Full Bayesian Significance Test for Coefficients of Variation

CARLOS ALBERTO de BRAGANCA PEREIRA
and JULIO MICHAEL STERN
University of Sao Paulo, Brasil

Abstract: New application of the Full Bayesian Significance Test (FBST) for precise hypotheses is presented. The FBST is an alternative to significance tests or, equivalently, to *p-values*. In the FBST we compute the evidence of the precise hypothesis. This evidence is the complement of the probability of a credible set "tangent" to the sub-manifold (of the parameter space) that defines the null hypothesis. We use the FBST to compare coefficients of variation, in applications arising in finance and industrial engineering.

Keywords: BAYES FACTOR; NUMERICAL INTEGRATION; GLOBAL OPTIMIZATION; P-VALUE; POSTERIOR DENSITY; FINANCE; INDUSTRIAL ENGINEERING.

1. INTRODUCTION

The Full Bayesian Significance Test (FBST) is presented in Pereira *et al.* (1999b) as a coherent Bayesian significance test. The FBST is intuitive and has a geometric interpretation. It can be easily implemented using modern numerical optimization and integration techniques. The method is "Full" Bayesian and consists in the analysis of credible sets. By Full we mean that we need only the knowledge of the parameter space represented by the posterior distribution, without the need for any adhockery, a term used by Good (1983), like a positive probability for the precise hypothesis, generating the Lindley's paradox effect. Another important aspect of the FBST is its consistency with the "benefit of doubt" juridical principle. These remarks will be understood in the sequel. Important issues concerning invariance and dimension reduction are addressed at the final remarks.

Significance tests, Cox (1977), are regarded as procedures for measuring the consistency of data with a null hypothesis by the calculation of a *p-value* (tail area under the null hypothesis). Previously defined Bayesian significance tests, like Bayes Factor or the posterior probability of the null hypothesis consider the *p-value* as a measure of evidence of the null hypothesis and present alternative Bayesian measures of evidence, Aitkin (1991), Berger *et al.* (1987) and (1997), Irony *et al.* (1986) and (1995), Pereira *et al.* (1993), Sellke *et al.* (1999). As pointed out in Cox (1977), the first difficulty in defining the *p-value* is the way the sample space is ordered under the null hypothesis. Pereira *et al.* (1993) suggested a *p-value* that always considers the alternative hypothesis. To each of these measures of evidence one could find a great number of counter arguments. The most important argument against Bayesian test for precise hypothesis is presented by Lindley (1957). There are many arguments in the literature against the classical *p-value*. The book by Royall (1997) and its review by Vieland *et al.* (1998) present interesting and relevant arguments to start statisticians thinking about new methods of measuring evidence. In a more philosophical terms, Good (1983) discuss, in a great detail, the concept of evidence.

2. THE EVIDENCE CALCULUS

Consider the random variable D that, when observed, produces the data d. The statistical space is represented by the triplet (Ξ, Δ, Θ) where Ξ is the sample space, the set of possible values of d, Δ is the family of measurable subsets of Ξ and Θ is the parameter space. We define now a prior model (Θ, B, π), which is a probability space defined over Θ. Note that this model has to be consistent, so that $Pr\{A \mid \theta\}$ turns out to be well defined. As usual after observing data d, we obtain the posterior probability model (Θ, B, π_d), where π_d is the conditional probability measure on *B* given the observed sample point, d. In this paper we restrict ourselves to the case where the function π_d has a probability density function, $f(\theta \mid d)$.

To define our procedure we should concentrate only on the posterior probability space (Θ, B, π_d). First we will define T_φ as the subset of the parameter space where the posterior density is greater than φ.

$$T_\varphi = \{\theta \in \Theta \mid f(\theta \mid d) > \varphi\}$$

The credibility of T_φ is its posterior probability,

$$\kappa = \int_{T_\varphi} f(\theta \mid d) d\theta = \int_\Theta f_\varphi(\theta \mid d) d\theta$$

where $f_\varphi(x) = f(x)$ if $f(x) > \varphi$ and zero otherwise.

