Bayesian Analysis (2008) **3**, Number 3, pp. 467–478

Rejoinder

Andrew Gelman*

Abstract. In the main article I presented a series of objections to Bayesian inference, written in the voice of a hypothetical anti-Bayesian statistician. Here I respond to these objections along with some other comments made by four discussants.

Keywords: Comparisons to other methods, Foundations

1 Introduction

I wrote my anti-Bayesian rant in a fit of April Fool's Day inspiration, but now is the time to be serious. To start with, I appreciate the four thoughtful discussions, two of which are Bayesian and two of which are non-Bayesian, but none of which are anti-Bayesian. I would like to attribute this, possibly optimistically, to an advance in statistical culture. It is still possible to be non-Bayesian, but, given the advances in practical Bayesian methods in the past two decades, *anti-Bayesianism* is no longer a serious option. Of the four discussions, Kadane's is closest to my own views (with Bernardo's a close second), but I learned from and would like to comment on points raised by all four.

Before going on, I wish to make some comments about the term "frequentist," a word which I was careful to avoid in my original article but which was used by Bernardo, Senn, and Wasserman in their discussions. Wasserman writes, "a particular analysis can be Bayesian or frequentist." Larry is also in the machine learning department so I assume that when he uses the word "or," it includes "and" as well. I worry, however, that readers may erroneously interpret it as "exclusive or," so let me clarify. Bayesian inference (or, more generally, Bayesian data analysis) is a method for summarizing uncertainty and making estimates and predictions using probability statements conditional on observed data and an assumed model. Frequentist statistics, as I understand the term, is an approach for evaluating statistical procedures conditional on some family of posited probability models.

In a nutshell: Bayesian statistics is about making probability statements, frequentist statistics is about evaluating probability statements. As Wasserman points out (following Box and Rubin before him), Bayesians can feel free, or even obliged, to evaluate the frequency properties of their procedures. Conversely, as Efron, Morris, Little, and others have pointed out, frequentists can feel free to use Bayesian methods to derive statistical procedures with good frequency properties. So, not only can a statistician

*Department of Statistics and Department of Political Science, Columbia University, New York, N.Y., http://www.stat.columbia.edu/~gelman

 DOI:10.1214/08-BA318REJ

be Bayesian and frequentist at different times, a single analysis can also be both at the same time.

That said, I have to admit that I don't spend a lot of time evaluating the frequency properties of my Bayesian procedures. It would be diplomatic of me to follow Ronald Reagan and say "Trust, but verify," recommending Bayesian methods with frequentist evaluations, but I usually don't get around to that. My actual practice is closer to "Trust, but check": assume a model, fit it, then check its fit to existing and future data.

One reason for this is that I typically work with hierarchical models, which have internal replication (multiple groups) which allows group-level parameters to be estimated and group-level models to be fit. The concern then comes in with hyperparameters, and indeed it is possible for casual Bayesian analyses to have poor frequency properties (see the discussion in Gelman, 2006, of prior distributions for hierarchical variance parameters). Frequency properties in practice might not be so bad as in theory—if a method performs really horribly, we might be lucky enough to notice the problem and stop right then—but they are a real concern. When data become more complicated, we can take our Bayesian models to the next level and add another layer of hyperparameters. My point is that, at intermediate levels of a Bayesian model, frequency properties typically take care of themselves. It is typically only at the top level of unreplicated parameters that we have to worry.

This is a point that Ripley made in his book on spatial statistics, explaining how internal replication often allows spatial and time series analyses to work fine in settings where independent realizations of a process are not available. Replication is, in practice, central to Bayesian and to frequentist approaches, appearing as multilevel (hierarchical) modeling in the Bayesian context and as long run evaluation for the frequentists.

2 Responses to my own criticisms of Bayesian methods

As several discussants have noted, the objections to Bayesian methods in my original article were not entirely sincere. Or, to put it another way, these are sincere objections that I have thought through and, I believe, have largely been resolved. I will briefly dispose of them here by referring to the relevant sections of our book on practical Bayesian statistics, Gelman et al. (2003). I refer incessantly to this book because it best reflects my own thinking in this area; similar and perhaps clearer expositions appear in the books of Berger (1985) and Carlin and Louis (2008).

