RMM Vol. 2, 2011, 67–78
Special Topic: Statistical Science and Philosophy of Science
Edited by Deborah G. Mayo, Aris Spanos and Kent W. Staley
http://www.rmm-journal.de/

Andrew Gelman

Induction and Deduction in Bayesian Data Analysis*

Abstract:

The classical or frequentist approach to statistics (in which inference is centered on significance testing), is associated with a philosophy in which science is deductive and follows Popper's doctrine of falsification. In contrast, Bayesian inference is commonly associated with inductive reasoning and the idea that a model can be dethroned by a competing model but can never be directly falsified by a significance test. The purpose of this article is to break these associations, which I think are incorrect and have been detrimental to statistical practice, in that they have steered falsificationists away from the very useful tools of Bayesian inference and have discouraged Bayesians from checking the fit of their models. From my experience using and developing Bayesian methods in social and environmental science, I have found model checking and falsification to be central in the modeling process.

1. The Standard View of the Philosophy of Statistics, and Its Malign Influence on Statistical Practice

Statisticians can be roughly divided into two camps, each with a clear configuration of practice and philosophy. I will divide some of the relevant adjectives into two columns:

Frequentist	Bayesian
Objective	Subjective
Procedures	Models
P-values	Bayes factors
Deduction	Induction
Falsification	Pr (model is true)

This division is not absolute—in particular, it does not capture my own mix of philosophy and statistical practice—but it reflects a common alignment of beliefs. I shall call it the *standard view of the philosophy of statistics* and abbreviate it as S.

* We thank Cosma Shalizi, Deborah Mayo, and Phil Price for helpful discussions and the National Science Foundation for grants SES-1023176 and SES-1023189, Institute of Education Sciences for grants R305D090006-09A and ED-GRANTS-032309-005, Department of Energy for grant DE-SC0002099, and National Security Agency for grant H98230-10-1-0184.

The point of this article is that S is a bad idea and that one can be a better statistician—and a better philosopher—by picking and choosing among the two columns rather than simply choosing one.

S is neither a perfect classfier nor does it encompass all of statistical practice. For example, there are objective Bayesians who feel comfortable choosing prior distributions via the same sorts of theory-based principles that non-Bayesians have commonly used when choosing likelihoods. And there are concepts and methods such as robustness, exploratory data analysis, and nonparametrics which do not traditionally fall in one or the other column. Overall, though, S represents a set of associations that are commonly held—and I argue that they are influential, and can be harmful, even to those practically-minded folk who have no interest in foundations or philosophy. The economist (and Bayesian) John Maynard Keynes famously remarked, "even the most practical man of affairs is usually in the thrall of the ideas of some long-dead economist", and this is the case for philosophy as well.

Before explaining why I think that S, the above two-column categorization of statistics, is *wrong*, I will argue that it has been *influential*. (After all, it would be silly for an applied statistician to fight against a scheme that had no practical implications.)

I see two distinct ways in which S has hurt statistical practice. From one direction, consider the harm done to falsificationists—those scientists influenced (correctly, in my view) by the ideas of Karl Popper and his followers to favor an approach to inference in which hypotheses can only be rejected, not confirmed, along with an objective framework in which informed scientific consensus is held to reflect reality rather than the reverse. S has influenced these falsificationists to disdain Bayesian inference in favor of the Neyman-Pearson or Fisherian approaches to hypothesis testing. In light of the many applied successes achieved by Bayesian methods in recent decades,[1] it seems a pity to abandon such a powerful approach because of philosophical qualms—especially if, as I argue here, these qualms are misplaced.

