Rev.R.Acad.Cienc.Exact.Fis.Nat. (Esp)
Vol. 93, N.º 3, pp 343-350, 1999
Monográfico: Métodos Bayesianos en las Ciencias

A PREDICTIVE APPROACH TO SOME HYPOTHESIS TESTING PROBLEMS

(Behrens-Fisher problem/exchangeability/highest predictive regions/hypothesis testing/predictive distributions/*p*-values)

F. J. GIRÓN*, M. L. MARTÍNEZ* AND E. M. PARRADO**

* Departamento de Estadística e I. O., Universidad de Málaga, Campus de Teatino s/n, Málaga, Spain. fj_giron@uma.es
** Departamento de Estadística y Econometría, Universidad de Málaga, Campus de El Ejido s/n, Málaga, Spain.

ABSTRACT

A new approach to testing statistical hypothesis is considered in this paper. The main departure from both classical and Bayesian testing is that the null and alternative hypothesis now refer to observable quantities, i.e. sample statistics, instead of model defining parameters. The idea for constructing the new tests for comparing two populations is based on the agreement between observed data or statistics and the corresponding predictive distributions, assuming that the resulting join sample, from the combination of the two samples, is exchangeable. This may result in that the null hypothesis may be neither rejected nor accepted, that is, the test may be inconclusive at some specified level or probabilistic content.

These ideas are illustrated and applied to some classical two-sample problems, where comparisons with the corresponding classical tests are considered, and extended to the testing of homegeneity of the sample medians of two populations and the classical Behrens-Fisher problem.

RESUMEN

Una aproximación predictiva para algunos problemas de contraste de hipótesis

En este trabajo se considera un nuevo enfoque al problema del contraste de hipótesis estadísticas. La diferencia más notable con el enfoque clásico y el bayesiano es la de considerar que, tanto la hipótesis nula como la alternativa, hacen referencia a cantidades observables, es decir, estadísticos, en vez de a los parámetros que definen un modelo estadístico. La idea para contribuir los nuevos contrastes en los problemas de dos muestras se basa en la bondad del ajuste o concordancia entre los datos observados o ciertos estadísticos calculados a partir de estos datos y sus correspondientes distribuciones predictivas, suponiendo que la muestra resultante de unir la dos muestras sea intercambiable. Este procedimiento puede producir contrastes que no aceptan ni rechazan la hipótesis nula y, por consiguiente pudieran ser no concluyentes a un nivel o contenido probabilístico prefijado.

Estas ideas se ilustran y se aplican a algunos problemas de las dos muestras clásicos, estableciéndose comparaciones con los contrastes clásicos correspondientes, y se extienden al problema de contraste de igualdad de medianas y al clásico problema de Behrens-Fisher.

1. INTRODUCTION

In this paper we deal with some two sample hypothesis testing problems from a new perspective based on the concepts of predictive distribution and highest predictive density regions.

It is well known that when testing one-sided hypothesis both classical and Bayesian methods are fairly in agreement and that, in this case, *p*-values are a fair approximation to posterior probabilities. However, difficulties appear when sharp or precise hypothesis are considered from either perspective; and that strong disagreement between *p*-values and posterior probabilities of the null hypothesis are generally expected. Many, see Nester (1996), have argued that null sharp hypothesis can never be exactly true.

From a Bayesian viewpoint, at least for exchangeable processes, parameters are almost sure limits of sample statistics; thus, the establishment of strict sharp hypothesis makes no sense. This simple consideration seems to have been overlooked in the large Bayesian literature devoted to the problem of testing null sharp hypothesis. Nevertheless, in model selection when, for instance, nested models are assumed, there is a need to formulate

simple models within the class of all contemplated models as a sharp null hypothesis.

The approach we adopt in this paper is the following: instead of testing hypothesis referring to population parameters we propose to test hypothesis referring to actual observable quantities. For testing these sort of hypothesis, which do not involve parameters, a Bayesian predictive perspective, see, e.g. Geisser (1993), is adopted. From this viewpoint, the parameters are but mere vehicles for modelling the data and, consequently, for computing predictive distributions involving observables, but they have no interest in themselves. Thus, for example, when one is trying to test the homogeneity of two samples from, say, normal populations, $\boldsymbol{x}_1 = (x_{11}, ..., x_{n_1})$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2n_2})$, respectively, one is not really interested in testing the equality of the defining parameters, i.e., $H_0 : (\mu_1, \sigma_1^2) = (\mu_2, \sigma_2^2)$ vs. $H_a : (\mu_1, \sigma_1^2) \neq (\mu_1, \sigma_2^2)$, which are in some sense chosen arbitrarily, but on testing if the resulting join sample $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ appears in some sense homogeneous, i.e., exchangeable.

An interesting feature of the procedures here proposed is that the new tests are based not on a single statistic but on two test statistics, thus opening up the possibility of neither accepting nor rejecting a null hypothesis, that is, of taking no action. This latter characteristic of the new approach has some relation —though the approach and interpretation are quite different— with the recent conditional tests proposed by Berger, Boukal and Wang (1997).

