

Measures of Surprise in Bayesian Analysis

by

M.J. Bayarri
University of Valencia
and Duke University

and

James O. Berger
Duke University

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

Abstract

Measures of surprise refer to quantifications of the degree of incompatibility of data with some hypothesized model H_0 without any reference to alternative models. Traditional measures of surprise have been the p -values, which are however known to grossly overestimate the evidence against H_0 . Strict Bayesian analysis calls for an explicit specification of all possible alternatives to H_0 so Bayesians have not made routine use of measures of surprise. In this report we CRITICALLY REVIEW the proposals that have been made in this regard. We propose new modifications, stress the connections with robust Bayesian analysis and discuss the choice of suitable predictive distributions which allow surprise measures to play their intended role in the presence of nuisance parameters. We recommend either the use of appropriate likelihood-ratio type measures or else the careful calibration of p -values so that they are closer to Bayesian answers.

Key words and phrases. Bayes factors; Bayesian p -values; Bayesian robustness; Conditioning; Model checking; Predictive distributions.

1. Introduction

Imagine a situation in which inferences are desired about some state of Nature and you, as the statistician, carefully assess a model that you believe most accurately reflects whatever is known before data is collected and incorporates the assumptions that seem reasonable

in that specific context. However, once the data x_{obs} is collected, it is indeed most natural (and sensible) for you to consider whether x_{obs} is compatible with the assumed model.

This is, of course, a very old question in Statistics: once a (null) model (or hypothesis) H_0 is formulated and x_{obs} observed, is data “surprising”? From a frequentist perspective, a traditional measure of “surprise” is the p -value, or probability of observing data “more extreme” than x_{obs} *if* the null model is correct. In so doing, no (explicit) formulation of “alternative” models is required, which is a feature that has been attractive to applied statisticians over the years.

From a strictly Bayesian perspective, especially if approached from the full decision theoretic point of view, there seems to be little, if any, role for measures of surprise. Bayesians should ideally explicitly list all the components of the problem, including all the alternative models when the null model is under suspicion, assess priors over all unknown features of the problem and derive posterior probabilities (or Bayes factors) for the different possible models under consideration (and choose the model with highest expected posterior utility, if utility functions are also assessed).

The literature in Bayesian model selection and Bayes factors is much too vast to be reviewed in this Report, and we shall not attempt to do so. Some good references, from which other ones might be obtained are Gelfand and Dey (1994), Kass and Raftery (1995), Laud and Ibrahim (1995), O’Hagan (1995), and Berger and Pericchi (1996). Model selection and model testing incorporating also utility functions is addressed in San Martini and Spezzaferri (1984), Bayarri (1986), Poskitt (1987), Bernardo and Smith (1994), Carota et al. (1996), Gutiérrez-Peña and Walker (1996), and Goutis and Robert (1998).

We certainly believe that this is the ideal approach to model checking, and that whenever a “shortcut” is taken, it should aim to producing answers as close as possible to a full Bayesian answer. However, the use of measures of surprise (and by that we mean an analysis in which the only model explicitly considered is the hypothesized one) is *very* appealing due to its considerable simplification when compared to the complexity of specifying a full plethora of alternative models with their associated priors over their parameters. We do not believe that a “surprise” analysis can ever replace a full Bayesian one. We do argue, however, that surprise measures have an important role to play as exploratory tools, in the sense that, if x_{obs} can be nicely explained by H_0 we might not

need to take the extra effort of the full analysis. If, however, x_{obs} is “surprising” then we *do* have to carefully specify alternative models to H_0 and carry out a Bayesian (or at least a default Bayesian) analysis. Also, as Good (1981) remarked,

“The evolutionary value of surprise is that it causes us to check our assumptions. Hence if an experiment gives rise to a surprising result, given some null hypothesis H , it might cause us to wonder whether H is true even in the absence of a vague alternative to H .”

Through most of this report, we shall operate solely under H_0 which is defined as $X \sim f(x|\theta)$. A priori $\theta \sim \pi(\theta)$ (often non-informative) will also be assumed. Since there is no explicit H_1 , no prior is assigned over it. In some few occasions we shall need to refer to the Bayesian answer under a specific alternative H_1 . In this case we shall denote by $f_1(x|\eta)$ and $\pi_1(\eta)$ the corresponding densities under H_1 .

In a sense, Bayesians have a very natural measure of surprise in the infimum of Bayes factors derived from Bayesian global robustness analyses. Indeed if the statistician is willing to approximately specify a rough H_1 and only some vague characteristics of the prior π_1 in the form of a (large) class of priors to which π_1 belongs, then a natural measure of surprise would be based on the infimum of the Bayes factor in favor of H_0 . We assume in such a case that specification of H_1 takes the form of a larger model $f(x|\theta, \xi)$ in which $f(x|\theta)$ is nested; we further assume that the marginal prior distribution, $\pi(\theta)$ for θ is the same under both hypotheses and that $\pi_1(\theta, \xi) = \pi_1(\xi)\pi(\theta)$, where it is only assumed that $\pi_1 \in ?$. Then, the lower bound on the Bayes factor of H_0 to H_1 is

$$\underline{B} = \inf_{\pi_1 \in \Gamma} \frac{\int f(x|\theta)\pi(\theta)d\theta}{\iint f_1(x|\theta, \xi)\pi(\theta)\pi_1(\xi)d\theta d\xi} \quad , \quad (1.1)$$

and data x_{obs} resulting in small \underline{B} would be considered surprising.

Two remarks are in order. First, notice that for surprise purposes, that is, exploratory, informal, purposes, the alternative hypothesis H_1 does not need to be carefully assessed: an informal statement of interesting or likely departures from H_0 would suffice. Also notice that, should x_{obs} be considered surprising, the robust Bayesian approach would give us grounds as to what kind of departures from H_0 are worth considering.

The approach just outlined is a middle road between a full Bayesian analysis and

a surprise analysis, and would serve us as a good reference point when judging some measures of surprise to be developed in situations where the robust Bayesian approach is feasible. However, our motivation here is to be able to quantify the “surprise” in the data without *explicit* resort to alternative hypotheses, and we shall be mainly concerned with this goal in the rest of this Report.

Several authors have worked on measures of surprise from a Bayesian perspective. Naturally, many of them defend the possibility of being “surprised” based solely on the null model. Thus, Box(1980) says,

“...In making this predictive check it was not necessary to be specific about an alternative model. This issue is of some importance for it seems a matter of ordinary human experience that an appreciation that a situation is unusual does not necessarily depend on the immediate availability of an alternative.”

Good, as it will be seen in subsequent sections, was a firm defender of surprise analyses and in ([24]) commented

“...in the forensic example ... [data] is surprising under *each* of the ... hypotheses so far considered ... To be surprised ... is to obtain a substantial Bayes factor in favor of the hypothesis that you have overlooked some hypothesis! ... Anyone, Bayesian or not, who ... insists on formulating *all* hypotheses in advance of the observations, would be embarrassed by the forensic example. I do not so insist, and nor do most scientists, judges, and juries.”

It should be noticed that, unlike some other authors, we are *not* saying that a surprising x implies rejecting H_0 . In fact, we are firmly convinced that rejecting needs more, basically that we “accept” an alternative model, and this requires an honest, full, Bayesian analysis. However, there are instances where a careful specification of alternatives to a model may be perceived as a waste of time unless the data does cast doubt on its validity. In this spirit, measures of surprise can be a very valuable tool in exploratory Bayesian analyses, and we believe that some such measures can be developed that are more in agreement with the Bayesian approach to inference than the usual classical p -values. Of course, we reiterate that, as Bayesian reasoning tells us, a model should not be rejected because the observed data is “surprising” unless a better explanation can be found.

This Report is organized in 8 Sections, this Introduction being the first one. In Section 2 we review the first surprise indices that were proposed, namely those of Weaver and

Good, which are based on likelihood-ratio comparisons. A different proposal is based in computing p -values, either for the prior predictive distribution (treated in Section 3), or for the posterior predictive distribution (Section 4). In Section 5 we review some other proposals to measure surprise in Bayesian contexts, specially that of Evans. In Section 6 we turn to the likelihood-ratio type of comparison, as formalized in Berger (1980/85); we explore its implications and propose some modifications in terms of explicit considerations of statistics measuring departure from H_0 . When there are nuisance parameters, it is difficult for the measures of surprise to separate between learning about the unknown parameter and detecting surprising features in the behavior of the data. In Section 6 we argue that some conditioning is needed in these situations, but that full conditioning (as is done by posterior predictive p -values) might work poorly. We explore some possibilities and make some proposals. Although likelihood-ratio types of comparison do not average over the sample space and are thus, closer in spirit to Bayesian computations, p -values are usually easier to compute. However, it is quite well known that they grossly overestimate the evidence against H_0 , so that if a p -value is computed, at the very least some kind of adjustment is necessary for it to be safely interpreted as a measure of surprise (or evidence against H_0). In the final Section¹ of this report, we use the global robustness idea to produce a very simple way to calibrate p -values: we basically defend the use of $-ep \log(p)$ as a rough approximation to the infimum of a Bayes factor, and thus as a more suitable measure of evidence against H_0 than the raw p .

2. Weaver's and Good's surprise indices.

The first proposal for a surprise index (other than the usual p -values) seems to have been given by Weaver (1948, 1963). He based his surprise index on the probability $f(x_{obs})$ of observing the data that eventually materialized. He stressed, however, that a small probability is not necessarily surprising unless it is small in comparison with the probability $f(x)$ of the other possible results. He chose to compare $f(x_{obs})$ with the average (expected) probability. His proposal (with a slightly different formulation) is as follows:

¹This Section is joint work with Tom Sellke, from Purdue University.

Suppose that the random variable (or random vector) X has a discrete distribution taking values x_1, x_2, \dots , with probabilities f_1, f_2, \dots , respectively. Then, the *surprise index* associated with the observed value x_{obs} is

$$\lambda_1 = \frac{E[f(X)]}{Pr(X = x_{obs})} = \frac{\sum_i f_i^2}{f_{obs}}.$$

The numerator ($\sum_i f_i^2$) is sometimes referred to as *Gini's index of homogeneity* (Good, 1988), and in fact, these indices of surprise are quite related to measures of diversity (Good, 1982). The generalization to continuous random variables is direct:

$$\lambda_1 = \frac{E[f(X)]}{f(x_{obs})} = \frac{\int f^2(x)dx}{f(x_{obs})}. \quad (2.1)$$

It can easily be seen from (2.1) that Weaver's index of surprise is multiplicative, that is, if X and Y are independent random variables, then

$$\lambda_1(x_{obs}, y_{obs}) = \lambda_1(x_{obs})\lambda_1(y_{obs}). \quad (2.2)$$

An undesirable feature of Weaver's surprise index is that, for continuous X , λ_1 is invariant *only* under linear transformations. Good (1956) addressed this difficulty by relating *surprise* with *simplicity*:

“... Thus the vagueness in the definition ... is seen to arise from the difficulty in measuring simplicity ... Since no one has yet thought of a satisfactory measure of simplicity, it seems unlikely that a really satisfactory measure of surprise can be given.”

Another possible difficulty with (2.1) is that the standard chosen to compare the observed $f(x_{obs})$ with, that is, its expected value $E[f(X)]$, might be considered somewhat arbitrary. Good (1953, 1956) proposed a single-parameter generalization of (2.1) which would also possess the multiplicative property (2.2) (as well as the lack of invariance under non-linear one-to-one transformations). His measures of surprise compare $f(x_{obs})$ with some sort of *geometric expectation*. For $c > 0$:

$$\lambda_c = \frac{\{E[f(X)]^c\}^{1/c}}{f(x_{obs})}. \quad (2.3)$$

Weaver's index λ_1 is, of course, given by (2.3) for $c = 1$. The limiting case, as $c \rightarrow 0$ gives

$$\lambda_0 = \frac{\exp\{E[\log f(X)]\}}{f(x_{obs})}. \quad (2.4)$$

Good (1988) still proposed a further generalization,

$$\frac{\phi^{-1}\{E[\phi(f(X))]\}}{f(x_{obs})},$$

where ϕ is a monotonic increasing function, which is multiplicative only in the case in which ϕ is a power or logarithm (so that it reduces to either (2.3) or (2.4)).

If an *additive* (as opposed to multiplicative) index of surprise is desired, Good proposed to take the logarithm of (2.3),

$$\Lambda_c = \log(\lambda_c), \quad c \geq 0$$

which he called *logarithmic surprise index*, and which has many connections with information theory. In fact, $\Lambda_c + \log(f(x_{obs}))$ is also called *Renyi's generalized entropy* (Renyi, 1961). In particular,

$$\Lambda_1 = \log(E[(f(X))] - \log(f(x_{obs})))$$

$$\Lambda_0 = E[\log(f(X))] - \log(f(x_{obs}))$$

are closely related to the quadratic and logarithmic scoring rules, respectively (Good, 1983, Chapter 14).

Among the infinite possibilities in (2.3), Good (references above) suggested as most "natural" the measures λ_1 , and λ_0 , and their corresponding Λ_1 and Λ_0 for the logarithmic indexes. His arguments were based on properties of the expected indices of surprise before the experiment is performed. Requiring $E[\lambda_c] = 1$ produces $c = 1$, but because of the skewness of the distribution of λ_c , which makes expectations artificial, Good defended as more natural the property that the expected log-surprise should be 0, which is true for Λ_0 . This, together with the close relationship of Λ_0 with measures of entropy and information, caused him to recommend use of Λ_0 as the most natural additive index, and of λ_0 as the most natural multiplicative index.

Example 2.1 Good (1956) computed these surprise indices for the multivariate normal distribution. Assume that $X \sim N_k(\theta, \Sigma)$, where both the mean vector and covariance matrix are assumed known under the null. Then the logarithmic index of surprise is

$$\Lambda_c = \frac{1}{2}(x - \theta)^t \Sigma^{-1}(x - \theta) - \frac{k}{2c} \log(c + 1) \quad ,$$

and, in particular

$$\Lambda_1 = \frac{1}{2}(x - \theta)^t \Sigma^{-1}(x - \theta) - \frac{k}{2} \log 2, \quad \Lambda_0 = \frac{1}{2}(x - \theta)^t \Sigma^{-1}(x - \theta) - \frac{k}{2},$$

where Λ_0 can be seen to be $(D^2 - k)/2$, where D is *Mahalanobis distance*. For the univariate normal ($k = 1$) they reduce to

$$\Lambda_c = \frac{1}{2}[z^2 - \log(c + 1)], \quad \Lambda_1 = \frac{1}{2}[z^2 - \log 2], \quad , \quad \Lambda_0 = \frac{1}{2}[z^2 - 1] \quad ,$$

where $z = |x - \theta|/\sigma$. The corresponding surprise indices are

$$\lambda_1 = 2^{-k/2} e^{\frac{1}{2}(x-\theta)^t \Sigma^{-1}(x-\theta)} \quad \lambda_0 = e^{-k/2} e^{\frac{1}{2}(x-\theta)^t \Sigma^{-1}(x-\theta)} \quad ,$$

and for the univariate normal,

$$\lambda_1 = \frac{1}{\sqrt{2}} e^{z^2/2} \quad \lambda_0 = \frac{1}{\sqrt{e}} e^{z^2/2}$$

Good also provided a table comparing these surprise indices with $1/p$ -values (Table 1).

