SORT 30 (1) January-June 2006, 3-28

On the frequentist and Bayesian approaches to hypothesis testing

Elías Moreno and F. Javier Girón

University of Granada and University of Málaga

Abstract

Hypothesis testing is a model selection problem for which the solution proposed by the two main statistical streams of thought, frequentists and Bayesians, substantially differ. One may think that this fact might be due to the prior chosen in the Bayesian analysis and that a convenient prior selection may reconcile both approaches. However, the Bayesian robustness viewpoint has shown that, in general, this is not so and hence a profound disagreement between both approaches exists.
In this paper we briefly revise the basic aspects of hypothesis testing for both the frequentist and Bayesian procedures and discuss the variable selection problem in normal linear regression for which the discrepancies are more apparent. Illustrations on simulated and real data are given.

MSC: 62A01; 62F03; 62F15

Keywords: Bayes factor, consistency, intrinsic priors, loss function, model posterior probability, *p*-values.

1 Introduction

In parametric statistical inference estimating parameters and hypothesis testing are two basic but different problems, although some Bayesian estimation devices have gained some popularity as hypothesis testing tools. For instance, high posterior probability regions have been used as acceptance regions for the null hypothesis (see, for instance, the analysis of variance solution proposed by Lindley 1970, Box and Tiao 1992). This is misleading as far as the hypotheses to be tested do not play any role in the construction of such acceptance regions.

Address for correspondence: Elías Moreno and F. Javier Girón. Department of Statistics and O.R. University of Granada and University of Málaga.
Received: November 2005

To distinguish between estimation and testing was strongly recommended by Jeffreys (1961, pp. 245-249), mainly because the methods commonly used for estimation are typically not suitable for hypothesis testing. Thus, we think it is timely to devote a separate paper to discuss the setting and tools devised for hypothesis testing from the frequentist and Bayesian perspectives. We want to thank the Editor of SORT for inviting us to contribute on this topic.

In this paper we deal with frequentist and Bayesian parametric hypothesis testing procedures. A third approach, based solely on the likelihood function, which we do not discuss here, is to be found in Royall (1997) and Pawitan (2001). As Pawitan (2001, p. 15) states: *The distinguishing view is that inference is possible directly from the likelihood function; this is neither Bayesian nor frequentist, and in fact both schools would reject such a view as they allow only probability-based inference*.

From the frequentist viewpoint two closely related methods have been developed. One is the Neyman-Pearson theory of significance tests and the other one is based on Fisher's notion of p-values. Here, we shall give arguments that make the p-values to be preferable to significance tests.

On the Bayesian side, robustness with respect to the prior showed that there is a strong discrepancy between the frequentist and Bayesian solutions to parametric hypothesis testing (Berger 1985, 1994, Berger and Delampady 1987, Berger and Sellke 1987, Berger and Mortera 1999, Casella and Berger 1997, Moreno and Cano 1998, Moreno 2005, among others). This means that the discrepancy is not due to the prior chosen for the Bayesian analysis but it is of a more fundamental nature which is inherent to the procedures. In particular, there is a marked difference on the way frequentist and Bayesian methods account for the sample size, and the dimensions of the null and the alternative parametric spaces.

Since subjective prior distributions for the parameters of the models involved in hypothesis testing are not generally available, and their use is perceived as the weak point in the Bayesian implementation, objective prior distributions will be employed in this paper. By objective priors we mean priors that only depend on the sampling model and theoretical training samples. This priors are called intrinsic priors (Berger and Pericchi 1996, Moreno 1997, Moreno *et al.* 1998) and their merits can be judged for each specific application. We remark that they have been proved to behave extremely well in a wide variety of problems (Casella and Moreno 2004, 2005, Girón and Moreno 2004, 2005, Moreno *et al.* 1999, 2000, 2003, 2005, Moreno and Liseo 2003).

The rest of the paper is organized as follows. The second section is devoted to briefly describing significance tests and p-values. Section 3 reviews the Bayesian testing machinery and justifies the need for objective priors. The challenging problem of testing whether the means of two normal distributions with unknown variances are equal is considered in Section 4. A comparison of the frequentist and Bayesian testing procedures for the normal linear regression model, including a discussion on some fundamental issues of the variable selection problem, is given in Section 5, and some conclusions and recommendations are given in Section 6.

2 Significance tests and p-values

Let X denote an observable random variable and $\mathbf{x} = (x_1, \ldots, x_n)$ an available sample from either the model $P_0(x)$ or $P_1(x)$ with probability densities $f_0(x)$ or $f_1(x)$, respectively. Suppose that we want to choose between either the null hypothesis $H_0 : f_0(x)$ or the alternative $H_1 : f_1(x)$. This is the simplest hypothesis testing formulation.

For this problem the well-known Neyman-Pearson theory of significance tests proposes a subset of the sample space $\mathcal{R}^n$, the so-called critical or rejection region,

$$W_\alpha = \left\{ \mathbf{y} : \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} \geq c_\alpha \right\}, \tag{1}$$

as the region containing evidence against the null hypothesis. The threshold c_α is determined so that the probability of the critical region under the null is α, that is

$$P_0(W_\alpha) = \int_{W_\alpha} f_0(\mathbf{y})\, d\mathbf{y} = \alpha.$$

Given the data $\mathbf{x}$, the null H_0 is rejected at the significance level α if $\mathbf{x} \in W_\alpha$, and accepted otherwise. The value α is usually taken to be small so that we have a small probability of rejecting the null when it is true. Typically α is chosen to be 0.05 or 0.01.

Fisher criticized the notion of significance of the Neyman-Pearson theory and proposed replacing c_α in (1) with the observed likelihood ratio $\lambda_n(\mathbf{x}) = f_1(\mathbf{x})/f_0(\mathbf{x})$, so that the above critical region becomes

$$W_n(\mathbf{x}) = \left\{ \mathbf{y} : \frac{f_1(\mathbf{y})}{f_0(\mathbf{y})} \geq \lambda_n(\mathbf{x}) \right\},$$

that now depends on the observed data $\mathbf{x}$. The probability of this region under the null, $p = P_0(W_n(\mathbf{x}))$ say, is called the p-value of the data $\mathbf{x}$, and it is interpreted in such a way that a small enough p-value implies evidence against H_0. The smaller the p-value the stronger the evidence against the null.

Significance tests and p-values substantially differ. While the p-value depends on the data thus giving a measure of how strongly the observed data reject the null, the significance test does not provide such a measure. This simple fact implies that, for the dicotomous testing problem, the p-value is consistent under the null and the alternative, while the significance test is consistent under the alternative but it is not under the null.

Indeed, when sampling from the null and the sample size grows to infinity the likelihood ratio statistic $\lambda_n(\mathbf{x})$ tends to zero under P_0, and the corresponding sequence of p-values tends to 1. This implies that, asymptotically, there is no evidence at all to reject the null. Alternatively, when the sample comes from the alternative hypothesis, the statistic $\lambda_n(\mathbf{x})$ tends to ∞ under P_0, and hence the corresponding sequence of p-

values tends to 0 showing an increasing evidence to reject the null. On the other hand, when sampling from the null, the significance test will reject H_0 with probability α even when the sample size increases to infinity. This means that the significance test is not a consistent procedure under the null, although it is consistent when sampling from the alternative (Casella and Berger 1990, pp. 381-382, Wilks 1962, pp. 411).

Under some regularity conditions on the likelihood, this theory can be extended to the more general situation of considering a family of probability densities $\{f(x|\theta), \theta \in \Theta\}$, where Θ might be a multidimensional parametric space, and the null $H_0 : \theta \in \Theta_0$ and the alternative $H_1 : \theta \in \Theta_1$ may contain more than one density. To preserve the good properties of the mentioned simplest hypothesis testing procedures, the hypotheses must be nested, that is $\Theta_0 \subset \Theta_1$, and also we must have $k_0 = \dim(\Theta_0)$ strictly smaller than $k_1 = \dim(\Theta_1)$. The critical region now is

$$W_n(\mathbf{x}) = \left\{ \mathbf{y} : \frac{f(\mathbf{y}|\hat{\theta}_1(\mathbf{y}))}{f(\mathbf{y}|\hat{\theta}_0(\mathbf{y}))} \geq \hat{\lambda}_n(\mathbf{x}) \right\},$$

where $\hat{\theta}_1(\mathbf{y})$, $\hat{\theta}_0(\mathbf{y})$ are the MLE's of θ in the spaces Θ_1 and Θ_0, respectively, and $\hat{\lambda}_n(\mathbf{x}) = f(\mathbf{x}|\hat{\theta}_1(\mathbf{x}))/f(\mathbf{x}|\hat{\theta}_0(\mathbf{x}))$.