The main idea of this new approach is to establish another bridge between the classical and Bayesian paradigms to hypothesis testing problems.

The paper is divided in sections. In section 2 we give some definitions and state the precise meaning given to the testing of *equality* or *homogeneity* of sample statistics under exchangeability assumptions. In section 3 we analyse the classical problem of testing the equality of the means of two normal populations in the homoscedastic case from the new perspective and compare the results of the new test with that of the classical test. The classical Behrens-Fisher problem is analysed in section 4. For this problem the definitions given in section 2 do not apply as the two samples cannot be regarded as exchangeable, due to heteroscedasticity, so that we propose a new approach to the Behrens-Fisher problem which, interestingly enough, turns out to be equivalent to the well known Bayesian solution to the problem. Finally, after a brief discussion section, the proof of the lemmas needed to prove theorems 1 and 2 are gathered in the Appendix.

2. DEFINITIONS AND NOTATION

The following example is aimed at introducing the problem of testing the homogeneity of two samples and of some sample statistics, and justifying some definitions and notation to be used in the sequel.

Suppose $\boldsymbol{x}_1 = (x_{11}, ..., x_{1n_1})$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2n_2})$ are two independent samples from Poisson distributions with unknown parameters λ_1 and λ_2, respectively. The usual way of testing the homogeneity of the two samples is to test the equality of the defining parameters, that is to test $H_0 : \lambda_1 = \lambda_2$ vs. $H_a : \lambda_1 \neq \lambda_2$. From a more realistic Bayesian viewpoint the homogeneity of the two samples could be assesed by saying that the join sample $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a realization from a exchangeable Poisson process. This implies that both samples come from the same Poisson process, a statement which could be interpreted as an alternative to the classical homogeneity test. From this perspective, one way to test the homogeneity of the two samples could be to compute the two predictive distributions of $\boldsymbol{x}_1 | \boldsymbol{x}_j$, for $i = 1, 2$, $j = 1, 2$, $i \neq j$ and see whether the observed data $\boldsymbol{x}_i$, $i = 1, 2$ conforms or not to these distributions in a similar way to the one adopted when testing if some set of observations are or are not outliers.

A second alternative approach to testing homogeneity would be *to test* the *equality* of some sample statistics related to the parameter of the Poisson distribution; for example, to test whether the sample means, $\bar{x}_1$ and $\bar{x}_2$ are homogeneous, in some sense to be made precise later. A third alternative could be to test whether the sample variances, s_1^2 and s_2^2 are also homogeneous, as both pair of sequences converge almost surely to λ_1 and λ_2, respectively, as n_1 and n_2 tend to infinity. The way of testing the *equality* of these statistics is similar to the one described above, i.e., to compute the predictive distributions of these statistics conditional to the other sample and ponder their agreement with the values of the observed statistic.

In these cases —when comparing, for example, the equality of the sample means $\bar{x}_1$ and $\bar{x}_2$, or of s_1^2 and s_2^2, as the dimension of the two statistics are the same— another alternative method to construct a test would be to compute some *distance* measure, e.g. the Kullback-Leibler divergence between the predictive distributions $\bar{x}_1 | \boldsymbol{x}_2$ and $\bar{x}_2 | \boldsymbol{x}_1$ and $s_1^2 | \boldsymbol{x}_2$ and $s_2^2 | \boldsymbol{x}_1$, respectively.

As the solutions we have presented to test the homogeneity of two samples are different, we also expect to use different tests, accordingly; though one would expect that the asymptotic behaviour of the last two tests to be the same. Yet, we want to stress that, unlike the traditional approach to testing the equality of two parameters, the different testing problems considered above describe quite different statistical problems and, consequently, their solutions may be different.

To clarify this point, note that in the first homogeneity test proposed, which involves the computation of two predictive distributions, the predictive distribution of $\boldsymbol{x}_i | \boldsymbol{x}_j$ not only depends on the sufficient statistic $\Sigma_{k=1}^{n_j} x_{jk}$

but also on the *predictive* statistic $\Pi_{k=1}^{n_i} x_{ik}$!; unlike either the classical or the Bayesian parametric tests which only depend on the usual sufficient statistics.

Definition 2.1 makes precise the idea of homogeneity of two populations.

Definition 2.1. *Two samples* $\boldsymbol{x}_1 = (x_{11}, ..., x_{1n_1})$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2n_2})$ *are said to be* ***homogeneous****, and will be denoted by* $\boldsymbol{x}_1 \overset{h}{=} \boldsymbol{x}_2$, *if the resulting join sample* $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ *is exchangeable.*

Basically, the preceding definition is equivalent to assuming that the two populations share the same parameters if it holds true for all sample sizes n_1 and n_2.

