Statistical Methods & Applications
https://doi.org/10.1007/s10260-024-00745-1

ORIGINAL PAPER

A new Bayesian discrepancy measure

Francesco Bertolino[1] · Mara Manca[1] · Monica Musio[1] · Walter Racugno[1] · Laura Ventura[2]

Accepted: 14 January 2024

Abstract
The aim of this article is to make a contribution to the Bayesian procedure of testing precise hypotheses for parametric models. For this purpose, we define the Bayesian Discrepancy Measure that allows one to evaluate the suitability of a given hypothesis with respect to the available information (prior law and data). To summarise this information, the posterior median is employed, allowing a simple assessment of the discrepancy with a fixed hypothesis. The Bayesian Discrepancy Measure assesses the compatibility of a single hypothesis with the observed data, as opposed to the more common comparative approach where a hypothesis is rejected in favour of a competing hypothesis. The proposed measure of evidence has properties of consistency and invariance. After presenting the definition of the measure for a parameter of interest, both in the absence and in the presence of nuisance parameters, we illustrate some examples showing its conceptual and interpretative simplicity. Finally, we compare a test procedure based on the Bayesian Discrepancy Measure, with the Full Bayesian Significance Test, a well-known Bayesian testing procedure for sharp hypotheses.

Keywords Bayesian test · Evidence · Precise hypothesis · Significance test · Full Bayesian significance test

1 Introduction

D. V. Lindley in Lindley (1965) (preface page xi) stated that

> " ...hypothesis testing looms large in standard statistical practice, yet scarcely appears as such in the Bayesian literature."

Since then things have changed and, in the last sixty years, there have been several attempts to build a measure of evidence that covers, in a Bayesian context, the role that the *p-value* has played in the frequentist setting. A prominent example is

Extended author information available on the last page of the article

Published online: 23 February 2024

the decision test based on the Bayes Factor and its extensions (see, for instance, Berger (1985)).

As an alternative to the Bayes Factor, another Bayesian evidence measure is provided in Pereira and Stern (1999) upon which the testing procedure Full Bayesian Signicance Test (FBST) is based. For a recent survey on the FBST see Pereira and Stern (2020).

The main aim of this paper is to give a contribution to the testing procedure of precise hypotheses. In particular, the proposed Bayesian measure of evidence, called Bayesian Discrepancy Measure (BDM), gives an absolute evaluation of a hypothesis H in light of prior knowledge about the parameter and observed data. The proposed measure of evidence has the desired properties of invariance under reparametrization and consistency for large samples.

Our starting point is the idea that a hypothesis may be more or less supported by the available evidence contained in the posterior distribution.

We do not adopt the hypothesis testing approach for which there is no test that can lead to the rejection of a hypothesis except by comparing it with another hypothesis (Neyman-Pearson in the frequentist perspective, Bayes factor in the Bayesian one), but rather the approach proposed by Fisher (see Christensen (2005) and Deni (2004)). Reference is made to a precise hypothesis H and no alternative is considered against such hypothesis. In this view different hypotheses made by several experts can be evaluated and using the information coming from the same data, some can be accepted others not. In this respect, in a broad sense, we can say that we return to Fisher's original idea of pure significance according to which "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher 1925).

The proposed measure of evidence can be seen as a Bayesian tool for model checking, that is, as a technique that can aid in the actual specification of a model, without the need to make explicit reference to alternative models or hypotheses. For an extensive discussion of this point and the difference with the procedure of Bayesian model selection see O'Hagan (2003).

The structure of the paper is as follows. In Sect. 2 the definition of the proposed measure is presented for a scalar parameter of interest, both in the absence or presence of nuisance parameters. In Sect. 3 different illustrative examples are discussed, involving one or two independent populations. Finally, in Sect. 4 we make a comparison between the Bayesian Discrepancy Test and the Full Bayesian Significance Test which is based on the e-value, a well-known Bayesian evidence index used to test sharp hypotheses. The last section contains conclusions and directions for further research.

2 The Bayesian discrepancy measure

Let $(\mathcal{X}, \mathcal{P}_{\theta}^{X}, \boldsymbol{\Theta})$ be a parametric statistical model where $X \in \mathcal{X} \subset \mathbb{R}^k$, $\mathcal{P}_{\theta}^{X} = \{f(x|\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is a class of probability distributions (Lebesgue integrable) defined on $\mathcal{X}$, depending on an unknown vector of continuous parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, an open subset of $\mathbb{R}^p$. Assume that

(a) the model is identifiable;
(b) $f(x|\boldsymbol{\theta})$ have support not depending on $\boldsymbol{\theta}$, $\forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}$;
(c) the log likelihood function must be at least twice differentiable;
(d) the operations of integration and differentiation with respect to $\boldsymbol{\theta}$ can be exchanged.

We assume a prior probability density $g_0(\boldsymbol{\theta})$ following Cromwell's Rule which states that "it is inadvisable to attach probabilites of zero to uncertain events, for if the prior probability is zero so is the posterior, whatever be the data. A probability of one is equally dangerous because then the probability of the complementary event will be zero" (see Section 6.2 in Lindley (1991)). We are then assuming that $g_0(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta}$ for you will need that assumption for claiming consistency (see Proposition 1 of the next Section).

First, we discuss the case of a scalar parameter. Then we discuss the case of a scalar parameter of interest in the presence of nuisance parameters.

2.1 The Bayesian discrepancy measure for a scalar parameter

In this section we assume that $k = p = 1$. Given an *iid* random sample $\boldsymbol{x} = (x_1, \dots, x_n)$ from $\mathcal{P}_\theta^X$, let $L(\theta|\boldsymbol{x})$ be the corresponding likelihood function based on data $\boldsymbol{x}$ and let $g_0(\theta)$ be a continuous prior distribution on $\Theta \subseteq \mathbb{R}$. The posterior probability density for θ given $\boldsymbol{x}$ is then

$$g_1(\theta|\boldsymbol{x}) \propto g_0(\theta)\, L(\theta|\boldsymbol{x}).$$

Moreover, given the posterior distribution function $G_1(\theta|\boldsymbol{x})$, the posterior median is any real number m_1 which satisfies the inequalities $G_1(m_1|\boldsymbol{x}) \geq \frac{1}{2}$ and $G_1^-(m_1|\boldsymbol{x}) \leq \frac{1}{2}$, where $G_1^-(m_1|\boldsymbol{x}) = \lim\limits_{\theta \uparrow m_1} G_1(\theta|\boldsymbol{x})$. In the case in which $G_1(\cdot|\boldsymbol{x})$ is continuous and strictly increasing we have $m_1 = G_1^{-1}(\frac{1}{2}|\boldsymbol{x})$. Under the assumptions made in the beginning of Sect. 2, posterior median m_1 is uniquely defined.

We are interested in testing the precise hypothesis

$$H\, :\, \theta = \theta_H. \tag{1}$$

In order to measure the discrepancy of the hypothesis (1) w.r.t. the posterior distribution, in the case $\Theta = \mathbb{R}$, we consider the following two intervals:

1. the *discrepancy interval*

$$I_H = \begin{cases} (m_1, \theta_H) & \text{if} \quad m_1 < \theta_H \\ \{m_1\} & \text{if} \quad m_1 = \theta_H, \\ (\theta_H, m_1) & \text{if} \quad m_1 > \theta_H \end{cases} \tag{2}$$

2. the *external interval*

$$I_E = \begin{cases} (\theta_H, +\infty) & \text{if} \quad m_1 < \theta_H \\ (-\infty, \theta_H) & \text{if} \quad \theta_H < m_1. \end{cases} \tag{3}$$

When $m_1 = \theta_H$, the external interval I_E can be $(-\infty, m_1)$ or $(m_1, +\infty)$. Note that, by construction, $\mathbb{P}(I_H \cup I_E) = \frac{1}{2}$ (see Fig. 1). If the support of the posterior is a subset of $\mathbb{R}$, the intervals I_H and I_E can be defined consequently.

Definition 1 Given the posterior distribution function $G_1(\theta|\boldsymbol{x})$, we define the Bayesian Discrepancy Measure of the hypothesis H as

$$\delta_H = 2\,\mathbb{P}(\theta \in I_H|\boldsymbol{x}) = 2 \int_{I_H} dG_1(\theta|\boldsymbol{x}). \tag{4}$$

The measure can be also computed by means of the external interval as

$$\delta_H = 1 - 2\,\mathbb{P}(\theta \in I_E|\boldsymbol{x}) = 1 - 2 \int_{I_E} dG_1(\theta|\boldsymbol{x}), \tag{5}$$

which can also be written as

$$\delta_H = 1 - 2 \min\{G_1^-(\theta_H|\boldsymbol{x}), 1 - G_1(\theta_H|\boldsymbol{x})\}, \tag{6}$$

where $G_1^-(\theta_H|\boldsymbol{x}) = \lim\limits_{\theta \uparrow \theta_H} G_1(\theta|\boldsymbol{x})$. In our case, since $G_1(\theta_H|\boldsymbol{x})$ is continuous, this simplifies to

$$\delta_H = 1 - 2 \min\{G_1(\theta_H|\boldsymbol{x}), 1 - G_1(\theta_H|\boldsymbol{x})\}. \tag{7}$$

Formulations (6) and (7) have the advantage of not involving the posterior median in the integral computation. Furthermore, one can interpret the quantity $\min\{G_1(\theta_H|\boldsymbol{x}), 1 - G_1(\theta_H|\boldsymbol{x})\}$ as the posterior probability of a “tail" event concerning only the precise hypothesis H. Doubling this “tail" probability, related to the precise hypothesis H, one gets a posterior probability assessment about how “central" the hypothesis H is and hence how it is supported by the prior and the data.

