

Discussion by
J.M. Robins,
pp. 67-70.

BAYESIAN STATISTICS 6, pp. 53-82
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)
© Oxford University Press, 1998

Quantifying Surprise in the Data and Model Verification

M. J. BAYARRI and JAMES O. BERGER
Universitat de València, Spain, and Duke University, USA

SUMMARY

P-values are often perceived as measurements of the degree of surprise in the data, relative to a hypothesized model. They are also commonly used in model (or hypothesis) verification, i.e., to provide a basis for rejection of a model or hypothesis. We first make a distinction between these two goals: quantifying surprise can be important in deciding whether or not to search for alternative models, but is questionable as the basis for rejection of a model. For measuring surprise, we propose a simple calibration of the *p*-value which roughly converts a tail area into a Bayes factor or 'odds' measure. Many Bayesians have suggested certain modifications of *p*-values for use in measuring surprise, including the *predictive p-value* and the *posterior predictive p-value*. We propose two alternatives, the *conditional predictive p-value* and the *partial posterior predictive p-value*, which we argue to be more acceptable from Bayesian (or conditional) reasoning.

Keywords: BAYES FACTORS; BAYESIAN *P*-VALUES; BAYESIAN ROBUSTNESS; CONDITIONING; MODEL CHECKING; PREDICTIVE DISTRIBUTIONS.

1. INTRODUCTION

The statistical tool that causes the most consternation among Bayesians is undoubtedly the *p*-value. Its use in model comparison and hypothesis testing is frequently reviled (cf. Edwards, Lindman, and Savage, 1963, Berger and Selke, 1987, and Berger and Delampady, 1987), but sometimes argued to be sensible (cf. Casella and Berger, 1987). Even Bayesians who decry the use of *p*-values in model comparison often defend their use in model checking (cf. Box, 1980).

In parametric statistical analysis of data X , one is frequently working, at a given moment, with an entertained model or hypothesis $H_0 : X \sim f(x|\theta)$. We assume $f(x|\theta)$ is a density, typically with respect to Lebesgue measure, and will often assume that, a priori, the unknown θ has (usually noninformative) prior density $\pi(\theta)$. A statistic $T(X)$ is chosen to investigate compatibility of the model with the observed data x_{obs} , with large values of T indicating less compatibility. *P*-values are defined as

$$p = \Pr(T(X) \geq T(x_{obs})), \quad (1.1)$$

with there being a number of different proposals for the probability distribution to be used in this computation when θ is unknown.

This paper consists of three rather distinct parts. Section 2 reviews the chief difficulty with a *p*-value as in (1.1), namely that it is a tail area rather than a likelihood-based measure, and discusses a simple calibration of *p* to partially overcome this problem. The calibration is easy to state: simply compute $\underline{B}(p) = -ep \log(p)$, when $p < e^{-1}$, and interpret this as a lower bound on the *Bayes factor* (or odds) of H_0 to H_1 , where H_1 denotes the (unspecified) alternative to H_0 . This calibration will be motivated from a robust Bayesian perspective.

While more readily interpretable than a p -value, the fact that $\underline{B}(p)$ is only a lower bound on the Bayes factor implies that it cannot be directly interpreted as a measure of the evidence against the given hypothesis H_0 . Rather, it should be viewed as a 'measure of surprise.' The idea is that one will often decide to entertain alternatives to H_0 only if the data seem surprising under H_0 . Thus $\underline{B}(p)$ is used to trigger a search for alternative models, but H_0 would only be rejected if an alternative model is found which is shown to be considerably better by a full Bayesian analysis. This is the subject of Section 3.

Section 4 of the paper discusses choice of the probability distribution used to compute (1.1) when θ is unknown. Three common suggestions for this distribution are $f(x|\hat{\theta})$, where $\hat{\theta}$ is the m.l.e. for θ ; the Bayesian predictive distribution (cf. Box, 1980); and the posterior predictive distribution (cf. Guttman, 1967, and Rubin, 1984). We propose two alternatives, the *conditional predictive p-value* and the *partial posterior predictive p-value*, arguing that they will yield answers considerably more compatible with intuition and with Bayesian reasoning.

As the problem considered here is among the most discussed problems in statistics, there is a huge literature on the subject. On the Bayesian side, it is often argued that choosing T is as difficult as choosing an alternative model and, hence, that one should stick to the 'pure' Bayesian approach of actually specifying an alternative to H_0 and performing a fully Bayesian analysis. Other Bayesians have proposed default Bayesian analyses based on automatic embedding of the model under question in a large family of 'default' alternatives. This embedding can be parametric (cf. Bayarri, 1986, and Delampady and Berger, 1990) or nonparametric (cf. Verdinelli and Wasserman, 1998). Space precludes discussion of these issues here.

An even larger literature exists on the development of other alternatives to p -values for measuring surprise. A few of the many references are Weaver (1948), Good (1956, 1983, 1988), Berger (1985), and Evans (1997); extensive discussion can be found in Bayarri and Berger (1997). The chief motivation for these alternative measures is the intuition that 'likelihood,' not 'tail areas,' should be the basis of any statistical measures. In this regard, note that the 'calibration' of p -values mentioned earlier can be considered to be a conversion of the tail area into a likelihood or odds measure. Attacking the problem in this way has the pragmatic advantage of allowing the methodology to be based on the familiar notion of p -values, and has the theoretical advantage of inheriting desirable properties of p -values, such as invariance to transformation (a property not shared by most likelihood measures of surprise).

2. CALIBRATION OF P -VALUES

This section reviews the developments in Sellke, Bayarri and Berger (1999). We assume here that the model under consideration is the simple $H_0 : X \sim f(x)$, so that the p -value in (1.1) is computed with respect to $f(x)$. In Section 4, we consider the reduction of a model with unknown parameters to a simple model, for the purpose of calculating a p -value.

The proposal for calibrating a p -value is to compute, when $p < e^{-1}$,

$$\underline{B}(p) = -ep \log(p) \quad (2.1)$$

and interpret this as a lower bound on the Bayes factor (or odds) of H_0 to H_1 , where H_1 denotes the (unspecified) alternative to H_0 . Note that $\underline{B}(p)$ typically differs by an order of magnitude from the p -value itself (cf. the first two rows of Table 1), indicating the substantial difference between p -values and Bayes factors. Both parametric and nonparametric arguments for this calibration are given in Sellke, Bayarri and Berger (1999). We first report a typical parametric example, and then give the nonparametric argument.

Example 1. Assume that X_1, X_2, \dots, X_n is an i.i.d sample from the $N(0, 1)$ null model. Suppose $T(X) = |\bar{X}|$, so that the p -value is given by $p = 2(1 - \Phi(\sqrt{n}|\bar{x}|))$, where $\Phi(\cdot)$ is the

standard normal c.d.f. Consider alternative models of the form $X_i|\theta \sim N(\theta, 1)$ with $\theta \sim \pi_1(\theta)$. Berger and Sellke (1987) provided lower bounds for the Bayes factor

$$B = \frac{(2\pi/n)^{-1/2} \exp\{-\frac{n}{2}\bar{x}^2\}}{\int (2\pi/n)^{-1/2} \exp\{-\frac{n}{2}(\bar{x} - \theta)^2\} \pi_1(\theta) d\theta}, \tag{2.2}$$

as $\pi_1(\theta)$ ranges over the following possible classes of priors:

$$\begin{aligned} \Gamma_{Normal} &= \{\pi_1 : \pi_1(\theta) = N(\theta|0, \tau^2), \tau^2 > 0\} \\ \Gamma_{US} &= \{\pi_1 : \pi_1(\theta) \text{ is unimodal and symmetric}\} \\ \Gamma_{Sym} &= \{\pi_1 : \pi_1(\theta) \text{ is symmetric}\}. \end{aligned}$$

The following table displays these lower bounds together with the p -values and the calibration $\underline{B}(p) = -ep \log p$; it is apparent that the calibration is quite consistent with these lower bounds.

p	0.1	0.05	0.01	0.001
$-ep \log p$	0.6259	0.4072	0.1252	0.01878
Γ_{Normal}	0.7007	0.4727	0.1534	0.02407
Γ_{US}	0.6393	0.4084	0.1223	0.01833
Γ_{Sym}	0.5151	0.2937	0.07296	0.008873

Table 1. Infimum of Bayes factors, p -values and their calibrations.

In the robust Bayesian literature, the class Γ_{US} is often perceived as particularly reasonable, not requiring specification of a functional form for the prior (as does Γ_{Normal}), and yet corresponding to natural shapes for $\pi_1(\theta)$ (which is not true of many priors in Γ_{Sym}). The close agreement between the lower bounds on the Bayes factor for the class Γ_{US} and the proposed calibration $\underline{B}(p) = -ep \log p$ is thus particularly appealing.

In ascertaining the extent to which $\underline{B}(p)$ agrees with lower bounds on Bayes factors in general, two obvious questions arise. The first is what happens in higher dimensional problems, and the second is what happens when the alternative is not a parametric alternative. In partial answer to these questions, we state the following results from Sellke, Bayarri and Berger (1999), the first of which is the generalization to higher dimensions of Example 1 and the second of which is the nonparametric result.

Theorem 1. Assume that the null model is $N_k(0, I)$, with I being the $k \times k$ identity matrix, and consider alternatives of the form $X_i|\theta \sim N_k(\theta, I)$, $i = 1, \dots, n$, with $\theta \sim N_k(0, \tau^2 I)$. Letting \underline{B} denote the lower bound on the Bayes factor over all τ^2 and assuming $p < 1/2$,

$$\lim_{k \rightarrow \infty} \frac{\underline{B}}{\underline{B}(p)} = \frac{2}{ez_p}, \tag{2.3}$$

where z_p is the $(1 - p)$ quantile of the standard normal distribution.

For the p -value equal to 0.1, 0.05, 0.01, and 0.001, the right hand side of (2.3) is 0.57, 0.45, 0.32, and 0.24, respectively. While the agreement between \underline{B} and $\underline{B}(p)$ is not as close here as in the one dimensional case of Example 1, it is rather startling that the two are even of the same order of magnitude as the dimension goes to infinity. This should lend some assurance as to the general applicability of the proposed calibration.

The nonparametric argument for the suggested calibration is based on the fact that, under the null hypothesis, the distribution of the 'random' p -value, $p(X)$, is uniform on $[0, 1]$. Instead

of assessing alternative distributions for X and prior distributions for the parameters of these alternative distributions, one can directly consider alternative distributions for $p(X)$. Indeed, we will propose a reasonable class of such alternative distributions and then compute the lower bound on the Bayes factor over this class. (Others have previously considered direct choice of alternatives for $p(X)$; see, for instance, Hodges, 1992.)

Instead of working with $p(X)$, it is more convenient to work with $Y = -\log p(X)$ and its distribution under the null hypothesis (the standard exponential distribution) and under the alternative hypothesis. A natural condition to impose on the distribution of Y under the alternative is that it has a decreasing (or, at least, non-increasing) failure rate. This is equivalent to requiring that the distribution of $Y - y \mid Y > y$ be stochastically increasing with y . In terms of $p = e^{-y}$, the requirement of decreasing failure rate for Y means that the distribution of $(p/p_0) \mid (p < p_0)$ is stochastically decreasing with p_0 , which, for instance, implies that, for any fixed ρ , the probability $\Pr\{p < \rho p_0 \mid p < p_0\}$ increases in p_0 ; this is a very reasonable kind of behavior. It is worth noting that the 'boundary' distributions having constant failure rate are the exponential distributions for Y , which correspond to Beta($\xi, 1$) distributions for $p(X)$.

Theorem 2. Consider testing

$$H_0 : p \sim Un(0, 1) \text{ versus } H_1 : p \sim g(p),$$

where $g(p)$ is such that $Y = -\log p$ has non-increasing failure rate. Then, for $p < e^{-1}$, $\underline{B}(p)$ in (2.1) is an (attainable) lower bound on the Bayes factor of H_0 to H_1 .

The calibration of p -values suggested here is specific to fixed sample size experiments involving testing of a precise hypothesis versus a larger alternative. Other types of testing, e.g. one-sided hypothesis testing, can have very different 'calibrations' between p -values and Bayes factors or odds (cf. Casella and Berger, 1987). The same is true for sequential experiments; the dependence of p -values on the stopping rule in such situations will make questionable the robust Bayesian justifications for $\underline{B}(p)$ (since Bayes factors do not depend on the stopping rule used). Interestingly, one might consider simply ignoring the stopping rule in such a situation and computing the p -value as if the experiment were a fixed sample size experiment; $\underline{B}(p)$ might then still be a reasonable calibration.

3. SURPRISE VERSUS MODEL VALIDATION

The logic behind the proposed calibration of p -values in Section 2 is simply that the resulting number, $\underline{B}(p)$, is on a Bayes factor or odds scale and can hence be more easily interpreted. Note, however, that $\underline{B}(p)$ only seems to correspond to a lower bound on the Bayes factor of H_0 to (the unspecified) H_1 , and one must realize the limitations of such lower bounds.

On the positive side, if this lower bound is large enough there is clearly no cause to question H_0 . For instance, $\underline{B}(p) = 0.5$ implies that there is at most 2 to 1 evidence against H_0 , which would not seem to be cause to seriously question H_0 . (Observe that $\underline{B}(p) = 0.5$ corresponds to $p = 0.07$, so such a p -value would also not seem to be cause to question H_0 .)

On the other hand, if $\underline{B}(p)$ is small, then it seems very reasonable to feel 'surprised' and to initiate consideration or development of alternative models. In that $\underline{B}(p)$ is only a lower bound, however, it would not be reasonable to reject H_0 based on only this evidence. Indeed, from the Bayesian robustness literature (cf. Berger, 1994) it is known that lower bounds on Bayes factors can be rather extreme, especially for large sample sizes. In Example 1, for instance, suppose the (sometimes recommended) 'default' $N(0, 1)$ prior distribution were used. If, say, $p = 0.01$ and sample sizes of 1, 10, 20, 50, 100, and 1000 are considered, simple computation then shows that the actual Bayes factors are 0.27, 0.16, 0.19, 0.28, 0.37, and 1.13, respectively.