The corresponding indices for random samples from these distributions are easily derived from the expressions above and the multiplicative property of the λ 's and the additive property of the Λ 's. □

z	0	1	2	3	4	5
$1/p$	1.00	3.10	22.0	370.0	16000	1740000
λ_0	0.61	1.00	4.5	54.6	1800	160000
λ_1	0.71	1.17	5.2	64.0	2100	187000

Table 1: Weaver and Good surprise indices and $1/p$ -values.

In order to compute all of the indices of surprise in this Section, the distribution of the observations under the null has to be completely specified, which is naturally the case

when H_0 is a simple statistical hypothesis. Otherwise, none of the authors is very clear about how to proceed; Good (1988) merely states that if H_0 is not simple, we need to be “sharp Bayesians” to have a sharp measure of surprise.

In spite of their very early introduction, these indices of surprise have not been very popular. The main alternatives among Bayesians to measure surprise, and by far the ones most commonly used, are the tail areas or *Bayesian p-values* corresponding to the observed data and computed for some suitable predictive distribution. The most vocal advocate of Bayesian *p-values* was Box (1980), who computed the tail areas in the prior predictive distribution, and with time, they have become to be called *prior predictive p-values*.

3. Prior predictive *p-values*.

If (under H_0) data $X \sim f(x|\theta)$, and, a priori, $\theta \sim \pi(\theta)$, then for Bayesians, the *prior predictive distribution*,

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta, \tag{3.1}$$

is, of course, always completely specified, and to many, it is the natural tool to quantify surprise. The importance of the prior predictive distribution and its use for model checking have been amply recognized for a long time (see, for instance, Roberts, 1965, Guttman, 1967, Geisser, 1975, and references given in the Introduction). Indeed, $m(x)$ is the probability (or density) of observing data x , so that a small value would indicate data that is unlikely to be observed (under the assumed model/prior specification). If the actual data x_{obs} produces a “small” $m(x_{obs})$, then it seems natural to consider them “surprising”.

In addition, a small value of $m(x_{obs})$ might flag a clash between prior and likelihood, which usually has associated considerable lack of posterior robustness. The argument can be formalized resorting to tools of *local sensitivity* analysis (O’Hagan, 1994).

Of course, when we refer to a “small” value of $m(x_{obs})$, we have to speak in relative terms, since there can not be an absolute measure of “smallness”: if the sample space is very large, then $m(x)$ would typically be very small for all possible sets of data x . Hence,

in order to get a feeling of how small $m(x_{obs})$ is, we have to compare it with some standard. It seems to us that the most natural comparison is in terms of “relative likelihoods”, in which case $m(x_{obs})$ is compared with some “typical” or “likely” $m(x)$. We postpone its discussion till Section 6. We concentrate here on the popular proposal by Box (1980), in which the smallness of $m(x_{obs})$ is measured by computing its associated tail area in the prior predictive $m(x)$.

For Box, an overall predictive check of a given model is provided by

$$\alpha = Pr\{m(X) < m(x_{obs})\}, \quad (3.2)$$

where the probability is computed with respect to the prior predictive distribution (3.1), and $1 - \alpha$ or $1/\alpha$ can be used as measures of surprise. Similarly, the surprise for some relevant checking function, $D(x_{obs})$ can be assessed by computing

$$Pr\{m[D(X)] < m[D(x_{obs})]\}. \quad (3.3)$$

Evans (1997) calls $1 - \alpha$ the *observed surprise*. Good (1988) also proposes to compute the tail areas probabilities associated to his various surprise indices (he refers to it as a surprise/Fisher compromise).

Example 3.1 Assume that $X = (X_1, X_2, \dots, X_n)$ is a random sample from a $N(\theta, \sigma^2)$ distribution, with σ^2 known. Let the prior for θ be $\pi(\theta) = N(\theta|m, v^2)$. Then the predictive distribution of X is

$$m(x) \propto \sigma^{-n-1} \left(\frac{\sigma^2}{n} + v^2 \right) \exp \left\{ -\frac{1}{2} \left[\frac{ns^2}{\sigma^2} + \frac{(\bar{x} - m)^2}{n^{-1}\sigma^2 + v^2} \right] \right\}, \quad (3.4)$$

where $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. It can be seen from (3.4) that $m(x)$ is a decreasing function of

$$T(x) = \frac{ns^2}{\sigma^2} + \frac{(\bar{x} - m)^2}{n^{-1}\sigma^2 + v^2}, \quad (3.5)$$

and $T(X)$ has a χ_n^2 distribution. Therefore, the overall predictive check (3.2) can be computed as

$$Pr\{m(X) < m(x_{obs})\} = Pr\{T(X) > T(x_{obs})\} = Pr\{\chi_n^2 > T(x_{obs})\}. \quad (3.6)$$

Partial predictive checks are obtained in a similar way from the predictive distributions of $D_1(X) = \bar{X}$, $D_2(X) = S^2$, and $D_3(X) = (R_1, R_2, \dots, R_{n-2})$, a vector of residual

quantities defined as

$$R_j = X_{j+1} / \left(\sum_{i=1}^j X_i^2 / j \right)^{1/2} .$$

with predictive distributions:

$$\begin{aligned} m(d_1) &= N(d_1 | m, v^2 + \frac{\sigma^2}{n}) , \\ m(d_2) &= Ga(d_2 | \frac{n-1}{2}, \frac{n}{2\sigma^2}) , \\ m(d_3) &\propto \prod_{i=1}^{n-2} \left(1 + \frac{r_i^2}{i} \right)^{-(i+1)/2} , \end{aligned} \tag{3.7}$$

D_1 and D_2 being independent. □

The main difficulty that we find with these measures of surprise is that they are, in spirit, very close to classical p -values in that they average over non observed values of X , namely, over *all* values of X that are less compatible with the assumed model than the observed x_{obs} . Thus they not only violate the Conditionality Principle (and hence the Likelihood Principle) but they also are based on values of X that provide a much stronger evidence against the null model than the observed one, thus providing an exaggerated measure of surprise.

Another undesirable feature of prior predictive p -values is that they are not invariant under one-to one transformations either, as the following example (taken from Evans, 1997) demonstrates:

Example 3.2 Assume that the prior predictive distribution is

$$m_X(x) = 2x \quad \text{for } 0 < x < 1 ,$$

so that surprise in x_{obs} would be measured according to

$$Pr^X[m_X(X) \geq m_X(x_{obs})] = 1 - x_0^2 ,$$

which equals 1 when $x_{obs} = 0$, and equals 0 when $x_{obs} = 1$. Now assume that we make the one-to-one transformation $Y = x^4$, then it follows that the predictive density of Y is

$$m_Y(y) = \frac{1}{\sqrt{y}} ,$$

so that the surprise in y_{obs} would be measured in

$$Pr^Y[m_Y(Y) \geq m_Y(y_{obs})] = \sqrt{y_{obs}} \quad ,$$

which, in complete contradiction with the previous result, equals 0 when y_{obs} (and hence x_{obs}) equals 0, and equals 1 when y_{obs} (and hence x_{obs}) equals 1. \square

Box (1980) recognized this lack of invariance, but he did not think it to be too crucial:

“These probabilities are, of course, affected by transformations ... I do not find this particularly disturbing. Slightly different questions can be expected to have slightly different answers.”

Also, prior predictive p -values can be completely useless (in the sense of providing a surprise of 0 for all observed values), and can also depend on the stopping rule used:

Example 3.3 Assume that Z_1, Z_2, \dots are a sequence of Bernoulli trials with probability of success θ . Assume also that the prior distribution for θ is uniform on $(0, 1)$, and that in n trials we observe $x_{obs} = \sum_{i=1}^n z_i$.

1. If n is assumed fixed, so that X is binomial, it can easily be checked that

$$m(x) = \frac{1}{n+1} \quad \text{for } x = 0, 1, \dots, n \quad ,$$

which is constant, and therefore useless for surprise purposes.

2. If x is fixed, so that $N - x$ is negative binomial, it can be checked that the prior predictive of n is

$$m(n) = \frac{y}{n(n+1)} \quad \text{for } n = x, x+1, \dots \quad ,$$

which is decreasing with n so that large values of n_{obs} would be considered surprising. Moreover, Box's index of surprise, $1 - \alpha$ is given by

$$Pr(N \leq n_{obs}) = x/n_{obs} = \hat{\theta}$$

so that the larger the MLE $\hat{\theta}$ the more surprising the data is. \square

If interest lies in checking a statistic $D = D(X)$, then some of these difficulties can be removed by directly computing a p -value for D instead of computing the p -value for $m(D)$. Indeed, the prior predictive p -value

$$P_{r^{m(x)}}\{D(X) > D(x_{obs})\}$$

may, in fact, be more often used than (3.3), Box's original proposal, and it is invariant under one-to-one transformations.

Another concern is that prior predictive p -values (both versions) are not defined if the prior $\pi(\theta)$ is not proper, since in this case the prior predictive $m(x)$ is also improper, as can easily be seen from (3.1). The concern is of some importance in surprise analyses, not only because improper priors are often used in preliminary investigations, but most importantly because, with a proper prior, all we can check is the joint model, $f(x|\theta)\pi(\theta)$. It is thus, most natural to use a non-informative prior for θ if the goal is to check the adequacy of $f(x|\theta)$. If p -values on a predictive distribution are still desired, a possibility that have been proposed is to use the *posterior predictive p-values*.

4. Posterior predictive p -values.

A very useful property of posterior predictive distributions,

$$m(z|x_{obs}) = \int f(z|\theta)\pi(\theta|x_{obs})d\theta, \tag{4.1}$$

when used as a tool for model checking, is that, unlike the prior predictive distributions (3.1) they are always proper, even when the prior $\pi(\theta)$ is improper (provided, of course, that the posterior distribution $\pi(\theta|x_{obs})$ is proper).

The first proposal for model checking based on posterior predictive distributions seems to be that of Guttman (1967). His proposal was to compare the observed empirical frequencies in a partition of the sample space with the "theoretical" frequencies computed from the posterior predictive distribution of a future observation. The comparison was carried out by means of a χ^2 -type of procedure and "surprise" was based on p -values.

A formal, general proposal of the use of posterior p -values is due to Rubin (1984), who argued for the use of tail area probabilities (computed for the posterior predictive distribution) corresponding to the observed value of some test statistic, $T(X)$:

$$p_B = Pr\{T(X) \geq T(x_{obs})|x_{obs}\}, \quad (4.2)$$

where the probability is computed with respect to the posterior predictive distribution $m(x|x_{obs})$ in (4.1).

Posterior predictive p -values have been further developed by Meng (1994) and Gelman, Meng and Stern (1996), by allowing T to also depend on the parameter θ , and by allowing conditioning in the value of some “auxiliary” statistic $A(x)$:

$$p_B = Pr\{T(X, \theta) \geq T(x_{obs}, \theta)|x_{obs}, A(x_{obs})\}, \quad (4.3)$$

where the probability is now computed with respect to the joint distribution:

$$p(\theta, x|x_{obs}, A(x_{obs})) = f(x|\theta, A(x) = A(x_{obs}))\pi(\theta|x_{obs}).$$

The idea of using posterior predictive distributions in lieu of prior predictive distributions is, in a sense, very similar to the motivating idea in Aitkin’s posterior Bayes factors (Aitkin, 1991); posterior predictive p -values are derived in absence of alternative models and use posterior predictive distributions to compute tail areas, whereas in Atkins scenarios, which include explicit specification of alternative models, posterior predictive distributions are used to compute Bayes factors. We might expect, therefore, similar difficulties in both approaches.

Example 4.1 Assume the same conditions as in Example 3.1, that is, a random sample from a normal distribution with unknown mean θ and known variance σ^2 , but let the prior for θ be the usual non-informative (improper) prior $\pi(\theta) \propto 1$. Then the posterior distribution is $N(\bar{x}, \sigma^2/n)$, and the posterior predictive distribution of X is given by (3.4) with $m = \bar{x}$, and $v^2 = \sigma^2/n$. Then, if we were to take the same test statistic as in Example 3.1, which in this case reduces to:

$$T(X) = \frac{n}{\sigma^2} \left[S^2 + \frac{1}{2}(\bar{X} - \bar{x}_{obs})^2 \right] ,$$

we would have the same expression for the measure of surprise as in (3.6) for the particular distribution for θ used here.

Notice however that, since $T(x_{obs}) = (ns_{obs}^2)/\sigma^2$ which is the minimum of (??) over all m and v^2 , the effect of using the posterior distribution as the distribution for θ with respect to which compute the predictive of T is to make the p -value (3.6) as *large* as

possible, that is, to make the data as unsurprising as possible. We defer discussion of other test statistics until Section 6. \square

A very attractive characteristic of posterior predictive p -values (4.2) (and its generalizations) is that they are very easy to compute with the usual outputs from Bayesian numerical analyses. Thus, for example, if MCMC is being used to carry out the desired analysis, resulting in a sequence of generated values of θ , it suffices to add a new step in the recursive MCMC simulation in which x is generated from $f(x|\theta)$: at the end of the run, we have a sample from the posterior predictive distribution $m(x|x_{obs})$ from which the desired tail areas are easily approximated. However, we do not find them to be completely convincing from a Bayesian perspective for two reasons:

Our first objection is to the use of the *full* posterior predictive distribution $m(x|x_{obs})$ as the distribution of X to be used to locate the observed x_{obs} (or $T(x_{obs})$ for that matter). It appears that data x_{obs} is being used twice: first to *train* the improper prior $\pi(\theta)$ into a proper distribution $\pi(\theta|x_{obs})$, and then for measuring “surprise” when locating the observed x_{obs} (or $T(x_{obs})$) in the predictive $m(x|x_{obs})$.

Defenders of posterior p -values will argue that $m(x|x_{obs})$ refers to the predictive distribution of *future replications* of the experiment under the same conditions, but it seems to us that this interpretation does not remove the difficulty, and we do agree with Kass and Wasserman (1991) when they comment on the “oddity of using hypothetical replications of data in place of data at hand”.

One possible effect of “too much” use of the observed data is inadequate behavior of posterior p -values as illustrated in the following simple example, which demonstrates that, for a fixed sample size, posterior p -values may not go to 0 as the observations become infinitely “extreme”.

Example 4.2 Assume that, under the null, X_i are iid $N(0, \sigma^2)$, with σ^2 unknown. Assume we take $T = T(X) = \bar{X}$ (which would be natural if we worry most about the location). With the usual non-informative prior for σ^2 : $\pi(\sigma^2) \propto 1/\sigma^2$, the posterior distribution is

$$\pi(\sigma^2|x_{obs}) = Ga^{-1}\left(\frac{n}{2}, \frac{v_{obs}^2}{2}\right) ,$$

where $v_{obs}^2 = v^2(x_{obs})$, and $v(x) = \sum_{i=1}^n x_i^2$. The posterior predictive distribution of a new

$T = \bar{X}$ is

$$\begin{aligned}
m(t|x_{obs}) &= \int N(t|0, \frac{\sigma^2}{n}) Ga^{-1}(\sigma^2|\frac{n}{2}, \frac{v_{obs}^2}{2})d\sigma^2 \\
&\propto \int \frac{1}{(\sigma^2)^{(n+3)/2}} \exp\{-\frac{1}{2\sigma^2}(v_{obs}^2 + nt^2)\}d\sigma^2 \\
&\propto (v_{obs}^2 + nt^2)^{-(n+1)/2} \propto St(t|0, \frac{v_{obs}^2}{n^2}, n) .
\end{aligned} \tag{4.4}$$

Hence, the posterior p -value is given by

$$p_B = Pr\{|\bar{X}| > |\bar{x}_{obs}|\} = 2 \left[1 - \Upsilon_n \left(\frac{\sqrt{n}\bar{x}_{obs}}{\sqrt{v_{obs}^2/n}} \right) \right] , \tag{4.5}$$

where Υ_n stands for the distribution function of a Student t with n degrees of freedom. Since $v(x) = \sum_{i=1}^n x_i^2 = nS^2 + n\bar{x}^2$, where $S^2 = \sum(x_i - \bar{x})^2/n$, then it follows from (4.5) that, as $|\bar{x}_{obs}| \rightarrow \infty$,

$$p_B = \text{posterior } p\text{-value} \rightarrow 2[1 - \Upsilon_n(\sqrt{n})] ,$$

which is a constant > 0 for any given fixed sample size n . (Note that Meng, 1994, also used this example as an illustration; (4.5) is equivalent to his (3.4).) \square

The following example (inspired by Goldstein, 1991) also demonstrates the inadequacies of using posterior predictive distributions to assess surprise.

