BAYESIAN STATISTICS 6, pp. 101–130
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Nested Hypothesis Testing: The Bayesian Reference Criterion

JOSÉ M. BERNARDO
Universitat de València, Spain

SUMMARY

It is argued that hypothesis testing problems are best considered as decision problems concerning the choice of a *useful* probability model. Decision theory, information measures and reference analysis, are combined to propose a non-subjective Bayesian approach to nested hypothesis testing, the *Bayesian Reference Criterion* (BRC). The results are compared both with frequentist based procedures, and with the use of Bayes factors. The theory is illustrated with stylized examples, where alternative approaches may easily be compared.

Keywords: DECISION THEORY; LOGARITHMIC DISCREPANCY; MODEL CHOICE; MODEL COMPARISON; NON-SUBJECTIVE BAYESIAN STATISTICS; PROPER SCORING RULES; REFERENCE ANALYSIS.

1. MOTIVATION

Let M_1 denote a probability model, $p_{\boldsymbol{x}}(. \mid \theta), \boldsymbol{x} \in \boldsymbol{X}$, which is currently assumed to provide an appropriate description of the probabilistic behaviour of an observable vector $\boldsymbol{x}$ in terms of some relevant quantity $\theta \in \Theta$ and, on this basis, let us consider whether the *null* model M_0 labeled by a particular value $\theta = \theta_0$ may —or may not— be judged to be *compatible* with an observed value of $\boldsymbol{x}$; the value θ_0 may have the support of a scientific theory (but some unknown experimental bias may be present), or it may just label a model which is easier to use, or simpler to interpret. For instance, one might have collected a set $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of n dichotomous observations with $r = \sum x_j$ successes assumed to be a subset of an exchangeable sequence; it then follows from de Finetti's representation theorem (see *e.g.*, Lindley and Phillips, 1976) that $\boldsymbol{x}$ is a random sample of n Bernoulli observations with some parameter θ, and we may wish to judge whether, given the exchangeability assumption, the particular value $\theta = \theta_0$ (maybe suggested by a scientific theory, maybe a number with political significance, or maybe just a simple approximation to a historical relative frequency) is *compatible* with the observed data (r, n).

Any Bayesian solution to the problem posed will obviously require a prior distribution $p(\theta)$ over Θ, and the result may well be very sensitive to the particular choice of such prior; note that, in principle, there is no reason to *assume* that the prior should *necessarily* be concentrated around a particular θ_0; indeed, for a judgement on the compatibility of a particular parameter value with the observed data to be useful for scientific communication, this should only depend on the assumed model and the observed data, and this requires some form of non-subjective prior specification for θ which could be argued to be 'neutral'; a sharply concentrated prior around a particular θ_0 would hardly qualify.

This research has been funded with Grant PB96-0754 of the DGICYT, Spain.

The conventional Bayesian approach to compare a 'null' model M_0 versus an alternative model M_1, on the basis of some data $\boldsymbol{x}$, is to compute the corresponding *Bayes factor* $B_{01}(\boldsymbol{x})$; indeed, the ratio $\Pr(M_0 \mid \boldsymbol{x})/\Pr(M_1 \mid \boldsymbol{x})$ of the posterior probabilities associated to each model may be written as $B_{01}(\boldsymbol{x})\Pr(M_0)/\Pr(M_1)$, where $B_{01}(\boldsymbol{x}) = p(\boldsymbol{x} \mid M_0)/p(\boldsymbol{x} \mid M_1)$, and therefore, the Bayes factor $B_{01}(\boldsymbol{x})$ seemingly encapsulates all the data have to say about the problem. Bayes factors have been the basis of most work on Bayesian hypothesis testing, and the relevant literature is huge, dating back to Jeffreys (1939); Kass and Raftery (1995) have provided an excellent review. If M_0 is a particular case of M_1, and M_0 is of smaller dimension than M_1, then the use of Bayes factors implicitly assumes that a *strictly positive* probability $\Pr(M_0)$ has been assigned to a set of *zero Lebesgue measure* under the larger model M_1 (which is assumed to be appropriate). The posterior probability $\Pr(M_0 \mid \boldsymbol{x})$ obtained from this *singular* prior may be shown to provide an approximation to the posterior probability associated to a small neighbourhood $\theta_0 \pm \epsilon$ of the null value obtained from a regular prior sharply concentrated around θ_0 (Berger and Delampady, 1987). However, for any fixed ϵ, this approximation *always* breaks down for sufficiently large samples; moreover, as mentioned above, it does not seem reasonable to *require* a sharply concentrated prior around θ_0 just to check the compatibility of θ_0 with the observed data.

Foundational issues aside, the use of singular priors may demonstrably have unpleasant consequences. The simplest illustration is provided by Lindley's famous paradox.

Lindley's paradox. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $\mathrm{N}(x \mid \mu, \sigma)$, with *known* variance σ^2 (model M_1), and let M_0 be the particular case which corresponds to $\mu = \mu_0$. The sample mean $\overline{x}$ is then sufficient and, if the prior distribution of μ is assumed to be $p(\mu) = \mathrm{N}(\mu \mid \mu_0, \sigma_1)$, then the Bayes factor $B_{01}(\boldsymbol{x}, \mu_0)$ in favour of the simpler model M_0 is easily found to be

$$B_{01}(\boldsymbol{x}, \mu_0) = B_{01}(z, n, \lambda) = \Big(1 + \frac{n}{\lambda}\Big)^{1/2} \exp\Big[-\frac{1}{2}\,\frac{n}{n+\lambda}\,z^2\Big],$$

in terms of the conventional statistic $z = z(\boldsymbol{x}, \mu_0) = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$, the sample size n, and the ratio $\lambda = \sigma^2/\sigma_1^2$ of the model variance to the prior variance. The following disturbing facts may then be established:

(i) As pointed out by Lindley (1957), for *any fixed prior and fixed* $z(\boldsymbol{x}, \mu_0)$, the Bayes factor $B_{01}(z, n, \lambda)$ increases as $\sqrt{n}$ with the sample size, so that 'evidence' in favour of the simpler model M_0 may become overwhelming as the sample size increases, *even for data sets extremely implausible under* M_0, such as those $(\overline{x}, n)$ leading to large $|z|$ values, which are however quite likely under alternative μ values, namely under those close to $\overline{x}$. The same phenomenon is observed for *any* other reasonable choice of the prior $p(\mu)$, including the conventional non-subjective (proper) Cauchy prior suggested by Jeffreys (1961, p. 274). We argue that this is an undesirable behaviour, inconsistent with accepted scientific practice; it may be avoided if posterior probabilities (rather than Bayes factors) are used *and* the prior probability of the null model is made to depend on the sample size n (Bernardo, 1980; Smith and Spiegelhalter, 1980), but this may well be regarded as a rather artificial solution.

(ii) As pointed out by Bartlett (1957), *for any fixed data*, and hence any fixed (z, n), the Bayes factor $B_{01}(z, n, \lambda)$ tends to infinity as σ_1 increases (and hence λ goes to 0), so that 'evidence' in favour of M_0 becomes overwhelming as the prior variance of μ gets large, a situation often thought to describe 'vague prior knowledge' about μ. In particular, this is true for data $(\overline{x}, n)$ such that $|z|$ is large enough to cause the 'null' model M_0 to be rejected at any arbitrarily prespecified level using a conventional frequentist test. Again, qualitatively similar results are obtained for *any* other reasonable choice for the family of priors $p(\mu)$; the

> Bayes factor exhibits an *extreme lack of robustness* with respect to the choice of the prior, and tends to infinity as the prior variance of the mean increases. In particular, no improper prior for μ may be used.

For further discussion of Lindley's paradox, see Smith (1965), Shafer (1982), Berger and Delampady (1987), Berger and Sellke (1987), Consonni and Veronese (1987), Moreno and Cano (1989), Berger and Mortera (1991) and Robert (1993).

Lindley's paradox already suggests that it may not be wise to use Bayes factors in nested hypothesis testing, but there is one further complication. Indeed, it is often argued that, at least in scientific contexts, prior specification should preferably be non-subjective, in the sense that the results obtained should only depend on the data and the models considered. It is also argued that, even when prior information is publicly available, a 'reference' non-subjective solution is necessary to gauge the actual importance of the prior in the final solution. Unfortunately, however, it is well known that one cannot directly use standard 'non-informative' priors in nested hypothesis testing because —contrary to the situation in estimation problems— the arbitrary constants which appear in the typically improper 'non-informative' priors do *not* cancel out and, as a consequence, the resulting Bayes factors are undetermined. The literature contains many attempts to circumvent this difficulty, thus providing some form of non-subjective Bayes factors. Some involve partitions of the sample into a 'training sample' to obtain a proper posterior and an 'effective sample' used to compute the Bayes factor, as in Lempers (1971, Ch. 6), or Berger and Pericchi (1995, 1996) with *intrinsic* Bayes factors; others propose alternative, *ad hoc* devices to 'fix' the arbitrary constants, as in Spiegelhalter and Smith (1982), O'Hagan (1995) with *fractional* Bayes factors, and Robert and Caron (1996) with *neutral* Bayes factors; Aitkin (1991) suggested a non-coherent sample reuse. All these are indeed automatic, non-subjective 'Bayes' factors, which often provide useful large sample approximations; for instance, the *geometric* intrinsic factor of Berger and Pericchi (1996) may be seen as an asymptotic Monte Carlo approximation to a real Bayes factor (Bernardo and Smith, 1994, p. 423). However, the behaviour of these proposals for small samples may be unsatisfactory and, more importantly, these 'Bayes' factors are generally *not* Bayesian, in that they typically do *not* correspond to a Bayesian analysis for any prior (proper or improper). This may have undesirable consequences; for example, as one would expect from the mathematical consistency which drives Bayesian inference, for all models M_1, M_2, M_3, all (proper) priors on their parameters, and any data $\boldsymbol{x}$, one *must* have $B_{12}(\boldsymbol{x}) = B_{21}^{-1}(\boldsymbol{x})$ and $B_{12}(\boldsymbol{x})\, B_{23}(\boldsymbol{x}) = B_{13}(\boldsymbol{x})$, but those minimal coherence requirements are often *not* honored by the proposals mentioned above; for details, see O'Hagan (1997).

One is thus lead to wonder whether the conventional (Bayes factor) formulation of Bayesian hypothesis testing may always be appropriate. In this paper, it is argued that nested hypothesis testing problems are better described as specific decision problems about the choice of a *useful* model and that, when formulated within the framework of decision theory, they do have a natural, fully Bayesian, coherent solution. Moreover, within such a formulation, *reference analysis* (Bernardo, 1979b; Berger and Bernardo, 1989, 1992) may successfully be used to provide a *non-subjective* Bayesian solution, which is consistent with accepted scientific practice. In Section 2, nested hypothesis testing is formally described as a precise decision problem, where the *terminal* utility function takes the form of a *proper scoring rule*. In Section 3, reference analysis and accepted scientific practice in a canonical situation, are respectively used to motivate the choice of the prior distribution and the choice of the utility threshold; as a consequence, a precise procedure for nested hypothesis testing, the *Bayesian Reference Criterion* (BRC), is formally proposed. In Section 4, the behaviour of BRC is explored in simple stylized examples, where alternative approaches may easily be compared.

2. NESTED HYPOTHESIS TESTING AS A DECISION PROBLEM

Let $\boldsymbol{x} \in \boldsymbol{X}$ be some available data, whose probabilistic behaviour is assumed to be appropriately described by the probability model $p_{\boldsymbol{x}}(. \mid \theta, \omega)$, $\theta \in \Theta$, $\omega \in \Omega$, and suppose that it is desired to 'test' whether or not those data are compatible with the 'null value' $\theta = \theta_0$, that is, whether, assuming that $p_{\boldsymbol{x}}(. \mid \theta, \omega)$ is appropriate, one could actually use a model of the form $p_{\boldsymbol{x}}(. \mid \theta_0, \omega_0)$, for some $\omega_0 = \omega_0(\theta_0, \theta, \omega) \in \Omega$ to be specified. Typically, the data $\boldsymbol{x}$ will consist of a random sample $\{x_1, \ldots, x_n\}$ from some model $p_x(. \mid \theta, \omega)$, but we will *not* need to make such an assumption. The problem proposed may formally be described as a decision problem with only two alternative strategies, namely

$$\begin{cases} a_0 = \text{for some } \omega_0(\theta_0, \theta, \omega) \in \Omega, \text{ act } \textit{as if} \text{ data were generated from } p_{\boldsymbol{x}}(. \mid \theta_0, \omega_0), \\ a_1 = \text{keep the assumed model } p_{\boldsymbol{x}}(. \mid \theta, \omega), \quad \theta \in \Theta, \quad \omega \in \Omega. \end{cases}$$

For coherent behaviour, it is then necessary (i) to specify a utility function $u(a_i, \theta, \omega)$ measuring the conditional desirability of each of those two possible decisions as a function of the parameter values (θ, ω), (ii) to specify a prior distribution $p(\theta, \omega)$ describing available prior information about those unknown parameters, and (iii) to choose that decision a_i which maximizes the corresponding posterior expected utility $\overline{u}(a_i \mid \boldsymbol{x})$.

