British Journal of Mathematical and Statistical Psychology (2013), 66, 8–38

The British Psychological Society

www.wileyonlinelibrary.com

Philosophy and the practice of Bayesian statistics

Andrew Gelman[1]* and Cosma Rohilla Shalizi[2]

[1]Department of Statistics and Department of Political Science, Columbia University, New York, USA

[2]Statistics Department, Carnegie Mellon University, Santa Fe Institute, Pittsburgh, USA

A substantial school in the philosophy of science identifies Bayesian inference with inductive inference and even rationality as such, and seems to be strengthened by the rise and practical success of Bayesian statistics. We argue that the most successful forms of Bayesian statistics do not actually support that particular philosophy but rather accord much better with sophisticated forms of hypothetico-deductivism. We examine the actual role played by prior distributions in Bayesian models, and the crucial aspects of model checking and model revision, which fall outside the scope of Bayesian confirmation theory. We draw on the literature on the consistency of Bayesian updating and also on our experience of applied work in social science. Clarity about these matters should benefit not just philosophy of science, but also statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework.

1. The usual story – which we don't like

> In so far as I have a coherent philosophy of statistics, I hope it is 'robust' enough to cope in principle with the whole of statistics, and sufficiently undogmatic not to imply that all those who may think rather differently from me are necessarily stupid. If at times I do seem dogmatic, it is because it is convenient to give my own views as unequivocally as possible. (Bartlett, 1967, p. 458)

Schools of statistical inference are sometimes linked to approaches to the philosophy of science. 'Classical' statistics – as exemplified by Fisher's *p*-values, Neyman–Pearson hypothesis tests, and Neyman's confidence intervals – is associated with the hypothetico-deductive and falsificationist view of science. Scientists devise hypotheses, deduce implications for observations from them, and test those implications. Scientific hypotheses

*Correspondence should be addressed to Andrew Gelman, Department of Statistics and Department of Political Science, 1016 Social Work Bldg, Columbia University, New York, NY 10027 USA (e-mail: gelman@stat.columbia.edu).

DOI:10.1111/j.2044-8317.2011.02037.x

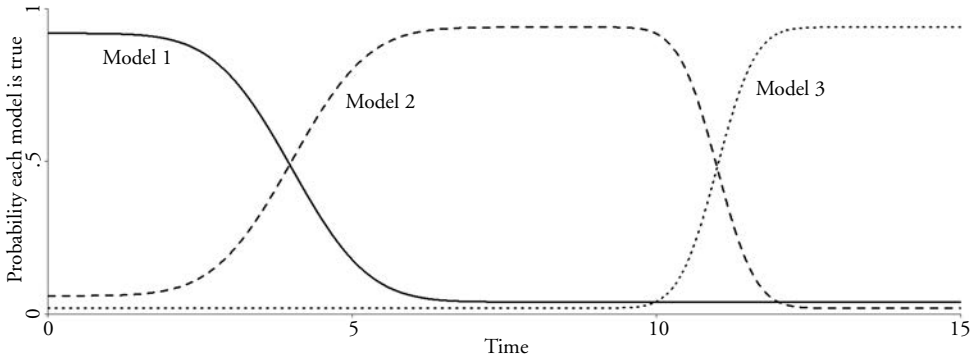

Figure 1. Hypothetical picture of idealized Bayesian inference under the conventional inductive philosophy. The posterior probability of different models changes over time with the expansion of the likelihood as more data are entered into the analysis. Depending on the context of the problem, the time scale on the *x*-axis might be hours, years, or decades, in any case long enough for information to be gathered and analysed that first knocks out hypothesis 1 in favour of hypothesis 2, which in turn is dethroned in favour of the current champion, model 3.

can be rejected (i.e., falsified), but never really established or accepted in the same way. Mayo (1996) presents the leading contemporary statement of this view.

In contrast, Bayesian statistics or 'inverse probability' – starting with a prior distribution, getting data, and moving to the posterior distribution – is associated with an inductive approach of learning about the general from particulars. Rather than employing tests and attempted falsification, learning proceeds more smoothly: an accretion of evidence is summarized by a posterior distribution, and scientific process is associated with the rise and fall in the posterior probabilities of various models; see Figure 1 for a schematic illustration. In this view, the expression $p(\theta|y)$ says it all, and the central goal of Bayesian inference is computing the posterior probabilities of hypotheses. Anything not contained in the posterior distribution $p(\theta|y)$ is simply irrelevant, and it would be irrational (or incoherent) to attempt falsification, unless that somehow shows up in the posterior. The goal is to learn about general laws, as expressed in the probability that one model or another is correct. This view, strongly influenced by Savage (1954), is widespread and influential in the philosophy of science (especially in the form of Bayesian confirmation theory – see Howson & Urbach, 1989; Earman, 1992) and among Bayesian statisticians (Bernardo & Smith, 1994). Many people see support for this view in the rising use of Bayesian methods in applied statistical work over the last few decades.[1]

[1] Consider the current (9 June 2010) state of the Wikipedia article on Bayesian inference, which begins as follows:

> Bayesian inference is statistical inference in which evidence or observations are used to update or to newly infer the probability that a hypothesis may be true.

It then continues:

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

We think most of this received view of Bayesian inference is wrong.[2] Bayesian methods are no more inductive than any other mode of statistical inference. Bayesian data analysis is much better understood from a hypothetico-deductive perspective.[3] 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. Indeed, crucial parts of Bayesian data analysis, such as model checking, can be understood as 'error probes' in Mayo's sense.

We proceed by a combination of examining concrete cases of Bayesian data analysis in empirical social science research, and theoretical results on the consistency and convergence of Bayesian updating. Social-scientific data analysis is especially salient for our purposes because there is general agreement that, in this domain, all models in use are wrong – not merely falsifiable, but actually false. With enough data – and often only a fairly moderate amount – any analyst could reject any model now in use to any desired level of confidence. Model fitting is nonetheless a valuable activity, and indeed the crux of data analysis. To understand why this is so, we need to examine how models are built, fitted, used and checked, and the effects of misspecification on models.

Our perspective is not new; in methods and also in philosophy we follow statisticians such as Box (1980, 1983, 1990), Good and Crook (1974), Good (1983), Morris (1986), Hill (1990) and Jaynes (2003). All these writers emphasized the value of model checking and frequency evaluation as guidelines for Bayesian inference (or, to look at it another way, the value of Bayesian inference as an approach for obtaining statistical methods with good frequency properties; see Rubin, 1984). Despite this literature, and despite the strong thread of model checking in applied statistics, this philosophy of Box and others remains a minority view that is much less popular than the idea of Bayes being used to update the probabilities of different candidate models being true (as can be seen, for example, by the Wikipedia snippets given in footnote 1).

A puzzle then arises. The evidently successful methods of modelling and model checking (associated with Box, Rubin and others) seem out of step with the accepted view of Bayesian inference as inductive reasoning (what we call here 'the usual story'). How can we understand this disjunction? One possibility (perhaps held by the authors of the Wikipedia article) is that the inductive Bayes philosophy is correct and that the model-building approach of Box and others can, with care, be interpreted in that way. Another possibility is that the approach characterized by Bayesian model checking and continuous model expansion could be improved by moving to a fully Bayesian approach centring on the posterior probabilities of competing models. A third possibility, which we advocate, is that Box, Rubin and others are correct and that the usual philosophical story of Bayes as inductive inference is faulty.

Nonetheless, some Bayesian statisticians believe probabilities can have an objective value and therefore Bayesian inference can provide an objective method of induction.

These views differ from those of, for example, Bernardo and Smith (1994) or Howson and Urbach (1989) only in the omission of technical details.

[2] We are claiming that most of the standard philosophy of Bayes is wrong, *not* that most of Bayesian inference itself is wrong. A statistical method can be useful even if its common philosophical justification is in error. It is precisely because we believe in the importance and utility of Bayesian inference that we are interested in clarifying its foundations.

[3] We are not interested in the hypothetico-deductive 'confirmation theory' prominent in philosophy of science from the 1950s to the 1970s, and linked to the name of Hempel (1965). The hypothetico-deductive account of scientific method to which we appeal is distinct from, and much older than, this particular sub-branch of confirmation theory.

We are interested in philosophy and think it is important for statistical practice – if nothing else, we believe that strictures derived from philosophy can inhibit research progress.[4] That said, we are statisticians, not philosophers, and we recognize that our coverage of the philosophical literature will be incomplete. In this presentation, we focus on the classical ideas of Popper and Kuhn, partly because of their influence in the general scientific culture and partly because they represent certain attitudes which we believe are important in understanding the dynamic process of statistical modelling. We also emphasize the work of Mayo (1996) and Mayo and Spanos (2006) because of its relevance to our discussion of model checking. We hope and anticipate that others can expand the links to other modern strands of philosophy of science such as Giere (1988), Haack (1993), Kitcher (1993) and Laudan (1996) which are relevant to the freewheeling world of practical statistics; our goal here is to demonstrate a possible Bayesian philosophy that goes beyond the usual inductivism and can better match Bayesian practice as we know it.

2. The data-analysis cycle

We begin with a very brief reminder of how statistical models are built and used in data analysis, following Gelman, Carlin, Stern, and Rubin (2004), or, from a frequentist perspective, Guttorp (1995).

The statistician begins with a model that stochastically generates all the data y, whose joint distribution is specified as a function of a vector of parameters θ from a space Θ (which may, in the case of some so-called non-parametric models, be infinite-dimensional). This joint distribution is the likelihood function. The stochastic model may involve other (unmeasured but potentially observable) variables $\tilde{y}$ – that is, missing or latent data – and more or less fixed aspects of the data-generating process as covariates. For both Bayesians and frequentists, the joint distribution of $(y, \tilde{y})$ depends on θ. Bayesians insist on a full joint distribution, embracing observables, latent variables and parameters, so that the likelihood function becomes a conditional probability density, $p(y|\theta)$. In designing the stochastic process for $(y, \tilde{y})$, the goal is to represent the systematic relationships between the variables and between the variables and the parameters, and as well as to represent the noisy (contingent, accidental, irreproducible) aspects of the data stochastically. Against the desire for accurate representation one must balance conceptual, mathematical and computational tractability. Some parameters thus have fairly concrete real-world referents, such as the famous (in statistics) survey of the rat population of Baltimore (Brown, Sallow, Davis, & Cochran, 1955). Others, however, will reflect the specification as a mathematical object more than the reality being modelled – t-distributions are sometimes used to model heavy-tailed observational noise, with the number of degrees of freedom for the t representing the shape of the distribution; few statisticians would take this as realistically as the number of rats.

Bayesian modelling, as mentioned, requires a joint distribution for $(y, \tilde{y}, \theta)$, which is conveniently factored (without loss of generality) into a prior distribution for the parameters, $p(\theta)$, and the complete-data likelihood, $p(y, \tilde{y}|\theta)$, so that $p(y|\theta) = \int p(y, \tilde{y}|\theta) d\tilde{y}$. The prior distribution is, as we will see, really part of the model. In practice, the various parts of the model have functional forms picked by a mix of substantive knowledge,

[4] For example, we have more than once encountered Bayesian statisticians who had no interest in assessing the fit of their models to data because they felt that Bayesian models were by definition subjective, and thus neither could nor should be tested.

scientific conjectures, statistical properties, analytical convenience, disciplinary tradition and computational tractability.

Having completed the specification, the Bayesian analyst calculates the posterior distribution $p(\theta|y)$; it is so that this quantity makes sense that the observed y and the parameters θ must have a joint distribution. The rise of Bayesian methods in applications has rested on finding new ways to actually carry through this calculation, even if only approximately, notably by adopting Markov chain Monte Carlo methods, originally developed in statistical physics to evaluate high-dimensional integrals (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Newman & Barkema, 1999), to sample from the posterior distribution. The natural counterpart of this stage for non-Bayesian analyses are various forms of point and interval estimation to identify the set of values of θ that are consistent with the data y.

According to the view sketched in Section 1 above, data analysis basically ends with the calculation of the posterior $p(\theta|y)$. At most, this might be elaborated by partitioning Θ into a set of models or hypotheses, $\Theta_1, \ldots, \Theta_K$, each with a prior probability $p(\Theta_k)$ and its own set of parameters θ_k. One would then compute the posterior parameter distribution within each model, $p(\theta_k|y, \Theta_k)$, and the posterior probabilities of the models,

$$p(\Theta_k|y) = \frac{p(\Theta_k)p(y|\Theta_k)}{\sum_{k'}(p(\Theta_{k'})p(y|\Theta_{k'}))} = \frac{p(\Theta_k)\int p(y, \theta_k|\Theta_k)d\theta_k}{\sum_{k'}(p(\Theta_{k'})\int p(y, \theta_k|\Theta_{k'})d\theta_{k'})}.$$

These posterior probabilities of hypotheses can be used for Bayesian model selection or Bayesian model averaging (topics to which we return below). Scientific progress, in this view, consists of gathering data – perhaps through well-designed experiments, designed to distinguish among interesting competing scientific hypotheses (cf. Atkinson & Donev, 1992; Paninski, 2005) – and then plotting the $p(\Theta_k|y)$ over time and watching the system learn (as sketched in Figure 1).

In our view, the account of the last paragraph is crucially mistaken. The data-analysis process – Bayesian or otherwise – does not end with calculating parameter estimates or posterior distributions. Rather, the model can then be *checked*, by comparing the implications of the fitted model to the empirical evidence. One asks questions such as whether simulations from the fitted model resemble the original data, whether the fitted model is consistent with other data not used in the fitting of the model, and whether variables that the model says are noise ('error terms') in fact display readily-detectable patterns. Discrepancies between the model and data can be used to learn about the ways in which the model is inadequate for the scientific purposes at hand, and thus to motivate expansions and changes to the model (Section 4.).

2.1. Example: Estimating voting patterns in subsets of the population

We demonstrate the hypothetico-deductive Bayesian modelling process with an example from our recent applied research (Gelman, Lee, & Ghitza, 2010). In recent years, American political scientists have been increasingly interested in the connections between politics and income inequality (see, for example, McCarty, Poole, & Rosenthal 2006). In our own contribution to this literature, we estimated the attitudes of rich, middle-income and poor voters in each of the 50 states (Gelman, Park, Shor, Bafumi, & Cortina, 2008). As we described in our paper on the topic (Gelman, Shor, Park, & Bafumi, 2008), we began by fitting a varying-intercept logistic regression: modelling votes (coded as $y = 1$ for votes for the Republican presidential candidate and $y = 0$

for Democratic votes) given family income (coded in five categories from low to high as $x = -2, -1, 0, 1, 2$), using a model of the form $\Pr(y = 1) = \text{logit}^{-1}(a_s + bx)$, where s indexes state of residence – the model is fitted to survey responses – and the varying intercepts a_s correspond to some states being more Republican-leaning than others. Thus, for example, a_s has a positive value in a conservative state such as Utah and a negative value in a liberal state such as California. The coefficient b represents the 'slope' of income, and its positive value indicates that, within any state, richer voters are more likely to vote Republican.

It turned out that this varying-intercept model did not fit our data, as we learned by making graphs of the average survey response and fitted curves for the different income categories within each state. We had to expand to a varying-intercept, varying-slope model, $\Pr(y = 1) = \text{logit}^{-1}(a_s + b_s x)$, in which the slopes b_s varied by state as well. This model expansion led to a corresponding expansion in our understanding: we learned that the gap in voting between rich and poor is much greater in poor states such as Mississippi than in rich states such as Connecticut. Thus, the polarization between rich and poor voters varied in important ways geographically.

We found this not through any process of Bayesian induction but rather through model checking. Bayesian inference was crucial, not for computing the posterior probability that any particular model was true – we never actually did that – but in allowing us to fit rich enough models in the first place that we could study state-to-state variation, incorporating in our analysis relatively small states such as Mississippi and Connecticut that did not have large samples in our survey.[5]

Life continues, though, and so do our statistical struggles. After the 2008 election, we wanted to make similar plots, but this time we found that even our more complicated logistic regression model did not fit the data – especially when we wanted to expand our model to estimate voting patterns for different ethnic groups. Comparison of data to fit led to further model expansions, leading to our current specification, which uses a varying-intercept, varying-slope logistic regression as a baseline but allows for non-linear and even non-monotonic patterns on top of that. Figure 2 shows some of our inferences in map form, while Figure 3 shows one of our diagnostics of data and model fit.

The power of Bayesian inference here is *deductive*: given the data and some model assumptions, it allows us to make lots of inferences, many of which can be checked and potentially falsified. For example, look at New York state (in the bottom row of Figure 3): apparently, voters in the second income category supported John McCain much more than did voters in neighbouring income groups in that state. This pattern is theoretically possible but it arouses suspicion. A careful look at the graph reveals that this is a pattern in the raw data which was moderated but not entirely smoothed away by our model. The natural next step would be to examine data from other surveys. We may have exhausted what we can learn from this particular data set, and Bayesian inference was a key tool in allowing us to do so.

3. The Bayesian principal–agent problem

Before returning to discussions of induction and falsification, we briefly discuss some findings relating to Bayesian inference under misspecified models. The key idea is that

[5] Gelman and Hill (2006) review the hierarchical models that allow such partial pooling.

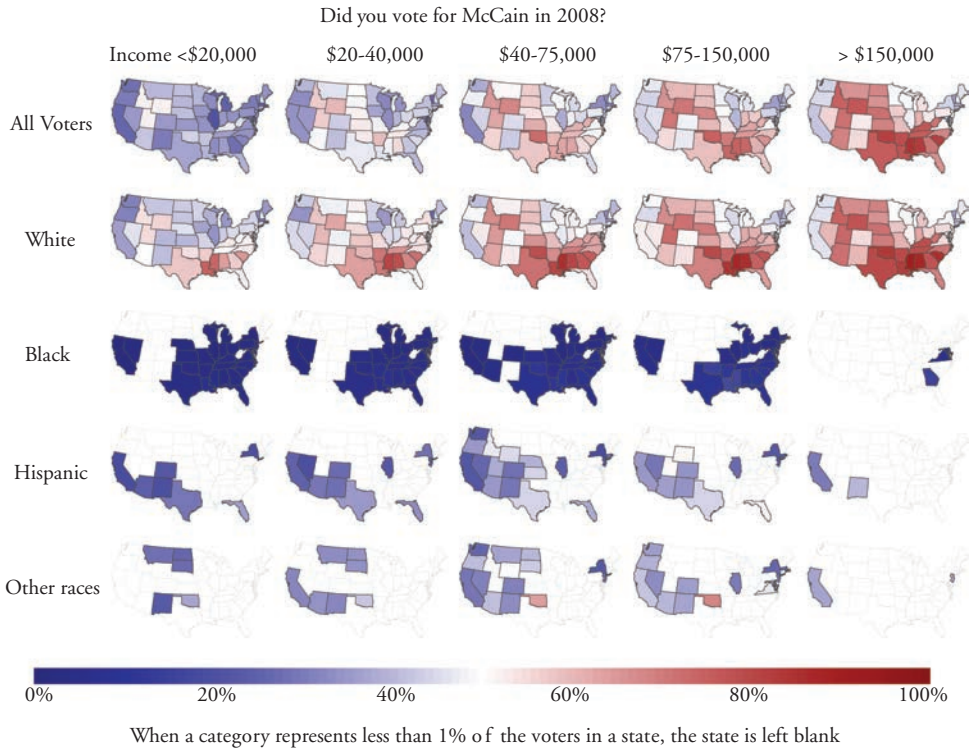

Figure 2. [Colour online]. States won by John McCain and Barack Obama among different ethnic and income categories, based on a model fitted to survey data. States coloured deep red and deep blue indicate clear McCain and Obama wins; pink and light blue represent wins by narrower margins, with a continuous range of shades going to grey for states estimated at exactly 50–50. The estimates shown here represent the culmination of months of effort, in which we fitted increasingly complex models, at each stage checking the fit by comparing to data and then modifying aspects of the prior distribution and likelihood as appropriate. This figure is reproduced from Ghitza and Gelman (2012) with the permission of the authors.

Bayesian inference for model selection – statements about the posterior probabilities of candidate models – does not solve the problem of learning from data about problems with existing models.

In economics, the 'principal–agent problem' refers to the difficulty of designing contracts or institutions which ensure that one selfish actor, the 'agent', will act in the interests of another, the 'principal', who cannot monitor and sanction their agent without cost or error. The problem is one of aligning incentives, so that the agent serves itself by serving the principal (Eggertsson, 1990). There is, as it were, a Bayesian principal–agent problem as well. The Bayesian agent is the methodological fiction (now often approximated in software) of a creature with a prior distribution over a well-defined hypothesis space Θ, a likelihood function $p(y|\theta)$, and conditioning as its sole mechanism of learning and belief revision. The principal is the actual statistician or scientist.

