International Journal for Uncertainty Quantification, 2 (2): 161–172 (2012)

RELATIONSHIP BETWEEN BAYESIAN AND FREQUENTIST SIGNIFICANCE INDICES

Marcio Diniz,[1,] Carlos A. B. Pereira,[2] Adriano Polpo,[1] Julio M. Stern,[2] & Sergio Wechsler[2]*

[1] *Department of Statistics, Federal University of Sao Carlos, Rod. Washington Luis, km 235, Sao Carlos, SP, 13565-905, Brazil*

[2] *Mathematics and Statistics Institute, University of Sao Paulo, Sao Paulo, SP, Brazil*

Original Manuscript Submitted: 06/29/2011; Final Draft Received: 02/05/2012

*The goal of this paper is to illustrate how two significance indices—the frequentist **p**-value and Bayesian **e**-value—have a straight mathematical relationship. We calculate these indices for standard statistical situations in which sharp null hypotheses are being tested. The **p**-value considered here is based on the likelihood ratio statistic. The existence of a functional relationship between these indices could surprise readers because they are computed in different spaces: **p**-values in the sample space and **e**-values in the parameter space.*

KEY WORDS: *Bayesian test, likelihood ratio, **e**-value, **p**-values, significance test*

1. INTRODUCTION

This paper relates two significance indices: the **p**-value and **e**-value, **p** for probability and **e** for evidence. The main objective of these indices is to measure the consistency of data with a sharp null hypothesis **H** (**A** representing the alternative). As heuristic definitions of the indices, we could say the following: the **p**-value is the probability of the set of points of the sample space that has densities smaller than the actual sample computed under **H**; the **e**-value is the posterior probability of the set of points of the parameter space that has posterior densities smaller than the maximal density within **H**.

The frequentist **p**-value has a long history because it appeared in the statistical literature. [1–3] seem to be the first to use the concepts of "tail" and "more extreme" sample points—now included in almost every statistics textbook. Their definition of **p**-value is the probability, under the null hypothesis, that the test statistic would take a value at least as extreme as the one in fact observed. More recently, [4, 5] strongly advocated the use of **p**-values as proper indices to evaluate significance. **p**-values have been by far the most used statistical tool in all fields of science.

The history of the Bayesian **e**-value is recent, being introduced by [6] and reviewed extensively by [7]. It has been applied in different fields, [8–10].

The comparison of frequentist to Bayesian tests is made by [11–13]. In these works, **p**-values are compared to Bayes factors. Usually, papers comparing Bayesian and classical tests reach the conclusion for accept/reject rules (see, for instance, [14]). A decision-theoretical approach to significance testing is developed by [15]. Here we are looking exclusively for the values of the significance indices, **p**- and **e**-values.

There is a clear duality between **p**- and **e**-values: while the former is a tail evaluation of the sampling distribution under the null hypothesis, the latter is a tail evaluation of the posterior distribution conditional on the actual sample observation. Furthermore, while the tail for **p**-value evaluation starts at the observed statistic value, the tail for **e**-value starts at the sharp null hypothesis. In other words, the **p**-value is the tail from $\boldsymbol{x}$ given **H** volume, while the **e**-value is the tail from **H** given $\boldsymbol{x}$ volume. More detailed definitions of both indices and examples are provided in the sequel.

*Correspond to Marcio Diniz, E-mail: marcio.alves.diniz@gmail.com

2152–5080/12/$35.00 © 2012 by Begell House, Inc.

We call the reader's attention to the way a tail should be defined. To have a meaningful concept of more extreme sample (parameter) points, an order must be defined in the sampling (parameter) space. We consider the order based on the likelihood ratio λ [posterior density $\pi(\theta)$]: a sample point $\boldsymbol{y}$ is more extreme than $\boldsymbol{x}$ if $\lambda(\boldsymbol{y}) < \lambda(\boldsymbol{x})$ [parameter θ_2 is more extreme than θ_1 if $\pi(\theta_2) < \pi(\theta_1)$]. As seen in [16] and [17], the likelihood ratio function λ (the posterior density π), defines a natural order on the sample (parameter) space, regardless of the dimension of the space. One should note that these orders do take into account the alternative hypothesis **A** to the null hypothesis **H**. Figures 1 and 2 illustrate this discussion displaying the two tails for the binomial with sample size $n = 10$, observation $\boldsymbol{x} = 3$, and null hypothesis $\mathbf{H} : \theta = 1/2$.

The above considerations suggest that there is a strong connection between **p** and **e**. Such a relation would depend on the chosen prior and on the dimension of the two spaces considered, sampling and parameter spaces. To fairly compare the indices, we use noninformative priors for **e**-values and the likelihood ratio statistic for **p**-values.

While searching for an analytical relation between the indices, we realized that a possible function would be model dependent, as the examples presented confirm. Some surprising results are shown, particularly for multiparameter cases.

Section 2 presents the background, notation, and definitions. Also, some illustrative examples are provided. Section 3 presents our conjectures about promising relations between **p**- and **e**-values. More examples are given in Sections 4 and 5, where intriguing problems are discussed. Finally, Section 6 provides additional comments and discussion.

2. MOTIVATION AND DEFINITIONS

For motivation, we consider a standard illustration: the normal mean test with variance equal to 1. Let $t = 3.9$ be the observed value of the minimal sufficient statistic: the sample mean for $n = 3$ observations. Suppose that the null hypothesis to be tested is $\mathbf{H} : \mu = 5$. Regard the normalized likelihood function as a normal density with mean equal to 3.9 and variance equal to $1/3$. Recall that the sampling density under **H** is a normal distribution with mean 5 and variance $1/3$. The **p**-value is 0.0567, twice the area of the tail starting at $t = 3.9$. Surprisingly, twice the area of the tail that starts at $\mu = 5$, in the posterior distribution, also equals 0.0567, $\mathbf{p} = \mathbf{e}$. This result is a consequence of the fact that the normal density depends on t and μ only on $(t - \mu)^2$.

