Wiley StatsRef: Statistics Reference Online

E-Value

By *Marcio A. Diniz*[1], *Carlos A. de Bragança Pereira*[2], *and Julio M. Stern*[2]

Keywords: *Bayesian inference, hypothesis testing, significance test, sharp hypothesis*

Abstract: This article describes a coherent Bayesian measure of evidence for precise or sharp null hypotheses, the evidence value, or e-value, derived from the fully Bayesian significance test (FBST), based solely on the posterior distribution of the parameters of the statistical model. The method can be easily implemented using modern numerical optimization and integration techniques. After illustrating its use on two nonstandard applications (Hardy–Weinberg equilibrium and the Behrens–Fisher problem), we list some of its properties and refer the interested reader to articles discussing further applications and properties of the e-value.

In several applied fields, theoretical models derived from first principles are based on specific parameter values or on functional relationships between them that must be empirically tested in order to validate such models and their predictions. When the mentioned tests address hypotheses in which all or some of the parameters must assume a specific set of values, they are called *significance tests*. These tests do not apply to situations in which it is necessary to test hypotheses that are more general than those related to specific parameter values: these hypotheses are called *sharp hypotheses*. The Full Bayesian Significance Test (FBST) was firstly proposed by Pereira and Stern[1] mainly to test sharp hypotheses. Other than its several interesting properties – discussed in Refs 2–4 – the procedure is completely based on posterior distributions, thus avoiding complications such as the elimination of nuisance parameters or the necessity to use prior distributions that assign positive probabilities to sets of zero Lebesgue measure.

Before proceeding with illustrative examples, let us first fix the notation in order to define sharp hypotheses and the other concepts needed to describe how to compute the FBST evidence value, or simply e-value, supporting a sharp hypothesis. We denote by $\mathbf{X}$ the vector of random variables to be observed in a given experiment, and by $\mathcal{X}$ the sample space, that is, the set of all possible values the random vector may assume. It is usually the case that $\mathcal{X} \subseteq \mathbb{R}^n$, where $n \in \mathbb{N}$ is the sample size of the experiment. The lowercase $\mathbf{x}$ is a specific sample point or simply referred to as the data observed from the experiment. In this article, we assume that the random vector is modeled by a parametric statistical model, $\mathcal{M}$, defined as

$$\mathcal{M} = \{f(\mathbf{x} \mid \theta) : \mathbf{x} \in \mathcal{X}, \theta \in \Theta\}$$

[1]Federal University of São Carlos, São Carlos, Brazil
[2]University of São Paulo, São Paulo, Brazil

DOI: 10.1002/9781118445112.stat08375

in which f is a probability mass or density function from a specified model (e.g., Gaussian, Gamma, Weibull) etc. and $\Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$, is the parameter space (set of all values the parameter(s) may assume). In this framework, we say the dimension of the parameter space is k, $\dim(\Theta) = k$.

Definition 1. *A sharp hypothesis* $\mathbf{H}$ *is the statement* $\theta \in \Theta_H$ *in which* $\Theta_H \subset \Theta$*, and the dimension of* Θ_H*, dim*$(\Theta_H) = h$*, is strictly smaller than the dimension* k *of* Θ*.*

Thus, as an example, considering the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $i = 1, \dots, n$ and ε_i i.i.d. $N(0, \sigma^2)$, hypotheses such as $H : \beta_0 = 0$ and $H : \beta_0 + \beta_1 = 1$ are both sharp, although only the first defines a significance test.

Following the Bayesian paradigm, let $\pi(\cdot)$ be a probability prior distribution over Θ, and $L(\cdot \mid \mathbf{x})$ the likelihood function derived from data $\mathbf{x}$ and model $\mathscr{M}$. In order to compute the evidence value or e-value that supports a sharp hypothesis $\mathbf{H}$ based on the FBST, the most important entity is the posterior distribution (*see* **Posterior Distribution: Introduction**) for θ given the sample $\mathbf{x}$, here denoted $g(\theta \mid \mathbf{x})$:

$$g(\theta \mid \mathbf{x}) \propto \pi(\theta) \cdot L(\theta \mid \mathbf{x})$$

Even though the procedure may be used when the parameter space is discrete, we must emphasize that it is when the posterior distribution over Θ is absolutely continuous that the FBST presents its most interesting properties to test sharp hypotheses. To simplify notation, we denote Θ_H by $\mathbf{H}$ in the sequel.

