Bayesian Analysis (2008) **3**, Number 1, pp. 79–100

Can a Significance Test Be Genuinely Bayesian?

Carlos A de B Pereira* , Julio Michael Stern† and Sergio Wechsler‡

Abstract. The Full Bayesian Significance Test, FBST, is extensively reviewed. Its test statistic, a genuine Bayesian measure of evidence, is discussed in detail. Its behavior in some problems of statistical inference like testing for independence in contingency tables is discussed.

Keywords: Bayes tests, Sharp hypotheses, Significance tests.
[62A15, 62C10, 62F03, 62F15]

1 Introduction

The present article deals with an old and controversial problem which has been central in statistical inference: significance testing of precise (or sharp) hypotheses. 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. We present and discuss another solution to the problem, the Full Bayesian Significance Test (FBST), which attempts to ease some of the questions met by frequentist and standard Bayes tests.

According to Cox (1977) and Kempthorne (1976) a significance test is a procedure for measuring the consistency of data with a null hypothesis. The basis of this old understanding of significance is an ordering of the sample space according to increasing inconsistency with the hypothesis. This goal of measuring consistency seems *prima facie* amenable to a Bayesian reading. In both, frequentist and Bayesian settings, consistency of data and parameter values is to be measured.

For the moment, let us restrict the discussion, as in Cox (1977), to univariate parameter and (sufficient statistic) sample spaces,

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

A sharp hypothesis H is 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 may 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 value in favor of H is for frequentists the usual p-value, $pv = \Pr\{x \in$

*Institute of Mathematics and Statistics, U São Paulo, Brazil, mailto:cpereira@ime.usp.br
†Institute of Mathematics and Statistics, U São Paulo, Brazil, mailto:jstern@ime.usp.br
‡Institute of Mathematics and Statistics, U São Paulo, Brazil, mailto:sw@ime.usp.br

$C \mid \theta_0\}$, while for Bayesians it should be $ev = 1 - \overline{ev}$, where $\overline{ev} = \Pr\{\theta \in T \mid x\}$. The smaller pv and ev, the stronger the evidence against H.

We point out that, in the general case, the posterior distribution is sufficient for ev to be calculated, without any complication due to the dimensionality of neither the parameter nor of the sample space. This feature avoids the need for eliminating nuisance parameters, a problem that disturbs some statisticians, see 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, other than numerical optimization and integration.

While the measure of evidence ev may be derived from this duality program, it has been realized, on the other hand, that it is not just a mere Bayesian counterpart of the ubiquitous pv. Instead, the inferential use of ev is a genuine Bayesian procedure. It is in fact a well-defined posterior probability of a subset of the parameter space. Hence, its use does not violate the paramount Likelihood Principle, see Basu (1975) and Birnbaum (1962). Furthermore, as the full parameter space is used in the computation of ev, the alternative hypothesis is always intrinsically considered. As pointed out by Pereira and Wechsler (1993), ordinary significance testing sometimes disregards the alternative hypothesis, troubling in this way the inference.

The above paragraph brings the fundamental Neyman-Pearson (NP) lemma into discussion. 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 help to answer the question "How small is significant?". For p-values, the NP lemma settles the question by means of subjective arbitration of bounds on the probability of first-kind errors, which is formally equivalent to introducing loss functions. 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's test.

The FBST possesses not only this Bayesian decision-theoretic quality but also complies with the time-honored Onus Probandi juridical principle (or *In Dubio Pro Reo* rule). In addition, the FBST satisfies logical requirements met by neither p-values nor Bayes Factors based tests (Stern (2003)).

The Bayes's significance test based on ev - the FBST - does not demand the adoption of a prior distribution which assigns positive probability for the subset that defines the sharp null hypothesis. This is a most relevant coherence feature of the FBST over Bayes Factor tests for sharp null hypotheses. Let us recall that Bayesian inference has long replaced p-values by Bayes Factors, see Jeffreys (1939). However, whenever the posterior is absolutely continuous and the null hypothesis sharp, the use of Bayes Factors for significance testing is controversial, as discussed by many authors, standing out Good (1983), Lindley (1957), Lindley (1997), and Shafer (1982). In addition, there are recommendations for Bayes Factor bounds in order to define decision rules (Kass and Raftery (1995)). However, as in the case of p-values, this seems to be rather arbitrary.

The FBST has been successfully applied to several relevant problems of statistical inference, such as: testing for homogeneity and independence in contingency tables; comparison of coefficients of variation; the multivariate Behrens-Fisher problem; Hardy-Weinberg equilibrium testing; variable selection; testing for independence in the Holgate (bivariate Poisson) distribution; mixture models; Weibull wear-out testing, see Irony et al. (2002), Lauretto et al. (2003), Madruga et al. (2003), Pereira and Stern (1999), Pereira and Stern (2001a), Rodrigues (2006) and Stern and Zacks (2002).

The FBST is presented formally in Section 2. Section 3 discusses the Neyman-Pearson lemma and its influence on the building of the Bayes Factor environment. Section 4 presents the Decision-Theoretic description as well as the invariant version of the FBST. In Section 5 FBST asymptotic properties are presented. Section 6 has many illustrations to motivate the reader. Section 7 lists the important properties of the FBST as a summary of the discussion presented in the previous sections.

2 FBST Definition

The original version of the FBST was introduced by Pereira and Stern (1999). It was created under the assumption that a significance test of a sharp hypothesis had to be performed. Testing sharp hypotheses is of course a rich matter of discussion and controversies. The different viewpoints go from the blunt refusal to test a hypothesis having (posterior as well) probability zero to the assignment of mass probability to it. We will return to this discussion at the final section. At this point we present a formal definition of a sharp hypothesis.

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

Definition 2.1. A **sharp** hypothesis H states that θ belongs to a sub-manifold Θ_H of smaller dimension than Θ.

The subset Θ_H then has null Lebesgue measure whenever H is sharp. A probability density on the parameter space also is an ordering system, notwithstanding giving every point probability zero. In the FBST construction, all sets of the same nature are treated accordingly in the same way. As a consequence, the sets that define sharp hypotheses keep having nil probabilities. Instead of changing the nature of H by assigning positive probability to it, we will look for the tangential set, T, of points having posterior density values higher than any in Θ_H. We then do not reject H if the posterior probability of T is small. We will formalize these ideas in the sequel.