Based on this definition and the preceding comments, a procedure for testing homogeneity would consist in deriving the predictive distribution of one sample given the other one and seeing the agreement between the real data and the predictive distribution. This could be done, for example, by computing a multivariate highest predictive density region of given probabilistic content and seeing whether the population belongs to that region. This implies that two comparisons are to be made in order to test homogeneity.

Thus, the proposed homogeneity test is the following: let R_1 and R_1 be highest predictive regions of $\boldsymbol{x}_1 | \boldsymbol{x}_2$ and $\boldsymbol{x}_2 | \boldsymbol{x}_1$ of a given probabilistic content $1 - \alpha$, respectively. Accept the null hypothesis of homogeneity of the two samples, $H_0 : \boldsymbol{x}_1 \overset{h}{=} \boldsymbol{x}_2$, at leve $1 - \alpha$ if $\boldsymbol{x}_1 \in R_1$ and $\boldsymbol{x}_2 \in R_2$. Reject the homogeneity hypothesis if $\boldsymbol{x}_1 \notin R_1$ and $\boldsymbol{x}_2 \notin R_2$. Otherwise, the test is inconclusive at the fixed probability content $1 - \alpha$, that is, the hypothesis is neither accepted nor rejected.

Obviously, the preceding test depends on the predictive distributions involved which, in turn, depend on the prior distribution assigned to common parameters. To establish links with the classical and Bayesian tests, we will assume throughout this paper the usual reference priors for the parameters involved.

As an application, we derive the following homogeneity test for two normal samples. Suppose $\boldsymbol{x}_1 = (x_{11}, ..., x_{1n_1})$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2n_2})$ are two exchangeable random samples from the same normal population $N(\mu, \sigma^2)$. Now, the predictive distribution of $\boldsymbol{x}_i | \boldsymbol{x}_j$, when the reference prior on μ, σ^2 is used, is the following multivariate Student t

$$\boldsymbol{x}_i | \boldsymbol{x}_j \sim t_{n_i}\left[\boldsymbol{x}_i | \mathbf{1}_{n_i}\bar{x}_j,\ s_j^2\left(\boldsymbol{I}_{n_i} + \frac{1}{n_j}\mathbf{1}_{n_i}\mathbf{1}_{n_i}'\right),\ v_j\right],$$

where, as usual, $v_k = n_k - 1$, $n_k\bar{x}_i = \Sigma_{k=1}^{n} x_{ik}$, $v_k s_i^2 = \Sigma_{k=1}^{n_i} (x_{ik} - \bar{x}_i)^2$. Then, the highest predictive region is

$$R_i = \left\{\boldsymbol{y} \in \mathbb{R}^{n_i} : (\boldsymbol{y} - 1_{n_i}\bar{x}_j)\left(\boldsymbol{I}_{n_i} + \frac{1}{n_j}\mathbf{1}_{n_i}\mathbf{1}_{n_i}'\right)^{-1} (\boldsymbol{y} - \mathbf{1}_{n_i}\bar{x}_j) \leqslant n_i s_j^2 F_{(n_i, v_j; 1-\alpha)}\right\},$$

where $F_{(n_i, v_j; 1-\alpha)}$ denotes the $1 - \alpha$ fractile of the F distribution with n_i and v_j degrees of freedom. After some algebraic manipulation, the homogeneity test becomes:

Accept (Reject) H_0 at level $1 - \alpha$ if, simultaneously, the following inequalities hold

$$v_1 s_1^2 + \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_2 - \bar{x}_1)^2 \leqslant (\geqslant)\ n_1 s_2^2 F_{(n_1, v_2; 1-\alpha)},$$

$$v_2 s_2^2 + \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2 \leqslant (\geqslant)\ n_2 s_1^2 F_{(n_2, v_1; 1-\alpha)},$$

otherwise, the test is inconclusive.

In Parrado (1996), several homogeneity tests for standard statistical models, derived from Definition 2.1, are given, along with other tests involving just one population.

Even though the two samples cannot be assumend homogeneous one may be interested in testing whether some sample statistics computed from the two samples are homogeneous as explained in the introductory example.

For any sample of arbitrary size $\boldsymbol{x} = (x_1, ..., x_n)$, let $T(\boldsymbol{x})$ be any fixed dimension statistic. We want to test if the value of this statistic is homogeneous for both samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which will be denoted by $H_0 : T(\boldsymbol{x}_1) \overset{h}{=} T(\boldsymbol{x}_2)$. Using the same idea as in the full homogeneity test, let C_1 and C_2 be highest predictive regions of $T(\boldsymbol{x}_1) | \boldsymbol{x}_2$ and $T(\boldsymbol{x}_2) | \boldsymbol{x}_1$ of a given probabilistic content $1 - \alpha$, respectively. Accept the null hypothesis of homogeneity of the two sample statistics, $H_0 : T(\boldsymbol{x}_1) \overset{h}{=} T(\boldsymbol{x}_2)$, at level $1 - \alpha$ if $T(\boldsymbol{x}_1) \in C_1$ and $T(\boldsymbol{x}_2) \in C_2$. Reject the homogeneity hypothesis if $T(\boldsymbol{x}_1) \notin C_1$ and $T(\boldsymbol{x}_2) \notin C_2$. Otherwise, the test is inconclusive, and the hypothesis is neither accepted nor rejected at the specified level.

