

**Relative Surprise Inferences and Computations  
For a Reliability Problem**

by

**Michael Evans  
Department of Statistics  
University of Toronto**

and

**Irwin Guttman  
Department of Statistics  
State University of New York at Buffalo**

and

**Tim Swartz  
Department of Mathematics  
and Statistics  
Simon Fraser University**

**Technical Report No. 0408 October 5, 2004**

**TECHNICAL REPORT SERIES**

**University of Toronto**

**Department of Statistics**



# Relative Surprise Inferences and Computations for a Reliability Problem

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

## Abstract

Relative surprise inferences are derived for a quantity of interest in reliability problems. Such inferences are generally appropriate when the goal is to de-emphasize the role of the prior and we want inferences invariant under relabellings of the parameter of interest. Computational techniques, that are appropriate for a much wider class of problems, are developed for implementing these inferences.

## 1 Introduction

We are concerned with making inferences about the probability

$$\tau = P(Y_2 > Y_1) \quad (1)$$

where  $Y_1$  and  $Y_2$  are independent random variables. Interest in such a quantity arises in reliability problems. 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. For example, the response variables may correspond to a measure of the strength of a metal and we may wish to compare the strength of this metal under different metallurgical conditions. Alternatively the  $Y$ 's may correspond to log-lifetimes of a processed food item manufactured under different conditions. Clearly (1), referred to as the *stress-strength reliability*, corresponds to the probability that the system has greater strength under condition 2 than under condition 1.

Suppose we have samples  $y_1 = (y_{11}, \dots, y_{1n_1})'$  and  $y_2 = (y_{21}, \dots, y_{2n_2})'$  from the distributions of the variables under the two conditions, respectively. A simple estimate of (1) is then given by

$$\hat{\tau} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{(0, \infty)}(y_{2j} - y_{1i}) \quad (2)$$

where  $I_{(0, \infty)}$  is the indicator function for  $(0, \infty)$ . Note that (2) is the Mann-Whitney test statistic and this fact can be used to derive further inferences, see Birnbaum (1956).

Of some interest, however, is to derive inferences for (1) when assumptions are made about the distributions of the  $Y$  variables and when covariates are taken into account. Several authors have considered this problem in both the frequency and Bayesian contexts. For example, see 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  $\mathbf{y}_1 \in R^{n_1}, \mathbf{y}_2 \in R^{n_2}$  with

$$\mathbf{y}_1 \sim N_{n_1}(X_1\beta_1, \sigma_1^2 I_{n_1 \times n_1}) \text{ and } \mathbf{y}_2 \sim N_{n_2}(X_2\beta_2, \sigma_2^2 I_{n_2 \times n_2}), \quad (3)$$

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

$$\pi(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2) \propto 1/(\sigma_1^2 \sigma_2^2)$$

was placed on the parameters  $\beta_1 \in R^{p_1}, \beta_2 \in R^{p_2}, \sigma_1^2 > 0, \sigma_2^2 > 0$ . Under these assumptions, the posterior distribution of  $\tau$  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. A sampling algorithm was derived under these assumptions, so that values could be generated from the posterior distribution of  $\tau$ , and then various characteristics of this distribution estimated.

In this paper we take a somewhat different Bayesian approach. First we note that it is not at all clear what the *a priori* assumptions about the fundamental parameter of the model, namely  $(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ , imply about  $\tau$ . This is because Jeffreys prior is improper, and as such, a marginal prior for  $\tau$  does not exist, at least as a probability distribution. One consequence of this is that it is not obvious how the data have changed our beliefs concerning the true value of  $\tau$ . As just noted, the use of Jeffreys prior does not even allow us to clearly state what our *a priori* beliefs about this quantity. Obviously we do not have an equal weighting on each of the possible values  $\tau \in [0, 1]$  (see the discussion in Example 2 for more on this). It seems reasonable to ask, in any inference problem, what the prior assumptions imply about a quantity of interest. This deficiency is a common characteristic of inferences when improper priors are used. Of course we could reparametrize, so that  $\tau$  is part of the basic parameter, and then choose a prior that reflected our prior beliefs about  $\tau$ , but if this prior is improper then the same criticism would apply to some other marginal parameter of interest.

A simple resolution of this problem, when we deem it necessary to look at the implications of *a priori* beliefs, is to restrict attention to proper priors and look at the marginal prior distribution of the parameter of interest. We will do this here using the prior structure

$$\begin{aligned} \beta_1 | \beta_2, \sigma_1^2, \sigma_2^2 &\sim N_{p_1}(\beta_{10}, \sigma_1^2 \Lambda_1), \\ \beta_2 | \sigma_1^2, \sigma_2^2 &\sim N_{p_2}(\beta_{20}, \sigma_2^2 \Lambda_2), \\ \sigma_1^{-2} | \sigma_2^2 &\sim \text{Gamma}(\alpha_1, \eta_1), \\ \sigma_2^{-2} &\sim \text{Gamma}(\alpha_2, \eta_2), \end{aligned} \quad (4)$$

where  $\beta_{10} \in R^{p_1}, \beta_{20} \in R^{p_2}, \Lambda_1 \in R^{p_1 \times p_1}, \Lambda_2 \in R^{p_2 \times p_2}, \alpha_1 \geq 0, \eta_1 > 0, \alpha_2 \geq 0, \eta_2 > 0$  are fixed hyperparameters and the Gamma( $\alpha, \eta$ ) density is given by

$$\frac{\eta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x/\eta)$$

for  $x > 0$ . Actually many other choices can be made for the prior for the approach discussed here, provided that we can obtain sampling algorithms for both the prior and posterior. We have chosen to use (4) because these sampling algorithms are relatively straight-forward and our focus is on other aspects of the problem.

For posterior inference, based on (4), we could follow Guttman and Papadonatos (1997), as Jeffreys prior was used there simply for convenience. We choose to proceed somewhat differently here, however, as inferences for  $\tau$  are based on how beliefs are changed from *a priori* to *a posteriori*. The proposed procedures are examples of relative surprise inferences as introduced in Evans (1997). Essentially these are based on how the data changes beliefs rather than basing the inferences only on *a posteriori* beliefs as in traditional Bayesian inferences.

