RACSAM (2021) 115:158
https://doi.org/10.1007/s13398-021-01098-0

ORIGINAL PAPER

A Bayesian solution to the Behrens–Fisher problem

Fco. Javier Girón[1] · **Carmen del Castillo**[2,3]

Received: 20 August 2020 / Accepted: 6 July 2021 / Published online: 19 July 2021

Abstract
A simple solution to the Behrens–Fisher problem based on Bayes factors is presented, and its relation with the Behrens–Fisher distribution is explored. The construction of the Bayes factor is based on a simple hierarchical model, and has a closed form based on the densities of general Behrens–Fisher distributions. Simple asymptotic approximations of the Bayes factor, which are functions of the Kullback–Leibler divergence between normal distributions, are given, and it is also proved to be consistent. Some examples and comparisons are also presented.

Keywords Behrens–Fisher problem · Bayes factor · Hierarchical models

1 Introduction

The Behrens–Fisher problem is that of testing the equality of the means of two independent normal populations with unknown and arbitrary variances. The problem arises when the quotient between the variances is unknown. If this quotient is known, the problem is easily solved from both the frequentist and Bayesian approaches.

More specifically, suppose that we have two samples $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1})$ and $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n_2})$ of the populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. The problem is then to test the hyphotesis $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ with unknown variances.

Under a frequentist point of view, the Behrens–Fisher problem has the difficulty that standard homoscedastic normal theory cannot be applied due to the presence of the two unrelated variances, so that there is no an exact p-value. Following this approach and trying to find an approximate solution to the problem, two different but closely related methods have been addressed: the Neyman-Pearson theory of significance tests (Bartlett [1] , Welch

✉ Carmen del Castillo
carmelina@uma.es

Fco. Javier Girón
fj_giron@uma.es

[1] Real Academia de Ciencias Exactas, Físicas y Naturales, Calle de Valverde, 24, 28004 Madrid, Spain

[2] Dpto. Análisis Matemático, Estadística e Investigación Operativa, Universidad de Málaga, Facultad de Ciencias, Campus de Teatinos, 29071 Malaga, Spain

[3] Instituto Universitario de Investigación en Telecomunicación (TELMA), Universidad de Málaga, CEI Andalucía TECH, E.T.S.I. Telecomunicación, Bulevar Louis Pasteur 35, 29010 Málaga, Spain

[21], Wald [19], Scheffé [17]) and that based on the notion of generalized p-values (Tsui and Weerahandi [18], Weerahandi [20] and Witkovsky [22]). An extensive bibliography of such frequentists methods appears in Kim [12].

In a Bayesian approach, the problem has been studied as one of interval estimation of the difference of $\mu_1 - \mu_2$. In this way, as Lindley [13] proposes, Jeffreys [11] provides a credible interval for $\mu_1 - \mu_2$ based on the posterior distribution of the difference, using as a prior $\pi^J(\mu_1, \mu_2, \sigma_1, \sigma_2) \propto (\sigma_1\sigma_2)^{-1}$. It is well known that the posterior distribution of this difference is a Behrens–Fisher distribution (Behrens [2], Fisher [6,7], Box-Tiao [4], Girón et al. [9]).

There is not an agreed Bayesian solution to the Behrens–Fisher testing problem based on Bayes factors. This is mainly due to the fact that as the priors commonly used—generally, reference priors—are improper, the Bayes factor to compare the null and the alternative hypotheses cannot be computed. To overcome this difficulty, in Moreno et al. [14] and Moreno and Girón [15], the problem is formulated as one of Bayesian model selection, and intrinsic and fractional prior distributions are generated in order to compute a proper Bayes factor.

In this paper the Behrens–Fisher problem is approached as a true testing problem based on the Bayes factor. To this end, the problem is presented as a problem of testing the homogeneity of the means of two normal populations and a simple solution is provided by formulating a hierarchical model under the alternative hypothesis of the problem. The use of such hierarchical model will allow us to derive a proper Bayes factor, thus avoiding the impossibility that arises in its calculation when only improper distributions are used to compare models of different dimensions. The most important fact in the paper is that the solution provided here is simpler than the one obtained with the use of intrinsic priors distributions, since it only involves one-dimensional integrals, and the numerical results of the Bayes factors obtained are very similar in both cases.

In addition to the simplicity of the solution obtained, it is important to note that the prior distributions considered for the hierarchical model parameters imply that they are basically the same as those in Jeffreys [11], which shows an interesting relationship between the proposed testing approach and the estimation approach. This fact possibly explains the fact that it is possible to obtain an expression for the proposed Bayes factor using the density of the Behrens–Fisher distribution.

The article is structured as follows. In Sect. 2, the problem is posed as a homogeneity test and a hierarchical model under the alternative hypothesis is formulated to derive a proper Bayes factor. In Sect. 3, relationship between the Bayes factor and the Behrens–Fisher distribution is considered. In Sect. 4, a simple asymptotic approximation of the Bayes factor, related to the Kullback–Leibler divergence is given, and the consistency of the Bayes factor is proven when both sample sizes grow to infinity at the same rate. Section 5 provides several examples and comparison with other Bayesian and frequentist approaches.

Finally, Sect. 6 discusses some of the findings in the paper and suggests a more general solution in line with other approaches like the one based on intrinsic priors. Possible extensions of the presented results to the case of more than two samples and to the problem of comparison, i.e., testing the homogeneity of regression coefficients under heteroscedasticity are also addressed.

2 The Behrens–Fisher problem

In this section, the Behrens–Fisher problem is formulated as a problem of homogeneity of the means of two independent normal distributions with unknown and unequal variances.

Suppose that we have two independent samples $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1})$ and $\mathbf{x}_2 = (x_{21}, \ldots, x_{2n_2})$ of populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. An homogeneity test for the means of the two sample problem can be stated as that of comparing the hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu; \sigma_1^2 \text{ and } \sigma_2^2 \text{ arbitrary vs. } H_1 : \mu_1 \neq \mu_2; \tau_1^2 \text{ and } \tau_2^2 \text{ arbitrary},$$

where the null hypothesis H_0 represents the equality of the mean parameters, whilst the alternative hypothesis states that the population means are different, irrespective of the variances of the normal populations, which are treated as nuisance parameters.