Example 4.3 Suppose that a random number between 1 and 1000 is generated and not revealed. This is the unknown θ , and the prior predictive is uniform on the integers $\{1, 2, \dots, 1000\}$. For data x we shall assume two different models: Under model 1, the data consists in revealing the value of θ ($x = \theta$) while under model 2, the value of θ is ignored and a new random number x is generated. Notice that data gives *no* information whatsoever that allows us to distinguish between the two models, and, as a matter of fact, prior predictive distributions $m_1(x)$ and $m_2(x)$ are identical. Posterior predictive distributions are, however, very different: the posterior predictive distribution under model 1 is degenerated at $y = y_{obs}$, while that under model 2 is uniform on $\{1, 2, \dots, 1000\}$. \square

We do not object to some conditioning when selecting the predictive distributions to use, as long as the data do not get used "twice", or, if a different (but equivalent) phrasing is desired, if resorting to "hypothetical" replications is not needed to assess the

surprisingness of the data at hand. An example of some “admissible” conditioning of this type is provided by *cross-validation*, in which part of the data is used to train the prior and *the rest* of the data to assess the degree of surprise (Geisser, 1980; O’Hagan 1991; Gelfand, Dey and Chang, 1992; Gelfand and Dey, 1994). Another example is the use of the predictive of data at time t given the previous data for checking model adequacy in times series (West, 1986; West and Harrison, 1997). Other proposals will be discussed in Section 7..

Our main objection to the use of posterior p -values is that they are too similar to classical p -values, and hence inherit the considerable inadequacies of the latter. To see this, notice from (4.2) that the posterior predictive p -value can be expressed as:

$$p_B = \int Pr\{T(X) \geq T(x_{obs})|\theta\}\pi(\theta|x_{obs})d\theta \quad , \quad (4.6)$$

that is, the posterior predictive p -value, p_B is nothing more than the expected value of the classical tail probability

$$p_c(\theta) = Pr\{T(X) \geq T(x_{obs})|\theta\} \quad ,$$

with respect to the posterior distribution. It follows that, for large sample sizes, $p_B \approx p_c(\hat{\theta})$, where, say, $\hat{\theta}$ is the MLE of θ , which is the classical estimated p -value, and the behavior of both measures will be similar. As a matter of fact, when H_0 completely specifies the distribution of data (that is, when there is no θ under H_0), then both the posterior predictive p -value and the classical p -value are *identical*. For discussion of the very misleading nature of classical p -values, see Berger and Sellke (1987), Berger and Delampady (1987), and the references therein. Note, also, that the same problems hold in cases of “model fit”: see Delampady and Berger (1990).

5. Other proposals to measure surprise.

The most comprehensive development of a measure of surprise for use in Bayesian analyses other than the ones treated previously is due to Evans (1997). One of his main motivations was to provide a measure of surprise that, unlike the ones proposed by Box and Good,

was invariant under one to one transformations.

His proposal is as follows: Assume that $\varphi = \varphi(\theta)$ is the parametrical function of interest. Then, Evans defines the *observed relative surprise* for testing the null hypothesis $H_0 : \varphi = \varphi_0$ against each of the possible alternative values φ in H_1 as:

$$Pr \left\{ \frac{\pi(\varphi|x_{obs})}{\pi(\varphi)} > \frac{\pi(\varphi_0|x_{obs})}{\pi(\varphi_0)} \right\}, \quad (5.1)$$

where the probability is computed with respect to the posterior distribution $\pi(\varphi|x_{obs})$. Evans proposed using (5.1) not only for hypothesis testing, but also for estimation (minimizing the observed relative surprise) and for confidence regions (α -relative surprise regions).

It can easily be seen from (5.1) that Evans' measure of surprise is invariant under one to one transformations (since the Jacobians occur in both numerator and denominator, and hence cancel). Another motivation that Evans gives in favor of (5.1) as an inferential tool is that "inferences are determined by how our beliefs change from a priori to a posteriori".

We see, however, two basic difficulties with the use of (5.1) as a measure of surprise. First, it is, again, using the data twice: once to produce the ratio of the posterior to the prior which is the basic quantity in this approach, then again when the posterior distribution is used to compute the probability that this ratio is larger than its hypothesized value.

The following argument further indicates this double use of the data by relating Evans' relative surprise (ERS) to Atkins posterior Bayes factors. It easily follows from (5.1) that ERS can alternatively be written as:

$$Pr^{\varphi|x_{obs}} \left\{ \frac{f(x_{obs}|\varphi)}{f(x_{obs}|\varphi_0)} > 1 \right\} = Pr^{B|x_{obs}} \{B > 1\}.$$

That is, ERS is nothing but the tail area above 1 of the posterior distribution of the Bayes factor B in favor of φ against φ_0 . Notice that the *expected value* of that distribution:

$$\frac{\int f(x_{obs}|\varphi)\pi(\varphi|x_{obs})d\varphi}{f(x_{obs}|\varphi_0)},$$

is nothing but Atkins posterior Bayes factor for H_1 .

However, the main reason why we would not adopt (5.1) as a measure of surprise is that it requires *both* the careful assessment of the alternatives to φ_0 and the complete specification of a prior distribution over such alternatives. Once these inputs are fully specified, it seems more natural (and preferable) to us to proceed in an entirely Bayesian fashion, and hence the need for new inferential methods based on such inputs is not very clear to us.

Evans (1997) also used the idea of relative surprise for model checking. In this case, his observed relative surprise, in the notation of previous sections, takes the form:

$$Pr \left\{ \frac{m(T(X)|x_{obs})}{m(T(X))} > \frac{m(T(x_{obs})|x_{obs})}{m(T(x_{obs}))} \right\}, \quad (5.2)$$

which, again, avoids the problem of the lack of invariance. Also, the definition (5.2) allows for the relative surprise method to be used for prediction as well. However, (5.2) is not very useful for model checking, as Evans (1997) pointed out, since the value of $m(T(x_{obs})|x_{obs})/m(T(x_{obs}))$ is likely to be large or even maximal, thus producing a measure of surprise (5.2) equal to 0 and hence useless. Evans proposes to then apply cross-validation ideas to his basic measure.

Martin and Meeden (1984) proposed a very attractive measure of surprise as a distance between the prior and the posterior distributions. They explored the Hellinger and the Kullback-Leibler distances. Their measure, however, requires the specification of the alternative hypothesis and a prior over it.

Good (1988) proposed still another measure of surprise related to complexity. Recall that his logarithmic measure of how surprising the occurrence of E_0 is, was given by

$$\Lambda_0 = -\log Pr(E_0) - \{-E[\log Pr(E_j)]\}.$$

His new measure replaces the entropy term $-E[\log Pr(E_j)]$ in Λ_0 by a *complexity term*:

$$S(E_0) = -\log_{10} Pr(E_0) - \chi(E_0),$$

where (quote) “ $\chi(E_0)$ is an additive measure, in decimal units, of the complexity of the part of E_0 that goes beyond what is known to follow from H_0 ”.

A basically different use of the concept of surprise is the one by Shackle (1949) who used the concept of *potential surprise* instead of subjective probability, claiming that the later is psychologically more basic. His development is similar to that of Bayesian decision theory, in that he proposed to order the degrees of beliefs by the potential degrees of surprise. Also, he considered that the “interestingness” of an outcome was a function of its desirability and of its potential surprise, and that people, when deciding on an action, usually concentrate on two “focus outcomes” of maximum interestingness, one desirable and the other undesirable. Krelle (1957) argued that there is a one to one relationship between potential surprise and subjective probability.

6. Relative maximized and expected measures of surprise.

Berger (1980/85) proposed two measures of surprise that are in the same spirit as the ones of Good and Weaver, in that they compare relative likelihoods, instead of computing p -values. His proposals were made in the context of global robustness analyses, and thus they were originally meant to flag a problem with the prior, but since the roles of prior and likelihood are symmetrical in the predictive distribution, they have an identical role to play in flagging a misspecification of the likelihood.

In Berger’s proposals, the key factor is, again, the prior predictive distribution $m(x)$, and data is surprising if $m(x_{obs})$ is “small”. Specifically, he proposed to compute one of the two relative likelihoods:

$$m^*(x_{obs}) = \frac{m(x_{obs})}{\sup_x m(x)} \quad , \quad (6.1)$$

or

$$m^{**}(x_{obs}) = \frac{m(x_{obs})}{E^{m(x)}[m(X)]} \quad . \quad (6.2)$$

These measures “seek to provide a base for comparison of the size of $m(x_{obs})$ not by averaging $m(x)$ over ‘extreme’ x that did not occur ... but by directly seeing if the observed x_{obs} is unusual compared to the ‘most likely’ x or the ‘average’ x , respectively”.

It can directly be seen from (6.2) that the surprise measure m^{**} is nothing but the

inverse of Weaver's and Good's index λ_1 as given in (2.1) when applied to the prior predictive distribution $m(x)$, and hence it shares the very same properties. Also, it follows from (2.3), that (under usually met regularity conditions) m^* is the limiting case, as $c \rightarrow \infty$, of the inverse of Good's surprise index λ_c .

Example 6.1 Assume that X is multivariate normal, $X|\theta \sim N_k(\theta, \Sigma)$, with Σ known, and that a priori, $\theta \sim N_k(m, V_0)$. Then, the marginal (predictive) distribution of X , $m(x)$ is given by

$$X \sim N_k(m, V_0 + \Sigma) = N_k(m, V_1) \quad . \quad (6.3)$$

Then, since the maximum value of the density (6.3) is

$$\sup_x N_k(x|m, V_1) = N_k(m|m, V_1) = (2\pi)^{-k/2} (\det V_1)^{-1/2} \quad ,$$

and its expected value,

$$\begin{aligned} E^{m(x)}[N_k(X|m, V_1)] &= \frac{1}{(2\pi)^k |V_1|} \int \dots \int \exp\left\{-\frac{1}{2}(x-m)^t (2V_1^{-1})(x-m)\right\} dx \\ &= \frac{1}{2^{k/2}} \frac{1}{(2\pi)^{k/2} |V_1|^{1/2}} \quad , \end{aligned}$$

it follows that (6.1) and (6.2) are given by

$$m^*(x_{obs}) = \exp\left\{-\frac{1}{2}(x_{obs}-m)^t V_1^{-1}(x_{obs}-m)\right\} \quad , \quad (6.4)$$

$$m^{**}(x_{obs}) = 2^{k/2} m^*(x_{obs}) \quad . \quad (6.5)$$

Of course, $m^{**}(x_{obs})$ is nothing but the inverse of Good and Weaver's index λ_1 as given in Example 2.1 when applied to the predictive (6.3). The particular case in which X is a random sample of size n from a univariate $N(\theta, \sigma^2)$ obtains from the argument above when $\Sigma = \sigma^2 I$. With $\theta \sim N(m, v_0^2)$, and $v_1^2 = v_0^2 + \sigma^2$, we have

$$m^*(x_{obs}) = e^{-\sum z_i^2/2} \quad \text{and} \quad m^{**}(x_{obs}) = 2^{n/2} m^*(x_{obs}) \quad , \quad (6.6)$$

where $z = \frac{x-m}{v_1}$. □

Example 6.2 Let X_1, X_2, \dots, X_n be a random sample of size n from an Exponential distribution with parameter λ (mean = $1/\lambda$). If the prior distribution of λ is $Ga(a, b)$, then the (joint) predictive distribution is

$$m(x) = \frac{?(n+a)}{(b+\sum x_i)^{n+a}} \frac{b^a}{?(a)} \quad . \quad (6.7)$$

Since (6.7) is decreasing with $\sum x_i$, it follows that

$$m^*(x) = \left(\frac{b}{b + \sum x_i} \right)^{n+a},$$

Also, it can be checked that

$$E^{m(x)}[m(X)] = [?(n+a)]^2 \int \dots \int \frac{1}{(b + \sum x_i)^{2(n+a)}} \frac{b^a}{?(a)} dx_1 \dots dx_n = \frac{[?(n+a)]^2}{?[2(n+a)]} \frac{?(n+2a)}{b^n [?(a)]^2},$$

so that

$$m^{**}(x) = \frac{?[2(n+a)]}{?(n+a)} \left(\frac{b}{b + \sum x_i} \right)^{n+a} \frac{?(a)}{?(n+2a)}.$$

□

The surprise measure (6.1) seems to be in agreement with the intuition of many Bayesians. As early as 1965, Roberts [43], when commenting on a proposal by Anscombe about using tail-areas of the (prior) predictive distributions of goodness-of-fit test statistics, suggested a “more nearly Bayesian approach”:

“For example, we might compare the actual predictive probability [of the observed sample] against the maximum probability that might have been observed. Such a procedure can be proposed only as a rough guide for and intuitive judgment about ‘surprisingness’, and defended only because it seems closer to Bayesian concepts than do tail-area computations of criterion statistics.”

Other authors have also suggested its use, as for example O’Hagan (1994) and Draper (1996). Pettit (1990) and Gelfand, Dey and Chang (1992) use a similar idea in the cross-validation predictive distribution $m(x_i|x_{-i})$.

A very interesting property of the measure of surprise m^* is its close connection with the robust Bayes approach outlined in the Introduction. As a matter of fact, if H_0 is simple (that is, there is no θ in the formulation) and $?$ is the class of *all* prior distributions $\pi_1(\xi)$ for the alternative values ξ , then it can be shown that the infimum of the Bayes factor (1.1) for the observed data is given by

$$\underline{B} = \frac{f(x_{obs})}{\sup_{\xi} f_1(x_{obs}|\xi)}. \quad (6.8)$$

Then, if ξ is a location parameter, it follows from (6.1), (6.8) and the fact that $f_1(x|\xi) = f(x - \xi)$, that the measure of surprise m^* is in fact equal to the infimum of the Bayes factor \underline{B} .

Unfortunately, these measures of surprise also suffer from a lack of invariance under non-linear, one-to-one transformations. Also, they might be difficult to interpret if the dimension is very large, or, in particular, if the number of observations n in a random (univariate) sample is very large. Berger (1985) caution us against the use of m^* in these situations, remarking that, in high dimensions, m^* can be very far from Bayes factors. The same, of course, is true of infimums of Bayes factors over too large a class of priors (Bayarri and Berger, 1994).

