BAYESIAN INFERENCE PROCEDURES DERIVED VIA THE CONCEPT OF RELATIVE SURPRISE

Michael Evans
Department of Statistics
University of Toronto
Toronto, Ontario M5S 3G3

Keywords: observed relative surprise; Bayesian hypothesis testing; principle of least relative surprise; model checking; change of variable problem; cross-validation.

ABSTRACT

We consider the problem of deriving Bayesian inference procedures via the concept of relative surprise. The mathematical concept of surprise has been developed by I.J. Good in a long sequence of papers. We make a modification to this development that permits the avoidance of a serious defect; namely, the change of variable problem. We apply relative surprise to the development of estimation, hypothesis testing and model checking procedures. Important advantages of the relative surprise approach to inference include the lack of dependence on a particular loss function and complete freedom to the statistician in the choice of prior for hypothesis testing problems. Links are established with common Bayesian inference procedures such as highest posterior density regions, modal estimates and Bayes factors. From a practical perspective new inference procedures arise that possess good properties.

1. INTRODUCTION

The mathematical concept of surprise has its origins in Weaver (1948, 1963) and in a sequence of papers by Good (1953, 1955, 1956, 1971, 1981, 1982a, 1982b, 1983a, 1983b, 1985, 1988, 1989). Further references to this concept can be found in Bartlett (1952), Kvalseth (1987) and Redheffer (1951). Also Box (1980) discussed the use of this concept in Bayesian model checking contexts. A primary thesis of this paper is that a modified version of surprise serves as a logical foundation for developing a range of estimation, hypothesis testing, model checking and model selection procedures within the context of a Bayesian model.

Before developing this modified version of surprise we establish some notation for the remainder of the paper. A probability model is denoted by $(\mathcal{X}, \mathcal{A}, P)$ where P has density f with respect to support measure μ. A Bayesian statistical model will be comprised of a basic statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta | \theta \in \Omega\})$, where P_θ has density f_θ with respect to support measure μ, and a prior probability model $(\Omega, \mathcal{B}, \Pi)$, where Π has density π with respect to support measure ν. We denote the posterior probability measure given x_0 by $\Pi(\cdot|x_0)$ and the posterior density with respect to support measure ν by $\pi(\theta|x_0) = f_\theta(x_0)\pi(\theta)/m(x_0)$ where $m(x) = \int f_\theta(x)\pi(\theta)\nu(d\theta)$ is the density, with respect to μ, of the marginal measure M of the data. If $T : (\Omega, \mathcal{B}) \to (\mathcal{T}, \mathcal{C})$ then we will denote the marginal prior and posterior measures induced on $(\mathcal{T}, \mathcal{C})$ by Π_T and $\Pi_T(\cdot \mid x_0)$ respectively and the marginal prior and posterior densities, with respect to some support measure ν_T, by π_T and $\pi_T(\cdot \mid x_0)$ respectively. Typically the function T arises as some marginal parameter that we wish to make inferences about.

We have given formal measure-theoretic definitions of the basic ingredients of the problems we will discuss. Our only purpose in doing this is to emphasize the dependence of the densities on the particular choices of support measures. This is an important part of some of our discussion and in particular when we are dealing with the change of variable problem. Thus we want to acknowledge this dependence explicitly. No measure theory is used in the paper, however.

The basic ideas behind the developments here are simple to state. Suppose that we observe data x_0 from a statistical model and that we have a prior distribution on the parameter of the model. Consider a set $\mathcal{T}$ of possible values for some quantity $T(\theta)$ depending on the parameter of the model. We totally order the elements of $\mathcal{T}$ as follows: *t_1 is strictly preferred to t_2 if the relative increase in belief for t_1, from a priori to a posteriori, is greater than the corresponding increase for t_2*. We translate this mathematically into strictly prefering t_1 to t_2 whenever

$$\frac{\pi_T(t_1|x_0)}{\pi_T(t_1)} > \frac{\pi_T(t_2|x_0)}{\pi_T(t_2)}. \tag{1}$$

We use this preference ordering to determine inferences.

In an estimation context, where we are required to select a value from $\mathcal{T}$ as an estimate, this leads to selecting a value in $\mathcal{T}$ that has the greatest relative increase in belief from a priori to a posteriori; i.e. select a value of t maximizing $\pi_T(t|x_0)/\pi_T(t)$. This estimator is computed by maximizing this ratio as a function of t. We call such an estimate a *least relative surprise* estimate and justify this terminology below.

In hypothesis testing contexts we have an hypothesized true value $t_0 \in \mathcal{T}$ for $T(\theta)$ and are required to assess this hypothesis using the evidence provided by the data. The above preference ordering leads to comparing the relative increase in belief for t_0, from a priori to a posteriori, with this increase for each of the other possible values in $\mathcal{T}$. If the increase for t_0 is small compared to the other increases then the data suggests that t_0 is surprising and we have evidence against the hypothesis. There are numerous ways in which this comparison can be made, generalizing ideas found in Good's papers, but we will use the posterior probability of obtaining a relative increase larger than that observed for t_0 and refer to this as the *observed relative surprise* hereafter. Therefore the observed relative surprise at t_0 is given by

$$\Pi_T \left(\frac{\pi_T(t|x_0)}{\pi_T(t)} > \frac{\pi_T(t_0|x_0)}{\pi_T(t_0)} \, |x_0 \right). \tag{2}$$

Notice that the value of t_0 minimizing (2) is the least relative surprise estimate. It is the value most supported by the data, and so least surprising, when the relative change in degree of belief from a priori to a posteriori is our criterion for assessing this. This motivates our choice of terminology.

The hypothesis testing approach via observed relative surprise can be inverted in a standard way to give *relative surprise regions* for the unknown true value in $\mathcal{T}$. An $\alpha-$relative surprise region for $T(\theta)$ is given by

$$C_\alpha(x_0) = \{t_0 \in \mathcal{T} \mid \Pi_T \left(\frac{\pi_T(t|x_0)}{\pi_T(t)} > \frac{\pi_T(t_0|x_0)}{\pi_T(t_0)} \, |x_0 \right) \leq \alpha\}. \tag{3}$$

This is the set of values in $\mathcal{T}$ whose observed relative surprise is no greater than α. For example, if $\mathcal{T} \subset \mathcal{R}$ then we compute an α-relative surprise interval for $T(\theta)$ by tabulating (2) as a function of t_0. Then when $T(\theta)$ is continuous we compute the interval (t_1, t_2) such that taking $t_0 = t_1$ and $t_0 = t_2$ makes (2) equal to α. When the parameter is discrete this procedure requires some obvious modifications as we are not guaranteed to get an exact α-relative surprise interval.