Recalling that $\underline{B}(p) = 0.1252$ in this situation, it is clear that one should not reject H_0 solely on the basis of $\underline{B}(p)$, at least for larger sample sizes.

In the above type of situation, it can be shown that the ratio of an actual Bayes factor to $\underline{B}(p)$ is roughly proportional to \sqrt{n} ; this explains why Good (1983) proposed a calibration of p -values that included multiplication by the factor \sqrt{n} . Were we proposing conversion of p -values into actual Bayes factors for the purpose of model rejection, such a multiplicative factor would be needed. Our view, however, is that such conversion is too difficult, and that a model should really only be rejected if an alternative and *a priori believable* model can be found and shown to be superior to H_0 by a full Bayesian analysis. Note the requirement that the alternative model must be a priori believable.

Example 2. One of the major astronomical mysteries is the source of extremely large gamma ray bursts that have been observed over the last 20 years. Efron (1996) reported an earlier analysis of 260 gamma ray bursts from the BATSE1B catalog. The analysis was an effort to detect if the spatial directions from which the bursts arrived were uniform (H_0). This is viewed as important for determining whether the sources are from outside the galactic halo (in which case the bursts would be uniformly distributed). A test to detect non-uniformity yielded a p -value of 0.027. This would correspond to $\underline{B}(p) = 0.27$, which has the formal interpretation of odds of about 4 to 1 against H_0 . This seems reasonably surprising, and suggests looking for alternative models. Unfortunately, no alternative astronomical model could be found that was believable and which would yield a substantial Bayes factor in its favor. This is thus a case where one would have been 'surprised' by the data, but would eventually conclude that there is no reason to reject H_0 . The story did not end here. Many more and better observations were later obtained (BATSE2 and BATSE3) and, indeed, the 'surprising' non-uniformity in the data disappeared. It turns out that there was probably a slight bias in the BATSE1 instruments, and this bias resulted in the apparent non-uniformity in the early data. Of course, had such instrumental bias been proposed as the alternative theory to H_0 , a Bayesian analysis might well have yielded significant evidence in favor of the alternative, but this would not have had any astronomical significance (besides the need to correct the instrument!)

We have barely scratched the surface of a number of fascinating issues. For instance, the relationship between robust Bayesian lower bounds and 'surprise in the data' can often be exploited to suggest alternative models to consider. Also, there are a number of interesting issues involving the appropriateness of 'full' Bayesian analysis of alternative models that were suggested by the data through a surprise analysis. (The short summary is that this is justifiable, providing a serious effort is made to determine what the prior probability of the model would have been, had one considered it before seeing the data.) Alas, further discussion of these issues would take us too far afield.

4. PREDICTIVE P -VALUES

For the typical case in which the null model $H_0 : X \sim f(x|\theta)$ contains unknown parameters, a p -value cannot be computed directly. Classical approaches to the problem include (i) replacing θ by an estimate before computing the p -value; and (ii) finding a statistic $U(x)$ such that, conditional on U , the distribution of T does not depend on θ and then using this conditional distribution to compute the p -value (as in the Fisher exact test for independence in a contingency table). Bayesian solutions to the problem involve elimination of θ by integration with respect to some distribution of θ .

4.1. *The prior and posterior predictive p-values*

Box (1980) popularized use of p -values based on the *prior predictive distribution*,

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta. \quad (4.1)$$

To a pure Bayesian, this is the actual predictive distribution of X and is, hence, the most natural distribution to use for measuring surprise. Also, $m(x_{obs})$ would be the numerator of any Bayes factor comparing H_0 with an alternative, lending familiarity to its use.

For a chosen model departure statistic $T(x)$, the *prior predictive p-value* would then be

$$p = \Pr^m(T(X) \geq T(x_{obs})). \quad (4.2)$$

One of the attractions of basing surprise on $m(x)$ is that there is then a natural statistic, $T(x)$, to consider, namely $T(x) = 1/m(x)$, since small values of $m(x)$ (and hence large values of $T(x)$) would correspond to data that is unlikely to be observed under the null model (and prior). This choice of T is not free from criticism (for instance, it is easy to see that the resulting p -value can change if one considers a transformation of the data), but its simplicity is appealing to many.

A general concern with prior predictive p -values is their dependence on the prior $\pi(\theta)$; in essence, $m(x)$ measures the likelihood of x relative to both the model and the prior, and an excellent model could come under suspicion if a poor prior distribution were used. For this reason, and because model checking is often considered at early stages of an analysis before careful prior elicitation is performed (and/or because a non-subjective analysis might be desired from the beginning), it is attractive to attempt to utilize noninformative priors. Unfortunately, noninformative priors are typically improper, in which case the prior predictive $m(x)$ would also be improper, precluding computation of (4.2).

The concerns mentioned in the last paragraph have led many Bayesians, beginning with Guttman (1967) and Rubin (1984), to eliminate θ from $f(x|\theta)$ by integrating with respect to the posterior distribution, $\pi(\theta|x_{obs})$, instead of the prior, before computing a p -value. The *posterior predictive p-value* is thus defined as

$$p = \Pr^{m(\cdot|x_{obs})}(T(X) \geq T(x_{obs})), \quad (4.3)$$

where

$$m(x|x_{obs}) = \int f(x|\theta)\pi(\theta|x_{obs})d\theta. \quad (4.4)$$

This overcomes the difficulties with the prior predictive p -value because (i) improper noninformative priors can readily be used (since $\pi(\theta|x_{obs})$ will typically be proper) and (ii) $m(x|x_{obs})$ will typically be much more heavily influenced by the model than by the prior; indeed, as the sample size goes to infinity, the posterior distribution will essentially concentrate at $\hat{\theta}$, the m.l.e. for θ , so that $m(x|x_{obs})$ will essentially equal $f(x|\hat{\theta})$, the classical distribution most commonly used to compute a p -value. In addition, posterior predictive p -values are typically very easy to compute using output from modern MCMC Bayesian analyses.

Generalizations of (4.3) were considered in Meng (1994) and Gelman, Meng and Stern (1996); in particular, $T(X)$ could be replaced by a function $T(X, \theta)$, and $f(x|\theta)$ in (4.4) could be replaced by $f(x|\theta, A)$, where A is some other statistic. It should also be noted that a similar notion was used in Aitkin (1991) for computing actual Bayes factors.

The apparent practical appeal of posterior predictive p -values is counterbalanced by the fact that they are not really Bayesian quantities. This is indicated by the fact, observed above,

that, for large sample sizes, the posterior predictive p -value will essentially equal the classical p -value, which (as discussed in Section 2) is quite non-Bayesian in character. Also troubling is the apparent "double use" of the data in (4.3), first to convert the (possibly improper) prior $\pi(\theta)$ into a proper distribution $\pi(\theta|x_{obs})$ for determining the reference distribution $m(x|x_{obs})$, and then for computing the tail area corresponding to $T(x_{obs})$. We will see an example of the unnatural behavior that this double use of the data can induce. Finally, defenders of the prior predictive, as opposed to the posterior predictive, point out that the former has natural Bayesian interpretations, while the latter does not. Here is an artificial, but illuminating, example, inspired by Goldstein (1991).

Example 3. Suppose θ is generated as a random integer between 1 and 1000. Thus the prior distribution is uniform on the integers $\{1, 2, \dots, 1000\}$. Consider two different models. Under M_1 , the data X equals θ with probability one while, under M_2 , the value of θ is ignored and a new random integer X between 1 and 1000 is generated. The prior predictive distributions, $m_1(x)$ and $m_2(x)$, are both clearly uniform on the integers $\{1, 2, \dots, 1000\}$, properly reflecting the fact that x_{obs} tells us nothing about which model is true. The posterior predictive distributions are, however, very different: the posterior predictive distribution under M_1 is degenerate at x_{obs} , while that under M_2 is uniform on $\{1, 2, \dots, 1000\}$.

4.2. The conditional predictive p -value

It would be very attractive if the advantages of the posterior predictive p -value and the prior predictive p -value could be achieved in the same procedure, without the disadvantages. Such a procedure would (i) be based on the prior predictive $m(x)$, which has natural Bayesian meaning; (ii) be more heavily influenced by model adequacy than prior adequacy; (iii) allow use of (improper) noninformative priors; and (iv) not involve a double use of the data. We will argue that these advantages can all be obtained by choice of an appropriate statistic $U(X)$ and use of the *conditional predictive distribution*, $m(t|u)$, to compute a p -value for the departure statistic T . This leads to the *conditional predictive p -value*, defined as

$$p = \Pr^{m(\cdot|u_{obs})}(T \geq t_{obs}), \quad (4.5)$$

where $u_{obs} = U(x_{obs})$, $t_{obs} = T(x_{obs})$, and (formally)

$$m(t|u) = \int f(t|u, \theta)\pi(\theta|u)d\theta, \quad (4.6)$$

assuming that

$$\pi(\theta|u) = \frac{f(u|\theta)\pi(\theta)}{\int f(u|\theta)\pi(\theta)d\theta} \quad (4.7)$$

is proper; here $f(t|u, \theta)$ and $f(u|\theta)$ are defined as the obvious conditional and marginal distributions of T and U under H_0 . This can achieve the goals mentioned above because (i) when $\pi(\theta)$ is proper, $m(t|u)$ is the conditional distribution of T given U arising from the prior predictive $m(x)$, and is hence a natural Bayesian quantity; (ii) with appropriate choice of U , (4.5) can be made to primarily reflect surprise in the model; (iii) noninformative priors can be used, as long as $\pi(\theta|u)$ is proper; and (iv) there is no double use of the data, since only part of the data (u_{obs}) is used to produce the posterior to eliminate θ , while another part (t_{obs}) is used when computing the tail area. We illustrate with an example from Meng (1994).

Example 4. Assume that, under the null, the X_i are iid $N(0, \sigma^2)$, with σ^2 unknown. The statistic $T(x) = |\bar{x}|$ is chosen to measure departure from the model (which would be natural

for detecting discrepancy in the mean of the model). Let $U(x) = s^2 = \sum(x_i - \bar{x})^2/n$, and consider the usual non-informative prior for σ^2 : $\pi(\sigma^2) \propto 1/\sigma^2$. Computation shows that the posterior distribution, $\pi(\sigma^2|s^2)$, is $\text{Ga}^{-1}((n-1)/2, ns^2/2)$ and that

$$m(\bar{x}|s_{obs}^2) = t_{n-1}(\bar{x} | 0, \frac{s_{obs}^2}{n-1}) \quad \text{or} \quad \frac{\sqrt{n-1} \bar{X}}{s_{obs}} \sim t_{n-1}(\cdot | 0, 1) . \quad (4.8)$$

(Here Ga^{-1} and t_{n-1} denote, respectively, the Inverse Gamma distribution and the t distribution with $n-1$ degrees of freedom.) The resulting conditional predictive p -value is simply

$$p = Pr\{|\bar{X}| > |\bar{x}_{obs}|\} = 2 \left[1 - \Upsilon_{n-1} \left(\frac{\sqrt{n-1} |\bar{x}_{obs}|}{s_{obs}} \right) \right] , \quad (4.9)$$

where Υ_{n-1} stands for the distribution function of the t distribution with $n-1$ degrees of freedom. This is perfectly satisfactory (and, indeed, equals the usual classical p -value for the problem).

A prior predictive p -value cannot be computed for this example, since the prior distribution is improper. The posterior predictive p -value can be computed; indeed, $\pi(\sigma^2|x_{obs})$ is $\text{Ga}^{-1}(n/2, n(s^2 + \bar{x}^2)/2)$, and the posterior predictive distribution of \bar{X} is

$$m(\bar{x}|x_{obs}) = t_n(\bar{x} | 0, \frac{s_{obs}^2 + \bar{x}_{obs}^2}{n}) \quad \text{or} \quad \frac{\sqrt{n} \bar{X}}{\sqrt{s_{obs}^2 + \bar{x}_{obs}^2}} \sim t_n(\cdot | 0, 1) . \quad (4.10)$$

It follows that the posterior predictive p -value is given by

$$p = Pr\{|\bar{X}| > |\bar{x}_{obs}|\} = 2 \left[1 - \Upsilon_n \left(\frac{\sqrt{n} |\bar{x}_{obs}|}{\sqrt{s_{obs}^2 + \bar{x}_{obs}^2}} \right) \right] . \quad (4.11)$$

This is unsatisfactory, as can be seen by letting $|\bar{x}_{obs}| \rightarrow \infty$. Then $p \rightarrow 2[1 - \Upsilon_n(\sqrt{n})]$, which is a positive constant for any n . For instance, when $n = 4$ this constant is 0.12, and the posterior predictive p -value never drops below this constant, no matter how many standard deviations \bar{x}_{obs} is from zero. The inadequacy of the posterior predictive p -value here can be directly traced to the double use of the data, in particular to the fact that \bar{x}_{obs} is involved in computing both the posterior and the tail area.

Choice of U . The key to the conditional predictive p -value is suitable choice of the conditioning statistic U . Different possible choices of U are explored in Bayarri and Berger (1997). (See also Evans, 1997, where the conditional predictive distribution is used to develop alternate measures of surprise, with U and T being chosen to be separate subsamples of the data.) The intuition behind suitable choice of U is that one wants U to contain as much information about θ as possible, so that $\pi(\theta|u_{obs})$ will effectively eliminate θ (via integration), subject to the constraint that U should not involve T , as this would entail a 'double use' of part of the data. In Example 4, for instance, $\sum x_i^2/n$ would contain all information about σ^2 (being a sufficient statistic under the presumed model), but does involve $T(x) = |\bar{x}|$. The obvious solution (used in Example 4) is to define $U(x) = s^2 = \sum(x_i - \bar{x})^2/n$, since this contains the information about σ^2 that is independent of $T(X)$.

