ACCEPTED

Comput Stat (XXXX)

ORIGINAL PAPER

Bayesian model selection approach to the one way analysis of variance under homoscedasticity

J. A. Cano · C. Carazo · D. Salmerón

Received: 15 June 2011 / Accepted: 6 June 2012
© Springer-Verlag 2012

Abstract An objective Bayesian model selection procedure is proposed for the one way analysis of variance under homoscedasticity. Bayes factors for the usual default prior distributions are not well defined and thus Bayes factors for intrinsic priors are used instead. The intrinsic priors depend on a training sample which is typically a unique random vector. However, for the homoscedastic ANOVA it is not the case. Nevertheless, we are able to illustrate that the Bayes factors for the intrinsic priors are not sensitive to the minimal training sample chosen; furthermore, we propose an alternative pooled prior that yields similar Bayes factors. To compute these Bayes factors Bayesian computing methods are required when the sample sizes of the involved populations are large. Finally, a one to one relationship—which we call the calibration curve—between the posterior probability of the null hypothesis and the classical p value is found, thus allowing comparisons between these two measures of evidence. The behavior of the calibration curve as a function of the sample size is studied and conclusions relating both procedures are stated.

J. A. Cano (✉)
Departamento de Estadística e Investigación Operativa,
Universidad de Murcia, Murcia, Spain
e-mail: jacano@um.es

C. Carazo
Escuela de Arquitectura e Ingeniería de la Edificación,
Universidad Católica San Antonio, Murcia, Spain

D. Salmerón
Department of Epidemiology, Regional Health Authority, Murcia, Spain

D. Salmerón
CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

Keywords Calibration curve · Bayes factor · Intrinsic priors · Model selection · Robustness · p values

1 Introduction

Let us consider k normal populations $N(x_1|\mu_1, \sigma^2), \ldots, N(x_k|\mu_k, \sigma^2)$ with means $\mu_1, \ldots, \mu_k$ and common variance σ^2 unknown. From population i we have a sample of size n_i, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})$, with sample mean and variance denoted as $\overline{x}_i$ and $s_i^2/n_i, i = 1, \ldots, k$, respectively, and we want to test whether the means μ_i are all equal to a constant μ. The frequentist solution can be found in Rohatgi (1984) and Casella and Berger (1990).

Lindley (1970) and Box and Tiao (1973) deal with the classical Bayesian viewpoint solution to the homoscedastic one way analysis of variance that is based on the posterior distribution of the parameter $(\lambda_1, \ldots, \lambda_k)$, where $\lambda_i = \mu_i - \mu, i = 1, \ldots, k$. For a specified significance level α the highest posterior density region with probability equal to $1 - \alpha$ is computed and any point in this region is accepted as plausible; then, the rule is applied to the null hypothesis $H_0 : \lambda_1 = \cdots = \lambda_k = 0$. Nevertheless this procedure can not be considered a Bayesian solution since the null hypothesis has a posterior probability equal to zero and as we can see from the description of the procedure is not taken into account to solve the problem. This is what happens when a testing problem is focused as an estimation problem. In Bayesian analysis we should distinguish between problems of estimation and problems of testing as recommended in Jeffreys (1961, pp. 245–249).

The objective here is to formulate the homoscedastic one way analysis of variance as a model selection problem using the Bayes factor as the main tool to solve it. The one way analysis of variance under heteroscedasticity was studied in Bertolino et al. (2000), where a test to determine if homoscedasticity is present was included, therefore it becomes necessary to study the case when homoscedasticity is present, developing specific solutions for this case since it is not a particular case but a special one that deserves special attention. Since the number of parameters in the homoscedastic case is lower we have to deal with a simpler model and we should take advantage of this developing intrinsic priors adapted to this situation. However, the minimal training sample for the homoscedastic case is of dimension $(k + 1)$ and one observation is needed from each population except for one of the populations for which two observations are needed to complete the minimal training sample; therefore, in the homoscedastic case the minimal training sample is not unique and a sensitivity analysis is then needed. We note that for the heteroscedastic case the minimal training sample is of dimension $2k$ and two observations are needed from each population and therefore it is unique and yields a unique intrinsic prior. Consequently, contrary to what happens in the frequentist analysis of this problem, from a computational point of view the homoscedastic case is more complicated than the heteroscedastic one and the following duality appears, while the frequentist solution for the heteroscedastic case uses a pooled variance of the populations variances our Bayesian analysis can use a pooled intrinsic prior that is presented later derived from the non-uniqueness of the minimal training sample above mentioned.