Thus, we have, in principle, seven unknown parameters $\mu, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1^2$ and τ_2^2 to assign prior information. Under H_0, reference priors are considered, whereas the use of a hierarchical model only for the means, and objective priors for the common mean parameter μ, and the unknown variances τ_1^2, τ_2^2 is proposed under H_1.

The general form of the likelihood function for the data $\mathbf{x}_1$ and $\mathbf{x}_2$ with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is

$$L(\mathbf{x}_1, \mathbf{x}_2; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \prod_{i=1}^{2} (2\pi\sigma_i^2)^{-n_i/2} \exp\left[-\frac{1}{2\sigma_i^2} \left(\nu_i s_i^2 + n_i(\mu_i - \bar{x}_i)^2 \right) \right],$$

where $\bar{x}_1, \bar{x}_2$ are the sample means, s_1^2 and s_2^2 are the unbiased estimates of the variances σ_1^2 and σ_2^2, respectively, and $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$, the degrees of freedom.

Under H_0, the marginal of the data $\mathbf{x}_1, \mathbf{x}_2$, conditional on μ, σ_1^2 and σ_2^2 is the likelihood function substituting μ_1 and μ_2 by μ, that is

$$f_0(\mathbf{x}_1, \mathbf{x}_2 | \mu, \sigma_1^2, \sigma_2^2) = \prod_{i=1}^{2} (2\pi\sigma_i^2)^{-n_i/2} \exp\left[-\frac{1}{2\sigma_i^2} \left(\nu_i s_i^2 + n_i(\mu - \bar{x}_i)^2 \right) \right].$$

This conditional marginal density of the data only depends on the nuisance parameters μ, σ_1^2 and σ_2^2, which are regarded to be independent and they are assigned reference priors, i.e.,

$$\mu \perp\!\!\!\perp \sigma_1^2 \perp\!\!\!\perp \sigma_2^2 \text{ and } h(\mu, \sigma_1^2, \sigma_2^2) = c_0 c_1 c_2 (\sigma_1^2)^{-1} (\sigma_2^2)^{-1},$$

where $\perp\!\!\!\perp$ means independence or conditional independence.

Therefore,

$$f_0(\mathbf{x}_1, \mathbf{x}_2) = c_0 c_1 c_2 \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \int_{\mathbb{R}} f_0(\mathbf{x}_1, \mathbf{x}_2 | \mu, \sigma_1^2, \sigma_2^2) (\sigma_1^2)^{-1} (\sigma_2^2)^{-1} d\sigma_1^2 d\sigma_2^2 d\mu.$$

Integrating first with respect to σ_1^2 and σ_2^2, and then with respect to μ, the marginal under H_0 turns out to be

$$f_0(\mathbf{x}_1, \mathbf{x}_2) = c_0 c_1 c_2 \int_{\mathbb{R}} \prod_{i=1}^{2} \pi^{-n_i/2} \Gamma\left(\frac{n_i}{2}\right) \left(\nu_i s_i^2 + n_i(\mu - \bar{x}_i)^2 \right)^{-n_i/2} d\mu.$$

Note that the integrals with respect to the population variances are analytic but the integral with respect μ has not a closed form formula.

Under H_1, the hierarchical model is given by setting the first hierarchy

$$\mu_1 | \mu, \tau_1^2 \perp\!\!\!\perp \mu_2 | \mu, \tau_2^2 \text{ and } \mu_i | \mu, \tau_i^2 \sim N(\mu_i | \mu, \tau_i^2),$$

the second is that

$$\tau_1^2|\sigma_1^2 \perp\!\!\!\perp \tau_2^2|\sigma_2^2 \text{ and } \tau_i^2|\sigma_i^2 \sim s(\sigma_i^2) \text{ where } s \text{ is any scale density function}$$

and the thirth, and last, hierarchy is, as before, that

$$\mu \perp\!\!\!\perp \sigma_1^2 \perp\!\!\!\perp \sigma_2^2 \text{ and } h(\mu, \sigma_1^2, \sigma_2^2) = c_0 c_1 c_2 (\sigma_1^2)^{-1} (\sigma_2^2)^{-1}.$$

Note that the first hierarchy is related to the g-priors of Zellner and Ziow [23] in the one-dimensional case when $g_i = n_i$, with the important difference that they are centered at μ, as the intrinsic priors, instead of 0, as is usually done.

According to the following Lemma 1, and taking into account the second and third hierarchy of the model, it is easy to deduce that the prior distribution of τ_i^2 is the same reference prior as that of σ_i^2 for $i = 1, 2$, with the same constant c_i.

Lemma 1 *If the prior density of σ_i^2 is $h(\sigma_i^2) = c_i/\sigma_i^2$ $(i = 1, 2)$, with c_i any positive constant, and the distribution of $\tau_i^2|\sigma_i^2 \sim s(\sigma_i^2)$, where $s(.)$ is any scale density function, then the prior distribution of τ_i^2 is $h(\tau_i^2) = c_i/\tau_i^2$.*

Proof If s_0 is the density generator of the scale family, then

$$s(\tau_i^2|\sigma_i^2) = \frac{1}{\sigma_i^2} \cdot s_0\left(\frac{\tau_i^2}{\sigma_i^2}\right).$$

If we do the change of variables $y = \tau_i^2/\sigma_i^2$ in the integral, then, after some obvious simplifications

$$h(\tau_i^2) = \int_{\mathbb{R}^+} s_0(y) \frac{c_i}{\tau_i^2} dy = \frac{c_i}{\tau_i^2} \int_{\mathbb{R}^+} s_0(y) dy = \frac{c_i}{\tau_i^2}.$$

□

Lemma 1 implies that the second and third hierarchy of the proposed model are

$$\mu \perp\!\!\!\perp \tau_1^2 \perp\!\!\!\perp \tau_2^2 \text{ and } h(\mu, \tau_1^2, \tau_2^2) = c_0 c_1 c_2 (\tau_1^2)^{-1} (\tau_2^2)^{-1}.$$