Two important difficulties now arise with the p-values. First, it is not clear how the probability of $W_n(\mathbf{x})$ under the null can be computed when either the null is composite or the distribution induced by $\hat{\lambda}_n(\mathbf{y})$ depends on some (nuisance) parameter, as in the Behrens-Fisher problem which we briefly deal with in Section 4.

Second, while a small p-value might contain evidence against the null for a dataset, the same p-value for a different sample size and/or for a null parameter space of different dimension might not contain the same evidence against the null. In fact, when we consider multiple tests, i.e. when we want to test that some subsets of regression coefficients are zero, as in variable selection, the frequentist literature (see, for instance Miller 2002) recognizes that the meaning of the p-value should depend on the dimensions of the null and the alternative parameter spaces and, consequently, provides a variety of methods to correct the p-value to account for this. It has also been recognized that the bigger the sample size the smaller the evidence of a given p-value against the null so that it is also necessary to correct the p-value to account for the sample size. These considerations have prompted the need to introduce frequentist criteria based on statistics that, in some way, adjust for all the varying parameters in the model such as the sample size and the dimensions of the null and the alternative hypothesis, such as the adjusted R^2, Mallow's C_p or the AIC criteria.

In summary, the meaning of the p-value is unfortunately unclear. Its interpretation should depend on the dimension of the null space, the dimension of the alternative space, and the sample size in an unknown, and probably complex and non-trivial, way; as a consequence, the calibration of a p-value is deemed to be a very difficult task, although some attempts for calibrating the p-values can be found in Sellke *et al.* (2001) and Girón *et al.* (2004).

Furthermore, although a p-value is derived as a probabilistic statement and, consequently, lies between zero and one, it cannot be interpreted as the posterior probability that the null is true – this is an instance of the well known *transposed conditional or prosecutor's fallacy*. However, practitioners of the p-values have very often this *wrong* probability interpretation in mind, maybe because this provides them with some sort of a (wrong) measurement devise for calibration.

3 Bayesian hypothesis testing

From a Bayesian viewpoint the testing problem is treated as follows. Consider the simplest testing problem where we have to choose either the model $M_0 : f_0(x)$ or $M_1 : f_1(x)$ based on the observations $\mathbf{x} = (x_1, \ldots, x_n)$. Let d_i denote the decision of choosing M_i and let P be the prior probability defined on the model space $\{M_0 , M_1\}$. Assume that a loss $L(d_i, M_j) = c_{ij}$, $i, j = 0, 1$, is incurred when we make the decision d_i and the true model is M_j (for other loss functions in model selection, see San Martini and Spezzaferri 1984, and Bernardo and Smith 1994).

Assuming that the loss for a correct decision is zero, that is $c_{ii} = 0$, and $c_{ij} > 0$ otherwise, d_1 is the optimal decision when the posterior risks satisfy $R(d_0|\mathbf{x}) > R(d_1|\mathbf{x})$. This implies the following inequality

$$P(M_0|\mathbf{x}) < \frac{c_{01}}{c_{01} + c_{10}}.$$

By Bayes theorem, this is equivalent to the inequality

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > \frac{c_{10}}{c_{01}} \frac{P(M_0)}{P(M_1)}.$$

Notice that the value of $P(M_0|\mathbf{x})$ is a measure, in a probability scale, of the strength we have in accepting the model M_0. We now observe a first important difference between the p-value and the Bayesian report. While the former is obtained by integration over a region of the sample space, namely the rejection region, the latter is obtained directly from the loss function and the prior probabilities assigned to the models.

The extension to the more realistic case of a parametric families of densities is straightforward. Consider the sampling models $\{f(x|\theta_0), \theta_0 \in \Theta_0\}$ and $\{f(x|\theta_1), \theta_1 \in \Theta_1\}$. A complete Bayesian specification of the models needs prior distributions for the parameter θ_0 and θ_1, that is

$$M_0 : \{f(x|\theta_0), \pi_0(\theta_0)\},$$

and

$$M_1 : \{f(x|\theta_1), \pi_1(\theta_1)\}.$$

Then under the loss function given above, assuming $c_{01} = c_{10}$ and $P(M_0) = P(M_1) = 1/2$, the model M_0 is to be rejected if

$$P(M_0|\mathbf{x}) = \frac{1}{1 + B_{10}(\mathbf{x})} < 1/2, \tag{2}$$

where $B_{10}(\mathbf{x})$, the Bayes factor for models $\{M_1, M_0\}$, is the ratio of the marginal density, sometimes called the integrated or marginal likelihood, of the data under the two models, that is

$$B_{10}(\mathbf{x}) = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} = \frac{\int f(x|\theta_1)\pi_1(\theta_1)\, d\theta_1}{\int f(x|\theta_0)\pi_0(\theta_0)\, d\theta_0}.$$

We note that a second important difference between the p-values and the Bayesian method is that while in the p-value approach the parameters of the null and the alternative hypothesis are estimated using the maximum likelihood method, in the Bayesian approach they are integrated out using prior distributions.

For nested models it can be shown that, under mild conditions, the Bayesian procedure chooses the correct model with probability that tends to one as the sample size increases, so that it is a consistent procedure under both the null and the alternative hypothesis (see, for instance, O'Hagan and Forster 2004, p.182).

This approach needs the specification of the losses c_{ij}, $i, j = 0, 1$, the prior distributions for parameters $\pi_0(\theta_0)$ and $\pi_1(\theta_1)$, and the model prior $(P(M_0), P(M_1))$. While the specification of the loss function and the model prior seems to be a plausible task in real applications, the specification of subjective priors for parameters is admittedly a hard task. For instance, when θ_1 represents the vector of regression coefficients and the variance error of a regression model, to specify the prior is far from trivial. More so when the null parameter θ_0 is a subvector of the set of regression coefficients of θ_1, thus indicating that a submodel is being considered plausible and a testing problem is called for.

This problem is an important example in which the use of objective priors is fully justified. Unfortunately, the priors considered for estimation, as the Jeffreys or the reference priors by Berger and Bernardo (1992), are typically improper so that they depend on arbitrary multiplicative constants that leave the Bayes factor ill-defined as the following simple example shows.

Example 1 Suppose that X is a random variable with distribution $N(x|\mu, \sigma^2)$, with both parameters unknown, and we want to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. This is equivalent to choosing between models $M_0 : \{x|\sigma_0 \sim N(x|0, \sigma_0), \pi_0(\sigma_0)\}$ and

$\{x|\mu,\sigma_1 \sim N(x|\mu,\sigma_1),\pi_1(\mu,\sigma_1)\}$. The reference prior for the parameter of the null model is $\pi_0(\sigma_0) = c_0/\sigma_0$, where c_0 is an arbitrary positive constant that cannot be specified because π_0 is improper. Likewise, the reference prior for the parameter of the alternative model is $\pi_1(\mu,\sigma_1) = c_1/\sigma_1$, where again c_1 is an arbitrary positive constant. Therefore, the Bayes factor $B_{10}(\mathbf{x})$ is defined up to the multiplicative constant c_1/c_0, whatever the data $\mathbf{x}$ might be.

3.1 Intrinsic priors

Consider the Bayesian models

$$M_0 : \{f(x|\theta_0),\pi_0^N(\theta_0)\},$$

and

$$M_1 : \{f(x|\theta_1),\pi_1^N(\theta_1)\}$$

where π_0^N and π_1^N are objective, or default, improper priors.

