Optimal Properties of Some Bayesian Inferences

by

Michael Evans
Department of Statistics
University of Toronto

and

Mohammed Shakhatreh
Department of Statistics
University of Toronto

Technical Report No. 0710, October 1, 2007

TECHNICAL REPORT SERIES

UNIVERSITY OF TORONTO

DEPARTMENT OF STATISTICS

Optimal Properties of Some Bayesian Inferences

M. Evans and M. Shakhatreh
Dept. of Statistics
U. of Toronto

Abstract: We consider various properties of Bayesian inferences related to repeated sampling interpretations, when we have a proper prior. While these can be seen as particularly relevant when the prior is diffuse, we argue that it is generally reasonable to consider such properties as part of our assessment of Bayesian inferences. We discuss the logical implications for how repeated sampling properties should be assessed when we have a proper prior. We develop optimal Bayesian repeated sampling inferences using a generalized idea of what it means for a credible region to contain a false value and discuss the practical use of this idea for error assessment and experimental design. We present results that connect Bayes factors with optimal inferences and develop a generalized concept of unbiasedness for credible regions. Further, we consider the effect of reparameterizations on hpd-like credible regions and argue that one reparameterization is most relevant, when repeated sampling properties and the prior are taken into account.

Key words and phrases: repeated sampling properties, relative surprise inferences, prior-data conflict, Bayes factors, relative belief ratios.

1 Introduction

Repeated sampling properties and Bayesian properties of inferences are sometimes seen as antithetical. In some ways this reflects different views of the interpretation of probability. Whether or not such different interpretations are relevant to the development of statistical theory is an issue we do not pursue further here. We note, however, that this does not play a role in the development of probability theory and one could argue that it really shouldn't in statistics either.

For example, it is commonly noted that a confidence coefficient does not correspond to a probability and posterior probabilities do not correspond to long-run relative frequencies. While Bayesian inferences possess good coherence properties, at least relative to confidence inferences, there is no denying the appeal that long-run relative frequencies have in validating inferences. It is

part of our argument here, however, that the way in which we assess repeated sampling properties is crucial.

One approach is to ask that inferences possess good repeated sampling properties uniformly over all sampling distributions prescribed in a model. While this has some appeal, we argue that this is often not appropriate as it does not take into account all the information available. For example, it is common that sampling models are used that contain some distributions for which we know it is virtually impossible for one of them to correspond to the true distribution. Perhaps it is thought that the data will effectively exclude these candidates in the inference process, and so the inclusion of such distributions is relatively harmless. For example, we may believe that the heights of individuals in a particular population are approximately normally distributed and so include all normal distributions in our model, even though we know that some of these are not possible. While this may seem innocuous, however, it does not seem appropriate to then treat all distributions as equivalent candidates when we assess the behavior of inferences. For example, if an inference procedure did well for those candidate distributions that we regarded as reasonable, but did very poorly for some distributions that we regarded as virtually impossible, we would not regard this as a reason not to use the procedure.

Perhaps the most convenient way to handle such situations is via a prior that is fairly diffuse or flat over the plausible distributions and which downweights the implausible values. We argue then, that the repeated sampling behavior of Bayesian inferences are still relevant, but our assessment of these must also reflect the prior. We develop this idea further in section 2.

One objection to this approach is that the prior is subjective, in the sense that it reflects beliefs about what the true value of the model parameter is. Of course, the sampling model is also subjective as it also reflects beliefs about the distribution that produced the data. Both the sampling model and the prior are subjective choices made by an analyst or, for that matter, by a committee of analysts. Rather than distinguishing one from the other in terms of their probabilistic interpretations, it seems more fruitful to simply ask whether or not these choices make sense in light of the data observed. After all, at least when it is collected correctly, the data is unequivocally objective. The sampling model is then acceptable if, for at least one distribution in the family prescribed, the observed data is not surprising. There is an extensive literature on checking the plausibility of the sampling model. If the sampling model is acceptable, it is then reasonable to ask if the prior makes sense. We characterize this as an absence of prior-data conflict, namely, the prior does not place most of its mass on distributions for which the observed data is surprising. An approach to assessing prior-data conflict is discussed in Evans and Moshonov (2006, 2007).

While the sampling model and prior are subjective choices, once they have been accepted, via model checking and checking for prior-data conflict, we then need to derive relevant inferences based on these ingredients together with the data. Bayes theorem implies that probability statements must be conditional on the observed data, i.e., based on the posterior distribution. This doesn't say anything, however, about what form our inferences should take, e.g., nothing

is prescribed about the form of a γ-credible region for an unknown. We will see that the results in sections 3, 4 and 5 lead to a particular class of Bayesian inferences that have optimal repeated sampling behavior and maximize Bayes factors. These inferences were discussed, from a different viewpoint, in Evans (1997) and called relative surprise inferences there, as the development was motivated by how to measure surprise in a Bayesian hypothesis assessment problem. See, for example, Good(1988), for some further discussion concerned with measuring surprise.

In essence, relative surprise regions are hpd regions where the base support measure is taken to be the prior measure. For various reasons it is natural to take this base measure to be the prior measure as opposed to another choice, such as volume measure. For example, when the parameter of interest does not take its values in a Euclidean space, we need an alternative to volume measure to determine inferences. Perhaps most significantly, however, the use of the prior measure as the base support measure leads to Bayesian inferences that are invariant under 1-1 transformations. Invariance under reparameterization is a very natural property to require of inferences and many inferences, frequentist or Bayesian, do not satisfy this requirement. Invariance can be seen as a basic coherence requirement of any approach to inference.

To discuss these inferences we first present some notation. Suppose we observe data x_0 from a statistical model $\{f_\theta : \theta \in \Theta\}$, 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 Θ. With these ingredients we have available the joint distribution of (θ, x), as given by the density $f_\theta(x)\,\pi(\theta)$ with respect to support measure $\nu \times \mu$, and the observed value x_0. The prior predictive measure of x is given by

$$M(B) = E_\Pi(P_\theta(B)) = \int_\Theta \int_B f_\theta(x)\pi(\theta)\,\mu(dx)\,\nu(d\theta) = \int_B m(x)\,\mu(dx)$$

with density $m(x) = \int_\Theta f_\theta(x)\pi(\theta)\,\nu(d\theta)$ with respect to μ. Further the posterior measure of θ is given by

$$\Pi(A \mid x) = \int_A \pi(\theta \mid x)\,\nu(d\theta)$$

with density $\pi(\theta \mid x) = f_\theta(x)\,\pi(\theta)/m(x)$ with respect to ν.

Consider now a quantity of interest $\tau = \Upsilon(\theta)$ taking values in a set $\mathcal{T}$. For this quantity we denote the marginal posterior and prior measures by $\Pi_\Upsilon(\cdot \mid x)$ and Π_Υ respectively, with corresponding densities $\pi_\Upsilon(\cdot \mid x)$ and π_Υ, taken with respect to some support measure $\nu_\mathcal{T}$ on $\mathcal{T}$. Relative surprise inferences are based on a particular approach to assessing a hypothesis such as $H_0 : \tau = \tau_0$. For this we compute the *observed relative surprise* (ORS) given by

$$\Pi_\Upsilon\left(\frac{\pi_\Upsilon(\tau \mid x_0)}{\pi_\Upsilon(\tau)} > \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\pi_\Upsilon(\tau_0)} \;\middle|\; x_0\right). \tag{1}$$

We see that (1) compares 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. For estimation, we consider (1) as a function of τ_0 and select a value which minimizes this quantity as the estimate. We refer to this estimate as the *least relative surprise (LRSE)* estimate. This estimate can also be obtained by maximizing $\pi_\Upsilon(\cdot \mid x_0)/\pi_\Upsilon(\cdot)$. To obtain a γ-credible region for τ we simply invert (1) in the standard way to obtain the *γ-relative surprise region*

$$C_\gamma(x_0) = \left\{ \tau_0 \in \mathcal{T} : \Pi_\Upsilon \left(\frac{\pi_\Upsilon(\tau \mid x_0)}{\pi_\Upsilon(\tau)} > \frac{\pi_\Upsilon(\tau_0 \mid x_0)}{\pi_\Upsilon(\tau_0)} \,\middle|\, x_0 \right) \leq \gamma \right\}. \qquad (2)$$

Note that (1) is in accord with Bayes theorem as probabilities are computed based on the posterior. It is also clear from (1) that relative surprise inferences are invariant under smooth reparameterizations.