Investigations in Bayarri and Berger (1997) also suggest that $U(x)$ should have the same dimension as θ . The simplest general algorithm that achieves these various aims, for the case of continuous data, is to define U to be the conditional m.l.e. of θ , given $T(x) = t$,

$$U(x) = \hat{\theta} = \arg \max f(x|t, \theta) = \arg \max \frac{f(x|\theta)}{f(t|\theta)} . \quad (4.12)$$

(The situation of discrete data is considerably more difficult; while $U(x)$ in (4.12) is still typically well defined, it will usually not be suitable as a conditioning statistic.) Note that $m(t|u)$ is unaffected by one-to-one transformations of $U(x)$, so that any one-to-one transformation of θ is satisfactory as the choice of U . In Example 4, for instance,

$$f(x|t, \sigma^2) \propto (\sigma^2)^{-\frac{n-1}{2}} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\},$$

which is maximized at $\hat{\sigma}^2 = ns^2/(n-1)$, choice of which is clearly equivalent to choosing $U(x) = s^2$, as was previously done. Here is another example.

Example 5. Assume that X_1, X_2, \dots, X_n is a random sample from the exponential $\text{Ex}(\lambda)$ distribution. We consider several possibilities for T , and derive the suggested U in each case. In the following, $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ denote the order statistics for the observations.

1. $T = X_{(n)} - X_{(n-1)}$ (which could be used for detecting departures from exponential decay of the upper tail). Then, since $T \sim \text{Ex}(\lambda)$,

$$f(x|t, \lambda) \propto \lambda^{n-1} \exp\left\{-\lambda \left[\sum_{i=1}^{n-1} x_{(i)} + x_{(n-1)}\right]\right\}, \quad (4.13)$$

and the U suggested by maximizing (4.13) is $\sum_{i=1}^{n-1} x_{(i)} + x_{(n-1)}$ (again recalling that any convenient one-to-one transformation of the conditional m.l.e can be chosen).

2. $T = X_{(n)}$ (which could be used to investigate the upper tail of the null distribution). Then

$$f(x|t, \lambda) \propto \left(\frac{\lambda}{1 - e^{-\lambda t}}\right)^{n-1} \exp\left\{-\lambda \left[\sum x_{(i)} - t\right]\right\}, \quad (4.14)$$

which would have to be numerically maximized over λ to determine U via (4.12).

3. $T = X_{(1)}$ (which could be used to investigate the lower tail of the null distribution). Here, an easy computation shows that

$$f(x|t, \lambda) \propto \lambda^{n-1} \exp\left\{-\lambda (\sum x_{(i)} - nt)\right\}, \quad (4.15)$$

which, upon maximization over λ , suggests use of $U = \bar{x} - x_{(1)}$.

4. $T = \prod_{i=1}^n X_i$ (which could be used to investigate the shape of the null distribution). In this case, the distribution of T is not available in closed form, and all computations would need to be performed numerically.

Computation. It is, unfortunately, rather rare to be able to analytically determine $m(t|u)$, so that simulation methods are typically required for computation of the conditional predictive p -value. In general, these simulations will be more difficult than those involved in computation of either the prior predictive p -value or the posterior predictive p -value.

If one can generate from $m(x|u_{obs})$, then computation of p is straightforward; for instance, one could simply estimate p by the fraction of generated random variables, X , for which $T(x)$ exceeds $T(x_{obs})$. Unfortunately, generating from $m(x|u_{obs})$ can itself be difficult. Indeed, it will typically be considerably more convenient to instead generate from $m(x| |u - u_{obs}| < \delta)$ which, for small enough δ , is simply an approximation to $m(x|u_{obs})$. Of course, smaller δ will typically require more computation time. In this regard, note that p -values computed with respect to $m(x| |u - u_{obs}| < \delta)$ are themselves meaningful measures of surprise, even for moderate δ ; indeed, they are conditional predictive p -values, simply being based on a somewhat

different U . As a matter of fact, as $\delta \rightarrow \infty$, it is clear that $m(x| |u - u_{obs}| < \delta) \rightarrow$ the prior predictive, $m(x)$. Thus one need not strive for exceedingly small δ .

If the prior distribution is proper, it will typically be relatively easy to generate from $m(x)$ itself. Doing so, and saving those generated variables for which $|u - u_{obs}| < \delta$, will obviously yield a sample from $m(x| |u - u_{obs}| < \delta)$. Unfortunately, we are primarily interested in cases in which (improper) noninformative priors are used, so it will be necessary to turn to MCMC algorithms.

First, notice that $m(x| |u - u_{obs}| < \delta)$ can be written as

$$m(x| |u - u_{obs}| < \delta) = \frac{\int f(x|\theta)\pi(\theta)1_{\{|u - u_{obs}| < \delta\}}d\theta}{Pr(\{|u - u_{obs}| < \delta\})}, \quad (4.16)$$

where $1_{\{A\}}$ is the indicator function on the set A . The denominator, $Pr(\{|u - u_{obs}| < \delta\})$, is a constant that will be irrelevant in the following MCMC schemes.

Gibbs. A Gibbs sampler based on (4.16) is quite easy to implement:

- *Step 1.* Generate $\theta \sim \pi(\theta|x)$.
- *Step 2.* Generate $X \sim f(x|\theta)1_{\{|u - u_{obs}| < \delta\}}$.
- *Step 3.* After sufficiently many iterations of *Step 1* and *Step 2*, estimate p by the fraction of the generated x 's for which $T(x)$ is greater than $T(x_{obs})$.

Notice that *Step 1* merely calls for generating from the usual posterior distribution of θ , not the typically more difficult posterior conditional on u . The easiest way to implement *Step 2*, for the given θ , is to repeatedly generate x 's until $u(x)$ is within δ of u_{obs} . (Of course, much more efficient schemes may well be available in specific problems).

The chain for this Gibbs sampler is built specifically for computation of the conditional predictive p -value, and hence must be done as a separate computation. In contrast, the following algorithm can be based on the outputs of a typical MCMC Bayesian analysis involving the null model, i.e., using the θ^i generated from $\pi(\theta|x_{obs})$.

Metropolis-Hastings. Generate a chain (x^j, θ^j) as follows. The proposal for a probing (or jumping) distribution is $f(x|\theta)\pi(\theta|x_{obs})1_{\{|u - u_{obs}| < \delta\}}$. Then, from (x^t, θ^t) at time t ,

- *Step 1.* Generate a candidate (x^*, θ^*) from the probing distribution by taking $\theta \sim \pi(\theta|x_{obs})$, simulating $x \sim f(x|\theta)$, and repeating this procedure until $u(x)$ is within δ of u_{obs} . Notice that, if $u(x)$ is not within δ of u_{obs} , a *new* θ has to be generated from $\pi(\theta|x_{obs})$; this was *not* required for the similar step in the Gibbs sampler, and may make the Metropolis-Hastings algorithm more expensive in practice.
- *Step 2.* Accept the candidate with probability

$$\alpha = \min \left\{ 1, \frac{\pi(\theta^*)}{\pi(\theta^*|x_{obs})} \frac{\pi(\theta^t|x_{obs})}{\pi(\theta^t)} \right\} = \min \left\{ 1, \frac{f(x_{obs}|\theta^t)}{f(x_{obs}|\theta^*)} \right\}.$$

- *Step 3.* After sufficiently many iterations of *Step 1* and *Step 2*, estimate p by the fraction of the x^j in the chain for which $T(x^j)$ is greater than $T(x_{obs})$.

The previous schemes can be easily implemented whenever U is explicitly available. When U is not available in closed form but $f(t|\theta)$ is explicitly available, numerical maximization of (4.18) to determine U will often add only modest cost to the above algorithms. The most difficult case is when $f(t|\theta)$ is not available in closed form; the following simple algorithm could then be used, for a given x^* , to compute the required $u = u(x^*)$ and $t^* = T(x^*)$ in the above algorithms:

- *Step 1.* Take a grid of θ values, in an adaptive way if needed.
- *Step 2.* For each θ , generate a sample of random variables $\{x^i\}$ from $f(x|\theta)$ and compute $r(\theta) = f(x^*|\theta)/\hat{f}(t^*|\theta)$, where $\hat{f}(t|\theta)$ is some estimate of the density $f(t|\theta)$. The crudest such estimate is $\hat{f}(t|\theta) = (\# T(x^i) \text{ within } \epsilon \text{ of } t)/(2M\epsilon)$; of course, a more sophisticated kernel estimator could be used.
- *Step 3.* Set u equal to the value of θ that maximizes $r(\theta)$ over the grid.

Notice that we only need those values of u that are within δ of u_{obs} . Thus, once u_{obs} has been (carefully) computed, all that is needed is a grid of θ values that are within δ of u_{obs} , and a rough check that would indicate whether the maximum of $r(\theta)$ occurs in this restricted grid. This last can be accomplished by checking whether or not the maximum occurs on the boundaries of the grid. This can result in a considerable computational simplification.

Connections with classical p-values. Many classical p -values are derived by conditioning on a statistic U for which $f(t|u, \theta) = g(t|u)$, i.e., the conditional distribution of T does not depend on θ . A famous example is the Fisher exact test for independence in a 2×2 contingency table, in which conditioning on the marginal totals of the table yields a parameter-free null distribution. In such a case, it is clear from (4.6) that the conditional predictive distribution exactly equals $g(t|u)$, so the conditional predictive p -value will exactly equal the classical p -value. Thus, from a classical perspective, one could simply view the conditional predictive p -value as a generalization that allows one to also handle situations in which $f(t|u, \theta)$ is not free of θ . And since the conditional predictive p -value can be based on noninformative prior distributions, it has as much claim to 'objectivity' as any classical p -value.

A less obvious advantage (for classical statisticians) of the conditional predictive p -value is that it allows for choice of different departure statistics, $T(X)$. When a statistic, U , can be found such that $g(t|u)$ is free from θ , the choice of T is typically forced upon the classical statistician. (The Fisher exact test is such a situation.) We hope to demonstrate elsewhere that better choice of T (combined with use of the conditional predictive p -value) gives a considerably better tool for model checking.

4.3. The partial posterior predictive p -value

While logically appealing, the conditional predictive p -value with the conditioning statistic U chosen as in (4.12) can be difficult to compute. An attractive alternative is to directly use $f(x|t, \theta)$ (see (4.12)) to integrate out θ , rather than simply using it to define U . This leads to the *partial posterior predictive p -value*, defined for a prior $\pi(\theta)$ (typically noninformative) as

$$p = \Pr^{m^*}(\cdot)(T \geq t_{obs}), \quad (4.17)$$

where m^* and the (assumed proper) *partial posterior* π^* are given by

$$m^*(t) = \int f(t|\theta)\pi^*(\theta)d\theta, \quad \pi^*(\theta) \propto f(x_{obs}|t_{obs}, \theta)\pi(\theta) \propto \frac{f(x_{obs}|\theta)\pi(\theta)}{f(t_{obs}|\theta)}. \quad (4.18)$$

Intuitively, this avoids the double use of the data that occurs in the posterior predictive p -value because the contribution of t_{obs} to the posterior is 'removed' before θ is eliminated by integration. Furthermore, the parallel with (4.12) suggests that the resulting answer will be very similar to the conditional predictive p -value. Indeed, in Example 4 it is easy to see that the answers are *exactly* the same.

To see the comparative simplicity of the partial posterior predictive p -value, we return to the situation of Example 5.

Example 6. Assume exponential data, as in Example 5. For the first three choices of T considered therein, we derive the partial posterior predictive p -value.

1. $T = X_{(n)} - X_{(n-1)}$. Then, from (4.13) and defining s_{obs} to be the sum of the observed x_i , the partial posterior arising from the usual noninformative prior, $\pi(\lambda) = 1/\lambda$, is easily seen to be

$$\pi^*(\lambda) = \frac{\lambda^{n-2} e^{-\lambda(s_{obs} - t_{obs})}}{\Gamma(n-1)(s_{obs} - t_{obs})^{-(n-1)}}.$$

It follows that m^* and the partial posterior predictive p -value are given, respectively, by

$$m^*(t) = \frac{(n-1)}{(s_{obs} - t_{obs})} \left(1 + \frac{t}{(s_{obs} - t_{obs})}\right)^{-n}, \quad p = \left(1 - \frac{t_{obs}}{s_{obs}}\right)^{n-1}.$$

This is eminently sensible. In contrast, the prior predictive p -value can be seen to be $(1 + t_{obs}/s_{obs})^{-n}$, which has the unattractive feature of converging to a nonzero constant as $t_{obs}/s_{obs} \rightarrow 1$, even though such data would unequivocally indicate that the exponential model is wrong. This same unattractive feature can be seen to hold for the plug-in p -value obtained by replacing λ by an estimate in the density of T .

2. $T = X_{(n)}$. Then, from (4.14) and using the noninformative prior $\pi(\lambda) = 1$, the partial posterior is easily seen to be

$$\pi^*(\lambda) = c \lambda^{n-1} (1 - e^{-\lambda t_{obs}})^{(n-1)} e^{-\lambda(s_{obs} - t_{obs})},$$

where c is the normalizing constant. (Interestingly, the usual noninformative prior, $1/\lambda$, cannot be used here as it would result in an improper partial posterior.) Computation yields the partial predictive p -value $p = 1 - c \Gamma(n) [(s_{obs} - t_{obs})^{-n} - s_{obs}^{-n}]$. Because c must be computed numerically, the behavior of p is more difficult to study than in case 1. One interesting situation is when $s_{obs}/t_{obs} \rightarrow 1$, in which case the model is clearly contraindicated by the data. It can be shown that then $p = O((s_{obs}/t_{obs} - 1)^n)$, so that the partial predictive p -value will sensibly discredit the model. In contrast, the posterior predictive p -value and the plug-in p -value converge to nonzero constants in this situation (roughly $n2^{-n}$ and ne^{-n} , respectively), which is not appropriate behavior.