We consider the nested sampling models

$$M_1 : f_1(\mathbf{z}|\theta_1) = \prod_{i=1}^{k} N_{n_i}(\mathbf{x}_i|\mu 1_{n_i}, \tau^2 I_{n_i}), \tag{1}$$

$$M_2 : f_2(\mathbf{z}|\theta_2) = \prod_{i=1}^{k} N_{n_i}(\mathbf{x}_i|\mu_i 1_{n_i}, \sigma^2 I_{n_i}) \tag{2}$$

where $\mathbf{z} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, $\theta_1 = (\mu, \tau)$ and $\theta_2 = (\mu_1, \ldots, \mu_k, \sigma)$. The conventional priors for μ and $\log \tau$ are uniform assuming that μ and τ are *a priori* independent, as we can see in Jeffreys (1961, p. 138), that is

$$\pi_1^N(\theta_1) = c_1/\tau. \tag{3}$$

Analogously, assuming that μ_i and σ are also *a priori* independent, the conventional priors for μ_i and $\log \sigma$ are uniform, that is

$$\pi_2^N(\theta_2) = c_2/\sigma, \tag{4}$$

therefore the prior distributions are improper.

In (3) and (4) c_1 and c_2 are positive constants that are undefined since $\pi_1^N(\theta_\mathbf{1})$ and $\pi_2^N(\theta_\mathbf{2})$ are improper functions hence the Bayes factor $B_{21}^N(\mathbf{z})$ is defined up to a multiplicative constant c_2/c_1. To avoid this difficulty we compute the Bayes factor of M_2 against M_1 for intrinsic priors.

The structure of this paper is the following. In Sect. 2 intrinsic priors for the homoscedastic one way analysis of variance are derived, their corresponding Bayes factors are computed and it is found that their expression depends on the minimal training sample chosen to obtain the intrinsic priors. Furthermore, when the sample sizes of the populations considered are large, these Bayes factors are computationally unfeasible and therefore an estimate of it based on Monte Carlo methods is required. In Sect. 3 we study empirical robustness of these Bayes factors as a function of the minimal training sample. Two examples show that they are not sensitive to the minimal training sample chosen and the use of an alternative pooled prior is proposed to derive an alternative Bayes factor. There is an extensive literature devoted to compare frequentist and Bayesian methodologies, Berger and Sellke (1987), Casella and Berger (1987), Girón et al. (2006) and our paper tries to add something in this context, for it, in Sect. 4 we prove that there exists a one-to-one relationship between the posterior probability of the null hypothesis and the p value that permits to define an increasing curve, which we call the calibration curve. Calibration is a simple means of establishing a map between both measures of evidence against the null to compare the frequentist and the default Bayesian approaches. Finally, in Sect. 5 we briefly summarize the results obtained in previous sections and we also give some concluding remarks.

2 Intrinsic priors and their corresponding Bayes factors

2.1 Intrinsic priors

The sampling model M_1 is nested in M_2 and for the prior given in (4) the minimal training sample is a $(k+1)-$dimensional random vector that consists of two observations from one of the populations and a single observation from the remaining $(k-1)$ populations, that is $\mathbf{x}(l) = (x_1, \ldots, x_{j1}, x_{j2}, \ldots, x_k)$, with $1 \leq j \leq k$.

Theorem 2.1 *The intrinsic priors for comparing M_1 and M_2, are $\{\pi_1^N(\theta_1), \pi_2^I(\theta_2)\}$, where $\pi_2^I(\theta_2) = \int \pi_2^I(\theta_2|\theta_1)\pi_1^N(\theta_1)d\theta_1$ and*

$$\pi_2^I(\theta_2|\theta_1) = N(\mu_j|\mu, (\tau^2+\sigma^2)/2)HC^+(\sigma|0,\tau)\prod_{i=1, i\neq j}^{k} N(\mu_i|\mu, \tau^2+\sigma^2), \quad (5)$$

where $HC^+(\sigma|0,\tau)$ denotes the half Cauchy density.