It is important to highlight that the hypothesis H induces the following partition

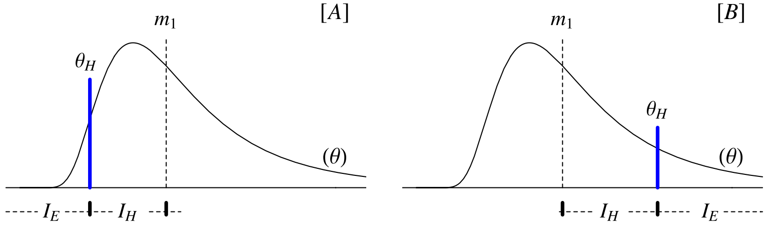

Fig. 1 Posterior density $g_1(\theta|\boldsymbol{x})$, the corresponding discrepancy interval I_H and external interval I_E when $\theta_H < m_1$ ([A]) and $\theta_H > m_1$ ([B])

$$\left\{\Theta_a = (-\infty, \theta_H),\ \Theta_H = \{\theta_H\},\ \Theta_b = (\theta_H, \infty)\right\} \tag{8}$$

of the parameter space Θ. Then formulations (6) and (7) can be equivalently expressed as

$$\delta_H = \ 1 - 2 \cdot \min_{a,b} \left\{\mathbb{P}(\theta \in \Theta_a|\boldsymbol{x}),\, \mathbb{P}(\theta \in \Theta_b|\boldsymbol{x})\right\}. \tag{9}$$

The last formula can be naturally extended to the case where, besides the scalar parameter of interest, nuisance parameters are also present. This issue will be developed in Sect. 2.2.

The following properties apply to the BDM, for a scalar parameter θ.

Proposition 1

(i) *δ_H always exists and, by construction, $\delta_H \in [0, 1]$;*

(ii) *δ_H is invariant under invertible monotonic transformations of the parameter θ;*

(iii) *if θ is an a.c. random variable, θ^* is the true value of the parameter and $\theta^* = \theta_H$, then δ_H converges asymptotically to a Unif$(\cdot|0, 1)$. Otherwise, if $\theta^* \neq \theta_H$, then $\delta_H \xrightarrow{p} 1$ (consistency property).*

Proof *(i)* The first property follows immediately from the fact that in (4) the posterior probability $\mathbb{P}(\theta \in I_H|\boldsymbol{x}) \in \left[0, \frac{1}{2}\right]$.

(ii) Let $\lambda = \lambda(\theta)$ be an invertible monotonic transformation of the parameter θ and let $K_1(\cdot)$ be the cumulative distribution function of the parameter λ. We denote with $\lambda_H = \lambda(\theta_H)$ and we notice that $m_1' = \lambda(m_1)$ thanks to the monotonic invariance of the median. Suppose, for simplicity, that $\theta_H > m_1$. Then

$$\delta_H = 2 \int_{m_1}^{\theta_H} dG_1(\theta|\boldsymbol{x}) \ = 2 \left| \int_{m_1'}^{\lambda_H} dK_1(\lambda|\boldsymbol{x}) \right|.$$

Therefore, the invariance of the BDM follows immediately from the invariance of the median under invertible monotonic transformations. Notice that if instead of the median m_1 we consider, for example, the posterior mean $E(\theta|\boldsymbol{x})$, which is not invariant under invertible monotonic reparametrizations, the property will not hold in general. Moreover, $E(\theta|\boldsymbol{x})$ for some models may not even exist.

(iii) We first examine the first part of the statement for which $\theta^* = \theta_H$. Let $J(\hat{\theta})$ be the observed Fisher information and let $\hat{\theta}$ be the maximum likelihood estimator of θ. Under suitable regularity and technical conditions

(see for instance Section 7, p. 129 in Lindley (1965) and Section 5.3.2, p. 287 in Bernardo and Smith (1994)), the asymptotic distribution of the "normalized" random quantity $W = \sqrt{J(\hat{\theta})}(\theta - \hat{\theta})$ is standard normal, both in the posterior, for fixed data and random θ, and in the sampling distribution, for fixed θ and random data. We have

$$\delta_H = 1 - 2\min\{G_1(\theta_H|\boldsymbol{x}), 1 - G_1(\theta_H|\boldsymbol{x})\}, \tag{10}$$

where

$$G_1(\theta_H|\boldsymbol{x}) = P(W \leq \sqrt{J(\hat{\theta})}(\theta_H - \hat{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}). \tag{11}$$

Since W is asymptotically standard normal, then $G_1(\theta_H|\boldsymbol{x})$ is asymptotically $\Phi\left(\sqrt{J(\hat{\theta})}(\theta_H - \hat{\theta})\right)$ (a function of the data through $\hat{\theta}$). But also, in the sampling distribution given $\theta^* = \theta_H$, $\sqrt{J(\hat{\theta})}(\theta_H - \hat{\theta})$ is asymptotically standard normal and thus, in view of the probability integral transform, $G_1(\theta_H|\boldsymbol{X})$ is asymptotically uniform on [0, 1] in this sampling distribution. Then

$$\mathbb{P}(\delta_H \leq t|\theta_H) = \mathbb{P}\left(\frac{1}{2}(1-t) \leq G_1(\theta_H|\boldsymbol{X}) \leq \frac{1}{2}(1+t)|\theta_H\right) \approx t,$$

so that δ_H is asymptotically uniform under θ_H. If, instead, $\theta^* \neq \theta_H$ and $n \to \infty$, under suitable regularity conditions (see for instance Section 7, p. 129 in Lindley (1965)) it is well known that $g_1(\theta|\boldsymbol{x})$ is concentrated in a neighbourhood whose size is of order $n^{-\frac{1}{2}}$ around θ^*. Then from equation 5, since the tail event $\theta \in I_E$ will have vanishingly small probability, we have that $\lim_{n\to\infty} \delta_H = 1$.

□

As pointed out before, the further θ_H is from the posterior median m_1 of the distribution function $G_1(\theta|\boldsymbol{x})$, the closer δ_H is to 1. It can then be said that H does *not conform* to $G_1(\theta|\boldsymbol{x})$. On the contrary, the smaller δ_H the stronger is the evidence in favor of H. Following this idea, we can construct a procedure to evaluate (and possibly reject) the hypothesis H, using the evidence measure δ_H.

Definition 2 The Bayesian Discrepancy Test (BDT) is the procedure for evaluating a hypothesis H based on the Bayesian Discrepancy Measure (BDM).

High values of δ_H provide strong evidence against the hypothesis H. On the other hand, if δ_H is small, the data are consistent with H.

Summarizing, when H is true, then, for large n, δ_H is roughly equally likely to fall anywhere between 0 and 1. By contrast, when H is false, δ_H is more likely to be near 1 than near 0. As for other measures of evidence (as for the Full Bayesian Significance Test or the frequentist p-value), a threshold could be chosen in order to interpret the observed value of δ_H. However, in the direction recommended in the ASA statement (see Wasserstein and Lazar (2016)) and in view of the debate on hypothesis testing (Benjamin et al. 2018; Benjamin and Berger 2019) and the recent studies about the reproducibility of experiments (Collaboration 2015; Johnson et al. 2017), we agree with Fisher (1973) that “no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas". Given the critical points related to the choice of a threshold, we think it is important to look for an applied measure of evidence that pushes the researcher to think more about the specific problem, and that avoids the use of standard receipes.

2.2 The Bayesian discrepancy measure in presence of nuisance parameters

Suppose that $p \geq 2$ and $k \geq 1$. Let $\varphi = \varphi(\boldsymbol{\theta})$ be a scalar parameter of interest, where $\varphi : \boldsymbol{\Theta} \to \Phi \subseteq \mathbb{R}$. Let us further consider a bijective reparametrization $\boldsymbol{\theta} \Leftrightarrow (\varphi, \boldsymbol{\zeta})$, where $\boldsymbol{\zeta} \in \boldsymbol{Z} \subseteq \mathbb{R}^{p-1}$ denotes an arbitrary nuisance parameter, which is determined on the basis of analytical convenience (note that the value of the evidence measure is invariant with respect to the choice of the nuisance parameter). We consider hypotheses that can be expressed in the form

$$H : \varphi = \varphi_H, \tag{12}$$

where φ_H is known as it represents the hypothesis that it is of interest to evaluate. The transformation φ must be such that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and for all $\varphi_H \in \Phi$, it can always be assessed whether φ is strictly smaller, strictly larger or equal to φ_H (i.e. $\varphi < \varphi_H$ either $\varphi > \varphi_H$, or $\varphi = \varphi_H$). Hypothesis (12) and transformation φ, with

$$\begin{array}{ll} \boldsymbol{\Theta}_a & = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \varphi < \varphi_H\right\} \\ \boldsymbol{\Theta}_H & = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \varphi = \varphi_H\right\}. \\ \boldsymbol{\Theta}_b & = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \varphi > \varphi_H\right\} \end{array} \tag{13}$$