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

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

We are of course restricted to the case where the posterior probability distribution over Θ is absolutely continuous, that is, $g_x(\theta)$ is a density over Θ. For simplicity we use H for Θ_H in sequel.

Definition 2.2 (Evidence). Consider a sharp hypothesis $H : \theta \in \Theta_H$ and let

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

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

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

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

Definition 2.3 (Test). FBST (Full Bayesian Significance Test) is the procedure that rejects H whenever ev is small.

The first 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 (1947).

Example 2.1 (McNemar). 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.

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

Table 1: Evaluation Results for McNemar

We have a four-fold classification with probabilities $p_{1,1}$, $p_{1,2}$, $p_{2,1}$ and $p_{2,2}$. Using standard notation, the hypothesis to be tested is $H : p_{1,\bullet} = p_{\bullet,1}$, which is equivalent to $H : p_{1,2} = p_{2,1}$ (against $A : p_{1,2} \neq p_{2,1}$). In order to have the likelihood function readily available, we will consider a uniform prior, i.e., a Dirichlet with parameter $[1, 1, 1, 1]$.

The first step to compute the value of ev is to obtain the point p^*, satisfying H, which maximizes the posterior density. That is,

$$p^* = (1/224)[62, 33, 33, 96] \text{ and } \widehat{p} = (1/224)[62, 41, 25, 96],$$

where $\widehat{p}$ is the posterior overall mode. Evaluating the (likelihood) posterior density at these maxima we find

$$g^* = g_x(p^*) = 622 \text{ and } \widehat{g} = g_x(\widehat{p}) = 4409.$$

A likelihood-oriented statistician would not be reluctant to reject H, since the ratio $L_x(p^*)/L_x(\widehat{p}) = 0.14$. The set T is of course defined as

$$T = \{p : g_x(p) > 622\}.$$

Note that T is a subset of

$$\Theta = \{p = (p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2}) : p_{1,1} + p_{1,2} + p_{2,1} + p_{2,2} = 1 \text{ and } p_{i,j} > 0\},$$

the standard simplex in four dimensions.

Finally, numerical integration yields $ev = 0.11$. The McNemar p-value for this data set is equal to 0.064. The value of the Bayes Factor under the same uniform prior is $BF = 0.95$. If one assigns probability $1/2$ to the sharp hypothesis H, its posterior probability attains $PP = 0.49$. Hence, the posterior probability, PP, barely differs from $1/2$, the probability previously assigned to H, while pv and ev seem to be more conclusive against H. While ev $= 0.11$ may seem to be a low value, the test can not be performed without a criterion. In other words, a decision is not made until ev is compared to a "critical value".

A strong disagreement among ev, pv and BF seldom occurs in situations where Θ is a subset of the real line. We suspect that this is related to the absence of nuisance parameters. In higher dimensions, elimination of nuisance parameters may become problematic as pointed by Basu (1977).

3 Brief Outline of Sharp Hypothesis Testing

We believe that Fisher (Fisher (1922), Fisher (1934)) was the introducer of modern tests of significance. However the imprecise form of judgment of an observed $p-value$ in favor or against a null hypothesis, H, led Neyman and Pearson (1936) and Wald (1939), Wald (1950), to create the Theory of Testing Statistical Hypotheses. In fact this was conceived with the goal of having an objective and precise decision-theoretical approach in lieu of the imprecise way the conclusions based on $p-values$ are taken still today. Their achievement - the celebrated Neyman-Pearson (NP) Lemma - is better appreciated in a more general version, which is also much closer to the optimization ideas of Abraham Wald. In the sequel, we state this version DeGroot (1975) naming it the NPW Lemma. The usual NP Lemma is in fact a corollary of this result.

Suppose that the probability density for the experiment being performed (x being the observation) is one of only two options, f_H or f_A. The null hypothesis, $H : f_H$, states that f_H is the true density while the alternative, $A : f_A$, states that f_A is the

true density. The Likelihood Ratio statistic LR is defined as:

$$LR = f_H(x)/f_A(x)$$

For any positive real number c, define the test τ_c as the binary function satisfying:

- $\tau_c(x) = 1$, i.e. accept H, if $LR \geq c$ $(or\ \ > c)$ and
- $\tau_c(x) = 0$, i.e. accept A, if $LR < c$ $(or\ \ \leq c)$.

We note that equality must hold in only one of the cases.

Lemma 3.1. (NPW lemma) For any other test δ, if α and β represent the first and second kind errors, we have that

$$\alpha(\delta) + c\beta(\delta) \geq \alpha(\tau_c) + c\beta(\tau_c).$$

The proof of this lemma is straightforward, as shown in DeGroot (1975). The following result, named after Wald, characterizes the optimal procedures τ_c.

Lemma 3.2. (W lemma) τ_c defined above is a Bayes rule.

By Bayes rule we understand a statistical rule that minimizes the risk function associated to a properly defined loss function. The proof of this Lemma is also straightforward. Let us consider a prior probability $\pi = \Pr(H) = 1 - \Pr(A)$ and λ_H and λ_A the losses associated to H and A. Taking c as the product of prior odds and loss ratio,

$$c = \frac{1-\pi}{\pi}\,\frac{\lambda_A}{\lambda_H}\ ,$$

we easily obtain the result by minimizing the risk or the expected overall loss, see DeGroot (1975). In the other direction, for a given c, there exist positive constants λ_H, λ_A and $\pi(< 1)$, satisfying the above equation.

Statisticians expected to obtain similar results for composite hypotheses. However, frequentist and Bayesian statisticians had to follow different paths to generalize the NPW lemma to composite hypotheses. Bayesians, in the case of non-sharp hypotheses, could define f_H and f_A by averaging the likelihood over the two hypotheses sets to obtain the Bayes Factor, BF. Frequentists, on the other hand, were able to define f_H and f_A by taking maxima of the likelihood over the two hypothesis sets, obtaining LR, the profile likelihood ratio. In both cases a projection operator, integration for BF and maximization for profile LR, are used to bring the problem back to a scenario similar to two simple hypotheses. Both projection operators accomplish a dimensionality reduction, closely related to the idea of nuisance parameters elimination. This is crucial for the merits or drawbacks of both approaches, in contrast to the FBST, which always maintains the full original parameter space. For additional discussion of alternative p-values see Dempster (1997) and Kempthorne and Folks (1971).

