Philosophy and the practice of Bayesian statistics in the social sciences[1]

Andrew Gelman, Dept of Statistics and Dept of Political Science, Columbia University
Cosma Rohilla Shalizi, Statistics Department, Carnegie Mellon University

19 December 2010[2]

Abstract. We view Bayesian data analysis--the iterative process of model building, posterior inference, and model checking--as fitting well within an error-statistics or hypothetico-deductive philosophy of science, with posterior inference playing the role of "normal science," model checking allowing falsification, and model building providing the potential for progress in normal science (for small changes in a model) or scientific revolutions (for larger re-evaluations).

Our practical experience and corresponding philosophy differs in from usual presentations of the philosophy of statistics in two ways. First, we consider posterior probabilities to be a form of scientific measurement rather than as subjective statements of belief. Second, we perform Bayesian model checking by comparing predictions to observed data (in a generalization of classical statistical hypothesis testing), rather than by attempting to compute posterior probabilities of competing models.

This chapter presents our own perspective on the philosophy of Bayesian statistics, based on our idiosyncratic readings of the philosophical literature and, more importantly, our experiences doing applied statistics in the social sciences and elsewhere. Think of this as two statistical practitioners' perspective on philosophical and foundational concerns. What we bring to the table are our substantive claims about actual social science research, and we attempt to explain those practices here.

We are motivated to write this chapter out of dissatisfaction with what we perceive as the standard view of the philosophical foundations of Bayesian statistics.

Here's what we take as the standard view:

- Bayesian inference--"inverse probability"--is subjective and inductive, learning about the general from the particular. The expression p(H|y) says it all: the Bayesian learns the probability that a hypothesis H is true, given data. In the conventional view, this is completely different from classical frequentist statistics which is based on null hypothesis testing, that is, falsification.[3]

[1] For the *Oxford Handbook of the Philosophy of the Social Sciences.* We thank Igor Carron, Deborah Mayo, and Harold Kinkaid for helpful discussions and the National Science Foundation, Institute of Education Sciences, Department of Energy, and National Security Agency for partial support of this work.

[2] The original Figure 1 had some errors. We inserted a corrected version on 12 June 2011.

[3] The so-called Bayesian hypothesis test, in which a null hypothesis H_0 and an alternative hypothesis H_1 are created, and then one computed their relative posterior probabilities, $p(H_0|y)/p(H_1|y)$, is an attempt to reproduce the aims of hypothesis testing in an inductive framework that is not based on falsification. For reasons we discuss in the present article, we do not think the Bayesian hypothesis test as described here makes much sense—it turns out to be completely sensitive to aspects of the model that are untestable from data and are in practice typically set based on

- The paradigmatic setting of Bayesian inference is the computation of the posterior probability of hypotheses, with scientific progress represented by the updating of these probabilities.

To give a quick sense of our position, we agree essentially entirely with Greenland (1998), who attributes to Karl Popper the following attitude toward inductive inference: "we never use any argument based on observed repetition of instances that does not *also* involve a hypothesis that predicts both those repetitions and the unobserved instances of interest." To put it another way, statistical models are a tool that allow us to do inductive reasoning in a deductive framework.

In taking a broadly Popperian perspective, we are not committing to many of Popper's specific ideas such as the existence of a logical criterion for demarcating science from conscience or that there is something called collaboration of a theory which is central to science but provides no evidence for the truth of a theory. Rather, what we view as central is the scientific--and statistical--ideal of building serious models that make strong predictions, leaving themselves wide open to being falsified by data. This is "severe testing" in the terminology of Mayo (1996), who works in a frequentist framework but whose ideas, we believe, are equally applicable to modern Bayesian data analysis..

What is salient about our social science research experiences is that the models are all wrong. As we like to say, our models are not merely falsifiable, they are also false! To put it another way, we are confident that, by gathering a moderate amount of data, we could reject any model we have ever fit to data.

From "null hypothesis testing" to "model checking"

We write that, in our applied work, we fit complex models and check them with data, and we characterize model checking as central to statistical learning. This perspective differs in both philosophy and practice from the two dominant conceptions of the philosophy of statistics:

1. Classical frequentist inference is centered upon *null hypothesis testing*, with the goal often being to reject the null, for example demonstrating that two distributions are not in fact identical or that a particular parameter is not exactly zero. In contrast, our models are not null hypotheses but rather complex structures within which one can learn about parameters of interest.

2. The dominant strain of Bayesian philosophy, as presented by Savage (1954) and others, subsumes inference within decision theory and treats all probabilities are subjective. In contrast, we agree that probabilities are model-dependent, but we treat statistical models as testable and thus as potentially objective as any other scientific measurements.

