Utility for Significance Tests

Nathália Demetrio Vasconcelos Moura and Sergio Wechsler

Abstract The range of possible readings among and within the statistical inference, in addition to the relevance of these in the applied context, justify the extensive literature analyzing and comparing the main methodologies. However, the fact that each approach is built upon their own structures, varying even the spaces in which they are evaluated, limit the conclusions to the specified scenarios. As a solution for that, in the context of hypotheses tests, we work with the decision theory, which provides a unique language to incorporate the logic of each existent philosophy. For such, after discussing the main points of the frequentist and Bayesian inference, the main approaches are presented, particularly regarding to precise hypotheses, and then unify by the decision-theoretic viewpoint. Additionally, by through this perspective we analyze, interpret and compare the loss functions of some precise approaches, in the context of significance tests.

Keywords Significance tests · Decision theory · FBST · Loss function · Bayes Fisher

1 Introduction

The main goal of Statistical Inference is to answer about random phenomena based on the available information. For such, it is possible to work with different paradigms, including likelihood-based, fuzzy, among others, with the Frequentist approach being by far the most used. For this school, the probability of an event is given by the limit of the relative frequencies, being such frequencies represented by an entity called parameter, defined according to infinite and hypothetical repetitions of the associated experiment. Particularly relevant, the parameter is responsible for specifying the

N. D. V. Moura (✉) · S. Wechsler
University of São Paulo, São Paulo, Brazil
e-mail: nathdvs@ime.usp.br

S. Wechsler
e-mail: sw@ime.usp.br

A. Polpo et al. (eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Springer Proceedings in Mathematics & Statistics 239, https://doi.org/10.1007/978-3-319-91143-4_17

behavior of the referenced random variable. Nevertheless, dealing with such limit as a fixed quantity, despite unknown, imposes some difficult to analysis. For instance, the need for an infinite sequence of repetitions of the experiment, carried out under the same conditions, or the violation of the Likelihood Principle.

To circumvent such limitations, we have the option of extending the analysis to the Bayesian understanding. In this, by looking at the parameter, the entity of interest, as a latent random entity, we obtain a harmonious reading with the way that uncertainty is commonly used. And the laws of probability being the structure according to which a coherent individual must express his uncertainty. Besides, the axioms of coherence [1], presupposed for such approach, are: simple, interpretable, and intuitive.

However, in practice, there are applications working with different readings, particularly with regard to the Hypothesis Tests, and even more to the Precise Hypothesis case. As a solution, we will address the Hypothesis Tests in a single language: decision theory, representing the main logics and objectives through the respective loss functions. Additionally, by through this perspective, we analyze, interpret and compare the loss functions of some precise approaches, in the context of significance tests.

2 Decision Theory

Aiming to structure a methodology that helps us choose the best action taking into account our objectives, circumstances and knowledge, we have the **decision theory**. In this, the action to be taken admits values in the decision space $\mathscr{D}$, and is influenced by the results of an entity involving uncertainty, called Ω. So, given the preferences of the decision agent, given by the loss function $L(\cdot)$ in relation to the possible consequences ($\mathscr{D} \times \Omega$), we get the optimal choice.

For this, we look for the decision so that the associated loss is minimal. However, since the choice must be made without knowledge of the state of nature, we assign probability to the set Ω, which is therefore seen as a random variable. Thus, we estimate its behavior through the understanding that the decision agent has on the parametric space, represented by the probability distribution $\pi(\theta)$, called priori.