In this way, the computation of the marginal under H_1 runs as follows

$$f_1(\mathbf{x}_1, \mathbf{x}_2|\mu, \mu_1, \mu_2, \tau_1^2, \tau_2^2) = \prod_{i=1}^{2} (2\pi\tau_i^2)^{-n_i/2} \exp\left[-\frac{1}{2\tau_i^2}\left(\nu_i s_i^2 + n_i(\mu_i - \bar{x}_i)^2\right)\right],$$

$$g(\mu_1, \mu_2|\mu, \tau_1^2, \tau_2^2) = \prod_{i=1}^{2} (2\pi\tau_i^2)^{-1/2} \exp\left[-\frac{1}{2\tau_i^2}\left(\mu_i - \mu\right)^2\right],$$

$$h(\mu, \tau_1^2, \tau_2^2) = \frac{c_0 c_1 c_2}{\tau_1^2 \tau_2^2}.$$

Multiplying the three equalities and rearranging the quadratic forms of the exponent, we get

$$f_1(\mathbf{x}_1, \mathbf{x}_2|\mu, \mu_1, \mu_2, \tau_1^2, \tau_2^2) = c_0 c_1 c_2 \prod_{i=1}^{2} (2\pi)^{-(n_i+3/2)} (\tau_i^2)^{-(n_i+3/2)/2}$$

$$\times \exp\left[-\frac{1}{2\tau_i^2}\left(\nu_i s_i^2 + (n_i+1)(\mu_i - \bar{x}_i)^2 + \frac{n_i}{n_i+1}(\mu - \bar{x}_i)^2\right)\right].$$

Now, the marginal of the data is obtained by integrating out this expression, first with respect to μ_1 and μ_2 and, then, with respect to τ_1^2 and τ_2^2. The last integral, with respect to μ has not an analytical closed form. Therefore, after computing all the integrals, the marginal turns out to be

$$f_1(\mathbf{x}_1, \mathbf{x}_2) = c_0 c_1 c_2 \times \int_{\mathbb{R}} \prod_{i=1}^{2} \frac{1}{\sqrt{n_i + 1}} \pi^{-n_i/2} \Gamma\left(\frac{n_i}{2}\right) \left(\nu_i s_i^2 + \frac{n_i}{n_i + 1}(\mu - \bar{x}_i)^2\right)^{-n_i/2} d\mu.$$

Finally, after simplifying common terms, the Bayes factor for testing H_0 vs. H_1 turns out to be

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) = \prod_{i=1}^{2} \sqrt{n_i + 1} \frac{\int \prod_{i=1}^{2} \left(\nu_i s_i^2 + n_i(\mu - \bar{x}_i)^2\right)^{-n_i/2} d\mu}{\int \prod_{i=1}^{2} \left(\nu_i s_i^2 + \frac{n_i}{n_i+1}(\mu - \bar{x}_i)^2\right)^{-n_i/2} d\mu},$$

which can be written in simplified form as

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) = \prod_{i=1}^{2} \sqrt{n_i + 1} \frac{\int \prod_{i=1}^{2} \left(1 + n_i \frac{(\mu - \bar{x}_i)^2}{\nu_i s_i^2}\right)^{-n_i/2} d\mu}{\int \prod_{i=1}^{2} \left(1 + \frac{n_i}{n_i+1} \frac{(\mu - \bar{x}_i)^2}{\nu_i s_i^2}\right)^{-n_i/2} d\mu}. \quad (1)$$

It can be remarked that the application of the Lemma 1, would have allowed us to formulate the hypothesis test —by abusing the notation just a little—with the same variances in both hypotheses, i.e., in the form

$$H_0 : \mu_1 = \mu_2 = \mu; \sigma_1^2 \text{ and } \sigma_2^2 \text{ arbitrary vs. } H_1 : \mu_1 \neq \mu_2; \sigma_1^2 \text{ and } \sigma_2^2 \text{ arbitrary.}$$

An interesting point that we would like to make to conclude this Section is that μ_1 and μ_2 have the same improper marginal as μ, as it is easily proved. Therefore, we can establish a close relationship between the proposed Bayesian approach and the one given by Jeffreys [11] as far as prior distributions are concerned. This is possibly the reason why an expression of the Bayes factor in (1) can be obtained in terms of the density function of the Behrens–Fisher distribution, which we will study in the next section of the paper.

3 Relation with the Behrens–Fisher distribution

It is interesting to see that the integrands in both the numerator and the denominator of the Bayes factor in (1) are proportional to the product of the densities of two Student t distributions. In this section, we will prove that these integrals can be evaluated through the general form of a Behrens–Fisher distribution. To do this, we begin by recalling the definitions of the standard and generalized Behrens–Fisher distributions.

Definition 1 A random variate b_0 follows a standard Behrens–Fisher distribution with degrees of freedom $f_1 > 0$, $f_2 > 0$ and angle $\phi \in [0, \pi/2]$ if $b_0 = t_1 \sin\phi - t_2 \cos\phi$, where t_1 and t_2 are independent random variates following Student t distributions with $f_1 > 0$ and $f_2 > 0$ degrees of freedom, respectively. It will be denoted by

$$b_0 \sim \text{Be-Fi}(f_1, f_2, \phi).$$

The extension of this distribution to a location-scale family is defined in Girón et al. [9] as follows.

Definition 2 A random variate b is said to be distributed as a generalized Behrens–Fisher distribution with location $\mu \in (-\infty, \infty)$, scale $\sigma > 0$, degrees of freedom $f_1 > 0$, $f_2 > 0$ and angle $\phi \in [0, \pi/2]$ if $b = \mu + \sigma b_0$. It will be denoted by

$$b \sim \text{Be-Fi}(\mu, \sigma^2, f_1, f_2, \phi).$$

The following theorems are given in Girón et al. [9]. Theorem 1 shows that the generalized Behrens–Fisher distribution is a convolution of two general Student t distributions. Note that the general form $t(\mu, \sigma^2, \nu)$ corresponds to a Student t distribution with location parameter μ, scale parameter σ^2 and degrees of freedom ν. Theorem 2 states that the generalized Behrens–Fisher distribution is a location mixture of Student's t distributions with mixing distribution a Student t.

