The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Coherent Hypothesis Testing

Victor Fossaluza, Rafael Izbicki, Gustavo Miranda da Silva & Luís Gustavo Esteves

To cite this article: Victor Fossaluza, Rafael Izbicki, Gustavo Miranda da Silva & Luís Gustavo Esteves (2017) Coherent Hypothesis Testing, The American Statistician, 71:3, 242-248, DOI: 10.1080/00031305.2016.1237893

To link to this article: https://doi.org/10.1080/00031305.2016.1237893

Accepted author version posted online: 28 Sep 2016.
Published online: 28 Sep 2016.

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 3, 242–248: Teacher's Corner
https://doi.org/10.1080/00031305.2016.1237893

Taylor & Francis
Taylor & Francis Group

TEACHER'S CORNER

Coherent Hypothesis Testing

Victor Fossaluza[a], Rafael Izbicki[b], Gustavo Miranda da Silva[a], and Luís Gustavo Esteves[a]

[a]Department of Statistics, University of São Paulo, São Paulo, Brazil; [b]Department of Statistics, Federal University of São Carlos, São Carlos - SP, Brazil

ABSTRACT

Multiple hypothesis testing, an important quantitative tool to report the results of scientific inquiries, frequently leads to contradictory conclusions. For instance, in an analysis of variance (ANOVA) setting, the same dataset can lead one to reject the equality of two means, say $\mu_1 = \mu_2$, but at the same time to not reject the hypothesis that $\mu_1 = \mu_2 = 0$. These two conclusions violate the coherence principle introduced by Gabriel in 1969, and lead to results that are difficult to communicate, and, many times, embarrassing for practitioners of statistical methods. Although this situation is common in the daily life of statisticians, it is usually not discussed in courses of statistics. In this work, we enrich the teaching and discussion of this important topic by investigating through a few examples whether several standard test procedures are coherent or not. We also discuss the relationship between coherent tests and measures of support. Finally, we show how a Bayesian decision-theoretical framework can be used to build coherent tests. These approaches to coherence enlighten when such property is appealing in multiple testing and provide means of obtaining it.

ARTICLE HISTORY
Received April 2015
Revised July 2016

KEYWORDS
Bayesian tests; Coherence; Loss functions; Measures of support; Simultaneous hypothesis testing

1. Introduction

Testing hypotheses has become a widespread quantitative tool in many fields of human knowledge to report the results of scientific experiments. Because of its importance, great advances have been made both on theoretical and practical aspects of multiple (or simultaneous) hypothesis testing—a situation where one aims at testing several hypotheses simultaneously (Shaffer 1995). These developments have been mostly focused on the construction of procedures satisfying statistical optimality criteria. For example, under the Bayesian approach, one typically seeks for procedures that minimize posterior expected loss functions. Similarly, under a frequentist perspective, the main paradigm is the control of various error rates. Detailed accounts on this matter may be found in Shaffer (1995), Hochberg and Tamhane (1987), Farcomeni (2008), and references therein. On the other hand, little emphasis has been given to the rational consistency of the simultaneous inferences drawn from the optimal solutions of such multiple test procedures. As a matter of fact, a practitioner, while analyzing a few hypotheses of interest, often obtains contradictory conclusions such as incoherent measures of support in Schervish's sense (by a coherent measure of support, Schervish 1996 means that "if hypothesis H implies hypothesis H', then there should be at least as much support for H' as there is for H") for the hypotheses under consideration (like p-values (Schervish 1996) and Bayes factors (Lavine and Schervish 1999)) and incompatible sets of rejected hypotheses (Lehmann 1957). For example, Patriota (2013) presented an analysis of variance (ANOVA) setting where the equality of two means, say $\mu_1 = \mu_2$, is rejected but a proper subset of it, $\mu_1 = \mu_2 = 0$, is not. That is, the observed sample point provides enough evidence against $\mu_1 = \mu_2$ but points to the plausibility of $\mu_1 = \mu_2 = 0$. Such peculiar conclusions puzzle practitioners of statistical methods. Thus, some authors argue that, in addition to meeting optimality criteria, simultaneous hypothesis testing should yield logically coherent results. As stressed by Hommel and Bretz (2008) "One could (...) argue that 'power is not everything'. In particular for multiple test procedures one can formulate additional requirements, such as, for example, that the decision patterns should be logical, conceivable to other persons, and, as far as possible, simple to communicate to nonstatisticians."

In this work, we review a logical relationship introduced by Gabriel (1969) that a practitioner may expect from multiple hypothesis testing, the *coherence* property. Such property states that if hypothesis H_0^1 implies hypothesis H_0^2 (i.e., $H_0^1 \subseteq H_0^2$), then the rejection of H_0^2 implies the rejection of H_0^1 (see Figure 1). Equivalently, nonrejection of H_0^1 implies nonrejection of H_0^2.

