Rev.R.Acad.Cienc.Exact.Fis.Nat. (Esp)
Vol. 93, N.º 3, pp 303-310, 1999
Monográfico: Métodos Bayesianos en las Ciencias

TESTING A PRECISE NULL HYPOTHESIS USING REFERENCE POSTERIOR ODDS

(Bayes-Frequentist interface/Bayes information criterion/posterior Bayes factor/realization factor/Schwarz criterion)

K. R. W. BREWER

Department of Statistics and Econometrics, Faculty of Economics and Commerce, Australian National University, A.C.T. 0200, Australia. ken.brewer@anu.edu.au

ABSTRACT

Testing a precise null hypothesis against a composite alternative presents a problem for Reference Bayesian inference. When the alternative prior is improper, any finite observation ensures that the Bayes Factor will be infinite. This paradox can be avoided by using a Reference Posterior Odds (RPO) ratio rather than the Bayes Factor. The RPO is closely related to the ratio of the Bayes Factor to its repeated sampling expectation, to Aitkin's Posterior Bayes Factor and also to the probability density or mass of the corresponding Frequentist test statistic, differing from all three principally by a factor $n^{1/2}$, When the observations are normally distributed, the logarithm of the RPO is exactly equal to the Schwarz Criterion, up to an arbitrary constant.

RESUMEN

Contraste de una hipótesis precisa mediante el cociente de referencia de probabilidades finales

El contraste de una hipótesis precisa frente a una alternativa compuesta presenta dificultades para la inferencia bayesiana objetiva, puesto que cuando la distribución inicial bajo la hipótesis alternativa es impropia, el factor Bayes correspondiente resulta infinito. Esta dificultad puede evitarse utilizando el cociente de referencia de probabilidades a posteriori (*reference posterior odds, RPO*). El *RPO* está muy relacionado con el cociente del factor de Bayes a su valor esperado en el muestreo, al factor Bayes a posteriori de Aitkin, y a a densidad de probabilidad en el muestreo del estadístico convencional de contraste, difiriendo de todos ellos en un factor del orden de $n^{1/2}$. Cuando las observaciones tienen una distribución normal, el logaritmo del *RPO* coincide exactamente con el criterio de Schwartz, excepto por una constante arbitraria.

1. INTRODUCTION

For those who follow the Reference Bayesian (RB) approach exemplified by such articles as Bernardo (1979) and Bernard (1996), the testing of a precise null hypothesis against a composite alternative has long been a serious problem. There is usually either one conjugate prior or a limited range of them that can be taken as formalizing neutrality. (Such priors are sometimes described as «noninformative», but since all priors can be viewed as supplying information of one kind or another, the term «reference prior» will be used here instead.) The basic problem is that many reference priors are improper and that for any improper alternative prior the resulting Bayes Factor is automatically infinite —no matter how small the sample or how extreme the observations (Bartlett, 1957).

Earlier attempts to overcome this problem are considered in Section 3 (especially Remark 3.1) and in Appendix 2. The counterintuitive behaviour of the Bayes Factor is here circumvented by using a Reference Posterior Odds ratio (RPO) instead. This RPO is defined, and some of its properties described, in the following Sections of this paper. As a preliminary, the ordinary Bayes Factor, its repeated sampling expectation, and the ratio of the two (the «Realization Factor», R) are discussed in Section 2. The RPO itself is defined in Section 3, and examples of it are derived there and in Appendix 1. In Section 4, Aitkin's Posterior Bayes Factor (PBF or A) is shown to be closely related both to R and to the RPO. In Section 5, all three of these are shown to be related to the Classical or Frequentist test statistic as well. Information criteria are considered in Section 6. (The logarithm of the RPO is exactly equal to the Schwarz Criterion, up to an arbitrary constant.)

2. THE BAYES FACTOR AND ITS EXPECTATION OVER REPEATED SAMPLINGS

2.1. The Bayes Factor

The ordinary or conventional Bayes Factor is defined as the ratio of the mean likelihood of the parameter θ over the null hypothesis, H_0, to its mean likelihood over the alternative hypothesis, H_1. If the prior distribution under H_i is $f_i(\theta)$, $i = 0, 1$, the Bayes Factor may be written,

$$B = \frac{\int_{\theta \in H_0} \text{Lik}(\theta; \mathbf{y}) f_0(\theta)\, d\theta}{\int_{\theta \in H_1} \text{Lik}(\theta; \mathbf{y}) f_1(\theta)\, d\theta},$$

Lik(.) being the likelihood operator, $\mathbf{y}$ the observation vector and $\int f_i(\theta)\, d\theta$ unity by definition. In the important special case where the null hypothesis is precise, i.e. of the form $H_0{:}\theta = \theta_0$, this expression becomes

$$B = \frac{.\text{Lik}(\theta_0; \mathbf{y})}{\int_{\theta \in H_1} \text{Lik}(\theta; \mathbf{y}) f_1(\theta)\, d\theta}.$$

So defined, it is unequivocally a measure of the empirical evidence supporting H_0 over H_1, and this has led to it being regarded as the appropriate test statistic for the choice between these two hypotheses. This is *prima facie* the case in situations where H_0 and H_1 are genuinely practical alternatives and meaningful subjective prior probabilities can be associated with them but, as mentioned in Section 1, if H_0 is precise and H_1 is improper, the Bayes Factor is necessarily infinite. Conventionally, π_0 (the prior probability associated with H_0) and its complement π_1 are each given the value one half, but if π_0 takes any finite value at all, the infinite Bayes Factor requires the posterior probability of H_0 to be unity—for any experimental outcome.