The second way in which I believe S has harmed statistics is by blinding many Bayesians to the benefits of predictive model checking. I vividly remember going from poster to poster at the 1991 Valencia meeting on Bayesian statistics and hearing from their presenters that, not only were they not interested

[1] Progress in statistical methods is uneven. In some areas the currently most effective methods happen to be Bayesian, while in other realms other approaches might be in the lead. The openness of research communication allows each side to catch up: any given Bayesian method can be interpreted as a classical estimator or testing procedure and its frequency properties evaluated; conversely, non-Bayesian procedures can typically be reformulated as approximate Bayesian inferences under suitable choices of model. These processes of translation are valuable for their own sake and not just for communication purposes. Understanding the frequency properties of a Bayesian method can suggest guidelines for its effective application, and understanding the equivalent model corresponding to a classical procedure can motivate improvements or criticisms of the model which can be translated back into better understanding of the procedures. From this perspective, then, a pure Bayesian or pure non-Bayesian is not forever doomed to use out-of-date methods, but at any given time the purist will be missing some of the most effective current techniques.

in checking the fit of the models, they considered such checks to be illegitimate. To them, any Bayesian model necessarily represented a subjective prior distribution and as such could never be tested. The idea of testing and p-values were held to be counter to the Bayesian philosophy. Bayesians have become more flexible in recent years but I still see some resistance to checking model fit.

It is not only Bayesians who avoid model checking. Quantitative researchers in political science, economics, and sociology (to name just three fields with which I happen to be familiar) regularly fit elaborate models without even the thought of checking their fit. Sometimes there is a bit of data exploration beforehand to suggest possible transformations and maybe tests of one or two assumptions (for example, checking for autocorrelation or clustering), but rarely the sort of open-ended exploration devoted to learning the limitations of a fitted model.

Model checking plays an uncomfortable role in statistics. A researcher is typically not so eager to perform stress testing, to try to poke holes in a model or estimation procedure that may well represent a large conceptual and computational investment. And the model checking that is done is often invisible to the reader of the published article. Problems are found, the old model is replaced, and it is only the new, improved version that appears in print.

So I am not claiming that S alone, or even mostly, has stopped statisticians from checking their models. But I do feel that S has served as a justification for many Bayesians to take the easy way out and, as a result, miss out on one of the most useful ideas in statistics: that it is possible to reject a model using the very data to which the model has been fit.

In short, S has led many falsificationists and others who are interested in objective scientific knowledge to shun Bayesian methods, and S has led many Bayesians to shun falsification.

2. The Standard View of the Philosophy of Statistics Does Not Describe How I Do Statistics

At the center of S is the view that Bayesian inference represents inductive reasoning about scientific hypotheses. Here is how Wikipedia puts it (at the time of this writing):[2]

> "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

[2] We use Wikipedia here not as an authoritative source but rather as a reflection of a general consensus. My views on induction, deduction, and Bayesian inference are not in agreement with this consensus, and so it is my duty to explain, first, why my views are right and the consensus wrong, and, second, if the consensus is so clearly wrong, how so many could intelligent people hold to it.

> 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 does not describe what I do in my applied work. I do go through models, sometimes starting with something simple and building up from there, other times starting with my first guess at a full model and then trimming it down until I can understand it in the context of data. And in any reasonably large problem I will at some point discard a model and replace it with something new (see Gelman and Shalizi 2011a,b, for more detailed discussion of this process and how it roughly fits in to the philosophies of Popper and Kuhn).

But I do not make these decisions on altering, rejecting, and expanding models based on the posterior probability that a model is true. Rather, knowing ahead of time that my assumptions are false, I abandon a model when a new model allows me to incorporate new data or to fit existing data better.

At a technical level, I do not trust Bayesian induction over the space of models because the posterior probability of a continuous-parameter model depends crucially on untestable aspects of its prior distribution. (For any parameters that are identifiable by the data, the behavior of the prior in the far tails of the distribution is irrelevant to inference within the model but can have arbitrarily large effects on the model's marginal posterior probability.)