In order to guarantee that the e-value is invariant to reparameterizations, it is necessary to specify a distribution on the parameter space called reference density, $r(\theta)$. With this density we obtain the *relative surprise function*, the ratio between the posterior density and the reference density, that is, $s(\theta) = g(\theta \mid \mathbf{x})/r(\theta)$ (*see* Good [5, p. 145–146] and **Surprise Index**). The surprise function guarantees the invariance under reparameterizations of θ, even when $r(\theta)$ is improper (*see* Stern [4, p. 253]). Thus, if $r(\theta)$ is proportional to any given constant the surprise function will be, in practical terms, equivalent to the posterior distribution. It is possible to compute the e-value using other reference densities such as neutral, invariant, maximum-entropy, or noninformative distributions whenever they are available and/or desirable for the problem under analysis.

Definition 2. *Given a sharp hypothesis* $\mathbf{H}$*:* $\theta \in \Theta_H$*, the tangent set of the hypothesis given data* $\mathbf{x} \in \mathcal{X}$ *is*

$$\mathbb{T}_{\mathbf{x}} = \{\theta \in \Theta : s(\theta) > s^*\}$$

in which $s^* = \sup_{\theta \in \mathbf{H}} s(\theta)$*.*

Thus, the tangent set is the subset of the parameter space with points whose relative surprise, $s(\theta)$, is larger than the relative surprise of any point in $\mathbf{H}$, being *tangential* to $\mathbf{H}$ in this sense. The e-value favoring a sharp hypothesis $\mathbf{H}$ is then defined as the posterior probability of the complementary set, regarding the parameter space, of the tangent set, that is, $\Theta - \mathbb{T}_{\mathbf{x}}$.

Definition 3. *The Bayesian e-value supporting a sharp hypothesis* $\mathbf{H}$ *is*

$$ev = 1 - P(\theta \in \mathbb{T}_{\mathbf{x}} \mid \mathbf{x}) = 1 - \int_{\mathbb{T}_{\mathbf{x}}} \mathrm{d}G_{\mathbf{x}}(\theta)$$

in which $G_{\mathbf{x}}$ *denotes the posterior distribution function of* θ*, and the integral is of the Riemann–Stieltjes type.*

Thus, the e-value considers that the posterior probability of all points of the parameter space whose relative surprise is at most as large as its supremum over $\mathbf{H}$ is the Bayesian evidence supporting $\mathbf{H}$. Given

DOI: 10.1002/9781118445112.stat08375

this, a large value of ev means that Θ_H lies in a region of large posterior probability, implying that the data strongly support the hypothesis. On the other hand, whenever ev is small, this means that Θ_H is in a region of the parameter space with low posterior probability, implying that the data is leading us to discredit $\mathbf{H}$. Nevertheless, ev is not evidence against a global alternative hypothesis $\mathbf{A} : \theta \notin \Theta_{\mathbf{H}}$, which is not sharp. Similarly, $1 - ev$ is not evidence supporting $\mathbf{A}$, even though it is against $\mathbf{H}$.

The procedure implied by Definitions 1–3 may be summarized by the following algorithm.

1. Specify the statistical model, the correspondent likelihood, and the prior distribution on Θ (*see* **Prior Distribution Elicitation**; **Bayesian Methods**).
2. Specify the reference density, $r(\theta)$, and derive the relative surprise function, $s(\theta)$.
3. Optimization: compute s^*, the maximum value of $s(\theta)$ under the constraint $\theta \in \mathbf{H}$.
4. Integration: compute the posterior probability on the complement of the tangent set, $\Theta - \mathbb{T}_{\mathbf{x}}$. This is the e-value supporting $\mathbf{H}$.