Proof It is obtained adapting the proof of Theorem 1 in Bertolino et al. (2000). Note that, as occurs in the heteroscedastic case, see Bertolino et al. (2000), Theorem 1, in the homoscedastic one way ANOVA the $\mu_i's$ conditionally on (μ, τ, σ) are independent and normally distributed and σ conditionally on τ is independent and half-Cauchy distributed. Furthermore, since μ_i means the treatment effect we have given the same mean, μ, to every treatment which is sensible; however, under homoscedasticity, the conditional distribution of θ_2 given θ_1 depends on the minimal training sample chosen and this implies a labelled problem to compute the Bayes factor. Note that although the imaginary minimal training sample is integrated out to obtain the conditional intrinsic prior, in this prior the mean of the population from where two observations were taken to obtain the minimal training sample has as variance half of the variances of the other means. Therefore, we propose the following pooled conditional distribution of θ_2 given θ_1 to eliminate this dependency

$$\pi_2^I(\theta_2|\theta_1) = \prod_{i=1}^{k} N(\mu_i|\mu, (1-1/(2k))(\tau^2+\sigma^2))HC^+(\sigma|0,\tau), \quad (6)$$

where the common variance $(1-1/(2k))(\tau^2+\sigma^2)$ is the average of the $\mu_i's$ variances for $i = 1, \ldots, k$, since

$$\frac{(k-1)(\tau^2+\sigma^2) + (\frac{\tau^2+\sigma^2}{2})}{k} = (1-1/(2k))(\tau^2+\sigma^2).$$

2.2 Bayes factors for intrinsic priors

The Bayes factor for intrinsic priors $\{\pi_1^N(\theta_1), \pi_2^I(\theta_2)\}$ for the sample $\mathbf{z}$ turns out to be

(i) when the marginal $m_2^I(\mathbf{z})$ comes from the conditional distribution (5),

$$B_{21}^I(\mathbf{z}) = \frac{2^{\frac{3}{2}} N^{\frac{1}{2}} \Gamma\left(\frac{N}{2}\right) \left(\sum_{i=1}^k \left(s_i^2 + n_i(\overline{x}_i - \overline{x})^2\right)\right)^{\frac{N-1}{2}}}{\pi^{\frac{3}{2}} \Gamma\left(\frac{N-1}{2}\right)} I_2, \tag{7}$$

where $I_2 = \int_{-\infty}^{\infty} I_1 d\mu$, with

$$I_1 = \int_0^{\pi/2} \frac{(h(n_j,\theta))^{-1/2} \prod_{i\neq j, i=1}^k (g(n_i,\theta))^{-1/2}}{(sen\theta)^{N-k} \left[\frac{S^2}{sen^2\theta} + \frac{2n_j(\overline{x}_j-\mu)^2}{h(n_j,\theta)} + \sum_{i\neq j, i=1}^k \frac{n_i(\overline{x}_i-\mu)^2}{g(n_i,\theta)}\right]^{\frac{N}{2}}} d\theta$$

and

$$h(n,\theta) = n + 2sen^2\theta,$$

$$g(n,\theta) = n + sen^2\theta,$$

(ii) when the marginal $m_2^I(\mathbf{z})$ comes from the conditional distribution (6),

$$B_{21}^I(\mathbf{z}) = \frac{2N^{\frac{1}{2}} \Gamma\left(\frac{N}{2}\right) \left(\sum_{i=1}^k \left(s_i^2 + n_i(\overline{x}_i - \overline{x})^2\right)\right)^{\frac{N-1}{2}}}{\pi^{\frac{3}{2}} \Gamma\left(\frac{N-1}{2}\right)} I_2, \tag{8}$$

where $I_2 = \int_{-\infty}^{\infty} I_1 d\mu$, with

$$I_1 = \int_0^{\pi/2} \frac{\prod_{i=1}^k (g(n_i,\theta,k))^{-1/2}}{(sen\theta)^{N-k} \left[\frac{S^2}{sen^2\theta} + \sum_{i=1}^k \frac{n_i(\overline{x}_i-\mu)^2}{g(n_i,\theta,k)}\right]^{\frac{N}{2}}} d\theta$$

and

$$g(n,\theta,k) = (1 - 1/(2k))\, n + sen^2\theta.$$

Note that in both expressions $S^2 = \sum_{i=1}^k s_i^2$ and $N = \sum_{i=1}^k n_i$.