3. $T = X_{(1)}$. Then, from (4.15) and using the usual noninformative prior $\pi(\lambda) = 1/\lambda$, the partial posterior distribution and partial posterior predictive p -value are easily seen to be

$$\pi^*(\lambda) = \frac{\lambda^{n-2} e^{-\lambda(s_{obs} - nt_{obs})}}{\Gamma(n-1)(s_{obs} - nt_{obs})^{-(n-1)}}, \quad p = \left(1 - \frac{nt_{obs}}{s_{obs}}\right)^{n-1}.$$

Here, the posterior predictive p -value and the plug-in p -value are given by $(1 + nt_{obs}/s_{obs})^{-n}$ and $\exp(-n^2 t_{obs}/s_{obs})$, respectively. The interesting situation is when $nt_{obs}/s_{obs} \rightarrow 1$, again being a case in which the model would clearly be contraindicated. The partial posterior predictive p -value reasonably goes to 0 in this situation, while the posterior predictive and plug-in p -values converge to the nonzero constants 2^{-n} and e^{-n} , respectively.

A point deserving of elaboration is the relationship between choice of $T(X)$ and choice of the noninformative prior. One may well need to modify usual noninformative priors so as to ensure that the partial posterior distribution is proper. This can be a serious problem if T is chosen inappropriately. As an extreme example, if $T(X)$ is sufficient for θ , then it is clear from (4.18) that the partial posterior will equal the prior, so that improper noninformative priors

cannot be used at all. Such T are not at all appropriate for model checking, however, and this problem should not be severe for properly chosen T . We expect that Example 6 is more typical; the usual noninformative prior will typically yield a proper partial posterior, but sometimes a different prior will be needed, as in Case 2 of the example. One could reasonably argue that, ideally, 'optimal' noninformative priors should be derived specifically for $f(x|t, \theta)$. In Example 6, for instance, one can easily compute the Jeffreys prior for λ in each of the situations considered. For Case 1 and Case 3, the Jeffreys prior is simply the usual $1/\lambda$. For Case 2, however, the Jeffreys prior can be seen to be

$$\pi(\lambda) = \left(\frac{1}{\lambda^2} - \frac{t_{obs}^2}{2[\cosh(\lambda t_{obs}) - 1]} \right)^{1/2}.$$

While computations with this prior would need to be done numerically, we would recommend its use in actual practice in Case 2.

Computation. Simulation methods will typically be required to compute the partial posterior predictive p -value. In general, these simulations will be somewhat more difficult than those involved in computation of either the prior predictive p -value or the posterior predictive p -value, but considerably easier than those involved in computation of the conditional predictive p -value.

Noting that the partial posterior predictive p -value is $p = \int \Pr(T \geq t_{obs}|\theta) \pi^*(\theta) d\theta$, an obvious strategy is to repeatedly generate θ from $\pi^*(\theta)$ and then T from $f(t|\theta)$ (which could, of course, be done by simply generating X from $f(x|\theta)$ and computing $T(X)$), and then estimate p by the fraction of generated T that are greater than t_{obs} . There are various possibilities for generating from $\pi^*(\theta)$. If generation from the full posterior $\pi(\theta|x_{obs})$ is easy, then a simple Metropolis chain will do the job: use $\pi(\theta|x_{obs})$ as the probing distribution to obtain a candidate θ^* , and then move from the current θ_j to the candidate with probability equal to the minimum of 1 and $f(t_{obs}|\theta_j)/f(t_{obs}|\theta^*)$.

Alternatives to such direct Monte Carlo computation of p include importance sampling schemes. For instance, if a sample $\{\theta_j, j = 1, \dots, m\}$ from $\pi(\theta|x_{obs})$ is available, then one could estimate p by

$$\hat{p} = \frac{\sum_{j=1}^m \Pr(T \geq t_{obs}|\theta_j)/f(t_{obs}|\theta_j)}{\sum_{j=1}^m 1/f(t_{obs}|\theta_j)}.$$

5. CONCLUSIONS

We view it to be reasonable to use p -values (calibrated by $\underline{B}(p) = -ep \log(p)$) to measure surprise in the data, presuming that 'surprise' is used only to suggest that further model development or elaboration should be attempted. (Once an alternative model is at hand, we would urge forgetting about the p -value or its calibration, and performing a real Bayesian model comparison.) The primary purpose of the proposed calibration is to transform the p -value to a Bayes factor or 'odds' scale, which can be more easily interpreted.

When the entertained model contains unknown parameters, we recommend use of the conditional predictive p -value or the partial posterior predictive p -value. We feel that the conditional predictive p -value is more truly Bayesian in nature than other predictive p -values, in that it is based on a natural Bayesian quantity, the predictive distribution, while employing the natural concept of conditioning to eliminate potential difficulties. We feel that the partial posterior predictive p -value is also sensible, avoiding a double use of the data and often being very similar to the conditional predictive p -value. Indeed, we are tempted to recommend the partial posterior predictive p -value for general use, in that it is usually considerably simpler computationally than the conditional predictive p -value. We hesitate to do so, however, because its motivation

is rather adhoc, and we prefer measures with a natural Bayesian interpretation. Except for the computational issue, implementation of both procedures is straightforward, requiring only specification of a departure statistic, T , and noninformative priors for the parameters.

Our perspective may appear to be an odd combination of Bayesian and non-Bayesian notions, but we would argue that this is only partly the case. We are presuming that one has available only the null model (typically with noninformative priors for the unknown parameters) and a statistic T to indicate possible departures from the null model. Given this premise, our goal was to derive measures of 'surprise' that would be as reasonable from the Bayesian perspective as possible. (Incidentally, the resulting measures seem to also be very reasonable from the classical perspective.) While some Bayesians may choose not to operate under this premise, many view these constraints as the common practical reality for model checking.

ACKNOWLEDGEMENTS

This work was supported in part by the NFS of USA under grants DMS-9303556 and DMS-9802261, and by the MEC of Spain under grant PB96-0776.

REFERENCES

- Aitkin, M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. B* 53, 111–142 (with discussion).
- Bayarri, M. J. (1986). A Bayesian goodness of fit test. *Proceedings of the 1985 Joint Statistical Meetings*. Washington, D.C.: ASA.
- Bayarri, M. J. and Berger, J. (1997). Measures of surprise in Bayesian analysis. *Tech. Rep.* 97–46, Duke University.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* 2nd Ed. New York: Springer-Verlag.
- Berger, J. (1994). An overview of robust Bayesian analysis. *Test* 3, 5–124 (with discussion).
- Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* 2, 317–352 (with discussion).
- Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *J. Amer. Statist. Assoc.* 82, 112–122.
- Box, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *J. Roy. Statist. Soc. A* 143, 383–430.
- Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* 82, 106–111 (with discussion).
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Delampady, M. and Berger, J.O. (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Ann. Statist.* 18, 1295–1316.
- Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70, 193–242.
- Efron, B. (1996). Talk at the Conference on *Statistical Challenges in Astronomy*, Penn State University, 1996.
- Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics* 26, 1125–1143.
- Gelman, A., Meng, X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733–807 (with discussion).
- Goldstein, M. (1991). Comment on 'Posterior Bayes factors' (by M. Aitkin). *J. Roy. Statist. Soc. B* 53, 134.
- Good, I. J. (1956). The surprise index for the multivariate normal distribution. *Ann. Math. Statist.* 27, 1130–1135.
- Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press.
- Good, I. J. (1988). Surprise index. In *Encyclopedia of Statistical Sciences* (S. Kotz, N. L. Johnson, and C. B. Reid, eds.) 7, 104–109.
- Gutman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. B* 29, 83–100.
- Hodges, J. (1992). Who knows what alternative lurks in the heart of significance tests? *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 247–266 (with discussion).

- Meng, X. L. (1994). Posterior predictive p -values. *Ann. Statist.* 22, 1142–1160.
- Roberts, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.* 60, 50–62.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* 12, 1151–1172.
- Selke, T., Bayarri, M. J. and Berger, J. (1999). Calibration of p -values for precise null hypotheses. *Tech. Rep.*, Duke University.
- Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness of fit testing using infinite dimensional exponential families. *Ann. Statist.* 26, 1215–1241.
- Weaver, W. (1948). Probability, rarity, interest and surprise. *Scientific Monthly* 67, 390–392.

DISCUSSION

JAMES M. ROBINS (*Harvard School of Public Health, USA*)

Berger and Bayarri (B&B) propose two new “Bayesian” p -values, the conditional predictive p -value and the partial posterior predictive p -value. They claim that, when one’s goal is to check the adequacy of a parametric model $f(x; \theta)$, these new p -values are superior to previously proposed “Bayesian” p -values – the prior predictive p -value of Box (1980), the posterior predictive p -value of Guttman (1967) and Rubin (1984), and the discrepancy p -value of Meng (1984) and Gelman et al. (1995). In my Valencia discussion, I disagreed with B&B’s claim. I was wrong. This note is intended as both apology and correction. B&B’s paper raises the following questions, which I consider below. 1). Are “Bayesian” p -values true p -values? 2). What good is a p -value that is not a p -value?

Definition: A candidate or potential p -value is a random variable “ p ” (X) with range $[0, 1]$.

Definition 1: If a potential p -value “ p ” (X) has a uniform $U[0, 1]$ distribution, we say “ p ” (X) is a p -value $p(X)$. More specifically, we say that “ p ” (X) is, respectively, a frequentist or prior predictive p -value if Definition 1 holds when $X \sim f(x; \theta^*)$ or $X \sim \pi(x) = \int f(x; \theta) \pi(\theta) d\theta$ where $\pi(\theta)$ is the prior and θ^* is the true value of θ .

In practice, we use small values of “ p ” (x_{obs}) to denote surprise, because we act as if “ p ” (X) is $U[0, 1]$ under the model. Here x_{obs} represents the observed value of X . Seriously conservative p -values (e.g., “ p ” (X) = 1/2 almost surely) will cause us never to be surprised, and thus we will fail to reject false models. Seriously anti-conservative potential p -values (e.g., “ p ” (X) = 0 almost surely) will cause us to reject the null model $f(x; \theta)$ even when true. Thus, potential p -values that are not p -values can be misleading and should be avoided. In general, we say a potential p -value (i) is conservative if $\Pr\{“p” (X) < t\} < t$ and (ii) is anti-conservative if $\Pr\{“p” (X) < t\} > t$ for all $t < 1/2$. If our goal is to check the model $f(x; \theta)$ rather than the prior $\pi(\theta)$, our procedures should perform adequately whatever the prior, including point-mass priors. This implies that we should require p -values to be frequentist p -values. Unfortunately, this requirement is generally unfulfillable. However, since the data dominate the prior as the sample size $n \rightarrow \infty$, the distribution of “ p ” (X) as $n \rightarrow \infty$ reflects only the model $f(x; \theta)$ and not the prior. Thus a potential natural, fulfillable, requirement on a Bayesian p -value is that it be an asymptotic frequentist p -value.

Definition 2: “ p ” (X) is an asymptotic frequentist p -value if its distribution converges to $U(0, 1)$ under $f(x; \theta^*)$.

We will thus determine which of various candidate p -values are asymptotic frequentist p -values. In general, “ p ” (X) depends on a test statistic $t(X)$, the null-model $H_0: f(x; \theta)$, and a reference density $m(x)$. Specifically, “ p ” (x_{obs}) = $\Pr^m[t(X) > t_{obs}]$ where $t_{obs} = t(x_{obs})$. That is, we compute, often by simulation, the probability that $t(X)$ exceeds the observed test

statistic t_{obs} when $X \sim m(x)$. We shall consider the following candidate p -values. The posterior predictive p -value uses $m(x) = \pi(x | x_{obs})$ where

$$\pi(x | x_{obs}) = \int f(x; \theta) \pi(\theta | x_{obs}) d\theta;$$

the bootstrap p -value uses $m(x) = f(x; \hat{\theta}_{MLE})$ where $\hat{\theta}_{MLE}$ is the maximizer of $f(x_{obs}; \theta)$; the partial posterior predictive p -value uses $m(x) = \int f(x; \theta) \pi^*(\theta) d\theta$ where

$$\pi^*(\theta) \propto f(x_{obs} | t_{obs}; \theta) \pi(\theta);$$

the conditional predictive p -value uses $m(x) = \int f(x | u_{obs}; \theta) \pi(\theta | u_{obs}) d\theta$, where $\pi(\theta | u_{obs}) \propto f(u_{obs}; \theta) \pi(\theta)$, and

$$u_{obs} \equiv \hat{\theta}_{cMLE} = \arg \max_{\theta} f(x_{obs} | t_{obs}; \theta)$$

is the conditional MLE of θ given t_{obs} ; the conditional bootstrap p -value uses

$$m(x) = f(x; \hat{\theta}_{cMLE}).$$

To study the large sample frequency properties of our candidate p -values, we consider the following canonical set-up. At sample size n , $X \equiv X_n = (Z_1, \dots, Z_n)$ is n iid copies of a random variable Z that follows a regular parametric model $f(z; \psi_n, \theta)$ with true parameter values ψ_n^* and θ^* . We shall consider the distribution of " p " (X) under the null hypothesis that $\psi_n^* = 0$ for each n and under uni-dimensional local alternatives with $\psi_n^* = k/\sqrt{n}$ for a constant k . We restrict attention to univariate test statistics $t(X)$ that are asymptotically linear with asymptotic mean $\nu(\theta)$ under the null. That is, for $Z \sim f(z; 0, \theta)$,

$$n^{1/2} [t(X) - \nu(\theta)] = n^{-1/2} \sum_{i=1}^n B_i(\theta) + o_p(1)$$

for some $B(\theta) = b(Z; \theta)$ with mean zero and finite variance and $o_p(1)$ indicates a random variable converging to zero. Thus, by the Central Limit Theorem and Slutsky's theorem, $n^{1/2} [t(X) - \nu(\theta)]$ is asymptotically $N(0, E[B(\theta)^2])$. We consider two types of test statistics.

