISTIC	HGLM	OLS	HGLM
GENDER	−.017	−.071	−.011	−.039	−.107
	(.031)	(.130)	(.152)	(.033)	(.164)
EDUCATION	−.083***	−.348***	−.383***	−.029	−.166
	(.018)	(.078)	(.093)	(.022)	(.108)
CONSTANT	.620	.497	.734	.633	.808
RANDOM EFFECTS VARIANCE			1.588		1.552
SECRET				−.081***	−.352**
				(.021)	(.111)
R^2	.021			.028	
AIC		1374.9	1264.5		1112.6

N = 1088; Number of districts = 117; *** $p < .001$; ** $p < .01$

district have a non-independent chance of getting in the sample. This is especially weighty because that means that they are voting for the same congressperson. "What do I do?" you ask, befuddled. He says, well, a robust standard error model could help with the non-independence of observations, but the "common congressperson" issue suggests that the best way is to add a random intercept at the district level.

So you take a minicourse on mixed models, and finally, you are able to fit what is in model 3: a hierarchical generalized linear model (HGLM). Your statistician friend (some friend!) was right—your coefficient for education has changed a little bit, and now its standard error is a bit bigger. But your results are all good! Pretty robust! But then you show it to me. It doesn't make sense to me that education would increase Republican vote by making people smarter (strike one) or by helping you understand economic issues (strike two). I tell you that I bet the problem is that educated people tend to be *richer*, not *smarter*. Your problem is a misspecification one, not a statistical one.

I have the data and quickly run an OLS and toss income in the mix

looped across them; in 1978 they were congressional districts. But your statistician friend is a bit foggy on these details. . . .

(model 4). The row marked "SECRET" is the income measure (I didn't want you to guess where this is going—but you probably did anyway). Oh no! Now your education coefficient has been reduced to a *thirteenth* of its original size! It really looks like it's *income*, and *not* education, that predicts voting. Your paper may need to go right in the trash. "Hold on!" you think. "Steady now. None of these numbers are *right*. I need to run a binary logistic HGLM model instead! That might be the ticket and save my finding!" So you do model 5. And it basically tells you the exact same thing.

At this point, you are seriously thinking about murdering your various statistician friends. But it's not their fault. They did *their* jobs. But never send in a statistician to do a sociologist's job. They're only able to help you get bestimates of the *right* parameters. But you don't know what they are. The lesson—I know you get it, but it needs to stick—is that it rarely makes any sense to spend a lot of time worrying about the bells and whistles, being like the fool mentioned by Denis Diderot, who was afraid of pissing in the ocean because he didn't want to contribute to drowning someone. Worry about the omitted variables. That's what's really drowning you.

So moving away from OLS might be important for you, but in most cases, that isn't your problem. Indeed, OLS turns out to be pretty robust to violations of its assumptions. Sure, it doesn't give you the *best* estimates, but it doesn't go bonkers when you have restricted count data, even (often) just 0s and 1s. Further, and more important, it has a close relation to some model-independent characteristics of the data. You can interpret a "slope" coefficient as some sort of estimate of a causal effect, if you really want to . . . or you can see it as a re-scaled partial correlation coefficient. And those descriptive interpretations can come in handy. Most methodologists these days are going to tell you to work closer and closer to a behavioral model. And I'm going to say that's one half of the story. Just like some politicians will say, "Work toward peace, prepare for war," I will say, "Work toward models, but prepare for description." And so I'm going to take a moment to lay out the theory of the use of data that guides the current work. But first, a little terminology.

Models, Measures, and Description

We're often a bit casual in talking about "models," "measures" and so on—statisticians aren't, and I think we should follow them here. A model is a statement about the real world that has testable implications: it can be a set of statements about *independencies*—certain variables can be treated as having no intrinsic association in the population—as in Leo Goodman's loglinear system. Or it can be a statement about mechanisms or processes—either causal pathways, or behavioral patterns.

Models usually have parameters in them. If the model is correct, these parameters *may* have a "real worldly" interpretation. For example, one of them might indicate the probability that persons of a certain type will do a certain type of thing. They might indicate the "elasticity" of an exchange between two types of resource. But they don't always need to have a direct real-world analogue. And our *estimates* of these parameters can be useful even when they aren't really very interpretable. In many cases, we use as a rule-of-thumb the statistical test of whether a parameter is likely non-zero in the population as a way of constraining the stories we tell about data. This approach has come in for a lot of hard knocks recently, perhaps deservedly, but I'll defend it below. For now, the point is that not all parameters have to be real-world interpretable for us to do something with them.

Let's go on and make a distinction between the *estimate* of a real-world parameter (should there be any) and *measures* that (as said in *TTM*) I'll use to refer to a process whereby we interact with the units of measurement (individually and singly, one might say) and walk away with information. Finally, when possible, we'll try to distinguish model parameters (and measurements) from *descriptive statistics*. While this distinction may get fuzzy in a few cases, the key is that descriptions are ways of summarizing information in a set of data that are *model-independent*. No matter what is going on in the world, a mean is a mean.[3] Not to be intentionally confusing, but what a mean *means* stays the same. In contrast, a parameter in a complex model (like a structural equation measurement model) has no meaning if the model is seriously off.

That's the nature of good description. What you do is figure out ways of summarizing your data to simplify it, give you a handle on it, and the best descriptive methods structure the data to help you understand some especially useful aspects, ones that have to do with the nature of your data and the nature of your questions, but still without making use of particular assumptions about the world. When it comes to continuous data expressed in a correlation matrix (which is what OLS among others deals with), conventional factor analysis is a classic descriptive approach.

3. If you're a sophist, you'll immediately try to come up with reasons why this isn't so, and you'll be wrong. If you have a single distribution of a continuous variable, with *N* observations, there are *N* moments to that distribution. The mean is the first, the standard deviation the second, and so on. If you use all of them, you recreate the whole distribution. The first two moments adequately characterize *some* distributions; for a wider family, you need the first four, but every distribution can be characterized with all of them. The best that you can do, oh beloved sophist, is to argue that there is a continuum, with some quantities useful for practically any model, and others useful for very few. I'll accept that, and it doesn't undermine my point.