We call any hypothesis of type (12), which identify a partition of the form (13), a *partitioning hypothesis*. It is easy to verify that many commonly used hypotheses are partitioning. In this paper we only consider hypotheses of this nature. In this setting, we express the BDM as

$$\begin{aligned} \delta_H &= 1 - 2 \cdot \min_{a,b} \left\{\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_a | \boldsymbol{x}), \, \mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_b | \boldsymbol{x})\right\} \\ &= 1 - 2 \cdot \int_{I_E} g_1(\boldsymbol{\theta}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{14}$$

where *the external set* is given by

$$I_E = \arg\min_{a,b} \left\{\mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_a | \boldsymbol{x}), \, \mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_b | \boldsymbol{x})\right\} . \tag{15}$$

In the particular scenario where the marginal posterior

$$h_1(\varphi|\boldsymbol{x}) = \int_{\varphi(\boldsymbol{\theta})=\varphi} g_1(\boldsymbol{\theta}|\boldsymbol{x}) \mathrm{d}\boldsymbol{\theta} \, , \quad \forall \varphi \in \Phi \, ,$$

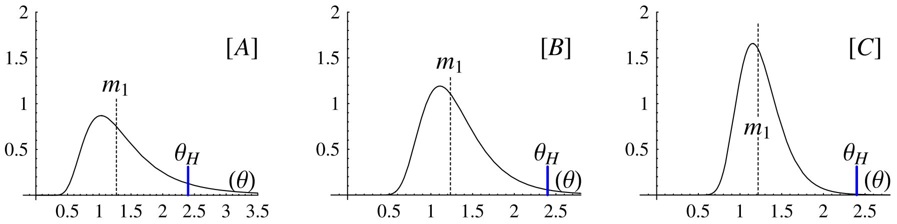

Fig. 2 Posterior density function $g_1(\theta|n\bar{x})$ and intervals $I_H = (m_1, \theta_H)$ and $I_E = (\theta_H, \infty)$, using data from Example 1

of the parameter of interest φ can be computed in a closed form, the hypothesis (12) can be easily treated using the methodologies seen in Subsection **2.1**, i.e. the BDM is computed by means of formula (4) or (5) applied to the marginal.

Properties reported in Proposition 1 naturally extend to the setting we just presented.

3 Illustrative examples

The simplicity of the BDT is highlighted by the following examples, some of which deal with cases not usually considered in the literature. Examples 1 and 2 focus on a scalar parameter of interest, while Examples 3, 4, 5, 6, 7 also contain nuisance parameters.

In all examples we have adopted a Jeffreys' prior (see Yang and Berger (1996) for a catalog of non-informative priors) for simplicity. However, other objective priors and, in the presence of substantive prior information, informative priors could equally be used.

3.1 Examples of the univariate parameter case

Example 1 *Exponential distribution* Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be an *iid* sample of size n from the Exponential distribution $X \sim Exp\big(x|\theta^{-1}\big)$, with $\theta \in \mathbb{R}^+$. We are interested in the hypothesis $H : \theta = \theta_H$. Assuming a Jeffreys' prior for θ, i.e. $g_0(\theta) \propto \theta^{-1}$, the posterior distribution is given by $g_1(\theta|\boldsymbol{x}) \propto \theta^{-n-1} \exp\{-n\bar{x} \cdot \theta^{-1}\}$, with $\bar{x}$ the sample mean.

Figure 2 shows the posterior density function as well as the discrepancy and the external intervals for $H : \theta = \theta_H = 2.4$ and the MLE $\bar{x} = 1.2$ for three sample sizes $[A]$ $n = 6$, $[B]$ $n = 12$, $[C]$ $n = 24$. In $[A]$ we have a posterior median $m_1 = 1.27$ and $\delta_H = 0.832$, while in $[B]$ $m_1 = 1.23$ and $\delta_H = 0.960$, in $[C]$ $m_1 = 1.22$ and $\delta_H = 0.997$.

While in case $[A]$ the data do not contradict H sufficiently, in case $[B]$ there is a weak evidence against H, which becomes stringer in $[C]$.

Note that in all scenarios considered, we find the following relation between δ_H and the *p-value*,

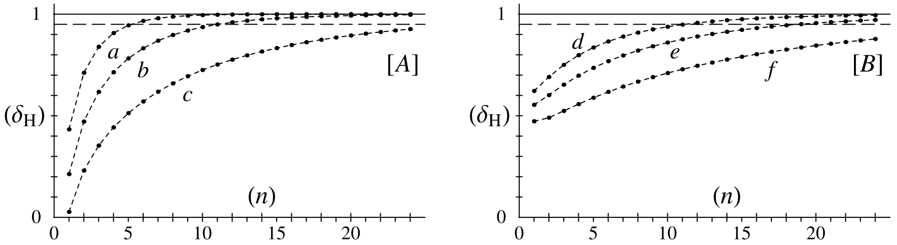

Fig. 3 BDM for n increasing and for different values of the MLE. Case [A] with MLE = 0.8 (**a**), 1.2 (**b**), 1.6 (**c**) and case [B] with MLE = 3.2 (**f**), 3.6 (**e**), 4 (**d**)

$$p\text{-}value = 1 - \delta_H \tag{16}$$

(in [A] $\delta_H = 0.832$ and *p-value*= 0.168, in [B] $\delta_H = 0.96$ and *p-value*= 0.04, while in [C] $\delta_H = 0.997$ and *p-value*= 0.003). This result depends clearly on the use of the Jeffreys' prior, which is a matching prior for a scalar parameter (see Ruli and Ventura (2021)).

Remark 1 The fact that classical and Bayesian procedures, under certain conditions, produce the same conclusions is well known (see, for instance, Lindley (1965)). The linear relationship (16) also occurs in other simple cases.

Finally, to conclude Example 1, it is useful to show the trend of the BDM when varying $n = 1, 2, \ldots, 25$ for six values of the MLE: (a) 0.8, (b) 1.2, (c) 1.6 (case [A]) and (d) 4.0, (e) 3.6, (f) 3.2 (case [B]), see Fig. 3. In order to explain the difference between the BDM trends in cases [A] and [B], consider that:

(i) in case [A] the posterior median $m_1 < \theta_H = 2.4$, whereas in case [B] $m_1 > \theta_H = 2.4$;
(ii) δ_H is monotonically increasing, both with respect to n, and with respect to the distance $|m_1 - \theta_H|$;
(iii) the posterior g_1 always has a positive asymmetry, which decreases as n increases;
(iv) the trend difference of the BDM in cases [A] and [B] depends on the fact that the posterior g_1 has 'small' tails on the left-hand side of m_1 and 'large' tails on the right-hand side.

Moving forward in the discussion, in order to highlight the evaluative nature of the BDT, it is worth pointing out that it allows the separate and simultaneous testing of $\ell \geq 2$ hypotheses

$$H_j\,:\,\varphi = \varphi_j, \quad j = 1, 2, \ldots, \ell, \tag{17}$$

as shown in Example 2. Remember that with the comparative approach, among the ℓ competing hypotheses, only one is accepted. On the contrary, under the evaluative approach, it may happen that several hypotheses are supported by the data, or even that all hypotheses must be rejected.

Example 2 - *Evaluation of some hypotheses made by several experts (Bernoulli distribution)* In the 1700 s, several hypotheses $H_j : \theta = \theta_j$ were formulated about the birth masculinity rate $\theta = \frac{M}{M+F}$. Among them we consider $\theta_1 = \frac{1}{2}$ (J. Bernoulli), $\theta_2 = \frac{13}{25}$ (J. Arbuthnot), $\theta_3 = \frac{1050}{2050}$ (J. P. Süssmilch), $\theta_4 = \frac{23}{45}$ (P. S. Laplace). We assume that the gender of each newborn is modeled as a $Bin(\cdot|1, \theta)$. Then, using data recorded in 1710 in London (see, for instance, Spiegelhalter (2019)), with 7640 males and 7288 females (the MLE is $\hat{\theta} = 0.512$) and assuming the Jeffreys' prior $Beta(\theta|1, 1)$, we compute δ_{H_j} using the Normal asymptotic approximation

$$\delta_{H_j} \cong 1 - 2 \cdot \int_{I_E^j} \tilde{g}_1\big(\theta|\hat{\theta}, \tfrac{1}{n}\hat{\theta}(1-\hat{\theta})\big)\mathrm{d}\theta, \quad j = 1, 2, 3, 4,$$

with $\tilde{g}_1$ the Normal distribution. Since $\delta_{H_1} = 0.996$, $\delta_{H_2} = 0.955$, $\delta_{H_3} = 0.079$, $\delta_{H_4} = 0.132$, we can conclude that there is sufficient evidence against the first two hypotheses, while there is not enough evidence agains the hypotheses made by Süssmilch and Laplace.

3.2 Examples of the more general case

The examples presented hereafter, can be distinguished by tests concerning a parameter or a parametric function of a single population, and tests concerning the comparison of two independent population parameters.

3.2.1 Tests involving a single population

Example 3 - *Test on the shape parameter, mean and variance of the Gamma distribution* Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be an *iid* sample of size n from $X \sim Gamma\big(x|\alpha, \beta\big)$, $(\alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+$. We denote by m_g the geometric mean of $\boldsymbol{x}$. The likelihood function for (α, β) is given by

$$L(\alpha, \beta|\boldsymbol{x}) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \cdot m_g^\alpha \cdot e^{-\bar{x}\cdot\beta}\right)^n.$$

For the fictitious data $\boldsymbol{x} = (0.8, 1.1, 1.2, 1.4, 1.8, 2, 4, 5, 8)$, we find that the MLEs are $\hat{\alpha} = 1.921$ and $\hat{\beta} = 0.7572$.

