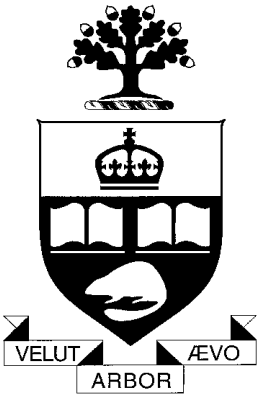

Optimality and Computations for Relative Surprise Inferences

by

M. Evans
University of Toronto

and

I. Guttmann
State University of New York

and

T. Swartz
Simon Fraser University

Technical Report No. 0508 September 1, 2005

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Optimality and Computations for Relative Surprise Inferences

M. Evans, I. Guttman and T. Swartz
U. of Toronto, State U. of New York, Simon Fraser U.

Abstract

Relative surprise inferences are based on how beliefs change from a priori to a posteriori. These inferences can be seen to be based on the posterior distribution of the integrated likelihood and, as such, are invariant under relabellings of the parameter of interest. In this paper we demonstrate that relative surprise inferences possess an optimality property. Further, computational techniques are developed for implementing these inferences that are applicable whenever we have algorithms to sample from the prior and posterior distributions.

1 Introduction

Suppose we observe data x_0 from a statistical model $\{f_\theta : \theta \in \Omega\}$, where f_θ is a density with respect to support measure μ on the sample space $\mathcal{X}$, and that we have a proper prior density π on θ, with respect to support measure υ on Ω. Consider a set $\mathcal{T}$ of possible values for some quantity of interest $\tau = \Upsilon(\theta)$ depending on the parameter of the model.

With these ingredients we have available the joint distribution of (x, θ), as given by the density $f_\theta(x)\,\pi(\theta)$ with respect to support measure $\mu \times \nu$, and the observed value x_0. A basic axiom of inference then says that probability statements about τ should be based on the conditional distribution of τ given the data x_0, otherwise known as the posterior distribution of τ and here denoted by the posterior density $\pi_\Upsilon(\cdot \mid x_0)$ with respect to some support measure $\upsilon_\mathcal{T}$ on $\mathcal{T}$. This is a particular application of conditional probability in a two-stage system where we observe the outcome from the second stage and want to make an inference about the concealed outcome from the first stage. This application is commonly referred to as Bayes theorem. We note, however, that the decision to use conditional probability is not compelled by a theorem, although there may be many theorems that suggest it is the appropriate thing to do, but rather it is an axiom that many agree is appropriate for inference in such a context.

This application of conditional probability could be viewed as the key step in what distinguishes a Bayesian analysis. The question remains, however, given the posterior distribution of τ, how should we use this to make inferences about the true unknown value of τ. It seems clear that Bayes theorem only says

that any probability statements we make about τ must be calculated using the posterior distribution. For example, if we wish to quote a set $C \subset \mathcal{T}$ that has a .95 probability of containing the true τ, then Bayes theorem does not tell us how to obtain the set C only that C must satisfy $\int_C \pi_\Upsilon(\tau \mid x_0)\, \upsilon_{\mathcal{T}}(d\tau) = .95$ and typically there are many such sets.

One possible response to this ambiguity is simply to say that the posterior distribution of τ is the outcome of the Bayesian inference process, namely, that one only needs to report this and it can be used as the user wishes. This seems somewhat inadequate, however, when we require a specific value of τ for further work, together with an assessment of the accuracy of this estimate, or wish to assess the plausibility of a hypothesized value τ_0 of τ as prescribed by some theory.

Another response is to say that the user needs to specify a set of inferences or actions A and a loss function L defined on $A \times \mathcal{T}$ and then choose the inference $a \in A$ that minimizes the posterior expected loss. While this has some appeal, in many applications it is difficult to specify L in a completely satisfactory way.

Considerations such as these have lead some to consider the possibility of stating a principle or axiom that would allow a statistician to determine the inferences by application of the principle to the ingredients of the problem, namely, the joint distribution of (x, θ) and the observed value x_0. To be Bayesian inferences, at least as we will use this terminology, these must only conform to the first principle, namely, that any probabilities quoted must be posterior probabilities.

So we ask what characteristics we would want such a principle to have beyond the restriction that the inferences produced be Bayesian. A general discussion of this problem is not our intent here, but we note that standard inferences about parameters could be taken to be estimation, together with the related problem of assessing the uncertainty in a quoted estimate, and testing problems. As such we might consider principles that lead, for any joint distribution, data and parameter of interest, to a class of regions $C_\gamma(x_0) \subset \mathcal{T}$ for $\gamma \in [0,1]$ with the property that

$$\int_{C_\gamma(x_0)} \pi_\Upsilon(\tau \mid x_0)\, \upsilon_{\mathcal{T}}(d\tau) \geq \gamma \tag{1}$$

and such that $C_{\gamma_1}(x_0) \subset C_{\gamma_2}(x_0)$ whenever $\gamma_1 \leq \gamma_2$. Typically we will require that there is an equality in (1) but in some cases this will not be possible. The nesting property seems quite natural as we could then take $\tau(x_0) \in C_0(x_0)$ as an estimate of τ with the full set of regions $C_\gamma(x_0)$, together with their posterior probabilities, serving as the quantification of uncertainty concerning the accuracy of $\tau(x_0)$. Further, assessing a hypothesized value τ_0 can be carried out by computing the Bayesian P-value $1 - \inf\{\gamma : \tau_0 \in C_\gamma(x_0)\}$.

There are many such principles with at least one in common use. We call this the *hpd principle* (highest posterior density principle) where a γ-hpd is given by $C_\gamma(x_0) = \left\{\tau : \pi_\Upsilon(\tau \mid x_0) \geq \pi_\Upsilon(\tau^*_\gamma \mid x_0)\right\}$ and τ^*_γ is determined so that (1) is satisfied. This leads to an estimate being a mode of $\pi_\Upsilon(\cdot \mid x_0)$. While this seems to lead to satisfactory inferences in many contexts, there is at least one concern

in the case when the posterior distribution of τ is continuous. For under a 1-1 reparameterization $\psi = \Psi(\tau)$ we do not have that hpd regions for τ transform to the corresponding hpd regions for ψ. We note that likelihood inferences about the full parameter θ satisfy the invariance property and it is generally required that likelihood-based inferences about marginal parameters should also possess this property.

We might look then for a general principle in the Bayesian context that possesses the invariance property. For example, when τ is real-valued we could consider taking $C_\gamma(x_0) = (\tau_{\alpha(1-\gamma)}, \tau_{1-(1-\alpha)(1-\gamma)})$ where τ_p is the p-th quantile of the posterior distribution of τ and α is fixed in $[0,1]$. We see that this prescribes an α-th quantile as an estimate of τ. Perhaps it is natural to take $\alpha = 1/2$ although this requires some justification as other choices are possible. We note, however, that this approach depends on τ being real-valued to be implemented and even then may seem unnatural when the posterior distribution is multimodal, as then we might like to discard other values of τ rather than just the tails. Also it is not at all clear how this should be generalized to situations where τ is not real-valued.

In Evans (1997) a principle of inference was stated, in the sense that a class of regions $C_\gamma(x_0)$ were specified with the appropriate properties, that generated Bayesian inferences possessing the invariance property. These inferences were referred to as relative surprise inferences as they were derived based on surprise as discussed in Good(1988, 1989). At least one implementation of Good's approach leads to hpd inferences which, as just noted, do note possess the invariance property and relative surprise inferences were devised as a method for recovering this.

We discuss relative surprise inferences in Section 2 and derive an optimality property for these inferences. In fact we show that relative surprise inferences are a particular example of what we call *hpd-like* inferences and in some ways are the most natural of these. Further we relate relative surprise inferences to likelihood inferences and the use of Bayes factors. We are not arguing here that the relative surprise principle is *the way* to determine inferences only that it is an intuitively reasonable way and the inferences possess some compelling properties.

Implementing relative surprise inferences requires that we be able to compute both the marginal prior and posterior densities for τ. In Section 3 we develop algorithms for implementing relative surprise inferences in situations where we may not have closed-form expressions for these quantities. The approach taken there shows that relative surprise inferences have some computational advantages over other hpd-like inferences as we can take advantage of the invariance to transform the parameter of interest to a compact parameter space. In Section 4 we consider some examples where these algorithms are used.

We note that the formulation of a Bayesian inference problem as stated here does not allow the prior π to be improper. This is because relative surprise inferences require the marginal prior for τ and, in general, an improper prior on θ does not marginalize in any obvious way to induce a prior on τ. For example, if $\theta = (\theta_1, \theta_2) \in R^2$ with $\tau = \theta_2$ and we put a flat prior on θ, then we cannot

simply integrate out θ_1 to get the marginal for τ. It seems highly desirable that whenever we start with a prior on the full parameter θ that we be able to examine what these prior beliefs say about the parameter of interest τ. Further, we cannot apply conditional probability as a principle when π is improper as the joint distribution is not a probability distribution so some other justification is needed for this step. Other objections have been raised about the use of improper priors, for example, Dawid, Stone and Zidek (1973).