Now when I was in graduate school, factor analysis was the one thing we despised even more than OLS. It was, as Mike Hout called it, "voodoo"—magic, in contrast to a theoretically informed model. That's because something always came out, and, as I'll argue in chapter 9, when something "always comes out" it's usually very bad. But in contrast to other techniques that always give an interpretable result, factor analysis turns out to be pretty robust. Is it perfect? Of course not. Should you accept it as a model of the data? Of course not—because it isn't a *model*. It's a description, a reduction of the data. And chances are, it's going to point out something about the nature of the data you have.

Now as Duncan (1984b, c) emphasized, knowing a correlation matrix doesn't always help you. The pattern of covariation in a set of data, he said, is far from a reliable guide to the structural parameters that should explain the data. But most of our methods are based on the same fundamental mathematical technique of singular value decomposition. This is a way of breaking up a data matrix into row and column spaces. As Breiger and Melamed (2014) have demonstrated, most of our techniques basically take these results and rescale them (like correspondence analysis) or project them (like regression). So our common linear model, rather than being an *alternative* to description, can be understood as a particular *projection* of the description for certain analytic purposes.

Many methodologists think that the best approach is one that precisely translates a set of behavioral assumptions into a model with parameters that quantify the linkages in the model. To them, the fact that OLS is close to description shows how primitive it is. I'm going to argue that we're best off attempting to use our data to eliminate theories using models that are actually as close as possible to description. So it's time I laid out the approach to statistics that guides this work.

What Is, What Are, and What Should Be, Statistics?

It seems that there isn't real agreement as to where the word "statistics" comes from, and it currently is ambiguous. We use it both to mean the raw materials—the numbers—that we analyze, as well as the set of tools that we use to analyze them. It seems pretty certain that the word first referred only to the former; it is also pretty certain that it comes from the root of words for "state," but it shares the ambiguity of that word itself, for it appears that it originally meant "those numbers that tell us the state of the state," that is, the condition of the government (see Pearson 1978 [1921–1933]; Stigler 1986).

But what we think of as statistics as a field of applied mathematics came from work that demonstrated that a finite, indeed, rather small,

sample could be used to estimate (1) the population value of some numerical characterization (such as average height); (2) the population variance; (3) the likely error of each of these estimates. That's still the heart of statistics—what we call the "central limit theorem." It isn't the limit that's "central," it's the theorem. It's the basis for what we think of as statistics. This enterprise of statistics, then, is all about inference to populations from samples.

This emphasis on inference continued as statistics began to move away from descriptions (like a mean or a correlation coefficient) to what were increasingly interpreted as models (like a set of slope coefficients). But this opened up a new form of error. There are three ways that a slope coefficient can be wrong. The first is that we've made a mistake of *calculation*. These days, this usually means a programming error. The second is that while our calculation is correct, the value in the sample isn't the same as that in the population, and it's the latter that we're trying to get at. If we had a complete sample, we'd get the right value. This, then, is a mistake of *inference*. The third is that there's nothing wrong with our calculation and inference, but our model is wrong. We have a mistake of *interrogation*—it's not that we have the wrong answer, but that we asked the world the wrong question. While we can make errors of calculation or of inference for descriptive statistics, errors of interrogation only arise when we move from description to models.

Our basic problem is that statistics, as generally taught, has relatively little to say about our problems of interrogation. We can't ask for the right estimates until we know what is going on, or at least, what might be going on—and until we actually have measures of these factors. So how can we proceed? We're often told that we should be using our data to test theories. I'm pretty sure that *that* approach hasn't worked out too well. Instead, I think we should start with a rudimentary theory of social science based on pragmatism (especially that of C. S. Peirce and John Dewey). If you aren't interested, feel free to skip to the next section. But I think we can derive a far more coherent theory of the relation of data to knowledge than we currently have, and one that will better help us with our actual practice, than our current orthodoxy. To do this, you need to imagine that there is something you are interested in, something that you don't know about. That's step one. Step two is that you think about the set of possible plausible explanations (this notion was famously laid out by the great geographer Chamberlain 1965 [1890]).

When we assemble this set of possibilities, you don't allow something that you call "theory" to lead you to ignore hypotheses or interpretations that draw some interest from competent social scientists. That is, if by "theory" you mean "what we already know," like one might talk about

"plate tectonic theory," then sure, your work should be theoretically informed. But if by "theory" you mean "my presuppositions" or "my claims," then any analysis that assumes it is a waste of all our time. Don't let anyone cow you into thinking that he has some special reason why he gets to ignore what he wants to ignore (we'll see examples of how this leads to bad practice in chapter 9).

This notion that you start not with *your* theory, but with the competing notions of a community of inquiry, doesn't quite fit conventional ways of thinking, though it is compatible with the neo-Chamberlainism of Anderson (2012). The pragmatist conception, however, differs even more fundamentally from our current vision of statistics.[4] In the conventional philosophy of science—one that lies at the bottom of our frequentist interpretations of most of our statistical practice—you start from scratch every time. You have a model, a theory of reality, and you want to test it. That means you have a billion things that you test all at the same time. It means you aren't just asking, "Does income affect vote?," but simultaneously, "Does this measure income? Is the central limit theorem applicable? Do other people have consciousness? Am I perhaps a brain in a vat?" A weakness in any one of the assumptions that you need to make your practice defensible undermines your conclusion.

In the Bayesian version, you try to fold more of your current understandings and beliefs into how you interpret the evidence the world gives you. The pragmatist idea is somewhat different. The consistent Bayesian will allow you to assign a .9 probability to the model that says that the presence of capitalist relations of production leads to greater psychological health than do socialist relations of production if you *really, really, really, really, really, really, really, really, really*, believe it.[5] The pragmatists, in contrast, understood science as the efforts of a *community*, and argued that no one could unilaterally declare something likely or unlikely.

Further, in the pragmatist conception, rather than start from the ground and build up, we start from where we are, and where we are is with what we, as regular people, already think that we know. Maybe you can't prove that any of this is true knowledge. Maybe we are hanging in the air without a foundation, but, still, *that's* where we are. Whether what is in our heads be philosophically justifiable or not, either way, science can

4. You might wonder why, given this pragmatist grounding, I don't focus on solving *practical* problems. I don't want to get side-tracked, but the simple story is that too often, our problems are due to *other people*. . . .

5. I recognize that there are now some non-subjectivist Bayesians, or so they think (with high subjective confidence). But they still accept the notion that a model *has a probability*. That actually isn't an easy notion to explain unless you are a subjectivist or believe that there are a large set of parallel universes constantly splitting from one another.

