On some assumptions of the null hypothesis statistical testing

Alexandre Galvão Patriota[*]

Departamento de Estatística, IME, Universidade de São Paulo
Rua do Matão, 1010, São Paulo/SP, 05508-090, Brazil

Abstract

Bayesian and classical statistical approaches are based on different types of logical principles. In order to avoid mistaken inferences and misguided interpretations, the practitioner must respect the inference rules embedded into each statistical method. Ignoring these principles leads to the paradoxical conclusions that the hypothesis $\mu_1 = \mu_2$ could be less supported by the data than a more restrictive hypothesis such as $\mu_1 = \mu_2 = 0$, where μ_1 and μ_2 are two population means. This paper intends to discuss and explicit some important assumptions inherent to classical statistical models and null statistical hypotheses. Furthermore, the definition of the p-value and its limitations are analyzed. An alternative measure of evidence, the s-value, is discussed. This paper presents the steps to compute s-values and, in order to illustrate the methods, some standard examples are analyzed and compared with p-values. The examples denunciate that p-values, as opposed to s-values, fail to hold some logical relations.

Key-words: Classical statistics, Inference, Logical Principles, P-value, Statistical hypothesis

1 Introduction

In social sciences, the majority of the events are contingent, full of uncertainties and permeated by nuisance variables. For instance, cognitive skills are affected by a number of factors such as education, culture, age, tiredness, genetics, etc. It is impractical to contemplate all factors that influence a specific cognitive skill. Probability and statistical models are mathematical tools used to handle contingent and uncertain events (Fisher, 1955; McCullagh, 2002; Kadane, 2011). These tools are defined in terms of sets and functions, which are fully consistent with the modern formulation of mathematics[1].

Statistical models are employed to make inferences about unknown quantities and to test the consistency of scientific statements with the observed data (Fisher, 1955). However, statistical models have domains of applicability, internal rules, principles, limitations and so on (Fisher, 1922; Hájek, 2008; Dempster, 1968). It is important to understand those internal features in

[*] email: patriota@ime.usp.br; fax: (+55 11) 3091-6130

[1] The formal apparatus needed for statistical models can be defined in a model of ZFC (Zermelo-Fraenkel with the choice axiom) set theory. The ZFC set theory is the most accepted axiomatic formalization of mathematics that prevents from trivial contradictions such as the Russel's paradox which can emerge from the vagueness of the naïve set theory (see, Terence, 2013, for more details). Natural, Rational and Real numbers, power sets, relations and so on can be derived from models of ZFC set theory (there are other formalizations, but ZFC is, at the present moment, the most studied and verified of them)

order to avoid inadequate interpretations obtained from prohibited inferential rules (Fisher, 1955; Kempthorne, 1976; Berger and Sellke, 1987; Lavine and Schervish, 1999).

The main goal of this paper is to discuss some hidden assumptions underlying the classical statistical models[2] and null hypotheses, see Sections 2 and 3. Section 4 discusses the formal definition of a p-value, Section 5 presents its limitations and reviews a new classical measure of evidence, called s-value, that overcomes some limitations of the p-value. Section 6 provides some standard examples on testing population averages that illustrate the following feature of p-values: they do not respect the reasoning of the logical consequence. The reasoning of logical consequence is: if one hypothesis H_{01} implies another one H_{02}, then, by the logical consequence, we would expect more evidence against H_{02} than that against H_{01}. For example, let μ_1 and μ_2 be two population means. From p-values, it is possible to obtain the following striking result: with the same observed data, it is possible to find more evidence against $\mu_1 = \mu_2$ than against $\mu_1 = \mu_2 = 0$, even though the latter necessarily implies the former. Section 7 concludes the paper resuming the main points discussed in the paper.

2 Statistical Models

It is difficult to introduce probability and statistical models by adopting an easy language without ambiguity. This paper avoids the set-theoretic notation and will not introduce the primary probability space where all quantities are well defined (e.g., random variables, statistics, estimators, induced spaces, etc.). The reader should be aware that the language used here is informal, and to avoid ambiguities it will be required to make many textual caveats. The reader is referred to Cox and Hinkley (1974), Schervish (1995), Lehmann and Casella (1998) and McCullagh (2002) for a detailed discussion on statistical models.

Roughly speaking, the steps before choosing a statistical model are:

1. Define the objectives of the study;
2. Define the population of interest;
3. Define the quantities of interest;
4. Define an adequate experiment to collect the sample.

The practitioner must have prior knowledge to construct an appropriate experiment to access the quantities of interest, for each field has its idiosyncrasies that must be taken into account. The experiment may be randomized in specific strata or layers or clusters (different treatments, genders, groups of risk and so on), and these considerations should guide the researcher to choose the class of probability distributions that will be considered in the statistical model. Typically, in scientific experiments, there are direct observable quantities (age, gender, measured height and weight, etc.) and unobservable quantities (intelligence, “feelings of morale”, “sense of belonging”, etc.). These quantities might be either random or non-random and are ingredients of a statistical model. All random quantities must be well-defined in a probability space.

In this paper, random observable quantities are denoted by upper-case Latin letters, say X or T, and their observed counterparts are denoted by lower-case Latin letters, say x or t. Random

[2] I prefer to use “classical” model rather than “frequentist” model, since the classical model is a mathematical structure that can be interpreted either inside or outside the frequentist paradigm, see Section 2 for more details.

and non-random unobservable quantities are denoted by the Greek letters γ and θ, respectively. The unobservable random quantities are called latent random variables (Bollen, 2002). Let us informally represent a statistical model by the triplet

$$(X, \gamma, \mathcal{M}), \tag{2.1}$$

where X represents the observable random variables, γ represents the latent random variables and $\mathcal{M}$ is a family containing joint probability (density) functions of the random variables, that is, $\mathcal{M} = \{g_\theta : \ \theta \in \Theta \subseteq \mathbb{R}^p\}$, $p \in \mathbb{N}$ where g_θ is a possible joint probability (density) function of (X, γ), for each $\theta \in \Theta$. It should be clear that $\theta \in \Theta$ is an indexer of possible probability distributions, it is not a random variable. Through residual analyses, one can verify empirically if the family $\mathcal{M}$ is adequate or inadequate to model the observable data. It is not possible to assure that the family $\mathcal{M}$ contains the generator mechanism of the data, that is, the mechanism that effectively generates the data. Furthermore, the data's generator mechanism might not even be translatable in terms of probability distributions.

When the probability distribution that governs the random quantities is known, then $\mathcal{M}$ contains only one element, namely $\mathcal{M} = \{g\}$, where $g(x, \gamma) \equiv f_\gamma(x) f_0(\gamma)$ is the joint probability (density) function of the observable and unobservable random variables, with

$$X|\gamma \sim f_\gamma \quad \text{and} \quad \gamma \sim f_0,$$

where f_γ is the probability (density) function of the random variable X given γ and f_0 is the probability (density) function of the random variable γ. Recall that in this latter example, it is assumed that the joint probability distribution that governs the random quantities is known. In this context, it is possible to provide full probabilistic descriptions of the random quantities (mean, variance, quantiles, marginal probabilities, joint probabilities, conditional probabilities, etc.). As aforementioned, in practice it is difficult (or even impossible) to known the generator of the random quantities and the family $\mathcal{M}$ typically has more than one element.