It is the case, however, that relative surprise inferences under improper priors can be obtained as the limit of relative surprise inferences under a sequence of proper priors which in some sense converge to the improper prior. Under these circumstances we can apply conditional probability and can consider the effect of the prior π on τ at each step in the sequence. Examples of this approach can be found in Evans (1997). Of course, much needs to be said about what are appropriate sequences, as in Berger and Bernardo (1992), but this is not our concern here.

2 Relative Surprise Inferences

Consider now the inference problem presented in Section 1. The relative surprise principle makes use of the following preference ordering on the possible values of τ. We totally order the elements of $\mathcal{T}$ so that τ_1 is strictly preferred to τ_2 if the relative increase in belief for τ_1, from *a priori* to *a posteriori*, is greater than the corresponding increase for τ_2. We translate this mathematically into strictly preferring τ_1 to τ_2 whenever

$$\frac{\pi_\Upsilon(\tau_1 \,|\, x_0)}{\pi_\Upsilon(\tau_1)} > \frac{\pi_\Upsilon(\tau_2 \,|\, x_0)}{\pi_\Upsilon(\tau_2)} \tag{2}$$

where π_Υ is the marginal prior density of τ, defined with respect to the support measure $\upsilon_\mathcal{T}$ on $\mathcal{T}$. Note that we can take the support measure to be the prior measure Π_Υ, so requiring densities is not a restriction. Notice also that the preference ordering given by (2) is precisely that specified by the function $\pi_\Upsilon(\cdot \,|\, x_0)/\pi_\Upsilon(\cdot)$ and referred to as the integrated likelihood in Kalbfleisch and Sprott (1970). Berger, Liseo and Wolpert (1999) argue for the use of integrated likelihood rather than other forms, such as profile likelihood, when making inferences about marginal parameters.

We use the preference ordering given by (2) to determine inferences. Note that (2) is invariant under smooth transformations of τ, i.e., if τ_1 is preferred to τ_2, then $\psi_1 = \Psi(\tau_1)$ is preferred to $\psi_2 = \Psi(\tau_2)$ for any smooth Ψ, as the Jacobian factor cancels in both the ratios involved.

In an estimation context, where we are required to select a value from $\mathcal{T}$ as an estimate, this ordering 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 τ maximizing $\pi_\Upsilon(\cdot \,|\, x_0)/\pi_\Upsilon(\cdot)$. This estimator is computed by maximizing this ratio as a function of τ. We call such an estimate a *least relative surprise (LRSE)* estimate.

In hypothesis testing contexts we have an hypothesized true value $\tau_0 \in \mathcal{T}$ for $\Upsilon(\theta)$ and we 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 τ_0, from *a priori* to *a posteriori*, with this increase for each of the other possible values in $\mathcal{T}$. If the increase for τ_0 is small compared to the other increases, then the data suggests that τ_0 is surprising and we have evidence against the hypothesis. We use the posterior probability of obtaining a relative increase larger than that observed for τ_0 and refer to this as the *observed relative surprise (ORS)*. Therefore the observed relative surprise at τ_0 is given by

$$\Pi\left(\frac{\pi_\Upsilon(\tau \mid x_0)}{\pi_\Upsilon(\tau)} > \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\pi_\Upsilon(\tau_0)} \;\middle|\; x_0\right) \tag{3}$$

where $\Pi(\cdot \mid x_0)$ is the posterior probability measure. Notice that the value of τ_0 minimizing (3) is the LRSE as in this case the ORS is 0. It is the value most supported by the data, and so least surprising from the point of view of the data, when the relative change in degree of belief from *a priori* to *a posteriori* is our criterion for assessing this.

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 $\mathcal{T}$. A γ-relative surprise region for τ is given by

$$C_\gamma(x_0) = \left\{\tau_0 \in \mathcal{T} : \Pi\left(\frac{\pi_\Upsilon(\tau \mid x_0)}{\pi_\Upsilon(\tau)} > \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\pi_\Upsilon(\tau_0)} \;\middle|\; x_0\right) \leq \gamma\right\}. \tag{4}$$

This is the set of values in $\mathcal{T}$ whose observed relative surprise is no greater than γ. Note that we could have equivalently proceeded by first defining the regions $C_\gamma(x_0)$ as in (4) and then determining the estimates and Bayesian P-value as described in Section 1. The approach taken here proceeds from taking (3) as an appropriate measure of surprise for the hypothesized value τ_0.

While relative surprise inferences may seem unusual, they are determined in a very familiar way. In fact, noting that $\pi_\Upsilon(\cdot \mid x_0)/\pi_\Upsilon(\cdot)$ is a proper density function for the posterior distribution with the support measure taken to be the prior, we see that $C_\gamma(x_0)$ in (4) is a γ-hpd region using this density. Of course, hpd regions are usually calculated using densities taken with respect to volume measure and, in such circumstances, we know that a γ-hpd region has the smallest volume among all regions having posterior probability content equal to γ. Corollary 4 below establishes a similar optimality result for a γ-relative surprise region and as such can be seen as a key justification for the use of such inferences.

We prove a general version of this optimality result for inferences determined by the ratio $\pi_\Upsilon(\cdot \mid x_0)/\lambda(\cdot)$ where λ is a nonnegative function that determines a σ-finite measure $\Lambda(C) = \int_C \lambda(\tau)\, \upsilon_{\mathcal{T}}$ on $\mathcal{T}$. We refer to such inferences as *hpd-like*. We define the γ-credible regions $B_\gamma(x_0)$ generated from Λ by

$$B_\gamma(x_0) = \left\{\tau_0 \in \mathcal{T} : \Pi\left(\frac{\pi_\Upsilon(\tau \mid x_0)}{\lambda(\tau)} > \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\lambda(\tau_0)} \;\middle|\; x_0\right) \leq \gamma\right\}.$$

We have the following result where $\Pi_\Upsilon(\cdot \,|\, x_0)$ is the posterior induced by Υ.

Lemma 1. $\Pi_\Upsilon(B_\gamma(x_0) \,|\, x_0) \geq \gamma$ with equality whenever the posterior distribution of $\pi_\Upsilon(\cdot \,|\, x_0)/\lambda(\cdot)$ has no atoms.
Proof: See the Appendix.

The optimality of the regions $B_\gamma(x_0)$ is as stated in the next theorem.

Theorem 2. The set $B_\gamma(x_0)$ minimizes $\Lambda(C)$ among all measurable sets $C \subset \mathcal{T}$ satisfying $\Pi_\Upsilon(C \,|\, x_0) \geq \Pi_\Upsilon(B_\gamma(x_0) \,|\, x_0)$.
Proof: See the Appendix.

Combining Lemma 1 and Theorem 2 we have the following corollary.

Corollary 3. When the posterior distribution of $\pi_\Upsilon(\cdot \,|\, x_0)/\lambda(\cdot)$ has no atoms, then $B_\gamma(x_0)$ minimizes $\Lambda(C)$ among all measurable sets $C \subset \mathcal{T}$ satisfying $\Pi_\Upsilon(C \,|\, x_0) \geq \gamma$.

Taking $\lambda = \pi_\Upsilon$ gives the result for relative surprise regions.

Corollary 4. The γ-relative surprise region $C_\gamma(x_0)$ has smallest prior measure among all measurable sets $C \subset \mathcal{T}$ satisfying $\Pi_\Upsilon(C \,|\, x_0) \geq \Pi_\Upsilon(C_\gamma(x_0) \,|\, x_0)$ and $\Pi_\Upsilon(C_\gamma(x_0) \,|\, x_0) = \gamma$ whenever the posterior distribution of $\pi_\Upsilon(\cdot|\, x_0)/\lambda(\cdot)$ has no atoms.

Note that, in the case that $\mathcal{T}$ is Euclidean and taking Λ to be volume measure, Theorem 2 gives the optimality of hpd inferences.

One might ask why we should put $\Lambda = \Pi_\Upsilon$. Consider, however, the general context where $\mathcal{T}$ is an arbitrary set, not necessarily Euclidean, e.g., infinite dimensional. In such a situation it may not be at all clear what to use as an appropriate measure Λ on $\mathcal{T}$. Also, suppose $\mathcal{T}$ is Euclidean and we take Λ to be volume measure. If $\psi = \Psi(\tau)$ where $\Psi : \mathcal{T} \to \mathcal{T}$ is a 1-1, onto, smooth transformation then do we use the measure Λ when determining inferences for ψ or do we use $\Lambda \circ \Psi^{-1}$? If minimizing volume is important then we must use Λ and we lose the invariance property. In general there does not seem to be a good reason to use $\Lambda \circ \Psi^{-1}$ so that the invariance property is retained. On the other hand if we take Λ to be the prior Π_Υ on τ then the appropriate prior on ψ is $\Pi_\Upsilon \circ \Psi^{-1}$. This is the essential reason why relative surprise inferences are invariant. In many ways the prior is the most natural choice of the measure Λ used to determine hpd-like inferences as it transforms appropriately under reparameterizations.