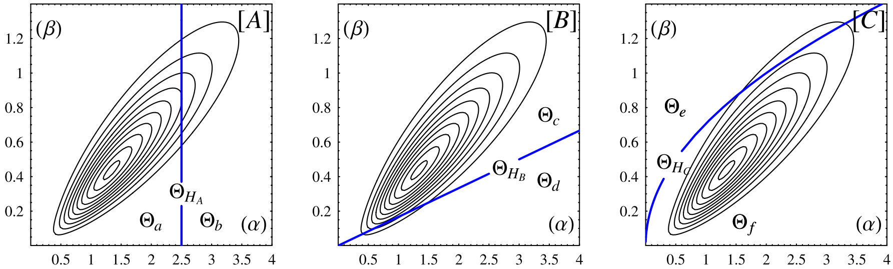

Fig. 4 Posterior density function $g_1(\alpha, \beta|\boldsymbol{x})$ from Example 3 and corresponding sets of the induced partition in the cases $[A]$, $[B]$ and $[C]$

We are interested in testing the hypotheses $[A]$ $H_A : \alpha = \alpha_H$, with $\alpha_H = 2.5$, $[B]$ $H_B : \mu = \mu_H$, with $\mu_H = 6$, and $[C]$ $H_C : \sigma^2 = \sigma_H^2$, with $\sigma_H^2 = 2$, where $\mu = \dfrac{\alpha}{\beta}$ and $\sigma^2 = \dfrac{\alpha}{\beta^2}$ denote the mean and the variance of X.

We suppose that the parameters α and β are independent and we assume the Jeffreys' prior for them (see Yang and Berger (1996)), i.e. $g_0(\alpha, \beta) = g_0^\alpha(\alpha) \cdot g_0^\beta(\beta)$ where $g_0^\alpha(\alpha) \propto \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1}$, $g_0^\beta(\beta) \propto \frac{1}{\beta}$, and $\psi^{(1)}(\alpha) = \sum_{j=0}^{\infty}(\alpha + j)^{-2}$ denotes the *digamma* function. Then, the posterior for (α, β) is given by $g_1(\alpha, \beta \mid \boldsymbol{x}) = k \cdot g_0^\alpha(\alpha) \cdot g_0^\beta(\beta) \cdot L(\alpha, \beta|\boldsymbol{x})$, with normalizing constant k.

- Case $[A]$ The hypothesis H_A identifies the vertical straight line of equation $\alpha = \alpha_H$ and two subsets $\boldsymbol{\Theta}_a = \{(\alpha, \beta) : \alpha < \alpha_H\}$ and $\boldsymbol{\Theta}_b = \{(\alpha, \beta) : \alpha > \alpha_H\}$ (see Fig. 4 [A]). Then we can compute

$$\begin{aligned}\mathbb{P}\big((\alpha, \beta) \in \boldsymbol{\Theta}_b \mid \boldsymbol{x}\big) &= \int_{\alpha_H}^{\infty} \int_0^{\infty} g_1(\alpha, \beta \mid \boldsymbol{x})\, \mathrm{d}\beta\, \mathrm{d}\alpha \\ &= k \cdot \int_{\alpha_H}^{\infty} \int_0^{\infty} \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1} \cdot \frac{1}{\beta} \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \cdot m_g^\alpha \cdot e^{-\bar{x} \cdot \beta} \right)^n \mathrm{d}\beta\, \mathrm{d}\alpha \\ &= k \cdot \int_{\alpha_H}^{\infty} \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\alpha)^n} \cdot \left(\frac{m_g}{n\,\bar{x}} \right)^{n\alpha} \mathrm{d}\alpha \;=\; 0.215\,,\end{aligned}$$

  and $\delta_H = 0.570$, indicating that there is not enough evidence against H.
- Case $[B]$ The hypothesis H_B identifies the straight line of equation $\beta = \frac{1}{\mu_H}\alpha$ in the $\alpha\beta$-plane (see Fig. 4 $[B]$) and the two subsets

$$\boldsymbol{\Theta}_c = \Big\{(\alpha, \beta) : \beta > \frac{1}{\mu_H}\alpha\Big\}\; and\, \boldsymbol{\Theta}_d = \Big\{(\alpha, \beta) : \beta < \frac{1}{\mu_H}\alpha\Big\}.$$

  We have

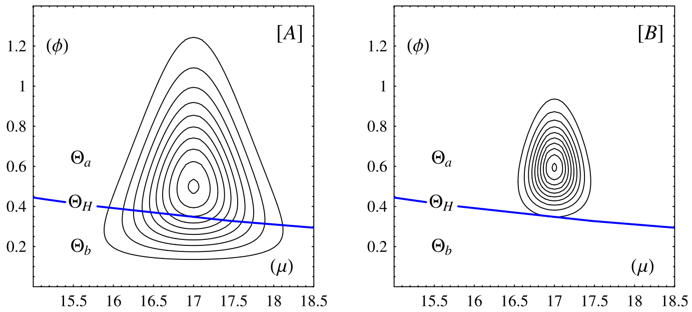

Fig. 5 Test on the coefficient of variation ψ of a Gaussian population. Data refers to Example 4. In the plots, the sets $\mathbf{\Theta}_a$, $\mathbf{\Theta}_b$ and $\mathbf{\Theta}_H$ are reported for $n = 10$ ([A]) and $n = 40$ ([B])

$$\mathbb{P}\big((\alpha,\beta) \in \mathbf{\Theta}_d \mid \boldsymbol{x}\big) \;=\; \int_{\mathbf{\Theta}_d} g_1(\alpha,\beta \mid \boldsymbol{x})\,\mathrm{d}\alpha\,\mathrm{d}\beta \;=\; 0.012\,,$$

and, since $\delta_H = 0.976$, we have strong evidence against H_B.

- Case [C] The hypothesis H_C identifies the parabola of equation $\beta = \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha}$, in the $\alpha\beta$-plane (see Fig. 4 [C]), and the two subsets

$$\mathbf{\Theta}_e = \Big\{(\alpha,\beta) \,:\, \beta > \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha}\Big\}\; and\, \mathbf{\Theta}_f = \Big\{(\alpha,\beta) \,:\, \beta < \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha}\Big\}.$$

We have

$$\mathbb{P}\big((\alpha,\beta) \in \mathbf{\Theta}_e \mid \boldsymbol{x}\big) \;=\; \int_{\mathbf{\Theta}_e} g_1(\alpha,\beta \mid \boldsymbol{x})\,\mathrm{d}\alpha\,\mathrm{d}\beta \;=\; 0.078\,.$$

Therefore $\delta_H = 0.846$, and so we have not strong evidence against H_C.

Example 4 - *Test on the coefficient of variation for a Normal distribution* Given an *iid* sample $\boldsymbol{x} = (x_1, \dots, x_n)$ from $X \sim N\big(x|\mu, \phi^{-1}\big)$, the parameter of interest is $\psi = \frac{\sqrt{Var(X)}}{\mid \mathbb{E}(X) \mid} = \frac{1}{\mid \mu \mid \sqrt{\phi}}$. We are interested in testing the hypothesis

$$H \,:\, \psi = \psi_H,$$

with $\psi_H = 0.1$. If we consider the Jeffreys' prior $g_0(\mu,\phi) \propto \phi^{-1} \cdot \mathbf{1}_{\mathbb{R}\times\mathbb{R}^+}$, the posterior distribution is the Normal-Gamma density

$$(\mu,\phi) \mid \boldsymbol{x} \;\sim\; NG\big(\mu,\phi \mid \eta,\nu,\alpha,\beta\big)\,,$$

with hyperparameters (η,ν,α,β), where $\eta = \bar{x}$, $\nu = n$, $\alpha = \frac{1}{2}(n-1)$, $\beta = \frac{1}{2}ns^2$, and density

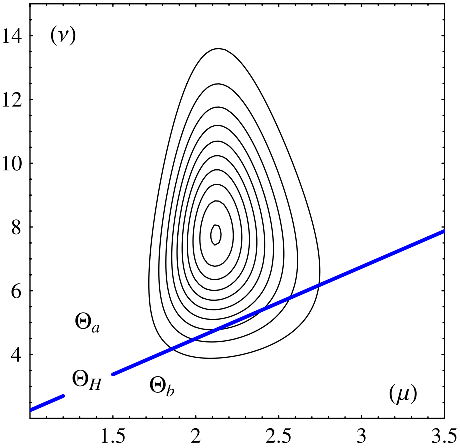

Fig. 6 Test on the skewness of the Inverse Gaussian distribution with $\gamma_H = 2$. In the plot the sets of the partition induced by H are reported. Data refers to Example 5

$$g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi}} \phi^{\alpha-1/2} e^{-\frac{\nu\phi}{2}(\mu-\eta)^2} e^{-\beta\phi}.$$

We consider the particular case in which $\bar{x} = 17$ and $s^2 = 1.6$ (so that the MLE is $\hat{\phi} = 0.074$) with two samples of size $n = 10$ (Fig. 5 [A]) and $n = 40$ (Fig. 5 [B]). In the $\mu\phi$–space, the hypothesis H is represented by the curve $\phi = \frac{1}{\psi_H^2}\mu^{-2}$ and determines the subsets $\mathbf{\Theta}_a$ and $\mathbf{\Theta}_b$ visualized in Fig. 5.