Let us recall the usual notation used in statistical inference. Let two non sharp hypotheses, H and its alternative A, be defined by the partition of the parameter space, Θ, into the two sets $H : \theta \in \Theta_H$ and $A : \theta \in \Theta_A$. Whenever either element of the partition is not a unitary set, the respective hypothesis is called composite. Let g be an absolutely continuous prior density defined over Θ with

$$0 < \pi(H) = \int_H g(\theta)d\theta = 1 - \int_A g(\theta)d\theta = 1 - \pi(A) < 1.$$

BF is then readily obtained as before:

$$BF(X) = \frac{f_H(X)}{f_A(X)} = \frac{\pi(A)}{\pi(H)} \times \frac{\int_H g(\theta)f(x|\theta)d\theta}{\int_A g(\theta)f(x|\theta)d\theta}.$$

Let us note that the above expressions for f_H and f_A are coherent under the calculus of probability and are the weighted mean likelihood functions. Furthermore, the errors α and β of a NPW test based on the statistic BF above are now weighted mean errors. The most important aspect of these extensions is the validity of the resulting extension of the NPW Lemma. One needs only to use the weighted means to minimize the linear combination of weighted mean errors. In order to compute BF, the prior density g is used unless both hypotheses are simple.

The BF is called the Bayes Factor for its multiplication by the prior odds, O_0, provides the posterior odds, O_x. Hence it is the factor that realizes the Bayesian operation:

$$O_x(H) = O_0(H)BF(X) = \frac{\pi(H)}{\pi(A)} \times \frac{f_H(X)}{f_A(X)}$$

Frequentists, on the other hand, could define f_H and f_A as the likelihood's maxima over the two hypotheses sets, to obtain a Likelihood Ratio, LR, and the likelihood ratio test. Nevertheless, an optimal test, in the sense of NPW, is not obtained. This is the procedure used also when H is sharp.

The trouble, for Bayesians, begins when sharp hypotheses are considered. If H is sharp and composite, the BF cannot be obtained as before. If H is sharp, g places probability zero on Θ_H. To overcome this difficulty, a density over the set Θ_H is placed in addition to a positive probability γ for H to be true. Such making of a modified probability measure may become polemical yielding for example Lindley's paradox, see Lindley (1957) and Shafer (1982). On the other hand, this procedure has become almost standard in Bayesian testing with a quite extensive literature on the choice of densities over Θ_H, see Dawid and Lauritzen (2001) and Pereira and Wechsler (1993). The aim of these papers is to obtain the Bayes Factor, the main object of the original NPW Lemma.

A solution placed under the calculus of probability by the use of weighed mean likelihood functions is given by Pereira and Wechsler (1993). The prior density g is used to obtain the weighed mean likelihood functions $f_H(x)$ and $f_A(x)$, even if H is

sharp - surface integrals are used in this case. Using $f_H(x)$ and $f_A(x)$, a Bayes Factor, BF, is obtained for every sample point. By ordering the sample space according to the values of $BF(x)$, a P-value, which explicitly regards A, is obtained from $f_H(x)$, the new statistical model, see Montoya-Delgado et al. (2001). Note that this P-value can be obtained regardless of the dimension of the sample and parameter spaces. Furthermore, this P-value - based on the (Neyman-Pearson) test statistic $BF(x)$ - needs no use of asymptotic distributions.

The objective of this section was to present the main test procedures for sharp hypotheses which are to be compared to the FBST.

4 FBST Theory

A major practical issue for the implementation of the FBST is the determination of how large the Bayesian evidence against H must be in order for one to decide for its rejection. As discussed in Section 1, the mere fact of ev being a statistic defined on a zero to one scale does not ease the matter (the same occurs with ordinary p-values). The formal identification of the FBST as a Bayes test of hypothesis yields critical values derived from the loss functions allowing such identification.

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 should lead to a declaration of significance (or not). Another viewpoint is to identify ev as an estimator of the indicator function $I(\theta \in \Theta_H)$. Madruga et al. (2001) show that there are loss functions the minimization of which makes ev a Bayes estimator of the indicator function, see Hwang et al. (1992). A much more philosophical rebuff to that position, based on a complete denial of Decision Theory can be found in Stern (2007).

A third point of view could demand the identification of the FBST as a Bayes test, since, for instance, its submission to the Likelihood Principle is clearly not sufficient to confer Bayesianity to the procedure. As is the case with p-values, Decision Theory again clears the picture: Madruga et al. (2001) prove that the FBST procedure is the posterior minimization of an expected loss function λ defined by

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

where a, b and d are positive real numbers.

It should be remarked that there are other loss functions the minimization of which is equivalent to performance of FBST, but they are just slight variations of the function λ defined above (Madruga et al. (2001)). Let us now discuss and analyze the loss function λ. Its interpretation may be given as the measure of embarrassment experienced by a statistician who - having accepted H - would be told that the parameter θ is in the tangential set T, the set of high posterior density. This is called a *stylized form of statistical inference* in Bernardo and Smith (1994). Under such an interpretation, the

loss function λ is also the measure of pride of the statistician who - having now rejected H - would be told that $\theta \in T$. The balance between embarrassment and pride is of course represented by the constants a, b, and d.

The careful examination of λ reveals that it is a loss function which depends on the action, on the parameter θ, and on the tangential set T. This last argument of λ makes it a loss function dependent on the observed sample point x *and* on the prior density for θ. Dependence on the former is not unusual in Statistical Decision Theory: Bettors in pari-mutuel horse races make their decisions after reading the board of totalized bets.(This example was kindly advanced by Prof. J.M. Bernardo during a discussion at the Chilean Bayesian Seminar held at Antofagasta in 1999.)

The dependence of λ on the prior π reveals that the performance of the FBST may not separate probability from utility. The FBST is therefore submitted to the weak system of rationality axioms of Rubin (1987), although it may violate more artificial systems.

Let us return to the matter of the first paragraph of this section. The operational FBST procedure is given by the criterion according to which H is to be rejected if, and only if, the Bayesian evidence value ev is smaller than $c = (b + d)/(a + d)$. One should notice that the Bayesian evidence value ev is the formal test statistic and that a positive probability for H is never required.