We can also interpret Corollary 4 in terms of hypothesis assessment. For suppose we choose to reject $H_0 : \tau = \tau_0$ whenever the ORS is greater than γ. This is equivalent to rejecting H_0 whenever $\tau_0 \in C_\gamma^c(x_0)$. This test has an optimal property when we consider other rejection regions $D \subset \mathcal{T}$.

Corollary 5. Among rejection regions $D \subset \mathcal{T}$ of $H_0 : \tau = \tau_0$ that satisfy $\Pi_\Upsilon(D \,|\, x_0) \leq \Pi_\Upsilon\left(C_\gamma^c(x_0) \,|\, x_0\right)$ the relative surprise test region $C_\gamma^c(x_0)$ maximizes $\Pi_\Upsilon(D)$.

Note that this result does not say that the region $C_\gamma^c(x_0)$ maximizes the prior probability of rejecting H_0, only that the region $C_\gamma^c(x_0)$ is the largest possible test region with posterior probability content no greater than $1 - \gamma$, where the size of a set is measured using the prior probability measure.

When $\tau = \theta$ the prior cancels in $\pi_\Upsilon(\cdot \mid x_0)/\pi_\Upsilon(\cdot)$ and the norming constant for the posterior cancels on both sides of the inequality in (3). Therefore $C_\gamma(x_0) = \{\theta_0 : \Pi(f(x_0 \mid \theta) > f(x_0 \mid \theta_0)) \mid x_0) \leq \gamma\}$ and relative surprise regions are likelihood regions. Corollary 4 then takes the following form.

Corollary 6. The likelihood region $C_\gamma(x_0)$ for θ minimizes $\Pi(C)$ among all measurable sets $C \subset \Omega$ satisfying $\Pi(C \mid x_0) \geq \Pi(C_\gamma(x_0) \mid x_0)$.

This result seems surprising because there is no connection between the prior and the likelihood. We note, however, that the link here is given by the choice of the posterior probability content of the set.

In Evans and Zou (2002) results were developed that show an increased robustness for (3) to the choice of prior when compared to computing the Bayesian P-value based on hpd regions. Certainly the fact that relative surprise regions for θ correspond to likelihood regions points to such a result. Similarly Wasserman (1989) develops robustness results for likelihood regions for θ showing that the posterior content of such regions are relatively insensitive to contaminations of the prior.

The following establishes a link between Bayes factors and relative surprise.

Example 1. *Bayes factors and the ORS*

Suppose we wish to assess the hypothesis that $\theta \in H_0 \subset \Omega$ where $\Pi(H_0) > 0$. So here we have that $\tau = \Upsilon(\theta) = H_0$ when $\theta \in H_0, \tau = \Upsilon(\theta) = H_0^c$ when $\theta \notin H_0$ and $\tau_0 = H_0$. Then (3) is equal to

$$\Pi\left(\frac{\pi_\Upsilon(\Upsilon(\theta) \mid x)}{\pi_\Upsilon(\Upsilon(\theta))} > \frac{\Pi(H_0 \mid x_0)}{\Pi(H_0)} \;\middle|\; x_0\right) = \begin{cases} 0 & \text{if} \quad BF_{H_0} \geq 1 \\ \Pi(H_0^c \mid x_0) & \text{if} \quad BF_{H_0} < 1 \end{cases} \tag{5}$$

where

$$BF_{H_0} = \frac{\Pi(H_0 \mid x_0)}{1 - \Pi(H_0 \mid x_0)} \Big/ \frac{\Pi(H_0)}{1 - \Pi(H_0)} \tag{6}$$

is the Bayes factor in favor of H_0. This is close to the usual recommendation of saying that we have evidence against H_0 when $\Pi(H_0^c \mid x)$ is small.

Sometimes just the Bayes factor is recommended in hypothesis testing with small values of (6) treated as evidence against H_0. Various calibrations have been suggested for the value of a Bayes factor but perhaps the most direct method is to compute the posterior probability of obtaining a value larger than (6) with large values of this probability indicating that H_0 is surprising. If we do this, we see that $\Pi\left(BF_{\Upsilon(\theta)} > BF_{H_0} \mid x_0\right)$ agrees exactly with the ORS. So in a sense we can think of the ORS as a generalization of the Bayes factor when we choose to calibrate the value of a Bayes factor using the tail probability.

Now suppose we have a sequence H_0^n where $\Pi(H_0^n) > 0$ for all n, and the H_0^n converge nicely to H_0 with $\Pi(H_0) = 0$. Typically such a structure will arise when we have a continuous parameter τ, we want to assess $H_0 = \{\tau_0\}$ and the H_0^n are shrinking neighborhoods of τ_0. Then it is easy to show that $BF_{H_0^n} \to \pi_\Upsilon(\tau_0 \mid x)/\pi_\Upsilon(\tau_0)$ as $n \to \infty$. So it is very natural to think of the ORS as a generalization of the Bayes factor approach. The ORS is comparing the limiting Bayes factor in favor of τ_0 with the limiting Bayes factors in favor of each of the other possible values of τ. ∎

One might compare relative surprise inferences to the hpd inferences. As we increase the amount of data the posterior density will, under suitable conditions, concentrate at the true value of the parameter. If the region where the posterior concentrates is small enough then the prior will typically look flat there and so, using Theorem 2, the γ-hdp regions and γ-relative surprise regions will be very similar. Notice, however, that this is not independent of the parameterization. In other words in some parameterizations the regions may be very similar but in other parameterizations, with the same data, the regions will be quite different. A simple example illustrates this.

Example 2. *Bernoulli*

Suppose we have a sample of n from a Bernoulli(θ) with $\theta \sim \text{Uniform}(0,1)$. Then it is easy to see that the LRSE and the posterior mode (the estimates obtained via relative surprise and hpd respectively) are both equal to x/n where x equals the number of 1's in the sample. Now consider the reparameterization $\psi = \theta^p$ where $p \leq 1$. The LRSE of ψ is $(x/n)^p$ while the posterior mode of ψ is $((x+1-p)/(n+1-p))^p$. When n is large enough then indeed the estimates of ψ will be similar, but how large n has to be to achieve a difference in the estimates of a given size, depends on p (and x). A comparison of the estimates when $x = 0$ is particularly informative in this regard. For example, when $n = 5, x = 0, p = .1$ then the LRSE is 0 and the posterior mode estimate of ψ is .828 and when $n = 100$ the posterior mode is .624! ∎

The departure between relative surprise inferences and hpd inferences can be much more radical as documented in Evans (1997). For example, suppose we observe $x = (x_1, \ldots, x_n)$ where $x_i \sim N(\theta_i, 1)$, with $x_1, \ldots, x_n$ statistically independent and $\theta_1, \ldots, \theta_n$ having independent $N(0, \sigma_0^2)$ prior distributions with σ_0^2 large and known. Suppose further that our interest is in estimating $\tau^2 = \sum_{i=1}^n \theta_i^2$. In this example the LRSE and the posterior mode differ substantially with the LRSE exhibiting much better repeated sampling behavior. This, and other examples, are more thoroughly discussed in Evans (1997).

3 Computations

Implementing relative surprise inferences requires that we be able to evaluate both $\pi_\Upsilon(\tau)$ and $\pi_\Upsilon(\tau \mid x_0)$ at many values of τ to obtain $\pi_\Upsilon(\cdot \mid x_0)/\pi_\Upsilon(\cdot)$. When we do not have closed-form expressions for these functions we must use numerical techniques.

We note that this problem become more difficult as the dimension of τ rises unless the problem is such that closed-form expressions can be obtained. Fortunately, however, the parameter of interest τ is typically 1-dimensional. Accordingly, we develop a computational approach suitable for such problems.

While we require τ to be 1-dimensional, no restriction is placed on the dimensionality of the full parameter θ. The only requirement is that we have sampling algorithms for both the prior and posterior distributions of θ so that we have the almost sure convergence of relative frequencies for intervals for τ. Obviously this will hold, via the strong law of large numbers, whenever we have algorithms

for generating i.i.d. samples from these distributions. More generally we can use MCMC algorithms and use the associated ergodic theorems to justify our results.

Given that inferences for τ are invariant under smooth reparameterization, we will assume in this section that we have transformed so that $\tau \in [0,1]$. Our computational methods can then be developed for this case and relative surprise inferences for the original parameter of interest obtained by applying the inverse of the transformation. Note that this is a distinct computational advantage for relative surprise over hpd inferences. For the computations for hpd inferences must be carried out in the original parameterization and it is much more difficult to deal with unbounded $\mathcal{T}$.

We denote the prior distribution function of τ by F_Υ and the posterior distribution function by $F_\Upsilon(\cdot \mid x_0)$. We use a sampling algorithm to generate a sample (exact or approximate) $\tau_1, \ldots, \tau_{N_1}$ from the prior distribution of τ and construct the estimate $\hat{F}_\Upsilon(\tau) = N_1^{-1} \sum_{i=1}^{N_1} I_{(-\infty,\tau]}(\tau_i)$. Similarly, we generate a sample (exact or approximate) $\tau_1^*, \ldots, \tau_{N_2}^*$ from the posterior distribution of τ and construct the estimate $\hat{F}_\Upsilon(\tau \mid x_0) = N_2^{-1} \sum_{i=1}^{N_2} I_{(-\infty,\tau]}(\tau_i^*)$.