We are not claiming that our approach to applied statistics is optimal or even, in many cases, qualitatively different from other methodologies. For example, there is a well-developed

conventional rules—but at this point all that is relevant is that, in the standard view, Bayesian inference is based on updating the probability that hypotheses are true, *not* on setting models up for potential falsification.

classical framework of estimation, prediction, and multiple comparisons which can handle complex models without recourse to the Bayesian formalism.

What we are claiming is Bayes need not be associated with subjectivity and inductive reasoning. It is possible to work within a model-checking, error-statistical framework without being tied to classical null hypotheses. And, conversely, one can use Bayesian methods without being required to perform typically meaningless task of evaluating the posterior probability of models.

The usual story

Is statistical inference inductive or deductive reasoning? What is the connection between statistics and the philosophy of science? Why do we care?

Schools of statistical inference are sometimes linked to philosophical approaches. "Classical" statistics--as exemplified by Fisher's p-values and Neyman's hypothesis tests--is associated with a deductive, Popperian, or error-statistical view of science: a hypothesis is made and then it is tested. It can never be accepted, but it can be rejected (that is, falsified). In the language of Mayo (1996), statistics and science interact via the design and analysis of experiments that yield severe tests of models.

Bayesian statistics--starting with a prior distribution, getting data, and moving to the posterior distribution--is associated with a formal inductive approach in which new information is smoothly integrated into scientific understanding via the updating of posterior probabilities of competing hypotheses, paradigms, or statistical models.

Our disagreement with the usual philosophical understanding of statistics can be conveniently expressed with reference to the following passage from the the Wikipedia[4] entry on Bayesian statistics:

> Bayesian inference uses aspects of the scientific method, which involves collecting evidence that is meant to be consistent or inconsistent with a given hypothesis. As evidence accumulates, the degree of belief in a hypothesis ought to change. With enough evidence, it should become very high or very low. . . . Bayesian inference uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed. . . . Bayesian inference usually relies on degrees of belief, or subjective probabilities, in the induction process and does not necessarily claim to provide an objective method of induction. Nonetheless, some Bayesian statisticians believe probabilities can have an objective value and therefore Bayesian inference can provide an objective method of induction.

This story does not fit applied Bayesian statistics as we have experienced it. Except in some very narrowly construed problems, we do not view “the degree of belief in a hypothesis” as a meaningful scientific statement, we do not consider probabilities to be subjective (or, at least, no

[4] We are citing Wikipedia not as an authoritative source on philosophy or statistics (let alone the combination of the two) but rather as a handy indicator of current consensus.

more subjective than any other aspects of scientific modeling), nor do we believe Bayesian inference to provide a method of induction, if that is meant to imply a direct mapping from data to posterior probabilities of hypotheses. We elaborate on these points below.

Our alternative story

We have no quarrel with the general idea that scientific knowledge builds upon the past, or that we can learn inductively from the particular to the general. As has been long realized, induction is most effective within the context of a good model.

Our key departure from the usually expressed Bayesian philosophy is that we have experienced two kinds of learning: inference about parameters within a model and decisive rejections of models that have forced us to improve or simply replace earlier paradigms. In our applied work, we have *not* found ourselves making gradual conceptual transitions based on evaluations of posterior probabilities of models.

Our progress in applied modeling has fit the hypothetico-deductive pattern pretty well: we build a model out of available parts and drive it as far as it can take us, and then a little farther. When the model breaks down, we take it apart, figure out what went wrong, and tinker with it, or else try a radically new design. In either case, we are using deductive reasoning as a tool to get the most out of a model, and we test the model--it is falsifiable, and when it is falsified, we alter or abandon it. To give this story a little Kuhnian flavor, we are doing “normal science” when we apply the deductive reasoning and learn from a model, or when we tinker with it to get it to fit the data, and occasionally enough problems build up that a “new paradigm” is helpful.

OK, all fine. But the twist is that we are using Bayesian methods, not classical hypothesis testing. We do not think of a Bayesian prior distribution as a personal belief; rather, it is part of a hypothesized model, which we posit as potentially useful and abandon to the extent that it is falsified.

Subjective Bayesian theory has no place for falsification of a prior distribution, except to the extent that it can be placed within a larger model in competition with other candidate prior distributions. But if we think of the model just as a set of assumptions, they can be falsified if their predictions--our deductive inferences--do not fit the data.

