# Quantifying Surprise in the Data and Model Verification

by

M.J. Bayarri and James O. Berger
University of Valencia and Duke University
Duke University

*Institute of Statistics and Decision Sciences*
*Durham, North Carolina 27708, USA*

**Abstract**

$P$-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.

**Key words and phrases**. 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 Sellke, 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, apriori, the unknown $\theta$ 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

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

with there being a number of different proposals for the probability distribution to be used in this computation when $\theta$ 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 $\underline{B}(p) = -e\, p \log(p)$, 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 $\theta$ is unknown. Three common suggestions for this distribution are $f(x|\hat{\theta})$, where $\hat{\theta}$ is the m.l.e. for $\theta$; 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, 1997). 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 (1998). 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 $p < e^{-1}$,

$$\underline{B}(p) = -e\,p\log(p) \tag{2.1}$$

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 (1998). We first report a typical parametric example, and then give the nonparametric argument.

**Example 2.1.** Assume that $X_1, X_2, \cdots, X_n$ 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

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

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 |
| $\Gamma_{Normal}$ | 0.7007 | 0.4727 | 0.1534 | 0.02407 |
| $\Gamma_{US}$ | 0.6393 | 0.4084 | 0.1223 | 0.01833 |
| $\Gamma_{Sym}$ | 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 $\Gamma_{US}$ is often perceived as particularly reasonable, not requiring specification of a functional form for the prior (as does $\Gamma_{Normal}$), and yet corresponding to natural shapes for $\pi_1(\theta)$ (which is not true of many priors in $\Gamma_{Sym}$). The close agreement between the lower bounds on the Bayes factor for the class $\Gamma_{US}$ 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 (1998), the first of which is the generalization to higher dimensions of Example 2.1. 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 $X_i|\theta \sim N_k(\theta, I)$, $i = 1, \ldots, n$, with $\theta \sim N_k(0, \tau^2 I)$. Letting $\underline{B}$ denote the lower bound on the Bayes factor over all $\tau^2$ and assuming $p < 1/2$,*

$$\lim_{k\to\infty} \frac{\underline{B}}{\underline{B}(p)} = \frac{2}{e\, z_p}\ , \tag{2.3}$$

*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 2.1., 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 $\frac{p}{p_0} \mid p < p_0$ is stochastically decreasing with $p_0$, which, for instance, implies that, for any fixed $\rho$, the probability $Pr\{p < \rho p_0 | p < p_0\}$ 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*

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

*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 2.1., 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 *apriori 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 apriori believable.

**Example 3.2.** 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 $\theta$ 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 $\theta$ 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 $\theta$ by integration with respect to some distribution of $\theta$.

### 4.1. The prior and posterior predictive $p$-values

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

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

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

$$p = \Pr^m(T(X) \geq T(x_{obs})). \tag{4.5}$$

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 (4.5).

The concerns mentioned in the last paragraph have led many Bayesians, beginning with Guttman (1967) and Rubin (1984), to eliminate $\theta$ 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

$$p = \Pr{}^{m(\cdot|x_{obs})}(T(X) \geq T(x_{obs})), \tag{4.6}$$

where

$$m(x|x_{obs}) = \int f(x|\theta)\pi(\theta|x_{obs})d\theta. \tag{4.7}$$

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 $\theta$, 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 (4.6) 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 (4.7) 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 (4.6), 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 4.3.** Suppose $\theta$ is generated as a random integer between 1 and 1000. Thus the prior distribution is uniform on the integers $\{1, 2, \ldots, 1000\}$. Consider two different models. Under $M_1$, the data $X$ equals $\theta$ with probability one while, under $M_2$, the value of $\theta$ 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 $\{1, 2, \ldots, 1000\}$, 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 $\{1, 2, \ldots, 1000\}$. □

### 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

$$p = \Pr^{m(\cdot|u_{obs})}(T \geq t_{obs}), \tag{4.8}$$

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

$$m(t|u) = \int f(t|u, \theta)\pi(\theta|u)d\theta, \tag{4.9}$$

assuming that

$$\pi(\theta|u) = \frac{f(u|\theta)\pi(\theta)}{\int f(u|\theta)\pi(\theta)d\theta} \tag{4.10}$$

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 $U$, (4.8) 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 $\theta$, while another part ($t_{obs}$) is used when computing the tail area. We illustrate with an example from Meng (1994).