In case [A] we have

$$\mathbb{P}\big((\mu, \phi) \in \mathbf{\Theta}_b \mid \boldsymbol{x}\big) = \int_{\mathbf{\Theta}_b} g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta)\, \mathrm{d}\mu\, \mathrm{d}\phi = 0.215,$$

where $g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta)$ is the Normal-Gamma density, so that $\delta_H = 0.570$ and there is not enough evidence against H. In case [B], we have $\mathbb{P}\big((\mu, \phi) \in \mathbf{\Theta}_b \mid \boldsymbol{x}\big) = 0.014$ and, since $\delta_H = 0.972$, there is strong evidence against H. Therefore in such a case, with different sample sizes, the inferential conclusions change (Fig. 6).

Example 5 - *Test on the skewness coefficient of the Inverse Gaussian distribution* Let us consider a Inverse Gaussian random variable X with density

$$f(x \mid \mu, \nu) = \sqrt{\frac{\nu}{2\pi x^3}} \exp\left\{ -\frac{\nu}{2}\left(\frac{x-\mu}{\mu\sqrt{x}}\right)^2 \right\} \cdot \mathbf{1}_{\mathbb{R}^+}(x),$$

where $(\mu, \nu) \in \mathbb{R}^+ \times \mathbb{R}^+$. The parameter of interest is the skewness coefficient $\gamma = 3\sqrt{\frac{\mu}{\nu}}$ and it is of interest to test the hypothesis $H : \gamma = \gamma_H$, where $\gamma_H = 2$. The Jeffreys' prior is

$$g_0(\mu, \nu) \propto \frac{1}{\sqrt{\mu^3 \nu}} \cdot \mathbf{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(\mu, \nu).$$

Given n observations, the posterior distribution of (μ, ν) is

$$g_1(\mu, \nu | \boldsymbol{x}) \ \propto \ \sqrt{\frac{\nu^{n-1}}{\mu^3}} \cdot \exp\left\{ -\frac{n\,\nu}{2} \cdot \left(\frac{\bar{x}}{\mu^2} - \frac{2}{\mu} + \frac{1}{a} \right) \right\} \cdot \mathbf{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(\mu, \nu),$$

where $\bar{x}$ and a are the arithmetic and harmonic mean, respectively.

We apply the procedure to the following rainfall data (inches) analyzed in Folks and Chhikara (1978) (p. 272):

1.01	1.11	1.13	1.15	1.16
1.17	1.17	1.20	1.52	1.54
1.54	1.57	1.64	1.73	1.79
2.09	2.09	2.57	2.75	2.93
3.19	3.54	3.57	5.11	5.62.

The hypothesis identifies in the parameter space $\boldsymbol{\Theta} = \mathbb{R}^+ \times \mathbb{R}^+$ the subsets

$$\begin{aligned} \boldsymbol{\Theta}_a &= \left\{ (\mu, \nu) \in \boldsymbol{\Theta} \, : \, 3\sqrt{\tfrac{\mu}{\nu}} < \gamma_H \right\}, \\ \boldsymbol{\Theta}_H &= \left\{ (\mu, \nu) \in \boldsymbol{\Theta} \, : \, 3\sqrt{\tfrac{\mu}{\nu}} = \gamma_H \right\}, \\ \boldsymbol{\Theta}_b &= \left\{ (\mu, \nu) \in \boldsymbol{\Theta} \, : \, 3\sqrt{\tfrac{\mu}{\nu}} > \gamma_H \right\}. \end{aligned}$$

We have that

$$\mathbb{P}\big((\mu, \nu) \in \boldsymbol{\Theta}_b \mid \boldsymbol{x}\big) \ = \ \int_{\boldsymbol{\Theta}_b} g_1(\mu, \nu | \boldsymbol{x}) \, \mathrm{d}\mu \, \mathrm{d}\nu = 0.078\,, \tag{18}$$

see Fig. 6, then we obtain $\delta_H = 0.844$. This result indicates that we do not have enough evidence against the hypothesis H.

3.2.2 Tests involving two independent populations

In this section we consider some examples concerning comparisons between parameters of two independent populations.

Example 6 - *Comparison between means and precisions of two independent Normal populations* Let us consider a case study on the dating of the core and periphery of some wooden furniture, found in a Byzantine church, using radiocarbon (see Casella and Berger (2001), p. 409). The historians wanted to verify if the mean age of the core is the same as the mean age of the periphery, using two samples of sizes $m = 14$ and $n = 9$, respectively, given by

core	1294	1279	1274	1264	1263	*periphery*	1284	1272	1256
	1254	1251	1251	1248	1240		1254	1242	1274
	1232	1220	1218	1210			1264	1256	1250

We assume that the age of the core X and of the periphery Y are distributed as

$$X \sim N(x|\mu_1, \phi_1^{-1}) \quad \text{and} \quad Y \sim N(y|\mu_2, \phi_2^{-1}),$$

where $Var(X) = \phi_1^{-1}$ and $Var(Y) = \phi_2^{-1}$, and we assume that the data are *iid* conditional on the parameters. We consider for (μ_i, ϕ_i) the Jeffreys' prior

$$g_0^i(\mu_i, \phi_i) \propto \phi_i^{-1} \cdot \mathbf{1}_{\mathbb{R}\times\mathbb{R}^+}\ , \ i = 1, 2\,.$$

We obtain $\bar{x} = 1249.86$, $\bar{y} = 1261.33$, $\bar{d} = \bar{x} - \bar{y} = -11.48$, while the MLEs for the sample standard deviations are $s_1 = 23.43$ and $s_2 = 12.51$. The posterior distribution for (μ_i, ϕ_i) is the Normal-Gamma law

$$(\mu_i, \phi_i) \mid \boldsymbol{x}, \boldsymbol{y} \ \sim \ NG\big(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i\big)\ , \ i = 1, 2,$$

with hyperparameters $\eta_1 = \bar{x}$, $\nu_1 = m$, $\alpha_1 = \frac{1}{2}(m-1)$, $\beta_1 = \frac{1}{2} m s_1^2$ and $\eta_2 = \bar{y}$, $\nu_2 = n$, $\alpha_2 = \frac{1}{2}(n-1)$, $\beta_2 = \frac{1}{2} n s_2^2$, and density

$$g_1^i(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i}\sqrt{\nu_i}}{\Gamma(\alpha_i)\sqrt{2\pi}} \phi_i^{\alpha_i - 1/2} e^{-\frac{\nu_i \phi_i}{2}(\mu_i - \eta_i)^2} e^{-\beta_i \phi_i}, \ i = 1, 2.$$

The hypothesis of interest

$$H_A\ :\ \mu_1 - \mu_2 = 0, \quad \forall \phi_1 > 0, \quad \forall \phi_2 > 0,$$

identifies the following subsets in the parameter space

$$\begin{aligned} \boldsymbol{\Theta}_a &= \Big\{\ \mathbb{R}^2 \times \mathbb{R}_+^2\ :\ \mu_1 < \mu_2\ \Big\}, \\ \boldsymbol{\Theta}_{H_A} &= \Big\{\ \mathbb{R}^2 \times \mathbb{R}_+^2\ :\ \mu_1 = \mu_2\ \Big\}, \\ \boldsymbol{\Theta}_b &= \Big\{\ \mathbb{R}^2 \times \mathbb{R}_+^2\ :\ \mu_1 > \mu_2\ \Big\}. \end{aligned}$$

Then we can compute

$$\begin{aligned} &\mathbb{P}\big((\mu_1, \mu_2, \phi_1, \phi_2) \in \boldsymbol{\Theta}_a \mid \boldsymbol{x}, \boldsymbol{y}\big) \\ &\qquad = \int_{\boldsymbol{\Theta}_a} \prod_{i=1}^{2} g_1^i(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i)\, \mathrm{d}\mu_1\, \mathrm{d}\mu_2\, \mathrm{d}\phi_1\, \mathrm{d}\phi_2 \\ &\qquad = \int_{\mu_1 < \mu_2} \prod_{i=1}^{2} \frac{\Gamma(\alpha_i + \frac{1}{2})}{\Gamma(\alpha_i)} \Big(\frac{\nu_i}{2\pi\beta_i}\Big)^{1/2} \Big[1 + \frac{\nu_i}{2\beta_i}(\mu_i - \eta_i)^2\Big]^{-(\alpha_i + \frac{1}{2})} \mathrm{d}\mu_1\, \mathrm{d}\mu_2 \\ &\qquad = 0.089\,, \end{aligned}$$

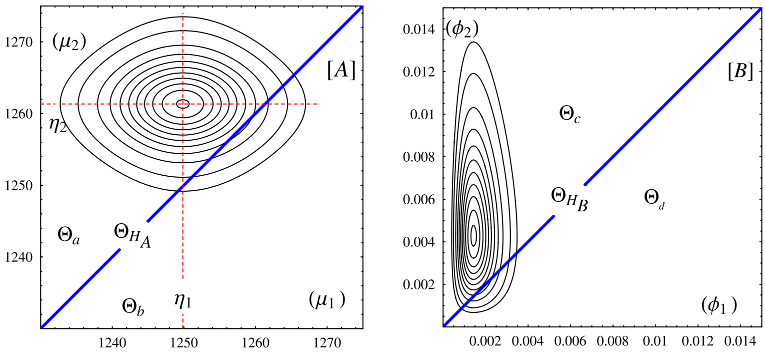

Fig. 7 Comparisons between means ([A]) and precisions ([B]) of independent normal populations for data in Example 6. For both cases we show the contour plots of the marginals of μ_j ([A]) and ϕ_j ([B]), and the partition sets associated with the corresponding hypotheses

so we have $\delta_H = 0.823$, a value that does not indicate evidence against the hypothesis. We exploited the fact that the marginal of each μ_i is a Generalized Student's t-distribution (denoted by *StudentG*) with hyperparameters $\left(\eta_i, \frac{\nu_i\alpha_i}{\beta_i}, 2\alpha_i\right)$.