Consider now a binomial sampling distribution with $n = 10$ and observation $\boldsymbol{x} = 3$. The interest is to test $\mathbf{H} : \theta = 1/2$ vs. $\mathbf{A} : \theta \neq 1/2$. Figures 1 and 2 illustrate the two kinds of tails discussed. Recall that in this case

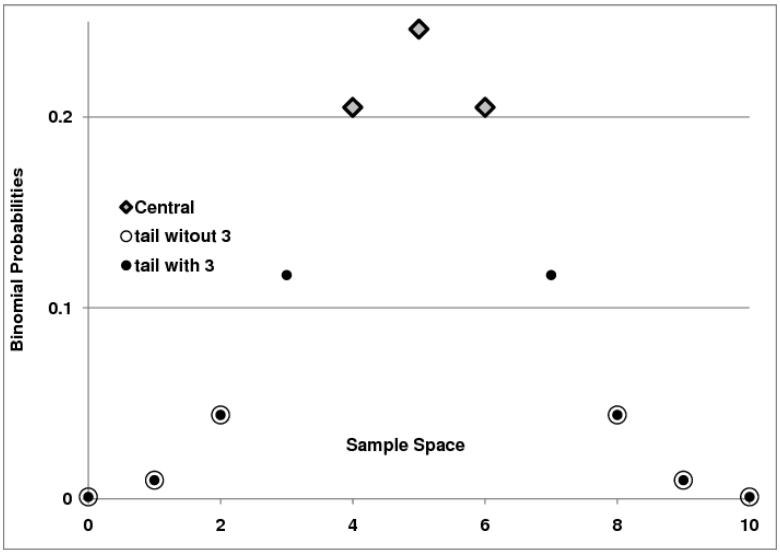

FIG. 1: Binomial sampling: posterior density under $n = 10$ and $\boldsymbol{x} = 3$.

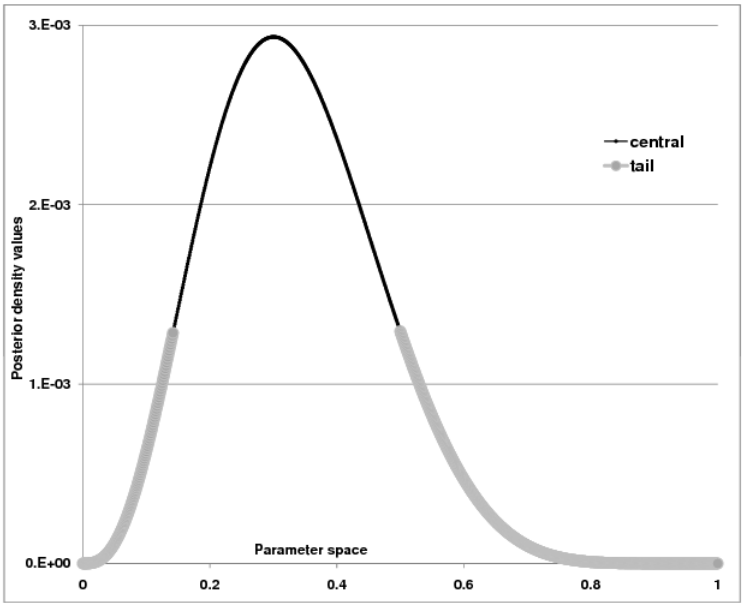

FIG. 2: Binomial sampling: posterior density under $n = 10$ and $\boldsymbol{x} = 3$.

the normalized likelihood is a beta density with parameter $(4, 8)$. The **p**-value is 0.109375 or 0.343750, depending on the observation $\boldsymbol{x} = 3$ being or not deemed to be extreme. On the other hand, the **e**-value is the sum of the right and left areas, $P[\theta > 0.5|\boldsymbol{x} = 3] + P[\theta < 0.14172|\boldsymbol{x} = 3]$ = 0.11328 +0.0582 = 0.17148. A curious fact is that twice this 0.11328 right tail area equals exactly the average of the two **p**-values: the one that includes $\boldsymbol{x} = 3$ and the other excluding it.

The two examples above illustrate that **p** and **e** may be, to some extent, functionally related even when they have different values. The aim of this paper is to investigate their relation in several particular cases and to suggest to some extent general results.

2.1 Basic Definitions

Our objective is to compare significance indices defined under different paradigms. Let us denote by Θ, the parameter space, $\mathcal{B}$ the sigma-algebra of subsets of Θ and π the prior probability measure. The prior probability model, which exists in the mind of the investigator, is the triple $(\Theta, \mathcal{B}, \pi)$. There is an observable random variable X with sample space $\mathcal{X}$ and an associated sigma-algebra $\mathcal{A}$. The distribution of X is supposed to belong to the family $\{P_\theta : \theta \in \Theta\}$. These elements form the statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$.