We now present two simple examples where this approach is seen to lead to inference procedures different from those obtained by some standard Bayesian and non-Bayesian approaches and which possess some good properties.

Example 1. *Estimating a sum of squared means.*

Efron in his discussion of Dawid, Stone and Zidek (1973) considered the following as showing that Bayesian inference with proper priors can lead to what he regards as a poor inference. Suppose we observe $x = (x_1, \ldots, x_n)$ where $x_i \sim N(\theta_i, 1)$ and $x_1, \ldots, x_n$ are statistically independent and $\theta_1, \ldots, \theta_n$ are given independent $N(0, \sigma^2)$ prior distributions with σ^2 large and known. Suppose further that our interest is in estimating $\tau^2 = \sum_{i=1}^n \theta_i^2$. Notice that the prior distribution of τ^2 is σ^2Chisquare$(n, 0)$ and the posterior distribution of τ^2 is $(1 + \frac{1}{\sigma^2})^{-1}$Chisquare$(n, (1 + \frac{1}{\sigma^2})^{-1}||x||^2)$ where Chisquare(n, δ) denotes the Chisquare distribution with n degrees of freedom and noncentrality δ. So to compute the the least relative surprise estimate we must maximize the ratio of these densities or equivalently, which is what we do here, find the value where the observed relative surprise is smallest.

For convenience we will consider the limiting case as $\sigma^2 \to \infty$ as this makes no difference when considering Efron's point. Then the limiting Bayes estimate of τ^2, when the loss function is squared error, is given by $||x||^2 + n$ with MSE given by $4\tau^2 + 2n + 4n^2$. The UMVU estimator by contrast is $||x||^2 - n$ with MSE given by $4\tau^2 + 2n$. Several other estimators are possible. For example, the plug-in MLE is $||x||^2$ with MSE equal to $4\tau^2 + 2n + n^2$ and this is equivalent to using the limiting posterior mode as a plug-in estimate. Observing that the

τ^2	n	UMVU	(UMVU)$_+$	MLE	Bayes	MPME	LRSE
0	5	10.00	6.59	35.00	110.00	61.88	8.30
1	5	14.00	9.64	39.00	114.00	65.25	11.86
2	5	18.00	13.36	43.00	118.00	68.75	15.96
5	5	30.00	26.33	55.00	130.00	79.76	29.57
10	5	50.00	48.54	75.00	150.00	98.99	51.50
100	5	410.00	406.19	435.00	510.00	454.50	407.18
0	20	40.00	22.91	440.00	1640.00	1430.20	25.21
1	20	44.00	25.96	444.00	1644.00	1432.34	28.48
2	20	48.00	29.73	448.00	1648.00	1435.21	32.49
5	20	60.00	44.10	460.00	1660.00	1444.24	47.43
10	20	80.00	71.12	480.00	1680.00	1460.56	74.95
100	20	440.00	441.00	840.00	2040.00	1799.22	442.24

Table 1: MSE's of estimators of τ^2 in Example 1.

UMVU estimator can be negative we could truncate this using $(||x||^2 - n)_+$. The MSE of the truncated UMVU estimator cannot be obtained in closed form. Alternatively we could use the mode of the limiting posterior distribution of τ^2 and we refer to this as the marginal posterior mode estimate (MPME). The MPME cannot be obtained in closed form. In Table 1 we present the results of a simulation study for these estimators together with the least relative surprise estimate (LRSE) described above. For the cases where the MSE cannot be calculated exactly the reported values are felt to be accurate with a relative error not exceeding 1%. We see immediately the generally poor performance of all the Bayesian estimates except for the LRSE. The LRSE has performance comparable to the UMVU or truncated UMVU. Further the LRSE is always nonnegative.

The LRSE in this example is equivalent to the estimator derived in Saxena and Alam (1982) and referred to there as the MLE although it would only be the maximum likelihood estimator when the observed data is $||x||^2$. That paper proves that (UMVU)$_+$ is uniformly better than the LRSE with respect to MSE and this is reflected in our table. We note again, however, that the LRSE has performance roughly comparable with (UMVU)$_+$. For example, from Table 1 we see that the largest relative error between the MSE of the LRSE and that of (UMVU)$_+$ is approximately 26% when $\tau^2 = 0$ and $n = 5$ and for most of the remaining cases it is considerably less. The estimation of τ^2 is also discussed in Perlman and Rasmussen (1975).

Example 2. *Interval for the sum of squared means*

We consider an example discussed in Stein (1959) concerning a strong contradiction between a frequency approach to the construction of intervals for a parameter and a fiducial approach. The fiducial interval in question also arises from a Bayesian analysis and it is in this context that we consider it here.

The ingredients of this example are basically the same as those given in Example 1 but we are now concerned with constructing an interval to contain τ^2 with high probability. An exact $\alpha-$confidence interval for τ^2 is given by the set $\{\tau^2|(1-\alpha)/2 \leq F_{\tau^2,n}(||x||^2) \leq (1+\alpha)/2\}$ where $F_{\tau^2,n}$ is the distribution function of the Chi-square(τ^2, n) distribution. This confidence interval has the unnatural property, however, of equaling the null set with positive probability for all parameter values. Bayesian intervals can be formed from the limiting marginal posterior of τ^2 either by using the highest posterior density (HPD) interval or by discarding $(1-\alpha)/2$ of the probability from each tail of the marginal posterior. We will refer to this latter interval as a Bayesian confidence interval (BCI). Finally we consider the relative surprise interval (RSI) described above. The RSI is computed, as earlier described, by tabulating the observed relative surprise using the prior and posterior distributions for τ^2 described in Example 1. Table 2 gives the result of a simulation of the coverage probabilities of these intervals based on 10^4 samples. Clearly the HPD and BCI perform very poorly in certain situations with respect to coverage. In particular in some cases the coverage of these intervals is virtually 0 while the posterior probability content is close to 1. This is what we mean by a strong contradiction. We note that the RSI interval, which is always finite and never null in this example, avoids the strong contradiction. The RSI does this by correcting for the extreme skewness in the posterior distribution of τ^2. In fact the right-hand end-points of the RSI intervals tend to be smaller than those of the other Bayesian intervals and the left-hand end-points are closer to 0. In the cases where the differences between the coverages are substantial there is typically a big difference in the left-hand end-points of the RSI and the other Bayesian intervals. We note that the coverage of the RSI interval is not always equal to the nominal posterior probability but in all the examples we have looked at, the error is always on the conservative side.