2.2. The Realization Factor

The paradox just described can only be fully resolved by recognizing that priors with $\pi_0 = \pi_1 = 1/2$ and an improper alternative distribution are not neutral between H_0 and H_1 but biased overwhelmingly towards the precise H_0. An appropriate adjustment, making the prior more genuinely neutral, will be described in Section 3. In the meantime, it is worth noting that for experiments of fixed size, the ratio of B to its repeated sampling expectation under H_0 is stable against small changes to the specification of H_1.

Denoting that ratio by R, we have by definition that

$$R = \frac{B}{E_p(B)} = \frac{\text{Lik}(\theta_0; \mathbf{y})/E_1\{\text{Lik}(\theta; \mathbf{y})\}}{\int [\text{Lik}(\theta_0; \mathbf{y})/E_1\{\text{Lik}(\theta; \mathbf{y})\}] f_p(\mathbf{y})\, d\mathbf{y}},$$

where E_1 denotes the expectation over θ under H_1, E_p is the repeated sampling expectation under H_0, and $f_p(\mathbf{y})$ is the sampling density of $\mathbf{y}$, also under H_0.

R is not the kind of statistic typically regarded as useful by Bayesian statisticians but, being directly proportional to B, it is at least arguably an alternative measure of the empirical evidence favouring H_0 over H_1. We refer to it here as «the Realization Factor» because it measures the extent to which the repeated sampling expectation of B under H_0 is realized in the experimental situation. It will be demonstrated in Section 5 that there is a close relationship between R and the Classical or Frequentist test statistic.

3. THE REFERENCE POSTERIOR ODDS

Another alternative to the Bayes Factor is the Reference Posterior Odds (RPO). The posterior odds ratio can be held stable, as the alternative prior distribution is made more and more diffuse, by adjusting the prior odds ratio π_0/π_1 appropriately. The limiting value of π_0 as the alternative prior becomes more and more diffuse is necessarily small. At the limit, where the alternative prior is improper, π_0 can be set equal to the amount of Lebesgue measure for the alternative prior in an interval of fixed length in the near proximity of θ_0, the posterior odds then being defined as the finite limit (the Reference Posterior Odds or RPO) which it was approaching while the alternative prior was still proper. As already indicated, the RPO is closely related to the Realization Factor.

The RPO is preferable to the Realization Factor from a Bayesian standpoint, because it does not depend upon the notion of repeated sampling, and so is not a function of the likelihoods of unobserved events. The derivation of the RPO used in this Section is very similar to that employed by Robert (1993, Section 2) and by Robert and Caron (1996, Section 2), but it avoids an important logical error (see Appendix 2).

Example 1: Normal mean, variance known. Let the observations be $N(\mu, \sigma^2)$ with σ^2 known. The natural choice of reference prior under the alternative hypothesis, H_1, is the (improper) uniform over the real line, but the Bayes Factor will then be infinite. Provided only that π_0 is finite, the posterior odds on $H_0{:}\mu = \mu_0$ over H_1 will then also be infinite. To overcome this problem, a proper uniform prior over the interval $(-C, C)$ can be used instead, and C allowed to tend to infinity. The Bayes Factor then also tends to infinity (for any given set of n observa-

tions) but the posterior odds on H_0 over H_1 can be kept constant by choosing π_0 to be a monotonically decreasing function of C. Specifically, while C is still finite, π_0 must be equal to the amount of alternative prior probability in the fixed length interval $[\mu_0 - \tau\sigma, \mu_0 + \tau\sigma]$ (where τ is an arbitrary constant); and similarly, in the limit where $C \to \infty$, π_0 must be equal to the amount of Lebesgue measure in that same interval. As long as C is still finite we may write $\pi_0/(1 - \pi_0) = \tau\sigma/C$, and the Bayes Factor as

$$B = \frac{(2\pi)^{-1/2}\sigma^{-1}n^{1/2}\exp\{-n(\bar{y} - \mu_0)^2/(2\sigma^2)}{\int_{-C}^{C}[(2\pi)^{-1/2}\sigma^{-1}n^{1/2}\exp\{-n(\bar{y} - \mu)^2(2\sigma^2)](2C)^{-1}\,d\mu}$$

where $\bar{y}$ is the mean of the n observations. For large C, the posterior odds tend to the limit $2\tau(2\pi)^{-1/2}n^{1/2}\exp\{-n(\bar{y} - \mu_0)^2/(2\sigma^2)\}$, and it is this limit that defines the Generalized Reference Posterior Odds (GRPO or G_{gr}).

The maximum value of the GRPO over $\bar{y}$ is therefore $2\tau(2\pi)^{-1/2}n^{1/2}$ or $kn^{1/2}$ where $k = \tau(2/\pi)^{1/2}$. To define the RPO itself, it only remains to choose a suitable value for the fixed length interval 2τ, or equivalently for k. Since $k = 2\tau(2\pi)^{-1/2}$, it is interpretable as the area of a rectangle whose width is $2\tau\sigma$ and whose height is $(2\pi)^{-1/2}\sigma^{-1}$, the maximum value of the density function for a single observation.

