The e-value and the Full Bayesian Significance Test: Logical Properties and Philosophical Consequences.

Julio Michael Stern* Carlos Alberto de Bragança Pereira†
Marcelo de Souza Lauretto‡ Luís Gustavo Esteves§
Rafael Izbicki¶ Rafael Bassi Stern||
Marcio Alves Diniz** Wagner de Souza Borges††

Dedicated to Walter Alexandre Carnielli, for his 70th birthday

Abstract: This article gives a conceptual review of the e-value, \( ev(\mathcal{H} | \mathcal{X}) \) – the epistemic value of hypothesis \( \mathcal{H} \) given observations \( \mathcal{X} \). This statistical significance measure was developed in order to allow logically coherent and consistent tests of hypotheses, including sharp or precise hypotheses, via the Full Bayesian Significance Test (FBST). Arguments of analysis allow a full characterization of this statistical test by its logical or compositional properties, showing a mutual complementarity between results of mathematical statistics and the logical desiderata lying at the foundations of this theory.

1 Introduction

The e-value, \( ev(\mathcal{H} | \mathcal{X}) \) – also named the epistemic-value of hypothesis \( \mathcal{H} \) given observations \( \mathcal{X} \), or the evidence-value of observations \( \mathcal{X} \) in favor (or in support) of hypothesis \( \mathcal{H} \) – is a Bayesian statistical significance measure introduced in 1999 by Carlos Alberto de Bragança Pereira and Julio Michael Stern, together with the FBST – the Full Bayesian Significance Test, see [16]. The definitions of e-value and the FBST were further refined and generalized by subsequent works of several researchers at the University of São Paulo (USP) and the Federal University of São Carlos (UFSCar), in Brazil, including Wagner Borges, Luís Gustavo Esteves, Rafael Izbicki, Regina Madruga, Rafael Bassi Stern, and Sergio Wechsler, see [2, 4, 5, 15, 17].

The e-value was specially designed to assess the logical truth value (a.k.a. statistical significance) of sharp (a.k.a. precise) hypotheses in the context of Bayesian statistics. The e-value has desirable

---

† cpereira@ime.usp.br, ORCID: 0000-0001-6315-7484, ResearcherID: O-5022-2015; IME-USP.
‡ marcelolaureto@usp.br, ORCID: 0000-0001-5507-2368, ResearcherID: J-1964-2012; EACH-USP - School of Arts, Sciences and Humanities of the University of São Paulo. Rua Arlindo Bêttio, 1000, 03828-000, São Paulo, Brazil.
§ lestevess@ime.usp.br, ID-Lattes: 8881931568548501; IME-USP.
¶ rizbicki@ufscar.br, ORCID: 0000-0003-0379-9690, ResearcherID: J-1160-2014; DS-UFSCar - Department of Statistics of the Federal University of São Carlos. Rodovia Washington Luís, km 235, 13565-905, São Carlos, Brazil.
|| rbstern@ufscar.br, ORCID: 0000-0002-4323-4515, ResearcherID: I-4745-2014; DS-UFSCar.
** marcio.alves.diniz@gmail.com, ORCID: 0000-0002-8239-4263, ID-Lattes: 8948404469003829; DS-UFSCar.
†† wborges@ime.usp.br, ORCID: 0000-0001-5520-5861, ID-Lattes: 0975898314511610; IME-USP.
asymptotic, geometrical (i.e., invariance), and logical (i.e., compositional) properties that allow consistent and coherent evaluation and testing of sharp statistical hypotheses. Furthermore, in applied modeling, the FBST offers an easy-to-implement and powerful statistical test that is fully compliant with Bayesian principles of good inference, like the likelihood principle, see [116, 160, 189, 230].

In the context of statistical test of hypotheses, a compositional logic is conveyed by an algebraic formalism that allows the evaluation of truth-functions of composite models and truth-values of composite hypotheses by algebraic operations on the corresponding truth functions of elementary models and truth values of elementary hypotheses. The e-value and the FBST have a rich, expressive, and intuitive compositional logic, while traditional truth-values and accompanying tests offered by either frequentist (classical) statistics, like the p-value, or by Bayesian statistics, like Bayes factors, have important and well-known deficiencies in this regard, specially in cases involving sharp statistical hypotheses. Furthermore, logically coherent evaluations and testing of sets and sub-sets of statistical hypotheses should render sequences of inferential reasoning that do not generate internal contradictions or anti-intuitive results. As expected, the e-value and the FBST comply with well-established rules of logical coherence, even in the case of sharp hypotheses, while traditional alternatives often fail to do so.

Asymptotically, sev(H|X) – the standardized e-value – shares several properties of the p-value, the well-known significance measure of frequentist statistics. This allows the use of the standardized e-value in frequentist-oriented applications, retaining, nevertheless, many theoretical characteristics of the Bayesian framework. Many theoretical developments and practical applications of the e-value and the FBST have already been published in the scientific literature, see [19]. Concise entries about the e-value and the FBST are available at the International Encyclopedia of Statistical Science and online at Wiley’s StatsRef, see [18, 66].

Section 2 reviews the Bayesian statistical framework; Section 3 defines the e-value; Sections 4, 5, and 6 explain the invariance, asymptotic, and compositional properties of the e-value; Section 7 defines the GFBST - the Generalized Full Bayesian Significance Test and its logical properties; Sections 8 and 9 comment on computational implementation and give a detailed numerical example in model selection; Section 10 lists a representative assortment of articles from many practical applications of the e-value and the FBST already published in the scientific literature. Section 11 considers the philosophical consequences of the aforementioned developments by briefly commenting on the Objective cognitive constructivism epistemological framework, which was specifically developed to accommodate the formal properties of the e-value and the FBST; and renders a naturalized approach to ontology and metaphysics. Section 12 presents some topics for further research at the interface between Logic and Statistics. One of the objectives of this paper is to foster work in areas of interface or overlap between Logic and Statistics and to stimulate greater cooperation between the two communities. Hence, for the sake of clarity and completeness, this paper includes some basic definitions and explanations in both areas that may be convenient to facilitate mutual understanding of more advanced topics.

2 Bayesian Framework

The basic definition of the e-value is given in the context of a standard parametric model in Bayesian statistics, where observable variables $x$ are generated with a probability density $p(x|\theta)$. The observable (vectors of) random variables $x \in \mathcal{X} \subseteq \mathbb{R}^s$ belong to the model’s sample space, while the latent (non-observable) vector $\theta \in \Theta \subseteq \mathbb{R}^p$ belongs to the model’s parameter space. Statistical inference aims to acquire information about the unknown parameter $\theta$ from a sequence of observations, $X = [x^{(1)}; x^{(2)}; \ldots; x^{(n)}]$, that for notational convenience are stacked in $n \times s$ matrix $X$.

Since the true value of the parameter, $\theta^0$, is unknown, $\theta$ is treated in Bayesian statistics as a (vector) random variable. Available previous information about the parameter is represented by its a priori density, $p_0(\theta)$. No previous information about $\theta$ is represented by a non-informative reference density, $r(\theta)$. From Bayes rule, it follows that, after having $n$ independent observations in dataset $X$, the available information about the parameter $\theta$ is represented by the posterior density

$$
p_n(\theta | X) = (1/c_n)p_0(\theta) \prod_{i=1}^n p(x^{(i)} | \theta) = (1/c_n)p_0(\theta)p(X | \theta),
$$

where $c_n$ is the appropriate normalization constant.
Whenever possible, we use a relaxed notation leaving implicit the conditionalization on the observed dataset, for example, writing $p_n(\theta)$ instead of $p_n(\theta | X)$. Further details about the Bayesian framework, including appropriate choices for non-informative reference densities, can be found in [31, 123, 160, 173, 176, 235].