Now, we define f^* as the maximum of the posterior density over the null hypothesis, attained at the argument θ^*,

$$\theta^* \in arg \max_{\theta \in \Theta_0} f(\theta \mid d) \ , \ \ f^* = f(\theta^* \mid d)$$

and define $T^* = T_{f^*}$ as the set "tangent" to the null hypothesis, H, whose credibility is κ^*.

The measure of evidence we propose in this article is the complement of the probability of the set T^*. That is, the evidence of the null hypothesis is

$$Ev(H) = 1 - \kappa^* \ \text{ or } \ 1 - \pi_d(T^*)$$

If the probability of the set T^* is "large", it means that the null set is in a region of low probability and the evidence in the data is against the null hypothesis. On the other hand, if the probability of T^* is "small", then the null set is in a region of high probability and the evidence in the data is in favor of the null hypothesis.

Although the definition of evidence above is quite general, it was created with the objective of testing precise hypotheses. That is, a null hypothesis for which the dimension is smaller than that of the parameter space, i.e. $dim(\Theta_0) < dim(\Theta)$.

3. NUMERICAL COMPUTATION

In this paper the parameter space, Θ, is always a subset of R^n, and the hypothesis is defined as a further restricted subset $\Theta_0 \subset \Theta \subseteq R^n$. Usually, Θ_0 is defined by vector valued inequality and equality constraints:

$$\Theta_0 = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}.$$

Since we are working with precise hypotheses, we have at least one equality constraint, hence $dim(\Theta_0) < dim(\Theta)$.

The computation of the evidence measure defined in the last section is performed in two steps, a numerical optimization step, and a numerical integration step. At the numerical steps, we operate with the posterior probability density function, $f(\theta \mid d)$, representing the state of uncertainty at the moment of evaluation. Hence, after the (posterior) Bayesian update process, the sample d plays no further role. Computationally, the posterior is optimized and integrating over the parameter space, whereas the observation data, d, or a sufficient statistic of it, is a constant list. To emphasize this point we write $f(\theta \mid d)$ as $f(\theta)$.

The numerical optimization step consists of finding an argument θ^* that maximizes the posterior density $f(\theta)$ under the null hypothesis. The numerical integration step consists of integrating the posterior density over the region where it is greater than $f(\theta^*)$. That is,

- Numerical Optimization step:

$$\theta^* \in arg \max_{\theta \in \Theta_0} f(\theta) \ , \ \ \varphi = f^* = f(\theta^*)$$

- Numerical Integration step:

$$\kappa^* = \int_\Theta f_\varphi(\theta) d\theta$$

where $f_\varphi(x) = f(x)$ if $f(x) > \varphi$ and zero otherwise.

Efficient computational algorithms are available for local and global optimization as well as for numerical integration in Fletcher (1987), Horst *et al.* (1995), Pinter (1996), Krommer *et al.* (1998), Nemhauser *et al.* (1989), and Sloan *et al.* (1994). Computer codes for several such algorithms can be found at software libraries such as NAG and ACM, or at internet sites as *www.ornl.org.*

We notice that the method used to obtain T^* and to calculate κ^* can be used under general conditions. Our purpose, however, is to discuss precise hypothesis testing, under absolute continuity of the posterior probability model, the case for which most solutions presented in the literature are controversial.

4. COEFFICIENT OF VARIATION APPLICATIONS

The Coefficient of Variation (CV) of a random variable X is defined as the ratio $CV(X) = \sigma(X)/E(X)$, i.e. the ratio of its standard deviation to its mean. We want to test the hypothesis that two normal random variables, with unknown mean and variance, have the same CV:

$$X^1 \sim N(m_1, \sigma_1) \ , \ X^2 \sim N(m_2, \sigma_2) \ , \quad H_0 : \ \sigma_1/m_1 = \sigma_2/m_2$$

We were faced with the necessity of testing this hypothesis in two occasions. The first application involves the calibration of a sensor device. The manufacturer claims that, within a certain range, the sensor readings have a constant CV. A proposed calibration procedure makes use of this hypothesis, so it is necessary to test it first. The second application concerns the certification of an automatic trading software. The software matches buy and sell orders at a proposed virtual stock exchange. The software is supposed to be “fair” in the sense of giving equal treatment to small and large orders, more precisely, the CV of orders’ execution price should be constant. We had full access neither to the sensor project nor the trading software source code, due to pending patents, industrial secrets, and security reasons. We could however get some experimental readings, in the first case, and have some information on executed “probe orders”, in the second case. Both cases imply testing the hypothesis stated above.