In Evans, Guttman and Swartz (2006) it was shown that relative surprise inferences possess an optimality property in the class of Bayesian inferences. In that development, (2) was taken as the basic concept. In particular, if we consider the class of all γ-credible regions for a function $\tau = \Upsilon(\theta)$ of the model parameter θ, then the γ-relative surprise region for τ has the smallest prior content among all γ-credible regions for this quantity. Hypothesis assessments and estimates are derived from relative surprise regions in a direct way and so also possess optimal properties. The LRSE is obtained by taking the region with $\gamma = 0$ and the ORS is obtained as $\inf\{\gamma : \tau_0 \in C_\gamma(x_0)\}$. In section 3 we show that this optimal property has a direct interpretation in terms of repeated sampling.

As discussed in Evans and Zou (2002) and Evans, Guttman and Swartz (2006), there is a close connection between relative surprise inferences and Bayes factors. In section 4 we establish some results that deepen this connection and show that relative surprise inferences lead to some optimal results when interpreted in terms of Bayes factors. Further, we introduce the relative belief ratio as an alternative method for measuring change of belief from *a priori* to *a posteriori*. In section 5 we discuss reparameterizations when considering the repeated sampling properties of credible regions. Finally, in section 6 we make some general comments about the assessment of repeated sampling properties and draw some conclusions.

2 Repeated Sampling Properties of Bayesian Regions

Suppose we have a parameter of interest $\tau = \Upsilon(\theta)$, and a rule based on the sampling model and prior, such that for each $\gamma \in [0, 1]$ and $x \in \mathcal{X}$, the rule determines a region $B_\gamma(x) \subset \mathcal{T}$ satisfying $\Pi_\Upsilon(B_\gamma(x) \mid x) \geq \gamma$. So $B_\gamma(x)$ is a γ-credible region for $\Upsilon(\theta)$. We might then consider the coverage probability $P_\theta(\Upsilon(\theta) \in B_\gamma(x))$ of this region, as a function of $\theta \in \Theta$. Particularly when Π is a diffuse prior, supposedly representing relative ignorance about the true value

of θ and thus τ, it might seem reasonable to require that $P_\theta(\Upsilon(\theta) \in B_\gamma(x)) \geq \gamma$ for all θ as well, or at least be such that the inequality hold in the limit as the diffuseness of Π increases. The repeated sampling interpretation of the coverage probability would in such a case provide greater force to the interpretation of the posterior probability.

From results of Joshi (1974), however, we know that it is generally not possible to simultaneously obtain both posterior probability content and coverage probability equal to γ. The following example, adapted from Stein (1959), illustrates that the discrepancy between the posterior content and coverage probability can be disturbingly wide.

Example 1.

Suppose that the coordinates of $x = (x_1, \ldots, x_n)$ are independent with $x_i \sim N(\theta_i, 1)$ for $i = 1, 2, \ldots, n$ and the θ_i are i.i.d. $N(0, \sigma^2)$ with σ^2 known. Then, if $\tau_n^2 = \sum_{i=1}^n \theta_i^2$ it is easy to show that the posterior distribution of $(1+1/\sigma^2)\tau_n^2$ is Chi-squared$(n, (1+1/\sigma^2)^{-1}||x||^2)$. From this we get an approximate γ-credible interval for τ_n^2/n, obtained by discarding $(1-\gamma)/2$ of the probability in each tail of the posterior, given by

$$\begin{aligned} &\left\{(1+1/\sigma^2)^{-1} + (1+1/\sigma^2)^{-2}(||x||^2/n)\right\} \\ &\pm\sqrt{2(1+1/\sigma^2)^{-2} + 4(1+1/\sigma^2)^{-3}(||x||^2/n)}(z_{(1+\gamma)/2}/\sqrt{n}) \end{aligned} \tag{3}$$

where $z_{(1+\gamma)/2}$ is the $(1+\gamma)/2$-quantile of the $N(0,1)$ distribution.

Assume that $\tau_n^2/n \to \tau_*^2$ as $n \to \infty$ and consider the interval given by

$$\begin{aligned} &\left\{(1+1/\sigma^2)^{-1} + (1+1/\sigma^2)^{-2}(1+\tau_*^2)\right\} \\ &\pm\sqrt{2(1+1/\sigma^2)^{-2} + 4(1+1/\sigma^2)^{-3}(1+\tau_*^2)}(z_{(1+\gamma)/2}/\sqrt{n}). \end{aligned} \tag{4}$$

Comparing (3) and (4) we see that the differences in the respective endpoints converge to 0, and the ratio of their lengths goes to 1, in probability as $n \to \infty$. So (3) and (4) are asymptotically equivalent. Now consider whether or not (4) contains $\tau_n^2/n \approx \tau_*^2$. An easy argument, see Evans and Shakhatreh (2007), shows that (4) always contains τ_*^2 when $\tau_*^2 = \sigma^2$ and when $\tau_*^2 \neq \sigma^2$, (4) will never contain τ_*^2 for all n large enough.

Accordingly, the interval given by (3) will generally have terrible coverage properties whenever $\tau_*^2 \neq \sigma^2$. This is borne out by the simulation results reported in Evans (1997).

One might question the necessity of having $P_\theta(\Upsilon(\theta) \in B_\gamma(x)) \geq \gamma$ for all θ. This uniformity property treats all θ values equally while the prior Π may not. For example, if we are sampling from a population of humans and measuring their heights in centimeters, and we assume the distribution of the heights is $N(\theta, \sigma^2)$ with $\theta \in R^1, \sigma^2 > 0$ unknown, then it doesn't make any sense to consider θ values that are too small (let alone negative) or too large when assessing the repeated sampling performance of intervals for θ. Perhaps the simplest way of dealing with this problem is to place a proper prior on the space that reflects, as best as possible, what we know about θ. In any case,

when we choose a proper prior Π we are typically saying some values of θ are effectively impossible and, if we know that this is the case, then it does not seem reasonable to demand good coverage properties for the region at such values.

One might argue that we never really want to rule out any values of θ as the prior may be wrong, in the sense that the true value of θ lies in a region of low prior probability. But this kind of error is similar to the possibility that the sampling model is wrong. In fact, we make a choice of a sampling model for a problem that we believe is sensible, and then proceed to check the model against the data. The model is acceptable if, for at least one distribution in the model, the observed data is not surprising. If we do not find evidence against the model, then we can proceed to use this model for inference. Similarly, we make a choice of a prior that we believe is sensible and then we check the prior against the data. Once we have accepted the sampling model and the prior as reasonable, then we are prepared to proceed to inference and we suppose that this involves the quotation of a region (or regions) such as $B_\gamma(x_0)$.

Accordingly, throughout the remainder of our discussion we will assume that the model and the prior are acceptable and then examine the implications of this for assessing the repeated sampling behavior of inferences such as B_γ. The first question we ask then is: how should we assess the repeated sampling behavior of B_γ? Note that, since $\Pi_\Upsilon(B_\gamma(x) \,|\, x) \geq \gamma$, then

$$\begin{aligned} \gamma &\leq E_M(\Pi_\Upsilon(B_\gamma(x) \,|\, x)) = \int_{\mathcal{X}} \int_\Theta I_{B_\gamma(x)}(\Upsilon(\theta))\, \Pi_\Upsilon(d\theta \,|\, x)\, M(dx) \\ &= \int_{\Theta \times \mathcal{X}} I_{B_\gamma(x)}(\Upsilon(\theta))\, P_\theta(dx)\, \Pi(d\theta) = E_\Pi(P_\theta(\Upsilon(\theta) \in B_\gamma(x))) \end{aligned} \tag{5}$$

and so the average coverage probability of B_γ, with respect to the prior Π, is always at least γ. Of course, when $\Pi_\Upsilon(B_\gamma(x) \,|\, x) = \gamma$ for all x, then we have equality in (5).