We further discuss the motivation for the various definitions associated with relative surprise and provide additional examples in section 2. In section 3 inference for prediction and the model checking procedures of Box (1980) are developed using relative surprise. The extensions are similar in spirit to those underlying cross-validation and the prequential inference of Dawid (1984). In section 4 we draw some conclusions and point to future research on this topic.

We note that the development here specifically excludes the inclusion of a loss function. In doing this we are not asserting the superiority of this approach to inference over a decision-theoretic one. Rather we are simply developing an approach to inference based on what we perceive to be an appealing concept: *inferences are determined by how our beliefs change from a priori to a posteriori.* The value in this approach can be measured by the appeal of this idea, the lack of the need to specify a loss function when this is not feasible and, most importantly, the properties of the inferences it generates. For other discussions of hypothesis testing and related problems in Bayesian inference see, for example, Aitkin (1991), Berger and Delampady (1987), Berger and Selke (1987) and Kass and Raftery (1995). Quantitative approaches to surprise also appear

τ^2	n	α	HPD	BCI	RSI
0	5	.95	0.0000	0.0000	1.0000
1	5	.95	0.4421	0.0324	1.0000
100	5	.95	0.9427	0.9369	0.9824
0	20	.95	0.0000	0.0000	1.0000
1	20	.95	0.0000	0.0000	1.0000
100	20	.95	0.6085	0.7722	1.0000
0	5	.99	0.0000	0.0000	1.0000
1	5	.99	0.7980	0.4941	1.0000
100	5	.99	0.9887	0.9894	0.9979
0	20	.99	0.0000	0.0000	0.9999
1	20	.99	0.0000	0.0000	1.0000
100	20	.99	0.8266	0.8232	1.0000

Table 2: Coverages of intervals for τ^2 in Example 2.

in the work of Levi (1972) and Shackle (1949) but these appear to have little relevance to our treatment.

2. SURPRISE, RELATIVE SURPRISE AND BAYESIAN INFERENCE

The intuitive idea behind the concept of surprise can be most easily expressed when all probability measures are discrete. Initially suppose then that we have a single probability model and that μ is counting measure. The basic idea is that the occurrence $x_0 \in \mathcal{X}$ is surprising if the value of $f(x_0) = P(\{x_0\})$ is small when compared to all the other possible values for $f(x)$. If we conclude that a surprising value has occurred then we have evidence against the validity of the probability model. Therefore an observation is not surprising simply because it has a small probability of occurence but it must also have a probability of occurrence that is small when compared to all the other probabilities of occurrence.

We want a numerical measure of surprise and there have been several proposed. For example, Weaver (1948) proposed using the *surprise index*

$$\lambda_1 = \frac{\mathrm{E}_P[f]}{f(x_0)} = \frac{\int f^2(x)\mu(dx)}{f(x_0)}. \tag{4}$$

The larger λ_1 is, the more surprising is the observation x_0. Good (1953) generalized this to a class of indices given by $\lambda_r = \left(\mathrm{E}_P[f^r]\right)^{1/r}/f(x_0)$, when $r > 0$, and $\lambda_0 = \exp\left(\mathrm{E}_P[\log f]\right)/f(x_0)$.

There are several problems with surprise indices. For example, for some models they may be identically infinite. More significantly we have no idea of how large a surprise index has to be to conclude that a surprising event has occurred. As a perhaps more natural measure of how relatively small $f(x_0)$ is we define the *observed surprise* as

$$P\left(f(X) > f(x_0)\right). \tag{5}$$

The observed surprise is simply the probability of observing an event whose probability of occurrence is greater than the probability of the event that has occurred. If this is large; i.e. close to 1, then a surprising event has occurred. We note that the observed surprise is a probability and hence we have a very natural scale on which to calibrate it. For example, as in hypothesis testing, we could choose a cut-off such as .95 to determine when a surprising event has occurred. We can interpret the observed surprise as a measure of the evidence against P provided by the observation just as we interpret a P-value. A suggestion virtually equivalent to using the observed surprise to assess surprise is made in Good (1988).

If we are required to predict a value in $\mathcal{X}$, given the probability model, then Good (1988) has suggested choosing the least surprising value and called this the *principle of least surprise.* When we choose the observed surprise as our measure this leads to using the value $x_0 \in \mathcal{X}$ that minimizes (5) or equivalently, maximizes f.

For the general situation we define the surprise index and the observed surprise as in (4) and (5) respectively. The interpretation is now somewhat ambiguous, however, as it is clear that the value of these quantities depends on the choice we have made for μ and in general there is not a natural choice as in the discrete case. It might be argued that in many contexts we should take μ to be Lebesgue measure as a canonical support measure. This does not avoid the difficulty, however. For if we make a change of variable, and still insist on having our densities be relative to μ, then the surprise indices and the observed surprise will generally change due to the Jacobian factor. We refer to this dependence on μ as the *change of variable problem.* As a simple example suppose that X has density $f_X(x) = 2x$ on $[0,1]$ and we make the transformation $Y = X^4$. Then Y has density $f_Y(y) = y^{-1/2}/2$ on $[0,1]$. From this we see that $x_0 = 0$ has observed surprise 1 under X and observed surprise 0 under Y while $x_0 = 1$ has observed surprise 0 under X and observed surprise 1 under Y.

One possible approach to avoiding the change of variable problem is to require that the statistician initially make a specific choice of μ and then interpret the value $f(x)$ as a relative probability of occurrence. By this we mean that if $f(x_1) > f(x_2)$ then the probability of x_1 occurring is greater than the probability of x_2 occurring. Provided the density is smooth then this interpretation seems appropriate. This induces a total ordering on the elements of $\mathcal{X}$. This kind of ordering is certainly the way in which densities are commonly interpreted when they are expressed with respect to Lebesgue measure. In fact some Bayesian inference procedures, such as highest posterior density regions and estimation via the posterior mode, depend implicitly on such an interpretation of a density. Then to keep the total ordering invariant under a change of variable, the Jacobian factor must be allocated with μ to create a new support measure μ^*. With this qualification the change of variable problem has been partially resolved as the observed surprise is now invariant under a change of variable.

There still remains, however, a significant problem. Our previous discussion has been based on comparing the probability of occurrence of x_0 with the probability of occurrence of any other $x \in \mathcal{X}$. Here we are thinking of $\mathcal{X}$ as representing a basic set of possible observations. It is often very natural, however, to base our evaluation on the value of $T(x_0)$ for some $T : \mathcal{X} \to \mathcal{T}$. The obvious way to do this in the discrete context is to compute the observed surprise $P\left(P(\{T(X)\}) > P(\{T(x_0)\})\right)$. If we have determined a canonical support measure on $\mathcal{X}$ it is not clear, however, what support measure we should use on $\mathcal{T}$. Intuitively such a support measure on $\mathcal{T}$ should be related to the support measure we placed on $\mathcal{X}$ and which determined the total orderings on the basic possible observations. If T is 1-1 then we can proceed as above but in general it is not obvious how to go about this even in discrete contexts.