The ideas of the Bayesian agent are much more precise than those of the actual scientist; in particular, the Bayesian (in this formulation, with which we disagree) is

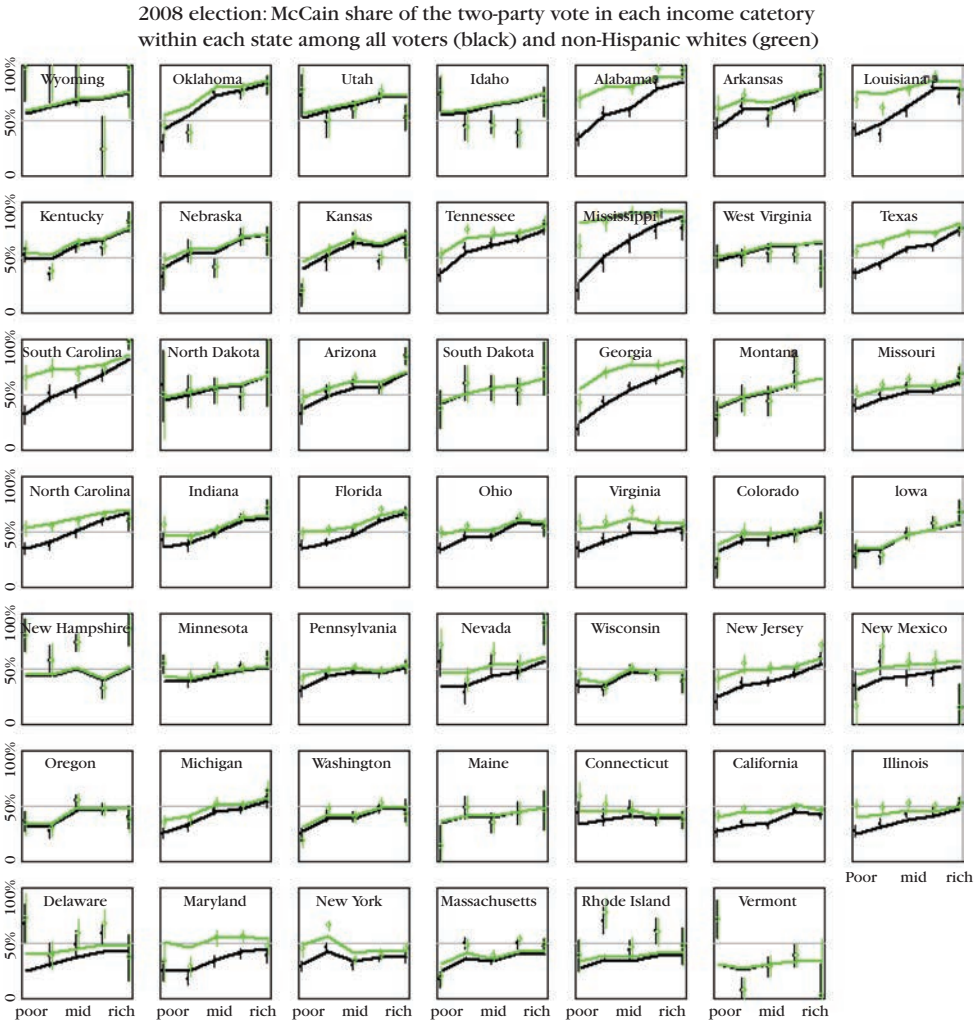

Figure 3. [Colour online]. Some of the data and fitted model used to make the maps shown in Figure 2. Dots are weighted averages from pooled June–November Pew surveys; error bars show $\pm$ 1 standard error bounds. Curves are estimated using multilevel models and have a standard error of about 3% at each point. States are ordered in decreasing order of McCain vote (Alaska, Hawaii and the District of Columbia excluded). We fitted a series of models to these data; only this last model fitted the data well enough that we were satisfied. In working with larger data sets and studying more complex questions, we encounter increasing opportunities to check model fit and thus falsify in a way that is helpful for our research goals. This figure is reproduced from Ghitza and Gelman (2012) with the permission of the authors.

certain that *some* θ is the exact and complete truth, whereas the scientist is not.[6] At some point in history, a statistician may well write down a model which he or she

[6] In claiming that 'the Bayesian' is certain that some θ is the exact and complete truth, we are not claiming that actual Bayesian scientists or statisticians hold this view. Rather, we are saying that this is implied by the philosophy we are attacking here. All statisticians, Bayesian and otherwise, recognize that the philosophical position which ignores this approximation is problematic.

believes contains all the systematic influences among properly defined variables for the system of interest, with correct functional forms and distributions of noise terms. This could happen, but we have never seen it, and in social science we have never seen anything that comes close. If nothing else, our own experience suggests that however many different specifications we thought of, there are always others which did not occur to us, but cannot be immediately dismissed *a priori*, if only because they can be seen as alternative approximations to the ones we made. Yet the Bayesian agent is required to start with a prior distribution whose support covers *all* alternatives that could be considered.[7]

This is not a small technical problem to be handled by adding a special value of θ, say θ^{∞} standing for 'none of the above'; even if one could calculate $p(y|\theta^{\infty})$, the likelihood of the data under this catch-all hypothesis, this in general would *not* lead to just a small correction to the posterior, but rather would have substantial effects (Fitelson & Thomason, 2008). Fundamentally, the Bayesian agent is limited by the fact that its beliefs always remain within the support of its prior. For the Bayesian agent the truth must, so to speak, be always already partially believed before it can become known. This point is less than clear in the usual treatments of Bayesian convergence, and so worth some attention.

Classical results (Doob, 1949; Schervish, 1995; Lijoi, Prünster, & Walker, 2007) show that the Bayesian agent's posterior distribution will concentrate on the truth with *prior* probability 1, provided some regularity conditions are met. Without diving into the measure-theoretic technicalities, the conditions amount to: (i) the truth is in the support of the prior; and (ii) the information set is rich enough that some consistent estimator exists (see the discussion in Schervish, 1995, Section 7.4.1). When the truth is *not* in the support of the prior, the Bayesian agent still thinks that Doob's theorem applies and assigns zero prior probability to the set of data under which it does not converge on the truth.

The convergence behaviour of Bayesian updating with a misspecified model can be understood as follows (Berk, 1966, 1970; Kleijn & van der Vaart, 2006; Shalizi, 2009). If the data are actually coming from a distribution q, then the Kullback–Leibler divergence rate, or relative entropy rate, of the parameter value θ is

$$d(\theta) = \lim_{n\to\infty} \frac{1}{n} \mathrm{E}\left[\log \frac{p(y_1, y_2, \ldots, y_n|\theta)}{q(y_1, y_2, \ldots, y_n)}\right],$$

with the expectation being taken under q. (For details on when the limit exists, see Gray, 1990.) Then, under not-too-onerous regularity conditions, one can show (Shalizi, 2009) that

$$p(\theta|y_1, y_2, \ldots, y_n) \approx p(\theta) \exp\left\{-n(d(\theta) - d^*)\right\},$$

with d^* being the essential infimum of the divergence rate. More exactly,

$$-\frac{1}{n} \log p(\theta|y_1, y_2, \ldots, y_n) \to d(\theta) - d^*,$$

[7] It is also not at all clear that Savage and other founders of Bayesian decision theory ever thought that this principle should apply outside of the small worlds of artificially simplified and stylized problems – see Binmore (2007). But as scientists we care about the real, large world.

q-almost surely. Thus the posterior distribution comes to concentrate on the parts of the prior support which have the lowest values of $d(\theta)$ and the highest expected likelihood.[8] There is a geometric sense in which these parts of the parameter space are closest approaches to the truth within the support of the prior (Kass & Vos, 1997), but they may or may not be close to the truth in the sense of giving accurate values for parameters of scientific interest. They may not even be the parameter values which give the best predictions (Grünwald & Langford, 2007; Müller, 2011). In fact, one cannot even guarantee that the posterior will concentrate on a single value of θ at all; if $d(\theta)$ has multiple global minima, the posterior can alternate between (concentrating around) them forever (Berk, 1966).

To sum up, what Bayesian updating does when the model is false (i.e., in reality, always) is to try to concentrate the posterior on the best attainable approximations to the distribution of the data, 'best' being measured by likelihood. But depending on *how* the model is misspecified, and how θ represents the parameters of scientific interest, the impact of misspecification on inferring the latter can range from non-existent to profound.[9] 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. It is by this non-Bayesian checking of Bayesian models that we solve our principal–agent problem.

4. Model checking

In our view, a key part of Bayesian data analysis is model checking, which is where there are links to falsificationism. In particular, we emphasize the role of posterior predictive checks, creating simulations and comparing the simulated and actual data. Again, we are following the lead of Box (1980), Rubin (1984) and others, also mixing in a bit of Tukey (1977) in that we generally focus on visual comparisons (Gelman *et al.*, 2004, Chapter 6).

Here is how this works. A Bayesian model gives us a joint distribution for the parameters θ and the observables y. This implies a marginal distribution for the data,

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

If we have observed data y, the prior distribution $p(\theta)$ shifts to the posterior distribution $p(\theta|y)$, and so a different distribution of observables,

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta,$$

where we use y^{rep} to denote hypothetical alternative or future data, a replicated data set of the same size and shape as the original y, generated under the assumption that

[8] More precisely, regions of Θ where $d(\theta) > d^*$ tend to have exponentially small posterior probability; this statement covers situations such as $d(\theta)$ only approaching its essential infimum as $\|\theta\| \to \infty$. See Shalizi (2009) for details.

[9] White (1994) gives examples of econometric models where the influence of misspecification on the parameters of interest runs through this whole range, though only considering maximum likelihood and maximum quasi-likelihood estimation.

the fitted model, prior and likelihood both, is true. By simulating from the posterior distribution of y^{rep}, we see what typical realizations of the fitted model are like, and in particular whether the observed data set is the kind of thing that the fitted model produces with reasonably high probability.[10]

If we summarize the data with a test statistic $T(y)$, we can perform graphical comparisons with replicated data. In practice, we recommend graphical comparisons (as illustrated by our example above), but for continuity with much of the statistical literature, we focus here on p-values,

$$\Pr\left(T(y^{\text{rep}}) > T(y)|y\right),$$

which can be approximated to arbitrary accuracy as soon as we can simulate y^{rep}. (This is a valid posterior probability in the model, and its interpretation is no more problematic than that of any other probability in a Bayesian model.) In practice, we find graphical test summaries more illuminating than p-values, but in considering ideas of (probabilistic) falsification, it can be helpful to think about numerical test statistics.[11]

Under the usual understanding that T is chosen so that large values indicate poor fits, these p-values work rather like classical ones (Mayo, 1996; Mayo & Cox, 2006) – they are in fact generalizations of classical p-values, merely replacing point estimates of parameters θ with averages over the posterior distribution – and their basic logic is one of falsification. A very low p-value says that it is very improbable, under the model, to get data as extreme along the T-dimension as the actual y; we are seeing something which would be very improbable if the model were true. On the other hand, a high p-value merely indicates that $T(y)$ is an aspect of the data which would be unsurprising if the model is true. Whether this is evidence *for* the usefulness of the model depends how likely it is to get such a high p-value when the model is false: the 'severity' of the test, in the terminology of Mayo (1996) and Mayo and Cox (2006).

Put a little more abstractly, the hypothesized model makes certain probabilistic assumptions, from which other probabilistic implications follow deductively. Simulation works out what those implications are, and tests check whether the data conform to them. Extreme p-values indicate that the data violate regularities implied by the model, or approach doing so. If these were strict violations of deterministic implications, we could just apply *modus tollens* to conclude that the model was wrong; as it is, we nonetheless have evidence and probabilities. Our view of model checking, then, is firmly in the long hypothetico-deductive tradition, running from Popper (1934/1959) back through Bernard (1865/1927) and beyond (Laudan, 1981). A more direct influence on our thinking about these matters is the work of Jaynes (2003), who illustrated how

[10] For notational simplicity, we leave out the possibility of generating new values of the hidden variables $\tilde{y}$ and set aside choices of which parameters to vary and which to hold fixed in the replications; see Gelman, Meng, and Stern (1996).

[11] There is some controversy in the literature about whether posterior predictive checks have too little power to be useful statistical tools (Bayarri & Berger, 2000, 2004), how they might be modified to increase their power (Robins, van der Vaart, & Ventura, 2000; Fraser & Rousseau, 2008), whether some form of empirical prior predictive check might not be better (Bayarri & Castellanos, 2007), etc. This is not the place to rehash this debate over the interpretation or calculation of various Bayesian tail-area probabilities (Gelman, 2007). Rather, the salient fact is that all participants in the debate agree on *why* the tail-area probabilities are relevant: they make it possible to reject a Bayesian model without recourse to a specific alternative. All participants thus *disagree* with the standard inductive view, which reduces inference to the probability that a hypothesis is true, and are simply trying to find the most convenient and informative way to check Bayesian models.

we may learn the most 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 model's assumptions.[12] And the better our probability model encodes our *scientific* or *substantive* assumptions, the more we learn from specific falsification.

In this connection, the prior distribution $p(\theta)$ 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. The prior is connected to the data, and so is potentially testable, via the posterior predictive distribution of future data y^{rep}:

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta = \int p(y^{\text{rep}}|\theta)\frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}d\theta.$$

The prior distribution thus has implications for the distribution of replicated data, and so can be checked using the type of tests we have described and illustrated above.[13] When it makes sense to think of further data coming from the same source, as in certain kinds of sampling, time-series or longitudinal problems, the prior also has implications for these new data (through the same formula as above, changing the interpretation of y^{rep}), and so becomes testable in a second way. There is thus a connection between the model-checking aspect of Bayesian data analysis and 'prequentialism' (Dawid & Vovk, 1999; Grünwald, 2007), but exploring that would take us too far afield.

One advantage of recognizing that the prior distribution is a testable part of a Bayesian model is that it clarifies the role of the prior in inference, and where it comes from. To reiterate, it is hard to claim that the prior distributions used in applied work represent statisticians' states of knowledge and belief before examining their data, if only because most statisticians do not believe their models are true, so their prior degree of belief in all of Θ is not 1 but 0. The prior distribution is more like a regularization device, akin to the penalization terms added to the sum of squared errors when doing ridge regression and the lasso (Hastie, Tibshirani, & Friedman, 2009) or spline smoothing (Wahba, 1990). All such devices exploit a sensitivity–stability trade-off: they stabilize estimates and predictions by making fitted models less sensitive to certain details of the data. Using an informative prior distribution (even if only weakly informative, as in Gelman, Jakulin, Pittau, & Su, 2008) makes our estimates less sensitive to the data than, say, maximum-likelihood estimates would be, which can be a net gain.

Because we see the prior distribution as a testable part of the Bayesian model, we do not need to follow Jaynes in trying to devise a unique, objectively correct prior distribution for each situation – an enterprise with an uninspiring track record (Kass & Wasserman, 1996), even leaving aside doubts about Jaynes's specific proposal (Seidenfeld, 1979, 1987; Csiszár, 1995; Uffink, 1995, 1996). To put it even more succinctly, 'the model', for a Bayesian, is the combination of the prior distribution and

[12] A similar point was expressed by the sociologist and social historian Charles Tilly (2004, p. 597), writing from a very different disciplinary background: 'Most social researchers learn more from being wrong than from being right – provided they then recognize that they were wrong, see why they were wrong, and go on to improve their arguments. Post hoc interpretation of data minimizes the opportunity to recognize contradictions between arguments and evidence, while adoption of formalisms increases that opportunity. Formalisms blindly followed induce blindness. Intelligently adopted, however, they improve vision. Being obliged to spell out the argument, check its logical implications, and examine whether the evidence conforms to the argument promotes both visual acuity and intellectual responsibility.'

[13] Admittedly, the prior only has observable implications in conjunction with the likelihood, but for a Bayesian the reverse is also true.

the likelihood, each of which represents some compromise among scientific knowledge, mathematical convenience and computational tractability.

This gives us a lot of flexibility in modelling. We do not have to worry about making our prior distributions match our subjective beliefs, still less about our model containing all possible truths. Instead we make some assumptions, state them clearly, see what they imply, and check the implications. This applies just much to the prior distribution as it does to the parts of the model showing up in the likelihood function.

4.1. Testing to reveal problems with a model

We are not interested in falsifying our model for its own sake – among other things, having built it ourselves, we know all the shortcuts taken in doing so, and can already be morally certain it is false. With enough data, we can certainly detect departures from the model – this is why, for example, statistical folklore says that the chi-squared statistic is ultimately a measure of sample size (cf. Lindsay & Liu, 2009). As writers such as Giere (1988, Chapter 3) explain, the hypothesis linking mathematical models to empirical data is not that the data-generating process is exactly isomorphic to the model, but that the data source resembles the model closely enough, in the respects which matter to us, that reasoning based on the model will be reliable. Such reliability does not require complete fidelity to the model.

The goal of model checking, then, is not to demonstrate the foregone conclusion of falsity as such, but rather to learn how, in particular, this model fails (Gelman, 2003).[14] When we find such particular failures, they tell us how the model must be improved; when severe tests cannot find them, the inferences we draw about those aspects of the real world from our fitted model become more credible. In designing a *good* test for model checking, we are interested in finding particular errors which, if present, would mess up particular inferences, and devise a test statistic which is sensitive to this sort of misspecification. This process of examining, and ruling out, possible errors or misspecifications is of course very much in line with the 'eliminative induction' advocated by Kitcher (1993, Chapter 7).[15]

All models will have errors of approximation. Statistical models, however, typically assert that their errors of approximation will be unsystematic and patternless – 'noise' (Spanos, 2007). Testing this can be valuable in revising the model. In looking at the red-state/blue-state example, for instance, we concluded that the varying slopes mattered not just because of the magnitudes of departures from the equal-slope assumption, but also because there was a pattern, with richer states tending to have shallower slopes.

What we are advocating, then, is what Cox and Hinkley (1974) call 'pure significance testing', in which certain of the model's implications are compared directly to the data, rather than entering into a contest with some alternative model. This is, we think, more in line with what actually happens in science, where it can become clear that even

[14] In addition, no model is safe from criticism, even if it 'passes' all possible checks. Modern Bayesian models in particular are full of unobserved, latent and unobservable variables, and non-identifiability is an inevitable concern in assessing such models; see, for example, Gustafson (2005), Vansteelandt, Goetghebeur, Kenward, & Molenberghs (2006) and Greenland (2009). We find it somewhat dubious to claim that simply putting a prior distribution on non-identified quantities somehow resolves the problem; the 'bounds' or 'partial identification' approach, pioneered by Manski (2007), seems to be in better accord with scientific norms of explicitly acknowledging uncertainty (see also Vansteelandt *et al.*, 2006; Greenland, 2009).

[15] Despite the name, this is, as Kitcher notes, actually a deductive argument.

large-scale theories are in serious trouble and cannot be accepted unmodified even if there is no alternative available yet. A classical instance is the status of Newtonian physics at the beginning of the twentieth century, where there were enough difficulties – the Michaelson–Morley effect, anomalies in the orbit of Mercury, the photoelectric effect, the black-body paradox, the stability of charged matter, etc. – that it was clear, even before relativity and quantum mechanics, that something would have to give. Even today, our current best theories of fundamental physics, namely general relativity and the standard model of particle physics, an instance of quantum field theory, are universally agreed to be ultimately wrong, not least because they are mutually incompatible, and recognizing this does not require that one have a replacement theory (Weinberg, 1999).

4.2. Connection to non-Bayesian model checking

Many of these ideas about model checking are not unique to Bayesian data analysis and are used more or less explicitly by many communities of practitioners working with complex stochastic models (Ripley, 1988; Guttorp, 1995). The reasoning is the same: a model is a story of how the data could have been generated; the fitted model should therefore be able to generate synthetic data that look like the real data; failures to do so in important ways indicate faults in the model.

For instance, simulation-based model checking is now widely accepted for assessing the goodness of fit of statistical models of social networks (Hunter, Goodreau, & Handcock, 2008). That community was pushed toward predictive model checking by the observation that many model specifications were 'degenerate' in various ways (Handcock, 2003). For example, under certain exponential-family network models, the maximum likelihood estimate gave a distribution over networks which was bimodal, with both modes being very different from observed networks, but located so that the expected value of the sufficient statistics matched observations. It was thus clear that these specifications could not be right even before more adequate specifications were developed (Snijders, Pattison, Robins, & Handcock, 2006).

At a more philosophical level, the idea that a central task of statistical analysis is the search for specific, consequential errors has been forcefully advocated by Mayo (1996), Mayo and Cox (2006), Mayo and Spanos (2004), and Mayo and Spanos (2006). Mayo has placed a special emphasis on the idea of *severe* testing – a model being severely tested if it passes a probe which had a high probability of detecting an error if it is present. (The exact definition of a test's severity is related to, but not quite, that of its power; see Mayo, 1996, or Mayo & Spanos, 2006, for extensive discussions.) Something like this is implicit in discussions about the relative merits of particular posterior predictive checks (which can also be framed in a non-Bayesian manner as graphical hypothesis tests based on the parametric bootstrap).

Our contribution here is to connect this hypothetico-deductive philosophy to Bayesian data analysis, going beyond the evaluation of Bayesian methods based on their frequency properties – as recommended by Rubin (1984) and Wasserman (2006), among others – to emphasize the learning that comes from the discovery of systematic differences between model and data. At the very least, we hope this paper will motivate philosophers of hypothetico-deductive inference to take a more serious look at Bayesian data analysis (as distinct from Bayesian theory) and, conversely, motivate philosophically minded Bayesian statisticians to consider alternatives to the inductive interpretation of Bayesian learning.

