Brazilian Journal of Probability and Statistics
2020, Vol. 34, No. 2, 230–250
https://doi.org/10.1214/19-BJPS431

Agnostic tests can control the type I and type II errors simultaneously

Victor Coscrato, Rafael Izbicki and Rafael B. Stern

Federal University of São Carlos

Abstract. Despite its common practice, statistical hypothesis testing presents challenges in interpretation. For instance, in the standard frequentist framework there is no control of the type II error. As a result, the non-rejection of the null hypothesis (H_0) cannot reasonably be interpreted as its acceptance. We propose that this dilemma can be overcome by using agnostic hypothesis tests, since they can control the type I and II errors simultaneously. In order to make this idea operational, we show how to obtain agnostic hypothesis in typical models. For instance, we show how to build (unbiased) uniformly most powerful agnostic tests and how to obtain agnostic tests from standard p-values. Also, we present conditions such that the above tests can be made logically coherent. Finally, we present examples of consistent agnostic hypothesis tests.

1 Introduction

Despite its common practice, statistical hypothesis testing presents challenges in interpretation. For instance, some understand that an hypothesis test can either accept or reject the null hypothesis, H_0. However, in this paradigm the probability of accepting H_0 can be high even when H_0 is false. Therefore, it is possible to obtain the undesirable result of accepting H_0 even when this hypothesis is unlikely.

In order to deal with this problem, others propose that an hypothesis test should either reject or *fail to reject* H_0 (Casella and Berger (2002), p. 374, and DeGroot and Schervish (2002), p. 545). Such a position can also lead to challenges in interpretation, since the practitioner often wishes to be able to assert H_0 (Levine et al. (2008)). For example, in regression analysis non-significant predictors are often considered to not affect the response variable and are removed from the model. More generally, scientists often wish to assert a theory (Stern (2011, 2017)).

Neyman (1976), p. 14, briefly introduces an alternative to the above paradigms to hypothesis testing. In this setting, an hypothesis test can have three outcomes: reject H_0, accept H_0, or remain in doubt about H_0—the agnostic decision. This third decision allows the hypothesis test to commit a less severe error (remain in doubt) whenever the data doesn't provide strong evidence either in favor or against the null hypothesis. This approach, which was called agnostic hypothesis testing, was further developed in Berg (2004), Esteves et al. (2016), Stern et al. (2017). This framework allows the acceptance of H_0 while simultaneously controlling the type I and II errors through the agnostic decision. As a result, it is possible to control the probability that H_0 is accepted when H_0 is false.

Although agnostic decisions have been used in classification problems with great success (Lei (2014), Jeske et al. (2017), Jeske and Smith (2017), Sadinle, Lei and Wasserman (2019)) the agnostic hypothesis testing framework has only started to be explored. Here, we generalize to arbitrary hypotheses the setting in Berg (2004), which applies only to hypotheses of

Key words and phrases. Hypothesis test, uniformly most powerful tests, logical consistency, three-decision problem.

Received June 2018; accepted January 2019.

the form: $H_i : \theta = \theta_i$, for $i \in \{0, 1\}$. This generalization allows the translation of standard concepts, such as level, size, power, p-value, unbiased tests, and uniformly most powerful test into the framework of agnostic hypothesis testing. Within this framework, we create new versions of standard statistical techniques, such as t-tests, regression analysis and analysis of variance, which simultaneously control type I and type II errors.

Section 1.1 formally defines agnostic tests and concepts that are used for controlling their error, such as level, size, power and consistency. Sections 2.1 and 2.2 use these definitions to generalize the framework in Berg (2004); they derive agnostic tests that are uniformly most powerful tests and unbiased uniformly most powerful tests. Since it can be hard to obtain the above tests in complex models, Section 3 derives a general approach for controlling the error of agnostic tests that is based on p-values. Section 4 advances results that were obtained in Esteves et al. (2016), Stern et al. (2017) and shows that agnostic tests can control type I and II errors while retaining logical coherence. Section 5 discusses how to control the type I and II errors while obtaining consistent agnostic tests. All proofs are in the Appendix.

1.1 Definitions and notation for agnostic tests

We consider a setting in which the hypotheses that are tested are propositions about a parameter, θ, that assumes values in the parameter space, $\Theta \subset \mathbb{R}^d$. Specifically, the null hypotheses, H_0, are of the form, $H_0 : \theta \in \Theta_0$, where $\Theta_0 \subset \Theta$. The alternative hypotheses, H_1, are of the form $H_1 : \theta \in \Theta_0^c$. In order to test H_0, we use data, $\mathbf{X}$, which assumes values on the sample space, $\mathcal{X}$. Also, $\mathbb{P}_{\theta_0}$ denotes the probability measure over $\mathcal{X}$ when $\theta = \theta_0 \in \Theta$.

Notation 1.1. The i-th element of θ is denoted by $\theta(i)$. This notation is useful because θ_i is often used to denote an element of H_i and not the i-th element of θ.

H_0 is tested through an agnostic test. An agnostic test is a function that, for each observable data point, determines whether H_0 should be rejected, accepted or remain undecided. Let $\mathcal{D} = \{0, \frac{1}{2}, 1\}$ denote the set of possible outcomes of the test: accept H_0 (0), reject H_0 (1), and remain agnostic ($\frac{1}{2}$).

Definition 1.2. An agnostic test is a function, $\phi : \mathcal{X} \to \mathcal{D}$.

Definition 1.3. An agnostic test, ϕ, is a standard test if $\text{Im}[\phi] = \{0, 1\}$.

An agnostic test can have 3 types of errors. The type I and type II errors of agnostic tests are defined in the same way as those of standard tests. That is, a type I error occurs when the test rejects H_0 and H_0 is true. Similarly, a type II error occurs when the test accepts H_0 and H_0 is false. A type III error occurs whenever the test remains agnostic. That is, contrary to type I and type II errors, one knows when type III errors occur. An agnostic test can be designed to simultaneously control the errors of type I and II.

Definition 1.4. An agnostic test, ϕ, has (α, β)-level if the test's probabilities of committing errors of type I and II are controlled by, respectively, α and β. That is,

$$\alpha_\phi := \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(\phi = 1) = \alpha$$

$$\beta_\phi := \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi = 0) = \beta$$

Similarly, ϕ has size (α, β) if the probabilities of committing errors of type I and II are upper bounded by α and β. That is, $\alpha_\phi \leq \alpha$ and $\beta_\phi \leq \beta$.