The problem, as it has been largely recognized in several Bayesian scenarios (Bayesian robustness, empirical Bayes, etc.) is that taking supremums (or expectations) over enormous spaces usually results in bad procedures. This can immediately be seen for the measure m^* when applied to data x that is (marginally) independent under the null as, for instance, when H_0 is a point null. In that case, as $n \rightarrow \infty$,

$$m^*(x) = \prod_{i=1}^n \frac{f(x_i)}{f(x_{max})} \longrightarrow 0, \quad \text{with probability 1,}$$

even when the observations arise from the correct model.

Also, the predictive distribution of the observed, raw, data, can not be expected to detect *all* possible departures from the specified model. Although we are avoiding the careful specification of explicit alternatives to the null model, we do propose (in the same spirit as many of the authors mentioned in this Report) to use a “natural”, low dimensional, statistic T , which will measure, the “distance” between the data and H_0 and apply the measures m^* or m^{**} to its predictive distribution (we shall further elaborate in the next Section). The use of “natural” T will, hopefully, alleviate the problem of the lack of invariance and the impact of high dimensions.

The next example makes explicit the inadequate dependence of the “raw” measures (that is, as applied to the raw vector of observations) with n , in contrast to their nice behavior when they are computed for a suitable statistic T .

Example 6.3 In the situation of Example 6.1, assume that $X = (X_1, X_2, \dots, X_n)$ is a

random sample from a $N(\theta, \sigma^2)$ distribution, with $\theta \sim N(m, v_0^2)$. Since σ^2 is assumed known, a natural statistic T to consider is $T = \bar{X}$. Then, since

$$m^*(\bar{x}_{obs}) = \exp\left\{-\frac{n}{2(\sigma^2 + nv_0^2)}(\bar{x}_{obs} - m)^2\right\} , \quad (6.9)$$

it follows from (6.6) that

$$m^*(x_{obs}) = \exp\left\{-\frac{ns_{obs}^2}{2(\sigma^2 + v_0^2)}\right\} m^*(\bar{x}_{obs}) , \quad (6.10)$$

where $s_{obs}^2 = S^2(x_{obs})$, and $S^2(x) = \sum_{i=1}^n (x_i - \bar{x})^2$. From (6.9) and (6.10) it follows that $m^*(x_{obs})$ is, in a sense, about $\exp\{-\frac{n}{2}\}$ too small. Indeed, if the null model is correct and $n \rightarrow \infty$, $m^*(\bar{x}_{obs}) \rightarrow 1$, while $m^*(x_{obs})$ goes to 0 at the very fast $\exp\{-n/2\}$ rate.

A similar effect occurs with m^{**} . It follows from (6.5), (6.6), and (6.10), that

$$\begin{aligned} m^{**}(\bar{x}_{obs}) &= \sqrt{2} m^*(\bar{x}_{obs}) \text{ and} \\ m^{**}(x_{obs}) &= 2^{n/2} \exp\left\{-\frac{ns_{obs}^2}{2(\sigma^2 + v_0^2)}\right\} m^*(\bar{x}_{obs}) . \end{aligned}$$

Hence,

$$m^{**}(x_{obs}) = \frac{1}{\sqrt{2}} \left(\frac{2}{e^{s_{obs}^2/\sigma^2}}\right)^{n/2} m^{**}(\bar{x}_{obs}) ,$$

which, again, is about $(2/e)^{n/2}$ too small. As before, under the null model, and as $n \rightarrow \infty$, $m^{**}(\bar{x}_{obs})$ goes to $\sqrt{2}$, whereas $m^{**}(x_{obs})$ goes to 0 at a the fast $(1.36)^{-n/2}$ rate.

On the other hand, $m^*(x_{obs})$ and $m^{**}(x_{obs})$ take into account features of the data other than \bar{x} and it might, thus, be better measures for surprise purposes in some instances. A possible, ‘‘ad hoc’’ method to adjust for the dimension effect could be to consider $[m^*(x_{obs})]^{1/n}$ or $[m^{**}(x_{obs})]^{1/n}$ as measures of surprise. We shall not explore these possibilities here. □

The following examples explore different choices for the statistic T (reflecting different ways of measuring departures from the null) and some consequences derived from those choices.

Example 6.4 Let $X = (X_1, X_2, \dots, X_n)$ be a random sample from a $N(\theta_0, \sigma^2)$, where θ_0 is a specified value and σ^2 is known. (We would encounter such a model when testing $H_0 : \theta = \theta_0$.)

Assume first that we take as our “departure” statistic $T_1 = \frac{\sqrt{n}}{\sigma} |\bar{x} - \theta_0|$. Then, $T \sim f_1(t)$, where $f_1(t) = e^{-t^2/2} \sqrt{2/\pi}$, whose maximum is $\sqrt{2/\pi}$. Also it can be checked that $E^{f_1(t)}[f_1(T)] = 1/\sqrt{\pi}$, so that

$$m_1^*(t_{obs}) = e^{-t_{obs}^2/2} \quad \text{and} \quad m_1^{**}(t_{obs}) = \sqrt{2} m_1^*(t_{obs}) \quad .$$

However, the effect of one-to-one transformations can be quite dramatic. For instance, if we consider $T_2 = T_1^2$ then the distribution of T_2 is χ^2 with 1 degree of freedom, and it can be checked that both $\sup_t f_2(t)$ and $E^{f_2(t)}[f_2(T)]$ equal infinity, so that neither m^* nor m^{**} is defined.

Last, we consider $T_3 = T_1^{1/3}$. Then, an easy derivation gives:

$$\begin{aligned} f_3(t) &= \frac{6}{\sqrt{2\pi}} t^2 e^{-t^6/2} \quad , \\ \sup_t f_3(t) &= \frac{6}{\sqrt{2\pi}} \left(\frac{2}{3e}\right)^{1/3} \quad \text{at } t^6 = \frac{2}{3} \quad , \\ E^{f_3(t)}[f_3(T)] &= \frac{6}{\pi} ? \left(\frac{5}{6}\right) \quad , \end{aligned}$$

so that

$$\begin{aligned} m_3^*(t_{obs}) &= \left(\frac{2}{3e}\right)^{-1/3} t_{obs}^2 e^{-t_{obs}^6/2} = \left(\frac{2}{3e}\right)^{-1/3} t_{obs}^2 m_1^*(t_{obs}^3) \quad , \\ m_3^{**}(t_{obs}) &= \left(\frac{2}{3e}\right)^{1/3} m_3^*(t_{obs}) = \frac{t_{obs}^2}{\sqrt{2}} m_1^{**}(t_{obs}^3) \quad . \end{aligned}$$

□

Example 6.5 Assume that, under the null, X is k -variate Normal, $X \sim N_k(\theta_0, \Sigma)$, where, as in Example 6.4, θ_0 and Σ are considered known. The measures of surprise for the sample in general were derived in Example 6.1, with the obvious change of notation. Assume instead that, as a measure of “departure” from the null, we use $T_1 = (X - \theta_0)^t \Sigma^{-1} (X - \theta_0)$. Then $T \sim \chi_k^2$, whose maximum density, achieved at $(k - 2)$, is finite for $k \geq 2$ and is given by

$$\sup_t f_1(t) = \frac{(1/2)^{k/2}}{? (k/2)} \max\{(k - 2), 1\}^{(k/2)-1} e^{-\frac{k-2}{2}} \quad ,$$

so that, for $k \geq 2$,

$$m_1^*(t_{obs}) = \left(\frac{e}{\max\{(k - 2), 1\}} \right)^{(k/2)-1} t_{obs}^{(k/2)-1} e^{-t_{obs}/2} \quad .$$

Recall that $m^*(x)$ was given by only the last factor in the expression above. Similarly, for $k \geq 2$,

$$E^{f_1(t)}[f_1(T)] = \left[\frac{(1/2)^{k/2}}{?(k/2)} \right]^2 \int t^{k-2} e^{-t} dt = \frac{1}{2^k} \frac{?(k-1)}{[?(k/2)]^2} ,$$

hence, for $k \geq 2$,

$$m_1^{**}(t_{obs}) = \frac{2^{k/2}?(k/2)}{?(k-1)} t_{obs}^{(k/2)-1} e^{-t_{obs}/2} .$$

Assume now that we consider the (quite natural) function $T_2 = \sqrt{T}$. Then its density is given by

$$f_2(t) = 2 \frac{(1/2)^{k/2}}{?(k/2)} t^{k-1} e^{-t^2/2} , \quad (6.11)$$

whose maximum is achieved at $\sqrt{k-1}$, and it is finite for all k , giving

$$m_2^*(t_{obs}) = \left(\frac{e t_{obs}^2}{\max\{(k-1), 1\}} \right)^{(k/2)-1} e^{-t_{obs}^2/2} \approx \left(\frac{e}{k-1} \right)^{1/2} t m_1^*(t_{obs}^2) ,$$

the last (approximate) equality being valid obviously only for $k \geq 2$.

The expected value of (6.11) is also finite for all k and is given by

$$E^{f_2(t)}[f_2(T)] = 2 \left(\frac{(1/2)^{k/2}}{?(k/2)} \right)^2 ? \left(\frac{2k-1}{2} \right) ,$$

so that

$$m_2^{**}(t_{obs}) = \frac{?(k/2)}{?(k-\frac{1}{2})} 2^{k/2} t_{obs}^{k-1} e^{-t_{obs}^2/2} = \frac{?(k-1)}{?(k-\frac{1}{2})} t_{obs} m_1^{**}(t_{obs}^2) ,$$

where, again, the last equality does not apply for $k = 1$. □

In a sense, T is a surrogate for the careful specification of alternative hypotheses. The following examples make the connection explicit and explore the consequences. (From now on, we content ourselves with the computation of m^* .)

Example 6.6 Assume that, under the null hypothesis, X_i , for $i = 1, 2, \dots, n$ are iid $Un(0, 1)$. Let $T = T(X) = X_{(1)} + X_{(n)}$, where, as usual, $X_{(k)}$ denotes the k -th order statistic. Such a T could indicate that we might feel doubtful about the location of the distribution. Since the joint distribution of $(X_{(1)}, X_{(n)})$ is given by

$$f(x_{(1)}, x_{(n)}) = n(n-1)(x_{(n)} - x_{(1)})^{n-2} ,$$

it follows that the distribution of T is

$$\begin{aligned} f(t) &= \frac{n}{2} t^{n-1} \quad \text{for } 0 < t < 1 \quad , \\ &= \frac{n}{2} (2-t)^{n-1} \quad \text{for } 1 < t < 2 \quad . \end{aligned} \quad (6.12)$$

It is immediate that the supremum of (6.12) is $\frac{n}{2}$, occurring at $t = 1$, so that

$$\begin{aligned} m^*(t_{obs}) &= t_{obs}^{n-1} \quad \text{for } 0 < t_{obs} < 1 \quad , \\ &= (2 - t_{obs})^{n-1} \quad \text{for } 1 < t_{obs} < 2 \quad . \end{aligned} \quad (6.13)$$

The p -value (since there is no θ , all p -values: prior predictive, posterior predictive and classical, are the same), computed as the two sided tail area exceeding $|t_{obs} - 1|$ has a similar behavior and is given by

$$\begin{aligned} p &= t_{obs}^n \quad \text{for } 0 < t_{obs} < 1 \quad , \\ &= (2 - t_{obs})^n \quad \text{for } 1 < t_{obs} < 2 \quad . \end{aligned} \quad (6.14)$$

To establish a robust Bayes correspondence, we first must ask which alternatives would be compatible with such a T . One possibility is to consider as the “candidate” departures from the $Un(0, 1)$ the $Un(\xi, \xi + 1)$ distributions, so that the null model would be the particular case $\xi = 0$. Then, under H_1 , the MLE of ξ is easily seen to be

$$\hat{\xi} = \frac{X_{(1)} + X_{(n)}}{2} - 0.5 \quad ,$$

so that $T = \hat{\xi}$ or $T = T(X) = X_{(1)} + X_{(n)}$ are indeed sensible statistics to measure departure from H_0 . In this case:

$$\begin{aligned} f_1(t|\xi) &= \frac{n}{2} (t - 2\xi)^{n-1} \quad \text{for } 2\xi < t < 2\xi + 1 \quad , \\ &= \frac{n}{2} [2 - (t - 2\xi)]^{n-1} \quad \text{for } 2\xi + 1 < t < 2\xi + 2 \quad . \end{aligned} \quad (6.15)$$

Since the infimum (over ξ) of (6.15) is n (for $\xi = t/2$) it follows that the infimum of the Bayes factor is

$$\underline{B} = \inf_{\text{all distributions } \pi_1(\xi)} = \frac{f_1(t|\xi = 0)}{\sup_{\xi} f_1(t|\xi)} = \frac{f_1(t|\xi = 0)}{n}$$

so that it is identical to the measure of surprise (6.13), as it should be, since ξ here is a location parameter.

□

Example 6.7 In the conditions of Example 6.6, assume now that the statistic chosen to measure departure from the $Un(0, 1)$ is $T = T(X) = \prod_{i=1}^n X_i$. This would be a natural T to consider if the interesting departures from the null model concerned its shape.

Since $-\log X_i \sim Ex(1)$, it follows that $-\log T \sim Ga(n, 1)$, so that the (null) density of T is

$$f(t) = \frac{1}{n!} (-\log t)^{n-1} \quad ,$$

which has an unbounded supremum so that m^* is not defined. If a Bayesian analysis had been performed instead, a set of alternatives compatible with the chosen T would be

$$f_1(x|\xi) = \frac{1-\xi}{x^\xi} \quad 0 \leq \xi < 1 \quad ,$$

so that the interesting alternatives consider the density to be decreasing, and the null model occurs for $\xi = 0$. Again, as in Example 6.6, it can be checked that the MLE of $(1 - \xi)$ is $n/(-\log T)$, so that T is a sensible measure of departure. In this case,

$$f_1(t) = \frac{(1-\xi)^n}{n!} (-\log t)^{n-1} t^{-\xi}$$

which, as a function of ξ (for a fixed t_{obs}) is maximized at the MLE of ξ , so that the infimum of the Bayes factor is

$$\underline{B} = \inf_{\text{all } \pi_1(\xi)} = \frac{f_1(t_{obs}|\xi=0)}{\sup_{\xi} f_1(t_{obs}|\xi)} = \frac{f_1(t_{obs}|\xi=0)}{[n!(-\log t_{obs})t_{obs}e^n/n^n]^{-1}} = \left(\frac{e}{n}\right)^n t_{obs} (-\log t_{obs})^n$$

which is well defined even though m^* is not.

Notice, however, that if we had chosen $T^* = 1/(-\log T)$, as the MLE of ξ suggests, then it can be shown that,

$$m^*(t_{obs}^*) = \left(\frac{e}{(n+1)t_{obs}^*}\right)^{n+1} e^{-1/t_{obs}^*}$$

and

$$\underline{B} = \left(\frac{e}{nt_{obs}^*}\right)^n e^{-1/t_{obs}^*}$$

which are in close agreement.

□

The choice of the statistic T is a crucial issue. The only general proposal seems to be that of Box, who, as we saw in Section 3, uses $T(X) = m(X)$, the (prior) predictive density; as mentioned there, however, such a choice suffers from a lack of invariance difficulty. (The proposals by Meng, 1994, of using a conditional likelihood ratio statistic and the usual generalized likelihood ratio test statistic involve explicit consideration of alternative values.) For specific problems, particular departure statistics are sometimes proposed as “natural”. Our own point of view is that, for problems in which such natural test statistics T exist, measures of surprise can be a very useful tool, long demanded by applied Bayesians. However, if considerable effort must be spent in looking for an appropriate T , then it might well be better to devote that effort to a careful specification of alternative hypotheses and quantification of prior distributions under them so that a Bayesian analysis can be performed.