It is known (Bernardo, 1979a, Bernardo and Smith 1994, Sec. 2.7 and 3.4) that 'pure' scientific inference about some random quantity $\boldsymbol{\phi}$ may formally be described as a decision problem where the decision space is the class $\{q_{\boldsymbol{\phi}}(.)\}$ of strictly positive probability densities of $\boldsymbol{\phi}$ with respect to some dominating measure, and where the utility function is a logarithmic (proper) score function of the form $u(q_{\boldsymbol{\phi}}(.), \boldsymbol{\phi}) = \alpha \log q_{\boldsymbol{\phi}}(\boldsymbol{\phi}) + \beta(\boldsymbol{\phi})$. Using model $q_{\boldsymbol{x}}(.)$ to describe the behaviour of $\boldsymbol{x}$ may be seen as an inference statement about the random quantity $\boldsymbol{x}$; thus, the utility of using some model $q_{\boldsymbol{x}}(.)$ with data $\boldsymbol{x}$ could reasonably be assumed to be of the form $u(q_{\boldsymbol{x}}(.), \boldsymbol{x}) = \alpha \log q_{\boldsymbol{x}}(\boldsymbol{x}) + \beta(\boldsymbol{x})$ and therefore, before the data $\boldsymbol{x}$ are actually observed, the expected utility of using some *parametric model* $q_{\boldsymbol{x}}(. \mid \theta, \omega)$, with data actually generated from $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$, will be of the form

$$u[q_{\boldsymbol{x}}(. \mid \theta, \omega), \theta, \omega] = \alpha \int p_{\boldsymbol{x}}(\boldsymbol{x} \mid \theta, \omega) \log[\, q_{\boldsymbol{x}}(\boldsymbol{x} \mid \theta, \omega)] \, d\boldsymbol{x} + \beta(\theta, \omega), \quad \alpha > 0, \qquad (1)$$

for some function $\beta(\theta, \omega) = \int \beta(\boldsymbol{x}) \, p_{\boldsymbol{x}}(\boldsymbol{x} \mid \theta, \omega) \, d\boldsymbol{x}$, which will turn out to be irrelevant. Moreover, there *must* be a definite advantage of using the simpler model when it is appropriate for, otherwise, one would always use the full model, which is *assumed to be appropriate*. This may be due, for instance, to the mathematical simplicity of M_0, or to the existence of a scientific theory which supports the simpler model M_0. Using the terminology introduced by Raiffa and Schlaifer (1961), we will further assume that the *utility* of using a model M_i, $i \in \{0, 1\}$, may be additively decomposed into the *terminal utility* of M_i, which measures its conditional value to explain the data, and the *cost* c_i to be expected from using M_i, taking into account its simplicity, scientific implications, or any other considerations; under this assumption, one *must* have $c_1 > c_0$. Thus, dropping the subindices from the densities to simplify the notation, a sensible utility structure for the proposed decision problem is

$$u(a_0, \theta, \omega) = \sup_{\omega_0 \in \Omega} \alpha \int p(\boldsymbol{y} \mid \theta, \omega) \log[p(\boldsymbol{y} \mid \theta_0, \omega_0)] \, d\boldsymbol{y} + \beta(\theta, \omega) - c_0,$$

$$u(a_1, \theta, \omega) = \alpha \int p(\boldsymbol{y} \mid \theta, \omega) \log[p(\boldsymbol{y} \mid \theta, \omega)] \, d\boldsymbol{y} + \beta(\theta, \omega) - c_1,$$

where $\omega_0 = \omega_0(\theta_0, \theta, \omega)$ specifies the best approximation to the assumed model under the null, and the (dummy) variable $\boldsymbol{y}$ is used to denote data obtained from the full model $p_{\boldsymbol{x}}(. \mid \theta, \omega)$. It

immediately follows that, given data $\boldsymbol{x}$, the best action is to keep the full model if, and only if, $\overline{u}(a_1 \,|\, \boldsymbol{x}) > \overline{u}(a_0 \,|\, \boldsymbol{x})$. The difference between these expected utilities,

$$\begin{aligned}\overline{u}(a_1 \,|\, \boldsymbol{x}) - \overline{u}(a_0 \,|\, \boldsymbol{x}) &= \alpha \iint \Big[u(a_1, \theta, \omega) - u(a_0, \theta, \omega) \Big] \, p(\theta, \omega \,|\, \boldsymbol{x}) \, d\theta d\omega \\ &= \alpha \iint \Big[\inf_{\omega_0 \in \Omega} \int p(\boldsymbol{y} \,|\, \theta, \omega) \log \frac{p(\boldsymbol{y} \,|\, \theta, \omega)}{p(\boldsymbol{y} \,|\, \theta_0, \omega_0)} \, d\boldsymbol{y} + c_0 - c_1 \Big] \, p(\theta, \omega \,|\, \boldsymbol{x}) \, d\theta d\omega,\end{aligned}$$

may therefore be usefully reexpressed as

$$\overline{u}(a_1 \,|\, \boldsymbol{x}) - \overline{u}(a_0 \,|\, \boldsymbol{x}) = \alpha \; d(\boldsymbol{x}) - (c_1 - c_0), \tag{2}$$

where

$$d(\boldsymbol{x}, \theta_0) = \iint \delta(\theta_0, \theta, \omega) \, p(\theta, \omega \,|\, \boldsymbol{x}) \, d\theta d\omega \tag{3}$$

is the posterior expected value of

$$\delta(\theta_0, \theta, \omega) = \inf_{\omega_0 \in \Omega} \int p(\boldsymbol{y} \,|\, \theta, \omega) \log \frac{p(\boldsymbol{y} \,|\, \theta, \omega)}{p(\boldsymbol{y} \,|\, \theta_0, \omega_0)} \, d\boldsymbol{y}. \tag{4}$$

The non-negative quantity $\delta(\theta_0, \theta, \omega)$ has several interesting interpretations. Indeed, it may simply be described as the *expected value* (under the assumed model) of the *log-likelihood ratio* of the assumed model to its closest approximation under the null; but it also measures the *minimum amount of information* which would be necessary to recover M_1 from M_0 (Kullback and Leibler, 1951), so that the utility constant α may actually be interpreted as the *value of one unit of information* about data generated from $p_{\boldsymbol{x}}(\cdot \,|\, \theta, \omega)$.

It follows from (2) that, in the stylized purely inferential situation described by the utility function (1), the decision criterion *must* be of the form

$$\textit{Reject} \text{ the null model } M_0 \text{ if, and only if,} \quad d(\boldsymbol{x}, \theta_0) > g, \qquad g = (c_1 - c_0)/\alpha, \tag{5}$$

where the utility ratio g is the *only* number which must be assessed for a complete specification of the utility structure.

Noting that the logarithmic discrepancy $\delta(\theta_0, \theta, \omega)$ is non-negative and vanishes if $\theta = \theta_0$, one has $\overline{u}(a_0 \,|\, \boldsymbol{x}, M_0) - \overline{u}(a_1 \,|\, \boldsymbol{x}, M_0) = c_1 - c_0$; thus, since α is the value of one unit of information, it follows that g is a strictly positive constant which measures, *in information units*, the *expected utility gain* from using the null model M_0 when it is true.

Summarizing, we have found that the utility structure of the stylized decision problem which describes nested hypothesis testing only depends on the unknown parameters through the corresponding logarithmic discrepancy $\delta(\theta_0, \theta, \omega)$, which therefore is the *quantity of interest*. As a consequence, deciding whether or not the simpler model M_0 has to be rejected as an acceptable proxy for the full model M_1 is reduced to the much simpler problem of deciding whether or not $d(\boldsymbol{x}, \theta_0)$, the posterior expectation of $\delta(\theta_0, \theta, \omega)$ is —or is not— too large.

The idea of using some form of the logarithmic discrepancy in model selection has a long history, pioneered by Good (1950) and Kullback (1959). The use of *some* posterior expected value of the logarithmic discrepancy as the basic 'test' statistic for Bayesian model selection was originally proposed by Bernardo (1982, 1985), and further developed by Bernardo and Bayarri (1985), Ferrándiz (1985), Bayarri (1987), Gutiérrez-Peña (1992), and Rueda (1992), using conventional non-subjective priors. However, both the appropriate choice of the prior and the specification of the threshold utility value g —which are crucial for any practical implementation of the idea— remained open. We now turn to propose a choice for these two elements, which is consistent with accepted scientific practice.

3. STANDARDISATION: THE BAYESIAN REFERENCE CRITERION

3.1. *The Choice of the Prior Distribution*

It has been often recognised that in scientific inference there is a pragmatically important need for a form of non-subjective, model based prior, which has a minimal effect, relative to the data, on the posterior inference. The use of non-subjective priors has been criticized by subjectivist Bayesians, who argue that the prior should be an honest expression of the analyst's prior knowledge and not a function of the model. However, non-subjective posteriors may be seen as an important element of the *sensitivity analysis* to assess the changes in the posterior of interest induced by changes in the prior which should be part of any good subjective Bayesian analysis: a non-subjective posterior tries to give an answer to the question of what *could* be said about the quantity of interest, *if* one's prior knowledge *about that quantity* were dominated by the data. In the long quest for these "baseline" non-subjective posterior distributions, a number of requirements have emerged which may reasonably be regarded as necessary properties of the proposed algorithm. These include invariance, consistent marginalization, consistent sampling properties, general applicability and limiting admissibility. The *reference analysis* algorithm, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive non-subjective posterior distributions which satisfy all these desiderata; and it is found that, within a given model, the appropriate joint *reference* prior *depends on the quantity of interest*. For a recent discussion of the many polemic issues involved in this topic, see Bernardo (1997). For an introduction to reference analysis, see Bernardo and Smith (1994, Ch. 5), or Bernardo and Ramón (1998).

The solution to *any* decision problem, conditional on data $\boldsymbol{x}$ for which a probability model $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$ has been assumed, only depends on $\boldsymbol{x}$ through the posterior expectation of some function of the parameters, which *defines* the *quantity of interest* in that decision problem. In our formulation of nested hypothesis testing, the decision criterion only depends on the data through the expected value of the non-negative function $\delta = \delta(\theta_0, \theta, \omega)$, which is therefore the relevant quantity of interest. Thus, we propose to use the *reference* prior $\pi_\delta(\theta, \omega)$ of (θ, ω) which corresponds to the quantity of interest $\delta = \delta(\theta_0, \theta, \omega)$. Consequently, to decide whether or not M_0 is an acceptable proxy to M_1, we propose to evaluate the *reference* posterior expectation of the logarithmic discrepancy $\delta(\theta_0, \theta, \omega)$,

$$d_r(\boldsymbol{x}, \theta_0) = \int \delta(\theta_0, \theta, \omega)\, \pi_\delta(\theta, \omega \mid \boldsymbol{x})\, d\theta d\omega,$$

where $\pi_\delta(\theta, \omega \mid \boldsymbol{x})$ is the posterior distribution which corresponds to the reference prior $\pi_\delta(\theta, \omega)$, and the suffix r in the resulting statistic $d_r(\boldsymbol{x}, \theta_0)$ indicates that expectation is taken with respect to the *reference* posterior. The 'test statistic', $d_r(\boldsymbol{x}, \theta_0)$ encapsulates all relevant information from the data; thus, the simpler model M_0 should be rejected if, and only if, $d_r(\boldsymbol{x}, \theta_0) \geq g$, that is if, and only if, the reference expected posterior discrepancy is larger than a utility constant g, which measures (in information units) the expected utility gain from using the null model when it is true.

We note that $d_r(\boldsymbol{x}, \theta_0)$ remains invariant if a sufficient statistic $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{x})$ is used instead of the full data; indeed, if $\boldsymbol{x} = \{\boldsymbol{s}, \boldsymbol{r}\}$, one could write $\delta(\theta_0, \theta, \omega)$ as

$$\iint p(\boldsymbol{s} \mid \theta, \omega)\, p(\boldsymbol{r} \mid \boldsymbol{s}) \log \frac{p(\boldsymbol{s} \mid \theta, \omega)\, p(\boldsymbol{r} \mid \boldsymbol{s})}{p(\boldsymbol{s} \mid \theta_0, \omega_0)\, p(\boldsymbol{r} \mid \boldsymbol{s})}\, d\boldsymbol{r}\, d\boldsymbol{s} = \int p(\boldsymbol{s} \mid \theta, \omega) \log \frac{p(\boldsymbol{s} \mid \theta, \omega)}{p(\boldsymbol{s} \mid \theta_0, \omega_0)}\, d\boldsymbol{s}$$

and, *a fortiori*, its expected value will remain invariant. Moreover, $d_r(\boldsymbol{x}, \theta_0)$ also remains invariant under one-to-one transformations of the parameters; indeed,

$$d_r(\boldsymbol{x}, \theta_0) = \iint \delta(\theta_0, \theta, \omega) \pi_\delta(\theta, \omega \mid \boldsymbol{x})\, d\theta d\omega = \iint \delta\Big(\phi(\theta_0), \phi(\theta), \omega(\omega)\Big)\, \pi_\delta(\phi, \omega \mid \boldsymbol{x})\, d\phi d\omega.$$

We notice that if the data consist of a *random sample* $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n from some underlying model $p_x(\cdot \,|\, \theta, \omega)$, then the logarithmic discrepancy simply becomes

$$\begin{aligned}\delta(\theta_0, \theta, \omega) &= \inf_{\omega_0 \in \Omega} \int p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta, \omega) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta, \omega)}{p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta_0, \omega_0)} \, d\boldsymbol{y} \\ &= n \inf_{\omega_0 \in \Omega} \int p_x(y \,|\, \theta, \omega) \log \frac{p_x(y \,|\, \theta, \omega)}{p_x(y \,|\, \theta_0, \omega_0)} \, dy.\end{aligned}$$

Observe, however, that this exchangeability assumption is *not* necessary to implement the methodology we are proposing.

Reference analysis has suggested a precise choice for the prior. To complete the specification of the decision problem, we now turn to consider the choice of the utility constant g.

3.2. *Calibration of the Utility Function*

Measuring is comparing with a standard. An *operational* definition of *any* form of quantification requires a standard *unit of measurement*, such as the *metre* for measuring lengths, or the *standard events* to measure probabilities. To define an appropriate utility threshold for model evaluation, we will use as our 'unit' a canonical example in standard scientific practice.

Under approximate normality, there seems to be a general agreement among scientists about the use of two standard error deviations as a signal of mild evidence against the null and three standard error deviations as a signal of significant evidence (see *e.g.*, Jaynes, 1980, p. 634, or Jeffreys 1980, p. 453). For a formal robust Bayesian justification of this practice, see Berger and Sellke (1987) and Berger and Delampady (1987).

In the situation already discussed, when it is desired to test the hypothesis $\mu = \mu_0$ given n normal observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, with unknown mean μ but *known* standard deviation σ, accepted scientific practice reduces to computing $z = z(\boldsymbol{x}, \mu_0) = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$ and rejecting the null if $|z| > c$, where c is typically chosen to be around 2 or 3. In frequentist terms, this corresponds, for $c = 1.96$ or $c = 3.00$, to rejecting the null when it is appropriate with probability not larger than 0.05 or 0.0027, respectively. In Bayesian terms, this corresponds to using the conventional uniform prior for estimating μ, and rejecting the null when μ_0 does not belong to the corresponding HPD intervals with posterior probabilities 0.95 or 0.9973, respectively; note that this Bayesian procedure does *not* use a singular prior and, hence, avoids Lindley's paradox. Precisely the same results are also obtained in this problem from the fiducial, the likelihood or the pivotal viewpoints. Thus, on this *canonical* example, there appears to be a basically universal *consensus* on what an appropriate procedure should be doing, with the remarkable exception of Bayes factors based on singular priors.