The formal statistical model is defined with sigma-fields and a family of probability measures (see, for instance, Lehmann and Casella, 1998; McCullagh, 2002; Lehmann and Romano, 2005; Patriota, 2013). The reader must keep in mind that model (2.1) is a simplified version that shall help us to understand some important features of the classical statistical model and the null hypothesis statistical testing.

3 Scientific and Statistical Hypotheses

In science, it is common to formulate statistical hypotheses to test scientific statements. A non-trivial step is to translate a scientific statement into statistical language. In the classical paradigm, a statistical hypothesis is a statement about probability distributions that potentially govern the experimental data. That is, in order to create a statistical hypothesis, one must be able to transform a scientific statement in terms of probability distributions. For instance, the statement "this coin is not biased" is typically transformed into "P(this coin turns up head) = 0.5", that is, the following is taken as a hidden principle:

> "This coin is not biased" AND "Theoretical assumptions" $\Leftrightarrow$ "P(this coin turns up head) = 0.5".

The theoretical assumptions are attained from the chosen experiment. One experiment may be performed by independently throwing n times the coin over a smooth surface. The observable random variable is the number of times the coin turned up heads. In this simplified version, no latent variables are considered. Assuming that the coin cannot land on its edge, one statistical model that can represent this experiment is the binomial model $(X, \mathcal{M})$, where $\mathcal{M} = \{g_\theta : \theta \in (0,1)\}$ with

$$g_\theta(k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}, \qquad \text{for } k = 1, \dots, n,$$

where $n!$ is the usual factorial notation, θ is the probability that the studied coin turns up head and $g_\theta(k)$ is the probability that the coin turns up heads exactly k times in the performed experiment. The scientific statement and its statistical counterpart are related by

“This coin is not biased” AND “Theoretical assumptions” $\Leftrightarrow$ “$\theta = 0.5$”.

The null hypothesis is then represented by $H_0 : \theta = 0.5$, that is H_0 is a statement about probabilities: “if the coin is not biased, then [by the above principle and model assumptions] the probability that the coin turns up head is 0.5”. Notice that, unless the practitioner is totally certain of the theoretical assumptions, evidence to reject H does not mean evidence to reject the scientific statement. Indeed, we have that not-H implies that either “This coin is biased” or “at least one of the theoretical assumptions is not adequate”.

Under the null hypothesis H_0, the statistical model reduces to $(X, \mathcal{M}_0)$, where $\mathcal{M}_0 = \{g_{0.5}\}$. In general, the alternative hypothesis is defined to be $H_1 : \theta \neq 0.5$ and under this alternative hypothesis the statistical model is $(X, \mathcal{M}_1)$, where $\mathcal{M}_1 = \{g_\theta : \theta \neq 0.5\}$. Notice that the union of both restricted families under H_0 and H_1 must be the original family, that is, $\mathcal{M}_0 \cup \mathcal{M}_1 = \mathcal{M}$. This means that the original statistical model can be partitioned into two separated statistical models, namely the one generated under H_0 and the other generated under H_1.

In the binomial model, it is implicitly assumed in the “Theoretical assumptions” that “P(this coin turns up head)” does not change over all throws. Of course, this assumption is oversimplified for actual processes, since in each throwing the coin is submitted to impacts causing microscopic cracks, warps and, consequently, modifications in “P(this coin turns up head)” over time. Other statistical models can be implemented by relaxing some of the imposed suppositions: 1) latent random variables can be incorporated to model dependence among the coin flips and 2) covariates may be inserted to model variations in θ. That is, by changing some “Theoretical assumptions”, many statistical models could be used to model the outcomes of the very same experiment.

The concept of coin bias can be further elaborated. One may prefer to relate the statement “this coin is not biased” with the structural topology of the coin, e.g., types of symmetries around the mass center of the coin, etc. Under this latter definition, it is possible to define degrees of bias based on a measure of symmetry and another completely different statistical model will emerge. This simple example illustrates the complexity of statistical models and the problem of translating a simple scientific hypothesis into a statistical language. This example is applied in problems with binary outcomes; for instance, the random variable X may be defined to be the number of allergic patients, out of n, who react positively to a specific treatment.

3.1 Logical relations between the null and alternative statistical hypotheses

In general, a full statistical model is initially specified $(X, \gamma, \mathcal{M})$. After establishing the null and alternative hypotheses H_0 and H_1, reduced statistical models emerge $(X, \gamma, \mathcal{M}_0)$ and $(X, \gamma, \mathcal{M}_1)$ under these hypotheses, respectively, where $\mathcal{M}_0 \cup \mathcal{M}_1 = \mathcal{M}$. The null hypothesis states "at least one marginal probability distribution listed in $\mathcal{M}_0$ generates the observable random variable". Notice that, the alternative hypothesis H_1 is not the negation of H_0. Moreover, the negation of the null hypothesis cannot be written in statistical terms, since not-H_0 includes all possible mechanisms, not necessarily probabilistic ones, that could generate the observable variables X. The negation of H_0 is

> not-H_0 : "It is not the case that 'at least one marginal probability distribution listed in $\mathcal{M}_0$ generates the observable random variable X' ".

Therefore, H_1 does imply not-H_0, but not-H_0 does not imply H_1. Therefore, the practitioner should be aware that a decision between H_0 and H_1 is very limited, since there is an option beyond the disjunction "H_0 OR H_1". As not-H_0 does not imply H_1, "not-H_0 AND not-H_1" is a valid third option. These logical relations lie at the core of many controversies about null hypothesis statistical testing. For instance, Bayesian procedures typically use a prior probability π such that $\pi(H_0 \text{ OR } H_1) = 1$. The problem with this latter procedure is that it gives the impression that the alternative hypothesis is the negation of the null hypothesis, since by the probability properties the following is a consequence: $\pi(H_0) = 1 - \pi(H_1)$, which implies probability zero to the logically valid third option "not-H_0 AND not-H_1"; which means, in some sense, that the practitioner is sure that this third option is not relevant for the statistical analysis. This is exactly what is considered in the analysis derived by Trafimow (2003), which will be discussed in this section.

The statistical hypotheses H_0 and H_1 are not necessarily exhaustive, because, as said previously, the family $\mathcal{M}$ might not contain the data's generator mechanism. Even after making post-data analyses to verify whether the model assumptions are adequate (through residual analyses, simulated envelopes and so on. See Atkinson, 1985; Cook, 1977, 1986, for more details), it is not possible to guarantee that "not-H_0 AND not-H_1" is not a relevant option. For the sake of analysis, let us assume that H_0 and H_1 are exhaustive and mutually exclusive hypotheses, then the following inference rules are valid:

- Empirical evidence to reject H_0 is empirical evidence to accept H_1: not-$H_0 \Rightarrow H_1$.
- Empirical evidence to reject H_1 is empirical evidence to accept H_0: not-$H_1 \Rightarrow H_0$.

However, if the disjunction "H_0 OR H_1" is not exhaustive, then the preceding inference rules are not valid anymore, rather we have the following

- Empirical evidence to reject H_0 is not necessarily empirical evidence to accept H_1: not-$H_0 \not\Rightarrow H_1$.
- Empirical evidence to reject H_1 is not necessarily empirical evidence to accept H_0: not-$H_1 \not\Rightarrow H_0$.