Although coherence is not the only rational desideratum for simultaneous hypothesis testing (see, e.g., Bickel 2014; Izbicki and Esteves 2015; Esteves et al. 2016 for other properties), it is by far the most emphasized one in the literature under both Bayesian and frequentist schools (see, e.g., Gabriel 1969; Hochberg and Tamhane 1987; Sonnemann 2008; Bickel 2012; Patriota 2013; da Silva et al. 2015; Izbicki and Esteves 2015, and references therein). Nevertheless, this property is often neglected in courses of statistical inference, mathematical statistics, and statistical methods in general. As a matter of fact, in various courses at different levels, we have almost always experienced reactions of surprise from many students when examples of incoherent tests are pointed out. The question of which tests yield coherent conclusions is then recurrent. In this article, we enrich the teaching and discussion of this important matter for both statisticians and practitioners by

CONTACT Luís Gustavo Esteves lesteves@ime.usp.br Department of Statistics, University of São Paulo, Rua do Matão, 1010 - Cidade Universitária, São Paulo 05508-020, Brazil.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

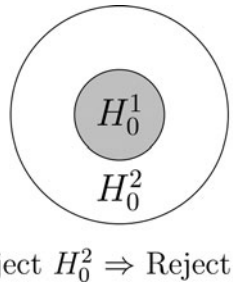

Figure 1. Coherence property: if hypothesis H_0^1 implies hypothesis H_0^2 (i.e., $H_0^1 \subseteq H_0^2$), rejection of H_0^2 should imply rejection of H_0^1. Equivalently, nonrejection of H_0^1 should imply nonrejection of H_0^2.

(i) providing three examples outside the usual scope of statistical linear models in which we examine the coherence of Bayesian tests, uniformly most powerful (UMP) tests, generalized likelihood ratio (GLR) tests, and significance tests (Section 2);
(ii) presenting a connection between coherent hypothesis testing and Schervish's notion of coherent measure of support (Section 3); and
(iii) exploring conditions for simultaneous test procedures to meet coherence under a Bayesian decision-theoretical standpoint (Section 4).

These analyses contribute to better decision-making, as well as to prevent practitioners from being faced with puzzling results from simultaneous tests. Final remarks are made in Section 5.

2. Examples

Next, we present examples of simultaneous tests of nested hypotheses. In each situation, a few solutions are investigated with respect to coherence. Additional details on the derivations of the tests may be found in the Appendix.

Example 1 (Lin et al. 2003). Consider that in a case–control study one measures the genotype in a certain *locus* for each individual of a sample. Results are shown in Table 1. These numbers were taken from a study presented by Lin et al. (2003) that had the aim of verifying the hypothesis that subunits of the gene $GABA_A$ contribute to a condition known as methamphetamine use disorder. Here, the set of all possible genotypes is $\mathbf{G} = \{AA, AB, BB\}$. Let $\boldsymbol{\theta} = (\theta_{AA}, \theta_{AB}, \theta_{BB})$, where θ_i is the probability that an individual from the case group has genotype $i \in \mathbf{G}$. Similarly, let $\boldsymbol{\pi} = (\pi_{AA}, \pi_{AB}, \pi_{BB})$, where π_i is the probability that an individual of the control group has genotype i.

In this context, two hypotheses are of interest: the hypothesis that the genotypic proportions are the same in both groups, $H_0^G : \boldsymbol{\theta} = \boldsymbol{\pi}$, and the hypothesis that the allelic proportions are the same in both groups $H_0^A : \theta_{AA} + \frac{1}{2}\theta_{AB} = \pi_{AA} + \frac{1}{2}\pi_{AB}$. Assuming independence between the groups, the p-values for H_0^G and H_0^A obtained using the standard chi-square approximations for the distributions of the corresponding (logarithms of) likelihood ratios are 0.152 and 0.069, respectively (see details in Izbicki et al. 2012). Hence, at the level of significance $\alpha = 10\%$, H_0^A is rejected, but H_0^G is not. That is, one concludes that the allelic proportions are not the same in both groups, but cannot reject that the genotypic proportions are the same. This is really bewildering: if the allelic proportions are *not* the same in both groups, the genotypic proportions *cannot* be the same either. Indeed, if the latter were the same, then $\gamma_i = \pi_i$, $\forall i \in \mathbf{G}$, and hence $\theta \in H_0^A$. Thus, the two conclusions drawn from these p-values are contradictory. These reports can truly confuse a practitioner, though the inferential and logical reasonings are not the same. It is then advisable the performance of simultaneous tests that avoid reaching to such confusions.

The same conflicting conclusions result from the calculation of Bayes factors (i.e., $P(\text{data}|H)/P(\text{data}|\Theta)$) taking into account independent uniform prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ and considering a common cutoff, say 1 or 2, for decision-making: the Bayes factor in favor of H_0^G (H_0^A) is 6.63 (0.28).

On the other hand, other Bayesian and Classical test procedures do cohere. For instance, in this setting, where both hypotheses of interest are precise, that is, have prior probabilities equal to zero (and hence Bayesian tests based on usual $0-1-c$ loss functions—see Table 2—cannot be applied satisfactorily due to zero posterior probabilities), one can use the full Bayesian significance tests (FBST) by Pereira and Stern (1999) instead of posterior probabilities. They yield evidences (e-values) of 0.434 and 0.493 for H_0^G and H_0^A, respectively. Note that coherence holds whichever cutoff is adopted. Similarly, the s-values—a classical coherent measure of evidence (Patriota 2013)—for these hypotheses are 0.444 and 0.503, respectively. Again, coherence holds whichever cutoff is adopted.