Consider now, for this canonical example, the logarithmic discrepancy of the null model $M_0 \equiv \mathrm{N}(x \,|\, \mu_0, \sigma)$ from the full model $M_1 \equiv \mathrm{N}(x \,|\, \mu, \sigma)$ which, assuming σ known, is

$$\delta(\mu_0, \mu, \sigma) = n \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu, \sigma) \log \frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \mu_0, \sigma)} \, dx = \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2. \tag{6}$$

Here, μ is the only unknown quantity and, as one might expect, the reference prior of μ when $\delta = n(\mu - \mu_0)^2/\sigma^2$ is the quantity of interest is the conventional uniform prior $\pi_\delta(\mu) = 1$. Hence, the corresponding posterior distribution of μ is $\pi_\delta(\mu \,|\, \boldsymbol{x}, \sigma) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})$, where $\overline{x}$ is the sample mean, and the *reference* expected posterior discrepancy is

$$d_r(\boldsymbol{x}, \mu_0) = \int_{-\infty}^{\infty} \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n}) \, d\mu = \tfrac{1}{2} \Big[1 + n \Big(\frac{\overline{x} - \mu_0}{\sigma} \Big)^2 \Big] = \tfrac{1}{2}(1 + z^2),$$

a one-to-one function of the 'consensus' test statistic $|z| = |\overline{x} - \mu_0|/(\sigma/\sqrt{n})$. Hence, the decision criterion becomes

$$\textit{Reject } \mu = \mu_0 \text{ if, and only if, } \quad d_r(\boldsymbol{x}, \mu_0) = \tfrac{1}{2}[1 + z^2] > g, \tag{7}$$

for some appropriately chosen utility constant g. As described before, accepted practice in this example suggests rejecting the null if $|z| > c$ for some c, usually chosen to be around 2 or 3; it follows from (7) that $|z| > 2$ when $d_r > 2.5$ and $|z| > 3$ when $d_r > 5$. Moreover, since the sampling distribution of $z = z(\boldsymbol{x}, \mu_0)$ is normal, centered at $(\mu - \mu_0)/(\sigma/\sqrt{n})$ and with standard deviation equal to one, the sampling distribution of z^2 is the non-central χ^2 with one degree of freedom and non-centrality parameter 2δ, where $\delta = \delta(\mu_0, \mu, \sigma)$ is given by (6) and, therefore, $\mathrm{E}[z^2 \,|\, \delta] = 1 + 2\delta$. It follows that when M_0 is true, and thus $\delta = 0$, one has $\mathrm{E}[z^2 \,|\, M_0] = 1$. Furthermore, if M_0 is *not* true, then δ is strictly positive and increases linearly with n; thus, the expected value of z^2 will then tend to infinity as the sample size increases. Since $d_r(\boldsymbol{x}, \mu_0)$ is a one-to-one function of z^2 this ensures a sensible large sample behaviour of the proposed procedure in the sense that

$$\mathrm{E}_{\boldsymbol{x} \,|\, \mu,\sigma}[d_r(\boldsymbol{x}, \mu_0) \,|\, M_0] = 1, \qquad \lim_{n\to\infty} \mathrm{E}_{\boldsymbol{x} \,|\, \mu,\sigma}[d_r(\boldsymbol{x}, \mu_0) \,|\, M_1 \cap \overline{M_0}] = \infty.$$

These results complete our motivation for the decision criterion being proposed.

3.3. *The Bayesian Reference Criterion*

The Bayesian Reference Criterion (BRC). *To decide whether or not some data $\boldsymbol{x}$ are compatible with the (null) hypothesis $\theta = \theta_0$, assuming that the data have been generated from the model $p_{\boldsymbol{x}}(\cdot \,|\, \theta, \omega)$, $\theta \in \Theta$, $\omega \in \Omega$:*

(i) compute the logarithmic discrepancy,

$$\delta(\theta_0, \theta, \omega) = \inf_{\omega_0 \in \Omega} \int p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta, \omega) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta, \omega)}{p_{\boldsymbol{x}}(\boldsymbol{y} \,|\, \theta_0, \omega_0)} \, d\boldsymbol{y},$$

between the assumed model and its closest approximation under the null.

(ii) derive the corresponding reference posterior expectation

$$d_r(\boldsymbol{x}, \theta_0) = \iint \delta(\theta_0, \theta, \omega) \, \pi_\delta(\theta, \omega \,|\, \boldsymbol{x}) \, d\theta d\omega;$$

(iii) for some d^*, reject the hypothesis $\theta = \theta_0$ if, and only if, $d_r(\boldsymbol{x}, \theta_0) > d^*$, where values such as $d^* = 2.5$ (mild evidence against θ_0) or $d^* = 5$ (significant evidence against θ_0) may conveniently be chosen for scientific communication.

The choice of d^* is formally determined by the *utility gain* which may be expected by using the null model when it is true; the larger that gain, the larger d^*. The analysis above suggests that a value $d_r(\boldsymbol{x}, \theta_0)$ close to 1 may be expected if M_0 is true, and scientific practice suggests that d_r-values over 2.5 should raise some doubts on the use of M_0, and that d_r-values over 5 should typically be regarded as significant evidence against the suitability of using M_0 as a proxy to M_1.

If $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a sufficiently *large* random sample from a *regular* model $p(x \,|\, \theta, \omega)$, the posterior distribution of (θ, ω) will concentrate on their maximum likelihood estimates $(\hat{\theta}, \hat{\omega})$, and thus the expected posterior discrepancy, $d_r(\boldsymbol{x}, \theta_0)$, will be close to $\delta(\theta_0, \hat{\theta}, \hat{\omega})$, the logarithmic discrepancy between the model identified by $(\hat{\theta}, \hat{\omega})$ and its closest approximation under the

null. Moreover, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from a model $p_x(x \mid \theta)$, where θ is one-dimensional and there are no nuisance parameters, then $\delta(\theta_0, \theta)$ will typically be a piecewise invertible function of θ and hence (see Proposition 1 in the Appendix) the relevant reference prior will simply be Jeffreys' prior, that is $\pi_\delta(\theta) \propto i(\theta)^{1/2}$, where $i(\theta)$ is Fisher's information function. Thus, in terms of the *natural* parametrization, defined as $\phi = \phi(\theta) = \int^\theta i(\theta)^{1/2} d\theta$, the reference prior $\pi_\delta(\phi)$ will be uniform. For large sample sizes, the corresponding reference posterior distribution of ϕ will then be approximately normal $\pi_\delta(\phi \mid \boldsymbol{x}) \approx \mathrm{N}(\phi \mid \hat{\phi}, 1/\sqrt{n})$, and will only depend on the data through its mle $\hat{\phi}$; moreover, the sampling distribution of $\hat{\phi}$, $p(\hat{\phi} \mid \phi)$ will also be approximately normal, $\mathrm{N}(\hat{\phi} \mid \phi, 1/\sqrt{n})$. Since the discrepancy function is invariant under one-to-one reparametrization, and hence $\delta(\phi_0, \phi) = \delta(\theta_0, \theta)$, one obtains, after some algebra,

$$d_r(\boldsymbol{x}, \theta_0) \approx \tfrac{1}{2}\left[1 + z^2(\hat{\theta}, \theta_0)\right], \quad z(\hat{\theta}, \theta_0) = \sqrt{n}\,[\phi(\hat{\theta}) - \phi(\theta_0)]\,. \tag{8}$$

This type of approximation may be extended to multivariate situations, with or without nuisance parameters; this provides a link to both Akaike's (1973, 1974) AIC, and Schwarz's (1978) BIC criteria. The results will be reported elsewhere.

4. EXAMPLES

4.1. *Testing a Normal Mean Value with Known Variance*

Let us first reconsider our canonical example. In this case, μ is the only unknown parameter and the logarithmic discrepancy is $\delta = n\theta^2$, with $\theta = (\mu - \mu_0)/\sigma$. It is well known that, when μ is the quantity of interest, the reference prior is the (improper) uniform prior $\pi_\mu(\mu) = 1$; moreover, given σ, θ is a one-to-one function of μ and, hence, from the invariance properties of the reference algorithm, the reference prior when θ is the quantity of interest is also $\pi_\theta(\mu) = 1$; besides, since θ^2 is piecewise invertible, if follows from Proposition 1 in the Appendix that the reference prior when θ^2 is the quantity of interest is still uniform and, since δ is a one-to-one function of θ^2, one finally has that the reference prior when δ is the quantity of interest is indeed the conventional uniform prior $\pi_\delta(\mu) = 1$. It follows that the corresponding posterior distribution of μ is $\pi_\delta(\mu \mid \boldsymbol{x}) = \mathrm{N}(\mu \mid \overline{x}, \sigma/\sqrt{n})$ and thus, as anticipated in Section 3, $d_r(\boldsymbol{x}, \mu_0) = (1+z^2)/2$ a one-to-one function of the 'consensus' test statistic $|z|$, where $z = \sqrt{n}(\mu_0 - \overline{x})/\sigma$.

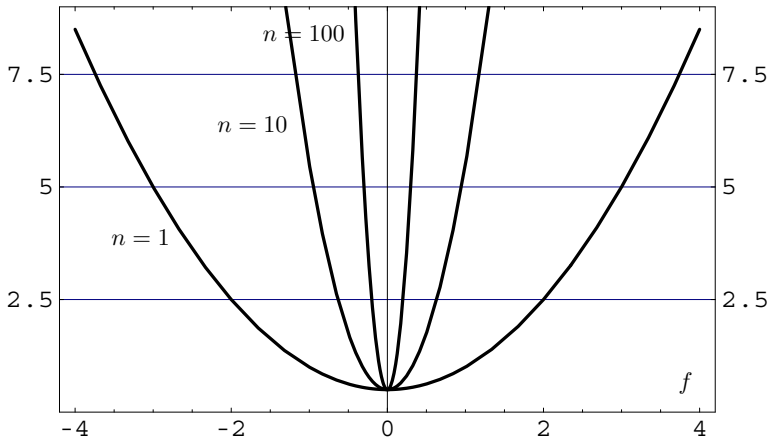

Figure 1. *Normal observations (known variance). Behaviour of the test statistic $d_r(\boldsymbol{x}, \mu_0) = d_r(f, n)$, as a function of the standardized distance $f = (\mu_0 - \overline{x})/\sigma$, for sample sizes $n = 1$, $n = 10$ and $n = 100$.*

Figure 1 describes, as a function of $f(\overline{x}) = (\mu_0 - \overline{x})/\sigma$ and the sample size n, the behaviour of $d_r(\boldsymbol{x}, \mu_0) = d_r(f, n)$. As one would expect, rejection —large $d_r(\boldsymbol{x}, \mu_0)$ values— is indicated

for progressively smaller values of f as n increases; indeed, as the sample size increases, one would require the standardized distance between $\overline{x}$ and μ_0 to decrease in order to accept working *as if* data had been generated with $\mu = \mu_0$.

Table 1. *Normal observations (known variance). Correspondence between the threshold value d^* of the test statistic $d_r(\boldsymbol{x}, \mu_0)$, and 'type 1' error probabilities.*

d^*	$P[d_r > d^* \mid \mu = \mu_0]$	d^*	$P[d_r > d^* \mid \mu = \mu_0]$
1.85277	0.10000	1.00	0.31731
2.42073	0.05000	2.00	0.08326
3.81745	0.01000	3.00	0.02535
4.43972	0.00500	4.00	0.00815
5.91378	0.00100	5.00	0.00270
6.55783	0.00050	6.00	0.00091
8.06835	0.00010	7.00	0.00031
8.72406	0.00005	8.00	0.00011
10.2557	0.00001	9.00	0.00004

The frequentist behaviour of the proposed test under the null is easily found. Indeed, if $\mu = \mu_0$, then the sampling distribution of $\overline{x}$ is $\mathrm{N}(\overline{x} \mid \mu_0, \sigma/\sqrt{n})$ and therefore, under M_0, $z^2 \sim \chi^2_1$ so that, the 'type 1' error probabilities $\Pr[d_r(\boldsymbol{x}, \mu_0) > d^* \mid \mu = \mu_0]$ are given, as a function of the threshold value d^*, by $\Pr[\chi^2_1 > 2d^* - 1]$. In particular, with the choice $d^* = 5$ the type 1 error probability is 0.0027 while, with $d^* = 2.42073$ it is the ubiquitous 0.05; Table 1 gives other values. As one would surely expect in this 'consensus' example, we here obtain, *for all sample sizes,* a one-to-one correspondence between d^*-values and frequentist significance levels. It is easily seen, however, that this *exact* correspondence is generally *not* to be expected.