4.3. Why not just compare the posterior probabilities of different models?

As mentioned above, the standard view of scientific learning in the Bayesian community is, roughly, that posterior odds of the models under consideration are compared, given the current data.[16] When Bayesian data analysis is understood as simply getting the posterior distribution, it is held that 'pure significance tests have no role to play in the Bayesian framework' (Schervish, 1995, p. 218). The dismissal rests on the idea that the prior distribution can accurately reflect our actual knowledge and beliefs.[17] At the risk of boring the reader by repetition, there is just no way we can ever have any hope of making Θ include all the probability distributions which might be correct, let alone getting $p(\theta|y)$ if we did so, so this is deeply unhelpful advice. The main point where we disagree with many Bayesians is that we do not see Bayesian methods as generally useful for giving the posterior probability that a model is true, or the probability for preferring model A over model B, or whatever.[18] Beyond the philosophical difficulties, there are technical problems with methods that purport to determine the posterior probability of models, most notably that in models with continuous parameters, aspects of the model that have essentially no effect on posterior inferences *within* a model can have huge effects on the comparison of posterior probability *among* models.[19] Bayesian inference is good for deductive inference within a model we prefer to evaluate a model by comparing it to data.

In rehashing the well-known problems with computing Bayesian posterior probabilities of models, we are not claiming that classical 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. From our model-building Bayesian perspective, the purpose of p-values (and model checking more generally) is not to reject a null hypothesis but rather to explore aspects of a model's misfit to data.

In practice, if we are in a setting where model A or model B might be true, we are inclined not to do *model selection* among these specified options, or even to perform *model averaging* over them (perhaps with a statement such as 'we assign 40% of our

---

[16] Some would prefer to compare the modification of those odds called the Bayes factor (Kass & Raftery, 1995). Everything we have to say about posterior odds carries over to Bayes factors with few changes.

[17] As Schervish (1995) continues: 'If the [parameter space Θ] describes all of the probability distributions one is willing to entertain, then one cannot reject [Θ] without rejecting probability models altogether. If one is willing to entertain models not in [Θ], then one needs to take them into account' by enlarging Θ, and computing the posterior distribution over the enlarged space.

[18] There is a vast literature on Bayes factors, model comparison, model averaging, and the evaluation of posterior probabilities of models, and although we believe most of this work to be philosophically unsound (to the extent that it is designed to be a direct vehicle for scientific learning), we recognize that these can be useful techniques. Like all statistical methods, Bayesian and otherwise, these methods are summaries of available information that can be important data-analytic tools. Even if none of a class of models is plausible as truth, and even if we are not comfortable accepting posterior model probabilities as degrees of belief in alternative models, these probabilities can still be useful as tools for prediction and for understanding structure in data, as long as these probabilities are not taken too seriously. See Raftery (1995) for a discussion of the value of posterior model probabilities in social science research and Gelman and Rubin (1995) for a discussion of their limitations, and Claeskens and Hjort (2008) for a general review of model selection. (Some of the work on 'model-selection tests' in econometrics (e.g., Vuong, 1989; Rivers & Vuong, 2002) is exempt from our strictures, as it tries to find which model is *closest* to the data-generating process, while allowing that all of the models may be misspecified, but it would take us too far afield to discuss this work in detail.)

[19] This problem has been called the Jeffreys–Lindley paradox and is the subject of a large literature. Unfortunately (from our perspective) the problem has usually been studied by Bayesians with an eye on 'solving' it – that is, coming up with reasonable definitions that allow the computation of non-degenerate posterior probabilities for continuously parameterized models – but we think that this is really a problem without a solution; see Gelman *et al.* (2004, Section 6.7).

posterior belief to A and 60% to B') but rather to do *continuous model expansion* by forming a larger model that includes both A and B as special cases. For example, Merrill (1994) used electoral and survey data from Norway and Sweden to compare two models of political ideology and voting: the 'proximity model' (in which you prefer the political party that is closest to you in some space of issues and ideology) and the 'directional model' (in which you like the parties that are in the same direction as you in issue space, but with a stronger preference to parties further from the centre). Rather than using the data to pick one model or the other, we would prefer to think of a model in which voters consider both proximity and directionality in forming their preferences (Gelman, 1994).

In the social sciences, it is rare for there to be an underlying theory that can provide meaningful constraints on the functional form of the expected relationships among variables, let alone the distribution of noise terms.[20] Taken to its limit, then, the idea of continuous model expansion counsels social scientists pretty much to give up using parametric statistical models in favour of non-parametric, infinite-dimensional models, advice which the ongoing rapid development of Bayesian non-parametrics (Ghosh & Ramamoorthi, 2003; Hjort, Holmes, Müller, & Walker, 2010) makes increasingly practical. While we are certainly sympathetic to this, and believe a greater use of nonparametric models in empirical research is desirable on its own merits (cf. Li & Racine, 2007), it is worth sounding a few notes of caution.

A technical, but important, point concerns the representation of uncertainty in Bayesian non-parametrics. In finite-dimensional problems, the use of the posterior distribution to represent uncertainty is in part supported by the Bernstein–von Mises phenomenon, which ensures that large-sample credible regions are also confidence regions. This simply fails in infinite-dimensional situations (Cox, 1993; Freedman, 1999), so that a naive use of the posterior distribution becomes unwise.[21] (Since we regard the prior and posterior distributions as regularization devices, this is not especially troublesome for us.) Relatedly, the prior distribution in a Bayesian non-parametric model is a stochastic process, always chosen for tractability (Ghosh & Ramamoorthi, 2003; Hjort *et al.*, 2010), and any pretense of representing an actual inquirer's beliefs abandoned.

Most fundamentally, switching to non-parametric models does not really resolve the issue of needing to make approximations and check their adequacy. All non-parametric models themselves embody assumptions such as conditional independence which are hard to defend except as approximations. Expanding our prior distribution to embrace *all* the models which are actually compatible with our prior knowledge would result in a mess we simply could not work with, nor interpret if we could. This being the case, we feel there is no contradiction between our preference for continuous model expansion and our use of *adequately checked* parametric models.[22]

---

[20] See Manski (2007) for a critique of the econometric practice of making modelling assumptions (such as linearity) with no support in economic theory, simply to get identifiability.

[21] Even in parametric problems, Müller (2011) shows that misspecification can lead credible intervals to have sub-optimal coverage properties – which, however, can be fixed by a modification to their usual calculation.

[22] A different perspective – common in econometrics (e.g., Wooldridge, 2002) and machine learning (e.g., Hastie *et al.*, 2009) – reduces the importance of models of the data source, either by using robust procedures that are valid under departures from modelling assumptions, or by focusing on prediction and external validation. We recognize the theoretical and practical appeal of both these approaches, which can be relevant to Bayesian inference. (For example, Rubin, 1978, justifies random assignment from a Bayesian perspective as a tool for obtaining robust inferences.) But it is not possible to work with *all* possible models when considering

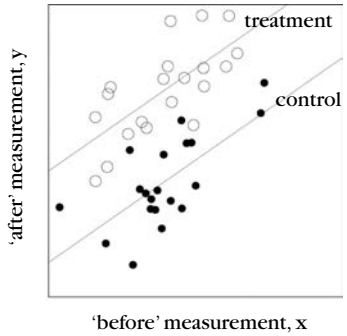

Figure 4. Sketch of the usual statistical model for before-after data. The difference between the fitted lines for the two groups is the estimated treatment effect. The default is to regress the 'after' measurement on the treatment indicator and the 'before' measurement, thus implicitly assuming parallel lines.

4.4. Example: Estimating the effects of legislative redistricting

We use one of our own experiences (Gelman & King, 1994) to illustrate scientific progress through model rejection. We began by fitting a model comparing treated and control units – state legislatures, immediately after redistricting or not – following the usual practice of assuming a constant treatment effect (parallel regression lines in 'before–after' plots, with the treatment effect representing the difference between the lines). In this example, the outcome was a measure of partisan bias, with positive values representing state legislatures where the Democrats were overrepresented (compared to how we estimated the Republicans would have done with comparable vote shares) and negative values in states where the Republicans were overrepresented. A positive treatment effect here would correspond to a redrawing of the district lines that favoured the Democrats.

Figure 4 shows the default model that we (and others) typically use for estimating causal effects in before–after data. We fitted such a no-interaction model in our example too, but then we made some graphs and realized that the model did not fit the data. The line for the control units actually had a much steeper slope than the treated units. We fitted a new model, and it had a completely different story about what the treatment effects meant.

The graph for the new model with interactions is shown in Figure 5. The largest effect of the treatment was not to benefit the Democrats or Republicans (i.e., to change the intercept in the regression, shifting the fitted line up or down) but rather to change the slope of the line, to reduce partisan bias.

Rejecting the constant-treatment-effect model and replacing it with the interaction model was, in retrospect, a crucial step in this research project. This pattern of higher before–after correlation in the control group than in the treated group is

---

fully probabilistic methods – that is, Bayesian inferences that are summarized by joint posterior distributions rather than point estimates or predictions. This difficulty may well be a motivation for shifting the foundations of statistics away from probability and scientific inference, and towards developing a technology of robust prediction. (Even when prediction is the only goal, with limited data bias–variance considerations can make even misspecified parametric models superior to non-parametric models.) This, however, goes far beyond the scope of the present paper, which aims merely to explicate the implicit philosophy guiding current practice.

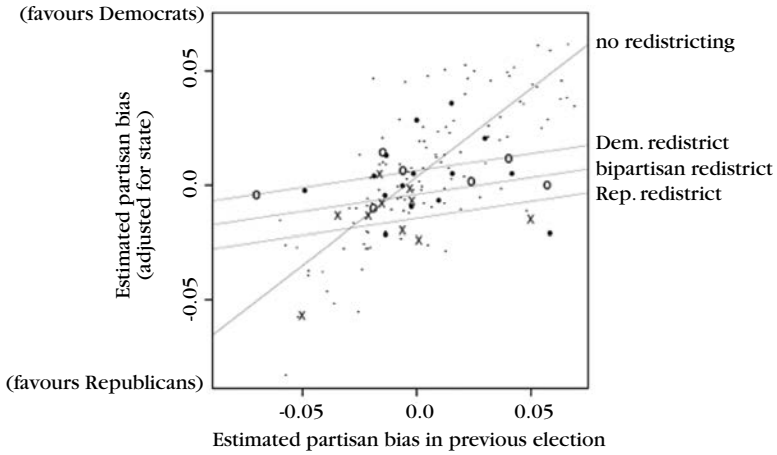

Figure 5. Effect of redistricting on partisan bias. Each symbol represents a state election year, with dots indicating controls (years with no redistricting) and the other symbols corresponding to different types of redistricting. As indicated by the fitted lines, the 'before' value is much more predictive of the 'after' value for the control cases than for the treated (redistricting) cases. The dominant effect of the treatment is to bring the expected value of partisan bias towards zero, and this effect would not be discovered with the usual approach (pictured in Figure 4), which is to fit a model assuming parallel regression lines for treated and control cases. This figure is re-drawn after Gelman and King (1994), with the permission of the authors.

quite general (Gelman, 2004), but at the time we did this study we discovered it only through the graph of model and data, which falsified the original model and motivated us to think of something better. In our experience, falsification is about plots and predictive checks, not about Bayes factors or posterior probabilities of candidate models.

The relevance of this example to the philosophy of statistics is that we began by fitting the usual regression model with no interactions. Only after visually checking the model fit – and thus falsifying it in a useful way without the specification of any alternative – did we take the crucial next step of including an interaction, which changed the whole direction of our research. The shift was induced by a falsification – a bit of deductive inference from the data and the earlier version of our model. In this case the falsification came from a graph rather than a p-value, which in one way is just a technical issue, but in a larger sense is important in that the graph revealed not just a lack of fit but also a sense of the direction of the misfit, a refutation that sent us usefully in a direction of substantive model improvement.

5. The question of induction

As we mentioned at the beginning, Bayesian inference is often held to be inductive in a way that classical statistics (following the Fisher or Neyman–Pearson traditions) is not. We need to address this, as we are arguing that all these forms of statistical reasoning are better seen as hypothetico-deductive.

The common core of various conceptions of induction is some form of inference from particulars to the general – in the statistical context, presumably, inference from

the observations y to parameters θ describing the data-generating process. But if *that* were all that was meant, then not only is 'frequentist statistics a theory of inductive inference' (Mayo & Cox, 2006), but the whole range of guess-and-test behaviors engaged in by animals (Holland, Holyoak, Nisbett, & Thagard, 1986), including those formalized in the hypothetico-deductive method, are also inductive. Even the unpromising-sounding procedure, 'pick a model at random and keep it until its accumulated error gets too big, then pick another model completely at random', would qualify (and could work surprisingly well under some circumstances – cf. Ashby, 1960; Foster & Young, 2003). So would utterly irrational procedures ('pick a new random θ when the sum of the least significant digits in y is 13'). Clearly something more is required, or at least implied, by those claiming that Bayesian updating is inductive.

One possibility for that 'something more' is to generalize the truth-preserving property of valid deductive inferences: just as valid deductions from true premises are themselves true, good inductions from true observations should also be true, at least in the limit of increasing evidence.[23] This, however, is just the requirement that our inferential procedures be consistent. As discussed above, using Bayes's rule is not sufficient to ensure consistency, nor is it necessary. In fact, every proof of Bayesian consistency known to us either posits that there is a consistent non-Bayesian procedure for the same problem, or makes other assumptions which entail the existence of such a procedure. In any case, theorems establishing consistency of statistical procedures make *deductively valid* guarantees about these procedures – they are theorems, after all – but do so on the basis of probabilistic assumptions linking future events to past data.

It is also no good to say that what makes Bayesian updating inductive is its conformity to some axiomatization of rationality. If one accepts the Kolmogorov axioms for probability, and the Savage axioms (or something like them) for decision-making,[24] then updating by conditioning follows, and a prior belief state $p(\theta)$ plus data y *deductively* entail that the new belief state is $p(\theta|y)$. In any case, lots of learning procedures can be axiomatized (all those which can be implemented algorithmically, to start with). To pick *this* system, we would need to know that it produces good results (cf. Manski, 2011), and this returns us to previous problems. To know that this axiom system leads us to approach the truth rather than become convinced of falsehoods, for instance, is just the question of consistency again.

Karl Popper, the leading advocate of hypothetico-deductivism in the last century, denied that induction was even possible; his attitude is well paraphrased by Greenland (1998) as: 'we never use any argument based on observed repetition of instances that does not also involve a hypothesis that predicts both those repetitions and the unobserved instances of interest'. This is a recent instantiation of a tradition of anti-inductive arguments that goes back to Hume, but also beyond him to al Ghazali (1100/1997) in the Middle Ages, and indeed to the ancient Sceptics (Kolakowski, 1968). As forcefully put by Stove (1982, 1986), many apparent arguments against this view of induction can be viewed as statements of abstract premises linking both the observed data and unobserved instances – various versions of the 'uniformity of nature' thesis have been popular, sometimes resolved into a set of more detailed postulates, as in

[23] We owe this suggestion to conversation with Kevin Kelly; cf. Kelly (1996, especially Chapter 13).

[24] Despite his ideas on testing, Jaynes (2003) was a prominent and emphatic advocate of the claim that Bayesian inference is the logic of inductive inference as such, but preferred to follow Cox (1946, 1961) rather than Savage. See Halpern (1999) on the formal invalidity of Cox's proofs.

Russell (1948, Part VI, Chapter 9), though Stove rather maliciously crafted a parallel argument for the existence of 'angels, or something very much like them'.[25] As Norton (2003) argues, these highly abstract premises are both dubious and often superfluous for supporting the sort of actual inferences scientists make – 'inductions' are supported not by their matching certain formal criteria (as deductions are), but rather by material facts. To generalize about the melting point of bismuth (to use one of Norton's examples) requires very few samples, provided we accept certain facts about the homogeneity of the physical properties of elemental substances; whether nature in general is uniform is not really at issue.[26]

Simply put, we think the anti-inductivist view is pretty much right, but that statistical models are tools that let us draw inductive inferences on a deductive background. Most directly, random sampling allows us to learn about unsampled people (unobserved balls in an urn, as it were), but such inference, however inductive it may appear, relies not any axiom of induction but rather on deductions from the statistical properties of random samples, and the ability to actually conduct such sampling. The appropriate design depends on many contingent material facts about the system we are studying, exactly as Norton argues.

Some results in statistical learning theory establish that certain procedures are 'probably approximately correct' in what is called a 'distribution-free' manner (Bousquet, Boucheron, & Lugosi, 2004, Vidyasagar 2003); some of these results embrace Bayesian updating (McAllister, 1999). But here 'distribution-free' just means 'holding uniformly over all distributions in a very large class', for example requiring the data to be independent and identically distributed, or from a stationary, mixing stochastic process. Another branch of learning theory does avoid making any probabilistic assumptions, getting results which hold universally across all possible data sets, and again these results apply to Bayesian updating, at least over some parameter spaces (Cesa-Bianchi & Lugosi, 2006). However, these results are all of the form 'in retrospect, the posterior predictive distribution will have predicted almost as well as the best individual model could have done', speaking entirely about performance on the past training data and revealing nothing about extrapolation to hitherto unobserved cases.

To sum up, one is free to describe statistical inference as a theory of inductive logic, but these would be inductions which are deductively guaranteed by the probabilistic assumptions of stochastic models. We can see no interesting and correct sense in which Bayesian statistics is a logic of induction which does not equally imply that frequentist statistics is also a theory of inductive inference (cf. Mayo & Cox, 2006), which is to say, not very inductive at all.

---

[25] Stove (1986) further argues that induction by simple enumeration is reliable *without* making such assumptions, at least sometimes. However, his calculations make no sense unless his data are independent and identically distributed.

[26] Within environments where such premises hold, it may of course be adaptive for organisms to develop inductive propensities, whose scope would be more or less tied to the domain of the relevant material premises. Barkow, Cosmides, and Tooby (1992) develop this theme with reference to the evolution of domain-specific mechanisms of learning and induction; Gigerenzer (2000) and Gigerenzer, Todd, and ABC Research Group (1999) consider proximate mechanisms and ecological aspects, and Holland *et al.* (1986) propose a unified framework for modelling such inductive propensities in terms of generate-and-test processes. All of this, however, is more within the field of psychology than either statistics or philosophy, as (to paraphrase the philosopher Ian Hacking, 2001) it does not so much solve the problem of induction as evade it.

6. What about Popper and Kuhn?

The two most famous modern philosophers of science are undoubtedly Karl Popper (1934/1959) and Thomas Kuhn (1970), and if statisticians (like other non-philosophers) know about philosophy of science at all, it is generally some version of their ideas. It may therefore help readers to see how our ideas relate to theirs. We do not pretend that our sketch fully portrays these figures, let alone the literatures of exegesis and controversy they inspired, or even how the philosophy of science has moved on since 1970.

Popper's key idea was that of 'falsification' or 'conjectures and refutations'. The inspiring example, for Popper, was the replacement of classical physics, after several centuries as the core of the best-established science, by modern physics, especially the replacement of Newtonian gravitation by Einstein's general relativity. Science, for Popper, advances by scientists advancing theories which make strong, wide-ranging predictions capable of being refuted by observations. A good experiment or observational study is one which tests a specific theory (or theories) by confronting their predictions with data in such a way that a match is not automatically assured; good studies are designed with theories in mind, to give them a chance to fail. Theories which conflict with any evidence must be rejected, since a single counter-example implies that a generalization is false. Theories which are not falsifiable by any conceivable evidence are, for Popper, simply not scientific, though they may have other virtues.[27] Even those falsifiable theories which have survived contact with data so far must be regarded as more or less provisional, since no finite amount of data can ever establish a generalization, nor is there any non-circular principle of induction which could let us regard theories which are compatible with lots of evidence as probably true.[28] Since people are fallible, and often obstinate and overly fond of their own ideas, the objectivity of the process which tests conjectures lies not in the emotional detachment and impartiality of individual scientists, but rather in the scientific community being organized in certain ways, with certain institutions, norms and traditions, so that individuals' prejudices more or less wash out (Popper, 1945, Chapters 23–24).

Clearly, we find much here to agree with, especially the general hypothetico-deductive view of scientific method and the anti-inductivist stance. On the other hand, Popper's specific ideas about testing require, at the least, substantial modification. His idea of a test comes down to the rule of deduction which says that if p implies q, and q is false, then p must be false, with the roles of p and q being played by hypotheses and data, respectively. This is plainly inadequate for statistical hypotheses, yet, as critics have noted since Braithwaite (1953) at least, he oddly ignored the theory of statistical hypothesis testing.[29] It is possible to do better, both through standard hypothesis tests and the kind of predictive checks we have described. In particular, as Mayo (1996) has emphasized, it is vital to consider the *severity* of tests, their capacity to detect violations of hypotheses when they are present.

Popper tried to say how science *ought* to work, supplemented by arguments that his ideals could at least be approximated and often had been. Kuhn's work, in contrast,

---

[27] This 'demarcation criterion' has received a lot of criticism, much of it justified. The question of what makes something 'scientific' is fortunately not one we have to answer; cf. Laudan (1996, Chapters 11–12) and Ziman (2000).

[28] Popper tried to work out notions of 'corroboration' and increasing truth content, or 'verisimilitude', to fit with these stances, but these are generally regarded as failures.