The various surprise indices, the observed surprise and the principle of least surprise can all be applied in the Bayesian context to yield inferences. For example, the principle of least surprise yields the mode of $\pi_T(\cdot \mid x_0)$ as the estimate of $T(\theta)$. The observed surprise for testing $H_0 = \{t_0\}$; i.e. we hypothesize $T(\theta) = t_0$, is equivalent to the Bayesian $P-$value $\Pi_T(\pi_T(t \mid x_0) > \pi(t_0 \mid x_0) \mid x_0)$. Inverting the observed surprise to form $\alpha-$surprise intervals gives HPD intervals. So applying surprise to the Bayesian model yields some standard inference techniques; see for example, Box and Tiao (1973). We note, however, that the difficulties for surprise associated with the change of variable problem apply to all of these methods.

These problems can be avoided in the Bayesian context, however, by using the observed relative surprise, as defined in (2), for hypothesis testing, constructing $\alpha-$relative surprise regions and least relative surprise estimates. It is immediate that these inferences are free of the change of variable problem as the Jacobian factor occurs in both the numerator and denominator of the density ratios and hence cancels. Thus the observed relative surprise is invariant under a change of variable, or equivalently, choice of support measure. For a given T a support measure always exists for the definition of the observed relative surprise; e.g. $\nu_{\mathcal{T}} = \Pi_T + \Pi_T(\cdot \mid x_0)$.

We note that the observed relative surprise is not just an arbitrary adjustment of the observed surprise to correct for the change of variable problem. Rather it arises in a natural inferential way when we ask about the relative change in our belief in t_0, from a priori to a posteriori, when compared to this change for other possible values for T. The observed relative surprise is the posterior probability of this change in relative belief being greater than that observed for the hypothesized value. The value t_0 is surprising when this posterior probability is large as this says that our current belief is that the relative increase in support for t_0 is not large when compared to that of other values. If the posterior degree of belief in t_0 is smaller than the prior degree of belief in t_0 then the data is providing evidence against the truth of t_0 and conversely. The size of the increase or decrease in the degree of belief is not sufficient information to base a conclusion concerning what the data is saying about t_0. We have to take into account the changes in the degree of belief for the other possible values of T as well. For suppose the degree of belief in t_0 increases. If the relative increase in the degree of belief is greater for a large proportion of the elements of $\mathcal{T}$ than it is for t_0 then we have no grounds for concluding that the data supports t_0 over the other possibilities.

The ultimate test of the relative surprise approach is through examples. We recall here the significant improvements in Bayesian inferences noted in Examples 1 and 2 at least when relative frequency considerations are raised. We will present a number of additional examples where the relative surprise approach leads to inferences in a Bayesian context that possess good properties again from the frequency point of view. We focus initially on Bayesian hypothesis testing.

Example 3. *Testing* H_0 *versus* H_0^c

Suppose we wish to test the hypothesis that the true value of θ is in $H_0 \subseteq \Omega$ versus θ is in H_0^c. Further suppose that $\Pi(H_0) \neq 0$ and $\Pi(H_0^c) \neq 0$. This is equivalent to testing $T(\theta) = 1$ versus $T(\theta) = 0$ where T is the indicator function for H_0. It is easy to show then that the observed relative surprise at $t_0 = 1$ is equal to $\Pi(H_0^c|x_0)$, whenever the Bayes factor against H_0 is greater than 1; i.e. when the ratio of the posterior odds against H_0 to the prior odds against H_0 is greater than 1, and the observed relative surprise is equal to 0 otherwise. Thus the relative surprise approach produces the usual Bayesian answer in this problem with the natural adjustment that we have no evidence against H_0 whatsoever; i.e. the observed relative surprise is 0, when the data has not produced a relative increase in our belief in H_0^c greater than the corresponding relative increase for H_0. This is a combination of the Bayes factor approach and just using the posterior probability $\Pi(H_0^c \mid x_0)$ to assess the evidence against an hypothesis. See Kass and Raftery (1995) for an extensive discussion of the Bayes factor approach to hypothesis testing.

In example 3 we have required that H_0 and H_0^c be non-null with respect to the prior. If $\Pi(H_0) = 0$ then $\Pi(H_0|x_0) = 0$ for every $x_0 \in \mathcal{X}$ and the posterior probability approach for assessing the evidence against H_0 would always reject while the Bayes factor is not defined. This phenomenon will occur, in spite of the fact that the prior density may be indicating a relatively high degree of belief in H_0, whenever H_0 is a lower dimensional subset of the parameter space. While the context where $\Pi(H_0) = 0$ is clearly a problem, this is only an extreme case of the situation where the prior assigns a small amount of probability to a set simply because it is a small set or assigns a large amount of probability to a set simply because it is a large set. For example, in testing the validity of a scientific theory H_0 it may be perfectly reasonable, and in fact preferable, to have a fairly diffuse prior that does not assign a significant amount of probability to H_0 to allow the data to dominate the inference. In such a situation we may require an inordinately large data set to conclude that H_0 is true using $\Pi(H_0|x_0)$, even when H_0 holds.

One approach to avoiding this problem is commonly advocated; namely modify the prior so that H_0 receives a sizeable amount p of prior probability. In many cases it may be appropriate to do this; i.e. the statistician truly believes that H_0 deserves p of the prior probability. But in general this requires modification of what may be a perfectly reasonable prior reflecting the statistician's degrees of belief concerning the parameter. So a method of assessing the evidence against H_0, that allows the null hypothesis to have 0 prior probability, is necessary in our view. A further difficulty for the approach of requiring a positive prior probability for H_0 is exemplified by the following example.

