RT-MAE 2014-03

LOGICAL CONSISTENCY IN SIMULTANEOUS TEST PROCEDURES

by

Rafael Izbicki
and
Luís Gustavo Esteves

Palavras-Chaves: Simultaneous Test Procedures, Monotonicity, Logical Coherence, Coherence Principle, Consonance Principle, Classes of Hypotheses Tests.

AMS Classification: 62A01; 62F03.

- Fevereiro de 2014 -

Logical Consistency in Simultaneous Test Procedures

Rafael Izbicki · Luís Gustavo Esteves

Abstract Many have argued that, when performing simultaneous test procedures, one should seek for solutions that are easier to communicate to non-statisticians. In particular, logical incoherences should be avoided when reporting the results of such tests: for example, if hypothesis A implies hypothesis B, the rejection of B should imply the rejection of A, a property not always met by multiple test procedures. In this paper we contribute to this discussion by exploiting how far one can go in requiring a test procedure to be (logically) coherent and still preserve statistical optimality. This is done by studying four types of logical consistency relations. We show that although the only procedures that satisfy more than (any) two of these properties are simple tests based on point estimation, it is possible to construct various interesting methods that fulfill one or two of them while preserving different statistical optimality criteria. This is illustrated with several Bayesian and frequentist examples.

Keywords Simultaneous Test Procedures · Monotonicity · Logical Coherence · Coherence Principle · Consonance Principle · Classes of Hypotheses Tests

AMS Classification: 62A01; 62F03.

1 Introduction

In many scientific problems, one is interested in testing several hypotheses simultaneously. Such a situation is called a *multiple* (or *simultaneous*) *hypotheses testing* problem (Shaffer (1995)). This is typical, for

This work was partially funded by *Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico* (grants 131982/2009-5 and 200959/2010-7) and *Fundação de Amparo à Pesquisa do Estado de São Paulo* (grant 2009/03385-5).

Rafael Izbicki

Statistics Department, Carnegie Mellon University, Pittsburgh, PA, USA

E-mail: rizbicki@stat.cmu.edu

Luís Gustavo Esteves

Instituto de Matemática e Estatística, Universidade de São Paulo, SP, Brasil

E-mail: lesteves@ime.usp.br

example, in clinical trials where one is interested in comparing the effectiveness of drugs and their side effects, or in genetic experiments involving microarrays. See more examples in Hochberg and Tamhane (1987).

There exist many methods that aim at creating optimal statistical tests for simultaneous procedures. From a frequentist perspective, rather than only controlling the level of significance α of each test individually, other criteria have been introduced. Among them, popular approaches are controlling the error rate per family (PFE), the family wise error rate (FWE) and the false discovery rate (FDR) (Shaffer (1995); Finner and Gontscharuk (2009)). Other approaches suggest that rather than control these (or related) quantities, one should estimate them (Pawitan et al (2006)). The reader is referred to Hochberg and Tamhane (1987), Shaffer (1995) and Farcomeni (2008) for a review on traditional methods for simultaneous tests.

Of particular interest are the so called closure methods (Marcus et al (1976); Sonnemann (1982, 2008)). Assume one is interested in testing a given set of hypotheses $\mathcal{A}$[1]. For each hypothesis, assign an α-level test, $\alpha \in (0, 1)$. The closure method for testing each of these hypotheses consists in rejecting hypothesis $H \in \mathcal{A}$ if, and only if,

1. H is rejected according to the α-level test.
2. All hypotheses in $\mathcal{A}$ that imply it (i.e. $\forall H' \subseteq H$, where $H' \in \mathcal{A}$) are rejected according to their respective α level tests.

Besides controlling the FWE (Sonnemann (1982)), this method has the advantage of satisfying the *coherence* property: if hypothesis H_0^1 implies hypothesis H_0^2 (i.e., $H_0^1 \subseteq H_0^2$) and H_0^2 is rejected, H_0^1 will also be rejected (Gabriel (1969)).

Although coherence is desirable, not all simultaneous procedures satisfy it. For instance, this is shown by Hommel and Bretz (2008) in a regression setting. Considering the linear model $E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$, they show that, for some samples, the Bonferroni-Holm procedure leads one to reject $\beta_2 = 0$, but not to reject $\beta_1 = \beta_2 = 0$. As $\beta_1 = \beta_2 = 0$ implies $\beta_2 = 0$, we have a logically incoherent procedure (in the sense of Gabriel (1969)). See also Raviv (2013) for an interesting example where one rejects the equality of means $\mu_1 = \mu_2$, but does not reject $\mu_1 = \mu_2 = 0$ in an ANOVA setting.

Although "not rejecting $\beta_1 = \beta_2 = 0$" is usually not understood as a definitive assertion "$\beta_1 = \beta_2 = 0$", reporting the results of an incoherent procedure is usually hard, and many times embarrassing (Templeton (2010); Zhao et al (2010); Romano et al (2011)). For instance, Schervish (1996) describes an example in the case E.E.O.C. vs. Federal Reserve Bank of Richmond (Russell (1983)) *"In this lively exchange the plaintiff's statistical experts tries to explain to a judge why one should use a one-sided test*

[1] Given a subset A of the parameter space Θ, we use "testing the hypothesis A" as a shorthand for the problem of testing $H_0 : \theta \in A$ versus $H_1 : \theta \notin A$.

(with P value 0.037 in this example) rather than a two-sided test (with P value 0.074). The significance of the choice of the hypothesis was quite apparent to the judge." The problem in this example is that while the two-sided hypothesis ($\mu = 0$) is not rejected at the level 5%, the one-sided ($\mu \leq 0$) is. However, $\mu = 0$ implies $\mu \leq 0$.

Some authors therefore argue that some times one should waive on only maximizing standard efficiency criteria (such as power of the tests) so as to produce logically coherent results that are easier to communicate to non-statisticians: *"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 non-statisticians."* (Hommel and Bretz (2008)).

Coherence is not the only logical relationship one might expect from simultaneous hypotheses tests (to avoid confusions, from here on we call this property *monotonicity* instead, and reserve the use of the term "coherent" for its meaning in Standard Logic, that is, the overall logical consistency among the conclusions from the hypotheses tests). Recently, much emphasis has been given to a different logical property named *consonance*, also introduced by Gabriel (1969). Informally, such a property states that when one rejects the intersection of several hypotheses, at least one of them should be rejected marginally (Sonnemann (1982, 2008); Rosenblum (2012)). Many closure methods that respect this property have been developed recently (Zhao et al (2010); Romano et al (2011)). Such procedures usually have smaller computational costs than traditional closure tests (Brannath and Bretz (2010)), and hence provide useful shortcuts in practice. Monotonic and consonant procedures also have better power properties than those that do not respect these consistency criteria (Sonnemann and Finner (1988); Romano et al (2011)). Finally, Lehmann (1957a) defined a different logical property which he named *compatibility* and will be revisited later in the paper. Several other coherence relationships can also be defined.

The main goals of this work are:

1. to formalize and characterize four of these properties,
2. to investigate simultaneous test procedures that satisfy them,
3. to examine how restrictive these properties are when put together.

For these purposes, in section 2 we introduce the concept of a *class of hypotheses tests*, which from now on we call *CHT*, a mathematical device that associates one test function to each hypothesis of interest. We also illustrate such concept with CHT's that will be used later on the paper. In section 3 we formalize four consistency relations one could desire from CHT's. They are *monotonicity*, *union consonance*, *intersection consonance* and *invertibility*. Next, we study some of their properties and consequences. We also investigate whether some common statistical procedures satisfy them. Finally, in Section 4, we study

how restrictive these four requirements are when put together. In particular, we compare them with compatible classes. Conclusions are presented in Section 5. In the Supplementary Material, we present the proofs of most of the results presented in this work. We omit trivial demonstrations.

2 Classes of hypotheses tests

We start by formally describing a class of hypotheses tests (CHT), a mathematical object that formalizes the notion that for each hypothesis of interest one assigns a hypothesis test (a test function). This raises the question of which are the hypotheses of interest for a given problem. This is problem dependent. However, as stated by Hochberg and Tamhane (1987), *"In some types of exploratory research it may be impossible to specify in advance the family of all potential inferences that may be of interest"*. In this work, we assume one has to assign a hypothesis test to each element of a given σ-field of the parameter space. This allows one to assign a test to each of the possible hypotheses that exist (by taking the σ-field to be the power set of the parameter space Θ), and also accommodates Bayesian procedures based on posterior probabilities, in which it is only possible to assign probabilities to some σ-fields of Θ[2]. Recall that a test function is a measurable function from the sample space $\mathcal{X}$ to $\{0,1\}$, where 1 represents the decision of rejecting the null hypothesis and 0 represents the decision of not rejecting it.