[29] We have generally found Popper's ideas on probability and statistics to be of little use and will not discuss them here.

was much more an attempt to describe how science had, in point of historical fact, developed, supported by arguments that alternatives were infeasible, from which some morals might be drawn. His central idea was that of a 'paradigm', a scientific problem and its solution which served as a model or exemplar, so that solutions to other problems could be developed in imitation of it.[30] Paradigms come along with presuppositions about the terms available for describing problems and their solutions, what counts as a valid problem, what counts as a solution, background assumptions which can be taken as a matter of course, etc. Once a scientific community accepts a paradigm and all that goes with it, its members can communicate with one another and get on with the business of solving puzzles, rather than arguing about what they should be doing. Such 'normal science' includes a certain amount of developing and testing of hypotheses but leaves the central presuppositions of the paradigm unquestioned.

During periods of normal science, according to Kuhn, there will always be some 'anomalies' – things within the domain of the paradigm which it currently cannot explain, or which even seem to refute its assumptions. These are generally ignored, or at most regarded as problems which somebody ought to investigate eventually. (Is a special adjustment for odd local circumstances called for? Might there be some clever calculational trick which fixes things? How sound are those anomalous observations?) More formally, Kuhn invokes the 'Quine–Duhem thesis' (Quine, 1961; Duhem, 1914/1954). A paradigm only makes predictions about observations in conjunction with 'auxiliary' hypotheses about specific circumstances, measurement procedures, etc. If the predictions are wrong, Quine and Duhem claimed that one is always free to fix the blame on the auxiliary hypotheses, and preserve belief in the core assumptions of the paradigm 'come what may'.[31] The Quine–Duhem thesis was also used by Lakatos (1978) as part of his 'methodology of scientific research programmes', a falsificationism more historically oriented than Popper's distinguishing between progressive development of auxiliary hypotheses and degenerate research programmes where auxiliaries become *ad hoc* devices for saving core assumptions from data.

According to Kuhn, however, anomalies can accumulate, becoming so serious as to create a crisis for the paradigm, beginning a period of 'revolutionary science'. It is then that a new paradigm can form, one which is generally 'incommensurable' with the old: it makes different presuppositions, takes a different problem and its solution as exemplars, redefines the meaning of terms. Kuhn insisted that scientists who retain the old paradigm are not being irrational, because (by the Quine–Duhem thesis) they can always explain away the anomalies *somehow*; but neither are the scientists who embrace and develop the new paradigm being irrational. Switching to the new paradigm is more like a bistable illusion flipping (the apparent duck becomes an obvious rabbit) than any process of ratiocination governed by sound rules of method.[32]

---

[30] Examples are Newton's deduction of Kepler's laws of planetary motion and other facts of astronomy from the inverse square law of gravitation, and Planck's derivation of the black-body radiation distribution from Boltzmann's statistical mechanics and the quantization of the electromagnetic field. An internal example for statistics might be the way the Neyman–Pearson lemma inspired the search for uniformly most powerful tests in a variety of complicated situations.

[31] This thesis can be attacked from many directions, perhaps the most vulnerable being that one can often find multiple lines of evidence which bear on either the main principles or the auxiliary hypotheses *separately*, thereby localizing the problems (Glymour, 1980; Kitcher, 1993; Laudan, 1996; Mayo, 1996).

[32] Salmon (1990) proposed a connection between Kuhn and Bayesian reasoning, suggesting that the choice between paradigms could be made rationally by using Bayes's rule to compute their posterior probabilities, with the prior probabilities for the paradigms encoding such things as preferences for parsimony. This has

In some way, Kuhn's distinction between normal and revolutionary science is analogous to the distinction between learning within a Bayesian model, and checking the model in preparation to discarding or expanding it. Just as the work of normal science proceeds within the presuppositions of the paradigm, updating a posterior distribution by conditioning on new data takes the assumptions embodied in the prior distribution and the likelihood function as unchallengeable truths. Model checking, on the other hand, corresponds to the identification of anomalies, with a switch to a new model when they become intolerable. Even the problems with translations between paradigms have something of a counterpart in statistical practice; for example, the intercept coefficients in a varying-intercept, constant-slope regression model have a somewhat different meaning than do the intercepts in a varying-slope model. We do not want to push the analogy too far, however, since most model checking and model reformulation would by Kuhn have been regarded as puzzle-solving within a single paradigm, and his views of how people switch between paradigms are, as we just saw, rather different.

Kuhn's ideas about scientific revolutions are famous because they raise so many disturbing questions about the scientific enterprise. For instance, there has been considerable controversy over whether Kuhn believed in any notion of scientific progress, and over whether or not he should have, given his theory. Yet detailed historical case studies (Donovan, Laudan, & Laudan, 1988) have shown that Kuhn's picture of sharp breaks between normal and revolutionary science is hard to sustain.[33] The leads to a tendency, already remarked by Toulmin (1972, pp. 112–117), either to expand paradigms or to shrink them. Expanding paradigms into persistent and all-embracing, because abstract and vague, bodies of ideas lets one preserve the idea of abrupt breaks in thought, but makes them rare and leaves almost everything to puzzle-solving normal science. (In the limit, there has only been one paradigm in astronomy since the Mesopotamians, something like 'many lights in the night sky are objects which are very large but very far away, and they move in interrelated, mathematically describable, discernible patterns'.) This corresponds, we might say, to relentlessly enlarging the support of the prior. The other alternative is to shrink paradigms into increasingly concrete, specific theories and even models, making the standard for a 'revolutionary' change very small indeed, in the limit reaching any kind of conceptual change whatsoever.

We suggest that there is actually some validity to both moves, that there is a sort of (weak) self-similarity involved in scientific change. Every scale of size and complexity, from local problem-solving to big-picture science, features progress of the 'normal science' type, punctuated by occasional revolutions. For example, in working on an applied research or consulting problem, one typically will start in a certain direction, then suddenly realize one was thinking about it incorrectly, then move forward, and so forth. In a consulting setting, this re-evaluation can happen several times in a couple of

---

at least three big problems. First, all our earlier objections to using posterior probabilities to chose between theories apply, with all the more force because every paradigm is compatible with a broad range of specific theories. Second, devising priors encoding those methodological preferences – particularly a non-vacuous preference for parsimony – is hard or impossible in practice (Kelly, 2010). Third, it implies a truly remarkable form of Platonism: for scientists to give a paradigm positive posterior probability, they must, by Bayes's rule, have always given it strictly positive prior probability, *even before having encountered a statement of the paradigm*.

[33] Arguably this is true even of Kuhn (1957).

hours. At a slightly longer time scale, we commonly reassess any approach to an applied problem after a few months, realizing there was some key feature of the problem we were misunderstanding, and so forth. There is a link between the size and the typical time scales of these changes, with small revolutions occurring fairly frequently (every few minutes for an exam-type problem), up to every few decades for a major scientific consensus. (This is related to but somewhat different from the recursive subject-matter divisions discussed by Abbott, 2001.) The big changes are more exciting, even glamorous, but they rest on the hard work of extending the implications of theories far enough that they can be decisively refuted.

To sum up, our views are much closer to Popper's than to Kuhn's. The latter encouraged a close attention to the history of science and to explaining the process of scientific change, as well as putting on the agenda many genuinely deep questions, such as when and how scientific fields achieve consensus. There are even analogies between Kuhn's ideas and what happens in good data-analytic practice. Fundamentally, however, we feel that deductive model checking is central to statistical and scientific progress, and that it is the threat of such checks that motivates us to perform inferences within complex models that we know ahead of time to be false.

7. Why does this matter?

Philosophy matters to practitioners because they use it to guide their practice; even those who believe themselves quite exempt from any philosophical influences are usually the slaves of some defunct methodologist. The idea of Bayesian inference as inductive, culminating in the computation of the posterior probability of scientific hypotheses, has had malign effects on statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework.

In our hypothetico-deductive view of data analysis, we build a statistical model out of available parts and drive it as far as it can take us, and then a little farther. When the model breaks down, we dissect it and figure out what went wrong. For Bayesian models, the most useful way of figuring out how the model breaks down is through posterior predictive checks, creating simulations of the data and comparing them to the actual data. The comparison can often be done visually; see Gelman *et al.* (2004, Chapter 6) for a range of examples. Once we have an idea about where the problem lies, we can tinker with the model, or perhaps try a radically new design. Either way, we are using deductive reasoning as a tool to get the most out of a model, and we test the model – it is falsifiable, and when it is consequentially falsified, we alter or abandon it. None of this is especially subjective, or at least no more so than any other kind of scientific inquiry, which likewise requires choices as to the problem to study, the data to use, the models to employ, etc. – but these choices are by no means arbitrary whims, uncontrolled by objective conditions.

Conversely, a problem with the inductive philosophy of Bayesian statistics – in which science 'learns' by updating the probabilities that various competing models are true – is that it assumes that the true model (or, at least, the models among which we will choose or over which we will average) is one of the possibilities being considered. This does

not fit our own experiences of learning by finding that a model does not fit and needing to expand beyond the existing class of models to fix the problem.

Our methodological suggestions are to construct large models that are capable of incorporating diverse sources of data, to use Bayesian inference to summarize uncertainty about parameters in the models, to use graphical model checks to understand the limitations of the models, and to move forward via continuous model expansion rather than model selection or discrete model averaging. Again, we do not claim any novelty in these ideas, which we and others have presented in many publications and which reflect decades of statistical practice, expressed particularly forcefully in recent times by Box (1980) and Jaynes (2003). These ideas, important as they are, are hardly ground-breaking advances in statistical methodology. Rather, the point of this paper is to demonstrate that our commonplace (if not universally accepted) approach to the practice of Bayesian statistics is compatible with a hypothetico-deductive framework for the philosophy of science.

We fear that a philosophy of Bayesian statistics as subjective, inductive inference can encourage a complacency about picking or averaging over existing models rather than trying to falsify and go further.[34] Likelihood and Bayesian inference are powerful, and with great power comes great responsibility. Complex models can and should be checked and falsified. This is how we can learn from our mistakes.

Acknowledgements

We thank the National Security Agency for grant H98230-10-1-0184, the Department of Energy for grant DE-SC0002099, the Institute of Education Sciences for grants ED-GRANTS-032309-005 and R305D090006-09A, and the National Science Foundation for grants ATM-0934516, SES-1023176 and SES-1023189. We thank Wolfgang Beirl, Chris Genovese, Clark Glymour, Mark Handcock, Jay Kadane, Rob Kass, Kevin Kelly, Kristina Klinkner, Deborah Mayo, Martina Morris, Scott Page, Aris Spanos, Erik van Nimwegen, Larry Wasserman, Chris Wiggins, and two anonymous reviewers for helpful conversations and suggestions.

References

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

al Ghazali, Abu Hamid Muhammad ibn Muhammad at-Tusi (1100/1997). *The incoherence of the philosophers = Tahafut al-falasifah: A parallel English-Arabic text*, trans. M. E. Marmura. Provo, UT: Brigham Young University Press.

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behaviour* (2nd ed.). London: Chapman & Hall.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Oxford: Clarendon Press.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.) (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press.

Bartlett, M. S. (1967). Inference and stochastic processes. *Journal of the Royal Statistical Society, Series A*, *130*, 457–478.

[34] Ghosh and Ramamoorthi (2003, p. 112) see a similar attitude as discouraging inquiries into consistency: 'the prior and the posterior given by Bayes theorem [*sic*] are imperatives arising out of axioms of rational behavior – and since we are already rational why worry about one more' criterion, namely convergence to the truth?

Bayarri, M. J., & Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58–80. doi:10.1214/088342304000000116

Bayarri, M. J., & Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, *22*, 322–343.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: Irreconcilability of *p*-values and evidence. *Journal of the American Statistical Association*, *82*, 112–122.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, *37*, 51–58. doi:10.1214/aoms/1177699597 Correction: *37* (1966), 745–746.

Berk, R. H. (1970). Consistency a posteriori. *Annals of Mathematical Statistics*, *41*, 894–906. doi:10.1214/aoms/1177696967

Bernard, C. (1865/1927). *Introduction to the study of experimental medicine*, trans. H. C. Greene. New York: Macmillan. First published as *Introduction à l'étude de la médecine experimentale*, Paris: J. B. Baillière. Reprinted New York: Dover, 1957.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Binmore, K. (2007). Making decisions in large worlds. Technical Report 266, ESRC Centre for Economic Learning and Social Evolution, University College London. Retrieved from http://else.econ.ucl.ac.uk/papers/uploaded/266.pdf

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures in machine learning* (pp. 169–207). Berlin: Springer.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, *143*, 383–430.

Box, G. E. P. (1983). An apology for ecumenism in statistics. In G. E. P. Box, T. Leonard & C.-F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 51–84). New York: Academic Press.

Box, G. E. P. (1990). Comment on 'The unity and diversity of probability' by Glen Shafer. *Statistical Science*, *5*, 448–449. doi:10.1214/ss/1177012024

Braithwaite, R. B. (1953). *Scientific explanation: A study of the function of theory, probability and law in science*. Cambridge: Cambridge University Press.

Brown, R. Z., Sallow, W., Davis, D. E., & Cochran, W. G. (1955). The rat population of Baltimore, 1952. *American Journal of Epidemiology*, *61*, 89–102.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Annals of Statistics*, *21*, 903–923. doi:10.1214/aos/1176349157

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins University Press.

Csiszár, I. (1995). Maxent, mathematics, and information theory. In K. M. Hanson& R. N. Silver (Eds.), *Maximum entropy and Bayesian methods: Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* (pp. 35–50). Dordrecht: Kluwer Academic.

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, *5*, 125–162. Retrieved from: http://projecteuclid.org/euclid.bj/1173707098

Donovan, A., Laudan, L., & Laudan, R. (Eds.), (1988). *Scrutinizing science: Empirical studies of scientific change*. Dordrecht: Kluwer Academic. Reprinted 1992 (Baltimore, MD: Johns Hopkins University Press) with a new introduction.

Doob, J. L. (1949). Application of the theory of martingales. In *Colloques internationaux du Centre National de la Recherche Scientifique*, Vol. 13 (pp. 23–27). Paris: Centre National de la Recherche Scientifique.

Duhem, P. (1914/1954). *The aim and structure of physical theory*, trans. P. P. Wiener. Princeton, NJ: Princeton University Press.

Earman, J. (1992). *Bayes or bust? A critical account of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Eggertsson, T. (1990). *Economic behavior and institutions*. Cambridge: Cambridge University Press.

Fitelson, B., & Thomason, N. (2008). Bayesians sometimes cannot ignore even very implausible theories (even ones that have not yet been thought of). *Australasian Journal of Logic*, *6*, 25–36. Retrieved from: http://philosophy.unimelb.edu.au/ajl/2008/2008_2.pdf

Foster, D. P., & Young, H. P. (2003). Learning, hypothesis testing and Nash equilibrium. *Games and Economic Behavior*, *45*, 73–96. doi:10.1016/S0899-8256(03)00025-3

Fraser, D. A. S., & Rousseau, J. (2008). Studentization and deriving accurate *p*-values. *Biometrika*, *95*, 1–16. doi:10.1093/biomet/asm093

Freedman, D. A. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, *27*, 1119–1140. doi:10.1214/aos/1017938917

Gelman, A. (1994). Discussion of 'A probabilistic model for the spatial distribution of party support in multiparty elections' by S. Merrill. *Journal of the American Statistical Association*, *89*, 1198.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, *71*, 369–382. doi:10.1111/j.1751-5823.2003.tb00203.x

Gelman, A. (2004). Treatment effects in before-after data. In A. Gelman& X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 191–198). Chichester: Wiley.

Gelman, A. (2007). Comment: 'Bayesian checking of the second levels of hierarchical models'. *Statistical Science*, *22*, 349–352. doi:10.1214/07-STS235A

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383. doi:10.1214/08-AOAS191

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

Gelman, A., Lee, D., & Ghitza, Y. (2010). Public opinion on health care reform. *The Forum*, *8*(1). doi:10.2202/1540-8884.1355

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733–807. Retrieved from: http://www3.stat.sinica.edu.tw/statistica/j6n4/j6n41/j6n41.htm

Gelman, A., Park, D., Shor, B., Bafumi, J., & Cortina, J. (2008). *Red state, blue state, rich state, poor state: Why Americans vote the way they do*. Princeton, NJ: Princeton University Press. doi:10.1561/100.00006026

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173.

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

Ghitza, Y., & Gelman, A. (2012). *Deep interactions with MRP: presidential turnout and voting patterns among small electoral subgroups*. Technical report, Department of Political Science, Columbia University.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. New York: Springer.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Good, I. J., & Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, *69*, 711–720.

Gray, R. M. (1990). *Entropy and information theory*. New York: Springer.

Greenland, S. (1998). Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology*, *27*, 543–548. doi:10.1093/ije/27.4.543

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science*, *24*, 195–210. doi:10.1214/09-STS291

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P. D., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*, 119–149. doi:10.1007/s10994-007-0716-7

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, *20*, 111–140. doi:10.1214/088342305000000098

Guttorp, P. (1995). *Stochastic modeling of scientific data*. London: Chapman & Hall.

Haack, S. (1993). *Evidence and inquiry: Towards reconstruction in epistemology*. Oxford: Blackwell.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

Halpern, J. Y. (1999). Cox's theorem revisited. *Journal of Artificial Intelligence Research*, *11*, 429–435. doi:10.1613/jair.644

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Retrieved from http://www.csss.washington.edu/Papers/wp39.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Hempel, C. G. (1965). *Aspects of scientific explanation*. Glencoe, IL: Free Press.

Hill, J. R. (1990). A general framework for model-based statistics. *Biometrika*, *77*, 115–126.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (Eds.), (2010). *Bayesian nonparametrics*. Cambridge: Cambridge University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*, 248–258. doi:10.1198/016214507000000446

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kass, R. E., & Vos, P. W. (1997). *Geometrical foundations of asymptotic inference*. New York: Wiley.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1370.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford: Oxford University Press.

Kelly, K. T. (2010). Simplicity, truth, and probability. In P. Bandyopadhyay & M. Forster (Eds.), *Handbook on the philosophy of statistics*. Dordrecht: Elsevier.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford: Oxford University Press.

Kleijn, B. J. K., & van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, *34*, 837–877. doi:10.1214/009053606000000029

Kolakowski, L. (1968). *The alienation of reason: A history of positivist thought*, trans. N. Guterman. Garden City, NY: Doubleday.

Kuhn, T. S. (1957). *The Copernican revolution: Planetary astronomy in the development of western thought*. Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lakatos, I. (1978). *Philosophical papers*. Cambridge: Cambridge University Press.

Laudan, L. (1996). *Beyond positivism and relativism: Theory, method and evidence*. Boulder, Colorado: Westview Press.

Laudan, L. (1981). *Science and hypothesis*. Dodrecht: D. Reidel.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Lijoi, A., Prünster, I., & Walker, S. G. (2007). Bayesian consistency for stationary models. *Econometric Theory*, *23*, 749–759. doi:10.1017/S0266466607070314

Lindsay, B., & Liu, L. (2009). Model assessment tools for a model false world. *Statistical Science*, *24*, 303–318. doi:10.1214/09-STS302

Manski, C. F. (2007). *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

Manski, C. F. (2011). Actualist rationality. *Theory and Decision*, 71. doi:10.1007/s11238-009-9182-y

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

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium* (pp. 77–97). Bethesda, MD: Institute of Mathematical Statistics.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, *71*, 1007–1025.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323–357. doi:10.1093/bjps/axl003

McAllister, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, *37*, 355–363. doi:10.1023/A:1007618624809

McCarty, N., Poole, K. T., & Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches*. Cambridge, MA: MIT Press.

Merrill III, S. (1994). A probabilistic model for the spatial distribution of party support in multiparty electorates. *Journal of the American Statistical Association*, *89*, 1190–1197.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092. doi:10.1063/1.1699114

Morris, C. N. (1986). Comment on 'Why isn't everyone a Bayesian?'. *American Statistician*, *40*, 7–8.

Müller, U. K. (2011). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, submitted. Retrieved from http://www.princeton.edu/~umueller/sandwich.pdf

Newman, M. E. J., & Barkema, G. T. (1999). *Monte Carlo methods in statistical physics*. Oxford: Clarendon Press.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*, 647–670. doi:10.1086/378858

Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, *17*, 1480–1507. doi:10.1162/0899766053723032

Popper, K. R. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. R. (1945). *The open society and its enemies*. London: Routledge.

Quine, W. V. O. (1961). *From a logical point of view: Logico-philosophical essays* (2nd ed.). Cambridge, MA: Harvard University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–196.

Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge: Cambridge University Press.

Rivers, D., & Vuong, Q. H. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, *5*, 1–39. doi:10.1111/1368-423X.t01-1-00071

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of *p* values in composite null models (with discussions and rejoinder). *Journal of the American Statistical Association*, *95*, 1143–1172.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58. doi:10.1214/aos/1176344064

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

Russell, B. (1948). *Human knowledge: Its scope and limits*. New York: Simon and Schuster.

Salmon, W. C. (1990). The appraisal of theories: Kuhn meets Bayes. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2, pp. 325–332). Chicago: University of Chicago Press.

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

Schervish, M. J. (1995). *Theory of statistics*. Berlin: Springer.