Figure 7 $[A]$ in the space (μ_1, μ_2) shows the contour lines of the distribution

$$StudentG\left(\mu_1 \middle| \eta_1, \frac{\nu_1 \cdot \alpha_1}{\beta_1}, 2\alpha_1\right) \cdot StudentG\left(\mu_2 \middle| \eta_2, \frac{\nu_2 \cdot \alpha_2}{\beta_2}, 2\alpha_2\right).$$

Note that the homoscedasticity assumption is not necessary. Consider now the hypothesis

$$H_B : \phi_1 - \phi_2 = 0, \quad \forall \mu_1, \mu_2,$$

which determines in the parameter space the subsets

$$\begin{aligned} \boldsymbol{\Theta}_c &= \left\{ \mathbb{R}^2 \times \mathbb{R}^2_+ : \phi_1 < \phi_2 \right\}, \\ \boldsymbol{\Theta}_{H_B} &= \left\{ \mathbb{R}^2 \times \mathbb{R}^2_+ : \phi_1 = \phi_2 \right\}, \\ \boldsymbol{\Theta}_d &= \left\{ \mathbb{R}^2 \times \mathbb{R}^2_+ : \phi_1 > \phi_2 \right\}. \end{aligned}$$

We have

$$\mathbb{P}\left((\mu_1, \mu_2, \phi_1, \phi_2) \in \boldsymbol{\Theta}_c \mid \boldsymbol{x}, \boldsymbol{y}\right) = \int_{\phi_1 < \phi_2} \prod_{i=1}^{2} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \phi_i^{\alpha_i - 1} e^{-\phi_i \beta_i} \, \mathrm{d}\phi_1 \, \mathrm{d}\phi_2 = 0.046,$$

from which it follows that $\delta_H = 0.908$ and there is strong evidence against the hypothesis H. To compute the integral we have used the fact that the marginal of each ϕ_i has Gamma distribution with parameters (α_i, β_i), $i = 1, 2$.

The contour lines of the law $Gamma(\phi_1|\alpha_1, \beta_1) \cdot Gamma(\phi_2|\alpha_2, \beta_2)$, in the space (ϕ_1, ϕ_2), are reported in Figure 7 $[B]$.

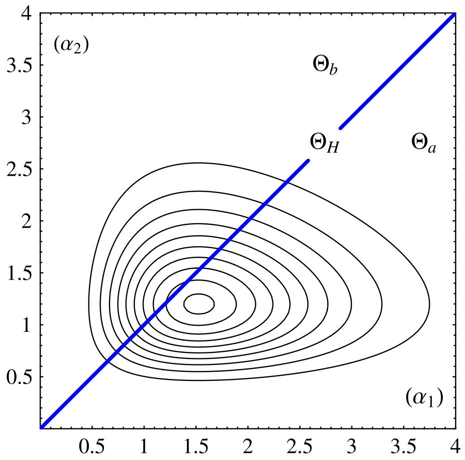

Fig. 8 Comparison of the shape parameters of two independent Gamma populations, using data of Example 7. The sets Θ_a, Θ_b and Θ_H of the partition are reported

Example 7 - *Comparison of the shape parameter of two Gamma distributions* Let us consider two *iid* Gamma populations $X_i \sim Gamma\big(\alpha_i, \beta_i\big)$, $(\alpha_i, \beta_i) \in \mathbb{R}^+ \times \mathbb{R}^+$, $i = 1, 2$, and let us consider two samples of sizes $n_1 = 9$ and $n_2 = 12$, respectively, with sample means $\bar{x}_1 = 2.811$ and $\bar{x}_2 = 1.973$, and geometric means $m_{g_1} = 2.116$ and $m_{g_2} = 1.327$.

We are interested in testing $H : \alpha_1 = \alpha_2$. The posterior distribution for $(\alpha_1, \beta_1, \alpha_2, \beta_2)$ is given by

$$g_1(\alpha_1, \beta_1, \alpha_2, \beta_2 | \boldsymbol{x}_1, \boldsymbol{x}_2) = g_1^1(\alpha_1, \beta_1 | \boldsymbol{x}_1) \cdot g_1^2(\alpha_2, \beta_2 | \boldsymbol{x}_2) \, ,$$

where

$$g_1^i(\alpha_i, \beta_i | \boldsymbol{x}_i) = k_i \cdot g_0(\alpha_i, \beta_i) \cdot L(\alpha_i, \beta_i \mid \boldsymbol{x}_i),$$

with normalizing constant k_i, $i = 1, 2$. Let $\boldsymbol{\Theta}_a = \big\{(\alpha_1, \alpha_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \, : \, \alpha_1 > \alpha_2\big\}$ and $\boldsymbol{\Theta}_b = \big\{(\alpha_1, \alpha_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \, : \, \alpha_1 < \alpha_2\big\}$ (see Figure 8). In order to test the hypothesis H, we compute the probability

$$\begin{aligned} &\mathbb{P}((\alpha_1, \alpha_2) \in \boldsymbol{\Theta}_b \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \\ &= \int_{\alpha_1 < \alpha_2} \int_{\mathbb{R}^+ \times \mathbb{R}^+} g_1^1(\alpha_1, \beta_1 | \boldsymbol{x}_1) \cdot g_1^2(\alpha_2, \beta_2 | \boldsymbol{x}_2) \, \mathrm{d}\beta_1 \mathrm{d}\beta_2 \, \mathrm{d}\alpha_1 \, \mathrm{d}\alpha_2 \\ &= \int_{\alpha_1 < \alpha_2} \prod_{i=1}^{2} k_i \cdot g_0^{\alpha}(\alpha_i) \cdot \frac{\Gamma(n_i \alpha_i)}{\Gamma(\alpha_i)^{n_i}} \cdot \left(\frac{m_{g_i}}{n_j \bar{x}_i} \right)^{n_i \alpha_i} \mathrm{d}\alpha_1 \mathrm{d}\alpha_2 = 0.311 \end{aligned}$$

and, since $\delta_H = 0.378$, there is evidence in favour of H.

4 Comparison with the FBST

In this section we present a comparison of the BDT with the Full Bayesian Significance Test (FBST) as presented in Pereira and Stern (2020), which provides an overview of the e-value.

In order to facilitate the discussion, let us briefly review the definition of the e-value and the related testing procedure. The FBST can be used with any standard parametric statistical model, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. It tests a sharp hypothesis H which identifies the null set Θ_H. The conceptual approach of the FBST consists of determining the e-value that represents the Bayesian evidence against H. To construct this measure, the authors introduce the posterior surprise function and its supremum, given respectively by

$$s(\boldsymbol{\theta}) = \frac{g_1(\boldsymbol{\theta}|\boldsymbol{x})}{r(\boldsymbol{\theta})} \quad \text{and} \quad s^* = s(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta}\in\Theta_H} s(\boldsymbol{\theta}),$$

where $r(\boldsymbol{\theta})$ is a suitable reference function to be chosen. Then, a tangential set is defined as

$$\overline{T}(s^*) = \{\boldsymbol{\theta} \in \Theta | s(\boldsymbol{\theta}) > s^*\},$$

to the sharp hypothesis H, also called a Highest Relative Surprise Set (HRSS), which includes all parameter values $\boldsymbol{\theta}$ that attain a larger surprise function value than the supremum s^* of the null set. Finally, the e-value, that represents the Bayesian evidence against H, is defined as

$$\overline{ev}(H) = \overline{W}(s^*) = \int_{\overline{T}(s^*)} g_1(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta}.$$

On the contrary, the e-value in support of H is $ev(H) = 1 - \overline{ev}(H_0)$, which is evaluated by means of the set $T(s^*) = \Theta \setminus \overline{T}(s^*)$ and the cumulative surprise function $W(s^*) = 1 - \overline{W}(s^*)$. In conclusion, the FBST is the procedure that rejects H whenever $\overline{ev}(H)$ is large.

As pointed out in Pereira and Stern (2020) (Section 3.2) "the role of the reference density is to make$\overline{ev}(H)$ *explicitly invariant under suitable transformations of the coordinate system*". A first non-invariant definition of this measure, which corresponds to the use of a flat reference function $r(\theta) \propto 1$ in the second formulation, has been given in Pereira and Stern (1999). The first version involved the determination of the tangential set $\overline{T}$ starting only from the posterior distribution, whereas in the second, a corrective element has been introduced by also including the reference function. Some of the suggested choices for the reference function are the use of uninformative priors such as "the uniform, maximum entropy densities, or Jeffreys' invariant prior" (see Pereira and Stern (2020), Section 3.2).

4.1 Similarities and differences between the procedures

The most striking similarity between the FBST and the BDT is that both tests, fully accepting the likelihood principle and relying on the posterior distribution of the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, are clearly Bayesian.

