Bayesian Analysis (2013) **8**, Number 3, pp. 569–590

Hypothesis Assessment and Inequalities for Bayes Factors and Relative Belief Ratios

Zeynep Baskurt*, and Michael Evans†

Abstract. We discuss the definition of a Bayes factor and develop some inequalities relevant to Bayesian inferences. An approach to hypothesis assessment based on the computation of a Bayes factor, a measure of the strength of the evidence given by the Bayes factor via a posterior probability, and the point where the Bayes factor is maximized is recommended. It is also recommended that the *a priori* properties of a Bayes factor be considered to assess possible bias inherent in the Bayes factor. This methodology can be seen to deal with many of the issues and controversies associated with hypothesis assessment. We present an application to a two-way analysis.

Keywords: Bayes factors, relative belief ratios, strength of evidence, a priori bias.

1 Introduction

Bayes factors, as introduced by Jeffreys (1935, 1961), are commonly used in applications of statistics. Kass and Raftery (1995) and Robert, Chopin, and Rousseau (2009) contain detailed discussions of Bayes factors.

Suppose we have a sampling model $\{P_\theta : \theta \in \Theta\}$ on $\mathcal{X}$, and a prior Π on Θ. Let T denote a minimal sufficient statistic for $\{P_\theta : \theta \in \Theta\}$ and $\Pi(\cdot \,|\, T(x))$ denote the posterior of θ after observing data $x \in \mathcal{X}$. Then for a set $C \subset \Theta$, with $0 < \Pi(C) < 1$, the *Bayes factor in favor of* C is defined by

$$BF(C) = \frac{\Pi(C \,|\, T(x))}{1 - \Pi(C \,|\, T(x))} \Big/ \frac{\Pi(C)}{1 - \Pi(C)}.$$

Clearly $BF(C)$ is a measure of how beliefs in the true value being in C have changed from *a priori* to *a posteriori*. Alternatively, we can measure this change in belief by the *relative belief ratio* of C, namely, $RB(C) = \Pi(C \,|\, T(x))/\Pi(C)$. A relative belief ratio measures change in belief on the probability scale as opposed to the odds scale for the Bayes factor. While a Bayes factor is the multiplicative factor transforming the prior odds after observing the data, a relative belief ratio is the multiplicative factor transforming the prior probability. These measures are related as we have that

$$BF(C) = \frac{(1 - \Pi(C))RB(C)}{1 - \Pi(C)RB(C)}, \; RB(C) = \frac{BF(C)}{\Pi(C)BF(C) + 1 - \Pi(C)}, \tag{1}$$

*Department of Statistics, University of Toronto, Toronto, Canada, zeynep@utstat.utoronto.ca
†Department of Statistics, University of Toronto, Toronto, Canada, mevans@utstat.utoronto.ca

 DOI:10.1214/13-BA824

and $BF(C) = RB(C)/RB(C^c)$. If it is hypothesized that $\theta \in H_0 \subset \Theta$, then $BF(H_0)$ or $RB(H_0)$ can be used as an assessment as to what extent the observed data has changed our beliefs in the truth of H_0.

Both the Bayes factor and the relative belief ratio are not defined when $\Pi(C) = 0$. In Section 2 we will see that, when we have a characteristic of interest $\psi = \Psi(\theta)$ where $\Psi : \Theta \to \Psi$ (we don't distinguish between the function and its range to save notation), and $H_0 = \Psi^{-1}\{\psi_0\}$ with $\Pi(H_0) = 0$, we can define the Bayes factor and relative belief ratio of H_0 as limits and the limiting values are identical. This permits the assessment of a hypothesis $H_0 = \Psi^{-1}\{\psi_0\}$ via a Bayes factor without the need to modify the prior Π by placing positive prior mass on ψ_0. Furthermore, we will show that the common definition of a Bayes factor, obtained by placing positive prior mass on ψ_0, is equal to our limiting definition in many circumstances.

The approach to defining Bayes factors and relative belief ratios as limits is motivated by the use of continuous probability distributions which can imply that $\Pi(H_0) = 0$ simply because H_0 is a set of lower dimension and not because we have no belief that H_0 is true. We take the position that all continuous probability models are employed to approximate something that is essentially finite and thus discrete. For example, all observed variables are measured to finite accuracy and are bounded and we can never know the values of parameters to infinite accuracy.

To avoid paradoxes it is important that the essential finiteness of statistical applications be taken into account. For example, suppose that Π is absolutely continuous on Θ with respect to Lebesgue (volume) measure with density π. Of course, π can be changed on a set of Lebesgue measure 0 and still serve as a density, but note that this completely destroys the meaning of the approximation $\Pi(A(\theta_0)) \approx \pi(\theta_0)Vol(A(\theta_0))$ when $A(\theta_0)$ is a neighborhood of θ_0 with small volume. The correct interpretation of the relative values of densities requires that such an approximation hold and it is easy to attain this by requiring that $\pi(\theta_0) = \lim_{A(\theta_0)\to\{\theta_0\}} \Pi(A(\theta_0))/Vol(A(\theta_0))$, where $A(\theta_0) \to \{\theta_0\}$ means that $A(\theta_0)$ converges 'nicely' (see, for example, Rudin (1974), Chapter 8 for the definition) to $\{\theta_0\}$. In fact, whenever a version of π exists that is continuous at θ_0, then $\pi(\theta_0)$ is given by this limit. As an example of the kind of paradoxical behavior that can arise by allowing for arbitrary definitions of densities, suppose we stipulated that all densities for continuous distributions on Euclidean spaces are defined to be 0 whenever a response x has all rational coordinates. Certainly this is mathematically acceptable, but now all observed likelihoods are identically 0 and so useless for inference. As noted, however, this problem is simple to avoid by requiring that densities be defined as limits.

In this paper the value of the Bayes factor $BF(H_0)$ or relative belief ratio $RB(H_0)$ is to be taken as the statistical evidence that H_0 is true. So, for example, if $RB(H_0) > 1$, we have evidence that H_0 is true and the bigger $RB(H_0)$ is, the more evidence we have in favor of H_0. Similarly, if $RB(H_0) < 1$, we have evidence that H_0 is false and the smaller $RB(H_0)$ is, the more evidence we have against H_0. There are several concerns with this. First, it is reasonable to ask how strong this evidence is and so we propose an *a posteriori* measure of strength. In essence this corresponds to a calibration of $RB(H_0)$. Second, we need to be concerned with the impact of our *a priori* assignments. As is well-known, a diffuse prior can lead to large values of Bayes factors for hypotheses and we need to protect against this and other biases. We discuss all these issues in Sections 3 and 4 and in Section 5 present an example.

There are some close parallels between the use of Bayes factors to assess statistical evidence, and the approach to assessing statistical evidence via likelihood ratios as discussed in Royall (1997, 2000). More general definitions have been offered for Bayes factors when improper priors are employed. O'Hagan (1995) defines fractional Bayes factors and Berger and Perrichi (1996) define intrinsic Bayes factors. In this paper we restrict attention to proper priors although limiting results can often be obtained when considering a sequence of increasingly diffuse priors. Lavine and Schervish (1999) consider the coherency behavior of Bayes factors.

The problem of assessing a hypothesis H_0 as considered here is based on the choice of a single prior Π on Θ. We will argue in Section 2 that the appropriate prior on $H_0 = \Psi^{-1}\{\psi_0\}$ is the conditional prior on θ given that $\theta \in H_0 = \Psi^{-1}\{\psi_0\}$. While there seem to be logical reasons for this choice, it has been noted that this can lead to anomalous behavior for Bayes factors and so not all authors agree with this approach. For example, Johnson and Rossell (2010) argue that priors should be separately chosen for H_0 and H_0^c and show that these can be selected in such a way that the resultant Bayes factors are better behaved with respect to their convergence properties as the amount of data increases. At least part of the purpose of this paper, however, is to show that the Bayes factor based on the single prior Π can be used effectively for hypothesis assessment. In particular, for the case when H_0 is nested within Θ, we feel that this represents a very natural approach.

It should also be noted that the approach to hypothesis assessment that we are advocating does not rule out the possibility of using a prior that places a discrete mass π_0 on H_0. So, for example, we might employ a prior such as $\pi_0\Pi_0 + (1 - \pi_0)\Pi$ where Π_0 is a prior concentrated on H_0. We acknowledge that there are situations where such a prior seems natural. Part of our purpose here, however, is to show that employing such a discrete mass to form a mixture prior is not *necessary* to obtain a logical approach to hypothesis assessment. Where we might differ from a mixture prior approach, however, is in the choice of the prior Π_0. We argue in Section 2 that, rather than allowing Π_0 to be completely free, it is appropriate to require that Π_0 be the conditional prior $\Pi(\cdot \mid \psi_0)$ on H_0 induced by a Ψ satisfying $H_0 = \Psi^{-1}\{\psi_0\}$. In fact, we show that, when we restrict Π_0 in this way, the usual definition of a Bayes factor agrees with our definition as a limit based on Π alone. There are differences, however, between what we are advocating and a common approach based solely on computing a Bayes factor to assess a hypothesis. For instance we add an additional ingredient involving assessing the strength of the evidence, given by the Bayes factor, via a posterior probability. As discussed in Section 3, this additional ingredient corresponds to a calibration of a Bayes factor and allows us to avoid some problems that have arisen with their use.

2 The Definitions of Bayes Factors and Relative Belief Ratios

We now extend the definition of relative belief ratio and Bayes factor to the case where $\Pi(H_0) = 0$. We assume that P_θ has density f_θ with respect to support measure μ, Π has density π on Θ with respect to support measure ν and $\pi(\cdot \,|\, T(x))$ denotes the posterior density on Θ with respect to ν. Suppose we wish to assess $H_0 = \Psi^{-1}\{\psi_0\}$ for some parameter of interest $\psi = \Psi(\theta)$.