Another possibility, which we do not explore in this paper, would consist in computing, for example, the Kullback-Leibler divergence between the two predictive distributions, $T(\boldsymbol{x}_1) | \boldsymbol{x}_2$ and $T(\boldsymbol{x}_2) | \boldsymbol{x}_1$, $\delta(p(T(\boldsymbol{x}_1) | \boldsymbol{x}_2), T(\boldsymbol{x}_2) | \boldsymbol{x}_1))$, so that the test would become.

$$\text{Accept } H_0 \text{ if } \delta(p(T(\boldsymbol{x}_1) | \boldsymbol{x}_2), T(\boldsymbol{x}_2) | \boldsymbol{x}_1)) \leqslant c; \text{ otherwise, reject } H_0$$

where c is some suitable positive constant.

Depending on the chosen statistic T, the tests will differ accordingly. Thus, for example if one is interested in comparing if two (one-dimensional) populations have an homogeneous location —that is, both samples are scattered roughly around the same position—, then the homogeneity of any location statistic, such as the sample mean $\bar{x}$ or the sample median $\tilde{m}$ could be tested. Note that if, for instance, the two statistical models are given by $f(x \mid \mu_1) = g(x - \mu_1)$ and $f(x \mid \mu_2) = g(x - \mu_2)$, where g(.) is symmetric, both parameters represent the population mean (assuming it exists) and the median, so that a standard parametric test of the null hypothesis $H_0 : \mu_1 = \mu_2$ would not distinguish between the two different situations.

3. TESTING HOMOGENEITY OF THE MEAN SAMPLE STATISTIC IN THE HOMOSCEDASTIC TWO-SAMPLE PROBLEM

Suppose $\boldsymbol{x}_1 = (x_{11}, ..., x_{1n_1})$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2n_2})$ are two exchangeable random samples from the same normal population $N(\mu, \sigma^2)$.

Consider first the case where the variance $\sigma^2 = \sigma_0^2$ is known. We want to test, according to the notation of the preceding section, the null hypothesis $H_0 : \bar{x}_1 \overset{h}{=} \bar{x}_2$.

The predictive distribution of $\bar{x}_i \mid \boldsymbol{x}_j$, when the reference prior for μ is used, is

$$\bar{x}_i \mid \boldsymbol{x}_j \sim N\left(\bar{x}_i \mid \bar{x}_j, \sigma_0^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

so that the highest predictive interval C_i for $\bar{x}_i$ is

$$C_i = \left(\bar{x}_j - z_{\alpha/2}\sigma_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_j + z_{\alpha/2}\sigma_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

where $z_{\alpha/2}$ denotes the $1 - \alpha/2$ fractile of the standard normal distribution. Thus, according to the procedure proposed in the preceding section, the new test becomes

Accept (Reject) H_0 at level $1 - \alpha$ if, simultaneously, the following inequalities hold

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sigma_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leqslant (\geqslant)\, z_{\alpha/2} \text{ and } \frac{|\bar{x}_2 - \bar{x}_1|}{\sigma_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leqslant (\geqslant)\, z_{\alpha/2};$$

otherwise, the test is inconclusive

But as two inequalities are the same, the test cannot be inconclusive, and it is identical to the classical test.

Suppose now that we want to test the equality of the sample medians $\tilde{m}_1$ and $\tilde{m}_2$, that is, now the null hypothesis is $H_0 : \tilde{m}_1 \overset{h}{=} \tilde{m}_2$. Assume that sample sizes n_1 and n_2 are large enough so that the asymptotic distribution of the sample median is approximately normal. Then the predictive distribution of $\tilde{m}_i \mid \boldsymbol{x}_j$, when the reference prior for μ is used, is

$$\tilde{m}_i \mid \boldsymbol{x}_j \approx N\left(\tilde{m}_i \mid \bar{x}_j, \sigma_0^2\left(\frac{\pi}{2n_i} + \frac{1}{n_j}\right)\right)$$

so that the approximate highest predictive interval C_i for $\tilde{m}_i$ is

$$C_i = \left(\bar{x}_j - z_{\alpha/2}\sigma_0\sqrt{\frac{\pi}{2n_i} + \frac{1}{n_j}}, \bar{x}_j + z_{\alpha/2}\sigma_0\sqrt{\frac{\pi}{2n_i} + \frac{1}{n_j}}\right).$$

Then, the test becomes

Accept (Reject) H_0 at level $1 - \alpha$ if, simultaneously, the following inequalities hold

$$\frac{|\tilde{m}_1 - \bar{x}_2|}{\sigma_0\sqrt{\frac{\pi}{2n_1} + \frac{1}{n_2}}} \leqslant (\geqslant)\, z_{\alpha/2} \text{ and } \frac{|\tilde{m}_2 - \bar{x}_1|}{\sigma_0\sqrt{\frac{\pi}{2n_2} + \frac{1}{n_1}}} \leqslant (\geqslant)\, z_{\alpha/2};$$

otherwise, the test is inconclusive

Note that the test for sample medians not only depends on these statistics but also on the sample means.