Considering also the case where the decision agent has access to a sample x, where the respective random variable X has its sigma-algebra of subsets of the sample space (χ) indexed by Ω, the decision agent starts to contemplate the knowledge of such evidences. And the action is specified according to a **decision rule** δ,

$$\begin{aligned} \delta_\pi : \chi &\to \mathscr{D} \\ x &\mapsto \delta_\pi(x). \end{aligned} \tag{1}$$

Thus, by means of the **priori expected loss**, or **risk against the priori**,

$$R_\pi(\delta) = E_\pi[L(\delta(X), \Theta)] = \int_\Omega \int_\chi L(\delta(x), \theta) f(x|\theta) \pi(\theta)\, dx\, d\theta. \tag{2}$$

Therefore, we seek for the strategy that minimizes the risk in relation to $\pi(\theta)$, so that $\delta^*_\pi = arg\ \min_{\delta \in \mathscr{D}} R_\pi(\delta)$. And, if the order of integration in (2) is alterable, the δ^*_π is equivalent to finding the rule that minimizes the expected loss a posteriori, or **risk against the posteriori** π, that is,

$$r_{\pi(\cdot|x)}(\delta) = \mathbb{E}_{\pi(\cdot|x)}[L(\delta(X), \Theta)] = \int_{\Omega} L(\delta(x), \theta)\pi(\theta|x)\, d\theta. \tag{3}$$

3 Hypothesis Testing

Hypothesis tests have the purpose of indicating the most plausible scenario among a collection of conjectures. However, it is usual to work with only two premises, so that they configure a partition of the parametric space Ω. Typically named as null and alternative, we have, respectively: $H_0 : \Theta \in \Omega_0$ and $H_1 : \Theta \in \Omega_1$.

In theoretical terms, the procedures are specified by a function φ, defined in class $\{\varphi : \chi \to \{0, 1\}\}$, so we decide by H_0 if $\varphi = 0$, e H_1 otherwise. Having further that the value of φ is determined by means of a Rejection Region, such a subset of the sample space is mathematically given by: $\varphi^{-1}(\{1\}) = \{x \in \chi : \varphi(x) = 1\}$. Regarding the specification of the hypotheses, there are two types of errors that can occur. The error of type I is given by $\alpha(\varphi) = \text{P}[\varphi(X) = 1|\Theta \in \Omega_0]$, which occurs when we incorrectly label the alternative hypothesis as true. On the other hand, the type II error is defined by $\beta(\varphi) = \mathbb{P}[\varphi(X) = 0|\Theta \in \Omega_1]$, in relation to the null hypothesis.

Additionally, in the frequentist context, it is usual to still work with the Power Function, or Power of Test. Such quantity associates the probability of rejecting the null hypothesis at each value of Θ. Then, we define the size of the test, given by the supreme power function, considering only the values of $\Theta \in \Omega_0$, i.e., $\alpha = \sup_{\Theta \in \Omega_0} \mathbb{P}_\varphi[\varphi = 1|\theta]$. Finally, we call the value α_0 the significance level, if this is the upper limitation for the other test sizes. Whereas, in the Bayesian context, we work directly with the posteriori probabilities of the hypotheses. Now we describe the main approaches.

3.1 Fisher and p-Value

The most widespread reading in relation to hypothesis testing, refers to the philosophy of Sir Karl Popper, disseminated in the statistics area by Sir Ronald Fisher. According to this, a hypothesis can never be proven by an empirical study. However, a counterexample is sufficient for its negation. In hypothesis testing, such a premise implies that we consider inductive reasoning, so regardless of the amount of evidence in favor of the premise in question, it should not be accepted [2]. Although it is not necessary to indicate whether we are dealing with the null or alternative hypothesis, it is usual to specify H_0, so this is the one that we attach the greatest importance.

For the context discussed, the descriptive level of observed significance (or p-value), introduced by Pearson, is presented as an appropriate tool. This is because the p-value searches from the unobserved samples for evidence at odds with the null hypothesis, considering for such, that the related experiment is fixed. Thus, by ordering the sample space given by H_0, we examine the probability of obtaining samples as extreme as that observed. However, this metric has a number of undesirable characteristics, such as its magnitude being dependent of sample size, or the difficulty of interpretation, since the conditional definition $\mathbb{P}(x|\Omega_0)$ is summarily intuited as a conditional probability $\mathbb{P}(\Omega_0|x)$. In any case, the principle of seeking evidence against H_0, instead of evaluating both hypotheses, is diffused to the point of having a specific class of tests, called **significance tests**.