**Example 4.4.** Assume that, under the null, the $X_i$ are iid $N(0,\sigma^2)$, with $\sigma^2$ unknown. The statistic $T(x) = |\overline{x}|$ is chosen to measure departure from the model (which would be natural for detecting discrepancy in the mean of the model). Let $U(x) = s^2 = \sum(x_i - \overline{x})^2/n$, and consider the usual non-informative prior for $\sigma^2$: $\pi(\sigma^2) \propto 1/\sigma^2$. Computation shows that the posterior distribution, $\pi(\sigma^2|s^2)$, is $Ga^{-1}((n-1)/2, ns^2/2)$ and that

$$m(\overline{x}|s^2_{obs}) = t_{n-1}(\overline{x} \mid 0, \frac{s^2_{obs}}{n-1}) \qquad \text{or} \qquad \frac{\sqrt{n-1}\ \overline{X}}{s_{obs}} \sim t_{n-1}(\,\cdot\,|\,0,1)\ . \tag{4.11}$$

(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

$$p = Pr\{|\overline{X}| > |\overline{x}_{obs}|\} = 2\left[1 - \Upsilon_{n-1}\left(\frac{\sqrt{n-1}\ |\overline{x}_{obs}|}{s_{obs}}\right)\right]\ , \tag{4.12}$$

where $\Upsilon_{n-1}$ 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 $Ga^{-1}(n/2, n(s^2+\overline{x}^2)/2)$, and the posterior predictive distribution of $\overline{X}$ is

$$m(\overline{x}|x_{obs}) = t_n(\overline{x} \mid 0, \frac{s^2_{obs}+\overline{x}^2_{obs}}{n}) \qquad \text{or} \qquad \frac{\sqrt{n}\ \overline{X}}{\sqrt{s^2_{obs}+\overline{x}^2_{obs}}} \sim t_n(\,\cdot\,|\,0,1)\ . \tag{4.13}$$

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

$$p = Pr\{|\overline{X}| > |\overline{x}_{obs}|\} = 2\left[1 - \Upsilon_n\left(\frac{\sqrt{n}\ |\overline{x}_{obs}|}{\sqrt{s^2_{obs}+\overline{x}^2_{obs}}}\right)\right]\ . \tag{4.14}$$

This is unsatisfactory, as can be seen by letting $|\overline{x}_{obs}| \longrightarrow \infty$. Then $p \longrightarrow 2[1-\Upsilon_n(\sqrt{n})]$, 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 $\overline{x}_{obs}$ 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 $\overline{x}_{obs}$ 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 $\theta$ as possible, so that $\pi(\theta|u_{obs})$ will effectively eliminate $\theta$ (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 4.4., for instance, $\sum x_i^2/n$ would contain all information about $\sigma^2$ (being a sufficient statistic under the presumed model), but does involve $T(x) = |\overline{x}|$. The obvious solution (used in Example 4.4.) is to define $U(x) = s^2 = \sum(x_i - \overline{x})^2/n$, since this contains the information about $\sigma^2$ that is independent of $T(X)$.

Investigations in Bayarri and Berger (1997) also suggest that $U(x)$ should have the same dimension as $\theta$. 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 $\theta$, given $T(x) = t$,

$$U(x) = \hat{\theta} = \arg\max f(x|t,\theta) = \arg\max \frac{f(x|\theta)}{f(t|\theta)} \quad . \tag{4.15}$$

(The situation of discrete data is considerably more difficult; while $U(x)$ in (4.15) 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 $\hat{\theta}$ is satisfactory as the choice of $U$. In Example 4.4., for instance,