Agnostic tests can be compared by means of their power. The power function of a test is the probability that it doesn't commit an error. That is, the probability that it accepts H_0 when H_0 is true or rejects H_0 when H_0 is false.

Definition 1.5. The power function of an agnostic test, ϕ, is denoted by $\pi_\phi(\theta)$.

$$\pi_\phi(\theta) = \begin{cases} \mathbb{P}_\theta(\phi = 0), & \text{if } \theta \in H_0 \\ \mathbb{P}_\theta(\phi = 1), & \text{if } \theta \in H_1 \end{cases}$$

The power function induces a partial order among hypothesis tests. If two agnostic tests are such that the power of the first is always at least as large as that of the second, then the first is at least as desirable as the second.

Definition 1.6. Let ϕ_1 and ϕ_2 be agnostic tests. We say that ϕ_1 is uniformly more powerful than ϕ_2 for H_0 and write $\phi_1 \succeq \phi_2$ if, for every $\theta \in \Theta$, $\pi_{\phi_1}(\theta) \geq \pi_{\phi_2}(\theta)$.

In some statistical models, the partial order given by Definition 1.6 is such that there exists a maximal element among hypothesis tests of a given size. That is, if one considers only tests that control type I and II errors by fixed values, then there exists a test that is more powerful than any other. This test is called uniformly most powerful.

Definition 1.7. An unbiased (α, β)-level agnostic test, ϕ^*, is uniformly most powerful (UMP) if, for every other (α, β)-size agnostic test, ϕ, $\phi^* \succeq \phi$.

One way to construct agnostic tests is based on a statistic T of how much the data is inconsistent with H_0. For instance, one could use the likelihood ratio statistic. Next, one builds a test that rejects H_0 when T is large, accepts H_0 when T is small and remains agnostic, otherwise. Such a test is presented in Definition 1.8 and can help to find UMP tests.

Definition 1.8. Let T be a statistic and $c_0 \leq c_1$. The agnostic test, ϕ_{T,c_0,c_1}, is

$$\phi_{T,c_0,c_1}(x) = \begin{cases} 0, & \text{if } T(x) \leq c_0 \\ 1, & \text{if } T(x) > c_1 \\ \dfrac{1}{2}, & \text{otherwise.} \end{cases}$$

UMP tests of the form in Definition 1.8 are presented in Section 2.1.

However, there often do not exist UMP tests among all tests of a given size. This often occurs because it is possible for a test to sacrifice power in a region of Θ in order to obtain a high power in another region. Such sacrifices can yield undesirable tests.

In order to define desirable tests, we consider a test that uses no data. If $\alpha + \beta \leq 1$ and $U \sim \text{Uniform}(0, 1)$, then $\phi^U := \phi_{U,\beta,1-\alpha}$ is called the trivial test of level (α, β), since it uses no data. Furthermore, for every $\theta_0 \in H_0$, $\pi_{\phi^U}(\theta_0) = \beta$ and also for every $\theta_1 \in H_1$, $\pi_{\phi^U}(\theta_1) = \alpha$. One can define that a test is desirable if it is more powerful than a trivial test of the same level. Such tests are usually called unbiased, as in Definition 1.9.

Definition 1.9. An agnostic test, ϕ, is unbiased if

$$\begin{cases} \inf\limits_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(\phi = 0) = \pi_\phi(\theta_0) \geq \sup\limits_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi = 0) = \beta_\phi \\ \inf\limits_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi = 1) = \pi_\phi(\theta_1) \geq \sup\limits_{\theta_0 \in H_0} \mathbb{P}_{\theta_1}(\phi = 0) = \alpha_\phi \end{cases}$$

Note that, if ϕ is unbiased, then $\alpha_\phi + \beta_\phi \leq 1$.

When there exists no uniformly most powerful test among the tests of a given level, a common strategy is to restrict the analysis to unbiased tests. In some such cases, there exists a uniformly most powerful test among unbiased tests of a given level. Such a test is called uniformly most powerful among unbiased tests (UMPU).

Definition 1.10. An (α, β)-level test is said to be uniformly most powerful among unbiased tests (UMPU) if, for every unbiased (α, β)-size test, ϕ, $\phi^* \succeq \phi$.

In order to construct unbiased agnostic hypothesis tests, it is often useful to consider a statistic, V, which assumes large absolute values when the data disagrees with H_0. One construct an agnostic test based on V by rejecting H_0 when V assumes extreme values, accepting H_0 when V is close to 0 and remaining agnostic, otherwise. Such a test is presented in Definition 1.11 and can help to find and describe UMPU tests.

Definition 1.11. Let $c_{0,l}, c_{1,l}, c_{0,r}, c_{1,r} \in \mathbb{R}$ be such that $c_{1,l} \leq c_{0,l} \leq c_{0,r} \leq c_{1,r}$.

$$\phi_{V,\mathbf{c}} = \begin{cases} 1, & \text{if } V < c_{1,l} \text{ or } V > c_{1,r} \\ 0, & \text{if } c_{0,l} \leq V \leq c_{0,r} \\ \dfrac{1}{2}, & \text{otherwise.} \end{cases}$$

UMPU tests of the form in Definition 1.11 are presented in Section 2.2.

However, in some statistical models an UMPU agnostic test might not exist or be hard to find. In such a situation, one might be satisfied by determining an arbitrary (α, β)-level test. A wide class of such tests can be obtained through the p-value of standard hypothesis tests. The definition of p-value is revisited below.

Definition 1.12. A nested family of standard tests for H_0, Φ, is such that

1. For every $\phi \in \Phi$, ϕ is a standard test.
2. The function $g : \Phi \to [0, 1]$, $g(\phi) = \alpha_\phi$ is bijective.
3. If $\phi_1, \phi_2 \in \Phi$ and $\alpha_{\phi_1} \leq \alpha_{\phi_2}$, then $\{x \in \mathcal{X} : \phi_1(x) = 1\} \subset \{x \in \mathcal{X} : \phi_2(x) = 1\}$.

Example 1.13. Let $\lambda(x) = -\log(\frac{\sup_{\theta_0 \in H_0} f_{\theta_0}(x)}{\sup_{\theta \in \Theta} f_\theta(x)})$. The collection of generalized likelihood ratio tests, $\Phi = \{\phi_{\lambda,k,k} : k \geq 0\}$, is a nested family of standard tests for H_0.