General objections (from Section 2 of my article, "Objections to Bayesian statistics"):

- "Bayesian methods are presented as an automatic inference engine": As we discuss on page 3 of our book, Bayesian data analysis has three stages: formulating a model, fitting the model to data, and checking the model fit. The second step—inference—gets most of the attention, but the procedure as a whole is not automatic. The challenge comes in constructing realistic models and in assessing their fit.

- "As scientists we should be concerned with objective knowledge rather than subjective belief": As Bernardo notes, Bayesian priors are objective in the same sense that classical methods are, "in that the final result only depends on the model assumed and the data obtained." Prior distributions can be informative while still being constructed from objective data: for two examples (football point spreads and accuracy of record linkage), see Sections 1.6 and 1.7 of *Bayesian Data Analysis*.

- "Bayesian methods seem to quickly move to elaborate computation": For better or worse, computation is becoming central to all statistical methods nowadays. Even the simplest methods become computationally intense when applied to gigabytes of data. I think the association of Bayesian methods with computation mostly arises from the overlay of the two time trends of increasing computation and increasing use of applied Bayesian methods. But I agree that right now we are at a point where we can easily bite off in modeling more than we can chew in computation.

Specific objections (from Section 3 of my article):

- "Subjective prior distributions don't transfer well from person to person": Again, I would distinguish between "informative" and "subjective." An informative prior distribution can be based on data relevant to a specific problem (as in Sections 1.6 and 1.7 of our book, as mentioned above), or its properties can be evaluated by cross-validation on a corpus of datasets (Gelman et al., 2008).

- "There's no good objective principle for choosing a noninformative prior ... Where do prior distributions come from, anyway?": A quick answer is that there is no general principle for choosing a likelihood either. Where did all those logistic regressions come from? From their different perspectives, Bernardo and Kadane both emphasize that, like all science, statistics is made of subjective procedures that yield objectively testable results.

- "Why should I believe your subjective prior? If I really believed it, then I could just feed you some data and ask you for your subjective posterior. That would save me a lot of effort!": I agree that this criticism reveals a serious incoherence with the subjective Bayesian framework as well with in the classical utility theory of von Neumann and Morgenstern (1947), which simultaneously demands that an agent can rank all outcomes a priori and expects that he or she will make utility calculations to solve new problems.

  The resolution of this criticism is that Bayesian inference (and also utility theory) are ideals or aspirations as much as they are descriptions. If there is serious disagreement between your subjective beliefs and your calculated posterior, then this should send you back to re-evaluate your model.

- "Bayesian theory requires a great deal of thought about the given situation to apply sensibly": I won't give the snappy answer that it's good to think about "the

given situation" because I recognize—as a teacher, a textbook writer, a researcher, and a practitioner—that we use default methods all the time. Rather, I'll retreat to the usual Bayesian answer that our default methods perform as well or better than classical default methods. Maximum likelihood maps to a flat prior distribution, and we can do better as needed.

- "The priors I see in practice are typically just convenient conjugate forms": Indeed, when choosing priors (and likelihoods), researchers must balance the goals of realism and convenience. As Kadane points out, Bayesians can (and do) use mixtures of conjugate families to combine computational convenience with generality.

- "I like unbiased inference and I like confidence intervals that really have their advertised confidence coverage": Regarding unbiased estimation, see the example on pages 248–249 of *Bayesian Data Analysis*.[1] I agree with Senn that it is unfortunate that the term "bias" has been stuck with its particular technical meaning.

- Bayesian simulation "seems stuck in an infinite regress of inferential uncertainty": As is just about all modern statistics, unfortunately. Computational messes are the price we pay for analyzing large and complex data sets.

- "People tend to believe results that support their preconceptions and disbelieve results that surprise them. Bayesian methods encourage this undisciplined mode of thinking": As Kadane notes, new methods get misused, and it is the responsibility of all of us to point out mistakes so that others can walk an easier path. Regarding the particular issue of priors that support one's preconceptions: it all depends on the available data. As a Bayesian, you are free to use a weak prior as a placeholder, adding more information ("preconceptions") as needed. As Bernardo points out, decisions need to be made somehow, and the notion that some results "support" preconceptions and others "surprise" them reflects the very real sense that science is built upon models as well as data.

