British Journal of Mathematical and Statistical Psychology (2013), 66, 57–64

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

The error-statistical philosophy and the practice of Bayesian statistics: Comments on Gelman and Shalizi: 'Philosophy and the practice of Bayesian statistics'

Deborah G. Mayo*

Department of Philosophy, Virginia Polytechnic Institute and State University, Blacksburg, USA

1. Introduction

I am pleased to have the opportunity to comment on this interesting and provocative paper. I shall begin by citing three points on which the authors happily depart from existing work on statistical foundations.

First, there is the authors' recognition that methodology is ineluctably bound up with philosophy. 'If nothing else, … strictures derived from philosophy can inhibit research progress' (Gelman & Shalizi, 2013, p. 11). They note, for example, the reluctance of some Bayesians to test their models because of their belief that 'Bayesian models were by definition subjective', or perhaps because checking involves non-Bayesian methods (p. 4, n. 4).

Second, they recognize that Bayesian methods need a new foundation. Although the subjective Bayesian philosophy, 'strongly influenced by Savage (1954), is widespread and influential in the philosophy of science (especially in the form of Bayesian confirmation theory…)', and while many practitioners perceive the 'rising use of Bayesian methods in applied statistical work' (p. 9), as supporting this Bayesian philosophy, the authors flatly declare that 'most of the standard philosophy of Bayes is wrong' (p. 10, n. 2). Despite their qualification that 'A statistical method can be useful even if its philosophical justification is in error', their stance will rightly challenge many a Bayesian.

This will be especially so when one has reached their third thesis, which seeks a new foundation that uses non-Bayesian ideas. Although the authors at first profess that their 'perspective is not new', but rather follows many other statisticians who emphasize 'the value of Bayesian inference as an approach for obtaining statistical methods with good

*Correspondence should be addressed to Deborah G. Mayo, Philosophy Department, 229 Major Williams Hall (0126), Virginia Tech, Blacksburg, VA 24061, USA (e-mail: mayod@vt.edu).

DOI:10.1111/j.2044-8317.2012.02064.x

frequency properties' (p. 10), they go on to announce they are 'going beyond the evaluation of Bayesian methods based on their frequency properties – as recommended by Rubin (1984), Wasserman (2006), among others – to emphasize the learning that comes from the discovery of systematic differences between model and data' (p. 21). Moreover, they suggest that 'Implicit in the best Bayesian practice is a stance that has much in common with the error-statistical approach of Mayo (1996), despite the latter's frequentist orientation.[1] Indeed, crucial parts of Bayesian data analysis, such as model checking, can be understood as "error probes" in Mayo's sense (2)', which might be seen as using modern statistics to implement the Popperian criteria for *severe tests*.

In the Popperian spirit, let me stick my neck out and conjecture that the authors are correct. This is not the place to detail the error-statistical account, but I will illustrate from among its themes where they pertain to the present paper (see Mayo & Spanos, 2011).

The idea that non-Bayesian ideas might afford a foundation for the many strands of Bayesianism is not as preposterous as it first seems. Supplying a foundation requires that we step back from formal methods themselves. That is what the error-statistical philosophy attempts to provide for such well-known ('sampling theory') tools as significance tests and confidence interval methods. But the idea of severe testing is sufficiently general to apply to any other methods on offer. On the face of it, any inference, whether to the adequacy of a model (for a given purpose) or to a posterior probability, can be said to be warranted just to the extent that the inference has withstood severe testing.

If the authors are right, several novel pathways for situating current work suddenly open up. But that is for another time. Here, I will point up some places where error-statistical methods might yield tools to promote the authors' ends, but also others where they will hold up large warning signs! In so doing I will often refer to the 'philosophical coda' in the last several pages of their paper. Leaving to one side quibbles about some of the philosophical positions they mention, their 'coda' contains many important philosophical insights that should be applied throughout.

2. Testing in their data-analysis cycle

The authors claim their statistical analysis is used 'not for computing the posterior probability that any particular model was true – we never actually did that' (p. 13), but rather 'to fit rich enough models' and upon discerning that aspects of the model 'did not fit our data' (p. 13), to build a more complex, better-fitting, model; which in turn called for alteration when faced with new data.