Definition 1.14. Let Φ denote a nested family of standard tests for H_0. The p-value of Φ against H_0, $p_{H_0,\Phi} : \mathcal{X} \to [0, 1]$ is such that $p_{H_0,\Phi}(x) = \inf\{\alpha_\phi : \phi \in \Phi \wedge \phi(x) = 1\}$.

If $p := p_{H_0,\Phi}$ is a p-value, then one might consider the agnostic test $\phi_{1-p,1-\alpha,\beta}$. Conditions under which such a test attains level (α, β) are explored in Section 3. The section also illustrates this result with a general linear hypothesis test in regression analysis and with a permutation test.

Another way of constructing agnostic tests is based on region estimators. One of the advantages of such tests is that they are logically consistent (Esteves et al. (2016)). The definitions of region estimator and agnostic tests based on a region estimator are presented below in Definitions 1.15 to 1.17.

Definition 1.15. A region estimator is a function $R : \mathcal{X} \to \mathcal{P}(\Theta)$.

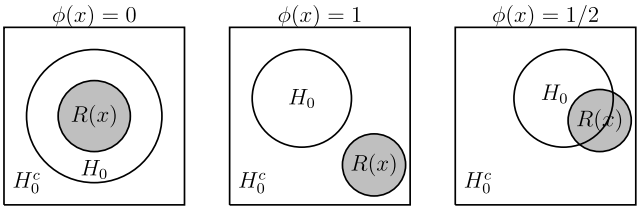

Figure 1 *$\phi(x)$ is an agnostic test based on the region estimator, $R(x)$, for testing H_0.*

Definition 1.16 (Agnostic test based on a region estimator). Let $R(x)$ be a region estimator and $H_0 \subseteq \Theta$. The agnostic test based on R for testing H_0, $\phi_{H_0,R}$ is such that

$$\phi_{H_0,R}(x) = \begin{cases} 0, & \text{if } R(x) \subseteq H_0 \\ 1, & \text{if } R(x) \subseteq H_0^c \\ \dfrac{1}{2}, & \text{otherwise.} \end{cases}$$

Figure 1 illustrates this procedure.

Definition 1.17. A collection of tests, $(\phi_{H_0})_{H_0 \in \mathcal{H}}$ is based on a region estimator if there exists a region estimator, $R(x)$, such that, for every $H_0 \in \mathcal{H}$, ϕ_{H_0} is based on R.

Section 4 shows that if a test is based on a region estimator and the region estimator is a confidence region, then the tests controls type I and type II errors. The section also shows that the unilateral tests in Section 2.1 are based on confidence regions. However, not every agnostic test is based on a region estimator. Section 4.2 shows that every agnostic test is a tested based on two nested region estimators and illustrates this result with the tests in Section 2.2.

One might wish that agnostic tests satisfy additional properties besides controlling the type I and II errors. For instance, one might wish that the power of the test goes to 1 as the sample size increases, that is, the probabilities of type I, type II and type III errors go to 0 as the sample size goes to infinity. Consistency is formalized in Definition 1.18.

Definition 1.18. A sequence of agnostic tests for H_0, $(\phi_n)_{n \in \mathbb{N}}$, is consistent if, for every $\theta \in \Theta$, $\lim_{n \to \infty} \pi_{\phi_n}(\theta) = 1$.

Section 5 shows that, for a wide class of statistical models, it is impossible to obtain consistent tests while uniformly controlling type I and type II errors. It also shows that, if one uses a more flexible control of these errors, then it is possible to obtain consistent tests.

2 The power of agnostic tests

2.1 Uniformly most powerful tests

We start by exploring UMP agnostic tests. Assumption 2.1 presents general conditions under which such tests can be found. These conditions are the same as the ones that are typically used in the standard frequentist framework (Casella and Berger (2002), p. 391).

Assumption 2.1.

1. For every $\theta \in \Theta$, $\mathbb{P}_\theta$ is absolutely continuous with respect to the Lebesgue measure, λ, and $f_\theta(x) := \frac{d\mathbb{P}_\theta}{d\lambda}(x) > 0$.

2. There exists a sufficient statistic for θ, T, and the likelihood is monotone non-decreasing over T.

Theorem 2.2 shows that, under Assumption 2.1, agnostic UMP tests for unilateral hypothesis can be created by checking if the statistic T is larger than some threshold, smaller than another threshold, or if it lies between such values.

Theorem 2.2. *Let* $H_0 = \{\theta \in \Theta : \theta \leq \theta^*\}$, $c_0 \in \mathbb{R}$ *be such that* $\sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(T(X) \leq c_0) = \beta$, *and* $c_1 \in \mathbb{R}$ *be such that* $\sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(T(X) > c_1) = \alpha$. *Under Assumption 2.1,*

1. *If* $c_0 \leq c_1$, *then* ϕ_{T,c_0,c_1} (*Definition 1.8*) *is an UMP* (α, β)*-size agnostic test.*
2. *If* α *and* β *are such that* $c_0 > c_1$ (*and thus* ϕ_{T,c_0,c_1} *is not well defined*), *then let* $\Phi = \{\phi_{T,c,c} : c_1 \leq c \leq c_0\}$. *For every* (α, β)*-size agnostic test,* ϕ, *there exists* $\phi^* \in \Phi$ *such that* $\phi^* \succeq \phi$.

Besides providing a framework for building agnostic UMP tests for unilateral hypothesis, Theorem 2.2 also generalizes several previous results in the literature. For example, if $\Theta = \{\theta_0, \theta_1\}$ and $T(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$, then the likelihood is monotone over T. In this setting, Berg (2004) shows that, if $c_0 \leq c_1$, then ϕ_{T,c_0,c_1} is the UMP agnostic test. Also, one can emulate the standard frequentist framework by not controlling the type II error, that is, by considering $(\alpha, 1)$-size tests. In this case, $\Phi = \{\phi_{T,c,c} : c \leq c_0\}$ is the set of α-size UMP tests in the standard frequentist framework (Casella and Berger (2002), p. 391).