We will assume that all our spaces possess sufficient structure, and the various mappings we consider are sufficiently smooth, so that the support measures are volume measure on the respective spaces and, as discussed in Section 1, that any densities used are derived as limits of the ratios of measures of sets converging to points. The mathematical details can be found in Tjur (1974), where it is seen that we effectively require Riemann manifold structure for the various spaces considered, and we note that these restrictions are typically satisfied in statistical problems. For example, these requirements are always satisfied in the discrete case, as well as in the case of the commonly considered continuous statistical models. One appealing consequence of such restrictions is that we get simple formulas for marginal and conditional densities. For example, putting $J_\Psi(\theta) = (\det(d\Psi(\theta))(d\Psi(\theta))^t)^{-1/2}$ where $d\Psi$ is the differential of Ψ, and supposing $J_\Psi(\theta)$ is finite and positive for all θ, then the prior probability measure Π_Ψ has density, with respect to volume measure ν_Ψ on Ψ, given by

$$\pi_\Psi(\psi) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta) J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta), \tag{2}$$

where $\nu_{\Psi^{-1}\{\psi\}}$ is volume measure on $\Psi^{-1}\{\psi\}$. Furthermore, the conditional prior density of θ given $\Psi(\theta) = \psi$ is

$$\pi(\theta \,|\, \psi) = \pi(\theta) J_\Psi(\theta)/\pi_\Psi(\psi) \tag{3}$$

with respect to $\nu_{\Psi^{-1}\{\psi\}}$ on $\Psi^{-1}\{\psi\}$. A significant advantage with (2) and (3) is that there is no need to introduce coordinates, as is commonly done, for so-called nuisance parameters. In general, such coordinates do not exist.

If we let $T : \mathcal{X} \to \mathcal{T}$ denote a minimal sufficient statistic for $\{f_\theta : \theta \in \Theta\}$, then the density of T, with respect to volume measure $\mu_{\mathcal{T}}$ on $\mathcal{T}$, is given by $f_{\theta T}(t) = \int_{T^{-1}\{t\}} f_\theta(x) J_T(x)\, \mu_{T^{-1}\{t\}}(dx)$, where $\mu_{T^{-1}\{t\}}$ denotes volume on $T^{-1}\{t\}$. The prior predictive density, with respect to μ, of the data is given by $m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, \nu(d\theta)$ and the prior predictive density of T, with respect to $\mu_{\mathcal{T}}$, is $m_T(t) = \int_\Theta \pi(\theta) f_{\theta T}(t)\, \nu(d\theta)$ $= \int_{T^{-1}\{t\}} m(x) J_T(x)\, \mu_{T^{-1}\{t\}}(dx)$. This leads to a generalization of the Savage-Dickey ratio result, see Dickey and Lientz (1970), Dickey (1971), as we don't require coordinates for nuisance parameters.

Theorem 1. *(Savage-Dickey)* $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi) = m_T(T(x) \,|\, \psi)/m_T(T(x))$.

Proof: The posterior density of θ, with respect to support measure ν, is $\pi(\theta \,|\, T(x)) = \pi(\theta) f_{\theta T}(T(x))/m_T(T(x))$, and the posterior density of $\psi = \Psi(\theta)$, with respect to ν_Ψ, is

$\pi_\Psi(\psi \,|\, T(x)) = \int_{\Psi^{-1}\{\psi\}} (\pi(\theta) f_{\theta T}(T(x))/m_T(T(x))) J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta) = \pi_\Psi(\psi) \int_{\Psi^{-1}\{\psi\}} \pi(\theta \,|\, \psi)\ (f_{\theta T}(T(x))/m_T(T(x)))\, \nu_{\Psi^{-1}\{\psi\}}](d\theta) = \pi_\Psi(\psi) m_T(T(x) \,|\, \psi)/m_T(T(x))$ where $m_T(\cdot \,|\, \psi)$ is the conditional prior predictive density of T, given $\Psi(\theta) = \psi$.

As T is minimal sufficient, $m_T(T(x) \,|\, \psi)/m_T(T(x)) = m(x \,|\, \psi)/m(x)$.

Since $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ is the density of $\Pi_\Psi(\cdot \,|\, T(x))$ with respect to Π_Ψ,

$$\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi) = \lim_{\epsilon \to 0} \Pi_\Psi(C_\epsilon(\psi) \,|\, T(x))/\Pi_\Psi(C_\epsilon(\psi)) \tag{4}$$

whenever $C_\epsilon(\psi)$ converges nicely to $\{\psi\}$ as $\epsilon \to 0$ and all densities are continuous at ψ, e.g., $C_\epsilon(\psi)$ could be a ball of radius ϵ centered at ψ. So $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ is the limit of the relative belief ratios of sets converging nicely to ψ and, if $\Pi(\Psi^{-1}\{\psi\}) > 0$, then $\pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ gives the previous definition of a relative belief ratio for $\Psi^{-1}\{\psi\}$. As such, we refer to $RB(\psi) = \pi_\Psi(\psi \,|\, T(x))/\pi_\Psi(\psi)$ as the *relative belief ratio* of ψ.

From (4) and (1) we have $BF(C_\epsilon(\psi)) \to (1 - \Pi(\Psi^{-1}\{\psi\}))RB(\psi)/(1 - \Pi(\Psi^{-1}\{\psi\})\, RB(\psi))$ as $\epsilon \to 0$ and this equals $RB(\psi)$ if and only if $\Pi(\Psi^{-1}\{\psi\}) = 0$. So, in the continuous case, $RB(\psi)$ is a limit of Bayes factors with respect to Π and so can also be called the Bayes factor in favor of ψ with respect to Π. If, however, $\Pi(\Psi^{-1}\{\psi\}) > 0$, then $RB(\psi)$ is not a Bayes factor with respect to Π but is related to the Bayes factor through (1). The following example demonstrates another important context where the relative belief ratio and Bayes factor are identical.

Example 1. *Comparison with Jeffreys' Bayes Factor.*

Suppose now that $H_0 = \Psi^{-1}\{\psi_0\}$ and $\Pi(H_0) = 0$. A common approach in this situation, due to Jeffreys (1961), is to modify the prior Π to the mixture prior $\Pi_\gamma = \gamma\Pi_0 + (1-\gamma)\Pi$ where Π_0 is a probability measure on H_0 and $0 < \gamma < 1$ so $\Pi_\gamma(H_0) = \gamma$. Then, letting m_{0T} denote the prior predictive density of T under Π_0, we have that the Bayes factor and relative belief ratio under Π_γ are given by $BF_{\Pi_\gamma}(\psi_0) = m_{0T}(T(x))/m_T(T(x))$ and $RB_{\Pi_\gamma}(\psi_0) = \{m_{0T}(T(x))/m_T(T(x))\}/\{1-\gamma+\gamma m_{0T}(T(x))/m_T(T(x))\}$ respectively, and these are generally not equal. We now show, however, that in certain circumstances $BF_{\Pi_\gamma}(\psi_0) = RB(\psi_0)$ where $RB(\psi_0)$ is the relative belief ratio with respect to Π.

The following result generalizes Verdinelli and Wasserman (1995) as we don't require coordinates for nuisance parameters.

Theorem 2. (*Verdinelli-Wasserman*) When $H_0 = \Psi^{-1}\{\psi_0\}$ for some Ψ and ψ_0 and $\Pi(H_0) = 0$, then the Bayes factor in favor of H_0 with respect to Π_γ is

$$m_{0T}(T(x))/m_T(T(x)) = RB(\psi_0) E_{\Pi_0}\left(\pi(\theta \,|\, \psi_0, T(x))/\pi(\theta \,|\, \psi_0)\right) \tag{5}$$

where E_{Π_0} refers to expectation with respect to Π_0.

Proof: We have $m_{0T}(T(x))/m_T(T(x)) = RB(\psi_0) m_{0T}(T(x))/m_T(T(x) \,|\, \psi_0)$ by Theorem 1 and

$$\frac{m_{0T}(T(x))}{m_T(T(x) \,|\, \psi_0)} = \frac{\int_{\Psi^{-1}\{\psi_0\}} \pi_0(\theta) f_{\theta T}(T(x))\, \nu_{\Psi^{-1}\{\psi_0\}}(d\theta)}{\int_{\Psi^{-1}\{\psi_0\}} \pi(\theta \,|\, \psi_0) f_{\theta T}(T(x))\, \nu_{\Psi^{-1}\{\psi_0\}}(d\theta)},$$

so the result follows from (3).

We then have the following consequence, where $\Pi(\cdot\,|\,\psi_0)$ denotes the conditional prior obtained from Π by conditioning on $\Psi(\theta) = \psi_0$.

Corollary 3. If $\Pi_0 = \Pi(\cdot\,|\,\psi_0)$, then $BF_{\Pi_\gamma}(\psi_0) = RB(\psi_0)$.

Proof: Since $\pi_0(\theta) = \pi(\theta\,|\,\psi_0)$ we have $E_{\Pi_0}\left(\pi(\theta\,|\,\psi_0, T(x))/\pi(\theta\,|\,\psi_0)\right) = 1$ which establishes the result.

In general, (5) establishes the relationship between the Bayes factor when using the conditional prior $\Pi(\cdot\,|\,\psi_0)$ on H_0 and the Bayes factor when using the prior Π_0 on H_0. The adjustment is the expected value, with respect to Π_0, of the conditional relative belief ratio $\pi(\theta\,|\,\psi_0, T(x))/\pi(\theta\,|\,\psi_0)$ for $\theta \in H_0$, given H_0. This can also be written as $E_{\Pi(\cdot\,|\,\psi_0, T(x))}\left(\pi_0(\theta)/\pi(\theta\,|\,\psi_0)\right)$ and so measures the discrepancy between the conditional priors given H_0 under Π and Π_γ. So when π_0 is substantially different than $\pi(\cdot\,|\,\psi_0)$, we can expect a significant difference in the Bayes factors. To maintain consistency in the prior assignments, we require here that Π_0 equal $\Pi(\cdot\,|\,\psi_0)$ for some smooth Ψ and ψ_0. In the discrete case it seems clear that choosing Π_0 not equal to $\Pi(\cdot\,|\,\psi_0)$ is incorrect. Also, in the continuous case, Jeffreys' approach requires completely different modifications of Π to obtain Bayes factors for different values of ψ_0. By contrast $RB(\psi_0)$ is defined for every value ψ_0 without any modification of Π. As discussed in Section 1, however, restricting the prior on H_0 in this way is not something that all statisticians agree with.