An immediately obvious choice for k is unity. It will be shown in Section 6 that defining the RPO with $k = 1$ is effectively the same as using the Schwarz Criterion or the Bayes Information Criterion (BIC). The choice $k = 1$ also corresponds to the situation where, given a single observation located exactly at μ_0, the experimenter would be indifferent between H_0 and H_1. The reference value proposed here, however, is $k = 2^{1/2}$. Some convenient results that follow from the use of that value will be encountered in Sections 4 and 5, but the choice remains essentially arbitrary. Robert (1993) and Robert and Caron (1996) advocate $\tau = 1/2$ or $k = 1/(2\pi)^{1/2}$, but their argument involves a confusion between probability mass and probability density (see Appendix 2). □

Two further examples of RPO derivations are given in Appendix 1.

When the null hypothesis is true, the posterior odds tend to increase proportionally with $n^{1/2}$. Hence for observations that are not intrinsically normal, but for which the Central Limit Theorem holds, it seems reasonable to choose the prior odds so as to ensure that, for large n, the maximum attainable value of the posterior odds would also be $kn^{1/2}$.

Formally, the generalized RPO, G_{gr}, can be defined by the requirement that when n is large and the observations are such that the posterior odds ratio attains its maximum possible value conditional on n, its leading term should be equal to $kn^{1/2}$, i.e.

$$\lim_{n\to\infty}\left\{\max_{y}(G_{gr})/kn^{1/2}\right\} = 1.$$

This definition can also be used more generally where the alternative prior density is not uniform but is still locally flat in the region of H_0. Let $\bar{y}_0$ be the value of $\bar{y}$ for which $\mathrm{Lik}(\theta_0; n, \bar{y})$ attains its maximum value (for any given n observations). Then we may write the generalized RPO as

$$G_{gr} = kn^{1/2}\frac{\mathrm{Lik}(\theta_0; n, \bar{y})}{\mathrm{Lik}(\theta_0; n, \bar{y}_0)},$$

and the proposed specific RPO, G_r, is the special case for which $k = 2^{1/2}$.

Remark 3.1. This is actually a revival of a much earlier idea. M. S. Bartlett, on reading Lindley (1957), had written to the author pointing out that when the prior distribution describing H_1 was improper, B became infinite and the «silly answer» ensued that the posterior probability of H_0 would be unity for any set of observations. Lindley's counter-suggestion was to make «the prior odds in favour of the null hypothesis against *any unit interval* of the alternative values [equal to a constant]» (Bartlett (1957)). At the time, however, Bartlett regarded this as «rather an artificial evasion of the difficulty» and commonly used Bayesian textbooks such as Press ((1989), p. 35) still warn that it may be impossible to conduct a meaningful hypothesis test of a simple null hypothesis against a diffuse alternative. The remedy usually recommended is to use a realistic subjective prior distribution over the space of H_1 instead.

However, a number of Reference Bayesians (including Jeffreys (1961), Smith and Spiegelhalter (1980), Aitkin (1991), O'Hagan (1995), Kass and Wasserman (1995), Berger (1995) and Berger and Pericchi (1996)) have attempted to tackle the problem by devising proper priors that deliver plausible values for B. An airing of diverse views on this topic can be found in the Discussion to O'Hagan (1993). □

4. THE POSTERIOR BAYES FACTOR

Concerned at the sensitivity of the Bayes Factor to the specification of the alternative prior, Aitkin (1991) introduced a statistic that he called the Posterior Bayes Factor (or PBF) and denoted by A. It differed from the conventional Bayes Factor in that the means of the likelihoods were taken over the posterior rather than the prior distributions, resulting in the expression

$$A = \frac{\int_{\theta\in H_0} \text{Lik}^2(\mathbf{y};\,\theta) f_0(\theta)\,d\theta \Big/ \int_{\theta\in H_0} \text{Lik}(\mathbf{y};\,\theta) f_0(\theta)\,d\theta}{\int_{\theta\in H_1} \text{Lik}^2(\mathbf{y};\,\theta) f_1(\theta)\,d\theta \Big/ \int_{\theta\in H_1} \text{Lik}(\mathbf{y};\,\theta) f_1(\theta)\,d\theta} =$$

$$= \frac{E_0\{\text{Lik}^2(\theta_0;\,\mathbf{y})\}/E_0\{\text{Lik}(\theta_0;\,\mathbf{y})\}}{E_1\{\text{Lik}^2(\theta_1;\,\mathbf{y})\}/E_1\{\text{Lik}(\theta_1;\,\mathbf{y})\}}.$$

This statistic did not meet with favour from orthodox Bayesians, mainly because, as a Bayes Factor, it corresponded only to one special and highly informative prior (Fearn (1991)), and because it used the same data twice to make a single inference in such a fashion as to be temporally incoherent (Cuzick, 1991; Cox, 1991). The following theorem, however, indicates a close relationship between A and the Realization Factor, R.

Theorem 4.1. Given that

1. H_0 is of the simple or precise form $\theta = \theta_0$,
2. $f_1(\theta)$ is flat,
3. there is a univariate sufficient statistic, $\hat{y}$, for θ and
4. $L(\theta;\,\hat{y})$ is origin invariant,

then A and R are equal.