Example 2 (Based on Schervish 1995). Suppose that according to technical specifications, the failure rate of a given component, θ, must lie in the interval $(1, 2)$. The following nested null hypotheses are thus relevant: $H_0^1 : \theta \leq 1$ and $H_0^2 : \theta \leq 1 \cup \theta \geq 2$. H_0^1 represents that the hazard rate is too small (at most a unit) and H_0^2 that it does not lie within the standard limits. Consider that a single component is tested. Let X be the lifetime of this component and assume that $X|\theta \sim \text{Exponential}(\theta)$, $\theta > 0$.

Under Neyman and Pearson's hypothesis testing, the UMP level α, $\alpha \in (0, 1)$, test for H_0^1 against its alternative consists in rejecting H_0^1 if, and only if, $x < -\log(1-\alpha)$. Moreover, the level α UMP test for H_0^2 consists in rejecting H_0^2 if, and only if, $-\log(\frac{1+\alpha}{2}) < x < -\log(\frac{1-\alpha}{2})$. For $\alpha = 0.05$ and $X = 0.7$, one rejects H_0^2 but accepts H_0^1! Indeed, for all $x \in (-\log(0.525), -\log(0.475))$ incoherence takes place.

Alternatively, one who adheres to Fisherian significance testing may determine p-values for H_0^1 and H_0^2. For the sample $X = 0.7$, the p-values derived from UMP tests (see the Appendix

Table 1. Genotypic sample frequencies.

	AA	AB	BB	Total
Case	55	83	50	188
Control	24	42	39	105

Table 2. The $0-1-c$ loss function for the hypothesis $\theta \in H$.

	State of Nature	
Decision	$\theta \in H$	$\theta \notin H$
0 (do not reject H)	0	1
1 (reject H)	c	0

for the derivation) are 0.5034 and 0.0068 for H_0^1 and H_0^2, respectively. Hence, for several standard levels of significance (e.g., 0.01, 0.05, 0.10), these p-values suggest that H_0^2 should be rejected while H_0^1 should not. That is to say, the datum supports the failure rate is too small, lying outside the standard limits; at the same time, it gives strong evidence that the specifications on it are satisfied. If one instead uses p-values corresponding to GLR tests, one obtains the values of 0.5034 and 0.0061 for H_0^1 and H_0^2, respectively. That is, such p-values do not cohere in Schervish's sense. Although p-values were designed to provide a plausibility measure for a given hypothesis of interest rather than comparing the plausibility among hypotheses, these quantities are widely used in multiple hypothesis testing and for the importance of p-values in such problems some criticism is directed to them (by several authors) for failing to meet the (desirable) coherence property.

Under a Bayesian perspective, considering $\theta \sim \text{Exp}(1)$, the posterior distribution of θ given $X = x$ is Gamma$(2, x + 1)$. For $X = 0.7$, the posterior probabilities of H_0^1 and H_0^2 are 0.5068 and 0.6536, respectively. Considering a common $0 - 1 - c$ loss function (see Table 2) for the null hypotheses, the resulting Bayesian tests will always be coherent, that is, it is not possible to reject H_0^2 if H_0^1 is not rejected (see Section 4 for a review of the general solution of Bayesian tests for such loss function). However, if a decision-maker believes the penalty for false acceptance of H_0^2 to be four times that for false rejection of it (expressed by the $0 - 1 - 1/4$ loss function) and that the errors of Types I and II are of the same degree of importance when testing H_0^1 (represented by the 0-1 loss function, see Table 2 with $c = 1$), the sample $X = 0.7$ will induce incoherent decisions, since $P(H_0^1|X = 0.7) \geq \frac{1}{1+1} = 1/2$ while $P(H_0^2|X = 0.7) < \frac{1}{1+\frac{1}{4}} = 4/5$ (thus one accepts H_0^1 but rejects H_0^2). Hence, coherence ultimately depends on the decision-maker's loss functions choice.

Example 3. Suppose three candidates are running for a majority election. The proportion of electors voting for candidate i is θ_i, $i = 1, 2, 3$, with $\theta_1 + \theta_2 + \theta_3 = 1$. Consider the null hypotheses to be tested are $H_0^1 : \theta \in \Theta_0^{(1)}$ and $H_0^2 : \theta \in \Theta_0^{(2)}$, where $\Theta_0^{(1)} = \cap_{i=1}^3 \{\theta_i \leq 1/2\}$ and $\Theta_0^{(2)} = \{\theta_1 \leq 1/2\}$ (see Figure 2). H_0^1 denotes that none of the candidates gets more than 50% of the valid votes (a relevant hypothesis under some voting systems where it implies the occurrence of a second round in the election) and H_0^2 that candidate 1 has at most 50% of the votes. To test these hypotheses, assume a public opinion poll is conducted with n electors. Let $X = (X_1, X_2, X_3)$, in which X_i is the number of electors in the sample that vote for candidate i.

For $n = 1500$, assuming a uniform prior for $\theta = (\theta_1, \theta_2, \theta_3)$ and $X|\theta \sim \text{Multinomial}(n, \theta)$, the posterior distribution for θ given $X = (784, 711, 5)$ is Dirichlet $(785, 712, 6)$. The posterior probabilities of H_0^1 and H_0^2 are, respectively, 0.0211 and 0.0419. If the same $0 - 1 - c$ loss function is chosen for both tests, the coherence property will be satisfied.