Marin and Robert (2010) question the validity of the Savage-Dickey result due to the arbitrariness with which densities can be defined on sets of measure 0. We note, however, that densities for us are not arbitrary and must be defined as limits as described in Section 1. With this restriction, Theorems 1 and 2 are valid results with interpretational value for inference and play a role in the results of Section 4.

3 Evidential Interpretation of Bayes Factors and Relative Belief Ratios

A Bayes factor or relative belief ratio for $H_0 = \Psi^{-1}\{\psi_0\}$ measures how our beliefs in H_0 have changed after seeing the data. The degree to which our beliefs have changed can be taken as the statistical evidence that H_0 is true. For if $RB(\psi_0) > 1$, then the probability of ψ_0 has increased by the factor $RB(\psi_0)$ from *a priori* to *a posteriori* and we have evidence in favor of H_0. Furthermore, the larger $RB(\psi_0)$ is, the more evidence we have in favor of H_0. Conversely, if $RB(\psi_0) < 1$, then the probability of ψ_0 has decreased by the factor $RB(\psi_0)$ from *a priori* to *a posteriori*, we have evidence against H_0 and the smaller $RB(\psi_0)$ is, the more evidence we have against H_0.

This definition of evidence leads to a natural total preference ordering on Ψ, namely, ψ_1 is preferred to ψ_2 whenever $RB(\psi_1) \geqslant RB(\psi_2)$ as the observed data have led to an increase in belief for ψ_1 at least as large as that for ψ_2. This total ordering in turn leads to the estimate of the true value of ψ given by $\psi_{\mathrm{LRSE}}(x) = \arg\sup RB(\psi)$ (*least relative* *surprise estimate*) and to assessing the accuracy of this estimate by choosing $\gamma \in (0,1)$, and looking at the 'size' of the γ-credible region $C_\gamma(x) = \{\psi_0 : RB(\psi_0) \geqslant c_\gamma(x)\}$ where $c_\gamma(x) = \inf\{k : \Pi_\Psi(RB(\psi) > k \,|\, T(x)) \leqslant \gamma\}$. The form of the credible region is determined by the ordering for, if $RB(\psi_1) \geqslant RB(\psi_2)$ and $\psi_2 \in C_\gamma(x)$, then we must have $\psi_1 \in C_\gamma(x)$. Note that $C_{\gamma_1}(x) \subset C_{\gamma_2}(x)$ when $\gamma_1 \leqslant \gamma_2$ and $\psi_{\text{LRSE}}(x) \in C_\gamma(x)$ for each γ that leads to a nonempty set. Of course 'accuracy' is application dependent and so a large $C_\gamma(x)$ for one application may in fact be small for another.

We cannot categorically state that $RB(\psi_0)$ is *the* measure of statistical evidence for the truth of H_0, but we can look at the properties of this measure, and the associated inferences, to see if these are suitable and attractive. Perhaps the most attractive property is that the inferences are invariant under smooth reparameterizations. This follows from the fact that, if $\omega = \Omega(\psi)$ for some 1-1, smooth function Ω, then $RB(\omega) = RB(\psi)$ as Jacobians cancel in the numerator and denominator. Furthermore, various optimality properties, in the class of all Bayesian inferences, have been established for $\psi_{\text{LRSE}}(x)$ and $C_\gamma(x)$ in Evans (1997), Evans, Guttman and Swartz (2006), Evans and Shakhatreh (2008) and Evans and Jang (2011c). For example, it is proved that among all subsets $B \subset \Psi$ satisfying $\Pi_\Psi(B \,|\, x) \geqslant \gamma$, both $BF(B)$ and $RB(B)$ are maximized by $B = C_\gamma(x)$ and these maximized values are always bounded below by 1 (a property not possessed by other rules for forming credible regions). So $C_\gamma(x)$ maximizes the increase in belief from *a priori* to *a posteriori* among all γ-credible regions and, as such, $C_\gamma(x)$ is letting the data speak the loudest among all such credible regions. Also, $C_\gamma(x)$ minimizes the *a priori* probability of covering a false value and this probability is always bounded above by γ when $\Pi_\Psi(C_\gamma(x) \,|\, x) = \gamma$. In this case, γ is also the prior probability that $C_\gamma(x)$ contains the true value, implying that $C_\gamma(x)$ is unbiased. The estimate $\psi_{\text{LRSE}}(x)$ is unbiased with respect to a general family of loss functions and, is either a Bayes rule or a limit of Bayes rules with respect to a simple loss function based on the prior.

While these results support the use of these inferences, we now consider additional properties of $RB(\psi_0)$ as a measure of the evidence in favor of H_0. The invariance of $RB(\psi_0)$ is certainly a necessary property of any measure of statistical evidence. Also, we have the following simple result.

Theorem 4. $RB(\psi_0) = E_{\Pi(\cdot\,|\,\psi_0)}(RB(\theta))$.

Proof: First we note that $RB(\theta) = f_{\theta T}(T(x))/m_T(T(x))$ and using (2) and (3), we have that

$$\begin{aligned} RB(\psi_0) &= \frac{\int_{\Psi^{-1}\{\psi_0\}} \pi(\theta) J_\Psi(\theta) (f_{\theta T}(T(x))/m_T(T(x)))\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)}{\int_{\Psi^{-1}\{\psi_0\}} \pi(\theta) J_\Psi(\theta)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta)} \\ &= \int_{\Psi^{-1}\{\psi_0\}} RB(\theta)\pi(\theta \,|\, \psi)\, \nu_{\Psi^{-1}\{\psi\}}(d\theta) = E_{\Pi(\cdot\,|\,\psi_0)}(RB(\theta)). \end{aligned}$$

This says that evidence in favor of H_0 is obtained by averaging, using the conditional prior given that H_0 is true, the evidence in favor of each value of the full parameter that makes H_0 true. Furthermore, based on the asymptotics of the posterior density, under quite general conditions, we will have that $RB(\psi_0) \to 0$ when H_0 is false and, in the continuous case, $RB(\psi_0) \to \infty$ when H_0 is true, as we increase the amount of data.

It is also reasonable to ask how strong the evidence given by $RB(\psi_0)$ is in a particular context. For example, how strong is the evidence in favor of H_0 when $RB(\psi_0) = 20$? So far we only know that this is more evidence in favor than when $RB(\psi_0) = 17$. Using a measure of evidence, without some assessment of the strength, does not seem appropriate as indeed different data sets can provide different amounts of evidence and with different strengths.

One way to answer this is to propose a scale on which evidence can be assessed. For example, Kass and Raftery (1995) discuss using a scale due to Jeffreys (1961). It is difficult, however, to see how such a universal scale is to be determined and, in any case, this does not tell us how well the data support alternatives to H_0. For example, when $H_0 = \Psi^{-1}\{\psi_0\}$ we can consider the relative belief ratios for other values of ψ. If a relative belief ratio for a $\psi \neq \psi_0$ is much larger than that for ψ_0, then it seems reasonable to at least express some doubt as to the strength of the evidence in favour of H_0. Note that we are proposing to compare $RB(\psi_0)$ to each of the possible values of $RB(\psi)$ as part of assessing H_0, as opposed to just considering the hypothesis testing problem H_0 versus H_0^c (see, however, Example 2). This is in agreement with a commonly held view as expressed, for example, in Gelman, Carlin, Stern and Rubin (2004), that hypothesis assessment is different than hypothesis testing as discussed, for example, in Berger and Delampady (1987).

Perhaps the most obvious way to measure the strength of the evidence expressed by $RB(\psi_0)$ is via the posterior tail probability

$$\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)\right). \tag{6}$$

This is the posterior probability that the true value of ψ has a relative belief ratio no greater than $RB(\psi_0)$. It is worth remarking that $C_\gamma(x) = \{\psi_0 : \Pi_\Psi(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)) \geqslant 1-\gamma\}$ and $\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)\right) = 1 - \inf\{\gamma : \psi_0 \in C_\gamma(x)\}$ so our measure of accuracy for estimation and our measure of strength for hypothesis assessment are intimately related. We now note that the interpretation of (6) depends on whether we have evidence against H_0 or evidence for H_0 and derive some relevant inequalities.

If $RB(\psi_0) < 1$, so that we have evidence against H_0, then a small value of (6) says there is a large posterior probability that the true value has a relative belief ratio greater than $RB(\psi_0)$. As such, this suggests that the evidence against H_0 is strong. We also have the following inequalities relevant to this case.

Theorem 5. When $RB(\psi_0) < 1$, then

$$\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)\right) \leqslant RB(\psi_0) \tag{7}$$

and $RB(RB(\psi) > RB(\psi_0)) > RB(\psi_0)$.

Proof: We have that

$$\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)\right) = \int_{\{RB(\psi) \leqslant RB(\psi_0)\}} RB(\psi)\pi_\Psi(\psi)\,\nu_\Psi(d\psi)$$
$$\leqslant \int_{\{RB(\psi) \leqslant RB(\psi_0)\}} RB(\psi_0)\pi_\Psi(\psi)\,\nu_\Psi(d\psi) = RB(\psi_0)\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0)\right)$$

which establishes (7). Furthermore, we have that

$$RB(\psi_0)\Pi_\Psi\left(RB(\psi) > RB(\psi_0)\right) = \int_{\{RB(\psi) > RB(\psi_0)\}} RB(\psi_0)\pi_\Psi(\psi)\,\nu_\Psi(d\psi)$$
$$\leqslant \int_{\{RB(\psi) > RB(\psi_0)\}} RB(\psi)\pi_\Psi(\psi)\,\nu_\Psi(d\psi) = \Pi_\Psi\left(RB(\psi) > RB(\psi_0) \,|\, T(x)\right)$$

with equality if and only if $\Pi_\Psi\left(RB(\psi) > RB(\psi_0)\right) = 0$. So equality will occur if and only if $\psi_0 = \hat{\psi}_{\text{LRSE}}(x)$. It is established in Evans and Shakhatreh (2008) that $RB(\hat{\psi}_{\text{LRSE}}(x)) \geqslant 1$ and since $RB(\psi_0) < 1$ by hypothesis, the inequality is strict. Dividing both sides of the inequality by $\Pi_\Psi\left(RB(\psi) > RB(\psi_0)\right)$ proves $RB(RB(\psi) > RB(\psi_0)) > RB(\psi_0)$.