At a philosophical level, I have been persuaded by the arguments of Popper (1959), Kuhn (1970), Lakatos (1978), and others that scientific revolutions arise from the *identification and resolution of anomalies*. In statistical terms, an anomaly is a misfit of model to data (or perhaps an internal incoherence of the model), and it can be identified by a (Fisherian) hypothesis test without reference to any particular alternative (what Cox and Hinkley 1974 call "pure significance testing"). True, one might argue that finding an anomaly via a graphical or numerical data summary involves at the very least some implicit consideration of alternatives—a sense of what directions of misfit are potentially interesting or important. But such an implicit choice of directions in which to test is a far cry from the fully specified probability model that would be required to perform a Bayes factor. At the next stage, we see science—and applied statistics—as resolving anomalies via the creation of improved models which often include their predecessors as special cases. This view corresponds closely to the error-statistics idea of Mayo (1996).

Where does induction fit into this story? Popper has argued (convincingly, in my opinion) that scientific inference is not inductive but deductive, that the way we generalize from particular cases is through the medium of models, and that

inference within a model is deductive (see also Greenland 1998). Our key departure from the mainstream Bayesian view (as expressed, for example, in the Wikipedia excerpt above) is that we do not attempt to assign posterior probabilities to models or to select or average over them using posterior probabilities. Instead, we use predictive checks to compare models to data and use the information thus learned about anomalies to motivate model improvements.

3. The Stability of the Consensus View of Bayesian Induction

If the falsificationist Bayesian view is so evidently correct (as I believe)—or, at the very least, not evidently wrong—how is it that the consensus of philosophically-minded statisticians and statistically-minded philosophers is so different. Jeffreys, Savage, De Finetti, and their modern followers ... these guys are not dumb!

I have a few thoughts on how these thoughtful, intelligent researchers have come to a philosophical position much different from mine.[3] Most important, at a practical level their methods work.[4] You can use Bayes factors to select and average over models, and when the resulting inferences make no sense, you can change the models. Many aspects of model checking can be performed informally without the need for any p-values or significance levels.

My second explanation for the tenacity of the subjective Bayesian approach (in the face of the much-noted general tendency toward Popperian objectivism among working scientists) is simple logic: the argument made by Keynes, Ramsay, R. T. Cox, Neumann, and others from the 1920s through 1940s that any complete set of inferences must be either Bayesian or incoherent (see Savage 1954). I believe this argument fails because of the imperfections of any statistical model—Bayesian or otherwise—in real-world settings; nonetheless, it has been a powerful motivation for the subjective inductive philosophy.

It is a bit easier to understand the consensus in the other direction, in which frequentist statisticians have come to consider Bayesian inference as anti-falsificationist. Frequentists just took subjective Bayesians at their word and quite naturally concluded that Bayesians had achieved the goal of coherence only by abandoning scientific objectivity. Every time a prominent Bayesian published an article on the unsoundness of p-values, this became confirming evidence of the hypothesis that Bayesian inference operated in a subjective zone bounded by the prior distribution. You don't have to be Georg Cantor to realize that no prior distribution or set of prior distributions, no matter how carefully

[3] And I'm sure they have some theories of their own about how I could be so wrong!

[4] I believe my methods work better, at least for the sorts of problems I've studied in social and environmental sciences (see Gelman et al. 2003), but I recognize that different statistical methods work well in different problems (Gelman 2011a). Again, my ultimate argument in the present article is not that my philosophical perspective is definitely correct but rather that it is not clearly wrong.

chosen, can sit there being ready for any data. Frequentists have reasonably concluded that Bayesians were unwilling to see their models falsified and have unfortunately not generally kept up with developments in the past fifteen years on Bayesian model checking.[5]

Bandyopadhyay and Brittan (2010) argue that neither subjective nor objective Bayesianism, as usually defined, are possible, in that it is unrealistic to imagine that prior probabilities represent personal degrees of belief *or* that they directly correspond to observable real-world frequencies. We agree and would make the argument more directly by pointing to the melange of normal distributions, Poisson distributions, Dirichlet processes, and the like that characterized statistics as it is actually practiced. It would be silly to suppose that these conventional choices represent subjective belief or objective reality. Given that even the top statisticians tend to construct their models Tinkertoy-style from previously manufactured parts, we are under a continuing obligation to check model fit.