A statistical hypothesis $H$ states that the parameter $\theta^0$ generating the observations $X$ belongs to the hypothesis’ set, $\Theta_H$, a region of $\Theta$ constrained by (vector) inequality and equality constraints,

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

It is common practice in statistics to use a relaxed notation, writing $H$ for both the hypothesis statement and the hypothesis’ set.

The presence of $q$ equality constraint makes the hypothesis sharp or precise, namely, $H$ becomes a proper surface (sub-manifold) of dimension strictly lower than the dimension of the parameter space, that is, $h = \dim(H) = t - q < t = \dim(\Theta)$. In particular, a point hypothesis is a hypothesis of dimension $h = 0$. Regarded as a subset of the $t$-dimensional parameter space, a sharp hypothesis has volume zero, and should therefore have zero posterior probability, at least for a regular (continuous and differentiable) posterior density, which is usually a natural assumption for statistical modeling, see [11]. These conditions make $\Pr(H)$, the probability of hypothesis $H$, an inappropriate measure to evaluate sharp hypotheses, unless a special probability measure is created for the hypothesis set, a situation that may be computationally cumbersome and theoretically challenging, see [35, 232]. Nevertheless, sharp hypotheses have a prominent role in science, for the most important statements in exact sciences are natural laws. These are formulated as equations in a theory of interest that, in turn, can be expressed by sharp hypotheses in statistical models used to empirically verify the same laws, see [5, 31, 32, 37, 42]. Sharp hypotheses also naturally arise in legal applications, like auditing models for regulatory compliance and concerns related to the burden of proof in related legal cases, see [25, 39, 93, 163, 190, 250, 254, 260]. This state of affairs was the main motivation for defining the e-value as a Bayesian significance measure specially well-suited for sharp hypotheses.

### 3 The e-value

$\text{ev}(H | X) \in [0, 1]$, the e-value, or the epistemic value of hypothesis $H$ given the observed data $X$, or the evidence given by the observed data $X$ in favor of hypothesis $H$, and its complement, $\overline{\text{ev}}(H | X) = 1 - \text{ev}(H | X)$, are defined as follows:

(i) $s(\theta)$, the surprise function in a statistical model is defined as the quotient between the posterior and the reference densities in the statistical model, see [15, 35, 139, 150, 202, 212, 213, 251],

$$s(\theta) = p_n(\theta) / r(\theta);$$

(ii) $s^*$, the maximum (or supremum) of the surprise function constrained to the hypothesis $H$, is defined as

$$s^* = \sup_{\theta \in H} s(\theta),$$

A maximizing argument, $\theta^* | s^* = s(\theta^*)$, is called a tangential point, see Figure 1;

(iii) $T(v)$, the closed lower $v$-cut of the surprise function, see [35, 139], and its complement, the open upper $v$-cut of the surprise function, $\overline{T}(v)$, are defined as

$$T(v) = \{ \theta \in \Theta | s(\theta) \leq v \}, \quad \overline{T}(v) = \{ \theta \in \Theta | s(\theta) > v \};$$

The upper $v$-cut at level $v = s^*$, $T(s^*)$, is called the tangential set, for its border corresponds to the contour line of the surprise function that is tangential to hypothesis $H$, see Figure 1.

(iv) $W(v)$, the truth function or Wahrheitsfunktion at level $v$, is defined as the posterior probability mass inside the lower $v$-cut of the surprise function, see [2, 35],

$$W(v) = \int_{T(v)} p_n(\theta) d\theta,$$

while its complement is defined as $\overline{W}(v) = 1 - W(v);$
Figure 1: Flat prior, \( r(\theta) \propto 1 \); Surprise function equal to posterior density, \( s(\theta) = p_n(\theta) \); Optimal \( \theta^* \in H \); Contour curve at level \( s^* = s(\theta^*) \); and Tangential highest probability density set, \( T(\theta^*) \).

(v) \( \text{ev}(H|X) \), the epistemic value of hypothesis \( H \) given the observed data \( X \), is defined as the truth function \( W(v) \) computed at level \( v = s^* \), see [2, 16, 17, 35], while its complement, \( \overline{\text{ev}}(H|X) \), the evidence given by the observed data \( X \) against hypothesis \( H \), has the complementary probability mass,

\[
\text{ev}(H|X) = W(s^*) , \quad \overline{\text{ev}}(H|X) = \overline{W}(s^*) = 1 - \text{ev}(H) ;
\]

For further details and remarks on the development of the e-value and its historical predecessors, see [16, 25, 35].

In the special case of the flat reference density, \( r(\theta) \propto 1 \), the surprise function coincides with the posterior probability density function. Its upper cuts, \( T(v) \), coincide with highest probability density sets (HPDS), making it easier to visualize all the elements defined in the last paragraph. Figure 1 depicts a simple example with a unimodal surprise function and a simply connected tangential set, see [16, Sec.4.3]. However, in general, the posterior probability density function may have multiple local maxima, the tangential set may have multiple connected components, and all these components must be taken into account when computing the e-value. In this respect, the e-value differs from traditional uses of HPDS that only consider a single connected component, usually the connected component containing \( \hat{\theta} \), the unconstrained maximum a posteriori estimator, see [35, 123, 140, 184, 208, 231].

4 Invariance and Reference Density

A reparameterization of the parameter space, that is, a change in the coordinate system used to map \( \Theta \), could “stretch” or “compress” the region surrounding a given location point in the map. This kind of effect is clearly visible comparing geographic maps using different cartographic projections. However, the probability mass inside a given region, as specified by a probability density function, should remain the same, regardless of the coordinate system in use. Hence, the value or “height” of a density function at a given point must change according to the coordinate system in use. This kind of reasoning explains the transformation rules specified by differential and integral calculus for modifying density functions according to transformations of the coordinate system, see [152, 154, 195]. Nevertheless, since the surprise function is defined as the ratio of two densities, \( s(\theta) = p_n(\theta) / r(\theta) \), its value remains unchanged by regular (continuous and differentiable) reparameterizations of the parameter space, and the e-value inherits this important invariance property, see [2, 35].
5 Asymptotic Consistency

As the number n of observations grows to infinity, that is, in the asymptotic limit n → ∞, we expect a consistent statistical procedure to reach the “correct” conclusion. This section analyses the asymptotic properties of the e-value that motivate its use to consistently evaluate a statistical hypothesis. It is easier to describe these asymptotic properties in terms of the standardized version of the e-value, defined as follows:

(i) $Q(d, z)$, the chi-square cumulative distribution with $d \in \mathbb{N}_+$ degrees of freedom for random variable $z \in [0, \infty]$, is defined by the following analytical expression using the incomplete gamma function:

$$Q(d, z) = \frac{\gamma(d/2, z/2)}{\gamma(d/2, \infty)}, \quad \gamma(d, z) = \int_0^z y^{d-1}e^{-y}dy;$$

(ii) $\sigma(t, h, c)$, the standardization function on arguments $t, h \in \mathbb{N}_+$ and $c \in [0, 1]$, is defined by the expression

$$\sigma(t, h, c) = Q\left(t-h, Q^{-1}(t, c)\right);$$

(iii) $\text{sev}(H|X)$, the standardized e-value of a hypothesis $H \subset \Theta$ of dimension $h = \dim(H) \leq t = \dim(\Theta)$, is defined as follows:

$$\text{sev}(H|X) = 1 - \frac{\chi^2(H|X)}{\chi^2(H|X)}, \quad \frac{\chi^2(H|X)}{\chi^2(H|X)} = \sigma(t, h, \chi^2(H|X));$$