For proof, see Appendix 3. □

Remark 4.1. Since Conditions 3 and 4 of the Theorem are satisfied by the normal distribution, the asymptotic equality of R and A can be invoked whenever H_0 is precise, $f_1(\theta)$ is locally uniform and the Central Limit Theorem is opeative. □

Whenever Theorem 4.1 holds, the PBF can also be interpreted as the posterior odds corresponding to the situation where the prior distribution is identical with the reference prior defined in Section 3, except that the amount of prior Lebesgue measure associated with H_0 is chosen specifically to ensure that $G_{gr} = A$. Such a prior may be viewed as reflecting the beliefs of a person who, faced with an experiment of size n in which the posterior odds take their maximum value given n, regards those posterior odds as being $2^{1/2}$:1 on. So $G_{gr} = A$ implies that k has been chosen to be $(2/n)^{1/2}$. If, however, k is already specified as taking the specific reference value, $2^{1/2}$, (i.e. $G_{gr} = G_r = An^{1/2}$) then $G_{gr} = A$ only for the special case where $n = 1$.

It might perhaps be regarded as unduly arbitrary to adopt a prior in which the experiment size, n, figures so prominently, but such a judgment could well be regarded as implicit in the (equally arbitrary) choice of the experiment size itself. Alternatively, one might argue that the prior can only be regarded as a reference prior for one particular value of n, and that the natural value to choose is the size of the experiment.

There is, however, a substantial difference between the interpretation of a hypothesis test based on the use of the statistic G_r (with $k = 2^{1/2}$) and that of one based on the use of A (or, almost equivalently, of R). If $k = 2^{1/2}$ is used, there is a quite definite *a priori* degree of belief associated with H_0. Although in the limit as $C \to \infty$ the value of π_0 is notionally zero, it is nevertheless equal to the amount of Lebesgue measure associated with the alternative hypothesis in the interval $(\theta_0 - \pi^{1/2}\sigma,\ \theta_0 + \pi^{1/2}\sigma)$ or, equivalently, in an interval of length $2\pi^{1/2}\sigma$ anywhere on the real line.

If A or R is used, however, the implied degree of belief in H_0 is proportional to $n^{-1/2}$. This can only be compatible with Bayesian inference if n is held fixed throughout the inference process. It would not be permissible, for instance, to stop the experiment halfway through and make a provisional inference based on the value of A obtained from that half of the experiment. The two inferences taken together would be incoherent.

Thus the use of A in the fashion recommended by Aitkin seems not to be fully Bayesian in spirit. Instead of there being a fixed prior degree of belief in H_0, there is the implicit assumption that H_0 can never really be trusted, and therefore that the only question at issue is whether or not there is already an experiment size large enough to demonstrate that it is false. The similarity between this approach and that of the Frequentist statistician is obvious, and it is not surprising to see that Aitkin himself drew attention to the repeated sampling properties of A (Aitkin [1991], Section 3).

5. THE CLASSICAL OR FREQUENTIST TEST STATISTIC

Under the conditions required for Theorem 4.1, A and R are interchangeable and, when considering relationships with the Frequentist test statistic, it is more convenient to think in terms of R. For all three examples (the first in Section 3 and the other two in Appendix 1) the choice of a uniform prior distribution over H_1 leads to a formula for R which is proportional either to the probability density or, in the discrete case, to the probability distribution of the corresponding Classical test statistic under H_0.

For Example 1 (the normal mean) R (i.e. $G_r n^{-1/2}$) is proportional to the ordinate of $N(\mu_0,\ \sigma^2/n)$, the distribution of the Classical test statistic.

For Example 2 (the normal variance), R may be interpreted as the ordinate of χ^2_{n-1+2a} standardized so as to have the expectation unity over repeated sampling when $\sigma^2 = \sigma_0^2$. Hence if the flat prior (the one for which $a = 0$) is

chosen, R is the similarly standardized ordinate of χ^2_{n-1}. The Classical test statistic is also proportional to χ^2_{n-1}, but has its ordinates standardized so as to ensure that the value of its definite integral over the alternative parameter space is equal to unity.

For Example 3 (the binomial parameter), G_r, B and R are all inversely proportional to $r!(n - r)!$ (r being the number of «successes») when the alternative prior is uniform. This is also the condition for proportionality to the probability distribution of the Classical test statistic.

While the correspondence between the Classical test statistic and the RB test statistic obtained when the alternative prior is uniform may not hold in general, it does hold wherever the Central Limit Theorem can be invoked. However, whereas the Classical test focuses on the repeated sampling distribution of the observations and uses tail areas, the RB test focuses on the possible parameter values and uses ordinates. Further, for discrete distributions such as the binomial where the meaningful tail areas are limited in number, the one to one correspondence is restricted to a finite number of possible comparisons.

There remains one more important difference between the Classical test and the specific RB test based on G_r (for which $k = 2^{1/2}$). This is the factor $n^{1/2}$, which has already been recognized as the rate at which B tends to increase with the size of the experiment when H_0 is true. It corresponds to the progressive elimination of alternatives that are nearly but not quite equivalent to H_0, and the consequent reduction in the amount of alternative prior probability able to support H_1 in any effective sense. It is this factor, combined with the hypersensitivity of B to the choice of alternative prior, that leads to the well-known Lindley Paradox (Lindley, 1957), namely that a Classical significance test for H_0 could be indicating **y** to be significant at the 5 % level, while at the same time the posterior probability of H_0, given **y**, could be as high as 95 %, even though the corresponding prior probability was quite small.