Lempers (1971, section 5.3) overcomes the difficulty that the Bayes factor for improper priors is not well defined by considering a partial Bayes factor. This is a Bayes factor constructed as follows. A part of the sample $\mathbf{x}$, the training part, is devoted to converting the reference improper priors for the parameters of the null and alternative models into proper posteriors. Then, with the rest of the sample the Bayes factor is computed using these proper posteriors as priors. That is, the whole sample $\mathbf{x}$ is split into $(\mathbf{x} = \mathbf{x}(t),\mathbf{x}(n-t))$, where $\mathbf{x}(t)$ is the training sample of the vector and $\mathbf{x}(n-t)$ the remaining one. Then, the posterior distributions for the above priors are

$$\pi_i(\theta_i|\mathbf{x}(t)) = \frac{f(\mathbf{x}(t)|\theta_i)\pi_i^N(\theta_i)}{\int f(\mathbf{x}(t)|\theta_i)\pi_i^N(\theta_i)\,d\theta_i}, \quad i = 0,1,$$

that are now proper. Now, using the above posteriors as priors, the Bayes factor for the rest of the data $\mathbf{x}(n-t)$ turns out to be

$$B_{10}^P(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi_1^N(\theta_1)\,d\theta_1}{\int f(\mathbf{x}|\theta_0)\pi_0^N(\theta_0)\,d\theta_0}\frac{\int f(\mathbf{x}(t)|\theta_0)\pi_0^N(\theta_0)\,d\theta_0}{\int f(\mathbf{x}(t)|\theta_1)\pi_1^N(\theta_1)\,d\theta_1} = B_{10}^N(\mathbf{x})B_{01}(\mathbf{x}(t)).$$

Note that the partial Bayes factor is a well defined Bayes factor that uses each of the components of the sample only once. However, it does depend on the specific training sample $\mathbf{x}(t)$ we choose.

To avoid the arbitrariness in choosing the training sample Berger and Pericchi (1996) suggested computing the partial Bayes factors for all possible training samples with minimal size $\mathbf{x}(\ell)$, and then computing the mean of those partial Bayes factors. The number ℓ is chosen so that the marginals of $\mathbf{x}(\ell)$ with respect to the improper priors are positive and finite so that the factor $B_{01}(\mathbf{x}(\ell))$ is well-defined up to the multiplicative constant c_0/c_1.

The resulting value was called the arithmetic intrinsic Bayes factor, which does not depend on any arbitrary constant nor the particular training sample. Then, the Bayes factor appearing in (2) is replaced with the arithmetic intrinsic Bayes factor for computing the null posterior probability "as if" it were a Bayes factor.

We observe that the arithmetic intrinsic Bayes factor is a mean of (partial) Bayes factors and hence it reuses the sample observations. We also observe that for some subsamples of minimal size it might be the case that the marginal $m_i(\mathbf{x}(\ell))$ could be zero or infinite. In that case, the Bayes factor is not well defined and we adopt the convention of not considering those subsamples. This implies that the arithmetic intrinsic Bayes factor might be quite unstable depending on the nature of the sample at hand. Some samples can have very many nice subsamples of minimal size but others may not have so many.

However, to use the arithmetic intrinsic Bayes factor "as if" it were a Bayes factor is, in our opinion, not the best use we can give to the arithmetic intrinsic Bayes factor. It can be better employed as a tool for constructing priors. In fact, the arithmetic intrinsic Bayes factor is not a Bayes factor although as the sample size increases it becomes more and more stable and tends to be a Bayes factor for the so called intrinsic priors. Thus, if we use theoretical training samples instead of actual samples along with a limiting procedure we end up with intrinsic priors (Moreno *et al.* 1998).

Given the model

$$M_0 : \{f(x|\theta_0), \pi_0^N(\theta_0)\}$$

and

$$M_1 : \{f(x|\theta_1), \pi_1^N(\theta_1)\},$$

where $f(x|\theta_0)$ is nested into $f(x|\theta_1)$ and π_1^N is improper, the following statements can be proven.

(i) The intrinsic prior for θ_1 conditional on an arbitrary but fixed point θ_0 is given by

$$\pi^I(\theta_1|\theta_0) = \pi_1^N(\theta_1) E_{X(\ell)|\theta_1} \frac{f(X(\ell)|\theta_0)}{\int f(X(\ell)|\theta_1)\pi_1^N(\theta_1)\, d\theta_1},$$

where $X(\ell)$ is a vector of dimension ℓ with i.i.d components and distribution $f(x|\theta_1)$, such that

$$0 < \int f(X(\ell)|\theta_1)\pi_1^N(\theta_1)\,d\theta_1 < \infty,$$

ℓ being the smallest natural number satisfying the above inequality. Roughly speaking, ℓ coincides with the dimension of θ_1 .

(ii) $\pi^I(\theta_1|\theta_0)$ is a probability density for θ_1, for any fixed θ_0.

(iii) If the default prior $\pi_0^N(\theta_0)$ is also improper, the ratio

$$B_{10}^I(\mathbf{x}) = \frac{\int f(\mathbf{x}|\theta_1)\pi^I(\theta_1|\theta_0)\pi_0^N(\theta_0)\,d\theta_0\,d\theta_1}{\int f(x|\theta_0)\pi_0^N(\theta_0)\,d\theta_0} \tag{3}$$

is the limit of the sequence of Bayes factors given by

$$B_i = \frac{\int_{C_i} f(\mathbf{x}|\theta_1)\pi^I(\theta_1|\theta_0)\pi_0^N(\theta_0|C_i)\,d\theta_0\,d\theta_1}{\int_{C_i} f(\mathbf{x}|\theta_0)\pi_0^N(\theta_0|C_i)\,d\theta_0},$$

where $\{C_i, i \geq 1\}$ is a covering monotone increasing sequence of sets in Θ_0. Of course it can be shown that the limiting value (3) does not depend on the chosen sequence $\{C_i, i \geq 1\}$.

In summary, intrinsic priors are well defined priors for testing problem involving nested models. For some particular non-nested models intrinsic priors can also be defined (Cano *et al*. 2004). The Bayes factor for intrinsic priors can be seen as the stabilized version of the arithmetic intrinsic Bayes factor. Further, as the sample size n tends to infinity the sequence of intrinsic posterior probabilities of model M_0

$$P(M_0|x_1,\ldots,x_n) = \frac{1}{1 + B_{10}^I(x_1,\ldots,x_n)}$$

tends to one when sampling from the null and tends to zero when sampling from the alternative, so that the intrinsic Bayesian procedure is consistent; for a result in this direction see Moreno and Girón (2005a) for the case of the general normal linear model.

4 The two sample problem

A p-value does not always exist, so that some sort of "approximation" is in that case necessary. A classical example in which this situation ocurrs is that of comparing the means of two normal distributions with unknown variances. Let $N(x_1|\mu_1,\sigma_1^2)$, $N(x_2|\mu_2,\sigma_2^2)$ be two normal distributions where the means μ_1, μ_2 and variances σ_1^2, σ_2^2 are unknown. Suppose that samples $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})$ and $\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2})$ have been drawn independently for each of the distributions, and we are interested in testing the null $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$.

Under the frequentist point of view this problem is easily solved when $\sigma_1 = \sigma_2$ or when $\sigma_1 = k\sigma_2$ and k is known. In fact, an exact p-value can be computed by using a test statistic which follows a t-distribution.

However, standard normal theory cannot be applied when the quotient betweeen the variances σ_1^2 and σ_2^2 is unknown. This is the well-known Behrens-Fisher problem and we emphasize the fact that an exact p-value does not exist. This is a theoretically relevant result that demonstrates that frequentist testing procedures cannot be applied to some important problems. Certainly this theoretical gap is not such a serious problem from the applications viewpoint since "good" approximations – to a nonexistent solution! – were given by Fisher (1936), Wald (1955), and Welch (1947).

Under the Bayesian viewpoint the problem was "solved" by regarding it as a problem of interval estimation of the parameter $\lambda = \mu_1 - \mu_2$. From the posterior distribution of λ a $(1 - \alpha)$ highest posterior interval was computed and *the result was declared to be significant if this interval did not contain the origin* (Lindley 1970, pp. 92-93).

Notice that the posterior distribution of λ on which the inference is based – a location-scale transformation of the standard Behrens-Fisher distribution (Girón *et al.* 1999) – is obtained under the condition that $\mu_1 \neq \mu_2$; otherwise, if $\mu_1 = \mu_2$ is assumed, the posterior distribution of λ would be a point mass on zero. Therefore, using this procedure the key function for testing the null $H_0 : \mu_1 = \mu_2$ is the posterior distribution of λ conditional on the alternative hypothesis which otherwise has a posterior probability equal to zero. Of course, this cannot be the solution to the Behrens-Fisher testing problem.