Another important similarity is that, asymptotically, both tests lead to the rejection of the hypothesis H when it is false (i.e. when we test $\theta_H \neq \theta^*$ where θ^* is the true value of the parameter). On the contrary, if $\theta^* = \theta_H$ they have a different asymptotic behaviour (see Proposition 1 for the BDM and Section 3.4 in Pereira and Stern (2020) for the e-value).

Certainly, the FBST has a more general reach than the BDT. Indeed, it examines the entire class of sharp hypotheses, whereas the extension of the BDT to such hypotheses is not straightforward and, currently, is limited to considering the subclass of the hypotheses expressed as $H : \varphi = \varphi_H$ that are able to partition the parameter space $\boldsymbol{\Theta}$ as $\left\{\boldsymbol{\Theta}_a, \boldsymbol{\Theta}_H, \boldsymbol{\Theta}_b\right\}$. Moreover, notice that while the integration sets $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ are determined exclusively by the hypothesis, the tangential set $\overline{T}$ depends on the hypothesis, the posterior density and the choice of the reference function. It is questionable, on the other hand, whether the e-value is as easily computable as the BDM is in cases where the parameter space has dimension higher than 1.

Unlike the BDM, the elimination of nuisance parameters is not recommended when using the e-value. In fact, this measure is not invariant with respect to marginalisations of the nuisance parameter and the use of marginal densities to construct credible sets may produce inconsistency.

It is easy to see that one can create an analogy between the *p-value*, the e-value and δ_H. Regarding frequentist *p-value*s, the sample space is ordered according to increasing inconsistency with the assumed null hypothesis H. The FBST instead orders the parameter space according to increasing inconsistency with the assumed null hypothesis H, based on the concept of statistical surprise. In the same way, it can be seen that the probability in (7) has to do with the posterior probability of exceeding θ_H in a direction in contrast with the data (namely, the side where there is more posterior probability).

Another similarity occurs when considering the reference density $r(\theta)$ as the (possibly improper) uniform density, since the first and second definitions of evidence define the same tangent set, i.e. the HRSS and the HPDS coincide. Then, for a scalar parameter θ, since the BDM is linked to the equi-tailed credible regions while the e-value is linked to the HPDS, we have that if:

- $g_1(\theta|\boldsymbol{x})$ is symmetric and unimodal, then $\overline{ev}(H) = \delta_H$;
- $g_1(\theta|\boldsymbol{x})$ is asymmetric and unimodal (for instance with positive skewness) and $m_1 < \theta_H$ $[\theta_H < m_1]$, then $\overline{ev}(H) > \delta_H$ $[\overline{ev}(H) < \delta_H]$. When $m_1 = \theta_H$ we have $0 = \delta_H < \overline{ev}(H)$.

Table 1 Exponential distribution: false positive rates for different sample sizes n and different thresholds ω.

	$\omega = 0.90$			$\omega = 0.95$			$\omega = 0.99$		
	n			n			n		
	10	100	1000	10	100	1000	10	100	1000
e-value	0.102	0.100	0.099	0.052	0.050	0.051	0.011	0.010	0.011
$r(\theta) \propto 1$	[0.031]	[0.089]	[0.100]	[0.012]	[0.043]	[0.049]	[0.001]	[0.009]	[0.009]
e-value	0.101	0.102	0.100	0.051	0.050	0.051	0.010	0.010	0.011
$r(\theta) = g_0(\theta)$	[0.064]	[0.098]	[0.101]	[0.033]	[0.048]	[0.050]	[0.009]	[0.010]	[0.010]
δ_H	0.103	0.102	0.101	0.053	0.049	0.052	0.010	0.009	0.011
	[0.091]	[0.099]	[0.102]	[0.045]	[0.050]	[0.049]	[0.009]	[0.011]	[0.009]

In each cell are reported the value of the rates using the Jeffreys' prior and the conjugate one [in brackets]

Table 2 Poisson distribution: false positive rates for different sample sizes n and different thresholds ω

	$\omega = 0.90$			$\omega = 0.95$			$\omega = 0.99$		
	n			n			n		
	10	100	1000	10	100	1000	10	100	1000
e-value	0.091	0.099	0.101	0.041	0.050	0.051	0.010	0.010	0.011
$r(\theta) \propto 1$	[0.412]	[0.138]	[0.102]	[0.299]	[0.080]	[0.052]	[0.099]	[0.022]	[0.011]
e-value	0.095	0.095	0.101	0.046	0.047	0.051	0.011	0.011	0.011
$r(\theta) = g_0(\theta)$	[0.013]	[0.072]	[0.095]	[0.006]	[0.032]	[0.047]	[0.002]	[0.006]	[0.009]
δ_H	0.091	0.099	0.103	0.043	0.045	0.051	0.011	0.011	0.010
	[0.299]	[0.126]	[0.102]	[0.205]	[0.070]	[0.051]	[0.080]	[0.015]	[0.011]

In each cell are reported the value of the rates using the Jeffreys' prior and the conjugate one [in brackets]

4.1.1 Simulation study

In order to determine the resulting false-positive rates of both the FBST and the BDT, we conduct a simulation study for specific sample sizes, considering a continuous (Exponential) and a discrete (Poisson) model and for each one two different choices for the prior distribution, the Jeffreys' and the conjugate priors. The last ones have been chosen to have mean “far" from the true hypothesized values for the parameters. Regarding the FBST we have considered two different choices for the reference function $r(\theta)$, the flat and the prior density.

Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be an *iid* sample of size n from the Exponential distribution $X \sim Exp\big(x|1/\theta^*\big)$, with $\theta^* = 1.2$. We are interested in testing the hypothesis $H : \theta_H = \theta^* = 1.2$. Assuming the Jeffreys' prior $g_0(\theta) \propto \theta^{-1}$, the posterior distribution is $InvGamma(\theta|n, \sum x_i)$ (see Example 1), while adopting a $InvGamma(\theta|\alpha_0, \beta_0)$ prior, with $\alpha_0 = 3$ and $\beta_0 = 6$, we have a posterior that is still $InvGamma(\theta|\alpha_1, \beta_1)$, with parameters $\alpha_1 = \alpha_0 + n$ and $\beta_1 = \beta_0 + \sum x_i$. Let now $\boldsymbol{y} = (y_1, \dots, y_n)$ be an *iid* sample of size n from a Poisson distribution $Y \sim Poi\big(y|\lambda^*\big)$, with $\lambda^* = 3$. Interest

Table 3 Results, for the three different cases examined in Example 1, of δ_H and the e-value considering, as a reference distribution, both a flat reference function and a Jeffreys' prior

	$\theta_H = 2.4$			$\theta_H = 0.7$		
	e-value		δ_H	***e-value***		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$		$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
[A] $n = 6$	0.909	0.866	0.832	0.646	0.847	0.886
[B] $n = 12$	0.978	0.968	0.960	0.899	0.957	0.968
[C] $n = 24$	0.999	0.998	0.997	0.991	0.997	0.997

is on the hypothesis $H : \lambda_H = \lambda^* = 3$. For both choices of the prior, the Jeffreys' $g_0(\lambda) \propto \lambda^{-\frac{1}{2}}$ and the conjugate $Gamma(\lambda|\alpha_0, \beta_0)$, we have a Gamma posterior $Gamma(\lambda|\alpha_1, \beta_1)$, with parameters respectively equal to $\alpha_1 = \sum y_i + \frac{1}{2}$, $\beta_1 = n$ and $\alpha_1 = \alpha_0 + \sum y_i$, $\beta_1 = \beta_0 + n$.

Table 1 shows the simulation results for three different values of the threshold $\omega = \{0.90, 0.95, 0.99\}$, for $S = 50000$ simulations and $D = 50000$ posterior draws for the Exponential model. Concerning the Exponential model with the Jeffreys' prior across the different sample sizes considered, the false-positive rates are very similar for both tests (two different version of the FBST and the BDM) and, as we expect since we are using objective priors (see Bayarri and Berger (2004)), they are close to the error of the first type $\alpha = \{0.10, 0.05, 0.01\}$, related to ω. With the conjugate prior the BDM seems to perform better w.r.t. the two versions of the FBST. Concerning the Poisson model, we have good results for large sample sizes, but also for smaller n expecially with the conjugate prior (see Table 2).

4.1.2 Some examples

In order to compare the BDM and the e-value, let us consider different situations and then examine the results.

Example 8 (*Continuation of Example* 1) As a first comparative scenario, consider the test performed in Example 1 in which $\theta_H = 2.4$ and additionally the case in which $\theta_H = 0.7$. Since the posterior $g_1(\theta|\boldsymbol{x})$ has a positive skewness and $m_1 < \theta_H = 2.4$ then $\overline{ev}(H) > \delta_H$, on the contrary, for $m_1 > \theta_H = 0.7$ then $\overline{ev}(H) < \delta_H$. Indeed, we find the results reported in Table 3.

The differences between the e-value and δ_H, which in this example appear to be modest, can actually become meaningful when the posterior has a greater asymmetry and heavier tails. In such case, comparing different hypotheses, the FBST always leads to favour the hypothesis with higher density. Moreover, the e-value may be more or less robust w.r.t. the position of θ_H, as it is highlighted in the example below.