3.2 Neyman–Pearson and Likelihood Ratio

The perspective advocated by Jerzy Neyman and Egon Pearson (N–P) complements the frequentist scenario regarding hypothesis testing. For this reading, we shall initially consider that the test consists of simple hypotheses, that is, $H_0 : \Theta = \theta_0$ and $H_1 : \Theta = \theta_1$. Thus, there is a critical region given in function of the ratio of probabilities evaluated in the respective subspaces Ω, that is, $\lambda(X) = f(X|\theta_0)/f(X|\theta_1)$. However, given the impossibility of simultaneously controlling the two errors involved, the analysis is limited to the family consisting of the significance level tests α_0. Formally, for $k \geq 0$,

$$\varphi^*(x) = \begin{cases} 1 \text{ se } \lambda(x) < k \\ 0 \text{ se } \lambda(x) > k. \end{cases} \tag{4}$$

In case one of the assumptions is compound, say H_1, we restrict the domain to the Uniformly Most Powerful (UMP) tests. In general, terms, to extend the analysis with some guaranteed properties, it will always be necessary to continue applying restrictions in the domain of tests. Additionally, there are some undesirable characteristics, like the imbalance between errors I and II when the sample size increases, sometimes reaching the inversion of the initially specified match. DeGroot reread the question from a broader perspective, working with the minimization of the linear combination of errors. And later, Pericchi and Pereira [3] generalized the idea, by weighing the likelihoods, obtaining a globally optimal test, plus a balance between the specified errors and the sample size.

3.3 Bayes and Conditional Measures

In contrast to the frequentist theory, which bases its conclusions on samples and unobserved events, Bayesian Inference presents conclusions derived directly from the parametric space. Thus, we can indicate a premise with greater chance of occurrence through the posteriors ratio (denominated Bayes Factor) and the loss function used, that is,

$$\frac{\mathbb{P}(\Theta \in \Omega_0|x)}{\mathbb{P}(\Theta \in \Omega_1|x)} = \frac{\mathbb{P}(X|\Theta \in \Omega_0)}{\mathbb{P}(X|\Theta \in \Omega_1)} \frac{\mathbb{P}(\Theta \in \Omega_0)}{\mathbb{P}(\Theta \in \Omega_1)} \geq k(L(d, \Theta)). \tag{5}$$

This reasoning is interesting, since it contemplates not only the acceptability of an isolated hypothesis, but also the circumstances of the said complement, without priorities.

4 Precise Hypotheses

Hypothesis tests have an important special case: when the conjecture of interest has Lebesgue measure zero, also known as precise hypothesis. The best-known example is the case where the parametric space is defined in the real line: $H_0 : \Theta = \theta_0$ versus $H_1 : \Theta \neq \theta_0$, circumstance which we will give emphasis.

The absence of probability in Ω, does not result in mathematical restrictions in the frequentist approach. However, in the context of significance tests, the constraint of the subspace Ω_0 assigns particular importance to the structure of Popper, given the limitation of the hypothesis in relation to the parametric space as a whole. Whereas in the Bayesian context, if the priori distribution on Ω is continuous, the posterior probability of the subset Ω_0 will be zero, invalidating the usual approaches, justifying, therefore, the development of other criteria. Following we introduce the main criteria.

4.1 Jeffreys

The Jeffreys Test, the most widespread approach, circumvented the problem of the posterior probability of Ω_0 by imposing the specification of a priori with positive probability for H_0. Thus, we started to have $\pi_{\theta_0}(\theta, \zeta)$ defined according to a combination of probabilities: $(1 - \zeta)\pi(\theta)$ to Ω_1 and ζ to Ω_0. Such rebalancing is not a problem if it is, in fact, the analyst's opinion. However, as usually it is only a practical palliative for a mathematical limitation, we violate the principle of coherence, in addition to requiring a greater amount of evidence against H_0 to enables its rejection.