Recall that, as discussed previously, to accept (or reject) H_0 is not the same as to accept (or reject) the scientific hypothesis, unless the practitioner is certain of the theoretical assumptions, which is scarcely the case. The above analysis explicits the main difference between *uncertain inference* and *decision theory* as professor Sir Ronald Fisher argued in some of his papers (Fisher, 1935, 1955). On the one hand, if the disjunction "H_0 OR H_1" is not exhaustive, we have uncertain inference and more difficulties arise, for the universe of possibilities is not closed (we have to deal with the third option). Under this context, the practitioner must not use the inferential rules "not-$H_1 \Rightarrow H_0$" and "not-$H_0 \Rightarrow H_1$". On the other hand, if the disjunction "H_0 OR H_1" is (assumed to be) exhaustive, we have decision theory and the space of decisions becomes well defined, for the inferential rules "not-$H_1 \Rightarrow H_0$" and "not-$H_0 \Rightarrow H_1$" are valid. It is important to note that the classical statistical model is sufficiently general to allow these two situations discussed above:

1. **The Fisherian procedure** considers that "H_0 OR H_1" is not necessarily exhaustive. P-values were initially defined to be used in this situation, they were designed to detect discrepancies between the null hypothesis and the observed data. It is not required even to define an alternative hypothesis; in this context, as aforementioned, some inference rules should not be employed. A very small p-value indicates a large discordance between the postulated null hypothesis and the observed data, however, a non-significant p-value does not indicate evidence in favor of the null hypothesis. Fisher (1955) says: "The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to 'decisions' in Wald's sense, originated in several misapprehensions and has led, apparently, to several more."

2. **The Neyman-Personian procedure** considers that "H_0 OR H_1" is exhaustive. This is the case for the statistical tests developed by Neyman and Pearson. They developed the most powerful test for a fixed significance level (the probability of rejecting the null when it is false). A rejection region is built based on this procedure and a decision is taken by verifying whether the observed sample lies or not in the rejection region. The Bayesian procedure is more aligned with the Neyman-Personian procedure than with the Fisherian, for at least some logical principles are shared between them. Naturally, regarding "H_0 OR H_1" as exhaustive is only an artificial assumption to resolve a statistical problem; the statistician may not consider this as True in an ontological sense.

The above two perspectives lead to different types of statistical inferences. Moreover one cannot be used to invalidate the other, since they use different principles (one considers that "H_0 OR H_1" is exhaustive and the other does not) which lead to different rules of inferences. Many papers in the scientific literature confound these two intrinsically different perspectives (see Hubbard *et al.*, 2003, and the references therein).

Recently, Trafimow (2003), by explicitly assuming that "H_0 OR H_1" is exhaustive, defined p-values by conditional probabilities and employed the rules of conditional probabilities to show that p-values are internally flawed. He wrote "the Bayesian analyses presented earlier not only suggest possible problems with null hypothesis significance testing procedure (NHSTP) but also demonstrate when these potential problems become actual problems and when they do not". Trafimow (2003) deliberately applied the Bayesian reasoning to analyze the p-values' behavior and to conclude that they are flawed. In a recent Editorial note published by "Basic and Applied Social Psychology" (BASP), Trafimow and Marks (2015) communicated that the NHSTP was banned from BASP. The Editorial note said that

> "prior to publication, authors will have to remove all vestiges of the NHSTP (p-values,

> t-values, F-values, statements about significant differences or lack thereof, and so on)."
> (Trafimow and Marks, 2015, , page 1, in Answer to Question 1)

The attempts of writing classical statistics with Bayesian notation is a strong source of misinterpretations and controversies. One reason, as discussed previously, is because their logical reasoning are different. Another reason is that some conditional statements in the classical statistics are not probabilistic statements. The p-value is formally defined in the next section; as the reader shall see, it has nothing to do with the formal definition of conditional probabilities and it is not connected directly with the Bayesian interpretation. In my view, the main problem with the subjective Bayesian approach is that it excludes all possible probability measures outside $\mathcal{M}$ from the very beginning of the statistical analysis[3].

4 Definition of p-values

A p-value is built with the purpose of capturing a disagreement between the observed data and the postulated null hypothesis. In this context, a first step is to define a positive real statistic $T \equiv T_{H_0}$, it is a function of the random sample X which depends on the null hypothesis H_0, such that: the larger its observed value t, the stronger is the disagreement between the observed data and the null hypothesis H_0 (Cox, 1977; Mayo and Cox, 2006; Patriota, 2013). The set $C_{H_0}(t) = \{x : T(x) \geq t\}$ describes all sample values which have stronger disagreements with the postulated null hypothesis H_0 than the observed one t. This set has three important elements, namely: the null hypothesis of interest H_0, the random statistic T and the observed statistic t. Note that T strongly depends on H_0.

If $C_{H_0}(t)$ is small compared to the total set $C_{H_0}(0)$, then the observed experiment provides strong evidence against H_0; this happens when the observed t is large enough to lie in the extreme right tail of the statistics T's distribution. One way to measure the size of $C_{H_0}(t)$ is through probabilities. As the null hypothesis states probability distributions that represent the scientific statement of interest, the p-value is computed for the case with the highest probability in H_0. Let us consider the model without latent variables $(X, \mathcal{M})$, where $\mathcal{M}_0 = \{g_\theta : \theta \in \Theta_0\}$ is the set of probability (density) functions restricted under the specifications of H_0. Let P_θ be the probability measure associated with g_θ, that is, if g_θ is a probability function, then $P_\theta(A) = \sum_{x \in A} g_\theta(x)$ and if g_θ is a probability density function then $P_\theta(A) = \int_A g_\theta(x)dx$, where $A \subseteq \mathcal{X}$ is a measurable set.

[3] Consider the null hypothesis $H_0 : \theta \in \Theta_0$ and the alternative hypothesis $H_1 : \theta \in \Theta_1$, where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \varnothing$. Any Bayesian procedures that use a single prior probability over Θ are saying explicitly that either the null or the alternative are the only possible hypotheses according to the prior distribution. These procedures exclude all possibilities outside the chosen statistical model *a priori*. There are many responses to this critic. Here we present two of them: 1) "it is possible to consider the family of all probability measures for the observed data given θ and use a prior probability over its subsets". However, this family maybe too large to be considered a well-formed set in the ZFC set-theory and the old known contradictions of set theory might arise. Furthermore, by the choice axiom, many subsets of any non-countable family are non-measurable in a probabilistic sense. That is to say that many hypotheses cannot be tested inside the Bayesian approach; and 2) "it is always possible to use another prior probability that gives positive mass for another family of probability measures for the observable data". However, all analyses based on this new prior probability will consider again that the chosen family of possible probability measures for the observable data is certain with probability one. That is, each fixed analysis is an analysis of certainty regarding the family of probability measures for the observable data. These features reflect in the permitted rules of inferences to reject or accept a hypothesis. On the other hand, the classical approach does not impose a prior belief over the class of possible probability measures of the observable data that excludes other possible statistical models. Instead, it says (implicitly) that the chosen family of probability measures for the data is only a possibility among others. This feature prohibits some inferential rules, as discussed in the paper.