Let Π be the product measure for the joint random object $\omega = (\theta, X)$, $\Omega = \Theta \times \mathcal{X}$, and $\mathcal{F} = \mathcal{B} \times \mathcal{A}$. That is, $(\Omega, \mathcal{F}) = (\Theta \times \mathcal{X}, \mathcal{B} \times \mathcal{A})$ is a measurable product space with a probability model $(\Omega, \mathcal{F}, \Pi)$. There is an important restriction in building such a complete model: P_θ must be a measurable function on $\mathcal{B}$ for every $\boldsymbol{x}$ on $\mathcal{X}$. Hidden behind the wings of this global probability model $(\Omega, \mathcal{F}, \Pi)$ are the following four operational models:

1. the prior model $(\Theta, \mathcal{B}, \pi)$
2. the statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$, a set of probability spaces
3. the posterior model $(\Theta, \mathcal{B}, \{\pi_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\})$
4. the predictive model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{P}$ is the marginal or predictive distribution of $\boldsymbol{x}$

If the probability spaces are absolutely continuous, then we have

1. the likelihood functions $D_x = \{f(\boldsymbol{x}|\theta) = L(\theta|\boldsymbol{x}); \theta \in \Theta\}$
2. the posterior density functions $P_{\boldsymbol{x}} = \{\pi(\theta|\boldsymbol{x}); \theta \in \Theta\}$

The likelihood ratio was chosen to order the sample space. An order that regards both the null and alternative hypotheses. This subject is discussed by [16] and [17]. The asymptotic distribution of the likelihood ratio statistics under the null hypothesis is given by [18]. On the other hand, by definition, **e**-values always take into account both hypotheses.

Definition 1.
A sharp null hypothesis **H** is the statement $\theta \in \Theta_H$ for $\Theta_H \subset \Theta$ in which the dimension h of Θ_H is smaller than the dimension m of Θ. The global alternative hypothesis **A** is the statement $\theta \in \Theta - \Theta_H$.

As can be seen in the sequel, the computation of **e**-values is based on $P_{\boldsymbol{x}}$ while **p**-values are based on D_x.

2.2 Likelihood Ratio p-Value

The likelihood ratio statistic for an observation $\boldsymbol{x}$ of $\mathcal{X}$ is defined as follows:

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_H} L(\theta|\boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta|\boldsymbol{x})}.$$

Definition 2.
The **p**-value at $\boldsymbol{x}$ is the probability, under **H**, of the event $S_x = \{y \in \mathcal{X} : \lambda(y) \leq \lambda(\boldsymbol{x})\}$.

For large samples and under well-behaved models $-2\ln\lambda(\boldsymbol{x})$ is asymptotically distributed as χ^2 with $m - h$ degrees of freedom (see [19]).

2.3 Evidence Index: e-Values

Definition 3.
Let $\pi(\theta|\boldsymbol{x})$ the posterior density of θ given the observed sample and $T(\boldsymbol{x}) = \{\theta \in \Theta : \pi(\theta|\boldsymbol{x}) \geq \sup_{\Theta_H} \pi(\theta|\boldsymbol{x})\}$. The measure of evidence favoring the hypothesis $\theta \in \Theta_H$ is defined as $Ev(\Theta_H, \boldsymbol{x}) = 1 - P(\theta \in T(\boldsymbol{x})|\boldsymbol{x})$.

The evidence value considers, favoring a sharp hypothesis, all points of the parameter space whose posterior density values are, at most, as large as its supremum over Θ_H. Therefore, a large value of $Ev(\Theta_H, \boldsymbol{x})$ means that the subset Θ_H lies in high (posterior) probability of Θ and this shows that the data strongly support the hypothesis. On the other hand, when $Ev(\Theta_H, \boldsymbol{x})$ is small this shows that Θ_H is in a low-probability region of Θ and the data would make us discredit **H**. An advantage of this procedure is that it overcomes the difficulty of dealing with sharp hypotheses because there is no need to introduce a prior probability for the null hypothesis, as in the usual Bayesian test, which uses Bayes factors. The Bayesian procedure that uses the evidence value to test sharp hypotheses is also known as the full Bayesian significance test (FBST) (see [7]).

2.4 Illustration

In this section, we consider the standard example of comparing two proportions θ_1 and θ_2, parameters of two independent binomial samples, x and y. Consider that the sample sizes are $m = 20$ and $n = 30$ and the observed data $x = 14$ and $y = 12$. The objective is to test **H**:$\theta_1 = \theta_2$, against **A**:$\theta_1 \neq \theta_2$.

The likelihood ratio test statistic in this case is 4.42, and the **p**-value is the tail of a χ^2 density with one degree of freedom, $p = 0.0355$. To compute the **e**-value, first we obtain the tangential set (T) described in Fig. 3 and obtain the posterior probability of this set, $P(\theta \in T(\boldsymbol{x})|\boldsymbol{x})$. Taking its complement, we obtain the **e**-value.

Let us consider independent uniform priors for the parameters θ_1 and θ_2. The two independent posteriors in this situation are $\pi(\theta_1|x = 14; m = 20) \sim$ Beta(15;7) and $\pi(\theta_2|y = 12; n = 30) \sim$ Beta(13;19). The point $\theta^* = 26/50$ is the parameter point that maximizes the posterior under Θ_H and defines the tangential set.