Theorem 1 *If* $t_i \sim t(\mu_i, \sigma_i^2, f_i)$, $i = 1, 2$ *and* t_1, t_2 *are independent, then*

$$b = t_1 \pm t_2 \sim \textit{Be-Fi}(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2, f_1, f_2, \phi),$$

where $\phi \in [0, \pi/2]$ *is such that* $\tan^2 \phi = \sigma_1^2/\sigma_2^2$.

Theorem 2 *If*

$$\begin{aligned} x|\mu &\sim t(\mu, \sigma_0^2, f_1) \\ \mu &\sim t(m_0, \tau_0^2, f_2) \end{aligned}$$

then $x \sim$ *Be-Fi*$(m_0, \sigma_0^2 + \tau_0^2, f_1, f_2, \phi)$ *where* ϕ *is such that* $\tan^2 \phi = \sigma_0^2/\tau_0^2$.

As a consequence of Theorem 2, the probability density function of x can be expressed as

$$\begin{aligned} f_x(x) = &\frac{\Gamma\left(\frac{f_1+1}{2}\right)\Gamma\left(\frac{f_2+1}{2}\right)}{\pi\,\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)\sqrt{f_1 f_2 \sigma_0^2 \tau_0^2}} \\ &\times \int_{\mathbb{R}} \left(1 + \frac{(x-\mu)^2}{f_1 \sigma_0^2}\right)^{-\frac{f_1+1}{2}} \left(1 + \frac{(\mu - m_0)^2}{f_2 \tau_0^2}\right)^{-\frac{f_2+1}{2}} d\mu \end{aligned} \tag{2}$$

Once we have introduced the above definitions and results, let us consider again the Bayes factor in expression (1).

Defining $d = \bar{x}_2 - \bar{x}_1$, the integral of the numerator in (1) is

$$\int_{\mathbb{R}} \left(1 + n_1 \frac{(\mu - \bar{x}_2 + d)^2}{\nu_1 s_1^2}\right)^{-n_1/2} \left(1 + n_2 \frac{(\mu - \bar{x}_2)^2}{\nu_2 s_2^2}\right)^{-n_2/2} d\mu.$$

From a simple change of variable $\delta = \mu - \bar{x}_2 + d$, the last equality can be written as

$$\int_{\mathbb{R}} \left(1 + n_1 \frac{\delta^2}{\nu_1 s_1^2}\right)^{-n_1/2} \left(1 + n_2 \frac{(\delta - d)^2}{\nu_2 s_2^2}\right)^{-n_2/2} d\delta.$$

Taking into account formula (2), the preceding last formula is, up to constants, the probability density function evaluated at zero of a random variable distributed as a Behrens–Fisher

distribution with location $\bar{x}_2 - \bar{x}_1$, scale $s_1^2/n_1 + s_2^2/n_2$, degrees of freedom ν_1, ν_2 and angle ϕ such as $\tan^2\phi = \dfrac{s_1^2/n_1}{s_2^2/n_2}$.

Thus, the integral can be written as

$$\frac{\pi\,\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1+1}{2}\right)\Gamma\left(\frac{\nu_2+1}{2}\right)}\sqrt{\nu_1\nu_2\frac{s_1^2}{n_1}\frac{s_2^2}{n_2}}\cdot f_{b_1}(0),$$

where f_{b_1} is the probability density function of a random variable b_1 distributed as

$$b_1 \sim \text{Be-Fi}\left(\bar{x}_2 - \bar{x}_1, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, \nu_1, \nu_2, \phi\right).$$

Following the same reasoning, the integral of the denominator in expression (1) can be expressed as

$$\frac{\pi\,\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1+1}{2}\right)\Gamma\left(\frac{\nu_2+1}{2}\right)}\sqrt{\nu_1\nu_2\frac{(n_1+1)s_1^2}{n_1}\cdot\frac{(n_2+1)s_2^2}{n_2}}\cdot f_{b_2}(0),$$

where f_{b_2} is the probability density function of a random variable b_2 distributed as

$$b_2 \sim \text{Be-Fi}\left(\bar{x}_2 - \bar{x}_1, \frac{(n_1+1)s_1^2}{n_1} + \frac{(n_2+1)s_2^2}{n_2}, \nu_1, \nu_2, \psi\right),$$

where ψ is an angle such that $\tan^2\psi = \dfrac{(n_1+1)s_1^2/n_1}{(n_2+1)s_2^2/n_2}$.

Finally, (1) turns out to be

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) = \frac{f_{b_1}(0)}{f_{b_2}(0)}. \tag{3}$$

4 Asymptotic behavior: approximations and consistency

In this section we will study some asymptotic properties of the proposed Bayes factor in order to know its behavior when sample sizes tend to infinity. We will also prove its consistency for the case where the sample sizes grow towards infinity at the same rate.

We will begin by giving an approximation of the Behrens–Fisher distribution when its degrees of freedom grow indefinitely, which will allow us to derive an approximation of the Bayes factor studied in this article for large sample sizes. Taking into account Theorem 1 and the well known result that the Student t distribution asymptotically follows a normal distribution if the degrees of freedom are large enough, the following theorem holds.

Theorem 3 *If $b \sim$ Be-Fi$(\mu, \sigma^2, f_1, f_2, \phi)$ and f_1, f_2 tends to ∞, then the distribution of b is approximately $N(\mu, \sigma^2)$.*

Considering last theorem, the distribution of the random variate b_1 in (3) can be approximated by a normal with mean $\bar{x}_1 - \bar{x}_2$ and variance $s_1^2/n_1 + s_2^2/n_2$, if the sample size is large enough, whilst the distribution of the random variate b_2 can be approximated by a normal with mean $\bar{x}_1 - \bar{x}_2$ and variance $s_1^2 + s_2^2$.