Now suppose the population variance σ^2 is unknown, and we want to test the null hypothesis $H_0 : \bar{x}_1 \overset{h}{=} \bar{x}_2$.

The predictive distribution of $\bar{x}_i \mid \boldsymbol{x}_j$, when the non-informative prior for the pair (μ, σ^2) is used, is the following Student t

$$\bar{x}_i \mid \boldsymbol{x}_j \sim t\left(\bar{x}_i \mid \bar{x}_j, \frac{n_1 + n_2}{n_1 n_2} s_j^2;\ \nu_j\right).$$

Thus, the highest predictive interval C_i for $\bar{x}_i$ is

$$C_i = \left(\bar{x}_j - t_{(\nu_j, \alpha/2)}s_j\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_j + t_{(\nu_j, \alpha/2)}s_j\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

where $t_{(\nu_j, \alpha/2)}$ denotes the $1 - \alpha/2$ fractile of the Student t distribution with ν_j degrees of freedom. Then, the new test becomes:

Accept (Reject) H_0 at level $1 - \alpha$ if, simultaneously, the following inequalities hold

$$t_1 = \frac{|\bar{x}_1 - \bar{x}_2|}{s_1\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leqslant (\geqslant)\ t_{(v_1,\,\alpha/2)}$$

and (1)

$$t_2 = \frac{|\bar{x}_1 - \bar{x}_2|}{s_2\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leqslant (\geqslant)\ t_{(v_2,\,\alpha/2)};$$

otherwise, the test is inconclusive

It seems interesting to compare this test with either the equivalent (parametric) classical test or the Bayesian test for the difference between two normal means based on the H.P.D. region of $\mu_1 - \mu_2$, assuming the reference prior for μ_1, μ_2, and σ^2.

The classical test is:

Accept (Reject) H_0 at level $1 - \alpha$ if,

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leqslant (\geqslant)\ t_{(v,\,\alpha/2)}, \qquad (2)$$

where $v = v_1 + v_2$ and s is the square root of the pooled variance estimator $s^2 = (v_1/v)s_1^2 + (v_2/v)s_2^2$.

The classical test statistic t is related to the statistics t_1 and t_2 by the following simple relations

$$t^2 = \frac{v_1 s_1^2}{vs^2}t_1^2 + \frac{v_2 s_2^2}{vs^2}t_2^2 \quad \text{and/or} \quad \frac{1}{t^2} = \frac{v_1}{v}\frac{1}{t_1^2} + \frac{v_2}{v}\frac{1}{t_2^2}, \qquad (3)$$

that is, t^2 is, at the same time, an arithmetic and a harmonic mean of t_1^2 and t_2^2, with weights depending on their degrees of freedom, and/or estimated variances, respectively.

The following theorem shows that whenever the new test rejects the null hypothesis so does the classical test, but not conversely. This obviously implies that, for small or moderate sample sizes, the classical test tends to reject the null hypothesis more often than new test as the latter may be inconclusive. However, acceptance of the null hypothesis by either the new test or the classical test does not necessarily imply the acceptance of the other one.

The proof of the theorem is based on the following lemma which states the stochastic dominance ordering of Student t distributions with respect to the degress of freedom.

Lemma 3.1. *For any two positive real degrees of freedom v and v' and for every $\alpha \in (0, 1)$, the following equivalence holds true*

$$t_{(v,\,\alpha/2)} < t_{(v',\,\alpha/2)} \Leftrightarrow v > v'.$$

Theorem 3.1. *If test (1) rejects the null hypothesis at level $1 - \alpha$, so does test (2) at the same level.*

Proof. If test (1) rejects the null hypothesis at level $1 - \alpha$, then $t_1^2 \geqslant t^2_{(v_1,\,\alpha/2)}$ and $t_2^2 \geqslant t^2_{(v_2,\,\alpha/2)}$. But, by lemma 1, $t^2_{(v_1,\,\alpha/2)} > t^2_{(v,\,\alpha/2)}$ and $t^2_{(v_2,\,\alpha/2)} > t^2_{(v,\,\alpha/2)}$. This implies that

$$\frac{1}{t^2} = \frac{v_1}{v}\frac{1}{t_1^2} + \frac{v_2}{v}\frac{1}{t_2^2} < \frac{v_1}{v}\frac{1}{t^2_{(v,\,\alpha/2)}} + \frac{v_2}{v}\frac{1}{t^2_{(v,\,\alpha/2)}} = \frac{1}{t^2_{(v,\,\alpha/2)}},$$

That is, $t^2 > t^2_{(v,\,\alpha/2)}$, and the classical test rejects the null hypothesis at the same level. □