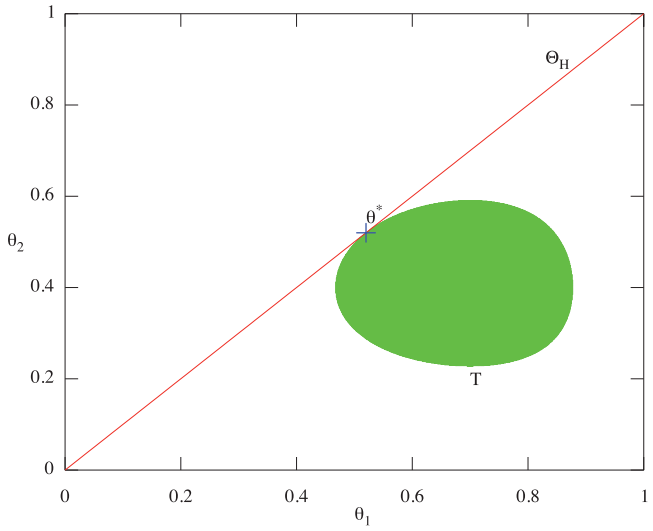

FIG. 3: Parameter space, Θ_H and tangent set (T) for two Binomial counts.

The **e**-value obtained with these posteriors is 0.0971. Taking the tail of a χ^2 density with 2 degrees of freedom, tail starting at 4.42, the value of the likelihood ratio statistic at the observation, we obtain the value 0.1097, which is close to our **e**-value.

3. RELATING THE TWO INDICES

The duality between **p** and **e**-values presented in Section 1 suggests that there might exist a functional relation between them. In principle, such a relation would operationally equate the significance indices turning unimportant the choice among them. There is, however, an important advantage of **e**-values over **p**-values: the use of the former never violates the paramount Likelihood principle, [20], while significance tests based on **p**-values may violate the principle for small samples. Furthermore, and indeed, the tentative functional relations would not be invariant to noninformative sampling stopping rules, even if **e**-values, by adhering to the Likelihood Principle, do not depend on the stopping rule. The relation between **e** and **p** would therefore depend on the entertained statistical model.

To not compromise the comparison, we designed it with **p**-values the order of which on the sample space takes into account both hypotheses, null and alternative: The Likelihood ratio test statistic orders the sample space regarding both hypotheses. Moreover, the use of **p**-values based on the Likelihood ratio test do not violate the Likelihood principle for large samples. The logical necessity of regarding both hypotheses is extensively discussed by [16] and amounts, at the end of the day, to the use of Neyman-Pearson Lemma in lieu of "null-only" significance tests (see also [15] on this matter). On the other hand, **e**-values, by their very definition, intrinsically regard both **H** and **A**. To explicitly enunciate the formal relation between the indices we present the next result. By recalling that $\dim(\Theta) = m$, $\dim(\Theta_H) = h$, $\overset{a}{\sim}$ is for asymptotical distribution and F_k being the distribution function of a χ^2_k, we have the following:

Theorem 1. *Assuming that the countour restrictions for asymptotic normality described by [21] are satisfied,* $-2\ln\lambda(\boldsymbol{x}) \overset{a}{\sim} \chi^2_{m-h}$, $Ev(\Theta_H, \boldsymbol{x}) \approx 1 - F_m[F^{-1}_{m-h}(1-\boldsymbol{p})]$.

Proof: Assume that all contour properties listed in [21] page 436, are satisfied. Relative to convergence of large samples, the normalized likelihood and the posterior density are indistinguishable.

Letting L, M, and m be, respectively, the normalized likelihood, the posterior mode, and the maximum restricted to Θ_H. Therefore, $L(m|\boldsymbol{x}) = \sup_{\theta\in\Theta_H} L(\theta|\boldsymbol{x})$, $L(M|\boldsymbol{x}) = \sup_{\theta\in\Theta} L(\theta|\boldsymbol{x})$ and the tangential set is $T = \{\theta \in \Theta : L(m|\boldsymbol{x}) \leq L(\theta|\boldsymbol{x})\}$. Because of the the good behavior of L, one may use the normal approximation in order to evaluate the posterior probability of any subset of interest, like T. Hence, using the standard norm notation $\|(\theta - M)\|$, for vector $(\theta - M)$, we have

$$\|(\theta - M)\|^2 = (\theta - M)'\Sigma^{-1}(\theta - M)$$

where Σ^{-1} is the (generalized) inverse of the posterior covariance matrix of θ. We can write the tangential set as

$$T = \{\theta \in \Theta : \|m - M\|^2 \geq \|\theta - M\|^2\}. \tag{1}$$

If $\dim(\Theta) = m$ (> 1) and using the normal approximation, then $\|\theta - M\|$ is asymptotically distributed as a χ^2 distribution with m degrees of freedom. Consequently, the evidence value is evaluated as

$$Ev(\Theta_H, \boldsymbol{x}) = 1 - P(T \in \Theta|\boldsymbol{x}) \approx 1 - F_m(\|m - M\|^2). \tag{2}$$

Let the relative likelihood be denoted by $\lambda(\theta) = L(\theta|\boldsymbol{x})/L(M|\boldsymbol{x})$. The tangential set has also the following representation:

$$T = \{\theta \in \Theta : \ln\lambda(m) \leq \ln\lambda(\theta)\} = \{\theta \in \Theta : -2\ln\lambda(m) \geq -2\ln\lambda(\theta)\}. \tag{3}$$

From (1) and (3), we have that $\|m - M\|^2 \approx -2\ln\lambda(m) = c_m$, and the asymptotic **p**-value is

$$\mathbf{p} = 1 - F_{m-h}(c_m).$$

Writing $c_m = F^{-1}_{m-h}(1 - \mathbf{p})$ and substituting in (2), we obtain $Ev(\Theta_H, \boldsymbol{x}) \approx 1 - F_m[F^{-1}_{m-h}(1 - \mathbf{p})]$. ■

Consequently, asymptotically the relation between the indices is simply a relation between two survival functions of χ^2 with different degrees of freedom [i.e. $\overline{F}_{m-h}(s) = 1 - F_{m-h}(s) =$ **p**-value and $\overline{F}_m(s) = 1 - F_m(s) =$ **e**-value, where s is the actual observation of S, the test statistic]. We did note by experience that there is a beta distribution function relating two χ^2 survival functions evaluated at the same point. The webpage www.ufscar.br/∼polpo/papers/indices presents 900 graphics illustrating the fitting quality of the beta distribution functions relating two χ^2 survival functions with degrees of freedom varying from 1 to 30. This fact favors the use of these beta functions to the detriment of the inverse of the χ^2 distribution function that does not have closed form. One can also observe that, in most of our standard examples, this kind of beta fitting works well even for small samples, using indices calculated exactly.