4.2. *Testing an Exponential Parameter Value*

We now consider a simple non-normal problem with continuous data. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, be a random sample of exponential observations with parameter θ, so that $p(\boldsymbol{x} \mid \theta) = \theta^n \exp[-n\overline{x}\theta]$, and the sample mean $\overline{x}$ is sufficient. To test whether or not the value $\theta = \theta_0$ is compatible with those observations, we first derive the corresponding logarithmic discrepancy,

$$\delta(\theta_0, \theta) = n\Big[\int_0^\infty \theta e^{-\theta x} \log \frac{\theta e^{-\theta x}}{\theta_0 e^{-\theta_0 x}}\, dx\Big] = n\Big[\frac{\theta_0}{\theta} - 1 - \log \frac{\theta_0}{\theta}\Big] .$$

This is a piecewise invertible function of θ and it is known, (see *e.g.*, Bernardo and Smith, 1994, p. 438) that the reference posterior distribution of θ is $\pi(\theta \mid \boldsymbol{x}) \propto \theta^{n-1} e^{-n\overline{x}\theta}$, a Gamma distribution $\mathrm{Ga}(\theta \mid n, n\overline{x})$, with a unique *mode* at $\tilde{\theta} = (n-1)/n\overline{x}$, whenever $n > 1$. Using the fact that if θ has a $\mathrm{Ga}(\theta \mid \alpha, \beta)$ distribution, then $\mathrm{E}[\log \theta] = \psi(\alpha) - \log \beta$, where $\psi(x)$ is the digamma function, the reference posterior expectation of the logarithmic discrepancy is found to be

$$d_r(\boldsymbol{x}, \theta_0) = n\left[\psi(n) - \log(n-1) + \frac{\theta_0}{\tilde{\theta}} - 1 - \log \frac{\theta_0}{\tilde{\theta}}\right], \quad n \geq 2.$$

Note that $d_r(\boldsymbol{x}, \theta_0)$ only depends on the data through the ratio $\theta_0/\tilde{\theta}$ and that the procedure suggests that no testing of the parameter value is possible in the exponential model with only one observation. Using Stirling's approximation for the digamma function, it is easily verified that, for large sample sizes, the expected posterior discrepancy is approximately given by $d_r(\boldsymbol{x}, \theta_0) \approx \delta(\theta_0, \tilde{\theta})$, the discrepancy of the model identified by θ_0 from the model identified

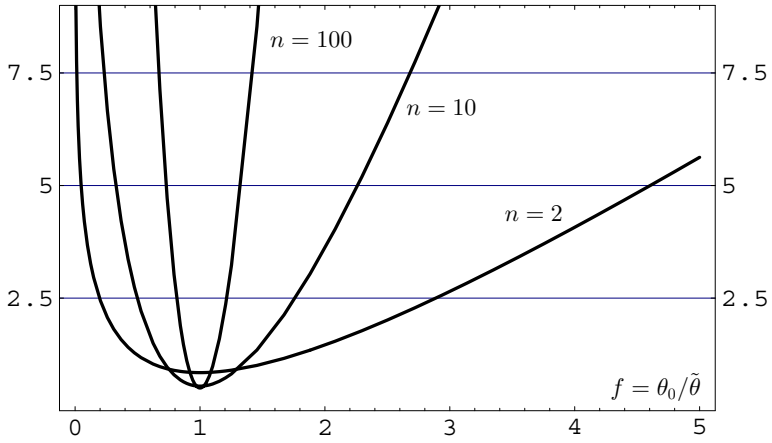

Figure 2. *Exponential observations. Exact behaviour of the test statistic $d_r(\boldsymbol{x}, \theta_0) = d_r(r, n)$, as a function of the ratio $f = \theta_0/\tilde{\theta}$, for sample sizes $n = 2$, $n = 10$ and $n = 100$.*

by $\tilde{\theta}$, and also by $d_r(\boldsymbol{x}, \theta_0) \approx \frac{1}{2}(1 + z^2)$, with $z = z(\boldsymbol{x}, \theta_0) = \sqrt{n}\, \log(\theta_0/\tilde{\theta})$, which is the approximation given by (8).

Figure 2 describes, for several sample sizes, the exact behaviour of $d_r(\boldsymbol{x}, \theta_0)$, as a function of the ratio $f = \theta_0/\tilde{\theta}$, and the sample size n. As one would expect, to accept the value $\theta = \theta_0$, the ratio f has to be progressively close to 1 as n increases.

Table 2. *Exponential times. Correspondence between the threshold value d^* of the test statistic $d_r(\boldsymbol{x}, \theta_0)$, and 'type 1' error probabilities, $P[d_r > d^* \mid H_0]$, for sample sizes 2, 10, 100 and 1000.*

d^*	$n = 2$	$n = 10$	$n = 100$	$n = 1000$
1.0000	0.71695	0.37004	0.32219	0.31780
2.0000	0.32020	0.10885	0.08552	0.08349
2.4207	0.24502	0.06867	0.05161	0.05016
3.0000	0.17325	0.03726	0.02634	0.02544
4.0000	0.09844	0.01347	0.00857	0.00819
5.0000	0.05723	0.00511	0.00287	0.00272
6.0000	0.03370	0.00190	0.00098	0.00092
7.0000	0.01193	0.00028	0.00012	0.00011
9.0000	0.00714	0.00011	0.00004	0.00004

The exact frequentist behaviour of the proposed test under the null may be obtained from the fact that if x has an exponential sampling distribution with parameter θ, then $\overline{x}$ has a Gamma sampling distribution, $\text{Ga}(\overline{x} \mid n, n\theta)$ and, therefore, $y = \theta/\tilde{\theta}$ has a Gamma sampling distribution $\text{Ga}(y \mid n, n-1)$. Table 2 reproduces the results obtained for several sample sizes. As could be expected from the asymptotic results described above, the frequentist behaviour observed for large samples is similar to that obtained for testing a normal mean value, encapsulated in Table 1, hence providing *asymptotic* agreement with frequentist hypothesis testing. Note, however, that there is not anymore a one-to-one correspondence between d^*-values and significance levels; indeed, our procedure recommends rejecting the null whenever $d_r > 5$, which implies 'type 1' error probabilities of 0.0572, 0.0051, 0.0029 and 0.0027 when the sample size is, respectively, 2, 10 100 and 1000; this is in agreement with the popular belief on decreasing the significance levels as the sample size increases.

4.3. *Testing a Binomial Parameter Value*

The proposed procedure is easily applied to *discrete* data, with none of the problems that plague frequentist hypothesis testing in that case. As an example, we will now consider the binomial case. Thus, let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, be a random sample of n Bernoulli observations with parameter θ, so that $p(\boldsymbol{x} \,|\, \theta) = \theta^r(1-\theta)^{n-r}$, and the number of successes, $r = \sum x_j$ is sufficient. To test whether or not the value $\theta = \theta_0$ is compatible with those observations, we have to derive the reference posterior expectation of the corresponding logarithmic discrepancy,

$$\delta(\theta_0, \theta) = n\Big[\theta \log \frac{\theta}{\theta_0} + (1-\theta) \log \frac{1-\theta}{1-\theta_0}\Big] \,.$$

This is a piecewise invertible function of θ, and it is known (see *e.g.*, Bernardo and Smith, 1994, p. 436), that the reference posterior of θ is a Beta distribution $\text{Be}(\theta \,|\, r + 1/2, n - r + 1/2)$, whose expected value is $\overline{\theta} = (r + 1/2)/(n + 1)$. Using the fact that, if θ has a $\text{Be}(\theta \,|\, \alpha, \beta)$ distribution, then $\text{E}[\theta \log \theta] = \alpha(\alpha+\beta)^{-1}[\psi(\alpha+1) - \psi(\alpha+\beta+1)]$, one finds

$$\begin{aligned} d_r(\boldsymbol{x}, \theta_0) = \int \delta(\theta_0, \theta)\, \pi_\delta(\theta \,|\, \boldsymbol{x})\, d\theta = d_r(\overline{\theta}, n) = n\overline{\theta}\Big[\psi\Big(1 + (n+1)\overline{\theta}\Big) - \log \theta_0\Big] \\ + n(1-\overline{\theta})\Big[\psi\Big(1 + (n+1)(1-\overline{\theta})\Big) - \log(1-\theta_0)\Big] - n\psi(n+2). \end{aligned}$$

Figure 3 describes, as a function of the *discrete* variable $\overline{\theta}$, and the sample size n, the exact behaviour of $d_r(\boldsymbol{x}, \theta_0)$, for $\theta_0 = 1/5$, and several sample sizes. As one would expect, no parameter value may be rejected with only a few observations; moreover, rejection is indicated for values of $\overline{\theta}$ increasingly close to θ_0 as n increases; indeed, as the sample size becomes larger, one would require $\overline{\theta}$ to be progressively close to θ_0 in order to accept the value $\theta = \theta_0$; for example, with $n = 5$, $\theta_0 = 1/5$ is only rejected ($d_r > 5$) if $r = 5$ while, with $n = 10$, it is rejected whenever $r \geq 7$.

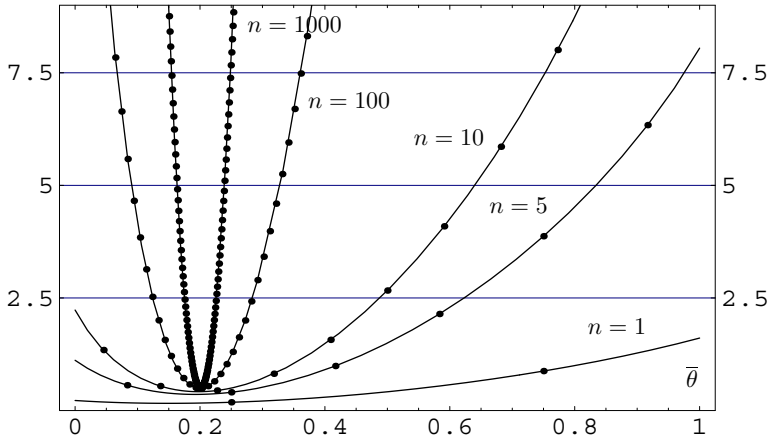

Figure 3. *Bernoulli counts. Exact behaviour of the test statistic $d_r(\boldsymbol{x}, \theta_0) = d_r(\overline{\theta}, n)$, for $\theta_0 = 1/5$, as a function of the reference expected posterior value of the parameter, $\overline{\theta} = (r + \frac{1}{2})/(n+1)$, for sample sizes $n = 1$, $n = 5$, $n = 10$, $n = 100$ and $n = 1000$.*

The particular case where $r = n$ (all successes) and $\theta_0 = 1/2$ may be specially illuminating. In that situation, it is found that the null value should be questioned $(d_r > 2.5)$ for all $n > 5$ and definitely rejected $(d_r > 5)$ for all $n > 9$; thus, a scientist analysing an experiment to test for ESP powers *on the sole strength of the data* should require about 6 consecutive perfect answers

before questioning the hypothesis of random guessing, and about 10 consecutive perfect answers before definitely rejecting such a hypothesis.

Using Stirling's approximation, it is found that, for large sample sizes, the function $d_r(\boldsymbol{x}, \theta_0)$ is well approximated by $\delta(\theta_0, \overline{\theta})$, the logarithmic discrepancy between the models identified by θ_0 and by $\overline{\theta}$, and also by $\frac{1}{2}(1 + z^2)$, with $z = z(\boldsymbol{x}, \theta_0) = \sqrt{n}\,[\phi(\overline{\theta}) - \phi(\theta_0)]$, where $\phi(\theta) = 2\text{ArcSin}(\sqrt{\theta})$, which is the approximation given by (8).

Table 3. *Bernoulli counts. Correspondence between the threshold value d^* of the test statistic $d_r(\overline{\theta}, n)$, and 'type 1' error probabilities, $P[d_r > d^* \mid H_0]$, for sample sizes* 5, 10, 100 *and* 1000.

d^*	$n = 5$	$n = 10$	$n = 100$	$n = 1000$
1.0000	0.05792	0.22825	0.31759	0.32300
2.0000	0.05792	0.03279	0.10274	0.08904
2.4207	0.00672	0.03279	0.05948	0.05264
3.0000	0.00672	0.00637	0.02382	0.02417
4.0000	0.00032	0.00637	0.00546	0.00806
5.0000	0.00032	0.00086	0.00241	0.00263
6.0000	0.00032	0.00008	0.00061	0.00089
7.0000	0.00000	0.00008	0.00023	0.00031
8.0000	0.00000	0.00008	0.00008	0.00010
9.0000	0.00000	0.00000	0.00004	0.00004

The exact frequentist behaviour of the proposed test under the null may be computed from the null model $p(x \mid \theta_0) = \theta_0^x(1 - \theta_0)^{1-x}$, $x \in \{0, 1\}$. Table 3 reproduces the results obtained with $\theta_0 = 1/5$ for several sample sizes. Note —and this is of course a crucial shortcoming of frequentist measures— that in discrete data problems confidence levels are barely meaningful for small sample sizes. As one would expect from the asymptotic results described before, the behaviour of BRC for large samples is similar again to that obtained for testing a normal mean value with known variance; however as indicated in Table 3, *differences may be huge* for the small sample sizes which are often found, for example, in drug testing, or in the quality assessment of expensive items.

4.4. *Testing a Normal Mean Value with Unknown Variance*

We finally consider an example with nuisance parameters, which is probably the most common example of nested hypothesis testing in scientific practice. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, be a random sample of n real valued observations, and suppose that it is desired to check whether or not they could be described as a random sample from *some* normal distribution with mean μ_0, *assuming* that they may be described as a random sample from *some* normal distribution.

The logarithmic discrepancy between the assumed model and its closest approximation under the null is

$$\begin{aligned}\delta(\theta_0, \mu, \sigma) &= \inf_{\sigma_0 \in [0,\infty]} n \int \mathrm{N}(x \mid \mu, \sigma) \log \frac{\mathrm{N}(x \mid \mu, \sigma)}{\mathrm{N}(x \mid \mu_0, \sigma_0)}\, dx \\ &= \inf_{\sigma_0^2 \in [0,\infty]} \frac{n}{2}\Big[\log \frac{\sigma_0^2}{\sigma^2} - 1 + \frac{\sigma^2}{\sigma_0^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\Big].\end{aligned}$$

The infimum is attained at $\sigma_0^2 = \sigma_0^2(\mu_0, \mu, \sigma) = \sigma^2 + (\mu - \mu_0)^2$ and, substituting, one has

$$\delta(\theta_0, \mu, \sigma) = \frac{n}{2}\Big[\log\Big(1 + \Big(\frac{\mu - \mu_0}{\sigma}\Big)^2\Big)\Big] = \frac{n}{2}\,\log(1 + \theta^2),$$

where $\theta = (\mu - \mu_0)/\sigma$. It follows that the required test statistic is

$$d_r(\boldsymbol{x}, \mu_0) = \int_{-\infty}^{\infty} \tfrac{n}{2}\, \log\left(1 + \theta^2\right)\, \pi_\delta(\theta \,|\, \boldsymbol{x})\, d\theta, \tag{9}$$

where $\pi_\delta(\theta \,|\, \boldsymbol{x})$ is the reference posterior of θ when δ is the quantity of interest. In this problem, (μ, σ) are unknown parameters, the quantity of interest, $\delta = \frac{n}{2}\log(1+\theta^2)$, is a piecewise invertible function of θ, and the pair (θ, σ) is a one-to-one transformation of the pair (μ, σ). In Proposition 2 of the Appendix, we prove that, in a normal model, the joint reference prior for (θ, σ) when θ is the parameter of interest is $\pi_\theta(\theta, \sigma) \propto (1 + \frac{1}{2}\,\theta^2)^{-1/2}\sigma^{-1}$; moreover, since $\delta = \frac{n}{2}\log(1+\theta^2)$ is a piecewise invertible function of θ, it follows from Proposition 1 in that Appendix that this is also the reference prior when δ is the quantity of interest. Hence, using Bayes' theorem and integrating out the nuisance parameter σ, one has

$$\pi_\delta(\theta \,|\, \boldsymbol{x}) \propto \left(1 + \frac{\theta^2}{2}\right)^{-1/2} \exp\Big[-\frac{n\theta^2}{2}\Big]\; I\Big[n, \Big(\frac{n}{n-1+t^2}\Big)^{1/2} t\,\theta\Big], \tag{10}$$

where $t = t(\boldsymbol{x}, \mu_0) = \sqrt{n}(\overline{x} - \mu_0)/s$, with $s^2 = \sum (x_j - \overline{x})^2/(n-1)$, is the conventional t statistic and, in terms of the ${}_1F_1$ hypergeometric function,

$$\begin{aligned} I[n, \gamma] &= \int_0^\infty \omega^{n-1} \exp[-\tfrac{1}{2}w^2 + \gamma\,\omega]\, d\omega \\ &= 2^{(n-3)/2}\left[\sqrt{2}\,\Gamma(\frac{n}{2})\,{}_1F_1(\frac{n}{2}, \frac{1}{2}, \frac{\alpha^2}{2}) + 2\alpha\,\Gamma(\frac{n+1}{2})\,{}_1F_1(\frac{n+1}{2}, \frac{3}{2}, \frac{\alpha^2}{2})\right]. \end{aligned}$$

It may be verified that the reference posterior (10) is proper whenever $n \geq 2$. The function $I[n, \gamma]$ may also be recursively evaluated in terms of the standard normal cumulative distribution function Φ.