This cycle, they rightly note, involves a 'non-Bayesian checking of Bayesian models' (p. 17), but they should not describe it as purely deductive; it is not. Nor should they wish to hold to that old distorted view of a Popperian test as 'the rule of deduction which says that if p implies q, and q is false, then p must be false' (with p and q the hypothesis and data, respectively) (p. 28). Having thrown off one oversimplified picture, they should avoid slipping into another. As Popper well knew, any observable predictions are derived only with the help of various auxiliary claims $A_1, \ldots, A_n$. Confronted with anomalous data one

[1] I refer to these methods as 'error-statistical' because of their focus on using sampling distributions to control and assess error probabilities. In contexts of scientific inference, error probabilities are used to evaluate severity and non-severity. The single concept of severity applies to both the usual rejections and non-rejections, but severity, which is data-dependent, is only in the same direction as power in the case of non-rejections. (This qualifies a point on p. 15 of Gelman & Shalizi 2013.)

may at most infer that either H or one of the auxiliaries is to blame: *Duhem's problem.* While mentioned in the philosophical coda (p. 31), they should be explicitly raising Duhemian concerns all along.

To infer evidence of a genuine anomaly is to make an inductive inference to the existence of a reproducible effect: Popper called it a *falsifying hypothesis*. Although falsification rules must be probabilistic in some sense, it is not enough to regard the anomaly as genuine simply because the outcome is highly improbable under a hypothesized model. Individual outcomes described in detail may easily have very small probabilities without being genuine anomalies.

Alluding to Mayo and Cox (2006), the authors suggest that any account that moves from data to hypotheses might be called a theory of inductive inference in our sense. Not at all. The requirements for reliable or severe tests must be met. Our point was to show that sampling theory methods, contrary to what has been supposed, satisfy these requirements, so long as they are suitably interpreted. Severity assignments are not posterior probabilities, but they do involve induction. Since the authors concur with the idea of 'a model being severely tested if it passes a probe which had a high probability of detecting an error if it is present' (Gelman and Shalizi, 2013, p. 21), it will be up to them to show they can satisfy this.

3. Significance tests and *p*-values in model checking

In probing the adequacy of statistical models, the authors recommend a method akin to 'pure significance testing' (p. 20), where no specific alternative models are considered. In frequentist significance testing for misspecifications, the 'null' hypothesis asserts, in effect, that a given model adequately captures the data-generating mechanism, and one constructs a relevant test statistic whose distribution may be computed, at least under the null hypothesis. The authors do something analogous, using what they call the posterior predictive distribution as the reference (or sampling) distribution of the chosen test statistic. Here, they build on a distinct strand in the 'Bayesian *p*-value' research programme, one of whose developers was Gelman.

Some claim that, at least for large sample sizes, their analysis leads essentially to 'rediscovering' frequentist *p*-values (Bayarri & Berger, 1998; Ghosh, Delampady, & Samatra, 2006, p. 182). But the authors are right to point out that all participants in the Bayesian *p*-value program implicitly '*disagree* with the standard inductive view' of Bayesianism (Gelman and Shalizi, 2013, p. 18, n. 11). Even if some use such tests only to infer the adequacy or inadequacy of an underlying model (with a view to later finding Bayesian posteriors), the reasoning employs hypothetical repetitions of the data in these inferences, thereby apparently violating the likelihood principle.[2] If the authors' approach is accused of producing a non-Bayesian animal, as has been alleged, so it seems do other Bayesian *p*-value appeals. (The qualifications that Berger and others propose to distinguish degrees of heresy do not seem to hold water.) More constructively, the value of employing a sampling distribution to represent statistically what it would be like were one or another assumption of the data-generating mechanism violated, argues for the validity of such non-Bayesian reasoning more generally.