4.1 *Conjugate Family*

It can be shown that the conjugate family for this problem is family of bivariate distributions, where the conditional distribution of the mean m, for a fixed precision $r = 1/\sigma^2$, is normal, and the marginal distribution of the precision r is gamma, DeGroot (1970), Lindley (1978). Using the standard improper priors, uniform on $]-\infty, +\infty[$ for m, and $1/r$ on $]0, +\infty[$ for r, we get the posterior joint distribution for m and r:

$$f(m, r \,|\, x) \propto \sqrt{r} exp(-nr(m - \bar{x})^2/2) exp(-br) r^{a-1}$$

$$x = [x_1 \ldots x_n] \ , \ a = \frac{n-1}{2} \ , \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ b = \frac{n}{2}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Given two samples, $x_1 = [x_{1,1} \ldots x_{1,n_1}]$ and, $x_2 = [x_{2,1} \ldots x_{2,n_2}]$ it is better, for numerical stability, to optimize the log-likelihood,

$$\begin{aligned} &fl(m_1, r_1, m_2, r_2 \,|\, n_1, \bar{x}_1, b_1, n_2, \bar{x}_2, b_2) = \\ &\qquad (n_1/2 - 1) log(r_1) - b_1 r_1 - (n_1 r_1/2)(m_1 - \bar{x}_1)^2 \\ &\qquad + (n_2/2 - 1) log(r_2) - b_2 r_2 - (n_2 r_2/2)(m_2 - \bar{x}_2)^2 \end{aligned}$$

the hypothesis being represented by the constraint

$$g(m_1, r_1, m_2, r_2) = m_1^2 r_1 - m_2^2 r_2 = 0$$

The gradients of $fl(\ )$ and $g(\ )$ have easy analytical expressions, that can be given to the optimizer:

$$dfl \ = \ [\, -n_1 r_1 (m_1 - \bar{x}_1) \ , \ (n_1/2 - 1)/r_1 - b_1 - (n_1/2)(m_1 - \bar{x}_1)^2 \ ,$$

$$
\begin{aligned}
& - n_2 r_2 (m_2 - \bar{x}_2) \ , \ (n_2/2 - 1)/r_2 - b_2 - (n_2/2)(m_2 - \bar{x}_2)^2 \ ] \\
dg \quad & = \quad [2 m_1 r_1 \ , \ m_1^2 \ , \ -2 m_2 r_2 \ , \ -m_2^2]
\end{aligned}
$$

4.2 *Numerical Examples*

Figure 1 illustrates the FBST construction in a simpler case, namely, testing the hypothesis that coefficient of variation is equal to a given constant, $\mathrm{H} : \mathrm{CV} = c$. We plot some level curves of the posterior density function, including the level curve tangent to the hypothesis manifold. At the tangency point, θ^*, the posterior density attains its maximum, f^*, on the hypothesis. The interior of the tangent level curve, T^*, includes all points with posterior density greater than f^*, i.e. it is the highest probability density set tangent to the hypothesis. In Figure 1 we give the FBST evidence, $\mathrm{Ev}(H)$, when testing $c = 0.1$ with 3 samples of size $n = 16$, mean $\bar{x} = 10$ and standard deviations $s = 1.0, \ 1.1, $ and 1.5.