Now choose a grid of values $0 = \hat{\tau}_1 < \cdots < \hat{\tau}_{N_3} = 1$. Then, when $F_\Upsilon(\hat{\tau}_{i+1}) - F_\Upsilon(\hat{\tau}_i) > 0$, we have that,

$$\frac{\hat{F}_\Upsilon(\hat{\tau}_{i+1} \mid x_0) - \hat{F}_\Upsilon(\hat{\tau}_i \mid x_0)}{\hat{F}_\Upsilon(\hat{\tau}_{i+1}) - \hat{F}_\Upsilon(\hat{\tau}_i)} \to \frac{F_\Upsilon(\hat{\tau}_{i+1} \mid x_0) - F_\Upsilon(\hat{\tau}_i \mid x_0)}{F_\Upsilon(\hat{\tau}_{i+1}) - F_\Upsilon(\hat{\tau}_i)} \tag{7}$$

almost surely as $N_1, N_2 \to \infty$. Now put

$$R(\tau_0) = \left\{ \hat{\tau}_{j+1} : \frac{F_\Upsilon(\hat{\tau}_{j+1} \mid x_0) - F_\Upsilon(\hat{\tau}_j \mid x_0)}{F_\Upsilon(\hat{\tau}_{j+1}) - F_\Upsilon(\hat{\tau}_j)} > \frac{F_\Upsilon(\hat{\tau}_{i+1} \mid x_0) - F_\Upsilon(\hat{\tau}_i \mid x_0)}{F_\Upsilon(\hat{\tau}_{i+1}) - F_\Upsilon(\hat{\tau}_i)} \right\}$$

where $i = i(N_3)$ is such that $\tau_0 \in (\hat{\tau}_i, \hat{\tau}_{i+1}]$. We have the following result.

Theorem 7. Suppose that $\pi_\Upsilon(\tau_0 \mid x_0)/\pi_\Upsilon(\tau_0)$ is a continuity point of the posterior distribution of $\pi_\Upsilon(\tau \mid x_0)/\pi_\Upsilon(\tau)$ and every open interval about $\pi_\Upsilon(\tau_0 \mid x_0)/\pi_\Upsilon(\tau_0)$ has positive posterior probability. If the grids $0 = \hat{\tau}_1 < \cdots < \hat{\tau}_{N_3} = 1$ are chosen so that $\sup\{\hat{\tau}_{i+1} - \hat{\tau}_i : i = 1, \ldots, N_3\} \to 0$ as $N_3 \to \infty$, then

$$\sum_{\hat{\tau}_{j+1} \in R(\tau_0)} (F_\Upsilon(\hat{\tau}_{j+1} \mid x_0) - F_\Upsilon(\hat{\tau}_j \mid x_0)) \tag{8}$$

converges to (3) as $N_3 \to \infty$.

Proof: See the Appendix.

Theorem 7 and (7) suggest that we approximate (3) by

$$\sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0)} \left(\hat{F}_\Upsilon(\hat{\tau}_{j+1} \mid x_0) - \hat{F}_\Upsilon(\hat{\tau}_j \mid x_0) \right) \tag{9}$$

where

$$\hat{R}(\tau_0) = \left\{ \hat{\tau}_{j+1} : \frac{\hat{F}_\Upsilon(\hat{\tau}_{j+1} \mid x_0) - \hat{F}_\Upsilon(\hat{\tau}_j \mid x_0)}{\hat{F}_\Upsilon(\hat{\tau}_{j+1}) - \hat{F}_\Upsilon(\hat{\tau}_j)} > \frac{\hat{F}_\Upsilon(\hat{\tau}_{i+1} \mid x_0) - \hat{F}_\Upsilon(\hat{\tau}_i \mid x_0)}{\hat{F}_\Upsilon(\hat{\tau}_{i+1}) - \hat{F}_\Upsilon(\hat{\tau}_i)} \right\}.$$

We have the following result.

Theorem 8. Suppose that $\epsilon > 0$ is given. Then, under the hypotheses of Theorem 7, there exist N_1, N_2, and N_3 so that, for all larger values of these quantities, (9) differs from (8) by at most ϵ almost surely.
Proof: See the Appendix.

Therefore, to approximate the ORS we choose the grid of points $\hat{\tau}_1 < \cdots < \hat{\tau}_{N_3}$, generate the samples from the prior and posterior, and finally compute (9). We can use the values $(\hat{F}_\Upsilon(\hat{\tau}_{i+1} \mid x_0) - \hat{F}_\Upsilon(\hat{\tau}_i \mid x_0))/(\hat{F}_\Upsilon(\hat{\tau}_{i+1}) - \hat{F}_\Upsilon(\hat{\tau}_i))$ to compute the value $\hat{\tau} = \hat{\tau}_{i+1}$ that maximizes this ratio, as an approximation to the LRSE. Once we have obtained the approximate value of the ORS at each value $\hat{\tau}_{i+1}$ we can use this to calculate an approximate γ-relative surprise interval as in (4).

4 Examples

Many examples of relative surprise inferences in contexts where we have expressions for the prior and posterior densities can be found in Evans (1997). We now consider several examples where we need the computational approach discussed in Section 3.

Example 3. *Stress-strength reliability*

This example is concerned with making inferences about the probability $\tau = P(Y_2 > Y_1)$, where Y_1 and Y_2 are independent random variables. Here the Y's are measurements of a variable that measures the strength of a system and the different Y's correspond to different conditions under which this measurement is taken. The parameter τ is called the *stress-strength reliability as it* corresponds to the probability that the system has greater strength under condition 2 than under condition 1. Inferences for τ are discussed in Birnbaum (1956), Simonoff, Hochberg and Reiser (1986), Guttman, Johnson, Bhattacharayya and Reiser (1988), Reiser and Guttman (1989), and Guttman and Papandonatos (1997).

In Guttman and Papandonatos (1997) it was assumed that the statistician has available the independent observations $y_1 \sim N_{n_1}\left(X_1\beta_1, \sigma_1^2 I_{n_1 \times n_1}\right)$ and $y_2 \sim N_{n_2}\left(X_2\beta_2, \sigma_2^2 I_{n_2 \times n_2}\right)$, where $X_1 \in R^{n_1 \times p_1}, X_2 \in R^{n_2 \times p_2}$ are both of full rank, and Jeffreys prior was placed on the parameters $\beta_1 \in R^{p_1}, \beta_2 \in R^{p_2}, \sigma_1^2 > 0, \sigma_2^2 > 0$. The posterior distribution of τ was approximated when $Y_1 \sim N(v_1^t\beta_1, \sigma_1^2)$ and $Y_2 \sim N(v_2^t\beta_2, \sigma_2^2)$ are future observations at potentially new values of the covariates. We assume this structure here so $\theta = \left(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2\right)$.

We note that it is not at all clear what Jeffreys prior on θ implies about the prior distribution of τ. This is because Jeffreys prior is improper, and as such, a marginal prior for τ does not exist, at least as a probability distribution. Also, as noted in Section 1 we need to have a proper prior in order to implement relative surprise inferences although we can still derive limiting relative surprise inferences as a sequence of proper priors converges to an improper prior.

For illustration purposes we assume the following conjugate prior structure

$$\begin{aligned} \beta_1 \,|\, \beta_2, \sigma_1^2, \sigma_2^2 &\sim N_{p_1}(\beta_{10}, \sigma_1^2 \Lambda_1), \quad \beta_2 \,|\, \sigma_1^2, \sigma_2^2 \sim N_{p_2}(\beta_{20}, \sigma_2^2 \Lambda_2), \\ \sigma_1^{-2} \,|\, \sigma_2^2 &\sim \text{Gamma}(\alpha_1, \eta_1), \quad \sigma_2^{-2} \sim \text{Gamma}(\alpha_2, \eta_2), \end{aligned} \tag{12}$$

where $\beta_{10} \in R^{p_1}, \beta_{20} \in R^{p_2}$, $\Lambda_1 \in R^{p_1 \times p_1}, \Lambda_2 \in R^{p_2 \times p_2}, \alpha_1 > 0, \eta_1 > 0, \alpha_2 > 0$, $\eta_2 > 0$ are fixed hyperparameters and the Gamma(α, η) density is given by $\eta^{-\alpha} \Gamma^{-1}(\alpha)\, x^{\alpha-1} \exp(-x/\eta)$ for $x > 0$.