Seidenfeld, T. (1979). Why I am not an objective Bayesian: Some reflections prompted by Rosenkrantz. *Theory and Decision*, *11*, 413–440. doi:10.1007/BF00139451

Seidenfeld, T. (1987). Entropy and uncertainty. In I. B. MacNeill& G. J. Umphrey (Eds.), *Foundations of statistical inference* (pp. 259–287). Dordrecht: D. Reidel.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, *3*, 1039–1074. doi:10.1214/09-EJS485

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153. doi:10.1111/j.1467-9531.2006.00176.x

Spanos, A. (2007). Curve fitting, the reliability of inductive inference, and the error-statistical approach. *Philosophy of Science*, *74*, 1046–1066. doi:10.1086/525643

Stove, D. C. (1982). *Popper and after: Four modern irrationalists*. Oxford: Pergamon Press.

Stove, D. C. (1986). *The rationality of induction*. Oxford: Clarendon Press.

Tilly, C. (2004). Observations of social processes and their formal representations. *Sociological Theory*, *22*, 595–602. Reprinted in Tilly (2008). doi:10.1111/j.0735-2751.2004.00235.x

Tilly, C. (2008). *Explaining social processes*. Boulder, CO: Paradigm.

Toulmin, S. (1972). *Human understanding: The collective use and evolution of concepts*. Princeton, NJ: Princeton University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in the History and Philosophy of Modern Physics*, *26B*, 223–261. doi:10.1016/1355-2198(95)00015-1

Uffink, J. (1996). The constraint rule of the maximum entropy principle. *Studies in History and Philosophy of Modern Physics*, *27*, 47–79. doi:10.1016/1355-2198(95)00022-4

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., & Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, *16*, 953–980.

Vidyasagar, M. (2003). *Learning and generalization: With applications to neural networks* (2nd ed.). Berlin: Springer.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

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

Weinberg, S. (1999). What is quantum field theory, and what did we think it was? In T. Y. Cao (Ed.), *Conceptual foundations of quantum field theory* (pp. 241–251). Cambridge: Cambridge University Press.

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge: Cambridge University Press.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Ziman, J. (2000). *Real science: What it is, and what it means*. Cambridge: Cambridge University Press.

Received 28 June 2011; revised version received 6 December 2011

British Journal of Mathematical and Statistical Psychology (2013), 66, 39–44

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

How to practise Bayesian statistics outside the Bayesian church: What philosophy for Bayesian statistical modelling?

Denny Borsboom[1]* and Brian D. Haig[2]

[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]Department of Psychology, University of Canterbury, New Zealand

Then I saw Tom Bayes – Now I'm a believer,
Without a trace – of doubt in my mind!
I'm a Bayesian (oooh) – Oh I'm a believer –
I couldn't *p* now if I tried!
- Brad Carlin, *Bayesian Believer*

1. Introduction

Unlike most other statistical frameworks, Bayesian statistical inference is wedded to a particular approach in the philosophy of science (see Howson & Urbach, 2006); this approach is called *Bayesianism.* Rather than being concerned with model fitting, this position in the philosophy of science primarily addresses theory choice. Naturally, in some cases there exists a relation between scientific theories and statistical models, and this relation can be so tight that choosing the model is tantamount to accepting the theory. However, in many cases of data analysis, the statistical model bears only an indirect relation to scientific theory, and in such cases the act of statistical modelling is distinct from the act of theory choice.

If one takes seriously the distinction between statistical modelling and theory evaluation, it becomes clear that one who takes a Bayesian approach in one of these areas need not take it in the other. Thus, one who utilizes statistical techniques that are known as 'Bayesian' (e.g., computes a Bayes factor) may not adhere to the philosophy of science that goes by the same name; and one committed to the philosophy of Bayesianism may, for a host of reasons, employ techniques that arise from non-Bayesian statistics (e.g., a classical *t*-test). The former stance, where one utilizes Bayesian machinery without adhering to the Bayesian account of science, has become quite commonplace in statistical

*Correspondence should be addressed to Denny Borsboom, Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands (e-mail: dennyborsboom@gmail.com).

DOI:10.1111/j.2044-8317.2012.02062.x

modelling. This is because Bayesian model-fitting routines occasionally offer considerable resources where frequentist approaches struggle, and thus there is a pragmatic reason to use them.

At conference dinners, modellers who use Bayesian statistics without adhering to Bayesianist philosophy are usually called 'practical Bayesians'. The paper by Gelman and Shalizi (2013) is perhaps the first attempt to systematically justify and underpin the position of the practical Bayesian, and thereby offers a welcome addition to the literature. Gelman and Shalizi's alternative philosophy is primarily shaped by Popper's view that scientific propositions are to be submitted to repeated criticism in the form of strong empirical tests. For them, best Bayesian statistical practice involves formulating models using Bayesian statistical methods, and then checking them through attempts to falsify and modify them. At the same time, Gelman and Shalizi reject the orthodox Bayesian view that statistical inference is inductive inference, which involves updating subjective probability estimates that hypotheses are true. Their hope is that the philosophical foundations they provide for Bayesian statistical practice will benefit both the practice of statistics and the philosophy of science.

We welcome Gelman and Shalizi's paper because it serves to explicate the idea that, to practise Bayesian statistics, one need not first be converted and baptized in the Bayesian church. Practical Bayesianism is a viable position that can be taken by one who subscribes to non-Bayesian philosophy of science without internal contradiction. As Gelman and Shalizi show, many of the procedures commonly followed in Bayesian statistics can be accommodated by alternative perspectives. However, the particular account that Gelman and Shalizi have is not the only alternative, and it is questionable whether it is adequate in all situations where Bayesian statistical modelling arises. In this commentary, we will draw attention to alternatives that may be more fruitful in accommodating Bayesian practice.

2. The philosophy of science and the philosophy of data analysis

Bayesianism, in the philosophy of science, holds that rational agents should update their belief in various theories according to the famous formula $P(T|E) = P(E|T)P(T)/P(E)$, where T denotes a theory and E the evidence. Bayesian statistics works on the idea that models should be evaluated via the analogous formula $P(M|D) = P(D|M)P(M)/P(D)$, in which M is a statistical model and D the data. In discussions on Bayesian statistics, it often tacitly assumed that $T = M$ and $E = D$, so that statistical modelling is a special case of Bayesian philosophy. However, in many cases, these identifications are not viable. In particular, it is hard to uphold that statistical model and substantive theory are typically one and the same.

For instance, we may entertain the theory that males and females do not structurally differ in intelligence. In statistical work, we may then subsequently analyse IQ scores and test the statistical model $F(\text{IQ}) = F(\text{IQ}|\text{Sex})$, where $F(\cdot)$ denotes the population distribution function. The statistical model is not identical to the theory, unless one is willing to assume that IQ and intelligence are interchangeable terms, which is a gross simplification on any of the psychometric theories of intelligence currently on offer. One can readily see the importance of this distinction by considering the appropriate theoretical move upon finding, in empirical work, that the statistical model is not appropriate for the data, so that the distributions of IQ are not invariant across sex. Clearly, one can accept this conclusion without necessarily accepting that intelligence differs across the sexes.

Because the theory is not identical to the statistical model, one has to fix a relation between them to get the scientific process going. Given that identity of T and M is clearly not an option, one may alternatively construct the relation to be one of logical implication (i.e., it is assumed that T entails M), but even this is, we think, is too much to honour when faced with the exceedingly messy process of statistical testing in science. Even in order to get the entailment of M from T, one has to fix many auxiliary hypotheses (e.g., the validity of the IQ test as measure of intelligence, the distribution of error scores, the dichotomous treatment of sex, etc.), and if one ponders this issue a little it is evident that there are not just many such hypotheses, but an infinity of them. As a result, our working hypothesis ought to be that statistical inference and theory choice are not games that are played in the same ball park.

Given this conclusion, one has to wonder whether it really makes any difference for our evaluation of the theoretical hypothesis in question whether we evaluate the statistical hypothesis $F(\text{IQ}) = F(\text{IQ}|\text{Sex})$ through Bayesian statistics or otherwise. Suppose John evaluates the hypothesis by computing a posterior distribution for the statistical hypothesis. Is John then automatically mandated to generalize his Bayesian behaviour to the evaluation of the theory? Suppose Jane simply evaluates the hypothesis using a frequentist t-test. Does this preclude Jane from entering the conclusion into a Bayesian chain of reasoning to find out what she should believe about the theory in question? We think it is clear that no such generalizations follow. John could evaluate his results according to a hypothetico-deductive scheme, while Jane could plug hers into the Bayesian philosophical framework. There is no reason suppose that statistical data analysis and theory evaluation should follow the same theoretical precept.

This, we think, is one of the most important conclusions to draw from the paper by Gelman and Shalizi, who work out the connection between a Bayesian data-analytic framework and a hypothetico-deductive account of theory evaluation in detail. However, their vigorous defence of this analysis, which claims that Bayesian practice sits 'much better' with hypothetico-deductive accounts across the board, suggests that Bayesian practitioner should now trade their Bayesian account for Gelman and Shalizi's hypothetico-deductive theory. This, we think, is a *non sequitur*, for two reasons. First, that Gelman and Shalizi are *able* to accommodate Bayesian practice in a hypothetico-deductive framework does not entail that such practice could not be accommodated by one of the many *other* philosophies of science that have been developed in the past centuries. Second, the philosophy of science that Gelman and Shalizi have in store is unnecessarily impoverished – so impoverished, in fact, that it is doubtful whether their version of hypothetico-deductivism is able to sustain their data-analytic work in the first place.

3. Popper and the hypothetico-deductive method

Gelman and Shalizi maintain that formulating, checking and revising models accords well with sophisticated forms of hypothetico-deductive inference. However, for them, sophisticated hypothetico-deductivism essentially amounts to a meagre version of Popper's (1969) falsificationist view of hypothetico-deductive method. By stripping away many of the elements of Popper's philosophy of critical rationalism that make his falsificationist theory of science a rich account of scientific inquiry, the authors are left with an impoverished and rather unsophisticated account of the hypothetico-deductive method. Notably, Gelman and Shalizi reject Popper's confirmation-theoretic notion of corroboration, but they offer nothing in its place. Further, perhaps influenced by Popper's

view that there is no logic to discovery, they offer no methodological account of model formulation. For them, hypothetico-deductive inquiry is little more than the injunction to engage in repeated strong testing of hypotheses about models, along with a commitment to the view that this should be done by exploiting deductive inference only. We acknowledge that Gelman and Shalizi speak of adding a neo-Popperian Lakatosian flavour to their thinking in related publications, but the ideas of Lakatos' methodology of scientific research programmes do no real work in their philosophy of Bayesian modelling.

In choosing a Popperian view of scientific confirmation, Gelman and Shalizi explicitly reject the confirmationist account of hypothetico-deductive method promoted by Carl Hempel. Actually, their reference to Hempel (1965) is to his early instance confirmation view of scientific confirmation, not to his later account of the hypothetico-deductive method. Hempel proposed the idea that, in scientific confirmation, hypotheses are confirmed by discovering their positive instances. In his formalization of this idea, Hempel required that the evidence entailed the development of the relevant hypothesis. This is quite different from hypothetico-deductive inference in which the evidence is deductively entailed by the hypothesis, and confirmation occurs through successful predictive testing.

More importantly, there are now available a number of sophisticated variants of the hypothetico-deductive method that Gelman and Shalizi might have made use of in formulating their own account of this method. Sprenger (2011a) provides a useful overview and defence of modern thinking about the hypothetico-deductive method. Interestingly, Sprenger (2011b) also recently proposed an account of confirmation that unifies Hempel's insight that hypotheses are confirmed by their instances and the core hypothetico-deductive idea that hypotheses are confirmed by their successful predictions. This modern hybrid account of confirmation has an important advantage over that outlined by Gelman and Shalizi: it allows for an objective notion of inductive support – something that we think Gelman & Shalizi's model testing strategy, in fact, requires. At the same time, it features strong hypothetico-deductive testing of a falsificationist kind, and it allows for the piecemeal testing of entire theories. Both of these are desirable features of scientific modelling for Gelman and Shalizi.

In addition to drawing from Popper, Gelman and Shalizi make brief heuristic use of some of Thomas Kuhn's (1970) ideas about science. They suggest that Kuhn's distinction between normal and revolutionary science is somewhat analogous to their distinction between learning within a Bayesian model and checking the model either to discard or expand it. However, they caution about pushing the analogy too far, correctly pointing out that most model checking and reformulation is puzzle-solving work, not revolutionary change, that takes place within a single paradigm. We think this disanalogy renders a serious appeal to Kuhn's theory of science as basically inappropriate for their particular philosophy, as indeed it is for the social sciences generally.

4. Modes of scientific inference

Gelman and Shalizi follow Popper in declaring that deductive inference is all there is to scientific inference. For them, this allows for the strong testing of Bayesian models by constantly checking them via their deductively derived predictions. However, unlike Popper, they acknowledge that science does trade in inductive inference of a material kind in which the premises and conclusion of inductive arguments contain reference to context-specific facts. Furthermore, they acknowledge that inductive statistical inferences to unobserved cases can be drawn on a background of deductive models. So it

would seem that, for them, science admits both deductive and inductive modes of reasoning.

We would go further and claim that in addition to deductive and inductive inference, science makes heavy use of abductive inference – a form of inference, moreover, that we think has a place in Gelman and Shalizi's view of model generation and model revision. In a word, abductive inference is explanatory inference, and it involves reasoning to, or from, hypotheses that explain relevant facts (e.g., Magnani, 2001). For example, the statistical method of exploratory factor analysis involves the abductive generation of latent factors to explain patterns in multivariate data. Further, inference to the best explanation (briefly mentioned by Gelman and Shalizi) is an abductive approach to theory appraisal in which explanatory reasoning forms the basis for evaluating rival theories.

Although Gelman and Shalizi describe their modelling philosophy as falsificationist in nature, it is more than this. For, when a model (or a component of a model) is confronted with negative evidence, the model (or its relevant parts) can be revised by means other than straight rejection or elimination. Often, this modification of a hypothesis will be seen to plausibly explain the anomalous data. We think that when scientists engage in such model revision, they engage in abductive reasoning, whether they know it or not. Thus, it would seem that Gelman and Shalizi's account requires an extension to abductive inference if it is to account for standard practices of model revision.

Two further comments on the inference forms involved in modelling are in order. Gelman and Shalizi regard the process of checking and ruling out possible misspecifications of a model as consistent with the strategy of eliminative induction. However, in this context, they think the word *induction* is a misnomer, and they enlist the support of Kitcher (1993) in maintaining that the strategy really embodies a deductive argument. However, Kitcher is concerned with the successive elimination of actual theories that rival the theory of interest, not with successive checks for possible inconsistencies in a single model. We think that both inductive and deductive eliminative strategies are used in science, and that because of the uncertainties in social science research, the aspect of model checking referred to here by Gelman and Shalizi is more realistically construed as an inductive strategy.

Our final comment on scientific inference is to point out that Gelman and Shalizi's bald characterization of classical Fisherian and Neyman–Pearsonian inference as deductive in nature is wrong, or at least simplistic. If anything, Fisher was an inductivist, and Neyman seemed to endorse both inductive and deductive forms of inference. However, a proper characterization of the forms of inference involved in these two statistical traditions is demanding and complex (Rivadulla, 1991).

5. Conclusion

Gelman and Shalizi have done the statistical community a great service by decoupling Bayesian data analysis and Bayesian philosophy of science. We fully agree with the conclusion that Bayesian data analysis can be justified in non-Bayesian ways. However, we think it is important to emphasize that the actual approach chosen by Gelman and Shalizi is but one of many. It should be recognized that one can consistently be a practical Bayesian without being a philosophical Bayesian, but it should also be recognized that one can maintain that consistency in a variety of ways, not just through hypothetico-deductivism. We ourselves have reservations about the suitability of Popper's theory-

centred falsificationist account of science for data modelling which, despite their comments to the contrary, we take to be Gelman and Shalizi's real focus.

In a sense, Gelman and Shalizi may trade one problem for another, not because they are substituting hypothetico-deductive philosophy for Bayesianism, but because they are attempting to characterize data analysis in such a grand theory in the first place. Perhaps the philosophy of data analysis cannot be tied uniquely to one of the major theories of scientific inference. The process of data analysis could be argued to feature elements that are reminiscent of many several theories in the philosophy of science – including inductive, deductive, and abductive accounts. However, it is also guided and constrained by strongly pragmatic concerns, ranging from the available money and time to the computational resources of computers, that are alien to Bayesianism, but also to the type of hypothetico-deductive theories that Gelman and Shalizi enlist. In fact it is improbable, in our view, that a one-to-one mapping between a philosophy of inference and a philosophy of data analysis could ever be achieved.

References

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

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. (3rd ed.) La Salle, IL: Open Court.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. New York: Kluwer/Plenum.

Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge*. (3rd ed.) London: Routledge & Kegan Paul.

Rivadulla, A. (1991). Mathematical statistics and metastatistical analysis. *Erkenntnis*, *34*, 211–236. doi:10.1007/BF00385721

Sprenger, J. (2011a). Hypothetico-deductive confirmation. *Philosophy Compass*, *6*, 497–508. doi:10.1111/j.1747-9991.2011.00409.x

Sprenger, J. (2011b). *A synthesis of Hempelian and hypothetico-deductive confirmation*. Manuscript submitted for review.

Received 19 January 2012

British Journal of Mathematical and Statistical Psychology (2013), 66, 45–56

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'

John K. Kruschke*

Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA

Bayesian inference is conditional on the space of models assumed by the analyst. The posterior distribution indicates only which of the available parameter values are less bad than the others, without indicating whether the best available parameter values really fit the data well. A posterior predictive check is important to assess whether the posterior predictions of the least bad parameters are discrepant from the actual data in systematic ways. Gelman and Shalizi (2013) assert that the posterior predictive check, whether done qualitatively or quantitatively, is non-Bayesian. I suggest that the qualitative posterior predictive check might be Bayesian, and the quantitative posterior predictive check should be Bayesian. In particular, I show that the 'Bayesian *p*-value', from which an analyst attempts to reject a model without recourse to an alternative model, is ambiguous and inconclusive. Instead, the posterior predictive check, whether qualitative or quantitative, should be consummated with Bayesian estimation of an expanded model. The conclusion agrees with Gelman and Shalizi regarding the importance of the posterior predictive check for breaking out of an initially assumed space of models. Philosophically, the conclusion allows the liberation to be completely Bayesian instead of relying on a non-Bayesian *deus ex machina*. Practically, the conclusion cautions against use of the Bayesian *p*-value in favour of direct model expansion and Bayesian evaluation.

1. Introduction

Bayesian inference is conditional on the space of models assumed by the analyst. Within that assumed space, the posterior distribution only tells us which parameter values are relatively less bad than the others. The posterior does not tell us whether the least bad parameter values are actually any good. Assessing the goodness of the least bad parameter values is the job of the *posterior predictive check*. In a posterior predictive check, the analyst assesses whether data simulated from credible parameter values resemble the

*Correspondence should be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington, IN 47405-7007, USA (e-mail: kruschke@indiana.edu).

DOI:10.1111/j.2044-8317.2012.02063.x

actual data, with 'resemblance' measured in any way that is meaningful in the applied context. If the resemblance is not good enough, then the analyst changes the model and does Bayesian inference on the modified model. This cycle repeats until the resemblance of the predicted data and the actual data is good enough for purposes of the application.

The posterior predictive check allows the analyst to solve the problem of being confined within the initially assumed space of models. Gelman and Shalizi (2012, 2013) emphasized that the posterior predictive check is a non-Bayesian process: 'It is by this non-Bayesian checking of Bayesian models that we solve our … problem' (Gelman & Shalizi, 2013, p. 17). In particular, the goodness of the resemblance, between simulated and actual data, is assayed in either of two non-Bayesian ways, qualitative or quantitative.

In the *qualitative* way of assessing resemblance between simulated and actual data, the analyst can visually examine graphical or tabular displays to look for structured patterns in the residuals between actual and simulated data. If there appears to be structure in the residuals that meaningfully informs the interpretation of the model, then the analyst can change the model so that it better captures the revealed trends. This intuitive assessment uses no explicit, formal Bayesian calculations.

Although intuitive assessment of pattern is not formally Bayesian, some leading theories in cognitive science assert that perception is well described as Bayesian inference. Essentially, these theories propose that the mind has a vast library of candidate perceptible patterns, with a distribution of prior credibilities across those patterns, and the observed residuals are used to infer, in a Bayesian manner, the posterior credibilities of candidate patterns for the residuals. Thus, when we perceive a pattern in the residuals, it is because that pattern has a reasonably high posterior credibility among the various patterns we have available in our perceptual space.

In the *quantitative* way of assessing resemblance between simulated and actual data, the analyst defines a formal measure of the magnitude of discrepancy between observed data y and predicted values $\hat{y}$, denoted $T(y,\hat{y})$. The observed data may be the actual data from the empirical research and denoted y^{act}, or the observed data may be simulated from the model and denoted y^{rep}, where the superscript rep refers to 'replication'. With many replications of data simulated from posterior parameter values, a sampling distribution of $T(y^{rep},\hat{y})$ is created. From that sampling distribution we compute the probability of obtaining a value of T as big as or bigger than the actual one: $p(T(y^{rep},\hat{y}) \geq T(y^{act},\hat{y}))$. This probability is also known as the 'Bayesian p-value'. If the Bayesian p-value is very small, then we reject the model and search for something better. Gelman and Shalizi (2012, 2013) point out that this procedure can reject a model without specifying an alternative model.