While (5) might be viewed as comforting, Example 1 illustrates that, in general, it does not tell us much about the coverage probability at the true value. For (5) says that if we generate a value of θ from Π, generate x from P_θ, then the probability is at least γ that $B_\gamma(x)$ will contain $\Upsilon(\theta)$. This doesn't say anything about the probability of $B_\gamma(x)$ containing the true value. In Example 1 the prior distribution of τ_n^2/σ^2 is Chi-squared$(n, 0)$ implying that τ_n^2/n has mean σ^2 and variance $2\sigma^4/n$. So the prior concentrates at σ^2 as n gets large and, when $\tau_n^2/n \approx \tau_*^2 \neq \sigma^2$, then (5) holds but the coverage of (3) is terrible.

One approach to dealing with this problem is to change the prior. This implies that there are good and bad priors for problems and, when dealing with proper priors at least, this seems a bit unnatural, particularly if the prior does not conflict with the data. Another approach is to use an inference methodology that leads to choosing the rule B_γ so that bad properties are avoided as much as possible. Without additional ingredients to the problem, there is nothing that forces us to form $B_\gamma(x_0)$, in the case of a 1-dimensional parameter, by discarding equal amounts of probability in the tails of $\Pi_\Upsilon(\cdot \,|\, x)$, or more generally to use an hpd region. One might, for example, ask for a choice of B_γ that possessed

good repeated sampling properties.

It is not unequivocally clear, however, what we want these repeated sampling properties to be. While the confidence property has its appeal, the plethora of absurd confidence regions, see Plante (1991) for some discussion, might at least lead one to doubt the wisdom of focusing too closely on this goal. We now discuss an alternative approach.

3 Minimizing the Probability of Covering False Values

In this section, we look for B_γ that will minimize the probability of covering false values in a somewhat different sense than that usually encountered in discussions of confidence regions. For this, we suppose that "nature", or some hidden process, chooses a value $\theta^* \in \Theta$ as the true value of the parameter. We allow this true value to be chosen quite generally according to a probability measure Π^* on Θ. For example, Π^* could be degenerate at some element of Θ. Naturally we would like B_γ to do a good job at covering the true value $\Upsilon(\theta^*)$. If $\Upsilon(\theta^*)$ is in the effective support of Π_Υ, which corresponds to no prior-data conflict, then (5) gives some support that this is the case.

Now suppose, however, that we test the coverage properties of B_γ by computing the probability that B_γ contains an independently generated $\tau \sim \Lambda$, for some probability measure Λ on $\mathcal{T}$. This probability is given by

$$\int_\Theta \int_\mathcal{T} P_{\theta^*}(\tau \in B_\gamma(x))\, \Lambda(d\tau)\, \Pi^*(d\theta^*). \tag{6}$$

Note that, when Λ is a continuous measure, then (6) is precisely the probability of B_γ covering a false value when $\theta^* \sim \Pi^*, x \sim P_{\theta^*}$ independent of $\tau \sim \Lambda$. More generally, it is the probability of covering a value of the parameter of interest that is generated in a way that is independent of the way the true value of θ and the data are generated. Suppose that we have two γ-credible regions B^1_γ and B^2_γ and (6) is smaller for B^1_γ than for B^2_γ. Clearly the value of τ has nothing to do with the true value $\Upsilon(\theta^*)$ and so B^2_γ has a greater chance of containing an arbitrary value of τ from Λ than B^1_γ. Therefore, $B^2_\gamma(x)$ can be considered a less accurate inference about $\Upsilon(\theta^*)$ than $B^1_\gamma(x)$.

So among all regions B_γ satisfying $\Pi_\Upsilon(B_\gamma(x) \,|\, x) \geq \gamma$ we look for the one which minimizes (6). For this we need some preliminary results. We will ignore measurability concerns although these can accommodated with minor modifications to our arguments. Suppose we have a probability measure P and a σ-finite measure Q on a set Ω. Further, suppose that P and Q are both absolutely continuous with respect to the same measure on Ω with respective densities p and q. For $\gamma \in [0,1]$, let

$$D_\gamma = \left\{ \omega_0 \in \Omega : P\left(\frac{p(\omega)}{q(\omega)} > \frac{p(\omega_0)}{q(\omega_0)} \right) \leq \gamma \right\}.$$

Then the following results are proved in Evans, Guttman and Swartz (2006). In that paper P is taken to be the posterior but otherwise the proofs are the same.

Lemma 1. $P(D_\gamma) \geq \gamma$ with equality whenever the distribution of $p(\omega)/q(\omega)$, with $\omega \sim P$, has no atoms.

Theorem 2. The set D_γ minimizes $Q(D)$ among all measurable sets $D \subset \Omega$ satisfying $P(D) \geq P(D_\gamma)$. Further, when the distribution of $p(\cdot)/q(\cdot)$ has no atoms, then D_γ minimizes $Q(D)$ among all measurable sets $D \subset \Omega$ satisfying $P(D) \geq \gamma$.

We then have the following result that establishes the optimality, with respect to (6), of hpd-like credible regions.

Theorem 3. Suppose that the probability distribution Λ is also absolutely continuous with respect to $\nu_{\mathcal{T}}$ on $\mathcal{T}$ with density λ. Then, in the Bayesian model given by $\Pi \times P_\theta$, the region $B_{\Lambda,\gamma}$ given by

$$B_{\Lambda,\gamma}(x) = \left\{ \tau_0 \in \mathcal{T} : \Pi\left(\frac{\pi_\Upsilon(\tau \,|\, x)}{\lambda(\tau)} > \frac{\pi_\Upsilon(\tau_0 \,|\, x)}{\lambda(\tau)} \,\middle|\, x \right) \leq \gamma \right\}$$

minimizes (6) among all regions B satisfying $\Pi_\Upsilon(B(x) \,|\, x) \geq \Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x)$. If $\Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x) = \gamma$ for each x, then $B_{\Lambda,\gamma}$ minimizes (6) among all γ-credible regions for $\Upsilon(\theta)$.

Proof: Putting $P = \Pi_\Upsilon(\cdot \,|\, x)$ and $Q = \Lambda$ in Theorem 2, implies that $B_{\Lambda,\gamma}(x)$ minimizes $\Lambda(B(x))$ among all sets $B(x)$ satisfying $\Pi_\Upsilon(B(x) \,|\, x) \geq \Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x)$. Now let M^* be the measure given by $M^*(A) = \int_\Theta P_{\theta^*}(A)\, \Pi^*(d\theta^*)$. Then, clearly $E_{M^*}(\Lambda(B(x)))$ is minimized, among all regions B satisfying $\Pi_\Upsilon(B(x) \,|\, x) \geq \Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x)$ for each x, by $B = B_{\Lambda,\gamma}$.

Now observe that

$$\begin{aligned}
E_{M^*}(\Lambda(B(x))) &= E_{M^*} E_\Lambda(I_{B(x)}(\tau)) \\
&= \int_\Theta \int_{\mathcal{X}} E_\Lambda(I_{B(x)}(\tau))\, P_{\theta^*}(dx)\, \Pi^*(d\theta^*) \\
&= \int_\Theta \int_{\mathcal{X}} \int_{\mathcal{T}} I_{B(x)}(\tau)\ \Lambda(d\tau)\, P_{\theta^*}(dx)\, \Pi^*(d\theta^*) \\
&= \int_\Theta \int_{\mathcal{T}} \int_{\mathcal{X}} I_{B(x)}(\tau)\ P_{\theta^*}(dx)\, \Lambda(d\tau)\, \Pi^*(d\theta^*) \\
&= \int_\Theta \int_{\mathcal{T}} P_{\theta^*}(\tau \in B(x))\, \Lambda(d\tau)\, \Pi^*(d\theta^*)
\end{aligned}$$

which is (6). This completes the proof.

When we specialize Π^* to be degenerate at the "true" value θ^*, we have the following result.

Corollary 4. In the Bayesian model given by $\Pi \times P_\theta$, the region $B_{\Lambda,\gamma}$ minimizes the probability of covering τ, when $x \,|\, \theta^* \sim P_{\theta^*}$ independent of $\tau \sim \Lambda$, among all credible regions B for $\Upsilon(\theta^*)$ satisfying $\Pi_\Upsilon(B(x) \,|\, x) \geq \Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x)$ for each x. If $\Pi_\Upsilon(B_{\Lambda,\gamma}(x) \,|\, x) = \gamma$ for each x, then $B_{\Lambda,\gamma}$ minimizes this probability among all γ-credible regions for $\Upsilon(\theta)$.

Note that Theorem 3 is independent of how we choose Π^*, i.e., the optimal region does not depend on how the true value of θ arises. The optimal region is dependent, however, on the choice of Λ. We have the following result.