Expressions (7) and (8) are obtained by direct integration on the μ_i and (τ, σ). These Bayes factors are the limit of sequences of *actual* Bayes factors, see Moreno et al. (1998), and consequently they share properties of *actual* Bayes factors as dependence on the sample through a sufficient statistics. Furthermore, note that in (7) and (8) $\sum_{i=1}^k \left(s_i^2 + n_i(\overline{x}_i - \overline{x})^2\right)$ is the classical *total sum of squares* that is decomposed into S^2 that is the *error sum of squares, ESS,* and $\sum_{i=1}^k n_i(\overline{x}_i - \overline{x})^2$, the *treatment sum of squares, TSS*, the quantities involved in the computation of the p value. Consequently,

under both methodologies (frequentist and Bayesian) the study of the homoscedastic one way analysis of variance depends on the decomposition of the total variation of the data although the Bayesian viewpoint involves more complex computations than the classical ANOVA table like integrals I_2. These integrals need numerical integration; although, this is not a serious inconvenience since it can be carried out with standard software, e. g. Mathematica. However, when the sample sizes of the populations under consideration are large, this approximation is unfeasible and an estimate of $m_2^I(\mathbf{z})$ based on Monte Carlo methods is required, see Robert and Casella (2001).

In these cases, for the imaginary minimal training sample

$$\mathbf{x}^*(l) = (x_1^*, x_2^*, \ldots, x_{j1}^*, x_{j2}^*, \ldots, x_k^*),$$

with $1 \leq j \leq k$, we obtain $m_2^I(\mathbf{z})$ as

$$m_2^I(\mathbf{z}) = \int \frac{p(\mathbf{z}|\mathbf{x}^*(l))\pi_1^N(\theta_1)}{g(\theta_1)} f_1(\mathbf{x}^*(l)|\theta_1) g(\theta_1) d\theta_1 d\mathbf{x}^*(l),$$

and the estimate is given by

$$\frac{1}{H}\sum_{i=1}^{H} \frac{p(\mathbf{z}|\mathbf{x}^{*i}(l))\pi_1^N(\theta_1^i)}{g(\theta_1^i)},$$

where θ_1^i, $i = 1, \ldots, H$, is a sample of size H of $g(\theta_1)$ and $x^{*i}(l) \sim f_1(\mathbf{x}^*(l)|\theta_1^i)$ for $i = 1, \ldots, H$.

Note that we have chosen $g(\theta_1)$ as the product of a Normal and a Inverse Gamma, that is $\mu \sim N(\mu_0, \tau_0)$, where μ_0 and τ_0 are arbitrary values and $\tau = v^{-1}$ with $v \sim Gamma(p, a)$, where p and a are arbitrary values too. Finally, it is easy to see that

$$p(\mathbf{z}|\mathbf{x}^{*i}(l)) = \frac{2|x_{j1}^* - x_{j2}^*|(2\pi)^{-\frac{N+1}{2}}}{\sqrt{n_j+2}\prod_{i\neq j, i=1}^{k}\sqrt{n_i+1}} 2^{\frac{N-1}{2}}\Gamma\left(\frac{N+1}{2}\right)\left(\sum_{i=1}^{k}\tilde{s}_i^2\right)^{-\frac{N+1}{2}},$$

where

$$\tilde{s}_j^2 = s_j^2 + n_j\left(\frac{2\overline{x}_j - x_{j1}^* - x_{j2}^*}{n_j+2}\right)^2 + \sum_{h=1}^{2}\left(x_{jh}^* - \frac{n_j\overline{x}_j + x_{j1}^* + x_{j2}^*}{n_j+2}\right)^2$$

and

$$\tilde{s}_i^2 = s_i^2 + n_i\left(\frac{\overline{x}_i - x_i^*}{n_i+1}\right)^2 + \left(x_i^* - \frac{n_i\overline{x}_i + x_i^*}{n_i+1}\right)^2$$

for $i = 1, \ldots, k$ with $i \neq j$.

3 Empirical robustness of the Bayes factor as a function of the minimal training sample

In Sect. 2 we have seen that several choices of priors are possible and therefore the Bayes factor for the homoscedastic one way analysis of variance depends on the minimal training sample considered, that is its expression is different according to the population chosen to take out two observations. This implies an *a priori* labelled problem; however, in this section we illustrate that the Bayes factor is robust with respect to the choice of the minimal training sample.

Example 1 We consider two groups each one consisting of three normal populations and we want to test whether the population means of each group are all equal.

1st group. We simulated from three normal populations $N(0, 1)$ with sizes $(n_1, n_2, n_3) = (10, 25, 50)$, and the following data were obtained:

$$(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (0.06, -0.09, -0.21), (s_1^2, s_2^2, s_3^2) = (5.418, 31.825, 51.249).$$

2nd group. We simulated from three normal populations with parameters

$$\{N(2.5, 1), N(1, 1), N(1, 1)\}$$

and sizes $(n_1, n_2, n_3) = (30, 20, 60)$, and the following data were obtained:

$$(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (2.204, 1.091, 1.086), (s_1^2, s_2^2, s_3^2) = (35.665, 20.660, 74.211).$$

In each row of Table 1 are displayed the p value, the corresponding values of the posterior probability of M_1 according to the chosen population $j = 1, 2, 3$ to take out two observations to obtain the minimal training sample and the value of the posterior probability of M_1 when the conditional distribution of θ_2 given θ_1 is (6).