Remark 1. For very large samples both tests accept and reject the null hypothesis at the same level. For if n_1 and n_2 are large enough then, for the usual (not too small) values of α, the fractiles $t^2_{(v_1,\,\alpha/2)} \approx t^2_{(v_2,\,\alpha/2)} \approx t^2_{(v,\,\alpha/2)} \approx z^2_{\alpha/2}$, where $z_{\alpha/2}$ denotes the $1 - \alpha/2$ fractile of the standard normal distribution. This implies that if the new test accepts the null hypothesis, i.e., $t_1^2 \leqslant z^2_{\alpha/2}$ and $t_2^2 \leqslant z^2_{\alpha/2}$ then, by (3), $t^2 \leqslant z^2_{\alpha/2}$, and the classical test accepts the null hypothesis.

Furthermore, it can be easily proved that the probability that the new test be inconclusive conditional on the homoscedasticity of the two populations tend to zero as $n_1 \to \infty$ and $n_2 \to \infty$, thus implying that the asymptotic behaviour of the new and the classical test are equivalent.

Next lemma is the same as lemma 3.1 but stated in terms of the F distribution.

Lemma 3.2. *For any two positive real degrees of freedom v and v' and for every real number x, the following holds*

$$F(x \mid 1, v) > F(x \mid 1, v') \Leftrightarrow v > v',$$

where $F(x \mid 1, v)$ denotes distribution function of the F distribution with 1 and v degrees of freedom.

Next theorem, which obviously implies Theorem 3.1, establish the relation between the p-value of the classical test based on the statistc t and the p-values associated to the single statistics t_1 and t_2.

Theorem 3.2. *Let p, p_1, and p_2 be the p-values associated to t, t_1, respectively, that is,*

$$p = 1 - \Pr(F(\cdot \mid 1, v) \leqslant t^2), \text{ and } p_i = 1 - \Pr(F(\cdot \mid 1, v_i) \leqslant t_i^2), \text{ for } i = 1, 2;$$

then, the following inequality holds

$$p < \left(\frac{v_1 s_1^2}{vs^2}\right)p_1 + \left(\frac{v_2 s_2^2}{vs^2}\right)p_2 \leqslant \max\{p_1, p_2\}.$$

Proof. From (3)

$$t^2 = \frac{v_1 s_1^2}{vs^2} t_1^2 + \frac{v_2 s_2^2}{vs^2} t_2^2.$$

Now, for every v the distribution function $F(x \mid 1, v)$ is a (strictly) concave function; hence

$$F(t^2 \mid 1, v) \geqslant \frac{v_1 s_1^2}{vs^2} F(t_1^2 \mid 1, v) + \frac{v_2 s_2^2}{vs^2} F(t_2^2 \mid 1, v).$$

But as $v > v_1$ and $v > v_2$, then from Lemma 3.2,

$$F(t^2 \mid 1, v) \geqslant \frac{v_1 s_1^2}{vs^2} F(t_1^2 \mid 1, v_1) + \frac{v_2 s_2^2}{vs^2} F(t_2^2 \mid 1, v_2),$$

which is equivalent to

$$1 - p > \left(\frac{v_1 s_1^2}{vs^2}\right)(1 - p_1) + \left(\frac{v_2 s_2^2}{vs^2}\right)(1 - p_2),$$

and the theorem follows. □

Remark 2. The p-value of the new test (1), say p_{new}, can be computed from the join sampling distribution of t_1 and t_2 conditional on the hypothesis of homoscedasticity, and is generally smaller than p. In particular —compare this result with the statement of theorem 2—, the inequality

$$p_{\text{new}} \leqslant \min\{p_1, p_2\}$$

follows easily from (1).

4. A PREDICTIVE APPROACH TO THE BEHRENS-FISHER PROBLEM

The preceding section dealt with the problem of testing the homogeneity of the sample mean statistic of two samples of normal populations when they were assumed to be exchangeable. On the other hand, the Behrens-Fisher problem deals with heteroscedastic populations, thus ruling out exchangeability and, consequently, the possibility of computing, as in the homoscedastic case, the predictive distribution of $\bar{x}_i$ give $\boldsymbol{x}_j$. Instead, we consider eliminating the nuisance parameter σ_i^2 by computing the predictive distribution of $\bar{x}_i$ given s_i^2, and $\boldsymbol{x}_j$. This predictive distribution turns out to be a generalised Behrens-Fisher distribution as defined in Girón, Martínez and Imlahi (1998).