Under the classical approach to hypothesis testing, one may use GLR tests as there is no UMP test for H_0^1. Fossaluza (2008) showed that for $\alpha = 0.053$, the GLR tests for these hypotheses are

$$\begin{cases} \text{reject } H_0^1 & \text{if } \max\{x_1, x_2, x_3\} \geq 788 \\ \text{not reject } H_0^1 & \text{if } \max\{x_1, x_2, x_3\} < 788 \end{cases}$$

and

$$\begin{cases} \text{reject } H_0^2 & \text{if } x_1 \geq 782 \\ \text{not reject } H_0^2 & \text{if } x_1 < 782. \end{cases}$$

Thus, from the observed vector $x' = (784, 711, 5)$, the decision-maker concludes the occurrence of a second round in the majority election (not rejection of H_0^1), as well as the election of candidate 1 in the first round (rejection of H_0^2)!

The inconsistency remains if one uses p-values instead (see the Appendix for the development). For the same data, one gets p-values of 0.084 and 0.042 for H_0^1 and H_0^2, respectively. Hence, at the level of significance 0.05, H_0^2 is rejected but H_0^1 is not. As in the previous examples, p-values are revealed to be incoherent measures of support of hypotheses (Schervish 1996). s-values, on the other hand, lead to coherent conclusions. Indeed, in this case both hypotheses have the same s-value of 0.214.

From these examples, one concludes that both UMP tests and GLR tests with the same level of significance can yield incoherent results. The same happens when performing tests based either on Bayes factors or on p-values. On the other hand, s-values and e-values lead to coherent tests. The relationship between coherent hypothesis testing and measures of support is the object of the next section.

3. Coherent Tests and Measures of Support

For the remainder of this manuscript, Θ represents the parameter space and $\mathcal{X}$ is the sample space. The set of hypotheses to be tested is denoted by $\mathcal{H}$. (To accommodate the Bayesian perspective, we assume all hypotheses in $\mathcal{H}$ are measurable. See Izbicki and Esteves (2015) for additional technical details.)

For each hypothesis $H \in \mathcal{H}$, let $\varphi_H : \mathcal{X} \to \{0, 1\}$ be a test function for the hypothesis $\theta \in H$ (H for short), where $\varphi_H(x) = 1$ ($\varphi_H(x) = 0$) denotes the rejection (acceptance) of H when observing $x \in \mathcal{X}$. Also, let $L_H : \{0, 1\} \times \Theta \to \mathbb{R}_+$ be a hypothesis testing loss function for H, that is, L_H satisfies $L_H(0, \theta) \leq L_H(1, \theta)$, $\theta \in A$, and $L_H(0, \theta) \geq L_H(1, \theta)$, $\theta \in A^c$ (Schervish 1995). (These inequalities characterize the natural premise that wrong decisions ought to be assigned penalties greater than those associated with correct decisions.)

In this setting, the tests $(\varphi_H)_{H \in \mathcal{H}}$ are said to be *coherent* if for all $A, B \in \mathcal{H}$ with $A \subseteq B$, $\varphi_B \leq \varphi_A$ (i.e., $\varphi_B(x) \leq \varphi_A(x)$, for every $x \in \mathcal{X}$). In words, rejection of B (i.e., $\varphi_B(x) = 1$) implies rejection of A (i.e., $\varphi_A(x) = 1$). (Coherence may be rewritten in several forms, such as $\varphi_{A \cup B} \leq \varphi_A \varphi_B$ or $\varphi_{A \cap B} \geq 1 - (1 - \varphi_A)(1 - \varphi_B)$, whenever $A \cup B, A \cap B \in \mathcal{H}$. These and other representations of coherence are detailed in Izbicki (2010).)

In Example 1 of the previous section, it is shown that the two tests based on p-values are not coherent. Likewise, it illustrates that incoherence also happens to tests having Bayes factors as test statistics. (Recall that a test statistic for hypothesis H is a function $T_H : \mathcal{X} \to I$, where I is a subset of $\mathbb{R}$.) On the other hand, the FBST, performed by means of the calculation of e-values, cohere. The same holds for s-values. In Example 2, p-values once again yield incoherent tests, while posterior probabilities lead to coherent decisions considering a common $0 - 1 - c$ loss function for both null hypotheses. These examples suggest the following question: what kinds of test statistics yield coherent simultaneous hypothesis testing? The short answer is that only measures of support that are coherent in

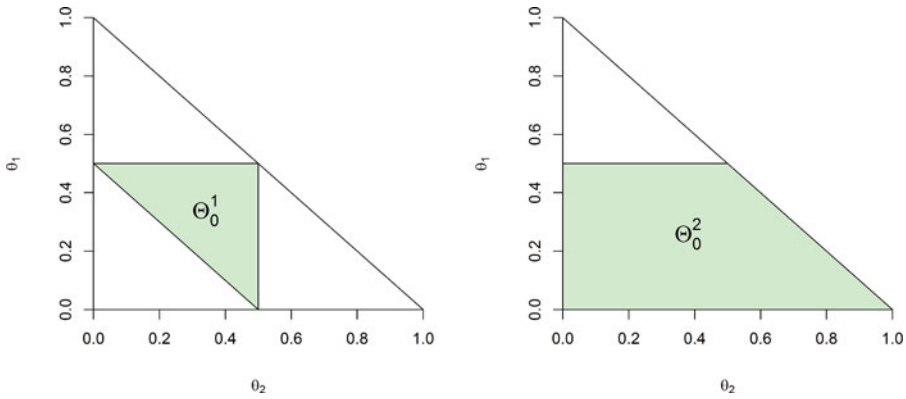

Figure 2. Null hypotheses H_0^1 (left) and H_0^2 (right) for Example 3.