We see that (7) says that, whenever we have a small value of $RB(\psi_0)$, then we have strong evidence against H_0 and, in fact, there is no need to compute (6). The inequality $RB(RB(\psi) > RB(\psi_0)) > RB(\psi_0)$ says that when we iterate relative belief, the evidence that the true value is in $\{\psi : RB(\psi) > RB(\psi_0)\}$ is strictly greater than the evidence that ψ_0 is the true value, when we have evidence against ψ_0 being true.

As previously discussed, when $\Pi(\Psi^{-1}\{\psi\}) = 0$, we can also interpret $RB(\psi_0)$ as the Bayes factor with respect to Π in favour of H_0 and so (6) is also an *a posteriori* measure of the strength of the Bayes factor. When ψ has a discrete distribution, we have the following result where we interpret $BF(\psi)$ in the obvious way.

Corollary 6. If Π_Ψ is discrete, then $\Pi_\Psi\left(BF(\psi) \leqslant BF(\psi_0) \,|\, T(x)\right) \leqslant BF(\psi_0) \times E_\Pi(\{1 + \pi_\Psi(\Psi(\theta))(BF(\psi_0) - 1)\}^{-1})$, the upper bound is finite and converges to 0 as $BF(\psi_0) \to 0$.

Proof: Using (1) we have that $BF(\psi) \leqslant BF(\psi_0)$ if and only if $RB(\psi) \leqslant BF(\psi_0)/\{1 + \pi_\Psi(\psi)(BF(\psi_0) - 1)\}$ and, as in the proof of Theorem 5, this implies the inequality. Also $1 + \pi_\Psi(\psi)(BF(\psi_0) - 1) \geqslant 1 + \max_\psi \pi_\Psi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) \leqslant 1$ and $1 + \pi_\Psi(\psi)(BF(\psi_0) - 1) \geqslant 1 + \min_\psi \pi_\Psi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) > 1$ which completes the proof.

So we see that a small value of $BF(\psi_0)$ is, in both the discrete and continuous case, strong evidence against H_0.

If $RB(\psi_0) > 1$, so that we have evidence in favor of H_0, and (6) is small, then there is a large posterior probability that the true value of ψ has an even larger relative belief ratio and so this evidence in favor of H_0 does not seem strong. Alternatively, large values of (6), when $RB(\psi_0) > 1$, indicate that we have strong evidence in favor of H_0

as $\{\psi : RB(\psi) \leqslant RB(\psi_0)\}$ contains the true value with high posterior probability and, based on the preference ordering, ψ_0 is the best estimate in this set.

While (7) always holds it is irrelevant when $RB(\psi_0) > 1$. Markov's inequality implies $\Pi_\Psi(RB(\psi) > RB(\psi_0) \,|\, T(x)) \leqslant E_{\Pi_\Psi(\cdot \,|\, T(x))}(RB(\psi))/RB(\psi_0)$ but this does not imply that large values of $RB(\psi_0)$ are strong evidence in favor of H_0. In particular, in many situations the upper bound never gets small because of the relationship between $RB(\psi_0)$ and $\Pi_\Psi(\cdot \,|\, T(x))$. We do, however, have the following result.

Theorem 7. When $RB(\psi_0) > 1$, then $RB(RB(\psi) < RB(\psi_0)) < RB(\psi_0)$.

Proof: As in the proof of Theorem 6 we have that $\Pi_\Psi(RB(\psi) < RB(\psi_0) \,|\, T(x)) \leqslant RB(\psi_0)\Pi_\Psi(RB(\psi) < RB(\psi_0))$ and equality occurs if and only if $\Pi_\Psi(RB(\psi) < RB(\psi_0))$ $= 0$ which implies $\Pi_\Psi(RB(\psi) < RB(\psi_0) \,|\, T(x)) = 0$ which implies $1 = \Pi_\Psi(RB(\psi) \geqslant RB(\psi_0) \,|\, T(x)) = \int_{\{RB(\psi) \geqslant RB(\psi_0)\}} RB(\psi)\pi_\Psi(\psi)\,\nu_\Psi(d\psi) \geqslant RB(\psi_0) > 1$ which is a contradiction.

So the evidence that the true value is in $\{\psi : RB(\psi) < RB(\psi_0)\}$ is strictly less than the evidence that ψ_0 is the true value, when we have evidence in favor of ψ_0 being true.

Consider the following example concerned with comparing H_0 to H_0^c.

Example 2. *Binary* Ψ.

Suppose $\Psi(\theta) = I_{H_0}$ and $0 < \Pi(H_0) < 1$. We have $\Pi_\Psi\left(BF(\psi) \leqslant BF(H_0) \,|\, T(x)\right) = \Pi(H_0 \,|\, T(x))$ when $BF(H_0) \leqslant 1$, and $\Pi_\Psi(BF(\psi) \leqslant BF(H_0) \,|\, T(x)) = 1$ otherwise, while $\Pi_\Psi(RB(\psi) \leqslant RB(H_0) \,|\, T(x)) = \Pi(H_0 \,|\, T(x))$ when $BF(H_0) \leqslant 1$, and $\Pi_\Psi(RB(\psi) \leqslant RB(H_0) \,|\, T(x)) = 1$ otherwise. So these give the same assessment of strength. This says that in the binary case $BF(H_0) < 1$ or $RB(H_0) < 1$ is strong evidence against H_0 only when $\Pi(H_0 \,|\, T(x))$ is small. By Corollary 6 and Theorem 5 this will be the case whenever $BF(H_0)$ or $RB(H_0)$ are suitably small. Furthermore, large values of $BF(H_0)$ or $RB(H_0)$ are always deemed to be strong evidence in favour of H_0 in this case. So if one has determined in an application that comparing H_0 to H_0^c is the appropriate approach, as opposed to comparing the hypothesized value of the parameter of interest to each of its alternative values, then (6) leads to the usual answers.

The interpretation of evidence in favor of H_0 is somewhat more involved than evidence against H_0 and the following example illustrates this.

Example 3. *Location normal.*

Suppose we have a sample $x = (x_1, \ldots, x_n)$ from a $N(\mu, 1)$ distribution, where $\mu \in R^1$ is unknown, so $T(x) = \bar{x}$, we take $\mu \sim N(0, \tau^2), \Psi(\mu) = \mu$, and we want to assess $H_0 : \mu = 0$. We have that

$$RB(0) = (1 + n\tau^2)^{1/2} \exp\{-n(1 + 1/n\tau^2)^{-1}\bar{x}^2/2\} \tag{8}$$

and

$$\begin{aligned}
&\Pi_\Psi\left(RB(\mu) \leqslant RB(0) \,|\, T(x)\right) \\
&= 1 - \Phi((1 + 1/n\tau^2)^{1/2}(|\sqrt{n}\bar{x}| + (n\tau^2 + 1)^{-1}\sqrt{n}\bar{x})) \\
&+ \Phi((1 + 1/n\tau^2)^{1/2}(-|\sqrt{n}\bar{x}| + (n\tau^2 + 1)^{-1}\sqrt{n}\bar{x})). \qquad (9)
\end{aligned}$$

From (8) and (9) we have, for a fixed value of $\sqrt{n}\bar{x}$, that $RB(0) \to \infty$ and $\Pi_\Psi(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)) \to 2(1 - \Phi(|\sqrt{n}\bar{x}|)$ as $\tau^2 \to \infty$. This encapsulates the essence of the problem with the interpretation of large values of a relative belief ratio or Bayes factor as evidence in favor of H_0. For, as we make the prior more diffuse via $\tau^2 \to \infty$, the evidence in favor of H_0 becomes arbitrarily large. So we can bias the evidence *a priori* in favor of H_0 by choosing τ^2 very large. It is interesting to note, however, that $RB(0)$ is behaving correctly in this situation because, as τ^2 gets larger and larger, we are placing the bulk of the prior mass further and further away from $\bar{x}$. As such, $\mu = 0$ looks more and more like a plausible value when compared to the values where the prior mass is being allocated. On the other hand the strength of this evidence may prove to be very small depending on the value of $2(1 - \Phi(|\sqrt{n}\bar{x}|)$. Given that this bias is induced by the value of τ^2, we need to address this issue *a priori* and we will present an approach to doing this in Section 4.

We note that $2(1 - \Phi(|\sqrt{n}\bar{x}|)$ is the frequentist P-value for this problem. It is often remarked that a small value of $2(1 - \Phi(|\sqrt{n}\bar{x}|)$ and a large value of $RB(0)$, when τ^2 is large, present a paradox (*Lindley's paradox*) because large values of τ^2 are associated with noninformativity and we might expect classical frequentist methods and the Bayesian approach to then agree. But if we accept (6) as an appropriate measure of the strength of the evidence in favor of H_0, then the paradox disappears as we can have evidence in favor of H_0 while, at the same time, this evidence is not strong.

It also follows from (8) and (9) that, for a fixed value of $RB(0)$, (6) decreases to 0 as n or τ^2 grows. Basically this is saying that a higher standard is set for establishing that a fixed value of $RB(0)$ is strong evidence in favour of H_0, as we increase the amount of data or make the prior more diffuse.