Corollary 5. If $\Lambda = \Pi_\Upsilon$, then $B_{\Lambda,\gamma} = C_\gamma$, i.e., the optimal region is the γ-relative surprise region.

So now consider a sequence $(\theta_i^*, x_i, \tau_i)$ for $i = 1, 2, \ldots$ of independent values from the joint distribution $\Pi^* \times P_{\theta^*} \times \Pi_\Upsilon$. Then Corollary 5 says that, among all γ-credible regions B_γ for $\Upsilon(\theta^*)$ formed from $\Pi \times P_\theta$, a γ-relative surprise region for $\Upsilon(\theta)$ minimizes the proportion of times the event $\tau_i \in B_\gamma(x_i)$ is true. If we take $\Pi^* = \Pi$, then we get the result that, in the sequence (θ_i, x_i, τ_i) for $i = 1, 2, \ldots$ of independent values from the joint distribution $\Pi \times P_\theta \times \Pi_\Upsilon$, a γ-relative surprise region for $\Upsilon(\theta)$ minimizes the proportion of times the event $\tau_i \in B_\gamma(x_i)$ is true among all γ-credible regions B_γ for $\Upsilon(\theta)$ formed from $\Pi \times P_\theta$ and, further, the event $\Upsilon(\theta_i) \in C_\gamma(x_i)$ is true at least γ of the time.

So we see that relative surprise regions have a general optimal repeated sampling property when $\Lambda = \Pi_\Upsilon$ and, in addition, have the correct coverage property when $\Pi^* = \Pi$. One might ask why we should choose $\Lambda = \Pi_\Upsilon$ and $\Pi^* = \Pi$. Consider now our discussion of how coverage should be assessed in Bayesian contexts when we have agreed, via checking for prior-data conflict, that the prior is acceptable. In such circumstances Π gives the appropriate weighting to the various possible true values of θ and similarly Π_Υ does this when considering false values of $\Upsilon(\theta)$. For example, it does not make sense to consider the coverage properties of a credible region at values where Π assigns little or even no probability. Similarly, we can ignore false values of $\Upsilon(\theta)$ where Π_Υ places little weight. Of course, there might be other ingredients added to a problem that would lead to a different choice of Λ and Π^*, but $\Lambda = \Pi_\Upsilon$ and $\Pi^* = \Pi$ would seem to be the most natural choices.

We also might wonder how sensible it is to minimize (6) with $\Lambda = \Pi_\Upsilon$, when the prior is concentrated at a value τ_0. If this is not the true value, and we have sufficient data, then we will detect a prior-data conflict at the prior checking stage and proceed no further. So certainly checking for prior-data conflict is a necessary part of our argument. If we check, and obtain that a prior-data conflict exists, then minimizing (6) with $\Lambda = \Pi_\Upsilon$ doesn't make sense. Of course, this reasoning also applies to the sampling model. We must check the sampling model before using it in the inference stage, as inferences derived from an inappropriate model are not generally valid. On the other hand when τ_0 is the true value, and Π_Υ is degenerate at τ_0, then the posterior $\Pi_\Upsilon(\cdot \mid x)$ is also degenerate at τ_0 and so, for any $\gamma > 0$, we have that $C_\gamma(x) = \{\tau_0\}$.

When $\Pi^* = \Pi, B_\gamma(x) = C_\gamma(x)$ and $\Lambda = \Pi_\Upsilon$, then (6) becomes

$$E_M\left(\Pi_\Upsilon(C_\gamma(x))\right) = \int_\Theta \int_\mathcal{T} P_\theta(\tau \in C_\gamma(x))\, \Pi_\Upsilon(d\tau)\, \Pi(d\theta) \tag{7}$$

and this is the probability that $C_\gamma(x)$ contains τ when $\theta \sim \Pi, x \sim P_\theta$ independently of $\tau \sim \Pi_\Upsilon$. While the optimality of C_γ is independent of the choice of Π^*, for a particular application, choosing $\Pi^* = \Pi$ makes sense as Π reflects

beliefs about what the true value of θ is. Now (7) would appear to have several practical uses. First we can quote this probability as a way of assessing the accuracy of C_γ with a given prior. If this probability is quite high, then we have a region with low accuracy. We see also that (7) can be used for experimental design purposes such as setting sample size.

Consider the following example as an illustration of the behavior of (7).

Example 2. *Location normal*

Suppose that $x = (x_1, \ldots, x_n)$ is a sample from the $N(\theta, 1)$ distribution and $\theta \sim N(0, \sigma^2)$. Then the posterior distribution of θ is $N((n + 1/\sigma^2)^{-1} n\bar{x}, (n + 1/\sigma^2)^{-1})$. The ratio of the posterior density to the prior density is, in this case, proportional to the likelihood $\exp\left\{-n(\theta - \bar{x})^2/2\right\}$. Therefore, a γ-relative surprise interval for θ is a likelihood interval and so takes the form $C_\gamma(x) = \bar{x} \pm k_\gamma(n, \bar{x}, \sigma^2)$ where $k_\gamma(n, \bar{x}, \sigma^2) \geq 0$ satisfies

$$\begin{aligned} \gamma &= \Phi\left((n + 1/\sigma^2)^{-1/2}\bar{x}/\sigma^2 + (n + 1/\sigma^2)^{1/2} k_\gamma(n, \bar{x}, \sigma^2)\right) - \\ &\quad \Phi\left((n + 1/\sigma^2)^{-1/2}\bar{x}/\sigma^2 - (n + 1/\sigma^2)^{1/2} k_\gamma(n, \bar{x}, \sigma^2)\right). \end{aligned} \tag{8}$$

The value $k_\gamma(n, \bar{x}, \sigma^2)$ is easily obtained numerically from (8).

Now $\Pi_\Upsilon(C_\gamma(x)) = \Phi\left((\bar{x} + k_\gamma(n, \bar{x}, \sigma^2))/\sigma\right) - \Phi\left((\bar{x} - k_\gamma(n, \bar{x}, \sigma^2))/\sigma\right)$ and $\bar{x} \sim N(0, \sigma^2 + 1/n)$. Therefore, (7) is given by

$$\begin{aligned} &\int_{-\infty}^{\infty} [\Phi\left(((\sigma^2 + 1/n)^{1/2} z + k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2))/\sigma\right) - \\ &\Phi\left(((\sigma^2 + 1/n)^{1/2} z - k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2))/\sigma\right)] \varphi(z)\, dz. \end{aligned} \tag{9}$$

Note that $k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2)$ satisfies

$$\begin{aligned} \gamma &= \Phi\left(z/(n^{1/2}\sigma) + (n + 1/\sigma^2)^{1/2} k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2)\right) - \\ &\quad \Phi\left(z/(n^{1/2}\sigma) - (n + 1/\sigma^2)^{1/2} k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2)\right). \end{aligned} \tag{10}$$

The value of (9) can be obtained via numerical integration or simulation.

Suppose that $n \to \infty$ in (10). Then we have that $(n + 1/\sigma^2)^{1/2} k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2) \to z_{(1+\gamma)/2}$ where $z_{(1+\gamma)/2}$ is the $((1+\gamma)/2)$-quantile of the $N(0, 1)$ distribution. Therefore, as $n \to \infty$, $k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2) \to 0$ so that the integrand in (9) converges to 0. Applying the dominated convergence theorem we have that (9) converges to 0. So by choosing n large enough we can make (9) as small as we like and so control the error in our inferences. For example, the following table gives some values of (9) when $\gamma = .95$ and $\sigma^2 = 1$, based on a Monte Carlo sample of size 10^3 with the standard errors in parentheses.

n	1	10	25	50
$E_M\left(\Pi_\Upsilon(C_\gamma(x))\right)$	.700 (.004)	.322 (.004)	.212 (.003)	.152 (.002)

The interpretation of $E_M\left(\Pi_\Upsilon(C_\gamma(x))\right) = .322$ when $n = 10$ is then as follows. When the true value of θ is generated from the $N(0, \sigma^2)$ distribution, the data

$x = (x_1, \ldots, x_n)$ is a sample generated from the $N(\theta, 1)$ distribution and an independent value θ' is generated from the $N(0, \sigma^2)$ distribution, then the probability that $\theta' \in C_\gamma(x)$ is .322.

