FULL BAYESIAN SIGNIFICANT TEST - FBST

Carlos Alberto de Bragança Pereira [1]

Professor, Head, Instituto de Matemática e Estatística
Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil

1. INTRODUCTION

Significance testing of precise (or sharp) hypotheses is an old and controversial problem: it has been central in statistical inference. Both frequentist and Bayesian schools of inference have presented solutions to this problem, not always prioritizing the consideration of fundamental issues such as the meaning of precise hypotheses or the inferential rationale for testing them. The Full Bayesian Significance Test, FBST, is an alternative solution to the problem, which attempts to ease some of the questions met by frequentist and standard Bayes tests based on Bayes factors. FBST was introduced by Pereira and Stern (1999) and reviewed by Pereira, Stern and Wechsler (2008).

The discussion here is restricted to univariate parameter and (sufficient statistic) sample spaces;

$$\Theta \subset \mathcal{R} \text{ and } X \subset \mathcal{R}$$

A sharp hypothesis H is then a statement of the form $H : \theta = \theta_0$ where $\theta_0 \in \Theta$. The posterior probability (density) for θ is obtained after the observation of $x \in X$. While a frequentist looks for the set, C, of sample points at least as inconsistent with θ_0 as x is, a Bayesian could look for the tangential set T of parameter points that are more consistent with x than θ_0 is. This understanding can be interpreted as a partial duality between sampling and Bayesian theories. The evidence in favor of H is for frequentists the usual p-value, while for Bayesian it should be $ev = 1 - \underline{ev}$:

$$pv = Pr\{x \in C|\theta_0\} \text{ and } ev = 1 - \underline{ev} = 1 - Pr\{\theta \in T|x\}.$$

The larger pv and ev, the stronger the evidence favoring H.

[1] Dr Carlos Pereira is a Professor and Head, Department of Statistics, University of São Paulo, Brazil. He is Past President of the Brazilian Statistical Society (1998–1990). He was the Director of the Institute of Mathematic and Statistics. São Paulo, Brazil (1994–1998). He was also Director of the Bioinformatic Scientific Center, University of SoPaulo (2006–2009). He is an Elected member of the International Statistical Institute. He has authored and co-authored more than 150 papers and 4 books, including *Bayesian Analysis* (in Portuguese) in 1982 – the first Bayesian book published in Latin America. Professor Pereira has received the Ralph Bradley award from Florida State University in 1981. He was a research engineer at IEOR in Berkeley at the University of California (1986–1988). He was Associate editor of *Entropy*, *Environmetrics*, and *Brazilian J of Probability and Statistics*. Currently, he is the Statistical editor of the *Brazilian J of Psychiatry*. He was a member of both the Envirometrics Society and Board of Directors of *Entropy*.

In the general case, the posterior distribution is sufficient for *ev* to be calculated, without any complication due to dimensionality of neither the parameter nor of the sample space. This feature ceases the need for nuisance parameters elimination, a problem that disturbs some statisticians (Basu, 1977). If one feels that the goal of measuring consistency between data and a null hypothesis should not involve prior opinion about the parameter, the normalized likelihood, if available, may replace the posterior distribution. The computation of *ev* needs no asymptotic methods, although numerical optimization and integration may be needed.

The fact that the frequentist and Bayesian measures of evidence, *pv* and *ev*, are probability values – therefore defined in a zero to one scale – does not easily help to answer the question "How small is *significant*?". For *p*-values, the NP lemma settles the question by means of subjective arbitration of critical values. For Bayesian assessment of significance through evaluation of *ev*, decision theory again clears the picture. Madruga et al. (2001) show that there exist loss functions the minimization of which render a test of significance based on *ev* into a formal Bayes test.

The FBST has successfully solved several relevant problems of statistical inference: see Pereira, Stern and Wechsler (2008) for a list of publications.

2. FBST DEFINITION

Significance FBST was created under the assumption that a significance test of a sharp hypothesis had to be performed. At this point, a formal definition of a sharp hypothesis is presented.

Consider general statistical spaces, where $\Theta \subset \mathcal{R}^m$ is the parameter space and $X \subset \mathcal{R}^k$ is the sample space.