It is instructive to consider the behavior of $RB(0)$ as $n \to \infty$. For this we have that

$$RB(0) \to \begin{cases} \infty & H_0 \text{ true} \\ 0 & H_0 \text{ false}, \end{cases}$$

$$\Pi_\Psi\left(RB(\mu) \leqslant RB(0) \,|\, T(x)\right) \to \begin{cases} U(0,1) & H_0 \text{ true} \\ 0 & H_0 \text{ false} \end{cases}$$

where $U(0, 1)$ denotes a uniform random variable on $(0, 1)$. So as the amount of data increases, $RB(0)$ correctly identifies whether H_0 is true or false and we are inevitably lead to strong evidence against H_0 when it is false. When H_0 is true, however, it is always the case that, while we will inevitably obtain evidence in favor of H_0, for some data sets this evidence will not be deemed strong, as other values of μ have larger relative belief ratios. We have, however, that $\mu_{\text{LRSE}}(x)$ converges to the true value of μ and so, in cases where we have evidence in favor of H_0 that is not deemed strong, we can simply look at $\mu_{\text{LRSE}}(x)$ to see if it differs from H_0 in any practical sense. Similarly, if we have evidence against H_0 we can look at $\mu_{\text{LRSE}}(x)$ to see if we have detected a deviation from H_0 that is of practical importance. This requires that we have a clear idea of the size of an important difference. It seems inevitable that this will have to be taken into account in any practical approach to hypothesis assessment. While we must always take into account practical significance when we have evidence against H_0, the value of (9) is telling us when it is necessary to do this when we have evidence in favor of H_0.

As a specific numerical example suppose that $n = 50, \tau^2 = 400$ and we observe $\sqrt{n}\bar{x} = 1.96$. Figure 1 is a plot of $RB(\mu)$. This gives $RB(0) = 20.72$ and Jeffreys scale says that this is strong evidence in favour of H_0. But (6) equals 0.05 and, as such, 20.72 is clearly not strong evidence in favour of H_0 as there is a large posterior probability that the true value has a larger relative belief ratio. In this case $\mu_{\text{LRSE}}(x) = 0.28$ and $RB(\mu_{\text{LRSE}}(x)) = 141.40$. Note that $\mu_{\text{LRSE}}(x) = 0.28$ cannot be interpreted as being close to 0 independent of the application context. If, however, the application dictates that a value of 0.28 is practically speaking close enough to 0 to be treated as 0, then it certainly seems reasonable to proceed as if H_0 is correct and this is supported by the value of the Bayes factor.

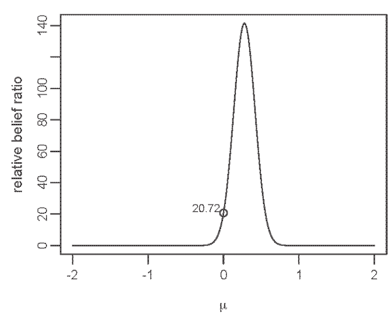

Figure 1: Plot of $RB(\mu)$ against μ when $n = 50, \tau^2 = 400$ and $\sqrt{n}\bar{x} = 1.96$ in Example 4.

Notice that, whenever ψ_0 is not true, then $RB(\psi_0) \to 0$ as the amount of data increases, and so (7) implies that $\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)\right) \to 0$ as well. As seen in Example 3, however, it is not always the case that $\Pi_\Psi(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)) \to 1$ when ψ_0 is true and this could be seen as anomalous. The following result, proved in the Appendix, shows that this is simply an artifact of continuity.

Theorem 8. Suppose that $\Theta = \{\theta_0, \ldots, \theta_k\}, \pi(\theta) > 0$ for each $\theta, H_0 = \Psi^{-1}\{\psi_0\}$ and $x = (x_1, \ldots, x_n)$ is a sample from f_θ. Then we have that $\Pi_\Psi(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)) \to 1$ as $n \to \infty$ whenever H_0 is true.

So if we think of continuous models as approximations to situations that are in reality finite, then we see that (6) may not be providing a good approximation. One possible solution is to use a metric d on Ψ and a distance δ such that $d(\psi, \psi') \leqslant \delta$ means that ψ and ψ' are practically indistinguishable. We can then use this to discretize Ψ and compute both the relative belief ratio for $H_0 = \{\psi : d(\psi, \psi_0) \leqslant \delta\}$ and its strength in this discretized version of the problem. Actually this can be easily implemented computationally and is implicit in our computations when we don't have an exact expression available for $RB(\psi)$. From a practical point-of-view, computing (6), and when this is small looking at $d(\psi_{\mathrm{LRSE}}(x), \psi_0)$ to see if a deviation of any practical importance has been detected, seems like a simple and effective solution to this problem.

To summarize, we are advocating that the evidence concerning the truth of a hypothesis $H_0 = \Psi^{-1}\{\psi_0\}$ be assessed by computing the relative belief ratio $RB(\psi_0)$ to determine if we have evidence for or against H_0. In conjunction with reporting $RB(\psi_0)$, we advocate reporting (6) as a measure of the strength of this evidence. It is important to note that (6) is not to be interpreted as any part of the evidence and, in particular, it is not a P-value. For if $RB(\psi_0) > 1$ and (6) is small, then we have weak evidence in favor of H_0, while if $RB(\psi_0) < 1$ and (6) is small, then we have strong evidence against H_0. It seems necessary to calibrate a Bayes factor in this way. We also advocate looking at $(\psi_{\mathrm{LRSE}}(x), RB(\psi_{\mathrm{LRSE}}(x)))$ as part of hypothesis assessment. The value $RB(\psi_{\mathrm{LRSE}}(x))$ tells us the maximum increase in belief for any value of ψ. If $RB(\psi_0) < 1$, and (6) is small, then the value of $\psi_{\mathrm{LRSE}}(x)$ gives an indication of whether or not we have detected a deviation from H_0 of practical significance. Similarly, if $RB(\psi_0) > 1$ and (6) is not high, then the value $\psi_{\mathrm{LRSE}}(x)$ gives us an indication of whether or not we truly do not have strong evidence or this is just a continuous scale effect. In general, it seems that the assessment of a hypothesis requires more than the computation of a single number.

It is clear that $RB(\psi_0)$ could be considered as a standardized integrated likelihood. But multiplying $RB(\psi_0)$ by a positive constant, as we can do with a likelihood, destroys its interpretation as a relative belief ratio, and thus its role as a measure of the evidence that H_0 is true, and we lose the various inequalities we have derived. Also, we have that $RB(\psi_0) \leqslant \sup_{\theta \in \Psi^{-1}\{\psi_0\}} f_{\theta T}(t)/m_T(T(x))$ which is a standardized profile likelihood at ψ_0. So the standardized profile likelihood also has an evidential interpretation as part of an upper bound on (6) although the standardized integrated likelihood gives a sharper bound. This can be interpreted as saying the integrated likelihood contains more relevant information concerning H_0 than the profile likelihood. This provides support for the use of integrated likelihoods over profile likelihoods as discussed in Berger, Liseo, and Wolpert (1999). Aitkin (2010) proposes to use something like (6) as a Bayesian P-value but based on the likelihood. We emphasize that (6) is not to be interpreted as a P-value.

4 Relative Belief Ratios A Priori

We now consider the *a priori* behavior of the relative belief ratio. First we follow Royall (2000) and consider the prior probability of getting a small value of $RB(\psi_0)$ when H_0 is true, as we know that this would be misleading evidence. We have the following result, where M_T denotes the prior predictive measure of the minimal sufficient statistic T.

Theorem 9. The prior probability that $RB(\psi_0) \leqslant q$, given that H_0 is true, is bounded above by q, namely,

$$M_T(\, m_T(t \,|\, \psi_0)/m_T(t) \leqslant q \mid \psi_0) \leqslant q. \tag{10}$$

Proof: Using Theorem 1 the prior probability that $RB(\psi_0) \leqslant q$ is given by

$$\begin{aligned}
&\Pi \times P_\theta \left(\frac{\pi_\Psi(\psi_0 \,|\, T(X))}{\pi_\Psi(\psi_0)} \leqslant q \,\middle|\, \psi_0 \right) = \Pi \times P_\theta \left(\frac{m_T(T(X) \,|\, \psi_0)}{m_T(T(X))} \leqslant q \,\middle|\, \psi_0 \right) \\
&= \int_{\left\{ \frac{m_T(t \,|\, \psi_0)}{m_T(t)} \leqslant q \right\}} m_T(t \,|\, \psi_0)\, \mu_{\mathcal{T}}(dt) \leqslant \int_{\left\{ \frac{m_T(t \,|\, \psi_0)}{m_T(t)} \leqslant q \right\}} q m_T(t)\, \mu_{\mathcal{T}}(dt) \leqslant q.
\end{aligned}$$

So Theorem 9 tells us that, *a priori*, the relative belief ratio for H_0 is unlikely to be small when H_0 is true.

Theorem 9 is concerned with $RB(\psi_0)$ providing misleading evidence when H_0 is true. Again following Royall (2000), we also need to be concerned with the prior probability that $RB(\psi_0)$ is large when H_0 is false, namely, when $\psi_0 \neq \psi_{\text{true}}$. For this we consider the behavior of the ratio $RB(\psi_0)$ when ψ_0 is a false value, as discussed in Evans and Shakhatreh (2008), namely, we calculate the prior probability that $RB(\psi_0) \geqslant q$ when $\theta \sim \Pi(\cdot \,|\, \psi_{\text{true}}), x \sim P_\theta$ and $\psi_0 \sim \Pi_\Psi$ independently of (ψ_{true}, x). So here ψ_0 is a false value in the generalized sense that it has no connection with the true value of the parameter and the data. We have the following result.

Theorem 10. The prior probability that $RB(\psi_0) \geqslant q$, when $\theta \sim \Pi(\cdot \,|\, \psi_0), x \sim P_\theta$ and $\psi_0 \sim \Pi_\Psi$ independently of (θ, x), is bounded above by $1/q$.