Type 1: $t(X) = \sum_i d(Z_i) / n$ for some $D = d(Z)$. Then $\nu(\theta) = E_{0,\theta}[D] \equiv \int d(z) f(z; 0, \theta) dz$ and $B(\theta) = d(Z) - \nu(\theta)$.

Type 2: $t(X) = \sum_i d(Z_i, \hat{\theta}) / n$ where $\hat{\theta}$ is either the MLE of θ or its posterior mean computed under the null model $f(z; 0, \theta)$, and $d(Z, \theta)$ is a function chosen by the analyst satisfying $E_{0,\theta}[d(Z, \theta)] = 0$ for all θ . Then $\nu(\theta) = 0$, $B(\theta) = D_{resid}(\theta)$, where

$$D_{resid}(\theta) \equiv d_{resid}(Z, \theta) \equiv d(Z, \theta) - E_{0,\theta}[d(Z, \theta) S_{\theta}^2(\theta)] \{E_{0,\theta}[S_{\theta}^2(\theta)]\}^{-1} S_{\theta}(\theta)$$

is the residual from the population linear regression of $d(Z, \theta)$ on $S_{\theta}(\theta)$, and where $S_{\theta}(\theta) = \partial \log f(Z; 0, \theta) / \partial \theta$ is the score for θ . Note for $t(X)$ of Type 2, $\hat{\theta}$ must be recomputed for each simulated data set X . A natural choice for $d(Z, \theta)$ would be the score for ψ , $S_{\psi}(\theta) = \partial \log f(Z; 0, \theta) / \partial \psi$, to insure power against local alternatives $\psi^* = k/\sqrt{n}$.

Quantifying Surprise in the Data

Under regularity conditions, $\partial \nu(\theta) / \partial \theta' = E_{0,\theta} [B(\theta) S_\theta(\theta)']$. Note for Type 1 statistics, in general, $\partial \nu(\theta) / \partial \theta' \neq 0$ while for Type 2 statistics, $\partial \nu(\theta) / \partial \theta' = 0$.

Example 1: Suppose $Z = (Y, V_1, V_2)$ and $Y = \psi V_1 + \theta V_2 + \varepsilon$, $\varepsilon \sim N(0, 1)$, $E(V_1) = E(V_2) = 0$, $E(V_1^2) = E(V_2^2) = 1$, $E[V_1 V_2] = \rho$, and $\theta^* = 0$. Then $S_\theta(\theta) = \varepsilon(\theta) V_2$ and $S_\psi(\theta) = \varepsilon(\theta) V_1$ where $\varepsilon(\theta) = Y - \theta V_2$. We shall consider the type 1 statistic, $T_1 = \sum_i D_i / n$ with $D = Y V_1$. Then $\nu_1(\theta) = \theta \rho$, $B_1(\theta) = Y V_1 - \nu_1(\theta)$, and $\partial \nu_1(\theta) / \partial \theta = E_{0,\theta} [B_1(\theta) S_\theta(\theta)] = \rho$. We also consider the type 2 statistic, $T_2 = \sum_i d(Z_i, \hat{\theta}) / n$ with $d(Z, \theta) = S_\psi(\theta)$; then $B_2(\theta) = \varepsilon(\theta) (V_1 - \rho V_2)$ and $\partial \nu_2(\theta) / \partial \theta = E_{0,\theta} [B_2(\theta) S_\theta(\theta)] = 0$.

Let $B \equiv B(\theta^*)$, $S_\theta \equiv S_\theta(\theta^*)$, $S_\psi = S_\psi(\theta^*)$, and $E[h(Z)] \equiv E_{0,\theta^*} [h(Z)]$ for any $h(Z)$.

Theorem 1. (Robins, 1998): Subject to regularity conditions, under law $f(z; k/\sqrt{n}, \theta^*)$, for each candidate p -value, " p " $(X) = 1 - \Phi(Q) + o_p(1)$ where $\Phi(t)$ is the $N(0, 1)$ CDF and $Q = q(X) \sim N(\mu, \sigma^2)$. Thus a candidate p -value is an asymptotic frequentist p -value if and only if $\mu = 0$ and $\sigma^2 = 1$ when $k = 0$. If, when $k = 0$, $\mu = 0$ and $\sigma^2 < 1$, the p -value is conservative. We now give σ^2 and μ for our candidates. Let $E(BS'_\theta) = \tau$, $\Omega = \tau [\text{var}(S_\theta)]^{-1} \tau'$ and $\Omega_c = \tau [\text{var}(S_{c\theta})]^{-1} \tau'$ where $S_{c\theta} = S_\theta - E(S_\theta B) \{\text{var}(B)\}^{-1} B$ is the approximate conditional score for θ , i.e., $n^{-1/2} \partial \log f[X | t(X); 0, \theta^*] / \partial \theta = n^{-1/2} \sum_i S_{c\theta,i} + o_p(1)$. Let $\omega = E(BS_\psi) - \tau \{\text{var}(S_\theta)\}^{-1} E(S_\theta S_\psi)$ and $\omega_c = E(BS_\psi) - \tau \{\text{var}(S_{c\theta})\}^{-1} E(S_{c\theta} S_\psi)$.

Bootstrap (boot): $\sigma_{boot}^2 = [\text{var}(B) - \Omega] / \text{var}(B)$, $\mu_{boot} = k\omega / \text{var}(B)^{1/2}$.

Posterior Predictive (pp): $\sigma_{pp}^2 = [\text{var}(B) - \Omega] / [\text{var}(B) + \Omega]$, $\mu_{pp} = k\omega / [\text{var}(B) + \Omega]^{1/2}$.

Partial Posterior Predictive (ppp): $\sigma_{ppp}^2 = 1$, $\mu_{ppp} = k\omega_c / [\text{var}(B) + \Omega_c]^{1/2}$.

Conditional Predictive (cp): $\sigma_{cp}^2 = \sigma_{ppp}^2$, $\mu_{cp} = \mu_{ppp}$.

Conditional Bootstrap (cboot): $\sigma_{cboot}^2 = [\text{var}(B) + \Omega_c] / \text{var}(B)$, $\mu_{cboot} = k\omega_c / \text{var}(B)^{1/2}$.

Interpretation of Theorem 1: Case 1: $E(BS'_\theta) = 0$. If, as for type 2 statistics, $E(BS'_\theta) = 0$, all five candidate p -values are asymptotic frequentist p -values. Further, for all candidates, if $d(Z, \theta) = S_\psi(\theta)$, the non-centrality parameter μ is $k\{E[S_{\psi, \text{resid}}^2]\}^{1/2}$ which is optimal against local alternatives $\psi_n = k/\sqrt{n}$. In contrast to this large sample result, Example 6 of B&B suggests that pp and bootstrap p -values as opposed to cp and ppp p -values are conservative with poor power in small samples.

Case 2: $E(BS'_\theta) \neq 0$. The cp and ppp p -values are asymptotically equivalent. If, as for most type one statistics, $E(BS'_\theta) \neq 0$, the bootstrap and pp p -values are conservative with the bootstrap the less conservative. Indeed, in Example 1, both p -values are converging to a point mass at 1/2 as the correlation $\rho = E[V_1 V_2]$ approaches 1. In contrast, the cp and ppp p -values are asymptotic frequentist p -values. Furthermore, they are locally optimal [i.e., $\mu = k\{E[S_{\psi, \text{resid}}^2]\}^{1/2}$] when $d(Z) = S_\psi$. It follows that statistic T_1 of Example 1 is locally optimal when $\theta^* = 0$ so $Y V_1 = S_\psi$. Consider the nominal one-sided α -level test that rejects whenever " p " (X) is less than α . Then, in Example 1, the ratio of the local power under alternatives $\psi^* = k/\sqrt{n}$ of the nominal bootstrap and pp tests to that of the ppp and cp tests goes to zero as the correlation ρ approaches 1. In contrast to the conservative bootstrap and pp p -values, the cboot p -value is anti-conservative. In particular, as $\rho \rightarrow 1$ in Example 1, the nominal one-sided α -level test will have actual level converging to 1/2.

In the next theorem, we follow the suggestion of B&B and examine the distribution of " p_{cp} " (X) under the conditional law $f(x | u_{obs}; k/\sqrt{n}, \theta^*)$ rather than under the unconditional law $f(x; k/\sqrt{n}, \theta^*)$ of Theorem 1.

Theorem 2: Under $f(x | u_{obs}; k/\sqrt{n}, \theta^*)$, " p_{cp} " (X) = $1 - \Phi(Q) + o_p(1)$ where $Q \sim N(\mu, \sigma^2)$, $\sigma^2 = \text{var}(B) / (\text{var}(B) + \Omega_c)$, $\mu = kE(BS_\psi) + \tau n^{\frac{1}{2}}(\mu_{obs} - \theta^*)$.

Thus, the cp p -value will not be $U[0, 1]$ under $f(x | u_{obs}; 0, \theta^*)$ but can be either conservative or anti-conservative depending on u_{obs} and θ^* . However, as noted by B&B, it is an exact (non-asymptotic) p -value under $\pi(x | u_{obs}) = \int f(x | u_{obs}; 0, \theta) \pi(\theta | u_{obs}) d\theta$. Further, by Theorem 1, it is an asymptotic frequentist p -value under $f(x; 0, \theta^*)$.

Meng (1994) and Gelman et al. (1995) proposed a discrepancy p -value based on the discrepancy $t(X, \theta) = \sum_i d(Z_i, \theta) / n$ rather than on $t(X) = \sum_i d(Z_i, \hat{\theta}) / n$ to avoid having to recompute the MLE or posterior mean $\hat{\theta}$ for each simulated data set. The discrepancy p -value is " p_d " (x_{obs}) = $pr^m [t(X, \theta) > t(x_{obs}, \theta)]$ with $m = m(\theta, x) = \pi(x, \theta | x_{obs}) = f(x; \theta) \pi(\theta | x_{obs})$. The following theorem of Robins (1998) implies that the discrepancy p -value can be quite conservative even though $E_{0, \theta} [d(Z, \theta)] = 0$. Also see Meng (1994).

Theorem 3: Suppose $E_{0, \theta} [d(Z, \theta)] = 0$. Under $f(z; k/\sqrt{n}, \theta^*)$, the discrepancy p -value " p_d " (X) = $1 - \Phi(Q_d) + o_p(1)$, where $Q_d \sim N(\mu_d, \sigma_d^2)$ with $\sigma_d^2 = \sigma_{pp}^2$ and $\mu_d = \mu_{pp}$ with both σ_{pp}^2 and μ_{pp} evaluated at $B \equiv d(Z, \theta^*)$. In particular, in Example 1, if $d(Z, \theta) = S_\psi(\theta) = \varepsilon(\theta) V_1$ then " p_d " (X) converges to a point mass at $1/2$ as the correlation $\rho \rightarrow 1$.

Example 2: The "skewness" discrepancy example of Gelman et al. (1995, p. 172) is $t(X, \theta) = |(\hat{\mu}_0 - \theta)| - |(\theta - \hat{\mu}_1)|$ where $\hat{\mu}_\alpha$ is the α quantile of the empirical distribution of $X = (Z_1, \dots, Z_n)$, and the Z_i are iid $N(\theta, 1)$. Robins (1998) proves that under the $N(\theta, 1)$ null model, " p_d " (X) = $1 - \Phi(Q_d) + o_p(1)$ where $Q_d \sim N(0, \sigma_d^2)$ with $\sigma_d^2 = \{.1 - \pi^{-1} \exp[-z_1^2]\} / \{.1 + \pi^{-1} \exp[-z_1^2]\} = .238$ where z_1 is the .1 quantile of a $N(0, 1)$ distribution. Therefore, the actual asymptotic level of a nominal one-sided .05 level test using " p_d " (X) is $pr[Q_d > 1.64] = .0004$.

One might hope that one could "fix" the conservativeness of " p_d " (X) by using the partial discrepancy p -value " p_{pd} " (X) that uses

$$m(\theta, x) = f(x; \theta) \pi^*(\theta), \quad \pi^*(\theta) \propto f(x_{obs} | t(x_{obs}, \theta); \theta) \pi(\theta).$$

Unfortunately, " p_{pd} " (X) has the same asymptotic distribution as " p_d " (X). However, " p_d " (X) can be made an asymptotic frequentist p -value by replacing, in $t(X, \theta)$, $d(Z, \theta)$ by $d_{resid}(Z, \theta)$, since $E(BS'_\theta) = 0$ with $B = d_{resid}(Z, \theta)$. Indeed, with $d(Z, \theta) = S_\psi(\theta)$, " p_d " (X) based on $S_{\psi, resid}(\theta)$ is locally optimal. The down side is that, for some models, the integrals $E_{0, \theta} [d(Z, \theta) S'_\theta(\theta)]$ and $E_{0, \theta} [S_\theta^{\otimes 2}(\theta)]$ in $d_{resid}(Z, \theta)$ would need to be evaluated numerically.

DONNA K. PAULER (Harvard School of Public Health, USA)

This paper highlights some of the limitations of previous approaches for measuring surprise or assessing goodness of fit via p -values (e.g., Dempster (1971), Box (1980), Rubin (1984)), and proposes two improvements, the *conditional predictive p-value* (CP) and *partial posterior predictive p-value* (PPP). The interesting debate over appropriateness of these frequency calculations for Bayesian inference will not be taken up here. The use of *ad hoc* diagnostic checks provides a vital ingredient in any model fitting process and the methods proposed in this paper provide a potential contribution. With this practical view of model checking in mind, my discussion focuses on three issues: calibration, computation, and the extension to more complex models.