[2] The likelihood principle (LP), despite following from Bayes' theorem, has become highly controversial. See Mayo, 2010 for a discussion of the flaw in Birnbaum's (1962) argument that the LP follows from frequentist principles. Since Gelman and Shalizi are rejecting inference by way of Bayes' theorem, they are not bound to the LP as other Bayesians are.

Nevertheless, the fact that the authors approve of reasoning akin to frequentist *p*-values does not automatically show that their methods enjoy the virtues that enable frequentist significance tests to reliably distinguish underlying sources of various observed discordancies.[3] Their examples, as presented, leave gaps that need to be filled in.

3.1. Reasonably large p*-value*

To compute 'whether the observed data set is the kind of thing that the fitted model produces with reasonably high probability' – assuming the replicated data are of the same size and shape as $\mathbf{y}_0$ – 'generated under the assumption that the fitted model, prior and likelihood both, is true' (p. 18), they check to see if the Bayesian *p*-value is reasonably high. If it is high, then the data are 'unsurprising if the model is true'. However, as the authors themselves note, 'Whether this is evidence *for* the usefulness of the model depends on how likely it is to get such a high *p*-value when the model is false, the "severity" of the test' (p. 18). But it is not clear how they are able to get this severity computation under the falsity of the model (a power-type assessment). A correct severity assessment with local tests would need to be qualified: the data may only indicate the absence of those violations that the test was at least reasonably capable of detecting, if present.

3.2. Small p*-value*

A small *p*-value, on the other hand, is taken as evidence of incompatibility between model and data (where their model includes the prior). The question that arises here is: what kind of incompatibility are we allowed to say this is evidence of? Even when it is warranted to infer there is evidence of a systematic departure from the assumed model and prior, the pure significance test would seem only to allow us to infer that there is a flaw somewhere either in the likelihood or prior. It would be fallacious to claim that one thereby has evidence for a specific alternative that 'explains' the effect – at least not without further work to pass the alternative with severity. (It is a kind of fallacy of rejection; see Mayo & Spanos, 2006, 2011).

Yet at times it appears that the authors will go from detecting an anomaly for the initial model (e.g., a logistic regression with varying intercept) to inferring a specific expansion to the model (e.g., one with both varying intercept and slope.) How have the other potential sources of misfit been probed and ruled out? I am not saying that they commit this common fallacy, only that we have not been told how they will avoid it. Aris Spanos calls it 'error fixing'. It is illustrated by a Durbin–Watson test that moves from evidence of some violation of independence to inferring the alternative hypothesis (autocorrelation) which describes just one of many types of dependence (Mayo & Spanos, 2004; Spanos, 2000, 2006). The test had little or no ability to identify other types of dependencies, and other model flaws.

A well-developed account of misspecification tests (under the error-statistical umbrella) exists, even though, admittedly, it is not used as often as it should be (Spanos, 1999). It is here that the authors could get real mileage from, as well as help to expand the use of, the error-statistical account of model-misspecification testing. At

[3] They claim their *p*-values are 'generalizations of classical *p*-values, merely replacing point estimates of parameters θ with averages over the posterior distribution' (p. 18).

the heart of the account is the recognition that significance tests must be used in a proper sequence to reflect the interdependence of the model assumptions. Judicious combinations of omnibus (non-parametric), directional (parametric) and simulation-based tests deliberately invoke dissimilar assumptions, and allow probing as broadly away from the model in question as possible. One must keep track of the assumptions each test requires to get going. It is very easy to show that even in the simplest models, such as the normal i.i.d. model, departures from dependence can misleadingly influence the result of testing for normality. An error statistician would worry about the authors jumping into the model validation task without first listing a complete set of probabilistic assumptions, for example, underlying their logistic regressions. This is particularly important for the subsequent task of respecifying the original model in light of the detected departures from the assumptions. Let me be clear that I can see no reason why (in principle) the authors could not avail themselves of this battery of tools, and this would be a fruitful avenue for future work; certainly more so than any one of the ongoing controversies about such things as which of the menu of Bayesian *p*-values has better asymptotic properties.