The strongest critique against the FBST was the lack of invariance with respect to smooth parameterizations. By using a reference density in the definition of the tangential set T, Madruga et al. (2003) obtained the invariant version of the FBST. Two kinds of invariance are usually required in statistical procedures:

1. Invariance with respect to the null hypothesis parameterization;

2. Invariance with respect to the parameter space parameterization.

The intuitive definition of the FBST is already invariant with respect to parameterizations of the null hypothesis. This is not a trivial issue because some statistical procedures do not satisfy this property. For instance, Pereira and Lindley (1987) discusses the problem of testing homogeneity of proportions showing how different parameterizations of the hypothesis may produce different answers. Pereira and Stern (2001a) show that, using just the resulting posterior density of the mean and the variance of a normal analysis, tests for the mean, the variance and the coefficient of variation can be easily performed.

An explicitly invariant definition of the FBST with respect to alternative parameterizations of the parameter space is given in the sequel. To state this generalization of the FBST we consider a reference density, $r(\theta)$ on Θ. This density is established taking into account the original parameter space where the prior was defined. For example, $r(\theta)$ could be a non-informative (possibly improper) density on Θ. Consider now the following notation, where $s_x(\theta)$ is known as the surprise function relative to the reference

density $r(\theta)$, see Good (1983):

$$\theta^* = arg\max_{\theta\in\Theta_H} s_x(\theta) \quad \text{and} \quad s^* = \max_{\theta\in\Theta_H} s_x(\theta) = \frac{g_x(\theta^*)}{r(\theta^*)}$$

Definition 4.1. (Invariant Evidence) Let the set tangential to Θ_H be defined as: $T = \{\theta \in \Theta | s_x(\theta) > s^*\}$. The evidence against H provided by the sample x is

$$1 - ev = \overline{ev} = \int_T g_x(\theta) d\theta \ .$$

Definition 4.2. (Invariant Version) The invariant version of the FBST is the procedure that rejects H whenever $ev = 1 - \overline{ev}$ is small.

Interpretations of the reference density may be found in Madruga et al. (2003) and Stern (2004). Note that the invariant definition agrees with the intuitive definition when the reference density is the (possibly improper) uniform density.

Under the decision-theoretic approach to inference, the Bayesianity of the FBST still holds after the introduction of the reference density for its invariant version. The loss functions, λ, remain valid with the invariant definition. It should be emphasized that with the invariant definition the loss functions depend on the observation, the prior density, *and* the reference density, (x, g, r). This points out the nonseparability between prior, reference, and utility.

It should be also understood that the elimination of nuisance parameters is not recommended for a Bayesian, Pereira and Lindley (1987). We end this section with a simple example that should convince the reader to work in the complete parameter space.

Example 4.1 (Bow Tie). Consider a bivariate parameter of interest (x, y) defined on $(x, y) : -1 < x < 1 \& -2|x| - 1 < y < 2|x| + 1$. The joint and marginal posterior densities are as follows:

$$f(x, y) = (|x| + 4)/(36|x| + 18) \qquad f(x) = (|x| + 4)/9.$$

Figures 1 and 2 illustrates that the credible set obtained using the joint (marginal) density is the center (the tails) of the parameter space. The conclusion is that the use of marginal densities to built credible sets may produce incoherence.

5 Asymptotic Considerations

Although we have shown all practical and positive aspects of the statistic *ev*, one could (and will) question about the convergence of it. That is, the sampling distribution of ev could be of interest to a frequentist statistician. To a genuine Bayesian, we believe that properties of sampling distributions of ev are irrelevant. However, there exist nice mathematical aspects which may be explored. Since the posterior distribution depends

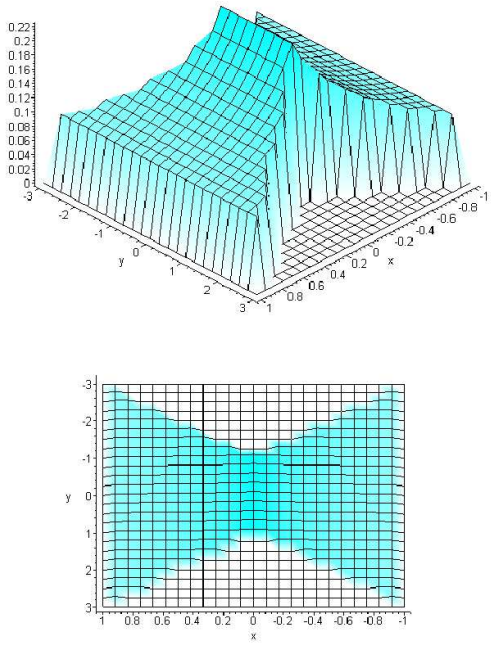

Figure 1: Overall and top views of the Bow Tie joint posterior density

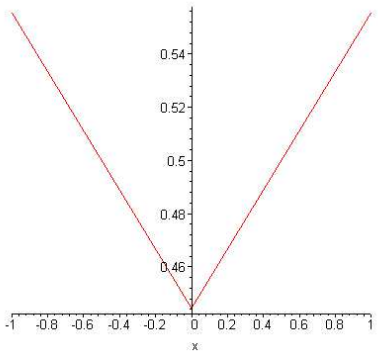

Figure 2: The Bow Tie marginal posterior density

on the sample point being observed, *ev* is of course a statistic, a function of the sample observation. For a frequentist, it is a natural need to obtain the sampling distribution of *ev*. The task is then to obtain, for all $0 < c < 1$, the (asymptotic) probabilities $Pr\{ev < c|\theta\}$.

We restrict ourselves to situations of well behaved likelihood and posterior densi-

ties, satisfying all contour properties listed in Schervish (1995), page 436. Relative to convergence of large samples, the normalized likelihood and the posterior density may be replaced one by each other. Letting L, M and m be respectively the normalized likelihood, the posterior mode and the maximum restricted to Θ_H, the tangential set T can be written as:

$$T = \{t \in \Theta | L(m) < L(t) < L(M)\}$$

Recalling the good behavior of L, one may make use of the normal approximation in order to evaluate the posterior probability of any subset of interest, T for instance. Hence, using the standard norm notation $\|(t-M)\|$, for vector $(t-M)$ we have

$$\|(t-M)\|^2 = (t-M)\Sigma^{-1}(t-M)'$$

where Σ^{-1} is the (generalized) inverse of the posterior covariance matrix, Σ, of Θ, we can write the tangential set as

$$T = \{t \in \Theta : \|m-M\|^2 > \|t-M\|^2\}$$

If k (> 1) is the dimension of Θ then, using the normal approximation, a posteriori, $\|\theta - M\|^2$ is asymptotically distributed as a χ^2 distribution with k degrees of freedom. Consequently, denoting the χ^2 distribution function with k degrees of freedom by F_k, the evidence value is evaluated as

$$ev = 1 - Pr\{T|x\} \approx F_k(\|m-M\|^2).$$

Recalling now that Θ and X are the parametric and sample spaces, we could look at this last probability as a conditional probability defined in the product space $\Theta \times X$ with product σ-algebra $B \times F$ and having $R \times F$ as the conditioning argument. That is, the event T is a set in the sub-σ-algebra $B \cap R$ and x is an event in the conditioning sub-σ-algebra. In the sequel an alternative representation of T is introduced.