The p-value is formally defined by

$$p(H_0, t) = \sup_{\theta \in \Theta_0} P_\theta(C_{H_0}(t)). \tag{4.2}$$

Therefore, as $p(H_0, t)$ is (greater than or equal to) the case with the highest probability in H_0, the smaller the value of $p(H_0, t)$, the larger is the evidence against H_0. Formula (4.2) explicitly says that the classical p-value is not a conditional measure in the probabilistic sense, it is instead a conditional measure in the *possibilistic* sense. The reader should notice that the usual representation p-value $= P(T \geq t|H_0)$ is inadequate, since (a) the probability P is meaningless in the context of classical statistical models and (b) the conditional probability is being misused, since its formal definition is being ignored. The conditional probability is defined by $P(A|B) = \frac{P(A\cap B)}{P(B)}$, where $P(B) > 0$ and A and B are events of the same type (they must be listed in the same sigma-field). As for random variables, the conditional probability is defined analogously for the probability (density) function g_θ. In classical statistics, the events $\{x : T(x) \geq t\}$ and H_0 are not of the same type, for they are not listed in the same sigma-field; otherwise it is a Bayesian-like analysis[4]. In classical statistics, there is not a probability distribution defined over the subsets of $\mathcal{M} = \{g_\theta : \theta \in \Theta\}$ and as $\mathcal{M}$ cannot (even ideally) list all possible measures, a probability measure over the subsets of $\mathcal{M}$ would be conceptually ill-defined[5].

> **Technical remark:** for each observed statistic t, the quantity $P_\theta(C_{H_0}(t))$ is fixed while $P_\theta(C_{H_0}(T))$ is random for each $\theta \in \Theta$. If, for each fixed t, $P_{\theta_1}(C_{H_0}(t)) = P_{\theta_2}(C_{H_0}(t))$ for all $\theta_1, \theta_2 \in \Theta_0$, then the statistic T will be (informally) said to be ancillary to Θ_0, and then the "sup" operation in (4.2) vanishes. This happens in many problems under normal distributions when the interest is centered in testing population means and/or variances. In this context, if T is a continuous random variable and it is ancillary to Θ_0, the distribution of $p(H_0, T)$ is uniform between 0 and 1. This allows the practitioner to interpret a p-value in terms of ideal replications of the performed experiment:
>
>> "if the performed experiment were repeated N times, then it is estimated that $p(H_0, t) \times N$ of those experiments would produce p-values smaller than the observed one."
>
> This interpretation of repeating sampling from the same population is criticized by Fisher (1955). The main argument follows: "if we possess a unique sample in student's sense on which significance tests are to be performed, there is always, ..., a multiplicity of populations to each of which we can legitimately regard our sample as belonging." (see Section 2 of Fisher, 1955, for more details)

[4] In my view, it is one of the many theoretical differences between classical and Bayesian approaches. Notice that, in the classical formulation, as H_0 and $\{x : T(x) \geq t\}$ are not the same type, it is not allowed to use the same measure for these two events.

[5] The main set must contain (ideally) all possible events. It is possible theoretically to define a probability measure over the subsets of a set containing not all possible events, but trivial problems arise as, for instance, giving probability zero to possible events. In the case of the set $\mathcal{M}$, simulated envelopes may be employed to verify if some outside distribution is more adequate than the ones specified in $\mathcal{M}$.

5 Problems of p-values and an alternative measure of evidence

The p-value is a coherent measure to verify a possible discrepancy between a fixed null hypothesis and the observed data. Nevertheless, there is a serious limitation in the use of p-values in nested hypotheses. Consider that the p-value's computation under $H_0^{(1)}$ is extremely complicated. Let $H_0^{(2)}$ be an auxiliary hypothesis such that $H_0^{(1)} \Rightarrow H_0^{(2)}$, i.e., if $H_0^{(1)}$ is *true*, then $H_0^{(2)}$ is *true*. By logical reasoning: if $H_0^{(2)}$ is false, then $H_0^{(1)}$ must also be false. The practitioner, led by this logical reasoning, would compute the p-value under $H_0^{(2)}$ and conclude that if there is evidence to reject $H_0^{(2)}$, i.e., the p-value computed under $H_0^{(2)}$ is significantly small, then there must be evidence to reject $H_0^{(1)}$. However, p-values do not allow this latter logical reasoning. That is, it is not guaranteed that $p(H_0^{(1)}, t) \leq p(H_0^{(2)}, t)$, see Section 6 for numerical examples. This happens because the test statistic T is built for a specific null hypothesis, therefore, the respective p-value is valid only for this specific null hypothesis; for more details, see, for instance, Schervish (1996) and Patriota (2013). In previous work, Patriota (2013) proposed an alternative classical measure of evidence that meets the above logical reasoning; it is called s-value and will be presented in what follows.

The general purpose of the s-value is almost the same as of the p-value: to verify a discrepancy of null hypotheses with the observed data, but maintaining all logical consequence among null hypotheses. In order to define s-values, let us consider the simplest statistical model without latent variables $(X, \mathcal{M})$, where $\mathcal{M} = \{g_\theta : \ \theta \in \Theta \subseteq \mathbb{R}^p\}$ and let P_θ be the probability measure associated with the probability (density) function g_θ. The likelihood-ratio statistic is

$$\lambda(\theta; x) = \frac{g_\theta(x)}{\sup_{\theta \in \Theta} g_\theta(x)},$$

provided that $\sup_{\theta \in \Theta} g_\theta(x) > 0$. Notice that, $0 \leq \lambda(\theta; x) \leq 1$ for all $\theta \in \Theta$. The likelihood-ratio confidence region with significance level α is defined by

$$\Lambda_\alpha(x) = \{\theta \in \Theta : \ \lambda(\theta; x) \geq c_\alpha(\theta)\},$$

where

$$P_\theta(\lambda(\theta; X) \geq c_\alpha(\theta)) \geq 1 - \alpha, \qquad \inf_{\theta \in \Theta} P_\theta(\lambda(\theta; X) \geq c_\alpha(\theta)) = 1 - \alpha$$

and $0 \leq c_\alpha(\theta) \leq 1$. The following equivalent notation may be used

$$P_\theta(\lambda(\theta; X) \geq c_\alpha(\theta)) \equiv P_\theta(\Lambda_\alpha(X) \ni \theta).$$

The quantity $P_\theta(\Lambda_\alpha(X) \ni \theta)$ is the probability of $\Lambda_\alpha(X)$ to contain θ, under the measure P_θ. This is the formal definition of a general confidence region for the parameter θ (Schervish, 1995).

For some statistical models (normal distribution in general), the following occurs:

$$P_\theta(\lambda(\theta; X) \geq c_\alpha(\theta)) = 1 - \alpha \quad \text{for all } \theta \in \Theta,$$

in this case, the confidence region is said to be exact. For exact confidence regions, the value $c_\alpha(\theta)$ is the $(1 - \alpha) \times 100\%$ quantile of the random variable $\lambda(\theta, X)$. Observe that $\Lambda_\alpha(x)$ contains all θ's that generate likelihood values greater than (or equal to) $c_\alpha(\theta)$ times the largest likelihood value, namely, $\sup_{\theta \in \Theta} g_\theta(x)$. This set is intuitive, for it contains the optimal values for $\theta \in \Theta$ according to the likelihood function. The definition of s-values follows.