Suppose in (9) that $\sigma^2 \to \infty$, so the prior is becoming more diffuse. Then again $(n + 1/\sigma^2)^{1/2} k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2) \to z_{(1+\gamma)/2}$ and so $k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2) \to n^{-1/2} z_{(1+\gamma)/2}$ and $k_\gamma(n, (\sigma^2 + 1/n)^{1/2} z, \sigma^2)/\sigma \to 0$. Thus the integrand in (9) converges to 0 as $\sigma \to \infty$ and the dominated convergence theorem implies that (9) converges to 0. Accordingly as the prior becomes more diffuse the probability of covering a false value generated from the prior converges to 0. This is exactly how we would want our region to behave, namely, the data become much more important in determining the inference as the prior becomes more diffuse. For example, $C_\gamma(x) \to \bar{x} \pm n^{-1/2} z_{(1+\gamma)/2}$ as $\sigma^2 \to \infty$. So for a very diffuse prior, $C_\gamma(x)$ has a very small probability of covering an independently generated value from the prior. Further, this proves that $\bar{x} \pm n^{-1/2} z_{(1+\gamma)/2}$, is optimal under this sequence of priors.

The above illustrates that we can't use (7) to compare priors. A more concentrated prior will give a higher value for (7) than one more diffuse. But really we have different regions $C_\gamma(x)$ and different distributions for the false values. This emphasizes the importance of a careful choice of the prior so that unrealistic values of the parameter are excluded and also checking for prior-data conflict to confirm our assumptions about what the unrealistic values are. If we choose the prior too diffuse, then we are in effect optimizing at unrealistic values of the parameter which we have argued doesn't make sense. Of course these comments are less relevant when we have large amounts of data.

In general (7) provides an *a priori* method for designing experiments. For example, if we have a regression model and will observe a response at n input predictor vectors $v_1, \ldots, v_n$ then we would choose these vectors to minimize (7). A tabulation of this minimum for each n then tells us how many observations we need to attain a prescribed error probability. We do not pursue this topic further here.

Corollary 5 can also be interpreted in terms of the ORS given by (1). For suppose we agree to reject the hypothesis $H_0 : \Upsilon(\theta) = \tau_0$ whenever the ORS is greater than γ. This is equivalent to rejecting H_0 whenever $\tau_0 \in C_\gamma^c(x)$. Now consider the class of tests specified by γ-credible regions B_γ, so we reject H_0 whenever $\tau_0 \in B_\gamma^c(x)$. In this case, we take $\Pi^* = \Pi(\cdot \mid \Upsilon(\theta) = \tau_0)$, $\Lambda = \Pi_\Upsilon$, and find B_γ maximizing 1 minus (6), namely,

$$\int_\Theta \int_\mathcal{T} P_\theta(\tau \in B_\gamma^c(x))\, \Pi_\Upsilon(d\tau)\, \Pi(d\theta \mid \Upsilon(\theta) = \tau_0). \tag{11}$$

This is the conditional probability, given that H_0 is true, that we would reject the hypothesis specified by τ, when τ is a value independently generated from the prior. We refer to (11) as the sensitivity of the test based on B_γ and it is clearly analogous to power in the frequentist context. Then Corollary 5 says that this is maximized by $C_\gamma^c(x)$ among all rejection regions with posterior content less than or equal to $1 - \gamma$ and so C_γ^c is most sensitive among such rejection

regions. We can compute (11), with $B_\gamma^c = C_\gamma^c$, for a specific H_0 to determine a sample size so that the test has a prescribed sensitivity.

4 Bayes Factors, Relative Belief Ratios and Unbiasedness

If $C \subset \mathcal{T}$, then the Bayes factor in favor of C is given by

$$BF_C(x) = \frac{\Pi_\Upsilon(C \,|\, x)}{1 - \Pi_\Upsilon(C \,|\, x)} \frac{1 - \Pi_\Upsilon(C)}{\Pi_\Upsilon(C)}.$$

If we let C shrink "nicely" (for example, see Rudin (1974)) to a point $\{\tau_0\}$, then $BF_C(x)$ converges to $\pi_\Upsilon(\tau_0 \,|\, x_0)/\pi_\Upsilon(\tau_0)$. So we can think of this quantity as an approximation to the Bayes factor associated with τ_0 and the ORS is a calibration of this value to determine if it is indeed small and thus evidence against τ_0 as a plausible value.

The Bayes factor in favor of C is a measure of the change in our belief that C contains the true value from *a priori* to *a posteriori*. Perhaps a simpler measure of this change in belief is given by what we will call the *relative belief ratio,* namely,

$$RB_C(x) = \frac{\Pi_\Upsilon(C \,|\, x)}{\Pi_\Upsilon(C)}.$$

Again, as C converges nicely to $\{\tau_0\}$, $RB_C(x)$ converges to $\pi_\Upsilon(\tau_0 \,|\, x_0)/\pi_\Upsilon(\tau_0)$. Note that $BF_C(x) = RB_C(x)/RB_{C^c}(x)$ and so BF_C is not a function of RB_C or conversely. They are measuring change in belief on different scales. Clearly the two will be approximately equal when $RB_{C^c}(x) \approx 1$ and this will occur whenever C is "small".

Now consider a γ-relative surprise region $C_\gamma(x)$ for τ. From (2), and the fact that the function $\Pi_\Upsilon\left(\pi_\Upsilon(\tau \,|\, x)/\pi_\Upsilon(\tau) > k \,|\, x\right)$ is right-continuous in k, there exists $k_\gamma(x)$ such that

$$C_\gamma(x) = \{\tau : \pi_\Upsilon(\tau \,|\, x)/\pi_\Upsilon(\tau) > k_\gamma(x)\}.$$

From this we have the following property for relative surprise regions. We assume throughout the remainder of this section that $\pi_\Upsilon(\tau) > 0$ for every $\tau \in \mathcal{T}$.

Lemma 6. The relative surprise region $C_\gamma(x)$ satisfies $RB_{C_\gamma(x)}(x) > k_\gamma(x)$.

Proof: We have that

$$\begin{aligned} \Pi_\Upsilon(C_\gamma(x) \,|\, x) &= \int_{C_\gamma(x)} \pi_\Upsilon(\tau \,|\, x)\, \nu_{\mathcal{T}}(d\tau) > k_\gamma(x) \int_{C_\gamma(x)} \pi_\Upsilon(\tau)\, \nu_{\mathcal{T}}(d\tau) \\ &= k_\gamma(x) \Pi_\Upsilon(C_\gamma(x)) \end{aligned}$$

and the result follows.

So Lemma 6 says that the ratio of posterior to prior probabilities of $C_\gamma(x)$ satisfies the same inequality that the respective densities do on this set.

We have the following additional inequalities satisfied by the Bayes factor and the relative belief ratio for $C_\gamma(x)$.

Lemma 7. The relative surprise region $C_\gamma(x)$ satisfies $BF_{C_\gamma(x)}(x) > 1$ and $RB_{C_\gamma(x)}(x) > 1$.

Proof: Clearly we have that $C_\gamma^c(x) = \{\tau : \pi_\Upsilon(\tau \,|\, x)/\pi_\Upsilon(\tau) \leq k_\gamma(x)\}$ and, as in Lemma 6, this implies that $RB_{C_\gamma^c(x)}(x) \leq k_\gamma(x)$. Combining this with Lemma 6 gives that $BF_{C_\gamma(x)}(x) > 1$. Since $BF_C(x) > 1$, then $1/\Pi_\Upsilon(C) > 1/\Pi_\Upsilon(C \,|\, x)$ and this implies that $RB_{C_\gamma(x)}(x) > 1$.

Note that, since $BF_{C^c}(x) = 1/BF_C(x)$, we have that $BF_{C_\gamma^c(x)}(x) < 1$ and $RB_{C_\gamma^c(x)}(x) < 1$ for a relative surprise region $C_\gamma(x)$. Accordingly the Bayes factor and the relative belief ratio always indicates an increase in belief in the set $C_\gamma(x)$ from *a priori* to *a posteriori*. In particular, the posterior probability content of $C_\gamma(x)$ is always greater than its prior content. These would seem to be very natural properties to ask for a set that we are going to quote as a candidate to contain the true value of the parameter of interest. We would like the data to lead to an increase in our belief that the true value is in such a set. It would be odd to quote a set, as a candidate to contain the true value, whose probability decreased after seeing the data.