4. Falsification and Bayesian Data Analysis

As a data analyst, the statistical methods I use most are graphical exploratory data analysis and Bayesian modeling. I do not think of confidence intervals as inverses of hypothesis tests. Rather, I tend to think of maximum likelihood and other classical estimation procedures as approximations to Bayesian posterior summaries, and I interpret exploratory data analysis and confirmatory hypothesis tests as graphical and numerical posterior predictive checks (see Gelman 2003, and chapter 6 of Gelman et al. 2003).

As a modeler, I'm comfortable with continuity. For example, I'd prefer to consider voters in the United States as falling on a continuous partisan scale rather than being discretely categorized as Republicans, independents, and Democrats. I recognize that survey responses will give us discrete data but I like to think of these as measures of a continuous underlying quantity. My preference for continuity is not shared by all. For example, many statisticians and social scientists when studying opinion will use clustering models to place people in discrete latent categories.

My preference for continuity and my experiences in applications have led me to want to include all predictors in a regression model. In the sort of social science problems I study, there are no true zeroes except by design or through a natural experiment, and I do not see the point of statistical methods that attempt to discover from data conditional independence patterns that cannot

[5] For example, Huber (2011) writes, "Bayesian statistics lacks a mechanism for assessing goodness-of-fit in absolute terms. [...] Within orthodox Bayesian statistics, we cannot even address the question whether a model Mi, under consideration at stage i of the investigation, is consonant with the data y." This statement reflects unawareness of posterior predictive checking, which is a fully Bayesian approach to checking the fit of a particular model to a particular dataset (see Gelman, Meng and Stern 1996 for a review).

exist (Gelman 2011b). Rather than say that a variable is zero or not, I'd rather use a Bayesian model that partially pools coefficients toward a larger model.

Moving to the philosophy of science, I follow Popper in believing that a model can be rejected, never accepted.[6] I will go even further and say that, realistically, all my models are wrong. The encouraging message of the present article is that we can have all the powerful data-analytic tools of Bayesian inference, and falsification too!

Here are two simple examples of Bayesian falsification (taken from Gelman et al., 2003), one theoretical and one applied. In the theoretical example, a series of 20 binary outcomes, 11000001111100000000, is modeled as independent with common probability θ, with a uniform prior distribution θ. For the data at hand, the questionable part of the model is the independence assumption, not the prior, and the model can be checked with an autocorrelation statistic. A simple runs test will work fine; we can define a test statistic T(y) to be the number of switches in the series. The observed T(y) = 3 for these data. We can perform a Bayesian test by first assuming the model is true, then obtaining the posterior distribution (in this case, θ |y ~ Beta (8,14)), and then determining the distribution of the test statistic under hypothetical replicated data under the fitted model. In practice we implement this sort of check via simulation,[7] which reveals in this case that the observed value of 3 switches is about one-third the number expected from the fitted model, with a p-value of 0.03 when considered as a two-sided test. The point of this example is that we can indeed check the fit of a model by comparing data to a fitted posterior distribution.

My second, applied, example comes from a regression model predicting elections for the U.S. Congress, given incumbency status and previous district-level election results. Concerned about outliers, we define a test statistic that is the proportion of district elections where the regression prediction is off by more

[6] I use 'Popper' to broadly represent the falsificationist approach to the philosophy of science expressed by Lakatos (1978). I am not well read in the philosophical literature, and I recognize that the ideas of falsification, scientific inference, research programmes, and scientific revolutions have seen many developments in the past forty years.