Proof: We have that this prior probability equals

$$\begin{aligned}
&\Pi(\cdot \,|\, \psi_{\text{true}}) \times P_\theta \times \Pi_\Psi \left(\frac{\pi_\Psi(\psi_0 \,|\, T(x))}{\pi_\Psi(\psi_0)} \geqslant q \right) \\
&= M_T(\cdot \,|\, \psi_{\text{true}}) \times \Pi_\Psi \left(\frac{\pi_\Psi(\psi_0 \,|\, t)}{\pi_\Psi(\psi_0)} \geqslant q \right) \\
&= \int_{\mathcal{T}} \int_{\{\pi_\Psi(\psi_0 \,|\, t)/\pi_\Psi(\psi_0) \geqslant q\}} \pi_\Psi(\psi_0) m_T(t \,|\, \psi_{\text{true}})\, \nu_\Psi(d\psi_0)\, \mu_{\mathcal{T}}(dt) \\
&\leqslant \frac{1}{q} \int_{\mathcal{T}} \int_{\{\pi_\Psi(\psi_0 \,|\, t)/\pi_\Psi(\psi_0) \geqslant q\}} \pi_\Psi(\psi_0 \,|\, t) m_T(t \,|\, \psi_{\text{true}})\, \nu_\Psi(d\psi_0)\, \mu_{\mathcal{T}}(dt) \leqslant \frac{1}{q}.
\end{aligned}$$

Theorem 10 says that it is *a priori* very unlikely that $RB(\psi_0)$ will be large when ψ_0 is a false value. This reinforces the interpretation that large values of $RB(\psi_0)$ are evidence in favor of H_0.

In Example 3, if we fix $\sqrt{n}\bar{x}$, then $RB(\mu) \to \infty$ for every μ as $\tau^2 \to \infty$. This suggests that in general it is possible that a prior induces bias into an analysis by making it more likely to find evidence in favor of H_0 or possibly even against H_0. The calibration of

$RB(\psi_0)$ given by (6) is seen to take account of the actual size of $RB(\psi_0)$ when we have either evidence for or against H_0. This doesn't tell us, however, if the prior induces an *a priori* bias either for or against H_0. It seems natural to assess the bias against H_0 in the prior by

$$M_T(\, m_T(t \,|\, \psi_0)/m_T(t) \leqslant 1 \,|\, \psi_0). \tag{11}$$

If (11) is large, then this tells us that we have *a priori* little chance of detecting evidence in favor of H_0 when H_0 is true. We can also use (11) as a design tool by choosing the sample size to make (11) small. Similarly, we can assess the bias in favor of H_0 in the prior by the probabilities

$$M_T(\, m_T(t \,|\, \psi_0)/m_T(t) \leqslant 1 \,|\, \psi_*) \tag{12}$$

for various values of $\psi_* \neq \psi_0$ that represent practically significant deviations from ψ_0. If these probabilities are small, then this indicates that the prior is biasing the evidence in favor of ψ_0. Again we can use this as a design tool by choosing the sample size so that (12) is large.

We illustrate this via an example.

Example 4 *Continuation of Example 3.*

From (8) we see that $RB(0) \to 1$ as $\tau^2 \to 0$. So attempting to bias the evidence in favor of H_0 by choosing a τ^2 that concentrates the prior too much about 0, simply leads to inconclusive evidence about H_0. Furthermore, choosing τ^2 small is not a good strategy as we have to be concerned with the possibility of prior-data conflict, namely, there is evidence that the true value is in the tails of the prior, as this leads to doubts as to whether or not the prior is a sensible choice. How to check for prior-data conflict, and what to do about it when it is encountered, is discussed in Evans and Moshonov (2006) and Evans and Jang (2011a, 2011b). Checking for prior-data conflict, along with model checking, can be seen as a necessary part of a statistical analysis, at least if we want subsequent inferences to be credible with a broad audience.

The more serious issue with bias arises when, in an attempt to be conservative, we choose τ^2 to be large, as this will produce large values for Bayes factors. Of course, this assigns prior mass to values that we know are not plausible and we could simply dismiss this as bad modelling. But even when we have chosen τ^2 to reflect what is known about μ, we have to worry about the biasing effect.

We have that the conditional prior predictive $M_T(\cdot \mid \mu)$ is given by $\bar{x} \,|\, \mu \sim N(\mu, 1/n)$. Putting $a_n = \{\max(0, (1 + 1/n\tau^2) \log((1 + n\tau^2))\}^{1/2}$, then

$$M_T(\, RB(0) \leqslant 1 \,|\, \mu) = 1 - \Phi(a_n - \sqrt{n}\mu) + \Phi(-a_n - \sqrt{n}\mu) \tag{13}$$

and, as $\tau^2 \to \infty$, (13) converges to 0 for any μ, reflecting bias in favor of H_0 when τ^2 is large and $\mu \neq 0$. In this case (11) equals $M_T(\, RB(0) \leqslant 1 \,|\, 0) = 2(1 - \Phi(a_n))$ and we have recorded several values in the first row of Table 1 when $n = 50$. We see that only when τ^2 is small is there any bias against H_0. In the subsequent rows of Table 1 we have recorded the values of (13) when H_0 is false and, of course, we want these to be large.

τ^2	0.04	0.10	0.20	0.40	1.00	2.00	400.00
$\mu = 0.0$	0.20	0.14	0.10	0.07	0.05	0.03	0.00
$\mu = 0.1$	0.31	0.24	0.19	0.15	0.10	0.08	0.01
$\mu = 0.2$	0.56	0.48	0.42	0.35	0.28	0.23	0.04
$\mu = 0.3$	0.79	0.74	0.69	0.63	0.55	0.48	0.15

Table 1: Values of $M_T(RB(0) \leqslant 1 \,|\, \mu)$ for various τ^2 and μ in Example 3 when $n = 50$.

We see that there is bias in favor of H_0 when τ^2 is large. Note that (13) converges to 1 as $\mu \to \pm\infty$.

For the specific numerical example in Example 3 we have $n = 50$ and $\tau^2 = 400$. So there is no *a priori* bias against H_0 but some bias for H_0. Recall that $RB(0) = 20.72$ is only weak evidence in favor *of* H_0 since (6) equals 0.05. Also we have that $\mu_{\text{LRSE}}(x) = 0.28$ and $M_T(\, m_T(t \,|\, 0)/m_T(t) \leqslant 1 \,|\, 0.28) = 0.12$ which suggests that there is *a priori* bias in favor of H_0 at values like $\mu = 0.28$. So it is plausible to suspect that we have obtained weak evidence in favor of H_0 because of the bias entailed in the prior, at least if we consider a value like $\mu = 0.28$ as being practically different from 0.

It should also be noted that, as $n \to \infty$, then (13) converges to 1 when $\mu \neq 0$ and converges to 0 when $\mu = 0$. So in a situation where we can choose the sample size, after selecting the prior, we can select n to make (13) suitably large at selected values of $\mu \neq 0$ and also make (13) suitably small when $\mu = 0$.

Overall we believe that priors should be based on beliefs and elicited, but assessments for prior-data conflict are necessary and similarly, when hypothesis assessment is part of the analysis, we need to check for *a priori* bias. Of course, this should be done at the design stage but, even if it is done *post hoc,* this seems preferable to just ignoring the possibility that such biasing can occur. Happily the reporting of (6) as a posterior measure of the strength of the evidence, can help to warn us when problems exist.

Vlachos and Gelfand (2003) and Garcia-Donato and Chen (2005) propose a method for calibrating Bayes factors in the binary case, as discussed in Example 2. This involves computing tail probabilities based on the prior predictive distributions given by m_{H_0} and $m_{H_0^c}$.

5 Two-way Analysis of Variance

To illustrate the results of this paper we consider testing for no interaction in a two way ANOVA. Suppose we have two categorical factors A and B, and observe $x_{ijk} \sim N(\mu_{ij}, \nu^{-1})$ for $1 \leqslant i \leqslant a, 1 \leqslant j \leqslant b, 1 \leqslant k \leqslant n_{ij}$. A minimal sufficient statistic is given by $T(x) = (\bar{x}, s^2)$ where $\bar{x} \sim N_{ab}(\mu, \nu^{-1}D^{-1}(n))$, with $D(n) = \text{diag}(n_{11}, n_{12}, \ldots, n_{ab})$, independent of $(n_{\cdot\cdot} - ab)s^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(x_{ijk} - \bar{x}_{ij})^2 \sim \text{Gamma}_{\text{rate}}((n_{\cdot\cdot} - ab)/2, (2\nu)^{-1})$. Suppose we use the conjugate prior $\mu \,|\, \nu \sim N_{ab}(\mu_0, \nu^{-1}\Sigma_0)$, with $\Sigma_0 = \tau_0^2 I$, and $\nu \sim \text{Gamma}_{\text{rate}}(\alpha_0, \beta_0)$. Then we have that the posterior is given by $\mu \,|\, \nu, x \sim$

$N_{ab}(\mu_0(x), \nu^{-1}\Sigma_0(x)), \nu \,|\, x \sim \text{Gamma}_{\text{rate}}(\alpha_0(x), \beta_0(x))$ where

$$
\begin{aligned}
\mu_0(x) &= \Sigma_0(x)(D(n)\bar{x} + \tau_0^{-2}\mu_0),\\
\Sigma_0(x) &= (D(n) + \tau_0^{-2}I)^{-1},\\
\alpha_0(x) &= \alpha_0 + (n_{\cdot\cdot} - ab)/2,\\
\beta_0(x) &= \beta_0 + (\bar{x} - \mu_0)'(D^{-1}(n) + \tau_0^2 I)^{-1}(\bar{x} - \mu_0)/2 + (n_{\cdot\cdot} - ab)s^2/2.
\end{aligned}
$$