In Moreno *et al.* (1999) it was shown that the Behrens-Fisher problem can be formulated as a model selection problem for nested models for which an intrinsic Bayesian solution exists. Indeed, under the null, the Bayesian default sampling model is

$$M_0 : f_0(x_1, x_2|\theta_0) = N(x_1|\mu, \tau_1^2)N(x_2|\mu, \tau_2^2),\ \pi_0^N(\theta_0) = \frac{c_0}{\tau_1\tau_2},$$

and under the alternative is

$$M_1 : f_1(x_1, x_2|\theta_1) = N(x_1|\mu_1, \sigma_1^2)N(x_2|\mu_2, \sigma_2^2),\ \pi_1^N(\theta_1) = \frac{c_1}{\sigma_1\sigma_2},$$

where $\theta_0 = (\mu, \tau_1, \tau_2)$, $\theta_1 = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, π_i^N is the reference prior, and c_0, c_1 are arbitrary positive constants.

Applying the standard intrinsic methodology to these models the intrinsic prior for the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ conditional on a null point μ, τ_1, τ_2, is shown to be

$$\pi^I(\theta_1|\theta_0) = \prod_{i=1}^{2} N\left(\mu_i|\mu, \frac{\tau_i^2 + \sigma_i^2}{2}\right) HC^+(\sigma_i|0, \tau_i),$$

where $HC^+(\sigma_i|0,\tau_i)$ denotes the half-Cauchy distribution on the positive part of the real line located at 0 and with scale parameter τ_i. Under the conditional intrinsic prior the μ_i's are independent and centered at the null parameter μ and the σ_i's are also independent. Hence, the unconditional intrinsic prior distribution for μ_i is a mixture of normal distributions which has no moments. A nice property to be expected from an objective prior.

For the samples $\mathbf{x}_1$, $\mathbf{x}_2$ having size, mean and variance $(n_1, \bar{x}_1, s_1^2)$, $(n_2, \bar{x}_2, s_2^2)$ respectively, the Bayes factor for the intrinsic priors $(\pi_0^N(\theta_0), \pi^I(\theta_1))$ is given by

$$B_{10}^I(\mathbf{x}_1,\mathbf{x}_2) = \frac{2}{\pi^{5/2}} \frac{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}{\Gamma(\frac{n_1+n_2-1}{2})} \frac{B}{A},$$

where $B = \int_{-\infty}^{\infty} \left[\prod_{i=1}^{2} I_i(\mu)\right] d\mu$,

$$I_i(\mu) = \int_0^{\pi/2} \frac{d\varphi_i}{b_i(\mu,\varphi_i)},$$

$$b_i(\mu,\varphi_i) = \frac{(\sin\varphi_i)^{n_i-1}}{\left(\frac{1}{2}+\frac{\sin^2\varphi_i}{n_i}\right)^{-1/2}} \left(\frac{n_i s_i^2}{\sin^2\varphi_i} + \frac{(\bar{x}_i-\mu)^2}{\left(\frac{1}{2}+\frac{\sin^2\varphi_i}{n_i}\right)^{n_i/2}}\right)^{n_i/2},$$

and

$$A = \int_0^{\pi/2} \frac{d\varphi}{a(\varphi)},$$

$$a(\varphi) = \frac{\sin^{n_1}\varphi\cos^{n_2}\varphi}{\left(\frac{\sin^2\varphi}{n_1}+\frac{\cos^2\varphi}{n_2}\right)^{-1/2}} \left(\frac{n_1 s_1^2}{\sin^2\varphi} + \frac{n_2 s_2^2}{\cos^2\varphi} + \frac{(\bar{x}_1-\bar{x}_2)^2}{\frac{\sin^2\varphi}{n_1}+\frac{\cos^2\varphi}{n_2}}\right)^{(n_1+n_2-1)/2}.$$

It is easy to see that the Bayes factor for intrinsic priors B_{10}^I depends on the sample $(\mathbf{x}_1,\mathbf{x}_2)$ through the statistic $(s_1^2, s_2^2, |\bar{x}_1-\bar{x}_2|, n_1, n_2)$. Since the p-value in the Welch's approximation also depends on the the sample through this statistic it follows that there is a one-to-one relationship between the p-values and the null model posterior probabilities.

As an illustration of this relationship, in Table 1 we display p-values and null model posterior probabilities for sample observations with

$$n_1 = 200,\ s_1^2 = 12,\ n_2 = 120,\ s_2^2 = 40,$$

Table 1: *Comparison of p-values and null posterior probabilities for* $n_1 = 200$ *and* $n_2 = 120$.

$\lvert\bar{x}_1 - \bar{x}_2\rvert$	t	p-value	$P(M_0\lvert\mathbf{x}_1, \mathbf{x}_2)$
0.0	0.00	1.00	0.94
0.5	0.79	0.43	0.92
1.2	1.91	0.06	0.72
1.3	2.06	0.04	0.65
1.4	2.22	0.03	0.57
1.5	2.38	0.02	0.49
2.0	3.17	0.001	0.10

and several values of the difference $|\bar{x}_1 - \bar{x}_2|$ and the corresponding t statistic defined as

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

From the numbers in Table 1 we conclude that when the values of $|\bar{x}_1 - \bar{x}_2|$ are close to zero or large enough both procedures make the correct decision. However, when the empirical evidence is not conclusive, a situation far beyond intuition for which statistical methods are unavoidable, there is a strong disagreement between the report provided by Welch's p-values and that of the intrinsic null posterior probabilities. For instance, for values of the t statistic between 2.00 and 2.35, the p-values show evidence against the null hypothesis, stronger as t increases, while the null posterior probabilities show the opposite, they still favour the null hypothesis.

Such a disagreement heavily depends on the sample size. If we reduce the above sample sizes to $n_1 = 20$ and $n_2 = 12$, while maintaining the values of s_1^2 and s_2^2, the frequentist and Bayesian reports are not so strongly contradictory, as seen from Table 2, and it may happen that a p-value accepts the null but the corresponding posterior probability of the null may be less than 0.5 and then the Bayesian test rejects it. For instance, for $|\bar{x}_1 - \bar{x}_2| = 4.22$ or $t = 2.04$, the p-value is 0.06 while the null posterior probability is smaller than 0.5.

Table 2: *Comparison of p-values and null posterior probabilities for* $n_1 = 20$ *and* $n_2 = 12$.

$\lvert\bar{x}_1 - \bar{x}_2\rvert$	t	p-value	$P(M_0\lvert\mathbf{x}_1, \mathbf{x}_2)$
0.00	0.00	1.00	0.83
2.20	1.06	0.30	0.75
4.22	2.04	0.06	0.46
5.00	2.42	0.03	0.32
10.00	4.80	0.002	0.008

In passing, we note that when $|\bar{x}_1 - \bar{x}_2| = 0$ we expect both the p-value and null posterior probability to be large. Since the sample size is finite we should not expect

the p-value to attain its maximum value of one, but it does. However, the null posterior probability is always strictly smaller than one.

5 Testing hypotheses in linear regression

A scenario where the discrepancies between the p-values and the objective Bayesian test are apparent is that of testing that some regression coefficients of a linear regression model are equal to zero. Suppose that the observable random variable y follows the normal linear model

$$y = \sum_{i=1}^{k} \alpha_i x_i + \varepsilon,$$

where the random error term $\varepsilon \sim N(\varepsilon|0, \sigma^2)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^t$ is the vector of regression coefficients, and $(x_1, \ldots, x_k)$ is a set of potential explanatory variables. Given n independent observations $\mathbf{y} = (y_1, \ldots, y_n)^t$ from the model and denoting the design matrix by

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ x_{21} & \ldots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix},$$

the likelihood function of $(\boldsymbol{\alpha}, \sigma)$ is given by the density of a $N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$, where it is assumed that $\mathbf{X}$ is of full rank k, $(k < n)$.

Consider the partition of $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha}^t = (\boldsymbol{\alpha}_0^t, \boldsymbol{\alpha}_1^t)$ and the corresponding partition of the columns of $\mathbf{X} = (\mathbf{X}_0|\mathbf{X}_1)$, so that $\mathbf{X}_0$ is of dimensions $n \times k_0$ and $\mathbf{X}_1$ is $n \times k_1$, where $k_1 = k - k_0$.