Numbers in Table 1 show that the posterior probability of M_1 is not sensitive to the minimal training sample chosen. For the first group the value of the posterior probability of M_1 is 0.96 for $j = 1, 2, 3$, whereas this value is 0.002 for the second group. Furthermore, these values are also the same if the prior distribution of θ_2 given θ_1 is (6). Finally, from Table 1 it is immediately clear that in both cases the p value and the posterior probability of M_1 convey the same reasonable message for each group. For the first group they accept the simpler model (the population means are all equal) whereas for the second group they reject the simpler model (the population means are not all equal). Next we generalize this example.

Table 1 P value, values of the posterior probability of M_1 according to the minimal training sample chosen and values of the posterior probability of M_1 when the conditional distribution of θ_2 given θ_1 is (6)

	P value	$P_{j=1}(M_1)$	$P_{j=2}(M_1)$	$P_{j=3}(M_1)$	$P(M_1)$
1st group	0.721	0.962	0.963	0.963	0.963
2nd group	0.00003	0.002	0.002	0.002	0.002

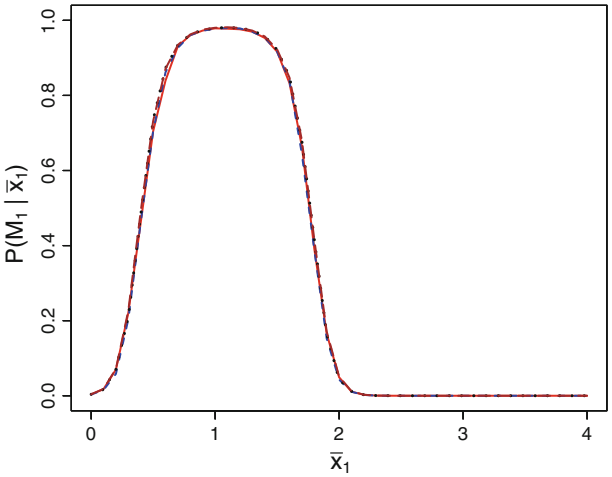

Fig. 1 Four *curves* corresponding to the value of the posterior probability of M_1 according to $\overline{x}_1$ when different prior distributions are considered: the *red/solid*, the *blue/dashed* and the *black/dotted curves* are obtained when the population chosen to take out two observations is the first population, the second one and the third one, respectively. The *brown/long dashed curve* is obtained when the prior is that given in (6)

Example 2 We consider three normal populations with sample sizes $(n_1, n_2, n_3) = (30, 20, 60)$, sample means $\overline{x}_1 \in (0, 4)$, $(\overline{x}_2, \overline{x}_3) = (1.091, 1.086)$ and $(s_1^2, s_2^2, s_3^2) = (35.665, 20.660, 74.211)$ respectively, and we want to test whether the population means are all equal for the infinite values of $\overline{x}_1$.

The four curves illustrated in Fig. 1 show the value of the posterior probability of M_1 as $\overline{x}_1$ varies over $(0, 4)$ and several intrinsic priors.

It is clear that the curves practically agree. The solid curve illustrate the values of the posterior probability of M_1 as $\overline{x}_1$ varies when the population chosen to take out two observations is the first one, that is $j = 1$; the dashed curve illustrate the values of this probability when the population chosen to take out two observations is the second one, that is $j = 2$; the dotted curve when the population is the third one; and finally, the long dashed curve when the intrinsic prior to obtain the Bayes factor is that given in (6).

So, again and more generally we illustrate that the posterior probability of M_1 is not sensitive to the minimal training sample chosen. Therefore, we can argue that to use the conditional prior (6) is a good alternative to avoid the dependence of the Bayes factor on the minimal training sample. Finally, from Fig. 1, it is clear that the posterior probability of M_1 is a good Bayesian measure of evidence for the homoscedasticity one way analysis of variance since only when $\overline{x}_1$ is close to 1 the value of the posterior probability of M_1 is close to 1 too.

Note that, in both examples, we have computed the posterior probabilities of M_1 using the prior probabilities $p_1 = p_2 = 1/2$ for models M_1 and M_2, respectively, and we have computed the Bayes factor using the software Mathematica.