Remark: While some argue the decision 0 should be interpreted as the definitive action "accept the hypothesis", others believe it is more appropriate to understand it as "not reject the hypothesis", suggesting a more cautious posture over decision-making (see, e.g., discussion in Mayo and Spanos (2006)). Such a distinction plays an important role in this paper: the coherence properties we define can be more or less appealing depending on which of the above positions is adopted by a practitioner. The reader should keep this in mind when judging how reasonable each of these properties is. We return to this point later in the paper.

Definition 1 *(Class of hypotheses tests (CHT)) Let $\sigma(\Theta)$ be a σ-field of Θ, $\sigma(\mathcal{X})$ be a σ-field of $\mathcal{X}$ and $\Psi = \{\phi : \mathcal{X} \to \{0,1\} : \phi$ is $\sigma(\mathcal{X})$-measurable$\}$ be the set of all test functions. A CHT is a function $\mathcal{L} : \sigma(\Theta) \to \Psi$ that, for each hypothesis $A \in \sigma(\Theta)$, associates the test $\mathcal{L}(A) \in \Psi$ for testing hypothesis A.*

Hence, for hypothesis $A \in \sigma(\Theta)$ and data $x \in \mathcal{X}$, $\mathcal{L}(A)(x) = 0$ represents the decision of not rejecting A, and $\mathcal{L}(A)(x) = 1$ of rejecting it. Examples 1 and 2 illustrate this concept by using classes induced by two traditional statistical tests. We denote the likelihood function at $\theta \in \Theta$ generated by the sample point $x \in \mathcal{X}$ by $L_x(\theta)$, which we assume to be always defined.

[2] Also, in the case where $\theta = (\theta_0, \theta_1)$, $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, where θ_1 are nuisance parameters (Casella and Berger (2002)), one can consider a σ-field of the form $\sigma(\Theta) = \sigma(\Theta_0) \times \Theta_1 = \{A \times \Theta_1 : A \in \sigma(\Theta_0)\}$. Hence, one can assign tests only to parameters of interest.

Example 1 *(Class of likelihood ratio tests of level α) Let $\Theta = \Re^d$ and $\sigma(\Theta) = \mathcal{P}(\Theta)$ be the power set of Θ. For each hypothesis $A \in \sigma(\Theta)$, let $\mathcal{L}(A) : \mathcal{X} \to \{0,1\}$ be defined by*

$$\mathcal{L}(A)(x) = \mathbf{I}\left(\frac{\sup_{\theta \in A} L_x(\theta)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c_A\right),$$

where $\mathbf{I}(B)$ is the indicator function that B holds and $c_A \in [0,1]$ is chosen so that each test has the same level $\alpha \in (0,1)$ previously fixed. This is the class that associates a likelihood ratio test of size α to each hypothesis $A \in \mathcal{P}(\Theta)$.

□

Example 2 *(Tests based on posterior probabilities) Assume the same setup as Example 1, but now with $\sigma(\Theta) = \mathcal{B}(\Theta)$, the Borelians of $\Re^d$. Assume that a prior probability measure $\mathbf{P}$ in $\sigma(\Theta)$ is fixed. For each $A \in \sigma(\Theta)$, let $\mathcal{L}(A) : \mathcal{X} \to \{0,1\}$ be defined by*

$$\mathcal{L}(A)(x) = \mathbf{I}\left(\mathbf{P}(A|x) < \frac{1}{2}\right),$$

where $\mathbf{P}(.|x)$ is the posterior distribution of θ, given x. This is the class that associates to each hypothesis $A \in \mathcal{B}(\Re^d)$, the test that rejects it when its posterior probability is smaller than 1/2.

□

From a Bayesian Decision-Theoretic perspective, a hypothesis test is derived, for each sample point, by minimizing the posterior expectation of a loss function[3] with respect to the posterior distribution of the parameters after observing the data DeGroot (1970). Hence, for a given probability measure for θ and for each $A \in \sigma(\Theta)$ and a specified loss function $L_A : \{0,1\} \times \Theta \to \Re$, one can derive a Bayes test for each of the hypotheses $A \in \sigma(\Theta)$. This procedure is formalized by the following definition:

Definition 2 *(CHT generated by a family of loss functions) Let $(\mathcal{X} \times \Theta, \sigma(\mathcal{X} \times \Theta), \mathbf{P})$ be a Bayesian statistical model. Let $(L_A)_{A \in \sigma(\Theta)}$ be a family of loss functions, where $L_A : \{0,1\} \times \Theta \to \Re$ is the loss function to test $A \in \sigma(\Theta)$. A CHT generated by the family of loss functions $(L_A)_{A \in \sigma(\Theta)}$ is any CHT $\mathcal{L}$ defined over the elements of $\sigma(\Theta)$ such that $\mathcal{L}(A)$ is a Bayes test for hypothesis A against $\mathbf{P}, \forall A \in \sigma(\Theta)$.*

A single family of loss functions can generate multiple CHT's. Example 3 illustrates this.

[3] A loss function for a test is a function $L : \{0,1\} \times \Theta \to \Re$ that associates for each $\theta \in \Theta$ the loss $L(d,\theta)$ for making the decision $d \in \{0,1\}$ of rejecting or not the null hypothesis. The Bayes test is given, for each $x \in \mathcal{X}$, by $\arg\min_{d \in \{0,1\}} \mathbf{E}[L(d,\theta)|X = x]$.

Example 3 *(Tests based on posterior probabilities) Assume the same scenario as Example 2 and that* $(L_A)_{A\in\sigma(\Theta)}$ *is a family of loss functions such that* $\forall A \in \sigma(\Theta)$ *and* $\forall\theta \in \Theta$,

$$L_A(0,\theta) = \mathbf{I}(\theta \notin A) \text{ and } L_A(1,\theta) = \mathbf{I}(\theta \in A),$$

that is, L_A *is the 0-1 loss for* A*. The class* $\mathcal{L}$ *defined in Example 2 is a CHT generated by this family of loss functions, as is the class* $\mathcal{L}'$ *defined by*

$$\mathcal{L}'(A)(x) = \mathbf{I}\left(\mathbf{P}(A|x) \leq \frac{1}{2}\right),\ \forall A \in \sigma(\Theta) \text{ and } \forall x \in \mathcal{X}.$$

□

Example 4 shows a class of Bayesian tests that is motivated by different epistemological considerations (see Stern (2011), but also see Madruga et al (2001) for a decision-theoretic motivation), the *Full Bayesian Significance Tests*, FBST, (Pereira and Stern (1999)). See also Patriota (2013) for a frequentist version of this test.

Example 4 *(Class of tests FBST) Let* $\Theta = \Re^d$, $\sigma(\Theta) = \mathcal{B}(\Re^d)$, *and* $f(\theta)$ *be the prior probability density function (p.d.f.) for* θ*. Suppose that, for each* $x \in \mathcal{X}$*, there exists* $f(\theta|x)$*, the p.d.f. of the posterior distribution of* θ*, given* x*. For each hypothesis* $A \in \sigma(\Theta)$*, let*

$$T_x^A = \{\theta \in \Theta : f(\theta|x) > \sup_{\theta\in A} f(\theta|x)\}$$

be the set tangent to the null hypothesis and let $ev_x(A) = 1 - P(\theta \in T_x^A|x)$ *be the Pereira-Stern evidence value for* A*. See Pereira and Stern (1999) for a geometric motivation. One can define a CHT* $\mathcal{L}$ *by*

$$\mathcal{L}(A)(x) = \mathbf{I}\left(ev_x(A) \leq c\right),\ \forall A \in \sigma(\Theta) \text{ and } \forall x \in \mathcal{X},$$

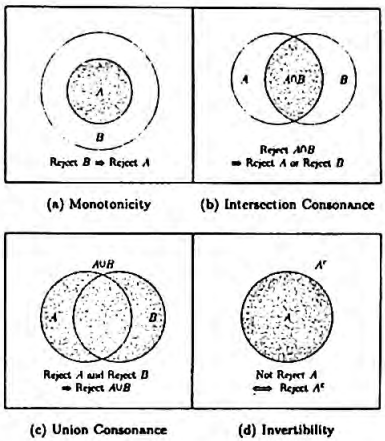
in which $c \in [0,1]$ *is fixed. In words, one does not reject the null hypothesis when its evidence is larger than* c*.*

□

We end this section by defining a CHT generated by a point estimation procedure, a concept that plays an important role when characterizing logically coherent procedures in Section 4.

Definition 3 *(CHT generated by a point estimation procedure) Let* $\widehat{\theta} : \mathcal{X} \longrightarrow \Theta$ *be a point estimator. The CHT generated by* $\widehat{\theta}$ *is defined by* $\mathcal{L}(A)(x) = \mathbf{I}(\widehat{\theta}(x) \notin A)$.

Hence, we reject hypothesis A after observing x if, and only if, the point estimate for θ, $\widehat{\theta}(x)$, is not in A.



(a) Monotonicity (b) Intersection Consonance

(c) Union Consonance (d) Invertibility

Fig. 1: Logical properties one might expect from hypotheses tests.