Of course, there may be other credible regions with these properties. For example, hpd regions often have these properties, although there does not seem to be an easy general proof of this. In any case, the following shows that relative surprise regions are best from this point-of-view.

Theorem 8. The set $C_\gamma(x)$ has maximal Bayes factor and maximal relative belief ratio among all measurable sets $C \subset \mathcal{T}$ satisfying $\Pi_\Upsilon(C \,|\, x) = \Pi_\Upsilon(C_\gamma(x) \,|\, x)$.

Proof: From Theorem 2 we know that $\Pi_\Upsilon(C\,)$ is minimized, among all measurable C satisfying $\Pi_\Upsilon(C \,|\, x) \geq \Pi_\Upsilon(C_\gamma(x) \,|\, x)$, by taking $C = C_\gamma(x)$. So $\Pi_\Upsilon(C\,)$ is also minimized by the same choice when we restrict to those C satisfying $\Pi_\Upsilon(C \,|\, x) = \Pi_\Upsilon(C_\gamma(x) \,|\, x)$. Since the function $f(x) = (1-x)/x$ is decreasing in x the result follows for the Bayes factor and the result for the relative belief ratio is obvious.

Theorem 8 is most relevant when there are a number of credible regions, including the relative surprise region $C_\gamma(x)$, with posterior content exactly equal to γ. Theorem 8 then says that $C_\gamma(x)$ is the best choice among these regions from the point of view of the Bayes factor and the relative belief ratio. For $C_\gamma(x)$ is the set containing γ of the posterior probability for which the data have lead to the largest increase in belief from *a priori* to *a posteriori*. A result relevant to repeated sampling considerations is then immediate from Theorem 8.

Corollary 9. Suppose that the true value of θ is selected according to Π^*. Then $E_{M^*}(BF_{B_\gamma(x)}(x))$ and $E_{M^*}(RB_{B_\gamma(x)}(x))$ are maximized, among all credible regions $B_\gamma(x)$ satisfying $\Pi_\Upsilon(B_\gamma(x) \,|\, x) = \Pi_\Upsilon(C_\gamma(x) \,|\, x)$ for all x, by $B_\gamma(x) = C_\gamma(x)$.

So, for example, no matter how the true value of θ is generated, the long-run average value of the Bayes factor of the γ-relative surprise region $C_\gamma(x)$ for $\tau = \Upsilon(\theta)$ is maximal among all credible regions with the same posterior content. Note that when Π^* is degenerate at θ^*, then $M^* = P_{\theta^*}$, which is the true distribution.

Consider the following example as an illustration.

Example 3. *Probability of joint success*

Suppose we observe x from a Binomial(n, θ_1), an independent y from a Binomial(n, θ_2) and we put independent uniform priors on θ_1 and θ_2. Suppose we are interested in making inference about $\psi = \theta_1\theta_2$. This is the probability of simultaneous success from tossing two coins where the coins have probability of heads equal to θ_1 and θ_2, respectively.

In Figure 1 we have plotted the prior density, the posterior density (when $n = 5, x = 4, y = 1$) and the ratio of these densities.

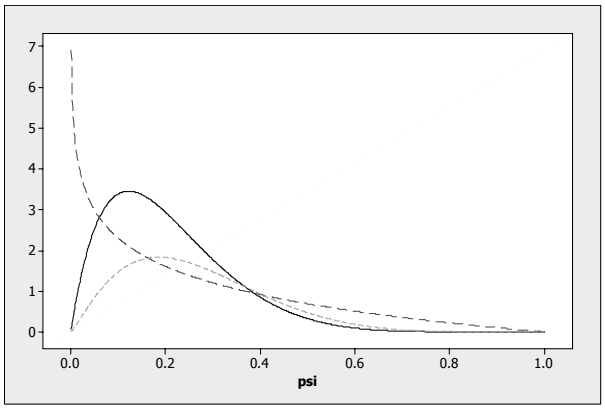

Figure 1: Plot of the prior – –, the posterior —, and ratio of posterior to prior - -, for ψ in Example 3 when $n = 5, x = 4$ and $y = 1$.

In the following table we give some γ-hpd intervals and γ-relative surprise intervals for ψ. We see that these intervals are quite different and that, considering the relative belief ratios and Bayes factors, the relative surprise intervals always dominate the hpd intervals, just as our results indicate they should. The estimate determined by the hpd approach is the mode and this is given by .122 while the LRSE is .186.

γ	hpd interval	RB	BF	rel. surp. interval	RB	BF
.95	(.008, .447)	1.25	5.99	(.028, .501)	1.32	7.35
.75	(.032, .293)	1.47	2.82	(.071, .361)	1.69	3.35
.50	(.059, .216)	1.57	2.16	(.110, .284)	1.74	2.48
.25	(.089, .163)	1.63	1.84	(.119, .270)	1.76	2.36

While the hpd intervals, in this example, always have RB > 1 and BF > 1, other methods of forming the intervals do not necessarily give intervals with

these properties. For example, if we took a left-tail of the posterior as a γ-credible interval for ψ, then the left-tail .4-credible interval has RB = .730 and BF = .640.

These results are somewhat typical for this example. For example, when $n = 20, x = 19, y = 19$, the .95-hpd interval is $(.675, .962)$ with RB = 16.16, and BF = 305.40, while the .95-relative surprise interval is $(.684, .990)$ with RB = 16.98, and BF = 322.59. In this case the posterior mode is .857 and the LRSE is .902.

In frequentist contexts, a confidence region is said to be unbiased, if the probability of the region containing a particular false value is always less than or equal to the probability of the region containing the true value. The following result shows that relative surprise regions are unbiased in a generalized sense.

Theorem 10. For a relative surprise region, the probability of containing an independent value generated from the prior is always less than the probability of containing the true value, when it is generated from the prior.

Proof: From Lemma 7 we have that $\Pi_\Upsilon(C_\gamma(x) \,|\, x) > \Pi_\Upsilon(C_\gamma(x))$ and so

$$E_M(\Pi_\Upsilon(C_\gamma(x))) < E_M(\Pi_\Upsilon(C_\gamma(x) \,|\, x))). \tag{12}$$

From (7), the left-hand side of (12) is the probability that $C_\gamma(x)$ contains τ when $\theta \sim \Pi, x \sim P_\theta$ independently of $\tau \sim \Pi_\Upsilon$. Now, by (5), the right-hand side of (12) is $E_\Pi(P_\theta(\Upsilon(\theta) \in C_\gamma(x)))$ which is the probability that $C_\gamma(x)$ contains the true value $\Upsilon(\theta)$ when $\theta \sim \Pi, x \sim P_\theta$.

When $\Pi_\Upsilon(C_\gamma(x) \,|\, x) = \gamma$ for all x then (12) says that the probability of covering a false value is strictly less than the posterior probability that we quote for the set $C_\gamma(x)$ to contain the true value.

5 Reparameterizations and Coverage

Suppose we focus on obtaining Bayesian credible regions whose posterior content will be close to their coverage probabilities. For this discussion we restrict our attention to those situations and rules for which we can obtain $\Pi_\Upsilon(B_\gamma(x) \,|\, x) = \gamma$ for every x, i.e., exact posterior content equal to γ. This will hold for hpd-like regions based on the measure Λ, whenever the distribution of $\pi_\Upsilon(\cdot \,|\, x)/\lambda(\cdot)$ has no atoms. This implies that $E_M(\Pi_\Upsilon(B_\gamma(x) \,|\, x)) = E_\Pi(P_\theta(\Upsilon(\theta) \in B_\gamma(x))) = \gamma$ and so the mean coverage under the prior is γ.

If our goal is to obtain a B_γ with coverage close to the posterior probability γ, then it would seem natural to prefer the γ-credible region for τ with the prior distribution of the coverage most concentrated about γ. A natural way to measure this concentration is via the variance, namely,

$$\begin{aligned} Var_\Pi(P_\theta(\Upsilon(\theta) \in B_\gamma(x))) &= E_\Pi((P_\theta(\Upsilon(\theta) \in B_\gamma(x))) - \gamma)^2) \\ &= E_\Pi(P^2_\theta(\Upsilon(\theta) \in B_\gamma(x))) - \gamma^2. \end{aligned}$$

So we look for the region B_γ with the smallest value of $E_\Pi(P_\theta^2(\Upsilon(\theta) \in B_\gamma(x)))$. In essence we can think of this as trying to find the γ-credible region for τ whose posterior content γ is closest to its coverage probability.