Let the relative likelihood and its natural logarithm be denoted, respectively, by $l(t) = L(t)/L(M)$ and $\lambda(t) = \ln l(t)$. The tangential set has also the following representation:

$$T = \{t \in \Theta : \lambda(m) < \lambda(t) < 0\} = \{t \in \Theta : -2\lambda(m) > -2\lambda(t) > 0\}.$$

If k and h are the dimensions of Θ and Θ_H and recalling that the sampling asymptotic distribution of $-2\lambda(m)$ is χ^2 with $k-h$ degrees of freedom then, the sampling distribution of T may be also obtained. Using the subscript 0 to indicate the observed value of the statistic, the event $\{ev < ev_0\}$, is equivalent to the event $\{-2\lambda(m) > -2\lambda(m_0)\}$. Using now the sampling distribution of T, the p-value associated with the ev statistic, when ev_0 is its effective observation, is the superior tail of the χ^2 density with $k-h$ degrees of freedom, starting from $-2\lambda(m_0)$. Using the symbols one can write

$$pv_0 = Pr\{ev < ev_0|\theta\} = 1 - F_{k-h}(-2\lambda(m_0)).$$

We end this section by observing that, after the sample have been observed, the two sample values $d_0 = \|m_0 - M_0\|^2$ and $-2\lambda(m_0)$ of the χ^2 statistics allow one to evaluate both Bayesian and frequentist significance values: $ev_0 = F_k(d_0)$ and $pv_0 = 1 - F_{k-h}(-2\lambda(m_0))$. For a Bayesian (frequentist) the decision is based on ev_0 (pv_0). Since ev and pv are two well defined statistics, it should be of some interest to obtain the one-to-one relationship between them. We leave this as a challenge to the reader. We also call attention to the fact that the pv used here (Wilks (1935) and Wilks (1938)) never violates the Likelihood Principle.

6 Illustrative Examples

The use of the FBST in complex structures has shown to be very successful. A wear-out reliability test for Weibull distributions is presented by Irony et al. (2002). A bioequivalence test comparing two bivariate normal distributions is discussed in Lauretto et al. (2003) and the comparison of normal coefficient of variations in Pereira and Stern (2001a). Recently, Rodrigues (2006) presented an elegant discussion on the problem of the zero-inflated Poisson distribution. Loschi et al. (2007) used FBST for testing genetical hypotheses in a very complex structure. The performance of the FBST in the variable selection for regression models, presented by Pereira and Stern (2001b), is also important. Here, we consider very simple and classical problems of statistics to better illustrate the FBST good performance.

Example 6.1 (Bernoulli sample). Consider a sample of exchangeable Bernoulli trials with observations $(x, y) = (12, 24)$. That is, 12 successes and 24 failures were observed. To test that the data was generated by a fair coin, $H : \pi = 0.5$ we evaluate the evidence and have obtained $ev = 0.041$. For a classical statistician considering the data as generated by a binomial model, its exact pv would be 0.065. Another classical statistician considering negative binomial sampling with parameter $k = 12$, would obtain a p-value of 0.139. With the FBST there is no violation of the Likelihood Principle. Figure 3 illustrates how the computation of ev is done.

Example 6.2 (Two Bernoulli samples). Consider two samples of exchangeable Bernoulli trials with observations $(x, y)_1 = (4, 16)$ and $(x, y)_2 = (10, 10)$. The tangential set for this problem is illustrated in Figure 4. Here the posterior densities are independent Betas with parameters $(5, 17)$ and $(11, 11)$. The value of the evidence in favor of $H : \pi_1 = \pi_2$ is $ev = 0.012$. For a classical statistician who believes that the data are from two binomial distributions, the homogeneity chi-squared p-value is $pv = 0.047$. Assuming now that the second sample is from a negative binomial and keeping the binomial for the first, the chi-squared p-value for H is $pv = 0.142$. It is interesting to note that in both cases we have exactly the same sample figures and the same unknown parameters. The difference is due only to model restrictions. This is a violation of the Likelihood Principle. For comparisons of p-values and Bayes Factors, see Irony and Pereira (1986).

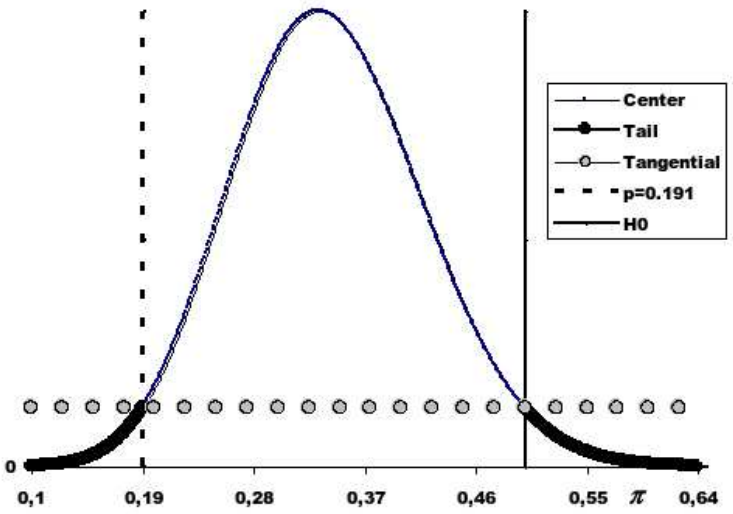

Figure 3: The evaluation of *ev* for the proportion example of a beta posterior density with parameters $(13, 25)$.