3 Consistency properties

In this section, we study four properties that one might expect from CHT's to induce logically coherent tests. Each formal definition in the sequence is preceded by an example for motivation. Figure 1 presents a visualization of the properties studied.

3.1 Monotonicity

The first property we describe is related to nested hypotheses. It states that if hypothesis A *implies* hypothesis B (i.e., $A \subseteq B$), the rejection of B must imply the rejection of A (equivalently, not rejecting A must imply not rejecting B).

Example 5 *Suppose 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.*

Table 1: Genotypic sample frequencies

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

Here, the set of all possible genotypes is $\{AA, AB, BB\}$. *Let* $\gamma = (\gamma_{AA}, \gamma_{AB}, \gamma_{BB})$, *where* γ_i *is the probability that an individual from the case group has genotype* i. *Similarly, let* $\pi = (\pi_{AA}, \pi_{AB}, \pi_{BB})$, *where* π_i *is the probability that an individual of control group has genotype* i. *The parameter space, sample space and likelihood function generated by the data are straightforwardly specified.*

In this context, two hypotheses are of interest: the hypothesis that the genotypic proportions are the same in both groups, $H_0^G : \gamma = \pi$, *and the hypothesis that the allelic proportions are the same in both groups* $H_0^A : \gamma_{AA} + \frac{1}{2}\gamma_{AB} = \pi_{AA} + \frac{1}{2}\pi_{AB}$. *The p-values obtained using chi-square tests for these hypotheses are, respectively, 0.152 and 0.069. 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 genotypic proportions are the same in both groups, but that the allelic proportions are not. This is absurd! If genotypic proportions are the same in both groups, allelic proportions must also be the same. Mathematically, if* $\theta \in H_0^G$, *then* $\gamma_i = \pi_i$, $\forall i \in \mathbf{G}$ *and hence* $\theta \in H_0^A$. *This example is further discussed in Izbicki et al (2012).*

□

This example motivates the following definition, first introduced by Gabriel (1969):

Definition 4 *(Monotonicity) A class of hypotheses tests* $\mathcal{L}$ *is monotonic (that is, satisfies monotonicity) if*

$$\forall A, B \in \sigma(\Theta),\ A \subseteq B \Rightarrow \mathcal{L}(A) \geq \mathcal{L}(B)^4.$$

In words, if after observing x, *a hypothesis is rejected, any hypothesis that implies it has to be rejected as well.*

Remark: it is straightforward to show that a class $\mathcal{L}$ is monotonic if, and only if, for all set of indices I,

$$\forall \{A_i\}_{i \in I} \subseteq \sigma(\Theta) \text{ such that } \cup_{i \in I} A_i \in \sigma(\Theta),\ \mathcal{L}(\cup_{i \in I} A_i) \leq \min\{\mathcal{L}(A_i)\}_{i \in I}.$$

Theorem 1 shows that monotonic classes have the advantage of controlling the FWE. Its proof is omitted as different versions of it were already provided in several works (e.g., Hochberg and Tamhane (1987) and Sonnemann (1982, 2008)).

Theorem 1 *Let* $\mathcal{L}$ *be a monotonic CHT and assume that* $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, *that is, the simple hypotheses are in* $\sigma(\Theta)$. *Then,*

$$FWE := \sup_{\theta \in \Theta} \mathbf{P}(\textit{Reject at least one correct } A \in \sigma(\Theta) | \theta) = \sup_{\theta \in \Theta} \mathbf{P}(\mathcal{L}(\{\theta\})(X) = 1 | \theta).$$

In particular, if each of the tests for the simple hypotheses is of size α, $FWE \leq \alpha$.

[4] i.e., $\forall x \in \mathcal{X}$, $\mathcal{L}(A)(x) \geq \mathcal{L}(B)(x)$.

Hence, if a monotonic class assigns a size α test for each simple hypothesis, we must have $FWE \leq \alpha$, and, in particular, each hypothesis test (for a simple or composite one) will also have size α.

As we saw in the introduction, closure procedures are monotonic. Moreover, Sonnemann (1982, 2008) show that any monotonic procedure can be constructed using the closure method. Sonnemann and Finner (1988) also showed that any non-monotonic procedure can be replaced by a monotonic one which is better in the sense that it has the same FWE as the original procedure, and rejects not only the hypotheses rejected by the first, but also potentially more of them.

Example 5 showed that p-values can yield non-monotonic classes. The use of Bayes Factors can also result in inconsistent conclusions (Lavine and Schervish (1999)). In fact, in Example 5, the Bayes Factor in favor of H_0^A is 0.28, while the Bayes Factor in favor of H_0^C is 6.63 (using independent uniform priors over the simplexes). Hence, inconsistency remains. Likelihood ratio tests with a fixed level α (Example 1) are also not monotonic (Izbicki et al (2012)). However, the likelihood ratio statistic is. This motivates the tests proposed by Gabriel (1969), which we recall in the next example.

Example 6 *(Likelihood Ratio Tests with fixed threshold) Let $c \in [0,1]$ and define $\mathcal{L}$ by*

$$\mathcal{L}(A)(x) = \mathbb{I}\left(\frac{\sup_{\theta \in A} L_x(\theta)}{\sup_{\theta \in \Theta} L_x(\theta)} \leq c\right), \forall A \in \sigma(\Theta) \text{ and } \forall x \in \mathcal{X}.$$

This class is monotonic. This follows from the fact that if $A, B \in \sigma(\Theta)$ are such that $A \subseteq B$ and $x \in \mathcal{X}$, then $\sup_{\theta \in A} L_x(\theta) \leq \sup_{\theta \in B} L_x(\theta)$.

□

In this example, in order to attain monotonic classes with likelihood ratio tests, one gives up on having common size α for each test. Some authors defend the use of the likelihood itself as a measure of evidence (Bickel (2008)). In these cases, the class defined in Example 6 is appropriate, with cutoffs being chosen by some predefined rules (e.g., Bickel (2008)).

The FBST class defined in Example 4 is in some sense the Bayesian counterpart of Example 6 and is also monotonic:

Example 7 *(Class of tests FBST) $\mathcal{L}$ defined in Example 4 is monotonic. In fact, let $A, B \in \sigma(\Theta)$ be such that $A \subseteq B$ and let $x \in \mathcal{X}$ be such that $\mathcal{L}(A)(x) = 0$. We have $\sup_B f(\theta|x) \geq \sup_A f(\theta|x)$. Hence, $T_x^B \subseteq T_x^A$, and, therefore, $ev_x(A) \leq ev_x(B)$, from which follows the monotonicity of the class.*

□

Bayesian tests based on posterior probabilities with a fixed common cutoff (as in Example 2, with cutoff $1/2$), generated by a family of $0-1-c$ loss functions, are monotonic. This follows from monotonicity

of probabilities. However, other families of loss functions may induce non-monotonic classes of tests: such loss functions lead to a different cutoff for each hypothesis test to be conducted. This is illustrated in Example 8.

Example 8 *Assume* $X \sim Ber(\theta)$, $\theta \in [0,1]$, *and that we are interested in testing the following hypotheses:*

$$H_0^A : \theta \leq 0.6, \text{ and } H_0^B : \theta \leq 0.7.$$

Notice that $H_0^A \subset H_0^B$. *Assume we use the loss functions from Table 2.*

Table 2: Loss function for tests of Example 8

	State of Nature	
Decision	$\theta \in H_0^A$	$\theta \notin H_0^A$
0	0	1
1	2	0

	State of Nature	
Decision	$\theta \in H_0^B$	$\theta \notin H_0^B$
0	0	1
1	1	0

The Bayes tests for testing H_0^A *and* H_0^B *considering these loss functions are, respectively,*

$$\mathcal{L}(H_0^A)(x) = \mathbf{1}\left(\mathbf{P}(\theta \in H_0^A | x) \leq 1/3\right) \text{ and } \mathcal{L}(H_0^B)(x) = \mathbf{1}\left(\mathbf{P}(\theta \in H_0^B | x) \leq 1/2\right).$$

If we assign a uniform prior for θ *and observe* $x = 1$, *we have* $\mathbf{P}(\theta \in H_0^A | x) = 0.36$ *and* $\mathbf{P}(\theta \in H_0^B | x) = 0.49$, *so that we do not reject* H_0^A, *but reject* H_0^B. *As* $H_0^A \subseteq H_0^B$, *we conclude monotonicity does not hold. Intuitively, this happens because the loss of rejecting* H_0^A *when* $\theta \in H_0^A$ *is twice as large as the loss of rejecting* H_0^B *when* $\theta \in H_0^B$. *Hence, we only reject* H_0^A *when there is very little evidence it holds (when compared to the amount of evidence needed to reject* H_0^B*).*

□

A question then arises. What conditions must be imposed on the loss functions so that the resultant CHTs are monotonic? Next, we study monotonicity under a Decision-Theoretic perspective by considering two properties for a family of loss functions $(L_A)_{A \in \sigma(\Theta)}$.