Now observe that

$$\begin{aligned} & E_\Pi(P_\theta^2(\Upsilon(\theta) \in B_\gamma(x))) \\ = & \int_\Theta \left\{ \int_{\mathcal{X}} I_{B_\gamma(x_1)}(\Upsilon(\theta))\, P_\theta(dx_1) \int_{\mathcal{X}} I_{B_\gamma(x_2)}(\Upsilon(\theta))\, P_\theta(dx_2) \right\} \Pi(d\theta) \\ = & \int_{\mathcal{X}\times\mathcal{X}} \left\{ \int_{\mathcal{T}} I_{B_\gamma(x_1)}(\tau) I_{B_\gamma(x_2)}(\tau)\, \Pi_\Upsilon\left(d\tau \,|\, x_1, x_2\right) \right\} M_2(d(x_1, x_2)) \\ = & E_{M_2}(\Pi_\Upsilon\left(B_\gamma(x_1) \cap B_\gamma(x_2) \,|\, x_1, x_2\right)) \end{aligned} \tag{13}$$

where M_2 is the prior predictive of a sample of size 2 from P_θ and $\Pi_\Upsilon\left(\cdot \,|\, x_1, x_2\right)$ is the posterior for τ having observed (x_1, x_2) when $\theta \sim \Pi$. So we want to find B_γ that minimizes (13) subject to the condition $\Pi_\Upsilon\left(B_\gamma(x) \,|\, x\right) = \gamma$ for every x. If we could find B_γ that minimized $\Pi_\Upsilon\left(B_\gamma(x_1) \cap B_\gamma(x_2) \,|\, x_1, x_2\right)$ for every x_1, x_2, then we would have the optimal γ-credible region, but such a solution does not seem generally possible. Note, however, that (13) can be used to evaluate $E_\Pi(P_\theta^2(\Upsilon(\theta) \in B_\gamma(x)))$.

As rules for forming credible regions, hpd-like regions based on some σ-finite measure Λ on $\mathcal{T}$ with density λ with respect to $\upsilon_{\mathcal{T}}$, provide an approach that can be used in almost any circumstance. For example, discarding probability in the tails only works with 1-dimensional parameters and, when the posterior is symmetric, corresponds to forming hpd intervals. When the posterior is not symmetric, it is not clear what amount should be discarded in each tail and, in such a case, hpd intervals are commonly used. Accordingly, we restrict our discussion to hpd-like credible regions and search for the measure Λ that minimizes $E_\Pi(P_\theta^2(\Upsilon(\theta) \in B_{\Lambda,\gamma}(x))$. We show that this problem is connected with reparameterizations.

We restrict attention to situations where $\mathcal{T}$ is an open subset of R^k. Let $\mathcal{D}_{\mathcal{T},\mathcal{T}}$ denote the class of transformations $\Psi : \mathcal{T} \to \mathcal{T}$ that are 1-1, onto, continuously differentiable and such that Ψ^{-1} is continuously differentiable. Suppose the rule being used, is to find the region that has smallest Λ measure. Then, by Theorem 3, the γ-credible region for $\psi = \Psi(\tau)$ that has minimal Λ content is given by the hpd-like region

$$B_{\Lambda,\gamma}^{\Psi}(x) = \left\{ \psi_0 \in \mathcal{T} : \Pi_\Upsilon \left(\begin{array}{c} \frac{\pi_\Upsilon(\tau \,|\, x) J_\Psi^{-1}(\tau)}{\lambda(\Psi(\tau))} > \\ \frac{\pi_\Upsilon(\Psi^{-1}(\psi_0) J_\Psi^{-1}\left(\Psi^{-1}(\psi_0)\right) |\, x)}{\lambda(\psi_0)} \end{array} \;\middle|\; x \right) \leq \gamma \right\} \tag{14}$$

where $J_\Psi\left(\tau\right)$ is the Jacobian of the transformation Ψ evaluated at τ.

Since $B_{\Lambda,\gamma}^{\Psi}(x)$ is a γ-credible region for $\psi = \Psi(\tau)$, then $\Psi^{-1} B_{\Lambda,\gamma}^{\Psi}(x)$ is a γ-credible region for τ. Let

$$[B_{\Lambda,\gamma}(x)] = \left\{ \Psi^{-1} B_{\Lambda,\gamma}^{\Psi}(x) : \Psi \in \mathcal{D}_{\mathcal{T},\mathcal{T}} \right\}$$

be the class of γ-credible regions for τ that arise via reparameterizations, when using the measure Λ to construct the credible regions. Each of the regions

$\Psi^{-1}B^{\Psi}_{\Lambda,\gamma}(x)$ is a plausible candidate as a γ-credible region for the parameter of interest. We have the following result.

Lemma 11. For $\Psi \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$ we have that $\Psi^{-1}B^{\Psi}_{\Lambda,\gamma}(x) = B_{\Lambda\circ\Psi,\gamma}(x)$ and so $[B_{\Lambda,\gamma}(x)] = \{B_{\Lambda\circ\Psi,\gamma}(x) : \Psi \in \mathcal{D}_{\mathcal{T},\mathcal{T}}\}$.
Proof: For $A \subset \mathcal{T}$, we have that $\Lambda \circ \Psi(A) = \Lambda(\Psi(A)) = \int_{\Psi(A)} \lambda(\tau)\, \nu_{\mathcal{T}}(d\tau) = \int_A \lambda(\Psi(\tau))J_{\Psi}(\tau)\ \nu_{\mathcal{T}}(d\tau)$ and so the density of $\Lambda\circ\Psi$ is $\lambda(\Psi(\tau))J_{\Psi}(\tau)$. Therefore, by Theorem 3 and (14), the result follows.

So it is equivalent to think of $[B_{\Lambda,\gamma}(x)]$ as containing all the γ-credible regions for τ obtained by minimizing the $\Lambda \circ \Psi$ content for some $\Psi \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$.

Now suppose that Λ is a probability measure on $\mathcal{T}$. Define $\Psi_{\Lambda} : \mathcal{T} \to [0,1]^k$ so that $\Psi_{\Lambda}(\tau) \sim \text{Uniform}\big([0,1]^k\big)$ when $\tau \sim \Lambda$, e.g., we can take Ψ_{Λ} to be the probability transform. Then, for probability measures Λ_1 and Λ_2, we can define $\Psi_{\Lambda_1,\Lambda_2} : \mathcal{T} \to \mathcal{T}$ by $\Psi_{\Lambda_1,\Lambda_2} = \Psi_{\Lambda_1} \circ \Psi_{\Lambda_2}^{-1}$ and thus $\Lambda_1 \circ \Psi_{\Lambda_1,\Lambda_2} = \Lambda_2$. Therefore, if $\Psi_{\Lambda_1,\Lambda_2} \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$, we have that $[B_{\Lambda_1,\gamma}(x)] = [B_{\Lambda_2,\gamma}(x)]$. We say that Λ_2 is obtained via a smooth reparameterization from Λ_1 when $\Psi_{\Lambda_1,\Lambda_2} \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$.

We have the following result.

Lemma 12. If $\Psi_{\Lambda}^{-1} \circ \Psi_{\Pi_{\Upsilon}} \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$ then $C_{\gamma}(x) \in [B_{\Lambda,\gamma}(x)]$.
Proof: This follows from Corollary 5.

Note that, when $\Psi_{\Pi_{\Upsilon}}, \Psi_{\Lambda}$ are the respective probability transforms, λ is continuous and positive and π_{Υ} is positive and continuous, then by the inverse function theorem, we must have that $\Psi_* = \Psi_{\Lambda}^{-1} \circ \Psi_{\Pi_{\Upsilon}} \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$ and $J_{\Psi_*}(\tau) = \pi_{\Upsilon}(\tau)/\lambda(\Psi_*(\tau))$. So, in very general circumstances, relative surprise regions arise via a reparameterization of the original problem, when the aim is to find a credible region with smallest Λ measure.