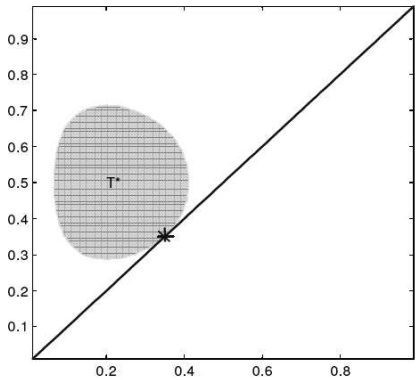

Figure 4: The evaluation of *ev* for the homogeneity test: The samples are $(4, 16)$ and $(10, 10)$.

A case of great importance in the statistical history is discussed next. We decided to present only the results of the FBST to illustrate how it is adequate in complex spaces and for complex hypotheses. The test for independence in a 2×2 contingency table is based in a collection of multivariate Bernoulli independent variables defined in the following set:

$$(0, 0, 0, 1); (0, 0, 1, 0); (0, 1, 0, 0); (1, 0, 0, 0)$$

The parameter is defined in the unity cube and taking values in the simplex represented by the set $\Pi = \{(\pi_1, \pi_2, \pi_3, \pi_4) : \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1\}$ The null sharp hypothesis is

defined by the following non linear sub-manifold:

$$\{(\alpha\beta, \alpha(1-\beta), (1-\alpha)\beta, (1-\alpha)(1-\beta)) : \alpha = \pi_1 + \pi_2; \beta = \pi_1 + \pi_3\} \subset \Pi$$

To measure the degree of dependence many indices were defined, see Goodman and Kruskal (1979) for instance. Many of such measures are functions of the cross product difference, δ: The difference between the product of the elements of the main diagonal and the product of the elements of secondary diagonal. Figure 5 is the diagram between this index and *ev*. We consider the case of a 2×2 contingency table with sample size $n = 20$. Note that the two indexes are in good agreement with the $|\delta|$ and *ev* being negative correlated.

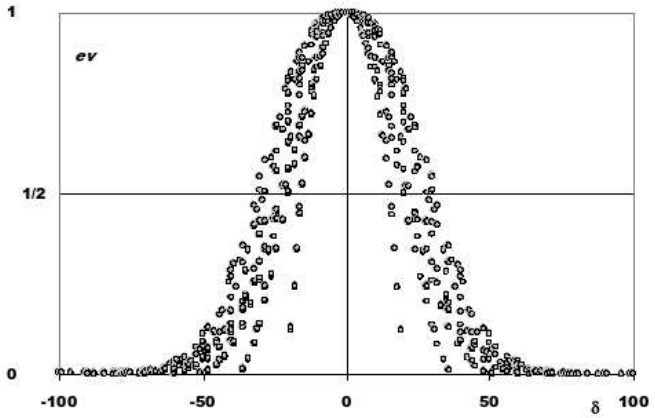

Figure 5: Diagram relating δ to *ev* in a 2×2 contingency table with $n = 20$.

We end this section with the most celebrated model in statistics, the normal distribution with unknown parameters, mean and variance. Here we calculate the evidence for a particular case.

Example 6.3 (Normal). Consider a sample from a normal distribution with unknown mean and variance, μ and $\sigma^2 = 1/\tau$ with τ being the precision parameter. After assigning a prior for $(\mu, \tau) \in \mathcal{R} \times \mathcal{R}$ the posterior density is given by

$$f(\mu, \tau) \propto \tau^{6,5} \exp\left\{-\frac{\tau\left[4 + 11\,(\mu - 0,9)^2\right]}{2}\right\}$$

The hypotheses and respective evidences are listed below:

The natural question is about the existence of alternative procedures to test hypotheses 2, 3 and 4, the hypotheses that are related to the coefficient of variation *cv*. In order to convince the reader about the strong power of the significance index *ev*, the

Table 2: Evidences against null hypotheses

H_0:	$\mu = 1.1$	$ev_0 = 0.51$
H_1:	$\tau = 2.5$	$ev_1 = 0.81$
H_2:	$cv = \sqrt{\mu^2\tau} = 0.5$	$ev_2 = 0.79$
H_3:	$\mu = 1.1 \ \wedge \ \tau = 2.5$	$ev_3 = 0.47$
H_4:	$\mu = 1.1 \ \wedge \ cv = 0.5$	$ev_4 = 0.49$
H_5:	$\tau = 2.5 \ \wedge \ cv = 0.5$	$ev_5 = 0.13$

following set of pictures from 6 to 10 is worth more than 1000 words. For a theoretical discussion on hypotheses composition see Borges and Stern (2007).

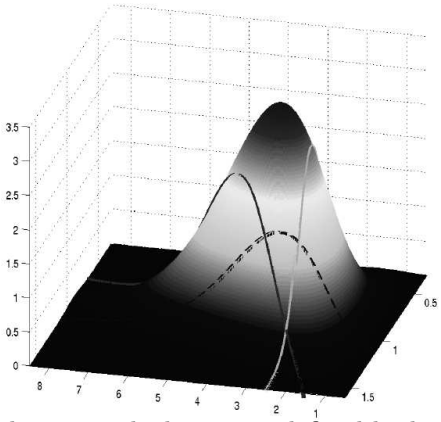

Figure 6: Posterior density with the curves defined by hypotheses 0, 1 and 2.

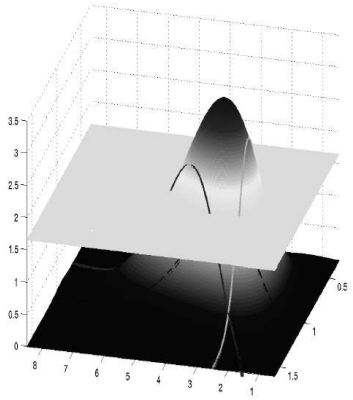

Figure 7: Cutting hyper-plane to define tangential set T_0

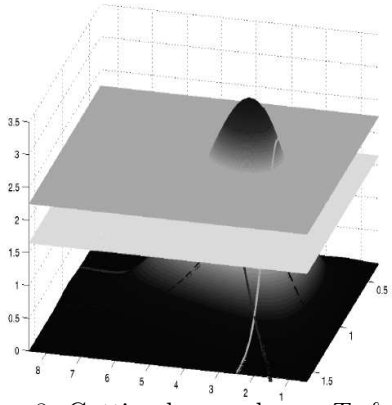

Figure 8: Cutting hyper-planes: T_0 & T_2.