For the marginal prior distribution of τ we note that $\tau = P_\theta(Y_2 - Y_1 > 0)$ and, writing $Y_i = v_i^t \beta_i + z_i$ with $z_i \sim N(0, \sigma_i^2)$, we have that $Y_2 - Y_1 = (v_2^t \beta_2 - v_1^t \beta_1) + (z_2 - z_1) \sim N(v_2^t \beta_2 - v_1^t \beta_1, \sigma_1^2 + \sigma_2^2)$. Therefore we can write

$$\tau = P_\theta(Y_2 - Y_1 > 0) = \Phi(\delta). \tag{13}$$

where $\delta = (v_2^t \beta_2 - v_1^t \beta_1) / \left(\sigma_1^2 + \sigma_2^2\right)^{1/2}$. The prior distribution of τ is then obtained from (13) using (12). For relative surprise inferences we need to obtain the prior density of τ and this would appear to be a non-trivial computation as the prior distribution of δ is non-standard. It is easily shown, however, that $\delta \,|\, \sigma_1^2, \sigma_2^2 \sim N(\mu_\delta(\sigma_1^2, \sigma_2^2), \sigma_\delta^2(\sigma_1^2, \sigma_2^2))$ where

$$\mu_\delta(\sigma_1^2, \sigma_2^2) = \frac{v_2^t \beta_{20} - v_1^t \beta_{10}}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \quad \sigma_\delta^2(\sigma_1^2, \sigma_2^2) = \frac{\sigma_1^2 v_1^t \Lambda_1 v_1 + \sigma_2^2 v_2^t \Lambda_2 v_2}{\sigma_1^2 + \sigma_2^2}$$

and this helps to simplify the simulation.

The posterior distribution of θ is like (12) but with different choices of the parameters. We have provided this in the Appendix. Again the posterior distribution of δ is nonstandard and this prevents the computation of the posterior density of τ in a simple closed form.

We note that in this problem there is some structure that allows us to improve on the computational approach discussed in Section 3. Notice that if $\left(\sigma_{11}^2, \sigma_{21}^2\right), \ldots, \left(\sigma_{1N_1}^2, \sigma_{2N_1}^2\right)$ is a sample from the prior distribution Π, then $\check{F}_{\Upsilon N_1}(\tau_0) = \sum_{i=1}^{N_1} \Pi\left(\tau \le \tau_0 \,|\, \sigma_{1i}^2, \sigma_{2i}^2\right) = \sum_{i=1}^{N_1} \Pi\left(\delta \le \Phi^{-1}(\tau_0) \,|\, \sigma_{1i}^2, \sigma_{2i}^2\right) = \sum_{i=1}^{N_1} \Phi((\Phi^{-1}(\tau_0) - \mu_\delta\left(\sigma_{1i}^2, \sigma_{2i}^2\right))/\sigma_\delta\left(\sigma_{1i}^2, \sigma_{2i}^2\right)) \to F_\Upsilon(\tau_0)$ almost surely as $N_1 \to \infty$. Accordingly we can substitute $\check{F}_{\Upsilon N_1}(\tau_0)$ for $\hat{F}_\Upsilon(\hat{\tau}_j)$ and a similar result holds for the posterior distribution function. It is easy to see that Theorems 7 and 8 also apply to this approximation. We refer to this hereafter as the Rao-Blackwellized approach.

We now implement our approach using simulated data with $p_1 = p_2 = 2$,

$$\beta_{11} = 1, \quad \beta_{12} = 0, \quad \sigma_1 = 1, \quad \beta_{21} = 2, \quad \beta_{22} = 1, \quad \sigma_2 = 1,$$

and $X_1 = X_2 \in R^{20 \times 2}$ with the first column entries all equal to 1 and second column equal to $(1, 2, \ldots, 20)$. We generated $n_1 = 20$ values $z_{11}, \ldots, z_{1n_1}$ from the $N(0, 1)$ distribution, putting $y_{1i} = 1 + z_{1i}$ and generated $n_2 = 20$ values $z_{21}, \ldots, z_{2n_2}$ from the $N(0, 1)$ distribution, putting $y_{2i} = 2 + i + z_{2i}$. Then we

let $v_1^t = v_2^t = (1,1)$ so that the exact value is $\tau = \Phi((3-1)/\sqrt{2}) = \Phi\left(\sqrt{2}\right) = 0.921348$. We selected a diffuse prior given by

$$\begin{aligned} \beta_{110} &= \beta_{120} = 0, \quad \Lambda_1 = \mathrm{diag}\,(2,2)\,, \quad \beta_{210} = \beta_{220} = 0, \quad \Lambda_2 = \mathrm{diag}\,(2,2)\,, \\ \alpha_1 &= 2, \quad \eta_1 = 1, \quad \alpha_{22} = 2, \quad \eta_2 = 1. \end{aligned}$$

In Figure 1 the solid line is the exact value of $\pi_\Upsilon(\cdot \mid x_0)/\lambda(\cdot)$. Based on $N_1 = N_2 = 5 \times 10^3$ (taking 34 seconds of computing time on a Sun workstation) the Rao-Blackwellized estimate is also plotted on this graph and, to the accuracy of the plotting, coincides with the exact value. The direct algorithm with $N_1 = N_2 = 10^5$ (taking 10 seconds of computing time), and where we have smoothed the prior and posterior cdf estimates by averaging a point with 2 points on both sides, is plotted in the dashed curve and we see that this also provides a reasonable approximation. With the same smoothing and $N_1 = N_2 = 5 \times 10^5$ (taking 3 minutes of computing time), this estimate becomes somewhat smoother but we see, in any case, that the methods of Section 3 provide acceptable approximations.

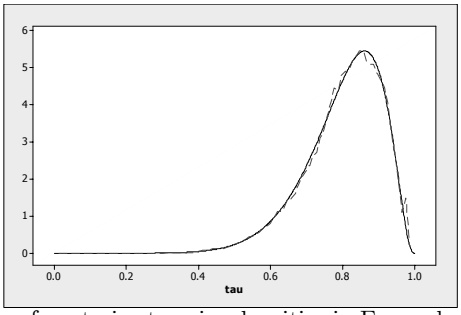

Figure 1: Plot of ratio of posterior to prior densities in Example 3 where the solid line denotes the exact ratio and the Rao-Blackwell estimate, and the dashed line denotes the brute force estimate based on $N_1 = N_2 = 10^5$.

For this data, the LRSE of τ is given by 0.861 and the .95-relative surprise interval for τ is given by $(0.598, 0.972)$. This compares with a posterior mode of 0.897 and a .95-HPD interval of $(0.625, 0.984)$. We note that these inferences are very similar and both intervals contain the true value. The HPD interval is always shortest but Corollary 4 tells us that the relative surprise interval is the smallest with respect to the prior and is also invariant.

The following data was analyzed in Guttman and Papandonatos (1997) and gives the results of measuring shear strength of spot welds for two different gauges of steel.

$\mathbf{y}_1$	350	380	385	450	465	185	535	555	590	605
$\mathbf{x}_1$	380	155	160	165	175	165	195	185	195	210
$\mathbf{y}_2$	680	800	780	885	875	1025	1100	1030	1175	1300
$\mathbf{x}_2$	190	200	209	215	215	215	230	250	265	250

The normal simple linear regression model is used for both Y_1 and Y_2, i.e. $p_1 = p_2 = 2$, measuring the *shear strength* of the two types of steel respectively,

and where X_1, X_2 are the same predictor, namely, *weld diameter*. Suppose now we want to compute τ when $X_1 = X_2 = 200$. For the prior we put

$$\begin{aligned} \beta_{110} &= \beta_{120} = 0, \quad \Lambda_1 = \text{diag}\,(2,2)\,, \quad \beta_{210} = \beta_{220} = 0, \quad \Lambda_2 = \text{diag}\,(2,2)\,, \\ \alpha_1 &= 10^{-1}, \quad \eta_1 = 10^{-1}, \quad \alpha_2 = 10^{-1}, \quad \eta_2 = 10^{-1}. \end{aligned}$$

In Figure 2 we have plotted the prior and posterior densities for τ. We note the concentration of this prior about 0 and 1 and the high concentration of the posterior near 1. This prior is not only dependent on the choice of prior for the basic parameters, but also is dependent on the value of the predictors $X_1 = X_2 = 200$. For this prior the LRSE is 0.994 and the .95-relative surprise interval is $(0.936, 1.000)$ while the posterior mode is 0.999 and the .95-HPD interval is $(0.946, 1.000)$.

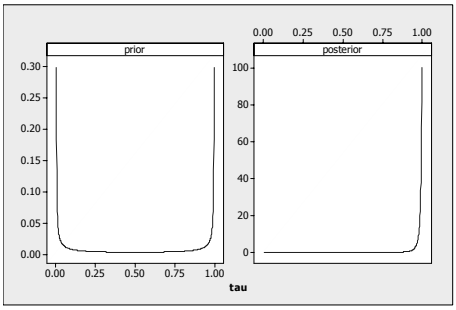

Figure 2: Plot of prior and posterior distribution of τ in Example 2.