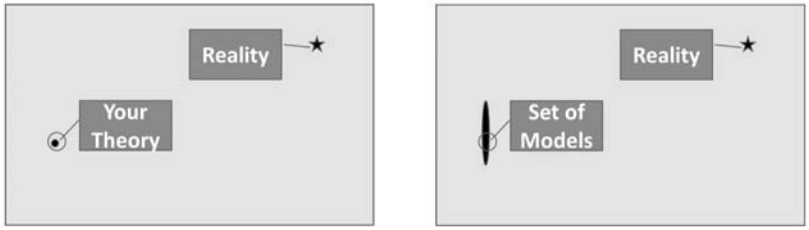

FIGURE 1.1. Conventional and Bayesian Approaches to the Use of Statistics

make it better—if we want it to. Some of our everyday knowledge might be wrong, but if it doesn't matter, let it be. If something matters, and our existing knowledge isn't good enough—that's when we say, "Stand back, I'm going to try science."

That means that if a statistical analysis improves on what we think we know from everyday life, we want to take that evidence seriously . . . even if our quantifications show that there's lots of imprecision. Our goal is to use the power of our data and the rigor of our methods to exert leverage on this process of deciding what interpretation to accept. How should we do that?

I propose that we adopt a strictly *falsificationist* ideal for our practice. This is an old-fashioned idea, and, even worse, associated with a stodgy liberal German philosopher of science, Karl Popper (1959 [1934]). There are lots of good critiques of his theory, and of falsificationism in general. But that isn't determinative—maybe it isn't a good philosophy of science. But it is good practice for *our* field. We should be using statistics not to estimate real world parameters, but to force us away from interpretations of the world that are inconsistent with the body of evidence as it stands.

Of the possible ways of formalizing this connection, falsification is best at working with our current state of statistics. Imagine that the space of all possible explanations of some phenomenon is a plane, like that drawn in figure 1.1, left. Conventional statistics work by us coming up with a single model. We try to estimate the parameters, and the better the estimates (the smaller the radius around the star, impressionistically speaking), the better. But if the truth is far away from where we even are, we don't get much traction on it. We're just tightening a lasso on a steer who's already got clear away.

Now you might respond that conventional methods *are* falsificationist. We're rejecting the null hypothesis, right? Didn't we learn that that's all we can do—falsify models? Well, not quite. Although some loglinear modelers still reject whole models, most of us don't. We might reject one *parameter* in a model, but we don't use that to say, "You know, maybe the *whole idea of regression isn't true* here." You might also say, well, Bayesian

methods get us past this "choke chain on nothing" view of statistics. And there's going to be a lot of promise in Bayesian methods, not only in terms of getting better maximum likelihood estimates, but in widening our view of the world. Bayesian methods can take into account a large class of possible models, and use this to choose a best model, or to incorporate our model uncertainty into our parameter estimates. (To be fair, information-theory based methods can do all of this as well, without leaving the Fisherian approach to statistics I'll advocate in chapter 5.)

Bayesian methods, however, also only explore a small fraction of this space. Even in some of the wilder approaches that try to examine multiple link functions as well as selections of variables, they are limited to a relatively small set of variables, and they certainly don't include the "unknown unknowns." (See fig. 1.1, right.) Although many Bayesian approaches can be used in service of the falsificationist approach I'm advocating here, I'm actually not going to discuss them here, for a simple reason: while Bayesianism as a philosophy is pretty old, most specific techniques are too new. You *should* be excited about the increasing ease of doing what are called "fully Bayesian" techniques that allow for taking one portion of our uncertainty—our distribution of belief across a set of models, both a priori and a posteriori—into account when we estimate parameters and their standard errors, as well as approaches that suggest how to trade off between fit and parsimony to choose reasonable models. But not only are none of these techniques going to offer us a silver bullet,[6] most are simply too new for us to have a "feel" for how they work in practice. Given that your tools will *all* be imperfect, it's better to use one whose tendencies and biases you understand than one you don't . . . of course, while you are also trying to develop new ones.

What's most important, then, isn't any particular number that you create, but the logic of your exploration of a set of substantively important alternative models. Rejecting false models is good, but since there are an infinite number of wrong ideas, we might just bounce from one to another. So we want to see whether, as we move away from a bad model to a different one, we are increasing our inferential scope. If so, we have reason to think that we're moving toward better ideas by throwing out the worse ones. To do this, all we really need are methods that have a better-than-

6. We can never take into account our true uncertainty, because the realm of possible models is actually an infinite one, and not restricted to the handful of predictors we happen to be looking at. Our biggest problems aren't about *estimation*, they're about *specification*, and what we need is rarely going to be solved by new statistics—rather, it will be solved by industrious workers applying the techniques they understand to get a good grasp of what's going on in their data.

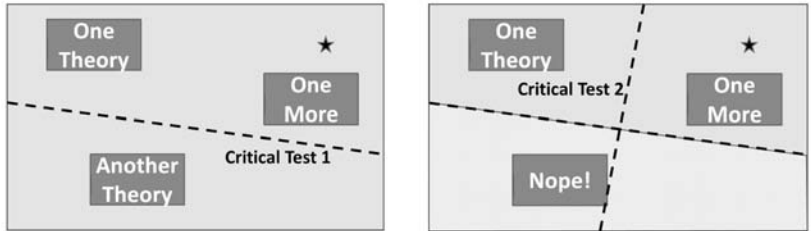

FIGURE 1.2. The Falsificationist Use of Statistics

random chance of leading us to give up on a false theory in favor of a different one that is more likely to be true.

And the cool thing about this is that we can actually say something useful even when there are "unknown unknowns"—relevant factors that affect our findings that we don't know about—so long as we understand something about the *other* possible options. So the way this falsificationist approach works is that we start with a set of possible explanations, and try to ask a question, the answer to which, divides up that space (see fig. 1.2, left). While this is classically called a "critical test," the point isn't that it's a one-shot, never-think-about-it-again answer. It's that it shifts the weight of the evidence from one possibility to another. Once we do that, we're going to want to see if we can further narrow down our options, by asking other relevant questions (see fig. 1.2, right). We might end up unable to narrow things down to more than a large section of this space. But this space has two important characteristics. First, it includes one of the three competing theories. And it includes reality.