The relative surprise approach has several advantages in certain circumstances. First, as discussed in Evans (1997), the inferences are invariant under smooth, 1-1 transformations of the parameter of interest and this is not the case for traditional Bayesian inferences such as posterior means, modes or highest posterior density regions. In fact, the inference approach we adopt here could be viewed as a general technique for obtaining invariant Bayesian inferences. Second, as discussed in Evans and Zou (2002) relative surprise inferences have superior robustness properties, with respect to the choice of the prior, when compared to purely posterior inferences. In essence, relative surprise inferences have advantages when strong prior opinion is not available and so we want to minimize its effect. Of course, there will be situations where we strongly believe in an informative prior but this would not appear to be the case in many applications.

Implementing relative surprise inferences in this problem requires the development of an appropriate computational approach because of the lack of closed form expressions for the marginal prior and marginal posterior densities for  $\tau$ . As such, the methods developed here are applicable to a much broader class of problems than the one discussed in this paper.

In Section 2 we discuss the application of relative surprise inferences to this problem. In Section 3 we develop an appropriate computational approach and in Section 4 we apply this methodology to some examples.

## 2 Relative Surprise Inferences

Suppose that we observe data  $x_0$  from a statistical model  $\{f_\theta : \theta \in \Omega\}$  and that we have a proper prior measure  $\Pi$  on the parameter  $\theta$  of the model. Consider a set  $\mathcal{T}$  of possible values for some quantity of interest  $\tau = \Upsilon(\theta)$  depending on the parameter of the model. We totally order the elements of  $\mathcal{T}$  as follows:  $\tau_1$

is strictly preferred to  $\tau_2$  if the relative increase in belief for  $\tau_1$ , from *a priori* to *a posteriori*, is greater than the corresponding increase for  $\tau_2$ . We translate this mathematically into strictly preferring  $\tau_1$  to  $\tau_2$  whenever

$$\frac{\pi_{\Upsilon}(\tau_1 | x_0)}{\pi_{\Upsilon}(\tau_1)} > \frac{\pi_{\Upsilon}(\tau_2 | x_0)}{\pi_{\Upsilon}(\tau_2)} \quad (5)$$

where  $\pi_{\Upsilon}$  is the marginal prior density of  $\tau$  defined with respect to some support measure on  $\mathcal{T}$  and  $\pi_{\Upsilon}(\cdot | x_0)$  is the marginal posterior density of  $\tau$  with respect to the same support measure. Note that we can always take the support measure to be the prior measure  $\Pi$ , so requiring densities is not a restriction here. We use the preference ordering given by (5) to determine inferences. Note that (5) is invariant under smooth transformations of  $\tau$ , i.e., if  $\tau_1$  is preferred to  $\tau_2$ , then  $\psi_1 = \Psi(\tau_1)$  is preferred to  $\psi_2 = \Psi(\tau_2)$  for any smooth  $\Psi$ , 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 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  $\tau$  maximizing  $\pi_{\Upsilon}(\tau | x_0) / \pi_{\Upsilon}(\tau)$ . This estimator is computed by maximizing this ratio as a function of  $\tau$ . 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  $\tau_0$ , from *a priori* to *a posteriori*, with this increase for each of the other possible values in  $\mathcal{T}$ . If the increase for  $\tau_0$  is small compared to the other increases then the data suggests that  $\tau_0$  is surprising and we have evidence against the hypothesis. We use the posterior probability of obtaining a relative increase larger than that observed for  $\tau_0$  and refer to this as the *observed relative surprise (ORS)*. Therefore the observed relative surprise at  $\tau_0$  is given by

$$\Pi \left( \frac{\pi_{\Upsilon}(\tau | x_0)}{\pi_{\Upsilon}(\tau)} > \frac{\pi_{\Upsilon}(\tau_0 | x_0)}{\pi_{\Upsilon}(\tau_0)} \mid x_0 \right) \quad (6)$$

where  $\Pi(\cdot | x_0)$  is the posterior probability measure. Notice that the value of  $\tau_0$  minimizing (6) is the least relative surprise estimate 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  $\mathcal{T}$ . An  $\alpha$ -relative surprise region for  $\tau$  is given by

$$C_{\alpha}(x_0) = \{\tau_0 \in \mathcal{T} \mid \Pi \left( \frac{\pi_{\Upsilon}(\tau | x_0)}{\pi_{\Upsilon}(\tau)} > \frac{\pi_{\Upsilon}(\tau_0 | x_0)}{\pi_{\Upsilon}(\tau_0)} \mid x_0 \right) \leq \alpha\}. \quad (7)$$

This is the set of values in  $\mathcal{T}$  whose observed relative surprise is no greater than  $\alpha$ .

In the context of the reliability problem discussed in Section 1, we have that the model for the data  $x_0 = (y_1, y_2)$  is given by (3). The parameter is  $\theta = (\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$  with the prior probability measure  $\Pi$  as specified in (4). The quantity of interest is then

$$\tau = P_\theta(Y_2 > Y_1)$$

where  $Y_1 \sim N(v_1^t \beta_1, \sigma_1^2)$  independent of  $Y_2 \sim N(v_2^t \beta_2, \sigma_2^2)$ .

For the marginal prior distribution of  $\tau$  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

$$\tau = P_\theta(Y_2 - Y_1 > 0) = \Phi\left(\frac{v_2^t \beta_2 - v_1^t \beta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \quad (8)$$

Then the prior distribution of  $\tau$  is obtained from (8) using (4). We note that to implement the relative surprise inference methods we need to obtain the prior density of  $\tau$  and this would appear to be a non-trivial computation as the prior distribution of

$$\delta = \frac{v_2^t \beta_2 - v_1^t \beta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (9)$$

is non-standard. It is easily shown that the conditional prior distribution of  $\delta$  given  $(\sigma_1^2, \sigma_2^2)$  is

$$\delta | \sigma_1^2, \sigma_2^2 \sim N(\mu_\delta(\sigma_1^2, \sigma_2^2), \sigma_\delta^2(\sigma_1^2, \sigma_2^2)) \quad (10)$$

where

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

and this simplifies the simulation. We discuss this further in Section 3.