Definition 5.1. *Let Θ_0 be a non-empty parameter subset related with H_0 and let $\Lambda_\alpha(x)$ be the likelihood-ratio confidence region with significance level α. Then, the s-value is defined by*

$$s(H_0; x) \equiv s(\Theta_0; x) \equiv \sup\{\alpha \in [0,1] : \ \Lambda_\alpha(x) \cap \Theta_0 \neq \varnothing\}.$$

If $\Theta_0 = \varnothing$, define

$$s(\varnothing; x) \equiv 0.$$

This general definition is valid for general hypotheses. Let Θ_{01} and Θ_{02} be two parameter subsets related with the hypotheses $H_0^{(1)}$ and $H_0^{(2)}$, respectively. In this context, if $H_0^{(1)} \Rightarrow H_0^{(2)}$, then $\Theta_{01} \subseteq \Theta_{02}$; Patriota (2013) showed that the following always occurs $s(\Theta_{01}; x) \leq s(\Theta_{02}; x)$. A possible interpretation for the s-value, under the regular conditions stated in Patriota (2013) and assuming that Θ_0 is non-empty and closed, reads

> "$s(\Theta_0, x)$ is equal to the maximum significance level α_M such that $\Lambda_{\alpha_M}(x)$ and Θ_0 have at least one element in common".

The smaller $s(\Theta_0, x)$ is, the more distant Θ_0 is from the maximum likelihood estimate of θ and, consequently, the more unlikely H_0 is according to the likelihood-ratio confidence region. Observe that, if $H_0 : \theta = \theta_0$, where θ_0 is a given vector (or number if $\Theta \subseteq \mathbb{R}$), then $\Theta_0 = \{\theta_0\}$ and the s-value reduces to

$$s(\{\theta_0\}; x) = \max\{\alpha \in [0,1] : \ \theta_0 \in \Lambda_\alpha(x)\}$$

and its interpretation reads

> "$s(\{\theta_0\}, x)$ is equal to the maximum significance level α_M such that $\Lambda_{\alpha_M}(x)$ contains θ_0".

Therefore, the farther away θ_0 is from the center of $\Lambda_\alpha(x)$, which in regular conditions is the maximum likelihood estimative, the more the observed evidence is against H_0. Patriota (2015) studied the likelihood-ratio statistic as a measure of evidence and compared it with the s-value and posterior distributions. González *et al.* (2016) employed the s-value to study confidence sets for observed samples.

5.1 Types of decisions

In this section, some types of decisions are studied. Let $\widehat{\theta}$ be the maximum likelihood estimative of θ, then, under regular conditions (Cox and Hinkley, 1974, Ch. 9), we have that $\widehat{\theta} \in \Theta$ and it exists.

First case: no alternative hypothesis is defined, then the general advice of this paper is to use the s-value as a thermometer of discrepancy between null hypotheses and the observed data. The smaller is $s(\Theta_0, x)$, the stronger is the evidence against H_0. Patriota (2013) showed that if $\widehat{\theta} \in \Theta_0$, then $s(\Theta_0, x) = 1$ and the observed data produce no evidence against H_0, which does not mean evidence in favor of H_0. In a working paper, we are showing that s-values are always greater than p-values (based on the likelihood-ratio statistic) for some specific models. This indicates that if a s-value is small, then the respective p-value must be even smaller. Therefore, one could just compute the s-value to verify discrepancies of the null hypothesis with the observed data. The use of s-values is also justified for general hypotheses, because p-values are much more difficult to compute than s-values and furthermore p-values do not satisfy the logical consequence.

Second case: an alternative hypothesis H_1 is defined and let Θ_1 be its related parameter space. Patriota (2013) showed that, on the one hand, if $\widehat{\theta} \in \Theta_0$, then $s(\Theta_0, x) = 1$; on the other hand if $\widehat{\theta} \in \Theta_1$, then $s(\Theta_1, x) = 1$. If the practitioner wants to decide between H_0 or H_1, then there are three possibilities

- If $s(\Theta_1, x) = 1$ and $s(\Theta_0, x) = a$, then reject H_0 and accept H_1 whenever a is sufficiently small.
- If $s(\Theta_1, x) = b$ and $s(\Theta_0, x) = 1$, then accept H_0 and reject H_1 whenever b is sufficiently small.
- If $s(\Theta_1, x) = s(\Theta_0, x) = 1$ and neither a nor b are sufficiently small, then neither reject nor accept H_0. More data are required.

The threshold values for a and b are being studied. They depend on the sample size, effect sizes, error of type I and II, power of the test, severity (Mayo and Spanos, 2006; Mayo and Cox, 2006), and/or other factors. Notice also that more than one alternative hypotheses $H_1, \ldots, H_k$ can be defined. It is possible to use the s-value in the latter context, but it is beyond the scope of this paper.

Izbicki and Esteves (2015) investigated some properties of statistical test procedures, namely: monotonicity, intersection consonance, union consonance and invertibility. According to Izbicki and Esteves (2015):

1. Monotonicity is a property related to nested hypothesis: if $H_0 \to H_0'$, then a testing scheme that rejects H_0' should also reject H_0.
2. Intersection consonance is a property related to conjunctions: if a testing scheme rejects "H_0 AND H_0'", then it should also reject at least one of the hypotheses H_0 or H_0'.
3. Union consonance is a property related to disjunctions: if a testing scheme rejects each of the hypotheses H_0 and H_0', then it should also reject the disjunction "H_0 OR H_0'".
4. Invertibility is a property related with the null and alternative hypotheses: if a testing scheme rejects the null hypothesis, then it should accept the alternative one and vice-verse.

The s-value satisfies the following property:

$$\forall\ \Theta_0 \subseteq \Theta,\ s(\Theta_0, x) = \sup_{\theta \in \Theta_0} s(\{\theta\}, x). \tag{5.3}$$

By property stated in Equation (5.3), the following property is entailed: for all $\Theta_0 \subseteq \Theta_0' \subseteq \Theta$, $s(\Theta_0, x) \leq s(\Theta_0', x)$. Provided that the hypotheses are statements regarding to the parameter space, namely $H_0 : \theta \in \Theta_0$ and $H_0' : \theta \in \Theta_0'$, we have that: (1) $H_0 \to H_0' \iff \Theta_0 \subseteq \Theta_0'$; (2) "$H_0$ AND H_0'" $\iff \theta \in \Theta_0 \cap \Theta_0'$; and (3) "$H_0$ OR H_0'" $\iff \theta \in \Theta_0 \cup \Theta_0'$. By property stated in Equation (5.3), it is straightforward to show that the testing scheme based on the s-value satisfies monotonicity, intersection consonance and union consonance. The testing scheme based on the s-value does not satisfy invertibility, since the s-value allows us to maintain both hypotheses whenever the observed evidence is not strong enough against at least one of the null or the alternative hypotheses.

Some alternative Bayesian measures of evidence can be seen in Diniz *et al.* (2012). The authors studied some relationships between Bayesian and frequentist significance indices. It is beyond the scope of this paper to compare the classical and Bayesian approaches.