L1 $\forall A \in \sigma(\Theta)$, $\theta \in A \Rightarrow L_A(0,\theta) \leq L_A(1,\theta)$ and $\theta \in A^c \Rightarrow L_A(0,\theta) \geq L_A(1,\theta)$

L2 $\forall A, B \in \sigma(\Theta)$ with $A \subseteq B$ and $\forall \theta \in \Theta$, $L_A(0,\theta) - L_A(1,\theta) \geq L_B(0,\theta) - L_B(1,\theta)$

In words, L1 means that by taking a correct decision we lose the same or less than by taking a wrong decision (as a matter of fact, some authors regard condition L1 in the early definition of a hypothesis testing problem, as Schervish (1997) does). Property L2 can be interpreted in three different cases.

Denoting by *relative loss* the difference between the losses of taking the wrong and the correct decisions, i.e., $L_A(1,\theta) - L_A(0,\theta)$ when $\theta \in A$, and $L_A(0,\theta) - L_A(1,\theta)$ when $\theta \in A^c$, we have that:

- If $\theta \in A$, both A and B are true. L2 describes the situation in which the relative loss is larger for B than for A. The rougher error of rejecting B compared to rejecting A should be assigned a greater relative loss.
- If $\theta \in B \backslash A$, this is a consequence of property L1.
- If $\theta \in B^c$, it can be interpreted in a similar way as the first case.

Example 9 *The following families of loss functions satisfy L1 and L2:*

- *Losses of the form of Table 3, with the restrictions that* $\forall A \in \sigma(\Theta)$, $a_A = b_{A^c}$, *and that* $\forall A, B \in \sigma(\Theta)$ *such that* $A \subseteq B$, $a_A \geq a_B \geq 0$.

Table 3: Example of loss function

	State of Nature	
Decision	$\theta \in A$	$\theta \in A^c$
0	0	a_A
1	b_A	0

- $L_A(0,\theta) = f(d(\theta, A))$ *and* $L_A(1,\theta) = f(d(\theta, A^c))$, *in which* $d(\theta, A)$ *is a distance between* θ *and* A *and* f *is a non-decreasing function in* $\Re_+$.

□

Theorem 2 establishes that L2 is a sufficient condition for producing monotonic classes, and that when L1 holds, L2 is, in some sense, necessary for monotonicity.

Theorem 2 *Let* $(L_A)_{A \in \sigma(\Theta)}$ *be a family of loss functions and* $\mathcal{L}$ *a CHT generated by this family. Suppose that* $\forall A \in \sigma(\Theta)$ *and* $\forall x \in \mathcal{X}$, $|\mathbf{E}[L_A(0,\theta)|x]| < \infty$ *and* $|\mathbf{E}[L_A(1,\theta)|x]| < \infty$. *Then:*

1. *If* $(L_A)_{A \in \sigma(\Theta)}$ *satisfies L2,* $\mathcal{L}$ *is monotonic, whatever the prior distribution for* θ *is.*
2. *If* $(L_A)_{A \in \sigma(\Theta)}$ *satisfies L1, but there exist* $A, B \in \sigma(\Theta)$, *with* $A \subset B$, *and* $\theta_1 \in A$ *and* $\theta_2 \in B^c$, *with* $\{\theta_1\}, \{\theta_2\} \in \sigma(\Theta)$, *such that* $L_A(0,\theta_i) - L_A(1,\theta_i) < L_B(0,\theta_i) - L_B(1,\theta_i)$, $i = 1,2$ *(and, therefore, L2 does not hold), and* $L_{\theta_1}(x), L_{\theta_1}(x) > 0$ $\forall x \in \mathcal{X}$, *then there exists a prior distribution for which* $\mathcal{L}$ *is not monotonic.*

See the Supplementary Material for a proof of part 2. L2 is not reasonable when one prefers "smaller" hypotheses, that is, when the cost of not rejecting a "large" hypothesis is greater than that of not rejecting

a "small" one, even when both are correct (as is the case of the loss in Example 8). Theorem 2 says that this is exactly when monotonicity may not hold. On the other cases, the theorem shows monotonicity will hold. Hence, any class derived from the loss functions of Example 9 is necessarily monotonic.

More properties of monotonic classes, such as further characterizations and its relationship to admissible classes, will be explored in a forthcoming paper.

3.2 Intersection Consonance

The second property we describe involves testing two hypotheses separately and then testing their intersection. From a standard logical point of view, if we reject the intersection of these hypotheses, we should reject *at least* one of the original hypotheses. The following example shows this is not always the case.

Example 10 *(ANOVA) Suppose that* $X_1, \ldots, X_{20}$ *are i.i.d.* $N(\mu_1, \sigma^2)$; $X_{21}, \ldots, X_{40}$ *are i.i.d.* $N(\mu_2, \sigma^2)$ *and* $X_{41}, \ldots, X_{60}$ *are i.i.d.* $N(\mu_3, \sigma^2)$. *Consider the following hypotheses:*

$$H_0^{(1,2,3)}: \ \mu_1 = \mu_2 = \mu_3$$
$$H_0^{(1,2)}: \ \mu_1 = \mu_2$$
$$H_0^{(1,3)}: \ \mu_1 = \mu_3$$

and suppose that we observe the following means and standard-deviations on the data: $\overline{X}_1 = 0.15$; $S_1 = 1.09$; $\overline{X}_2 = -0.13$; $S_2 = 0.5$ $\overline{X}_3 = -0.38$; $S_3 = 0.79$. *Using the likelihood ratio statistics, we have the following p-values for these hypotheses:*

$$p_{H_0^{(1,2,3)}} = 0.0498 \qquad p_{H_0^{(1,2)}} = 0.2564 \qquad p_{H_0^{(1,3)}} = 0.0920.$$

Therefore, at the level of significance $\alpha = 5\%$, *we reject* $H_0^{(1,2,3)}$ *but do not reject either* $H_0^{(1,2)}$ *or* $H_0^{(1,3)}$. *Hence, we conclude that at least two of the three groups have different means. However, when we compare the first with the second, we don't reject that they have the same mean, as well as when we compare the first with the third. Therefore, there is a contradiction.*

□

This contradiction is named as a *consonance* contradiction by Gabriel (1969). Here, we call this property *intersection consonance*, as later we will introduce the *union consonance*. Several variations of intersection consonance were defined in the literature (Bickel (2008); Rosenblum (2012)). Here, we present the definition of $|S|$-intersection consonance, where we use $|S|$ to denote the cardinality of set S.

Definition 5 *(|S|-intersection consonance) A CHT $\mathcal{L}$ satisfies the |S|-intersection consonance if for all sets of indices I with cardinality $|I| \leq |S|$,*

$$\forall \{A_i\}_{i \in I} \subseteq \sigma(\Theta) \text{ such that } \cap_{i \in I} A_i \in \sigma(\Theta),$$

$$\mathcal{L}(\cap_{i \in I} A_i) \leq \max\{\mathcal{L}(A_i)\}_{i \in I}.$$

In words, if we don't reject any of the hypotheses $\{A_i\}_{i \in I}$, we should also not reject their intersection.

In Section 4, we will specially be interested in three cases of intersection consonance, namely

- **finite intersection consonance.** In this case, $S = \{0, 1\}$, and we only require such property to hold for a *finite* number of hypotheses[5].
- **countable intersection consonance.** In this case, $S = \mathbf{N}$, and we only require such property to hold for a *countable* number of hypotheses.
- **complete intersection consonance.** In this case, $S = \Theta$, and we require such property to hold for any set of hypotheses with the same cardinality of the parameter space.

It can be shown that although complete intersection consonance implies countable intersection consonance which implies finite intersection consonance, the reverse is *not* true.

Example 11 *For each $A \in \sigma(\Theta)$, let*

$$\mathcal{L}(A)(x) = \mathbf{1}\left(R(x) \not\subseteq A\right), \forall x \in \mathcal{X},$$

in which $R : \mathcal{X} \longrightarrow \sigma(\Theta)$ is a region estimator of θ. In words, we reject a hypothesis if, and only if, the estimated region is not fully contained in (i.e., is not a subset of) the hypothesis of interest.[6] $\mathcal{L}$ satisfies both the $|\Theta|$-intersection consonance and monotonicity.

□

Many simultaneous hypotheses procedures developed satisfy intersection consonance (see e.g. Sonnemann (2008), as well as Romano et al (2011), who also discusses optimal power properties of such procedures). As noted by Gabriel (1969), tests that satisfy monotonicity and intersection consonance are related to union-intersection tests. The following Theorem establishes this relation in the context of CHTs, and is useful when one wants to build classes that respect these properties. As the proof of this result is essentially the same as that from Gabriel (1969), it is omitted.

[5] It is possible to check that taking $S = \{0, 1\}$ yields the same classes as taking $S = \{0, \ldots, n\}$ for any finite natural n.

[6] This is different from the test that rejects a hypothesis when a region estimate does not intercept it, see Example 13.