For the posterior distribution of  $\tau$  we can use (8) and the posterior distribution of  $(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ . From the conjugacy of (4) (see the Appendix) we have that

$$\begin{aligned} \beta_1 | \beta_2, \sigma_1^2, \sigma_2^2, y_1, y_2 &\sim N_{p_1}(\beta_{10}(y_1, y_2), \sigma_1^2 (X_1^t X_1 + \Lambda_1^{-1})^{-1}), \\ \beta_2 | \sigma_1^2, \sigma_2^2, y_1, y_2 &\sim N_{p_2}(\beta_{20}(y_1, y_2), \sigma_2^2 (X_2^t X_2 + \Lambda_2^{-1})^{-1}), \\ \sigma_1^{-2} | \sigma_2^2, y_1, y_2 &\sim \text{Gamma}\left(\frac{n_1 + p_1}{2} + \alpha_1, \eta_1(y_1, y_2)\right), \\ \sigma_2^{-2} | y_1, y_2 &\sim \text{Gamma}\left(\frac{n_2 + p_2}{2} + \alpha_2, \eta_2(y_1, y_2)\right), \end{aligned} \quad (11)$$

where

$$\begin{aligned}
\beta_{10}(\mathbf{y}_1, \mathbf{y}_2) &= (X_1^t X_1 + \Lambda_1^{-1})^{-1} (X_1^t \mathbf{y}_1 + \Lambda_1^{-1} \beta_{10}), \\
\beta_{20}(\mathbf{y}_1, \mathbf{y}_2) &= (X_2^t X_2 + \Lambda_2^{-1})^{-1} (X_2^t \mathbf{y}_2 + \Lambda_2^{-1} \beta_{20}) \\
\eta_1(\mathbf{y}_1, \mathbf{y}_2) &= \left( \frac{1}{\eta_1} + \frac{1}{2} \left( \frac{\mathbf{y}_1^t \mathbf{y}_1 + \beta_{10}^t \Lambda_1^{-1} \beta_{10} - (X_1^t \mathbf{y}_1 + \Lambda_1^{-1} \beta_{10})^t (X_1^t X_1 + \Lambda_1^{-1})^{-1} (X_1^t \mathbf{y}_1 + \Lambda_1^{-1} \beta_{10})}{(X_1^t \mathbf{y}_1 + \Lambda_1^{-1} \beta_{10})} \right) \right)^{-1} \\
\eta_2(\mathbf{y}_1, \mathbf{y}_2) &= \left( \frac{1}{\eta_2} + \frac{1}{2} \left( \frac{\mathbf{y}_2^t \mathbf{y}_2 + \beta_{20}^t \Lambda_2^{-1} \beta_{20} - (X_2^t \mathbf{y}_2 + \Lambda_2^{-1} \beta_{20})^t (X_2^t X_2 + \Lambda_2^{-1})^{-1} (X_2^t \mathbf{y}_2 + \Lambda_2^{-1} \beta_{20})}{(X_1^t \mathbf{y}_1 + \Lambda_1^{-1} \beta_{10})} \right) \right)^{-1}.
\end{aligned}$$

Again the posterior distribution of  $\delta$  is nonstandard and this prevents the computation of the posterior density of  $\tau$  in a simple closed form. We have, however, that the conditional posterior distribution of  $\delta$  given  $(\sigma_1^2, \sigma_2^2)$  is

$$\delta \mid \sigma_1^2, \sigma_2^2, \mathbf{y}_1, \mathbf{y}_2 \sim N(\mu_\delta(\sigma_1, \sigma_2, \mathbf{y}_1, \mathbf{y}_2), \sigma_\delta^2(\sigma_1, \sigma_2, \mathbf{y}_1, \mathbf{y}_2)) \quad (12)$$

where

$$\begin{aligned}
\mu_\delta(\sigma_1^2, \sigma_2^2, \mathbf{y}_1, \mathbf{y}_2) &= \frac{v_2^t \beta_{20}(\mathbf{y}_1, \mathbf{y}_2) - v_1^t \beta_{10}(\mathbf{y}_1, \mathbf{y}_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \\
\sigma_\delta^2(\sigma_1^2, \sigma_2^2, \mathbf{y}_1, \mathbf{y}_2) &= \frac{\sigma_1^2 v_1^t (X_1^t X_1 + \Lambda_1^{-1})^{-1} v_1 + \sigma_2^2 v_2^t (X_2^t X_2 + \Lambda_2^{-1})^{-1} v_2}{\sigma_1^2 + \sigma_2^2}
\end{aligned}$$

and this simplifies the simulation. We discuss this further in Section 3.

### 3 Computations

The essence of the computational difficulty that we need to overcome, to implement the inferences discussed in Section 2, is the computation of the ratio

$$\frac{\pi_\Upsilon(\tau \mid x_0)}{\pi_\Upsilon(\tau)} \quad (13)$$

for arbitrary values of  $\tau$ . We note that, while (13) involves densities with respect to the support measure on  $\mathcal{T}$ , we need only compute their ratio. Actually this is the density of the posterior distribution of  $\tau$  with respect to the prior distribution of  $\tau$ . In contexts where we cannot evaluate this density in closed form we need to approximate (13) numerically for many values of  $\tau$ . In essence we need to numerically differentiate the posterior measure with respect to the prior measure.

To do this, for the problem under discussion, we can proceed via simulation. There are several possibilities for this. First we describe a brute force approach.