Whenever the sharp hypothesis is a single point, the null hypothesis space has nil dimension and the **p**- and **e**-values are asymptotically equal. The two indices can be exactly evaluated under various Gaussian statistical models. An exception is the Behrens-Fisher problem, discussed in Section 5.2, for which exact **p**-values are not available.

The examples discussed in the sequel illustrate the functional relation under several statistical models.

4. SIMPLE ILLUSTRATIVE EXAMPLES

This section illustrates situations with continuous random variables for which we know the likelihood ratio distributions or some transformations of them. In such cases, one can find the **p**-values without the χ^2 approximation that are used for large samples.

4.1 Mean Test for Normal Random Samples

We simulated 1000 normal random samples with nine observations each, zero mean, and variance 2. Considering the variance unknown, the LR is equivalent to the Student's t procedure for $H_0 : \mu = 0$. For the FBST, we assume a noninformative prior, $\pi(\mu, \sigma) \propto 1/\sigma$. Figure 4 plots the **p**-values in the horizontal axis and **e**-values in the vertical axis.

The fitted line was obtained by finding the beta parameters, which minimized the sum of squared differences between the data and the beta[1] cdf. The estimated values for the beta parameters were $a = 1.1004$ and $b = 2.0884$.

[1]We are considering the following expression for the beta density of a random variable x: $f(x|a, b) = [1/B(a, b)]x^{a-1}(1 - x)^{b-1}$ for $x \in [0, 1]$, a and b positive real parameters.

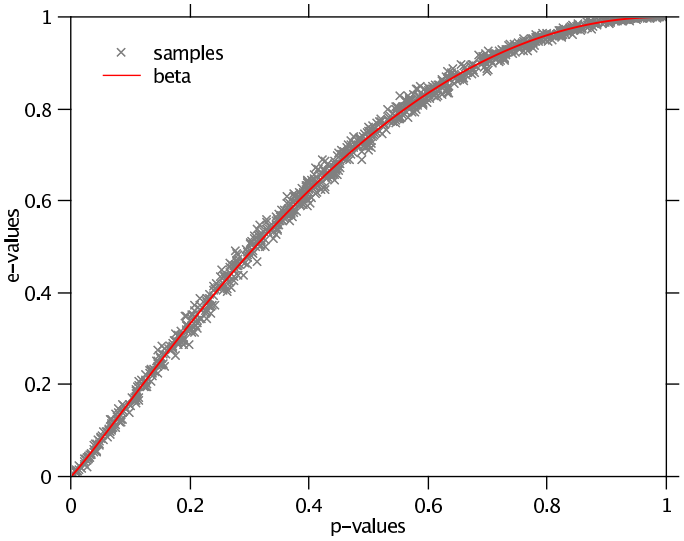

FIG. 4: Mean test with normal random samples.

4.2 Exponential Random Samples

Consider the case of two samples of independent and identically exponential distributions with parameters δ_1 and δ_2. Our interest is to test **H**:$\delta_1 = \delta_2$.

For this exercise, we simulated 1000 samples of size $n = 20$ and 1000 of size $k = 25$ of an exponential distribution with the parameter equal to 2. Taking these two groups of samples, we consider the first as random samples of X and the remaining as random samples of Y: X and Y representing random variables. Under **H**, the statistic $T = \sum_{i=1}^{n} x_i/(\sum_{i=1}^{n} x_i + \sum_{i=1}^{k} y_i)$ follows a beta distribution with parameters $n = 20$ and $k = 25$. Therefore, we can find the distribution of some function of the likelihood ratio statistic, $\lambda(X, Y)$, because this statistic is a function of T.

For the FBST we adopted noninformative priors for δ_i, $\pi(\delta_i) \propto 1/\delta_i$, where $i = 1, 2$. The fitted line in Fig. 5 is the accumulated beta distribution function with estimated parameters $a = 0.79378$ and $b = 1.9943$.

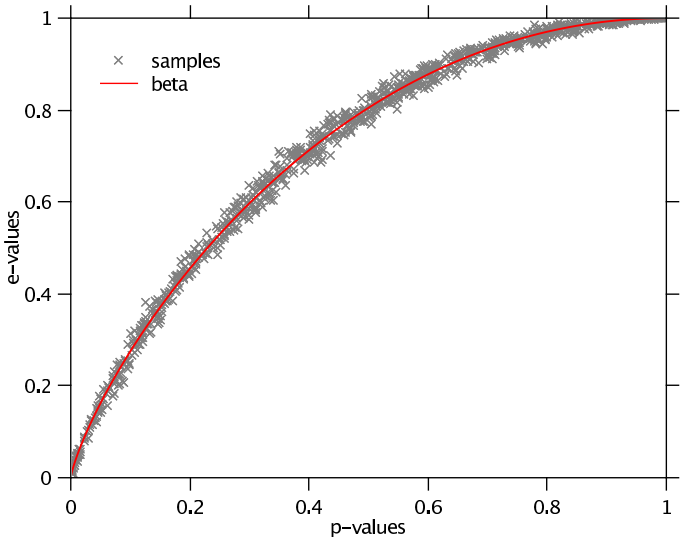

FIG. 5: Exponential random samples test.