Some remarks, specially about steps 3 and 4, seem in order. In step 3, one should find the point of the parameter space in $\mathbf{H}$ that maximizes $s(\theta)$, that is, to solve a problem of constrained maximization. In several applications it is not possible to find a closed-form solution for these problems, requiring the use of numerical optimizers. Step 4 requires the integration of the posterior distribution on a subset of Θ, the tangent set $\mathbb{T}_{\mathbf{x}}$, that can be highly complex. As in the previous step, since in many cases it is difficult to find an explicit expression for $\mathbb{T}_{\mathbf{x}}$, the use of numerical techniques to compute the integral is the easiest way to proceed. If it is possible to generate random samples from the posterior distribution, Monte Carlo integration (*see* **Bayesian Analysis and Markov Chain Monte Carlo Simulation**) provides an accurate estimate of ev. It is also possible to use approximation techniques, such as those proposed in Tierney and Kadane[6], based on Laplace approximations. Refs 7, 8 show how to implement such approximations for unit root and cointegration tests of time series. We now illustrate the computation of the e-value in some applications.

Example 1. (Hardy–Weinberg equilibrium law). Consider a sample of $n \in \mathbb{N}$ individuals of the same species randomly drawn from their population. We represent by x_1 and x_3 the two homozygote sample counts of a given genetic locus, and by $x_2 = n - x_1 - x_3$ the heterozygotic individuals in the sample, such that $\mathbf{x} = (x_1, x_2, x_3)$. In this trinomial model, the parameter space is the simplex:

$$\Theta = \{\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3 : \theta_1 + \theta_2 + \theta_3 = 1, \theta_i \geq 0, i = 1, 2, 3\}$$

where θ_i is the population relative frequency of individuals with genotype $i = 1, 2, 3$. The Hardy–Weinberg law (HWL),[9], assumes that, under equilibrium, these frequencies are such that $\theta_3 = (1 - \sqrt{\theta_1})^2$. Thus, the hypothesis to test the HWL is $\mathbf{H} : \Theta \in \Theta_H$ in which

$$\Theta_H = \{\theta \in \Theta : \theta_3 = (1 - \sqrt{\theta_1})^2\}$$

defines a subset of the parameter space with zero Lebesgue measure. To keep this numerical illustration as simple as possible, let us adopt as prior distribution over Θ the uniform distribution on the simplex, that is, a Dirichlet density with its three parameters equal to one, and as reference density also the uniform distribution on the simplex such that the surprise function is proportional to the likelihood

$$L(\theta \mid \mathbf{x}) \propto \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}$$

In this framework it is possible to find a closed-form solution to $\sup_{\theta \in \mathbf{H}} s(\theta)$, and the computation of ev is carried by Monte Carlo integration from independent vectors sampled from a Dirichlet distribution with parameters $x_1 + 1$, $x_2 + 1$, and $x_3 + 1$. Assuming, for instance, $n = 20$, $x_1 = 5$, and $x_3 = 5$, the estimated

DOI: 10.1002/9781118445112.stat08375

Table 1. E-values for the Behrens–Fisher problem with different sample means m_2

m_2	ev_J	ev_u
100	0.00	0.00
101	0.08	0.07
102	0.53	0.50
103	0.89	0.87
104	0.98	0.98
105	1.00	1.00

e-value is 0.91. The problem of testing the Hardy–Weinberg equilibrium law using Bayes factors requires the specification of a mixed prior on Θ_H. Frequentist alternatives are the Chi-square goodness-of-fit test (with continuity correction) and the asymptotic likelihood ratio test, both described in Lauretto *et al.*[10] and references therein.