7. Conditional predictive distributions.

Recall the two difficulties with direct use of the prior predictive distribution $m(x) = \int f(x|\theta)\pi(\theta)d\theta$: (i) Improper (and even vague proper) priors $\pi(\theta)$ cannot typically be used; (ii) The effects of model inadequacy and prior inadequacy cannot be separated. Even use of a departure statistic T may not help with these difficulties, as the following example indicates:

Example 7.1 Assume that $X = (X_1, X_2, \dots, X_n)$ is a random sample from a $N(0, \sigma^2)$ distribution, and that the aspect of the model under question is the zero mean, with σ^2 being viewed simply as a nuisance parameter. A (proper) $Ga^{-1}(a, b)$ prior for σ^2 is elicited, and computation yields $m(x) = t_n(0, \frac{b}{a}I, 2a)$, so that

$$m(x) \propto [2b + n\bar{x}^2 + ns^2]^{-\left(\frac{n}{2}+a\right)} . \quad (7.1)$$

This can be small either because $|\bar{x}|$ is larger than expected or because s^2 is larger than expected. Even if one uses the natural departure statistic $T(x) = |\bar{x}|$, the resulting predictive distribution is

$$m(t) \propto [2b + n\bar{x}^2]^{-a} ; \quad (7.2)$$

while the confounding with s^2 is gone, surprise can still only be measured relative to the prior for the ‘nuisance parameter’ σ^2 . \square

Sometimes use of T can alleviate the problem, as when T can be chosen to be ancillary, but such cases are very special.

We will argue that the solution to these difficulties is not to abandon $m(x)$, as (for instance) is done by the posterior predictive p -value, but rather to condition on an appropriate statistic U , to obtain a *conditional predictive*, $m(t|u)$, for the departure statistic T . That this can solve the difficulty with improper priors can be seen by writing

$$m(t|u) = \int f(t|u, \theta)\pi(\theta|u)d\theta \quad (7.3)$$

if an improper prior is utilized, we will require that the statistic U be such that $\pi(\theta|u)$ is proper, and then $m(t|u)$ will be a proper distribution. Indeed, with an improper prior, we will view this equation as the *definition* of $m(t|u)$. Note that, in contrast with the posterior predictive p -value, there will be no double use of data here; the part of the data represented by U will be used to eliminate the nuisance parameter, while that represented by T will be used for measuring surprise.

The second difficulty mentioned above, that of separating surprise in the model and surprise in the prior, will also be seen to be reduced, at least if U is chosen appropriately. The rest of this section is primarily devoted to discussion of appropriate choice of U .

Note that, once $m(t|u)$ is obtained, we would argue for its use in any of the surprise measures discussed. One could compute associated p -values (or their calibrations in Section 8). Or one could use these in the relative surprise measures (6.1) and (6.2), yielding

$$m^*(t_{obs}|u_{obs}) = \frac{m(t_{obs}|u_{obs})}{\sup_t m(t|u_{obs})} \quad , \quad (7.4)$$

$$m^{**}(t_{obs}|u_{obs}) = \frac{m(t_{obs}|u_{obs})}{E^{m(t|u_{obs})}[m(T|u_{obs})]} \quad , \quad (7.5)$$

In illustrations in this section we limit ourselves to consideration of (7.4).

Example 7.2 In the situation of Example 7.1, let $T = \bar{X}$ and $U = S^2$. Also, consider the usual non-informative prior $\pi(\sigma^2) \propto 1/\sigma^2$ for σ^2 . An easy computation shows that

the posterior distribution $\pi(\sigma^2|s^2)$ is $Ga^{-1}(\frac{n-1}{2}, \frac{ns^2}{2})$ and that

$$m(\bar{x}|s_{obs}^2) = t(\bar{x} | 0, \frac{s_{obs}^2}{n-1}, n-1) \quad \text{or} \quad \frac{\sqrt{n-1} \bar{X}}{s_{obs}} \sim t_{n-1} \quad . \quad (7.6)$$

This is a most sensible distribution to consider, as large values of $|\bar{x}_{obs}|/s_{obs}$ would clearly be “surprising” . The distribution (7.6) should be contrasted with that obtained by conditioning on *all* the data, that is, the posterior predictive distribution, which in this case is

$$m(\bar{x}|x_{obs}) = t(\bar{x} | 0, \frac{s_{obs}^2 + \bar{x}_{obs}^2}{n}, n) \quad \text{or} \quad \frac{\sqrt{n} \bar{X}}{\sqrt{s_{obs}^2 + \bar{x}_{obs}^2}} \sim t_n \quad . \quad (7.7)$$

The distribution (7.7) is not very appropriate to measure departures from 0, the hypothesized mean. In fact, if s_{obs}^2 is very small or $|\bar{x}_{obs}|$ is very large, then

$$v_{obs} = \frac{\sqrt{n} \bar{x}_{obs}}{\sqrt{s_{obs}^2 + \bar{x}_{obs}^2}}$$

will be close to \sqrt{n} . As a concrete example, take $n = 4$, $s_{obs} = 0.1$ and $\bar{x}_{obs} = 10$. Then $\frac{\bar{x}_{obs}}{s_{obs}} = 100$, most incompatible with a zero mean assumption, but yet $v_{obs} = 2$, clearly compatible with a t_4 distribution, whether the compatibility is judged by (posterior predictive) p -values or by any other method. It follows that we would not be “surprised” by an observation 100 standard deviations away from its hypothesized mean! □

We now turn to discussion of some possible choices of U . (Note that $m(t|u)$ is unchanged under one-to-one transformations of U .)

Through one-to-one transformations of X .

In a sense, the most obvious choice of U is to take “the rest” of the data, relative to T . That is, if (T, X^*) is a one-to-one transformation of X , then take $U = X^*$. Notice that the dimension of U , $n - \dim(T)$, changes with n . The principal advantage of this choice is that it is *very* easy to implement; in fact, in this case $m(t, u)$ can easily be obtained from $m(x)$: it merely suffices to multiply by the Jacobian (no integrations are involved),

so that we can compute the measure of surprise (7.4) as

$$m^*(t_{obs}|u_{obs}) = \frac{m(t_{obs}|u_{obs})}{\sup_t m(t|u_{obs})} = \frac{m(t_{obs}, u_{obs})}{\sup_t m(t, u_{obs})} , \quad (7.8)$$

and $m(t|u)$ does not actually have to be derived. Notice that the procedure can formally be carried out even though $m(x)$ would usually be improper, since $m(t|u)$ is proper and the arbitrary constants cancel.

Example 7.3 Assume that X_1, X_2, \dots, X_n is a random sample from a $N(\theta, 1)$ distribution, with the usual non-informative prior for θ , $\pi(\theta) \propto 1$. Assume that $T = X_{(n)}$ (we might wonder about the tail of the distribution or about a possible outlier). Without loss of generality we can assume that data X are the order statistics, so that natural choice of U is $U = (X_{(1)}, X_{(2)}, \dots, X_{(n-1)}) = (U_1, U_2, \dots, U_{n-1})$. Then

$$m(t, u) \propto \exp \left\{ -\frac{1}{2} \left[\frac{n-1}{n} t^2 + \sum_{i=1}^{n-1} u_i^2 - \frac{(\sum_{i=1}^{n-1} u_i)^2}{n} \right] + \frac{t \sum_{i=1}^{n-1} u_i}{n} \right\} \quad \text{for } u_1 < \dots < u_{n-1} < t ,$$

which, as a function of t is maximized at $t = \sum_{i=1}^{n-1} u_i / (n-1) = \bar{u}$. Hence,

$$m^*(t_{obs}|u_{obs}) = \frac{m(t_{obs}, u_{obs})}{m(\bar{u}_{obs}, u_{obs})} = \exp \left\{ -\frac{1}{2} \frac{n-1}{n} (t - \bar{u})^2 \right\} .$$

Avoiding the computation of $m(t|u)$ is a great simplification. Of course, if p -values instead of relative likelihoods are used to locate t_{obs} , then $m(t|u)$ would have to be explicitly derived. In this example it turns out to be,

$$m(t|u) = \sqrt{\frac{n-1}{2\pi n}} \frac{\exp\{-\frac{1}{2} \frac{n-1}{n} (t - \bar{u})^2\}}{1 - \Phi\left(\frac{u_{n-1} - \bar{u}}{\sqrt{n/(n-1)}}\right)} , \quad \text{for } t > u_{n-1} , \quad (7.9)$$

that is, a $N(\bar{u}, \frac{n}{n-1})$ distribution, truncated at u_{n-1} . Notice that, since $u_{n-1} > \bar{u}$, (7.9) is just the upper tail of a renormalized normal distribution. □

Example 7.4 Assume that X_1, X_2, \dots, X_n is a random sample from an $Ex(\lambda)$ distribution, with the usual non-informative prior for λ , $\pi(\lambda) \propto 1/\lambda$. As in Example 7.3, take

$T = X_{(n)}$, and $U = (X_{(1)}, X_{(2)}, \dots, X_{(n-1)}) = (U_1, U_2, \dots, U_{n-1})$. Then

$$m(t, u) \propto \frac{1}{(t + \sum_{i=1}^{n-1} u_i)^n} \quad \text{for } u_1 < u_2 < \dots < u_{n-1} < t, \quad (7.10)$$

which is a decreasing function of t , so that

$$m^*(t_{obs} | u_{obs}) = \left(\frac{u_{n-1} + \sum_{i=1}^{n-1} u_{obs_i}}{t_{obs} + \sum_{i=1}^{n-1} u_{obs_i}} \right)^n,$$

and, as $n \rightarrow \infty$,

$$m^*(t_{obs} | u_{obs}) \longrightarrow \exp\left\{-\frac{x_{(n)} - x_{(n-1)}}{\bar{x}}\right\}.$$

□

Example 7.5 Under the conditions of Example 7.4, take now $T = X_{(1)}$ (which would be natural if we feel uncertain about the lower tail of the distribution of X) and $U = (X_{(2)}, X_{(3)}, \dots, X_{(n)}) = (U_1, U_2, \dots, U_{n-1})$. Then $m(t, u)$ is again given by (7.10) in the appropriate set, and now it is maximized at $t = 0$, so that

$$m^*(t_{obs} | u_{obs}) = \left(\frac{\sum_{i=1}^{n-1} u_{obs_i}}{t_{obs} + \sum_{i=1}^{n-1} u_{obs_i}} \right)^n,$$

and, as $n \rightarrow \infty$,

$$m^*(t_{obs} | u_{obs}) \longrightarrow \exp\left\{-\frac{x_{(1)}}{\bar{x}}\right\}.$$

□

Example 7.6 Under the conditions of Example 7.4, take now $T = \prod_{i=1}^n X_i$ (which might indicate that we are concerned about the shape of the distribution) and $U = (X_2, X_3, \dots, X_n)$. Then

$$m(t, u) \propto \left(\prod_{i=2}^n x_i \right)^{n-1} \left(t + \prod_{i=2}^n x_i \sum_{i=2}^n x_i \right)^{-n},$$

which is maximized at $t = 0$ giving

$$m^*(t_{obs} | u_{obs}) = \frac{(\prod_{i=2}^n x_{obs_i} \sum_{i=2}^n x_{obs_i})^n}{(t + \prod_{i=2}^n x_{obs_i} \sum_{i=2}^n x_{obs_i})^n},$$

and, as $n \rightarrow \infty$,

$$m^*(t_{obs} | u_{obs}) \longrightarrow \exp\left\{-\frac{x_1}{\bar{x}}\right\}. \quad (7.11)$$

Here $m^*(t|u)$ clearly depends on the particular choice of U ; if U consisted of all observations except X_i , then (7.11) would be the same but with X_i replacing X_1 . Also, (7.11) might be an indication that T is not appropriate or that we are conditioning too much: much information is used to train the prior and little (only the ratio of one of the observations to the mean) is left to judge “surprise”. \square

As Example 7.6 indicates, choosing U to be X^* might well result in too much conditioning; lower dimensional U may provide better answers. The following example is another demonstration.

Example 7.7 Assume that X_1, X_2, \dots, X_n is a random sample from a $N(0, \sigma^2)$ distribution, with the usual non-informative prior for σ^2 , $\pi(\sigma^2) \propto 1/\sigma^2$. Take $T = \bar{X}$. We treated this situation in Example 7.2 arguing that the appropriate conditioning is $U^* = S^2$. The measure of surprise corresponding to this choice of U can be computed to be

$$m^*(t_{obs}|u_{obs}^*) = \left[1 + \frac{\bar{x}_{obs}}{s_{obs}^2} \right]^{-n/2} . \quad (7.12)$$

Assume instead, in the spirit of this subsection, that we had chosen $U = (X_2, X_3, \dots, X_n)$. Then

$$m^*(t_{obs}|u_{obs}) = \left[1 + \frac{x_{1obs}^2}{\sum_{i=2}^n x_{iobs}^2} \right]^{-n/2} , \quad (7.13)$$

which again relies on merely one of the observations to measure surprise. Furthermore, the denominator in (7.13) ties together \bar{x}_{obs} and s_{obs}^2 , so that we anticipate the same kind of difficulties as when using the posterior predictive distribution $m(t|x_{obs})$. As a matter of fact, with this election of U

$$T|u \sim t\left(0, \frac{\sum_{i=2}^n x_{iobs}^2}{n(n-1)}, n-1 \right), \quad \text{or} \quad \frac{\sqrt{n} \bar{X}}{\sqrt{s_{u_{obs}}^2 + \bar{u}_{obs}^2}} \sim t_{n-1} , \quad (7.14)$$

where \bar{u}_{obs} and $s_{u_{obs}}^2$ are the sample mean and variance for the last $n-1$ observations. (7.14) can be seen to exhibit a behavior analogous to that of $m(t|x_{obs})$ as discussed in Example 7.2, clearly indicating again too much conditioning. \square

It seems plausible that this effect of “too much conditioning” could be alleviated if an appropriate “orthogonal” transformation can be found such that T and “the rest” of the data, X^* , are independent, or nearly independent. In such situations, the election of U as X^* might be quite sensible.

Example 7.8 Consider again the conditions of Example 7.7, but now choose $U = (X_2 - \bar{X}, \dots, X_n - \bar{X})$. Then T is conditionally independent of U . The joint marginal distribution is:

$$m(t, u) \propto [nt^2 + \sum_{i=1}^{n-1} u_i^2 + (\sum_{i=1}^{n-1} u_i)^2]^{-n/2} ,$$

which, for fixed u , is maximized at $t = 0$. Therefore, the relative maximized surprise is

$$m^*(t_{obs}|u_{obs}^*) = \left[1 + \frac{nt_{obs}^2}{\sum_{i=1}^{n-1} u_{i_{obs}}^2 + (\sum_{i=1}^{n-1} u_{i_{obs}})^2} \right]^{-n/2} ,$$

which can be checked to be the same as (7.12), that is, the “right” answer for this situation. \square

Although interesting, we shall not pursue this direction any further. Instead, we shall look for low-dimensional conditioning statistics U . Note that, since $\pi(\theta|u)$ must be proper, it will typically be necessary for U to be at least of the same dimension as θ . Indeed, natural choices of U are often statistics of the same dimension as θ , such as estimates of θ .

Ideal choice of U.