This is, it should be emphasized, an *ideal*. We'd like to push one whole side of this space into "nope!" Often we can't, and so what we are instead doing is assigning some sort of credence to various hypotheses—shifting the weight of the evidence back and forth. That sort of endeavor is mathematized in various information criteria, but I'm arguing that we need to be doing something like that as a community of investigators, even if we can't mathematize the answer (for example, when the evidence for one theory comes from ethnographies, and that for another from statistics). There aren't easy answers. But there might be good ones. Let's translate this to practice.

Enter Statistics

What we want, then, is to use statistical practice to help us in this goal. Let's take the example of studying individual political behavior with data on votes. If you regress political vote on income and education, you can, of course, be trying to get a sense of what the causal effect of education is

on vote. I have a hard time understanding what the hell that could possibly mean, and even if there was such a thing, I doubt there's any way of getting at it using these sorts of approaches. Many stats books would say, then go do something else. I disagree. Regression results aren't useful because they capture the actual real world causal parameters . . . but because they have descriptive utility for the project of adjudication.

Let's go back to our example of the person who noticed that Americans with more education are more likely to vote Republican than those with less education. Our researcher might have an explanation, namely that those with education are more likely to have an accurate grasp of the complexities of economic policy, and the need for responsible budgeting. As this set of economic policies is (rightly or wrongly) more strongly associated with the Republican party, those with more education are more likely to support the Republicans.

That's a story. Now you might think about education as a "cause" and something that might have "confounders" or all that sort of thing. And you could work very hard at establishing the correct estimate of this causal parameter. But you don't have to. You could stick with this as a comparison—those of more versus less education. You might notice that the logic of this person's claim implies that if you were, say, to stratify your sample by income level, within every income level, those with more education should be more likely to vote Republican. That is, there's no reason why his explanation wouldn't continue to hold. So let's say that you do this—you divide up income into deciles, and you look at the difference in Republican vote between those with more and those with less education *within* each decile. As you might guess from before, this will actually lead the relation to disappear or reverse; those with more education, in any income category, will vote *less* Republican!

We're going to throw out that first guy's story. And we're going to have an understanding of *why* it should be thrown out—in our sample, education is associated with income, and those with higher household income are more likely to vote Republican. Said somewhat differently, the relation between education and vote is very different from the relation between education-conditional-on-income and vote. You don't know much about American politics unless you understand both of these.

So we could split the sample into deciles like I said here, but if we found that in all deciles, the relationship between education and vote was relatively similar, then we might make a single number that expresses this relation, instead of ten different numbers. That's what a regression slope is. It's a way of taking a complex data space, rotating it so that we can ignore trivial dimensions, and then projecting it in such a way that it answers a question about a complex comparison.

We do a regression like this when we are going to *rely upon* the assumption that there is a simple relationship between subcategories formed by the cross-classification of many variables. We do that—and we do it *justifiably*—when some of the following are true: 1) in the past, we've found that sort of simple relationship; 2) we don't really have enough data to do anything else; 3) we don't really care all that much about our findings—we won't be in deep trouble if we're wrong about this. (This might seem shockingly casual, but I think we'll see that as we do serious research, we often have to budget our forces reasonably—we can't chase down every possibility fully.) Of course, if the comparison we're trying to get out of our regression is *too* complex, there's indeed a good chance that our regression won't give us what we want. One assumption of a linear relation is pretty innocuous. Add a second, and an assumption of independence of the two predictors, and it gets a bit shaky. Add a dozen, and you're talking Vegas odds. In general, we want our models to do two things: first, let us see patterns in the data that are relatively robust, and second, rule out some interpretations of those patterns.

Now of course, there *could* be a reason why our first person's theory turns out to still be true. It could be that there's *another* variable that we can use to stratify further that reverses our conclusions. Or it could be there's a really good reason, when we think about it carefully, that the person could still be right, though maybe we can't show it with these data. But the burden of proof has certainly shifted. Until this person comes back with something really strong, *that* story is out. Do we understand the association between education and vote? We don't. But we won't learn anything by just *assuming* that it *causes* vote in some mysterious way, and then trying to *estimate* the strength of this causal pathway.

In sum, statistics aren't there to estimate the parameters from your story. They're there to *eliminate* stories, to falsify some claims. We want to understand the possible explanations for simple descriptive patterns, and then bring data to bear on them to so that we can throw some of those out. And the task of statistics is to make sure that we're more likely to eliminate the interpretations that *should* be eliminated than those that shouldn't. If what we're left with really looks like it is pretty close to the "true" story, then let's break out the champagne.

Let me give an "existence proof" of the plausibility of this approach in the form of a true story. When my first book on *Social Structures* came out, it was given a very thorough, but ultimately negative, review by Neil Gross in *Contemporary Sociology*.[7] How delighted I was, then, to find that Gross—who is not an expert statistician—was writing a book on an obviously im-

7. "'Lack of insight!' I'll lack-of-insight him, I will!" I went around muttering for a week.

possible question, "Why are professors liberal?" All I had to do was hang back, watch him fall flat on his face, and then step daintly over.

Actually, he pulled it off (2013). How? First, he doesn't try to hang everything on a single coefficient in a single model. Instead, he starts with a robust finding—that professors *are* disproportionately liberal—and a small set of well-formulated alternative explanations of this fact: self-selection, discouragement of others, capture, and conversion. He judiciously goes over the different bits of evidence that we have, doing multivariate analyses where he can find something relatively robust, and continually assessing where the balance of the evidence lies. And, most important, while doing this, he attempts to root his investigations in what we can figure out about the concrete processes whereby careers unfold over the lifecourse. He never rests on one clever silver bullet to come up with "the" right estimate of a parameter from a simplified model.

To conclude, we've been a little bit too scared of ambiguity. Because our statistics would be totally defensible *if* the world was such-and-such a way, we've been tempted to *pretend* that it is in fact that way (cf. Gigerenzer 1991).[8] I'm saying that if we are willing to stop pretending, we have a different way of working, one that involves trying to learn from data by pursuing different interpretations and conducting multiple tests. And, sadly, many of us have been taught that this is the *worst* thing we can do!