Definition 1. A ***sharp*** hypothesis H states that θ belongs to a submanifold Θ_H of smaller dimension than Θ.

The subset Θ_H has null Lebesgue measure whenever H is sharp. A probability density on the parameter space is an ordering system, notwithstanding having every point probability zero. In the FBST construction, all sets of same nature are treated accordingly in the same way. As a consequence, the sets that define sharp hypotheses keep having nil probabilities. As opposed to changing the nature of H by assigning positive probability to it, the tangential set T of points, having posterior density values higher than any θ in Θ_H, is considered. H is rejected if the posterior probability of T is *large*. The formalization of these ideas is presented below.

Let us consider a standard parametric statistical model; i.e., for an integer m, the parameter is $\theta \in \Theta \subset \mathcal{R}^m$, $g(\bullet)$ a probability prior density over Θ, x is the observation (a scalar or a vector), and $L_x(\bullet)$ is the likelihood generated by data x. Posterior to the observation of x, the sole relevant entity for the evaluation of the Bayesian evidence *ev* is the posterior probability density for θ given x, denoted by

$$g_x(\theta) = g(\theta|x) \propto g(\theta)L_x(\theta).$$

Of course, one is restricted to the case where the posterior probability distribution over Θ is absolutely continuous; i.e., $g_x(\theta)$ is a density over Θ. For simplicity, H is used for Θ_H in the sequel.

DEFINITION 2. (EVIDENCE)**:** Consider a sharp hypothesis $H : \theta \in \Theta_H$ and

$$g^* = \sup_H g_x(\theta) \text{ and } T = \{\theta \in \Theta : g_x(\theta) > g^*\}.$$

The ***Bayesian evidence value against*** H is defined as the posterior probability of the tangential set, i.e.,

$$\underline{ev} = Pr\{\theta \in T|x\} = \int_T g_x(\theta)d\theta.$$

One must note that the evidence value supporting H, $ev = 1 - \underline{ev}$, is not an evidence against A, the alternative hypothesis (which is not sharp anyway). Equivalently, $\underline{ev}$ is not evidence in favor of A, although it is against H.

DEFINITION 3. (TEST)**:** The ***FBST*** (Full Bayesian Significance Test) is the procedure that rejects H whenever $ev = 1 - \underline{ev}$ is small.

The following example illustrates the use of the FBST and two standard tests, McNemar and Jeffreys' Bayes Factor. Irony et al. (2000) discuss this inference problem introduced by McNemar (1955).

Example 1. McNemar vs FBST.

Two professors, Ed and Joe, from the Department of Dentistry evaluated the skills of 224 students in dental fillings preparation. Each student was evaluated by both professors. The evaluation result could be approval (A) or disapproval (F). The Department wants to check whether the professors are equally exigent. Table 1 presents the data.

Table 1. Results of the evaluation of 224 students.

	Joe		
Ed	A	F	Total
A	62	41	103
F	25	96	121
Total	87	137	224

This is a four-fold classification with probabilities p_{11}, p_{12}, p_{21}, and p_{22}. Using standard notation, the hypothesis to be tested is $H : p_{1\cdot} = p_{\cdot 1}$ which is equivalent to $H : p_{12} = p_{21}$ (against $A : p_{12} \neq p_{21}$). In order to have the likelihood function readily available, we will consider a uniform prior, i.e., a *Dirichlet* density with parameter $(1, 1, 1, 1)$.

The McNemar exact significance for this data set is $pv = .064$. Recall that this test is based in a partial likelihood function, a binomial with $p = p_{12}(p_{12} + p_{21})^{-1}$

and $n = 66$. With the normal approximation, the *pv* become .049 with the partial likelihood used by McNemar, the FBST evidence is $ev = .045$. The value of the Bayes Factor under the same uniform prior is $BF = .953$. If one assigns probability 1/2 to the sharp hypothesis H, its posterior probability attains $\pi = .488$. Hence, the posterior probability π barely differs from 1/2, the probability previously assigned to H, while *pv* and *ev* seem to be more conclusive against H. While, in the three dimension full model, $ev = 0.265$ may seem to be a not low value and the test cannot be performed without a criterion. In other words, a decision is not made until *ev* is compared to a "critical value". The derivation of such a criterion – resulting from the identification of the FBST as a genuine Bayes procedure – is the subject of Madruga et al. (2001).