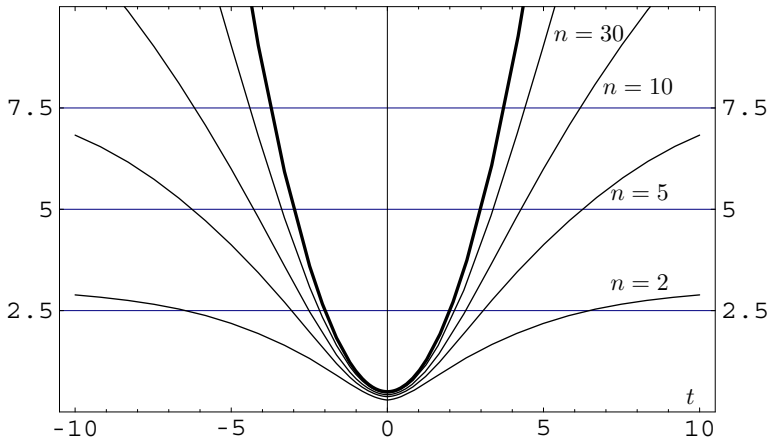

Figure 4. *Normal observations. Exact behaviour of the test statistic $d_r(\boldsymbol{x}, \mu_0) = d_r(t, n)$, as a function of the conventional t statistic, for sample sizes 2, 5, 10 and 30, and its limiting behaviour as $n \to \infty$ (solid line).*

Figure 4 describes the exact behaviour of the reference posterior expected discrepancy $d_r(\boldsymbol{x}, \mu_0)$, numerically computed from (9), as a function of the conventional statistic t, and the sample size n. For moderate sample sizes, a good approximation to d_r is provided by

$$d_r(\boldsymbol{x}, \mu_0) = d_r(t, \mu_0) \approx \tfrac{n}{2}\log(1 + \tfrac{t^2}{n}) + \tfrac{1}{2}\,\cdot$$

Table 4. *Normal observations. Correspondence between the threshold value d^* of the test statistic $d_r(\boldsymbol{x}, \mu_0)$, and 'type 1' error probabilities, $P[d > d^* \mid H_0]$, for sample sizes* 2, 5, 10, 30, 100 *and* 1000.

d^*	$n = 2$	$n = 5$	$n = 10$	$n = 30$	$n = 100$	$n = 1000$
1.0000	0.32299	0.23752	0.25310	0.28736	0.30706	0.31623
2.0000	0.14400	0.06540	0.06225	0.07131	0.07884	0.08278
2.4207	0.10471	0.04093	0.03683	0.04191	0.04691	0.04966
3.0000	0.05141	0.02232	0.01850	0.02067	0.02349	0.02514
4.0000	0.00000	0.00841	0.00601	0.00638	0.00740	0.00806
5.0000	0.00000	0.00335	0.00206	0.00205	0.00240	0.00266
6.0000	0.00000	0.00135	0.00074	0.00067	0.00080	0.00090
7.0000	0.00000	0.00045	0.00027	0.00023	0.00027	0.00031
8.0000	0.00000	0.00004	0.00010	0.00008	0.00009	0.00011
9.0000	0.00000	0.00000	0.00004	0.00003	0.00003	0.00004

The limiting function as n increases is found to be $\frac{1}{2}(1 + t^2)$ so that, as one might expect, the solution converges asymptotically to that obtained for the known variance case.

The exact frequentist behaviour of the proposed test under the null may easily be obtained from the fact that the sampling distribution of t is standard Student with $n-1$ degrees of freedom. Table 4 reproduces the results obtained for several sample sizes. As could be expected from the asymptotic results described above, the frequentist behaviour observed for large samples approaches that obtained for testing a normal mean value with known variance. Note that, although BRC also uses the conventional t statistic, one does *not* have anymore a correspondence between d^*-values and significance levels. However, as demonstrated in Table 4, qualitatively similar results are obtained for moderate and large sample sizes.

ACKNOWLEDGEMENTS

The author is very grateful to Raúl Rueda for his stimulating comments, and for his warm hospitality in Mexico city, where part of this research was done. Thanks are also due to Jim Berger, Michael Lavine, Dennis Lindley, Elías Moreno, Tony O'Hagan and Luca Tardella for their comments to an earlier version of this paper.

REFERENCES

Aitkin, M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. B* **53**, 111–142 (with discussion).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd. Int. Symp. Information Theory*. Budapest: Akademia Kaido, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19**, 716–727.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

Bayarri, M. J. (1987). Comment to Berger and Delampady. *Statist. Sci.* **3**, 342–344.

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–352 (with discussion).

Berger, J. O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *Internat. Statist. Rev.* **59**, 337–353.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.* **82**, 112–133 (with discussion).

Berger, J. O. and Pericchi, L. R. (1995). The intrinsic Bayes factor for linear models. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 25–44 (with discussion).

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229–263.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 605–647 (with discussion).

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J. M. (1985). Análisis Bayesiano de los contrastes de hipótesis paramétricos. *Trab. Estadist.* **36**, 45–54.

Bernardo, J. M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189, (with discussion).

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. *Model Choice* (J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, eds.). Brussels: Pub. Fac. Univ. Saint Louis, 43–59.

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Consonni, G. and Veronese, P. (1987). Coherent distributions and Lindley's paradox. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 111–120.

Ferrándiz, J. R. (1985). Bayesian inference on Mahalanobis distance: an alternative approach to Bayesian model testing. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 645–654.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London : Griffin; New York: Hafner Press.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 669–674.

Jaynes, E. T. (1980). Discussion to the session on hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 618–629. Reprinted in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. (R. D. Rosenkranz, ed.). Dordrecht: Kluwer(1983), 378–400.

Jeffreys, H. (1939). *Theory of Probability* (Third edition in 1961). Oxford: Oxford University Press.

Jeffreys, H. (1980). Some general points in probability theory. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 451–453.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley. Second edition in 1968, New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*. Rotterdam: University Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.* **30**, 112–119.

Moreno, E. and Cano, J. A. (1989). Testing a point null hypothesis: asymptotic robust Bayesian analysis with respect to priors given on a sub-sigma field. *Internat. Statist. Rev.* **57**, 221-232.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. B* **57**, 99–138 (with discussion).

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factor. *Test* **6**, 101–118.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 603–608.

Robert, C. P. and Caron N. (1996). Noninformative Bayesian testing and neutral Bayes factors. *Test* **5**, 411–437.

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1**, 61-67.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Shafer, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–351 (with discussion).

Smith, C. A. B. (1965). Personal probability and statistical analysis. *J. Roy. Statist. Soc. A* **128**, 469–499.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B* **42**, 213–220.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B* **44**, 377–387.

APPENDIX. SOME RESULTS ON REFERENCE DISTRIBUTIONS

Proposition 1. *Let $p(\boldsymbol{x} \mid \theta)$, $\theta \in \Theta \subset \Re$, be a regular one-parameter model. If the quantity of interest $\phi = \phi(\theta)$ is piecewise invertible, then the corresponding reference prior $\pi_\phi(\theta)$ is the same as if θ were the parameter of interest.*

Outline of proof. Let $\phi = \phi(\theta)$, with $\phi(\theta) = \phi_i(\theta)$, $\theta \in \Theta_i$, where each of the $\phi_i(\theta)$'s is one-to-one in Θ_i; thus, $\theta = \{\phi, \omega\}$, where $\omega = i$ iff $\theta \in \Theta_i$. The reference prior $\pi_\phi(\theta)$ only depends on the asymptotic posterior of θ which, for sufficiently large samples, will concentrate on that subset Θ_i of the parameter space to which the true θ belongs. Since $\phi(\theta)$ is one-to-one within Θ_i, and reference priors are consistent under one-to-one reparametrizations, the stated result follows. $\triangleleft$

Proposition 2. *Consider a normal model $N(x \mid \mu, \sigma)$ with both parameters unknown and, for some $\mu_0 \in \Re$, let $\theta = (\mu - \mu_0)/\sigma$ be the quantity of interest. Then, in terms of (θ, σ), the reference prior is $\pi_\theta(\theta, \sigma) \propto (1 + \theta^2/2)^{-1/2}\sigma^{-1}$.*

Proof. In terms of (θ, σ), Fisher's information matrix $H(\theta, \sigma)$ and its inverse $S(\theta, \sigma)$ are

$$H(\theta, \sigma) = \begin{pmatrix} 1 & \theta/\sigma \\ \theta/\sigma & (2 + \theta^2)/\sigma^2 \end{pmatrix}, \qquad S(\theta, \sigma) = H^{-1}(\theta, \sigma) = \begin{pmatrix} 1 + \theta^2/2 & -\theta\sigma/2 \\ -\theta\sigma/2 & \sigma^2/2 \end{pmatrix}.$$

The natural compact approximation to the nuisance parameter space is $\{\log \sigma \in [-i, i]\}$, which does not depend on θ, and both h_{22} and s_{11} factorise as functions of θ and σ; thus, (Bernardo and Smith, 1994, p. 328)

$$\pi(\sigma \mid \theta) \propto \sigma^{-1}, \qquad \pi(\theta) \propto (1 + \theta^2/2)^{-1/2}$$

and, hence, $\pi_\theta(\theta, \sigma) \propto (1 + \theta^2/2)^{-1/2}\sigma^{-1}$, as stated. $\triangleleft$

DISCUSSION

GAURI S. DATTA *(University of Georgia, USA)*

It is my pleasure to discuss a very stimulating paper by Professor Bernardo. He has presented another interesting article on the development of reference priors that are useful to carry out objective Bayesian analyses in scientific investigations. The author, with a number of eminent collaborators, has made many important contributions in default Bayesian analyses through reference priors in the last two decades since the publication of his pioneering paper on the subject. While in the majority of his works on reference priors he considered the estimation aspect of the Bayesian statistical inference, in the present article Professor Bernardo considers development of reference priors for Bayesian hypothesis testing and model selection.

In many respects Bayesian solutions, especially noninformative Bayesian solutions, to hypotheses testing are different from those for estimation problems. Unlike Bayesian estimation problems with improper priors, where the normalising constant for a single model gets cancelled in the final answer (of course, assuming all required integrals exist), Bayesian testing and model

selection deal with more than one model, where the normalising constants for different models are not readily comparable for improper noninformative priors. Thus a noninformative Bayesian solution to hypotheses testing needs careful attention. Often a hypothesis testing problem concerns selecting a model nested within a larger model. Bayesian testing of nested hypotheses through Bayes factors based on improper priors faces many difficulties and sometimes produces paradoxical results (e.g., Lindley's paradox).

To circumvent some of the problems associated with Bayes factors there have been several attempts to suitably modify the Bayes factors. Professor Bernardo in this paper takes a decision theoretic approach to developing an objective Bayes solution to test for nested hypotheses. He obtains a noninformative prior via Berger-Bernardo reference prior algorithm by treating $\delta(\theta_0, \theta, \lambda)$, the expected log-likelihood ratio under the full model, as the parameter of interest. The Bayesian reference criterion (BRC) that is suggested as a test statistic by the author is given by the expectation of $\delta(\theta_0, \theta, \lambda)$ under the posterior derived from this reference prior. I will examine in my discussion the proposed method through three examples.

Example 1. Let $f(x;\theta) = a(x)\exp\{\theta_1 u_1(x) + \theta_2 u_2(x) + c(\theta_1, \theta_2)\}$ be the density function of a two-parameter exponential distribution. Define $\eta_i = E_\theta(u_i(X))$, $i = 1, 2$. It is known that the mixed parameterisation (θ_1, η_2) introduces an orthogonal reparameterisation of (θ_1, θ_2). We assume $\theta_2 = -\theta_1\phi'(\eta_2)$ for some function ϕ. Bar-lev and Reiser (1982) showed that $c(\theta_1, \theta_2)$ and $\eta_1(\theta_1, \theta_2)$ can be expressed as $c(\theta_1, \eta_2) = \theta_1\chi(\eta_2) - M(\theta_1)$ and $\eta_1 = \phi(\eta_2) + M'(\theta_1)$, where $\chi(\eta_2) = \eta_2\phi'(\eta_2) - \phi(\eta_2)$ and $M(\theta_1)$ is an infinitely differentiable function with $M''(\theta_1) > 0$ and $\phi''(\eta_2) \neq 0$. To test H_0: $\theta_1 = \theta_{10}$ vs. H_1: $\theta_1 \neq \theta_{10}$, it can be checked that $\delta(\theta_{10}, \theta_1, \eta_2) = n(\theta_1 - \theta_{10})M'(\theta_1) - n\{M(\theta_1) - M(\theta_{10})\}$ is a function of θ_1 alone. Although $\delta(\theta_{10}, \theta_1, \eta_2)$ is not a one-to-one function of θ_1, reference analysis as proposed in the paper can be carried out by following the Berger-Bernardo algorithm, treating θ_1 as the parameter of interest and η_2 as a nuisance parameter. The information matrix is $I(\theta_1, \eta_2) = \text{Diag}(M''(\theta_1), -\theta_1\phi''(\eta_2))$. It follows from Berger (1992) or Datta and Ghosh (1995a) that the reference prior for $\{\theta_1, \eta_2\}$ is $\pi_\delta(\theta_1, \eta_2) = \sqrt{M''(\theta_1)|\phi''(\eta_2)|}$, which is also a first-order joint-probability-matching prior for θ_1 and η_2 (see Datta 1996 and Sun and Ye 1996).

