ESTADISTICA (2005), **57**, 168 y 169, pp. 1-9
©Instituto Interamericano de Estadística

POWER OF FBST: STANDARD EXAMPLES

MARIA REGINA MADRUGA
Departamento de Estatística,
Universidade Federal do Pará, Brazil.
madruga@ufpa.br

CARLOS ALBERTO DE BRAGANCA PEREIRA
Instituto de Matemática e Estatística,
Universidade de São Paulo, Brazil.
cab3p@aol.com

ABSTRACT

The Full Bayesian Significance Test (FBST) introduced in Pereira & Stern (1999) is the Bayesian alternative to the traditional significance test based on p-values. The FBST is based on e-values which is also an index that measures the consistency of the data with the null sharp hypothesis. This paper considers standard situations, for which the p-values are celebrated to compare the power of the two significance tests.

Key words

Empiric power, Empiric significance level, Evidence, p-value.

1. INTRODUCTION

Testing precise hypothesis is the problem discussed in this paper. By precise hypothesis we understand a hypothesis that belongs to a subspace whose dimension is smaller than the dimension of the original parameter space. Berger & Selke (1987) and Berger & Delampady (1987) suggest that, when testing precise hypothesis, there exists a conflict among their (Bayesian) measure of conditioning evidence and the p-value, the classical measure of significance (or evidence). In fact they conclude that "p-values can be highly misleading measures of the evidence provided by the data against the null hypothesis". We believe that the introduction, in the prior distribution, of a positive probability to the precise hypothesis (a set of null Lebesgue measure) seems not appropriate and could be, as Lindley Paradox, the heart of the conflict. The Full Bayesian Significance Test (FBST) introduced by Pereira & Stern (1999) evaluates the evidence in favor of the hypothesis after comparing densities of points inside the precise hypothesis with the ones in its complement. To use the FBST there is no need to introduce a positive probability for a set that naturally has null Lebesgue measure. Using this genuinely Bayesian procedure we show that in most standard cases p-value are very closed to the values obtained by this evidence measure.

Our objective is to compare, through simulation studies, the performance of the FBST with the classical standard procedure. We choose 3 simple celebrated examples for this comparison. Section 2 describes the FBST in detail. Section 3 introduces the empirical quantities used in the comparison. The case of testing the equilibrium of a coin is presented Section 4. The comparison of the means of two normal distributions is discussed in Section 5 for the case of equal variances and the Behrens-Fisher in Section 6.

2. FBST

Pereira & Stern (1999) introduced the following measure of evidence in favor of a precise hypothesis,

Definition 1: *Consider a parametric statistical model, i.e., a quintet* $(\mathcal{X}, \mathcal{A}, \mathbf{F}, \Theta, \pi)$, *where* $\mathcal{X}$ *is a sample space,* $\mathcal{A}$ *is a suitable sigma-algebra of subsets of* $\mathcal{X}$, $\mathbf{F}$ *is a class of probability distributions on* $\mathcal{A}$ *indexed on a parametric space* Θ *and* π *is a prior density over (a sigma-algebra of)* Θ. *Suppose a subset* Θ_0 *of* Θ *having null Lebesgue measure (wrt* Θ*) is of interest. Let* $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ *be the posterior density of* $\boldsymbol{\theta}$, *given the sample observation* $\boldsymbol{x}$, *and* $T(\boldsymbol{x}) = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\boldsymbol{x}) > \sup_{\Theta_0} \pi(\boldsymbol{\theta}|\boldsymbol{x})\}$. *The measure of evidence, e-value, is defined as* $EV(\Theta_0, \boldsymbol{x}) = 1 - Pr[\boldsymbol{\theta} \in T(\boldsymbol{x})|\boldsymbol{x}]$ *and a test (or procedure) is to accept* Θ_0 *whenever* $EV(\Theta_0, \boldsymbol{x})$ *is "large".*

As we can see from Definition 1, the measure of evidence considers, in favor of a precise hypothesis, all points of the parametric space whose posterior density values are, at most, as large as its supremum over Θ_0; roughly speaking, it considers all points which are less "probable" than some point in Θ_0. Also, we should remember that, according to Pereira & Stern (1999), a large value of $Ev(\Theta_0, \boldsymbol{x})$ means that the subset Θ_0 lies in a high-probability region of Θ and, therefore, the data support the null hypothesis; on the other hand, a small value of $Ev(\Theta_0, \boldsymbol{x})$ points out that Θ_0 is in a low-probability region of Θ and the data would make us discredit the null hypothesis.