The more diffuse we made the prior on the full parameter the more the marginal prior on τ concentrated about 0 and 1. Actually a fairly wide class of priors on τ is available based on how we choose the hyperparameters. For example, it is easy to see that the choices

$$\begin{aligned} \beta_{110} &= \beta_{120} = 0, \quad \Lambda_1 = \text{diag}\left(2.5 \times 10^{-5}, 2.5 \times 10^{-5}\right), \\ \beta_{210} &= \beta_{220} = 0, \quad \Lambda_2 = \text{diag}\left(2.5 \times 10^{-5}, 2.5 \times 10^{-5}\right), \\ \alpha_1 &= 1, \quad \eta_1 = 10^{-4}, \quad \alpha_2 = 1, \quad \eta_2 = 10^{-4}, \end{aligned}$$

produce a uniform prior distribution for τ, for which the hpd and relative surprise inferences are identical. For this choice the LRSE is 0.935 and the .95-relative surprise interval is $(0.779, 0.989)$. The point estimates are close to what the previous prior gave but we note the much wider intervals. We are not advocating necessarily the uniform prior for τ. Our point here is simply that we should look at the implications of any prior assignment for the full parameter on the parameter of interest. We note that this necessitates using proper priors or, when improper priors are to be used, considering a sequence of proper prior distributions that converge to the improper prior. Ultimately the correct method for the selection of prior distributions is through elicitation and this has nothing to do with the methods used to make inferences about model components. For further discussion on this point see O'Hagan (2005). ∎

In the following example we do not have exact algorithms for generating from the prior and posterior but must use MCMC methods. See, for example, Evans and Swartz (2000) for a discussion of these algorithms. The point of this example is to show that success in implementing the computational methods of Section 3 is largely dependent on good sampling algorithms for the full model parameter and, when this is the case, the dimensionality of the problem is not an issue.

Example 4. *Variance components*

In Gelfand, Hills, Racine-Poon and Smith (1990) a variance component model was considered of the form $Y_{ij} = \theta_i + e_{ij}$ where $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The sampling distribution is given by $e_{ij} \,|\, \sigma_e^2 \sim N(0, \sigma_e^2)$ independent of $\theta_i \,|\, \mu, \sigma_\theta^2 \sim N(\mu, \sigma_\theta^2)$. The prior structure is given by $\mu \sim N(\mu_0, \sigma_0^2)$, $\sigma_\theta^{-2} \sim \text{Gamma}(a_1, b_1)$, $\sigma_e^{-2} \sim \text{Gamma}(a_2, b_2)$ where μ, σ_θ^2 and σ_e^2 are mutually independent. Suppose our interest is in making inference about the intraclass correlation coefficient given by $\tau = \sigma_\theta^2 / \left(\sigma_\theta^2 + \sigma_e^2\right)$.

In this situation it is simple to simulate directly from the prior distribution of τ by generating σ_θ^2 and σ_e^2. The posterior distribution, however, requires that we integrate over $(\theta_1, \ldots, \theta_I), \mu, \sigma_\theta^2$ and σ_e^2, which has dimension $I + 3$, and we cannot generate sample directly from the posterior. Gelfand, Hills, Racine-Poon and Smith (1990)record the full conditional distributions of these parameters, and so we can implement a Gibbs sampler for this problem.

To assess the viability of the method recorded in Section 3 we generated data y_{ij} with $I = 30, J = 5$ with $\mu = 0, \sigma_\theta^2 = 2$ and $\sigma_e^2 = 1$ so the true value of the parameter of interest is $\tau = 2/3$. For the prior we used the values similar to those used in Gelfand et. al. (1990), namely, $\mu_0 = 0, \sigma_0^2 = 10^{12}, a_1 = 0.5, b_1 = 1.0$, and $a_2 = 0.1, b_2 = 0.1$. Then, setting $N_1 = N_2 = 2 \times 10^4$ and $N_3 = 5 \times 10^2$ (taking 40 seconds of computing time) and smoothing the prior and posterior density estimates, we obtained the plot of $\pi_\Upsilon(\cdot \,|\, x_0)/\pi_\Upsilon(\cdot)$ as in Figure 3. The graph obtained when $N_1 = N_2 = 2 \times 10^5$ is almost coincident with although somewhat smoother.

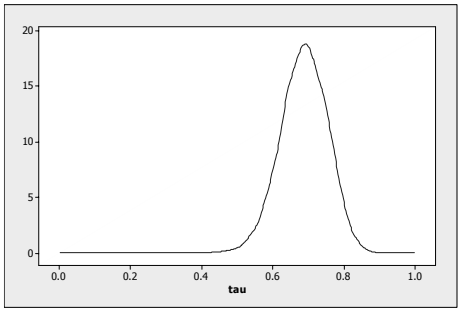

Figure 3: Plot of ratio of posterior to prior densities in Example 4.

The LRSE of τ is given by 0.706 while the posterior mode is given by 0.708. A .95-relative surprise interval for τ is given by $(.558, .813)$ while a .95-hpd interval is given by $(.567, .822)$ so the inferences are very similar and both intervals contain the true value. ∎

5 Conclusions

We have shown that relative surprise inferences arise as optimal inferences when the optimality criterion is the minimization of the prior content of a credible region. They are particular examples of hpd-like inferences that arise when we determine credible regions by the minimization of some measure of size of the region. In many ways, the prior is the most natural measure to use for this. In particular, this choice leads to inferences that are invariant over all possible smooth reparameterizations and this is an important property for inferences to possess. Relative surprise inferences are also strongly related to likelihood inferences and to the use of Bayes factors. Relative surprise inferences can be seen to be based on how the data changes beliefs from *a priori* to *a posteriori* and this emphasis on the data seems very natural in many applications. Further implementing relative surprise inferences has some computational advantages as we are completely free to choose the parameterization and this is not the case for hpd inferences. This enables us to transform so that the parameter of interest ranges over a compact interval in this situation we get the convergence results provided by Theorems 7 and 8.

We feel that one should choose inferences based on a principle or axiom. Choosing inferences on an ad hoc or arbitrary basis doesn't seem satisfactory, as we need a consistent scientific rationale for the selections made to provide support that they are sensible. The relative surprise principle is just one possibility and, as with any axiom we cannot ultimately claim that it is *the* way to do this. We do feel, however, that the properties relative surprise inferences possess, and their relationship with likelihood inferences, provide strong support for their use in Bayesian inference problems.

Another aspect of this paper has been the computations necessary to implement relative surprise inferences. This requires the calculation of both the prior and posterior densities of the parameter of interest. The computational strategies developed here have been shown to be feasible in several problems where the parameter of interest is 1-dimensional and we do not have closed form expressions for the marginal prior and posterior density. The dimensionality of the model does not affect the feasibility of our approach provided we have good sampling algorithms (exact or approximate) for the prior and posterior.

While many parameters of interest are 1-dimensional, e.g., probabilities of subsets of the sample space, this is certainly not always the case. The obvious analog of the methods described here will work generally for fairly low dimensional quantities of interest, but clearly other methods will have to be developed when these are high-dimensional. Happily this seems like a fairly rare occurrence. We also note that the algorithms we have developed can be used in problems where the integrated likelihood for a parameter of interest needs to be calculated without using the associated relative surprise inferences.

Acknowledgements

The authors would like to thank the Associate Editor and two referees whose comments led to substantial improvements in the paper.

6 Appendix

Proof of Lemma 1

Let G denote the posterior cdf of $\pi_{\Upsilon}(\cdot \,|\, x_0)/\lambda(\cdot)$. Then $\tau_0 \in B_\gamma(x_0)$ if and only if $G\left(\pi_{\Upsilon}(\tau_0 \,|\, x_0)/\lambda(\tau_0)\right) \geq 1-\gamma$ which holds if and only if $\pi_{\Upsilon}(\tau_0 \,|\, x_0)/\lambda(\tau_0)$ $\geq G^{-1}\left(1-\gamma\right) = \inf\left\{r : 1-\gamma \leq G(r)\right\}$ so $\Pi_{\Upsilon}\left(B_\gamma(x_0) \,|\, x_0\right) = 1-G(G^{-1}\left(1-\gamma\right))$ $+ \left(G(G^{-1}\left(1-\gamma\right)) - G(G^{-1}\left(1-\gamma\right)-0)\right) \geq \gamma$. Note we now have the simpler definition $B_\gamma(x_0) = \left\{\tau : \pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) \geq c_\gamma\right\}$ where $c_\gamma = G^{-1}\left(1-\gamma\right)$. ∎

Proof of Theorem 2

The proof of this result is very similar to the proof of the fundamental lemma of hypothesis testing as in Lehmann (1986). Let $C \subset \mathcal{T}$ be such that $\Pi_{\Upsilon}\left(C \,|\, x_0\right) \geq \Pi_{\Upsilon}\left(B_\gamma(x_0) \,|\, x_0\right)$. Put $\mathcal{T}_0 = \left\{\tau : I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) = 0\right\}$, $\mathcal{T}_1 = \left\{\tau : I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) < 0\right\}$, $\mathcal{T}_2 = \left\{\tau : I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) > 0\right\}$ and note that $\left\{\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2\right\}$ is a partition of $\mathcal{T}$. We have that

$$\mathcal{T}_1 = \left\{\tau : I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) < 0, \pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) \leq c_\gamma\right\}$$

because $\pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) > c_\gamma$ implies $I_{B_\gamma(x_0)}\left(\tau\right) = 1$, which implies $I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) = 1 - I_C\left(\tau\right) \geq 0$ a contradiction to the definition of $\mathcal{T}_1$. Also

$$\mathcal{T}_2 = \left\{\tau : I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) > 0, \pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) \geq c_\gamma\right\}$$

because $\pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) < c_\gamma$ implies $I_{B_\gamma(x_0)}\left(\tau\right) = 0$, which implies $I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) = -I_C\left(\tau\right) \leq 0$.