As a concrete application of Example 1, we consider the testing of a normal variance σ^2 when the mean μ is a nuisance parameter. Here

$$\delta(\sigma_0^2, \sigma^2, \mu) = \frac{n}{2}\left[(\frac{\sigma^2}{\sigma_0^2} - 1) - \log(\frac{\sigma^2}{\sigma_0^2})\right],$$

and

$$d_r(\boldsymbol{x}, \sigma_0^2) = \frac{n}{2}\left[\psi(\frac{n-1}{2}) - \log(\frac{S^2}{2\sigma_0^2}) + \frac{S^2}{(n-3)\sigma_0^2} - 1\right],$$

with $S^2 = \sum_1^n (x_i - \bar{x})^2$, are very similar to the corresponding quantities defined in Example 4.2. In general, this does not lead to the UMPU test.

Example 2. Balanced one-way random effects models: Let $y_{ij} = \mu + a_i + e_{ij}$, $j = 1, \ldots n$, $i = 1, \ldots, k$ where a_i and e_{ij} are independently distributed with $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. Writing $\theta = \sigma_a^2$, $\lambda = (\mu, \sigma_e^2)$, to test H_0: $\sigma_a^2 = 0$ vs. H_1: $\sigma_a^2 > 0$, the discrepancy function

$$\delta(\theta_0, \theta, \lambda) = \frac{k}{2}\Big[\frac{n\sigma_a^2}{\sigma_e^2} - \log(1 + \frac{n\sigma_a^2}{\sigma_e^2})\Big]$$

is only a function of the ratio of the two variances (here $\theta_0 = 0$). Defining $\sigma_e^{-2} = r$ and $\sigma_e^2(n\sigma_a^2 + \sigma_e^2)^{-1} = u$, it follows the reference prior for testing H_0: $\sigma_a^2 = 0$ is given by

$\pi(r, u, \mu) = (ru)^{-1}$. This prior was obtained earlier as a reference and probability-matching prior by Datta and Ghosh (1995b); see also Datta (1996). It can be checked that for priors of the form $r^{-b_1}u^{-b_2}$, the BRC is a strictly increasing function of the usual F-statistic, thereby leading to a test equivalent to the frequentist test.

Marginalisation Paradox: Notwithstanding the successful handling of many difficult problems in presence of nuisance parameters, the Berger-Bernardo algorithm can produce priors for certain group orderings of the parameters which fail to avoid marginalisation paradoxes (see Datta and Ghosh 1995c). We will give an example to show that the BRC also suffers from this pitfall.

Example 3. We consider testing H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$ in a bivariate normal distribution with density $N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Writing $\theta = \rho$, $\lambda = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, it can be shown that $\delta(\theta_0, \theta, \lambda) = -n \log(1 - \rho^2)/2$, which is only a function of ρ. The two-group reference prior for $\{\theta, \lambda\}$ is $\pi_\delta(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \sigma_1^{-2}\sigma_2^{-2}(1 - \rho^2)^{-1}$ (see Datta and Ghosh 1995c), which neither is probability-matching for ρ nor does it avoid the marginalisation paradox. It is also shown by these authors that further splitting of the last group results in a reference prior given by $\{\sigma_1\sigma_2(1 - \rho^2)\}^{-1}$ for parameter grouping $\{\rho, (\mu_1, \mu_2), (\sigma_1, \sigma_2)\}$ or $\{\rho, \mu_1, \mu_2, \sigma_1, \sigma_2\}$, which is probability-matching for ρ and avoids the paradox.

BRUNERO LISEO *(Università di Roma "La Sapienza", Italy)*

Let me start this discussion with a warm Thanks! to the Organizing Committee for putting on my (and Datta's) shoulders the responsibility of criticizing our host. We will do our best to make Valencia 7 still possible!

I will focus my discussion on three main points: (i) the role of probability and Bayes factors in hypothesis testing, (ii) the construction of the utility function, and (iii) the comparison of BRC with other approaches.

1. *The role of probability and Bayes factors in hypothesis testing.* Professor Bernardo says

> ...it may not be wise to use Bayes Factors in nested hypothesis testing

If one thinks to the immediate consequence of this statement, it is compulsory to say that it may not be wise to use probability in nested hypothesis testing! My view is somewhat different and here I will try to illustrate it. Models have different roles in statistics. Cox (1990) and Lehmann (1990) basically distinguish between *empirical* and *mechanistic* models. In the case of empirical models, we know that no one of them will be true and our aim is simply to select the model which best describes the phenomenon under study. Models are used as a guide to action and, in this sense I found the Bernardo's approach very sensible. However, I consider his scheme more adapt to analyze situations where different models are competing, as alternative tools to approximately describe the phenomenon, as, for example, in non-nested situations. In this case

> The question of truth of a mathematical hypothesis does not arise, only that of its use as a calculating tool. (Bishop Berkeley (1734), taken from Lehmann, 1990.)

On the other hand, there are completely different situations where a 'precise' null hypothesis makes sense (see Berger and Delampady, 1987): in these cases I cannot see any alternative way to use probability (and Bayes factors) statements on the truthfulness of the null hypothesis. All in all I challenge Bernardo's conclusion that the BRC is well suited for nested situations. I would rather suggest to check its applicability with non-nested models. Of course, in this case, the mathematics are going to be more involved and the posterior expected utility difference will lose its interpretation in terms of divergence.

2. *The Construction of the Utility Function.* Professor Bernardo starts from a well known and accepted utility function to be used in pure scientific inference about a random quantity ϕ, namely

$$u(q_\phi(\cdot), \phi) = \alpha \log q_\phi(\phi) + \beta(\phi), \tag{1}$$

which is proper and local (Bernardo and Smith, 1994, Ch 3). To adapt this utility to his problem, Professor Bernardo proposes the following modification

$$u(q_x(\cdot), \theta, \omega) = \alpha \int p_x(\boldsymbol{y} \,|\, \theta, \omega) \log(q_x(\boldsymbol{y})) d\boldsymbol{y} + \beta(\theta, \omega). \tag{2}$$

This is neither a particular case of (1) nor its consequence, and its use as a utility function would deserve more justification. To me it is not clear whether the first argument of the utility function, $q_x(\cdot)$, is a predictive distribution, free of the parameters, or it is a generic sampling distribution (belonging to M_0 or M_1). Note that expression (2) would remain a proper utility function only in the second case. Then, in its final step towards the transformation of the problem into a decision one, Professor Bernardo actually introduces a somewhat different utility function. The utilities of the two possible decisions a_0 and a_1 are in fact

$$u(a_0, \theta, \omega) = \alpha \sup_{\omega_0 \in \Omega} \int p_x(\boldsymbol{y} \,|\, \theta, \omega) \log(p(\boldsymbol{y} \,|\, \theta_0, \omega_0)) d\boldsymbol{y} + \beta(\theta, \omega) - c_0, \tag{3}$$

$$u(a_1, \theta, \omega) = \alpha \int p_x(\boldsymbol{y} \,|\, \theta, \omega) \log(p(\boldsymbol{y} \,|\, \theta, \omega)) d\boldsymbol{y} + \beta(\theta, \omega) - c_1. \tag{4}$$

Some questions arise:

(i) *Where do c_0 and c_1 come from?* It is true that we need them, otherwise the larger model, assumed to be true, will always be preferred. It is also true, as Professor Bernardo stresses, that the von Neumann-Morgenstern theory is compatible with an additive decomposition of the utility, but here we do not have a decomposition. We simply have an extra-component c_j added to the utility function. This modification makes it questionable, at least formally, whether the use of the expected utility is a coherent criterion for choosing among decisions.

(ii) *Sampling or predictive distributions?* In expressions (3) and (4) utilities of each single member of the families M_0 and M_1 are calculated for each single (θ, ω). In a sense, this seems to be too optimistic since each single sampling distribution is evaluated at the 'right' value of the parameters. It sounds like profiling the problem, by not considering the influence of the nuisance parameter. In a Bayesian model comparison, would it not be more realistic to use

$$m_0(\boldsymbol{y}) = \int p(\boldsymbol{y} \,|\, \theta_0, \omega) \pi(d\omega) \qquad \text{and} \qquad m_1(\boldsymbol{y}) = \iint p(\boldsymbol{y} \,|\, \theta, \omega) \pi(d\theta, d\omega)$$

instead of, respectively, $p(\boldsymbol{y} \,|\, \theta_0, \omega)$ and $p(\boldsymbol{y} \,|\, \theta, \omega)$? Of course this approach would imply the use of a prior distribution inside the utility, as Professor Herman Rubin (see, for example Rubin and Sethuraman, 1966) has often suggested. Clearly, this proposal needs to also be analyzed in detail as a utility function but it seems to me more naturally consistent with (2), if not with (1).

This way, granted the use of c_0 and c_1, formula (4) in the paper would become

$$\delta(\theta_0, \theta, \omega) = \int p(\boldsymbol{y} \,|\, \theta, \omega) \log \frac{m_1(\boldsymbol{y})}{m_0(\boldsymbol{y})} d\boldsymbol{y}. \tag{5}$$

Note that the priors to be used in this context cannot be improper. Is it surprising? No, I think not. Coherent Bayesian model selection needs proper priors. To see what happens in this case,

let us consider Example 1 (Lindley's Paradox). After some algebra, and assuming a conjugate prior $N(\mu_0, \sigma_1^2)$ for μ under the larger model, it turns out that BRC selects M_1 if and only if

$$z^2(\boldsymbol{x}) \geq 2\left[g - \log\left(\sigma/\sqrt{\sigma^2 + n\sigma_1^2}\right) - \frac{n\sigma_1^2(\sigma^2 + 2n\sigma_1^2)}{2(\sigma^2 + n\sigma_1^2)^2}\right]\left(\frac{n\sigma_1^2}{\sigma^2 + n\sigma_1^2}\right)^{-3} \tag{6}$$

As σ_1^2 goes to infinity all the quantities in the left-hand side of (6) remain bounded; the only exception is

$$\log\left(\sigma/\sqrt{\sigma^2 + n\sigma_1^2}\right).$$

This means that the Lindley's paradox appears again! From the above analysis it is clear that BRC avoids the paradox simply because the variance of $m_1(\boldsymbol{x})$, $\sigma^2/n + \sigma_1^2$ is replaced by the variance of $p(\overline{x} \mid \mu)$, which is σ^2/n independently of μ.

3. *Comparison of BRC with other approaches.* From an operational viewpoint, a new tool for Bayesian model comparison should be compared with the more important existing one, namely the Bayes factor and its ramifications. Now I will elaborate this point in the simple scenario of Example 1 (Lindley's Paradox). It is well known that, using a conjugate prior $N(\mu_0, \sigma_1^2)$ on μ under the larger model, we get a Bayes factor which, as n (or σ_1^2) goes to infinity always selects the simpler model. How can the BRC avoid this behavior and still remain a Bayesian criterion? BRC selects the larger model if and only if

$$z^2(\boldsymbol{x}, \mu_0) = \frac{n(\bar{x} - \mu_0)^2}{\sigma^2} > 2g - 1. \tag{7}$$

On the other hand a proper conjugate Bayesian analysis will select the larger model if and only if $m_1(\boldsymbol{x})/m_0(\boldsymbol{x}) > 1$, that is, when

$$z^2(\boldsymbol{x}, \mu_0) = \frac{n(\bar{x} - \mu_0)^2}{\sigma^2} > \frac{n\sigma_1^2 + \sigma^2}{n\sigma_1^2}\log\left(\frac{n\sigma_1^2 + \sigma^2}{\sigma^2}\right). \tag{8}$$

Note, also, that the "intrinsic" priors arising from the expected arithmetic intrinsic Bayes factor (Berger and Pericchi, 1996) and from the fractional Bayes factor (O'Hagan, 1995) are special cases of conjugate priors. By equating thresholds in (7) and (8) one obtains

$$1 + \frac{n\sigma_1^2}{\sigma^2} = \exp\left\{\frac{n\sigma_1^2}{n\sigma_1^2 + \sigma^2}(2g - 1)\right\} \tag{9}$$

That means that, for fixed n, there is a one to one relation between g and σ_1^2. Choosing a level g in terms of utility amounts to choose a conjugate with the appropriate variance. Also, Equation (9) shows that, as n increases, BRC avoids the paradox by decreasing the prior variance. In a sense the "intrinsic" prior of the BRC depends on n.

4. *Concluding remarks.* A general concern that I have with the BRC is that it is difficult to use a (inference tailored) utility function in a hypotheses testing set-up. The collapse of the action space into only two points makes it difficult for a utility function to remain proper. Consequently, the use of the logarithmic discrepancy turns out to be suspect.

Conclusions obtained with BRC are very close to a frequentist test, in the spirit of reference analysis. Whereas it can be valuable in estimation problems, it is going to be a problem in testing, especially when testing a precise null hypothesis. Berger and Delampady (1987) develop this point.

DENNIS V. LINDLEY *(Minehead, UK)*

The world that we inhabit is complicated. We know a few things about it, either through our personal experiences or from those experiences we share with others. Despite this knowledge, most aspects of our world are uncertain for us. One of the great achievements of mankind is the demonstration that this uncertainty must be described by quantities that obey the rules of the probability calculus. I personally learnt this from Harold Jeffreys, but others have given alternative demonstrations that lead to essentially the same conclusion: the inevitability of probability. Our knowledge is primary probabilistic. We therefore need to describe our uncertain world in probabilistic terms. A model refers to part of this description, and data can assist in determining modifications to a model.

In addition to knowledge about the world, we need to act in face of the uncertainty of that knowledge. Action, or decision-making, requires an evaluation of our individual preferences. These are expressed, again in terms of probability, through a utility function. Action is achieved by maximization of expected utility. Jeffreys did not concern himself with decisions, but these conclusions easily follow from the demonstration that probability is the appropriate language.

Jeffreys was concerned with uncertainty in science. A key concept in the scientific method is that of a theory, or hypothesis. Jeffreys pointed out that many theories can be put in the form that a parameter θ takes a particular value θ_0. More generally, it has proved useful to study situations in which a hypothesis that $\theta = \theta_0$ is proposed, which is than tested against $\theta \neq \theta_0$. As in this paper, I confine myself to the one parametric dimension of interest, recognizing that other nuisance parameter ω may be present. Combining this formulation with the major point about probability, Jeffreys formulated hypothesis testing as the calculation of the probability that θ is equal to θ_0, rather than to some other value. Hypothesis testing is, in principle, very simple, merely the calculation of $\Pr(\theta_0 \,|\, \boldsymbol{x})$, for data $\boldsymbol{x}$. It is part of our total expression of uncertainty about the world. In this view, it is part of our appreciation of the world, and has no element of decision-making in it.