An advantage of this procedure is that it overcomes the difficulty of dealing with a precise hypothesis because there is no need of to introduce a prior positive probability as in the usual Bayesian test, that use the Factor of Bayes. Pereira & Stern (1999) claim that the use of $EV(\Theta_0, \boldsymbol{x})$ to assess the evidence of Θ_0 is a "Bayesian" procedure, as only the posterior density is involved. Madruga et al. [2001] presented loss functions that turn the FBST a legitimate "Bayesian" procedure, because one would call "Bayesian" a procedure which minimizes expected loss functions - the coherent solution to the decision problem. Like this, the FBST consists in

- Reject H_0 if $EV(\Theta_0, \boldsymbol{x}) \leq K$
- Accept H_0 if $EV(\Theta_0, x) > K$

where K is a cutoff whose value depends on the loss function chosen. For instance, Madruga et al. (2001) consider $D = \{Accept\ H_0\ (d_0), Reject\ H_0\ (d_1)\}$ the decision space and define the loss function: $L : \boldsymbol{D} \times \Theta \rightarrow IR^+$, $L(Reject\ H_0,\ \theta) = a[1 - \mathbf{1}(\theta \in T(\boldsymbol{x}))]$ e $L(Accept\ H_0,\ \theta) = b + c\mathbf{1}(\theta \in T(\boldsymbol{x}))$, $a, b, c > 0$. They prove that for this loss function the cutoff value is $K = \frac{b+c}{a+c}$.

3. SIGNIFICANCE AND POWER

Pereira & Stern (1999) presented solutions for some known statistical problems using their measure, $EV(\Theta_0, \boldsymbol{x})$. In most of these applications there was an apparent agreement among the solution given by the $EV(\Theta_0, \boldsymbol{x})$ and the classic solution given by the p-value. Of course, the comparison among these solutions is difficult, once the classic solution possesses a decision rule well defined, that is, the comparison of the p-value with a pre-established significance level, while the decision rule for the FBST depends on the loss function chosen by the researcher. A more effective form of comparison among these solutions is through simulation studies, being observed the empiric power of the two solutions. In order to continue with this comparative study we will define the following empiric measures:

Definition 2: *The empiric significance level of a test, α, is given by the proportion of times that we rejected the null hypothesis when it is true.*

Definition 3: *The empiric power of a test, $\rho(\boldsymbol{\theta})$, is given by the proportion of times that we rejected the null hypothesis when it is false.*

With these definitions, we will compare the empiric power of the classic procedure, using the appropriate p-value, with the empiric power of the FBST for some problems of test of hypotheses known. Initially consider the test of the parameter of Binomial. In this case the empiric significance level will be fixed in $\alpha = 0.038$. Following, the comparison of two averages of normal populations with same variance will be considered. In this problem we will admit 5 % of rejections of the true null hypothesis, in other words, the empiric significance level will be $\alpha = 0.05$. We will also make the comparison of the empiric results in the test of equality of the normal averages supposing different variances, that is in the problem of Behrens-Fisher. In this case also we will admit $\alpha = 0.05$. In each one of these problems the rule of decision will be based on the fixed empiric significance level.