We have that,

$$\begin{aligned}
&\Lambda\left(B_\gamma(x_0)\right) - \Lambda\left(C\right) = \int \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right) \\
= &\int_{\mathcal{T}_1} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right) + \int_{\mathcal{T}_2} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right).
\end{aligned}$$

Now note that $\Lambda\left(d\tau\right) = \lambda\left(\tau\right) \upsilon_{\mathcal{T}}\left(d\tau\right)$, $\Pi_{\Upsilon}(d\tau \,|\, x_0) = \pi_{\Upsilon}(\tau \,|\, x_0)\, \upsilon_{\mathcal{T}}\left(d\tau\right)$ and because $I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) < 0$ and $\pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) \leq c_\gamma$ when $\tau \in \mathcal{T}_1$ then

$$\begin{aligned}
\int_{\mathcal{T}_1} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right) &\leq c_\gamma^{-1} \int_{\mathcal{T}_1} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Pi_{\Upsilon}(d\tau \,|\, x_0) \\
&= c_\gamma^{-1} E_{\Pi(\cdot \,|\, x_0)}\left(I_{\mathcal{T}_1}\left(I_{B_\gamma(x_0)} - I_C\right)\right).
\end{aligned}$$

Similarly, because $I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right) > 0$ and $\pi_{\Upsilon}(\tau \,|\, x_0)/\lambda(\tau) \geq c_\gamma$ when $\tau \in \mathcal{T}_2$, we have that

$$\begin{aligned}
\int_{\mathcal{T}_2} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right) &\leq c_\gamma^{-1} \int_{\mathcal{T}_2} \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Pi_{\Upsilon}(d\tau \,|\, x_0) \\
&= c_\gamma^{-1} E_{\Pi(\cdot \,|\, x_0)}\left(I_{\mathcal{T}_2}\left(I_{B_\gamma(x_0)} - I_C\right)\right).
\end{aligned}$$

Therefore we have that $\Lambda\left(B_\gamma(x_0)\right) - \Lambda\left(C\right) = \int \left(I_{B_\gamma(x_0)}\left(\tau\right) - I_C\left(\tau\right)\right) \Lambda\left(d\tau\right)$ $\leq c_\gamma^{-1} E_{\Pi(\cdot \,|\, x_0)}\left(\left(I_{B_\gamma(x_0)} - I_C\right)\right) = c_\gamma^{-1}\left(\Pi\left(B_\gamma(x_0) \,|\, x_0\right) - \Pi\left(C \,|\, x_0\right)\right) \leq 0$ because we have assumed that $\Pi\left(C \,|\, x_0\right) \geq \Pi\left(B_\gamma(x_0) \,|\, x_0\right)$. We conclude that $\Lambda\left(B_\gamma(x_0)\right) \leq \Lambda\left(C\right)$. ∎

Proof of Theorem 7

Let $F = F_\Upsilon, G = F_\Upsilon(\cdot \mid x_0)$, define the random variable $X = X(\tau) = \pi_\Upsilon(\tau \mid x_0)/\pi_\Upsilon(\tau)$ and, for the grid $0 = \hat{\tau}_1 < \cdots < \hat{\tau}_{N_3} = 1$, define the random variable $X_{N_3} = X_{N_3}(\tau) = (G(\hat{\tau}_{j+1}) - G(\hat{\tau}_j)) / (F(\hat{\tau}_{j+1}) - F(\hat{\tau}_j))$ whenever $\tau \in (\hat{\tau}_j, \hat{\tau}_{j+1}]$. Note that X_{N_3} is not defined when $F(\hat{\tau}_{j+1}) = F(\hat{\tau}_j)$ but we can ignore this because this implies $G(\hat{\tau}_{j+1}) = G(\hat{\tau}_j)$ (the posterior is absolutely continuous with respect to the prior) and a similar comment applies to the definition of X when $\pi_\Upsilon(\tau) = 0$.

Now suppose that we choose the grids so that $\sup\{\hat{\tau}_{j+1} - \hat{\tau}_j : j = 1, \ldots, N_3\} \to 0$ as $N_3 \to \infty$. Then we have that $X_{N_3} \to X$ as $N_3 \to \infty$ almost surely with respect to the posterior $\Pi(\cdot \mid x_0)$. This implies that X_{N_3} converges in distribution to X as $N_3 \to \infty$.

If $i = i(N_3)$ is such that $\tau_0 \in (\hat{\tau}_i, \hat{\tau}_{i+1}]$, then,as $N_3 \to \infty$,

$$\nu_{N_3} = \frac{G(\hat{\tau}_{i+1}) - G(\hat{\tau}_i)}{F(\hat{\tau}_{i+1}) - F(\hat{\tau}_i)} \to \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\pi_\Upsilon(\tau_0)} = \nu_0.$$

Therefore, if $\eta > 0$, then $\nu_{N_3} \in (\nu_0 - \eta, \nu_0 + \eta)$ for all N_3 large enough and

$$|\Pi(X_{N_3} \leq \nu_{N_3} \mid x_0) - \Pi(X \leq \nu_0 \mid x_0)| \leq \max\left\{ \left| \begin{array}{c} \Pi(X_{N_3} \leq \nu_0 \pm \eta \mid x_0) - \\ \Pi(X \leq \nu_0 \mid x_0) \end{array} \right| \right\}.$$

We can choose η so that $\nu - \eta$ and $\nu + \eta$ are continuity points for the distribution of X and so the right-hand side of the above inequality converges to $\max\{|\Pi(X \leq \nu_0 \pm \eta \mid x_0) - \Pi(X \leq \nu_0 \mid x_0)|\}$. Since ν_0 is a continuity point of the distribution of X, this can be made as small as we like by an appropriate choice of η and so $\limsup_{N_3 \to \infty} |\Pi(X_{N_3} \leq \nu_{N_3} \mid x_0) - \Pi(X \leq \nu_0 \mid x_0)|$ can be made as small as we like. This completes the proof. ∎

Proof of Theorem 8

We use some of the notation and results from the proof of Theorem 7 and in addition let $\hat{F} = \hat{F}_\Upsilon, \hat{G} = \hat{F}_\Upsilon(\cdot \mid x_0)$. Since ν_0 is a continuity point for the distribution of X we can choose $\eta > 0$ so that $\nu_0 - \delta$ and $\nu_0 + \delta$ are continuity points for X and $\Pi(X \in (\nu_0 - \eta, \nu_0 + \eta] \mid x_0) = \Pi(X \leq \nu_0 + \eta \mid x_0) - \Pi(X \leq \nu_0 - \eta \mid x_0)$ is smaller than $\epsilon/4$. Since $\nu_{N_3} \to \nu_0$ we have that $\nu_{N_3} \in (\nu_0 - \eta, \nu_0 + \eta]$ for all N_3 large enough. Then, because X_{N_3} converges in distribution to X, we have that $\Pi(X_{N_3} \in (\nu_0 - \delta, \nu_0 + \delta] \mid x_0) = \Pi(X_{N_3} \leq \nu_0 + \eta \mid x_0) - \Pi(X_{N_3} \leq \nu_0 - \eta \mid x_0) < \epsilon/2$ for all N_3 large enough. Putting

$$S(\tau_0) = \left\{ \hat{\tau}_{j+1} : \frac{G(\hat{\tau}_{j+1}) - G(\hat{\tau}_j)}{F(\hat{\tau}_{j+1}) - F(\hat{\tau}_j)} = \nu_{N_3} \right\},$$

then we have that $\sum_{\hat{\tau}_{j+1} \in S(\tau_0)} (G(\hat{\tau}_{j+1}) - G(\hat{\tau}_j)) = \Pi(X_{N_3} = \nu_{N_3} \mid x_0) \leq \Pi(X_{N_3} \in (\nu_0 - \delta, \nu_0 + \delta] \mid x_0) < \epsilon/2$ for all N_3 large enough.