In this setting, an important problem consists in testing that some of the covariates have no influence on the variable y. That is, we are interested in testing the null $H_0 : \boldsymbol{\alpha}_0 = 0$ versus the alternative $H_1 : \boldsymbol{\alpha}_0 \neq 0$. This is the natural way of reducing the complexity of the original linear model proposed. If the null is accepted the implication is that the covariates $x_1, \ldots, x_{k_0}$ will not be considered as explanatory variables.

5.1 The uniformly most powerful test

The frequentist testing procedure, derived from the likelihood ratio test, is based on the distribution of the ratio $\mathcal{B}_n = \mathrm{SS}/\mathrm{SS}_1$ of the quadratic forms

$$\mathrm{SS} = \mathbf{y}^t(\mathbf{I}_n - \mathbf{H})\mathbf{y}, \quad \mathrm{SS}_1 = \mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ and $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t$ are the hat matrices of the full and the reduced models, respectively. It is well known, from standard linear model theory, that the sampling distribution of the $\mathcal{B}_n$ statistic under the null $H_0 : \boldsymbol{\alpha}_0 = 0$ is

$$\mathcal{B}_n|H_0 \sim Be\left(\cdot|\frac{n-k}{2}, \frac{k_0}{2}\right),$$

where $Be(\cdot|\alpha, \beta)$ denotes the beta distribution with parameters α and β.

When sampling from the alternative $H_1 : \boldsymbol{\alpha}_0 \neq 0$, the corresponding distribution is

$$1 - \mathcal{B}_n|H_1 \sim Be'\Big(\cdot|\frac{k_0}{2}, \frac{n-k}{2}; \delta\Big),$$

where

$$\delta = \boldsymbol{\alpha}_0^t\mathbf{X}_0^t(\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_0\boldsymbol{\alpha}_0$$

and $Be'(\cdot|\alpha, \beta; \delta)$ denotes the noncentral beta distribution with parameters α and β and noncentrality parameter δ. If $H_0 : \boldsymbol{\alpha}_0 = 0$ is true, then $\delta = 0$, and the noncentral distribution reduces to the central one.

The UMP test of size α (Lehmann 1986, theorem 5, pp. 300 and pp. 369) has the following critical region

$$\text{Reject } H_0 \text{ when } \mathcal{B}_n \leq I_\alpha^{-1}\Big(\frac{n-k}{2}, \frac{k_0}{2}\Big),$$

where $I_\alpha^{-1}((n-k)/2, k_0/2)$ denotes the α fractile of the Beta distribution $Be(\cdot|((n-k)/2, k_0/2)$.

Likewise, for a given sampling value $\mathcal{B}_n$, the p-value is given by

$$p = \int_0^{\mathcal{B}_n} be\left(z|\frac{n-k}{2}, \frac{k_0}{2}\right) dz.$$

where $be(z|(n-k)/2, k_0/2)$ denotes the density of the corresponding beta distribution.

We remark that the p-value is an increasing function of $\mathcal{B}_n$ so that small values of $\mathcal{B}_n$ contain evidence against the null hypothesis.

This test is usually written in terms of the F-statistic, which is related to $\mathcal{B}_n$, by

$$F = \frac{n-k}{k_0}\frac{1-\mathcal{B}_n}{\mathcal{B}_n},$$

although for numerical illustrations it is more convenient to use the bounded $\mathcal{B}_n$ statistic instead of the unbounded F.

5.2 The objective Bayesian test

The default Bayesian formulation of this testing problem would be that of choosing between the Bayesian models

$$M_0 : N_n(\mathbf{y}|\mathbf{X}_1\gamma_1, \sigma_0^2\mathbf{I}_n),\ \pi_0^N(\gamma_1, \sigma_0) = \frac{c_0}{\sigma_0},$$

and

$$M_1 : N_n(\mathbf{y}|\mathbf{X}\alpha, \sigma_1^2\mathbf{I}_n),\ \pi_1^N(\boldsymbol{\alpha}, \sigma_1) = \frac{c_1}{\sigma_1},$$

where π^N represents the usual improper reference prior (Berger and Bernardo, 1992) for estimating the regression coefficients and the standard error. Unfortunately these priors are improper and hence cannot be used for solving the above testing problem.

Application of the standard intrinsic methodology (Moreno, Girón and Torres 2003, Girón *et al*. 2004) renders the intrinsic priors of $(\boldsymbol{\alpha}, \sigma_1)$ conditional on (γ_1, σ_0) as

$$\pi^I(\boldsymbol{\alpha}, \sigma_1|\gamma_1, \sigma_0) = \frac{2}{\pi\sigma_0(1 + \sigma_1^2/\sigma_0^2)} N_k(\boldsymbol{\alpha}|\tilde{\boldsymbol{\gamma}}_1, (\sigma_0^2 + \sigma_1^2)\mathbf{W}^{-1}),$$

where $\tilde{\gamma}_1^t = (\mathbf{0}^t, \gamma_1^t)$ and $\mathbf{W}^{-1}$ is

$$\mathbf{W}^{-1} = \frac{n}{k+1}(\mathbf{X}^t\mathbf{X})^{-1}.$$

We note that the conditional intrinsic prior for the parameter of the alternative $\boldsymbol{\alpha}$ is centered at the null parameter $\tilde{\gamma}_1$. Further, the conditional intrinsic prior for σ_1 is a half Cauchy located at zero and with scale parameter σ_0. This implies that the conditional intrinsic prior has no moments, a desirable property for a default prior. The unconditional intrinsic prior for $(\boldsymbol{\alpha}, \sigma_1)$ is given by

$$\pi^I(\boldsymbol{\alpha}, \sigma_1) = \int \pi^I(\boldsymbol{\alpha}, \sigma_1|\gamma_1, \sigma_0)\pi_0^N(\gamma_1, \sigma_0)d\gamma_1 d\sigma_0.$$

Of course, this prior is fully automatic, i.e. does not depend on any tuning parameters nor processes any subjective prior information.

Using the so called pair of intrinsic priors $\pi_0^N(\gamma_1, \sigma_0)$ and $\pi^I(\boldsymbol{\alpha}, \sigma_1)$, the intrinsic posterior probability of model M_0 is given by

$$P(M_0|\mathbf{y}, \mathbf{X}) = \frac{1}{1 + B_{10}}$$

where

$$B_{10} = \frac{2(k+1)^{k_0/2}}{\pi} \int_0^{\pi/2} \frac{\sin^{k_0}\varphi \, (n + (k+1)\sin^2\varphi)^{(n-k)/2}}{(n\mathcal{B}_n + (k+1)\sin^2\varphi)^{(n-k_1)/2}} d\varphi. \tag{4}$$

From this expression, and also from the frequentist analysis of the testing problem in subsection 5.1, it follows that for fixed values of the sample size n, the number of covariates k and the dimension of the null hypothesis k_1, the statistic $\mathcal{B}_n$ is a sufficient statistic for the testing problem, as the Bayes factor for the intrinsic priors does not depend on other ancillary statistics such as happens with other Bayes factors for linear models found in the literature, which depend on the quotient of the determinants $|\mathbf{X}'\mathbf{X}|$ and $|\mathbf{X}_1'\mathbf{X}_1|$.

Bayesian testing procedures different from the above one have been given by Berger and Pericchi (1996), O'Hagan (1995) and Zellner (1986) who proposed the use of the arithmetic intrinsic Bayes factor, the fractional Bayes factor and the Bayes factor derived from the g-priors, respectively. Except for the arithmetic intrinsic Bayes factor, the other two proposals depend on some tuning parameters which have to be adjusted.

Let us mention that for normal linear models the O'Hagan fractional Bayes factor provides sensible *fractional priors* for testing problems in a similar asymptotic way as the arithmetic intrinsic Bayes factor provides *intrinsic priors* (Moreno 1997). For so doing, the tuning parameter in the fractional Bayes factor is fixed as the quotient m/n, where m is the minimal training sample size. The results obtained when using Bayes factors for fractional priors are very close to those provided by Bayes factors for intrinsic priors, and hence only intrinsic priors are being considered here.