Example 4. *Lindley's Paradox*

Suppose that $x_0 = (x_1, \ldots, x_n)$ is a sample from a $N(\theta, 1)$ distribution where $\theta \in R^1$ is unknown and we want to test the hypothesis $\theta \in H_0 = \{0\}$. As a possible prior for θ we could use the $N(0, \sigma^2)$ distribution which we denote by Π_1. Taking σ very large reflects diffuse prior knowledge concerning the true value of θ. We note that we cannot use Π_1 if we want to proceed as in Example 3. So, following the above discussion, we also consider the prior Π_2 that places a positive mass at 0 obtained by mixing Π_1 with a degenerate distribution at 0 to obtain $\Pi_2 = p\delta_0 + (1-p)\Pi_1$. Then under Π_1 the posterior distribution of θ is $N((n + 1/\sigma^2)^{-1} n\bar{x}_0, (n + 1/\sigma^2)^{-1})$. Under Π_2 the posterior is a mixture of $\Pi_1(\cdot|x_0)$ with the distribution degenerate at 0 with probability $(1-p)m(x_0)/(pf(x_0) + (1-p)m(x_0))$ where f is the marginal density of the data when H_0 holds and m is the marginal density of the data when H_0^c holds.

Using the prior Π_2 the Bayes factor against H_0 equals

$$\mathrm{BF} = (n\sigma^2 + 1)^{-1/2} \exp\left\{n^2\sigma^2(n\sigma^2 + 1)^{-1}\bar{x}_0^2\right\}$$

and $\Pi_2(H_0|x_0) = (1 + \frac{1-p}{p}\mathrm{BF})^{-1}$. Now suppose that we fix $\sqrt{n}\bar{x}_0$ and let $\sigma \to \infty$. Then $BF \to 0$, $\Pi_2(H_0|x_0) \to 1$ and, by Example 3, the observed relative surprise goes to 0. Thus by choosing σ very large we can virtually guarantee that we will accept H_0 by any of these three approaches to hypothesis testing. If, however, $\sqrt{n}\bar{x}_0 = 3$ then the classical Z-test will categorically reject H_0 but the Bayesian tests will accept when σ is large. As we expect the classical and Bayesian approaches to be similar under a diffuse prior, this conflict between the two approaches to inference is known as *Lindley's paradox*; see Lindley (1957). There are various points of view concerning this paradox and some argue strongly in favour of the Bayesian approaches; see for example Berger and Delampady (1987) and Berger and Selke (1987). There have also been suggested resolutions, see Aitkin (1991) and Robert (1993), but these lie outside proper Bayesian inference and for this reason they are not satisfying. Sometimes the complaint is made that it is artificial to allow $\sigma \to \infty$. The paradox also applies, however, as $n \to \infty$.

We now consider the application of relative surprise to this problem when we use the prior Π_1 and take $T(\theta) = \theta$ with $t_0 = 0$. In this case the observed relative surprise equals

$$P\left(\chi_1^2(\frac{\bar{x}_0^2}{\sigma^2(n\sigma^2+1)}) < \bar{x}_0^2(n + 1/\sigma^2)\right) \tag{6}$$

where $\chi_1^2(\delta)$ is distributed Chisquare(δ, 1). If we fix $\sqrt{n}\bar{x}_0$ and let $\sigma \to \infty$ then we see that (6) converges to the observed level of significance of the classical two-sided Z-test. Also as $n \to \infty$ the same limiting observed relative surprise is obtained. Therefore the test of H_0 via the observed relative surprise agrees with the classical test asymptotically.

It cannot be claimed that the resolution of the paradox occurs because of the use of the observed relative surprise. As noted in Casella and Berger (1987) the paradox is caused by the discrete prior mass at 0. Other measures of evidence which avoid the need for H_0 to have positive prior probability may also avoid the paradox. For example, it is easy to show that using the observed surprise with Π_1 also leads to the classical Z-test as $\sigma^2 \to \infty$ or $n \to \infty$. We recall, however, that the observed surprise suffers from the change of variable problem. The question then is what is an appropriate measure of the evidence in a Bayesian hypothesis testing problem? Of course we are suggesting that the observed relative surprise is a good candidate. It is satisfying that the observed relative surprise avoids the paradox, allows complete freedom to the statistician in the choice of prior and essentially reproduces the standard Bayesian answer when H_0 does have positive prior mass.

We note an additional difference between the two tests of hypothesis discussed in Example 4. For when using Π_2, and proceeding as in Example 4, we are comparing the evidence for H_0 with the evidence for H_0^c. On the other hand when using Π_1 and T we are comparing the evidence for t_0 with the evidence for each of the other possible values for T. While it is sometimes argued that we must carry out the hypothesis testing problem using the binary partition $\{H_0, H_0^c\}$, in many problems H_0 arises as the specification of the value t_0 of a particular function T of interest. In such a situation it seems natural to us to evaluate the evidence by comparing what the data says about t_0 with what it says about each of the other possible values. Of course we are not suggesting that one should never use the binary partition, as there are contexts where it is appropriate, but it seems too restrictive to us to limit Bayesian hypothesis testing procedures to only using this approach.

It is sometimes claimed that it is not appropriate to use point null hypotheses as we never really believe that this holds exactly but rather only that the true value is in some small interval about the hypothesized point. The following example shows that the observed relative surprise approach has a satisying property with respect to this consideration.

Example 5. $\epsilon -$ *partitions*

Suppose that $\Omega = R^1$, $\epsilon > 0$ and $H_i^\epsilon = [\theta_0 + (i - 1/2)\epsilon, \theta_0 + (i + 1/2)\epsilon]$ for $i \in Z$. Define $T^\epsilon : \Omega \to Z$ by $T^\epsilon(\theta) = i$ when $\theta \in H_i^\epsilon$. Further suppose that $\Pi(H_i^\epsilon) > 0$ for all i and $\epsilon > 0$. If $t_0 = 0$ then the observed relative surprise at t_0 equals $\Pi\left(\cup_{i \in Z^\epsilon} H_i^\epsilon \mid x_0\right)$ where

$$Z^\epsilon = \left\{ i : \frac{\Pi(H_i^\epsilon | x_0)}{\Pi(H_i^\epsilon)} > \frac{\Pi(H_0^\epsilon | x_0)}{\Pi(H_0^\epsilon)} \right\}.$$

Then, provided that $\pi(\theta \mid x_0)/\pi(\theta)$ is continuous in θ it can be shown that

$\Pi\left(\cup_{i \in Z^\epsilon} H_i^\epsilon \mid x_0\right)$ converges to (2), with $T(\theta) \equiv \theta$, as $\epsilon \to 0$. Therefore the test of the null hypothesis $H_0 = \{\theta_0\}$, via the observed relative surprise, can be thought of as an approximation to some particular discrete partition of the parameter space into small intervals.

Example 5 can be straight-forwardly generalized to a much wider class of models. It illustrates that there is no fundamental conflict in the relative surprise approach between hypothesis testing via point null hypotheses or via quantizations into non-overlapping subsets. The point null approach will typically have the advantage, however, of being somewhat simpler computationally. As with most uses of continuous probability the observed relative surprise for a point null can be thought of as an approximation to a discrete reality.