Thus, if $n_1 \to +\infty$ and $n_2 \to +\infty$, an approximation of the Bayes factor in (3) is given by

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) \simeq \frac{\sqrt{s_1^2 + s_2^2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \exp\left\{ -\frac{1}{2(s_1^2/n_1 + s_2^2/n_2)} (\bar{x}_1 - \bar{x}_2)^2 \right\} \times \exp\left\{ \frac{1}{2(s_1^2 + s_2^2)} (\bar{x}_1 - \bar{x}_2)^2 \right\}.$$

An interesting fact about the asymptotic Bayes factor is that it can be expressed as a function of the Kullback–Leibler divergences of normal distributions. In fact, if we take into account that the Kullback Leibler divergence between the probability distributions of two normal distributions $X \sim N(m_1, v_1^2)$ and $Y \sim N(m_2, v_2^2)$ is

$$\delta_{KL}(f_X || f_Y) = \frac{1}{2v_2^2} \left[(m_1 - m_2)^2 + (v_1 - v_2)(v_1 + v_2) + 2v_2^2 \log\left(\frac{v_2}{v_1}\right) \right],$$

then, the above asymptotic expression of the Bayes factor can be written as

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) \simeq \frac{\exp\left\{ \frac{1}{2(s_1^2/n_1 + s_2^2/n_2)} \right\}}{\exp\left\{ \frac{1}{2(s_1^2 + s_2^2)} \right\}} \exp\left\{ \delta_{KL}(f_Z || f_X) - \delta_{KL}(f_Z || f_Y) \right\},$$

where f_X, f_Y and f_Z are the probability density functions of the random variates X, Y and Z normally distributed as follows

$$\begin{aligned} X &\sim N\left(\bar{x}_1 - \bar{x}_2, s_1^2 + s_2^2\right) \\ Y &\sim N\left(\bar{x}_1 - \bar{x}_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right) \\ Z &\sim N(0, 1) \end{aligned}$$

One important property of the Bayes factor for the Behrens–Fisher is that it is consistent in the sense that, as n_1 and n_2 tend to infinity at the same rate, then, under the null hypothesis, the probability of the true model goes to 1, or equivalently, the Bayes factor goes to infinity, and under the alternative hypothesis it goes to 0. Next theorem states this result in a more precise form. As a byproduct, we also obtain a simple approximation to the Bayes factor for large values of n_1 and n_2 when the growing rate is of the same order.

Theorem 4 *The Bayes factor is consistent under the null and the alternative hypotheses, in the sense that if the model we are sampling from is the true one, the Bayes factor goes to infinity, in probability. More precisely:*

$$\lim_{n_1, n_2 \to +\infty} B_{01}(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \infty & \text{if } \mu_1 = \mu_2 \\ 0 & \text{if } \mu_1 \neq \mu_2. \end{cases}$$

Proof Assuming that the convergence towards infinity of n_1 and n_2 is of the same order, i.e., $n_1 = m$ y $n_2 = a \cdot m$, $(a > 0)$, the Bayes factor can be written as

$$B_{01}(\mathbf{x}_1, \mathbf{x}_2) \simeq \frac{\sqrt{am}\sqrt{s_1^2 + s_2^2}}{\sqrt{as_1^2 + s_2^2}} \exp\left(-\frac{am}{2(as_1^2 + s_2^2)}(\bar{x}_1 - \bar{x}_2)^2\right) \times \exp\left(\frac{1}{2(s_1^2 + s_2^2)}(\bar{x}_1 - \bar{x}_2)^2\right).$$

Considering that s_1^2 and s_2^2 are consistent estimators of σ_1^2 and σ_2^2, respectively; the following expectation

$$E\left[\frac{\sqrt{am}\sqrt{s_1^2 + s_2^2}}{\sqrt{as_1^2 + s_2^2}} \exp\left(-\frac{am(\bar{x}_1 - \bar{x}_2)^2}{2(as_1^2 + s_2^2)}\right) \exp\left(\frac{(\bar{x}_1 - \bar{x}_2)^2}{2(s_1^2 + s_2^2)}\right)\right],$$

can be approximated for large values of m by

$$\frac{\sqrt{am}\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{a\sigma_1^2 + \sigma_2^2}} E\left[\exp\left(-\frac{am}{2(a\sigma_1^2 + \sigma_2^2)}(\bar{x}_1 - \bar{x}_2)^2 + O\left(\frac{1}{\sqrt{m}}\right)\right) \times \exp\left(\frac{1}{2(\sigma_1^2 + \sigma_2^2)}(\bar{x}_1 - \bar{x}_2)^2 + O\left(\frac{1}{\sqrt{m}}\right)\right)\right].$$

Taking into account that the distribution of the difference of means is approximated by

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{a\sigma_1^2 + \sigma_2^2}{am}\right)$$

then, under the null hyphotesis $\mathrm{H}_0 : \mu_1 = \mu_2$,

$$\bar{x}_1 - \bar{x}_2 \sim N\left(0, \frac{a\sigma_1^2 + \sigma_2^2}{am}\right).$$

Thus,

$$\lim_{m\to\infty}\left\{\frac{\sqrt{am}\sqrt{\sigma_1^2 + \sigma_2^2}}{a\sigma_1^2 + \sigma_2^2} E\left[\exp\left(-\frac{am}{2(a\sigma_1^2 + \sigma_2^2)}(\bar{x}_1 - \bar{x}_2)^2 + O\left(\frac{1}{\sqrt{m}}\right)\right) \times \exp\left(\frac{1}{2(\sigma_1^2 + \sigma_2^2)}(\bar{x}_1 - \bar{x}_2)^2 + O\left(\frac{1}{\sqrt{m}}\right)\right)\right]\right\}$$
$$= \lim_{m\to\infty}\left\{\frac{\sqrt{am}\sqrt{\sigma_1^2 + \sigma_2^2}}{a\sigma_1^2 + \sigma_2^2}\sqrt{\frac{am(\sigma_1^2 + \sigma_2^2)}{a(2m-1)\sigma_1^2 + (2am-1)\sigma_2^2}}\right\} = \infty.$$