Similarly to this case in which $\beta = 1$, the second condition in Theorem 2.2 occurs whenever the control over α and β is sufficiently weak so that there exist standard tests of size (α, β) and there is no need of using the agnostic decision. In this case, the tests in Φ cannot be uniformly more powerful than one another because of a trade-off in the power in each region of Θ. If $c_2 < c_3$, $\phi_2 = \phi_{T,c_2,c_2}$ and $\phi_3 = \phi_{T,c_3,c_3}$, then the comparison of the critical regions of ϕ_2 and ϕ_3 reveals that the power of ϕ_2 is higher over H_1 and the power of ϕ_3 is higher over H_0. That is, the choice between the elements in Φ depends on the desired balance between the power over H_0 and over H_1.

In the following, Example 2.3 presents an application of Theorem 2.2.

Example 2.3 (Agnostic z-test). Let $X_1, \ldots, X_n$ be an i.i.d. sample with $X_i \sim N(\mu, \sigma^2)$, where $\mu \in \mathbb{R} := \Theta$ and σ^2 is known. Let $H_0 = \{\mu \in \Theta : \mu \leq \mu_0\}$ and $T = \bar{X}$ be the sample mean. Note that the conditions in Assumption 2.1 are satisfied. Furthermore, if $\alpha + \beta \leq 1$, then by taking $c_0 = \mu_0 - \sigma n^{-0.5}\Phi^{-1}(0.5(1-\beta))$ and $c_1 = \mu_0 - \sigma n^{-0.5}\Phi^{-1}(0.5\alpha)$, one obtains that $c_0 \leq c_1$, $\sup_{\theta \in H_0} \mathbb{P}_\theta(T > c_1) = \alpha$ and $\sup_{\theta \in H_1} \mathbb{P}_\theta(T \leq c_0) = \beta$. Therefore, it follows from Theorem 2.2 that ϕ_{T,c_0,c_1} is an UMP (α, β)-level agnostic test.

Figure 2 illustrates the probability of each decision of this test as well as its power function when $\sigma = 1$, $n = 10$ and $\alpha = \beta = 0.05$.

2.2 Unbiased uniformly most powerful tests

If only unbiased tests are considered, it possible to find uniformly most powerful tests in statistical models that are more general than the ones considered in the previous section. In the following, Assumptions 2.4 and 2.6 present general conditions under which there exist tests that are uniformly most powerful among the unbiased tests. These conditions are the same as the ones that are typically used in the standard frequentist framework (Lehmann and Romano (2006), p. 151). Using these conditions, Theorems 2.5 and 2.7 derive UMPU tests for unilateral and bilateral hypotheses, respectively.

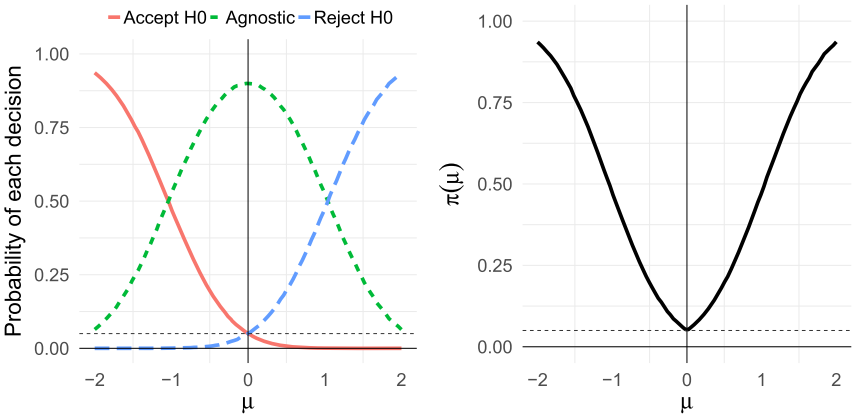

Figure 2 *Probability of each decision for the UMP* $(0.05, 0.05)$*-level agnostic test for* $H_0 : \mu \leq 0$ *(left) and power function for this test (right). The gray dashed horizontal line shows the values* $\alpha = \beta = 0.05$.

Assumption 2.4.

1. For every $\theta \in \Theta$, $\mathbb{P}_\theta$ is absolutely continuous with respect to the Lebesgue measure, λ, and $f_\theta(x) := \frac{d\mathbb{P}_\theta}{d\lambda}(x) > 0$.
2. $\theta \in \mathbb{R}^n = \Theta$ and $f_\theta(x)$ is in the exponential family, that is, there exists $h : \mathbb{R} \to \mathbb{R}^n$ such that $f_\theta(x) = c(x)\exp(\theta \cdot h(x) - d(\theta))$.
3. Let $T(X) = (h_2(X), \ldots, h_n(X))$, where h_i is the i-th component of h. There exists $V(h(X))$ such that V is increasing in $h_1(X)$ and T and V are independent when $\theta(1) = \theta^*$.

Theorem 2.5. *Let* $H_0 = \{\theta \in \Theta : \theta(1) \leq \theta^*\}$, $\bar{\theta} \in \Theta$ *be such that* $\bar{\theta}(1) = \theta^*$, $\alpha + \beta \leq 1$, *and* $c_0, c_1 \in \mathbb{R}$ *be such that* $\mathbb{P}_{\bar{\theta}}(V \leq c_0) = \beta$ *and* $\mathbb{P}_{\bar{\theta}}(V > c_1) = \alpha$, *where* V *is a statistic. Under Assumption 2.4,* ϕ_{V,c_0,c_1} *(Definition 1.8) is an UMPU* (α, β)*-level test.*

Theorem 2.5 shows that, under Assumption 2.4, the test that consists in rejecting H_0 if V is large, accepting it if V is small and remaining agnostic otherwise, is an UMPU unilateral tests on the exponential family. Under the stronger conditions in Assumption 2.6 it is also possible to derive UMPU bilateral tests, as presented in Theorem 2.7.