[7] To make the example fully self-contained we code the data and test in the open-source statistics package R:

```
y <- c (1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0)
test <- function (y){
n <- length (y)
n.switch <- sum (y[2:n] != y[1:(n-1)])
return (n.switch)
}
alpha <- 1
beta <- 1
n.sims <- 10000
theta <- rbeta (n.sims, alpha + sum(y==1), beta + sum(y==0))
T.rep <- rep (NA, n.sims)
for (i in 1:n.sims){
y.rep <- rbinom (length(y), 1, theta[i])
T.rep[i] <- test (y.rep)
}
p.value <- mean (T.rep >= test(y))
```

than 20% of the vote. The observed frequency of these 'outliers' in the observed data is 0.026 (that is, 2.6%) in open-seat elections and 0.008 in elections with incumbents running for reelection. Simulation of replicated data from the fitted model yields a posterior predictive distribution for these test statistics under which, if the model were true, the expected proportion of outliers is only 0.004, and the simulations reveal that the observed rates are far outside of what could plausibly happen under the normal model.

This sort of test is natural in statistical practice (some classic simulation-based model checks in the statistical literature appear in Bush and Mosteller 1955, and Ripley 1988), and it fits in fine with Bayesian inference. We can feel much more comfortable with probability models to new data if we are able and willing to check the fit. I associate this Bayesian approach of 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.

5. Reacting to 'All Models Are Wrong'

The celebrated dictum of Box (1976), "all models are wrong but some are useful", can lead an applied Bayesian (at least) two different directions: *Bayes factors* or *posterior predictive checks*.

The Bayes factor approach abandons the Popperian idea of hypothesis testing entirely. What is the point of rejecting a model if we know ahead of time that it is false? Instead you compute the relative posterior probabilities of competing models, with the progress of science corresponding to formerly high-probability models being abandoned in favor of new models that are more supported by the data (and, more generally, by the improvement or replacement of existing models by alternatives that currently or with future data will have higher posterior probability).

To me, Bayes factors correspond to a discrete view of the world, in which we must choose between models A, B, or C (or a weighted average of A, B, and C, using the related idea of discrete model averaging as in Madigan and Raftery 1994). I prefer the idea of model expansion, using posterior predictive checks as a guide to where and how efforts at expansion should be targeted.

In contrast, posterior predictive checks (as illustrated in the previous section) embrace rejection but with the goal of understanding what aspects of the data are not fit well by the model. As with Bayes factors, the goal is not rejection itself. What rejection tells us is not that a model is false or even likely so—we know our models are false even before gathering any data—but rather that certain potentially important aspects of the data are not captured by the model.

Formally, posterior predictive checks are based on a comparison of data y to hypothetical replicated data y^{rep}, averaging over the posterior distribution

of the unknown parameters θ. In this formulation (Gelman, Meng and Stern 1996), the joint distribution of all these quantities is $p(y,y^{rep},\theta) = p(\theta)p(y \mid \theta)$ $p(y^{rep} \mid \theta)$; that is, y and y^{rep} are two independent instances of the data. In a predictive check, there is some test statistic T(y), and it is compared to its distribution under hypothetical replication, y^{rep}. The p-value is $\Pr(T(y^{rep}) \geq T(y))$. If the test statistic is *pivotal* (that is, if the distribution of T does not depend on θ), we can stop right there. In general, though, the distribution of the test statistic will depend on unknown parameters, hence the p-value must be written conditionally as p-value $(\theta) = \Pr(T(y^{rep}) \geq T(y) \mid \theta)$. In classical tests a common fix at this point is to plug in a point estimate of θ and then adjust the p-value if necessary to account for the ensuing additional variability. Our Bayesian approach is to average over θ, integrating p-value (θ) over the posterior distribution $p(\theta \mid y)$, which results in the posterior predictive p-value, $\Pr(T(y^{rep}) \geq T(y) \mid y)$.