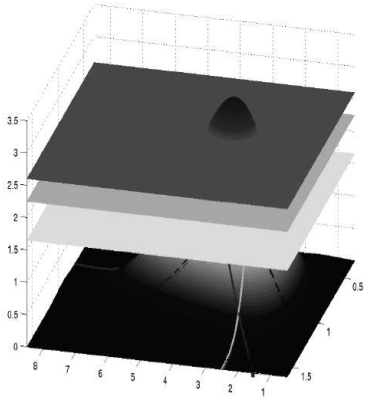

Figure 9: Cutting hyper-planes: T_0; T_1 & T_2.

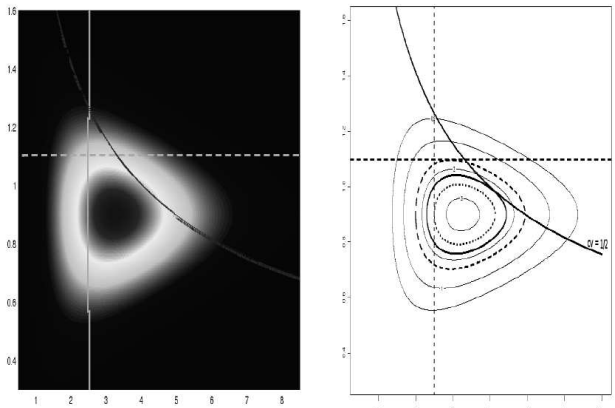

Figure 10: Level curves defining the tangential sets: T_0, T_1, T_2, T_3, T_4 & T_5.

7 Final Remarks

A list of several desirable properties of a test statistic that are satisfied by *ev* and its resulting FBST, is presented:

1. *ev* is a probability value derived from the posterior distribution on the full parameter space. *p*-values are also probabilities but defined on the sample space. Bayes Factors may be derived from marginal or conditional models due to the need for nuisance parameters elimination. For questions about the use of partial likelihoods see Pereira and Lindley (1987).
2. Both *ev* and FBST possess a version which is invariant against alternative parameterizations
3. The need of approximations in the computation of *ev* is restricted to numerical maximization and integration.
4. The Likelihood Principle is not violated by the FBST procedure.
5. The FBST requires neither the assignment of positive prior probabilities to sets of zero Lebesgue measure nor the elimination of nuisance parameters.
6. The FBST is a formal Bayes test and therefore has critical values easily obtained from elicited loss functions. This intrinsically makes critical sets and sample size dependent. Consequently, critical limits may change with sample size.
7. *ev* is a possibilistic (Darwiche and Ginsberg (1992)) support for sharp hypotheses, complying with the *Onus Probandi* juridical principle (*In Dubio Pro Reo* rule), see Stern (2003).
8. Being derived from the full posterior distribution, *ev* may be straightforwardly calculated for testing any precise hypothesis in the same parameter space. It is a homogeneous computation calculus with the same two steps: constrained optimization and integration with the posterior density.
9. The FBST, as a general inferential procedure, has performed successfully in several applications ranging from univariate parametric testing to multivariate model selection, without the need for any adaptations in the computation of *ev*.
10. *Ceteris paribus*, procedures that are computationally light, or even require only common statistical tables, are preferable to procedures requiring heavy or timely consuming computations. This does not favor the FBST, a computationally intensive procedure. Computing time was not a great burden in the many problems that the FBST was used. However, the sophisticated numerical algorithms used could be considered a more serious obstacle to the popularization of the FBST.

We can only end this article by addressing the question on its title. In a genuinely Bayesian way, the three authors have different answers to the question - their respective names are left for the reader to (probabilistically) guess.

Author A feels that Bayesians ought to be listened and that *ev* must replace *pv* in scientific literature. Author B thinks that any statistical procedure has to be explained in a formal logic way and really believes in the existence of sharp hypotheses. Finally author C still sees the FBST with some reluctance as he feels that sharp hypotheses should seldom be tested. Such reluctance is weakened by the removal of positive prior probabilities on sharp null hypotheses and also by the identification of significance testing as a decision problem. Author A agrees with the other two authors and helped the development of the theoretical properties presented and the three of them see the FBST as the Genuine Bayesian Significance Test for Sharp Hypotheses. Pereira et al. (2006), Stern (2007), and Wechsler (1993) may help the reader guessing.

In 1998, at Florida State University, while discussing Basu (1975) the late Professor Oscar Kempthorne challenged the first author to present a coherent and general Bayesian significance test procedure for sharp hypotheses. The idea was to have a Bayesian alternative to replace p-values while maintaining its most desirable (known or perceived) properties in practical use. From the list presented above, the authors believe they have responded successfully to Professor Kempthorne's 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

Basu, D. (1975). "Statistical Information and Likelihood (with discutions)." *Sankhya A*, 37: 1–71. 80, 97

— (1977). "On the elimination of nuisance parameters." *Journal of the American Statistical Association*, 72: 355–366. 80, 83

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. NY: J Wiley. 86

Birnbaum, A. (1962). "On the foundations of statistical inference (with discussion)." *Journal of the American Statistical Association*, 57: 269–326. 80

Borges, W. and Stern, J. (2007). "The rules of logic composition for the Bayesian epistemic e-values." *Logic J IGPL*, 15: in press. 94

Cox, D. (1977). "The role of significance tests." *Scandinavian J of Statistics*, 4: 49–70. 79

Darwiche, A. and Ginsberg, M. (1992). "A Symbolic Generalization of Probability Theory." In *AAAI92 Proc.*. American Association for Articial Intelligence. 96

Dawid, A. and Lauritzen, S. (2001). "Compatible Prior Distributions." In *Bayesian Methods with Applications to Science, Policy, and Official Statistics, EI George (editor)*, 109–118. Creta, Greece: ISBA. 85

DeGroot, M. (1975). *Probability and Statistics*. NY: Addison–Wesley, second edition. 83, 84

Dempster, A. (1997). "The direct use of likelihood for significance testing." *Statistics and Computing*, 7: 247–252. 84

Fisher, R. (1922). "The mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society*, 222: 309–368. 83

— (1934). *Statistical Methods for Research Workers*. Edinburgh, GB: Oliver and Boyd. 83

Good, I. (1983). *Good Thinking: the Foundations of Probability and its Applications*. Mineapolis, MN: University of Minnesota Press. 80, 88