As is common in this situation, we test first for interactions between A and B and, if no interactions are found we proceed next to test for any main effects. For this we let $C_A = (c_{A1}\, c_{A2}\, \ldots\, c_{Aa}) \in R^{a\times a}, C_B = (c_{B1}\, c_{B2}\, \ldots\, c_{Bb}) \in R^{b\times b}$ denote contrast matrices (orthogonal and first column constant) for A and B, respectively, and put $C = C_A \otimes C_B = (c_{11}\, c_{12}\, \ldots\, c_{ab})$ where $c_{ij} = c_{Ai} \otimes c_{Bj}$ and $\otimes$ denotes direct product. The contrasts $\alpha = C'\mu$, where $\alpha_{ij} = c_{ij}'\mu$, have joint prior distribution $\alpha \,|\, \nu \sim N_{ab}(C'\mu_0, \nu^{-1}C'\Sigma_0 C) = N_{ab}(C'\mu_0, \nu^{-1}\Sigma_0)$, since C is orthogonal, and posterior distribution $\alpha \,|\, \nu, y \sim N_{ab}(C'\mu_0(y), \nu^{-1}C'\Sigma_0(x)C)$. From this we deduce that the marginal prior and posterior distributions of the contrasts are given by

$$
\begin{aligned}
\alpha &\sim \text{Student}_{ab}(2\alpha_0, C'\mu_0, (\beta_0/\alpha_0)C'\Sigma_0 C),\\
\alpha \,|\, x &\sim \text{Student}_{ab}(2\alpha_0(x), C'\mu_0(x), (\beta_0(x)/\alpha_0(x))C'\Sigma_0(x)C),
\end{aligned} \tag{14}
$$

where we say $w \sim \text{Student}_k(\lambda, m, M)$ with $m \in R^k$ and $M \in R^{k\times k}$ positive definite, when w has density

$$
\frac{\Gamma((\lambda+k)/2)}{\Gamma(\lambda/2)\Gamma^k(1/2)}(\det(M))^{-1/2}(1 + (w-m)'M^{-1}(w-m)/\lambda)^{-(\lambda+k)/2}\lambda^{-k/2}
$$

on R^k. Recall that, if $w \sim \text{Student}_k(\lambda, m, M)$ then, for distinct i_j with $1 \leqslant j \leqslant l \leqslant k$, we have that $(w_{i_1}, \ldots, w_{i_l}) \sim \text{Student}_l(\lambda, m(i_1, \ldots, i_l), M(i_1, \ldots, i_l))$ where $m(i_1, \ldots, i_l)$ and $M(i_1, \ldots, i_l)$ are formed by taking the elements of m and M as specified by $(i_1, \ldots, i_l)$.

We have that no interactions exist between A and B if and only if $\alpha_{ij} = 0$ for all $i > 1, j > 1$. So to assess the hypothesis H_0, we set $\psi = \Psi(\mu, \nu^{-1}) = (\alpha_{22}, \alpha_{23}, \ldots, \alpha_{ab}) \in R^{(a-1)(b-1)}$ and then $H_0 = \Psi^{-1}\{0\}$. From (14), and the marginalization property of Student distributions, we get an exact expression for $RB(0)$ and we can compute $\Pi_\Psi\,(RB(\psi) \leqslant RB(0) \,|\, T(x))$ by simulation.

To assess the *a priori* bias against H_0 based on a given prior, we need to compute $M_T(\,RB(0) \leqslant 1 \,|\, \alpha_{ij}$ for all $i > 1, j > 1)$. For this we need to be able to generate $T(x) = (\bar{x}, s^2)$ from the conditional prior predictive $M_T(\cdot \,|\, \alpha_{ij}$ for all $i > 1, j > 1)$. This is easily accomplished by generating (μ, ν) from the conditional prior given α_{ij} for all $i > 1, j > 1$, and then generating $\bar{x} \sim N_{ab}(\mu, \nu^{-1}D^{-1}(n))$ independent of $(n_{\cdot\cdot} - ab)s^2 \sim \text{Gamma}_{\text{rate}}((n_{\cdot\cdot} - ab)/2, (2\nu)^{-1})$. For this we need the conditional prior distribution of μ given ν and α_{ij} for all $i > 1, j > 1$. We have that $\alpha = C'\mu$ and $\mu = C\alpha$. As noted above, $\alpha \,|\, \nu \sim N_{ab}(C'\mu_0, \nu^{-1}C'\Sigma_0 C)$ and so we can generate μ from this conditional distribution by generating α from the conditional distribution obtained from

τ_0^2	0.01	0.05	0.08	0.10	0.50	5.00	10.00	100.00
$\alpha_{22}^2+\alpha_{32}^2=0.00$	0.53	0.28	0.26	0.24	0.16	0.10	0.09	0.06
$\alpha_{22}^2+\alpha_{32}^2=0.05$	0.74	0.58	0.51	0.47	0.30	0.19	0.15	0.11
$\alpha_{22}^2+\alpha_{32}^2=0.10$	0.85	0.77	0.71	0.67	0.46	0.27	0.24	0.17
$\alpha_{22}^2+\alpha_{32}^2=0.20$	0.95	0.93	0.91	0.90	0.74	0.50	0.43	0.31
$\alpha_{22}^2+\alpha_{32}^2=0.30$	0.98	0.98	0.98	0.97	0.90	0.69	0.62	0.45
$\alpha_{22}^2+\alpha_{32}^2=0.40$	0.99	0.99	0.99	0.99	0.97	0.84	0.78	0.61
$\alpha_{22}^2+\alpha_{32}^2=0.50$	1.00	1.00	1.00	1.00	0.99	0.93	0.89	0.73
$\alpha_{22}^2+\alpha_{32}^2=0.60$	1.00	1.00	1.00	1.00	1.00	0.97	0.95	0.84
$\alpha_{22}^2+\alpha_{32}^2=0.80$	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.95
$\alpha_{22}^2+\alpha_{32}^2=1.00$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99

Table 2: Values of $M_T(RB(0) \leqslant 1 \mid \alpha_{22}^2+\alpha_{32}^2=\delta)$ for various δ and τ_0^2 in two-way analysis.

the $N_{ab}(C'\mu_0, \nu^{-1}\Sigma_0)$ distribution by conditioning on α_{ij} for all $i>1, j>1$ and putting $\mu = C\alpha$. Note that the contrasts are *a priori* independent given ν so we just generate $\alpha_{i1} \mid \nu \sim N(c_{i1}'\mu_0, \nu^{-1}\tau_0^2)$ for $i=1,\ldots,a$, generate $\alpha_{1j} \mid \nu \sim N(c_{1j}'\mu_0, \nu^{-1}\tau_0^2)$ for $j=2,\ldots,b$, fix α_{ij} for all $i>1, j>1$ and set $\mu = C\alpha$.

As a specific numerical example suppose $a=3, b=2, (n_{11}, n_{12}, n_{21}, n_{22}, n_{31}, n_{32}) = (55, 50, 45, 43, 56, 48), \mu_0 = 0, \alpha_0 = 3, \beta_0 = 3$ and the contrasts are

$$C_A = \begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 0 & 2/\sqrt{6} \end{pmatrix}, C_B = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}.$$

Then the hypothesis H_0 of no interaction is equivalent to assessing whether or not $\psi = \Psi(\mu, \nu^{-1}) = (\alpha_{22}, \alpha_{32}) = (0, 0)$.

The prior for ν^{-1} has mean 1.5 and variance 2.25 and we now consider the choice of τ_0^2 as this has the primary effect on the *a priori* bias for H_0. In the first row of Table 2 we present the values of the *a priori* bias against H_0 for several values of τ_0^2 and see that the bias against H_0 is large when τ_0^2 is small. In the subsequent rows of Table 2 we present the bias in favor of H_0 when H_0 is false. For this we record $M_T(RB(0) \leqslant 1 \mid \alpha_{22}^2+\alpha_{32}^2=\delta)$ for various δ so we are averaging over all $(\alpha_{22}, \alpha_{32})$ that are the same distance from H_0. To generate $T(x) = (\bar{x}, s^2)$ from $M_T(\cdot \mid \alpha_{22}^2+\alpha_{32}^2=\delta) = \int_{\{\alpha_{22}^2+\alpha_{32}^2=\delta\}} M_T(\cdot \mid \alpha_{22}, \alpha_{32})\pi(\alpha_{22}, \alpha_{32} \mid \alpha_{22}^2+\alpha_{32}^2=\delta)\, d\alpha_{22}d\alpha_{32}$, we generate $(\alpha_{22}, \alpha_{32})$ from the conditional prior given $\alpha_{22}^2+\alpha_{32}^2=\delta$, and this is a uniform on the circle of radius $\delta^{1/2}$, and then generate from $M_T(\cdot \mid \alpha_{22}, \alpha_{32})$ as previously described. As expected, we see that there is bias in favor of H_0 only when τ_0^2 is large and we are concerned with detecting values of $(\alpha_{22}, \alpha_{32})$ that are close to H_0.

Suppose now that our prior beliefs lead us to choose $\tau_0^2 = 0.10$. In Table 3 we present some selected cases of assessing H_0 based on simulated data sets where the data is generated in such a way that we know there is no prior-data conflict. Recall that

Case	ψ_{true}	$RB(0)$	(6)	$\psi_{\text{LRSE}}(x)$	$RB(\psi_{\text{LRSE}}(x))$
1	$(0.00, 0.00)$	3.50	0.62	$(0.10, 0.11)$	5.10
2	$(0.00, 0.00)$	3.16	0.22	$(-0.10, -0.13)$	12.76
3	$(0.00, 0.00)$	5.11	0.55	$(-0.02, -0.12)$	8.62
4	$(0.00, 0.00)$	1.22	0.17	$(-0.14, -0.32)$	5.59
5	$(0.01, 0.01)$	3.07	0.55	$(-0.09, -0.16)$	4.94
6	$(0.01, 0.01)$	0.09	0.00	$(-0.22, 0.18)$	25.60
7	$(0.10, 0.10)$	0.02	0.00	$(0.36, 0.05)$	24.75
8	$(0.10, 0.10)$	1.96	0.35	$(0.24, -0.17)$	4.42
9	$(0.20, 0.20)$	0.04	0.00	$(0.19, 0.35)$	19.28
10	$(0.20, 0.20)$	1.84	0.11	$(0.13, 0.15)$	14.74
11	$(0.30, 0.30)$	0.27	0.02	$(0.22, 0.23)$	14.55
12	$(0.30, 0.30)$	0.00	0.00	$(0.23, 0.31)$	32.12

Table 3: Values of $RB(0), \Pi_\Psi(RB(\psi) \leqslant RB(\psi_0) \,|\, T(x)), \psi_{\text{LRSE}}(x)$ and $RB(\psi_{\text{LRSE}}(x))$ in various two-way analyses.