4.2 FBST

In order to develop a Bayesian significance test that holds the coherence assumptions, Pereira and Stern [4] introduced the Full Bayesian Significance Test (FBST). Although this test is feasible for applications in different spaces, its contribution is more expressive in the context of the precise hypotheses, as it is developed based on the principle of least surprise, aiming for evidence in favor of the null hypothesis.

For such, it sorts the parametric space according to the posteriori probability, and seeks the θ^* that belongs to the region of H_0 and its density is maximum. Then, we form the tangent set to the null hypothesis, configured by all points with density lower than the obtained θ^*. Formally,

Definition 1 For the tangent set $T(x) = \{\theta : \pi(\theta|x) > \sup_{\Omega_0} \pi(\theta|x)\}$ the FBST evidence measure in favor of H_0 is: $EV(\Omega_0, x) = 1 - \int_{T(x)} \pi(\theta|x) d\theta$.

For high values of $EV(\Omega_0, x)$, or *e-value* as it is also known, θ_0 will be among the most likely points a posteriori, and will favor the null hypothesis. Additionally, this approach presents advantages as: intuitive logic, geometric interpretation, consistency, and invariance under one-to-one parameter transformations.

5 Loss Function

In order to approach the tests of significance according to a single language, we work with decision theory. For this, we consider the space of decisions $\mathscr{D}$, given by $\{d_0, d_1\}$, where d_i denotes the action of accepting the hypothesis $H_i : \Theta \in \Omega_i$, with $i \in \{0, 1\}$, and losses L_0 and L_1, respectively. In addition, assuming that there is a differentiated posture in relation to the null hypothesis, the decision is presented in relation to the H_0, this is, d_1 is read as rejection of H_0. Besides, that conservative behavior is incorporated into the analysis through the loss function. Thus, for a sample x, we will have

$$\varphi_\pi(x) = \begin{cases} d_0 \text{ if } \dfrac{\pi(\Theta \in \Omega_0|x)}{1 - \pi(\Theta \in \Omega_0|x)} > \dfrac{L_0}{L_1} \\ d_1 \text{ if } \dfrac{\pi(\Theta \in \Omega_0|x)}{1 - \pi(\Theta \in \Omega_0|x)} < \dfrac{L_0}{L_1}, \end{cases} \tag{6}$$

and the conclusion will be given according to the already known Factor of Bayes. Note that assigning randomness to the set Ω does not invalidate the generalization of the analyzes, since the interest is in replicating the philosophy of each approach and not the system itself. Considering this perspective, follows the description of the FBST, and the Popper‘s perspective (essence of the p-value).

Table 1 Loss function of the FBST test

	Accept H_0	Reject H_0
$\theta \notin T(x)$	b	a
$\theta \in T(x)$	$b + c\,[\mathbb{I}\{\theta \in T(x)\}]$	0

Table 2 Risk for some cases of evidence

$EV(\Omega_0, x)$	$r_{\pi(\cdot\|x)}(d_0)$	$r_{\pi(\cdot\|x)}(d_1)$
0	$b + c$	0
0.5	$b + 0.5\,c$	$0.5\,a$
1	b	a

5.1 FBST and Madruga et al.

By making use only of the information contained in the posterior density, the FBST has been classified as full Bayesian since its genesis. However, only in the work of Madruga et al. [5] this measure was analyzed according to the decision theory, being obtained by minimizing the loss function given from positive a, b and c (Table 1).

Note that, in this case, unlike classical theory, we consider a broader class, wherein the observed sample is also incorporated into the loss function. Thus, assuming that the tangent space of the FBST is defined from the sample, we have, from the minimization of $L(d, \theta, x)$, that the acceptance of the hypothesis H_0 will occur if, and only if, $EV(\Omega_0, x) > \frac{b+c}{a+c}$.