Let us denote the prior distribution function of  $\tau$  by  $F_\Upsilon$  and the posterior distribution function by  $F_\Upsilon(\cdot | x_0)$ . Using (4), (8) and (10) we then generate a sample  $\tau_1, \dots, \tau_{N_1}$  from the prior distribution of  $\tau$  and construct the estimate

$$\hat{F}_\Upsilon(\tau) = \frac{1}{N_1} \sum_{i=1}^{N_1} I_{(-\infty, \tau]}(\tau_i).$$

Similarly, we use (8), (11) and (12) to generate a sample  $\tau_1^*, \dots, \tau_{N_2}^*$  from the posterior distribution of  $\tau$  and construct the estimate

$$\hat{F}_\Upsilon(\tau | x_0) = \frac{1}{N_2} \sum_{i=1}^{N_2} I_{(-\infty, \tau]}(\tau_i^*).$$

It is worth noting here that, based on our comments in Section 2, we need only simulate from the prior and posterior distributions of  $(\sigma_1^2, \sigma_2^2, \delta)$  and then use (8) to get the sampled values of  $\tau$ , i.e., we have a reduction in dimension from  $2 + p_1 + p_2$  to  $2 + 1 = 3$ .

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

$$\frac{\hat{F}_\Upsilon(\hat{\tau}_{i+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_i | x_0)}{\hat{F}_\Upsilon(\hat{\tau}_{i+1}) - \hat{F}_\Upsilon(\hat{\tau}_i)} \rightarrow \frac{F_\Upsilon(\hat{\tau}_{i+1} | x_0) - F_\Upsilon(\hat{\tau}_i | x_0)}{F_\Upsilon(\hat{\tau}_{i+1}) - F_\Upsilon(\hat{\tau}_i)} \quad (14)$$

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

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

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

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

$$\sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0)} (F_\Upsilon(\hat{\tau}_{j+1} | x_0) - F_\Upsilon(\hat{\tau}_j | x_0)) \quad (15)$$

converges to (6) as  $N_3 \rightarrow \infty$ .

**Proof:** See the Appendix.

Theorem 1 and (14) suggest that we approximate (6) by

$$\sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0)} (\hat{F}_\Upsilon(\hat{\tau}_{j+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_j | x_0)) \quad (16)$$

where

$$\hat{R}(\tau_0) = \left\{ \hat{\tau}_{j+1} : \frac{\hat{F}_Y(\hat{\tau}_{j+1} | \mathbf{x}_0) - \hat{F}_Y(\hat{\tau}_j | \mathbf{x}_0)}{\hat{F}_Y(\hat{\tau}_{j+1}) - \hat{F}_Y(\hat{\tau}_j)} > \right\}. \quad (17)$$

We have the following result.

**Theorem 2.** Suppose that  $\epsilon > 0$  is given. Then, under the hypotheses of Theorem 1, there exist  $N_1, N_2$ , and  $N_3$  so that, for all larger values of these quantities, (16) differs from (15) by at most  $\epsilon$  almost surely.

**Proof:** See the Appendix.

So to approximate the ORS we choose the grid of points  $\hat{\tau}_1 < \dots < \hat{\tau}_{N_3}$ , generate the samples from the prior and posterior, and finally compute (16). Typically we will need to generate from the posterior first, however, to determine the grid of points so that the values

$$\frac{\hat{F}_Y(\hat{\tau}_{i+1} | \mathbf{x}_0) - \hat{F}_Y(\hat{\tau}_i | \mathbf{x}_0)}{\hat{F}_Y(\hat{\tau}_{i+1}) - \hat{F}_Y(\hat{\tau}_i)} \quad (18)$$

can provide an adequate approximation to the ratio (13), for  $N_1$  and  $N_2$  large enough, in the region where there is appreciable posterior probability. Further we will need to smooth the values given by (18), as we can see that choosing the intervals  $(\hat{\tau}_i, \hat{\tau}_{i+1})$  to be quite short will require very large values of  $N_1$  and  $N_2$  to obtain an accurate approximation to (18) at each point. We discuss these issues further in Section 4.

We note that we can use the values given by (18) 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 then use this to calculate an approximate  $\alpha$ -relative surprise interval for  $\tau$  as in (7).

We also consider another approach to approximating the relative surprise inferences for  $\tau$  that uses some additional features of the problem we are considering here. Notice that if  $(\sigma_{11}^2, \sigma_{21}^2), \dots, (\sigma_{1N_1}^2, \sigma_{2N_1}^2)$  is a sample from the prior distribution  $\Pi$ , then

$$\begin{aligned} \sum_{i=1}^{N_1} \Pi(\tau \leq \tau_0 | \sigma_{1i}^2, \sigma_{2i}^2) &= \sum_{i=1}^{N_1} \Pi(\delta \leq \Phi^{-1}(\tau_0) | \sigma_{1i}^2, \sigma_{2i}^2) \\ &= \sum_{i=1}^{N_1} \Phi\left(\frac{\Phi^{-1}(\tau_0) - \mu_\delta(\sigma_{1i}^2, \sigma_{2i}^2)}{\sigma_\delta(\sigma_{1i}^2, \sigma_{2i}^2)}\right) \\ &\rightarrow F_Y(\tau_0) \end{aligned}$$

almost surely as  $N_1 \rightarrow \infty$ . Accordingly we can substitute

$$\sum_{i=1}^{N_1} \Phi\left(\frac{\Phi^{-1}(\hat{\tau}_j) - \mu_\delta(\sigma_{1i}^2, \sigma_{2i}^2)}{\sigma_\delta(\sigma_{1i}^2, \sigma_{2i}^2)}\right)$$

for  $\hat{F}_\tau(\hat{\tau}_j)$  and a similar result holds for the posterior distribution function. It is easy to see that Theorems 1 and 2 also apply to this approximation. We discuss this approach, referred to hereafter as the Rao-Blackwellized approach, further in Section 4.

Using the invariance of relative surprise inferences we can avoid the need to evaluate  $\Phi^{-1}$  in the above computations. For relative surprise inferences about  $\delta$  can be transformed into the corresponding inferences about  $\tau$  by the transformation  $\Phi$ . If  $\hat{\delta}$  is the LRSE of  $\delta$ , then  $\hat{\tau} = \Phi(\hat{\delta})$  is the LRSE of  $\tau$ , the ORS for  $\tau$  at  $\tau_0$  is the same as the ORS for  $\delta$  at  $\delta_0 = \Phi^{-1}(\tau_0)$  and the value  $\delta_0$  is in an  $\alpha$ -relative surprise region for  $\delta$  if and only if  $\tau_0$  is in an  $\alpha$ -relative surprise region for  $\tau$ . So it is simpler to carry out the necessary computations for  $\delta$  and then transform the final inferences for  $\delta$  to get the inferences for  $\tau$ .