4. TEST OF THE PARAMETER OF THE BINOMEAL DISTRIBUTION

Let $X \sim \text{Binomial}(n, \theta)$. Based on observed result x, we want to test the hypotheses

$$H_0 : \theta = 0.5 \qquad \text{against} \qquad H_1 : \theta \neq 0.5$$

Taking the prior distribution to θ as the uniform density in $(0, 1)$, the posterior distribution is given by

$$\theta | x \sim \text{Beta}(x + 1, n - x + 1).$$

Table 1 presents the empiric power, $\rho(\theta)$, in order to test the hypotheses above, considering some values of θ. The results are based on 1000 samples of size $n = 20$ of X. The decision rules to the p-value (p) and the $EV(\Theta_0, x)$, based on the fixed empiric significance level, $\alpha = 3.8\%$, are given by

- p-value (p) : Reject H_0 if $p < 0.0254$.
- Measure of evidence $(EV(\Theta_0, x))$: Reject H_0 if $EV(\Theta_0, x) < 0.0219$.

Table 1 we observe that the empiric power is the same for two tests.

TABLE 1

Empiric power to the Binomial test

θ	Empiric power	
	Classic	$FBST$
0.05	1.000	1.000
0.10	0.987	0.987
0.20	0.776	0.773
0.30	0.410	0.410
0.40	0.147	0.146
0.50	0.038	0.038
0.60	0.124	0.124
0.70	0.412	0.412
0.80	0.793	0.793
0.90	0.990	0.990
0.95	1.000	1.000

5. COMPARASION OF TWO AVERAGES OF NORMAL POPULATIONS

Let $X \sim N(\mu_1, \sigma^2)$ e $Y \sim N(\mu_2, \sigma^2)$ be independents random variables. Based on the observed values of X and Y, $\mathbf{z} = (\boldsymbol{x}, \mathbf{y})$, with $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n})$ and $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_m})$, we want to test the hypotheses:

$$H_0 : \mu_1 = \mu_2 \qquad \text{against} \qquad H_1 : \mu_1 \neq \mu_2$$

Using the standard improper prior, $\pi(\mu_1, \mu_2, \sigma^2) \propto \frac{1}{\sigma}$, for the normal parameters, in order to get a fair comparison with p-values (DeGroot (1970)), the conditional posterior distribution of $\gamma = \mu_1 - \mu_2$, given σ^2, is

$$\gamma|(\sigma^2, \mathbf{z}) \sim N\left(\bar{x} - \bar{y}\ ;\ \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

Table 2 presents the empiric power, $\rho(\gamma)$, considering some values of γ. The presented results are based on 1000 pairs of samples of variables X and Y, with sizes $n = 10$ and $m = 13$, respectively, generated from normal distributions with variances equal to 1 and averages $\mu_1 = 0$ and $\mu_2 = \mu_1 - \gamma$. The rules of decision to the p-value (p) and the $EV(\Theta_0, x)$, based on the fixed empiric significance level, $\alpha = 5\%$, are given by

- p-value (p) : Reject H_0 if $p < 0.05$.
- Measure of evidence ($EV(\Theta_0, \mathbf{z})$) : Reject H_0 if $EV(\Theta_0, \mathbf{z}) < 0.1740$.

Again the results indicate an agreement of the empiric power of two tests.

TABLE 2

Empiric power to the test of equality of means

	Empiric power	
γ	Classic	*FBST*
-2.0	0.995	0.995
-1.5	0.918	0.914
-1.0	0.640	0.637
-0.5	0.202	0.198
0.0	0.050	0.050
0.5	0.209	0.207
1.0	0.654	0.648
1.5	0.929	0.927
2.0	0.994	0.994

6. THE PROBLEM OF BEHRENS-FISHER

In this Section we consider again the test of equality of two normal averages but with different variances. In the classic approach (or frequentist) this problem has some approximate solutions that are discussed and compared in Mehta & Srinivasan (1970). If the ratio between the variances is known there are exact solutions of the classic point of view (Lehmann (1959)). The standard Bayesian procedure for tests of hypotheses is not applied in the problem of Behrens-Fisher, maybe because the usual improper prior for normal parameters dont allow the calculation of the Factor of Bayes. Due to this, in general, the bayesians treat the problem of Behrens-Fisher as a problem of estimation, whose objective is to estimate $D = \mu_1 - \mu_2$. Moreno et al. (1999) present Bayesian solutions to test the equality of the averages, based on Bayes Intrinsic Bayes Factor and on Fractional Bayes Factor, proposed by Berger & Pericchi (1996) and O'Hagan (1995), respectively.