4.3 Shape Parameter for Pareto Random Samples

Considering a random sample from a Pareto distribution with shape parameter γ and scale parameter ν, we focus on $H_0 : \gamma = 1$, ν being unknown. In order to calculate the LR exact **p**-value, we use the fact that $2T \sim \chi^2_{2(n-1)}$, where n is the sample size and $T = \log\left[\prod_{i=1}^{n} x_i/(x_{(1)})^n\right]$, where $x_{(1)}$ is the sample minimum. For the FBST, we again propose the noninformative priors $\pi(\gamma) \propto 1/\gamma$ and $\pi(\nu) \propto 1/\nu$.

We found the **p**- and **e**-values for 1000 simulated random samples with 20 observations each, $\gamma = 1$ and $\nu = 2$. Figure 6 shows the results with the beta parameters estimated at $a = 1.4712$ and $b = 2.428$.

5. INTRIGUING CASES

By performing simulations for another test whose **p**-values can be calculated exactly, we found a situation where the proposed relationship fails to happen, at least for small samples. This led our thoughts to another question: what is the necessary number of observations for the relationship to be valid? For the above-proposed tests, and the tests to follow, we reached the answer by simply following practical considerations and observing the simulation results. As a final example we consider the Behrens-Fisher problem to illustrate the use of asymptotical **p**-values.

5.1 Variance Test for Normal Random Samples

We simulated 1000 normal random samples with 20 observations each, zero mean and variance two. The LR gives us the F test for $H_0 : \sigma^2 = 2$. For the FBST, we use the prior $\pi(\mu, \sigma) \propto 1/\sigma$.

Figure 7 below shows the plot of **p**- and **e**-values. Faced with this umbrella graph, we increased the sample size hoping to get the beta relation as above. Therefore, we performed the same exercise with samples with 100, 500, and 1000 observations each and eventually arrived at the beta relationship. Figure 8 shows the results of the accumulated beta distribution with estimated parameters $a = 0.84161$ and $b = 2.0182$ for samples with 1000 observations. Evaluating the **e**-values through the relation given by Theorem 1, and drawing the line in Fig. 8, it is not possible distinguish between the asymptotic and empirical relation estimated by the beta distribution. To show how close these two lines are, we evaluated the maximum pointwise absolute distance between the two lines and obtained the value of 0.005403. The graphs for samples with 100 and 500 observations are available at the website www.ufscar.br/∼polpo/papers/indices.

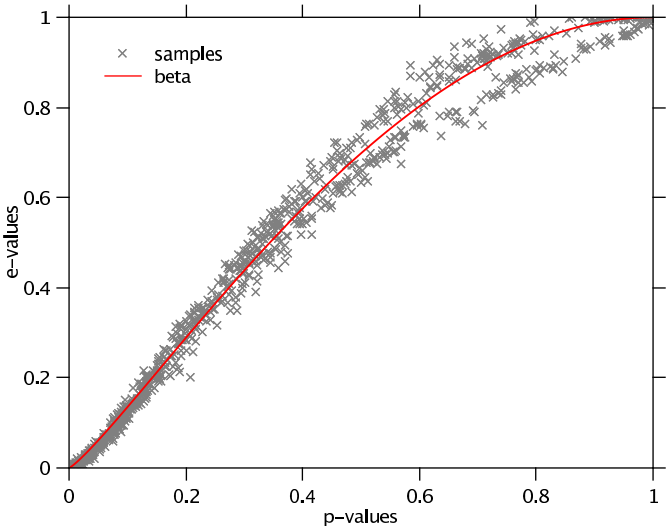

FIG. 6: Shape parameter test for Pareto random samples.

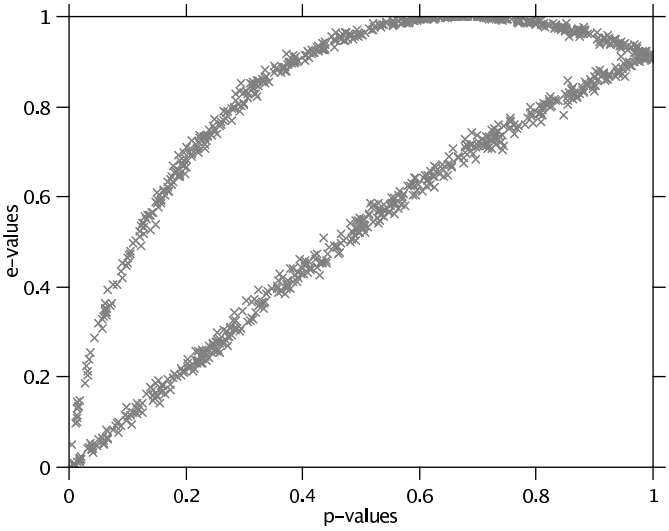

FIG. 7: Variance test for normal random samples with 20 observations.