## 4 Examples

We consider two examples. Our first example is a generated example where we know the exact value of  $\tau$  so that we can examine the accuracy of the inferential and computational methodology discussed in the previous sections. The second example is the analysis of a real data set.

We note that the dimensions of the  $\beta$ -vectors will not affect the accuracy of our approximations as we are able to effectively integrate out  $p_1 + p_2 - 1$  variables. Accordingly, we can consider low dimensional examples in assessing how well the methodology discussed here works.

### Example 1. Simulated data

We consider the case where  $p_1 = p_2 = 2$ ,

$$\begin{aligned}\beta_{11} &= 1, & \beta_{12} &= 0, & \sigma_1 &= 1, \\ \beta_{21} &= 2, & \beta_{22} &= 1, & \sigma_2 &= 1,\end{aligned}$$

and  $X_1 = X_2 \in R^{20 \times 2}$  with the first column entries all equal to 1 and second column equal to  $(1, 2, \dots, 20)$ . We generated  $n_1 = 20$  values  $z_{11}, \dots, z_{1n_1}$  from the  $N(0, 1)$  distribution and put  $y_{1i} = 1 + z_{1i}$  and generated  $n_2 = 20$  values  $z_{21}, \dots, z_{2n_2}$  from the  $N(0, 1)$  distribution and put  $y_{2i} = 2 + i + z_{2i}$ . Then we take  $v_1^t = v_2^t = (1, 1)$  so that

$$\tau = \Phi\left(\frac{3-1}{\sqrt{2}}\right) = \Phi(\sqrt{2}) = 0.921348.$$

For the prior we put

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

so the prior is fairly diffuse.

It turns out that the Rao-Blackwellized approach to computing is much more efficient here than the first method described in Section 3 (the brute force method). So, to obtain the exact value of the function of  $\tau$  given by (13), we ran the Rao-Blackwell algorithm for many hours with  $N_3 = 10^2$ . A plot of (13) appears in Figure 1 as the solid line and represents the exact value of this function of  $\tau$ . 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 brute force algorithm with  $N_1 = N_2 = 10^5$  (taking 10 seconds of computing time), where we have smoothed the prior and posterior cdf estimates by averaging a point with 2 points on both sides (a 5-point average), is plotted in the dashed curve and we see that this provides a reasonable approximation, although it is a bit rough. In Figure 2 the dashed line corresponds to the brute force method with the same smoothing and  $N_1 = N_2 = 5 \times 10^5$  (taking 3 minutes of computing time), and we see that this provides a more acceptable approximation.

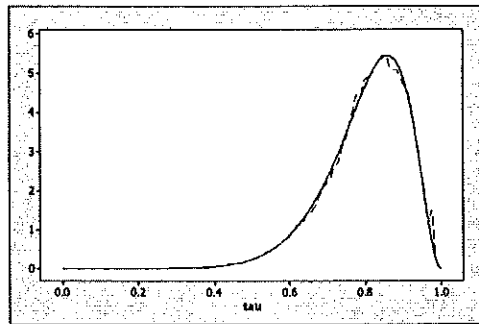


Figure 1: Plot of ratio of posterior to prior densities in Example 1 where the solid line denotes the exact ratio and the Rao-Blackwell estimate, and the dashed line denotes the brute force estimate based on  $N_1 = N_2 = 10^5$ .

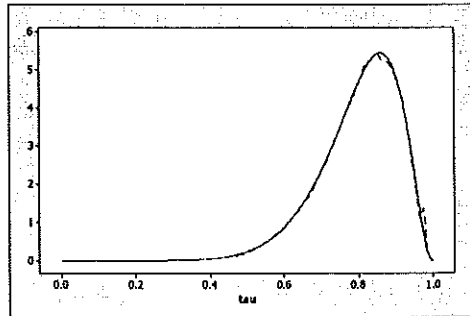


Figure 2: Plot of ratio of posterior to prior densities in Example 1 where the solid line denotes the exact ratio and the Rao-Blackwell estimate, and the dashed line denotes the brute force estimate based on  $N_1 = N_2 = 5 \times 10^5$ .

It is clear that the Rao-Blackwellized approach provides a more efficient approach to computation. Of course, this approach will not necessarily be available in another problem and so it is comforting that the brute force approach, which is always available, will work within feasible computing times.

Using the results of the above computations we can determine relative surprise inferences. For example, the LRSE of  $\tau$  is given by 0.861 and the .95-relative surprise interval for  $\tau$  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. The HPD intervals will always be the shortest .95-credible intervals, but they do not possess the invariance property, i.e., if we want a .95-HPD interval for  $\psi(\tau)$ , where  $\psi$  is 1-1, then we must redo the computations, while the .95-relative surprise interval is obtained merely by applying  $\psi$  to the end-points of the interval for  $\tau$ . The invariance property for relative surprise inferences is a consistency property (the analysis of a problem does not depend on the parameterization) but it is also satisfying that, in this example at least, not much is lost in our assessment of the accuracy of the inference via the length of the interval. Note that both intervals contain the true value.

**Example 2. Shear strength of welds**

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. The normal simple linear regression model  $Y_i = \beta_{i0} + \beta_{i1}X_i + \epsilon_i$  is used for both  $Y_1$  and  $Y_2$ , 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  $\tau$  when  $X_1 = X_2 = 200$ .

$y_1$	$x_1$	$y_2$	$x_2$
350	380	680	190
380	155	800	200
385	160	780	209
450	165	885	215
465	175	875	215
185	165	1025	215
535	195	1100	230
555	185	1030	250
590	195	1175	265
605	210	1300	250

For the prior we put

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

and we note that these choices increase the diffuseness of the prior considerably over that in Example 1. In Figure 3 we have plotted the prior and posterior

densities for  $\tau$  that resulted. We note the very strange prior that is induced for  $\tau$  and the high concentration of the posterior near 1. Note that 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).