4 Calibration of p values

In the homoscedastic one way ANOVA the classical measure of evidence against the null commonly used is the p value, while the corresponding Bayesian measure is the posterior probability of the null. Calibration is a simple means of establishing a map between the p value and a meaningful measure of evidence as it is the posterior probability of the null. In this section we calibrate the p values used in the homoscedastic one way ANOVA in a similar way to the one in objective testing procedures in linear models considered by Girón et al. (2006). The p value and the Bayes factor for intrinsic priors in the homoscedastic one way ANOVA depend on the sample through the same sufficient statistics. This result allows us to define a calibration curve by the parametric equations

$$y = P(M_1|\overline{\mathbf{x}}, \mathbf{s^2}, \mathbf{n}), \tag{9}$$

$$p = P_{M_1}(T \geq t|\overline{\mathbf{x}}, \mathbf{s^2}, \mathbf{n}), \tag{10}$$

where the parameters are the sample means, the sum of squares of the differences with respect to the mean and the sample sizes. These equations provide an easy way of calibrating the $p$value against the corresponding posterior probability of model M_1. To illustrate this idea we obtain calibration curves by varying in (9) and (10) a sample mean in an interval while the remaining arguments are kept fixed; of course, we can vary any component of the sufficient statistics in an interval while keeping fixed the remaining arguments to obtain another calibration curves.

An important feature of a calibration curve is what can be termed the *disagreement region*. For any cutoff value α_0 of the p values, say 0.05 or 0.01, and any cutoff value of the posterior probability P_0, say 0.5, we define the disagreement region as the arc of the calibration curve intersecting one of the quadrants $Q_L = \{(x, y); x > \alpha_0, y < P_0\}$ or $Q_U = \{(x, y); x < \alpha_0, y > P_0\}$. If the calibration curve is a strictly monotonic increasing function of the p value, when the cutoff point (α_0, P_0) happens to lie in the calibration curve, the disagreement region is empty. If the point (p, y) lies in the disagreement region, then both testing criteria disagree; otherwise, both tests accept or reject model M_1, simultaneously. Therefore, an obvious use of the calibration curve is for finding the size of the UMP test to get both methodologies (frequentist and Bayesian) accept or reject model M_1, simultaneously. This value is given by the x-value of the intersecting point of the calibration curve with the line $y = 1/2$. Typical calibration curves corresponding to the homoscedastic one way ANOVA are shown in the next examples.

Example 3 We consider three normal populations with sample sizes $(n_1, n_2, n_3) = (10, 20, 25)$, with sample values $(\overline{x}_1, \overline{x}_2, \overline{x}_3) = (0.022, 0.101, 0.057)$ and $(s_1^2, s_2^2, s_3^2) = (7.965, 17.327, 24.992)$, respectively, and we want to test whether the population means are all equal.

For the given data set we compute the p value and the posterior probability of model M_1 and the values obtained are 0.97 and 0.96, respectively. As both values are closer to one, the frequentist and the Bayesian methodology favor the simpler model. If we link those values with the x-axis and the y-axis, respectively, we obtain the point

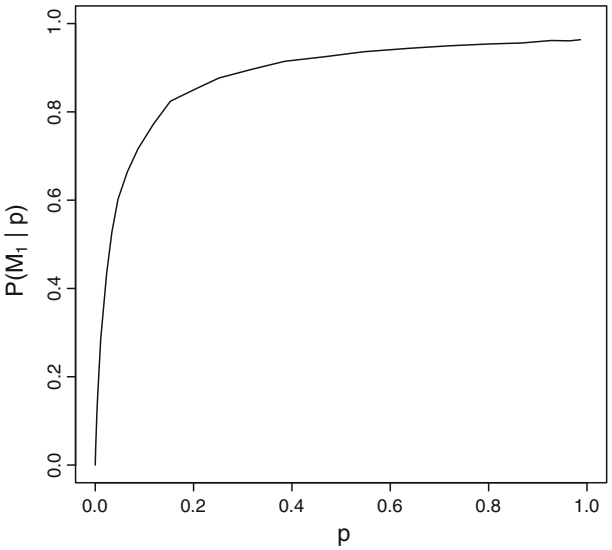

Fig. 2 Calibration *curve* for three normal populations with sizes $(n_1, n_2, n_3) = (10, 20, 25)$, samples values $\overline{x}_1 \in (-4, 1)$, $(\overline{x}_2, \overline{x}_3) = (0.101, 0.057)$ and $(s_1^2, s_2^2, s_3^2) = (7.965, 17.327, 24.992)$, respectively

Table 2 Changes of the sample size and the sum of squares of the differences with respect to the mean corresponding to the first population

	n_1^*	s_1^{*2}
Dashed curve	20	15.93
Dotted curve	80	63.72

$P(0.97, 0.96)$ that lies in the agreement region $Q^A = \{(x, y); x > \alpha_0, y > P_0\}$ for the standard values of α_0 and P_0.