Theorem 3 *Let $\mathcal{L}$ be a CHT constructed in the following way: for each $\theta \in \Theta$, a test $\mathcal{L}(\Theta\backslash\{\theta\})$ is fixed*[7]*. For each $A \in \sigma(\Theta)$, define*

$$\mathcal{L}(A) = \max_{\theta \in A^c} \mathcal{L}(\Theta\backslash\{\theta\}),$$

the union-intersection test for A based on tests for the hypotheses $\{\theta\}$, $\theta \in A$ (Casella and Berger (2002)). Then

1. *$\mathcal{L}$ satisfies $|\Theta|$-intersection consonance and monotonicity.*
2. *Let $\mathcal{L}'$ be a CHT that satisfies monotonicity, with $\mathcal{L}'(\Theta\backslash\{\theta\}) = \mathcal{L}(\Theta\backslash\{\theta\})$, $\forall\theta \in \Theta$. If $\mathcal{L}'$ also satisfies $|\Theta|$-intersection consonance, we must have $\mathcal{L}' = \mathcal{L}$.*

While part *1.* of the above theorem provides a way to construct CHTs that are monotonic and satisfy intersection consonance, part *2.* shows that it is not possible to have two distinct CHTs that are monotonic and satisfy intersection consonance while preserving the tests assigned for hypotheses of the type $\Theta\backslash\{\theta\}$. See Gabriel (1969) and Hochberg and Tamhane (1987) for power considerations of union-intersection tests as described in this theorem. Notice also that, because of monotonicity, these tests control the FWE (Sonnemann (2008)).

3.3 Union Consonance

The third property we describe is similar to intersection consonance, however it involves testing the union of two hypotheses. From a logical point of view, if we reject each of the hypotheses A and B, we should also reject their union $A \cup B$. This is equivalent to stating that if we *don't* reject the union of the hypotheses, we should also *not* reject *at least* one of them. The following example shows this is not always the case.

Example 12 *Suppose three candidates are running for a majority election. The proportion of electors voting for each candidate are θ_1, θ_2 and θ_3, with $\sum_{i=1}^{3} \theta_i = 1$. We are interested in testing the following four hypotheses:*

$$H_0^0 : \bigcup_{i=1}^{3} \left\{\theta_i > \frac{1}{2}\right\}, \qquad H_0^1 : \left\{\theta_1 > \frac{1}{2}\right\}$$
$$H_0^2 : \left\{\theta_2 > \frac{1}{2}\right\}, \qquad H_0^3 : \left\{\theta_3 > \frac{1}{2}\right\}.$$

Hence, the null hypothesis H_0^0 is the hypothesis that one of the candidates has more than 50% of the votes, while the null hypothesis H_0^i, for $i = 1, 2, 3$, is the hypothesis that the i^{th} candidate has more than 50% of the votes. Assume we observe a sample of 410 electors. Let $X = (X_1, X_2, X_3)$, in which X_i

[7] We assume that $\{\theta\} \in \sigma(\Theta)$, $\forall\theta \in \Theta$.

is the number of electors in the sample that vote for candidate i, $i = 1, 2, 3$. Assuming a multinomial distribution for $X|\theta$ and using a uniform prior for θ, if the observed sample is $x = (200, 200, 10)$, we have

$$\mathbf{P}\left(\bigcup_{i=1}^{3}\left\{\theta_i > \frac{1}{2}\right\}\middle|x\right) = 0.588; \qquad \mathbf{P}\left(\left\{\theta_1 > \frac{1}{2}\right\}\middle|x\right) = 0.294;$$
$$\mathbf{P}\left(\left\{\theta_2 > \frac{1}{2}\right\}\middle|x\right) = 0.294; \qquad \mathbf{P}\left(\left\{\theta_3 > \frac{1}{2}\right\}\middle|x\right) = 0.000.$$

When using the CHT described in Example 2, we don't reject H_0^0 but reject H_0^i, $i = 1, 2, 3$. From a logical point of view, we have a contradiction: we don't reject that one of the candidates has at least 50% of the votes (i.e., $\theta_i > 1/2$ for some i), however, separately, we conclude that each of the candidates have at most 50% of the votes (i.e., $\theta_i \leq 1/2$ for all i).

□

We call this inconsistency lack of *union consonance*, which we formally define in what follows:

Definition 6 *(|S|-union consonance) A CHT $\mathcal{L}$ satisfies the $|S|$-union consonance if for all sets of indices I with cardinality $|I| \leq |S|$,*

$$\forall\{A_i\}_{i\in I} \subseteq \sigma(\Theta) \text{ such that } \cup_{i\in I} A_i \in \sigma(\Theta),$$
$$\mathcal{L}(\cup_{i\in I} A_i) \geq \min\{\mathcal{L}(A_i)\}_{i\in I}.$$

In words, if we don't reject the union of the hypotheses $\{A_i\}_{i\in I}$, we should also not reject at least one of them.

As with intersection consonance, we are mostly interested in the cases $S = \{0, 1\}, \mathbf{N}$ and Θ.

Example 13 *For each $A \in \sigma(\Theta)$, let*

$$\mathcal{L}(A)(x) = \mathbf{I}\left(R(x) \bigcap A = \emptyset\right), \forall x \in \mathcal{X},$$

in which $R : \mathcal{X} \longrightarrow \sigma(\Theta)$ is a region estimator of θ. In words, we reject a hypothesis if, and only if, the estimated region does not intersect the hypothesis of interest. This very intuitive procedure was proposed by Aitchison (1964) focusing on classical confidence regions. It is straightforward to show $\mathcal{L}$ satisfies $|\Theta|$-union consonance. Also, Hochberg and Tamhane (1987) noticed it is monotonic and hence controls FWE (Theorem 4). In particular, if $R(X)$ has confidence $1 - \alpha$ (i.e., $\mathbf{P}(\theta \in R(X)|\theta) = 1 - \alpha$, $\forall\theta \in \Theta$), the tests for each of the simple hypotheses $\{\theta\}$ have level α.

□

As discussed in the last section, CHT's that are monotonic and satisfy intersection consonance are, to some extent, derived from union-intersection tests. An analogous relationship between union consonance and intersection-union tests holds. More precisely, the following theorem shows that to create a CHT that satisfy monotonicity and union consonance simultaneously, it is only necessary to define the tests for the simple hypotheses (that is, for each $\{\theta\} \in \sigma(\Theta)$) and consider intersection-union tests derived from them.

Theorem 4 *Let $\mathcal{L}$ be a CHT constructed as follows: for each $\theta \in \Theta$, fix a test $\mathcal{L}(\{\theta\})$[8]. For each $A \in \sigma(\Theta)$, define*

$$\mathcal{L}(A) = \min_{\theta \in A} \mathcal{L}(\{\theta\}),$$

the intersection-union test for A based on the hypotheses $\{\theta\}$, $\theta \in A$ (Casella and Berger (2002)). Then,

1. *$\mathcal{L}$ satisfies the $|\Theta|$-union consonance, as well as monotonicity.*
2. *Let $\mathcal{L}'$ be a CHT that satisfies monotonicity, with $\mathcal{L}'(\{\theta\}) = \mathcal{L}(\{\theta\})$, $\forall \theta \in \Theta$. If $\mathcal{L}'$ also satisfies $|\Theta|$-union consonance, we must have $\mathcal{L}' = \mathcal{L}$.*

The proof of this theorem is shown in the Supplementary Material. As is the case of closure procedures (Shaffer (1995)), classes created according to Theorem 4 control the FWE. This follows from Theorem 1. In particular, if each of the tests for the simple hypotheses is of size α, then the FWE is also α. Notice that the class presented in Example 13 is composed of intersection-union tests based on tests of the form $\mathcal{L}(\{\theta\})(x) = \mathbf{I}(\theta \notin R(x))$, $\theta \in \Theta$, as in Theorem 4. The second part of Theorem 4 asserts that such class is the unique extension of the above-mentioned tests assigned to simple hypotheses to a CHT that is monotonic and satisfies union consonance.

In practice, procedures that satisfy both union consonance and monotonicity are usually easier to implement than the traditional closure method described in the introduction. This is because only tests for the simple hypotheses have to be conducted. If Θ is finite, it requires only $|\Theta|$ operations (instead of $2^{|\Theta|}$, as in the case of the closure method when all hypotheses are rejected). Such procedures are also easy to implement when Θ is continuous if confidence regions can be easily built, as in the following example of Analysis of Variance (ANOVA).