One could argue that our goal should be to find $\Psi \in \mathcal{D}_{\mathcal{T},\mathcal{T}}$ such that $E_{\Pi}(P^2_{\theta}(\Upsilon(\theta) \in B_{\Lambda\circ\Psi,\gamma}(x)))$ is smallest. Suppose, however, that we can find such a Ψ. Then, in the sequence (θ_i, x_i, τ_i) for $i = 1, 2, \ldots$ of independent values from the joint distribution $\Pi \times P_{\theta} \times \Lambda \circ \Psi$, we have that $B_{\Lambda\circ\Psi,\gamma}$ satisfies, $\Upsilon(\theta_i) \in B_{\Lambda\circ\Psi,\gamma}(x_i)$ for a proportion γ of the i, minimizes the proportion of times $\tau_i \in B_{\Lambda\circ\Psi,\gamma}(x_i)$, and also minimizes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (P_{\theta_i}(\Upsilon(\theta_i) \in B_{\Lambda\circ\Psi,\gamma}(x)) - \gamma)^2 \tag{15}$$

among all exact γ-credible regions for τ.

We note several things about this result. First, it is not at all clear that minimizing the proportion of times $\tau_i \in B_{\Lambda\circ\Psi,\gamma}(x_i)$ is relevant to the particular application, because $\Lambda \circ \Psi$ may not place its mass primarily on appropriate values of τ. We have argued that the relevant measure to use here is Π_{Υ}, as this weights false values appropriately based on prior information. Second, while finding a Bayesian credible region, whose posterior content is close to its coverage probability seems like an appropriate goal, (15) shows that this really entails considering a doubly infinite sequence $(\theta_i, x_{ij}, \tau_i)$ for $i, j = 1, 2, \ldots$, where now

$x_{i1}, x_{i2}, \ldots$ is a sequence of i.i.d. values from P_{θ_i} and $B_{\Lambda\circ\Psi,\gamma}$ minimizes

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\Big(\lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^{m} I_{B_{\Lambda\circ\Psi,\gamma}(x_{ij})}(\Upsilon(\theta_i)) - \gamma\Big)^2. \tag{16}$$

The relevance of (16) for applications seems somewhat remote. So, for the reasons cited we conclude that, with no other ingredients to the problem, the most relevant element of $[B_{\Lambda,\gamma}(x)]$ to be used for inference is $C_\gamma(x)$. It also at least worth noting that the worst case behavior of an element of $[B_{\Lambda,\gamma}(x)]$ is bounded below by $C_\gamma(x)$ since

$$\sup\left\{E_\Pi(P^2_\theta(\Psi\circ\Upsilon(\theta)\in B_{\Lambda\circ\Psi,\gamma}(x))) : \Psi\in\mathcal{D}_{\mathcal{T},\mathcal{T}}\right\} \geq E_\Pi(P^2_\theta(\Upsilon(\theta)\in C_\gamma(x)).$$

So far we have restricted the discussion to probability measures Λ. Suppose, however, that Λ is a bounded measure on $\mathcal{T}$. Now it is immediate that for any positive constant b we have that $B_{b\Lambda,\gamma}(x) = B_{\Lambda,\gamma}(x)$. So we can take $b = 1/\Lambda(\mathcal{T})$ and simply treat Λ as a probability measure, as we get the same set of credible regions and $C_\gamma(x) \in [B_{\Lambda,\gamma}(x)]$. Also dividing (6) by $\Lambda(\mathcal{T})$ gives the same repeated sampling interpretation for this quantity.

For an unbounded measure Λ that assigns finite measure to compact subsets of $\mathcal{T}$, e.g., volume measure, we can consider an expanding sequence of compact sets $\mathcal{T}_k$ and let Λ_k be the restriction of Λ to $\mathcal{T}_k$, so that Λ_k is a bounded measure. Then, we have that $C_\gamma(x) \in [B_{\Lambda_k,\gamma}(x)]$ for each k, and $B_{\Lambda_k,\gamma}(x) \to B_{\Lambda,\gamma}(x)$ as $k\to\infty$ since $\lambda_k \to \lambda$. When (6), with Λ replaced by Λ_k and divided by $\Lambda(\mathcal{T}_k)$, converges as $k\to\infty$, then $B_{\Lambda,\gamma}$ has a limiting optimality property with respect to repeated sampling.

For any parameter of interest there are many γ-credible regions and we are forced to choose among them. The arguments presented in this section show that a γ-relative surprise region arises naturally from any hpd-like γ-credible region when we consider reparameterizations and the repeated sampling properties of regions with respect to coverage probabilities.

6 Conclusions

The consideration of repeated sampling behavior of inferences is quite natural. The proper Bayesian formulation of statistical problems seems almost antithetical to this. We have argued here, however, that consideration of repeated sampling properties represents a natural way to choose among various Bayesian procedures, provided we take into account prior information, i.e., we don't want to provide much weight at all to the repeated sampling behavior of a procedure at those θ values that receive little prior support.

The validity of any inference method in an application, depends essentially on the relevance of the particular ingredients chosen by the statistician. If, for example, the sampling model does not include a distribution for which the observed data is not surprising, then any inferences drawn using the model are at least suspect. Hence model checking is an intrinsic and necessary part of a

statistical analysis. Good repeated sampling properties for a procedure derived from a particular model are of no value if the model is wrong.

Similarly, checking for prior-data conflict is a necessary part of a statistical analysis. Not only does a prior that passes such checks inspire more confidence in any inferences drawn, but this seems like an essential step if we are going to take into account the prior when considering repeated sampling behavior of inference procedures. As we have shown in this paper, when we incorporate the prior into the assessment, then relative surprise inferences arise in a very natural way. In addition, we have shown that relative surprise inferences have optimal properties with respect to Bayes factors and arise also when we consider reparameterizations and focus on coverage probabilities.

When discussing repeated sampling properties of inferences, we have to decide on the appropriate sequence to consider. Traditional frequentist statistics considers a sequence of independent values $x_1, x_2, \ldots$ from P_{θ^*} where θ^* is the true value and then looks for procedures that do uniformly well for all θ^*. One argument against such a sequence is that imagining a sequence from a fixed P_{θ^*} seems somewhat unrealistic in practice. Further, we have argued here that considering all values of θ^* as equivalent is somewhat unrealistic in an application, where we typically have further information about the true value. Accordingly, we have argued in favor of considering sequences of the form $(\theta_i^*, x_i, \tau_i)$ for $i = 1, 2, \ldots$ of independent values from the joint distribution $\Pi^* \times P_{\theta^*} \times \Pi_\Upsilon$ or (θ_i, x_i, τ_i) for $i = 1, 2, \ldots$ of independent values from the joint distribution $\Pi \times P_{\theta^*} \times \Pi_\Upsilon$, where Π^* represents the distribution from which a true value is selected, and Π is the prior distribution selected by the statistician. In both cases relative surprise inferences arise when looking for optimal behavior. Just as with the sampling model, however, this entails that the prior makes sense, if our inferences are to have any validity. So we must check for prior-data conflict, e.g., we don't want to find evidence that the true value of the parameter lies in the tails of the prior.

7 References

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., Guttman, I. and Swartz, T. (2006). Optimality and computations for relative surprise inferences. Canadian Journal of Statistics, Vol. 34, No. 1, 113-129.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. Bayesian Analysis, Volume 1, Number 4, 893-914.

Evans, M. and Moshonov, H., (2007) Checking for prior-data conflict with hierarchically specified priors. Bayesian Statistics and its Applications,

eds. A.K. Upadhyay, U. Singh, D. Dey, Anamaya Publishers, New Delhi, 145-159.

Evans, M. and Shakhatreh, M. (2007) Consistency of Bayesian estimates for the sum of squared normal means with a normal prior. Technical Report 0607, Dept. of Statistics, University of Toronto.

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.

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.

Joshi, V.M. (1974) A note on the incompatibility of one-sided Bayes and frequency confidence intervals. Journal of the Royal Statistical Society, Series B, Vol. 36, No. 2, 237-242.

Plante, A. (1991) An inclusion-exclusion approach to the problem of absurd confidence statements; 1. Consistent exact confidence-interval estimation. Canadian Journal of Statistics, Vol. 19, No. 4, 389-397.

Rudin, W. (1974) Real And Complex Analysis, Second Edition. McGraw-Hill, New York.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals, Ann. Math. Statist, 30, 877-880.