5.2 Steps to compute the s-value

The steps to compute the s-value are:

1. Define the statistical model $(X, \mathcal{M})$. Remember that X represents the observable sample and contains n random variables, namely $X = (X_1, \ldots, X_n)$;

2. Define the null hypothesis H_0 and its related set Θ_0;

3. If required, define the alternative hypothesis H_1 and its related set Θ_1;

4. Compute the likelihood-ratio statistic $\lambda(\theta; x)$;

5. Compute $c_\alpha(\theta)$;

6. Compute $\Lambda_\alpha(x)$;

7. Compute $s(\Theta_0, x)$.

8. If required, compute $s(\Theta_1, x)$.

The step 5 is somewhat difficult to execute for some complex statistical models, since for those models the distribution of $\lambda(\theta, X)$ is not trivial and may depend on θ. In those cases, under regular conditions (Cox and Hinkley, 1974), the practitioner may apply the limiting distribution of $-2\log\big(\lambda(\theta, X)\big)$, which is a chi-squared distribution with p degrees of freedom, where $\dim(\Theta) = p$. Then, step 4 reduces to

$$c_\alpha(\theta) = \exp\bigg(-\frac{1}{2}\chi^2_{p,1-\alpha}\bigg), \quad \text{for all } \theta \in \Theta,$$

where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha) \times 100\%$ quantile of a chi-squared distribution with p degrees of freedom. This approximation reduces the complexity, since $c_\alpha(\theta)$ does not depend on θ. Under this asymptotic approximation, the "asymptotic" s-value, denoted by s_a, reduces simply to

$$s_a(\Theta_0, x) = 1 - \inf_{\theta \in \Theta_0} F_{\chi^2_p}\bigg(-2\log(\lambda(\theta, x))\bigg) = 1 - F_{\chi^2_p}\bigg(-2\log\big(\sup_{\theta\in\Theta_0} \lambda(\theta, x)\big)\bigg),$$

where $F_{\chi^2_p}$ is the cumulative distribution of a chi-squared distribution with p degrees of freedom and log is the natural logarithm function. If $\Theta_0 = \{\theta_0\}$, then the asymptotic p-value (i.e., the asymptotic approximation for the p-value) based on the likelihood-ratio statistic coincides with the above asymptotic s-value. Nevertheless, if $\dim(\Theta_0) > 0$ (the Lebesgue dimension), the asymptotic p-value and asymptotic s-value will probably differ from each other. In the asymptotic p-value, the degree of freedom of the chi-squared distribution varies with the dimension of Θ_0; more precisely, the asymptotic p-value based on the likelihood-ratio statistic is

$$p_a(\Theta_0; x) = 1 - F_{\chi^2_q}\bigg(-2\log\big(\sup_{\theta\in\Theta_0} \lambda(\theta, x)\big)\bigg),$$

where $q = \dim(\Theta) - \dim(\Theta_0)$, where $\dim(\Theta) = p$. That is, the cumulative distribution function $F_{\chi^2_q}$ varies with the chosen null hypothesis, whereas for the s-value $F_{\chi^2_p}$ does not vary with the chosen null hypothesis. Patriota (2013, 2014) showed that the asymptotic s-value and p-values (based on the likelihood-ratio statistic) are connected through the following relation

$$s_a(\Theta_0, x) = 1 - F_{\chi^2_p}\big(F^{-1}_{\chi^2_q}(1 - p_a(\Theta_0, x))\big).$$

That is, from a p-value (based on the likelihood-ratio statistic) we can compute the s-value *via* the above formulae. If $p = q$, then $s(\Theta_0, x) = p(\Theta_0, x)$.

6 Numerical examples

In this section, the s-value is applied for univariate and bivariate normal distributions. We consider known variances (and covariances) to maintain the simplicity. All required steps are computed.

Example 6.1. *(Normal distribution, variance known: z test) Let $X = (X_1, \dots, X_n)$ be a sample from a normal distribution with population mean θ and variance 1. Let $H_0 : \theta = \theta_0$ be the null hypothesis of interest. The statistical model is $(X, \mathcal{M})$, where $\mathcal{M} = \{g_\theta : \theta \in \mathbb{R}\}$ and*

$$g_\theta(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\bigg).$$

The likelihood-ratio statistic is

$$\lambda(\theta, x) = \exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\bigg) = \exp\bigg(-\frac{n}{2}(\bar{x} - \theta)^2\bigg),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the maximum likelihood estimate for θ. It is known that

$$-2\log(\lambda(\theta, X)) \overset{P_\theta}{\sim} \chi^2_1,$$

where the symbol "$\overset{P_\theta}{\sim} \chi^2_p$" means "follows a chi-squared distribution with p degrees of freedom, under the law P_θ". Then,

$$c_\alpha(\theta) = \exp\bigg(-\frac{1}{2}\chi^2_{1,1-\alpha}\bigg)$$

and

$$\Lambda_\alpha(x) = \bigg\{\theta \in \mathbb{R} :\ n(\bar{x} - \theta)^2 \leq \chi^2_{1,1-\alpha}\bigg\} = \bigg[\bar{x} - \sqrt{\frac{1}{n}\chi^2_{1,1-\alpha}}, \bar{x} + \sqrt{\frac{1}{n}\chi^2_{1,1-\alpha}}\bigg].$$

The quantity $\sqrt{\chi^2_{1,1-\alpha}}$ coincides with the normal $(1-\alpha/2)$-quantile $z_{1-\alpha/2}$, for instance, for $\alpha = 0.05$, we have $\sqrt{\chi^2_{1,0.95}} = z_{0.97} \approx 1.96$. That is, in this example, Λ_α is the usual $(1-\alpha)$-confidence interval for the population mean.

Let $H_0 : \theta = \theta_0$ be the null hypothesis of interest. The s-value is computed by finding the α-value such that the border of the observed confidence interval $\Lambda_\alpha(x)$ is θ_0. The solution is

$$s(\{\theta_0\}, x) = 1 - F_{\chi^2_1}\big(n(\bar{x} - \theta_0)^2\big).$$

As aforementioned, for this simple null hypothesis, the s-value is precisely the p-value based on the *likelihood-ratio statistic and coincides with the famous z-test. Table 1 depicts numerical s-values to illustrate the univariate normal distribution example for $n = 10$ and $\sigma^2 = 1$. The null hypothesis is $H_0 : \theta = \theta_0$, where $\theta_0 = -1, 0, 1$.*

Example 6.2. *(Bivariate Normal distribution, with known variances and covariances) Let $X = (X_1, \ldots, X_n)$ be a sample from a bivariate normal distribution with population mean $\theta = (\mu_1, \mu_2)^\top$ and covariance-variance matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. The statistical model is $(X, \mathcal{M})$, where $\mathcal{M} = \{g_\theta : \theta \in \mathbb{R}^2\}$ and*

$$g_\theta(x) = \frac{1}{(2\pi)^n} \exp\bigg(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^\top (x_i - \theta)\bigg).$$