We conclude this section with an example of the application of the principle of least relative surprise to a Bayesian inference problem.

Example 6. *Estimating cell probabilities*

Suppose we observe $(f_1, \ldots, f_k) \sim \text{Multinomial}(n, p_1, \ldots, p_k)$ and $(p_1, \ldots, p_k)$ has prior given by a Dirichlet$(1, \ldots, 1)$ distribution. Also suppose that we are interested in estimating p_1. Then the posterior distribution of p_1 is Dirichlet$(f_1 + 1, f_2 + \ldots + f_k + k - 1)$. The posterior mean of p_1 is given by $(f_1 + 1)/(n + k)$ and the marginal posterior mode is $f_1/(n + k - 2)$. Both of these estimates have the property of depending on the number of classes k. As such, when k is very large these estimates will be very biased. The least relative surprise estimate, however, is given by f_1/n, and this is the UMVU estimate of p_1. Thus the relative surprise approach leads to the standard frequentist estimate without having to modify the prior. For situations where the statistician does not want prior beliefs to strongly influence the estimate this seems like an appropriate inference. An improper prior $\prod_{i=1}^k p_i^{-1}$ is needed for the posterior mean to produce f_1/n as the estimate.

3. PREDICTION AND MODEL CHECKING

In Box (1980) it was suggested that the validity of the Bayesian model could be checked by comparing the observed value $R(x_0)$ to the distribution of $R(X)$ when $X \sim M$ for various functions $R : (\mathcal{X}, \mathcal{A}) \to (\mathcal{R}, \mathcal{D})$. Further this paper suggested using something equivalent to the observed surprise, at least in the continuous case, to determine whether or not the data x_0 contradict the Bayesian model. The observed surprise, when R is the identity, is given by

$$M\left(m(X) > m(x_0)\right). \tag{7}$$

In Box (1980) the support measure is always taken to be Lebesgue measure so for a general R we simply replace m in (7) by the density of R with respect to Lebesgue measure on the appropriate space. Notice that (7) can be used for prior prediction as well; namely before observing data the principle of least surprise leads to choosing a value x_0 that minimizes (7) as our prediction. As we discussed earlier these inferences all suffer from the change of variable problem. In many problems, under appropriate changes of variable, (7) can take any value between 0 and 1. So this problem cannot be ignored.

To avoid the change of variable problem we could approach the problem just as in Section 2. The basic idea again is to compare the relative change in the prior degree of belief in $R(x_0) = r_0$ to the posterior degree of belief in r_0, relative to changes in the degree of belief for any other possible value $R(x) = r$. For this let $f_{\theta R}$ be the density of $R(X)$ with respect to some support measure $\mu_{\mathcal{R}}$ on $\mathcal{R}$ when $X \sim f_\theta$. Then $m_R(r) = \int f_{\theta R}(r)\pi(\theta)\nu(d\theta)$ is the *prior predictive density* of $R(X)$ at r with respect to $\mu_{\mathcal{R}}$. The *posterior predictive density* of $R(X)$ at r with respect to $\mu_{\mathcal{R}}$ is given by $m_R(r|x_0) = \int f_{\theta R}(r)\pi(\theta|x_0)\nu(d\theta)$. We then measure the relative change in the degree of belief in r by $m_R(r|x_0)/m_R(r)$. Thus the observed relative surprise at r_0 is equal to

$$M_R\left(\frac{m_R(r|x_0)}{m_R(r)} > \frac{m_R(r_0|x_0)}{m_R(r_0)} \mid x_0\right). \tag{8}$$

As before this is completely independent of any choices we have made for support measures and the change of variable problem is resolved. Further we can use (8) for prediction of a future value of R by an application of the principle of least relative surprise.

Unfortunately (8) suffers from a defect if it is to be used for model checking. For because $m_R(\cdot|x_0)$ is based on the data x_0, the value of $m_R(R(x_0) \mid x_0)/m_R(R(x_0))$ is often high or even maximal, particularly when prior beliefs are diffuse. Thus in many contexts (8) is identically 0 so that we never would conclude that the data provide evidence against the model. This is not always the case but it is difficult to characterize the situations where (8) will be useful for model checking. This phenomenon has also been observed for Box's approach to model crticism; see Geisser (1993) and example 7 below.

One approach that corrects for this defect is to use cross-validation. For this we suppose that we have a 1-1 transformation $(R, S) : (\mathcal{X}, \mathcal{A}) \to (\mathcal{R} \times \mathcal{S}, \mathcal{D} \times \mathcal{E})$ where R and S are statistically independent for every θ and the marginal model for S is indexed by the parameter θ. This ensures that a posterior for θ can be determined from the value $S(x_0)$ and that there is no information about $R(x_0)$ in the observation $S(x_0)$. For example, if x corresponds to an i.i.d. sample $x = (x_1, \ldots, x_n)$ then we could take $R(x) = (x_{n_*+1}, \ldots, x_n)$ and $S(x) = (x_1, \ldots, x_{n_*})$ for $n_* < n$ but more general splits are possible. We then calculate the posterior density $\pi(\cdot \mid S(x_0))$ for θ using the prior for θ and the marginal model for S, calculate the prior predictive density $m_R(\cdot)$ using the prior for θ and the marginal model for R and finally calculate the posterior predictive density $m_R(r \mid S(x_0)) = \int_\Omega f_{\theta R}(r)\pi(\theta \mid S(x_0))\nu(d\theta)$. These densities can be calculated using any support measures provided the same support measure is always used on a particular space. The *cross-validational observed relative surprise* is then given by

$$M_R\left(\frac{m_R(r|S(x_0))}{m_R(r)} > \frac{m_R(r_0|S(x_0))}{m_R(r_0)} \mid S(x_0)\right) \tag{9}$$

and, with the above proviso for the choice of support measures, it is free of the change of variable problem.

We consider several applications of this.

Example 7. *Checking a Bernoulli model*

Suppose that $x = (x_1, \ldots, x_n)$ is a sample from a Bernoulli(p) and $p \sim U(0,1)$ is the prior. We have that the density of x with respect to counting measure is

$$m(x) = \left[(n+1)\left(\begin{array}{c} n \\ \sum_{i=1}^n x_i \end{array}\right)\right]^{-1}.$$

Then (7) leads to the seemingly anomalous result that samples with values of $R(x) = \sum_{i=1}^n x_i$ near 0 or n are less surprising than samples with values of $R(x)$ near $n/2$. If instead we use $R(x)$ to assess the model then the density with respect to counting measure on $\{0, \ldots, n\}$ is $m_R(r) = 1/(n+1)$ and, as pointed out in Geisser (1993), (7) is useless for model criticism as the observed surprise is identically 0; i.e. we never reject the model. Although not quite as easy to see it turns out that (8) is also not useful for this purpose here. In simulation studies it leads to very low values for the observed relative surprise even with very extreme data sets.