  There is a class of problems where, for ethical or security reasons, it is illegal or inappropriate to use all available information. For example, the Census is not allowed to release fine-grained cross-tabulations that could be used to deduce information about individual people; racial profiling cannot be used in making mortgage decisions; and students' Calculus 1 grades and SAT scores would not be used in determining their grades in Calculus 2, even though these pieces of

[1] The problem is to estimate θ, the height of an adult daughter, from her mother's height, y, given data showing mothers' and daughters' heights to be jointly normally distributed with means 160 centimeters, equal standard deviations, and correlation 0.5. The Bayesian estimate of θ given y is $\mathrm{E}(\theta|y) = 160 + 0.5(y - 160)$, the familiar "regression to the mean" that would be used by Galton and every statistician since then. But this is *not* an unbiased estimate of θ in the classical sense of $\mathrm{E}(\widehat{\theta}|\theta) = \theta$. In this problem, the actual unbiased estimate is $\widehat{\theta} = 160 + 2(y - 160)$, a truly ridiculous estimate that, instead of shrinking to the mean, doubles the error. This example is a cousin to the negative variance estimators and other messes left in the wake of "unbiasedness." As with the negative variance estimate, nobody would ever actually use the anti-shrinkage estimate of the daughter's height. I mention this example not to claim a defect in classical statistical *practice* but rather to dramatize the limitations of the *principle* of unbiasedness.

information could be informative. So there are settings where preconceptions don't get used, but this is more of a concern of law and ethics than of statistical inference.

- "Bayesian techniques motivate even the best-intentioned researchers to get stuck in the rut of prior beliefs": I'd prefer to say that Bayesian techniques allow prior beliefs to be tested (using posterior predictive checks, as discussed in chapter 6 of *Bayesian Data Analysis*) and discarded as appropriate.

- "Bayesianism assumes: (a) Either a weak or uniform prior, in which case why bother?, (b) Or a strong prior, in which case why collect new data?, (c) Or more realistically, something in between, in which case Bayesianism always seems to duck the issue": The fallacy here comes from the assumption of fixed data and a choice of prior distribution. More realistic, especially in the context of hierarchical models, is a fixed prior distribution that is applied to different data.

  Section 2.8 of *Bayesian Data Analysis* considers an example of estimation of rates of a rare cancer in different counties in the United States, with a Gamma(α, β) prior distribution representing the set of true cancer rates in the 3000 counties. Given a county with y cancers out of n people, the Bayes estimate of the underlying cancer rate is $(\alpha + y)/(\beta + n)$. For a county with large population, the Bayes estimate is approximately the raw rate, y/n (case (a) above). If the population is low, the estimate is close to the prior mean, α/β (case (b)). Otherwise, the estimate is a weighted average of the two. But this is not "ducking the issue," any more than least squares "ducks the issue" of finding a regression line that balances among all possible errors.

- Empirical and hierarchical Bayes methods "rely on an assumption of exchangeability": This was a big concern back in the 1970s when hierarchical models were first being systematically applied in statistics; see, for example, Oscar Kempthorne's discussion of Lindley and Smith (1972). My quick response is that classical alternatives to hierarchical modeling are also exchangeable in the sense of treating all the groups exchangeably. My longer response is that if information is available to distinguish the groups, then it can and should be added to the model—this is called a multilevel regression—so that it is only the errors that are exchangeable, just as in classical regression and just about every other statistical method.[2]

- "I'd rather let the data speak without applying a probability distribution to something like the 50 states which are neither random nor a sample": This is no worse than fitting a regression model to the 50 states. The errors in such a regression represent model errors or deviations from linearity, not random sampling.

[2]In my original article, I wrote, "The 50 states aren't exchangeable." From a distance, North and South Dakota may appear exchangeable, and so may Mississippi and Alabama. A fuller model could allow spatial autocorrelation and state-level predictors to capture this structure. But however the model is set up, the key is that we can and do include additional information in the group-level regression equation.

- "Hierarchical Bayes methods use the data twice": As Senn states, using the data to set the prior is cheating. But such methods can often be reformulated as hierarchical models which are fully Bayesian (conditioning on the data only once). The set-the-prior-from-data approach is best viewed as a possibly useful approximation (as Bernardo notes) or as a way of getting reasonable point estimates, as discussed by Efron and Morris in the 1970s.