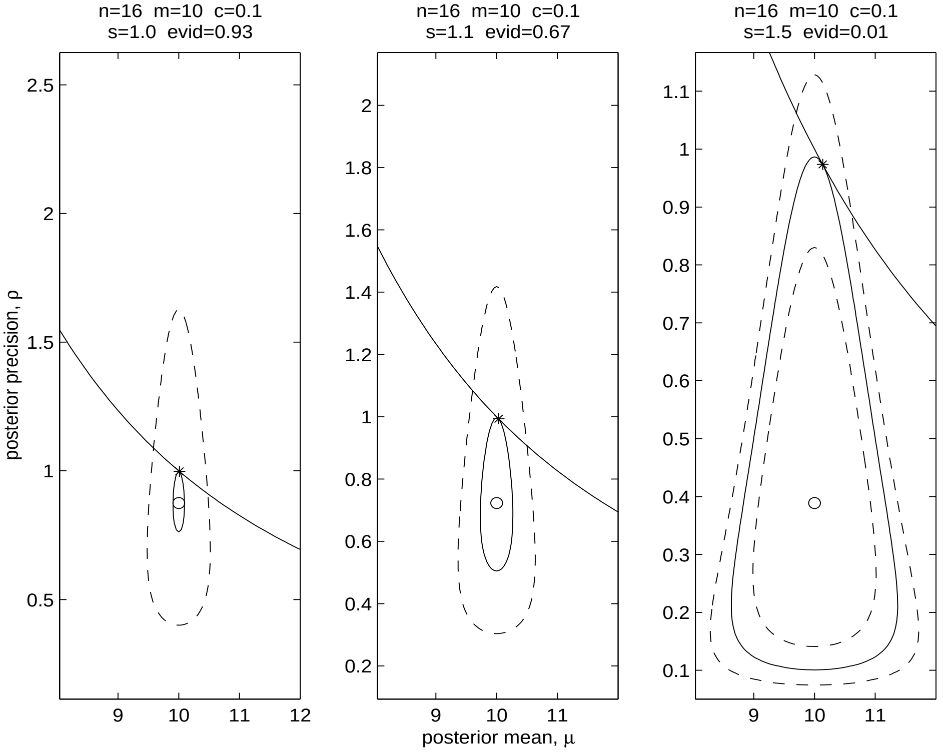

Figure 1. *FBST for H: CV=0.1*

The next two examples in this section come from the applications mentioned at the beginning of this section. Let $[x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}]$ and $[x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}]$ be samples of two normal distributed data. A sufficient statistic for the normal distribution is the vector $[n, \bar{x}, s]$, i.e. the sample size, the observations mean and standard deviation, where $s^2 = (1/n) \sum (x_i - \bar{x})^2$. We always have $n > 2$, so the posterior likelihood function is integrable on the parameter space.

Tables 1 and 2 present sensor readings and value of executed orders, two different samples in each case. The tables also present the FBST evidence for the constant CV hypothesis. The first example favors the null hypothesis, whereas the second is against it.

Table 1. *Sensor readings*

Sample 1				
9.43	9.81	10.60	10.84	9.28
9.28	9.80	9.98	10.28	11.06
mean$_1$ $= 10.03$			std$_1$ $= 0.61$	
Sample 2				
21.24	16.50	21.39	21.62	21.27
22.62	19.65	-	-	-
mean$_2$ $= 20.62$			std$_2$ $= 1.86$	
Evidence				
$CV_1/CV_2 = 0.66$			$Ev(H) = 0.98$	

Table 2. *Order values*

Sample 1				
8.65	9.70	6.10	10.95	10.23
8.82	8.69	7.84	9.90	10.76
10.31	10.10	9.87	10.66	10.90
mean$_1$ $= 9.57$			std$_1$ $= 1.31$	
Sample 2				
19.34	19.00	19.93	19.65	18.09
22.59	20.88	22.56	19.00	18.76
mean$_2$ $= 19.98$			std$_2$ $= 1.48$	
Evidence				
$CV_1/CV_2 = 1.86$			$Ev(H) = 0.10$	

5. FINAL REMARKS

The theory presented in this paper, grew out of the necessity of the authors' activities in the role of audit, control or certification agents, Pereira and Stern (1999a). These activities made the authors (sometimes painfully) aware of the benefit of doubt juridical principle, or safe harbor liability rule. This kind of principle establishes that there is no liability as long as there is a reasonable basis for belief, effectively placing the burden of proof on the plaintiff, who, in a lawsuit, must prove false a defendant's misstatement. Such a rule also prevents the plaintiff from making any assumption not explicitly stated by the defendant, or tacitly implied by existing law or regulation. The use of an a priori point mass on the null hypothesis, as on standard Bayesian tests, can be regarded as such an ad hoc assumption.