Schervish's sense do. As a matter of fact, a set of tests $(\varphi_H)_{H\in\mathcal{H}}$ is coherent if, and only if, there exist

(i) a set of test statistics $(T_H)_{H\in\mathcal{H}}$, all taking values on $I \subseteq \mathbb{R}$, with $T_A(x) \leq T_B(x)$ for every $x \in \mathcal{X}$ whenever $A \subseteq B$, and

(ii) a common cutoff $\alpha \in I$

such that, for all hypotheses $H \in \mathcal{H}$ and samples $x \in \mathcal{X}$,

$$\varphi_H(x) = 1 \Leftrightarrow T_H(x) < \alpha. \tag{1}$$

In words, H is rejected if, and only if, the test statistic T_H has a small value. The direction ($\Rightarrow$) is immediate: take $I = \{0, 1\}$, $T_H = 1 - \varphi_H$, and $\alpha = 1$. For the ($\Leftarrow$) part, let $A, B \in \mathcal{H}$ such that $A \subseteq B$ and for each $H \in \mathcal{H}$, let φ_H be given by Equation (1) assuming (i) and (ii). Thus,

$$\varphi_A(x) = 0 \;\Rightarrow\; T_A(x) \geq \alpha \;\Rightarrow\; T_B(x) \geq \alpha \;\Rightarrow\; \varphi_B(x) = 0,$$

and coherence holds.

This result (also proved by Gabriel (1969) in a particular case) states that the only decision rules based on test statistics $(T_H)_{H\in\mathcal{H}}$ that yield coherent simultaneous tests are those for which $T_A \leq T_B$ whenever $A \subseteq B$. $T_H(x)$ can be interpreted as the amount of evidence in favor of the hypothesis H brought by the data x (relative to the postulated model). Under this reading, the test statistics $(T_H)_{H\in\mathcal{H}}$ associated with coherent tests may be seen as coherent measures of support in the sense that "the larger the hypothesis is, the more support there is" (Schervish 1996).

Practitioners usually adopt a measure of support when testing statistical hypotheses as it provides more information about the hypotheses of interest than the methodology by Neyman and Pearson does. The equivalence between coherent tests and measures of support we show makes clear that this approach to multiple testing is only successful in cases of coherent measures of support. Thus, p-values and Bayes factors cannot be regarded as appropriate numerical descriptions of the degree to which the data corroborate a statistical hypothesis (see further discussion on this matter in Schervish (1996) and Lavine and Schervish (1999), for instance). On the other hand, well-known measures of support generate coherent tests: posterior probabilities of hypotheses, posterior odds (i.e., $P(H|x)/P(H^c|x)$), e-values by Pereira and Stern (1999), likelihood ratio test statistics (Bickel 2012), and s-values proposed by Patriota (2013) are a few of them. In addition, the relationship between coherent tests and measures of support indicates that to create reasonable coherent tests, one only needs to first choose a coherent measure of support, and then choose α according to some optimality criteria. For instance, one may choose to use the measure of evidence given by the likelihood ratio statistic, $T_H(x) = \frac{\sup_{\theta\in H} L_x(\theta)}{\sup_{\theta\in\Theta} L_x(\theta)}$, and then define a common cutoff $\alpha \in (0, 1)$ to test each hypothesis H of interest. A way of choosing α was described by Bickel (2008). Alternatively, one may choose the posterior probability as the test statistic and derive $\alpha \in (0, 1)$ under a decision-theoretical framework (see the details in Section 4).

The result also points out that checking whether a given measure of support is coherent is an important step toward obtaining coherent tests: if the chosen test statistics $(T_H)_{H\in\mathcal{H}}$ are not coherent, the practitioner has to keep in mind he may obtain incoherent decision patterns when testing the hypotheses of interest.

It should also be mentioned that at least one coherent measure of support can be created from any coherent tests $(\varphi_H)_{H\in\mathcal{H}}$, namely, $T_H = 1 - \varphi_H$, $H \in \mathcal{H}$, which may be seen as the least informative measure of support since it simply indicates the hypotheses supported by the data.

4. Coherent Bayesian Hypothesis Testing

In Example 2 of Section 2, the performance of Bayesian tests with $0-1-1/4$ and $0-1$ loss functions yields incoherent decisions. Intuitively, incoherence happens in that example because the loss of falsely rejecting H_0^1 is four times the loss of falsely rejecting H_0^2, while the corresponding errors of Type II are of the same magnitude. Hence, these loss functions reveal that the decision-maker is more reluctant to reject H_0^1 than to reject H_0^2, in such a way that he only needs little evidence to accept H_0^1 (posterior probability greater than 1/2) when compared with the amount of evidence needed to accept H_0^2 (posterior probability greater than 4/5). Thus, even though H_0^1 implies H_0^2, it is not surprising in this case that coherence may not hold.