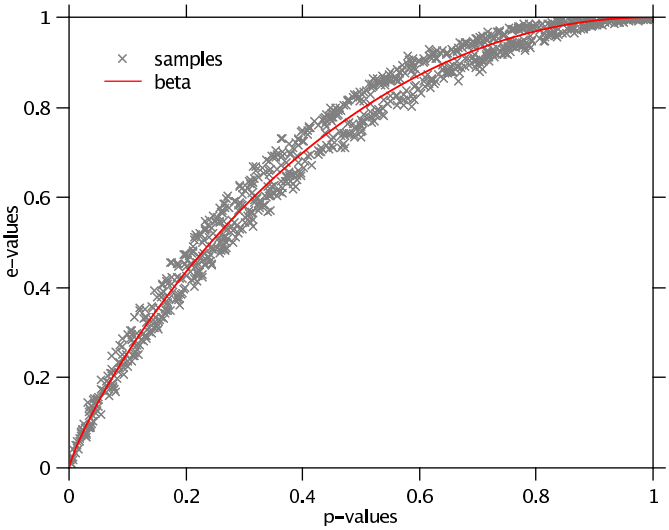

FIG. 8: Variance test for normal random samples with 1000 observations.

We can see in Fig. 9 why the relation is not obeyed for small samples when we plot the sample variance against the respective **p**- and **e**-values. The horizontal axis presents the sample variance, and the vertical axis the correspondent **p**- and **e**-values both evaluated for the 20 observations samples. The left legs of both curves display the samples that are on the straight line in Fig. 7, and the right legs display the samples that form the umbrella shape.

5.2 Behrens-Fisher Problem

For the Behrens-Fisher problem we simulated 1000 samples of a normal random variable X with 9 observations each, mean μ and variance σ_x^2. At the same time, we generated 1000 samples of a normal random variable Y with 20 observations each, mean μ and variance σ_y^2. For each sample, the value of μ was generated from a normal distribution with zero mean and variance 100. The standard deviations were obtained from γ distributions with expected values of 7 and 20 for σ_x and σ_y, respectively.

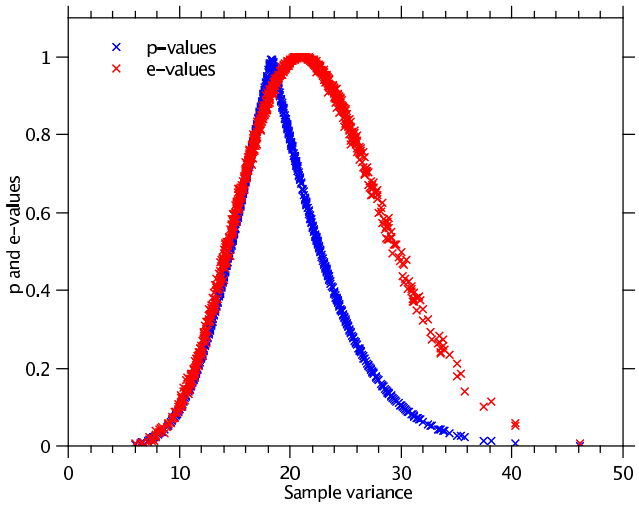

FIG. 9: Sample variance against **p**- and **e**-values for normal samples with 20 observations.

In this case the LR does not provide an exact **p**-value. Therefore, we present the asymptotic **p**-value calculated with the χ^2 approximation. Figure 10 shows the results of the accumulated beta adjusted with estimated parameters $a = 0.9121$ and $b = 1.2843$.

We also used the relationship described in Section 3 to estimate **p**- from **e**-values. To find it, we computed the inverse cdf of a χ^2 distribution with four degrees of freedom, which is $\dim(\Theta)$ in this case, for the **e**-value obtained. Then, we evaluated this result with a χ^2 cdf with three degrees of freedom, $\dim(\Theta) - \dim(\Theta_H)$. This would be our approximation for the **p**-value. Figure 11 shows the results with the accumulated beta adjusted with estimated parameters $a = 0.8535$ and $b = 1.3868$.

6. CONCLUDING REMARKS

p- and **e**- values handle the same problems in a very similar way: (i) the **p**-value is a tail from ***x***, the observation, given **H**, the null hypothesis; and (ii) the **e**-value is a tail from **H** given ***x***. This introduces a duality between them. The following questions can be naturally asked:

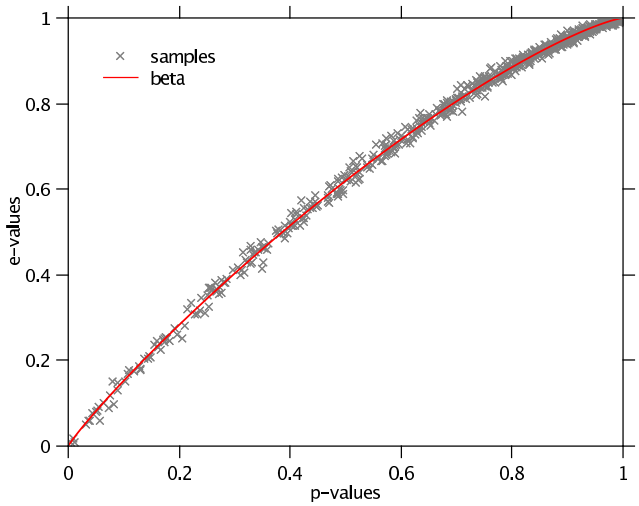

FIG. 10: Behrens-Fisher problem with asymptotic **p**-values.