Example 14 *(ANOVA) Suppose that $X_{k,1}, \ldots, X_{k,n_k}$ are i.i.d. $N(\mu_k, \sigma^2)$, $k = 1, \ldots, g$, conditionally on $\mu_1, \ldots, \mu_g, \sigma^2$, and that $X_{i,j}$ is independent of $X_{k,l}$ $\forall i \neq k$ and $\forall l$. Here $X_{i,j}$ represents the measurement made on the $j-th$ sample unit of the $i-th$ group. A confidence region for $(\mu_1, \ldots, \mu_g)$ of confidence*

[8] We assume that $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$.

at least $1-\alpha$ presented by Johnson and Wichern (2007) associates to the sample point x the region

$$R(x) = \left\{ (\mu_1, \ldots, \mu_g) \in \Re^g : \forall k \neq l \ \mu_k - \mu_l \in \left[\overline{x}_k - \overline{x}_l \ \pm t_{n-g}\left(\frac{\alpha}{g(g-1)}\right) \sqrt{\frac{s^2}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)} \right], \right.$$
$$\left. k, l = 1, \ldots, g \right\},$$

where $n = n_1 + \ldots + n_g$, $\overline{x}_k$ is the sample average of the k-th group, $s^2 = (n_1 - 1)s_1^2 + \ldots + (n_g - 1)s_g^2$, where s_k^2 is the sample variance of k-th group, and $t_d(\alpha)$ denotes the α percentile of a t distribution with d degrees of freedom. Plugging the region estimator R above in the CHT defined in Example 13 yields a CHT that

1. *is monotonic,*
2. *satisfies $|\Theta|$-union consonance,*
3. *controls the FWE.*

Hence, it is possible to test all hypotheses of interest in an Analysis of Variance problem while preserving these properties. Notice that we are treating σ^2 as a nuisance parameter (see footnote 2 in Section 2).

□

Next example provides another application of Theorem 4, and also illustrates the class of Example 13 for a particular situation.

Example 15 ***(Uniformly Most Powerful Unbiased Tests)*** *Let $X_1, \ldots, X_n | \theta \sim N(\theta, 1)$ and assume one wants to test each of the simple null hypotheses $\{\theta\} \subseteq \Theta$ with the Uniformly Most Powerful Unbiased (UMPU) Tests of level α for them (Casella and Berger (2002)), and extend these tests to $\mathcal{B}(\Theta)$ while preserving monotonicity and complete union consonance. This can be accomplished by using the class of Example 13 with $R(x) = \left(\overline{x} - z_{1-\alpha/2}\frac{1}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{1}{\sqrt{n}}\right)$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal distribution. The extended class of intersection-union tests of Theorem 4 assigns to hypotheses of the form $(-\infty, \theta_0]$, $\theta_0 \in \Theta$, the Uniformly Most Powerful Tests of level $\alpha/2$, and not of level α. Theorem 4 part 2. guarantees that this extension is unique if one desires monotonicity and $|\Theta|$-union consonance starting from level-α UMPU tests for simple hypotheses. Hence, it is not possible to build a CHT composed of UMPU tests for simple hypotheses and for hypotheses of the form $(-\infty, \theta_0]$ (one-sided hypotheses) with common level α that satisfies both monotonicity and $|\Theta|$-union consonance. Therefore, to preserve logical properties, one should consider level-α tests for simple hypotheses, and level-$\alpha/2$ tests for one-sided hypotheses. In a sense, part 2. of Theorem 4 may be seen as an impossibility result for the construction of common level-α UMPU tests in this example.*

□

If one is not interested in controlling the size of the tests, other procedures can be built. Two examples are shown below.

Example 16 *(Likelihood Ratio Tests with fixed threshold) The CHT of Example 6 was already shown to satisfy monotonicity. It also satisfies $|\Theta|$-union consonance. In fact, let $A \in \sigma(\Theta)$. We have*

$$\mathcal{L}(A)(x) \overset{ui}{=} \mathbb{I}\left(\frac{\sup_{\theta\in A} L_x(\theta)}{\sup_{\theta\in\Theta} L_x(\theta)} \leq c\right) = \min_{\theta_0\in A} \mathbb{I}\left(\frac{L_x(\theta_0)}{\sup_{\theta\in\Theta} L_x(\theta)} \leq c\right) \overset{ui}{=} \min_{\theta_0\in A} \mathcal{L}(\{\theta_0\})(x).$$

The result follows from the first part of Theorem 4.

□

There are also Bayesian tests that are in accordance with union consonance. Although classes based on posterior probabilities with a fixed threshold (Example 12) do not respect union consonance, classes of tests FBST do satisfy it:

Example 17 *(Class of tests FBST) Example 4 shows that a FBST class satisfies monotonicity. It can also be shown that it satisfies $|\Theta|$-union consonance, provided that $\forall x \in \chi$ and $\forall a \in \Re^+$, $\mathbb{P}(\{\theta : f(\theta|x) = a\}|x) = 0$ (Izbicki (2010)). It is also interesting to note that this class is a particular case of the CHT's described in Example 13: it can be shown that this CHT is equivalent to*

$$\mathcal{L}(A)(x) = \mathbb{I}\left(A \bigcap HPD_c^x = \emptyset\right),$$

where HPD_c^x is the Highest Probability Density region (Jaynes (1976)) with probability $1-c$, based on observation x. Hence, the FBST procedure can be efficiently implemented by constructing the posterior $(1-c)-HPD$ for θ and not rejecting all hypotheses that intercept it[9]. In a sense, a class of tests FBST extends Lindley's tests for simples hypotheses (Lindley (1965)), according to intersection-union procedures in Theorem 4.

□

3.4 Invertibility

The following example is traditional in introductory statistics courses and illustrates the difference that exists between choosing the labels "null hypothesis" and "alternative hypothesis" under the classical approach to inference.

[9] Note that this procedure is more easily implemented only for the purpose of testing the hypotheses, and not for calculating their measures of evidence.

Example 18 *Suppose that $X|\theta \sim Normal(\theta, 1)$ and that one wants to test the following null hypotheses:*

$$H_0^{\leq} : \theta \leq 0$$
$$H_0^{>} : \theta > 0$$

The Uniformly Most Powerful Tests for these hypotheses have the following critical regions, at the level 5%, respectively:

$$\{x \in \Re : x > 1.64\} \text{ and } \{x \in \Re : x < -1.64\}.$$

Hence, if we observe $x = 1.0$, we do not reject either that the mean is less than or equal to 0 ($H_0^{\leq}$) or that it is greater than 0 ($H_0^{>}$). That is, on one hand, $x = 1.0$ does not bring enough evidence in favor of $\Re_+^*$ ($\Re_-$ is preferred to $\Re_+^*$ in the first test); on the other hand, it suggests $\Re_+^*$ cannot be rejected ($\Re_+^*$ is preferred to $\Re_-$ in the second problem). Therefore, the conclusion drawn from the sample observation about a hypothesis of interest (here $\Re_+^*$, for instance) strongly depends on whether it is considered as the null or the alternative hypothesis. We note that if the level of significance was taken to be any $\alpha > 50\%$, observing $x = 0$ would lead one to reject both $H_0^{\leq}$ and $H_0^{>}$ simultaneously.

□

Many authors believe that the asymmetry between the null and the alternative hypotheses is somewhat unnatural (e.g., Robert (2007), Section 5.3). The next definition formalizes the notion of simultaneous tests independent of the labels "null" and "alternative" for the hypotheses of interest.

Definition 7 *(Invertibility) A CHT $\mathcal{L}$ satisfies invertibility if*

$$\forall A \in \sigma(\Theta),\ \mathcal{L}(A) = 1 - \mathcal{L}(A^c).$$

In words, it is irrelevant which hypothesis is labeled as null and which is labeled as alternative.

Example 19 *Suppose that $(L_A)_{A \in \sigma(\Theta)}$ is a family of loss functions with*

$$L_A(0, \theta) = a_A \mathbf{I}(\theta \notin A) \text{ and } L_A(1, \theta) = b_A \mathbf{I}(\theta \in A),\ \forall \theta \in \Theta,$$

with $a_A = b_{A^c} > 0$, $\forall A \in \sigma(\Theta)$. Let $\theta_0 = \theta_0(x) \in \Theta$ and $\mathcal{L}$ be defined as

$$\mathcal{L}(A)(x) = \mathbf{I}\left(\mathbf{P}(A|x) < \frac{a_A}{a_A + b_A}\right) + \mathbf{I}\left(\mathbf{P}(A|x) = \frac{a_A}{a_A + b_A} \text{ and } \theta_0 \notin A\right),$$

$\forall A \in \sigma(\Theta)$ *and* $\forall x \in \mathcal{X}$. *In words, we reject* A *whenever its posterior probability is smaller than* $\frac{a_A}{a_A+b_A}$, *or its posterior probability is* $\frac{a_A}{a_A+b_A}$ *and* θ_0 *(which may depend on* x*) is not in* A. $\mathcal{L}$ *is a Bayesian CHT generated by the family* $(L_A)_{A\in\sigma(\Theta)}$. *This CHT satisfies both invertibility and monotonicity.*[10]

□

Example 19 can be generalized. In fact, one can verify that any family of loss functions $(L_A)_{A\in\sigma(\Theta)}$ that satisfies $L_A(0,\theta) = L_{A^c}(1,\theta)$, $\forall A \in \sigma(\Theta)$ and $\forall \theta \in \Theta$, generates CHTs that respect invertibility (see, e.g., Silva (2010)). This restriction on the loss functions implies that a type I error for testing A has to be penalized in the same way as a type II error for testing A^c.