On the other hand, many intuitive loss functions yield coherent simultaneous Bayesian tests. In this section, we present some of them. To the best of the authors' knowledge, coherence and other logical properties in multiple Bayesian tests have not been deeply studied yet (except for Schervish (1996), who examined coherence for a few loss functions and connected it to admissibility/Bayesian optimality in certain situations).

First, recall that, under a Bayesian decision-theoretical approach to inference, a Bayes test for the hypothesis H is given by

$$\text{Reject } H \;\Leftrightarrow\; E[L_H(1,\theta)|X=x] \;<\; E[L_H(0,\theta)|X=x],$$

where expectations are evaluated against the posterior distribution for θ given $X = x$. That is, H is to be rejected if and only if the posterior risk of acceptance is greater than that of rejection.

Next, suppose that for each $H \in \mathcal{H}$, a Bayes rule is built from a $0-1-c_H$ loss function (see Table 2). Under these conditions, the Bayes test for H is to reject it if $P(\theta \in H|X=x) < \frac{1}{1+c_H}$ (Schervish 1995).

If for any pair of nested hypotheses $A, B \in \mathcal{H}$, with $A \subseteq B$, we have that $c_A \leq c_B$, then the corresponding Bayes tests will cohere. Indeed,

$$\begin{aligned}\text{Not Reject } A \Rightarrow P(\theta \in A|X=x) \;&\geq\; \frac{1}{1+c_A}\\ \Rightarrow\; P(\theta \in B|X=x) &\geq P(\theta \in A|X=x)\\ \geq\; \frac{1}{1+c_A} &\geq \frac{1}{1+c_B} \Rightarrow \text{Not Reject } B.\end{aligned}$$

Hence, any constants $(c_H)_{H\in\mathcal{H}}$, $c_H \in \mathbb{R}_+$, such that $c_A \leq c_B$ when $A \subseteq B$, generate coherent simultaneous Bayesian tests, whatever the (proper) prior distribution is.

As an example, if Θ is finite, one can define c_H as the number of elements of H. The resultant Bayes tests will be coherent. Analogously, if $\Theta \subset \mathbb{R}$ is bounded and c_H is the length of H, coherence will take place for any prior distribution for θ too. More generally, the choice of any nonnegative monotone set function u defined on $\mathcal{H}$ (i.e., $A, B \in \mathcal{H}$, $A \subseteq B \Rightarrow u(A) \leq u(B)$) for $(c_H)_{H\in\mathcal{H}}$ leads to coherent tests.

More elaborate losses that express the level of inaccuracy of a decision, such as nondecreasing functions of distances between the parameter and the chosen hypothesis, also produce coherent tests. For example, if $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}^*$, consider

$$L_H(0,\theta) = g(d(\theta, H)) \;\text{ and }\; L_H(1,\theta) = g(d(\theta, H^c)), \qquad (2)$$

where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing function and $d(\theta, H)$ denotes the distance between the point $\theta \in \Theta$ and the set $H \subseteq \Theta$ (see, e.g., Kolmogorov and Fomin 1975). That is, the greater the distance between the parameter and the wrong decision, the heavier the penalty determined by L_H. The fact that tests based on loss functions like Equation (2) are coherent can be seen by noting that if one rejects B, then $E[L_B(1,\theta)|X=x] < E[L_B(0,\theta)|X=x]$. As

$$L_B(1,\theta) = g(d(\theta, B^c)) \geq g(d(\theta, A^c)) = L_A(1,\theta)$$

and

$$L_B(0,\theta) = g(d(\theta, B)) \leq g(d(\theta, A)) = L_A(0,\theta),$$

it follows that

$$E[L_A(1,\theta)|X=x] \;<\; E[L_A(0,\theta)|X=x],$$

and therefore one rejects A. Coherence then holds for every (proper) prior distribution over Θ. As an illustration, we return to Example 2, where $\Theta = \mathbb{R}_+$. If one uses the loss of Equation (2) with $g(z) = z$ and $d(\theta, H) = \inf_{\theta_0 \in H} |\theta - \theta_0|$ in Example 2, one obtains using standard calculus or Monte Carlo integration

$$E[L_{H_0^1}(0,\theta)|X=0.7] = 0.39,\;\; E[L_{H_0^1}(1,\theta)|X=0.7] = 0.22$$

and

$$E[L_{H_0^2}(0,\theta)|X=0.7] = 0.08,\; E[L_{H_0^2}(1,\theta)|X=0.7] = 0.33.$$

Thus, we reject H_0^1, but we do not reject H_0^2. As expected, there is no incoherence in these results.

More general necessary and sufficient conditions (detailed in Silva 2014) can be imposed on loss functions to ensure coherent tests against all (proper) prior distributions. Thus, a broad class of appealing loss functions can be used by Bayesians that consider coherence to be an important feature for simultaneous hypothesis testing.

5. Conclusions

The coherence property in simultaneous hypothesis testing is reviewed through new examples, which enlighten the concept and motivate the discussion about which usual test procedures meet such logical requirement. It is shown that coherent tests are unavoidably associated with test statistics that are coherent measures of support in Schervish's sense, such as posterior probabilities, posterior odds, *e*-values, likelihood ratio test statistics, and *s*-values. On the other hand, Bayes factors and *p*-values cannot be regarded as measures of support. The connection between coherent tests and measures of support suggests a way of constructing simultaneous tests that yield conclusions that are easier to communicate to nonstatisticians: the recommended routine is to first evaluate coherent measures of support for the relevant hypotheses and then to reject those with measures smaller than a common fixed threshold. Finally, loss functions that induce coherent tests under a Bayesian decision-theoretical framework are also explored. It is stressed that simple and intuitive loss functions enable a practitioner to perform coherent Bayes tests. Several examples are given herein. These findings contribute to the discussion on and the improvement of inferences made from simultaneous hypothesis testing.