Bernardo takes a different view of model choice of hypothesis-testing. Let us look at the stages in this approach.

(i) It is treated as a decision problem. That is, it is not regarded as just one aspect of our appreciation of the world, but goes beyond it in contemplating action in that uncertain world.

(ii) The decision is not about the parameter but about data: "act as if $p_{\boldsymbol{x}}(\cdot \,|\, \theta_0, \omega)$ were true". This is restrictive since theories are general statements, not confined to data sets either present, $\boldsymbol{x}$, or to future data sets of the same type $\boldsymbol{y}$, referred to in the paper.

(iii) The interpretation of hypothesis-testing in decision terms requires a utility function. The twin requirement that this be both proper and local, means a logarithmic form. Locality may be queried because values near θ_0 play an important role, as will be seen below. This leads to the logarithmic discrepancy as the quantity whose expectation has to be found.

(iv) The theory of reference priors is then used to obtain the prior appropriate to the logarithmic discrepancy, from whence the expected utilities can be calculated. The test then becomes the choice of the better decision according to the expectation criterion.

Having summarized both approaches, let us compare them. One clear distinction is the complexity of Bernardo's method in comparison with the simplicity of Jeffreys's. It requires four procedures, as against a single calculation of probability. However, this is not necessarily a serious objection since once the analysis has been performed (and the paper provides several examples) the resulting test is just as easy to use as Jeffreys's. The user need not fear the complexity.

Another distinction, which I consider more important, is that the decision procedure is *automatic* once the sample space and its associate probability structure are given, whereas the probability approach requires the user to *think* about additional probabilities. Bernardo suffers from what I call the Greek-letter syndrome. Nowhere, in constructing the test, does he, or the user, need to ask what θ means. It is just a Greek letter. Jeffreys had the syndrome to a lesser degree, because he tried to find automatic priors; for example the Cauchy in the familiar, normal case discussed in Section 1. A subjective Bayesian, following de Finetti, requires one to think, not about a Greek letter, but about the feature of the world it attempts to describe.

This leads to a third distinction. Bernardo's automatic procedure leads to a unique answer, whereas the subjective approach does not; indeed, testing, unlike estimation, is disturbingly sensitive to the original probability distribution over values of θ other than θ_0. Jeffreys noticed that his invariant distributions, with their automatic element, produced nonsense with tests. So do reference priors for θ, but it is a triumph of this paper to suggest, and than to prove, that those for the logarithmic discrepancy are sensible. This feature alone makes this an important paper: a beautiful goal for Spain whilst England is confused mid-field.

To explore this contrast further, consider the normal case of Section 1. Since this provides a good approximation for a wide variety of testing problems with a single parameter of interest, what happens here is a good guide to most behaviour. Everyone agrees that the analysis depends on $z(\boldsymbol{x}, \theta_0) = (\overline{x} - \theta_0)/(\sigma/\sqrt{n})$. Disagreement lies on what to do with z. Bernardo agrees with the standard practice and rejects the null value θ_0 if z^2 exceeds a constant c that does not depend on n or σ. Jeffreys, or a subjectivist, would reject (with possible reservations on the word 'reject') if the probability of θ_0 was sufficiently small. If θ, given $\theta \neq \theta_0$, is normal, centred a θ_0, with variance σ_1^2, this leads to rejection if

$$z^2 > \left\{ \log\left(1 + \frac{n}{\lambda}\right) + c' \right\} \frac{n+\lambda}{n}, \qquad \lambda = \frac{\sigma^2}{\sigma_1^2}, \tag{1}$$

for constant c'. This follows form the first, displayed equation in the paper. Here, λ depends on the variance of the normal prior and exhibits the sensitivity referred to above. Bernardo says that (1) results in "undesirable behaviour, clearly inconsistent with accepted scientific practice". I disagree; so much the worse for scientific practice.

Consider first the influence of the variance σ_1^2 of the prior, expressed through λ. Equation (1) shows that is substantial. Suppose that the experiment yielding $\overline{x}$ concerns ESP and that θ_0 is the value that would arise were ESP absent. Then it is reasonable to suppose that θ cannot differ much form θ_0 because otherwise good evidence for ESP would have been demonstrated before. It has not. So σ_1 is small, λ is large and it is hard to reject θ_0. Contrast this with an experiment on a drug which is expected to do well in comparison with a placebo. Here the effect could be large and values of θ, the difference between drug and placebo, substantially different from $\theta = \theta_0$ quite reasonable. Hence λ small. Practical considerations like these seem entirely reasonable to me. We should not look at $\overline{x}$ in isolation. We should not ignore the meaning of Greek letters. For Bernardo, the real world does not appear to matter: telepathy, drug testing, it is all the same to the adherent of reference priors. Indeed he goes so far as to search for analysis in which the data dominates. That is, data sets are analysed in isolation. This is not how science works; it is through lots of different types of data that theories came to be accepted. Statisticians have slowly come to recognize this and introduce the topic of meta-analysis. In other words, I contend that the sensitivity to the prior, the mid-field confusion, is not a defect but a reasonable reflection of reality.

Another difference between (1) and $z^2 > c$ lies not in the dependence of λ, but on n. For large n, (1) behaves like

$$z^2 > c'' + \log n, \tag{2}$$

so that as n increases it becomes increasingly difficult to reject θ_0 in comparison with the standard $z^2 > c$, though the increase is logarithmic, and therefore slow. This can be defended using an Occam's razor type of argument, that values of the parameter other than the null value should not be introduced without due cause. A better defence, to my mind, is the remark that if sampling is continued (n increased) until $z^2 > c$, this will certainly happen even if $\theta \neq \theta_0$. Proof of this uses the law of the iterated logarithm. Introduction of the $\log n$ term in (2) prevent this happening. It is no longer possible to sample to a foregone conclusion (that $\theta \neq \theta_0$). This results connects with the likelihood principle, which (1) satisfies. It is not clear to me whether the decision method of this paper does. The appearance of integrals over sample space in the logarithmic discrepancy suggests it need not. The differences between the two approaches is most noticeable when n is large and then only when z is small, yet not too small. To express this differently, $p(\theta \mid \overline{x}, \theta \neq \theta_0)$ is normal with very small variance and mean near to θ_0. In those circumstances, locality may be influential: whatever θ obtains, it is almost surely near to θ_0.

Another possible objection to the use of $z^2 > c$ might arise in passing from (1) to higher dimensions. It is well known that, using standard, tail area significance tests, it can happen that, with (x, y) normal about (θ, ϕ), a test of $\theta = 0$ can lead to rejection at the same time as one for $\phi = 0$, whereas the test for $\theta = \phi = 0$ can result in acceptance. We will have to wait for Bernardo to extend his results to the bivariate situation before it can be seen whether his procedure avoids that difficulty.

I remain unconvinced by the wizardry of this paper. Yet, if we honestly compare the two models for hypothesis testing, that of the subjective Bayesian with that of this paper, we cannot, in our present state of knowledge, reject either. The more understanding we gain of this original, ingenious and stimulating approach, the easier it will be to achieve the ultimate goal; a sensible, probabilistic description of our uncertain world.

MICHEL MOUCHART *(Université Catholique de Louvain, Belgium)*

From a Bayesian point of view, whether two models are nested or not should depend not only on the sampling specification (the data density) but also on the prior specification. In Example 3.2.1 of Florens and Mouchart (1993) we produce a situation where the prior specification on the regression coefficient of a given explanatory variable should clearly depend on the model although, from a frequentist approach, one is nested in the other one. Such an issue might be kept hidden in a pure reference analysis.

It is quite interesting to realize that the author is developing a class of examples of the "encompassing principle" : this gives historical support for his work; for more detail, see my comments to Geweke's paper in this volume.

TONY O'HAGAN *(University of Nottingham, UK)*

Professor Bernardo's paper is interesting and provocative, but I have serious doubts about the development in Section 2 of the criterion $d(\boldsymbol{x}, \theta_0)$. In the discussion leading to (1), the analysis is said to be before observing data $\boldsymbol{x}$, and is based on the utility of predicting those future data $\boldsymbol{x}$, using the same model. In the next displayed equations, $\boldsymbol{x}$ is replaced by $\boldsymbol{y}$, which is described as a dummy variable. The next sentence contains the phrase "given data $\boldsymbol{x}$", so now Bernardo is looking at analysis after observing $\boldsymbol{x}$. What now is the status of $\boldsymbol{y}$? Just after (4), he refers to $d(\boldsymbol{x}, \theta_0)$ as being concerned with the amount of information about future observations, so apparently $\boldsymbol{y}$ is a future observation which will be made after observing the data $\boldsymbol{x}$. Presumably, $\boldsymbol{y}$ could be any future data, a single observation, many observations, or perhaps even an infinite sequence of future observations. But, remarkably, towards the end of Section 3.1, it seems that $\boldsymbol{y}$ is $\boldsymbol{x}$ again, because when $\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{x})$ is sufficient, Bernardo writes $\boldsymbol{x} = \{\boldsymbol{s}, \boldsymbol{r}\}$ and then proceeds as if $\boldsymbol{y} = \{\boldsymbol{s}, \boldsymbol{r}\}$ also. Throughout the rest of the paper it seems

that $\boldsymbol{y} = \boldsymbol{x}$ also. Now the original definition of $d(\boldsymbol{x}, \theta_0)$ is nonsense if we strictly interpret the statement $\boldsymbol{y} = \boldsymbol{x}$ because $\boldsymbol{y}$ is integrated out in (4) and yet appears again in (3) as $\boldsymbol{x}$. It seems that $\boldsymbol{y}$ is indeed future data, but with exactly the same number of observations and the same structure as $\boldsymbol{x}$. In effect, Bernardo's criterion relates to predicting a *replicate* of the data $\boldsymbol{x}$. There is no explanation or justification of this curious choice.

Another puzzling claim is that the equations right at the end of Section 3.2 demonstrate consistency. First, consistency in model choice is usually interpreted as meaning selecting the true model with probability one (as $n \to \infty$). These equations do not guarantee that, and indeed the behaviour when the null model is true is exactly that of a frequentist fixed-size hypothesis test: the null is falsely rejected with a probability that does not tend to zero. Further frequentist thinking is evident in the expectation being conditioned on the parameters, where one would have wanted a preposterior expectation.

Finally, I object to Bernardo's description in Section 1 of the fractional Bayes factor as a method of obtaining a non-subjective Bayes factor. I did not develop it as such, have never referred to it in that way, and dislike intensely the perversion of Bayesian statistics that is implied by the adjective 'non-subjective'.

CHRISTIAN P. ROBERT *(CREST-INSEE and Université de Rouen, France)*

The BRC approach adopted for model selection is quite convincing, especially in nested models. The choice of the Kullback-Leibler divergence has been stressed as a non-informative criterion in Robert (1996), since it encompasses all possible and future uses of the chosen model. Another advantage of the BRC criterion is that the whole analysis is done in terms of the full model and does not require to define a prior on each submodel, as pointed out in Goutis and Robert (1998) and Dupuis and Robert (1997). Moreover, this also allows for improper priors to be used in a regular Bayesian fashion, while avoiding the "dilution" phenomenon mentioned in George (1998).

The difficulty I have with the BRC method lies in the choice of a "golden standard" threshold. In Goutis and Robert (1998) and Dupuis and Robert (1997), we proposed alternatives which depend on the sampling model, the prior and/or the data at hand. It would seem that $d^* = 5$ could work in a limited set of models like exponential families in dimension one. Higher dimension models or setups with covariates could require more specific calibration.

Also, how does the BRC method extend to non-nested cases? There are many caveats related to encompassing pointed out in the econometric literature (see Goutis and Robert, 1997, for references) and I wonder whether the BRC criterion may suffer from those.

REPLY TO THE DISCUSSION

I am extremely grateful to all discussants for their thought-provoking comments. I will first try to give specific answers to their queries, and I will then attempt to summarize what I perceive to be the main conclusions.

1. *Reply to Datta.* I am very grateful to Datta for providing further examples which illustrate the behaviour of the BRC criterion. There are however two points in his comments which need clarification:

(i) In the one-way random effects model, he points out that the BRC statistic is a strictly increasing function of the usual F-statistic, and concludes that BRC leads to a test equivalent to the frequentist test. This is not so; the sampling distribution of F depends on the sample size and, therefore, using a fixed quantile of the corresponding sampling distribution as the cutoff point is *not* equivalent to using a fixed utility constant (independent of n) as suggested by BRC. There would be, however, asymptotic agreement.

(ii) Datta correctly points out that the prior obtained by grouping together the nuisance parameters in the coefficient of correlation example leads to a prior which is not probability matching and does not avoid the marginalisation paradox. I should stress however that, although the reference algorithm may technically be used for any grouping, we explicitly stated (Berger and Bernardo, 1992, Section 3.3; Bernardo, 1997) that 'the' reference prior should be that sequentially obtained by considering the nuisance parameters *one at a time*. In the coefficient of correlation example this leads to the reference prior $\sigma_1^{-1}\sigma_2^{-1}(1-\rho^2)^{-1}$ which has long been known to be *the* reference prior for this problem (Bayarri, 1981); as Datta mentions, this is probability matching for ρ, and avoids the marginalisation paradox. Priors based on grouping the nuisance parameters should *not* be referred to as reference priors.

2. *Reply to Liseo.* Liseo correctly stresses that there are situations where a 'precise' null hypothesis makes sense, but then he adds that he cannot see an alternative probability-based method to Bayes factors to analyze them. I would argue that the special status of a precise null may be incorporated through the *utility* function: the Bayesian analysis of the corresponding decision problem the naturally requires the derivation of the posterior distribution of the parameters; but this is done using a *regular* prior instead of a singular prior. The probability-based mechanism used to incorporate the information provided by the data is precisely the same, namely Bayes theorem followed by appropriate marginalization. The point is that, in my view, in nested hypothesis testing problems one should *not* try to find the posterior *probability* of the null (which must be zero if a regular prior is used), but one should either derive the posterior probabilities associated to interesting regions of the parameter space (which may or may not include the null) or, more to the point, or one should *judge* whether or not the null model provides a good enough explanation of the observed data, using a regular, possibly non-subjective prior.