The strong disagreement among the values of ev, pv, and BF seldom occurs in situations where Θ is a subset of the real line. The speculation is that this is related to the elimination of nuisance parameters: By conditioning in McNemar case and by marginalization in the Bayes Factor case. In higher dimension, elimination of nuisance parameters seems to be problematic, as pointed by Basu (1977).

3. FBST THEORY

From a theoretical perspective, on the other hand, it may be propounded that if the computation of *ev* is to have any inferential meaning, then it ought to proceed to a declaration of significance (or not). To this – in a sense – simultaneously NPW and Fisherian viewpoint can be opposed the identification of ev as an estimator of the indicator function $\phi = I(\theta \in \Theta_H)$. In fact, Madruga et al. (2001) show that there are loss functions the minimization of which makes *ev* a Bayes estimator of ϕ (see Hwang et al., 1992).

Madruga et al. (2001) prove that the FBST procedure is the posterior minimization of an expected loss λ defined as follows:

$$\lambda(\text{Rejection of } H, \theta) = a\{1 - I[\theta \in T]\} \text{ and}$$
$$\lambda(\text{Acceptance of } H, \theta) = b + dI[\theta \in T].$$

Here, a, b and d are positive real numbers. The operational FBST procedure is given by the criterion according to which H is to be rejected if, and only if, the evidence *ev* is smaller than $c = (b+d)/(a+d)$. One should notice that the evidence *ev* is the Bayesian formal test statistic and that positive probability for H is never required. A complete discussion of the above approach can be found in Pereira, Stern and Wechsler (2008).

4. FINAL REMARKS

The following list states several desirable properties attended by *ev* :

1. *ev* is a probability value derived from the posterior distribution on the full parameter space.

2. Both *ev* and FBST possesses versions which are invariant for alternative parameterizations.
3. The need of approximations in the computation of *ev* is restricted to numerical maximization and integration.
4. FBST does not violate the Likelihood Principle.
5. FBST neither requires nuisance parameters elimination nor the assignment of positive prior probabilities to sets of zero Lebesgue measure.
6. FBST is a formal Bayes test and therefore has critical values obtained from considered loss functions.
7. *ev* is a possibilistic support for sharp hypotheses, complying with the Onus Probandi juridical principle (In Dubio Pro Reo rule), Stern (2003).
8. Derived from the full posterior distribution, *ev* is a homogeneous computation calculus with the same two steps: constrained optimization and integration with the posterior density.
9. Computing time was not a great burden whenever FBST was used. The sophisticated numerical algorithms used could be considered a more serious obstacle to the popularization of the FBST.

ev was developed to be the Bayesian *pv* alternative, while maintaining the most desirable (known or perceived) properties in practical use. The list presented above seems to respond successfully to the challenge: the FBST is conceptually simple and elegant, theoretically coherent, and easily implemented for any statistical model, as long as the necessary computational procedures for numerical optimization and integration are available.

References

[1] Basu, D. (1977). On the elimination of nuisance parameters. *JASA*, **72**, 355-66.

[2] Hwang, J. T., G. Casella, C. Robert, M. T. Wells, R. G. Farrel (1992). Estimation of accuracy in testing. *Annals of Statistics,* **20**, 490–509.

[3] Irony, T. Z., C. A. de B. Pereira, R.C. Tiwari (2000). Analysis of Opinion Swing: Comparison of Two Correlated Proportions, American Statistician 54(1), 57–62.

[4] McNemar, Q. (1955). *Psychological Statistics.* Wiley, New York.

[5] Madruga, M. R., L. G. Esteves, S. Wechsler (2001). On the Bayesianity of Pereira-Stern tests. *Test,* **10**, 291–9.

[6] Pereira, C. A. de B., J. M. Stern (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy* **1**, 69–80.

[7] Pereira, C. A. B., J. M. Stern, S. Wechsler (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis,* **3**(1), 79–100.

[8] Stern, J. M. (2007). Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Cybernetics* & *Human Knowing,* **14**(1), 9–36.