6. THE CHOICE OF INFORMATION CRITERION

Aitkin (1991, Section 3) showed that for a comparison between two nested hypotheses with v additional parameters in H_1, the multivariate version of A implied a particular form of information criterion within a general class of penalized likelihood ratio test statistics discussed by Smith and Spiegelhalter (1980). Using λ to denote the usual likelihood test statistic, this general class may be written $A(m) = \lambda - mv$ and, since $-2 \log A = \lambda - v \log 2$, the choice of multivariate A as a test statistic implies the choice $m = \log 2 = 0.693$. This is a smaller value for m than is customarily used in practice. For comparison, the Akaike Information Criterion or AIC uses $m = 2$, while the Bayes Information Criterion [together with the effectively identical Schwarz Criterion, S] uses $m = \log n$ (Schwarz, 1978).

The choice $m = \log 2$ for all n results in decisions to favour complex models to a counterintuitive extent. For example, suppose that there were v potential regressors for a given regressand, all spurious, and that each in turn was to be judged as either «real» or «spurious» depending on whether or not the value of λ obtained by adding each candidate in turn to the equation exceeded log 2 (or 0.693). Since the probability that χ^2_1 exceeds 0.693 is about 0.43, an expected 43 per cent of these spurious parameters would be judged to be «real», regardless of the size of the experiment. If they were to be tested as a group, the likely outcome would be even more extreme. For $v = 12$, the probability that χ^2_{12} exceeds 12 log 2 is just greater than 3/4, so the entire group would be judged «real» with a probability of more than 75 per cent.

By contrast, the use of G_{gr} in place of A leads to the formula $m = \log (k^2 n)$. The choice $k = 1$ yields $m = \log n$ and therefore corresponds to the Schwarz Criterion/BIC, for which m exceeds the value 2 when $n \geqslant 8$ (Schwarz [1978]). The choice of G_r, with $k = 2^{1/2}$ and $m = \log (2n)$, is more parsimonious still, m exceeding 2 when $n \geqslant 4$.

It would nevertheless appear that the BIC, and even more the RPO, could be unduly parsimonious (i.e. prone to underfitting) if used to decide between the use of H_0 and H_1 in practice. Unlike any Frequentist test statistic, the RPO is intended to reflect the odds on H_0 that a reasonable person might hold, after observing a sample, if that person had no strong beliefs beforehand as to the value of the relevant unknown parameter, conditional on H_1 being true. The BIC test statistic implies a posterior odds smaller than the RPO only by the factor $2^{v/2}$, and therefore carries fairly similar implications. It seems difficult to justify, for reference purposes, any posterior odds markedly larger or smaller than these.

There is, however, an important difference between the inference that H_0 is more probable than H_1 and the decision to use H_0 rather than H_1 in practice. It is fairly obviously unsafe to use H_0 in circumstances where it is only marginally more probable than H_1, since the latter makes greater use of the information provided by the data. A penalty function is therefore indicated, but one operating in a direction opposite to that used by Frequentists. Pending a comprehensive empirical investigation, the logarithmic penalties 2, 3 and 4 might be used to define a «parsimonious», a «standard» and a «generous» test; e.g. the standard test would indicate the use of H_0 only if $2 \log G_r > 3v$. These tests would be more parsimonious than the AIC for $n \geqslant 28$, 75 and 202 respectively.

Kass and Wasserman (1995) have further suggested that, for the BIC, the relevant value of n in the formula

for m is not the number of vector observations but the number of vector components in the observations, nv. If this is correct, the use of G_r should lead to the choice $m = \log(2nv)$, which is more parsimonious still. However the notion that this result should hold even when each component within the matrix of observations, **Y**, is almost perfectly correlated with every other is counterintuitive, and it seems wise to suspend judgment on this issue for the time being.

7. ACKNOWLEDGEMENTS

I am indebted to Prof. M. Aitkin for the initial stimulus to tackle this problem, to Dr. D. S. Poskitt for early encouragement and to Profs. R. L. Chambers and D. V. Lindley for pointing out pitfalls along the way.

Appendix 1. Two further derivations of Reference Posterior Odds

Example 2: Normal variance, mean unknown. Let the n observations y_i be distributed $N(\mu, \sigma^2)$ with μ initially distributed uniformly over the real line. The precise null hypothesis, H_0, is that $\sigma^2 = \sigma_0^2$ and the composite alternative, H_1, that $\sigma^2 \neq \sigma_0^2$, the alternative prior having the distribution $k_{a,\varepsilon,\omega} = \left[\int_{\varepsilon}^{\omega} \sigma^{-2a}\, d\sigma^2\right]^{-1}$.