Although the following ideal scenario will rarely, if ever, be applicable, it is interesting to consider. The ideal scenario has sufficient statistics, (T, U) , which are low dimensional and conditionally *independent*. In this ideal situation,

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

and data gets used just once, with independent pieces of it used to learn about the nuisance parameter and to detect surprising features. we have encountered such a situation in Example 7.2.

While, we shall rarely encounter such a favorable scenario, this “ideal” can guide us in searching for conditioning a statistic U with good behavior; see Examples 7.7, 7.12, and 7.13, for instance.

Asymptotic independence.

If actual independence between T and U cannot be achieved, one possibility is to seek a U that is asymptotically independent of T . That is, under the usual regularity conditions, choose U such that

$$\begin{pmatrix} T \\ U \end{pmatrix} \doteq N\left(\begin{pmatrix} m \\ \theta \end{pmatrix}, \Sigma\right) .$$

with Σ block diagonal. Often U could be chosen as the MLE, $\hat{\theta}$, of some linear transformation of $(T, \hat{\theta})$. The idea is, in principle, quite attractive, but the following example shows that the use of such U can still result in too much conditioning.

Example 7.9 In the situation of Example 7.2, we have $T = \bar{X}$, and a natural procedure would be to apply the above argument with U being proportional to the MLE of σ^2 . Accordingly, we choose $U = \sum_{i=1}^n x_i^2$. Then \bar{X} and U are asymptotically independent, so this argument has failed to suggest the optimal choice S^2 . Indeed, as we shall later see (Example 7.11), choosing $U = \hat{\sigma}^2$ results in too much conditioning. □

Sufficiency for the nuisance parameter.

One of the original motivations behind conditioning on some statistic U was the need to separately learn about the “nuisance” parameter θ . Hence, a natural idea is to choose U to be a sufficient statistic for θ . In this situation, the conditional distribution $f(x|u, \theta) = f(x|u)$, will not involve θ (and hence neither will $f(t|u)$). This type of conditioning has been proposed in the frequentist scenario (see Cox and Hinkley, 1974) to produce confidence regions whose coverage is independent of the nuisance θ (called *similar* regions), and also to derive testing procedures whose power functions do not depend on θ (uniformly most powerful *similar* tests).

With such a choice of U , $m(t|u)$ is simply given by $f(t|u)$ and, since no θ is involved, there is no need for prior distributions. If we were to use p -values computed from $m(t|u)$, then we would obtain exactly the same answers as from classical *similar* tests. Here is an example:

Example 7.10 Consider two independent Bernoulli populations with probability of success p_1 and p_2 respectively. A random sample of size n_i from population i results in S_i successes and $F_i = n_i - S_i$ failures. Arranged in a two-by-two contingency table, we have

	Population 1	Population 2	Totals
successes	S_1	S_2	$S = S_1 + S_2$
Failures	F_1	F_2	$F = F_1 + F_2$
Totals	n_1	n_2	n

Reparameterizing in terms of $\psi = p_1/p_2$ and $\theta = p_2$ gives

$$f(S_1, S_2 | \psi, \theta) = \binom{n_1}{S_1} \binom{n_2}{S_2} \psi^{S_1} (1 - \theta\psi)^{n_1 - S_1} \theta^{S_1 + S_2} (1 - \theta)^{n_2 - S_2} \quad . \quad (7.15)$$

Suppose we want to test the null hypothesis $H_0 : p_1 = p_2$, that is $H_0 : \psi = 1$. For simplicity, assume that $p_1 \leq p_2$ (or $\psi \leq 1$), although an analogous argument could be made without this restriction.

Consider, first, the frequentist, *similar* analysis. Under the null value $\psi = 1$, it can easily be seen from (7.15) that the *marginal total* $S = S_1 + S_2$ is sufficient for the “nuisance” parameter θ , so it is a natural conditioning statistic. It can be seen that the conditional distribution of (S_1, S_2) , given $S = S_1 + S_2$, is independent of θ and is characterized by the hypergeometric distribution

$$f(S_1 | S) = \frac{\binom{n_1}{S_1} \binom{n_2}{S - S_1}}{\binom{n}{S}} \quad \text{for } S_1 = \max\{0, S - n_2\}, \dots, \min\{S, n_1\} \quad ,$$

so that the classical (*similar*) p -value is

$$\sum_{j=S_{1_{obs}}}^{\min\{S_{obs}, n_1\}} \binom{n_1}{j} \binom{n_2}{S_{obs} - j} / \binom{n}{S_{obs}} \quad ,$$

which is the *Fisher exact test*. Since the distribution of $T = S_1$ given the sufficient statistic $U = S$ does not involve θ , this could be the distribution to consider for “surprise” analysis also. This example will be revisited in Example 7.16

In this example, since an alternative hypothesis is readily available, a Bayesian analysis could be performed. Indeed, with prior distributions $p_1 \sim Be(a_1, b_1)$ and $p_2 \sim Be(a_2, b_2)$, the posterior distribution is easily seen to be the product of the two independent marginal posteriors $p_1 \sim Be(S_1 + a_1, n_1 - S_1 + b_1)$ and $p_2 \sim Be(S_2 + a_2, n_2 - S_2 + b_2)$. It is well

known that testing of point nulls, as here, typically produce radically different answers from the classical and Bayesian perspectives. The testing of one-sided hypotheses with non-informative priors, on the other hand, sometimes reproduces classical p -values. Interestingly, it can be checked that the one-sided posterior probability $Pr\{p_1 \leq p_2\}$ is given by Fisher exact test if $a_1 = 0$, $b_1 = 1$, $a_2 = 1$, $b_2 = 0$, that is, if the prior for (p_1, p_2) is the improper

$$\pi(p_1, p_2) = p_1^{-1}(1 - p_2)^{-1} \quad .$$

As this is a questionable prior, giving more mass close to $p_1 = 0$ and $p_2 = 1$, this may reveal possible inadequacies in the Fisher exact test and, therefore, in conditioning on a U that is sufficient for the nuisance parameter. □

In addition to the concern mentioned in Example 7.10, conditioning on a sufficient statistic may also result in an excess of conditioning. The following example demonstrates this.

Example 7.11 In the situation of Example 7.9, we have $T = \bar{X}$, and $U = \sum_{i=1}^n x_i^2 = \|X\|^2$, which is sufficient for σ^2 . For $u_{obs} = \|x_{obs}\|^2$, the distribution of X given u_{obs} is uniform on the set $\{x, s.t. \|x\|^2 = \|x_{obs}\|^2\}$, so that

$$Pr(\bar{X} > \bar{x}_{obs}) = Pr\left(\frac{\bar{X}}{\|x_{obs}\|} > \frac{\bar{x}_{obs}}{\|x_{obs}\|}\right) = Pr\left(\bar{Z} > \frac{\bar{x}_{obs}}{\sqrt{n\bar{x}_{obs}^2 + s_{obs}^2}}\right), \quad (7.16)$$

where Z has a uniform distribution on the set $\{z, s.t. \|z\|^2 = 1\}$. For every fixed sample size n , it follows from (7.16) that, as $\|x_{obs}\|^2 \rightarrow \infty$,

$$Pr(\bar{X} > \bar{x}_{obs}) \longrightarrow Pr\left(\bar{Z} > \frac{1}{\sqrt{n}}\right),$$

which is a non 0 constant. This is unsatisfactory, since very large \bar{x}_{obs} are clearly “surprising”. □

An attractive choice of U .

In our search for a suitable conditioning statistic U , recall that the goal is to use the information in the data, other than that contained in T , to eliminate the nuisance parameter

θ . The distribution $f(x|t, \theta)$ removes the information provided by T from the likelihood for θ so, in looking for an adequate U , $f(x|t, \theta)$ is an obvious distribution to explore.

One possibility would be to take U to be a low-dimensional sufficient statistic of this conditional distribution. However, sufficient statistics may not exist or may be difficult to identify. We shall therefore propose a general definition of U that will, in particular, identify a sufficient statistic having the same dimension as θ , should such exist. Our proposal is to define

$$U = \hat{\theta} = \arg \max f(x|t, \theta) = \arg \max \frac{f(x|\theta)}{f(t|\theta)} \quad \text{for } T(x) = t \quad . \quad (7.17)$$

Example 7.12 As in Example 7.7, assume that X_1, X_2, \dots, X_n is a random sample from a $N(0, \sigma^2)$ distribution and that $T = \bar{X}$. Then

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

which is maximized at $\hat{\sigma}^2 = \frac{n}{n-1}S^2$, where, as before, $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$; as we have previously discussed, this is the optimal choice of U for this situation. \square

Example 7.13 Assume that X_1, X_2, \dots, X_n is a random sample from a $N(\theta, 1)$ distribution and that $T = S^2$. Then, since $S^2 \sim Ga(\frac{n-1}{2}, \frac{n}{2\sigma^2})$,

$$f(x|t, \theta) \propto (S^2)^{-\frac{n-3}{2}} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} \quad ,$$

which is maximized at $\hat{\theta} = \bar{x}$, again providing the optimal choice of U . \square

Example 7.14 Assume that X_1, X_2, \dots, X_n is a random sample from a $Un(0, \theta)$ distribution and that we are uncertain about the shape, so we choose $T = \prod_{i=1}^n X_i$. Then, since

$$f(t|\theta) = \frac{1}{\theta^n \Gamma(n)} \left(-\log \frac{t}{\theta^n}\right)^{n-1} \quad , \quad \text{for } 0 \leq t \leq \theta^n \quad ,$$

it follows that

$$f(x|t, \theta) = \frac{\Gamma(n)}{\left(-\log \frac{t}{\theta^n}\right)^{n-1}} \quad \text{for } \theta \geq x_{(n)} \quad ,$$

which is decreasing with θ thus producing $U = X_{(n)}$, an intuitive and natural choice. \square

Example 7.15 As in Examples 7.4, 7.5 and 7.6 assume that X_1, X_2, \dots, X_n is a random sample from an $Ex(\lambda)$ distribution. We shall consider several possibilities for T :

1. $T = \sum_{i=1}^n X_i$, which is sufficient for the nuisance parameter. Clearly, this is not a reasonable choice for T since it contains the information about λ , and not that concerning departures from the assumed exponential model. Indeed $f(x|t, \lambda)$ then does not depend on λ and U is not needed to eliminate the nuisance parameter.
2. $T = X_{(n)} - X_{(n-1)}$ (which could aim at detecting departures from the exponential decay of the upper tail). Then, since $T \sim Ex(\lambda)$,

$$f(x|t, \lambda) \propto \lambda^{n-1} \exp\{-\lambda [\sum_{i=1}^{n-1} x_{(i)} + x_{(n-1)}]\} \quad , \quad (7.18)$$

and the U suggested by maximizing (7.18) is $\sum_{i=1}^{n-1} X_{(i)} + X_{(n-1)}$ (recall that $m(t|u)$ is unaffected by one-to-one transformations of U).

3. $T = X_{(n)}$. Then

$$f(x|t, \lambda) \propto \left(\frac{\lambda}{1 - e^{-\lambda t}} \right)^{n-1} e^{-\lambda [\sum x_{(i)} - t]} \quad , \quad \text{for } \max\{x_i\} = t \quad ,$$

which would have to be numerically maximized over λ to determine U via (7.17) .

4. $T = X_{(1)}$. Here, an easy computation shows that

$$f(x|t, \lambda) \propto \lambda^{n-1} e^{-\lambda (\sum x_{(i)} - nt)} \quad \text{for } \min\{x_i\} = t \quad ,$$

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

5. $T = \prod_{i=1}^n X_i$. In this case, the distribution of T is not available in closed form, so we would have to rely solely on numerical computation of U via (7.17) .

□

Example 7.16 Consider the contingency table situation treated in Example 7.10, where the null hypothesis stated the equality of two Bernoulli parameters, $p_1 = p_2 = p$. Assume we take $T = S_1$, the number of successes for the first population. It follows that, as a function of p , $f(S_1, S_2|t, p)$ is proportional to the $Bi(S_2|n_2, p)$ density , which is maximized at its MLE $\hat{p} = S_2/n_2$, thus suggesting the choice $U = S_2$. Fisher's exact test was derived in Example 7.10 by conditioning on $S_1 + S_2$, which made the conditional distribution of

the observations free of the parameter. Here, this is not the case. Interestingly, however, the conditional predictive distribution $m(t|u)$ (with a uniform prior on p) is given in this case by

$$m(S_1|S_2) = \int f(S_1|p)\pi(p|S_2)dp = \frac{n_2 + 1}{n_1 + n_2 + 1} \frac{\binom{n_1}{S_1} \binom{n_2}{S_2}}{\binom{n_1 + n_2}{S_1 + S_2}},$$

so that p-values in this distribution would reproduce the Fisher exact test. Notice, however, that in order to do so, we had to choose a most unreasonable statistic to measure departure from the null, namely S_1 . The proposed procedure for choosing U and conditioning upon it does alleviate the terrible choice of T , but more sensible results should be expected from a more sensible T (for instance, $T = S_1/n_1 - S_2/n_2$). \square

In the previous examples, we have been considering “casual” choices of T . We next turn to consideration of explicit alternative models as a guide to choice of T and then derive an appropriate U for such meaningful choices of T .

Example 7.17 As in Example 7.15, let X_1, X_2, \dots, X_n be a random sample from an $Ex(\lambda)$ distribution. Suppose that we are unsure about the shape of the density, and so choose $Ga(\xi, \lambda)$ alternatives, with the null model obtaining for $\xi = 1$. In this case, using the usual non-informative prior for λ , we can integrate out λ from $f(x, \lambda|\xi)$ to obtain

$$f(x|\xi) = \frac{(\prod_{i=1}^n x_i)^{\xi-1}}{[?(\xi)]^n} \frac{?(n\xi)}{(\sum_{i=1}^n x_i)^{n\xi}}.$$

The MLE for ξ is a function of

$$T = \frac{\prod_{i=1}^n x_i}{(\sum_{i=1}^n x_i)^n},$$

which is thus suggested as a good statistic to measure departure from $\xi = 1$. Indeed, this is a natural choice of T , since it is ancillary for the nuisance parameter λ (i.e., $(f(t|\lambda, \xi) = f(t|\xi))$). Under the null model (which is really all we assume is actually available), T could thus be used directly to measure surprise; there is no need to consider any U . As a matter of fact, choice of U via (7.17) does reproduce this intuition. Indeed, in this case, $f(x|t, \lambda) \propto f(x|\lambda)$, so that U would be the usual MLE for λ , that is, a sufficient statistic for λ . Then, under some regularity conditions, sufficient and ancillary statistics are independent (see Lehman, 1959), so that $f(t|u, \lambda) = f(t)$, and, according to intuition, $m(t|u) = f(t)$. \square

Example 7.18 Assume, as in Example 7.14, that X_1, X_2, \dots, X_n is a random sample from a $Un(0, \theta)$ distribution and that we are uncertain about the shape, contemplating alternative densities that are decreasing. One possibility would be

$$f(x|\theta, \xi) = \frac{1-\xi}{\theta^{1-\xi}} \frac{1}{x^\xi} \quad \text{for } 0 \leq x \leq \theta \quad ,$$

with $0 \leq \xi < 1$, so that the null model obtains for $\xi = 0$. To choose an appropriate T , we again integrate θ out of the joint $f(x, \theta|\xi)$ with a non-informative prior $\pi(\theta) \propto 1/\theta$, giving

$$f(x|\xi) \propto (1-\xi)^{n-1} \left(\prod_{i=1}^n x_i \right)^{-\xi} (x_{(n)})^{-n(1-\xi)} \quad ,$$

whose maximization suggests use of

$$T = \frac{\prod_{i=1}^n x_i}{x_{(n)}} \quad .$$

Since this choice of T is again ancillary for θ , there is no compelling reason to choose a U upon which to condition. \square

Example 7.19 Assume that the *full* model (not available really in surprise contexts) is a Pareto $X \sim Pa(\alpha, \lambda)$ distribution, so that, if $Y = \log X$, then $Y - \log \alpha \sim Ex(\lambda)$. Consider testing:

1. $H_0 : \alpha = 1$. (Note that this can be considered to be a new set of alternatives for the situation in Example 7.17, where in this case we are concerned about the lower bound of Y). Integrating out λ , it can be checked that

$$f(x|\alpha) \propto \frac{\alpha^n}{\prod_{i=1}^n x_i}, \quad \text{for } \alpha \leq x_{(1)} \quad .$$

Hence, a natural choice of T is $T = X_{(1)}$, agreeing with intuition. Since

$$\frac{\alpha}{T} \sim Be(n\lambda, 1) \quad ,$$

it follows that

$$f(x|t, \lambda) \propto \left(\frac{t}{\prod_{i=1}^n x_i} \right) \left(\frac{t^n}{\prod_{i=1}^n x_i} \right)^\lambda \lambda^{n-1} \quad .$$

Maximization over λ yields

$$U = \frac{\prod_{i=1}^n x_i}{t^n} \quad .$$

2. $H_0 : \lambda = 1$. (This is equivalent to the situation discussed in Example 7.18). Here, integration of α with $\pi(\alpha) \propto 1/\alpha$ gives

$$f(x|\lambda) \propto \frac{\lambda^{n-1}}{(\prod_{i=1}^n x_i)^{\lambda+1}} (x_{(1)})^{n\lambda} ,$$

which, upon maximization, suggests

$$T = \frac{\prod_{i=1}^n x_i}{(x_{(1)})^n} .$$

This is ancillary, and so no U is needed.