1. *Calibration.* The p -values discussed in this paper are subject to the usual criticisms of classical p -values such as lack of direct interpretation, overestimation of evidence against the null hypothesis, and uncertainties in how to combine results from multiple comparisons. The proposed calibration via $\underline{B}(p)$ only addresses the first of these issues. Therefore, I would more strongly emphasize the authors' warning that *neither* p nor $\underline{B}(p)$ be interpreted as formal measures of surprise but rather regarded with the same degree of suspicion normally assigned p -values.

2. *Computation.* As discussed in Section 4.2, computation of the CP or PPP usually requires Monte Carlo sampling beyond the machinery already in place to fit the model. In the context of a simple example, I will highlight some points deserving caution in the proposed algorithms and offer suggestions for reducing the computational burden. Suppose under H_0 , X_1, X_2, \dots, X_n are independent and follow an exponential distribution with rate λ and prior specification, $\lambda \sim \pi(\lambda) \propto 1/\lambda$. It is often of interest to check for overdispersion, which can occur, for example, in medical applications when the failure rate varies randomly between individuals. A sensible test statistic to measure for overdispersion is the difference between maximal and minimal order statistics, $T = X_{(n)} - X_{(1)}$.

Operationally, to compute the CP one samples from $m(t||u - u_{obs}| < \delta)$, where $U = U(x)$ is given by (4.12), and δ is chosen as small as computationally possible. The authors suggest that p -values for moderate δ are of interest in their own right too. A word of caution is in order here. The following table shows the CP calculated for 20 observations simulated from a Pareto (overdispersed exponential) distribution using the Gibbs and Metropolis algorithms in Section 4.2 for various choices of δ . Conclusions drawn from this table differ dramatically depending on δ , emphasizing the importance of choosing δ to appropriately match the *scale* of U .

Table 1. The CP with U given by (4.12) for 20 observations simulated from a Pareto distribution, calculated using Gibbs and Metropolis Hastings algorithms, for different choices of δ .

δ	2	1	.1	.05	.01
CP (Gibbs)	.085	.036	.036	.033	.028
CP (Met.)	.120	.045	.025	.039	.026

This simple example requires a one-dimensional numerical maximization of (4.12) for every replicated value of T , anticipating the computational burden for multi-parameter models. Besag and Clifford (1991) suggest sequential methods for reducing the costs of Monte Carlo tests. The general idea is that sampling should continue until the standard error of the estimate of the p -value falls within some constant fraction of p . Large p -values can be identified immediately and accuracy beyond the customary report $p > .10$ is not required. Alternatively, a short initial Gibbs run can be used to determine the number of samples required to achieve a prespecified accuracy. In the example of Table 1, to obtain a standard error within .25 of p , an initial Gibbs run showed that $N = 1000$ iterations were necessary but for data simulated with a p -value equal to .85, only $N = 5$.

The computational expense of the CP with U given by (4.12) raises the question of whether more *ad hoc* but feasible methods similar to the CP are preferable. Besag and Clifford (1989) give a number of examples in the spatial statistics literature where conditioning on sufficient statistics for the nuisance parameter leads to manageable hypothesis tests. For the data analyzed in Table 1, conditioning on the sufficient statistic $U = 1/\bar{X}$ only requires sampling from a uniform density on $\{x : U(x) = u_{obs}\}$ and yields the p -value, .025, which is very similar to the CP.

Two other easily calculated measures discussed in this paper, the PPP and posterior predictive p -value, give values .031 and .094, respectively.

3. *Extensions to Complex Models.* Presumably, the methods proposed in this paper are to be most beneficial in more complex modelling scenarios, where significance testing becomes analytically intractable, and a large number of local checks are required. For example consider the following two-stage model:

$$\begin{aligned} y_{ij} | \theta_i, \sigma^2 &\sim p_1(g_1(x_{ij}; \theta_i), \sigma^2) \\ \theta_i | \beta, D &\sim p_2(g_2(z_i; \beta), D) \end{aligned}$$

for $j = 1, \dots, n_i, i = 1, \dots, I$ where p_1 and p_2 are probability distributions, g_1 and g_2 are functions, at the first and second stage, x_{ij} and z_i are vectors of covariates, θ_i is a vector of random effects, β is a vector of fixed effects, and σ^2 and D are a variance and covariance matrix, respectively. Here there are a number of places where the model can be inadequate including nonconstant variance at the first stage and inappropriate distributional assumptions or mean functions at the first or second stages. What test statistics can Bayesians construct and are the methods in this paper appropriate or feasible?

To answer these questions we begin by defining residuals. Let $\theta = (\theta_1, \dots, \theta_n)$ and $\gamma = (\beta, \sigma^2, D)$. Individual first stage residuals and multivariate second stage residuals are given by $\epsilon_{ij}(\theta_i, \gamma) = (y_{ij} - g_1(x_{ij}; \theta_i))/\sigma$ and $R_i(\theta_i, \gamma) = D^{-1}(\theta_i - g_2(z_i; \beta))$, respectively (Weiss (1996), Hodges (1998)). Formal single discrepancy measures can be defined by counting the number of residuals greater than some prescribed value, $T_1(\theta, \gamma) = \sum_{i=1}^I \sum_{j=1}^{n_i} 1[|\epsilon_{ij}(\theta, \gamma)| > \tau]$, where $1[\cdot]$ denotes the indicator function, or by measuring their total squared size, $T_2(\theta, \gamma) = \sum_{i=1}^I R_i'(\theta_i, \gamma) R_i(\theta_i, \gamma)$. Weiss (1996) studies the posterior distribution of statistics similar to T_1 and T_2 and Gelman *et al.* (1996), the posterior predictive distribution. For the methods proposed in this paper, test statistics need to be observable. This can be accomplished by evaluating T_1 and T_2 at the maximum likelihood estimates or posterior modes of the unknown parameters, or averaging over all draws from the posterior distribution; see also Davison (1998).

I would like to point out just how quickly the CP with U given by (4.12) and the PPP become computationally challenging in this case. Both the CP and PPP require an expression for $f(t|\gamma)$, the density of T_1 or T_2 , which must be estimated by a kernel density estimate over a grid of γ values. The dimension of γ can be large, depending on the number of covariates and random effects in the model, and also contains constrained parameters. Similarly, maximization of (4.12) over this high-dimensional constrained space at every replication soon becomes infeasible. To avoid these problems it is tempting to choose a convenient U , as proposed by Besag and Clifford (1989). An obvious choice of u for T_1 or T_2 is a posterior mean or modal estimate of γ . This choice still requires fitting the entire model at each replication, necessitating some of the sequential Monte Carlo methods discussed in the previous section.

JOSÉ M. BERNARDO (*Universitat de València, Spain*)

The notion of testing whether or not data are compatible with a given model without specifying any alternative is indeed very attractive but, unfortunately, it seems to be beyond reach. In this paper, a statistic $T(X)$ is to be chosen "to investigate compatibility with the observed data $x_{obs} = \{x_1, \dots, x_n\}$, with a hypothesis $H_0 : X \sim f(x|\theta)$, with large values indicating less compatibility". It seems to me that the choice of that statistic is precisely a disguised introduction of an alternative: if, say $|\bar{x}|$ is used to measure compatibility, one is obviously testing a location at the origin against other possible locations.

If it is conceded that the choice of a universal 'default' statistic to measure compatibility is simply not possible, I believe it follows that there is a definite advantage in being explicit

about the alternatives one has in mind. Indeed, (i) this attitude has a solid decision-theoretical foundation, (ii) it makes it possible to provide a constructive definition of the compatibility statistic, and (iii) if the original model is judged not to be compatible with the observed data, then one has an immediate hint about a possible modification of that model that would increase compatibility.

If the envisaged alternatives are structured as an encompassing model $f_e(x|\theta, \omega)$ which generalizes H_0 , then the BRC statistic introduced in this conference (Bernardo, 1998), *i.e.*, the expected posterior of the logarithmic discrepancy,

$$d_r(x_{obs}) = \int \delta(\theta) \pi(\theta | x_{obs}) d\theta, \quad \delta(\theta) = n \inf_{\omega \in \Omega} \int f_e(x|\theta, \omega) \log \frac{f_e(x|\theta, \omega)}{f(x|\theta)} dx,$$

provides an statistic which is specifically tailored to check the compatibility of x_{obs} with $f(x|\theta)$ against the alternatives encapsulated in $f_e(x|\theta, \omega)$.

An stylised decision-theoretical analysis of this problem suggests that d_r values about 2.5 should be considered 'mildly surprising', while d_r values larger than 5 should be regarded as 'very surprising'. It could be interesting to the readers of this very interesting paper to compare this analysis with the calibrated p -value that Bayarri and Berger have suggested. In the canonical Example 1 (where the alternatives have been specified), $d_r = 2.4207$ corresponds to a p -value 0.05, calibrated to $-ep \log p = 0.4071$, so that you should only bet about 2.5 to 1 against H_0 (mild evidence against H_0), while $d_r = 5$ corresponds to a p -value 0.0027, calibrated to $-ep \log p = 0.0434$ so that you should be prepared to bet at least about 23 to 1 against H_0 (strong evidence against H_0). This suggests a basic agreement between BRC and calibrated p -values in this canonical example and, hence, in any one-dimensional regular problem with sufficiently large samples.

The implications of both approaches will however be different for small samples when the sampling distribution of the test statistic depends on the sample size and, therefore, there is no longer a one-to-one relation between d_r and the p -value: in continuous regular problems, a fixed p -value will correspond to smaller d_r values as the sample size decreases, suggesting that the 'surprise' against the null indicated by a fixed p -value should decrease with the sample size. Would the authors claim that the calibrated p -value should be a one-to-one function of p or, as suggested above, the 'surprise' associated to a fixed p -value should rather decrease with n ? Similarly, should not a 'proper' calibration of a p -value also depend on the dimensionality of the problem?

BRADLEY P. CARLIN (*University of Minnesota, USA*)

First, congratulations to the authors on an intriguing paper in a research area that offers a perennial challenge for Bayesians. I simply wish to remark that, when calculating Bayesian p -values, another possible approach intermediate to the use of the prior and posterior predictives,

$$m(x) = \int f(x|\theta) \pi(\theta) d\theta \quad \text{and} \quad m(x|x_{obs}) = \int f(x|\theta) \pi(\theta|x_{obs}) d\theta,$$

would be the *cross-validatory* predictive, as described for example in Gelfand, Dey, and Chang (1992). The components of this distribution are

$$m(x_i|x_{(i)}) = \int f(x_i|\theta) \pi(\theta|x_{(i)}) d\theta,$$

where $x_{(i)} \equiv x \setminus x_i$ is the dataset with the i^{th} point deleted, so there is no "double use" of the observed x_i when its predictive is computed. Notice that, unlike the prior predictive $m(x)$,

$m(x_i|x_{(i)})$ will be proper if $\pi(\theta|x_{(i)})$ is. In addition, the collection of conditional predictive densities $\{m(x_i|x_{(i)}), i = 1, \dots, n\}$ is equivalent to $m(x)$ when both exist (Besag, 1974), encouraging the use of the former even when the latter is undefined.

Such a cross-validatory approach can be carried out more or less automatically in any given model setting, thus avoiding the need to specify an appropriate $U(x)$ statistic. Also, computation with $m(x_i|x_{(i)})$ is straightforward using Monte Carlo methods. For example, draws $\{x_i^{(g)}, g = 1, \dots, G\}$ can be made from $m(x_i|x_{(i)})$ for $i = 1, \dots, n$, possibly with the help of importance sampling adjustments so separate MC runs need not be obtained for each of the n "leave-one-out" posteriors. This in turn produces $\{T(x^{(g)}), g = 1, \dots, G\}$ values, which can be histogrammed and compared to t_{obs} for p -value computation.

JULIÁN DE LA HORRA (*Universidad Autónoma de Madrid, Spain*)

I would like to add two comments to this very interesting paper.

1. It is important to know the asymptotic behaviour of any alternative to the classical p -value. For instance, it has been proved in De la Horra and Rodríguez-Bernal (1997) that the posterior predictive p -value is asymptotically distributed as a uniform over the interval $(0, 1)$, when the null hypothesis is true.

2. A calibration of the p -value can be used as an estimate of the probability of the model, given the data. This use was proposed in De la Horra and Rodríguez-Bernal (1999) in the following way:

Let X be a random sample obtained, perhaps, from a density of the model $M = \{f(x|\theta) : \theta \in \Theta\}$. If the model M is true, we want to test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. So, we are faced two problems:

- a) The problem of accepting or rejecting the model M ;
- b) If M is accepted, the problem of accepting or rejecting the null hypothesis $H_0 : \theta \in \Theta_0$.

These two problems can be simultaneously analyzed in a setting of decision theory, as it is next described:

The new parameter space is $\Omega = \{M^c, (M, \Theta_0), (M, \Theta_0^c)\}$, where $M^c =$ "The model M is false", $(M, \Theta_0) =$ "The model M is true and $\theta \in \Theta_0$ " and $(M, \Theta_0^c) =$ "The model M is true and $\theta \in \Theta_0^c$ ".

The action space is $\mathcal{A} = \{a_1, a_2, a_3\}$, where $a_1 =$ "Reject M ", $a_2 =$ "Accept M and H_0 " and $a_3 =$ "Accept M and H_1 ".

If the loss function is given by a suitable matrix, the Bayes action is that minimizing the posterior expected loss; for doing that, we elicit the prior probabilities $P(M)$ and $P(M^c)$, and the prior density $\pi(\theta)$. The problem arises when we try to compute posterior probabilities; for example,

$$P(M^c|x) = \frac{P(M^c)f(x|M^c)}{P(M^c)f(x|M^c) + P(M) \int_{\Theta} \pi(\theta)f(x|\theta)d\theta}$$

and $f(x|M^c)$ is impossible to know (in general).