4. Some puzzles

With this in mind, it is puzzling that the authors claim to 'find graphical test summaries more illuminating than *p*-values' (p. 18). Although useful, particularly in getting ideas for discrepancies to probe, exclusive reliance on eyeballing loosens, rather than tightens, the required constraints demanded to ensure that one knows which model violations any given test can or cannot discern with severity. The choice of which residuals to look at, as with the choice of test statistic, already implies the type or direction of departure. Data plots that seem to indicate one flaw, say non-normality, can easily be the result of an entirely different assumption, say independence, being at fault; but the given graphical discernment may have had little chance to reveal this.

Perhaps the disparaging of *p*-value reasoning by Bayesians leads them to champion something less advertently non-Bayesian, such as graphical analysis. They emphasize 'we are not claiming that classic *p*-values are the answer. As is indicated by the literature on the Jeffreys–Lindley paradox (notably Berger & Sellke, 1987), *p*-values can drastically overstate the evidence against a null hypothesis'. My puzzle here is that the allegations in Berger and Sellke, and more recently in Berger (2003), are based on assuming a Bayesian inference of the sort the authors have said they were rejecting. From the error statistician's perspective, what these Bayesians regard as problematic for frequentist *p*-values is actually problematic for their 'conditional *p*-values' (for two-sided tests): highly significant results are construed as no evidence against the null, or even evidence in favour of the null (the posterior to the null going up in value) (Mayo, 2003). Talk about low power. But the relevant point here is simply that the authors should not see the choice as between an unsophisticated use of significance tests and eyeballing. They need the full battery of misspecification tests.

It is true that allegations of double-counting are frequently heard when the 'same' data are used to arrive at as well as to check model discrepancies. That may be another reason they prefer to stick with something more informal (such as graphical methods). However, it is precisely the effect on the test's error probability that will tell us whether double-counting is problematic or not. With misspecification tests, correctly applied, it is not problematic.

Having heretically announced that they seek a non-Bayesian (error-statistical) foundation for Bayesian methods, the authors might as well take advantage of the mileage it can afford.

5. The role of priors and testing priors

In many Bayesian accounts the prior probability distribution is assumed as a given, either as a way of introducing prior beliefs into the analysis (as with subjective Bayesians) or, conversely, to avoid introducing prior beliefs (as with the appeal to reference or default priors). In contrast, the authors claim that their methods provide ways of testing priors. To check if something has satisfied its role, however, we had best be clear on what its intended role is.

The authors tell us what a prior need *not* be. It will not, or need not, be a default prior. Because their prior is testable, they are freed from finding the unique objectively correct prior, unlike the default Bayesians.

Nor need the prior represent a statistician's beliefs. The prior distribution, the authors claim, is one of the assumptions of the model and does not need to represent the statistician's personal degree of belief in alternative parameter values. (Suppose it does, however. I wonder if in that case the approach focuses only on checking the likelihood, assuming the prior?)

Elsewhere we hear that the model 'is the combination of the prior distribution and the likelihood, each of which represents some compromise among scientific knowledge, mathematical convenience, and computational tractability' (p. 20). So what does it mean to say we have tested the prior and it fails? It could mean the prior represents false beliefs, or it is not so convenient after all, or …?

At other times the authors claim that they view the prior as 'a regularization device', making fitted models less sensitive to certain details of the data. I do not pretend to be clear on why the likelihood here needs smoothing or regularizing; but accepting that it does, I am unclear as to how checking the prior-likelihood model can be seen as checking the regularization device. (Perhaps when the prior serves to regularize, then, once again, there is no reason to check; they do not say.) Again, Duhemian problems loom large; there are all kinds of things one might consider changing to make it all fit.

There is no problem with the prior serving many functions, so long as its particular role is pinned down for the case at hand. The error-statistical account would suggest first checking the likelihood portion of the model, and then turning to the prior. If a battery of tests is available (with or without priors) it is hard to see that there is any advantage to their forgoing them. This leads to my last key point.