Posterior predictive checks are disliked by some Bayesian statistics because of their low power arising from their allegedly "using the data twice" (Bayarri and Berger 2000; see also Mayo 2008). Here is not the place to debate this issue (see Bayarri and Castellanos 2007, and Gelman 2007, for two views on Bayesian predictive checking) but I will briefly note that statistical power is not a big concern for us: unlike in classical testing, the goal is not to reject a model (we could do that before we started on a priori grounds) but rather to understand aspects of lack of fit. If a certain posterior predictive check has zero or low power, this is not a problem for us: it simply represents a dimension of the data that is automatically or virtually automatically fit by the model.

6. Discussion

What can statistics learn from philosophy? Falsification and the notion of scientific revolutions can make us willing and even eager to check our model fit and to vigorously investigate anomalies rather than taking a naively positivistic approach that would treat prediction as the only goal of statistics.

What can the philosophy of science learn from statistical practice? The success of inference using elaborate models, full of assumptions that are certainly wrong, demonstrates the power of deductive inference, and posterior predictive checking demonstrates that ideas of falsification and error statistics can be applied in a fully Bayesian environment with informative likelihoods and prior distributions.

Nowadays 'Bayesian' is often taken to be a synonym for rationality, and I can see how this can irritate thoughtful philosophers and statisticians alike: To start with, lots of rational thinking—even lots of rational statistical inference—does not occur within the Bayesian formalism. And, to look at it from the other direction, lots of self-proclaimed Bayesian inference hardly seems rational at all. And in what way is 'subjective probability' a model for rational scientific inquiry? On the contrary, subjectivity and rationality are in many ways opposites!

The goal of this paper is to break the link between Bayesian modeling (good, in my opinion) and subjectivity (bad). From this perspective, the irritation of falsificationists regarding exaggerated claims of Bayesian rationality are my ally. Being Bayesian is no guarantee of rationality or even of coherence. Our appropriate willingness to discard and improve models that poorly fit data has the effect of destroying all the theorems of Bayesian coherence. What we are left with is an approach that is coherent—deductive—within a model and which is an effective tool for model checking through its ability to generate probabilistic predictions about anything. A Bayesian model makes strong claims and is thus falsifiable.

I admit, however, that there is a philosophical incoherence in my approach! Consider a simple model with independent data y_1, y_2, .., $y_5 \sim N(\theta,\sigma^2)$, with a prior distribution $\theta \sim N(0,10^2)$ and σ known and taking on some value of approximately 10. Inference about θ is straightforward, as is model checking, whether based on graphs or numerical summaries such as the sample variance and skewness.

But now suppose we consider θ as a random variable defined on the integers. Thus θ = 0 or 1 or 2 or 3 or ... or -1 or -2 or -3 or ..., and with a discrete prior distribution formed by the discrete approximation to the $N(0,10^2)$ distribution. In practice, with the sample size and parameters as defined above, the inferences are essentially unchanged from the continuous case, as we have defined θ on a suitably tight grid.

But from the philosophical position argued in the present article, the discrete model is completely different: I have already written that I do not like to choose or average over a discrete set of models. This is a silly example but it illustrates a hole in my philosophical foundations: when am I allowed to do normal Bayesian inference about a parameter θ in a model, and when do I consider θ to be indexing a class of models, in which case I consider posterior inference about θ to be an illegitimate bit of induction? I understand the distinction in extreme cases—they correspond to the difference between normal science and potential scientific revolutions—but the demarcation does not cleanly align with whether a model is discrete or continuous.

Another incoherence in Bayesian data analysis, as I practice it, arises after a model check. Judgment is required to decide what to do after learning that an aspect of data is not fitted well by the model—or, for that matter, in deciding what to do in the other case, when a test does not reject. In either case, we must think about the purposes of our modeling and our available resources for data collection and computation. I am deductively Bayesian when performing inference and checking within a model, but I must go outside this framework when making decisions about whether and how to alter my model.