The statistics $\bar{y} = n^{-1}\sum_{j=1}^{n} y_j$ and $s^2 = (n-1)^{-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$ are sufficient for μ and σ^2 respectively. Integrating over μ, the likelihood of σ^2 indexed on s^2 only may be written

$$\mathrm{Lik}(\sigma^2; s^2) = \sigma^{-(n-1)} \exp\,[-(n-1)s^2/(2\sigma^2)].$$

The expectation of this expression over the parameter space of H_1 is

$$\int_{H_1} \mathrm{Lik}(\sigma^2; s^2) f_1(\sigma^2)\, d\sigma^2$$

$$\cong k_{a,\varepsilon,\omega}\Gamma[(n-3+2a)/2][2/\{n-1)s^2\}]^{(n-3+2a)/2}.$$

This approximation can be made arbitrarily close—in the sense that the ratio of the two sides can be made to approach unity—by choosing ε to be small enough and ω to be large enough. With this approximation in mind, the Bayes Factor in favour of H_0 over H_1 is

$$B_a \cong \frac{\sigma_0^{2(a-1)}[(n-1)s^2/(2\sigma_0^2)]^{(n-3+2a)/2} \exp[-(n-1)s^2/(2\sigma_0^2)]}{k_{a,\varepsilon,\omega}\Gamma[(n-3+2a)/2]}.$$

This expression, and hence also the posterior odds ratio, attains its maximum value over s^2 when $(n-1)s^2/\sigma_0^2 = n - 3 + 2a$, in which case

$$\max_{s^2} B_a$$

$$\cong \frac{\sigma_0^{2(a-1)}[(n-3+2a)/2]^{(n-3+2a)/2} \exp[-(n-3+2a)/2]}{k_{a,\varepsilon,\omega}\Gamma[(n-3+2a)/2]}.$$

Equating $\max_{s^2} B_a(\pi_0/\pi_1)_{\mathrm{gr}}$ asymptotically to $kn^{1/2}$ and applying Stirling's formula, the generalized reference prior odds ratio is given by

$$(\pi_0/\pi_1)_{\mathrm{gr}} = kn^{1/2}(\max_{s^2} B_a)^{-1} \cong 2k\pi^{1/2}k_{a,\varepsilon,\omega}\sigma_0^{-2(a-1)}.$$

The corresponding generalised RPO is

$$G_{\mathrm{gr}} = \lim_{\varepsilon \to 0, \omega \to \infty} \{B_a(\pi_0/\pi_1)_{\mathrm{gr}}\}$$

$$= \frac{2k\pi^{1/2}\{(n-1)s^2/(2\sigma_0^2)\}^{(n-3+2a)/2} \exp[-(n-1)s^2/(2\sigma_0^2)]}{\Gamma[(n-3+2a)/2]}.$$

or, for large n (and again applying Stirling's formula),

$$G_{\mathrm{gr}} \cong \frac{(2n)^{1/2}k \exp[-(n-1)s^2/(2\sigma_0^2)]}{\exp[-(n-5+2a)/2]}.$$

Example 3: Binomial parameter. Let the probability of success on a single trial be P. Consider the test of $H_0{:}P = P_0$ against $H_1{:}P \neq P_0$, the prior density over the alternative parameter space being $\pi_1 f_1(P) = \pi_1 P^{\alpha-1}(1-P)^{\beta-1}\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)$. Suppose r successes are observed in n trials. The posterior odds ratio is then

$$G = \frac{\pi_0}{\pi_1} B = \frac{\pi_0}{\pi_1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\frac{\Gamma(n+\alpha+\beta)}{\Gamma(r+\alpha)\Gamma(n-r+\beta)} P_0^r(1-P_0)^{n-r}.$$

This attains its maximum value over r when $(r+\alpha)/(n+\alpha+\beta) = P_0$. In the limit as $n \to \infty$, this implies $r/n \cong P_0$ and, since the leading term in G_{gr} must be $kn^{1/2}$,

$$G_{\mathrm{gr}}n^{-1/2} \cong k = (\pi_0/\pi_1)_{\mathrm{gr}}\Gamma(\alpha)\Gamma(\beta)\{\Gamma(\alpha+\beta)\}^{-1}$$

$$\lim_{n \to \infty} [\Gamma(n+\alpha+\beta)\{\Gamma(r+\alpha)\Gamma(n-r+\beta)\}^{-1} r^r(n-r)^{n-r}n^{-(2n+1)/2}].$$

Applying Stirling's formula, we have

$$\left(\frac{\pi_0}{\pi_1}\right)_{\mathrm{gr}} \cong k(2\pi)^{1/2}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} P_0^{\alpha-0.5}(1-P_0)^{\beta-0.5}$$

and the generalized RPO is

$$G_{gr} = \left(\frac{\pi_0}{\pi_1}\right)_{gr} B$$

$$\cong (2n)^{1/2}\{nP_0/r\}^{r+\alpha-0.5}\{n(1-P_0)/(n-r)\}^{n-r+\beta-0.5}$$

Three priors having some claim to be neutral are of special interest:

1) $\alpha = \beta = 1$, the uniform prior;

2) $\alpha = \beta = 0.5$, the Jeffreys prior (Jeffreys (1961), p. 125) and

3) $\alpha = \beta = 0$, the Haldane (1948) improper prior proportional to $P^{-1}(1-P)^{-1}$.

Comparing the formulae for G_{gr} in the three cases, the RPO can be shown to be asymptotically stable against such variations in α and β, provided only that P_0 is neither zero nor one. It is least stable when either r or $n-r$ is small.

The expression for $(\pi_0/\pi_1)_{gr}$ is not stable against changes in α or β and tends to zero as $\alpha, \beta \to 0$. When $\alpha = = \beta = 0.5$, however, it takes a particularly simple form, $k(2/\pi)^{1/2}$, which is uniquely invariant over P_0 (cf. Jeffreys (1961), p. 188).