Goodman, L. and Kruskal, W. (1979). *Measures of Association for cross classifications*. New York: Springer. 93

Hwang, J., Casella, G., Wells, M., and Farrell, R. (1992). "Estimation of accuracy in testing." *The Annals of Statistics*, 20: 490–509. 86

Irony, T., Lauretto, M., Pereira, C., and Stern, J. (2002). "A Weibull wearout test: full Bayesian approach." In *System Bayesian Reliability. Hayakawa, Irony, and Xie (Editors): In honor of Richard Barlow*, 287–300. Singapure: World Scientific. 81, 91

Irony, T. and Pereira, C. (1986). "Exact Bayes tests for equality of two proportions: Fisher versus Bayes." *J Statistical Comptation and Simulation*, 25: 93–114. 91

Irony, T., Pereira, C., and Tiwari, R. (2000). "Analysis of Opinion Swing: Comparison of Two Correlated Proportions." *The American Statistician*, 54(1): 57–62. 82

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Claredon Press. 80

Kass, R. and Raftery, A. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 777–795. 80

Kempthorne, O. (1976). "Of what use are tests of significance and tests of hypothesis." *Communications in Statistics - Theory and Methods*, 8(A5): 763–777. 79

Kempthorne, O. and Folks, L. (1971). *Probabilistic, Statistics, and Data Analysis*. Ames, IO: The Iowa U Press. 84

Lauretto, M., Pereira, C., Stern, J., and Zacks, S. (2003). "Comparing parameters of two bivariate normal distributions using invariant Full Bayesian Significance Test." *Brazilian J of Probability and Statistics*, 17: 147–168. 81, 91

Lindley, D. (1957). "A statistical paradox." *Biometrika*, 44(3): 187–192. 80, 85

— (1997). "Some comments on Bayes Factors." *Journal of Statistical Planning and Inference*, 61: 181–189. 80

Loschi, R., Monteiro, J., Rocha, G., and Mayrink, V. (2007). "Testing and estimating the non-conjunction fraction in meiosis I using reference priors." *Biometrical Journal*, 49(6): in press. 91

Madruga, M., Esteves, L., and Wechsler, S. (2001). "On the Bayesianity of Pereira-Stern tests." *Test*, 10: 291–299. 80, 86

Madruga, M., Pereira, C., and Stern, J. (2003). "Bayesian Evidence Test for Precise Hypotheses." *Journal of Statistical Planning and Inference*, 117: 185–198. 81, 87, 88

Mcnemar, Q. (1947). "Note on the sampling error of the differences between correlated proportions or percentages." *Psychometrika*, 12: 153–157. 82

Montoya-Delgado, L., Irony, T., Pereira, C., and Whottle, M. (2001). "An unconditional exact test for the Hardy-Weimberg Equilibrium Law: Sample space ordering using the Bayes Factor." *Genetics*, 158: 875–883. 86

Neyman, J. and Pearson, E. (1936). "Sufficient statistics and uniformly most powerful tests of statistical hypotheses." *Statistics Research Memoirs*, 1: 133–137. 83

Pereira, C. and Lindley, D. (1987). "Examples questioning the use of partial likelihood." *The Statistician*, 36: 15–20. 87, 88, 96

Pereira, C., Nakano, F., Stern, J., and Whittle, M. (2006). "Genuine Bayesian multiallelic significance test for the Hardy-Weinberg equilibrium law." *Genetics and Molecular Research*, 5(4): 619–631. 97

Pereira, C. and Stern, J. (1999). "Evidence and credibility: full Bayesian significance test for precise hypotheses." *Entropy*, 1: 69–80. 81

— (2001a). "Full Bayesian Significance Tests for Coefficients of Variation." In *Bayesian Methods with Applications to Science, Policy, and Official Statistics, EI George (editor)*, 391–400. Creta, Greece: ISBA. 81, 87, 91

— (2001b). "Model Selection: Full Bayesian Approach." *Environmetrics*, 12(6): 559–568. 91

Pereira, C. and Wechsler, S. (1993). "On the concept of P-value." *Brazilian J Probability and Statistics*, 7: 159–177. 80, 85

Rodrigues, J. (2006). "Full Bayesian significance test for zero-inflated distributions." *Communications in Statistics-Theory and Methods*, 35: 1–9. 81, 91

Rubin, H. (1987). "A weak system of axioms for rationality behavior and the non-separability of utility from prior." *Statistical Decisions*, 5: 47–58. 87

Schervish, M. (1995). *Theory of Statistics*. NY: Springer. 90

Shafer, G. (1982). "Lindley's paradox (with comments)." *Journal of the American Statistical Association*, 77: 325–351. 80, 85

Stern, J. (2003). "Significance tests, Belief Calculi, and Burden of Proof in legal and Scientific Discourse." *Frontiers in Artificial Intelligence and its Applications*, 101: 139–147. 80, 96

— (2004). "Paraconsistent sensitive analysis for Bayesian significance tests." *Lecture Notes in Artificial Intelligence*, 3171: 134–143. 88

— (2007). "Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses." *Cybernetics & Human Knowing*, 14(1): 9–36. 86, 97

Stern, J. and Zacks, S. (2002). "Testing the independence of Poisson variables under the Holgate bivariate distribution." *Statistics and Probability Letters*, 60: 313–320. 81

Wald, A. (1939). "Contributions to the theory of statistical estimation and testing hypotheses." *Annals of Probability and Statistics*, 10: 299–326. 83

— (1950). *Statistical Decision Functions*. N York: Wiley. 83

Wechsler, S. (1993). "Exchangeability and Predictivism." *Erkenntnis; International J Analytic Philosophy*, 38(3): 343–350. 97

Wilks, S. (1935). "The likelihood test of independence in contingency tables." *Annals of Mathematical Statistics*, 6: 190–196. 91

— (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses." *Annals of Mathematical Statistics*, 9: 60–62. 91

Acknowledgments

The authors gratefully acknowledge the support of the Brazilian funding agencies, FAPESP (2003/10105-2), CNPq and CAPES, which provide the research environment they live in. They also thank the Editors and Referees for their corrections and suggestions, which enriched very much the paper. The nice pictures used in this paper are due to Ricardo Vêncio. Last but not least the authors thank Professor Dennis Lindley for his important comments on a previous version of the paper.