- "A Bayes estimator is a statistical estimator that minimizes the average risk, but when we do statistics, we're not trying to 'minimize the average risk,' we're trying to do estimation and hypothesis testing": This I actually agree with (probably in disagreement with Kadane, Bernardo, and many other Bayesians). My own position here is explained in chapter 22 of *Bayesian Data Analysis*, where I present loss functions and decision analysis as being relevant to decisions but not to statistical inference. Bayesian inference is about creating the posterior distribution, which can then be used as appropriate. I don't see any role for squared error loss, minimax, or the rest of what is sometimes called "statistical decision theory."

- "If the Bayesian philosophy of axiomatic reasoning implies that we shouldn't be doing random sampling, then that's a strike against the theory right there": Luckily, the Bayesian philosophy does not imply this; as discussed in chapter 7 of *Bayesian Data Analysis*, Bayesian inference is perfectly consistent with classical sampling and experimental design. So we're off the hook here.

3 Responses to others' criticisms of Bayesian methods

Unfortunately the editor did not solicit comments from any of the extreme anti-Bayesians of the sort that I have encountered over the years (and to some extent parodied in my original article). However, Senn and Wasserman in their discussions do make a few comments about Bayesian methods to which I would like to respond.

Incoherence of Bayesian inference in practice

Senn writes that "adopting Jeffreys's [objective Bayes] approach to parameter estimation without adopting his approach to significance testing is a recipe for disaster ... The Jeffreys-subjective synthesis now common betrays a much more dangerous confusion than the Neyman-Pearson-Fisher synthesis as regards significance/hypothesis tests." I completely agree with Senn here and have no patience for statistical methods that, Jeffreys-like, assign positive probability to point hypotheses of the $\theta = 0$ type that can never actually be true (at least in my own lines of applied research) and then turn around and try to give subjective or decision-theoretic interpretations to posterior probabilities of hypotheses. Gelman and Rubin (1995) discuss some problems with this hybrid approach for social science problems. I prefer to work within a model, checking and improving it as necessary.

Senn follows up by criticizing subjective Bayes as being "self-contained and logical,

but . . . impossible to apply." I agree (hence my comment that, if you really believe in subjective priors, you can cut out the middleman of statistical analysis and jump straight to your subjective posterior). However, the mathematics of subjective probability works well in combining information from multiple sources, hence the move from "subjective Bayes" to "hierarchical Bayes."

Finally, Senn writes, "the gloomy conclusion to which I am drawn on reading de Finetti (1974) is that ultimately the Bayesian theory is destructive of any form of public statistics." My quick response here is that a lot of progress has been made since 1974! De Finetti made great contributions but we are allowed to move forward from the methods of thirty-five years ago. I fear that one of the consequences of Bayesian statistics being given a proper name is that it encourages too much historical deference from people who think that the bibles of Jeffreys, de Finetti, Jaynes, and others have all the answers. To reply more specifically to Senn's comment, I think of Bayesian inference as a generalization of least squares and maximum likelihood, with prior distributions and multilevel models (which can also be viewed as simultaneous equations or measurement-error models) as a way of regularizing or obtaining more stable estimates. Public statistics performed this way can be more effective than classical estimates that commonly need to be interpreted with one eye on the sample size.

Frequency evaluation in theory and practice

Wasserman contrasts Bayesian and frequentist approaches but, I fear, is a bit too confident about what classical statistical methods can do.

He writes, "The particle physicists have left a trail of such confidence intervals in their wake. Many of these parameters will eventually be known (that is, measured to great precision). Someday we can count how many of their intervals trapped the true parameter values and assess the coverage. The 95 percent frequentist intervals will live up to their advertised coverage claims. A trail of Bayesian intervals will, in general, not have this property."

As a Bayesian, I won't try to predict the future with the certainty that Larry claims. But I can look at the past, and in fact the frequentist intervals of physical constants did *not* in general live up to their advertised coverage claims (see Youden, 1962, and Henrion and Fischhoff, 1986). It is possible that researchers are more careful now then they were in the 1950s, but I would guess that systematic error will always be with us, a point that in recent years has been emphasized by Sander Greenland.

What happened here? Maybe the Law of Large Numbers and the Central Limit Theorem weren't actually true back then? No. The mathematics was fine, but the models were not. A little thing called systematic error got in the way. Or, to put it in Bayesian terms, the assumed likelihoods were wrong, and, as a result, fewer than 95% of the 95% intervals contained the true values.