The standardized e-value has the following asymptotic properties, under usual continuity and differentiability regularity conditions, see [2]:

(a) If $H$ is true, i.e. $\theta^0 \in H$; and $H$ is slack, i.e. $\dim(H) = \dim(\Theta)$; and $\theta^0$ is in the topological interior of $H$; Then ev $(H|X)$ converges (in probability) to $1$, as the number of observations increases;

(b) If $H$ is true, i.e. $\theta^0 \in H$; and $H$ is sharp, i.e. $\dim(H) < \dim(\Theta)$; Then $\text{sev}(H|X)$ converges (in distribution) to $U[0, 1]$, the uniform distribution in the unit interval. The standardized e-value shares these asymptotic properties with the p-value – the well-known and widely used significance measure of frequentist statistics. Hence, in a frequentist-oriented application, one could use the e-value as a Bayesian replacement for the p-value, preserving already familiar decision procedures and their interpretations. Moreover, in several modeling applications, the e-value exhibits better convergence characteristics than those of the p-value, see [12, 13, 55, 79, 80]. Further consistency properties of the e-value and higher order asymptotic approximations have been studied and developed in [3, 20–22, 45–47, 94].

6 Compositional Logic of e-values

The e-value and its truth function yield simple algebraic expressions that facilitate the study of a composite model, $M$, build by serial coupling of $k$ independent statistical models, indexed on $j = 1 \ldots k$, $M^{(j)} = \{\Theta^{(j)}, p_0^{(j)}, p_n^{(j)}, r^{(j)}\}$. In this setting, $q$ alternative (or parallel) sets of hypotheses, $H^{(i,j)}$, indexed on $i = 1 \ldots q$, are provided for evaluation. This situation is somewhat analogous to the analysis of parallel-serial systems in reliability theory, see [114, 122, 179] and Figure 2. Let us first consider a pure serial system, consisting of $k$ individual models, $M^{(j)}, j = 1 \ldots k$, each one contemplating a single hypothesis, $H^{(1,j)}$.

Since the individual models are independent, the joint posterior density, reference, and surprise function of the composite model are the product of the corresponding individual functions, that is,

$$p_n(\theta) = \prod_{j=1}^k p_n^{(j)}(\theta^{(j)}), \quad r(\theta) = \prod_{j=1}^k r^{(j)}(\theta^{(j)}),$$
The Mellin convolution of the cumulative distributions for (scalar) random variables $x$ and $y$, $F(x) \otimes G(y)$, gives the cumulative distribution of the product $z = xy$. Hence, the composite model’s truth function is the Mellin convolution of the individual truth functions, that is,

$$W(v) = \bigotimes_{1 \leq j \leq k} W(j) = W(1) \otimes W(2) \ldots \otimes W(k)(v).$$

Further properties and interpretations of the Mellin convolution, and detailed analytical procedures and numerical algorithms for its efficient computation, can be found in [124, 175, 220, 233]. Matlab code for computing the composite model’s truth function in the discrete case, i.e. for step ladder cumulative functions, is presented in [1].

Next, let us consider a composite hypothesis $H$ expressed in conjunctive form, where the conjunction symbol ($\land$) stands for the and logical operator,

$$H = \bigwedge_{j=1}^{k} H^{(j)} = H^{(1,1)} \land H^{(1,2)} \ldots \land H^{(1,k)}.$$

In the conjunctive hypothesis, the surprise function attains a maximum (or supremum) value equal to the product of individual maxima (or suprema). Therefore, the e-value of the conjunctive composite hypothesis is given by

$$\text{ev}(H) = W(s^*) = \bigotimes_{1 \leq j \leq k} W(j) \left( \prod_{j=1}^{k} s^{*(j)} \right).$$

Finally, let us consider a composite hypothesis $H$ expressed in disjunctive normal form, where the disjunction symbol ($\lor$) stands for the or logical operator,

$$H = \bigvee_{i=1}^{q} \bigwedge_{j=1}^{k} H^{(i,j)} = \left( \bigwedge_{j=1}^{k} H^{(1,j)} \right) \lor \ldots \lor \left( \bigwedge_{j=1}^{k} H^{(q,j)} \right).$$

The corresponding e-value requires the conjunctive surprise to be maximized over the finite set of $q$ disjunctive alternatives,

$$\text{ev}(H) = \text{ev} \left( \bigvee_{i=1}^{q} \bigwedge_{j=1}^{k} H^{(i,j)} \right) = W(s^*) = \max_{1 \leq j \leq k} \left( \prod_{j=1}^{k} s^{*(j)} \right).$$

Interestingly, in the last expression, if all elementary hypotheses have either null or full epistemic value, that is, if, $\forall (i, j), \text{ev}(H^{(i,j)}) \in \{0, 1\}$, the evaluation of $\text{ev}(H)$ reduces to the the evaluation in classical logic of a corresponding expression preserving the same normal structure [2]. Another interesting aspect of the last expression is how conjunction operators are translated into product operations, while disjunction operators are translated into maximization operations. This kind of translation characterizes the e-value as a possibilistic abstract belief calculus. For further details and comments, see [2, 25, 26, 35, 134, 135, 139, 142, 180, 223, 224].
In the practice of statistics, it is often necessary to make a decision to either reject, or to remain undecided, or to accept a hypothesis \( H \) (in statistics, the alternative undecided is also called agnostic). In multi-valued logics, these three alternatives are traditionally encoded by the logical values \{0, 1/2, 1\}, see [147, 148, 162, 197]. In modal logics, these three alternative attitudes consider the hypothesis to be impossible, contingent, or necessary, see [126, 196, 249]. Table 1 presents the corresponding modal operators together with some relevant compositions obtained by negation (\( \neg \), the not operator), disjunction (\( \lor \), the or operator) and conjunction (\( \land \), the and operator).

The modalities in Table 1 are related by logical relations depicted in diagrammatic form by the hexagon of opposition in Figure 3, where: Arrows (\( \rightarrow \)) indicate logical implication; Dashed-lines (\( \neg \neg \)) indicate contrariety, connecting modalities that cannot both be true; Dotted-lines (\( \cdots \cdots \)) indicate sub-contrariety, connecting modalities that cannot both be false; and Double-lines (\( \equiv \equiv \)): indicate contradiction, connecting modalities that can neither both be true nor both be false. The vowels at the vertices of the hexagon are historical labels inherited from medieval logic, see [243].

The logical relations depicted in the hexagon of opposition are considered basic principles of rational argumentation, intuitive to, and tacitly assumed by most users and, sometimes, even soft coded in popular culture or hard coded in natural language, see [8, 43, 44, 118, 121, 243, 248]. Therefore, departing from these principles may make arguments counterintuitive, create barriers to the natural flow of communication, or even lead to misunderstandings. This situation justifies the development of inference methods and supports ways of reasoning that are fully compliant with these basic principles.

A statistical test of hypotheses is a statistical procedure used to make a required decision to either
reject, or remain undecided, or accept a hypothesis $H$. It is convenient to use a statistical test based on an already available significance measure, $\mu(H) \in [0, 1]$. Such a statistical test can be regarded as a discretization map, $\delta(\mu(H))$, that collapses the interval $[0, 1]$ into the ternary set $\{0, 1/2, 1\}$. Some applications may only allow decisions in a binary subset, like $\{0, 1/2\}$, although a full logical analysis of these tests requires (at least) a ternary decision space, see [6, 8, 9, 23, 41].