Assumption 2.6. Besides the conditions in Assumption 2.4, also assume that there exist functions $a, b : \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$ such that

$$V\big(h(X)\big) = a\big(T(X)\big)h_1(X) + b\big(T(X)\big)$$

Theorem 2.7. *Let* $H_0 = \{\theta \in \Theta : \theta(1) = \theta^*\}$, V *be a statistic,* $\bar{\theta} \in \Theta$ *be such that* $\bar{\theta}(1) = \theta^*$, $\alpha + \beta \leq 1$, *and for each* $\gamma \in (0, 1)$, *let* $c_{\gamma,l}$ *and* $c_{\gamma,r}$ *be constants such that*

$$1 - \mathbb{P}_{\bar{\theta}}(c_{\gamma,l} \leq V \leq c_{\gamma,r}) = \gamma$$
$$\mathbb{E}_{\bar{\theta}}\big[V\big(1 - \mathbb{I}(c_{\gamma,l} \leq V \leq c_{\gamma,r})\big)\big] = \gamma\mathbb{E}_{\bar{\theta}}[V]$$

Let $\mathbf{c} = (c_{1-\beta,l}, c_{\alpha,l}, c_{\alpha,r}, c_{1-\beta,r})$. *Under Assumption 2.6,* $\phi_{V,\mathbf{c}}$ *(Definition 1.11) is an UMPU* (α, β)*-level test.*

The conditions from Theorem 2.7 are quite general, and allow the computation of UMPU agnostic tests in several standard statistical problems. In what follows, we illustrate how Theorems 2.5 and 2.7 can be applied to the t-test and linear regression.

Example 2.8 (Agnostic t-test). Let $X_1, \ldots, X_n$ be an i.i.d. sample with $X_i \sim N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$. Let $H_0^{\leq} = \{(\mu, \sigma^2) \in \Theta : \mu \leq \mu_0\}$ and also $H_0^{=} = \{(\mu, \sigma^2) \in \Theta : \mu = \mu_0\}$. Let $V = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sqrt{(n-1)^{-1}\sum_{i=1}^n (X_i-\mu_0)^2}}$. It follows from Lehmann and Romano (2006), p. 153, that V satisfies the conditions in Assumptions 2.4 and 2.6 for testing $H_0^{\leq}$ and $H_0^{=}$. Therefore, if $\alpha + \beta \leq 1$, then it follows from Theorems 2.5 and 2.7 that ϕ_{V,c_0,c_1} and $\phi_{V,\mathbf{c}}$ are the UMPU tests for $H_0^{\leq}$ and $H_0^{=}$. Moreover, by defining $T(X) = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sqrt{(n-1)^{-1}\sum_{i=1}^n (X_i-\bar{X})^2}}$, it follows from Lehmann and Romano (2006), p. 155, that ϕ_{V,c_0,c_1} and $\phi_{V,\mathbf{c}}$ are such that

$$\phi_{V,c_0,c_1}(x) = \begin{cases} 0 & T(x) \leq t_{n-1}(\beta) \\ 1 & T(x) > t_{n-1}(1-\alpha) \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

$$\phi_{V,\mathbf{c}}(x) = \begin{cases} 0, & \text{if } |T(x)| \leq t_{n-1}\big(0.5(1+\beta)\big) \\ 1, & \text{if } |T(x)| > t_{n-1}(1-0.5\alpha) \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

where $t_{n-1}(p)$ is the p-quantile of a Student's t-distribution with $n-1$ degrees of freedom. Figure 3 illustrates the probability of each decision for ϕ_{V,c_0,c_1} and $\phi_{V,\mathbf{c}}$ when $\mu_0 = 0$, $\sigma^2 = 1$, $n = 10$ and $\alpha = \beta = 0.05$. The power of both tests at $\mu_0 = 0$ is β. Indeed, it follows from Assumption 2.4 that the power of a (α, β)-size test at the border points of H_0 cannot be higher than $\min(\alpha, \beta)$.

Example 2.9 (Agnostic linear regression). Consider a linear regression setting, that is, $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $d < n$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I}_d)$, $\mathbb{X}$ is a $n \times d$ design matrix of rank d and $\boldsymbol{\beta}$ is the $d \times 1$ vector with coefficients. For a fixed $k \in \mathbb{R}^d$ and $c \in \mathbb{R}$, let $H_0^{\leq} : k \cdot \boldsymbol{\beta} \leq c$ and $H_0^{=} : k \cdot \boldsymbol{\beta} = c$. Let $\alpha + \beta \leq 1$. By taking $\hat{\boldsymbol{\beta}} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t Y$, the least squares estimator for $\boldsymbol{\beta}$, it follows from Shao (2003), p. 416, that $V = \frac{k^t\hat{\beta}-c}{\sqrt{k^t(\mathbb{X}^t\mathbb{X})^{-1}k\|Y\|_2^2(n-d)^{-1}}}$ satisfies the conditions in

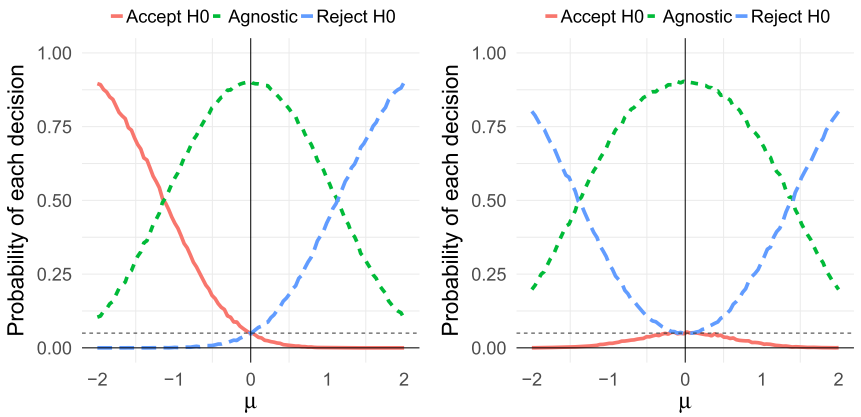

Figure 3 *Probability of each decision for* ϕ_{V,c_0,c_1} *(left) and* $\phi_{V,\mathbf{c}}$ *(right) when* $\mu_0 = 0$, $\sigma^2 = 1$, $n = 10$ *and* $\alpha = \beta = 0.05$.

Assumptions 2.4 and 2.6. Therefore, the UMPU tests, ϕ_{V,c_0,c_1} and $\phi_{V,\mathbf{c}}$, are such that

$$\phi_{V,c_0,c_1}(x) = \begin{cases} 0 & V(x) \leq t_{n-d}(\beta) \\ 1 & V(x) > t_{n-d}(1-\alpha) \\ \frac{1}{2}, & \text{otherwise,} \end{cases}$$

$$\phi_{V,\mathbf{c}}(\mathbf{x}) = \begin{cases} 0, & \text{if } |V(x)| \leq t_{n-d}(0.5(1+\beta)) \\ 1, & \text{if } |V(x)| > t_{n-d}(1-0.5\alpha) \\ \frac{1}{2}, & \text{otherwise,} \end{cases}$$

where $t_{n-d}(q)$ denotes the q quantile of Student's t-distribution with $n-d$ degrees of freedom.