5.3 Comparing the frequentist and Bayesian tests

The dependence of the p-value and the posterior probability of the null on the sufficient statistic $(\mathcal{B}_n, n, k, k_1)$, where n, k, and k_1 are ancillary makes possible the comparison of the frequentist and objective Bayesian test.

For fixed values of the ancillaries n, k and k_1, the p-value and the intrinsic posterior probability of the null model $P(M_0|\mathcal{B}_n, n, k, k_1)$ are monotone increasing functions of the $\mathcal{B}_n$ statistic. This permits us to establish a one-to-one relation between both measures of evidence through the parametric equations

$$\begin{aligned} y &= P(M_0|b, n, k, k_1) \\ p &= I_b\left(\frac{n-k}{2}, \frac{k-k_1}{2}\right), \end{aligned} \tag{5}$$

where the parameter b, the sufficient statistic, ranges in the interval $[0, 1]$.

The separate behaviour of y and p as the sufficient statistic b goes to zero or one is as follows. The null posterior probability and the p-value go to zero as b tends to zero, whatever the values of the ancillaries, as

$$\lim_{b\to 0} P(M_0|b,n,k,k_1) = 0 \quad \text{and} \quad \lim_{b\to 0} I_b\Big(\frac{n-k}{2}, \frac{k-k_1}{2}\Big) = 0.$$

If $\mathcal{B}_n = 0$, then the residual sum of squares of the full model SS is also 0; this means that there is no uncertainty in the full model, i.e. it is deterministic; thus, the reduced model M_0 has zero posterior probability and the p-value is also zero, so that the full model is obviously accepted.

When $\mathcal{B}_n$ tends to one, then the p-value tends to one whatever the values of the ancillaries, but the null posterior probability tends to a number strictly smaller that one (Theorem 2.2 in Girón *et al* 2004) which, on the other hand, tends to one as n goes to infinity.

In this case, as $\mathcal{B}_n = 1$, the residual sum of squares of the full SS and the reduced null model SS_1 are the same so that the data favour either model equally; however, the frequentist evidence in favour of the reduced model M_0 is one as if there were no uncertainty about what model to choose but, on the other hand, the Bayesian test accounts for the uncertainty inherent in the data rendering a posterior probability of M_0 greater than 1/2 but strictly less than 1; hence, the Bayesian test chooses the simpler model, which is a consequence of the built-in Occam's razor implicit in the objective Bayesian test.

From the above equations (5) we can eliminate the parameter b to obtain a explicit equation of the null posterior probability as a function of the p-value, n, k and k_1,

$$y = P(M_0|I_p^{-1}(n-k/2, k-k_1/2), n, k, k_1).$$

From this equation it follows that for fixed values of n, k and k_1, the null posterior probability is an increasing function of the p-value. Therefore, the difference between the frequentist and Bayesian measures of evidence is a simple calibration problem. For this reason this curve is given the name of *calibration curve* in Girón *et al* (2004).

Unfortunately, the null posterior probability also depends on n, k, k_1 so that the properties of the calibration curve have to be established in a case-by-case basis. When simultaneous hypothesis testing are considered we have to jump among different calibration curves hence losing the monotonicity between the null posterior probabilities and the p-values. In this latter case calibration is no longer possible.

In the remainder of this section the number of possible regressors k will be keep fixed. In Figures 1 and 2 we display the typical behavior of the calibration curves, first, for different values of the sample size n and fixed k_1, and second, for different values of k_1 and fixed n.

The calibration curves in Figure 1 correspond to $k = 10$ and $k_1 = 9$ and $n = 20$, 50 and 90; they indicate that for a given p-value the null posterior probability increases as the sample size increases. Further, from the consistency of the Bayes factor for intrinsic priors it follows that the slope of the calibration curve at the origin tends to infinity as n increases. This shows that the evidence against the null conveyed by the p-values should be diminished as the sample size n increases in order to reconcile the frequentist and Bayesian test. Otherwise, we would reject a null hypothesis that has a very high posterior probability.

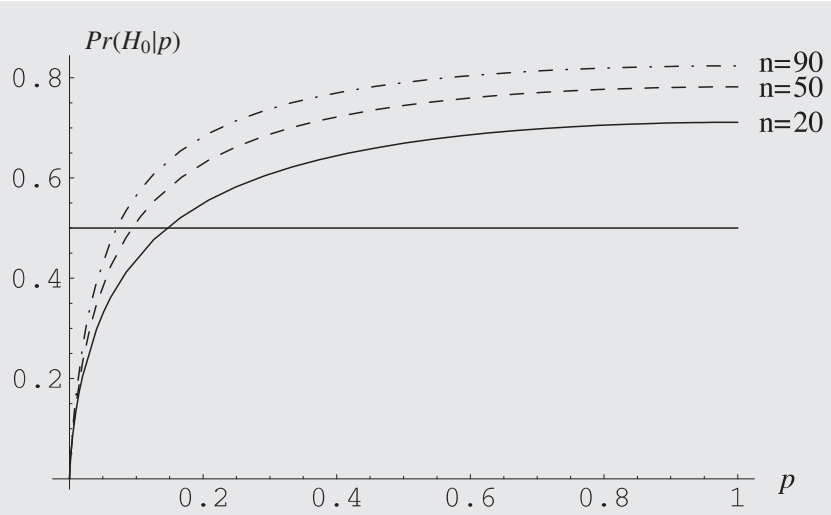

Figure 1: *Three calibration curves for different sample sizes* $n = 20$, $n = 50$ *and* $n = 90$, *when the number of regressors is* $k = 10$, *and* $k_1 = 9$.

The curves in Figure 2, for a fixed value of the sample size n, indicate that small p-values correspond to small posterior probabilities when k_1 is large, and also that the posterior probability increases as k_1 decreases. This implies that for small values of k_1 a p-value would reject a null hypothesis that has a large posterior probability, a fact which is generally acknowledged in the literature. But, on the other hand, for large values of k_1 a p-value would accept a null hypothesis that has a small posterior probability. Notice, in Figure 2, that for the curve with $k_1 = 9$ there is an interval of p-values larger than 0.05 whose corresponding posterior probabilities of the null are smaller than 1/2. This important fact, which is generally overlooked by the feeling that p-values tend to reject the null hypothesis more often than the Bayesian tests, also reveals that the opposite may happen sometimes; namely, that a null hypothesis may be not rejected by the frequentist p-value and rejected by the Bayesian test for the same data.

Figure 2 also indicates that, when comparing several nulls of different dimension k_1 that convey the same frequentist evidence, the Bayesian test chooses the simplest

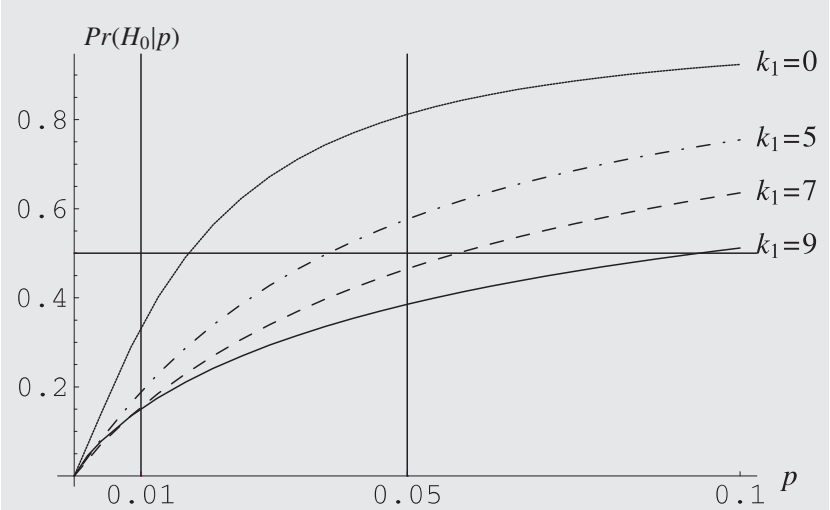

Figure 2: *Four calibration curves, plotted for values of p-values in the interval* $(0, 0.10)$*, for different choices of* $k_1 = 0, 5, 7, 9$*, when the sample size is* $n = 50$*, and the number of regressors is* $k = 10$*.*

hypothesis or model. As before, this is another instance of the automatic Occam's razor property implicit in the objective Bayesian test.