6. Error statistics is piecemeal

A central feature of the error-statistical philosophy of science is in its distinguishing substantive scientific questions from various statistical ones. In almost all cases these are distinct; while hypotheses that appear in standard null hypothesis tests may be far too simple to represent the main or primary scientific question at hand, for the tasks of checking for errors and discerning systematic effects in data, they are just the ticket. However, there are several places where the authors do not avail themselves of this important distinction. They instead infer from the fact that, strictly speaking, our models of the world may be false, that therefore all inferences to statistical models are false.

A hypothesis that Einstein's model of light deflection fully captures light deflection phenomena is false, but claims that radio-astronomical data are genuinely anomalous for a Newtonian deflection are true, and have been known to be true, at least since the 1970s.

By the authors' own lights, the statistical model is supposed to capture the systematic statistical information in the model, relative to the aspects or questions the model is trying to capture. To their credit, the authors emphasize that they wish to reject models if they do not account for all the systematic (statistical) information patterns in the data (Spanos, 2007). However, if all models incorrectly captured the statistical information, one forfeits the very idea of severely ruling out specific ways a model can fail for the problem at hand. 'Since we are quite sure our models are wrong, we need to check whether the misspecification is so bad that inferences regarding the scientific parameters are in trouble' (p.17). This assumes that claims about being in trouble may be correct. If they have split things off properly, error statisticians can pinpoint the trouble: we determine how badly a violation would distort the error probabilities for a statistical inference that will rely on the model.

7. Concluding remark

The authors have provided a radical and important challenge to the foundations of current Bayesian statistics, in a way that reflects current practice. Their paper points to interesting new research problems for advancing what is essentially a dramatic paradigm change in Bayesian foundations. While their examples involve survey sampling, they clearly see themselves as advancing a general conception.

I hope that Gelman and Shalizi's paper will motivate Bayesian epistemologists in philosophy to take note of foundational problems in Bayesian practice, and that it will inspire philosophically-minded frequentist error statisticians to help craft a new foundation for using statistical tools – one that will afford a series of error probes that, taken together, enable stringent or severe testing.

Acknowledgement

I gratefully acknowledge the insights of Aris Spanos on misspecification testing, and his very useful comments on earlier drafts of this paper.

References

Bayarri, M. J., & Berger, J. O. (1998). Robust Bayesian analysis of selection models. *Annals of Statistics*, *26*, 645–659. doi:10.1214/aos/1028144852

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*(1), 1–32. doi:10.1214/ss/1056397485

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p*-values and evidence (with discussion). *Journal of the American Statistical Association*, *82*, 112–139. doi:10.2307/2289131

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, *57*, 269–326. doi:10.1037/h0044139

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8–38. doi:10.1111/j.2044-8317.2011.02037.x

Ghosh, J. K., Delampady, M., & Samatra, T. (2006). *An introduction to Bayesian analysis: Theory and methods*. New York: Springer.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? Commentary on J. Berger's Fisher address. *Statistical Science*, *18*(1), 19–24. doi:10.1214/ss/1056397485

Mayo, D. G. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In D. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 305–314). Cambridge, UK: Cambridge University Press.

Mayo, D. (2011). Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)? *Rationality, Markets and Morals*, *2*, 79–102. Retrieved from http://www.rmm-journal.de/htdocs/st01.html

Mayo, D., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 153–198). Oxford, UK: Elsevier.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172. doi:10.1214/aos/1176346785

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Spanos, A. (1999). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.

Spanos, A. (2000). Revisiting data mining: 'Hunting' with or without a license. *Journal of Economic Methodology*, *7*, 231–264. doi:10.1080/13501780050045119

Spanos, A. (2006). Econometrics in retrospect and prospect. In T. C. Mills & K. Patterson (Eds.), *Palgrave handbook of econometrics* (Vol. 1, pp. 3–58). Basingstoke, UK: Macmillan.

Spanos, A. (2007). Curve fitting, the reliability of inductive inference, and the error-statistical approach. *Philosophy of Science*, *74*(5), 1046–1066.

Wasserman, L. (2006). Frequentist Bayes is objective. *Bayesian Analysis*, *1*(3), 451–456. doi:10.1214/06-BA116H

Received 8 February 2012