3 General agnostic tests of a given level

An intuitive agnostic procedure consists in rejecting H_0 if the p-value is small, accepting it if the p-value is large and remaining agnostic otherwise. If H_0 is rejected whenever the p-value is smaller than α, then the type I error is controlled by α. Similarly, one might expect that if H_0 is accepted whenever the p-value is larger than $1-\beta$, then the type II error is controlled by β. Theorem 3.1 provides conditions under which this reasoning is valid, and therefore this intuitive agnostic procedure leads to an (α, β)-level test for H_0.

Theorem 3.1. *Let Φ be a nested family of standard tests for H_0 such that, for every $\phi \in \Phi$, ϕ is an unbiased test. Assume that Θ is a connected space and that, for every $x \in \mathcal{X}$, $\mathbb{P}_\theta(p_{H_0,\Phi}(x) \leq t)$ is a continuous function over θ. Let $p = p_{H_0,\Phi}$. Then, the test $\phi_{1-p,\beta,1-\alpha}$, i.e.,*

$$\phi_{1-p,\beta,1-\alpha}(x) = \begin{cases} 0, & \textit{if } p(x) \geq 1-\beta \\ 1, & \textit{if } p(x) < \alpha \\ \frac{1}{2}, & \textit{otherwise} \end{cases}$$

is a (α, β)-level test for H_0.

Next, we apply Theorem 3.1 to the problems of testing the general linear hypothesis, as well as for permutation tests.

Example 3.2 (General Linear Hypothesis in Regression Analysis). Consider the linear regression setting (Example 2.9) and the general linear hypothesis

$$H_0 : \mathbb{K}\boldsymbol{\beta} = \boldsymbol{\gamma}_0$$

where $\mathbb{K}$ is a $q \times d$ matrix and $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$. A particular case of this problem is the ANOVA test (Neter et al. (1996)). There exists no UMPU test for H_0 (Geisser and Johnson (2006)). However, the F-statistic

$$F = \frac{(\mathbb{K}\widehat{\boldsymbol{\beta}} - \boldsymbol{\gamma}_0)^t (\mathbb{K}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{K}^t)^{-1} (\mathbb{K}\widehat{\boldsymbol{\beta}} - \boldsymbol{\gamma}_0) q^{-1}}{(\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbb{X}\boldsymbol{\beta})(n-p)^{-1}}$$

is such that, for every $k \geq 0$, $\phi_{F,k,k}$ is unbiased for H_0 (Monahan (2008)). Furthermore, it can be shown that $p_{H_0,\Phi} = F_{q,n-1}(F)$, where $F_{q,n-1}(\cdot)$ denotes the cumulative distribution function of a Snedecor's F-distribution random variable with $(q, n-1)$ degrees of freedom. Since all conditions in Theorem 3.1 are satisfied, $\phi_{1-F_{q,n-1}(F),\beta,1-\alpha}$ is a (α, β)-level test.

Example 3.3 (Permutation Test). Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be i.i.d. samples from continuous distributions, F_X and F_Y. Also, consider that $H_0 : F_X = F_Y$ and $\Theta = \{(F_X, F_Y) : F_X \text{ is stocastically larger than } F_Y\}$. Let $p_{H_0}(\mathbf{X}, \mathbf{Y})$ be a p-value based on a permutation test such that, if $\mathbf{Y}' = (Y'_1, \ldots, Y'_n)$ is such that, for every $i = 1, \ldots, n$, $y'_i \geq y_i$, then $p_{H_0}(\mathbf{X}, \mathbf{Y}') \geq p_{H_0}(\mathbf{X}, \mathbf{Y})$. It follows from Lehmann and Romano (2006), Lemma 5.9.1, that p_{H_0} is unbiased for H_0. Also, under the topology induced by the total variation metric, Θ is connected and $\mathbb{P}_\theta(p_{H_0} \leq t)$ is continuous over θ. Conclude from Theorem 3.1 that $\phi_{1-p_{H_0},\beta,1-\alpha}$ is a (α, β)-level agnostic test.

4 Connections to region estimation

There exist several known equivalences between standard tests and region estimators (Bickel and Doksum (2015), p. 241). For example, every region estimator is equivalent to a collection of bilateral standard tests. Also, standard tests for more general hypothesis can be obtained as the indicator that the hypothesis intercepts a region estimator. These connections are useful for providing a method of obtaining and interpreting standard hypothesis tests.

The following subsections show that similar results hold for the agnostic tests that were obtained previously. Section 4.1 presents a general method for obtaining agnostic tests from confidence regions. Furthermore, it shows how this method relates to logical coherence and to the unilateral tests in Section 2. Section 4.2 presents an equivalence equivalence between nested region estimators and collections of bilateral agnostic tests.

4.1 Agnostic tests based on a region estimator

An agnostic test can have other desirable properties besides controlling both the type I and type II errors. For instance, Esteves et al. (2016), Stern et al. (2017) show that agnostic tests can be made logically consistent. That is, it is possible to test several hypothesis using agnostic hypothesis tests in such a way that it is impossible to obtain logical contradictions between their conclusions. This property generally cannot be obtained using standard tests (Izbicki and Esteves (2015)). Logically consistent agnostic tests are connected to region estimators, as summarized below.

Theorem 4.1 (Esteves et al. (2016)). *Let $(\phi_{H_0})_{H_0 \in \sigma(\Theta)}$ be a collection of agnostic tests such that $\sigma(\Theta)$ is a σ-field over Θ and, for every $\theta \in \Theta$, $\{\theta\} \in \sigma(\Theta)$. $(\phi_{H_0})_{H_0 \in \sigma(\Theta)}$ is logically consistent if and only if it is based on a region estimator.*

It follows from Theorem 4.1 that the collection of tests based on a region estimator is logically consistent. Theorem 4.2 shows that, if this region estimator has confidence $1 - \alpha$, then the tests based on it also control both the type I and II errors by α.