Here’s an example of what we consider to be a flaw of the standard view of Bayesian learning about models. In the 1800s, physicists believed that Newton’s laws were true, an attitude potentially expressible as Pr (Newton | information available as of 1890) = 0.99. But after Einstein’s theory came out and was tested by experiment, the evidence moved to Pr (Einstein | information available as of 1920) = 0.999 and Pr (Newton | information available as of 1920) = 0.001, or something of the sort.

We see two big problems with this standard approach. First, a lot of data and analysis showed critical problems with Newton’s laws--even if the relativistic alternative had not been proposed. Quantum states, the blackbody paradox, the stability of the atom, and all the rest. We would like our philosophical model to be able to “reject” Newton’s model--to reveal serious problems with it—without the need for an alternative that beats it under a likelihood ratio criterion or any other.[5]

The second problem with this framing is that physicists do not think Einstein's theory, as stated, is true either. Relativity has problems too, and researchers have been working for decades to come up with possible alternatives. In this search for new models, the various falsifications of classical relativity theory have been crucial. What is important is not Einstein's theory has been "rejected" but rather the particular ways that it does not correspond to reality.[6]

Why does this matter?

Philosophy matters, even to practitioners, because philosophy is used as a guide to practice. We believe that the idea of Bayesian inference as inductive, culminating in the computation of the posterior probability of scientific hypotheses, has had malign effects on statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their conventional framework.

In contrast, a philosophy of deductive reasoning, accompanied by falsifiability, gives us a lot of flexibility in modeling. We do not have to worry about making our prior distributions match our subjective knowledge, or about our model containing all possible truths. Instead we make some assumptions, state them clearly, and see what they imply. Then we try to falsify the model--that is, we perform posterior predictive checks, creating simulations of the data and comparing them

[5] One might think it possible to simply include an alternative, "not-Newton," whose probability would gradually rise toward 1 as new evidence arise to contradict Newton's theory. But this would not work, for two reasons: First, to apply the Bayesian formulation, one would need to specify a specific model for "not-Newton," and this would be nothing so simple as merely taking a parameter in the model and allowing it to vary. Our whole point in this example is that Newton's model was falsified in decisive and interesting ways *before* any coherent alternative had been specified. The second reason why the "not-Newton" gambit would not work, at least nowadays, is that in the modern era we recognize that all scientific models have holes in them. Einstein's relativity has not yet been integrated with quantum mechanics; the familiar probability calculations of genetics are, given the complexities of live cells, only approximations; the laws governing chemical interactions need modification under extreme conditions; and so on. Social-science models tend to be even more contingent and full of holes--consider, for example, supply and demand curves, spatial models of voting, and probability distributions for social networks--and the "not-Newton" alternative in any of these fields, were it even possible to express such a thing, could safely be assigned the probability of 1 without the need to gather any data at all.

Our point here is not that it is impossible to compare models but merely that, under the traditional Bayesian paradigm in which one evaluates the posterior probability of models, it is not possible to reject a model without comparing it to a specific alternative. We find this aspect of the theory unfortunately, given that in our applied research we routinely reject models by themselves. In fact, it is often the information gained in a falsification that gives us direction in searching for a reasonable improvement to the model. Flexible, open-ended model checking is well known in statistical graphics (see, for example, Tukey, 1977) but has no formal place in the traditional Bayesian philosophy of statistics. See Gelman (2003) for further discussion of this issue.

[6] As Sober (1991) notes, concepts such as simplicity of a model are domain specific and cannot merely be expressed in terms of mathematical structure. This relates to our point that a model check necessarily relates to some aspect of interest of observable reality; even a so-called omnibus test of a model represents some projection in a high-dimensional space. Similarly, Day and Kincaid (1994) argue that the choice of best explanation in any field is domain specific, which is consistent with the statistical idea that model choice must be based on utilities rather than posterior probabilities alone.

to the actual data. The comparison can often be done visually; see chapter 6 of *Bayesian Data Analysis* (Gelman et al., 2003) for lots of examples.

We associate this “objective Bayes” approach--making strong assumptions and then testing model fit--with the work of the philosophically-minded physicist E. T. Jaynes. As he has illustrated (Jaynes, 1983, 1996), the biggest learning experience can occur when we find that our model does not fit the data--that is, when it is falsified--because then we have found a problem with our underlying assumptions.

Conversely, a problem with the inductive philosophy of Bayesian statistics--in which science “learns” by updating the probabilities that various competing models are true--is that it assumes that the true model is one of the possibilities being considered. This can does not fit our own experiences of learning by finding that a model doesn't fit and needing to expand beyond the existing class of models to fix the problem.[7]

We fear that a philosophy of Bayesian statistics as subjective, inductive inference can encourage a complacency about picking or averaging over existing models rather than trying to falsify and go further. Likelihood and Bayesian inference are powerful, and with great power comes great responsibility. Complex models can and should be checked and falsified.