While this quantitative process is non-Bayesian, I show that its results are ambiguous and undertaking it is unnecessary. Instead, a Bayesian procedure can yield clearer results. Specifically, the analyst should create a formal model that addresses the perceived discrepancy, and the expanded model can be assessed in a Bayesian fashion. This approach avoids ambiguity in T, which could be a signature of many different underlying structures. The expanded model is assessed by Bayesian parameter estimation, and does not necessarily rely on Bayesian model comparison, which has problems of hypersensitivity to priors (as pointed out by Gelman & Shalizi, 2013).

The rest of this article expands on the two-pronged argument outlined above. Examples of regression analysis are provided to illustrate ambiguous implications from T and p, but clearer conclusions from Bayesian estimation of specific expanded models. The argument agrees that a posterior predictive check is an important step in Bayesian data analysis, but avers that a posterior predictive check need not be inherently non-Bayesian. Whether the check of resemblance is qualitative or quantitative, it should be consummated by a formal

specification of structure in an expanded or new model, with parameters estimated in a Bayesian fashion. Thus, a posterior predictive check can and should be Bayesian.

2. A qualitative posterior predictive check can be Bayesian

In a qualitative posterior predictive check, the analyst displays the original data along with the posterior predictions in some way that highlights potentially systematic discrepancies. The display could be tabular or graphical, and can accentuate or attenuate different aspects of the data and posterior predictions. Regardless of the exact nature of the display, the human analyst must perceive systematic patterns in the discrepancies. What are the possible patterns that a human can perceive? And of all those possible patterns, which ones are most likely to be perceived when observing a display?

Some leading theories in cognitive science describe perception as Bayesian inference (e.g., Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Shams & Beierholm, 2010; Yuille & Kersten, 2006). According to this theoretical perspective, the mind has a vast repertoire of possible descriptions of the world, with innate or previously learned knowledge providing a prior distribution over that space of perceptible patterns. When new stimuli impinge upon the senses, the mind infers the most likely distal objects that may have produced the sensory stimulus. The inference relies heavily on prior knowledge, and formal Bayesian models have successfully accounted for many aspects of human perception.

One of the simplest examples of prior knowledge deployed in perception is the interpretation of three-dimensional shape from observable shading on the object. Consider Figure 1, which shows two circular regions spanned by gradients of grey. When the light end of the gradient is at the top, we perceive the circular region as a protuberance, but when the light end of the gradient is at the bottom, we perceive the circular region as an indentation. This difference in perceptual interpretation is explained by the mind applying prior knowledge: illumination usually comes from above, as from the sun and sky. As another example, consider the learning of functional relationships between input and output values, such as drug dosage (input value) and symptom severity (output value). People are tasked with learning the relationship between the variables by observing many examples, and then their learned responses are used to teach new learners. After a few generations, the learned and retaught function evolves into a linear relationship, regardless

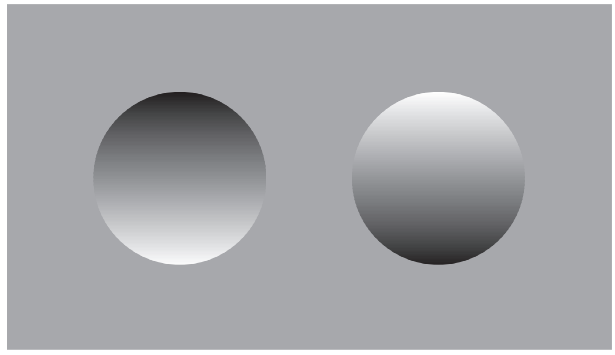

Figure 1. The circular region on the left is perceived as an indentation, while the circular region on the right is perceived as a protuberance, even though the gradients of grey are identical except for orientation. Perception apparently employs prior knowledge that illumination comes from above, and that the surface itself has constant colour.

of how it started, which reveals that linear relations are weighed heavily in learner's prior knowledge (Kalish, Griffiths, & Lewandowsky, 2007). While recent theories have given explicit formal expression to the idea of perception as Bayesian inference, informal theories of perception as inference go back at least to Helmholtz (1867), although it is doubtful that Helmholtz had any explicitly Bayesian notions (Westheimer, 2008).

A variety of other aspects of cognition and learning have been modelled as Bayesian inference (for overviews, see Chater, Tenenbaum, & Yuille, 2006; Jacobs & Kruschke, 2010). Recent work has shown that human perception of accidental coincidences versus causes can be modelled as Bayesian inference (Griffiths & Tenenbaum, 2007), and human interpretation of many different data structures can be modelled as Bayesian inference (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The issue of how the mind or brain might implement Bayesian inference is one of current discussion and debate. Some theorists suggest that the mind merely approximates Bayesian inference (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), but this particular approach may be unsatisfying because many non-Bayesian algorithms are 'approximately' Bayesian without having any necessary relation to Bayesian computation (Kruschke, 2010a). Another approach suggests that the mind might be well described as Bayesian only within certain levels of analysis, while larger-scale behaviour is not (Kruschke, 2006). Whatever the domain or level of analysis, the goal for genuinely Bayesian models of cognition is discovering functional forms and priors that closely mimic human behaviour.

Regardless of the ultimate veracity of any specific Bayesian model of perception or cognition, the point of this section is that intuitive assessment of patterned discrepancies could be Bayesian. There is nothing necessarily non-Bayesian in a qualitative posterior predictive check. On the other hand, I am not claiming that qualitative posterior predictive checking is in fact, or must be, well described as Bayesian inference. Indeed, even if human perception and cognition – the engine of qualitative posterior predictive checking – ultimately proves to be impossible to adequately model as Bayesian inference, it is still appropriate to formally analyse scientific data with Bayesian methods (Kruschke, 2010b), and it is still the case that *quantitative* posterior predictive checking can and should be Bayesian, as the next section illustrates.

3. A quantitative posterior predictive check should be Bayesian

As a concrete example to frame discussion, consider the data displayed in Figure 2. For every individual, we measure a criterion value y that we wish to predict from a value x. The conventional first approach would be simple linear regression with normally distributed noise, expressed formally as

$$\hat{y} = \beta_0 + \beta_1 x, \tag{1}$$

$$y \sim N(\hat{y}, \sigma), \tag{2}$$

where $\hat{y}$ is the predicted value of y for predictor value x, β_0 is the intercept, β_1 is the slope, and σ is the standard deviation of the normal distribution.

For Bayesian estimation of the three parameters in equations (1) and (2), I began with vague priors that had minimal influence on the posterior distribution. The analysis used Markov chain Monte Carlo (MCMC) sampling by JAGS (Plummer, 2003) called from R (R Development Core Team, 2011) via package rjags, with programs created in the style of

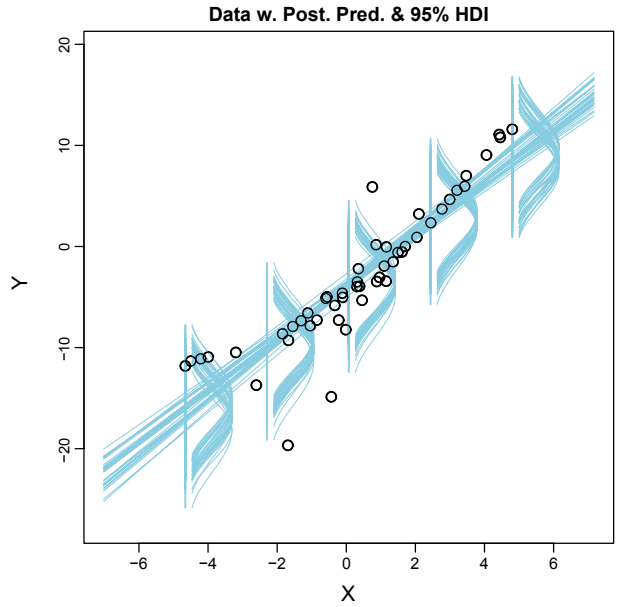

Figure 2. Data with posterior predictions, using linear regression with normally distributed likelihood as defined in equations (1) and (2). The lines extending from left to right show a smattering of credible regression lines from the MCMC chain. The vertical segments show 95% highest density intervals (HDIs) with normal density functions (plotted sideways) having corresponding credible standard deviations. The data appear to be too tightly clustered within vertical slices, relative to the spread of the posterior predicted normal distributions. The data also appear to have a slight non-linear trend.

Kruschke (2011b). Figure 2 shows plots of 30 credible regression lines superimposed on the data. Displayed with each line are sideways plots of a normal distribution with the corresponding standard deviation. The slope, intercept, and standard deviation of the 30 plots came from every (200,000/30)th step in the MCMC chain of 200,000 steps.

Visual inspection of the posterior estimates in Figure 2 suggests at least two discrepancies between data and model. First, the data appear to be too tightly clustered within vertical slices, relative to the spread of the posterior predicted normal distributions. Second, the data also appear to have a slight upward curvature relative to the linear predictions of the model.

Having noticed possible systematic discrepancies between the data and the posterior predictions, what should we do next? One possibility is to create some measure of discrepancy, $T(y, \hat{y})$, that somehow captures the seemingly anomalous discrepancy. The measure T does not need to express an alternative model; it merely needs to quantify the discrepancy. We then generate the sampling distribution of T from the posterior distribution, and assess whether $p(T(y^{rep}, \hat{y}) \geq T(y^{act}, \hat{y}))$ is sufficiently small that we are justified to look for a better model of the data. Gelman and Shalizi (2013, footnote 11) say 'the tail-area probabilities are relevant [because] they make it possible to reject a Bayesian model without recourse to a specific alternative' and 'What we are advocating, then, is what Cox and Hinkley (1974) call "pure significance testing", in which certain of the model's implications are compared directly to the data, rather than entering into a contest with some alternative model' (Gelman & Shalizi, 2013, p. 20).

For example, suppose we want to define a measure of upward curvature for the data in Figure 2. For purposes of defining the measure of discrepancy, we will index the 50 observations from smallest x-value to largest x-value. Thus, $\langle x_1, y_1 \rangle$ is the leftmost point, and $\langle x_{50}, y_{50} \rangle$ is the rightmost point. Upward curvature implies that the left-hand end and right-hand end points tend to be above the linear prediction, while middle points, namely $\langle x_{25}, y_{25} \rangle$ and $\langle x_{26}, y_{26} \rangle$, tend to be below the linear prediction. This signature of curvature could be formalized as, say,

$$T(y, \hat{y}) = (y_1 - \hat{y}_1) + (y_{50} - \hat{y}_{50}) - (y_{25} - \hat{y}_{25}) - (y_{26} - \hat{y}_{26}). \quad (3)$$

Defining T in terms of ranked data has precedents in Gelman, Carlin, Stern, and Rubin (2004). For example, when modelling a set of data with a normal distribution and assessing leftward skew or outliers, one definition for T was simply $T(y, \hat{y}) = y_1 = \min(y)$ (Gelman *et al.*, 2004, p. 160). For the same set of data, another definition for T was $T(y, \hat{y}) = |y_{61} - \hat{y}| - |y_6 - \hat{y}|$ (Gelman *et al.*, 2004, p. 164). Thus, the form of definition of T in equation (3) is consistent with standard practice.

The value of T in equation (3) for the actual data in Figure 2 is greater than zero. In fact, across all the credible parameter values in the 200,000-step MCMC chain, the average value of $T(y^{act}, \hat{y})$ is 7.52, as shown in the left panel of Figure 3. This distribution and the others in Figure 3 were created by generating a complete set of random data from the model at every step in the 200,000-step MCMC chain, and computing $T(y^{rep}, \hat{y})$, $T(y^{act}, \hat{y})$, and $T(y^{rep}, \hat{y}) - (y^{act}, \hat{y})$ at every step. The fact that $T(y^{act}, \hat{y})$ is robustly greater than zero indicates that it is a plausible signature of upward curvature. The expected value of $T(y^{rep}, \hat{y})$, however, must be zero, because randomly generated values will be above or below the prediction line equally often. This expected value is verified in the middle panel of Figure 3. The right panel of Figure 3 shows the sampling distribution of $T(y^{rep}, \hat{y}) - (y^{act}, \hat{y})$, where it can be seen that the Bayesian p-value is .116, which is not very small. In other words, from the definition of curvature in equation (3), we would *not* reject the linear model.

What are we to conclude about curvature in the data, in light of this failure to reject the linear model using a Bayesian p-value? Not much (in my opinion), because the visual impression of discrepancy is very strong, and we can always try some other definition of T.

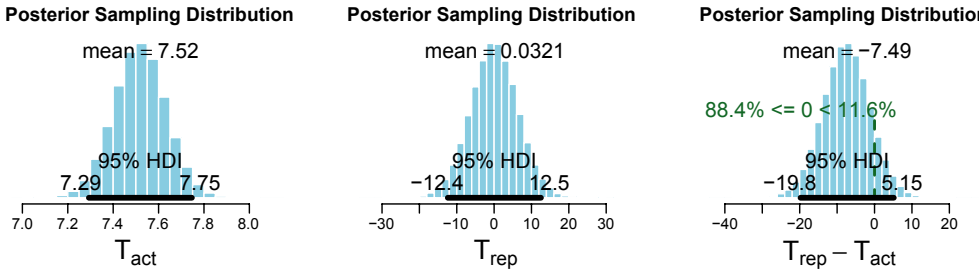

Figure 3. Posterior sampling distributions of $T(y^{act}, \hat{y})$, $T(y^{rep}, \hat{y})$, and $T(y^{rep}, \hat{y}) - T(y^{act}, \hat{y})$ for T defined in equation (3), from the posterior and data of Figure 2. 'HDI' denotes highest density interval. In the right panel, the text '88.4% $\leq$ 0 < 11.6%' means that 88.4% of the distribution falls below zero, and 11.6% of the distribution falls above zero. (Theoretically, $T(y^{rep}, \hat{y})$ is symmetric with a mean of 0.0. The histogram in the middle panel deviates slightly from the theoretical characteristics because of random sampling noise.)

Analysts who harbour a desire to reject the model can keep trying until they find a definition of T for which p is small, while analysts who harbour a desire not to reject the model can stop when they find a definition of T for which p is not very small.

Importantly, the goal of posterior predictive checking is not merely to reject the model, because, as Gelman and Shalizi (2012, 2013) and Gelman *et al.* (2004) have emphasized, we know in advance that the descriptive model is almost surely wrong for real data. The goal of posterior predictive checking is to come up with a more satisfying descriptive model of the data. Therefore we can simply side-step the process of arbitrarily defining T, generating its sampling distribution and struggling with its ambiguous implications. Instead, we should expand the descriptive model with explicit structural terms that capture the trends in which we are interested.

The apparent discrepancy in Figure 2 can be directly expressed in an expanded model that allows for non-linear trend and outliers. For example, we can directly express a quadratic trend and a heavy-tailed distribution as

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{4}$$

$$y \sim t(\hat{y}, \sigma, \nu), \tag{5}$$

where β_2 is the coefficient of quadratic trend and $\nu \geq 1$ is the degrees of freedom parameter for the t distribution. The t distribution is often used as a convenient descriptive distribution for data with outliers (e.g., Damgaard, 2007; Jones & Faddy, 2003; Lange, Little, & Taylor, 1989; Meyer & Yu, 2000; Tsionas, 2002). When ν is large (e.g., 100), the t distribution is very nearly normal. When ν gets close to 1, the t distribution is strongly kurtotic.

Figure 4 shows the results from Bayesian estimation of the five parameters in equations (4) and (5). As before, the prior distributions were minimally informed, and the analysis used MCMC sampling by JAGS (Plummer, 2003) called from R (R Development Core Team, 2011) with programs in the style of Kruschke (2011b). Visual inspection of the posterior estimates suggests that the model describes the data well: the data tend to be tightly clustered near the quadratic curve, with only a few outliers accommodated by the heavy-tailed distribution. (In fact, the data were randomly generated from exactly such a model, and the Bayesian estimates recovered the generating values well. But we never know the true generating model for real data.)

How do we know that the expanded model is better than the original model? In principle, we could do Bayesian model comparison. But in practice, Bayesian model comparison can be hypersensitive to the choice of prior distributions in the models, as Gelman and Shalizi (2013) remind us. Therefore Bayesian model comparison is to be avoided unless we have well-informed priors that put the two models on equal footing (e.g., Kruschke, 2011a; Liu & Aitkin, 2008; Vanpaemel, 2010), which we do not have in this case. Instead, because the models are nested in this case, we can simply see whether the posterior estimates of the additional parameters are credibly non-zero. Figure 5 displays the marginals of the posterior distribution, where it can be seen that the quadratic coefficient β_2 is robustly non-zero. Thus, despite the fact that the Bayesian p-value did *not* reject the linear model (recall Figure 3), an expanded model with an explicit quadratic trend strongly *does* implicate non-linearity in the data.

As another illustration of the peril of allowing arbitrary definitions of T without a specific alternative model, suppose we look at Figure 4 with the aim of finding even more discrepancy and rejecting the expanded model (perception of pattern is linked to motivation; e.g.,

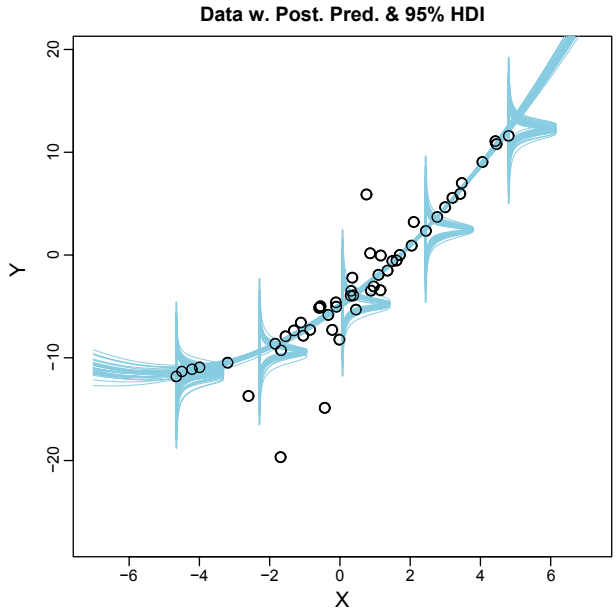

Figure 4. Data with posterior predictions from quadratic regression with *t*-distributed likelihood as defined in equations (4) and (5). The curves extending from left to right show a smattering of credible regression lines from the MCMC chain. The vertical lines show 95% highest density intervals (HDIs) and *t* density functions with corresponding standard deviations and degrees of freedom. The data appear to be well described by the posterior prediction (which is fortunate, because this form of model actually generated the data).

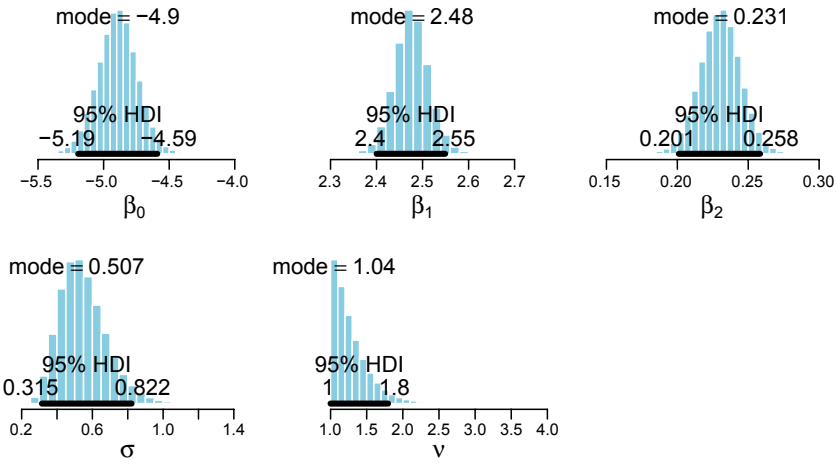

Figure 5. Marginals of the posterior distribution for the five parameters of equations (4) and (5), for the data displayed in Figure 4. The histograms summarize an MCMC chain of 200,000 steps. The top right-hand histogram shows that the quadratic coefficient β_2 is credibly greater than zero. (The displayed modes are approximated by a kernel density smoother.)

Whitson & Galinsky, 2008). It appears that there is counterclockwise 'torsion' in the residuals, such that there are more outliers in the lower-left and upper-right quadrants near the median of x. Because I am not sure what I mean by this, in terms of an actual structural trend expressed in a functional form, I will define a signature of the torsion as

$$T(y,\hat{y}) = -\sum_{i=6}^{17}(y_i - \hat{y}_i) + \sum_{i=28}^{29}(y_i - \hat{y}_i). \tag{6}$$

The expression in equation (6) merely sums the residuals in a particular range below the median of x and subtracts the result from the sum of residuals in a particular range above the median of x. A posterior predictive check produces the posterior sampling distributions shown in Figure 6. The Bayesian p-value is small, just .032. According to conventional p-value criteria, this result should lead us to reject the model, without recourse to a specific alternative.

But this conclusion seems unwarranted. In this case, we know that the data were actually generated by the model that has been rejected, but this conflict is not the reason for being sceptical, because for real data we do not know the true generator of the data. The scepticism arises because the definition of T was cherry-picked from a universe of all possible definitions of T without any motivation other than trying to prove the model wrong.