Since we want to calibrate both procedures, now we consider the same example varying the value of $\overline{x}_1$. In particular, we consider $\overline{x}_1 \in (-4, 1)$. For the different values of $\overline{x}_1$ together with the others data, we obtain the respective p value and posterior probability of model M_1. If again we link those values with one point, we obtain the infinite points which allows us to define a calibration curve corresponding to this example. Figure 2 illustrate this calibration curve. From Fig. 2 it is clear that this calibration curve is a monotonic increasing function of the p value.

Now we study the behavior of this curve depending on the sample size of the populations. We illustrate this point with two examples.

Example 4 Calibration curves corresponding to the test of equality of means of the normal populations considered in Example 3 when the sample size of the first population is varied. For this we have changed some data from the first population. These data are displayed in Table 2.

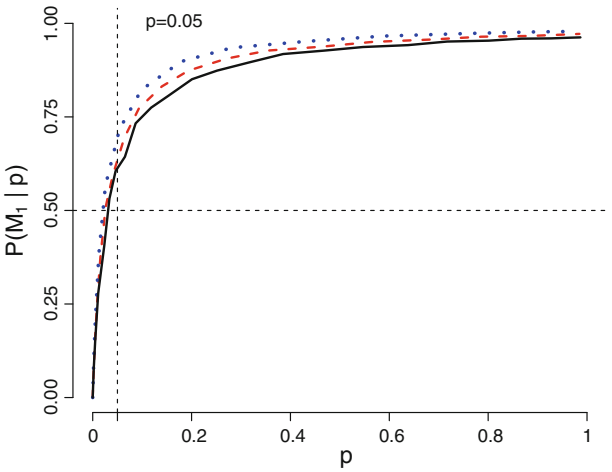

Fig. 3 Three calibration *curves* corresponding to three normal populations for different sample sizes $n_1 = 10$ (*black/solid*), $n_1 = 20$ (*red/dashed*) and $n_1 = 80$ (*blue/dotted*). The *horizontal* and *vertical lines* at 0.50 and 0.05, respectively, show that the disagreement region is in the quadrant Q_U

Note that the s_1^{*2} appearing in Table 2 are proportional to the s_1^2 previously considered through the ratio

$$s_1^{*2} = s_1^2 \frac{n_1^*}{n_1}.$$

The three calibration curves corresponding to these data are in Fig. 3. The solid curve here is the same curve appearing in Fig. 2. So, again these calibration curves are monotonic increasing functions of the p value. Another important feature is that its *disagreement region* is in the quadrant Q_U for $\alpha_0 = 0.05$ and $P_0 = 0.5$. Then, if we want both tests (frequentist and Bayesian) to give the same answer we should consider the size α_0 as the value of the x-axis at the intersecting point of the curve with the line $y = 1/2$ and this size will be smaller than the usually recommended test size $\alpha_0 = 0.05$. Furthermore, when n_1 increases the calibration curve increases as a function of the p value as it is shown in Fig. 3 and it has the following practical implication: when the sample size increases, the corresponding calibrated p values decreases and the size α_0 should be diminished accordingly if we want both tests to accept or reject model M_1 simultaneously.

Example 5 Calibration curves corresponding to the test of equality of means of three normal populations depending on their sample size. For this we consider three groups of three normal populations whose sample data are illustrated in Table 3.

Since $\overline{x}_1 \in (-4, 1)$ we obtain a calibration curve for each group. These curves are in Fig. 4.

Note that the solid curve is the calibration curve corresponding to group 1, the dashed curve is the calibration curve corresponding to group 2 and the dotted curve is the calibration curve corresponding to group 3, respectively.