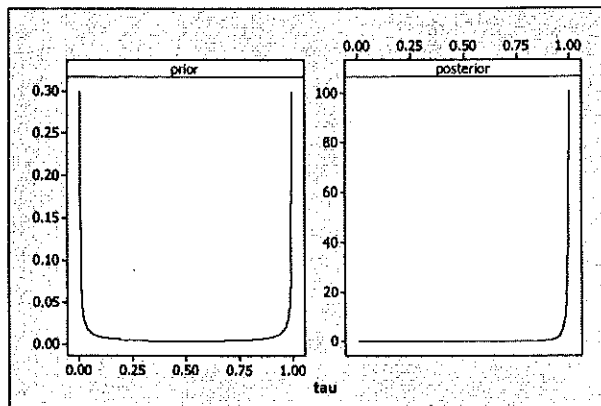


Figure 3: Plot of prior and posterior distribution of  $\tau$  in Example 2.

Given the somewhat unnatural look of the prior for  $\tau$ , however, we experimented with different choices. Working directly with the prior on the model parameters, lead to the conclusion that the more diffuse we made the prior on the model parameters the more the prior on  $\tau$  exhibited the tendency to place all its mass near 0 or 1. Looking at (8)-(10), however, we see that the choices

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

produce a uniform prior distribution for  $\tau$ . For this choice the LRSE is 0.935 and the .95-relative surprise interval is (0.779, 0.989) while the posterior mode is 0.935, and the .95-HPD interval is (0.779, 0.989). In this case the inferences are the same because of the uniformity of the prior for  $\tau$ . The point estimates are close to what the previous prior gave but we note the much wider intervals, i.e., there is much more uncertainty concerning the true value of  $\tau$ .

## 5 Conclusions

We have examined the use of relative surprise inferences in a problem where the prior and posterior densities of the quantity of interest are not available in

closed form. This necessitates some additional computations over those required when obtaining traditional Bayesian inferences. When the prior and posterior densities of the quantity of interest are available in closed form, the computational requirements for relative surprise inferences are similar to those needed to implement more typical Bayesian inferences.

We have demonstrated two effective methods of approaching this problem when the quantity of interest is 1-dimensional. Note that these methods and the convergence results are applicable quite generally. While many quantities of interest in inference problems are 1-dimensional, 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. Note that the effectiveness of the methods described here does not depend on the dimension of the model parameter, here  $(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ , but rather on the dimension of the quantity of interest, here  $\tau = \Upsilon(\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)$ .

Another limitation of the methodology here is the need for algorithms to sample from the prior and posterior distributions on the model parameter. In this paper we have chosen the prior so that there are algorithms for sampling directly from these distributions, i.e., i.i.d. samples, but more generally we can use MCMC methods such as Gibbs sampling or Metropolis-Hastings, for example, as described in Evans and Swartz (2000).

Given the similarity of relative surprise inferences to posterior inferences one might wonder at the need for the relative surprise approach. First we note that in a number of problems, see Evans (1997), there is a great difference in the inferences. This seems to occur mostly in problems where posterior inferences seem to be irregular in some respect. Second there is the satisfying invariance property that relative surprise inferences possess. This can be viewed as a kind of consistency requirement in the sense that two different statisticians proceeding from the same ingredients (sampling model, prior, data and inference criterion), with possible 1-1 relabellings of these ingredients, should produce inferences that are consistent under the relabelling. For example, in the frequentist domain likelihood methods for the full parameter accomplish this, while other methods do not. In the Bayesian context, relative surprise methods possess this invariance. Third, as documented in Evans and Zou (2002) relative surprise inferences are more robust to the choice of the prior than traditional Bayesian inferences. In problems where we feel that the prior information is rather weak this seems like a satisfying property. Relative surprise inferences are based on how the data changes beliefs from *a priori* to *a posteriori* and as such are more data dependent than pure posterior inferences.

When prior information is weak one might wonder why we don't simply use a noninformative prior. In the problem we have been discussing, however, a noninformative prior will be improper and in that case their use becomes at least somewhat debatable. For example, the inability to assess the prior distribution on a parameter of interest seems to us to be of concern. At the very least the debate about the use of improper priors remains open and, at this point, we do not view the dependence of relative surprise methods on proper priors

as a weakness. As discussed in Evans (1997), it is always possible to consider limiting relative surprise inferences as priors become increasingly diffuse as one approach to recovering the completely noninformative prior formulation.

## 6 Appendix

### Posterior Distribution of the Parameter

Writing a general quadratic form  $x^t A x$  as  $x^t A(\cdot)$  we have that the likelihood times prior is proportional to