On the other hand, under the alternative hyphotesis $\mathrm{H}_1 : \mu_1 \neq \mu_2$,

$$\lim_{m\to\infty}\left\{\frac{\sqrt{am}\sqrt{\sigma_1^2 + \sigma_2^2}}{a\sigma_1^2 + \sigma_2^2} E\left[\exp\left(-\frac{am}{2(a\sigma_1^2 + \sigma_2^2)}(\bar{x}_1 - \bar{x}_2)^2 + O\left(\frac{1}{\sqrt{m}}\right)\right)\right.\right.$$

$$\times \exp\left(\frac{1}{2(\sigma_1^2+\sigma_2^2)}(\bar{x}_1-\bar{x}_2)^2+O\left(\frac{1}{\sqrt{m}}\right)\right)\Bigg]\Bigg\}$$

$$= \lim_{m\to\infty}\left\{\frac{\sqrt{am}\sqrt{\sigma_1^2+\sigma_2^2}}{a\sigma_1^2+\sigma_2^2}\sqrt{\frac{am(\sigma_1^2+\sigma_2^2)}{a(2m-1)\sigma_1^2+(2am-1)\sigma_2^2}}\right.$$

$$\left.\times \exp\left[-\frac{am(\mu_1-\mu_2)^2}{4}\left(\frac{1}{a\sigma_1^2+\sigma_2^2}-\frac{1}{a(2m-1)\sigma_1^2+(2am-1)\sigma_2^2}\right)\right]\right\}=0$$

□

5 Examples: simulation study and calibration curves

If we define the statistic $d=\bar{x}_2-\bar{x}_1$ as in Sect. 3, then the Bayes factor of equation (1), after a simple change of variable, can be rewritten as a function of d and the rest of sufficient statistics n_1, n_2, s_1^2, s_2^2

$$B_{01}(d,n_1,n_2,s_1^2,s_2^2) = \prod_{i=1}^{2}\sqrt{n_i+1}\,\frac{\int\left(1+n_1\frac{\delta^2}{\nu_1 s_1^2}\right)^{-n_1/2}\left(1+n_2\frac{(\delta-d)^2}{\nu_2 s_2^2}\right)^{-n_2/2}d\delta}{\int\left(1+\frac{n_1}{n_1+1}\frac{\delta^2}{\nu_1 s_1^2}\right)^{-n_1/2}\left(1+\frac{n_2}{n_2+1}\frac{(\delta-d)^2}{\nu_2 s_2^2}\right)^{-n_2/2}d\delta} \tag{4}$$

and, the posterior probability of the null hypothesis, assuming that both hypotheses are equaly likely, is

$$\Pr(H_0|d,n_1,n_2,s_1^2,s_2^2)=\frac{B_{01}(d,n_1,n_2,s_1^2,s_2^2)}{1+B_{01}(d,n_1,n_2,s_1^2,s_2^2)}.$$

From these formulas, it follows an interesting property of the Bayes factor and, consequently, of the posterior probability of the null hypothesis, which is that both are symmetric and unimodal functions of $d=\bar{x}_2-\bar{x}_1$, irrespective of the values of the sample sizes n_1 and n_2, and the unbiased estimates of the variances s_1^2 and s_2^2. This implies that the acceptance regions of the Bayes test, as functions of the statistic d, are always symmetric intervals around 0. Thus, the acceptance region is $A(d;n_1,n_2,s_1^2,s_2^2)=\{d: B_{01}(d,n_1,n_2,s_1^2,s_2^2)\geq 1\}$ or, equivalently, $A(d;n_1,n_2,s_1^2,s_2^2)=\{d:\Pr(H_0|d,n_1,n_2,s_1^2,s_2^2)\geq 1/2)\}$.

This behavior is illustrated with a well known example taken from Box and Tiao [4, pp 107-109]: In a spinning modification experiment involving independent data from two normal distributions, the following results for the sufficient statistics were obtained:

$$\begin{aligned}\bar{x}_1&=50 \quad n_1=20 \quad s_1^2=12\\ \bar{x}_2&=55 \quad n_2=12 \quad s_2^2=40\end{aligned} \tag{5}$$

The value of the observed statistic $d=\bar{x}_2-\bar{x}_1=55-50=5$. It then follows, from formula (4), that the numerical value of Bayes factor is

$$B_{01}(\bar{x}_1,n_1,s_1^2,\bar{x}_2,n_2,s_2^2)=0.306118,$$

and the posterior probability of the null hypothesis is

$$Pr(H_0|\bar{x}_1,n_1,s_1^2,\bar{x}_2,n_2,s_2^2)=0.234373,$$

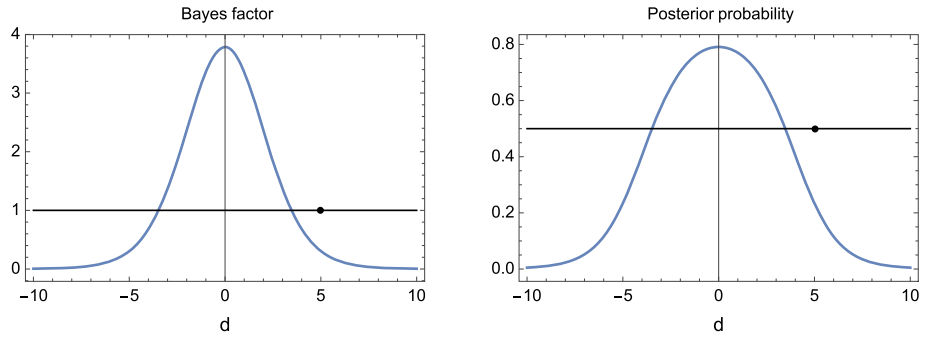

Fig. 1 Bayes factors and posterior probabilities of the null hypothesis as functions of the d statistic for the Box and Tiao example with $n_1 = 20$, $s_1^2 = 12$, $n_2 = 10$ and $s_2^2 = 40$. The point located at the horizontal lines corresponds to the observed value of the d statistic, which is outside of the acceptance interval

Table 1 Posterior probabilities of the null hypothesis for the proposed hierarchical model ($P^H(H_0|\mathbf{x}_1, \mathbf{x}_2)$), for the intrinsic approach ($P^I(H_0|\mathbf{x}_1, \mathbf{x}_2)$) and Welch's p-value for the Box and Tiao example with $n_1 = 20$, $s_1^2 = 12$, $n_2 = 10$ and $s_2^2 = 40$ and different values of the d statistic

d	$P^H(H_0\|\mathbf{x}_1, \mathbf{x}_2)$	$P^I(H_0\|\mathbf{x}_1, \mathbf{x}_2)$	p-value
0.00	0.79	0.83	1.00
2.20	0.68	0.75	0.30
4.22	0.37	0.46	0.06
5.00	0.23	0.32	0.03
10.00	0.005	0.008	0.002

meaning that the null hypothesis is rejected.