Appendix: Proofs

A.1 Example 2

The fact that $\varphi_\alpha^{(1)} : \mathcal{X} \to \{0, 1\}$ given by

$$\varphi_\alpha^{(1)}(x) = \begin{cases} 1, & \text{if } x < -\log(1-\alpha) \\ 0, & \text{if } x \geq -\log(1-\alpha) \end{cases},$$

is the UMP level α, $\alpha \in (0, 1)$, test for H_0^1 follows from Karlin and Rubin's theorem.

Moreover, from theorem 4.82, p. 249 of Schervish (1995), it follows that the level α UMP test for H_0^2, $\varphi_\alpha^{(2)}$, is given by

$$\varphi_\alpha^{(2)}(x) = \begin{cases} 1, & \text{if } -\log\left(\dfrac{1+\alpha}{2}\right) < x < -\log\left(\dfrac{1-\alpha}{2}\right) \\ 0, & \text{otherwise} \end{cases}.$$

The *p*-value of the observation x relative to the set of UMP tests $\{\varphi_\alpha^{(i)} : \alpha \in (0, 1)\}$ for hypothesis H_0^i, $p_{H_0^i}^{(\text{UMP})}(x)$, is (Schervish 1995)

the greatest lower bound of the set of all significance levels for which the corresponding UMP tests lead to rejection of H_0^i when x is observed, $i = 1, 2$, that is,

$$p_{H_0^i}^{(\mathrm{UMP})}(x) = \inf\left\{\alpha \in (0,1) : \varphi_\alpha^{(i)}(x) = 1\right\}.$$

For the hypothesis H_0^2,

$$\alpha \in \left\{\alpha' \in (0,1) : \varphi_{\alpha'}^{(2)}(x) = 1\right\} \Leftrightarrow \varphi_\alpha^{(2)}(x) = 1$$
$$\Leftrightarrow -\log\left(\frac{1+\alpha}{2}\right) < x < -\log\left(\frac{1-\alpha}{2}\right) \Leftrightarrow \alpha > |2e^{-x} - 1|.$$

Thus, $p_{H_0^2}^{(\mathrm{UMP})}(x) = \inf\{\alpha \in (0,1) : \varphi_\alpha^{(2)}(x) = 1\} = \inf(|2e^{-x} - 1|, 1) = |2e^{-x} - 1|$.

Following the steps above, the reader obtains $p_{H_0^1}^{(\mathrm{UMP})}(x) = 1 - e^{-x}$.

A.2 Example 3

Denote the maximum of $\{x_1, x_2, x_3\}$ by $x_{(3)}$. For $j \in \mathbb{N}$, let $\varphi_j^{(1)}$ be the test function that rejects H_0^1 if and only if $x_{(3)} \geq j$, and $\varphi_j^{(2)}$ be the one that rejects H_0^2 if $x_1 \geq j$. It can be proved (see Fossaluza 2008) that $\varphi_j^{(1)}$ and $\varphi_j^{(2)}$ are GLR tests for H_0^1 and H_0^2, respectively. Also, let $\alpha_j^{(i)}$ be the size of test $\varphi_j^{(i)}$, $i = 1, 2$ and $j \in \mathbb{N}$. According to Schervish's definition of p-value, $p_{H_0^i}^{(\mathrm{GLR})}(x)$ is given by

$$p_{H_0^i}^{(\mathrm{GLR})}(x) = \inf\left\{\alpha_j^{(i)} \in [0,1] : \varphi_j^{(i)}(x) = 1\right\}, i = 1, 2.$$

As $\varphi_j^{(1)}(x_1, x_2, x_3) = 1 \Leftrightarrow x_{(3)} \geq j$, we have, for $j \leq x_{(3)}$ and all θ, that

$$P(\max\{X_1, X_2, X_3\} \geq j|\theta) \geq P(\max\{X_1, X_2, X_3\} \geq x_{(3)}|\theta),$$

which implies that $\alpha_j^{(1)} \geq \alpha_{x_{(3)}}^{(1)}$. Thus,

$$p_{H_0^1}^{(\mathrm{GLR})}(x) = \inf\left\{\alpha_0^{(1)}, \ldots, \alpha_{x_{(3)}}^{(1)}\right\} = \alpha_{x_{(3)}}^{(1)}$$
$$= \sup_{\theta \in \Theta_0^{(1)}} P(\max\{X_1, X_2, X_3\} \geq x_{(3)}|\theta).$$

Shifeng and Guoying (2005) proved that $P(\max\{X_1, X_2, X_3\} \geq x_{(3)}|\theta)$ attains its maximum over $\Theta_0^{(1)}$ at the points $(1/2, 1/2, 0)$, $(1/2, 0, 1/2)$, and $(0, 1/2, 1/2)$, which yields for $x_{(3)} > \frac{n}{2}$

$$p_{H_0^1}^{(\mathrm{GLR})}(x) = 2\sum_{i=x_{(3)}}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n.$$

Proceeding in the same way, the p-value for H_0^2 is

$$p_{H_0^2}^{(\mathrm{GLR})}(x) = \sum_{i=x_1}^{n}\binom{n}{i}\left(\frac{1}{2}\right)^n.$$

Acknowledgments

The authors are thankful for Carlos Alberto de Bragança Pereira, Fábio Gagliardi Cozman, Márcio Alves Diniz, Julio Michael Stern, Paulo Cilas Marques Filho, Rafael Bassi Stern, Sergio Wechsler, and Verónica Andréa González-López for fruitful discussions and important comments and suggestions that improved the manuscript.