Overfitting and Learning from Data

When I was in graduate school, we were severely warned against "data mining," which meant trawling through our data, looking for interesting things. The reason for this is that all the statistics we were learning required that one first (without looking at the data) construct a hypothesis, and then test it. If you didn't "call it in the air," the statistics were meaningless. If you *really* needed to look at your data first, you could have *one* peek, but only if you split your data, and peeked at one half, and then tested on the other.

In case you're rusty on this, and because sociologists can be casual in their language, let's review our basic terms. *Fit* is how close our predictions are to the data we used. *Model selection* is about choosing which model to believe. Often the best model is not the best-fitting one. Why?

8. I often hear the response, "But if you can think it, you can model it." Indeed, if you devote five minutes to thinking about something about the social world, you should be able to write down a fitable model for it. But if you spend two hours *really* thinking about it, you should be able to figure out why the model isn't right. And if you spend *eight* hours, you should be able to figure out why you *can't* do a plausible model that doesn't assume almost everything you're supposedly trying to find out.

Because the one that stretches itself to best fit *these* data might be terrible on the *next* set of data. It's trained itself to fight the last war, as it were. Of equally well-fitting models, we tend to prefer those with fewer parameters, not because we arbitrarily prefer simplicity, but because we think it's less likely to have parameters that only fit the particularities of *this* sample. We use the combination of fit and parsimony in the quest of model selection—our decision as to what we think holds in the population from which we've sampled. Fit, then, is about *this sample*, and when we are interested in inferences, we have to control our desire to fit. If we spend too much time looking at our data, we will increase our fit—but decrease our capacity to make inferences, because we choose the wrong model for the population. Hence the rule: don't choose your models based on what the data tell you! It's a basic notion that goes back to Francis Bacon.[9]

That follows logically from the basic axioms of our statistics. But no one I know who went on to be a skilled data analyst did this. Guiltily, we pored over our data, looking for what the hell was going on. Yes, that means we sometimes "over-fit"—that is, we came up with "false positives" where it looked like our theory was true, but it really was just the "good" luck due to the sampling fluctuations in *this* data set. . . . in next year's, say, we wouldn't come up with the same finding.[10] But we over-fit a lot less than you would think. In fact, I think those of us who combed through the data had *fewer* false positives than the rigid "testers." Remember that if your statistics teacher urges you to take a more orthodox approach, that's because in his class of 40, he is fine with the idea that two or three of you will go to press with a *totally* wrong finding just due to sampling error. You do things his way, that's as good as you can expect. You look more closely, and I think you're less likely to be one of those false positives.

Here's why. The classic "tester" has a finding she wants to be true. Let's say that she wants to argue against those who say that "feminism" (in the sense of being against traditional gender roles) is intrinsically a pro-

9. "In forming our axioms from induction, we must examine and try, whether the axiom we derive, be only fitted and calculated for the particular instances from which it is deduced, or whether it be more extensive and general. If it be the latter, we must observe, whether it confirm its own extent and generality by giving surety, as it were, in pointing out new particulars, so that we may neither stop at actual discoveries nor with a careless grasp catch at shadows and abstract forms instead of substances of a determinate nature and as soon as we act thus well authorized hope may with reason be said to beam upon us" (*Novum Organum*, 106; 1901 [1620]: 128). Thus statisticians correctly caution us against overfitting.

10. Now even then, truth be told, there were some suitable ways of determining the statistical significance of a parameter when one conducts multiple tests and so on, and now it's more common for people to carry these out.

woman position. So she takes data from a survey (say the General Social Survey [GSS]) and shows that men and women don't differ significantly in their support for less traditional gender roles for women (and indeed, that's what you'd get, at least circa 1993 when I last did this). But a critic might say, "Maybe this is due to education differences between men and women." Sweating, our researcher adds "education" as a control, crosses her fingers, and . . . ah! No change. The gender parameter is still insignificant. Thank God.

What this approach (theory testing) is doing is preventing our researcher from using the data to learn much. She's so intent on preserving her own theory, that she can't see what she didn't already think of . . . unless it was suggested by someone else. And even then, rather than push us to be increasingly clear about the contours of the processes or patterns we are trying to uncover, this way of prompting research blurs them. The game she plays is that she tries to hang onto her finding, while doing justice to everyone else's idea by tossing in more and more controls. All these controls make it hard to know what the hell we are doing. (We'll return to these issues in chapter 4.)

And what this sort of approach encourages is that we *stop* our investigation at a convenient place. That's not a "conservative" strategy at all! A conservative strategy is one where we take our current leading interpretation (even if it's "nothing is going on") and see whether it implies any *other* testable hypotheses. Our researcher won't stop with just finding an insignificant gender coefficient, because she's trying to paint a picture of what men and women are like. And then she'll think, "Well, if men and women really are no different, if I split the sample by sex, within each group, all the predictors should have the same coefficients." And she'll do that and find that the education coefficient is the same for both men and women. But if she throws in a measure of income, suddenly, men and women *aren't* the same.

Why not? Now our researcher isn't interested in testing, but learning. Hmmm. . . . that measure of income was total *household* income. Maybe it matters *whose* income it is. What if I look only at married people here, and only at women who have jobs, and men whose wives have jobs, and break up the income into two portions, one, her income, and the other his. Guess what jumps out at you? Every dollar that the woman makes predicts *decreased* traditionalism on her part. Every dollar that the man makes predicts *increased* traditionalism on her part.

Don't stop here. Is it the dollars, the percentage she contributes, or just the fact of working, that is the best predictor? Can we do the model separately by year? Are the parameters changing? And so on. When you're done with this, you're more likely to have a true finding, despite the "data

mining," than if you tested. Why? Because you follow the trail pointed to by a working hypothesis, and you follow it until the trail dissolves. Everything seems to be coming together, and that implies you should see *X* . . . and there's no *X*. So you have to revise your working hypothesis and go on. That's what we call learning. In other words, we put robustness (the same finding appears when we do things differently) and internal validity (the findings all support one another in interpretation) above that magical * sign that indicates statistical significance.

Further, this approach helps you remember what you have—cases, usually people. Instead of jumping to test an abstract theory, you start thinking concretely. It's a mystery you need to solve, but like a good detective, you want to determine your suspects before you start worrying about their motives. Who is driving my finding? Who has high values on my key variables? Who has low? Is the finding driven more by lows or highs, or both?