Table 4 For the two different hypotheses examined in Example 9, the table shows δ_H and the e-value considering, as a reference distribution, both a flat reference function and a Jeffreys' prior

	e-value		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
$H_A : \mu = 2.5$	0.803	0.848	0.975
$H_B : \mu = 12$	1	1	0.907

Example 9 - *Test on the mean of the Inverse Gaussian distribution* Consider a random variable X with Inverse Gaussian distribution $X \sim IG(x|\mu, \nu_0)$, $\mu \in \mathbb{R}^+$ and ν_0 known. Given an *iid* sample $\boldsymbol{x}$ of size n, the likelihood function for μ is $L(\mu|\boldsymbol{x}) \propto \exp\left\{-n\nu_0 \cdot \left(\frac{\bar{x}}{2\mu^2} - \frac{1}{\mu}\right)\right\}$. Adopting the Jeffreys' prior $g_0(\mu) \propto \frac{1}{\sqrt{\mu^3}}$, we obtain the posterior

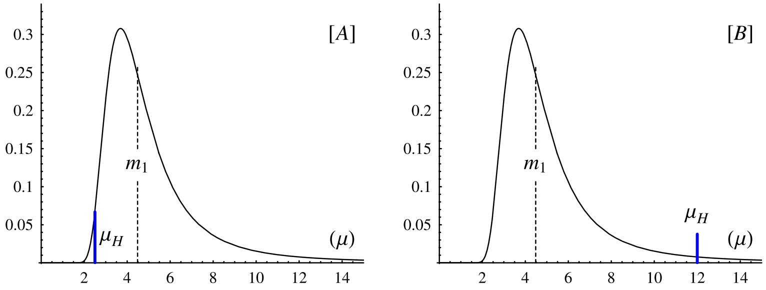

Fig. 9 Posterior density function $g_1(\mu|\boldsymbol{x})$ associated to Example 9. In $[A]$ we have $\mu_H = 2.5 < m_1$, while in $[B]$ $\mu_H = 12 > m_1$.

$$g_1(\mu|\boldsymbol{x}) \propto \frac{1}{\sqrt{\mu^3}} \cdot \exp\left\{-n\nu_0 \cdot \left(\frac{\bar{x}}{2\mu^2} - \frac{1}{\mu}\right)\right\}.$$

We are interested in testing the hypothesis $H : \mu = \mu_H$ and we consider a sample of size $n = 8$ for which $\bar{x} = 4.2$ and $m_1 = 4.483$. For $\nu_0 = 5$, we choose to test $H_A : \mu = 2.5$ and $H_B : \mu = 12$. The results of the analysis are displayed in Table 4 and Fig. 9. If we choose $\omega = 0.95$ as a rejection threshold in both cases, and with both references, we are lead to opposite inferential conclusions.

Example 10 (*Continuation of Examples* 3, 4, 5) Let us now compare the results obtained with the FBST and the BDT for the Examples 3, 4 and 5, when fixing a value of 0.95 as a rejection threshold.

The conclusions reached with the FBST and with the BDT for Example 3, which can be seen in Table 5, are the same (for both reference functions considered) although, in some cases, there are substantial differences between the values of the evidence measures. To summarise, the hypothesis H_B has to be rejected while not enough evidence is available for the rejection of the hypotheses H_A and H_C.

Table 5 Results of the Example 3 on the test on the shape parameter, mean and variance of the Gamma distribution

	e-value		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
$H_A : \alpha = 2.5$	0.557	0.186	0.570
$H_B : \mu = 6$	0.984	0.963	0.976
$H_C : \sigma^2 = 2$	0.784	0.562	0.846

For the *e*-value we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior

Table 6 Results of the Example 4 on the test of the coefficient of variation for a Normal distribution

	e-value		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
$[A]\ n = 10$	0.364	0.999	0.570
$[B]\ n = 40$	0.924	1	0.972

For the *e*-value we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior

Table 7 Results of the Example 5 on the test of the skewness coefficient of the Inverse Gaussian distribution

	e-value		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
$H : \gamma = 2$	0.650	0.691	0.844

For the *e*-value we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior

Moving on to Example 4 we can say that the analysis of the findings with the two different tests appears to be more complex than the previous one, see Table 6. In case [*A*], for both BDT and FBST with the flat reference function, there is not enough evidence to reject the hypothesis. On the contrary, if one considers the FBST with the Jeffreys' prior as reference function, one is led to reject this hypothesis. In case [B], by rejecting the hypothesis, the BDT is in agreement with the FBST with the Jeffreys' reference function in contrast to the FBST with the flat reference function for which there is not enough evidence to reject it.

Finally, in the case illustrated in Example 5, the conclusion reached with the FBST and with the BDT is the same (for both reference functions considered), i.e. there is not enough evidence to reject the hypothesis (see Table 7). It should be noted that, again, there are substantial differences between the values of the evidence measures.

The calculation of the FBST for a scalar parameter of interest without nuisance parameters, has been carried out through the function defined in the 'fbst' package for R (Kelter 2022). Instead, tangential sets $\overline{T}$ and its integrals, for Examples 3, 4 and 5, were determined by means of the *Mathematica* software. Browsing through the code that leads to the calculation of these measures (see Manca (2022)), it is evident that more work is required for the calculation of the integration region related to the FBST. In this sense, the BDT appears to be easier to apply.

5 Conclusions

We propose a new measure of evidence in a Bayesian perspective. From an examination of the examples illustrated, the conceptual simplicity of the proposed method is evident as well as its theoretical consistency. We have presented some simple cases where the computation of the BDM is straightforward.

In some situations, the BDM can be usefully applied adopting a subjective prior. It is indeed interesting the situation where one or more statisticians choose the hypothesis H and the prior according to his or their knowledge. In such cases the BDT would have a confirmatory value. The use of subjective priors must be accompanied by a robustness study especially in the case of small sample sizes.

So far we have considered only hypotheses that induce a partition on the parameter space, but the extension of the definition and the analysis of the BDT to more complex hypotheses is under investigation. Theoretical and computational developments in more general contexts are also being explored.

Acknowledgements The authors thank the Associate Editor and the Referees for helpful comments and suggestions that improved the quality of the paper. Research partially supported by the Fondazione di Sardegna 2020.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

References

Bayarri MJ, Berger JO (2004) The interplay of Bayesian and frequentist analysis. Stat Sci 19(1):58–80

Benjamin DJ, Berger JO (2019) Three reccomandations for improving the use of p-values. Am Stat 73:186–191

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C et al (2018) Redefine statistical significance. Nat Human Behav 2(1):6–10

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer, New York

Bernardo JM, Smith AFM (1994) Bayesian theory. John Wiley & Sons Inc., New York

Casella G, Berger RL (2001) Statistical inference, 2nd edn. Duxbury, Pacific Grove (CA)

Christensen R (2005) Testing Fisher, Neyman, Pearson, and Bayes. Am Stat 59(2):121–126

Collaboration OS (2015) Estimating the riproducibility of psychological science. Science 349(aac4716)

Denis DJ (2004) The modern hypothesis testing hybrid: R. A. Fisher's fading influence. Journal de la société française de statistique 145(4):5–26

Fisher RA (1925) Statistical methods for research workers. Oliver & Boyd, Edinburgh

Folks JL, Chhikara RS (1978) The inverse Gaussian distribution: theory methodology and applications—a review. J R Stat Soc Ser B 40(3):263–289

Johnson VE, Payne RD, Wang T, Acher A, Mandal S (2017) On the riproducibility of psychological science. JASA 112(517):1–10

Kelter R (2022) fbst: an R package for the full Bayesian significance test for testing a sharp null hypothesis against its alternative via the e-value. Behav Res Methods 54(3):1114–1130

Lindley DV (1965) Introduction to probability and statistics from a Bayesian viewpoint. Cambridge University Press, Cambridge, UK

Lindley DV (1991) Making decisions, 2nd edn. Wiley, Louisville

Manca M (2022) maramanca/A_new_Bayesian_Discrepancy_Measure: A new Bayesian Discrepancy Measure. Zenodo. https://doi.org/10.5281/zenodo.7317122

O'Hagan A (2003) HSSS model criticism. In: Green PJ, Hjort NL, Richardson ST (eds) Highly structured stochastic systems. Oxford University, Oxford, pp 423–445

Pereira C, Stern JM (1999) Evidence and credibility: full bayesian significance test for precise hypotheses. Entropy 1(4):99–110

Pereira C, Stern JM (2020) The *e*-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program. Sao Paulo J. Math, Sci

Ruli E, Ventura L (2021) Can Bayesian, confidence distribution and frequentist inference agree? Stat Methods Appl 30(1):359–373

Spiegelhalter D (2019) The art of statistics: learning from data. Penguin Books, London

Wasserstein RL, Lazar NA (2016) The asa statement on p-values: context, process, and purpose. Am Stat 70(2):129–133

Yang R, Berger JO (1996) A catalog of noninformative priors. Duke University, Durham, Institute of Statistics and Decision Sciences

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Francesco Bertolino[1] · **Mara Manca**[1] · **Monica Musio**[1] · **Walter Racugno**[1] · **Laura Ventura**[2]

✉ Monica Musio
mmusio@unica.it

Francesco Bertolino
bertolin@unica.it

Mara Manca
mara.manca@unica.it

Walter Racugno
racugno@unica.it

Laura Ventura
ventura@stat.unipd.it

[1] Department of Mathematics and Computer Science, University of Cagliari, via Ospedale, 72, 09124 Cagliari, Italy

[2] Department of Statistical Sciences, University of Padua, via Cesare Battisti, 241, 35121 Padua, Italy