$\psi = (\alpha_{22}, \alpha_{32})$ and (6) is measuring the strength of the evidence that $\psi = 0$. For the first 4 cases H_0 is true and we always get evidence in favor of H_0. Notice that in case 4, where we only have marginal evidence in favor, the strength of this evidence is also quite low (recall that strong means (6) is small when we have evidence against and (6) is big when we have evidence in favor). In cases 5 and 6 the hypothesis H_0 is marginally false and in only one of these cases do we get evidence against and this evidence is deemed to be strong. The other cases indicate that we can still get misleading evidence (evidence in favor when H_0 is false) but the strength of the evidence is not large in these cases. Also, as we increase the effect size, the evidence becomes more definitive against H_0 and also stronger. Overall we see that, based on the sample sizes and the prior, we never get evidence in favor of H_0 that can be considered extremely strong when H_0 is false. In case 3 we get the most evidence in favor of H_0 but (6) says that the posterior probability of the true value having a larger relative belief value is 0.45. The best estimate of the true value in this case is $\psi_{\text{LRSE}}(x) = (-0.02, -0.12)$ with $RB(\psi_{\text{LRSE}}(x)) = 8.62$. Depending on the application, these values can add further support to accepting H_0 as being effectively true.

6 Conclusions

We have shown that, when a hypothesis H_0 has 0 prior probability with respect to a prior on Θ, a Bayes factor and a relative belief ratio of H_0 can be sensibly defined via limits, without the need to introduce a discrete mass on H_0. In general, we have argued that computing a Bayes factor, a measure of the strength of the evidence given by a Bayes factor via a posterior tail probability, and the point where the Bayes factor is maximized together with its Bayes factor, provides a logical, consistent approach to hypothesis assessment. Various inequalities were derived that support the use of the Bayes factor in assessing either evidence in favour of or against a hypothesis. Furthermore, we have presented an approach to assessing the *a priori* bias induced by a particular prior, either in favor of, or against a hypothesis, and have shown how this can be controlled via experimental design.

Acknowledgements

The authors thank the editors and referees for many valuable comments.

References

Aitkin, M. (2010) Statistical Inference: An Integrated Bayesian/Likelihood Approach. CRC Press, Boca Raton.

Berger, J.O. and Delampady, M. (1987) Testing Precise Hypotheses. Statistical Science, 2, 317-335.

Berger, J.O., Liseo, B., and Wolpert, R.L. (1999) Integrated likelihood methods for eliminating nuisance parameters. Statistical Science, 14, 1, 1-28.

Berger, J.O. and Perrichi, R.L. (1996). The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91:10-122.

Dickey, J.M. and Lientz, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. Annals of Mathematical Statistics, 41, 1, 214-226.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Annals of Statistics, 42, 204-223.

Evans, M. Bayesian inference procedures derived via the concept of relative surprise. (1997) Communications in Statistics, Theory and Methods, 26, 5, 1125-1143, 1997.

Evans, M. Guttman, I. and Swartz, T. (2006) Optimality and computations for relative surprise inferences. Canadian Journal of Statistics, 34, 1, 113-129.

Evans, M. and Jang, G-H. (2011a) Weak informativity and the information in one prior relative to another. Statistical Science, 2011, 26, 3, 423-439.

Evans, M. and Jang, G-H. (2011b) A limit result for the prior predictive applied to checking for prior-data conflict. Statistics and Probability Letters, 81, 1034-1038.

Evans, M. and Jang, G-H. (2011c) Inferences from prior-based loss functions. Technical Report No. 1104, Department of Statistics, University of Toronto.

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

Evans, M. and Shakhatreh, M. (2008) Optimal properties of some Bayesian inferences. Electronic Journal of Statistics, 2, 1268-1280.

Garcia-Donato, G. and Chen, M-H. (2005) Calibrating Bayes factor under prior predictive distributions. Statistica Sinica, 15, 359-380.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) Bayesian Data Analysis, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.

Jeffreys, H. (1935) Some Tests of Significance, Treated by the Theory of Probability. Proceedings of the Cambridge Philosophy Society, 31, 203- 222.

Jeffreys, H. (1961) Theory of Probability (3rd ed.), Oxford University Press, Oxford.

Johnson, V.E., Rossell, D. (2010) On the use of non-local prior densities in Bayesian hypothesis tests. Journal of the Royal Statistical Society: Series B, 72, 2, 143–170.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. Journal of the American Statistical Association, 90, 430, 773-795.

Lavine, M. and Schervish, M.J. (1999) Bayes Factors: what they are and what they are not. The American Statistician, 53, 2, 119-122.

Marin, J-M., and Robert, C.P. (2010) On resolving the Savage-Dickey paradox. Electronic Journal of Statistics, 4, 643-654.

O'Hagan, A. (1995) Fractional Bayes factors for model comparisons (with discussion). Journal of the Royal Statistical Society B 56:3-48.

Robert, C.P., Chopin, N. and Rousseau, J. (2009) Harold Jeffreys's Theory of Probability Revisited (with discussion). Statistical Science, 24, 2, 141-172.

Royall, R. (1997) Statistical Evidence. A likelihood paradigm. Chapman and Hall, London.

Royall, R. (2000) On the probability of observing misleading statistical evidence (with discussion). Journal of the American Statistical Association, 95, 451, 760-780.

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

Tjur, T. (1974) Conditional Probability Models. Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.

Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. Journal of the American Statistical Association, 90, 430, 614-618.

Vlachos, P.K. and Gelfand, A.E. (2003) On the calibration of Bayesian model choice criteria. Journal of Statistical Planning and Inference, 111, 223-234.

Appendix

Proof of Theorem 8 We have that

$$\frac{RB(\psi)}{RB(\psi_0)} = \frac{\pi_\Psi(\psi)}{\pi_\Psi(\psi_0)} \frac{\sum_{\theta:\Psi(\theta)=\psi} \pi(\theta \mid \psi) f_{\theta,n}(x)}{\sum_{\theta:\Psi(\theta)=\psi_0} \pi(\theta \mid \psi_0) f_{\theta,n}(x)}$$

and, for θ_0 such that $\Psi(\theta_0) = \psi_0$, let $A_n(\theta_0) = \{\theta : n^{-1} \log(RB(\Psi(\theta))/RB(\psi_0)) \leqslant 0\}$. Note that $\theta_0 \in A_n(\theta_0)$. Now

$$\begin{aligned} \frac{1}{n} \log\left(\frac{RB(\psi)}{RB(\psi_0)}\right) &= \frac{1}{n} \log\left(\frac{\pi_\Psi(\psi)}{\pi_\Psi(\psi_0)}\right) + \frac{1}{n} \log\left(\frac{f_{\theta(\psi),n}(x)}{f_{\theta(\psi_9),n}(x)}\right) \\ &\quad + \frac{1}{n} \log\left(\frac{\sum_{\theta:\Psi(\theta)=\psi} \pi(\theta \mid \psi) f_{\theta,n}(x)/f_{\theta(\psi),n}(x)}{\sum_{\theta:\Psi(\theta)=\psi_0} \pi(\theta \mid \psi) f_{\theta,n}(x)/f_{\theta(\psi_0),n}(x)}\right) \end{aligned} \tag{15}$$

where $f_{\theta(\psi),n}(x) = \sum_{\theta:\Psi(\theta)=\psi} f_{\theta,n}(x)$. Observe that, as $n \to \infty$, the first term on the right-hand side of (15) converges to 0 as does the third term since $0 < \min\{\pi(\theta \mid \psi) : \Psi(\theta) = \psi\} \leqslant \sum_{\theta:\Psi(\theta)=\psi} \pi(\theta \mid \psi) f_{\theta,n}(x)/f_{\theta(\psi),n}(x) \leqslant \max\{\pi(\theta \mid \psi) : \Psi(\theta) = \psi\} < 1$. Now putting $f_{\hat\theta_n(\psi),n}(x) = \max\{f_{\theta,n}(x) : \Psi(\theta) = \psi\}$, the second term on the right-hand side of (15) equals

$$\frac{1}{n} \log\left(\frac{f_{\hat\theta_n(\psi),n}(x)}{f_{\theta_0,n}(x)}\right) - \frac{1}{n} \log\left(\frac{f_{\hat\theta_n(\psi_0),n}(x)}{f_{\theta_0,n}(x)}\right) + \frac{1}{n} \log\left(\frac{f_{\theta(\psi),n}(x)}{f_{\hat\theta_n(\psi),n}(x)} \frac{f_{\hat\theta_n(\psi_0),n}(x)}{f_{\theta(\psi_0),n}(x)}\right). \tag{16}$$

Note that the third term in (16) is bounded above by $n^{-1} \log(\#\{\theta : \Psi(\theta) = \psi\})$ which converges to 0. Now by the strong law, when θ_0 is true, then $n^{-1} \log(f_{\theta,n}(x)/f_{\theta_0,n}(x)) \to E_{\theta_0}(\log(f_\theta(X)/f_{\theta_0}(X)))$ as $n \to \infty$. By Jensen's inequality $E_{\theta_0}(\log(f_\theta(X)/f_{\theta_0}(X))) \leqslant \log E_{\theta_0}(f_\theta(X)/f_{\theta_0}(X)) = 0$ and the inequality is strict when $\theta \neq \theta_0$ while $E_{\theta_0}(\log(f_{\theta_0}(X) /f_{\theta_0}(X))) = 0$. Therefore, using $\#\{\theta : \Psi(\theta) = \psi\} < \infty$, the first term in (16) converges to $\max\{E_{\theta_0}(\log(f_\theta(X)/f_{\theta_0}(X))) : \Psi(\theta) = \psi\}$ while the second term converges to $\max\{E_{\theta_0}(\log(f_\theta(X)/f_{\theta_0}(X))) : \Psi(\theta) = \psi_0\} = 0$. Therefore, we have that there exists n_0 such that $A_n(\theta_0) = \Theta$ for all $n \geqslant n_0$ and so $\Pi_\Psi\left(RB(\psi) \leqslant RB(\psi_0) \mid T(x)\right) = \Pi_\Psi(A_n(\theta_0) \mid T(x)) = 1$.