Funding

This work was partially supported by *Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico* [131982/2009-5, 200959/2010-7] and *Fundação de Amparo à Pesquisa do Estado de São Paulo* [2009/03385-5, 2014/25302-2].

References

Bickel, D. R. (2008), "The Strength of Statistical Evidence for Composite Hypotheses with an Application to Multiple Comparisons," *COBRA Preprint Series*, 22, paper no. 49. [245]

——— (2012), "The Strength of Statistical Evidence for Composite Hypotheses: Inference to the Best Explanation," *Statistica Sinica*, 1147–1198. [242,245]

——— (2014), "Self-Consistent Confidence Sets and Tests of Composite Hypotheses Applicable to Restricted Parameters," Technical Report of the University of Ottawa, Ottawa, Canada. [242]

da Silva, G. M., Esteves, L. G., Fossaluza, V., Izbicki, R., and Wechsler, S. (2015), "A Bayesian Decision-Theoretic Approach to Logically-Consistent Hypothesis Testing," *Entropy*, 17, 6534–6559. [242]

Esteves, L. G., Izbicki, R., Stern, J. M., and Stern, R. B. (2016), "The Logical Consistency of Simultaneous Agnostic Hypothesis Tests," *Entropy*, 18, paper no. 256. [242]

Farcomeni, A. (2008), "A Review of Modern Multiple Hypothesis Testing, with Particular Attention to the False Discovery Proportion," *Statistical Methods in Medical Research*, 17, 347–388. [242]

Fossaluza, V. (2008), "Hypothesis Testing in Majoritarian Elections (in Portuguese)," Master's thesis, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil. [244,247]

Gabriel, K. R. (1969), "Simultaneous Test Procedures - Some Theory of Multiple Comparisons," *The Annals of Mathematical Statistics*, 40, 224–250. Available at *http://www.jstor.org/stable/2239213* [245]

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: Wiley. [242]

Hommel, G., and Bretz, F. (2008), "Aesthetics and Power Considerations in Multiple Testing–A Contradiction?" *Biometrical Journal*, 50, 657–666. [242]

Izbicki, R. (2010), "Classes of Hypothesis Tests (in Portuguese)," Master's thesis, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil. [244]

Izbicki, R., and Esteves, L. G. (2015), "Logical Consistency in Simultaneous Statistical Test Procedures," *Logic Journal of IGPL*, 23, 732–758. [242,244]

Izbicki, R., Fossaluza, V., Hounie, A. G., Nakano, E. Y., and Pereira, C. A. B. (2012), "Testing Allele Homogeneity: The Problem of Nested Hypotheses," *BMC Genetics*, 13, paper no. 103. doi:10.1186/1471-2156-13-103. [243]

Kolmogorov, A. N., and Fomin, S. V. (1975), *Introductory Real Analysis*, Englewood Cliffs, NJ: Dover. [246]

Lavine, M., and Schervish, M. J. (1999), "Bayes Factors: What they are and what they are not," *The American Statistician*, 53, 119–122. [242,245]

Lehmann, E. L. (1957), "A Theory of Some Multiple Decision Problems, i," *The Annals of Mathematical Statistics*, 28, 1–25. [242]

Lin, S. K., Chen, C. K., Ball, D., Liu, H. C., and Loh, E. W. (2003), "Gender-Specific Contribution of the Gabaa Subunit Genes on 5q33 in Methamphetamine use Disorder," *The Pharmacogenomics Journal*, 3, 349–355. [243]

Patriota, A. G. (2013), "A Classical Measure of Evidence for General null Hypotheses," *Fuzzy Sets and Systems*, 233, 74–88. [242,243,245]

Pereira, C. A. B., and Stern, J. M. (1999), "Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses," *Entropy*, 1, 99–110. [243,245]

Schervish, M. J. (1995), *Theory of Statistics*, New York: Springer. [243,244,246]

——— (1996), "P Values: What They are and What They are Not," *The American Statistician*, 50, 203–206. [242,244,245]

Shaffer, J. P. (1995), "Multiple Hypothesis Testing," *Annual Review of Psychology*, 46, 561–584. [242]

Shifeng, X., and Guoying, L. (2005), "Testing for the Maximum Cell Probabilities in Multinomial Distributions," *Science in China Series A: Mathematics*, 48, 972–985. [247]

Silva, G. M. (2014), "Logical Properties of Classes of Hypothesis Tests (in Portuguese)," Ph.D. dissertation, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil. [246]

Sonnemann, E. (2008), "General Solutions to Multiple Testing Problems," *Biometrical Journal*, 50, 641–656. [242]