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

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 4.5.** Assume that $X_1, X_2, \ldots, X_n$ is a random sample from the $Exponential(\lambda)$ distribution. We consider several possibilities for $T$, and derive the suggested $U$ in each case. In the following, $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ 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 $T \sim Exp(\lambda)$,

$$f(x|t,\lambda) \propto \lambda^{n-1} \exp\{-\lambda \, [\sum_{i=1}^{n-1} x_{(i)} + x_{(n-1)}]\} \quad , \tag{4.16}$$

   and the $U$ suggested by maximizing (4.16) 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

$$f(x|t,\lambda) \propto \left(\frac{\lambda}{1 - e^{-\lambda t}}\right)^{n-1} \exp\{-\lambda \, [\sum x_{(i)} - t]\} \quad , \tag{4.17}$$

   which would have to be numerically maximized over $\lambda$ to determine $U$ via (4.15) .

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

$$f(x|t,\lambda) \propto \lambda^{n-1} \exp\{-\lambda \, (\sum x_{(i)} - nt)\}, \tag{4.18}$$

   which, upon maximization over $\lambda$, 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 $\delta$, is simply an approximation to $m(x|u_{obs})$. Of course, smaller $\delta$ 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 $\delta$; indeed, they are conditional predictive $p$-values, simply being based on a somewhat different $U$. As a matter of fact, as $\delta \to \infty$, it is clear that $m(x|\ |u-u_{obs}|<\delta) \to$ the prior predictive, $m(x)$. Thus one need not strive for exceedingly small $\delta$.

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

$$m(x|\ |u-u_{obs}|<\delta) = \frac{\int f(x|\theta)\pi(\theta)1_{\{|u-u_{obs}|<\delta\}}d\theta}{Pr(|u-u_{obs}|<\delta)} \quad , \tag{4.19}$$

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 (4.19) 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 $\theta$, not the typically more difficult posterior conditional on $u$. The easiest way to implement *Step 2*, for the given $\theta$, is to repeatedly generate $x$'s until $u(x)$ is within $\delta$ 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 $\theta^i$ generated from $\pi(\theta|x_{obs})$.

**Metropolis-Hastings** . Generate a chain $(x^j, \theta^j)$ 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 $(x^t, \theta^t)$ at time $t$,

- *Step 1.* Generate a candidate $(x^*, \theta^*)$ 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 $\delta$ of $u_{obs}$. Notice that, if $u(x)$ is not within $\delta$ of $u_{obs}$, a *new* $\theta$ 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

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

- *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 (4.15) 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 $\theta$ values, in an adaptive way if needed.
- *Step 2.* For each $\theta$, 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

$$\hat{f}(t|\theta) = (\#\ T(x^i) \text{ within } \epsilon \text{ of } t)\ /\ (2M\epsilon) \quad ;$$

of course, a more sophisticated kernel estimator could be used.

- *Step 3.* Set $u$ equal to the value of $\theta$ that maximizes $r(\theta)$ over the grid.

Notice that we only need those values of $u$ that are within $\delta$ of $u_{obs}$. Thus, once $u_{obs}$ has been (carefully) computed, all that is needed is a grid of $\theta$ values that are within $\delta$ 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 $\theta$. A famous example is the Fisher exact test for independence in a $2 \times 2$ 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 (4.9) 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 $\theta$. 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 $\theta$, 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 (4.15) can be difficult to compute. An attractive alternative is to directly use $f(x|t,\theta)$ (see (4.15)) to integrate out $\theta$, 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

$$p = \Pr^{m^*(\cdot)}(T \geq t_{obs}), \tag{4.20}$$

where $m^*$ and the (assumed proper) *partial posterior* $\pi^*$ are given by

$$m^*(t) = \int f(t|\theta)\pi^*(\theta)d\theta, \qquad \pi^*(\theta) \propto f(x_{obs}|t_{obs},\theta)\pi(\theta) \propto \frac{f(x_{obs}|\theta)\pi(\theta)}{f(t_{obs}|\theta)}. \tag{4.21}$$

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 $\theta$ is eliminated by integration. Furthermore, the parallel with (4.15) suggests that the resulting answer will be very similar to the conditional predictive $p$-value. Indeed, in Example 4.4. 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 4.5.