If I were forced to define a functional form for the structural trend of 'torsion in outliers', I might attempt to use a likelihood distribution that has a skew parameter, with the skew parameter functionally linked to the value of x, so that the skew is negative when x is just below its median, but positive when x is just above its median. This expanded form involves new parameters for skew and for the functional relation between skew and x, and we would also have to specify a prior on the parameters of the expanded model. A prior that would be agreeable to a sceptical audience might favour null values on the expanded parameters because the model is so unusual. Even without a sceptical prior on the extra parameters, there is increased uncertainty in the higher-dimensional parameter space, hence it is less likely that the estimates of the extra parameters would be credibly non-zero. Even though an expanded model might be deemed arbitrary like T, Bayesian evaluation of the expanded model incorporates penalties for arbitrariness, unlike T. There

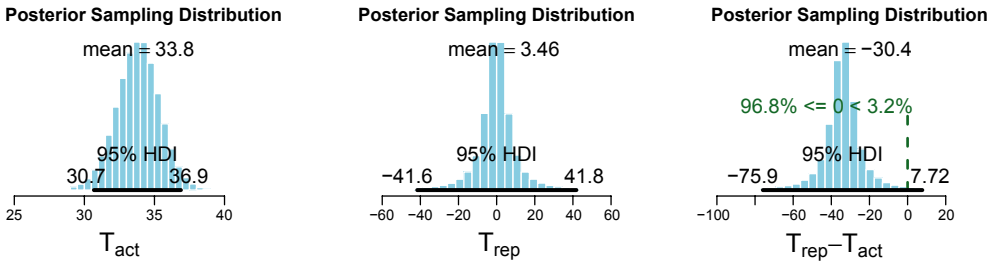

Figure 6. Posterior sampling distributions of $T(y^{act},\hat{y})$, $T(y^{rep},\hat{y})$, and $T(y^{rep},\hat{y}) - T(y^{act},\hat{y})$ for T defined in equation (6), from the posterior and data of Figure 4. 'HDI' denotes highest density interval. In the right panel, only 3.2% of the distribution falls above zero. (Theoretically, $T(y^{rep},\hat{y})$ is symmetric with a mean of 0.0. The histogram in the middle panel deviates slightly from the theoretical characteristics because of random sampling noise in the extreme tails of the distribution.)

is a penalty from a sceptical prior and from increased uncertainty in a higher-dimensional parameter space. Moreover, if Bayesian model comparison is undertaken with appropriate caution, the diluted prior on the higher-dimensional parameter space automatically penalizes the more complex model to fend off overfitting (often referred to as the Bayesian Occam's razor effect, e.g., MacKay, 2003).

I have presented two examples in which the conclusion from a Bayesian p-value conflicted with the conclusion from a Bayesian estimation of an expanded model. In general, the conclusions from Bayesian estimation of an expanded model supersede the conclusions of a corresponding Bayesian p-value. If the conclusions agree, the expanded model and explicit posterior distribution provide rich structural definition that is more specific than the ambiguous signature expressed by T. If the conclusions disagree, then again we look to the explicit structural form of the expanded model, and its estimated parameters, to better understand the data. If a Bayesian p-value is small and rejects a model, it merely confirms a foregone conclusion, and we still need an explicit structural form to understand why. If a Bayesian p-value is large and does not reject a model, it might be merely because the definition of T does not capture the structural form of the discrepancy which would be apparent when estimated in an explicit expanded model.

4. Summary and conclusion

In typical research, the models we use to describe data are selected because of their familiarity from previous training, tractability in computation, and prior probability of describing trends we care about in the specific application. But we know in advance that the models are merely descriptive, and that the data were almost surely *not* generated by such a model. Gelman and Shalizi (2013, p. 20) say 'The goal of model checking, then, is not to demonstrate the foregone conclusion of falsity as such, but rather to learn how, in particular, this model fails'. My argument above is completely consistent with this perspective. The argument, bolstered with examples, said merely that the *ad hoc* construction of a measure T such that $p(T^{rep} \geq T^{act})$ is an exercise in a foregone conclusion. Moreover, the implications are ambiguous because the measure T does not entail a specific structural form for an expanded model. Instead of going through the foregone conclusion and ambiguous implication of Bayesian p values, we should instead define an expanded model and evaluate it with Bayesian estimation.

I have also suggested that a qualitative posterior predictive check may be Bayesian, insofar as perception and cognition themselves may be Bayesian. There is no inherent necessity for model checking to be non-Bayesian. Formal Bayesian calculations are conditional on a particular model space, but there are a variety of ways to provoke the analyst to consider other model spaces. The provocation can come from a posterior predictive check, or the provocation can come from learning about other types of models in other applications and wondering whether there is an analogous application, or the provocation can come from simply wanting to prove a competing theorist wrong. But whatever the provocation, the space of possible alternatives is still governed by the mental prior in the analyst's mind.

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, *10*(7), 287–344.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Damgaard, L. H. (2007). Technical note: How to use WinBUGS to draw inferences in animal models. *Journal of Animal Science*, *85*, 1363–1368.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics in the social sciences. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science*. Oxford: Oxford University Press.

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

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Helmholtz, H. L. (1867). *Handbuch der physiologischen Optik*. Leipzig: L. Voss.

Jacobs, R. A., & Kruschke, J. K. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 8–21.

Jones, M. C., & Faddy, M. J. (2003). A skew extension of the *t*-distribution, with applications. *Journal of the Royal Statistical Society, Series B*, *65*(1), 159–174.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(4), 677–699.

Kruschke, J. K. (2010a). Bridging levels of analysis: comment on McClelland et al. and Griffiths et al. *Trends in Cognitive Sciences*, *14*(8), 344–345.

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, *84*(408), 881–896.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

MacKay, D. J. C. (2003). *Information theory, inference & learning algorithms*. Cambridge: Cambridge University Press.

Meyer, R., & Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal*, *3*(2), 198–215.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (dsc 2003)*, Vienna.)

R Development Core Team (2011). *R: A language and environment for statistical computing* [computer software manual]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*, 425–432.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tsionas, E. G. (2002). Bayesian inference in the noncentral Student-*t* model. *Journal of Computational and Graphical Statistics*, *11*(1), 208–221.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Westheimer, G. (2008). Was Helmholtz a Bayesian? a review. *Perception*, *37*, 642–650.

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science*, *322*, 115–117.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Received 30 January 2012

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

British Journal of Mathematical and Statistical Psychology (2013), 66, 65–67

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

Comment on Gelman and Shalizi

Stephen Senn*

Competence Center for Methodology and Statistics, CRP Santé, Strassen, Luxembourg

As an applied statistician, one of the complaints I make about my more theoretically minded colleagues is that many of my problems that are solved by them in principle are not solved in practice. This complaint is sometimes addressed to Bayesians (Senn, 2011), but frequentists are not exempt (Senn, 1998). (In fact, my subjective impression is that they are frequently worse.) One of the challenges I make to anybody telling me how to do better is 'don't tell, do'. Unfortunately, I cannot do better than Gelman and Shalizi (2013, henceforth GS) with the examples they provide. Thus, if I am to avoid being hypocritical I have to concede that the solution is beyond (my) criticism.

What I am reduced to doing is raising the hackneyed quibble, 'it may work in practice but does it work in theory?'. A common claim for Bayesian inferences is that it is a theory of everything. Clearly, however, much of what GS are doing is not covered by the standard theory of coherent Bayesian updating of prior to posterior probability statements using data. Model checking (at least) has to be added to the mix. Of course, one has to be careful here; to be Bayesian means many different things to different people, and Jack Good famously determined that Bayesians came in 46,656 varieties (Good, 1983, pp. 20–21). Perhaps there is no standard theory of what it is to be Bayesian.

If, however, model checking is an essential part of the (or a) Bayesian mix, it raises the question as to what the status is of the 'final' analysis that GS produce. (For simplicity, I will only consider the analysis of the 2008 voting data but the argument carries over to the more complicated cases.) Consider another Bayesian, one who so firmly believed in the model that GS eventually chose that he or she had no doubts whatsoever as to its veracity. Suppose that his or her prior distribution under the model had been the same as that of GS. The posterior 'statement' is now the same: can they both be valid? Well, perhaps in this case, there would be very little to choose between them in terms of validity. This is because in the end GS accepted a rather richer model. The varying intercepts model is a special case of the varying slopes model: in Bayesian terms, one might say that it corresponds to taking the random slopes model but having a completely informative prior that the variance of the slopes is zero. By moving from the intercepts model to the slopes

*Correspondence should be addressed to Stephen Senn, Competence Center for Methodology and Statistics, CRP Santé, Strassen, Luxembourg (e-mail: stephen.senn@crp-sante.lu).

DOI:10.1111/j.2044-8317.2012.02065.x

model GS have clearly added some uncertainty (the reverse of what one expects examining data to do according to the Bayesian account!). So perhaps everything is (as it turns out) approximately all right. They were in danger of using a prior distribution that was too informative. Model checking has saved them from the error and because they have not fallen into the trap their inferences are now reasonable.

However, things might have turned out differently. GS must believe this is possible, otherwise it is hard to see why they started where they did. Suppose that instead the model checking had revealed no or little problem. In that case they might have been tempted to use the random intercepts model. However, although the data might be compatible with the simpler (intercept) model they would be compatible with some value of the richer (slopes) model. Proceeding to use the intercept model as if one always knew it were true must underestimate the uncertainties.

Then I worry about the predictive checks they are undertaking – analogous to what Good (1983) calls the 'device of imaginary results'. What exactly is going on here? They seem to assume that we agree that there is a strong distinction between prior distribution and data. Is it not supposed to be a strength of the Bayesian approach that prior and data are exchangeable? Consider that notorious example of frequentist irrationality, analysis of sequential trials. Suppose we have a trial with 200 patients and decide to look after 100 patients. Denote by D_1 the data of the first 100, and by D_2 the data of the second 100. Let P_0 be the posterior distribution after updating based on D_1 alone and P_2 the posterior distribution after seeing D_1 and D_2. Then the following schematic algebra of Bayesian inference applies:

$$P_0 + D_1 \rightarrow P_1, \quad P_1 + D_2 \rightarrow P_2, \tag{1}$$

or

$$P_0 + D_1 + D_2 \rightarrow P_2. \tag{2}$$

In (1) the inference is performed in two steps and P_1 carries out the dual role of being the posterior distribution after seeing D_1 and the prior distribution before seeing D_2. In (2) we see D_1 and D_2 together and reach P_2 without need of P_1. What I worry about is how this pans out when model checking is added to the mix. GS take the point of view that the way in which we judge the adequacy of a prior distribution is by constructing predictive distributions for data sets of the same size as the data we have added to the prior. So this seems to imply that in case (1) we compare D_1 to P_1 and compare D_2 to P_2 (if the model survives this far!), and in case (2) we compare $D_1 + D_2$ to P_2. To take the argument further, suppose that we have as many steps as there are data points, in the spirit of Philip Dawid's prequential inference (Dawid, 1997); does GS model checking get us where we want to be?

Perhaps it does. Perhaps coherent model checking can be added to coherent updating. Perhaps, however, inference is a much messier business than the builders of grand systems suppose. Of course, GS might argue that I am being unfair here, that I am mixing up two kinds of prior probability: the P_0 sort, which is not based on local data but on vague notions, and the P_1 sort, which is based (partially) on data. However, I suspect that some members of many of the 46,656 tribes would find this a very slippery slope to tread.

So, to sum up, I am convinced that what GS achieve is excellent applied statistics. I am not convinced by their explanation as to why it works. It works in practice, but does it work in theory?

References

Dawid, A. P. (1997). Prequential analysis. In S. Kotz, C. B. Read & D. L. Banks (Eds.), *Encyclopedia of statistical sciences update* (pp. 464–470). New York: Wiley.

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

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Senn, S. J. (1998). Mathematics: Governess or handmaiden? *The Statistician*, *47*, 251–259. doi:10.1111/1467-9884.00130

Senn, S. J. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, *2*, 48–66. doi:10.1038/clpt.2010.128

Received 30 April 2012

British Journal of Mathematical and Statistical Psychology (2013), 66, 68–75

The British Psychological Society

www.wileyonlinelibrary.com

Commentary

The humble Bayesian: Model checking from a fully Bayesian perspective

Richard D. Morey[1]*, Jan-Willem Romeijn[1] and Jeffrey N. Rouder[2]

[1]University of Groningen, The Netherlands
[2]University of Missouri, USA

Gelman and Shalizi (2013) criticize what they call the 'usual story' in Bayesian statistics: that the distribution over hypotheses or models is the sole means of statistical inference, thus excluding model checking and revision, and that inference is inductivist rather than deductivist. They present an alternative hypothetico-deductive approach to remedy both shortcomings. We agree with Gelman and Shalizi's criticism of the usual story, but disagree on whether Bayesian confirmation theory should be abandoned. We advocate a humble Bayesian approach, in which Bayesian confirmation theory is the central inferential method. A humble Bayesian checks her models and critically assesses whether the Bayesian statistical inferences can reasonably be called upon to support real-world inferences.

1. Comparison with Gelman and Shalizi

Modern statistics is a diverse field with disagreements about even basic foundational issues. Savage (1972) noted 60 years ago that there were scarcely any accepted facts about the foundations of statistics, and Gelman and Shalizi's (2013) article (henceforth GS) is proof that disagreements exist even today. But GS also reveal that in spite of these disagreements, or perhaps rather because of them, statisticians continue to develop useful ways of learning from data.

We agree with their critique of what they call the 'usual story' in Bayesian statistics, and also acknowledge the usefulness of the procedures they advocate. But rather than abandon the traditional Bayesian framework, we promote a perspective on Bayesian statistics that is strengthened through the use of model-checking procedures.

1.1. Overconfidence is wrong, but Bayes is right

GS introduce the usual story of Bayesian data analysis: that all information necessary for inference is contained in Bayesian quantities such as posterior distributions or model

*Correspondence should be addressed to Richard D. Morey, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands (e-mail: r.d.morey@rug.nl).

DOI:10.1111/j.2044-8317.2012.02067.x

posteriors. In this story, model checking is not performed at all; posterior quantities tell us the relative plausibilities of parameter values or models. A Bayesian of this stripe is what we refer to as an 'overconfident' Bayesian. To our mind, the overconfident Bayesian is an extreme point in the spectrum of Bayesians. We ourselves routinely perform model checks in our own work (Morey, Rouder, & Speckman, 2008, 2009; Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008b) and we believe that most practising Bayesian statisticians worry about the appropriateness of their models and hence engage in model checking.

One reason for the impression that overconfident Bayes features prominently in the philosophy of statistics may be that, in philosophy, Bayesian inference is often considered as part of a logic (de Finetti, 1995; Howson, 2001; Romeijn, 2011). Philosophers of statistics focus on the correctness of the inferential step rather than on the truth or falsity of the premises. In other words, the focus is on the Bayesian data analysis and not on the appropriateness of the model. However, as indicated by the parallel interest in model selection among philosophers of statistics (Forster & Sober, 1994; Kieseppä, 2001; Romeijn & van de Schoot, 2008; Romeijn, van de Schoot, & Hoijtink, 2012), the focus on correctness should not be taken to indicate that, according to philosophers of statistics, valid inference is all there is to good statistical practice.

In contrast to overconfident Bayesianism, the scheme that GS propose for model checking is not Bayesian. Their view is what they call hypothetico-deductive or, in other places, falsificationist: models are judged by how well they accommodate the data and then retained or discarded. The core of the view seems to be that model checking is not regulated by an inductive, but rather by a deductive mode of inference. Statistical models entail probabilistic empirical consequences, and they do so deductively, as a matter of mathematical fact. These probabilistic consequences can then be compared to data to arrive at a judgement on the model.

We accept that model checking is an integral part of good statistical practice. The overconfident Bayesian is wrong. But we believe that if model checking is to become a primary method of statistical inference, more detail is needed on how it is supposed to be done. In other words, we require a theory of inference using model checking. Although GS offer a number of tools and procedures and a general philosophy, they do not offer a theory of inference. But a suitable theory of inference already exists: the Bayesian confirmatory framework. A reasonable Bayesian can use model checking alongside traditional Bayesian analyses, casting the model checking itself in a Bayesian light.

1.2. Conceptual issues for Gelman and Shalizi

GS briefly discuss arguments for the Bayesian consistency and rationality but they do not seem pursuaded. To our mind, it is a major advantage of a Bayesian approach to model checking that it inherits the conceptual clarity and coherence of Bayesian theory generally. We provide some detail on Bayesian model checking below. Here we note two conceptual issues for GS.

As GS indicate, model checking typically proceeds by finding out that the model under scrutiny is false, as its empirical consequences do not match the data. Strictly speaking, statistical models cannot of course be falsified, since probabilistic consequences cannot be contradicted by data. Much like Mayo (1996) and Mayo and Spanos (2011), it seems that GS speak of falsificationism and deductivism by proxy: highly improbable data are somehow considered close enough to impossible data to effect a form of falsification. For philosophers and statisticians who champion the validity of the inferences this attitude is somewhat puzzling, especially since a valid inferential framework is already available in

Bayesian theory. Now it may be that GS are simply not bothered by these concerns, but we think they should be, and that the broad strokes in which GS's deductivism is painted need to be revisited with a finer brush.

Furthermore, GS's abandonment of the Bayesian framework has consequences for their proposed method of model checking. They imply that they have abandoned the Bayesian framework even to the extent of rejecting a probabilistic interpretation of the Bayesian prior, which to them is 'more like a regularization device, akin to the penalization terms added to the sum of squared errors when doing ridge regression and the lasso … or spline smoothing'. This rejection, however, has consequences. The probabilistic interpretation of the posterior arises from the probabilistic interpretation of the prior. Abandoning the probabilistic interpretation of the prior threatens the interpretation of the corresponding posterior, and thus the interpretation of the posterior predictive *p*-values.[1] Since posterior predictive *p*-values are one of the primary methods GS have advocated for model checking, it is important that these *p*-values be interpretable.

Summing up, we largely agree with GS in their dislike of overconfident Bayes and on the importance of model checking. But we feel that GS need to provide a theory. Their approach compares unfavourably to the coherence and conceptual clarity of Bayesianism.

2. Departing from Gelman and Shalizi

Apart from being Bayesian, our perspective differs from that of GS in two ways: one pragmatic, and the other philosophical. Pragmatically, we believe that although their approach is likely to be successful for the types of problems they encounter, it is not ideal for questions we commonly encounter. Philosophically, we disagree with GS that all models are wrong.

2.1. The importance of invariances

Statistics is such a diverse field in part because of the wide variety of questions that statistics is required to address. Differences in goals and applications lead to immediate differences in statistical philosophy. GS (p. 11) state: 'The statistician begins with a model that stochastically generates all the data y, whose joint distribution is specified as a function of a vector of parameters θ from a space Θ (which may, in the case of some so-called non-parametric models, be infinite-dimensional)'. In contrast, we start from a theoretical question about a scientific phenomenon of interest. We are almost always interested in assessing invariances: those elements of structure or constancy in a complex relationship among variables. A classic example of invariances are Kepler's laws of planetary motion. Although the trajectories of the planets seem complicated when viewed from Earth, Kepler was able to deduce a set of three simple constraints that governed the relationship among the observables. The search for simplifying structure is ubiquitous in the the natural sciences and in many experimental social sciences (Morey & Rouder, 2011; Rouder, Lu, Morey, Sun, & Speckman, 2008a; Rouder & Morey, 2011). Theoretical differences over models amount to different constraints on data.

[1] One could argue that the use of improper priors also threatens the interpretation of Bayesian quantities as well, since they do not have a ready probability interpretation. However, many improper priors are limits of proper priors. The interpretation of the prior in this case is quite different from a non-probabilistic 'regularization device'. The debate over the use of improper priors is ongoing and interesting (Berger, 2006; Goldstein, 2006) but we do not wish to engage in it here.

For problems like the ones GS handle, which often involve linear (or generalized linear) models of varying complexity, the model-checking approach is feasible. However, for many of the questions we encounter, it is difficult to imagine how model checking could serve as the primary mode of inference. If two theoretical positions are represented by radically different stochastic models (as will often be the case in psychology), both models will likely misfit in different ways, and it may not be obvious how to compare the two. We seek methods for moving either towards less complex models embedding more invariances, or across classes of models embedding different invariances. The traditional Bayesian framework offers a natural way of answering the questions we face, with the benefit that it comes with a formal inferential framework.

2.2. Models are neither true nor false

GS focus on the issues of 'false' models in statistical inference, but we believe the idea of 'false' models to be unhelpful. As representations, scientific models, including statistical models, are neither true nor false (Bailer-Jones, 2003; Hughes, 1997; Hutten, 1954), unlike the propositions about the world that they represent. We believe that Box's (1979) famous dictum that 'all models are false but some are useful' could be shortened to 'some models are useful' without any loss. The main question to us is to what extent the inferences made using representations can be applied to corresponding inferences about the world. Useful statistical models and procedures will provide inferences that can be 'interpreted' in such a way as to be useful for inference in the real world.

Hughes (1997) describes a framework that can be used to understand the process of how scientific models are used, which he called the 'denotation, demonstration, interpretation' (DDI) framework (Figure 1). To help answer the researcher's question, the statistician will develop a statistical model. This move from the 'real world' into the model world Hughes calls 'denotation'. The statistical model, by necessity, is an idealized representation. The researcher's hypotheses about the real world are not answered directly; instead, questions about parameters of the statistical model are answered. Inference about the mean value of a population, for instance, is replaced by inference about the normal distribution, which is a representation of the population of interest. Hughes called this process of acting on representations 'demonstration'. Finally, inferences with respect to the representation must be translated back into the world through interpretation of the statistical inference. Hughes's conception of the role of models is central to how we view Bayesian analysis.