If $\hat{\tau}_{j+1} \in R(\tau_0)$, then

$$\frac{G(\hat{\tau}_{j+1}) - G(\hat{\tau}_j)}{F(\hat{\tau}_{j+1}) - F(\hat{\tau}_j)} > \frac{G(\hat{\tau}_{i+1}) - G(\hat{\tau}_i)}{F(\hat{\tau}_{i+1}) - F(\hat{\tau}_i)}.$$

Therefore, since

$$\frac{\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)}{\hat{F}\left(\hat{\tau}_{j+1}\right)-\hat{F}\left(\hat{\tau}_{j}\right)} \rightarrow \frac{G\left(\hat{\tau}_{j+1}\right)-G\left(\hat{\tau}_{j}\right)}{F\left(\hat{\tau}_{j+1}\right)-F\left(\hat{\tau}_{j}\right)}$$

almost surely as $N_1, N_2 \rightarrow \infty$, we have that $\hat{\tau}_{j+1} \in \hat{R}\left(\tau_0\right)$ for all N_1, N_2 large enough. Similarly, if $\hat{\tau}_{j+1}$ is such that

$$\frac{G\left(\hat{\tau}_{j+1}\right)-G\left(\hat{\tau}_{j}\right)}{F\left(\hat{\tau}_{j+1}\right)-F\left(\hat{\tau}_{j}\right)} < \frac{G\left(\hat{\tau}_{i+1}\right)-G\left(\hat{\tau}_{i}\right)}{F\left(\hat{\tau}_{i+1}\right)-F\left(\hat{\tau}_{i}\right)}$$

then $\hat{\tau}_{j+1} \notin \hat{R}\left(\tau_0\right)$ for all N_1, N_2 large enough. Then, because there are only finitely many values $\hat{\tau}_{j+1}$, for all N_1, N_2 large enough,

$$\begin{aligned}\sum_{\hat{\tau}_{j+1} \in \hat{R}\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right) &= \sum_{\hat{\tau}_{j+1} \in R\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right)+ \\ &\quad \sum_{\hat{\tau}_{j+1} \in \hat{R}\left(\tau_0\right) \cap S\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right).\end{aligned}$$

Since there are only finitely many terms in these sums we have that

$$\sum_{\hat{\tau}_{j+1} \in R\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right) \rightarrow \sum_{\hat{\tau}_{j+1} \in R\left(\tau_0\right)}\left(G\left(\hat{\tau}_{j+1}\right)-G\left(\hat{\tau}_{j}\right)\right)$$

and

$$\begin{aligned}\sum_{\hat{\tau}_{j+1} \in \hat{R}\left(\tau_0\right) \cap S\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right) &\leq \sum_{\hat{\tau}_{j+1} \in S\left(\tau_0\right)}\left(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right)\right) \\ &\rightarrow \sum_{\hat{\tau}_{j+1} \in S\left(\tau_0\right)}\left(G\left(\hat{\tau}_{j+1}\right)-G\left(\hat{\tau}_{j}\right)\right)\end{aligned}$$

almost surely as $N_1, N_2 \rightarrow \infty$. Then for all N_1, N_2 large enough we have that (9) is within $\epsilon/4$ of (8) and $\sum_{\hat{\tau}_{j+1} \in S\left(\tau_0\right)}(\hat{G}\left(\hat{\tau}_{j+1}\right)-\hat{G}\left(\hat{\tau}_{j}\right))$ is within $\epsilon/4$ of $\sum_{\hat{\tau}_{j+1} \in S\left(\tau_0\right)}\left(G\left(\hat{\tau}_{j+1}\right)-G\left(\hat{\tau}_{j}\right)\right)$. This completes the proof. ∎

Posterior Distribution in Example 3

Some straightforward calculation gives the posterior distribution of θ as

$$\begin{aligned}\beta_1 \,|\, \beta_2, \sigma_1^2, \sigma_2^2, y_1, y_2 &\sim N_{p_1}(\beta_{10}\left(y_1, y_2\right), \sigma_1^2\left(X_1^t X_1+\Lambda_1^{-1}\right)^{-1}), \\ \beta_2 \,|\, \sigma_1^2, \sigma_2^2, y_1, y_2 &\sim N_{p_2}(\beta_{20}\left(y_1, y_2\right), \sigma_2^2\left(X_2^t X_2+\Lambda_2^{-1}\right)^{-1}), \\ \sigma_1^{-2} \,|\, \sigma_2^2, y_1, y_2 &\sim \operatorname{Gamma}\left(\frac{n_1+p_1}{2}+\alpha_1, \eta_1\left(y_1, y_2\right)\right), \\ \sigma_2^{-2} \,|\, y_1, y_2 &\sim \operatorname{Gamma}\left(\frac{n_2+p_2}{2}+\alpha_2, \eta_2\left(y_1, y_2\right)\right),\end{aligned}$$

where

$$\begin{aligned}
\beta_{10}(y_1,y_2) &= \left(X_1^tX_1+\Lambda_1^{-1}\right)^{-1}\left(X_1^ty_1+\Lambda_1^{-1}\beta_{10}\right), \\
\beta_{20}(y_1,y_2) &= \left(X_2^tX_2+\Lambda_2^{-1}\right)^{-1}\left(X_2^ty_2+\Lambda_2^{-1}\beta_{20}\right) \\
\eta_1(y_1,y_2) &= \left(\frac{1}{\eta_1}+\frac{1}{2}\left(\begin{array}{c} y_1^ty_1+\beta_{10}^t\Lambda_1^{-1}\beta_{10}- \\ \left(X_1^ty_1+\Lambda_1^{-1}\beta_{10}\right)^t\left(X_1^tX_1+\Lambda_1^{-1}\right)^{-1} \\ \left(X_1^ty_1+\Lambda_1^{-1}\beta_{10}\right)\end{array}\right)\right)^{-1} \\
\eta_2(y_1,y_2) &= \left(\frac{1}{\eta_2}+\frac{1}{2}\left(\begin{array}{c} y_2^ty_2+\beta_{20}^t\Lambda_2^{-1}\beta_{20}- \\ \left(X_2^ty_2+\Lambda_2^{-1}\beta_{20}\right)^t\left(X_2^tX_2+\Lambda_2^{-1}\right)^{-1} \\ \left(X_1^ty_1+\Lambda_1^{-1}\beta_{10}\right)\end{array}\right)\right)^{-1}.
\end{aligned}$$

Further, $\delta \,|\, \sigma_1^2,\sigma_2^2,y_1,y_2 \sim N\left(\mu_\delta(\sigma_1,\sigma_2,y_1,y_2),\sigma_\delta^2(\sigma_1,\sigma_2,y_1,y_2)\right)$ where

$$\begin{aligned}
\mu_\delta\left(\sigma_1^2,\sigma_2^2,y_1,y_2\right) &= \frac{v_2^t\beta_{20}(y_1,y_2)-v_1^t\beta_{10}(y_1,y_2)}{\sqrt{\sigma_1^2+\sigma_2^2}} \\
\sigma_\delta^2\left(\sigma_1^2,\sigma_2^2,y_1,y_2\right) &= \frac{\sigma_1^2v_1^t\left(X_1^tX_1+\Lambda_1^{-1}\right)^{-1}v_1+\sigma_2^2v_2^t\left(X_2^tX_2+\Lambda_2^{-1}\right)^{-1}v_2}{\sigma_1^2+\sigma_2^2}.
\end{aligned}$$

7 References

Berger, J.O. and Bernardo, J.M. (1992) On the development of the reference prior method. Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, Clarendon Press, Oxford, U.K., 35-60.

Berger, J.O., Liseo, B. and Wolpert, R.L. (1999) Integrated likelihood methods for eliminating nuisance parameters. Statistical Science, Vol. 14, No.1, 1-28.

Birnbaum, Z.W. (1956) On the use of the Mann-Whitney statistic. Proceedings of the Third Symposium on Mathematical Statistics and Probability, Volume 1, U. of California Press, Berkeley, 13-17.

Dawid, A.P., Stone, M. and Zidek, J.V. (1973) Marginalization paradoxes in Bayesian and structural inference. Journal of the Royal Statistical Society B, 35, 189-233).

Evans, M. (1997) Bayesian inference procedures derived via the concept of relative surprise. Communications in Statistics - Theory And Methods, Vol. 26, No. 5, 1125-1143.

Evans, M., and Swartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press.

Evans, M., and Zou, T. (2002) Robustness of relative surprise inferences to choice of prior. Recent Advances in Statistical Methods, Proceedings of Statistics 2001 Canada: The 4th Conference in Applied Statistics Montreal, Canada 6 - 8 July 2001, Yogendra P. Chaubey (ed.), 90-115, Imperial College Press.

Gelfand A.E., Hills S.E., Racine-Poon, A. and Smith, A.F.M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association, Vol. 85, No. 412, 972-985.

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.

Guttman, I., Johnson, R.A., Bhattacharayya, G.K., and Reiser, B. (1988) Confidence limits for stress-strength models with explanatory variables. Technometrics, 30, 161-168.

Guttman, I., and Papandonatos, G.D. (1997) A Bayesian approach to a reliability problem: theory, analysis and interesting numerics. Canadian Journal of Statistics, Vol. 25, No. 2, 143-158.

Kalbfleisch, J.D. and Sprott, D.A. (1970) Application of likelihood methods to models involving large numbers of parameters. Journal of the Royal Statistical Society, Series B, 32, 175-208.

Lehmann, E.L. (1986) Testing Statistical Hypotheses, Second Edition. John Wiley and Sons, New York.

O'Hagan, A. (2005). Research in elicitation. Research Report No. 557/05, Department of Probability and Statistics, University of Sheffield to appear in Bayesian Statistics and its Applications.

Reiser, B., and Guttman, I. (1989) Sample size choice for reliability verification on stress-strength models. Canadian Journal of Statistics, Vol. 17, 253-259.

Simonoff, J.S., Hochberg, Y., and Reiser, B. (1986) Alternative estimation procedures for $P(X < Y)$ in categorical data. Biometrics, 42, 895-907.

Wasserman, L.A. (1989) A robust Bayesian interpretation of likelihood regions. The Annals of Statistics, Vol. 17, Np. 3, 1387-1393.