The likelihood-ratio statistic is

$$\lambda(\theta, x) = \exp\bigg(-\frac{1}{2}(\bar{x} - \theta)^\top (\bar{x} - \theta)\bigg),$$

where $\bar{x} = (\bar{x}_1, \bar{x}_2)^\top$ is the maximum estimate for θ, where $\bar{x}_1$ and $\bar{x_2}$ are the sample averages of the bivariate sample. Observe that, here $p = 2$. It is also known that

$$-2\log(\lambda(\theta, X)) \overset{P_\theta}{\sim} \chi^2_2.$$

Then,

$$c_\alpha(\theta) = \exp\bigg(-\frac{1}{2}\chi^2_{2,1-\alpha}\bigg)$$

and

$$\Lambda_\alpha(x) = \bigg\{\theta \in \mathbb{R} : \ n(\bar{x} - \theta)^\top (\bar{x} - \theta) \le \chi^2_{2,1-\alpha}\bigg\}.$$

Null hypothesis 1: *Let $H_0^{(1)} : \theta = \theta_0$ be the null hypothesis of interest, where $\theta_0 = (\mu_{10}, \mu_{20})^\top$ is a given vector; then $\Theta_{01} = \{\theta_0\}$. The s-value is computed by finding the α-value such that the border of the observed confidence interval $\Lambda_\alpha(x)$ is θ_0. The solution is (which is also equal to the p-value based on the likelihood-ratio statistic)*

$$s(\{\theta_0\}, x) = 1 - F_{\chi^2_2}\big(n(\bar{x} - \theta_0)^\top (\bar{x} - \theta_0)\big).$$

Null hypothesis 2: *Let $H_0^{(2)} : \mu_1 = \mu_2$ be the null hypothesis of interest, then $\Theta_{02} = \{\theta \in \mathbb{R}^2 : \mu_1 = \mu_2\}$. The s-value is computed by finding the maximum α-value such that*

$$\Lambda_\alpha(x) \cap \Theta_{02} = \{\theta \in \Theta_{02} : n(\bar{x} - \theta)^\top (\bar{x} - \theta) \le \chi^2_{2,1-\alpha}\}$$

has at least one element. The solution is (which is not equal to the p-value based on the likelihood-ratio statistic)

$$s(\Theta_{02}, x) = 1 - F_{\chi^2_2}\big(n \min_{\theta \in \Theta_{02}} (\bar{x} - \theta)^\top (\bar{x} - \theta)\big).$$

Notice that

$$\min_{\theta \in \Theta_{02}} (\bar{x} - \theta)^\top (\bar{x} - \theta) = \min_{\mu \in \mathbb{R}} [(\bar{x}_1 - \mu)^2 + (\bar{x}_2 - \mu)^2] = \frac{n}{2}(\bar{x}_1 - \bar{x}_2)^2.$$

Then,

$$s(\Theta_{02}, x) = 1 - F_{\chi^2_2}\bigg(\frac{n}{2}(\bar{x}_1 - \bar{x}_2)^2\bigg).$$

Recall that the p-value based on the likelihood-ratio statistic is

$$p(\Theta_{02}; x) = 1 - F_{\chi_1^2}\left(\frac{n}{2}(\bar{x}_1 - \bar{x}_2)^2\right)$$

Table 2 presents numerical s-values to illustrate the bivariate normal distribution example for $n = 10$ *and covariance-variance matrix* $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. *The null hypotheses considered are* $H_{01} : \mu_1 = \mu_2 = 0$ *and* $H_{02} : \mu_1 = \mu_2$ *for which it is expected to find more evidence against* H_{01} *than* H_{02}. *The s-values were defined to hold this expected behavior. We purposely choose values for* $\bar{x}_1$ *and* $\bar{x}_2$ *such that p-values are problematic. The figures of Table 2 show that all p-values fail to hold the logical condition for all sample, except for* $\bar{x}_1 = \bar{x}_2 = 0$.

The behavior of p-values depicted in Tables 1 and 2 is not restricted to the examples where dispersion parameters are known. This feature happens also for unknown dispersion parameters, other test statistics, and other statistical models. Here, we consider likelihood-ratio statistics, since we are interested in comparing the p-value with the s-value. The distribution of $-2\log(\lambda(\theta; X))$ is not trivial when the dispersion parameters are unknown and in order to avoid cumbersome computations, we consider only the case with known dispersion parameters.

7 Conclusion

This paper discusses some conceptual and technical problems related to the null hypothesis statistical testing. The scientific and statistical hypotheses and the theoretical assumptions are connected by rules of inferences called *modus ponnes* and *modus tollens*, as studied in Section 3. Unless the practitioner is totally certain of the theoretical assumptions, evidence to reject the null statistical hypothesis does not mean evidence to reject the scientific hypothesis, since the assumptions of a statistical model interfere in this process. Types of decisions in null hypothesis statistical testing depend on important assumptions that are not always made explicit. On the one hand, if the practitioner considers the null and alternative statistical hypotheses are mutually exclusive and exhaustive, then procedures to accept-reject the null statistical hypothesis are justifiable (e.g., Neyman-Pearsonian and Bayesian procedures). On the other hand, if the practitioner considers that the null and alternative statistical hypotheses are mutually exclusive but not exhaustive, then procedures to reject the null statistical hypothesis are preferable (e.g., Fisherian procedures or some other procedures that do not use a belief measure that excludes all possibilities outside the null or alternative hypotheses), since a third option "not-H_0 AND not-H_1" must be taken into account. A statistical procedure developed under one assumption will certainly fail to be appropriated under the other, therefore an extra caution must be taken when comparing different statistical procedures (classical *versus* Bayesian). By construction, p-values do not respect the following logical reasoning: if $H_{01} \Rightarrow H_{02}$, then p-value($H_{02}$) $\not\leq$ p-value(H_{01}). That is, the practitioner must not use the p-value to extrapolate the inference made for H_{02} to H_{01}. This is not a defect in the classic statistical reasoning, because s-values do respect this logic and can be employed in the place of p-values. Asymptotic versions of s-values are simpler to compute than p-values. S-values can be used as a complementary measure of evidence and, as any other statistical measure, some care is needed when using it to make inferences; rules of thumb must be avoided, the inferential conclusions must be always complemented with other statistical tools.

My personal view is that models are useful tools, they can be adequate or inadequate in specific contexts. As for null hypotheses, they can be compatible or incompatible with the observed data; their degree of (in)compatibility with the observed data can be verified through measures of evidence (p-values, s-values, etc.). Statistical analyses have hard philosophical issues that should not be taken for granted, namely: translation problems, meaning of uncertainty, domain of applicability of each method, underlying (philosophical, scientific, logical and statistical) principles and so on. My impression is that science would be more trustful if these issues were taken seriously into account in the statistical analyses. For instance, a p-value (or any other quantitative measure of evidence) smaller than a certain threshold (e.g., 0.05) should not be used directly to reject a scientific hypothesis without further investigations regarding model assumptions, test statistics, sample size, scientific relevance, rules of inferences, adopted principles and so on.

8 Acknowledgements

The author gratefully acknowledges grant from FAPESP (2014/25595-0, Brazil). The author also thanks Prof. Dr. Denis Cousineau, Dr. Jonatas Eduardo Cesar and two anonymous referees for their valuable comments and suggestions that led to an improved version of this paper.

References

Atkinson, A.C. (1985). *Plots, transformations and regression : an introduction to graphical methods of diagnostic regression analysis.* Oxford Science Publications, Oxford.