Finally, if you're still queasy about this idea of really plumbing your data, because your statistics teacher has drummed into your head that only cheaters look at the data before conducting tests, ask your teacher if *the entire scientific community* should do only one test for each research question. Presumably, she or he will say absolutely not. We should use the results of past research to come up with new hypotheses and test them. Now ask why *you* can't do that on the basis of your *own* past results. Why is *your p*-value wrong when you pursue the logical next step, when the *p*-value of the person in the office next door isn't?

Science does, just like the squares will teach you, work by constructing hypotheses. But, as I said in *TTM*, this is something you should be doing *a dozen times a week*. Otherwise, you just make the rest of us work through the implications of your claims. And that slows us all down. If the scientific community as a whole can do it, you can try to do it too.

Standards of Proof

There's one last issue about how we compare our explanations, which has to do with our standards of evidence. I think we need to understand that different questions, and different comparisons, require different standards. We usually first think about this when we face the problem that classical statistics is able only to reject a null hypothesis—not to establish your own. The null hypothesis is something boring and dumb, like "everything is completely random" or "this variable doesn't predict that one at all!" or "maybe we just pulled a weird sample." In the rare chance that we test our *own* theory, as opposed to the null, we want to make it easier to reject. That's because it's a bit unfair to imply that you are right

simply because you can stump the chump. That doesn't establish which of the infinite number of *other* hypotheses is worth believing in. It's sort of like Albert Einstein claiming he's right, because he's shown that Mortimer Snerd is wrong.[11]

Approaches that look at the weight of evidence across a class of models seem a lot better: they can include a null model, but aren't restricted to this. But they come with their own problems; such equal weight across all models isn't always what we want, because we can't always determine the proper set of possible explanations, and how we populate this set influences our conclusions (I'll return to this shortly). Instead of looking for a general numerical solution to this problem, we need to think about the different types of claims we might be investigating.

Here, for want of a better term, I'm going to borrow a notion from Boltanski and Thénevot (2006 [1991]) and say that what researchers are trying to do with their statistics is to "qualify" something in the world. For example, they may say that employers discriminate against African-Americans. If successful, the admittedly obscure object "employers" now has a new quality—*discriminatory*. Others may attempt to use evidence to prevent or remove this qualification.

Here we can draw upon a different way that we have learned to think about this process of qualification, one coming from our legal system.[12] First, let's consider what I will call the *criminal* understanding of the relation between claim and evidence. This is an asymmetric understanding, where the burden of proof is against the claims-maker. Such asymmetry makes sense in some settings: it makes sense for criminal trials, given the asymmetry between a state and a citizen. The *p*-value test and the 95% confidence interval are our "beyond a reasonable doubt." When is this approach appropriate for social science? Perhaps when there is a single contending qualifier of the world (that is, someone who says, "the world is like this") and a single world to be qualified, where we want desperately to avoid being wrong, but don't particularly mind not being right. And when might that be? Very often, when someone is making a claim to support an *intervention* (e.g., a new policy). The deck *should* be stacked against her—at least, to the extent that her argument is made using statistical reasoning.

11. Mortimer Snerd was a literal dummy, used by ventriloquist Edgar Bergen. The great sociological theorist and grouch Pitrim Sorokin used him as an example of a real idiot. Because Sorokin probably listened to Bergen on the radio, I'm not sure whether he knew that Snerd was made of wood. If not, who was the real dummy?

12. If you would like some pedigree here, C. S. Peirce proposed just this way of thinking about differences between intellectual endeavors and their standards of proof. See "The Logic Notebook," 1985 [1865–1866], *Writings*, vol. 1, p. 337–350; "The Logic of Science; Or, Induction and Hypothesis, Lecture III, Lowell Lectures of 1866," 1985 [1866]: p. 357–504.

In contrast is a *civil* understanding, in which we have a balanced burden, and whichever side is over 50% wins. This arises in law when we have not a state against a citizen, but two citizens arguing over where the fence line should be. It has to be *somewhere*, and it would be unfair to imagine that whoever was the first to complain should face a higher burden of proof. When is this standard of "over 50%" the reasonable one? We might at first suspect that it is necessary in conditions where we are oriented by a need for practice: where we must do something—to do, or to forebear. But in a moment, I'll try to argue that this *isn't* the case; rather, this is most useful when we have a problem that is relatively disconnected from practice, and where the alternatives are few and data scarce.

For example, historians may have narrowed down their theory of the basis of the emerging party system in revolutionary America to two general theories. One is that it has to do with, basically, class relations (though not everyone uses these terms!). Some elites were tied to land, others to commerce; some had an interest in westward expansion, and others didn't. The second is that it was historically specific—whichever faction of elites was able to capture the governorship alienated the elites in different networks, as the former monopolized patronage opportunities. For this question, there is no pressing need to come to closure on our decision. Data is difficult to bring to bear on the question; the participants are long dead, and written statements of motivation are suspect. Each new bit of analysis brought by adherents of one theory can push the property line a bit further in the other direction, as it were.

But why isn't the same thing true if we are deciding, say, whether or not to decriminalize cocaine use? Either we do or we don't (cf. Peirce 1985 [1865–1866], "Logic Notebook," 339). So we should assess the evidence at hand, and whichever way it points, that's the way we should act, right? Unfortunately, this way of thinking about the issue is misleading, and here I'd like to review the identical problem as it defeated a number of early theories of probability (and I want to rely on C. S. Peirce's masterful work here).

Imagine we are hoping to test whether education affects income. We might think, well, either education causes income or it does not. Whichever side has the most evidence is the side I will pick. And (to use Bayesian terminology) you might think that you should therefore assume that your prior estimate of the probability of "no effect" is .5. But rephrase this as "I wonder whether the effect of education on income is precisely zero." Then the a priori chance of a zero effect is vanishingly small, so small that you'd have to put a rather crazy distribution on the priors to avoid it being completely unnecessary to carry out any research at all, because nothing could shake your commitment to a non-zero effect. Peirce

pointed out that a number of previous attempts to mathematize probability had foundered on this problem, confusing our ignorance regarding *one particular enunciation of a question*—such as, "will the die turn up a one or not?"—with a defensible assignment of probabilities. We may indeed be completely unsure as to the next roll, argued Peirce, but the *only* stable denominator for the construction of a probability is not the number of alternatives we happen to be thinking about, but *the number of possible states of the world*.[13]