The remainder of this section examines logical consistency conditions for such tests, and provides further justification for the use of the e-value as an appropriate significance measure. The preceding sections paid special attention to sharp hypotheses. In contrast, this section looks at sharp or slack hypotheses with equal interest. Although motivated by the need to evaluate sharp hypotheses, we should remark that nothing prevents the e-value from being computed for slack hypotheses.

As already examined in Section 2, by definition, a statistical hypothesis $H$ states that the true parameter value belongs to the hypothesis set, that is, that $\theta^0 \in H$. From this definition, in conjunction with classical principles or rational argumentation encoded in the hexagon of opposition, we now present three sets of conditions for logical consistency that ought to be required from a rational hypothesis test, namely, invertibility, monotonicity, and consonance, see [4, 5, 8, 41].

**Invertibility:** Every hypothesis $H \subset \Theta$ automatically defines its complement, $\overline{H} = \Theta - H$. Hence, the true value of the parameter, $\theta^0 \in \Theta$, must either belong to $H$ or belong to its complement. Therefore, the following implications ought to hold:

- **(I.i) Necessity inversion:** $H$ is necessary if and only if $\overline{H}$ is impossible, that is, $\square H \iff \neg \diamond \overline{H}$;
- **(I.ii) Possibility inversion:** $H$ is possible if and only if $\overline{H}$ is unnecessary, that is, $\diamond H \iff \neg \square \overline{H}$; and
- **(I.iii) Contingency inversion:** $H$ is contingent if and only if $\overline{H}$ is contingent, that is, $\forall H \iff \forall \overline{H}$.

Figure 4 (left) gives a diagrammatic representation of invertibility relations (and their complements obtained by negation) using fragments of the hexagon of opposition.

**Monotonicity:** Let us consider a larger hypothesis $H'$ that contains the original hypothesis $H$, that is $H' \supset H$. In this setting, the following implications ought to hold:

- **(M.i) Monotonic necessity:** If $H$ is necessary, so must be $H'$, that is, $\square H \Rightarrow \square H'$; and
- **(M.ii) Monotonic possibility:** If $H$ is possible, so must be $H'$, that is, $\diamond H \Rightarrow \diamond H'$.

Figure 4 (right) gives a diagrammatic representation of these monotonicity relations (and their complements obtained by negation) using fragments of the hexagon of opposition.

**Consonance:** Let $H_i$, for $i \in I$, be a collection of hypotheses referenced by the index set $I$. The following implications ought to hold:

- **(C.i) Union consonance:** If the join hypothesis, made by the union of all indexed hypotheses, is possible, then at least one of the joining hypotheses is possible, that is,
  $\diamond (\cup_{i \in I} H_i) \Rightarrow \exists i \in I | \diamond H_i$.
- **(C.ii) Intersection consonance:** If the meet hypothesis, made by the intersection of all indexed hypotheses, is unnecessary, then at least one of the meeting hypotheses is unnecessary, that is,
  $\neg \square (\cap_{i \in I} H_i) \Rightarrow \exists i \in I | \neg \square H_i \iff \forall i \in I, \square H_i \Rightarrow \square (\cap_{i \in I} H_i)$.

Figure 5 gives a diagrammatic representation of consonance relations for a collection of three hypotheses, $\{A, B, C\}$.

Arguments of analysis lead to a characterization of test procedures that comply with the aforementioned logical conditions of monotonicity, invertibility, and consonance, see [4, 9]. These tests must be based on a region estimator, $S \subset \Theta$, for the location of the true parameter value, $\theta^0$. Such a region estimator test accepts $H$ if $S \subset H$, rejects $H$ if $S \subset \overline{H}$, and remains undecided if $S$ properly intersects both $H$ and $\overline{H}$, as depicted in Figure 6.

As mentioned in the beginning of this section, it is convenient to use a statistical test, $\delta(\mu(H)) \in \{0, 1/2, 1\}$, that is based on an already available significance measure, $\mu(H) \in [0, 1]$. Nevertheless, the
following logical compatibility condition, between the statistical test and its underlying significance measure, ought to hold:

$$\mu(H^{(1)}) \geq \mu(H^{(2)}) \Rightarrow \delta(\mu(H^{(1)})) \geq \delta(\mu(H^{(2)})) .$$

This logical compatibility condition is inspired by several legal principles, like the ones known in the juridical literature as onus probandi, in dubio pro reo, and principle of proportionality, see [25, 163, 201, 234, 237, 244, 250, 252, 254, 260]. The onus probandi and in dubio pro reo principles can be interpreted as requiring the monotonicity and consonance properties examined in this section, while the principle of proportionality can be interpreted as follows: – Judging the acceptability of hypotheses $H^{(1)}$ and $H^{(2)}$, if the former has a better grade (a higher significance measure) than the latter, then fair judgments must render verdicts (decisions) that are as least as favorable to the former than to the latter.

From the definition of the e-value, it is clear that an upper $v$-cut $\bar{T}(v)$ can be used as a region estimator for $\theta_0$. Accordingly, the corresponding region estimator test rejects $H$ if its e-value stays below the threshold $c = W(v)$, that is, $H$ is rejected if $\text{ev}(H) < c$. By invertibility, $H$ is accepted if $\text{ev}(\overline{H}) < c$. This is the GFBST - the Generalized Full Bayesian Significance Test, see [4, 7, 8, 14, 23, 24, 41]. One can check that the GFBST generates logical modalities that are compatible with all logical requirements examined in this section, and also that it is logically compatible with its underlying significance measure, the e-value. Moreover, one can also check that all good invariance and asymptotic properties of the e-value are inherited by the GFBST. Furthermore, it is interesting to remark that, under usual regularity conditions (continuity and differentiability), if $\text{ev}(H) < 1$, a sharp hypothesis $H$ may either remain undecided or be rejected by the GFBST, but never be accepted. Finally, for simplicity and limitations of space, the exposition given in this article first defines the e-value and the GFBST, and then derives or explains their logical properties. In contrast, in [4, 5, 8, 9, 23, 41], the authors show that it is possible to go the other way around, demonstrating that the aforementioned logical properties render a complete characterization of the GFBST as a region estimator test.
Figure 6: Testing $H$ with region estimator $S$: Accept if $S \subset H$, reject if $S \subset \overline{H}$, undecided otherwise.

8 Computational Implementation

Numerical computation of the $e$-value can be accomplished in two basic steps, namely,

Integration step – computing an approximation of the truth function $W(v)$. In most cases, this task can be accomplished by using a computational condensation procedure, like [175, 233], on a numerical sequence, $s^{(k)} = s(\theta^{(k)})$, obtained by sampling from the posterior distribution, $p_{\theta}(\theta)$. Sampling sequences can, in turn, be generated using Markov Chain Monte-Carlo methods, see [125, 161, 166], or variations thereof, like the Hit-and-Run algorithm, see [129, 178];

Optimization step – finding the optimum of the objective function $s(\theta)$ under the constraints imposed by the hypothesis. In most cases, this task can be accomplished by standard methods of continuous constrained optimization, like Generalized Reduced Gradient, Sequential Quadratic Programming, Generalized Augmented Lagrangian, or Proximal Point methods, see [119, 151, 177, 187, 188]. The following numerical optimization solvers are readily available and potentially applicable:

TANGO – Trustable Algorithms for Nonlinear General Optimization, a joint project from USP – the University of São Paulo, and UNICAMP – the University of Campinas, that offers excellent code under General Public License (GNU); and

Gurobi – a robust, high-performance, and versatile optimization software that can currently be licensed for free academic use.