Table 3 Sample data from three groups of normal populations

	n_i	$\overline{x}_i$	s_i^2
Group 1			
Population 1	10	$\overline{x}_1 \in (-4, 1)$	11.813
Population 2	15	0.116	12.901
Population 3	20	0.145	19.918
Group 2			
Population 1	30	$\overline{x}_1 \in (-4, 1)$	28.312
Population 2	45	−0.127	45.899
Population 3	60	−0.062	61.096
Group 3			
Population 1	90	$\overline{x}_1 \in (-4, 1)$	79.339
Population 2	135	−0.039	149.993
Population 3	180	0.040	182.871

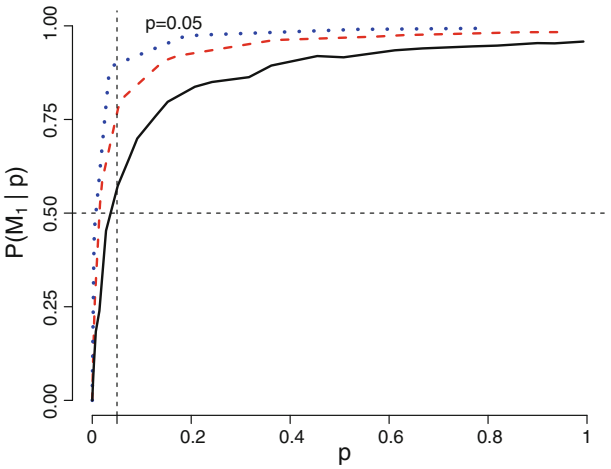

Fig. 4 Three calibration *curves* corresponding to the groups populations in Example 5: the *black/solid*, the *red/dashed* and the *blue/dotted curves* correspond to groups 1, 2 and 3, respectively. The *horizontal* and *vertical lines* at 0.50 and 0.05, respectively, show that the disagreement region is in the quadrant Q_U

From Fig. 4 it is clear that when the sample sizes increase, in particular we have triplicated them two times, the corresponding calibrated p values decrease. Therefore, when the sample sizes increase, then the size α of the usual UMP test should be diminished accordingly. Otherwise, if we stick to the usual recommended test sizes, regardless of the sample sizes, for large sample sizes we could reject hypotheses that have high posterior probabilities. This implies the same practical consequence of Example 4. Note that the behavior of the calibration curves as a function of the sample sizes in Examples 4 and 5 is similar to the behavior of the calibration curves regarded as a function of the same parameter in the study of objective testing procedures in linear models developed by Girón et al. (2006).

5 Summary and conclusions

We have developed a solution for the homoscedastic one way analysis of variance under the Bayesian viewpoint as a problem of objective Bayesian model selection using the Bayes factor as the main tool. We have continued the study of the heteroscedastic case developed in Bertolino et al. (2000) obtaining intrinsic priors for the homoscedastic case.

The Bayes factor obtained depends on the minimal training sample chosen, this implies a labelled problem which we study through two examples. In both examples we illustrate that this Bayes factor is not sensitive to the minimal training sample chosen.

In comparing both measures of evidence, p values and posterior probabilities of the null, following Girón et al. (2006) that study the calibration of p values for testing hypotheses about regression coefficients, we prove that in the case of the homoscedastic one way ANOVA there is a one-to-one relationship between them that permits to define an increasing curve that we call the *calibration curve*. This curve depends on the sample sizes, and when the sample sizes are large, it is apparent from the corresponding curves that the critical value for the p value has to be very small, otherwise we would reject the null when the posterior probability is large. Therefore, as the sample sizes increase the critical p value has to be diminished accordingly. The order of this diminution is obtained from the curve itself. A similar behavior was obtained by Girón et al. (2006) studying the calibration of the p values for testing hypothesis about regression coefficients.

References

Berger J, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of P-values and evidence. J Am Stat Assoc 82:112–122

Bertolino F, Racugno W, Moreno E (2000) Bayesian model selection approach to analysis of variance under heteroscedasticity. Stat 49(4):503–517

Box G, Tiao G (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading

Casella G, Berger RL (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J Am Stat Assoc 82:106–111

Casella G, Berger RL (1990) Statistical inference. Wadsworth and Brooks/Cole, Pacific Grove

Girón FJ, Martínez ML, Moreno E, Torres F (2006) Objective testing procedures in linear models: calibration of the P-values. Scand J Stat 33(4):765–784

Jeffreys H (1961) Theory of probability. Oxford University Press, Oxford

Lindley DV (1970) An introduction to probability and statistics from a Bayesian viewpoint. Cambridge University Press, Cambridge

Moreno E, Bertolino F, Racugno W (1998) An intrinsic limiting procedure for model selection and hypotheses testing. J Am Stat Assoc 93:1451–1460

Robert CP, Casella G (2001) Montecarlo statistical methods. Springer, New York

Rohatgi VK (1984) Statistical inference. Wiley, New York