Berger, J.O., Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence, *Journal of the American Statistical Association*, **82**, 112–122.

Bollen, K.A. (2002). Latent variables in psychology and the social sciences, *Annual review of psychology*, **53**, 605–634.

Cook, R.D. (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook, R.D. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society B*, **48**, 133–169.

Cox, D.R. (1977). The role of significant tests (with discussion), *Scandinavian Journal of Statistics*, **4**, 49–70.

Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Dempster, A.P. (1968). A generalization of Bayesian inference, *Journal of the Royal Statistical Society, Series B*, **30**, 205–247.

Diniz, M., Pereira, C.A.B., Polpo, A., Stern, J.M., Wechsler, S. (2012). Relationship between Bayesian and Frequentist significance indices, *International Journal for Uncertainty Quantification*, **2**, 161–172.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309–368.

Fisher, R.A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society. Series B*, **98**, 39–82.

Fisher, R.A. (1955). Statistical methods and statistical induction, *Journal of the Royal Statistical Society, series B*, **17**, 69–78.

Ginzález, J.A., Castro, L.M., Lachos, V.H. (2016). A confidence set analysis for observed samples: a fuzzy set approach, *Entropy*, **18**, 211.

Hájek, A. (2008). Arguments for-or-against-Probabilism?, *British Journal for the Philosophy of Science*, **59**, 793–819.

Hubbar, R., Bayarri, M.J., Berk, K.N., Carlton, M.A. (2003). Confusion over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing, *The American Statistician*, **57**, 171–182.

Izbicki, R., Esteves, L.G. (2015). Logical Consistency in Simultaneous Statistical Test Procedures, *Logic Journal of IGPL*, Online.

Kadane, J.B. *Principles of Uncertainty*, Chapman & Hall/CRC Texts in Statistical Science, 2011.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypothesis, *Communications in Statistics – Theory and Methods*, **8**, 763–777.

Lavine, M., Schervish, M.J. (1999). Bayes factors: What they are and what they are not, *The American Statistician*, **53**, 119–122.

Lehmann, E.L., Casella, G. *Theory of Point Estimation.* 2th Edition, Wiley, New York, 1998.

Lehmann, E.L., Romano, J.P. *Testing Statistical Hypotheses.* 3th Edition. Springer, New York, 2005.

McCullagh, P. (2002). What is a statistical model, *The Annals of Statistics*, **30**, 1225–1310.

Mayo, D.G, Cox, D.R. Frequentist statistics as a theory of inductive inference, in: Second Lehmann Symposium – Optimality IMS Lecture Notes – Monographs Series, 2006.

Mayo, D.G., Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction, *The British Journal for the Philosophy of Science*, **57**, 323–357.

Patriota, A.G. (2013). A classical measure of evidence for general null hypotheses, *Fuzzy Sets and Systems*, **233**, 74–88.

Patriota, A.G. (2014). Uma medida de evidência alternativa para testar hipóteses gerais, *Ciência e Natura*, **36**, 14–22. `http://www.ime.usp.br/~patriota/medida_evi.pdf`

Patriota, A.G. (2015). A measure of evidence based on the likelihood-ratio statistics. `http://arxiv.org/abs/1510.02950`.

Schervish, M.J. *Theory of Statistics.* Springer Series in Statistics, 1995.

Schervish, M.J. (1996). P Values: What they are and what they are not, *The American Statistician*, **50**, 203–206.

Terence, T. *Compactness and contradiction*, American Mathematical Society, Providence, RI, 2013, (pg. 156). `https://terrytao.files.wordpress.com/2011/06/blog-book.pdf`

Trafimow, D.(2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayess Theorem, *Psychological Review*, **110**, 526–535.

Trafimow D., Marks, M. (2015). Editorial, *Basic and Applied Social Psychology*, **37**, 1–2.

Table 1: S-values for testing $H_0 : \theta = \theta_0$, where $\theta_0 = 0, 1$ for some observed values of $\bar{x}$ when $n = 10$.

$\bar{x}$	$\theta_0 = 0$	$\theta_0 = 1$
0.0	1.0000	0.0016
0.1	0.7518	0.0044
0.2	0.5271	0.0114
0.3	0.3428	0.0269
0.4	0.2059	0.0578
0.5	0.1138	0.1138
0.6	0.0578	0.2059
0.7	0.0269	0.3428
0.8	0.0114	0.5271
0.9	0.0044	0.7518
1.0	0.0016	1.0000
1.1	0.0005	0.7518
1.2	0.0001	0.5271
1.3	<0.0001	0.3428
1.4	<0.0001	0.2059
1.5	<0.0001	0.1138
1.6	<0.0001	0.0578
1.7	<0.0001	0.0269
1.8	<0.0001	0.0114
1.9	<0.0001	0.0044
2.0	<0.0001	0.0016

Table 2: S-values and p-values for testing $H_{01} : \mu_1 = \mu_2 = 0$ (the s-values and p-values are identical) and $H_{02} : \mu_1 = \mu_2$ (the s-values and p-values differ) for some observed values of $(\bar{x}_1, \bar{x}_2)$ that generate problematic p-values (showing that p-values do not respect the logical consequence). The sample size is $n = 10$.

		$H_{01} : \mu_1 = \mu_2 = 0$		$H_{02} : \mu_1 = \mu_2$	
$(\bar{x}_1, \bar{x}_2)$	$\bar{x}_1 - \bar{x}_2$	s-value	p-value	s-value	p-value
(0.00, 0.00)	0.0	1.0000	1.0000	1.0000	1.0000
(0.05,-0.05)	0.1	0.9753	0.9753	0.9753	0.8231
(0.09,-0.11)	0.2	0.9039	0.9039	0.9048	0.6547
(0.14,-0.16)	0.3	0.7977	0.7977	0.7985	0.5023
(0.19,-0.21)	0.4	0.6697	0.6697	0.6703	0.3711
(0.23,-0.27)	0.5	0.5331	0.5331	0.5353	0.2636
(0.28,-0.32)	0.6	0.4049	0.4049	0.4066	0.1797
(0.33,-0.37)	0.7	0.2926	0.2926	0.2938	0.1175
(0.37,-0.43)	0.8	0.2001	0.2001	0.2019	0.0736
(0.42,-0.48)	0.9	0.1308	0.1308	0.1320	0.0442
(0.47,-0.53)	1.0	0.0813	0.0813	0.0821	0.0253
(0.51,-0.59)	1.1	0.0478	0.0478	0.0486	0.0139
(0.56,-0.64)	1.2	0.0269	0.0269	0.0273	0.0073
(0.61,-0.69)	1.3	0.0144	0.0144	0.0146	0.0037
(0.65,-0.75)	1.4	0.0073	0.0073	0.0074	0.0017
(0.70,-0.80)	1.5	0.0035	0.0035	0.0036	0.0008
(0.75,-0.85)	1.6	0.0016	0.0016	0.0017	0.0003
(0.79,-0.91)	1.7	0.0007	0.0007	0.0007	0.0001
(0.84,-0.96)	1.8	0.0003	0.0003	0.0003	0.0001
(0.89,-1.01)	1.9	0.0001	0.0001	0.0001	<0.0001
(0.93,-1.07)	2.0	<0.0001	<0.0001	<0.0001	<0.0001