There are more than two options for drug policy. Whenever we take *our own* preferred one, and put it up against all others, we're coming up with an incredibly overblown estimate of its probability, just like a would-be Bayesian who claimed that the prior probability of rolling a 1 versus a non-1 in a conventional die was ½. That said, we can make the inverse error—rather than using our interest in one particular option to push it to the front, as in this example of drug policy, we use our interest to push it off the table completely! How do we do this? We confuse our failure to reject a null hypothesis with our having established it. For example, as we'll see in chapter 3, it's quite true that, as many statistical methodologists now emphasize, observational studies often can't really identify putatively causal effects of environmental factors in the presence of strong selectivity. That fact can be used to deny the claim to make a policy to affect some outcome. But just as the capacity to reject the null hypothesis in one data set doesn't mean that the hypothesis held by the investigator is demonstrated to be true, so the failure to reject doesn't mean it's false. In many cases, the failure to reject the null hypothesis comes not because our results show that the true effect is close to zero, but because they simply tell us that our estimates are extremely imprecise. If the data is saying "this effect could be zero, or it could be very, very strong indeed," it's irresponsible for us to look at the data, and then turn around and tell people, "statistical analysis demonstrated that there was no effect."

How do we balance these two different errors? First, weak statistical evidence shouldn't lead us to throw out our belief in processes that we have other sorts of evidence for—even if that evidence is of the kind that doesn't qualify as social science, it's still evidence (for example, personal experience, hearsay, common sense—all of which we'd like to improve upon, but none of which we can live without).

Second, I would suggest that we will want to rely more and more on

13. For this reason, our conclusions can be changed depending on how we populate our set of possibilities: if, instead of considering three hypotheses, A, B and C, we divide A into two variants, A1 and A2, many methods will now give A = A1 ∪ A2 more a priori weight!

Bayesian approaches, in which we consider which model among a set to prefer, when our problems are more like *civil* ones, and more on classical approaches when our problems are more like *criminal* ones. And most important, what we want is to move to situations in which civil understandings are appropriate. That means that we, as a scientific community, have narrowed down the set of possible alternatives to a manageable number, and we are doing different types of research to assess the relative weight of the evidence for each. Where we *aren't* in that situation, either because there are far too many possibilities (put differently, we don't really know much) or because we are proposing some very new theory, criminal standards—and therefore classical statistics—are more appropriate. Because this is rapidly becoming a minority position, I'd like to defend it explicitly.

A Mindful Defense of Mindlessness

If you've ever read Evans-Pritchard's (1974 [1956]) book on *Nuer Religion*—and you should if you haven't!—you'll be familiar with a poison oracle. This oracle, like many others, is a way of taking luck and using it to make a difficult binding decision. Many societies do this where there is an important matter of life and death and no one can really be sure what is the right thing to do. In such cases, sometimes using a Magic 8-Ball turns out to be a better way of making the decision than trying to think it through. At least then no one is responsible for making a really bad decision.

This particular oracle involves preparing a special poison in a dose that is as likely as not to kill a fowl. The bird is made to ingest the poison. You ask it a question. It answers by dying or living. Think of this as the equivalent to running a regression and getting a significant or non-significant *p* value. It is a ritual that tells us whom (or, in our version, whose career) we should go off and kill.

There are all sorts of good arguments against using statistical significance as a criterion. Most impressively, the American Statistical Association recently came out with a strongly worded rejection of the use of such tests for the purposes of making claims (see Wasserstein and Lazar 2016). The arguments *against* are pretty good—it describes a practice we don't actually carry out (one-sided tests) to solve a problem we rarely have (rejecting the inference that a null hypothesis holds in the population). The arguments *for* are a lot weaker. If that's so, how can I defend it? Easy. If I had reason to think that the poison oracle has a non-zero correlation with the true answer, and I couldn't think of anything better, I'd say we should use *that*.

It's for this reason, when we are conducting research that needs to be judged according to a *criminal* standard, I think it makes sense to use

p-values in the old-style. The virtue of the *p*-value is that it is arbitrarily rigid. There are two findings, one with *p*-value = .0499999 and the other with *p*-value = .0500000. Accept the first, reject the second. Why? Because social science is a discipline that does not produce technology (the way physics produces refrigerators), we cannot be trusted to bargain with nature.[14] We need formulae like this, at least for the everyday, run-of-the-mill research that we are going to do. We should *increase* the rigidity and formulaic nature of *this* side of our work—just we should *decrease* the rigidity and formulaic nature of our work when it can be judged according to *civil* standards of proof.

In the civil understanding, we have a set of researchers who know what the relevant possible interpretations are. They use the compiled data in effect to bargain with each other about which to accept. In the criminal understanding, each researcher is out on her own . . . trying to bargain not with nature but only with herself. We have had too much sophistication, special pleading, bargaining, and "nuance." We need more constraint on our thinking, because we have far too many false positives. If you have a finding with a significance of $p = .07$, even if you have a story for why it's still something worth investigating, I'm going to counsel moving on to something more robust. And if you want to bargain with me I'll sit it out with the same glazed smile I use on the Maoists on Telegraph Avenue. Whatever.

Of course, you should put learning from your data above conducting tests. And that means I don't endorse the orthodox view of how to interpret the *p*-values. The statistics behind these are usually for a *single* test you have run on data you have not peeked at. We *never* do that, and you *never* should. When we are exploring, we might use the *p*-value as a rough guide, but there are other ones (raw numbers, percentages, change in slopes, and sub-sample sizes). So in data exploration, you should use the *p*-values the way meteorologists use their models—as one bit of information among others.

However, when we're done, we should have something that we believe in, and that is also significant according to conventional standards. It's quite true that because we haven't done the one-shot, no peeking, test, the *p*-value isn't quite right—our procedures are too liberal. But as if to

14. And this leads to a serious ethical issue about borderline results, which I explore in our conclusion. It also leads to what is called "*p*-hacking," where analysts tweak their models to get that .053 *p*-value down to .049. I think my emphasis on testing the implications of an interpretation is the best counter, but you should also know, it's pretty common for analysts to find that *nothing* they do can get the *p*-value below .053. (Not that I've tried.)

make up for it, we conduct two-tailed tests by convention, though we really have one-sided theories, and that's being overly conservative.[15]

It's worth emphasizing that this defense makes sense only for conventional data analysis—where we have an omnibus data set, collected for general purposes, and accessible to many different researchers for replications on basically even terms (as opposed to the original researchers collecting data that are already impregnated with their own assumptions). I certainly don't defend the use of *p*-values for experimental work, like that done in social psychology.