Appendix 2. The contributions of Robert and Caron

Robert (1993) considered H_0:$\theta = 0$ for a single incoming observation from the distribution $N(\theta, 1)$, and conjugate priors $N(0, \sigma^2)$ rather than the uniform over $(-C, C)$, but otherwise his approach was very similar to that described in Section 3. For a sequence of increasingly diffuse proper priors he proposed the constraint that (in the notation of this paper) required that π_0 be held equal to the amount of alternative probability mass in a fixed interval either side of 0. (He actually suggested the 99 % HPD region of the incoming observations.) At the point of transition from proper to improper priors, however, he made a logical error. He described his constraint as «too strong to hold when σ goes to infinity, since the prior probability of any fixed interval must go to 0». As an alternative he suggested that the densities for H_0 and H_1 be equal at 0. Had he carried this idea through consistently, it would have resulted in the ratio of π_0 to the alternative probability mass in any finite fixed interval going to zero. He interpreted his suggestion, however, as requiring equality between the probability *mass* π_0 and the probability *density* of the conjugate alternative prior at 0. This was equivalent to the requirement in Section 3 of this paper that defines the limiting value of $\pi_0 C$ as $C \to \infty$ to be $k(2\pi)^{1/2}$ observational standard deviations, with the value of k set at $(2\pi)^{-1/2}$ or 0.3989. It consequently defined the fixed interval to be half a standard deviation on either side of zero, covering approximately the 38.3 % HPD of incoming observations under H_0. (For $k = 2^{1/2}$, the relevant interval extends $\pi^{1/2}$ standard deviations each side of μ_0 and approximates the 92.4 % HPD region for the observations. To arrive at the 99 % HPD, it would have been necessary to set k equal to $2.5758(2/\pi)^{1/2}$ or 2.0552—see Fig. 1).

The consequence of choosing such a small value as 0.3989 for k is a test distinctly lacking in parsimony when considered against other Bayesian tests. In general, the maximum value for the posterior odds (obtained when all the n standard normal observations happen to be exactly zero) is $kn^{1/2}$. Thus taking the value $k = 1$ supplied by the BIC it is just possible—with probability zero—to achieve indifference between H_0 and H_1 for a sample of one. It is also possible, with a sample of size two, to achieve a posterior odds $\geqslant 1$ with probability 0.5 when H_0 is true. With $k = (2\pi)^{-1/2}$, however, the minimum possible sample size necessary to achieve a posterior odds $\geqslant 1$ is seven, and the minimum possible sample size necessary to achieve a posterior odds $\geqslant 1$ with probability greater than 0.5 when H_0 is true is 13. In terms of required sample size, the posterior odds test proposed by

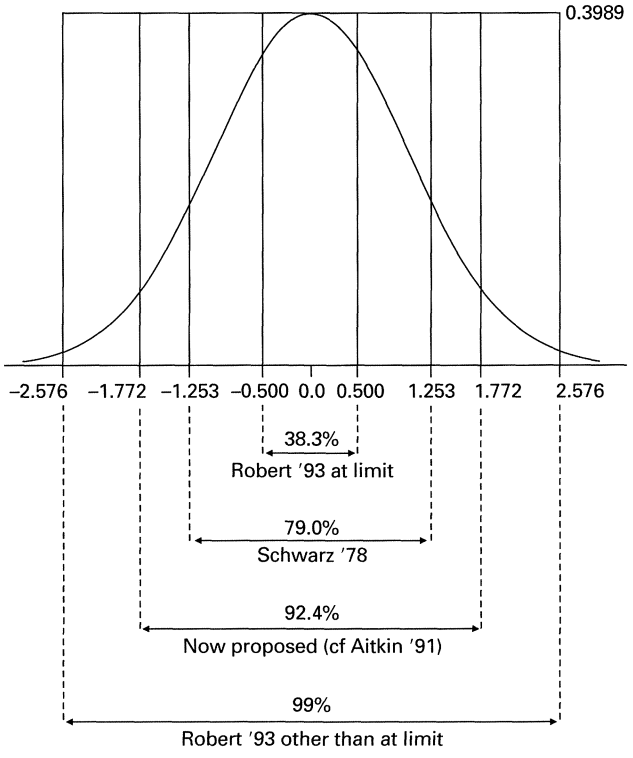

Figura 1. HPD regions for $N(0, 1)$ observations. The height of the rectangles is $(2\pi)^{-1/2}$ and $k = \tau(2/\pi)^{1/2}$ is the area of a rectangle of width 2τ. The four values of τ shown are 1/2, $(\pi/2)^{1/2}$, $\pi^{1/2}$ and $\Phi^{-1}(0.99)$ or 2.576..., corresponding to $k = (2\pi)^{-1/2}$, 1, $2^{1/2}$ and 2.055... respectively.

Robert is 2π or about 6.28 times less parsimonious than the BIC. For $n < 47$, it is less parsimonious than the AIC.