□

Computational issues.

Typically, calculations involving surprise measures will have to be carried out numerically. In Bayesian analysis, it has become standard to base inferences on samples generated from the target distribution via MC or MCMC methods. Here we propose simple strategies to perform the required computations. We assume that T and U have dimension 1 (the generalization to larger dimensions is direct).

First notice that, if $m(t|u_{obs})$ is not available, but we have at our disposal a simulated sample, x^1, x^2, \dots, x^M of size M from $m(x|u_{obs})$, then we can easily compute:

1. p -values:

$$Pr\{T(X) \geq t_{obs} | u_{obs}\} = \frac{\# \text{ of } T(x^i) \geq t_{obs}}{M} .$$

2. relative maximized surprise:

$$m^*(t_{obs} | u_{obs}) = \frac{\# \text{ of } T(x^i) \text{ within } \epsilon \text{ of } t_{obs}}{\text{maximum } \# \text{ of } T(x^i) \text{ in an interval of length } 2\epsilon} .$$

3. relative expected surprise:

$$m^{**}(t_{obs} | u_{obs}) = \frac{\# \text{ of } T(x^i) \text{ within } \epsilon \text{ of } t_{obs}}{\sum_{j=1}^M \# \text{ of } T(x^i) \text{ within } \epsilon \text{ of } T(x^j) / M} .$$

Of course, the last two use rather naïve estimators of a density, and more sophisticated estimators (e.g., kernel estimators or Rao-Blackwellized ones) could be used if greater accuracy is needed. Notice that the computations above can also be performed if $m(x)$ itself is the basis of the surprise measure.

We still need an algorithm to produce a simulated sample from $m(x|u)$. Assume first that an explicit expression for U is available. We shall suggest two algorithms to use, both easy to implement, one based on a Gibbs scheme and the other on a Metropolis-Hasting approach.

Both algorithms will actually generate, not from $m(x|u_{obs})$, but from $m(x| |u - u_{obs}| < \delta)$. Of course, for δ small enough, $m(x| |u - u_{obs}| < \delta)$ is simply an approximation to $m(x|u_{obs})$, but larger δ 's allow faster computation and also allow *less* conditioning than that provided by u_{obs} , if so desired. As a matter of fact, notice that, as $\delta \rightarrow \infty$, $m(x| |u - u_{obs}| < \delta) \rightarrow m(x)$ thus providing the answers corresponding to *prior* predictive distributions. Even when improper priors are used, $m(x| |u - u_{obs}| < \delta)$ will usually be proper for any δ , but corresponding measures of surprise may be meaningless if δ is too large.

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

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

where $1_{\{A\}}$ is the indicator function of the set A , and the denominator, $Pr(|u - u_{obs}| < \delta)$ is a constant, irrelevant in both Gibbs and Metropolis schemes.

Gibbs . A Gibbs sampler chain is quite simple to implement:

- *Step 1.* Generate $\theta \sim \pi(\theta|x)$.
- *Step 2.* Generate $X \sim f(x|\theta)1_{\{|u-u_{obs}|<\delta\}}$.

Notice that *Step 1* merely calls for generating from the usual posterior distribution. The easiest way to generate from *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). In the limit, (x, θ) thus generated would be a realization of $f(x, \theta \mid |u - u_{obs}| < \delta)$.

The chain for the Gibbs scheme just presented needs to be built specifically for “surprise” evaluations. The following algorithm, in contrast, can be based on the outputs, θ^i , of a usual Bayesian analysis, that is, as generated from $\pi(\theta|x_{obs})$.

Metropolis-Hastings . We generate a chain (x^j, θ^j) as follows. Our proposal for a probing (or jumping) distribution is

$$f(x|\theta)\pi(\theta|x_{obs})\mathbf{1}_{\{|u-u_{obs}|<\delta\}} \quad .$$

Then, from (x^t, θ^t) at time t ,

- 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 the condition fails, a *new* θ has to be generated from $\pi(\theta|x_{obs})$; this was not required for the similar step in the Gibbs sampler.
- Accept the candidate with probability

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

This scheme has the advantage that it can be readily incorporated into the usual MCMC for posterior analyses. However, it might require many draws from the usual $\pi(\theta|x_{obs})$ to get a sample from $\pi(\theta \mid |u - u_{obs}| < \delta)$.

The previous schemes can easily be implemented whenever the explicit form of U is available. If, however, U is computed via (7.17), additional computation is needed. When $f(t|\theta)$ is available in closed form, computing u from x reduces to a numerical maximization (over θ) of $\frac{f(x|\theta)}{f(t|\theta)}$. If $f(t|\theta)$ is not available in closed form, the following simple algorithm would compute the required $u = u(x^*)$ for a given x^* and $t^* = T(x^*)$:

- *Step 1.* Take a grid of θ values, in an adaptive way if needed.

- *Step 2.* For each θ generate a sample x^i from $f(x|\theta)$ and compute

$$r(\theta) = \frac{f(x^*|\theta)}{\hat{f}(t^*|\theta)},$$

where $\hat{f}(t|\theta)$ is some estimate of the density $f(t|\theta)$. The crudest such estimate is,

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

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

- *Step 3.* Take u as the value of θ maximizing $r(\theta)$ over the grid.

Notice, however, 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 $\max r(\theta)$ occurs in this restricted grid. This last could be accomplished by checking whether or not the maximum occurs on the boundaries of this grid. This is a remarkable simplification that can make this computation feasible.

As mentioned previously, we believe that measures of surprise based on likelihood ratios are more in accord with Bayesian reasoning than ones based on tail areas or p -values. However, tail areas are often easier to compute, are invariant under one to one transformations and can be applied to *discrepancy measures* (following the terminology in Gelman, Meng and Stern, 1996), that is, functions of both x and θ . For these reasons, it makes practical sense to compute the tail area of the observed $T(x_{obs})$ in the predictive distribution $m(t|u)$. Nevertheless, the fact remains that p -values are highly misleading measures of evidence against H_0 . In the next Section we *calibrate* the p -values so that they are closer to Bayesian answers (as would be provided, say, by an infimum of a Bayes factor, were alternatives actually specified).

8. Calibration of p-values.²

For ease of understanding, in this Section we shall switch to a more standard notation and denote the model under the null hypothesis by $f(x)$. It should be kept in mind, however, that for surprise purposes, $f(x)$ will usually be $m(t|u)$, some partially conditional predictive distribution of a test statistic t .

The purpose of this Section is to investigate the possibility of developing an adjustment to the p -value that would render it closer to an infimum of Bayes factors. Therefore, we have to explicitly consider alternatives to the null model, that, as usual, will take the form of a more elaborate model $f(x|\xi)$ in which $f(x)$ is nested, so that $f(x) = f(x|\xi_0)$. Assuming that large values of x reflect departure from the null model, the p -value will be given by $p = p(x_{obs})$, where

$$p(x) = \int_x^\infty f(z|\xi_0) dz \quad . \quad (8.1)$$

Also, as usual,

$$m(x) = \int f(x|\xi)\pi(\xi)d\xi \quad , \quad (8.2)$$

so that the Bayes factor in favor of ξ_0 is given by

$$B = \frac{f(x_{obs}|\xi_0)}{m(x_{obs})} \quad .$$

It can immediately be checked that the ratio of the Bayes factor, B , and the p -value, p , is

$$\frac{\text{Bayes Factor}}{p\text{-value}} = \frac{B}{p} = \frac{h_0(x_{obs})}{m(x_{obs})} \quad , \quad (8.3)$$

where $h_0(x)$ is the *hazard rate* or *failure rate* function of the null model:

$$h_0(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x|\xi_0)}{\int_x^\infty f(y|\xi_0) dy} \quad . \quad (8.4)$$

Also, it can easily be seen from (8.2), that $m(x) \leq f(x|\hat{\xi})$, where $\hat{\xi}$ is the MLE of ξ , so

²This Section is joint work with Tom Sellke, Purdue University.

that

$$\frac{\text{Bayes Factor}}{p\text{-value}} \leq \frac{\text{null failure - rate at } x_{obs}}{f(x_{obs}|\hat{\xi})} . \quad (8.5)$$

Since failure rate functions are well studied for standard models, and since they often have simple approximate expressions³ for large values of x_{obs} , (8.5) does provide a quick, easy fix for p -values on such problems. However, it is useless for surprise purposes, since alternatives are not defined, so that $f(x_{obs}|\hat{\xi})$ is not available. We shall thus take a different approach.

It is well known (and can easily be shown) that, under the null hypothesis, the distribution of the p -value $p(X)$ is uniform on $[0, 1]$. Now, instead of assessing alternative conditional distributions for X , and prior distributions for the added parameter, thus implicitly assessing $m(x)$, the alternative (marginal) distribution for X , the idea is to directly look for suitable alternative distributions for $p(X)$ and compute the infimum of the Bayes factor in favor of the uniform. In other words, we shall test

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

and compute the infimum of the Bayes factor in favor of H_0 over possible priors for ξ . (Others have previously considered direct choice of alternatives for $p(X)$; see, for instance, Hodges, 1992).

For any observed $p_0 = p(x_{obs})$, the distributions for $p = p(X)$ that are most unfavorable to the null hypothesis would concentrate all their mass on $[0, p_0]$. These are, however, obviously unreasonable as general alternatives. It does seem natural to require the distribution of the p -values to be decreasing with p and we shall so assume.

Example 8.1 Assume that, under H_0 , $X \sim N(0, 1)$, and, under H_1 , $X \sim N(0, v^2)$, with $v > 1$. (This would be the case if under H_1 , $X|\theta \sim N(\theta, 1)$ and $\theta \sim N(0, v^2 - 1)$ so that the predictive distribution is as indicated). Then, for $p = p(x) = 2[1 - \Phi(x)]$, and under H_1 ,

$$Pr\{p(X) > p\} = \Phi \left[\frac{\Phi^{-1}(1 - \frac{p}{2})}{v} \right] , \quad (8.6)$$

³For example, for large values of x , $h(x) \approx x$ for the normal, $h(x) = \text{constant}$ for the exponential, and $h(x) \approx 1/x$ for the Cauchy.

where Φ is the cdf of a standard normal. It follows from (8.6) that the density of p is given by

$$f(p) = \frac{1}{v} \exp\left\{ \frac{1}{2} \left(1 - \frac{1}{v^2}\right) \left[\Phi^{-1}\left(1 - \frac{p}{2}\right)\right]^2 \right\} ,$$

which is decreasing with p . □

A class of alternatives for p that is very easy to work with is the class of $Be(\xi, 1)$ distributions, where we take $0 \leq \xi \leq 1$ so that the distributions are nonincreasing:

$$f(p|\xi) = \xi p^{\xi-1} = \frac{\xi}{p^{1-\xi}} . \quad (8.7)$$

The uniform distribution obtains for $\xi = 1$. (We have used this distribution before in Example 6.7)

Instead of working with p and its distribution $f(p|\xi)$ it is more convenient to work with $Y = -\log p$ and its distributions under the null and alternative hypothesis. It can easily be checked that, if p has the $Be(\xi, 1)$ distribution given in (8.7), then

$$Pr\{Y > y\} = Pr\{p < e^{-y}\} = e^{-\xi y} ,$$

so that Y has an $Ex(\xi)$ distribution (and, of course, the null hypothesis again obtains for $\xi = 1$). Therefore, the infimum of the Bayes factor over all priors for ξ is

$$\underline{B} = \inf_{\text{all } \pi'_1s} \frac{f(y|\xi_0)}{\int f(y|\xi)\pi_1(\xi)d\xi} = \frac{Ex(y|1)}{\sup_{\xi} Ex(y|\xi)} = e^{-y} \quad \text{for } y > 1 , \quad (8.8)$$

and $\underline{B} = 1$ otherwise. Substituting $p = e^{-y}$ in the lower bound (8.8) gives, for $p < e^{-1} \approx 0.37$,

$$\underline{B} = -e p \log p \quad (8.9)$$

The next table gives some p -values and the values of this lower bound.

The calibration provided for p -values by (8.9) assumes that alternative models and priors are such that the distribution of $Y = -\log p$ is exponential, that is, has a constant failure rate. We would like to relax this assumption but still require that the distribution of p should decrease sufficiently fast so that it “piles up” mass close to 0.

p	.2	.1	.05	.01	.005	.001
$-ep \log p$.870	.625	.407	.125	.072	.0188

Table 2: Calibration of p -values as infimum of Bayes factors.

A natural requirement is that the distribution of Y has a decreasing (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 , which, for instance, implies that, for any fixed p_0 , the probability $Pr\{p < \frac{1}{2} \mid p < p_0\}$ increases as p_0 goes to 0, which is the kind of behavior we are looking for.

What we shall show next is that the simple calibration (8.9) is still valid when we only assume that the distribution of Y has a decreasing (that is, non-increasing) failure rate. Notice that we shall not be making *any* parametric assumption about the alternative models nor priors.

Assume, accordingly, that the failure rate function

$$h_1(y) = \frac{f_1(y)}{\int_y^\infty f_1(z) dz}$$

of the distribution of Y under the alternative (predictive) model f_1 has a decreasing failure rate (of course, the null distribution is still $Ex(1)$). Then

$$f_1(y) = h_1(y) e^{-\int_0^y h_1(z) dz} \leq h_1(y) e^{-y h_1(y)} ,$$

so that a lower bound on the Bayes factor (as we let the alternatives vary in the class providing marginal distributions of Y with decreasing failure rate) is given by

$$B = \frac{e^{-y}}{f_1(y)} \geq \frac{e^{-y}}{h_1(y) e^{-y h_1(y)}} \geq e y e^{1-y} \quad \text{for } y \geq 1 ,$$

and $\underline{B} = 1$ otherwise, the inequalities being sharp. Notice that this is exactly the same bound as (8.8).