In practical terms, the loss function is structured in order to evaluate the tangent set, that is, we describe our expectations regarding the information brought by the sample. However, once it is a measure of significance, the penalty associated with rejection of the null hypothesis when $\theta \in T(x)$ is smaller. For a better understanding, follows some examples (Table 2).

Usually, by specifying the loss function according to the real values of the parametric space, without worrying about our knowledge or the results of the sample, we are working with the penalty related to the phenomenon, rather than the study itself. Because such scenario is a simplification, we fail to incorporate our preferences regarding the reflexes of the study, such as psychological, financial, and even social issues. Whereas, when dealing with losses according to what we are actually going to obtain, and not with a utopian scenario of absolute knowledge, we have a more realistic reading, as well as a full analysis of the situation.

5.2 Popper and Rice

Aiming at a loss function that reflected Popperian philosophy, Rice [6] reread the decision space so that d_0 represents the decision "Nothing to declare", while for d_1

Table 3 Effects for different values of γ

γ	Penalty for reporting	Penalty for not reporting	Inverse of variability
0.01	10	0.10	99
0.1	3.16	0.32	9.00
0.3	1.83	0.55	2.33
0.5	1.41	0.71	1.00
0.7	1.20	0.84	0.43
0.99	1.01	0.994	0.01

we provide results indicating the rejection of H_0. Thus, by choosing to make such a statement, the conclusions are presented by means of an estimate $\hat{\theta}$, and the said loss is evaluated by the usual quadratic loss. On the other hand, by omitting the results, the loss occurs in terms of no longer known unfoldments and how informative they might be. This loss is represented proportionally to the distance between θ and θ_0. Thus, both losses are maintained on the same scale, allowing the decision agent to specify his/her opinion to the relation between them, by means of a factor γ, hence nothing to declare implies that $\gamma^{1/2}\,(\theta_0 - \theta)^2$, and the rejection of H_0 implies $\gamma^{-1/2}\,(d - \theta)^2$. Therefore, according to Bayes rule, we will report results ($\varphi_\pi^R(x) = d_1$) if,

$$R(\Omega_0, x) = \frac{\mathbb{E}^2_{\pi(\cdot|x)}[\Theta - \theta_0]}{Var_{\pi(\cdot|x)}[\Theta]} \geq \frac{1-\gamma}{\gamma}. \tag{7}$$

In practical terms, we reject H_0 when the estimate is "far" from the tested value. Otherwise, we do not have conclusions. The hesitation related to the statements is represented by the value γ, so that the smaller this quantity, the more skeptical the analysis, and the penalty for reporting results will always be higher than the alternative. Noting also that the product of the weights is fixed, they follow the effects for different values γ (Table 3).

Thus, considering the correspondence between the units of imprecision of the estimate d and the loss inherent in the lack of results, the agent should look for the point of balance between such entities. It should also be noted that the conclusion is taken on the basis of the inverse of a measure of variability, similar to the square of the coefficient of variation, but with a focus on the a posteriori parametric space. Thus, when such a measure of variability is significant, we have indicatives about the lack of accuracy and, consequently, the results are not reported.

In the reading made by Rice, we identify several important gains in relation to Fisher's test: lack of assumptions about repeatability and stopping rules of the experiment, mandatory incorporation of the alternative hypothesis, coherence between rejection of the hypothesis and its subsets and even informational results for large samples.