**Example 4.6.** Assume exponential data, as in Example 4.5. 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 (4.16) 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

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

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 $t_{obs}/s_{obs} \to 1$, 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 $\lambda$ by an estimate in the density of $T$.

2. $T = X_{(n)}$. Then, from (4.17) 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 $s_{obs}/t_{obs} \to 1$, 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 (4.18) 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

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

   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 $nt_{obs}/s_{obs} \to 1$, 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 noinformative 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 $\theta$, then it is clear from (4.21) 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 4.6. 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 4.6., for instance, one can easily compute the Jeffreys prior for $\lambda$ 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 can be rewritten

$$p = \int \Pr(T \geq t_{obs}|\theta)\pi^*(\theta)d\theta,$$

an obvious strategy is to repeatedly generate $\theta$ 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 $\theta^*$, and then move from the current $\theta_j$ to the candidate with probability $\min\{1, f(t_{obs}|\theta_j)/f(t_{obs}|\theta^*)\}$.

Alternatives to such direct Monte Carlo computation of $p$ include importance sampling schemes. For instance, if a sample $\{\theta_j, j = 1, \ldots, m\}$ 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 $\underline{B}(p) = -e\,p\log(p)$) 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 National Science Foundation (U.S.A.) under grants DMS-9303556 and DMS-9802261, and by the Ministry of Education and Culture (Spain) under grant PB96-0776.

## References

[1] Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society B*, **53**, 111-142.

[2] Bayarri, M.J. (1986). A Bayesian goodness of fit test. In *Proceedings of the 1985 Joint Statistical Meetings*, American Statistical Association, Washington, D.C.

[3] Bayarri, M.J. and Berger, J. (1997). Measures of surprise in Bayesian analysis. *ISDS Discussion Paper* 97-46, Duke University.

[4] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition.* New York: Springer-Verlag.

[5] Berger, J. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, **3**, 5–124.

[6] Berger, J.O., and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science*, **2**, 317-352.

[7] Berger, J.O., and Sellke (1987). Testing a point null hypothesis: the irreconciability of $p$ values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.

[8] Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A*, **143**, 383-430.

[9] Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Amer. Statist. Assoc.* **82**, 106–111.

[10] Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

[11] Delampady, M., and Berger, J.O., (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Annals of Statistics*, **18**, 1295-1316.

[12] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev.*, **70**, 193–242.

[13] Efron, B. (1996). Talk at the Conference on Statistical Challenges in Astronomy, Penn State University, June 2-5, 1996.

[14] Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics*, **26**, 1125-1143.

[15] Gelman, A., Meng, X.L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733-807.

[16] Goldstein, M. (1991). Comment on 'Posterior Bayes factors' (by M. Aitkin). *Journal of the Royal Statistical Society B*, **53**, 134.

[17] Good, I.J. (1956). The surprise index for the multivariate normal distribution. *Annals of Mathematical Statistics* **27**, 1130-1135.

[18] Good, I.J. (1983). *Good Thinking: The Foundations of Probability and its Applications.* Minneapolis: University of Minnesota Press.

[19] Good, I.J. (1988). Surprise index. *Encyclopedia of Statistical Sciences* (Kotz, S., Johnson, N.L., and Reid, C.B., eds.) **7**, 104-109.

[20] Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B*, **29**, 83-100.

[21] Hodges, J. (1992). Who knows what alternative lurks in the heart of significance tests? In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) 247-266. London: Oxford University Press.

[22] Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142-1160.

[23] Roberts, H.V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, **60**, 50-62.

[24] Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.

[25] Sellke, T., Bayarri, M. J., and Berger, J. (1998). Calibration of $p$-values for measuring surprise. *ISDS Discussion Paper,* Duke University.

[26] Verdinelli, I., and Wasserman, L. (1996). Bayesian goodness of fit testing using infinite dimensional exponential families. *Technical Report*, Carnegie Mellon University.

[27] Weaver, W. (1948). Probability, rarity, interest and surprise. *Scientific Monthly*, **67**, 390-392.