Example 20 *Any CHT generated by a point estimation procedure (recall Definition 3) is invertible. Moreover, such CHTs also satisfy monotonicity,* $|\Theta|$*-intersection and* $|\Theta|$*-union consonances.*

□

4 How restrictive are the consistency properties?

In Section 3, we studied four logical properties one may expect for classes of hypotheses tests. We also provided results and examples with useful tests that respect two of these conditions simultaneously (e.g, Theorems 3 and 4, Examples 13, 16 and 19). In this section, we will show that requiring more than two of such properties to hold simultaneously is very restrictive: under quite general conditions, CHTs that satisfy them are always generated by point estimation procedures.

We start by recalling the concept of compatibility of a multiple test procedure, introduced by Lehmann (1957a). Here we define this property adapted to the framework of CHTs.

Definition 8 *(Compatible Class) A CHT* $\mathcal{L}$ *is compatible (or generally consistent) if* $\forall x \in \mathcal{X}$

$$\bigcap_{A\in\sigma(\Theta)} A^{\mathcal{L}(A)(x)} \neq \emptyset,$$

where $A^0 \stackrel{def}{=} A$ *and* $A^1 \stackrel{def}{=} A^c$, *for* $A \in \sigma(\Theta)$.

Compatibility has been considered too strong by many authors (Sonnemann (2008)), including Lehmann himself (Lehmann (1957b)), who provides less stringent definitions motivated by the fact that one might interpret the result of a test $\phi(x) = 0$ as "not reject" rather than "accept". In fact,

[10] When $\mathbb{P}(A|x) = \frac{a_A}{a_A+b_A}$, the decision to not reject A has the same expected loss as the decision of rejecting A. This CHT was chosen because among all classes generated by $(L_A)_{A\in\sigma(\Theta)}$, which are equivalent from a decision-theoretic point of view, it satisfies invertibility. Of course, other CHT's derived from $(L_A)_{A\in\sigma(\Theta)}$ do as well.

when $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, it is straightforward to show that $\mathcal{L}$ is compatible if, and only if, $\mathcal{L}$ is generated by a point estimation procedure (recall Definition 3). Such classes do not allow any kind of logical contradiction among conclusions obtained after testing each of the hypotheses.

We will now put together the properties presented in Section 3 with the goal of understanding how restrictive such requirements are when compared to those of a compatible class. We begin with the following definition:

Definition 9 *(CHT of type $|S|$) We say a CHT is of type $|S|$ if it satisfies the four properties from Section 3: monotonicity, $|S|$-intersection consonance, $|S|$-union consonance and invertibility.*

The following theorem shows alternative characterizations of classes of type $|S|$.

Theorem 5 *Let S be $\{0,1\}, \mathbf{N}$ or Θ. The following are equivalent:*

1. *$\mathcal{L}$ is of type $|S|$;*
2. *$\mathcal{L}$ satisfies monotonicity, $|S|$-intersection consonance and invertibility;*
3. *$\mathcal{L}$ satisfies monotonicity, $|S|$-union consonance and invertibility;*
4. *$\mathcal{L}(\emptyset) = 1$[11], $\mathcal{L}(\Theta) = 0$, and $\mathcal{L}$ satisfies $|S|$-intersection consonance and $|S|$-union consonance;*

For the case $S = \{0,1\}$, we also have the additional equivalence:

5. *$\forall \{A_1, \ldots, A_n\}$ finite measurable partition of Θ,*

$$\sum_{i=1}^{n} (1 - \mathcal{L}(A_i)) = 1.$$

That is, one, and only one, A_i is not rejected.

Moreover, for the case $S = \mathbf{N}$, we also have the equivalence:

5. *$\forall \{A_1, A_2, \ldots\}$ countable measurable partition of Θ,*

$$\sum_{i \geq 1} (1 - \mathcal{L}(A_i)) = 1.$$

That is, one, and only one, A_i is not rejected.

A sketch of the proof of these facts can be found in the Supplementary Material. Of particular interest are characterizations *4.* and *5.*, which do not involve invertibility, controversial among advocates of frequentist methods. Moreover, characterizations *2.* and *3.* show that under invertibility and monotonicity, requiring union consonance is equivalent to requiring intersection consonance. Hence, the definition of a CHT of type $|S|$ may be reduced by requiring either intersection or union consonance.

[11] i.e., $\mathcal{L}(\emptyset)(x) = 1, \forall x \in \mathcal{X}$.

It is possible to show that when $\{\theta\} \in \sigma(\Theta)$, $\forall \theta \in \Theta$, a class is of type $|\Theta|$ if, and only if, it is compatible. Although the proof of this fact is intuitive, we omit it here for the sake of brevity. Hence, *the only examples of CHT's of type* $|\Theta|$ *are those generated by point estimation procedures.* In the remaining of this section, we will investigate whether this is also true when $S = \mathbf{N}$ or $S = \{0,1\}$.

The following theorem shows that, under some conditions, the only classes of type $|\mathbf{N}|$ are *also* the ones generated by a point estimation procedure. Hence, under these conditions, classes of type $|\mathbf{N}|$ are the same as compatible classes, which are, as we argued, the same as classes of type $|\Theta|$.

Theorem 6 *Assume there exists a metrizable topology* $\tau \subseteq \sigma(\Theta)$ *which is Lindelöf*[12]. *Then* $\mathcal{L}$ *is of type* $|\mathbf{N}|$ *if, and only if, it is generated by a point estimation procedure.*

The proof of this theorem is presented in the Supplementary Material.

Corollary 1 *If* $\Theta = \Re^d$, *a class* $\mathcal{L}$ *defined over any sigma-field* $\sigma(\Theta) \supseteq \mathcal{B}(\Theta)$ *is of type* $|\mathbf{N}|$ *if, and only if, it is generated by a point estimation procedure.*

Hence, under some conditions on Θ and $\sigma(\Theta)$, we have that compatible classes, classes of type $|\Theta|$, classes of type $|\mathbf{N}|$ and classes generated by a point estimation procedure are equivalent. Theorem 6 also formally links, in a sense, point estimation and hypothesis testing. In the vast statistical literature, these two celebrated problems are most of the times treated separately[13]. This theorem asserts that a practitioner that desires to use classes of type $|\mathbf{N}|$ cannot decide, for example, that an unknown proportion of interest is at most 50% and estimate it as 52% on the basis of the same sample information.

Are classes of type $|\mathbf{N}|$ in fact more restrictive than classes of type $|\{0,1\}|$? The following theorem, whose proof is presented in the Supplementary Material, shows that the answer is yes.

Theorem 7 *Assume that* $\Theta = \Re^d$ *and* $\mathcal{B}(\Theta) \subseteq \sigma(\Theta)$. *There exists a CHT of type* $|\{0,1\}|$ *which is not of type* $|\mathbf{N}|$. *In particular, if* $\sigma(\Theta) = \mathcal{P}(\Theta)$, *this existence is equivalent to the existence of a nontrivial ultrafilter over* Θ.

It is not possible to prove the existence of a nontrivial ultrafilter using only the Zermelo-Fraenkel axioms. One needs more axioms such as e.g. the Axiom of Choice (Engelking (1989)). Hence, it is not possible to construct "explicit examples" of such classes (see, e.g., Schechter (1996)). Therefore, when $\sigma(\Theta) = \mathcal{P}(\Theta)$, essentially all classes of type $|\{0,1\}|$ that can be built are classes generated by point estimation procedures. It is still an open question whether this is true when $\sigma(\Theta) \subsetneq \mathcal{P}(\Theta)$. Figure 2 summarizes the relationships between the different types of CHT's studied here.

[12] A topology over Θ is Lindelöf if all open covers of Θ admit countable subcovers (Engelking (1989)).

[13] This is not the case of region estimation, as discussed in Section 3.

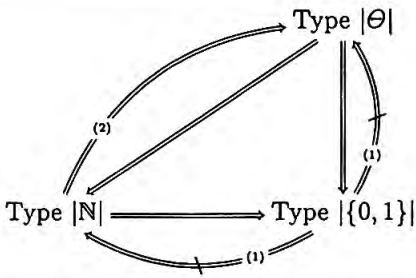


Fig. 2: Summary of relationships between the types of CHT. If the σ-field under consideration contains the singletons, compatible classes and classes generated by point estimation procedures are equivalent to CHT's of type $|\Theta|$. (1) doesn't hold if $\Theta = \Re^d$ and $\mathcal{B}(\Theta) \subseteq \sigma(\Theta)$, (2) holds if there exists a Lindelöf metrizable topology contained in $\sigma(\Theta)$.