5.4 Variable selection in regression

Although the purpose of this paper is to critically revise the similarities and, above all, the enormous differences between the two approaches to hypothesis testing, we want to devote this last section to the important problem of variable selection in normal linear regression, where we have to consider simultaneously a large number of multiple tests of different dimensions k_1, and then provide an ordering of the plausibility of the different models considered. A recent contribution to the subject, mainly from a frequentist perspective, is the revised version of the classical monograph by Miller (2002), where a chapter on Bayesian variable selection has been added.

We do not discuss here sequential, or non-exhaustive, frequentist variable selection procedures such as forward, backward and stepwise selection methods nor the more recent ones such as the lasso or shrinkage methods, in order to concentrate our discussion on exhaustive search methods from the frequentist and Bayesian perspectives. Neither do we discuss here the well known asymptotic *BIC* criterion for model selection, because we can dispense with it as we have the non-asymptotically based *from above* and *from below* Bayesian criteria.

The *all subsets selection criterion* is based on the idea of classifying the set of all models, say $\mathcal{M}$, in k disjoint classes, where k is the number of covariates of the full model, according to the number of covariates j, i.e. $\mathcal{M} = \bigcup_{j=1}^{k} \mathcal{M}_j$. Within each class,

the model with minimum residual sum of squares SS_j is chosen, so that we end up with *k maximal* models, one for each class. Note that *p*-values or the equivalent F statistic provide the same ordering within the class $\mathcal{M}_j$ and, consequently, the same sets of maximals, but these criteria turn to be useless when comparing models with different number of covariates, a fact which is well recognized by frequentist statisticians.

Hence, the problem of choosing the *best* model within the maximals, is far from trivial and a large number of procedures have been proposed in the literature. The underlying idea is to correct the SS_j, or a simple function of it and the ancillaries to account for the different number of covariates, as do the well known adjusted R^2, Mallows C_p, and the AIC criteria. Unfortunately, these corrections do not always work properly.

From the objective Bayesian viewpoint, which is mainly based on the results of Section 5.2, two procedures for model selection have been proposed. The main difference between these procedures relies on the form of encompassing – that is, of nesting – the class of all possible submodels. The so called *from above* procedure (Casella and Moreno 2005, Girón *et al.* 2004) is based on comparing all the submodels with the full model, and ordering them according to the posterior probabilities of all submodels using the formulae of Section 5.2. We denote these posterior probabilities by $P_{fa}(M_i|\mathcal{S}_n)$ for any submodel M_i, where $\mathcal{S}_n = (\mathcal{B}_n, n, k, k_1)$. The interpretation of these probabilities is that the model with highest posterior probability represents the most plausible reduction in complexity from the full model, the second highest the second most plausible model, and so on.

As these posterior probabilities are monote increasing functions of the $\mathcal{B}_n$ statistic for fixed ancillaries, this means that within each class of models $\mathcal{M}_j$ the ordering provided is the same as the one based on the residual sum of squares SS_j. Thus, the Bayesian solution *from above* is the maximal model having the largest posterior probability of its corresponding model. No need for extra adjustment!

The so called *from below* procedure (Girón *et al.* 2005b and Moreno and Girón 2005), based on the simple fact that the intercept only model is nested into any other possible model as far as it includes the intercept, produces a possibly different ordering of all the submodels of the full model. The ordering is now based on the Bayes factors resulting from comparing the current submodel with the intercept only model. Further, it turns out that this procedure provides a coherent set of model posterior probabilities on the set of all possible submodels denoted by $P_{fb}(M_i|\mathcal{S}_n)$, and these coherent probabilities are monote increasing functions of the R^2 statistic as now $\mathcal{B}_n = 1 - R^2$, which in turn, is also a monote decreasing function of the residual sum of squares SS_j of the corresponding submodel. This means, as with the *from above* Bayesian criterion, that the model chosen by the from below criterion is also the maximal model having the largest Bayes factor or, equivalently, the highest model posterior probability $P_{fb}(M_i|\mathcal{S}_n)$.

The main conclusion derived from these comparisons is, first, that the two Bayesian criteria always choose a maximal model, i.e. they are compatible with the best subsets

partial ordering and, second, that they are fully automatic in the sense that no tuning of extra parameters, neither the use of outside information nor additional criteria, is needed.

Table 3: *Comparison of different variable selection criteria for Hald's data*

Models	**From below** $P_{fb}(M_i\|\mathcal{S}_n)$	**From above** $P_{fa}(M_i\|\mathcal{S}_n)$	R^2	**Adjusted** R^2	**Mallows** C_p
$\{x_1, x_2\}$	0.5466	0.7407	0.9787	0.9744	2.6782
$\{x_1, x_4\}$	0.1766	0.5364	0.9725	0.9670	5.4958
$\{x_1, x_2, x_4\}$	0.0889	0.7231	0.9823	0.9764	3.0182
$\{x_1, x_2, x_3\}$	0.0879	0.7211	0.9823	0.9764	3.0413
$\{x_1, x_3, x_4\}$	0.0708	0.6809	0.9813	0.9750	3.4968
$\{x_2, x_3, x_4\}$	0.0165	0.3780	0.9728	0.9638	7.3375

Table 3 compares the results of several model selection criteria for the famous Hald's data on the composition of cement. The analysis illustrates the Occam's razor property of the Bayesian criteria. Note also that, for these data, the adjusted R^2 does not adjust the ordering provided by the original R^2 for the most plausible models.

A large simulation study, see Moreno and Girón (2005b) for the description and extent of the study, has shown that the adjusted R^2 performs very poorly in almost all situations, a well known fact. Mallow's C_p and the AIC criteria perform in a very similar way – another well known fact – but they show a poorer behaviour when compared with either Bayesian criteria in most circumstantes. This suggests that the Bayesian criteria account for the difference in dimensionality in some automatic way, hidden in the formulae of their corresponding Bayes factors, in the same manner they automatically obey Occam's razor principle.

Table 4 illustrates these comments for a medium size linear model, $k = 6$, i.e. with five covariates excluding the intercept, sample size $n = 40$ and values of k_1 ranging from 2 to 6.

Table 4: *Comparison of different variable selection criteria for the simulated data.*

	N.º of covariates				
Criterion	**1**	**2**	**3**	**4**	**5**
From below	0.901	0.910	0.962	0.977	1.000
From above	0.573	0.657	0.793	0.927	1.000
Mallows C_p	0.500	0.563	0.692	0.850	1.000
Adjusted R^2	0.234	0.292	0.452	0.716	1.000

The model considered for simulation is

$$\mathbf{y} = \mathbf{X}\alpha + \varepsilon$$

where $\mathbf{y}$ is a vector of length 40, $\mathbf{X}$ is a 40×6 matrix whose entries were obtained by simulation from a standard normal distribution $N(0, 1)$, except the entries in the first

column which were set equal to 1 to include the intercept, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_6)^t$ is a vector of length 6. The error terms in $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. $\varepsilon_i \sim N(0, 1)$.

After fixing $\mathbf{X}$, samples of size 5000 were simulated from the model for five different settings of the vector of regression coefficients $\boldsymbol{\alpha}$ including 1, 2 3, 4 and 5 non zero coefficients. In particular, we set

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= (-1, -2, 0, 0, 0, 0) \\
\boldsymbol{\alpha}_2 &= (-1, -2, 2, 0, 0, 0) \\
\boldsymbol{\alpha}_3 &= (-1, -2, 2, -3, 0, 0) \\
\boldsymbol{\alpha}_4 &= (-1, -2, 3/2, -2, 2, 0) \\
\boldsymbol{\alpha}_5 &= (-1, -2, 3/2, -2, 2, -1).
\end{aligned}$$

The entries in Table 4 represent the proportion of times that the true model is selected in the first place in the 5000 simulations according to the four criteria and to the number of nonzero regression coefficients in the model.