Example 2. (Behrens–Fisher problem). Two samples are observed from independent populations, both with Gaussian distributions with unknown parameters, that is, sample 1 from $N(\mu_1, \sigma_1^2)$ and sample 2 from $N(\mu_2, \sigma_2^2)$. Thus,

$$\Theta = \{(\mu_1, \sigma_1, \mu_2, \sigma_2) \in (\mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+)\}$$

and the hypothesis being tested is $\mathbf{H} : \mu_1 = \mu_2$. Madruga *et al.*[11] computed the e-value for the Behrens–Fisher problem adopting standard (independent) improper priors for the means, μ_1, μ_2, and the precisions, $1/\sigma_1^2, 1/\sigma_2^2$. For the reference density they used Jeffreys' prior density (*see* **Jeffreys' Prior Distribution**) and also the (improper) uniform density over Θ to compare the results.

They assumed a sample from population 1 with size, sample mean, and sample standard deviation, respectively, of $n_1 = 16, m_1 = 100$, and $s_1 = 3$ and a sample from population 2 with size $n_2 = 20$, sample standard deviation $s_2 = 3$, and different values for the sample mean, m_2. The results are displayed in Table 1, where ev_J denotes the e-value computed with Jeffrey's prior as reference density, and ev_u the e-value computed with the (improper) uniform on Θ. Frequentist alternatives to test the Behrens–Fisher hypothesis are the asymptotic likelihood ratio test and the Welsh approximation of the t-test [12, p. 208–209].

An important practical problem that must be addressed in applications of the FBST is the search of a threshold value below which one may reject **H**. In principle, one may use the sampling distribution of ev to find this threshold value: this can be done because ev is formally a statistic whose distribution can be derived from the adopted statistical model. If the likelihood and the posterior distribution satisfy certain regularity conditions, mentioned in Schervish [13, p. 436], Diniz *et al.*[14] proved that, asymptotically, there is a relationship between ev and p-values obtained from the likelihood ratio test used to test the same hypotheses. This result provides an alternative way to compute, at least for large samples, a critical value to ev to reject the hypothesis being tested.

In a recent review, Stern and Pereira[3], the authors discuss different ways to compute a threshold for ev. Among these alternatives, we highlight the standardized e-value, which follows, asymptotically, the uniform distribution on (0,1). See Borges and Stern[15] for more on the standardized version of ev.

Another alternative is to define the FBST as a Bayes test derived from a particular loss function (*see* **Utility Function**) and the respective minimization of the posterior expected loss. Following this strategy, Madruga *et al.*[16] showed that there are loss functions that result in the e-value as a Bayes estimator of $\phi = \mathbb{I}_H(\theta)$, where $\mathbb{I}_A(x)$ denotes the indicator function, being equal to 1 if $x \in A$, and 0 otherwise, $x \notin A$. Thus, the FBST is in fact a Bayes procedure in the formal sense as defined in Wald[17].

DOI: 10.1002/9781118445112.stat08375

Concluding Remarks

We have briefly defined and illustrated the use of the FBST e-value, or evidence value, which provides a measure of statistical support of sharp hypotheses. The e-value has several desirable properties, of which we underline:

1. it provides an intuitive and simple measure of support for the hypothesis in test, ideally, a probability defined directly on the original parameter space;
2. it requires neither the elimination of nuisance parameters nor *ad hoc* artifices such as the assignment of positive prior probabilities to zero measure sets and the setting of arbitrary initial belief ratios between hypotheses;
3. it obeys the likelihood principle, that is, the information conveyed by the sample should be represented by, and only by, the likelihood function[18];
4. it is invariant for alternative parameterizations;
5. it is an exact procedure, that is, it is not necessary to use of "large sample" asymptotic approximations to compute the e-value;
6. it is a formal Bayes test, and as such, its critical values may be obtained from the adopted loss function;
7. it is a possibilistic support measure for sharp hypotheses, complying with the *Onus Probandi* juridical principle (*In Dubio Pro Reo* rule);
8. it is a homogeneous computation calculus with the same two steps, constrained optimization and integration of the posterior density.

The e-value has been used in several applied works done in the past two decades, of which we highlight those on economics, biology and medicine, systems reliability, signal processing and detection of acoustic events, astronomy, and astrophysics. This list is far from exhaustive, but the interested reader may consult reference Stern and Pereira[3] for an extensive catalog of applications and of theoretical articles that discuss statistical and logical properties of the e-value. It is important to mention that, up to date, in all the mentioned applications computing time was not a great burden whenever FBST was used. More recently, Ly and Wagenmakers[19] examined some potential shortcomings of the e-value, while Kelter[20] has critically reevaluated their article, replying to some of their conclusions.