Example: Estimating voting patterns in subsets of the population

In recent years, political scientists have been increasingly interested in the connections between politics and income inequality (see, for example, McCarty, Poole, and Rosenthal, 2006). In our own contribution to this literature, we estimated the attitudes of rich, middle-income, and poor voters in each of the fifty states (Gelman, Park, et al., 2008). As we described in our article on the topic (Gelman, Shor, et al., 2008), we began by fitting a varying-intercept logistic regression: modeling votes (coded as y=1 for votes for the Republican presidential candidate or y=0 for Democratic votes) given family income (coded in five categories from low to high as -2, -1, 0, 1, 2), using a model of the form, logit $\Pr(y=1) = a_s + bx$, where s indexes state of residence--the model is fit to survey responses--and the varying intercepts a_s correspond to some states being more Republican-leaning than others. Thus, for example a_s has a positive value in a conservative state such as Utah and a negative value in a liberal state such as California. The coefficient b represents the “slope” of income, and its positive value indicates that, within any state, richer voters are more likely to vote Republican.

Anyway, it turned out that this varying-intercept model did not fit our data, as we learned by making graphs of the average survey response and fitted curves for the different income categories within each state. We had to expand to a varying-intercept, varying-slope model: logit $\Pr(y=1) = a_s + b_s x$, in which the slopes b_s varied by state as well. This model expansion led to a corresponding expansion in our understanding: we learned that the gap in voting between

[7] See the earlier footnote for an explanation why one cannot simply augment Bayesian inference with a “catchall” hypothesis that represents the space of possible alternatives to the models being studied. In short, there is generally now way to set up such a catchall as a probability model; and if it were possible to formulate a catchall hypothesis, it would (in all the applications we have ever worked on) have a probability of 1: all the models we have ever used have been wrong.

rich and poor is much greater in poor states such as Mississippi than in rich states such as Connecticut. Thus, the polarization between rich and poor voters varied in important ways geographically.

We found this not through any process of Bayesian induction as usually defined (via the computation of the posterior probabilities of competing hypotheses, paradigms, or scientific models) but rather through model checking. Bayesian inference was crucial, not for computing the posterior probability that any particular model was true--we did not do that--but in allowing us to fit rich enough models in the first place that we could study state-to-state variation, incorporating in our analysis relatively small states such as Mississippi and Connecticut that did not have large samples in our survey.

Life continues, though, and so do our statistical struggles. After the 2008 election, we wanted to make similar plots, but this time we found that even our more complicated logistic regression model did not fit the data--especially when we wanted to expand our model to estimate voting patterns for different ethnic groups. Comparison of data to fit led to further model expansions, leading to our current specification, which uses a varying-intercept, varying-slope logistic regression as a baseline but allows for nonlinear and even non-monotonic patterns on top of that. Figure 1 shows some of our inferences in map form, and Figure 2 shows some of the data and model fit.

Again, the power of Bayesian inference is deductive: given the data and some model assumptions, it allows us to make lots of inferences, many of which can be checked and potentially falsified. For example, look at New York State (in the bottom row of Figure 2): apparently, voters in the second income category supported John McCain much more than did voters in neighboring income groups in that state. This pattern is possible but it arouses suspicion. A careful look at the graph reveals that this is a pattern in the raw data which was moderated but not entirely smoothed away by our model. The natural next step could be to examine data from other surveys. We may have exhausted what we can learn from this particular dataset, and Bayesian inference was a key tool in allowing us to do so.

Inductive inference and the Kuhnian paradigm

We see Bayesian data analysis--or, really, applied statistical analysis in general--as fitting well into the falsificationist approach--as long as we recognize that data analysis includes model checking (as in Chapter 6 of Gelman et al., 2003) as well as inference (which, as far as we can see, is purely deductive--proceeding from assumptions to conclusions). Yes, we “learn,” in a short-term sense, from Bayesian inference--updating the prior to get the posterior--but this is more along the lines of what a Kuhnian might call “normal science.” The real learning comes in the model-checking stage, when we can reject a model and move forward. The inference is a necessary stage in this process, however, as it creates the strong conclusions that are falsifiable.

We don't see Bayesian statistics as subjective (any more than any science is subjective in the choice of problems to study and data to focus on). We see a Bayesian model as a clear set of assumptions, whose implications are then deductively evaluated. The assumptions can be subjective but they need not be--except to the extent that all statistical procedures are “subjective” in requiring some choice of what to do.