To conclude, our goal is not the estimation of parameters. That is a *means* to our goal. Our goal is seeing whether there are some interpretations of the world that we can, for now, dismiss. Does that seem a bit weak to you? Maybe, but knowing your limitations is a big part of science. If you want a more inspiring vision to strive for, the (optional) coda to this chapter describes the notion of mathematical sociology, which I think should be, if not our goal in actual practice, our orienting dream. Our reach must exceed our grasp, else what's a heaven for?

Coda: A Plea for Mathematical Sociology

There are good reasons to imagine that in the current world, the field of statistics would be at death's door, while mathematical sociology would be rising like a phoenix. Statistics as a discipline was oriented around issues of inference from samples, and, in a world where we have millions of observations, many of the issues that classical statistics was best at solving have retreated in importance. In contrast, the penetration of social network analysis by physicists has led to an infusion of much more rigorous mathematics to a number of the questions that historically were central to mathematical sociology.

Yet not only is mathematical sociology *not* reviving, I think it's in danger of completely being forgotten. You, dear reader, *may not even understand what I am talking about*. That is, you won't understand the opposition between the two approaches.

Mathematical sociology is about looking for the mathematics that (might) underlie social processes and structures. It's perhaps a quixotic quest for a true social science. The great breakthroughs here were Levi-

15. There should be *no actual* two-tailed tests, which makes around as much sense as Newton saying that his theory is right if objects either accelerate constantly *downwards* or constantly *upwards*. Pathological research can be identified by tests of a set of coefficients that are allowed to go in *either* direction, and get a story no matter which way. But using the statistics as if we *were* conducting two-sided tests has turned out to be a good practice.

Strauss's adoption of structural models, and then their mathematization by Andre Weil and Harrison White (as well as other attempts by people like George A. Lundberg and of course George K. Zipf!); the later work of White and collaborators like Scott Boorman and Ronald Breiger, and their collaborators Philippa Pattison and her student Carter Butts. Also important is the work by Paul Lazarsfeld on static models and his student James Coleman on processes.

I'll give one image for mathematical sociology—it's a lot like early crystallography. Before the age of the electron microscope, chemists tried to understand something about the structure of molecules by growing them as crystals. Water forms hexagonal crystals (that's why snowflakes look like they do), salt forms rectangular prisms, and so on. To make such crystals with recognizable forms, you generally need a pure solution of the molecule in question. If someone were to criticize you because you didn't have a good "random" sample of naturally occurring salt water, say (some from the Atlantic, some from the Pacific, some from the Indian Ocean, and so on), you would look at them as if they were crazy, right? It's like Allen Barton's (1968) analogy of an anatomist who takes a random sample of cells from an organism and studies them all together. But we often allow statisticians to get us to do just that—to make it hard for us to see structure (if there is any) because we need to make *inferences*. That's their job, and kudos to them. But they're not really there to solve our problems of having a social science.

A statistician, as I've said, can help you figure out whether you have correctly inferred the returns to education on average for American men. You can estimate that perhaps to four significant digits. But it's not much to build a social science on. Why? Because if we divide the country at the Mississippi, the chances that we'll get the same values West and East are pretty low. You might suggest that there are then two "isotopes" of American men; but if we split each of these again, we'll get four different numbers, and so on. These numbers are artifacts, in other words; they are the outcomes of our conventional manipulations of measurements. They aren't actually numbers that reflect something like invariants; rather, they are the result of throwing together all the *real invariants which we do not understand and which might well be mathematizable* (though not necessarily parameterizable). Remember: statistics gets you the best estimates of parameters—whether those are meaningless or not is a totally different story.

Speaking of stories, let me give you one involving a nice beef between two of my heroes, Harrison White and Leo Goodman. It had to do with a seemingly arcane issue, namely how to model the distribution of group sizes that might spontaneously form in a larger setting (such as conversational groups at an embassy party shortly before an Arnold Swarzeneg-

ger character crashes through the window in a tank). This nicely encapsulates the difference in habitus between mathematical sociologists and statisticians, and why statisticians can always win. White (1962) began from statistical mechanics and developed a rigorous approach to trying to solve this problem, one that built on and transcended some earlier work done by Coleman, by the way. But Goodman (1964) showed that there were problems in the way in which White had approached this.[16] And the most damning issue was that White had treated the expected number of singletons as a parameter, as opposed to a random variable. That is, he had ignored the fact that it had a sampling variance.

Now I get it, and I basically am convinced, or convinced enough. But a mathematical sociologist wants to be able to mathematize *this* group, with its number of isolated people *right here, right now*. If that doesn't do justice to the population of all possible sets of groups, then so be it. Mathematical sociology *isn't* about inference in this sense of sampling, and we shouldn't let statisticians come in and smash the more delicate constructions that we need to make. Or, at least, that would be true if mathematical sociologists were still making models of structure.

But instead, mathematical sociology began to fall apart with the influx of simulation work, which encouraged a lot of usually young, usually male, usually students to consider themselves mathematical sociologists because they were using a *computer*. 'Tain't so, not nohow. But this stuff flooded the journals, pushing out mathematical work, which is much harder to review. And these people were unable to do that sort of review, so even as the mathematical rigor of much work in sociological fields increased, there was less exciting mathematical sociology. But mathematical sociology should be the grail that we are searching for—it may be a myth, but if we stop believing that there are mathematical properties of social interaction, we should leave off all this number crunching (for an example, see Newman, Strogatz, and Watts 2001).

And so in this book, I'm going to be talking about how to get reasonable parameters, but my position—strongly—is that if there *is* a mathematical *order* to social life, we should uncover it wherever it appears, and if there *isn't*, we shouldn't fetishize inferential models; rather, we should *use* whatever we can to rule out wrong ideas . . . at least, the ones that we *should* rule out. We're going to try to get some principles to do this in the next few chapters.

16. Oh, did I mention that White also developed a theoretically defensible approach to modeling mobility tables, again, based on statistical mechanics, and Goodman blew him out of the water with the loglinear system, which had no real theoretical grounding but behaved much better statistically?