Following the cross-validation approach, however, with $S(x) = (x_1, \ldots, x_{n_*})$ and $R(x) = (x_{n_*+1}, \ldots, n)$, the cross-validational observed relative surprise gives sensible answers. For example if $n = 100, n_* = 50$ and the first 50 observations has 35 heads while the second 50 has 20 heads then the cross-validational observed relative surprise equals 1.00 so the test rejects the model. If the second 50 observations had 29 heads the observed relative surprise is .800, while with 35 heads it is 0, with 41 heads it is .828 and with 45 heads it is .988.

Example 8. *Checking a linear model*

Suppose that $y = X\beta + \sigma e$ where we observe $y \in R^n$, $X \in R^{n \times k}$ is known, $(\beta, \sigma) \in R^k \times R^+$ is unknown and $e \sim N_n(0, I)$. We take a conjugate prior structure for the parameter; namely $\beta|\sigma \sim N_k(0, \tau\sigma^2)$ and $\sigma^{-2} \sim$ Gamma(α, η). For convenience we will consider limiting inferences as the hyperparameters τ and η go to infinity to reflect diffuse knowledge about the parameters. Under these conditions the limiting posterior distribution of the parameters is given by $\beta|\sigma \sim N_k(b_X, \sigma(X'X)^{-1})$ and $\sigma^{-2} \sim$ Gamma$(\frac{n-k}{2} + \alpha, 2||y - Xb_X||^{-2})$ where $b_X = (X'X)^{-1}X'y$. Taking $\alpha = 2$, which we will do hereafter, gives a posterior equivalent to that obtained via Jeffreys prior.

For the cross-validation we let $S(y)$ denote some subset of n_* observations and $R(y)$ denote the remaining $n_{**} = n - n_*$ observations. Let X_* and X_{**} denote the corresponding n_* and n_{**} rows of X respectively. The limiting posterior predictive distribution of $R|S(y)$ is then given by $R|S(y), \sigma \sim N_{n_{**}}(X_{**}b_{X_*}, \sigma^2(I + X_{**}(X_*'X_*)^{-1}X_{**}'))$ and $\sigma^{-2} \sim$ Gamma$(\frac{n_*-k}{2} + 2, 2||y_* - X_*b_{X_*}||^{-2})$. This gives a simple prescription for simulating from the limiting posterior predictive distribution. If we put $d(n_*, n_{**}) = (\frac{n_{**}-k}{2} + 2)/(\frac{n-k}{2} + 2)$

and

$$\begin{aligned}\psi(R) &= d(n_*, n_{**}) \log\left\{\frac{||R - X_{**}b_{**}||^2}{||S(y) - X_* b_{X_*}||^2}\right\} - \\ &\quad \log\left\{1 + \frac{(R - X_{**}b_{X_*})'(I + X_{**}(X_*'X_*)^{-1}X_{**}')^{-1}(R - X_{**}b_{X_*})}{||S(y) - X_* b_{X_*}||^2}\right\}\end{aligned}$$

then it is easy to show that the limiting cross-validational observed relative surprise is given by $M_R(\psi(R) > \psi(R(y)) \mid S(y))$ and this can be calculated via simulation. Notice that the observed surprise is carrying out a somewhat complicated comparison of the residuals $R(y) - X_{**}b_{X_{**}}$ and $R(y) - X_{**}b_{X_*}$.

In Rice (1995, problem 14.35) twenty-one (x, y) values are given where x denotes the volume of a fluid in kiloliters in a tank, which also contains a variety of mechanical devices, and y denotes pressure in pascals. The data are plotted in Figure 1. While the data look remarkably linear a plot of the residuals reveals some systematic variation beyond the linear term; e.g. see the solutions in Rice. As a test of the above methodology we fit the model with an intercept and linear term to the first 15 observations, ordered by x, and used the last 6 observations as $R(y)$. The cross-validational observed relative surprise is plotted as a function of $\psi(R(y))$ in Figure 2. For this particular data set we obtained the value $\psi(R(y)) = .11$ and the observed relative surprise is 1.00. Thus this approach rejects the possibility that the linear model fits.

Example 9. *Estimating hyperparameters*

In a given problem the prior may depend on hyperparameters as in example 4 where the prior Π_1 depended on σ. Often it is difficult to specify a value for this precisely and various non-Bayesian devices are used to make sensible choices. For example, one possibility is to maximize the marginal density of the observed data x_0 as a function of the hyperparameter. This is sometimes called Type II maximum likelihood, see for example Good (1965). This is equivalent to estimating the hyperparameter using the LRSE whenever the hyperparameter has a prior distribution. Another possibility, to avoid the possibility of over-fitting to the data, is to select the value of the hyperparameter as an application of the principle of least relative surprise applied to the cross-validational observed relative surprise used for model checking. The selected prior is informative to the extent that the specific class used expresses information about the unknown value of the model parameter.

It is not entirely clear how we should choose (R, S) in a particular problem. This will rely on the statistician's judgement as to what deviations might be expected. We note that there are some decidely poor choices for R. For example, if R is ancillary then (9) is identically 0. The important point is that the cross-validational observed relative surprise is a useful inferential tool for the statistician in checking the validity of the model. Finally we note that there are many other problems to which this approach can be applied that we have not discussed here; e.g. model choice, outlier detection, etc.

There is some recent work in the literature that has some relationship with what we are proposing in this section. Model checking based on the posterior predictive distribution of a statistic is discussed in Gelman, Meng and Stern (1995). The use of cross-validational predictive densities for model selection is discussed in Gelfand and Dey (1994).

4. Conclusions

There is still much work to be done on the application of the concept of relative surprise to problems of Bayesian inference. For example, there are some difficult computational issues that will need to be addressed as in general we require the values of densities. For many traditional problems, however, these densities can be evaluated in closed form. In such a context there are typically no special computational difficulties in the relative surprise approach. General approximation techniques are a current research problem.

This paper has shown that the relative surprise approach offers additional insight and flexibility to the Bayesian statistician. What might seem to be artificial restrictions in the choice of prior for hypothesis testing problems are removed. Further we have shown, via a number of examples, that the relative surprise approach leads to inferences that have some satisfying properties not possessed by other Bayesian techniques in contexts where prior beliefs are felt to be diffuse.