In our view, a key part of model-based data analysis is model checking. This is where we see the link to falsification. To take the most famous example, Newton's laws of motion really have been falsified. And we really can use methods such as chi-squared tests to falsify little models in statistics. Now, once we've falsified, we have to decide what to do next, and that isn't obvious. A falsified model can still be useful in many domains (once again, Newton's laws are the famous example). But we like to know that it's falsified.

In our own philosophy of statistics (derived, in part, from our own readings, interpretations, and extrapolations of Jaynes, Popper, and Lakatos), the point of falsifying a model is not to learn that the model is false--certainly, all models that we've ever considered are false, which is why the chi-squared test is sometimes described as a “measure of sample size”--but rather to learn the ways in which a model is false. Just as with exploratory data analysis, the goal is to learn about aspects of reality not captured by the model (see Gelman, 2003, for more on this).

So, yes, the goal of “falsification” (as we see it), is not to demonstrate falsity but rather to learn particular aspects of falsity. We are interested in learning about flaws in a model without needing a new model to compare it to.

A standard view of Bayesian model comparison is that you just throw a new model into your class of explanations, and see what comes out having the best posterior odds. This doesn't really work for us, at least not in the problems we've worked on.

Example: Estimating the effects of legislative redistricting

Our stories of model rejection and new models are more on the lines of: We fitted a model comparing treated and control units (in the particular example that comes to mind, these are state legislatures, immediately after redistrictings or not), and assumed a constant treatment effect (in this case, parallel regression lines in “after” vs. “before” plots, with the treatment effect representing the difference between the lines). We made some graphs and realized that this model made no sense. The control units had a much steeper slope than the treated units. We fit a new model, and it had a completely different story about what the treatment effects meant. The graph falsified the first model and motivated us to think of something better. The graph for the new model with interactions is shown in Figure 3. For us, falsification is about plots and predictive checks. Not about Bayes factors or posterior probabilities of candidate models.

Connections to Popper’s falsificationism

To get back to the philosophers: We suspect that our Popperianism follows the ideas of an idealized Popper (following Lakatos, 1978, who introduced the concepts of Popper_0, Popper_1, and Popper_2 to capture different extrapolations of his mentor’s ideas).

Apparently the actual Popper didn't want to recognize probabilistic falsification. When one of us read (or attempted to read) *The Logic of Scientific Discovery* 20 years or so ago, we skipped over the probability parts because they seemed full of utopian and probably unrealizable ideas such as philosophical definitions of randomness. But the idea of falsification--and the dismissal of “inductive inference”--well, that part resonated. We’re on Popper's side in not believing in “induction” as a mode of inference. We don't mind “induction” in the sense of prediction and minimum-description-length (with more data in a time series, we should be able to better form an accurate prediction rule using fewer bits to describe the rule), but "induction" doesn't fit our understanding of scientific (or social-scientific) inference.

The main point where we disagree with many Bayesians is that we do *not* think that Bayesian methods are generally useful for giving the posterior probability that a model is true, or the probability for preferring model A over model B, or whatever. Bayesian inference is good for deductive inference within a model, but for evaluating a model, we prefer to compare it to data (what Cox and Hinkley, 1974, call “pure significance testing”) without requiring that a new model be there to beat it.

The fractal nature of scientific revolutions

With all this discussion of Kuhn and scientific revolutions, we have been thinking of the applicability of these ideas to our own research experiences.

At the risk of being trendy, we would characterize scientific progress as self-similar (that is, fractal). Each level of abstraction, from local problem solving to big-picture science, features progress of the “normal science” type, punctuated by occasional revolutions. The revolutions themselves have a fractal time scale, with small revolutions occurring fairly frequently (every few minutes for an exam-type problem, up to every few years or decades for a major scientific consensus). This is related to but somewhat different from the fractality-in-subject-matter discussed by Abbott (2001).

At the largest level, human inquiry has perhaps moved from a magical to a scientific paradigm. Within science, the dominant paradigm has moved from Newtonian billiard balls, to Einsteinan physics, to biology and neuroscience and perhaps to developments such as nanotechnology. Within, say, psychology, the paradigm has moved from behaviorism to cognitive psychology. But even on smaller scales, we see paradigm shifts. For example, in working on an applied research or consulting problem, we typically will start in a certain direction, then suddenly realize we were thinking about it wrong, then move forward, and so forth. In a consulting setting, this reevaluation can happen several times in a couple of hours. At a slightly longer time scale, we commonly reassess our approach to an applied problem after a few months, realizing there was some key feature we were misunderstanding.

Thus, we see this normal-science and revolution pattern as fundamental. Which, we think, ties nicely into our perspective of deductive inference (Bayesian or otherwise) as normal science and model checking as potentially revolutionary.