The relation between the p-values, or any equivalent model selection procedure, and the Bayesian model posterior probabilities in the variable selection problem can be summarized as follows. For a fixed sample size n, models with the same number of regressors k_1 are ordered in the same manner by all criteria: p-values, R^2 and adjusted R^2, Mallows C_p, AIC and the two Bayesian procedures. However, when comparing models with different number of regressors all frequentist and Bayesian criteria generally provide different orderings of the models. But, as we have learned from the simulations, the frequentist behaviour of the Bayesian criteria generally outperforms that of the frequentist ones. Comparisons between the *from below* and *from below* Bayesian criteria are discussed in Moreno and Girón (2005b). The conclusion in that paper is that for models with a small or medium number of relevant covariates, as the one illustrated in Table 3, the *from below* criterion performs better that the *from above* one, but for models with a large number of influential covariates, the opposite may happen.

6 Discussion

Two measures of evidence for hypothesis testing, frequentist and Bayesian, have been considered and compared in this paper for some important testing problems. The case of the standard normal model is not dealt with in the paper as it is a particular case of the normal linear model with no covariates; in this case the sample size is the only ancillary, and the frequentist-Bayesian comparison or calibration just depends on the sample size.

We have first recalled that standard normal-theory does not apply to the Behrens-Fisher problem of testing the equality of the means of two normal populations under heteroscedasticity, as there is no clear-cut p-value and, consequently, frequentist theory has to resort to computing an approximate p-value by adjusting the degrees of freedom of a t-distribution on which the solution of the homoscedastic case is based.

For this problem, we have illustrated the fact that p-values reject the null hypothesis for data for which the Bayesian inference accepts, more markedly as both sample sizes increase. To make p-values more unsatisfactory we have also seen that for small sample sizes p-values would accept the null for data for which the Bayesian rejects.

Therefore, we have to admit that the usual interpretation of a p-value as a measure of evidence against the null regardless the sample size, though it entails a notable simplification, may produce wrong answers. It is clear that the enormous success of the p-values in the realm of applications is partially due to their simplicity for scientific communication, but the bad news is that such a simplicity may be misleading.

The study of the relation between the two measures of evidence has been extended to normal data in the presence of covariates, that is to the normal linear model. Here two new ancillaries, in addition to the sample size, arise: the number of covariates and the dimension of the null. We have illustrated that the dimension of the null hypothesis is another fundamental ancillary to be taken into account when interpreting p-values. The disregard of this ancillary may produce the rejection of null hypotheses that have high posterior probabilities or the acceptation of nulls that have low posterior probabilities.

Finally, we have considered the variable selection problem, a challenging multiple testing problem, because for this problem the sample size and the dimension of the null play a very important role. For selecting variables we necessarily have to jump among models whose null parameter spaces have different dimensions. To overcome this difficulty, the frequentist approach has to adjust some statistic, usually the one based on the ratio of sums of the residuals under the full and null models, in several ways to account for the different number of covariates involved. This accommodation, however, is not very convincing. The objective Bayesian approach seems to deal with this problem in a more appropiate way than the frequentist counterpart because all the ancillaries in the problem are properly taken into account. Furthermore, the objective Bayesian solution works in a fully automatic way in the sense that there is no need for adjusting any tuning parameters.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J.O. (1994). An overview of robust Bayesian analysis (with discussion). *Test*, 3, 5-124.

Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F:M. Smith, eds. Oxford University Press: Oxford, pp. 35-60.

Berger, J.O. and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94, 542-554.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91, 109-122.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence (with discussion). *Journal of the American Statistical Association*, 82, 112-139.

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

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

Cano, J.A., Kessler, M. and Moreno, E. (2004). On intrinsic priors for nonnested models. Test, 13, 445-463.

Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82, 106-111.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Belmont: Wadsworth.

Casella, G. and Moreno, E. (2005). Intrinsic meta analysis of contingency tables, *Statistic in Medicine*, 24, 583-604.

Casella, G. and Moreno, E. (2005). Objective Bayesian variable selection. *Journal of the American Statistical Association* (to appear).

Fisher, R.A. (1936). The fiducial arguments in statistical inference. *Annals of Eugenics*, 6, 391-422.

Girón, F.J., Martínez, L. and Imlahi, L. (1999). A characterization of the Behrens-Fisher distribution with applications to Bayesian inference. *Comptes rendus de l'Académie des sciences de Paris, Ser I,* 701-706.

Girón, F., Martínez, L., Moreno, E. and Torres, F. (2003). Bayesian analysis of matched pairs in the presence of covariates. In *Bayesian Statistics 7*. J.M. Bernardo *et al.* (eds.), 553-563, Oxford: Oxford University Press.

Girón, F.J., Martínez, L., and Moreno, E. (2003). Bayesian analysis of matched pairs. *Journal of Statistical Planning and Inference*, 113, 49-66.

Girón, F.J., Martínez, M.L., Moreno, E. and Torres, F. (2004). Objective testing procedures in linear models. Calibration of the *p*-values. Submitted.

Girón, F.J., Moreno, E. and Martínez, L. (2005). An objective Bayesian procedure for variable selection in regression. In *Advances on distribution theory, order statistics and inference*. Eds. N. Balakrishnan *et al.*, Birkhauser Boston, (to appear).

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

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

Lehman, E. (1986). *Testing Statistical Hypotheses*, (second edition). New York: Wiley.

Lindley, D.V. (1970). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* (Vol. 2). Cambridge: Cambridge University Press.

Miller, A.J. (2002). *Subset Selection in Regression. 2nd edition*. London: Chapman and Hall.

Moreno, E. (1997). Bayes Factor for Intrinsic and Fractional Priors in Nested Models: Bayesian Robustness. *IMS Lectures Notes-Monograph Series*, 31, 257-270.

Moreno, E. (2005). Objective Bayesian analysis for one-sided testing, *Test*, 14, 181-198.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93, 1451-1460.

Moreno, E., Bertolino, F. and Racugno, W. (1999). Default Bayesian analysis of the Behrens-Fisher problem. *Journal of Statistical Planning and Inference*, 81, 323-333.

Moreno, E., Bertolino, F. and Racugno, W. (2000). Bayesian model selection approach to analysis of variance under heterocedasticity, *Journal of the Royal Statistical Society, Series D (The Statistician)*, 49, 1-15.

Moreno, E., Bertolino, F. and Racugno, W. (2003). Bayesian inference under partial prior information, *Scandinavian Journal of Statistics*, 30, 565-580.

Moreno, E. and Cano J.A. (1989). Testing a point null hypothesis: asymptotic robust Bayesian analysis with respect to the priors given on a subsigma field. *International Statistical Review*, 57, 221-232.

Moreno, E. and Girón, F.J. (2005*a*). Consistency of Bayes factors for linear models, *Comptes rendus de l'Académie des sciences de Paris, Ser I*, 911-914.

Moreno, E. and Girón, F.J. (2005*b*). Comparison of Bayesian objective procedures for variable selection in regression. Submitted.

Moreno, E., Girón, F.J., and Torres, F. (2003). Intrinsic priors for hypothesis testing in normal regression models. *Revista de la Real Academia de Ciencias Serie A,. Mat.*, 97, 53-61.

Moreno, E. and Liseo, B. (2003). A default Bayesian test for the number of components of a mixture. *Journal of Statistical Planning and Inference*, 111, 129-142.

Moreno, E., Torres, F., and Casella, G. (2005). Testing the equality of regression coefficients in heteroscedastic normal regression models. *Journal of Statistical Planning and Inference*, 131, 117-134.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *Journal of the Royal Statistical Society Series B*, 57, 99-138.

O'Hagan, A. and Forster, J. (2004). *Bayesian Inference.* Kendall's Advances Theory of Statistics (Vol. 2B). London: Arnold.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press: Oxford.

San Martini, A. and Spezaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Sociery, Series B*, 46, 296-303.

Sellke, T., Bayarri, M.J. and Berger, J. (2001). Calibration of *p*-values for testing precise null hypotheses. *The American Statistician*, 55, 62-71.

Wald, A. (1955). Testing the difference between the means of two normal populations with unknown standard deviations. In *Selected papers in Statistics and Probability*, T.W. Anderson *et al*. (eds.), 669-695, Stanford University Press.

Wilks, S.S. (1962). *Mathematical Statistics*. New York: Wiley

Welch, B.L. (1951). On the comparison of several means values: an alternative approach. *Biometrika*, 38, 330-336.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, P.K. Goel and A. Zellner (eds.), 233-243, Amsterdam: Elsevier.