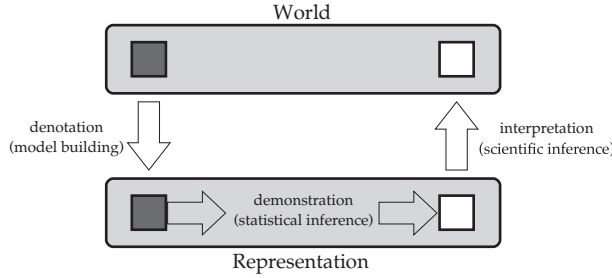

Figure 1. Hughes's (1997) DDI model of scientific representation. 'World' squares represent phenomena (dark) or propositions about phenomena (light); 'representation' squares represent models (dark) or inferences about models (light).

3. The humble Bayesian

With these differences in perspective in place, we now spell out how the practice of model checking can be aligned with Bayesian inference, as long as we are suitably humble in applying our inferences. We call our view 'humble Bayes',[2] but we make no claims as to its novelty. We note that GS's list of those who advocate some form of model checking is a veritable who's who of twentieth-century Bayesian statisticians, and we suspect that most Bayesians adhere to a similar philosophy, without giving it a name.

3.1. Open-minded inference

In the foregoing we noted that GS's falsificationism is not easily incorporated in a coherent theory of model checking. But for our Bayesian theory, we borrow from falsificationism what we take to be its greatest virtue: its open-mindedness. To its credit, there is no suggestion in the approach of GS that the models presently under consideration are in some sense true, and new models can enter the arena at any stage of investigation. By contrast, a Bayesian who is pondering over a fixed set of models seems ultimately closed-minded. She has a prior probability over the set which expresses her belief in each of the options available, and these priors sum to unity, meaning that the disjunction of the models is believed with absolute certainty (cf. Dawid, 1982).

If, on the other hand, we decide to employ odds as expressions of relative belief,[3] then it is left open whether or not the probabilities of the models under consideration sum to unity. A Bayesian who employs odds is silent on whether or not she is in possession of the true model, and, in fact, need not acknowledge the existence of a true model at all. But such a Bayesian is nevertheless able to incorporate prior beliefs into the inference. With minor interpretative adjustments, it is possible to incorporate openmindedness in the Bayesian inferential framework.

The primary inferential machinery in humble Bayesianism is thus traditionally Bayesian, using posterior distributions, model odds, and Bayes factors, the choice of which is largely driven by the research question. These Bayesian quantities are used to perform inferences within the statistical models at hand, but also to evaluate models and compare them to one another. This is unlike the overconfident Bayesian, simply because the models are questioned. Model checking serves two roles, which we can spell out in terms of the perspective on models given above: determination of the extent to which inferences can be carried from the statistical representation into the real world, and support of the generation of new models for comparison.

3.2. Model checking assures applicability of Bayesian inferences

In scientific settings, the quantities of interest are not quantities in any statistical model; rather, a researcher has questions about a particular population or process. These questions, if they are well formed, can be reframed in terms of propositions about the world that are either true or false. There are uncountably many statistical models that could be used to help answer the researcher's questions, but we emphasize that the original question is not itself a question about a statistical parameter or model.

---

[2] Readers who have interacted with Bayesians may find the term 'humble Bayesian' oxymoronic.

[3] 'Belief' in this sense may be part of the statistical representation, and may or may not reflect the analyst's belief in the corresponding proposition in the real world.

With this understanding, Bayesian confirmation theory still provides meaningful inferences. The goal of humble Bayesian confirmation theory is not to confirm a 'true' model. Because the model itself is not true (nor is it false), neither confirming it nor falsifying it can be our goal. However, we can take a confirmation as indicating something important about the world. GS mention that Bayes factors and posterior probabilities can be useful as long as they 'are not taken too seriously'. Our reason for not taking them too seriously is not that the underlying models are false; rather, it is that they are not the ultimate target for inference. The Bayes factor or posterior probability must be interpreted. If models are useful, statements about statistical parameters will correspond to statements about the world, but this correspondence will not be exact.

Overconfident Bayes is problematic because it lacks the necessary humility that accompanies the understanding that inferences are based on representations. We agree that there is a certain silliness in computing a posterior odds between model A and model B, seeing that it is in favour of model A by 1 million to one, and then declaring that model A has a 99.9999% probability of being true. But this silliness arises not from model A being false. It arises from the fact that the representation of possibilities is quite likely impoverished because there are only two models. This impoverished representation makes translating the representational statistical inferences into inferences pertaining to the real world difficult or impossible.[4] For this reason, we prefer to speak of 'model comparison' rather than 'model selection': models need never be selected as true, but they can be compared in meaningful and informative ways.

The key, then, is to ensure that our statistical inferences can be interpreted in a useful way into real-world inferences. What must we do to ensure that demonstrations in the representation realm can be interpreted in such a way that they correspond in a useful way to statements about the world? This is a difficult question to answer, but at minimum we believe it requires that the ancillary assumptions used to generate models are not unreasonable. Even in a fully Bayesian framework, model checks are necessary. Model checks help to assure ourselves that interpretation of the results is possible in a way that is useful for real-world inferences.

3.3. Model checking helps generate new models

In addition to helping assure ourselves that our Bayesian quantities are useful, model checks also spur the creation of new models, which can then be tested within the standard Bayesian model testing framework. GS describe model testing as being outside the scope of Bayesian confirmation theory, and we agree. This should not be seen as a failure of Bayesian confirmation theory, but rather as an admission that Bayesian confirmation theory cannot describe all aspects of the data analysis cycle. It would be widely agreed that the initial generation of models is outside Bayesian confirmation theory; it should then be no surprise that subsequent generation of models is also outside its scope.

Generating multiple models in Hughes's denotation phase allows for a richer representation of the world. Because the quality of our inferences is related to the richness of our representation of the world (or possible worlds), generation of new models is essential to ensuring that our Bayesian inferences are applicable to real-world scenarios. Statistical inferences, including Bayesian ones, are only as useful as the underlying representation admits.

[4] On the other hand, in some research scenarios inference from impoverished representations may be possible. The usefulness of a two-model comparison is highly dependent on the phenomenon and research question.

We therefore believe that model checking complements the Bayesian confirmatory approach to statistical inference. To a humble Bayesian, models are not true or false, but are representations. The humility in the humble Bayesian approach comes from understanding that these models are not the ultimate target of inference, and that model checking helps to ensure that we can bridge the gap between the representational world and the real world.

4. Conclusion

The humble Bayesian approach we have sketched out here has the advantage that it retains the core of the Bayesian confirmatory method with its formal inferential theory, something that GS's approach lacks. It avoids many of the criticisms of GS by keeping an open mind through model checking, and through humility, by understanding that Bayesian quantities must be interpreted. The applicability of Bayesian quantities will be determined by the quality of the statistical representation, which can be checked using the methods GS advocate.

References

Bailer-Jones, D. M. (2003). When scientific models represent. *International Studies in the Philosophy of Science*, *17*, 59–74. doi:10.1080/02698590305238

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 385–402. doi:10.1214/06-BA115

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics: Proceedings of a workshop* (pp. 201–236). New York: Academic Press.

Dawid, P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379), 605–610. doi:10.2307/2287720

de Finetti, B. (1995). The logic of probability. *Philosophical Studies*, 77, 181–190. doi:10.1007/BF00996317

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal of the Philosophy of Science*, *45*(1), 1–35. doi:10.1093/bjps/45.1.1

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

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*, 403–420. doi:10.1214/06-BA116

Howson, C. (2001). The logic of Bayesian probability. In D. Corfield & J. Williamson (Eds.), *Foundations of Bayesianism* (pp. 137–159). Dordrecht: Kluwer.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science*, *64*, S325–S336. doi:10.1086/392611

Hutten, E. H. (1954). The rôle of models in physics. *British Journal for the Philosophy of Science*, *4*, 284–301. doi:10.1093/bjps/IV.16.284

Kieseppä, I. (2001). Statistical model selection criteria and Bayesianism. *Philosophy of Science (Proceedings)*, *68*(3), S141–S152. doi:10.1086/392904

Mayo, D. (1996). *Error and the growth of scientific knowledge*. Cambridge, MA: MIT Press.

Mayo, D., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of the philosophy of science, Vol. 7: Philosophy of statistics* (pp. 153–198). London: Elsevier.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. doi:10.1037/a0024377

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36. doi:10.1016/j.jmp.2007.09.007

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2009). A truncated-probit item response model for estimating psychophysical thresholds. *Psychometrika*, *74*, 603–618. doi:10.1007/s11336-009-9122-3

Romeijn, J.-W. (2011). Statistics as inductive inference. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of the philosophy of science, Vol. 7: Philosophy of statistics* (pp. 751 –775). London: Elsevier.

Romeijn, J.-W., & van de Schoot, R. (2008). A philosopher's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 329 –357). New York: Springer.

Romeijn, J.-W., van de Schoot, R., & Hoijtink, H. (2012). One size does not fit all: Derivation of a prior-adapted BIC. In D. Dieks, W. Gonzalez, S. Hartmann, M. Stöltzner, & M. Weber (Eds.), *Probabilities, laws, and structures* (Vol. 3, pp. 87–105). Dordrecht: Springer.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008a). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389. doi:10.1037/0096-3445.137.2.370

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689. doi:10.3758/s13423-011-0088-7

Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008b). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*, 1201–1208. doi:10.3758/PBR.15.6.1201

Savage, L. J. (1972). *The foundations of statistics*. (2nd ed.) New York: Dover.

Received 10 March 2012

British Journal of Mathematical and Statistical Psychology (2013), 66, 76–80

The British Psychological Society

www.wileyonlinelibrary.com

Author response

Rejoinder to discussion of 'Philosophy and the practice of Bayesian statistics'

Andrew Gelman[1]* and Cosma Shalizi[2]

[1]Department of Statistics and Department of Political Science, Columbia University, New York, USA

[2]Department of Statistics, Carnegie Mellon University, Pittsburgh, USA

Different views of Bayesian inference

The main point of our paper was to dispute the commonly held view that Bayesian statistics is or should be an algorithmic, inductive process culminating in the calculation of the posterior probabilities of competing models. Instead, we argued that effective data analysis – Bayesian or otherwise – proceeds more messily through a jagged process of formulating research hypotheses, exploring their implications in light of data, and rejecting aspects of our models in light of systematic misfits compared to available data or other sources of information. We associate this last bit with Popper's falsification or (weakly) with Kuhn's scientific revolutions, but these connections with classical philosophy of science are not crucial. Our real point is that Bayesian data analysis, in the form that we understand and practise, requires the active involvement of the researcher in constructing and criticizing models, and that from this perspective the entire process of Bayesian prior-to-posterior inference can be seen as an elaborate way of understanding the implications of a model so that it can be effectively tested. Just as a chicken is said to be nothing but an egg's way of making another egg, so posterior inference is a way that a model can evaluate itself. But this inference and evaluation process, in our view, has essentially nothing to do with calculations of the posterior probability of competing models. As we discuss in our paper, for technical, philosophical, and historical reasons we tend not to trust such marginal posterior probabilities. (See Figure 1 of our paper for the sort of reasoning that we do *not* like.)

Given all this, the discussion of our paper is remarkably uncontentious: none of the five discussants express support for the standard (according to Wikipedia) view of an overarching inductive Bayesian inference, and all agree with us that the messiness of

*Correspondence should be addressed to Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027, USA (e-mail: gelman@stat.columbia.edu).

DOI:10.1111/j.2044-8317.2012.02066.x

real-world data analysis is central to statistical reasoning, not a mere obstacle to be cleaned up by means of a better prior distribution.

The discussants present different perspectives, but a common theme is that our own recommended approach of Bayesian analysis and posterior predictive checking is itself limited or, at the very least, is only one of many ways to approach statistical inference and decision-making.

We agree, and we briefly respond to each discussant in turn and then summarize our points.

Response to specific comments

Denny Borsboom (2013) points out that there are other Bayesian philosophies beyond the two discussed in our paper. We considered the 'usual story' based on computing the posterior probabilities of competing models and our preferred falsificationist attitude. Borsboom argues that a fuller philosophy of statistics – Bayesian or otherwise – should also account for confirmation and construction of models as well as inference and criticism of models that have already been proposed. We agree that our philosophy is incomplete and welcome such additions. We have had ideas of models for the model-building process using a recursive language-like framework in which a model is built from existing pieces (by analogy with the stepwise curve-fitting algorithm of Schmidt & Lipson, 2009; and reviewed by Gelman, 2009). But such models are a distant approximation to how we actually put models together. We hope that the philosophical approaches suggested by Borsboom lead us to a better understanding of the interaction between inference and the construction of models.

John Kruschke (2013) argues in the opposite direction, that our philosophy is not Bayesian enough, and that with a careful re-expression we can integrate predictive model checking into the inductive Bayesian fold. Kruschke notes that the act of interpreting a model check is itself a form of inference, perhaps with our brain's visual system performing some version of Bayesian decision-making. Indeed, one way to conceptualize the incompleteness of our philosophical framework is to imagine trying to program Bayesian data analysis in an artificial intelligence system. Inference would be no problem (at least for the large class of models that can be fitted in reasonable time using Markov chain Monte Carlo, variational Bayes, or some other existing computational approach). And we could just about imagine model expansion being performed algorithmically using some alphabet of models. But how would the artificial intelligence perform posterior predictive checks, if the program does not have a 'homunculus' to assess discrepancies between observed and predicted data? Kruschke is perhaps correct that this sort of comparison could itself be performed Bayesianly, and we are interested in the potential of this process being automated.[1] Deborah Mayo (2013) explains how, in our efforts to model our modelling process, we have oversimplified the philosophies of Popper and others. Here we fear we fall into a long tradition of scientists who attempt to develop philosophical principles via introspection – but without introspecting carefully enough.

[1] We disagree, however, with Kruschke's view that it is desirable to penalize complex models, automatically or otherwise. Instead we prefer the following dictum from Radford Neal (1996): Sometimes a simple model will outperform a more complex model . . . Nevertheless, I [Neal] believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

This is a good place for us to repeat our belief and hope that we are clearing the air by describing how we do Bayesian inference without comparing the posterior probabilities of models. Describing the philosophy of what we *do* do – that is more of a challenge. In particular, Mayo picks up on a hole in our philosophy that matches a similar gap in classical statistics: how do we decide when a discrepancy between replications and data is 'statistically significant', and what do we do about it? It is all well and good for us to emphasize practical significance and our concern for aspects of model misfit that are substantively important, but we still end up looking at *p*-values (or their graphical equivalent) one way or another. We are still struggling with this issue in our applied work: this gap in our philosophy represents a gap in our practical tools as well.

Richard Morey and colleagues (2013) argues that we are too quick to abandon Bayesian philosophy: by considering priors as 'regularization devices' rather than as true probability distributions, we are abandoning 'the interpretation of the corresponding posterior' and thus diminishing the value of the simulations of predictive data that we are using to check our model. In our paper we frame the strong assumptions of Bayesian inference as a feature rather than a bug: the stronger the assumptions, the more ways the model can be checked. This is a Popperian idea, that the best models make lots of predictions and are ready to be refuted, with the act of refutation being the spur to improvement. Morey, Romeijin, and Rouder is going one more step, noting that to the extent that we equivocate about the probabilistic nature of our priors, we are reducing our ability to learn from falsification. Belief is the foundation of scepticism, and by refusing to commit we are also losing an opportunity to refute. In that spirit, Morey, Romeijin, and Rouder would like to preserve the marginal probability calculation giving the relative (although not absolute) posterior probabilities of competing models, thus taking a half-way point between the standard view (in which new evidence causes the better model to dominate, with no need for the steps of model checking and improvement) and our view (in which these marginal probabilities and Bayes factors are so dependent on arbitrary aspects of the model as to be useless; see Section 4.3 of our paper).

Stephen Senn (2013) likes our applied work but points out some holes in our theory. This is important because undoubtedly we could have achieved similar results using other statistical approaches. Bayes is fine but other regularization methods could also do the job. In practice the following seem to be important in developing a method to solve problems in applied statistics: (1) the method must have a way to incorporate diverse sources of data (e.g., survey responses, demographic information, and election outcomes in the vote modelling problem); (2) when large amounts of data come in, the method must be flexible enough to expand, either non-parametrically or through a sieve-like set of increasingly dense forms; and (3) the estimation must be regularized to avoid overfitting. Bayesian inference has some particular advantages in that it automatically unifies inference and prediction (and its predictive simulations can be directly used to check model fit), but these are second-order benefits. Other modes of inference can be hacked to yield probabilistic predictions as needed. In any case, Senn points out a problem in our philosophy as well as other formulations of realistic statistical practice: we choose our model based on its fit to the data, thus the statistical properties of the *model* we choose are not the same as the (generally unstated) statistical properties of our full *procedure.* We do not know how important this is, but we suppose that a useful start would be to investigate this difference in some special cases in which a class of models is fitted and some rule is used to stop or go forward (we will not say 'accept or reject') given the results of a posterior predictive check.

As can be seen from our comments, the discussants raise complementary points. In our philosophy, neither model building nor model checking is fully formulated. Our weak defence is that in practice these steps are not so well understood but are part of any serious applied modelling, thus we prefer to include model building and checking as open-ended steps. We want our framework to catch up with statistical practice, but we find it difficult to devise a philosophy that anticipates future methods. But this is only a weak response: all the discussants raise important ideas that point towards potentially useful research in modelling the process of applied statistics.

As we say in our paper, the philosophy of statistics is not a mere game: wrong philosophies can trap people in relatively ineffective methods (this is how we feel about many of the applications of Bayes factors), whereas forward-looking philosophies can point towards methodological improvements (such as the ideas for Bayesian model building and model checking raised by some of the discussants here).

Looking forward

In summary, our goal in writing our paper was not to say that Bayesian inference is 'better' but to delineate what is done when we do Bayes, as compared to the 'party line' of what people say is done.

When we were beginning our statistical educations, the word 'Bayesian' conveyed membership in an obscure cult. Statisticians who were outside the charmed circle could ignore the Bayesian subfield, while Bayesians themselves tended to be either apologetic or brazenly defiant. These two extremes manifested themselves in ever more elaborate proposals for non-informative priors, on the one hand, and declarations of the purity of subjective probability, on the other.

Much has changed in the past 30 years. 'Bayesian' is now often used in casual scientific parlance as a synonym for 'rational', the anti-Bayesians have mostly disappeared, and non-Bayesian statisticians feel the need to keep up with developments in Bayesian modelling and computation. Bayesians themselves feel more comfortable than ever constructing models based on prior information without feeling an obligation to be non-parametric or a need for priors to fully represent a subjective state of knowledge.

In short, Bayesian data analysis has become normalized. Our paper is an attempt to construct a philosophical framework that captures applied Bayesian inference as we see it, recognizing that Bayesian methods are highly assumption-driven (compared to other statistical methods) but that such assumptions allow more opportunities for a model to be checked, for its discrepancies with data to be explored.

We felt that a combination of the ideas of Popper, Kuhn, Lakatos, and Mayo covered much of what we were looking for – a philosophy that combined model building with constructive falsification – but we recognize that we are, at best, amateur philosophers. Thus we feel our main contribution is to consider Bayesian data analysis worth philosophizing about.

Bayesian methods have seen huge advances in the past few decades. It is time for Bayesian philosophy to catch up, and we see our paper as the beginning, not the end, of this process.

Acknowledgements

We thank the editors of this journal for organizing the discussion and the US National Science Foundation for partial support of this work.

References

Borsboom, D. (2013). How to practise Bayesian statistics outside the Bayesian church: what philosophy for Bayesian statistical modelling? *British Journal of Mathematical and Statistical Psychology*, *66*, 39–44. doi:10.1111/j.2044-8317.2011.02062.x

Gelman, A. (2009). Equation search, part 2. Statistical Modeling, Causal Inference, and Social Science blog, 8 December. Retrieved from http://andrewgelman.com/2009/12/equation_search_1/

Kruschke, J. (2013). Posterior predictive checks can and should be Bayesian: comment on Gelman and Shalizi, 'Philosophy and the practice of Bayesian statistics'. *British Journal of Mathematical and Statistical Psychology*, *66*, 45–56. doi:10.1111/j.2044-8317.2011.02063.x

Mayo, D. (2013). The error-statistical philosophy and the practice of Bayesian statistics: comments on Gelman and Shalizi: philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 57–64. doi:10.1111/j.2044-8317.2011.02064.x

Morey, R., Romeijin, J.-W., & Rouder, J.N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*, 68–75. doi:10.1111/j.2044-8317.2011.02067.x

Neal, R. (1996). *Bayesian learning for neural networks*. New York: Springer.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, *324*, 81–85. doi:10.1126/science.1165893

Senn, S. (2013). Comment on Gelman and Shalizi. *British Journal of Mathematical and Statistical Psychology*, *66*, 65–67. doi:10.1111/j.2044-8317.2011.02065.x

Received 8 June 2012