Some optimization problems with many local maxima can be solved indirectly by repeated local optimization from candidate starting points selected at the integration step, or directly by stochastic optimization methods based on Simulated Annealing, see [194, 218, 221, 225, 226].

9 Numerical Example in Model Selection

Figure 7 (left) depicts the polynomial fitting problem for the classical Sakamoto et al. benchmark dataset presented in [207, ch.8]. This dataset, given in Table 1, was generated by a simulation from the i.i.d. (independent and identically distributed) stochastic process

$$y_i = g(x_i) + N(0, 0.1^2) , \quad g(x) = \exp((x - 0.3)^2) - 1 ,$$

that is, the points $y_i$ were generated at the grid $x_i = (i - 1)0.05$, for $i = 1, \ldots, 21$, by adding to the exponential target function, $g(x)$, a Gaussian random noise with mean $\mu = 0$ and standard deviation of $\sigma = 0.1$.

The linear regression polynomial model of order $k$ explains vector $y$ in the dataset, using as explanatory variables integer powers up to order $k$ of the grid points in vector $x$, plus an i.i.d. Gaussian random noise with standard deviation $\sigma$, that is,

$$y = \beta_0 x^0 + \beta_1 x^1 + \beta_2 x^2 + \ldots + \beta_k x^k + N(0, \sigma I) .$$

Using the weakly informative prior density, $p_0(\beta, \sigma) = 1/\sigma$, the posterior density for this statistical model can be conveniently written, see [136, 168, 235], in terms of the matrix of explanatory variables,
\[ X = [x^0, x^1, \ldots, x^k], \] the maximum a posteriori estimators of the model’s parameters, \( \hat{\beta} = (X'X)^{-1}X'y, \) and the auxiliary quantities \( \hat{y} = X\hat{\beta}, \) and \( s^2 = (y - \hat{y})'(y - \hat{y})/(n - k), \)

\[
p_n(\beta, \sigma | y, x^0, \ldots, x^k) = \frac{1}{\sigma^{n+1}} \exp \left( -\frac{1}{2\sigma^2} \left( (n - k)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \right) \right).
\]

The model selection problem that naturally arises in this context can be stated as the question: From a statistical point of view, which order \( k \) gives the best polynomial fit for this dataset? The first column of Table 3 presents the quadratic norm empirical error, \( R_{\text{emp}} = ||\hat{y} - y||_2, \) of each model, see [136, 168, 235] for further details. Polynomials of increasing order have increasing flexibility, making \( R_{\text{emp}} \) monotonically decrease until it reaches zero for order \( k = n - 1, \) when the model becomes just an interpolating polynomial for the dataset.

![Figure 7: Benchmark (left) and alternative (right) data points, ◯, for Sakamoto et al. [207, ch.8] polynomial fitting problem; Exponential target function, ⋄; Best fitted polynomials, — , of order 0 to 4; and Order 2 polynomial, *, that renders the smallest regularized error for both datasets.](image)

<table>
<thead>
<tr>
<th>Order</th>
<th>( R_{\text{EMP}} )</th>
<th>( R_{\text{FPE}} )</th>
<th>( R_{\text{SCB}} )</th>
<th>( R_{\text{GCV}} )</th>
<th>( R_{\text{SMS}} )</th>
<th>( R_{\text{AIC}} )</th>
<th>( \text{ev}(H) )</th>
<th>( \text{sev}(H) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.03712</td>
<td>0.04494</td>
<td>0.04307</td>
<td>0.04535</td>
<td>0.04419</td>
<td>-07.25</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>0.02233</td>
<td>0.02964</td>
<td>0.02787</td>
<td>0.03025</td>
<td>0.02858</td>
<td>-20.35</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.01130</td>
<td>0.01661</td>
<td>0.01534</td>
<td>0.01724</td>
<td>0.01560</td>
<td>-32.13</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.01129</td>
<td>0.01835</td>
<td>0.01667</td>
<td>0.01946</td>
<td>0.01667</td>
<td>-30.80</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.01088</td>
<td>0.01959</td>
<td>0.01751</td>
<td>0.02133</td>
<td>0.01710</td>
<td>-29.79</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.01087</td>
<td>0.02173</td>
<td>0.01913</td>
<td>0.02445</td>
<td>0.01811</td>
<td>-27.86</td>
<td>0.013</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2: Sakamoto et al. [207, ch.8] benchmark dataset for the polynomial fitting problem

| Table 3: Model selection by regularized empirical errors and \( \text{ev}(\beta_k = 0 | y, [x^0; x^1; \ldots; x^k]) \) |

Nevertheless, the predictive power of these models does not monotonically increase with the decrease of empirical error. From Figure 7 (right), it is easy to see that, if the order of the polynomial is too
high, the model becomes excessively complex and over-fits the data, resulting in poor generalizations (predictions) outside the grid points.

Columns 2 to 6 in Table 3 present a variety of penalized errors, $R_{\text{pen}} = r(d,n)R_{\text{emp}}$, where the regularization factor, $r(d,n)$, penalizes the statistical model complexity. Specifically, the factor, $r(d,n)$, increases with the dimension of the model’s parametric space, $d$, relative to the number of available data points, $n$. For the linear regressions at hand, having parameters $\sigma$ and $\beta_0, \beta_1, \ldots, \beta_k$, $d = k + 2$.

The penalized or regularized errors in Table 3 are defined by the following regularization factors, based on the quotient $q = (d/n)$. The corresponding model selection criteria chose the model with smallest regularized error; see [130, 207] for theoretical or heuristic justifications and further details.

- Akaike’s final prediction error, $r_{\text{FPE}} = (1 + q)/(1 - q)$;
- Schwartz’ Bayesian criterion, $r_{\text{SBC}} = 1 + \ln(n)q/(2 - 2q)$;
- Generalized cross validation, $r_{\text{GCV}} = (1 - q)^{-2}$;
- Shibata model selector $r_{\text{SMS}} = 1 + 2q$.

The last columns in Table 3 present, in the context of a linear regression polynomial model of order $k$, the e-value and its standardized version for the statistical hypotheses stating that the parameter of higher order is null, that is, $\text{ev}(H) = \text{ev}(\beta_k = 0 | y, [x_0, x_1, \ldots, x_n])$ and $\text{sev}(H)$. This example demonstrates that the e-value and the FBST can replace traditional model selection criteria based on penalized prediction errors. Moreover, using a selection criterion based on the FBST renders a decision process that is invariant, asymptotically consistent, and logically coherent, as explained in the previous sections. In contrast, selection criteria based on penalized prediction errors lack even the most basic invariance properties, even though alternative model selection criteria exist that are invariant and have other desirable statistical properties, see for example [170]. Furthermore, the use of the e-value and the FBST for model selection can be extended to a variety of non-nested (separate) or nested families of hypotheses, including Bayesian classifiers, as analyzed in [12, 13, 55, 57, 69, 79, 80, 87].

## 10 Applications

The development of statistical significance measures and tests may be motivated by their intended theoretical properties, which, in turn, may be inspired by epistemological desiderata. Nevertheless, these significance measures and tests must also prove themselves on the battlefields of science and technology as effective, efficient, robust, and reliable tools for the trade. A collection of over a hundred published applications of the e-value has been compiled in the 2020 survey [19]. This section gives a selection of references from this survey, organized by application area.