$$\begin{aligned}
& \sigma_1^{-n_1} \exp \left\{ -\frac{1}{2\sigma_1^2} (y_1 - X_1 \beta_1)^t (y_1 - X_1 \beta_1) \right\} \times \\
& \sigma_2^{-n_2} \exp \left\{ -\frac{1}{2\sigma_2^2} (y_2 - X_2 \beta_2)^t (y_2 - X_2 \beta_2) \right\} \times \\
& \sigma_1^{-p_1} \exp \left\{ -\frac{1}{2\sigma_1^2} (\beta_1 - \beta_{10})^t \Lambda_1^{-1} (\beta_1 - \beta_{10}) \right\} \times \\
& \sigma_2^{-p_2} \exp \left\{ -\frac{1}{2\sigma_2^2} (\beta_2 - \beta_{20})^t \Lambda_2^{-1} (\beta_2 - \beta_{20}) \right\} \times \\
& (\sigma_1^2)^{-\alpha_1+1} \exp \left\{ -\frac{1}{\eta_1 \sigma_1^2} \right\} (\sigma_2^2)^{-\alpha_2+1} \exp \left\{ -\frac{1}{\eta_2 \sigma_2^2} \right\} \\
= & \exp \left\{ \begin{array}{l} -\frac{1}{2\sigma_1^2} \left( \beta_1 - (X_1^t X_1 + \Lambda_1^{-1})^{-1} (X_1^t y_1 + \Lambda_1^{-1} \beta_{10}) \right)^t \\ \quad (X_1^t X_1 + \Lambda_1^{-1}) \\ \left( \beta_1 - (X_1^t X_1 + \Lambda_1^{-1})^{-1} (X_1^t y_1 + \Lambda_1^{-1} \beta_{10}) \right) \end{array} \right\} \times \\
& \exp \left\{ \begin{array}{l} -\frac{1}{2\sigma_2^2} \left( \beta_2 - (X_2^t X_2 + \Lambda_2^{-1})^{-1} (X_2^t y_2 + \Lambda_2^{-1} \beta_{20}) \right)^t \\ \quad (X_2^t X_2 + \Lambda_2^{-1}) \\ \left( \beta_2 - (X_2^t X_2 + \Lambda_2^{-1})^{-1} (X_2^t y_2 + \Lambda_2^{-1} \beta_{20}) \right) \end{array} \right\} \times \\
& (\sigma_1^2)^{-\frac{n_1+p_1}{2}-\alpha_1+1} \exp \left\{ -\frac{1}{\eta_1 (y_1, y_2) \sigma_1^2} \right\} \times \\
& (\sigma_2^2)^{-\frac{n_2+p_2}{2}-\alpha_2+1} \exp \left\{ -\frac{1}{\eta_2 (y_1, y_2) \sigma_2^2} \right\}
\end{aligned}$$

This verifies the result presented in Section 2 for the posterior distribution of the parameter. ■

### Proof of Theorem 1

Define the random variable

$$X = X(\tau) = \frac{\pi_{\mathcal{T}}(\tau | x_0)}{\pi_{\mathcal{T}}(\tau)}$$



and, for the grid  $0 = \hat{\tau}_1 < \dots < \hat{\tau}_{N_3} = 1$ , define the random variable

$$X_{N_3} = X_{N_3}(\tau) = \frac{F_Y(\hat{\tau}_{j+1} | x_0) - F_Y(\hat{\tau}_j | x_0)}{F_Y(\hat{\tau}_{j+1}) - F_Y(\hat{\tau}_j)}$$

whenever  $\tau \in (\hat{\tau}_j, \hat{\tau}_{j+1}]$ . Note that  $X_{N_3}$  is not defined when  $F_Y(\hat{\tau}_{j+1}) = F_Y(\hat{\tau}_j)$  but we can ignore this because this implies  $F_Y(\hat{\tau}_{j+1} | x_0) = F_Y(\hat{\tau}_j | x_0)$  (the posterior is absolutely continuous with respect to the prior) and a similar comment applies to the definition of  $X$  when  $\pi_Y(\tau) = 0$ .

Now suppose that as  $N_3 \rightarrow \infty$  we choose the grids so that

$$\sup \{\hat{\tau}_{j+1} - \hat{\tau}_j : j = 1, \dots, N_3\} \rightarrow 0.$$

Then we have that  $X_{N_3} \rightarrow X$  as  $N_3 \rightarrow \infty$  almost surely with respect to the posterior  $\Pi(\cdot | x_0)$ . This implies that  $X_{N_3}$  converges in distribution to  $X$  as  $N_3 \rightarrow \infty$ .

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

$$\nu_{N_3} = \frac{F_Y(\hat{\tau}_{i+1} | x_0) - F_Y(\hat{\tau}_i | x_0)}{F_Y(\hat{\tau}_{i+1}) - F_Y(\hat{\tau}_i)} \rightarrow \frac{\pi_Y(\tau_0 | x_0)}{\pi_Y(\tau_0)} = \nu_0$$

as  $N_3 \rightarrow \infty$ . Therefore, if  $\eta > 0$ , then for all  $N_3$  large enough

$$\nu_{N_3} \in (\nu_0 - \eta, \nu_0 + \eta)$$

and

$$\begin{aligned} & |\Pi(X_{N_3} \leq \nu_{N_3} | x_0) - \Pi(X \leq \nu_0 | x_0)| \\ & \leq \max \{ |\Pi(X_{N_3} \leq \nu_0 + \eta | x_0) - \Pi(X \leq \nu_0 + \eta | x_0)|, \\ & \quad |\Pi(X_{N_3} \leq \nu_0 - \eta | x_0) - \Pi(X \leq \nu_0 - \eta | x_0)| \}. \end{aligned}$$

We can choose  $\eta$  so that  $\nu_0 - \eta$  and  $\nu_0 + \eta$  are continuity points for the distribution of  $X$  and so the right-hand side of the above inequality converges to

$$\max \{ |\Pi(X \leq \nu_0 + \eta | x_0) - \Pi(X \leq \nu_0 + \eta | x_0)|, |\Pi(X \leq \nu_0 - \eta | x_0) - \Pi(X \leq \nu_0 - \eta | x_0)| \}. \quad (19)$$

Since  $\nu_0$  is a continuity point of the distribution of  $X$ , (19) can be made as small as we like by an appropriate choice of  $\eta$  and so

$$\limsup_{N_3 \rightarrow \infty} |\Pi(X_{N_3} \leq \nu_{N_3} | x_0) - \Pi(X \leq \nu_0 | x_0)|$$

can be made as small as we like. This completes the proof. ■

### Proof of Theorem 2

We use some of the notation and results from the proof of Theorem 1. Since  $\nu_0$  is a continuity point for the distribution of  $X$  we can choose  $\eta > 0$  so that  $\nu_0 - \eta$  and  $\nu_0 + \eta$  are continuity points for  $X$  and

$$\Pi(X \in (\nu_0 - \eta, \nu_0 + \eta) | x_0) = \Pi(X \leq \nu_0 + \eta | x_0) - \Pi(X \leq \nu_0 - \eta | x_0)$$

is smaller than  $\epsilon/4$ . Since  $\nu_{N_3} \rightarrow \nu_0$  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