Denoting by $p(x)$ the posterior predictive p -value, we used $p(x)^{1/3}$ as calibration function for estimating $P(M|x)$, obtaining reasonable results in some examples. Perhaps, the calibration function $-ep(x) \log p(x)$, for $p(x) \leq 1/e$, can give better results.

MICHAEL EVANS (*University of Toronto, Canada*)

The use of data splitting to avoid overly conservative inferences in model checking procedures with surprise is introduced in Evans (1997) where it is shown to work well in several examples. While Evans (1997) introduced the use of a general split into functions (r, s) , corresponding to the notation (T, U) of this paper, where r and s were required to be statistically

independent, the applications were restricted to situations where the data $x = (x_1, \dots, x_n)$ were split according to $s(x) = (x_1, \dots, x_{n_s})$ and $r(x) = (x_{n_s+1}, \dots, x_n)$. The approach taken in this paper will likely produce better choices of (r, s) for model checking in a number of situations but the use of the splits used in Evans (1997), perhaps with random splitting, will likely be the only feasible ones in many contexts.

The use of the tail probabilities $P(T(X) \geq t_{obs})$, for some probability measure P , does not work well in capturing the idea of a surprising observation having occurred when this probability is small. This is because t_{obs} could be an extreme value in the left tail or, in general be in a low probability region such as near an anti-mode and $P(T(X) \geq t_{obs})$ will not indicate this. For this reason authors such as Weaver, Good and Box have worked instead with the density of T as this corrects for this problem. Using the density, however, destroys the invariance of the measure. The observed relative surprise, introduced in Evans (1997) corrects for both of these problems.

DENNIS V. LINDLEY (*Minehead, UK*)

Objections are often raised to the Bayesian approach because of its dependence on the prior. It is not so often recognized that the p -value can equally be criticized because of its dependence on the sample space. One can produce, for a given data set, a range of p -values by varying the sample space. It follows that since, in most practical cases, the sample space for a given data set is ill-defined, the p -value is also ambiguous. In particular, it can only be considered as a measure of surprise when the sample space is unambiguous. My personal view is that p -values should be relegated to the scrap heap and not considered by those who wish to think and act coherently.

XIAO-LI MENG (*The University of Chicago, USA*) and
ANDREW GELMAN (*Columbia University, USA*)

A paper containing multiple ideas is always fun to read. The main idea of Section 2, namely converting a p -value into a lower bound for Bayes factors is quite intriguing, especially considering that Bayes factors and p -value type measures answer two different statistical questions – a model can have a high Bayes factor compared to its stated competitors but still poorly fit important aspects of observed data. What's unclear to us, in general, is which p -value can be used to construct a useful lower bound given that a p -value is a functional of test statistics (or more generally *discrepancies*, i.e., $T(X, \theta)$), choices of replications (see below), etc. Perhaps the p -value from the likelihood ratio test (or the conditional likelihood ratio as defined in Meng (1994))?

Regarding the central theme of Section 3, we view p -value as a measure of *discrepancy* between the posited model and the data being analyzed, as we emphasized strongly in Meng (1994) and Gelman, Meng, and Stern (1996). While it might be a semantic matter to some, we prefer the term *discrepancy* because it honestly reflects what a small p -value tells us: the data and the model do not seem to see eye to eye in a specified way, but we cannot tell you which to blame! The phrase "surprise in the data" seems to carry the impression that the problem is with the data (e.g., an "unlucky" sample) and the standard practice with hypothesis testing always emphasizes the rejection of the posited hypothesis, not the data. While it is obviously desirable to pinpoint the sources of the discrepancy, p -value type measures simply cannot tell us whether the problem is with the data, or the model, or, as is more likely, both! If one's goal is simply to fit the data, then of course the source is always the model. But with scientific inferences, the problem can be far more complicated – Example 2 is a good illustration.

Viewed as discrepancy measure, Example 4 can be readily understood. A large value of $|\bar{x}_{obs}|$ does not necessarily indicate a large discrepancy between the data and the posited model

$N(0, \sigma^2)$ unless we know for sure the value of σ^2 , which is precisely why the lower bound on the posterior predictive p -value given in (4.11) monotonically decreases to zero as n approaches infinity. This also suggests that a relative measure such as $|\bar{x}|/s$ would be more useful for detecting the discrepancy in the mean. Indeed, with this choice of discrepancy the posterior predictive p -value would be identical to that from the classical (two-sided) t test, i.e., the p -value given in (4.9).

Incidentally, (4.9) can be obtained under the posterior predictive framework even when using $|\bar{x}|$ if we ignore the null value $\mu = 0$ when computing the posterior for σ^2 under the same model as used in the paper. Note that although Example 4 states that the null is $N(0, \sigma^2)$, any model checking procedure based solely on \bar{x} and s cannot check the normality assumption – indeed, the classical t test and other methods discussed in Example 4 are robust to the normality assumption (unless n is very small). So it would be better to cast this problem as checking the mean parameter $\mu = 0$ versus $\mu \neq 0$, which was the original formulation given in Meng (1994). With this formulation, the classical answer is obtained if we use the marginal posterior $p(\sigma^2|x_{obs})$ instead of the conditional one $p(\sigma^2|x_{obs}, \mu = 0)$; see Meng (1994) for details. The problem with using $p(\sigma^2|x_{obs}, \mu = 0)$, from the point of view of testing $\mu = 0$ (not of checking discrepancy in $|\bar{x}_{obs} - 0|$), is that it can grossly overestimate σ^2 when $\mu \neq 0$ and thus leads to the very conservative nature of the p -value given in (4.11).

This example makes it clear that the choice of test/discrepancy is important and is confounded with the choice of replication. Throughout the literature of posterior predictive checking (e.g., Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996), these points are always emphasized. For example, in Gelman, Meng and Stern (1996), we explicitly define $f(x|\theta, A)$ with A being an auxiliary statistics, as the authors mentioned. What is proposed in Section 4 is to condition on such an A (or in authors' notation, U) instead of the full data when finding the posterior of θ . Such conditioning is in the same spirit as conditioning on the classical ancillary statistics (for the parameter fixed by the null model, i.e., μ , not σ^2 in Example 4), which is in the right direction for a frequentist in the mind of a Bayesian because it is towards full conditioning, but is in the opposite direction when one is already doing Bayesian full conditioning. It would be better to resolve the "power" issue through the choices of discrepancy and the sampling replication $f(x|\theta, A)$. Even the marginal (e.g., $p(\sigma^2|x_{obs})$) versus conditional (e.g., $p(\sigma^2|x_{obs}, \mu = 0)$) approach is unsatisfactory because once we allow ourselves to not fully condition on the null model when computing the p value, we would need a principle to decide to what extent the null should be conditioned upon.

Of course, mathematically speaking, having more flexibility implies possibly better optimality in terms of frequentist operating characteristics of the resulting procedures. We look forward to seeing more convincing examples of the utility of the conditional predictive approach (incidentally we think the term *partially conditional predictive* would be more precise than *conditional predictive* because the posterior predictive approach is conditional, in fact, full conditional predictive approach with the authors' use of the word "conditional"). Example 4 would be theoretically more revealing if the "perfectly satisfactory" answer (i.e., the classical t test) could only be obtained under the proposed partial conditioning approach – with the current example, the answer can be obtained via any predictive approach, from no conditioning (i.e., prior predictive) to full conditioning (i.e., posterior predictive), since $|x|/s$ is pivotal.

MICHEL MOUCHART (*Université Catholique de Louvain, Belgium*)

(i) That a quantification of "surprise" depends on the prior specification might be desirable for several reasons among which one should mention : a) in a Bayesian model, i.e. a joint probability on the observations and the parameters, the "sampling" component is as subjective, and liable to "doubts", as the "prior" components, b) as argued in p.316 (remark (iii)) in Florens

and Mouchart (1993), a possible motivation for model choice is to introduce the statistician's own doubt about (her)his own model, once (s)he has received a particular new piece of information; it has been shown that this leads to operate a "critical partition" on the product space "parameters \times observations" to be calibrated with respect to the joint probability (involving eventually the prior specification c) a closely related issue, treated in section 2 of Florens and Mouchart (1989) consists in considering Bayesian testing as a treatment of "Unreliable message".

(ii) a question: How far is the concept of "default" alternative different from the concept of "implicit" alternative which is defined with respect to a particular test statistic (in a frequentist approach) and developed, among others, by Davidson and McKinnon?

(iii) with respect to the conditional predictive p -value, it has been shown (Mouchart and Scheihing (1993)) that the loss of information in the exact Fisher test (due to inadequate conditioning) is increasing with the sample size, at least when evaluated from a bayesian point of view. Can something similar happen with what you suggest?

TONY O'HAGAN (*University of Nottingham, UK*)

In presenting this paper at the conference, Berger stressed that he and Bayarri were addressing the problem of assessing the fit of a proposed model in the absence of an alternative model. Now the strict Bayesian view is that this is nonsense: if it is not possible to think of any alternative then asking about model fit is superfluous; if plausible alternatives exist they should be formulated and their posterior probabilities computed. Despite this, I believe that there is a real problem to be addressed.

In practice there is rarely any difficulty in thinking of alternative models. Rather, the difficulty is that one can all too easily and quickly think of a huge number of them. The full Bayesian approach in which all possible alternatives are formally considered in one enormous analysis is then completely impractical. Nor is this task worth attempting.

The question is: would such a full Bayesian analysis produce inferences that differ in any important way from those we obtain from the simple analysis, using the originally proposed model? Broadly speaking, if the model fits the data well, then the alternatives (which begin with lower prior probabilities) will have low posterior probabilities and the full analysis will not produce importantly different answers. If, on the other hand, alternatives exist which are plausible a priori, fit the data better, and produce inferences differing in an important way from those obtained from the original model, then the full analysis will also produce importantly different inferences. But in this case the full analysis is unnecessary—we do not need to consider *all* alternatives, just these "important alternatives". The challenge is to recognise when important alternatives exist, and then to identify them.

This is where measures of surprise and other forms of model criticism are important. But the key point where I disagree with Bayarri and Berger is that I do not believe the two tasks can be separated. Any measure which helps to diagnose whether important alternatives exist implicitly points to the kind of alternatives which we should consider. There cannot be a single, portmanteau measure of fit that will always diagnose the need to consider alternative models. Therefore we need to have a diversity of model criticism tools in our practical Bayesian toolbox. We also need to understand the kinds of alternatives that each tool is effective in diagnosing.

LUIS RAÚL PERICCHI (*Universidad Simón Bolívar, Caracas, Venezuela.*)

We should congratulate Professors Bayarri and Berger for very original pieces of work that will give further fresh air to Bayesian strategies, the current most active approach to theoretical (and practical) statistics. My first question is about the calibration of p -values. Their formula (2.1) is a neat and simple improvement over flawed p -values, likely to be added to current packages overnight. However, as the authors warn us, its interpretation is not clearcut. The

corrections of p -values needed for model validation are bound to depend on the sample size and dimension of the model. The same value of their surprise index might indicate very different intensities of concern about the model for small sample sizes than for large ones. It is not possible to give some general approximate guidelines, regarding sample sizes and dimensions, by considering for example, natural alternative hypotheses and default Bayes Factors? My second question is about conditional predictive p -values. Their discussion of merits and demerits of prior and posterior predictive p -values is very incisive. They want to get the best of both, resorting to conditioning upon a statistics U orthogonal to the departure statistics T . This is certainly an improvement over both prior (often unachievable) and posterior (illegal) predictive p -values. There are two problems with this, for which the authors give considerable attention and useful suggestions: the choice of U and the computation of the conditional predictive. However there is a third aspect that deserves attention, namely that in the computation of the conditional predictive there is the assumption that the null model is true. An alternative, which is a potentially more robust method (in the sense that the computation of the tail probability remains approximately valid under departures of the base model) is a resampling or data splitting method. Take a training sample $x(l)$ and consider for the remaining observations $x(-l)$ the posterior predictive $m(t(x(-l))|x(l))$ p -value, and then average (for example) the log-conditional p -values over all training samples. This might be a more robust measure of the departure that matters more from the base model. For instance, in the Example 4 it is implicitly assumed (by the choice of T) that it is a change of location what really matters, rather than the Normal shape. This observation is still highly speculative, but is in line with what seems to be occurring with Bayes Factors, namely that resampling Bayes Factors appear to be more robust than non-resampling ones, in the sense of being meaningful even if the sampling model is outside the candidate set.

DAVID A. VAN DYK (*Harvard University, USA*)

I would like to begin by thanking the authors for a well written summary of many of the issues involved with "Bayesian p -values." As an applied statistician, I find the sampling properties of a procedure (i.e., how well I can expect it to perform in practice) much more compelling than issues such as "non-Bayesian character" or (apparent) lack of "natural" Bayesian interpretation. Thus, I will comment on the sampling properties of the various procedures.

In my experience, posterior predictive p -values, when used with proper care, are practical tools for model checking and building. A particularly powerful feature of these (and other) Bayesian p -values is that the choice of test statistic allows the procedure to be easily tailored to the question at hand. This flexibility comes at a cost: using a test statistic that is poorly suited to particular question can lead to a poor procedure. As a criterion for choosing the test statistic, we can define the power, $\beta(\theta) = \Pr(p < \alpha|\theta)$. (As pointed out by Meng (1994), $\sup_{\theta \in \Theta_0} \beta(\theta)$ can be less than α , where Θ_0 corresponds to H_0 , so the p -values are conservative.) Example 4 illustrates that a poor choice of the test statistic can result in a very weak procedure. In particular, for $\alpha < 2[1 - \Upsilon_n(\sqrt{n})]$, $\beta(\theta) = 0$ for all μ , where $\theta = (\mu, \sigma^2)$ with μ the mean of X_i . This is, however, not unexpected since neither small nor large values of the chosen test statistic, $T_0(x) = |\bar{x}|$, are surprising under the null model. What is surprising under the null model are values of s^2 which are much smaller than $\sum_i x_i^2$ or values of $T_0(x)$ which are much larger than s/\sqrt{n} . These observations lead to alternative statistics such as $T_1(x) = s^2$ or $T_2(x) = \sqrt{n}\bar{x}/s$. Figure 1 shows $\beta(\theta = (\mu, \sigma^2))$ as a function of μ , with $\sigma^2 = 1$ for $T_i(x)$ for $i = 0, 1, 2$ and the conditional predictive p -value suggested in Example 4. Note that $T_2(x)$ is as powerful as the conditional predictive p -value. It is also clear that the test based on $T_0(x)$ quickly becomes more powerful as n grows.