Table 4 Example 1: $R(\Omega_0, x) \times EV(\Omega_0, x), n = 10$

# Successes	Rice: $\varphi_{\pi}^{R}(x) = d_1$	FBST: $\varphi_{\pi}^{E}(x) = d_0$
0	7.989	0.020
1	1.664	0.123
2	0.173	0.450
3	0.065	1.000
4	0.728	0.471
5	2.080	0.156
6	4.294	0.036
7	7.865	0.006
8	14.040	0.001
9	26.624	0.000
10	64.716	0.000

6 FBST × Rice

Although both FBST and Rice methodologies share the same principle as *Onus Probandi*, where the defendant is to be presumed innocent, the Rice test has as its central concern whether or not to reject the hypothesis, while the Madruga et al. measures its consistency.

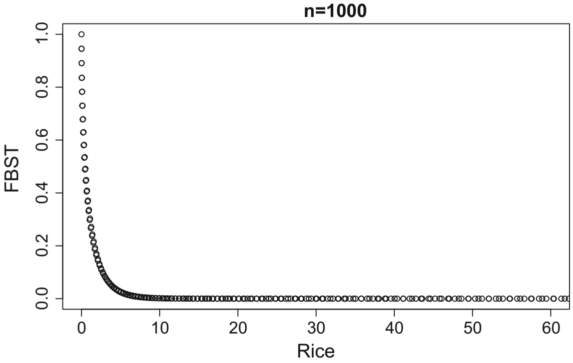
Despite the differences between the two approaches, when $\Omega \subset \mathbb{R}$ the tests are essentially equivalent in the frequentist sense of having a one-to-one relationship between test statistics, for such, follows an illustration.

Example 1 Consider a random sample i.i.d. of Bernoulli's, conditioned in the parameter θ, and the interest in testing the hypotheses $H_0 : \theta = 0.3 \times H_0 : \theta \neq 0.3$, assuming a priori Beta$(1, 1)$. Thus, knowing that all the information of the 2^n possible results can be examined by means of the number of successes obtained, we have the following values for the statistics of Rice and FBST for a sample size 10 (Table 4).

Considering the same analysis for a sample size 1000, we can see in the Fig. 1 the coherence between both results, by means of a negative association, as expected, since we are comparing opposite decisions. Note that we choose to report the results, regardless of the sample size, in the case of γ less than 0.015.

In the inferential context, where the goal is to learn about the parameter, not reporting results seems inappropriate. However, cautious behavior is nothing more than a characteristic of an philosophy, that is, a perspective that is perfectly valid for certain scenarios, such as when the contradiction of the tested hypothesis is not absolute, the sample presents itself as a too limited tool, or the analyst can simply prefer a more cautious stance.

Fig. 1 Example 1: $R(\Omega_0, x) \times EV(\Omega_0, x)$, $n = 1000$



7 Conclusion

In this paper, we discuss the main approaches to hypothesis testing, particularly with regard to significance tests, and reading according to the statistical decision theory. From the point of view of coherence, we concluded that, between the approaches evaluated, we have two satisfactory options from the point of view of coherence: FBST and Rice. Finally, by comparing both readings, we obtain harmonious results with the respective proposals. Besides, they are consistent with each other. For future works, the proposal is to extend the Rice test to parametric spaces larger than one and analyze practical applications of the discussed approaches.

References

1. Kadane, J.B.: Principles of Uncertainty. Chapman & Hall, Boca Raton (2011)
2. Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edinburgh (1935)
3. Pericchi, L.R., Pereira, C.A.B.: Changing the paradigm of fixed significance levels: testing hypothesis by minimizing sum of errors type I and type II. Braz. J. Probab. Stat. **1310.0039** (2013)
4. Pereira, C.A.B., Stern, J.M.: Evidence and credibility: full bayesian significance test for precise hypotheses. Entropy **1**(4), 99–110 (1999). https://doi.org/10.3390/e1040099
5. Madruga, M.R., Esteves, L.G., Wechsler, S.: On the Bayesianity of Pereira-Stern tests. Sociedad de Estadistica e Investigacion Operativa **10**, 291–299 (2001)
6. Rice, K.: A decision-theoretic formulation of Fisher's approach to testing. Am. Statis. **64**(4), 345–349 (2010)