We shall next compare the bound $-ep \log p$ with actual lower bounds for several classes of alternatives in the simplest normal example, so as to indicate that this bound is comparable with actual robust Bayesian bounds from parametric setups.

Example 8.2 Assume that the null model is $N(0, 1)$. We consider alternatives of the type $X|\theta \sim N(\theta, 1)$ with $\theta \sim \pi_1(\theta)$. Berger and Sellke (1987) provided lower bounds for the Bayes factor when π_1 belonged to the following possible classes of priors:

$$\begin{aligned} ?_{Normal} &= \{\pi_1 : \pi_1(\theta) = N(\theta|0, v^2 - 1), v > 1\} \\ ?_{US} &= \{\pi_1 : \pi_1(\theta) \text{ is unimodal and symmetrical}\} \\ ?_{Sym} &= \{\pi_1 : \pi_1(\theta) \text{ is symmetrical}\} . \end{aligned}$$

The following table displays these lower bounds together with the p -values and the calibrations $-ep \log p$.

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

Table 3: Infimum of Bayes factors, p -values and their calibrations.

A striking feature of Table 3 is the close agreement between the lower bounds of the Bayes factors for the class $?_{US}$ and the proposed calibration $-ep \log p$, which seems to indicate that the hazard rate function for the models giving such infimum are nearly constant. The results for the conjugate priors $?_{Normal}$ are consistent with lower bounds with decreasing failure rate, the calibration being smaller than the corresponding infimum of the Bayes factors (recall that the bound is actually attained with a constant failure rate). $?_{Sym}$ falls outside the conditions under which the calibration holds, but is arguably too large a class of priors. □

If the alternative model and prior have been assessed, there is a relatively simple way to check whether the predictive distribution under the alternative assessments is such that the distribution of Y has decreasing failure rate: Assume, as in the beginning of the section, that, under the null, $X \sim f(x)$ and, under the alternative, the marginal

distribution of X is $m(x)$. We shall denote by F and M the corresponding c.d.f.'s. If $p = p(X)$ is the p-value under the null as given in (8.1), then the survival function of $Y = -\log(p(X))$ under the alternative can be computed as

$$Pr\{Y > y\} = Pr\{p < e^{-y}\} = 1 - M[F^{-1}(1 - e^{-y})], \quad (8.10)$$

so that its density is given by

$$f_1(y) = \frac{m[F^{-1}(1 - e^{-y})]}{e^y f[F^{-1}(1 - e^{-y})]}. \quad (8.11)$$

The hazard rate function of Y is given by the ratio of (8.11) and (8.10), and it can easily be seen that it is decreasing if and only if

$$\frac{m(x)}{1 - M(x)} / \frac{f(x)}{1 - F(x)} \quad (8.12)$$

that is, the ratio of the alternative hazard rate to the null hazard rate, is decreasing.

Example 8.3 In the conditions of Example 8.2, assume first that the alternative hypothesis consists of the $N(0, \theta)$ model and some (unspecified) prior $N(0, v^2 - 1)$ with $v > 1$ (that is, the class $?_{Normal}$ in Example 8.2). Then, the infimum of the Bayes factor will be reached at some alternative normal model $N(x|0, v^2)$, and condition that (8.12) is decreasing reduces to requiring that

$$\frac{R(x)}{c R(\frac{x}{c})} \quad (8.13)$$

be decreasing, where $R(x)$ is *Mill's ratio*, or the inverse of the hazard rate function of the standard normal. Figure 1 shows the functions (8.13) for various values of c ; these decrease to their limiting value $1/c^2$.

Assume next that the only requirement on the prior over the alternative is for it to be unimodal and symmetric, that is, that it belongs to the class $?_{US}$ of Example 8.2. Then the infimum of Bayes factors will be reached for some uniform prior on $(-b, b)$, with unspecified b , for which alternatives (8.12) can be shown to reduce to

$$\frac{\Phi(x+b) - \Phi(x-b)}{\int_{x-b}^{x+b} [1 - \Phi(z)] dz} / \frac{\varphi(x)}{1 - \Phi(x)}, \quad (8.14)$$

where φ and Φ are the pdf and cdf, respectively, of the standard normal. This function

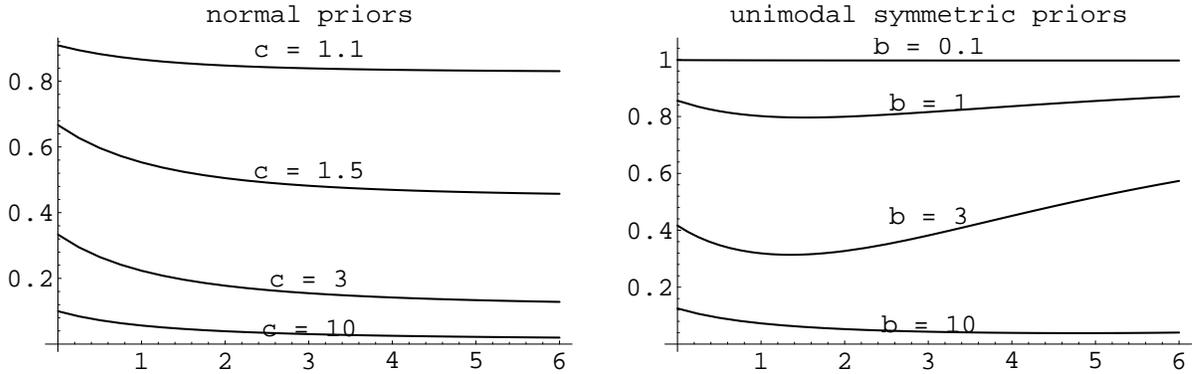


Figure 1: Ratio of hazard rate functions: (8.13) for normal priors (left) and (8.14) for unimodal symmetric priors (right).

is not monotone for all values of b (see Figure 1), but it is fairly constant, so that the approximation to the lower bound on the Bayes factor provided by $-e p \log p$ is good. \square

It is well recognized that p -values (and infimums of Bayes factors over “too large” classes) can behave poorly in high dimensions. An interesting question is, thus, whether or not the calibration of p -values that we are proposing in this Section is still roughly valid for dimensions higher than 1.

Example 8.4 This is the k -variate version of Example 8.2. Assume that the null model is $N_k(0, I)$ and that the marginal distribution of X under the alternative is $m(x) = N_k(x|0, v^2 I)$. Then the Bayes factor in favor of the null model is:

$$B = v^k e^{-\frac{1}{2}\|x\|^2(1-v^{-2})} ,$$

which can be shown to be minimized at $v^2 = \|x\|^2/k$ yielding an infimum for the Bayes factor equal to

$$\underline{B} = \left[\frac{\|x\|^2}{k} \right]^{\frac{k}{2}} e^{-\frac{1}{2}(\|x\|^2 - k)} .$$

For large k , an (asymptotic) approximation to the $\|x\|^2$ yielding a given p -value, p , is

$$\|x\|^2 \approx k + z_p \sqrt{2k} + \frac{2}{3}(z_p^2 - 1) + o(1) ,$$

where z_p is the $(1 - p)$ quantile of the standard normal. Furthermore, it can be shown that

$$\left[1 + z_p \sqrt{\frac{2}{k}} + \frac{2}{3} \frac{z_p^2 - 1}{k} + \frac{1}{k} o(1)\right]^{k/2} = e^{\sqrt{\frac{k}{2}} z_p - \frac{2+z_p^2}{6}} (1 + o(1)) \quad ,$$

obtained by expanding its logarithm. Using these, it can be shown that, for large k , the infimum of the Bayes factor is given by

$$\underline{B} \approx e^{-\frac{1}{2} z_p^2} (1 + o(1)) \quad .$$

By further approximating

$$p = \frac{\varphi(p)}{z_p} (1 + o(1)) \quad ,$$

we finally obtain,

$$\underline{B} = \left(\frac{2}{e z_p}\right) (-e p \log p) \quad ,$$

so that the ratio of \underline{B} to the calibrated p -value is a bounded, moderate quantity $2/(e z_p)$.

The actual infimum of Bayes factor for this situation were derived in Delampady and Berger (1990) for two classes of priors: $?_{normal}$ of k -variate conjugate priors, and $?_{USS}$ of all unimodal, spherically symmetrical priors for the alternative θ mean. In the following table we display some of these lower bounds, along with the p -values and their calibrated values.

p	0.1		0.05		0.01		0.001	
$-ep \log p$	0.6259		0.4072		0.1252		0.01878	
$k = 3$.5818	.5396	.3784	.3259	.1142	.0902	.0165	.0119
$k = 6$.5058	.5473	.3419	.3023	.0988	.0807	.0129	.0097
$k = 15$.5078	.4826	.3108	.2875	.0859	.0752	.0113	.0093
$k = 30$.4879	.4725	.2950	.2809	.0803	.0735	.0105	.0092
$k = \infty$.2585	.4682	.2950	.2809	.0803	.0735	.0105	.0092
	$?_{Normal}$	$?_{USS}$	$?_{Normal}$	$?_{USS}$	$?_{Normal}$	$?_{USS}$	$?_{Normal}$	$?_{USS}$

Table 4: \underline{B} , p -values and their calibrations for various dimensions k .

□

Some comments are in order here. First notice that reporting $-e p \log p$ is still better than reporting the raw p -values p and interpreting them as measuring evidence against

H_0 . Second, the fact that $-e p \log p$ is moderately *larger* than what it appears it should be might, actually, not be such an undesirable behavior, since we are completely ignoring the effect of the sample size, n . Also, in this situation it is not very clear that the classes of priors $?_{Normal}$ and $?_{USS}$ are always appropriate in high dimensions, since they turn out to be increasingly informative with increasing dimensions, piling most of the mass on the sphere. Last, if we use the calibration to adjust p -values computed from a $m(t|u_{obs})$ in the surprise context that we have been studying, then we do not really have to worry about very high dimensions, since typically the test statistic T is low dimensional.

References

- [1] Atkins, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society B*, **53**, 111-142 (with discussion).
- [2] Bayarri, M.J. (1986). A Bayesian Goodness of fit test. Proceedings of the 1985 Joint Statistical Meetings (ASA, IMS, ENAR and WNAR). Also, Carnegie Mellon University Tech. Rept.
- [3] Bayarri, M.J., and Berger, J.O. (1994). Applications and limitations of robust Bayesian bounds and Type II MLE. In *Statistical Decision Theory and Related Topics V* (S.S. Gupta, and J.O. Berger, eds.) 121-134. New York, Springer-Verlag.
- [4] Berger, J.O. (1980/85). *Statistical Decision Theory and Bayesian Analysis. Second Edition*. New York: Springer-Verlag.
- [5] Berger, J.O., and Delampady, M. (1987). Testing precise hypothesis. *Statistical Science*, **2**, 317-352 (with discussion).
- [6] Berger, J.O., and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.
- [7] Berger, J.O., and Sellke (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.
- [8] Bernardo, J.M., and Smith, A.F.M. (1994). *Bayesian Theory*. New York: John Wiley.
- [9] Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A*, **143**, 383-430.

- [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] Draper, D. (1996). Comment: Utility, sensitivity analysis, and cross-validation in Bayesian model checking. *Statistica Sinica*, **6** , 760-767.
- [13] Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics*, **26**, 1125-1143.
- [14] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320-328.
- [15] Geisser, S. (1983). *Proceedings of the 1982 Meeting of the International Statistical Institute*. 925-927.
- [16] Gelfand, A.E., and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society B*, **56**, 501-514.
- [17] Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 147-167 (with discussion). Oxford: University Press.
- [18] 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).
- [19] Goldstein, M. (1991). Comment on Posterior Bayes factors (by M. Atkins). *Journal of the Royal Statistical Society B*, **53**, 134.
- [20] Good, I.J. (1953). The appropriate mathematical tools for describing and measuring uncertainty. In *Uncertainty and Business Decisions* (C.F. Carter, G.P. Meredith, and G.L.S. Shackle, eds.), 19-34. Liverpool: University Press.
- [21] Good, I.J. (1956). The surprise index for the multivariate normal distribution. *Annals of Mathematical Statistics* **27**, 1130-1135.
- [22] Good, I.J. (1981). Some logic and history of hypothesis testing. In *philosophical Foundations of Economics* (J.C. Pratt, ed.), 149-174.

- [23] Good, I.J. (1982). Comment on diversity as a concept and its measurement, by Patil and Taillie. *Journal of the American Statistical Association*, **77**, 561-563.
- [24] Good, I.J. (1982b). Comment on Lindley's Paradox, by G. Shaffer. *Journal of the American Statistical Association*, **77**, 342-344.
- [25] Good, I.J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: University of Minnesota Press.
- [26] Good, I.J. (1988). Surprise index. *Encyclopedia of Statistical Sciences* (Kotz, S., Johnson, N.L., and Reid, C.B., eds.) **7**, 104-109.
- [27] Goutis, C., and Robert, C. (1998). Model choice in generalized linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika* (to appear)
- [28] Gutiérrez-Peña, E., and Walker, S.G. (1996). A Bayesian predictive approach to model selection. Technical Report TR-96-14. Department of Mathematics, Imperial College.
- [29] 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.
- [30] 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.
- [31] Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 928-934.
- [32] Kass, R.E., and Wasserman, L. (1996). Comment on "Posterior predictive assessment of model fitness via realized discrepancies" (by Gelman, A., Meng, X.L., and Stern, H.). *Statistica Sinica*, **6**, 774-779.
- [33] Krelle, W. (1957). *Econometrica*, **25**, 618-619.
- [34] Laud, P.W., and Ibrahim. J.G. (1995). Predictive model selection. *Journal of the Royal Statistical Society B*, **57**, 247-262.
- [35] Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley and Sons.
- [36] Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142-1160.

- [37] Martin, C.L., and Meeden, G. (1984). The distance between the prior and the posterior distributions as a measure of surprise. Unpublished Manuscript.
- [38] O'Hagan, A. (1994). *Kendall's Advance Theory of Statistics, Volume 2B: Bayesian Inference*. London: Edward Arnold.
- [39] O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, **57**, 99-138 (with discussion).
- [40] Pettit, L.I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society B*, **52**, 175-184.
- [41] Poskitt, D.S. (1987). Precision, Complexity and Bayesian model determination. *Journal of the Royal Statistical Society B*, **49**, 199-208.
- [42] Renyi, A. (1961). (????). *Proc. Fourth Berkeley Symposium Math. Statist. Prob.*, **1**, 547-561.
- [43] Roberts, H.V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, **60**, 50-62.
- [44] Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.
- [45] Sackle, G.L.S. (1949). *Expectations in Economics*. London: Cambridge University Press.
- [46] San Martini, A., and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society B*, **46**, 296-303.
- [47] Verdinelli, I., and Wasserman, L. (1996). Bayesian Goodness of fit testing using infinite dimensional exponential families. *Tech. Rep.* 1996. Carnegie Mellon University.
- [48] Weaver, W. (1948). Probability, rarity, interest and surprise. *Scientific Monthly*, **67**, 390-392.
- [49] Weaver, W. (1963). *Lady Luck: The Theory of Probability*. New York: Doubleday.