- Testing covariance structures in multivariate Normal models, treating in a unified way several alternative hypotheses (often treated as special cases in the literature): [79, 106];
- Testing unit root and cointegration hypotheses in time series, using plain and simple forms of prior information like flat or Jeffreys priors (no need for artificial priors): [61, 63, 64, 107];
- Solving Bayesian classification problems and testing nested and non-nested or separate hypotheses: [12, 13, 51, 52, 80, 193];
- Analyzing systems’ reliability from failure datasets: [73, 87, 98]
- Testing dependence structures using statistical copulas: [68];
- Testing (non)-informative sampling conditions in statistical surveys: [102];
- Model selection for generalized Poisson distributions: [69, 104];
- Model selection for generalized jump diffusion and Brownian motions, extremal distributions, and persistent memory processes: [49, 54, 78, 95, 96];
- Testing independence structures in contingency tables and multinomial models: [50, 55, 90, 192];
- Software certification according to compliance conditions: [93];
- Testing market equilibrium conditions for fundamental and financial derivative asset prices: [58];
- Testing hypotheses in empirical economic studies: [60];
- Event identification in acoustic signal processing: [70–72];
Testing Hardy-Weinberg equilibrium in genetics: [56, 74, 81, 83, 86, 89, 97, 108];
Testing hypotheses in biological sciences, including cases in ecology, environmental sciences, medical diagnostics, efficacy evaluation of medical procedures, psychology and psychiatry: [48, 57, 76, 82, 85, 99, 100, 103, 191];
Testing hypotheses in astronomy and astrophysics: [59, 75].

11 Epistemology, Ontology, and Metaphysics

Historically, the development of statistical significance measures (or logical truth-values), and the corresponding tests of hypotheses used in statistical science, have been reciprocally influenced by the epistemological frameworks in which they are presented. Rev. Thomas Bayes (1701-1761), whose work was communicated posthumously by Rev. Richard Price (1723-1791), developed the first methods in this area with clear goals in mind, as stated in [238], also quoted in [236, p.84] and [40, p.245]:

The purpose, is to shew [show] what reason we have for believing that there are in the constitution of things, fixed laws according to which events happen, and that, therefore, the frame of the world must be the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity [...]
It will be easy to see that the problem solved in this essay is more directly applicable to this purpose; for it shews [shows] us, with distinctness and precision, in every case of any particular order or recurrence of events, what reason there is to think that such recurrence or order is derived from stable causes or regulations in nature, and not from any of the irregularities of chance.

The next two generations in the development of probability and statistics, led by, among others, Pierre-Simon de Laplace (1749-1827) and George Boole (1815-1864), kept these core goals unchanged. While theological questions lost interest over time, the emphasis of statistical research remained essentially metaphysical – in the (gnoseological) sense of searching for and justifying causal explanations for manifested phenomena. Moreover, these causal links were ideally expressed as natural laws in the form of precise or exact quantitative relations, see [31, 32, 40]. Karl Pearson (1857-1936) was the founder and leader of frequentist statistics, the dominant school of thought in statistical science during the XX century. Influenced by the Inverted Spinozism philosophy of Johann Gottlieb Fichte (1762-1814) and the Positivist ideas of Auguste Comte (1798-1857), K.Pearson radically changed the goals of statistical science. He deprecated any form of causal reasoning, the conception or verification of natural laws, or the use of metaphysical (non-observable) entities. Instead, frequentist statistics, following a strict Positivist agenda, only aims to produce good-fitted empirical models able to describe or predict directly observed quantities. Accordingly, the very use of probability calculus is restricted to variables in the sample space, and strictly forbidden for (latent) variables in the parameter space, see [40].

Bruno de Finetti (1906–1985) is responsible for the resurgence of Bayesian statistics in the second half of the XX century, see [153, 245]. De Finetti reintroduced probability calculus in the parameter space. Latter on, this expanded use of probability language proved to be very useful for it could accommodate new means and methods provided by computer science, like Markov Chain Monte Carlo and other probabilistic algorithms, see [161, 166, 198]. Nevertheless, philosophically, the de Finetti revolution was a very conservative one, remaining always amenable to the Positivist agenda. This was accomplished by making probabilities in the parameter space quantities of subjective and ephemeral character (integration variables), entities of low ontological status used only at intermediate steps in the computation of predictive probabilities for observable variables in the sample space, see [153, 245].

The Objective cognitive constructivism epistemological framework was developed to host the e-value, the FBST, and their formal properties, including their ability to evaluate and test sharp statistical hypotheses. It also provides a naturalized way to ontology and metaphysics in empirical sciences via statistics, that is, a natural way to evaluate empirical support for natural laws and their accompanying causal explanations, and to validate the “objective” use of non-directly observable (or metaphysical) entities, in accordance with the goals of the original works of Bayes and Laplace. For further details on the Objective cognitive constructivism epistemological framework and the way in which, on the one
hand, it accommodates the e-value, the FBST and their logical properties, and, on the other hand, it provides a naturalized approach to ontology and metaphysics, see [27–29, 31–33, 36–38, 40–44].

12 Future Research and Final Remarks

In December 2018, at CLE-UNICAMP, Walter Carnielli organized the workshop *Induction, Probability and their Dilemmas*, where the first author gave the presentation *The Problem of Induction in Statistical Science*. During this workshop, we discussed several topics for further research, most of them motivated by areas of common interest that, nevertheless, are usually approached quite differently by the communities of Logic and Statistics. Four of these topics are presented in the sequel, followed by some additional topics for further research.

Statistics in (Un)Countable Sentential Probability

There is a long standing tradition in Logic to formalize probabilistic reasoning over finitary or countable sets of sentences in a language. This approach is appealing for its theoretical simplicity, for allowing computer-efficient implementations of specific models, etc. For historical analyzes of this approach, see [133, 164, 165, 205], for recent works from Walter’s group following this line, see [200, 239]. Notwithstanding their usefulness, simplicity, and popularity, finitary or countable sentential formalisms also have their limitations, for example, being unable to express measure-theoretic arguments used in mathematical statistics.

Sections 6 and 7 sowed how the statistical definitions of the e-value and the GFBST characterize their logical properties, and hinted at how arguments of mathematical analysis enable one to travel the other way around. In this context, it is a topic for further research to explore how much of this theory and its applications can be expressed using the sentential probability approach, either by expanding the underlying languages (like infinitary, second order, or fixed-point logics), see [146, 158, 159, 164, 165, 167, 169], or by exploring relevant properties in specific statistical models (including topological properties, like countability and compactness, and regularity conditions of constraints and distributions, like bounded continuity and differentiability), see [2, 11, 144, 153, 186, 210].

Functional Compositionality Structures

As already noticed in Section 6, the compositional structure of e-values can be characterized as a *possibilistic abstract belief calculus*, as analyzed in [2, 25, 35, 134, 135, 139, 142, 180]. This particular algebraic formalism resembles in many ways the formalism used in statistical reliability theory for the analysis of complex systems assembled by serial/parallel composition of simple(r) elements, see [114, 120, 179, 214]. As noticed in [2, Sec.1], this formal resemblance was a source of inspiration at early stages of this research program. At the same time, the truth-function interpretation of the cumulative surprise function, $W(v)$, was inspired by the work of Ludwig Wittgenstein [242, 264]. Finally, the concept of pragmatic hypotheses and its use in conjunction with the GFBST in testing scientific theories was inspired by sensitivity analysis, as used in optimization and systems’ theory, and its interpretation in terms of fuzzy or paraconsistent logics, see [5, 26]. While the many theoretical results already obtained in this research program justify (in our opinion) the analogies we made and the conceptual links we established between fragments of formal structures used in distinct (and often faraway) research areas, we believe that these intuitive connections deserve and can benefit from an even more rigorous and general setting. In this context, the tools of category theory or other formal abstraction methods, see [126, 127, 155, 183, 219], offer an opportunity for further research in the study of basic logical properties of the aforementioned systems, specially concerning the investigation of their essential compositionality structures and possible generalizations.