In fairness, I agree with Larry that Bayesian inferences would probably not have the advertised coverage either. Like the frequentist intervals, they're only as good as their

models.

Wasserman pithily writes, "Frequentist methods have coverage guarantees; Bayesian methods don't. In science, coverage matters." I think I'll dispute all three of these claims. I can dispose of the first two with a reference to Agresti and Coull (1998), who show that, in the case of inference for a binomial probability, Bayesian inference based on simple estimates such as $(y + 1)/(n + 2)$ have better frequency properties than the so-called Fisher exact test.

To give a slightly longer response: I'm not quite sure what a "frequentist method" is, but I will assume that the term refers to any statistical method for which a frequency evaluation has been performed (analytically or via simulation) conditional on some family of models that is considered relevant to the frequentist doing the evaluation. In any case, frequentist methods have coverage guarantees in some simple problems, but in general there is no coverage guarantee because frequency properties depend on nuisance parameters which can only be ignored in some special cases of pivotal test statistics.

I also dispute Wasserman's claim that "In science, coverage matters." To borrow a line from Bernardo, a Bayesian could reply: "In science, nonnegative variance parameters matter" or "In science, it matters that 95% confidence intervals not contain the whole real line." Unfortunately, when we deal with scientists, statisticians are often put in a setting reminiscent of Arrow's paradox, where we are asked to provide estimates that are informative and unbiased and confidence statements that are correct conditional on the data and also on the underlying true parameter. Larry Wasserman feels that scientists are truly frequentist, and Don Rubin has told me how he feels that scientists interpret all statistical estimates Bayesianly. I have no doubt that both Larry and Don are correct. Voters want lower taxes and more services, and scientists want both Bayesian and frequency coverage; as the saying goes, everybody wants to go to heaven but nobody wants to die.

Finally, I disagree with Wasserman's statement that randomization methods "don't really have a place in Bayesian inference." Beyond Kadane's discussion and reference, I'd also point the reader to chapter 7, in particular section 7.6, of *Bayesian Data Analysis*. Given the occasional excesses of Bayesian writings over the years, I can see why it makes sense to treat Bayes with at least the same skepticism as is given to other statisitcal methods. But please don't tell me that something "doesn't really have a place" when it's sitting right there in chapter 7 of our book!

I have to admit, though, that cross-validation is a more difficult question. I've often used cross-validation in my own work—there are times where it convinces like nothing else—but I struggle to figure out how it fits into any formal statistical theory (Bayesian, frequentist, or otherwise). My current thought is that there may be a link through hierarchical modeling: a key concept in cross-validation is that the model that works for part of the data may not be best for another part. As Kadane writes, I hope that more Bayesians will take up the challenge, etc., perhaps following the decision-theoretic framework of Vehtari and Lampinen (2002).

4 Model-based vs. non-model-based approaches to statistics

Hal Stern once told me that the real dividing line in statistics is not Bayesian vs. frequentist but rather model-based vs. non-model-based. Another way I've heard it is, generative vs. non-generative models. I'd like to compare the strengths and weaknesses of these approaches, first starting from one more example from Wasserman's article and then more generally.

In Section 3 of his discussion, Wasserman presents a simple and general example that illustrates the best and worst of non-Bayesian methods. He recommends estimating a univariate distribution function by the sample distribution of n data points. If the goal is to estimate the median of the distribution, this is a reasonable procedure, losing some efficiency but gaining some robustness compared to model-based methods.

But the empirical distribution might not work so well for estimating the mean or the upper 1% of the distribution. Consider insurance problems, for example. As the history of insurance can tell us, estimates of the upper 1% can be wrong. But in this setting, giving up is not an option. The modern Bayesian approach is to propose a model and go from there, checking the model where possible.

Wasserman's example illustrates how classical methods can sometimes work well while requiring less effort than the corresponding Bayesian approach. Similarly, I have several times used jackknife and bootstrap methods in my own work (for example, Alpert et al., 1991, Lu and Gelman, 2003, and Casella, Ehrenberg, Gelman, and Shen, 2008), for problems where I didn't see the point of setting up a full probability model. And I've used flat prior distributions on regression coefficients even when I knew that some Bayesian shrinkage would be appropriate, just because I wasn't set up to create the Bayesian model (for example, Gelman, Stevens, and Chan, 2003). But I'm not proud of this, and I realize I could do better by including more information in my models.