Definition 4.1. *A random variable b is said to be distributed as a generalised Behrens-Fisher distribution with location* $\mu \in (-\infty, \infty)$, *scale* $\sigma \in (0, \infty)$, *degrees of freedom* v_1, v_2, *and angle* $\phi \in [0, \pi/2]$, *and will be denoted by*

$$b \sim BeFi(b \mid \mu, \sigma^2, v_1, v_2, \phi),$$

if

$$b_0 = \frac{b - \mu}{\sigma} \sim BeFi(b_0 \mid v_1, v_2, \phi);$$

where $BeFi(b_0 \mid v_1, v_2, \phi)$ *denotes the standard Behrens-Fisher distribution, so that standard Behrens-Fisher distribution has location and scale paramenter* $\mu = 0$ *and* $\sigma = 1$, *respectively.*

Theorem 4.1. *If a priori the parameters* μ, σ_1^2, σ_2^2 *are independent, i.e.* $\mu \perp\!\!\!\perp \sigma_1^2 \perp\!\!\!\perp \sigma_2^2$ *and follow the usual non informative prior, i.e.,* $f_0(\mu, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2}\sigma_2^{-2}$, *then for* $i, j = 1, 2$ $i \neq j$, *we have i):* $\mu \perp\!\!\!\perp \sigma_i^2 \mid s_i^2$, *and ii):* $\mu \perp\!\!\!\perp s_i^2 \mid \boldsymbol{x}_j$. *Furthermore,*

$$\bar{x}_i \mid \mu, s_i^2 \sim t(\bar{x}_i \mid \mu, s_i^2/n_i, v_i), \tag{4}$$

$$\mu \mid \boldsymbol{x}_j, s_i^2 \sim t(\mu \mid \bar{x}_j, s_j^2/n_j, v_j); \tag{5}$$

and the predictive distribution of $\bar{x}_j$ *given* s_i^2 *and* $\boldsymbol{x}_j$ *is given by*

$$\bar{x}_i \mid s_i^2, \boldsymbol{x}_j \sim \text{BeFi}(\bar{x}_i \mid \bar{x}_j, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, v_i, v_j, \phi_{ij}), \text{ where } \phi_{ij}$$

is such that $\tan^2 \phi_{ij} = \dfrac{s_i^2}{n_i} \Big/ \dfrac{s_j^2}{n_j}$.

Proof. By hypothesis $\mu \perp\!\!\!\perp \sigma_i^2$. Now, it is well known from normal sampling theory that $s_i^2 \mid \mu, \sigma_i^2 \overset{\text{d}}{=} s_i^2 \mid \sigma_i^2$, that is, $\mu \perp\!\!\!\perp s_i^2 \mid \sigma_i^2$. Using the fundamental property of conditional independence, e.g. see Girón, Kadane and Moreno (1997), we obtain $\mu \perp\!\!\!\perp s_i^2$ and $\mu \perp\!\!\!\perp \sigma_i^2 \mid s_i^2$, thus proving i).

By examining the conditional distribution of $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ given μ, that can be easily derived from Box and Tiao (1973), which turns out to be improper, it can be proven that $\boldsymbol{x}_i \perp\!\!\!\perp \boldsymbol{x}_j \mid \mu$, and this, in turn, implies $s_i^2 \perp\!\!\!\perp \boldsymbol{x}_j \mid \mu$. As from above $\mu \perp\!\!\!\perp s_i^2$, then using again the fundamental property of conditional independence we get $s_i^2 \perp\!\!\!\perp \boldsymbol{x}_j$ and ii), that is, $\mu \perp\!\!\!\perp s_i^2 \mid \boldsymbol{x}_j$. From this, it follows that the distribution of $\mu \mid \boldsymbol{x}_j, s_i^2$ is equal to that of $\mu \mid \boldsymbol{x}_j$; that is, equation (5) holds.

The computation of (4) is somewhat more involved. We first compute the join conditional distribution of the pair $\bar{x}_i$, σ_i^2 given μ, s_i^2, and finally the marginal of $\bar{x}_i \mid \mu, s_i^2$. But the join distribution of $\bar{x}_i$, $\sigma_i^2 \mid \mu, s_i^2$ can decomposed in that of $\bar{x}_i \mid \mu, \sigma_i^2, s_i^2$ and that of $\sigma_i^2 \mid \mu, s_i^2$. By Fisher's theorem

$$\bar{x}_i \mid \mu, \sigma_i^2, s_i^2 \overset{\text{d}}{=} \bar{x}_i \mid \mu, \sigma_i^2 \sim N(\mu, \sigma_i^2/n_i),$$

and by i),

$$\sigma_i^2 \mid \mu,\ s_i^2 \overset{d}{=} \sigma_i^2 \mid s_i^2 \sim \mathrm{Ga}^{-1}(v_i/2,\ v_i s_i^2/2).$$

From this, recalling that the marginal of a normal-inverted-gamma is a Student, (4) holds. Finally, the predictive distribution of $\bar{x}_i$ given s_i^2, $\boldsymbol{x}_j$ follows directly from theorem 2 in Girón, Martínez and Imlahi (1998). □

As a consequence of this theorem we obtain that the highest predictive interval C_i for $\bar{x}_i$ is

$$C_i = \left(\bar{x}_j - b_{(v_i, v_j, \phi_{ij}, \alpha/2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \bar{x}_j + b_{(v_i, v_j, \phi_{ij}, \alpha/2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right),$$

where $b_{(v_i, v_j, \phi_{ij}, \alpha/2)}$ denotes the $1 - \alpha/2$ fractile of the standard Behrens-Fisher distribution with v_i, v_j degrees of freedom and angle ϕ_{ij}. Then, the new test becomes:

Accept (Reject) $H_0 : \bar{x}_1 \overset{d}{=} \bar{x}_2$ at level $1 - \alpha$ if, simultaneously, the following inequalities hold

$$b = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leqslant (\geqslant)\ b_{(v_1, v_2, \phi_{12}, \alpha/2)}$$

and (6)

$$b = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leqslant (\geqslant)\ b_{(v_2, v_1, \phi_{21}, \alpha/2)};$$

otherwise, the test is inconclusive

But as $b_{(v_1, v_2, \phi_{12}, \alpha/2)} = b_{(v_2, v_1, \phi_{21}, \alpha/2)}$ for all $v_1 > 0$, $v_2 > 0$, ϕ_{12}, $\phi_{21} = \pi/2 - \phi_{12}$ and $\alpha \in (0, 1)$, by the properties of the standard Behrens-Fisher distribution, then both inequalities are the same, so that the test is always conclusive and is identical to the parametric Bayesian test, based on H.P.D.'s intervals (see, Box and Tiao (1973), pp. 104-109), for testing the equality of the two population means $H_0 : \mu_1 = \mu_2$ in the heteroscedastic case.

Remark 3. The way of eliminating the nuisance parameter σ_i^2 in order to compute a proper predictive distribution for $\bar{x}_i \mid \boldsymbol{x}_j$ is simplified by the fact that Fisher's theorem states the conditional independence of $\bar{x}_i$ and s_i^2 given μ and σ_i^2. In more complex situations it is not at all clear how to proceed, this being a subject for further research.

5. DISCUSSION

The idea of constructing parametric tests for testing sharp null hypothesis by using H.P.D.'s regions of fixed probability content has been advocated in the past by many authors, mainly Lindley (1970) and Box and Tiao (1973), but has been criticized by many authors as a valid alternative to hypothesis testing, since the work of Berger and Delampady (1987), Casella and Berger (1987), and Berger and Sellke (1987), as addressing a different inferential problem, and propose as solutions using lower bounds of Bayes factors for some classes of prior distributions.

However, in this paper we address a different problem from that of testing a sharp parametric null hypothesis; namely, that of testing hypothesis referring to observable statistics, which, in general, cannot be regarded as sharp hypothesis. On the other hand, for observable quantities, we use the equivalent of paramentric H.P.D.'s regions: namely, highest predictive density regions to construct our tests, but other different approaches are possible though not explored in the paper. One advantage of using predictive distributions for testing is that, when no prior information is available, there is no need to resort to other reference priors —such as intrinsic priors or mixed priors when the null hypothesis is sharp—, different from those generally used for inference purposes, say point or confidence estimation.

One intriguing consequence of this approach is that the resulting tests may be inconclusive, a characteristic which is somehow related to the fact that the *p*-values of these tests are generally smaller than the corresponding ones of classical tests, a fact which points out at the use of smaller values of the *p*-values than customary in order to reject the null hypothesis. This is in agreement with the recommendations put forward in the references listed in the preceding paragraph.

ACKNOWLEDGMENTS

This paper has been partially supported by the *Dirección General de Investigación Científica y Técnica* (DGICYT) as part of the project PB97-1403 C03-02 and by *La Consejería de Educación de la Junta de Andalucía* as part of one *Marroquí-Andaluz* project.

REFERENCES

1. Berger, J. O. and Delampady, M. (1987). Testing precise hypothesis. *Statist. Sci.*, **3**, 317-352.
2. Berger, J. O., Boukal, B. and Wang, Y (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statist. Sci.*, **12**, 133-160.

3. Berger, J. O, and Sellke, T. (1987). Testing a point null hypothesis: the irreconciability of significance levels and evidence (with discussion). *J. Amer. Statist. Assoc.,* **82**, 112-122.

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

5. Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.,* **82**, 106-111.

6. Geisser, S (1993). *Predictive Inference: An Introduction.* Chapman and Hall: New York.

7. Girón, F. J., Kadane, J. B. and Moreno, E. (1997). Independence issues in imprecise data models: a Bayesian approach. *C. R. Acad. Sci. Paris,* **324**, Série 1, 1149-1153.

8. Girón, F. J., Martínez, M. L. and Imlahi, L. (1999). A characterization of the Behrens-Fisher distribution with applications to Bayesian inference. *C. R. Acad. Sci. Paris,* **325**, Série 1, 701-706.

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

10. Nester, M. R. (1996). An applied statistician's creed. *J. Roy. Statist. Soc. C,* **45**, 401-410.

11. Parrado, E. M. (1996). *Aplicaciones de la distribución predictiva a la robustez de los modelos dinámicos jerárquicos y a los contrastes de hipótesis.* Ph. D. dissertation. Málaga.