Robert and Caron [1996] compared Robert's posterior probabilities for H_0 first with the Frequentist p-values and then with another statistic they called the Neutral Bayes Factor. They showed that although the p-values and the posterior probabilities had different asymptotic behaviours, they took roughly similar values over the range where the standard normal observation x was between one and three. They concluded that for practical purposes their posterior probabilities yielded much the same inferences as the Frequentist test. This conclusion, however, ignored the factor $n^{1/2}$ that enters into the posterior probabilities for samples larger than a single observation. It was also dependent on the choice of the value $(2\pi)^{-1/2}$ for k which, as has been seen, leads to a test distinctly less parsimonious than the BIC.

Their Neutral Bayes Factor was actually the probability density of a proper conjugate prior distribution centred on zero and chosen such that the ordinary Bayes Factor, for any given standard normal observation, would be unity. This probability density was then interpreted as a Bayes Factor because it could «be interpreted as the maximum weight one [could] give to H_0^c [the complement of H_0] for H_0 to be accepted (i.e., more rigorously, for the probability of H_0 to increase from *a priori* to *a posteriori*)». (Robert and Caron [1996], p. 425). For a single standard normal observation, this Neutral Bayes Factor was asymptotically equivalent to their posterior odds as $|x| \to \infty$, but it took the values unity at $x = 0$ and infinity in the ranges $0 < |x| \leqslant 1$. This last result was interpreted by the authors as either «strong support for H_0 or, alternatively, as the impossibility to assess quantitatively the validity of H_0 when x is too close to 0» (p. 426). It seems unlikely that this statistic could bear the weight of the interpretations given to it by its authors without the value $k = (2\pi)^{-1/2}$ being accepted as normative first.

Appendix 3. Proof of Theorem 4.1

Since H_0 is precise,

$$A = \mathrm{Lik}(\theta_0;\, \hat{y})/[E_1\{\mathrm{Lik}^2(\theta;\, \hat{y})\}/E_1\{\mathrm{Lik}(\theta;\, \hat{y})\}], \text{ and}$$

$$R = \frac{[\mathrm{Lik}(\theta_0;\, \hat{y})/E_1\{\mathrm{Lik}(\theta;\, \hat{y})\}]}{\int [\mathrm{Lik}(\theta_0;\, \hat{y})/E_1\{\mathrm{Lik}(\theta;\, \hat{y})\}] f_p(\hat{y})\, d\hat{y}}$$

Since $\mathrm{Lik}(\theta;\, \hat{y})$ is origin invariant and $f_1(\theta)$ is flat, $E_1\{\mathrm{Lik}(\theta;\, \hat{y})\}$ is invariant with $\hat{y}$ and can be cancelled out, leaving

$$R = \mathrm{Lik}(\theta_0;\, \hat{y}) \Big/ \int \mathrm{Lik}(\theta_0;\, \hat{y}) f_0(\hat{y})\, d\hat{y}$$

$$= \mathrm{Lik}(\theta_0;\, \hat{y}) \Big/ \left[\int \mathrm{Lik}^2(\theta_0;\, \hat{y})\, d\hat{y} \Big/ \int \mathrm{Lik}(\theta_0;\, \hat{y})\, d\hat{y} \right] =$$

$$= \frac{\mathrm{Lik}(\theta_0;\, \hat{y}) \int_{\theta \in H_1} \mathrm{Lik}(\theta;\, \hat{y}) f_1(\theta) d\theta}{\int_{\theta \in H_1} \mathrm{Lik}^2(\theta;\, \hat{y}) f_1(\theta) d\theta}$$

REFERENCES

1. Aitkin, M. (1991). Posterior Bayes Factors (with discussion). *J. R. Statist. Soc.* 5 **53,** 111-142.
2. Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44,** 533-534.
3. Berger, J. O. (1995). Recent developments and applications of Bayesian analysis. *Bull. Int. Statist. Inst.* **LVI,** 3-14.
4. Berger, J. O. and Pericchi, L. (1996). Intrinsic Priors for model selection and prediction. *J. Am. Statist. Assoc.* **91,** 109-122.
5. Bernard, J. M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *Am Statist.,* **50,** 7-13.
6. Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. R. Statist. Soc.* B, **41,** 113-147.
7. Cox, D. R. (1991). Comment in the Discussion of Aitkin ([**10**], pp. 131-132).
8. Cuzick, J. (1991). Comment in the Discussion of Aitkin ([**10**], pp. 135-136).
9. Fearn, T. (1991). Comment in the Discussion of Aitkin ([**10**], p. 134).
10. Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35,** 297-303.
11. Jeffreys, H. (1961). *Theory of Probability,* 3rd edn. Oxford: Oxford University Press.
12. Kass, R. E. and Wasserman, L. (1995). A Reference Bayesian test for nested hypotheses and its relation to the Schwarz Criterion. *J. Am. Statist. Assoc.* **90,** 928-934.
13. Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44,** 187-192.
14. O'Hagan, A. (1995). Fractional Bayes Factors for model comparison (with discussion). *J. R. Statist. Soc.* B **57,** 99-138.
15. Press, S. J. (1989). *Bayesian Statistics: Principles, Models and Applications.* New York: John Wiley.
16. Robert, C. P. (1993). A note on Jeffreys-Lindley Paradox. *Statistica Sinica,* **3,** 601-608.
17. Robert, C. P. and Caron, N. (1996). Noninformative Bayesian testing and Neutral Bayes Factors. *Test* **5,** 411-437.
18. Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6,** 461-464.
19. Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes Factors and choice criteria for linear models. *J. R. Statist. Soc.* B7 **42,** 213-220.