On the other hand, the acceptance intervals for the d statistic are obtained from the intersection of the Bayes factor or the posterior probability of the null hypothesis—regarded as functions of d—with the horizontal lines located at 1 and 1/2 values, respectively, as shown in Fig.1. The acceptance interval is $A(d; n_1, n_2, s_1^2, s_2^2) = (-3.47197, 3.47197)$. As the observed value of $d = 5$ does not belong to the acceptance interval, the null hypothesis is rejected.

Next, we will establish a comparison of the proposed hierarchical Bayes factor with both the intrinsic Bayes factor, appearing in Moreno et al. [14] and Moreno and Girón [15], and the most commonly used frequentist test based on the Welch statistic.

The only common statistic to the intrinsic, the hierarchical Bayes factors and the p-values is d. For this, Table 1 displays, for the values of s_1^2, s_2^2, n_1 and n_2 considered in (5), the posterior probabilities of the null hypothesis obtained through them and the resulting p-value of the Welch's t-test for the values of the d statistic reported in Moreno and Girón [15].

From Table 1 we can conclude that the two Bayesian procedures produce very similar results, with the intrinsic ones being slightly higher than the hierarchical ones. It seems to indicate that the intrinsic Bayes factor slightly favors the null hypothesis, but there is a general agreement between the report provided by both procedures about accepting or rejecting the null hypothesis. More extensive analysis on the comparison of the three procedures for various sample sizes and different variance estimates should be carried out to have more evidence about their behavior.

A striking difference of the intrinsic and hierarchical based Bayes factors for the Behrens–Fisher problem with the corresponding ones for the two-sample normal homoscedastic

problem is that, in the latter case, both Bayes factors are simple functions of the standard t statistic. This does not happen for the Behrens–Fisher problem. In fact, the test statistic for the Welch tests is

$$t = \frac{d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Nevertheless, the asymptotic approximation of tour Bayes factor does depend on t but also depends on the following statistic

$$u = \frac{d}{\sqrt{s_1^2 + s_2^2}}$$

and the ancillaries.

Only in the case of equal sample sizes, there is a one-to-one relationship between the asymptotic Bayes factor and the Welch statistic.

Next, we will try to further explore how an increase in sample sizes can influence the possible disagreements provided by the frequentist and Bayesian approaches presented in this article.

We know that the t Welch statistic is distributed as a Student distribution with approximately ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}}.$$

However, a calibration curve—the one that measures the relationship between p-values and posterior probabilities of the null—which was introduced for the normal linear models in Girón et al. [10], can be extended to the Behrens–Fisher problem using the common d statistic. Usually, the calibration curves vary with the sample sizes of the two samples—the ancillary statistics—, and the values of the unbiased estimates of the variances, s_1^2 and s_2^2. This implies that calibration curves for the Behrens–Fisher problem, depend on too many parameters. Thus, to illustrate their form and behavior, we have chosen to compute only those that depend on the common varying sample size $n = n_1 = n_2$ for fixed values of s_1^2 and s_2^2.

Figure 2 illustrates the shapes of the calibration curves for $s_1^2 = 1$ and $s_2^2 = 25$ and increasing values of $n = 5, 10, 30$ and 100. It also points out to the fact that, as the sample size grows to infinity, the calibration curve converges to a constant equal to 1 for all p-values in $(0, 1]$, except for $p = 0$, meaning that the posterior probability of the null hypothesis goes to 1 when sampling from it.

One important conclusion from the plot is, that for the same value of the posterior probability, the p-values decrease with the sample size. This fact has very important consequences in the usual statistical practice: in order to match Bayesian and frequentist procedures, the α-level for accepting/rejecting the null hypothesis, the sample size should be taken into account, in the direction that for very large sample sizes, the α-level should be substantially decreased.

6 Extensions and conclusions

This paper provides a simple Bayesian solution to the Behrens–Fisher problem based on Bayes factors. A simple hierarchical model for testing homogeneity of the means under

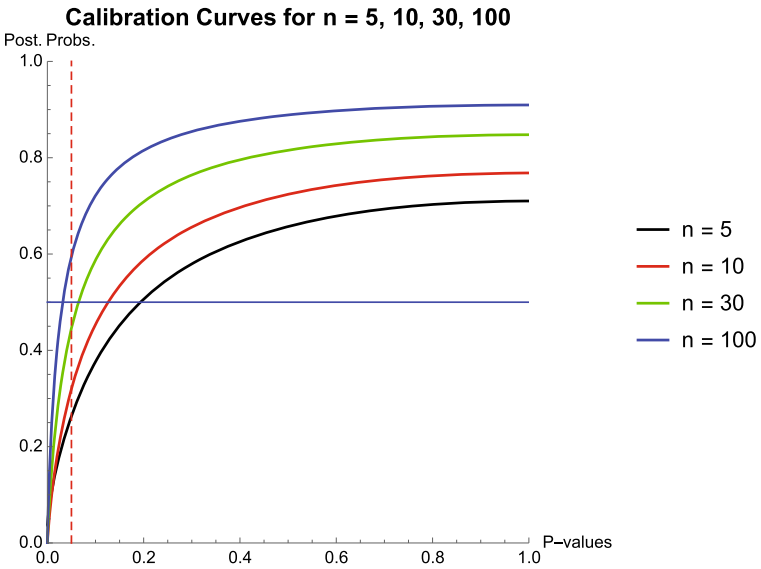

Fig. 2 Calibration curves for different values of the common sample size $n = 5, 10, 30$ and 100, for fixed values of the statistics $s_1^2 = 1$ and $s_2^2 = 25$

heteroscedasticity is considered to obtain a Bayes factor which is shown to be closely related to the densities of the general Behrens–Fisher distributions.