Liseo has some queries about the definition of the utility function. Actually, only the utility difference $u(a_1, \theta_0, \theta, \omega) - u(a_0, \theta_0, \theta, \omega)$ need be specified to solve the problem, and it is natural to assume

$$u(a_1, \theta_0, \theta, \omega) - u(a_0, \theta_0, \theta, \omega) = \alpha\, \delta(\theta_0, \theta, \omega) - (c_1 - c_0),$$

where $\delta(\theta_0, \theta, \omega)$ is some measure of the 'distance' between θ and θ_0 as possible 'explanations' of the observed data, assuming that $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$ is true, and where $c_0 - c_1$ is a measure of the utility increase of using the null model when it is true. The optimal action will be to reject working as if θ were equal to θ_0 if $d(\boldsymbol{x}, \theta_0)$, the expected posterior value of $\delta(\theta_0, \theta, \omega)$, is larger that the constant $g = (c_1 - c_0)/\alpha$. Thus, c_1 and c_0 are just part of the proposed utility function.

A particular solution to the problem posed will be found for each choice of the discrepancy function. For the reasons discussed in the paper, I propose using the logarithmic divergence

$$\delta(\theta_0, \theta, \omega) = \inf_{\omega_0 \in \Omega} \int p_{\boldsymbol{x}}(\boldsymbol{y} \mid \theta, \omega) \log \left[\frac{p_{\boldsymbol{x}}(\boldsymbol{y} \mid \theta, \omega)}{p_{\boldsymbol{x}}(\boldsymbol{y} \mid \theta_0, \omega_0)} \right] d\boldsymbol{y},$$

between the assumed model $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$ and its closest approximation under the null, identified by $\omega_0 = \omega_0(\theta_0, \theta, \omega)$. This specific choice has two very important properties: (i) it measures the 'distance' between θ and θ_0 in terms of *how different are the corresponding models*, rather than in terms of the (largely irrelevant), say Euclidean distance between θ and θ_0; and (ii) it is invariant under reparametrization, so testing whether or not $\theta = \theta_0$ will produce the same result as testing $\phi(\theta) = \phi(\theta_0)$ for any one-to-one transformation $\phi = \phi(\theta)$. Liseo gives a powerful argument against the use of predictives in the definition of $\delta(\theta_0, \theta, \omega)$, when he shows that this would bring back Lindley's paradox.

Anyone unconvinced by either the scoring rule argument used to motivate the use of the logarithmic discrepancy, or by its attractive properties, may indeed choose another definition

for $\delta(\theta_0, \theta, \omega)$. Any such choice (together with some prior on the parameters) will produce a completely Bayesian, coherent answer to the decision problem posed. I think, however, that the suggested choice provides a very good candidate for general use.

Liseo correctly points out that, in the context of Example 1, for fixed n there is a one to one relation between the utility constant g and the prior variance σ_1. It is indeed known (Bernardo, 1980; Smith and Spiegelhalter, 1980) that Lindley's paradox may be avoided using singular priors if the prior probability of the null is made to depend on the sample size n; however, as mentioned before, singular priors may only be justified as an approximation to strong prior subjective beliefs, and a subjective prior may hardly be assumed to depend on the sample size.

3. *Reply to Lindley.* Lindley is obviously right when he insists on using context dependent subjective priors in any Bayesian analysis but, as mentioned before, this is certainly compatible with the use of reference priors. Indeed, if at all possible, reference posteriors should not be used on their own, but compared with subjective-based posteriors in order to be able to gauge the actual importance of prior information in the final analysis. Moreover, it is only the context and the related subjective information, which will allow a proper *interpretation* of the results.

For example, Jahn *et al.* (1987) report the result of an experiment in parapsychological research where an electronic device is used to produce a random sequence of 0's and 1's with theoretical equal probability for each of two outcomes, and a subject attempts to 'influence' the random event generator to obtain a sequence of results with a different distribution; this results in $r = 52,263,471$ observed 1's, out of $n = 104,490,000$ performed trials. The hypothesis to be tested is $\theta = \theta_0 = 1/2$. The automatic analysis provided by BRC leads to $d_r(\theta_0, r, n) = 7.03$, thus suggesting that the true value of θ *cannot* be assumed to be $1/2$, (in sharp contrast with the corresponding Bayes factor analysis: see Jefferys, 1990), but the interpretation of this fact (whether this is due to ESP, it is due to some undetected bias, or it indicates the need of a more refined physics theory), is obviously a context dependent, subjective issue over which the data *cannot* provide any information whatsoever.

It should also be mentioned here that *restricted reference priors*, obtained by maximizing the missing information within the class of priors compatible with assumed prior knowledge (see *e.g.*, Bernardo and Smith, 1994, Section 5.4.3) may actually be used as a powerful mechanism to *elicit* prior subjective knowledge. Thus, if it is desired to incorporate some knowledge in the analysis of a precise hypothesis testing situation, (say the mean and variance of θ) then the (unrestricted) reference prior should be replaced by the restricted reference prior which corresponds to this assumed knowledge, and the resulting BRC statistic will automatically incorporate this further assumption. The more information included, the closest the result will be to a strictly subjective analysis; restricted reference priors provide a continuum of solutions, ranging from the conventional reference posterior to the posterior which corresponds to any (regular) subjective prior.

Lindley mentions that with BRC (as with frequentist testing) it is possible to sample to a foregone conclusion, in the sense that, allowing optional stopping, one can sample until BRC exceeds g, and this will eventually happen with probability one. This is a mathematical fact, but has to be seen in perspective. With a regular prior, any data set, however unlikely, has a positive prior predictive density so that, if one is allowed to sample indefinitely, one would eventually get to one of these unlikely data sets which suggest the wrong decision; this makes perfect sense to me.

Finally, Lindley worries about the behaviour of BRC in higher dimensions. Work in progress with Raúl Rueda (which includes many examples) indicates that this is appropriate, but further results are definitely needed there.

4. *Reply to Mouchart.* Mouchart suggests that in some situations the prior should depend on the model. Indeed, one may well consider situations when the 'null' is of the form $M_0 = \{p(\boldsymbol{x} \mid \theta_0, \omega), p_0(\omega)\}$, and the alternative $M_1 = \{p(\boldsymbol{x} \mid \theta, \omega), p(\omega \mid \theta)\}$ but, if this is the case, then the nuisance parameter may simply be integrated out and one is left with a standard situation of the type $p(\boldsymbol{x} \mid \theta_0)$ versus $p(\boldsymbol{x} \mid \theta)$, to which the proposed method may directly be applied.

5. *Reply to O'Hagan.*

O'Hagan seems to be confused by my use of the dummy integration variable $\boldsymbol{y}$ to define the logarithmic discrepancy. As mentioned above (in the reply to Liseo), what is required is a measure of the discrepancy between the model $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$ and its best approximation under the null, as a possible description of the probabilistic mechanism that has generated the *observed* data $\boldsymbol{x}$; thus, the $\boldsymbol{y}$ in the definition of the discrepancy is simply a possible observation from the *same* probabilistic mechanism that has generated $\boldsymbol{x}$.

O'Hagan points out that with BRC the null is falsely rejected with a probability that does not tend to zero as n increases. I do not find this disturbing, but rather an expected consequence of the proposed decision-oriented approach: the choice of the threshold d^* which controls such asymptotic 'error' probability is a trade-off between missing a possible opportunity for simplification and early detection of the unsuitability of the null. Since, in my statement of the problem, one is only trying to check whether or not the null model is a suitable proxy for the *correct* model $p_{\boldsymbol{x}}(\cdot \mid \theta, \omega)$, a 'false' rejection only means that one misses an opportunity of *simplifying* the model, but the model used will nevertheless be *correct*. In contrast, if the null is false, the BRC statistic increases linearly with n so that, for sufficiently large samples, a false null will *always* be rejected. Thus, with BRC and sufficiently large samples, one will *never* be lead to using a *wrong* model.

I take note of O'Hagan's uneasiness of my description of his fractional Bayes factor as a *non-subjective* Bayes factor. Fractional Bayes factors are *not* coherent (and thus, hardly Bayesian) but, on re-reading his papers I cannot find a single attempt at using subjective prior information in the examples he considers: in my limited use of the English language, a procedure which does *not* use subjective input *is* non-subjective.

The use of the adjective 'non-subjective' to describe attempts to describe the many published procedures which try to provide a solution to the definition of an 'origin' for Bayesian inference was voted in preference to other proposed alternatives (automatic, conventional, default, fair, neutral, objective, reference, standard) by those who attended an international workshop on that topic held at Purdue University in November 1996; I personally prefer the adjective 'reference' (introduced by Box and Tiao, 1962, p. 420), but this is now mostly used to refer to the procedures I introduced in the 70's. I am afraid that, in spite of O'Hagan's 'intense dislike' for what he considers nothing less than a 'perversion', non-subjective Bayesian methods are here to stay and, I would add, for many good reasons. I would refer those interested on this important foundational issue to Bernardo (1997), ensuing discussion, and references therein.

6. *Reply to Robert.* I appreciate Robert's s positive attitude towards BRC, and support his defense of the Kullback-Leibler divergence as a most appropriate intrinsic loss function.

He wonders on the generality of the value $d^* = 5$ as an standard for scientific communications in higher dimension models. I have already mentioned work in progress with Raúl Rueda which suggests that indeed, this standard may also be used in higher dimensions; however, further work is necessary.

Robert also asks about the extension of the BRC method to non-nested cases. We have been looking at non-nested cases using simple encompassing procedures. For instance, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a set of exchangeable observations, and two alternative models, $M_1 \equiv p_1(\boldsymbol{x} \mid \boldsymbol{\theta})$

and $M_2 \equiv p_2(\boldsymbol{x} \mid \boldsymbol{\theta})$, are considered in terms of some common parameter vector $\boldsymbol{\theta}$ defined, as exchangeability requires, as the limit $\boldsymbol{\theta} = \lim_{n\to\infty} \boldsymbol{f}(x_1, \ldots, x_n)$ of some function $\boldsymbol{f}$ of the observations, then BRC may be used with the encompassing model

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}, \phi) = [p_1(\boldsymbol{x} \mid \boldsymbol{\theta})]^{\phi}[p_2(\boldsymbol{x} \mid \boldsymbol{\theta})]^{1-\phi}, \quad \phi \in \{0, 1\},$$

obtained by incorporating the *discrete* parameter ϕ. This effectively allows testing either of the two models, assuming that one of the two is correct. The derivation of the appropriate reference prior here is involved, for it requires a very careful analysis of the necessary compact approximations, but we have worked out some examples (including the 'canonical' Exponential versus Poisson problem) and the results are very encouraging.

7. *Conclusions.* Many authors have stressed that precise hypothesis testing problems are important for scientists, but the procedures proposed to give a solution to these problems have always been subject to polemic. It seems clear from the discussion that some of this polemic is due to the fact that two *different* problems are often addressed under the common heading of precise hypothesis testing. In some situations, a scientist may have reasons to have a prior distribution on the quantity of interest sharply concentrated around some specific null value; if this is the case then, after the data have been observed, he will naturally be interested in the posterior probability that the parameter lies within an small enough interval around the null; if (*and only if*) the sample size is not too big, then such probability may be approximated by the posterior probability which may be deduced from a singular prior with a mass of probability on the null. On the other hand, in many other situations, the scientist is interested instead in checking the *compatibility* of the data with the particular model identified by the null; this is the problem that BRC tries to address.

Many of us have often advocated the systematic use of decision analysis to provide reasonable, coherent solutions to any problem, including those often considered as 'pure inference' problems. In this paper, I have tried to demonstrate that a decision-oriented analysis of the problem of checking the compatibility of data with a null model suggests that the special role of the null should be incorporated in the utility structure, *not* in the prior distribution. This implies that *regular* priors should used to obtain, by maximizing the expected posterior utility, sensible coherent criteria for model criticism. An important consequence of this approach is that (possibly improper) non-subjective priors *may* indeed be used to facilitate scientific communication.

Naturally, the logarithmic discrepancy is not the only possible loss function; indeed, the discrepancy functions derived from other proper scoring rules may well be worth exploring. I believe however, that the attractive properties and the information-theoretical interpretation of the logarithmic discrepancy makes it a natural first choice. Similarly, the procedure described may be used with any prior, subjective or not. As a matter of fact a totally subjective testing procedure may be achieved by computing the subjective posterior expectation of the discrepancy, and comparing this with a subjective threshold which measures the decision-maker level of preference for the simple model when it is true. I believe, however, in the convenience of establishing some standard which may be used for scientific communication, and the combined use of the appropriate reference prior and a threshold calibrated with a canonical example provides, I think, such a convenient standard.

ADDITIONAL REFERENCES IN THE DISCUSSION

Bar-lev, S. K. and Reiser, B. (1982). An exponential subfamily which admits UMPU tests based on a single test statistic. *Ann. Statist.* **10**, 979–989.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18–31.

Berger, J. (1992). Discussion of Ghosh and Mukerjee. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 205–206.

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* **49**, 419–432.

Cox, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5**, 169–174.

Datta, G. S. (1996). On priors providing frequentist validity for Bayesian inference for multiple parametric functions. *Biometrika* **83**, 287–298.

Datta, G. S. and Ghosh, J. K. (1995). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114.

Datta, G. S. and Ghosh, M. (1995a). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.

Datta, G. S. and Ghosh, M. (1995b). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Planning and Inference* **45**, 399–411.

Dupuis, J. and Robert, C. P. (1997). Model choice in qualitative regression models. *Tech. Rep.* **9717**, CREST-INSEE, France.

Florens, J.-P. and Mouchart, M. (1993). Bayesian testing and testing Bayesians. *Handbook of Statistics*, (G. S. Maddala and C. R. Rao, eds.), Amsterdam: North-Holland, Ch. 11.

George, E. (1998). Discussion of Clyde's paper. *In this volume*.

Goutis, C. and Robert, C. P. (1997). Selection between hypotheses using estimation criteria. *Ann. Econom. Stat.* **46**, 1–22.

Goutis, C. and Robert, C. P. (1998). Model choice in generalized linear models: a Bayesian approach via Kullback–Leibler projections. *Biometrika* **85**, 29–37.

Jahn, R. G., Dunne, B. J. and Nelson R. D. (1987). Engineering anomalies research. *J. Scientific Exploration* **1**, 21–50.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *J. Scientific Exploration* **4**, 153–169.

Lehmann, E .L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statist. Sci.* **5**, 160–168.

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision* **40**, 191–214.

Rubin, H. and Sethuraman, J. (1966). Bayes risk efficiency. *Sankhyā A* **27**, 347–356.

Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83**, 55–65.