5 Discussion and Conclusions

We introduced the concept of a class of hypotheses tests. Such a concept allows one to define several coherence properties that might be expected from simultaneous hypotheses tests. In particular, we studied four properties: monotonicity (also known as *coherence*, Gabriel (1969)), intersection consonance, union consonance and invertibility. Among these, monotonicity is the one that has been most emphasized in the literature (in particular due to closure procedures), followed by intersection consonance. We showed necessary and sufficient conditions for a class to be monotonic from a Bayesian decision-theoretic perspective. We also gave examples of classes of tests that satisfy each of the properties that were defined. Moreover, we gave general procedures that allow one to build classes that satisfy monotonicity and consonance (both for union and intersection) simultaneously. Finally, we showed that when put together, these properties are very restrictive: classes of hypotheses tests that satisfy (any) three of these properties are essentially equivalent to classes generated by point estimation procedures.

The fact that the consistency properties are too restrictive when put together suggests that a practitioner may abandon two or more of these properties when performing simultaneous tests procedures, and then choose a class that combines attainment of some optimality criteria (e.g., controlling the FWE or requiring the CHT to be a Bayesian class derived from an adequate family of loss functions) with agreement to the logical consistency properties he finds more important. We provided several examples that illustrate how this can be done. Alternatively, he might want to use a class based on a sensible point estimation procedure if monotonicity, invertibility and consonance are all of primary importance.

Several problems are open. From a Bayesian decision-theoretic perspective, an alternative way to proceed when dealing with several hypotheses tests is to consider a single decision problem with decision

space $\{0,1\}^{\sigma(\Theta)}$ taking into account joint loss functions rather than CHT's. This is done by e.g. Lavine and Schervish (1999) and Duncan (1965) for a finite number of hypotheses. Which constraints are necessary on such loss functions so that logical properties of interest are preserved?

A different approach that can be taken is that instead of considering decisions in the space $\{0,1\}$, one can create rules taking values on a decision space with three elements: accept a hypothesis of interest, reject it, or do not accept or reject it, the so called "agnostic" tests. See for example Ripley (1996). One can then ask which coherence properties are expected in this framework, which is similar to the one presented by Levi (1967). This approach also seems to be interesting as it naturally deals with the question of how (and to what extent) "not rejecting H" is different from "accepting H", maybe allowing a conciliation between properties expected by different practitioners.

Acknowledgements The authors are thankful for Victor Fossaluza, Jay B. Kadane, Verónica Andrea González-López, Tiago Mendonça, Carlos Alberto de Bragança Pereira, Teddy Seidenfeld, Gustavo Miranda da Silva, Julio Michael Stern, Lea Veras and Sergio Wechsler for their interesting comments and suggestions on this paper. They are especially grateful for Rafael Bassi Stern for the discussions and for helping in the generalization that resulted in Theorem 6, originally Corollary 1.

References

Aitchison J (1964) Confidence-region tests. Journal of the Royal Statistical Society Series B 26(3):462–476 15

Bickel D (2008) The strength of statistical evidence for composite hypotheses with an application to multiple comparisons. COBRA Preprint Series 22(49) 9, 12

Brannath W, Bretz F (2010) Shortcuts for locally consonant closed test procedures. Journal of the American Statistical Association 105(490):660–669 3

Casella G, Berger R (2002) Statistical inference. Duxbury advanced series in statistics and decision sciences, Thomson Learning 4, 14, 16, 17

DeGroot MH (1970) Optimal Statistical Decisions. McGraw-Hill, New York 5

Duncan DB (1965) A Bayesian approach to multiple comparisons. Technometrics 7(2):171–222 24

Engelking R (1989) General topology. Sigma series in pure mathematics, Heldermann Verlag 22

Farcomeni A (2008) Statistical methods in medical research. Biometrika 17(4):347–88 2

Finner H, Gontscharuk V (2009) Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71(5):1031–1048 2

Gabriel KR (1969) Simultaneous test procedures - some theory of multiple comparisons. The Annals of Mathematical Statistics 41(1):224–250 2, 3, 8, 9, 12, 13, 14, 23

Hochberg Y, Tamhane AC (1987) Multiple comparison procedures. John Wiley & Sons, Inc., New York, NY, USA 2, 4, 8, 14, 15

Hommel G, Bretz F (2008) Aesthetics and power considerations in multiple testing – a contradiction? Biometrical Journal 50(5):657–666 2, 3

Izbicki R (2010) Classes de testes de hipóteses (in Portuguese). Master's thesis, University of São Paulo 18

Izbicki R, Fossaluza V, Hounie AG, Nakano EY, Pereira CAdB (2012) Testing allele homogeneity: The problem of nested hypotheses. BMC Genetics 8, 9

Jaynes ET (1976) Confidence Intervals vs Bayesian Intervals. Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science 18

Johnson RA, Wichern DW (2007) Applied Multivariate Statistical Analysis (6th Edition). Prentice Hall 17

Lavine M, Schervish M (1999) Bayes factors: what they are and what they are not. The American Statistician 53:119–122 9, 24

Lehmann EL (1957a) A theory of some multiple decision problems, i. The Annals of Mathematical Statistics 28(1):1–25 3, 20

Lehmann EL (1957b) A theory of some multiple decision problems, ii. The Annals of Mathematical Statistics 28(3):547–572 20

Levi I (1967) Gambling with truth: an essay on induction and the aims of science. MIT Press Classic 24

Lin S, Chen C, Ball D, Liu H, Loh E (2003) Gender-specific contribution of the $gaba_a$ subunit genes on 5q33 in methamphetamine use disorder. Pharmacogenomics Journal 3:349–355 7

Lindley D (1965) introduction to probability and statistics from Bayesian viewpoint, part 2. Cambridge University Press 18

Madruga M, Esteves L, Wechsler S (2001) On the Bayesianity of Pereira-Stern tests. Test 10:291–299 6

Marcus R, Eric P, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63(3):655–660 2

Mayo D, Spanos A (2006) Severe testing as a basic concept in a neyman-pearson philosophy of induction. British journal for the philosophy of science 57:323–357 4

Patriota AG (2013) A classical measure of evidence for general null hypotheses. Fuzzy Sets and Systems Available online 6

Pawitan Y, Calza S, Ploner A (2006) Estimation of false discovery proportion under general dependence. Bioinformatics 22(24):3025–3056 2

Pereira CAdB, Stern JM (1999) Evidence and credibility: Full bayesian significance test for precise hypotheses. Entropy 1(4):99–110 6

Raviv E (2013) On p-value. URL `http://eranraviv.com/blog/on-p-value/` 2

Ripley BD (1996) Pattern Recognition and Neural Networks. Cambridge University Press 24

Robert C (2007) The Bayesian choice: from decision-theoretic foundations to computational implementation, 2nd edn. Springer 19

Romano JP, Shaikh AM, Wolf M (2011) Consonance and the closure method in multiple testing. The International Journal of Biostatistics 7(1) 2, 3, 13

Rosenblum M (2012) Tests that reject at least one subpopulation null hypothesis after rejecting for overall population. Johns Hopkins University, Dept of Biostatistics Working Papers (236):347–88 3, 12

Russell D (1983) Equal employment opportunity commission v. federal reserve bank of richmond. In: 698 Federal Reporter 2d Series, United States Court of Appeals, Fourth Circuit, pp 633–675 2

Schechter E (1996) Handbook of Analysis and Its Foundations. Elsevier Science 22

Schervish MJ (1996) P Values: What They Are and What They Are Not. The American Statistician 50(3):203–206 2

Schervish MJ (1997) Theory of statistics. Springer New York 10

Shaffer J (1995) Multiple hypothesis testing. Annual Review of Psychology 46:561–584 1, 2, 16

Silva GM (2010) Monotonicidade em testes de hipóteses (in Portuguese). Master's thesis, University of São Paulo 20

Sonnemann E (1982) Allgemeine Lösungen multipler Testprobleme. Institut für mathematische Statistik und Versicherung 2, 3, 8, 9

Sonnemann E (2008) General solutions to multiple testing problems. Biometrical Journal 50(6):641656 2, 3, 8, 9, 13, 14, 20

Sonnemann E, Finner H (1988) Vollständigkeitssätze für multiple testprobleme. In: Bauer P, Hommel G, Sonnemann E (eds) Multiple Hypothesenprüfung, Springer, Berlin, pp 121–135 3, 9

Stern J (2011) Constructive verification, empirical induction, and falibilist deduction: A threefold contrast. Information 2:635–650 6

Templeton AR (2010) Coherent and incoherent inference in phylogeography and human evolution. Proceedings of the National Academy of Sciences of the United States of America 107(14):6376–81 2

Zhao H, Wang B, Cui X (2010) General solutions to consistency problems in multiple hypothesis testing. Biometrical Journal 52(6):735746 2, 3

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2014-01 - KOLEV, N. PINTO J. Sibuya-type Bivariate Lack of Memory Property, 23p. (RT-MAE-2014-01)

2014-02 - PINTO, J., KOLEV, N. Extreme value properties of the extended marshall-olkin model.12p. (RT-MAE-2014-02)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05314-970 - São Paulo, Brasil