The Table 3 presents the results based on 1000 pairs of samples of sizes 10 and 13, generated by two normal distributions with parameters $\mu_1 = 0$, $\sigma_1^2 = 1$, μ_2 and

$\sigma_2^2 = 2$. The decision rules based on the fixed empiric significance level, $\alpha = 5\%$, are given by

- p-value (p) : Reject H_0 if $p \leq 0.0520$.
- Measure of evidence $(EV(\Theta_0, \mathbf{z}))$: Reject H_0 if $EV(\Theta_0, \mathbf{z}) \leq 0.5188$.

Now, if we maintain the same values for μ_1 and σ_1^2 and we take $\sigma_2^2 = 4$, the decision rules changes to

- p-value (p) : Reject H_0 if $p \leq 0.0555$.
- Measure of evidence $(EV(\Theta_0, \mathbf{z}))$: Reject H_0 if $EV(\Theta_0, \mathbf{z}) \leq 0.5235$.

In this case, the results are presented in the Table 4 and again we have similarities for the two tests. Comparing the results of the two tables, we observed that the measure of evidence is so sensitive to the increase of the variance as the p-value, because the number of rejections in the test decreases considerably when we maintain the same values for the averages of the two populations and we increased the variance of one of them.

TABLE 3

The problem of Behrens-Fisher $(\mu_1 = 0, \sigma_1^2 = 1, \sigma_2^2 = 2)$

	Empiric power	
μ_2	Classic	$FBST$
0.5	0.119	0.120
1.0	0.328	0.327
1.5	0.604	0.602
2.0	0.840	0.839

TABLE 4

The problem of Behrens-Fisher ($\mu_1 = 0, \sigma_1^2 = 1, \sigma_2^2 = 4$)

	Empiric power	
μ_2	Classic	*FBST*
0.5	0.067	0.071
1.0	0.137	0.146
1.5	0.243	0.256
2.0	0.368	0.373

7. FINAL REMARKS

The performance of the FBST measured by its powers is at least as good as the tests based on p-values. In most of the situations studied here, the power of the alternative tests are equivalent. For low dimension problems e-values and p-values are very similar. For higher dimension parameter spaces they could differ drastically. However in almost all problems there are functions describing the relationship among these indexes. Whenever the threshold that states the accept/reject rules, one may compute the empirical power of the test. By minimizing a linear function of the two kinds of errors, one may find the best decision threshold. In our case we minimize the sum of the two kinds of errors.

We notice that the FBST has no restriction of dimensionality for its use if appropriate computation facilities are available. Another interesting fact is that since FBST is based only on the posterior distribution, it does not depend on the stopping rule used. It depends only on the likelihood function. This fact shows that FBST does not violate the likelihood principle as most p-values do.

REFERENCES

BERGER, J. O. and SELKE, T. (1987). "Testing a Point Null Hypothesis: The Irreconcilability of P values and Evidence". *Journal of the American Statistical Association*, **82**: 112-139.

BERGER, J. O. and DELAMPADY, M. (1987). "Testing Precise Hypotheses". *Statistical Science*, **2**: 317-352.

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

DEGROOT, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

JEFFREYS, H. (1961). *Theory of Probability.* University Press, Oxford.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

MADRUGA, M. R., ESTEVES, L. G. and WECHSLER, S. (2001). "On the Bayesianity of Pereira-Stern tests". *Test*, **10 (2)**.

MEHTA, J. S. and SRINIVASAN, R. (1970). "On the Behrens-Fisher problem". *Biometrika*, **57**: 649-655.

MORENO, E., BERTOLINO, F. and RACUGNO, W. (1999). "Default Bayesian analysis of the Behrens-Fisher problem". *Journal of Statistical Planning and Inference*, **81**: 323-333.

O'HAGAN, A. (1995). "Fractional Bayes factor for model comparison". *Journal of Royal Statistical Society B*, **57**: 99-138.

PEREIRA, C. A. de B. and STERN, J. (1999). "Evidence and Credibility: a full Bayesian test of precise hypothesis". *Entropy*, **1**: 99-110.