As audit, control or certification agents, the authors had to check compliance with given requirements and specifications, formulated as precise hypotheses on contingency tables. In Pereira *et al.* (1999b) we describe several applications based on contingency tables, comparing the use of FBST with standard Bayesian and Classical tests. The applications presented in this paper are very similar in spirit, but we are not aware of any standard exact test in the literature. Some tests for simpler hypothesis on the CV are given in Lehmann (1986). The analytical derivation of these tests is quite sophisticated, whereas the implementation of FBST is immediate and trivial, as long as good numerical optimization and integration programs are at hand. In the applications in this paper, as well in those in Pereira *et al.* (1999b), it is desirable or necessary to use a test with the following characteristics:

- Be formulated directly in the original parameter space.
- Take into account the full geometry of the null hypothesis as a manifold (surface) imbedded in the whole parameter space.
- Have an intrinsically geometric definition, independent of any non-geometric aspect, like the particular parameterization of the (manifold representing the) null hypothesis being used.
- Be consistent with the benefit of doubt juridical principle (or safe harbor liability rule), i.e. consider in the "most favorable way" the claim stated by the hypothesis.
- Consider only the observed sample, allowing no ad hoc artifice (that could lead to judicial contention), like a positive prior probability distribution on the precise hypothesis.
- Consider the alternative hypothesis in equal standing with the null hypothesis, in the sense that increasing sample size should make the test converge to the right (accept/reject) decision.
- Give an intuitive and simple measure of significance for the null hypothesis, ideally, a probability in the parameter space.

FBST has all these theoretical characteristics, and straightforward (computational) implementation. Moreover, as shown in Madruga *et al.* (2000), the FBST is also in perfect harmony with the Bayesian decision theory of Rubin (1987), in the sense that there are specific loss functions which render the FBST.

Finally, we remark that the evidence calculus defining the FBST takes place entirely in the parameter space where the prior was assessed by the scientist, Lindley (1983). We call it the "original" parameter space, although acknowledging that the parameterization choice for the statistical model semantics is somewhat arbitrary. We also acknowledge that the FBST is not invariant under general change of parameterization.

The FBST is in sharp contrast with the traditional schemes for dimensional reduction, like the elimination of so called "nuisance" parameters. In these "reduced" models the hypothesis is projected into a single point, greatly simplifying several procedures. Interesting new work in this approach can be found in Evans (1997). However, problems with the traditional approach are presented in Pereira and Lindley (1987). The traditional reduction or projection schemes are also incompatible with the benefit of doubt principle, as stated earlier. In fact, preserving the original parameter space, in its full dimension, is the key for the intrinsic regularization mechanism of the FBST, when it is used in the context of model selection, Pereira and Stern (2000 a and b).

Of course, there is a price to be paid for working with the original parameter space, in its full dimension: A considerable computational work load. But computational difficulties can be overcome with the use of efficient continuos optimization and numerical integration algorithms. Large problems can also benefit from program vectorization and parallelization techniques. Dedicated vectorized or parallel machines may be expensive and not always available, but most of the algorithms needed can benefit from asynchronous and coarse grain parallelism, a resource easily available, although rarely used, on any PC or workstation network through MPI, Portable Parallel Programming Message-Passing Interface, or similar distributed processing environments, Wilson and Lu (1996).

REFERENCES

Aitkin, M. (1991). Posterior Bayes Factors. *J R Statist Soc B* 1, 111–142.

Berger, J.O. and M Delampady, M. (1987). Testing precise hypothesis. *Statistical Science* 3, 315–352.

Berger, J.O. Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science* 3, 315–352.

Cox, D.R. (1977). The role of significance tests. *Scand J Statist* 4, 49–70

DeGroot, M.H. (1970). *Optimal Statistical Decisions*. NY: McGraw-Hill.