Rough and Fuzzy Sets

Essential logical properties of the GFBST can be explained regarding the e-value as a transformation between probability and possibility measures, see [35]. The underlying interconnections between al-
ternative representations of uncertainty at the core of this theory provide a general motivation to more specific topics of further research presented in this section.

In the GFBST framework, a sharp hypothesis can be either rejected or remain undecided, but can never be accepted, see Section 7. Aiming to allow the acceptance of surrogate hypotheses of interest, the authors showed how to enlarge an underlying sharp hypothesis into a slack pragmatic hypothesis, see [5]. This is accomplished by taking into account (im)precisions of measurement equipment, (un)certainties of fundamental constants, and other relevant metrological or methodological error bounds. It should be remarked that the pragmatic hypotheses defined in [5] are crisp sets.

In the context of exact sciences, the sharp (and, hence, crisp) nature of the original statistical hypothesis of interest is well-supported by the Objective cognitive constructivism epistemological framework; see [27–29, 31–33, 36–38, 40–42]. In contrast, the definition of the associated pragmatic hypothesis as a crisp set may be considered an over-simplification. Rough and fuzzy sets offer a theoretical framework that allows the construction of generalized pragmatic hypotheses, see [139–142, 180, 181, 208]. Rough or fuzzy pragmatic hypotheses should be able to overcome artificial limitations imposed by crisp set representations, support natural and intuitive interpretations in varied contexts of application, and still preserve the best logical properties of the GFBST.

The theoretical framework of paraconsistent logic can be used to interpret sensitivity analyses of the e-value concerning changes in the statistical model’s prior or reference measures, see [26]. The same article shows how to integrate such sensitivity tests into (crisp) confidence intervals. Following the same rationale used in the last paragraph, these crisp intervals could be generalized to rough or fuzzy sets able to provide better representations of pertinent uncertainties.

Law, Complexity, and (In)Consequence in Social Systems

Early scientists used the expression natural law as a metaphor, that ferried to the domain of nature the normative character that a law has over human behavior. Now that mankind learned a good deal about natural laws, including their mathematical, logical, epistemic, ontological, and metaphysical characteristics, we could perhaps travel the metaphorical path taken by early scientists in the opposite direction. We could do so in an attempt to use lessons learned about natural laws to better understand the nature of laws intended to establish norms for human behavior, relations, and interactions in social systems. We believe this task can be accomplished within the framework provided by Niklas Luhmann (1927–1998) Sociological theory of law, see [39, 240, 246, 247, 255–257, 261, 263, 265].

In Luhmann’s theory of law, the main purpose of the legal system is congruent generalization of normative behavior expectations, see [255, pp.77,82]. This emblematic statement can be interpreted as follows, see [39]: In Luhmann’s view, norms are neither preexisting conditions nor a priori factual realities. Instead, norms are conceived as intentional projects or idealized models of how society should be; see [255, p.40]. Moreover, in such idealized models, social harmony is based on establishing well-defined, stable, sustainable, and reliable behavioral patterns, also known as eigen-behaviors or eigen-solutions, see [39, 247]. Furthermore, laws and regulations of a society should reflect its norms, stimulating/penalizing forms of conduct that sustain/disrupt virtuous eigen-behaviors. Finally, in Luhmann’s view, social norms are not static but essentially dynamic, co-evolving (over long runs) with the behavioral patterns in the society that they simultaneously try to describe, regulate and stabilize (in their present forms, over short runs). In this context, we can try to reinterpret some lessons about model selection and their complexity learned in Section 9, claiming that, as in the case of good empirical laws, good social laws should follow the golden path of equilibrium, avoiding extremes of scarcity and excess, see also [28, 32].

On the one hand, oversimplified social laws fail to capture important distinctions considered necessary or relevant for establishing sustainable eigen-behaviors. On the other hand, excessively complex legislation creates all sorts of misinterpretations, unforeseen loopholes, and other unintended consequences. Moreover, such spurious side effects may not only obstruct virtuous eigen-behaviors, but even induce vicious ones (that must then be detected, identified, and inhibited). Furthermore, increasing legal complexities imply increasing processing times, delayed justice, and greater economic costs to operate the legal system – all burdens to be paid by the society the same system serves.

The areas of Logic and Computer science have developed several methods to measure computational complexity, either by counting processing operations in an algorithm or by accessing the
code length of its description, see [146, 149, 172, 199, 227, 228]. Some of the methods used to
measure complexity in statistical models have already been explained in Section 9. Meanwhile, In-
formation science and Systems’ theory have focused on entropy related measures of complexity, see
[31, 34, 105, 109, 113, 143, 176]. In contrast, the mathematical treatment of complexity in theoretical
and empirical legal studies is still incipient, see [203, 204, 215, 215, 241, 253, 258, 259, 262]. Naïve
adaptations of complexity measures artificially borrowed from other areas often fail to capture relevant
aspects of the legal environment.

Although this topic of further research can potentially benefit from all the sophisticated technical
developments and approaches to complexity theory already mentioned, it also requires a healthy dose
of critical thinking applied to the human condition in daily life, yet another area of interest of Walter’s
research group, see [128].

Individual vs. Collective Liability, Legal Burden of Proof, etc.

As commented at the end of Section 2, the Onus Probandi and In Dubio Pro Reo principles, as they are
used in the legal system, were a source of inspiration for the definitions of the e-value and the FBST,
see [25, 39, 93, 163, 190, 250, 254, 260]. Moreover, it has long been recognized that many decision
procedures traditionally used in Bayesian statistics are incompatible with these principles. Dennis
Lindley uses the famous gatecrasher example to argue that, under certain circumstances, collective
responsibility is a preferable principle of justice, and that legal decision procedures implying collective
liabilities and punishments should be accepted, see [132, 185]. Furthermore, Lindley shows how legal
decisions based on Bayes Factors are compatible with such a collective liability principle.

The right tool for the right job! In accordance with this dictum, further research could contrast the
legal and social implications of alternative liability principles, their potential or desirability, and their
compatibility with alternative inference theories, decision procedures, and epistemological frameworks.

**e-value and FBST under Non-Standard Regularity Conditions**

The asymptotic consistency and standardization procedures presented in Section 5 rely on standard
regularity conditions that include continuity and differentiability of several functions defining the sta-
tistical model, and also the location of the true value of the parameter at the topological interior of the
parameter space. However, many interesting models present non-standard conditions. For example,
mixture models and separate hypotheses models often present hypotheses at the (topological) border of
the parameter space, see [12, 13], whereas eigenvalue and eigenvector inference problems may present
far more challenging conditions, see [209]. The methods presented in [131, 138, 211] and references
therein provide tools to generalize the e-value standardization procedure and the corresponding asymp-
totic results to non-standard regularity conditions.

Other non-standard inference conditions arise in semi-parametric models like, for example, Hilbert
space models based on Fourier, orthogonal polynomials, wavelets, and other infinite functional bases.
In these cases, it is often necessary to use weakly informative priors that dampen high-frequency modes
that effectively reduce the size of resulting truncated models, see [42, Sec.6] and references therein.
Further research should explore convenient forms and characteristics of such weakly informative priors
and related topics.