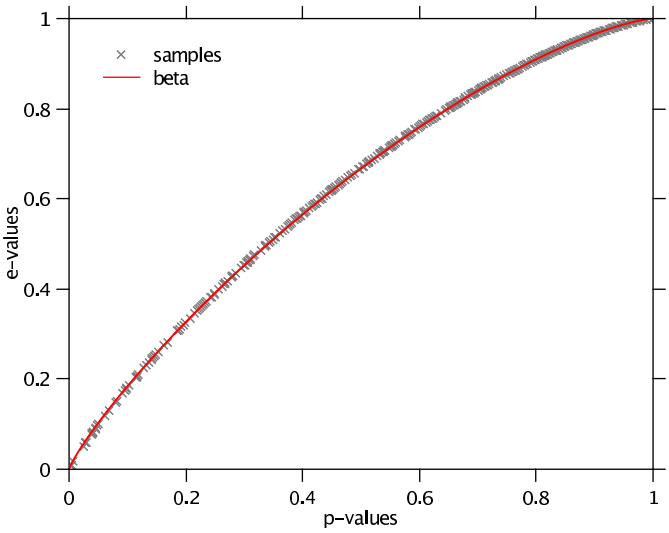

FIG. 11: Behrens-Fisher problem with **p**-values obtained from the relationship reported in section 3.

1. Is there any deterministic function relating the two indices?
2. What kind of difficulties arise for the case of discrete sample spaces?
3. What are the advantages to favor one index in detriment of the other?

These questions encouraged the authors to write this paper and tentative answers are presented below. Answering the two first questions, one should be able to respond the last one appropriately.

Answer to question 1. Theorem 1 shows that there exists a relationship that is valid asymptotically. We cannot write it explicitly because the χ^2 distribution function, and its inverse, does not have a closed form. Simulation exercises indicates that the relation between **p** and **e** is well approximated by the accumulated beta distribution function. The beta fitting does not describe the relation between the indices for the variance test when the sample size is small.

Answer to question 2. Recall, from Section 2, that we are restricted to the cases where all posterior probability spaces are absolutely continuous; hence, there are densities. Whenever we must work with discrete sample space, problems arise in the **p**-value computation: some include the observed value in the tail and some exclude it; see the binomial example in Section 2.

Answer to question 3. The two indices are very similar, but the computation of exact **p**-values usually depend on the whole sample space: it may, in contrast to the **e**-value, violate the Likelihood principle. Considering these answers and the results shown in this paper, we recommend that one may comfortably always use the **e**-value.

The **e**-value has other desirable properties for a Bayesian significance test, namely: it is a probability value derived from the posterior distribution on the full parameter space; an invariant for alternative parametrizations; neither requires nuisance parameters elimination nor the assignment of positive prior probabilities to sets of zero Lebesgue measure, contrary to tests based on Bayes factors; and is a formal Bayesian test and, as such, its critical values can be obtained from considered loss functions.

REFERENCES

1. Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling, *Philos. Mag.*, 50:157–175, 1990.
2. Fisher, R. A., On the mathematical foundations of theoretical statistics, *Philos. Trans. Royal Soc., A*, 222:309–368, 1922.
3. Neyman, J. and Pearson, E. S., On the use and interpretation of certain test criteria for purposes of statistical inference: Part I, *Biometrika*, 20:175–240, 1928.

4. Cox, D. R., The role of significant test (with discussion), *Scand. J. Stat.*, 4:49–70, 1977.
5. Kempthorne, O., Of what use are tests of significance and tests of hypothesis, *Commun. Stat. Theory Methods*, 5:763–777, 1976.
6. Pereira, C. A. B. and Stern, J. M., Evidence and credibilidty: Full Bayesian significance test of precise hypothesis, *Entropy J.*, 1:99–110, 1999.
7. Pereira, C. A. B., Stern, J. M., and Wechsler, S., Can a Significance Test Be a Genuinely Bayesian?, *Bayesian Anal.*, 3:79–100, 2008.
8. Chakrabarti, D., A Novel Bayesian Mass Determination Algorithm, *AIP Conf. Proc.*, 1082:317–323, 2008.
9. Johnson, R., Chakrabarti, D., O'Sullivan, E., and Raychaudhury, S., Comparing *x*-ray and dynamical mass profiles in the early-type galxy NGC-4636, *Astrophys. J.*, 706:980–994, 2009.
10. del Rincon, S. V., Rogers, J., Widschwendter, M., Sun, D., Sieburg, H. B., and Spruk, C., Development and validation of a method for profiling post-translational modification activities using protein microarrays, *PLoS ONE*, 5:e11332, 2010.
11. Berger, J. and Delampady, M., Testing precise hypotheses (with discussion), *Stat. Sci.*, 2:317–352, 1987.
12. Berger, J., Boukai, B., and Wang, Y., Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Stat. Sci.*, 12:133–160, 1997.
13. Aitikin, M., Boys, R., and Chadwick, T., Bayesian point null hypothesis testing via the posterior likelihood ratio, *Stat. Comput.*, 15:217–230, 2005.
14. Irony, T. Z. and Pereira, C. A. B., Exact tests for equality of two proportions: Fisher v. Bayes, *J. Stat. Comput. Simul.*, 25:93–114, 1986.
15. Rice, K., A decision-theoretic formulation of Fishers approach to testing, *Am. Stat.*, 64(4):345–349, 2010.
16. Pereira, C. A. B. and Wechsler, S., On the concept of *p*-value, *Brazil. J. Prob. Stat.*, 7:159–177, 1993.
17. Dempster, A. P., The direct use of likelihood for significance testing, *Stat. Comput.*, 7:247–252, 1997.
18. Wilks, S. S., The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann. Math. Stat.*, 9:60–62, 1938.
19. Casella, G. and Berger, R., *Statistical Inference*, Duxbury, Pacific Grove, 2002.
20. Berger, J. O. and Wolpert, R. L., *The Likelihood Principle*, IMS Lecture-Notes, Hayward, 1988.
21. Schervish, M., *Theory of Statistics*, Springer, New York, 1995.