Evans, M. (1997). Bayesian Inference Procedures Derived via the Concept of Relative Surprise. *Communications in Statistics* 26, 1125–1143.

Fletcher, R. (1987). *Practical Methods of Optimization*. Essex: J Wiley.

Good, I.J. (1983) *Good thinking: The foundations of probability and its applications*. University of Minnesota Press.

Gropp, W.D. Lusk, E. and Skjellum, A. (1994). *Using MPI: Portable Parallel Programming with Message-Passing Interface*. MIT Press.

Horst, R., Pardalos, P.M. and Thoai, N.V. (1995). *Introduction to Global Optimization.* Boston: Kluwer.

Irony, T.Z. and Pereira, C.A.B. (1986). Exact test for equality of two proportions: Fisher×Bayes. *J Statist Comp & Simulation* 25, 93–114.

Irony, T.Z. and Pereira, C.A.B. (1986). Bayesian Hypothesis Test: Using surface integrals to distribute prior information among hypotheses. *Resenhas* 2, 27–46.

Krommer, A.R. and Ueberhuber, C.W. (1998). *Computational Integration*. Philadelphia: SIAM.

Lehmann, E.L. (1986). *Testing Statistical Hypothesis*. NY: Wiley.

Lindley, D.V. (1957). A Statistical Paradox. *Biometrika* 44, 187–192.

Lindley, D.V. (1978). The Bayesian Approach. *Scand J Statist* 5, 1-26.

Lindley, D.V. (1983). *Lectures on Bayesian Satistics*, Univ. of São Paulo.

Marshall, A. and Prochan, F. (1972). Classes of Distributions Applicable in Replacement, with Renewal Theory Implications. *Proc. 6th Berkeley Symp. Math. Statist. Prob.* 395–415.

Madruga, M.R. Esteves, L.G. and Wechsler, S. (2000). On the Bayesianity of Pereira-Stern Tests, *TEST*, to appear.

Nemhauser, G.R. Rinnooy-Kan, A.H.G. and Todd, M.J. editors (1989). *Optimization, Handbooks in Operations Research Vol 1*. Amsterdam: North-Holland.

Pereira, C.A.B. and Lindley, D.V. (1987). Examples Questioning the Use of Partial Likelihood. *The Statistician* 36, 15–20.

Pereira, C.A.B. and Wechsler, S. (1993). On the Concept of *p-value*. *Braz J Prob Statist* 7, 159–177.

Pereira, C.A.B. Stern, J.M. (1999a). A Dynamic Software Certification and Verification Procedure. *Proc. ISAS-99 - International Conference on Information Systems Analysis and Synthesis* II, 426–435.

Pereira, C.A.B. Stern, J.M. (1999b). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy* 1, 69–80.

Pereira, C.A.B. Stern, J.M. (2000a). *Intrinsic Regularization in Model Selection using the Full Bayesian Significance Test*. Technical Report RT-MAC-2000-6, Dept. of Computer Science, University of Sao Paulo.

Pereira, C.A.B. Stern, J.M. (2000b). Model Selection: Full Bayesian Approach, *Environmetrics*, to appear.

Pinter, J.D. (1996). *Global Optimization in Action. Continuos and Lipschitz Optimization: Algorithms, Implementations and Applications*. Boston: Kluwer.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Rubin, H. (1987). A Weak System of Axioms for “Rational” Behaviour and the Non-Separability of Utility from Prior. *Statistics and Decisions* 5, 47–58.

Sellke, T. Bayarri, M.J. and Berger, J. (1999). Calibration of p-values for Testing Precise Null Hypotheses. *ISDS Discussion Paper 99-13*.

Sloan, I.R. and Joe, S. (1994). *Latice Methods for Multiple Integration*. Oxford: Oxford University Press.

Vieland, V.J. and Hodge, S.E. (1998). Book Reviews: *Statistical Evidence* by R Royall (1997). *Am J Hum Genet* 63, 283–289.

Wilson, G.V. and Lu, P. (1996). *Parallel Programming Using C++*. Cambridge: MIT Press.