Theorem 4.2. *If $R(x)$ is a region estimator for θ with confidence $1 - \alpha$ and $\phi_{H_0,R}$ is an agnostic test for H_0 based on R (Definition 1.16), then $\phi_{H_0,R}$ is a (α, α)-size test.*

Theorem 4.2 therefore provides a way of constructing agnostic tests that control types I and II error probabilities.

The unilateral tests that were developed in Sections 2 and 3 are based on confidence regions. In order to present such regions, Theorem 4.5 uses Assumptions 4.3 and 4.4.

Assumption 4.3. Let $H_{0,\theta^*} : \theta(1) \leq \theta^*$. $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$ is a collection of agnostic tests such that

(a) If $\theta_1 \leq \theta_2$ and $\phi_{H_{0,\theta_1}}(x) = 0$, then $\phi_{H_{0,\theta_2}}(x) = 0$
(b) If $\theta_1 \leq \theta_2$ and $\phi_{H_{0,\theta_2}}(x) = 1$, then $\phi_{H_{0,\theta_1}}(x) = 1$.

Assumption 4.4. Let $H_{0,\theta^*} : \theta(1) \leq \theta^*$. $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$ is a collection of agnostic tests such that for every $\theta \in \Theta$ such that $\theta(1) = \theta^*$, $\mathbb{P}_\theta(\phi_{H_{0,\theta^*}} = \frac{1}{2}) \geq 1 - 2\alpha$.

Assumption 4.3 requires that a collection of unilateral tests satisfy a weak form of logical coherence. That is, if $\theta_1 \leq \theta_2$ and the collection of tests accepts that $\theta \leq \theta_1$, then it accepts that $\theta \leq \theta_2$. Similarly, if $\theta_1 \leq \theta_2$ and the collection of tests rejects that $\theta \leq \theta_2$, then it also rejects that $\theta \leq \theta_1$. Assumption 4.4 requires that, for every test in the collection, the probability of the no-decision alternative in the border point of H_0 is at least $1 - 2\alpha$. Theorem 4.5 shows that a collection of unilateral tests that satisfy Assumptions 4.3 and 4.4 is based on a confidence region of confidence $1 - 2\alpha$.

Theorem 4.5. *For each, θ^*, let $H_{0,\theta^*} : \theta(1) \leq \theta^*$. If $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$ satisfies Assumption 4.3, then there exists a region estimator, $R(x)$, such that, for every θ^*, $\phi_{H_{0,\theta^*}}$ is based on $R(x)$. Furthermore, if Assumption 4.4 holds, then $R(x)$ is a confidence region for θ with confidence $1 - 2\alpha$.*

It is possible to use Theorems 4.2 and 4.5 in order to extend a collection of unilateral tests to a larger collection of tests. If the collection of unilateral tests satisfies Assumptions 4.3 and 4.4, then it follows from Theorem 4.5 that these tests are based on a region estimator, $R(X)$, with confidence $1 - 2\alpha$. Therefore, it follows from Theorem 4.2 that, for every H_0 of the type $\theta(1) \in \Theta_0 \subseteq \mathbb{R}$, the test for H_0 based on $R(X)$ has size $(2\alpha, 2\alpha)$. Furthermore, it follows from Theorem 4.1 that the collection of these tests is logically coherent. Corollary 4.6 summarizes these conclusions.

Corollary 4.6. *For each θ^*, let $H_0 : \theta(1) \leq \theta^*$. Also, assume that $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$ satisfies Assumptions 4.3 and 4.4. Let $R(X)$ be such as in Theorem 4.5. Consider the collection of agnostic tests $(\phi_{H_{0,\Theta_0},R})_{\Theta_0 \subset \mathbb{R}}$, where $H_{0,\Theta_0} : \theta(1) \in \Theta_0$ (recall Definition 1.16). Then*

(i) *this collection is logically coherent,*
(ii) *each test is this collection has size $(2\alpha, 2\alpha)$, and*
(iii) *this collection is an extension of the collection $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$.*

Under weak conditions, the tests that were developed in Theorems 2.2 and 2.5 satisfy Assumptions 4.3 and 4.4. As a result, they can be used in Theorem 4.5 and Corollary 4.6. These results are presented in Corollaries 4.7 and 4.10 and illustrated in Examples 4.8 and 4.11.

Corollary 4.7. *Consider the setting of Theorem 2.2, and let $H_{0,\theta^*} : \theta \leq \theta^*$. The collection $\phi_{H_{0,\theta^*}}$ of UMP (α, α)-level test presented in Theorem 2.2 is based on a region estimator, $R(X)$. Furthermore, if T is such that $\mathbb{P}_\theta(T \leq t)$ is continuous over θ, then $R(X)$ has confidence $1 - 2\alpha$ for θ.*

Example 4.8 (Agnostic z-test). Consider again Example 2.3. For each $\mu^* \in \mathbb{R}$, let $H_{0,\mu^*} : \mu \leq \mu^*$. Let $\alpha \leq 0.5$ and $(\phi_{H_{0,\mu^*}})_{\mu^* \in \mathbb{R}}$ be the collection of UMP (α, α)-level tests in Example 2.3. By defining the constants $a_1 = \sigma n^{-0.5}\Phi^{-1}(1 - \alpha)$ and $a_2 = \sigma n^{-0.5}\Phi^{-1}(\alpha)$, note that $\phi_{H_{0,\mu^*}} = \phi_{\bar{X},\mu^*-a_1,\mu^*-a_2}$. It follows that $(\phi_{H_{0,\mu^*}})_{\mu^* \in \mathbb{R}}$ is based on the region estimator $R(X) = [\bar{X} - a_1, \bar{X} - a_2]$, which is a $1 - 2\alpha$ confidence interval for μ.

Assumption 4.9. For each $\theta^* \in \mathbb{R}$, let V_{θ^*} be such as in Assumption 2.4 when $\theta(1) = \theta^*$. There exists a function, $g(v, \theta)$, which is decreasing over θ and such that $g(V_\theta, \theta)$ is ancillary.