It is possible in a problem that there is no obvious function T to use for testing a hypothesis H_0 of interest. Of course we could just use the binary partition but this may require us to modify our prior and, in general, this is not satisfying to us. There is, however, an approach to generating a sensible T in many such problems that does not require the modification of the prior. This is discussed in Evans, Gilula and Guttman (1993) and is based on measuring the extent to which the posterior distribution concentrates around the hypothesis H_0 and comparing this with the concentration of the prior about this hypothesis. The addition of the observed relative surprise to this approach gives an objective way of comparing these concentrations.

We are not claiming here that inference via the observed relative surprise, or even Bayesian inference itself, is *the* way to do inference. We do claim, however, that the motivation provided in this paper and the examples demonstrate, that relative surprise is a valuable addition to the set of statistical inference tools.

ACKNOWLEDGEMENTS

The author was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

Aitkin, M. (1991), "Posterior Bayes factors" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, Vol. 53, No. 1, 111-142.

Bartlett, M.S.(1952), "The statistical significance of odd bits of information", *Biometrika*, Vol. 39, 228-237.

Berger, J. and Delampady, M.(1987), "Testing precise hypotheses" (with discussion), *Statistical Science*, Vol. 2, No. 3, 317-352.

Berger, J. and Selke, T.(1987), "Testing a point null hypothesis. The irreconcilability of P-values and evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112-139.

Box, G.E.P. (1980), "Sampling and Bayes' inference in scientific modelling and robustness", *Journal of the Royal Statistical Society*, Ser. A, 143, Part 4, 383-430.

Box, G.E.P. and Tiao, G.T. (1973), *Bayesian Inference in Statistical Analysis*, Reading: Addison-Wesley.

Casella, G. and Berger, R.L. (1987), "Reconciling Bayesian and frequentist evidence in the one-sided testing problem", *Journal of the American Statistical Association*, 82, 397, 106-111.

Dawid, A.P. (1984), "Present position and potential developments: some personal views, statistical theory, the prequential approach". *Journal of the Royal Statistical Society*, Ser. A, 147, 278-292.

Dawid, A.P., Stone, M. and Zidek, J.V. (1973), "Marginalization paradoxes in Bayesian and structural inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 2, 189-233.

Evans, M., Gilula, Z. and Guttman, I. (1993), "Computational issues in the Bayesian analysis of categorical data: loglinear and Goodman's RC model", *Statistica Sinica*, 3, 391-406.

Geisser, S. (1993), *Predictive Inference: An Introduction.* New York: Chapman and Hall.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society B*, 56, 398-409.

Gelman, A., Meng, X. and Stern, H. (1995). Posterior predictive assessment of model fitness via realized discrepancies. Manuscript.

Good, I.J.(1953), "The appropriate mathematical tools for describing and measuring uncertainty", in *Uncertainty and Business Decisions*, eds. C.F. Carter, G.P. Meredith and G.L.S. Shackle, Liverpool, U.K.: University Press.

Good, I.J.(1955), "A new finite series for Legendre polynomials", *Proceedings of the Cambridge Philosophical Society*, 51, 385-388.

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

Good, I.J.(1965), *The Estimation of Probabilities*, Cambridge: The M.I.T. Press.

Good, I.J.(1971), "The probabilistic explication of information, evidence, surprise, causality, explanation and utility", in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, 108-141, Toronto: Holt, Rinehart and Winston.

Good, I.J.(1981), "Some logic and history of hypothesis testing", in *Philosophical Foundations of Economics*, ed. J.C. Pratt, Dordrecht, Holland: Reidel.

Good, I.J.(1982a), "Comment on Shafer, Constructive Probability", *Journal of the American Statistical Association*, Vol. 77, No. 378, 342-344.

Good, I.J.(1982b), "Comment on Patil and Taillie, Diversity as a concept and its measurement", *Journal of the American Statistical Association*, Vol. 77, No. 379, 561-563.

Good, I.J.(1983a), *Good Thinking, The Foundations of Probability and Its Applications*, Minneapolis: U. of Minnesota Press.

Good, I.J.(1983b), "Antisurprise', *Journal of Statistical Computation and Simulation*, 69-71.

Good, I.J.(1985), "A new measure of surprise", *Journal of Statistical Computation and Simulation*, 21, 88-89.

Good, I.J.(1988), "Surprise index", in *Encyclopaedia of Statistical Sciences, Vol. 7*, eds. S. Kotz, N.L. Johnson and C.B. Reid, New York: John Wiley and Sons.

Good, I.J.(1989), "Surprise indices and p-values", *Journal of Statistical Computation and Simulation*, 32, 90-92.

Kass, R.E. and Raftery, A.E. (1994), "Bayes factors", *Journal of the American Statistical Association*, 90, 430, 773-795.

Kvalseth, T.O.(1987), "Stimulus probability, surprise and reaction time", *Proceedings of the Human Factors Society*, Vol. 1, 147-150.

Levi, I.(1972), "Potential surprise in the context of inquiry. Uncertainty and Expectations in Economics", in *Essays in Honour of G.L.S. Shackle*, eds. C.F Carter and J.L. Ford, Oxford, U.K.: Oxford Press.

Lindley, D.V. (1957), "A statistical paradox", *Biometrika*, 44, 187-192.

Perlman, M.D. amd Rasmussen, U.A. (1975). Some remarks on estimating a noncentrality parameter. *Communications in Statistics*, 4(5), 455-468.

Redheffer, R.M.(1951), "A note on the surprise index", *Annals of Mathematical Statistics*, Vol. 22, 128-130.

Rice, J.A. (1995), *Mathematical Statistics and Data Analysis. Second Edition*, Belmont: Duxbury Press.

Robert, C.P. (1993), "A note on Jeffreys-Lindley paradox", *Statistica Sinica*, 3, 601-608.

Saxena, K.M.L and Alam, K. (1982). Estimation of the non-centrality parameter of a Chi-squared distribution. *Annals of Statistics*, Vol. 10, No. 3, 1012-1016.

Shackle, G.L.S. (1949), *Expectation in Economics*, Cambridge, U.K.: Cambridge University Press.

Stein, C. (1959), "An example of wide discrepancy between fiducial and confidence intervals", *Annals of Mathematical Statistics*, 30, 877-880.

Weaver, W. (1948), "Probability, rarity, interest and surprise", *Scientific Monthly*, Vol. 67, 6, 390-392.

Weaver, W.(1963), *Lady Luck, The Theory of Probability*, Science Study Series, Garden City: Doubleday and Co. Inc.