**Efficient Computational Methods and Implementations**

The e-value and FBST research program would greatly benefit from computational tools tailor-made
for computational efficiency, reliability, and ease of use in the numerical tasks described at Section 8.
Most beneficial would be the construction of an open-source and user-friendly environment combining
specially adapted versions of the following techniques: (1) Efficient Monte-Carlo integrators based on
Hit-and-Run, Nested Sampling and similar algorithms, see [129, 178, 216, 217]; (2) Efficient computa-
tional condensation procedures used in conjunction with the aforementioned Monte-Carlo procedures,
see [175, 233]; (3) Efficient stochastic optimization methods working in conjunction with the aforemen-
tioned Monte-Carlo procedures, see [194, 218, 221, 225, 226]; (4) Higher order asymptotic approxi-
mations used either to develop stand-alone fast e-value calculation procedures or to develop variance
reduction techniques for the aforementioned Monte-Carlo procedures, see [3, 20–22, 45–47, 94]; (5) Efficient numerical convolution procedures, used in conjunction with the aforementioned condensation procedures, as required in the analysis of complex models as described in Section 6.

Acknowledgments

The authors take this opportunity to thank Prof. Walter Carnielli for, along the years, inviting, welcoming, and stimulating investigations, discussions, and interactions at areas of overlap between logic, probability, statistics, formal methods in science, philosophy of science, and general philosophy. The authors have benefited greatly from Walter Carnielli’s encyclopedic knowledge, open-mindedness, positive attitude, and active engagement in fostering research in all these areas, either at his home base at CLE - the Center for Logic, Epistemology, and History of Science of UNICAMP, or throughout Brazil and around the world. The authors are grateful to Profs. Abilio Rodrigues, Alfredo Freire and Henrique Antunes for organizing a Festschrift celebrating Walter’s 70th birthday.

The authors thank the organizers of the following conferences, where this paper was presented: CLMPST-2023, the 17th Congress on Logic, Methodology and Philosophy of Science and Technology, at FCE-UBA, Buenos Aires, July 24-29, 2023; Principia-2023, the 13th Principia International Symposium, at UFSC, Florianópolis, August 14-17, 2023; RatioLog-2024, the 2nd Workshop on Logic, Rationality, and Probability, at CLE-Unicamp, Campinas, April 22-25, 2024; and MaxEnt-2024, the 43rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, at U-Ghent, July 1-5, 2024.

Support, Funding, and Conflicts of Interest

The authors have received direct or indirect support from the following institutions: USP – University of São Paulo (JMS, CABP, MSL, LGE, and WSB); UFSCar – Federal University of São Carlos (RI, RBS); ABJ – Brazilian Jurimetrics Association (JMS, RBS); CNPq – Brazilian National Counsel of Technological and Scientific Development, grants PQ 303290/2021-8 (JMS), PQ 302767/2017-7 (CABP), and PQ 309607-2020-5 (RI); FAPESP – State of São Paulo Research Foundation, grants 2019/11321-9 (RI), CEPID CeMEAI – Center for Mathematical Sciences Applied to Industry 2013/07375-0 (JMS, MSL), and CEPID Shell-RCGI – Research Center for Greenhouse Gas Innovation 2014/50279-4 (JMS); and CAPES – Coordination for Improvement of Higher Education Personnel, finance code 001 (RI, RBS, MAD). Concerning this article, the authors have no conflicts of interest to declare.

References

- e-value Theory and Epistemology


[4] Esteves, Luis Gustavo; Izbiicki, Rafael; Stern, Julio Michael; Stern, Rafael Bassi. The logical consistency of simultaneous agnostic hypothesis tests. Entropy, 18, 256, 2016


[9] Izbiicki, Rafael; Stern, Julio Michael; Stern, Rafael Bassi; Esteves, Luis Gustavo. Logical Coherence in


– e-value Use in Empirical Science


[53] Barahona, Manuel; Rifo, Laura; Sepúlveda, Maritza; Torres, Soledad. A Simulation-Based Study on Bayesian Estimators for the Skew Brownian Motion. Entropy, 18, 7, 241-1, 1-14, 2016.


[56] Brentani, Helena; Nakano, Eduardo Y; Martins, Camila B; Izbicki, Rafael; Pereira, Carlos Alberto. Disequilibrium Coefficient: A Bayesian Perspective. Statistical Applications in Genetics and Molecular Biology, 10, 1, 22, 1-24, 2011.


[76] Kelter, Riko. Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Medical Research Methodology, 20, 88, 1-18, 2020


[82] Lima, Adriano R; Mello, Marcelo; Andreoli, Sérgio; Fossaluza, Victor; Araújo, Célia de; Jackowski, Andrea; Bressan, Rodrigo; Mari, Jair. The Impact of Healthy Parenting As a Protective Factor for Posttraumatic Stress Disorder in Adulthood: A Case-Control Study. PLOS ONE, 9, 1, 1-9, e87117, 2014.

Loschi, Rosangela Helena; Santos, Cristiano C; Arellano-Valle, Reinaldo B. Test procedures based on combination of Bayesian evidences for $H_0$. *Brazilian J. of Probability and Statistics*, 26, 4, 450-473, 2012.

Mathis, Maria Alice de; Rosario, Maria C.do; Diniz, Juliana Belo; Torres, Albina R; Shavitt, Roseli G; Ferrão, Ygor A; Fossaluza, Victor; Pereira, C.A.B; Miguel, Euripides Constantino. Obsessive-compulsive disorder: Influence of age at onset on comorbidity patterns. *European Psychiatry*, 23, 3, 187-194, 2008.


Rincón, Sonia V. del; Rogers, Jeff; Widenschwender, Martin; Sun, Dahui; Sieburg, Hans B; Spruck, Charles. Development and Validation of a Method for Profiling Post-Translational Modification Activities Using Protein Microarrays. *PLoS ONE*, 5, 6, e11332, 1-11, 2010.


Santos, Natalia C.L; Dias, Rosa; Alves, Diego; Melo, Brian Ganassin, Maria; Gomes, Luiz; Severid, Willia; Agostinho, Angelo. Trophic and limnological changes in highly fragmented rivers predict the decreasing abundance of detritivorous fish. *Ecological Indicators*, 110, 105933, 1-8, 2020.

Seixas, Andre Augusto Anderson; Hounie, Ana G; Fossaluza, Victor; Curi, Mariana; Alvarenga, Pedro G; Mathis, Maria A; Vallada, Homero; Pauls, David; Pereira, C.A.B; Miguel, Euripides Constantino. Anxiety Disorders and Rheumatic Fever: Is There an Association? *CNS Spectrums*, 13, 12, 1039-1046, 2008.

Shavitt, Roseli G; Requena, Guaraci; Alonso, Pino; Zai, Gwyneth; Costa, Daniel L.C; Pereira, C.A.B; Rosário, Maria C; Morais, Ivanil; Fontenelle, Leonardo; Cappi, C; Kennedy, James; Menchon, Jose M; Miguel, Euripides; Richter, Peggy M.A. Quantifying dimensional severity of obsessive-compulsive disorder for neurobiological research. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 79, 206-212, 2017.

Sikov, Anna; Stern, Julio M. Application of the full Bayesian significance test to model selection under informative sampling. *Statistical Papers*, 60, 89-104, 2019.


Stern, Julio Michael; Zacks, Shelemyahu. Test procedures based on Bayesian evidences for the independence of Poisson variates under the Bayes Factor. *CNS Spectrums*, 13, 12, 1039-1046, 2008.


Vosseler, Alexander; Weber, Enzo. Bayesian


– General References in Logic and Statistics


Statistical and In-


– General References in Law and Philosophy


26