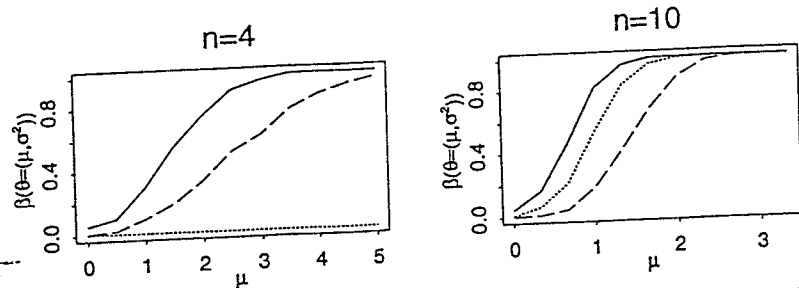


Figure 1. The power of several procedures for the test in Example 4, with $\alpha = 0.05$. The posterior predictive p -values generated with $T_i(x)$ for $i = 0, 1, 2$, correspond to the dotted line, the dashed line, and the solid line respectively. The solid line also corresponds to the conditional predictive p -value described in Example 4.

Clearly with careful selection of the test statistic, the posterior predictive p -value can be as powerful as a conditional predictive p -value (at least in this example). Moreover, even if a posterior predictive p -value is relatively weak it still has an important computational advantage. As recognized by the authors in the choice of δ , if a low-power easy-to-compute test gives strong evidence against the null model, there is no need to invest heavily (see Section 4.2 and Pauler's discussion) in using a more powerful procedure.

REPLY TO THE DISCUSSION

We thank the discussants for their considerable insights; if we don't mention particular points in the discussions, it is because we simply agree with those points.

We are, of course, delighted to accept the apology of *Dr. Robins* concerning his verbal discussion at the Valencia meeting, since his written discussion provides more extensive frequentist validation of our proposals than we expected. We would not have guessed that the conditional predictive and posterior predictive p -values are asymptotically uniform under the (true) null model, for any value of the parameters, and that they are the only predictive p -values that possess this property in general. In retrospect, however, perhaps we should not have been so surprised at this validation; our goal was to develop predictive p -values that make as much sense from the conditional Bayesian perspective as possible while utilizing noninformative priors, and we should by now be used to the fact that success in such endeavors on the Bayesian side will also yield success on the frequentist side. We should also not be surprised that the predictive p -values which violate Bayesian intuition in terms of conditioning will also fail frequentist validation, as Robins has shown. (This is also our rejoinder to the first paragraph of the discussion of *Dr. Van Dyk*.)

In addition to helping to answer the question of which predictive p -values should be utilized by Bayesians, the results of Robins can be viewed as providing strong arguments for use of the conditional and partial posterior predictive p -values by classical statisticians. Indeed, the 'standard' plug-in p -value (called the bootstrap p -value by Robins) also fails the asymptotic frequentist evaluation. It also seems likely that, for small samples, the Bayesian procedures will have superior frequentist properties to classical p -values, but this investigation will be pursued elsewhere. The paper De la Horra and Rodríguez-Bernal (1997a), mentioned by *Dr. De la Horra*, is also relevant to this issue, although the assumptions of that paper must be somewhat restrictive since the discussion of Robins shows that, in general, the posterior predictive p -value is not uniform under the null model.

An additional feature of the discussion of Robins is that it fills two gaps in our paper. The first gap is the rather ad hoc nature of the partial posterior predictive p -value, which we had proposed primarily on an intuitive basis. By showing the asymptotic equivalence of this procedure with the conditional predictive p -value, Robins has provided a validation of the intuition. Perhaps even more interesting is the manner in which the results of Robins fill the gap in the paper between the proposed calibration, $\underline{B}(p) = -e p \log(p)$, and the proposed predictive p -values. Our general argument in support of the calibration was explicitly based on having a (random) p -value which is uniform under the null hypothesis, but we never showed that this assumption was met by the p -values we proposed; Robins has shown this to be the case, at least asymptotically. Hence the calibration can be used with confidence for these predictive p -values. Note that De la Horra refers to a different possible calibration of p -values.

A final comment about the discussion of Robins is that it contains a wealth of results on other p -value methodologies. Space precludes further discussion of these fascinating results here.

We were very glad to see the analyses of *Dr. Pauler*, since our paper was definitely lacking in examples of application of the new predictive p -values. The overall conclusion of Pauler, that the conditional predictive p -value can be hard to compute, is certainly an important consideration and may well prevent its routine use by Bayesians in complicated situations. Clearly, however, the partial posterior predictive p -value is much easier to work with; indeed, when $f(t_{obs}|\theta)$ is available in closed form, computation of the partial posterior predictive p -value will be very straightforward, as discussed in the paper. When $f(t_{obs}|\theta)$ is not available in closed form, Pauler rightly observes that its estimation by kernel methods may well be expensive; note, however, that T is typically a one-dimensional statistic and estimation of a one-dimensional density at a point is usually not excessively difficult.

Pauler mentions the possibility of conditioning on sufficient statistics and points out that ensuing conditional p -values can be fairly easy to compute. This is certainly a reasonable possibility and the resulting p -values are clearly a version of conditional predictive p -values. Of course, sufficient statistics often do not exist and, even when they do, it is typically easier to operate with the partial posterior predictive p -value than with the distribution conditioned on the sufficient statistic. We also suspect that conditioning on the sufficient statistic loses power, in that the sufficient statistic also involves T , yielding a confounding double use of that part of the data. Pauler also observes that allowing T to depend on unknown parameters can be useful and that our methods, as currently formulated, do not allow for this option; we have not considered this issue enough to venture an opinion at this time.

In the example of exponential overdispersion that Pauler considered, we were actually rather encouraged by the stability of the conditional predictive p -value with respect to choices of δ . Indeed, the answers did not seem to vary much as δ varied from 0.01 to 1.0.

Several of the contributed discussions, especially those of *Dr. van Dyk* and *Dr. Carlin*, echoed these computational concerns. We are certainly not unsympathetic; the statistical tools we use are always limited by computational considerations. While we would argue for the conditional or partial posterior predictive p -values as being optimal, when both are too difficult to compute in a given situation we would certainly consider use of the predictive p -values discussed by van Dyk and Carlin.

Dr. Bernardo begins by reiterating Bayesian arguments in favor of developing alternative models; as we said in the paper, we were refraining from taking up this issue, and we will still refrain from doing so! Bernardo asks how we view the calibration of p -values in Section 2 in terms of varying sample size. For the reasons discussed at the beginning of Section 3, we feel that the calibration is likely to be reasonable for small n but that, when n is large, one might worry

about the quality of the calibration. *Dr. Pericchi* emphasizes the need for concern with this issue and asks if general guidelines can be given for varying sample sizes and dimensionalities. Our view is that such guidelines might one day be available but that they may be of such complexity that it may well be simpler to directly develop alternatives and use (default) Bayes factors.

Dr. Meng and *Dr. Gelman* pose a different question about calibration, asking which choices of T , the sample space, etc., yield p -values whose calibrations actually correspond to reasonable lower bounds on Bayes factors. Our hope in this regard is that the choice of T and the sample space might be a reasonable surrogate to choice of an actual alternative, and that the resulting calibrated p -value might then be a reasonable surrogate to the corresponding lower bound on the Bayes factor (at least if the CPP or PPPP is used). Each calibrated p -value would then correspond to a lower bound on the Bayes factor for a different (implicit) alternative. Of course, our results only hint that this might be so. *Dr. Lindley* also mentions the possible arbitrary nature of the sample space in determination of a p -value. Indeed, if the sample space is ill-specified, then embarking on this program would be ill-advised.

Meng and Gelman defend the posterior predictive p -value by arguing that the notion of discrepancy between the posited model and data must be interpreted properly, and that one must choose the statistic T carefully to obtain greatest sensitivity. In this regard, Meng and Gelman (and van Dyk) point out that, in the situation of our Example 4, Meng (1994) had (in addition to $T = \bar{X}$) considered $T = \bar{X}/S$, in which case the posterior predictive p -value is the same as that in (4.9).

While we agree with much of what Meng and Gelman say, the notion that one can overcome the problems with the posterior predictive p -value by careful choice of T raises two issues. The first is that, if one is going to spend considerable effort in choosing T , wouldn't the time be better spent in developing alternative models and computing real Bayes factors? Recall that we are recommending this enterprise only as a 'quick and dirty' alternative to actual Bayesian analysis, and this would (in our minds) preclude the expenditure of great effort to choose T . Related is the second issue, that finding a 'good' T is by no means an easy enterprise. Even in the simple Example 4, where concern about departure from a zero mean immediately suggests consideration of $T = |\bar{X}|$, realizing that (with the posterior predictive p -value) one should instead use $T = |\bar{X}|/S$ requires a considerable degree of statistical sophistication. In the frequentist language of Robins, one must take the provisional T and appropriately recenter by an estimate of its mean under the true model. In complicated problems, the determination of such 'good' T is likely to be extremely formidable. In contrast, the conditional and posterior predictive p -values require only the specification of the intuitive T , with the appropriate recentering taking place automatically as part of the procedure. Thus, in Example 4, using the 'naive' $T = |\bar{X}|$ still results in the correct answer (4.9) when the conditional or posterior predictive p -value are used.

Carlin reminds us that the cross-validated predictive is another predictive worthy of serious consideration, and we agree. Computational comparisons with this predictive are a bit murky since, when $f(t|\theta)$ is available in closed form, the partial posterior predictive p -value will often be considerably easier to compute than the cross-validated predictive p -value. In other situations, however, the reverse may well be true. It is also unclear if the cross-validated predictive p -value has desirable properties such as asymptotic uniformity under the null, as discussed by Robins. Our guess is that it will not, in part based on analysis of the only example we could do in closed form, namely Case 3 of Example 6. For this situation, the cross-validated predictive p -value can be computed to be

$$p = \prod_{i=1}^n \left(1 + \frac{t_{obs}}{s_{obs} - x_i} \right)^{-(n-1)},$$

which has the same difficulty as the posterior predictive p -value and the plug-in p -value, namely that it converges to a nonzero constant (roughly $e^{(0.5-n)}$) as $nt_{obs}/s_{obs} \rightarrow 1$. Again, this is a situation where the evidence against the model is clearly overwhelming, and the cross-validatory predictive p -value does not reflect this. Dr. Pericchi proposes study of averages of other 'data-splitting' predictives, based on favorable experience with such-in development of default Bayes factors. We would be interested in seeing and studying specific proposals in this direction, although there will be difficulties in actually defining p -values in this way. The results mentioned by Dr. Evans are certainly relevant here.

Evans mentions some of the problems with use of tail areas and we would not disagree, except to note that there is some hope of addressing these problems through choice of T . We do not know the answers to the questions of Dr. Mouchart, although we have been investigating the exact Fisher test and it seems that the p -values we propose (when used with sensible choices of T) are more powerful in various senses than the usual p -value. In reply to Dr. O'Hagan, we should also mention that we are not proposing a single portmanteau measure of fit; each possible choice of T provides a different measure of fit, and carries with it implicit understanding as to where to look for alternatives, should fit be lacking. We should also emphasize that the advantage of this class of measures of fit is interpretability; the results of Robins, which indicate that our p -values are true p -values, and the calibration in Section 2 (for true p -values) together suggest that the proposed measures of fit will carry a consistent and accessible meaning.

ADDITIONAL REFERENCES IN THE DISCUSSION

- Bernardo, J. M. (1998). Nested hypothesis testing: The Bayesian reference criterion. *In this volume*.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B* 36, 192–236 (with discussion).
- Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* 76, 633–642.
- Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p -values. *Biometrika* 78, 301–304.
- Davison, A. C. (1998). Comment to 'Some algebra and geometry for hierarchical models, applied to diagnostics', by J. S. Hodges, *J. Roy. Statist. Soc. B* 60, 529–530.
- De la Horra, J. and Rodríguez-Bernal, M. T. (1997). Asymptotic behaviour of the posterior predictive p -value. *Commun. Statist.-Theory Meth.* 26, 2689–2699.
- De la Horra, J. and Rodríguez-Bernal, M. T. (1999). The posterior predictive p -value for the problem of goodness of fit. *Test* (to appear).
- Dempster, A. P. (1971). Model searching and estimation in the logic of inference. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) Hoit, Rinehart, and Winston: Toronto, 56–81.
- Florens, J.-P. and Mouchart, M. (1993). Bayesian testing and testing Bayesians. *Handbook of Statistics* (G. S. Maddala and C. R. Rao, eds.). Amsterdam: North-Holland.
- Florens, J.-P. and Mouchart, M. (1989). Bayesian specification tests. *Contributions in Operations Research and Economics* (B. Cornet and H. Tulkens, eds.). Cambridge: MIT Press, 467–490.
- Gelman, A., Carlin, J. B., Stern, H.S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 147–167 (with discussion).
- Robins, J. M. (1998). Frequency properties of Bayesian p -values. *Tech. Rep.*, Harvard School of Public Health.
- Weiss, R. E. (1996). Bayesian model checking with applications to hierarchical models. *Tech. Rep.*, UCLA.
- Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *J. Roy. Statist. Soc. B* 60, 497–536 (with discussion).