In conclusion, we emphasize that this chapter is our attempt to connect modern Bayesian statistical practice with falsificationist philosophy of science, and does not represent anything like a comprehensive overview of the philosophy-of-science literature. Because we feel the status quo perception of Bayesian philosophy is wrong, we thought it more helpful to present our own perspective forcefully, with the understanding that this is only part of the larger philosophical picture.

References

Abbott, A. (2001). *Chaos of Disciplines.* University of Chicago Press.

Beller, M. (2000). *Quantum dialogue*. Oxford University Press.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. New York: Chapman and Hall.

Dawid, A. P. (2004). Probability, causality and the empirical world: A Bayes-de Finetti-Popper-Borel synthesis. *Statistical Science* 19, 44-57.

Day, T., and Kincaid, H. (1994). Putting inference to the best explanation in its place. *Synthese* 98, 271-295

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 2, 369-382.

Gelman, A. (2008). Objection to Bayesian statistics (with discussion). *Bayesian Analysis* 3, 445-478.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A., and King, G. (1994). Enhancing democracy through legislative redistricting. *American Political Science Review* 88, 541-559.

Gelman, A., Lee, D., and Ghitza, Y. (2010). A snapshot of the 2008 election. *Statistics, Politics, and Policy*.

Gelman, A., Park, D., Shor, B., Bafumi, J., and Cortina, J. (2008). *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do.* Princeton University Press.

Gelman, A., Shor, B., Park, D., and Bafumi, J. (2008). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science* 2, 345-367.

Greenland, S. (1998). Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology* 27, 543-548.

Jaynes, E. T. (1983). *Papers on Probability, Statistics, and Statistical Physics*, ed. R. D. Rosenkrantz. Dordrecht, Netherlands: Reidel.

Jaynes, E. T. (1996). *Probability Theory: The Logic of Science.* Cambridge University Press.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*, second edition. University of Chicago Press.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.

McCarty, N., Poole, K. T., and Rosenthal, H. (2006). *Polarized America: The Dance of Ideology and Unequal Riches.* MIT Press.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Sober, E. (1991). *Reconstructing the Past*. MIT Press.

Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics* 2, 492-494.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wikipedia (2009). Bayesian inference. Accessed 9 June 2010.

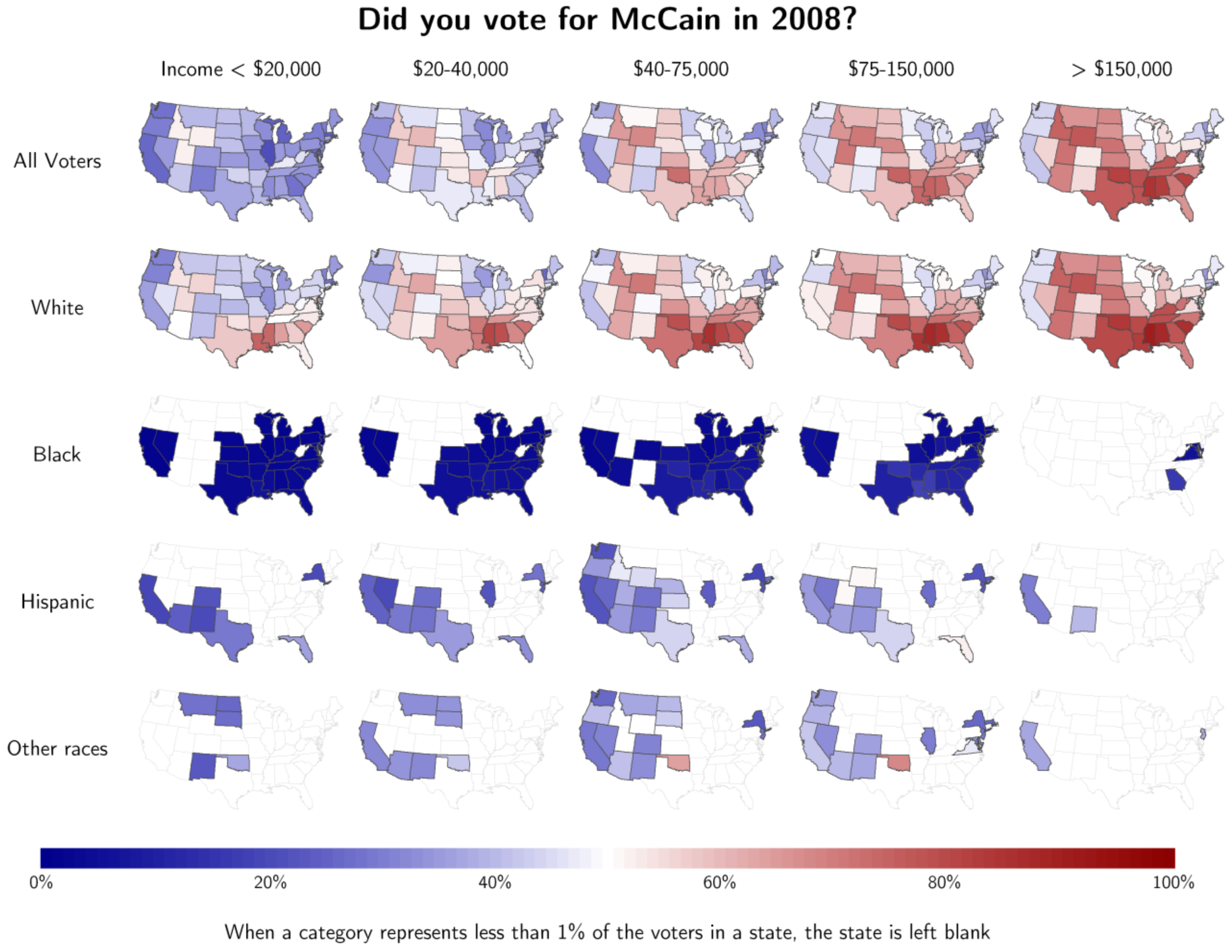
Did you vote for McCain in 2008?
Income < $20,000
$20-40,000
$40-75,000
$75-150,000
> $150,000
All Voters
White
Black
Hispanic
Other races
0%
20%
40%
60%
80%
100%
When a category represents less than 1% of the voters in a state, the state is left blank