Related Articles

Posterior Distribution: Introduction; **Surprise Index**; **Prior Distribution Elicitation**; **Bayesian Methods**; **Bayesian Analysis and Markov Chain Monte Carlo Simulation**; **Jeffreys' Prior Distribution**; **Utility Function**

References

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

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

[3] Stern, J.M. and Pereira, C.A.B. (2020) *The e-Value: A Fully Bayesian Significance Measure for Precise Statistical Hypotheses and its Research Program*. São Paulo Journal of Mathematical Sciences, *https://link.springer.com/article/10.1007%2Fs40863-020-00171-7* (accessed 20 August 2020).

[4] Stern, J.M. (2010) *Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses in Natural Sciences*. arXiv: 1006.5471, *https://arxiv.org/abs/1006.5471* (accessed on 20 August 2020).

DOI: 10.1002/9781118445112.stat08375

[5] Good, I.J. (1983) *Good Thinking: The Foundations of Probability and Its Applications*, University of Minnesota Press, Minneapolis, MN.

[6] Tierney, L. and Kadane, J.B. (1986) Accurate approximation for posterior moments and marginal densities. *J. Am. Stat. Assoc.*, **81**, 82–86.

[7] Diniz, M.A., Pereira, C.A.B., and Stern, J.M. (2011) Unit roots: Bayesian significance test. *Commun. Stat. Theory Meth.*, **40**, 4200–4213.

[8] Diniz, M.A., Pereira, C.A.B., and Stern, J.M. (2012) Cointegration: Bayesian significance test. *Commun. Stat. Theory Meth.*, **41**, 3562–3574.

[9] Hardy, G.H. (1908) Mendelian proportions in a mixed population. *Science*, **28**, 49–50.

[10] Lauretto, M.S., Nakano, F., Faria Jr., S.R. *et al*. (2009) A straightforward multiallelic significance test for the Hardy-Weinberg equilibrium law. *Genet. Mol. Biol.*, **32**, 619–625.

[11] Madruga, M.R., Pereira, C.A.B., and Stern, J.M. (2003) Bayesian evidence test for precise hypotheses. *J. Stat. Plan. Inference*, **117**, 185–198.

[12] Lehmann, E.L. (1986) *Testing Statistical Hypothesis*, Wiley, New York, NY.

[13] Schervish, M. (1995) *Theory of Statistics*, Springer, New York, NY.

[14] Diniz, M.A., Pereira, C.A.B., Polpo, A. *et al*. (2012) Relationship between Bayesian and frequentist significance indices. *Int. J. Uncertain. Quantif.*, **2**, 161–172.

[15] Borges, W. and Stern, J.M. (2007) The rules of logic composition for the Bayesian epistemic E-values. *Log. J. IGPL*, **15**, 401–420.

[16] Madruga, M.R., Esteves, L.G., and Wechsler, S. (2001) On the Bayesianity of Pereira–Stern tests. *Test*, **10**, 291–299.

[17] Wald, A. (1950) *Statistical Decision Functions*, John Wiley and Sons, New York, NY.

[18] Berger, J.O. and Wolpert, R.L. (1988) *The Likelihood Principle*, Institute of Mathematical Statistics, Hayward, CA.

[19] Ly, A. and Wagenmakers, E.J. (2021) A critical evaluation of the FBST ev for Bayesian hypothesis testing. *Comput. Brain Behav.*, DOI: 10.1007/s42113-021-00109-y.

[20] Kelter, R. (2021) On the measure-theoretic premises of bayes factor and full Bayesian significance tests: a critical reevaluation. Commentary to Ly and Wagenmakers. *Comput. Brain Behav.*, DOI: 10.1007/s42113-021-00110-5.

DOI: 10.1002/9781118445112.stat08375