In my defense, I see comparable incoherence in all other statistical philosophies:

- Subjective Bayesianism appears fully coherent but falls apart when you examine the assumption that your prior distribution can completely reflect prior knowledge. This can't be, even setting aside that actual prior

distributions tend to be chosen from convenient parametric families. If you could really express your uncertainty as a prior distribution, then you could just as well observe data and directly write your subjective posterior distribution, and there would be no need for statistical analysis at all.

- Classical parametric statistics disallows probabilistic prior information but assumes the likelihood function to be precisely known, which can't make sense except in some very special cases. Robust analysis attempts to account for uncertainty about model specification but relies on additional assumptions such as independence.
- Classical nonparametric methods rely strongly on symmetry, translation invariance, independence, and other generally unrealistic assumptions.

My point here is not to say that my preferred methods are better than others but rather to couple my admission of philosophical incoherence with a reminder that there is no available coherent alternative.

In conclusion, I am arguing here and elsewhere that Bayesian inference need not be subjective (beyond the subjective elements of human choice in any statistical method and any scientific endeavor) nor must it be inductive in the sense of resulting in posterior probabilities of models being true. The Bayesian data analysis that I practice is deductive within a model, with predictive falsification used to compare models. I hope that philosophers who are interested in falsification and error statistics will see the compatibility of my brand of Bayesian inference with their philosophy, and I hope that practicing Bayesians will recognize that falsification and model checking are consistent with a larger Bayesian approach. If you want to follow scheme S (see the chart at the beginning of this article), feel free to do so, but realize that is a choice, not a necessity.

References

Bandyopadhyay, P. S. and G. Brittan (2010), "Two Dogmas of Strong Objective Bayesianism", *International Studies in the Philosophy of Science* 24, 45–65.

Bayarri, M. J. and M. E. Castellanos (2007), "Bayesian Checking of the Second Levels of Hierarchical Models" (with discussion), *Statistical Science* 22, 322–343.

— and J. O. Berger (2000), "P-values for Composite Null Models" (with discussion), *Journal of the American Statistical Association* 95, 1127–1142.

Box, G. E. P. (1976), "Science and Statistics", *Journal of the American Statistical Association* 71, 791–799.

Bush, R. R. and F. Mosteller (1955), *Stochastic Models for Learning*, New York: Wiley.

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

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing", *International Statistical Review* 2, 369–382.

— (2007), Discussion of "Bayesian Checking of the Second Levels of Hierarchical Models", by M. J. Bayarri and M. E. Castellanos, *Statistical Science* 22, 349–352.

— (2011a), Discussion of "The Future of Indirect Evidence", by B. Efron, *Statistical Science* 25, 162–165.

— (2011b), "Causality and Statistical Learning", *American Journal of Sociology*, in press.

—, J. B. Carlin, H. S. Stern and D. B. Rubin (2003), *Bayesian Data Analysis*, 2nd ed., London: CRC Press.

—, X. L. Meng and H. S. Stern (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistical Science* 6, 733–807.

— and C. R. Shalizi (2011a), "Philosophy and the Practice of Bayesian Statistics in the Social Sciences", in: Kinkaid, H. (ed.) *The Oxford Handbook of the Philosophy of Social Sciences*, Oxford: Oxford University Press.

— and — (2011b), "Philosophy and the Practice of Bayesian Statistics", technical report, Department of Statistics, Columbia University.

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

Huber, P. J. (2011), *Data Analysis: What Can Be Learned from the Past 50 Years?*, New York: Wiley.

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

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

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

Madigan, D. M. and A. E. Raftery (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window", *Journal of the American Statistical Association* 89, 1335–1346.

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

— (2008), "How to Discount Double-counting When It Counts: Some Clarifications", *British Journal of Philosophy of Science* 59, 857–879.

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

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

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, Cambridge: Cambridge University Press.

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

Wikipedia (2009), Bayesian Inference, URL: http://en.wikipedia.org/wiki/Bayesian_Inference (09.06.2010).