However, I don't quite understand Wasserman's statement, "If the Bayes estimator has good frequency behavior then we might as well use the frequentist method." As far as I know, there is no "frequentist method" for coming up with an estimator. As noted in section 1 above, the frequentist method, as I understand it, is an approach for evaluating inferences—in which case, I have no problem with Wasserman or anyone else taking a Bayesian inference and labeling it as "frequentist." The practical point here is that the Bayesian approach is a useful way to come up with an estimator in complicated problems with structured data.

Another way to distinguish the two approaches to statistics is not in terms of what they do but rather what they'd like to do. Consider two inferential strategies:

1. Set up a strong model with many assumptions. The ultimate goal of the statistical analysis might be to reject the model and replace it with something better. In the terminology of Kuhn (1969), Bayesian inference conditional on the model is "normal science," and rejection through posterior predictive checking is the stuff of

"scientific revolutions." The key here is: the stronger the model, the more directly it can be falsified. I view this as a unification of the Popperian and Kuhnian philosophies of science; for the present discussion, what is relevant is that, in this framework, it is a positive feature of a model to have strong assumptions.

2. Assume no model or a partial model or only a model for the data collection, not the data-generating process. Methods here include second-order inference, proportional hazards models, various signal processing approaches (for example, wavelet shrinkage and lasso regression), and the jackknife and bootstrap. The key idea here is: any model will certainly be wrong, so it's best to anticipate this and develop robust methods that perform well without strong assumptions.

Debates between these two philosophies echo beyond statistics into other fields. For example, the mainstream of econometrics seems to me to follow the second approach, but Heckman (2007) argues the opposite, that it is through strong assumptions that we learn about social science. Biomedical statistics features, at one extreme, elaborate multi-compartment pharmacological models and, at the other, generalized estimating equations that fit models to hierarchical data structures while avoiding modeling the individual data points. In the long run there may be a synthesis of highly complex models that have many of the features of nonparametric approaches (for example, the additive regression trees of Chipman, George, and McCulloch, 2005), but for now I am content to recognize the successes of both these schools of statistics. I plan to continue doing most of my work using the Bayesian paradigm because it presents a more continuous connection between the data, models, and phenomena that I study.

Additional references

Agresti, A., and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Alpert, N., Barker, W. C., Gelman, A., Weise, S., Senda, M., and Correia, J. A. (1991). The precision of positron emission tomography: theory and measurement. *Journal of Cerebral Blood Flow and Metabolism* **11**, A26–30.

Carlin, B. P., and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*, third edition. Boca Raton, Fla.: CRC Press.

Casella, A., Ehrenberg, S., Gelman, A., and Shen, J. (2008). Protecting minorities in binary elections: A test of storable votes using field data. National Bureau of Economic Research Working Paper 14103.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2005). Bayesian additive regression trees. Technical report.

Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors: An empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

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

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, to appear.

Gelman, A., and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. Discussion of "Bayesian model selection in social research," by A. Raftery. *Sociological Methodology 1995*, 165–173.

Gelman, A., Stevens, M., and Chen, V. (2003). Regression models for decision making: a cost-benefit analysis of incentives in telephone surveys. *Journal of Business and Economic Statistics* **21**, 213–225.

Greenland, S. (2008). Hierarchical models for biases in observational studies. Technical report, Department of Epidemiology, University of California, Los Angeles.

Heckman, J. J. (2007). Rejoinder: Response to Sobel. *Sociological Methodology*, 135–162.

Henrion, M., and Fischoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics* **84**, 791–798.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lu, H., and Gelman, A. (2003). A method for estimating design-based sampling variances for surveys with weighting, post-stratification, and raking. *Journal of Official Statistics* **19**, 133–151.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Ripley, B. D. (1981). *Spatial Statistics*. New York: Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**, 267–288.

Vehtari, A., and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation* **14**, 2439–2468.

von Neumann, J., and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

Youden, W. J. (1962). *Experimentation and Measurement*. National Science Teachers' Association. Reprinted (1997) as National Institute of Science and Technology Special Publication 672. U.S. Department of Commerce.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.