Figure 1. Based on a model fit to survey data: states won by John McCain and Barack Obama among different categories of income and ethnicity (Gelman, Lee, and Ghitza, 2010). States colored deep red and deep blue indicate clear McCain and Obama wins; pink and light blue represent wins by narrower margins, with a continuous range of shades going to pure white for states estimated at exactly 50/50.

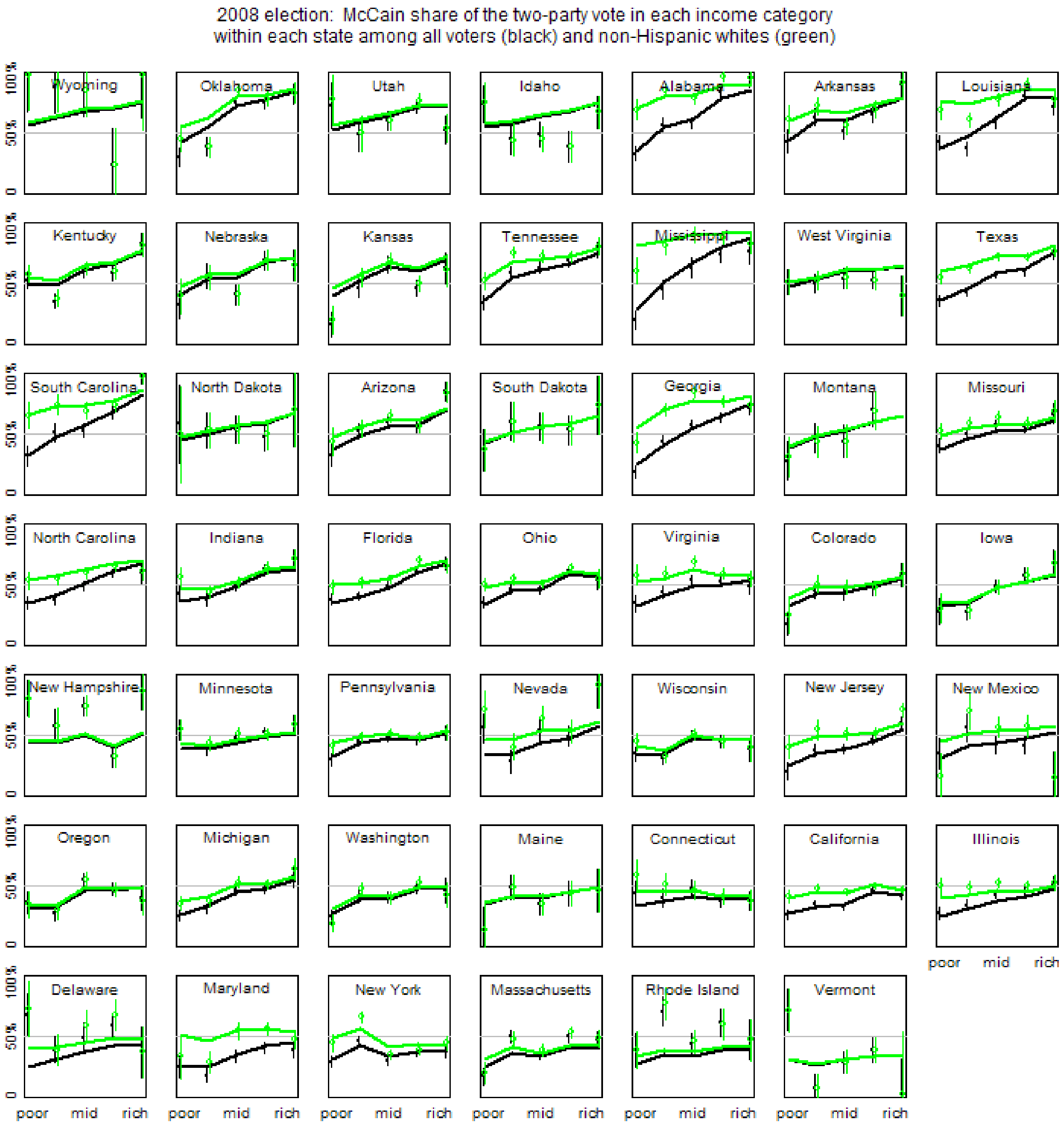

Figure 2. Data and fitted model used to make the maps shown in Figure 1. Dots are weighted averages from pooled June-Nov Pew surveys; error bars show +/- 1 standard error bounds. Curves are estimated using multilevel models and have a standard error of about 3% at each point. States are ordered in decreasing order of McCain vote (Alaska, Hawaii, and D.C. excluded). We fit a series of models to these data; only this last model fit the data well enough that we were satisfied. In working with larger datasets and studying more complex questions, we encounter increasing opportunities to check model fit and thus falsify in a way that is helpful for our research goals.

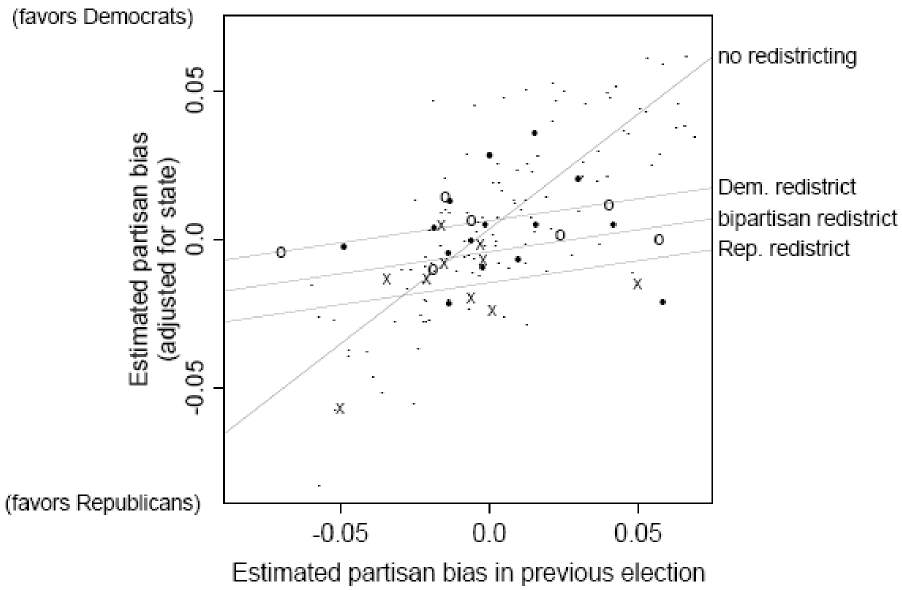

Figure 3. Effect of redistricting on partisan bias. Each symbol represents a state election year, with dots indicating controls (years with no redistricting) and the other symbols corresponding to different types of redistricting. As indicated by the fitted lines, the “before” value is much more predictive of the “after” value for the control cases than for the treated (redistricting) cases. The dominant effect of the treatment is to bring the expected value of partisan bias toward 0, and this effect would not be discovered with the usual approach, which is to fit a model assuming parallel regression lines for treated and control cases.

The relevance of this example to the philosophy of statistics is that we began by fitting the usual regression model with no interactions. Only after visually checking the model fit--and thus falsifying it in a useful way without the specification of any alternative--did we take the crucial next step of including an interaction, which changed the whole direction of our research. The paradigm shift was induced by a falsification--a bit of deductive inference from the data and the earlier version of our model.