$$\begin{aligned} & \Pi(X_{N_3} \in (\nu_0 - \delta, \nu_0 + \delta] | x_0) \\ &= \Pi(X_{N_3} \leq \nu_0 + \eta | x_0) - \Pi(X_{N_3} \leq \nu_0 - \eta | x_0) < \epsilon/2 \end{aligned}$$

for all  $N_3$  large enough. Putting

$$S(\tau_0) = \left\{ \hat{\tau}_{j+1} : \frac{F_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - F_{\Upsilon}(\hat{\tau}_j | x_0)}{F_{\Upsilon}(\hat{\tau}_{j+1}) - F_{\Upsilon}(\hat{\tau}_j)} = \nu_{N_3} \right\},$$

the above implies that for all  $N_3$  large enough

$$\begin{aligned} \sum_{\hat{\tau}_{j+1} \in S(\tau_0)} (F_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - F_{\Upsilon}(\hat{\tau}_j | x_0)) &= \Pi(X_{N_3} = \nu_{N_3} | x_0) \\ &\leq \Pi(X_{N_3} \in (\nu_0 - \delta, \nu_0 + \delta] | x_0) \\ &< \epsilon/2. \end{aligned}$$

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

$$\frac{F_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - F_{\Upsilon}(\hat{\tau}_j | x_0)}{F_{\Upsilon}(\hat{\tau}_{j+1}) - F_{\Upsilon}(\hat{\tau}_j)} > \frac{F_{\Upsilon}(\hat{\tau}_{i+1} | x_0) - F_{\Upsilon}(\hat{\tau}_i | x_0)}{F_{\Upsilon}(\hat{\tau}_{i+1}) - F_{\Upsilon}(\hat{\tau}_i)}.$$

Therefore, since

$$\frac{\hat{F}_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - \hat{F}_{\Upsilon}(\hat{\tau}_j | x_0)}{\hat{F}_{\Upsilon}(\hat{\tau}_{j+1}) - \hat{F}_{\Upsilon}(\hat{\tau}_j)} \rightarrow \frac{F_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - F_{\Upsilon}(\hat{\tau}_j | x_0)}{F_{\Upsilon}(\hat{\tau}_{j+1}) - F_{\Upsilon}(\hat{\tau}_j)}$$

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

$$\frac{F_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - F_{\Upsilon}(\hat{\tau}_j | x_0)}{F_{\Upsilon}(\hat{\tau}_{j+1}) - F_{\Upsilon}(\hat{\tau}_j)} < \frac{F_{\Upsilon}(\hat{\tau}_{i+1} | x_0) - F_{\Upsilon}(\hat{\tau}_i | x_0)}{F_{\Upsilon}(\hat{\tau}_{i+1}) - F_{\Upsilon}(\hat{\tau}_i)}$$

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

$$\begin{aligned} & \sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0)} \left( \hat{F}_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - \hat{F}_{\Upsilon}(\hat{\tau}_j | x_0) \right) \\ &= \sum_{\hat{\tau}_{j+1} \in R(\tau_0)} \left( \hat{F}_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - \hat{F}_{\Upsilon}(\hat{\tau}_j | x_0) \right) \\ &+ \sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0) \cap S(\tau_0)} \left( \hat{F}_{\Upsilon}(\hat{\tau}_{j+1} | x_0) - \hat{F}_{\Upsilon}(\hat{\tau}_j | x_0) \right). \end{aligned}$$

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

$$\begin{aligned} & \sum_{\hat{\tau}_{j+1} \in R(\tau_0)} \left( \hat{F}_\Upsilon(\hat{\tau}_{j+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_j | x_0) \right) \\ \rightarrow & \sum_{\hat{\tau}_{j+1} \in R(\tau_0)} \left( F_\Upsilon(\hat{\tau}_{j+1} | x_0) - F_\Upsilon(\hat{\tau}_j | x_0) \right) \end{aligned}$$

and

$$\begin{aligned} & \sum_{\hat{\tau}_{j+1} \in \hat{R}(\tau_0) \cap S(\tau_0)} \left( \hat{F}_\Upsilon(\hat{\tau}_{j+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_j | x_0) \right) \\ \leq & \sum_{\hat{\tau}_{j+1} \in S(\tau_0)} \left( \hat{F}_\Upsilon(\hat{\tau}_{j+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_j | x_0) \right) \\ \rightarrow & \sum_{\hat{\tau}_{j+1} \in S(\tau_0)} \left( F_\Upsilon(\hat{\tau}_{j+1} | x_0) - F_\Upsilon(\hat{\tau}_j | x_0) \right) \end{aligned}$$

almost surely as  $N_1, N_2 \rightarrow \infty$ . Then for all  $N_1, N_2$  large enough we have that (16) is within  $\epsilon/4$  of (15) and  $\sum_{\hat{\tau}_{j+1} \in S(\tau_0)} (\hat{F}_\Upsilon(\hat{\tau}_{j+1} | x_0) - \hat{F}_\Upsilon(\hat{\tau}_j | x_0))$  is within  $\epsilon/4$  of  $\sum_{\hat{\tau}_{j+1} \in S(\tau_0)} (F_\Upsilon(\hat{\tau}_{j+1} | x_0) - F_\Upsilon(\hat{\tau}_j | x_0))$ . This completes the proof. ■

## 7 References

- 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, p. 13-17.
- Evans, M. (1997) Bayesian inference procedures derived via the concept of relative surprise. *Commun. Statist.-Theory And Methods*, Vol. 26, No. 5, p. 1125-1143.
- Evans, M., and Swartz, T. (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, 352 pages.
- 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.), p. 90-115, Imperial College Press.
- Guttman, I., Johnson, R.A., Bhattacharayya, G.K., and Reiser, B. (1988) Confidence limits for stress-strength models with explanatory variables. *Technometrics*, 30, p. 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, p. 143-158.

- Reiser, B., and Guttman, I. (1989) Sample size choice for reliability verification on stress-strength models. *Canadian Journal of Statistics*, Vol. 17, p. 253-259.
- Simonoff, J.S., Hochberg, Y., and Reiser, B. (1986) Alternative estimation procedures for  $P(X < Y)$  in categorical data. *Biometrics*, 42, p. 895-907.