A comparison with the Bayes factor for intrinsic priors shows minor differences which basically produce the same Bayesian answers to the Behrens–Fisher problem. The inclusion of the calibration curve for the hierarchical Bayes factor shows disagreement with frequentist p-values when the sample sizes increase, a common characteristic of many Bayes factors.

Another possible Bayes factor that could be envisaged, following the steps suggested in Subsection 4.9.1 of [8], would involve a modification of the hierarchical model presented by introducing a certain hyperparameter σ^2 in the model and a new hierarchy by linking the variances with this hyperparameter by means of a weakly informative prior scale distribution, say $s(\cdot)$, an approach related to the one considered in Berger et al. [3] and similar to that of Moreno et al. [16], as follows: The hierarchy of the means μ_i given μ and their respective variances remains as before, and we add the following hierarchy of the variances given the common parameter σ^2:

$$\sigma_i^2 \perp\!\!\!\perp \sigma_2^2 \mid \sigma^2 \text{ and } \sigma_i^2|\sigma^2 \sim s(\sigma_i^2|\sigma^2),$$
$$\tau_i^2 \perp\!\!\!\perp \tau_2^2 \mid \sigma^2 \text{ and } \tau_i^2|\sigma^2 \sim s(\tau_i^2|\sigma^2).$$

In this way, the parameters on which to assign improper reference priors would be μ and σ^2 under both hypotheses, so the resulting Bayes factor would be proper. This solution is a little less simple than the one proposed in this paper since the the Bayes factor expression obtained involves the computation of two-dimensional integrals. Simulation studies carried out using Gamma, Inverted Gamma, Half normal and Half Cauchy distributions as links show robustness in the results obtained and a high degree of similarity with those obtained with the Bayes factor described here. This new Bayes factor would be more in line with other

objective or default Bayes factors, but the one presented here has the merit of being simple and does not differ much numerically from other possible solutions.

Finally, some possible extensions of the the simple hierarchical Bayes factor obtained in the paper could be easily applied to the problem of comparing the means of more than two samples of normal populations with unequal and unknown variances, and to the problem of comparing the regression coefficients of two, and more than two, heteroscedastic normal linear models.

Acknowledgements We want to thank an anonymous referee for her/his indications and comments which have greatly improved the paper.

Funding Open access funding provided by Universidad de Málaga/CBUA.

References

1. Bartlett, M.S.: The Information available in small samples. Proc. Cambridge Phil. Soc. **32**, 560–566 (1936)
2. Behrens, W.H.V.: Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. Landwirtsch Jahrbucher **68**, 807–837 (1929)
3. Berger, J.O., Pericchi, L.R.: Objective Bayesian methods for model selection. IMS Lecture Notes-Monograph Series **38**, 135–207 (2001)
4. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Addison-Wesley (1973)
5. Finney, D.J.: Statistical method in biological assay. Hafner, New York (1952)
6. Fisher, A.: The fiducial argument in statistical inference. Ann. Eugenics **6**, 391–398 (1935)
7. Fisher, A.: The comparison of samples with probability unequal variances. Ann. Eugenics **9**, 174–180 (1939)
8. Girón, F.J.: Bayesian Testing of Statistical Hypotheses. Arguval, Málaga (2021)
9. Girón, F.J., Martínez, M.L., Imlahi, L.: A characterization of the Behrens-Fisher distribution with applications to Bayesian inference. Comptes Rendus de l'Acadmie des Sciences, Series I, Mathematics **328**(8), 701–706 (1999)
10. Girón, F.J., Martínez, M.L., Moreno, E., Torres, F.: Objective testing procedures in linear models: calibration of the p-values. Scand. J. Stat. **33**, 765–784 (2006)
11. Jeffreys, H.: Theory of Probability. Oxford University Press (1961)
12. Kim, S.H., Cohen, A.H.: On the Behrens-Fisher problem. A review. J. Educ. Behav. Stat. **23**(4), 356–377 (1998)
13. Lindley, D.V., Scott, W.F.: New Cambridge elementary statistical tables (2nd ed.). Cambridge University Press, Cambridge (1995)
14. Moreno, E., Bertolino, F., Racugno, W.: Default Bayesian analysis of the Behrens-Fisher problem. J. Stat. Planning Inference **81**(2), 323–333 (1999)
15. Moreno, E., Girón, F.J.: On the frequentist and Bayesian approaches to hypothesis testing (with discussion). SORT **30**, 1–54 (2006)
16. Moreno, E., Girón, F.J., Vázquez-Polo, F.J.: Cost-effectiveness analysis for heterogeneous samples. Eur. J. Oper. Res. **254**, 127–137 (2016)
17. Scheffé, R.: Practical solutions of the Behrens–Fisher problem. J. Am. Stat. Assoc. **65**, 1501–1508 (1970)
18. Tsui, K.W., Weerahandi, S.: Generalized p values in significance testing of hypotheses in the presence of nuisance parameters. J. Am. Stat. Assoc. **84**, 602–607 (1989)

19. Wald, A.: Testing the difference between the means of two normal populations with unknown standard deviations. *Selected papers in Statistics and Probability*. T.W. Anderson *et al.* (eds.), 669-695, Stanford University Press (1955)
20. Weerahandi, S.: Generalized confidence intervals. J. Am. Stat. Assoc. **88**, 899–905 (1993)
21. Welch, B.L.: The generalization of students problem when several different population variances are involved. Biometrika **34**, 28–35 (1947)
22. Witkovsky, V.: Exact test of variance components using generalized p-values. *Folia, Fac. Sci. Nat. Universitatis Masarykianae Brunensis, Mathematica 9, Proceedings of the Summer School Datastat*, **99**, 119–125 (2001)
23. Zellner, A., Siow, A.: Posterior odds ratios for selected regression hypotheses (with discussion). In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (eds.) Bayesian Statistics, pp. 585–603. University Press, Valencia (1980)