Corollary 4.10. *For each $\theta^* \in \mathbb{R}$, let $H_{0,\theta^*} : \theta(1) \leq \theta^*$. Under Assumption 2.4 and $\alpha \leq 0.5$, let $\phi_{H_{0,\theta^*}}$ be the UMPU (α, α)-level test presented in Theorem 2.5. Under Assumption 4.9, the collection $(\phi_{H_{0,\theta^*}})_{\theta^* \in \Theta}$ is based on a region estimator, $R(X)$, which has confidence $1 - 2\alpha$ for θ.*

Example 4.11 (Agnostic t-test). Consider again Example 2.8. For each $\mu^* \in \mathbb{R}$, let $H_{0,\mu^*} : \mu \leq \mu^*$. Let $\alpha \leq 0.5$ and $(\phi_{H_{0,\mu^*}})_{\mu^* \in \mathbb{R}}$ be the collection of UMP (α, α)-level tests in Example 2.8. By defining $S = \sqrt{(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}$, $a_1 = n^{-0.5} S t_{n-1}^{-1}(1 - \alpha)$ and $a_2 = n^{-0.5} S t_{n-1}^{-1}(\alpha)$, note that $\phi_{H_{0,\mu^*}} = \phi_{\bar{X}, \mu^* - a_1, \mu^* - a_2}$. It follows that $(\phi_{H_{0,\mu^*}})_{\mu^* \in \Theta}$ is based on the region estimator $R(X) = [\bar{X} - a_1, \bar{X} - a_2]$, which is a $1 - 2\alpha$ confidence interval for μ.

4.2 Agnostic tests based on nested region estimators

Contrary to the unilateral tests, the bilateral tests in Section 2 are not based on region estimators. Indeed, while these bilateral tests can accept a precise hypothesis, this feature cannot be obtained in tests based on region estimators. However, similarly to the case for standard tests, there exists an equivalence between collections of bilateral agnostic tests and pairs of nested region estimators. Indeed, it is possible to obtain from one another a nested pair of $1 - \alpha$ and β confidence regions and a collection of bilateral (α, β)-size tests. Definition 4.12 prepares for this equivalence, which is established in Theorem 4.14.

Definition 4.12 (Agnostic test based on nested region estimators). Let $R_1(x)$ and $R_2(x)$ be region estimators such that, $R_1(x) \subseteq R_2(x)$ and $H_0 \subseteq \Theta$. The agnostic test based on R_1 and R_2 for testing H_0, ϕ_{H_0, R_1, R_2}, is

$$\phi_{H_0, R_1, R_2}(x) = \begin{cases} 0, & \text{if } H_0 \subseteq R_1 \\ 1, & \text{if } R_2 \subseteq H_0^c \\ \dfrac{1}{2}, & \text{otherwise.} \end{cases}$$

Figure 4 illustrates ϕ_{H_0, R_1, R_2} when $H_0 : \theta = \theta_0$.

Example 4.13 (Agnostic t-test). Consider Example 2.8. For each $\mu^* \in \mathbb{R}$, let $H_{0,\mu^*} : \mu = \mu^*$. The UMPU agnostic test is based on the region estimators

$$R_1(x) = \left[\bar{X} - t_{n-1}(0.5(1+\beta))\sqrt{S^2/n}, \bar{X} + t_{n-1}(0.5(1+\beta))\sqrt{S^2/n}\right], \quad \text{and}$$

$$R_2(x) = \left[\bar{X} - t_{n-1}(1 - 0.5\alpha)\sqrt{S^2/n}, \bar{X} + t_{n-1}(1 - 0.5\alpha)\sqrt{S^2/n}\right]$$

Theorem 4.14. *For each θ^*, let $H_{0,\theta^*} : \theta(1) = \theta^*$.*

1. *If $R_1(x) \subseteq R_2(x)$ are confidence regions for θ with confidence $1 - \beta$ and α, then for every $\theta^* \in \mathbb{R}$, $\phi_{H_{0,\theta^*}, R_1, R_2}$ is a (α, β)-size test.*
2. *Let $(\phi_{H_{0,\theta^*}})_{\theta^* \in \mathbb{R}}$ be a collection of (α, β)-size tests. If for every $\theta \in \Theta$ such that $\theta(1) = \theta^*$, $\mathbb{P}_\theta(\phi_{H_{0,\theta^*}} = 0) = \beta$ and $\mathbb{P}_\theta(\phi_{H_{0,\theta^*}} = 1) = \alpha$, then there exist region estimators, $R_1(x)$ and $R_2(x)$, such that $R_1(x) \subseteq R_2(x)$, $R_1(x)$ and $R_2(x)$ are confidence regions for θ with, respectively, confidence $1 - \beta$ and α and such that $\phi_{H_{0,\theta^*}}$ is based on $R_1(x)$ and $R_2(x)$.*

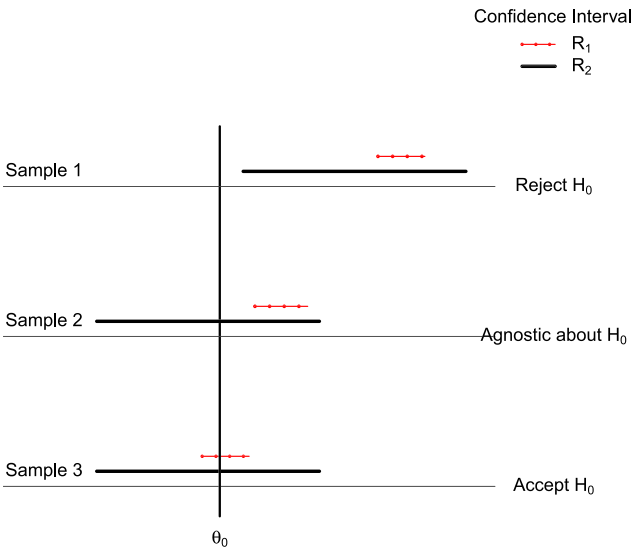

Figure 4 *Illustration of the agnostic test based on* R_1 *and* R_2 *(Definition 4.12) when* $H_0 : \theta = \theta_0$.

5 Consistency and agnostic tests

Under a wide variety of models, it is impossible to obtain consistent agnostic tests. A class of such models is described in Assumption 5.1.

Assumption 5.1 (Non-separability between H_0 and H_1).

1. Θ is connected.
2. $H_0 \notin \{\varnothing, \Theta\}$.
3. $(\phi_n)_{n\in\mathbb{N}}$ is a sequence of agnostic tests for H_0 such that, for every $n \in \mathbb{N}$ and $i \in \{0, \frac{1}{2}, 1\}$, $\mathbb{P}_\theta(\phi_n = i)$ is continuous over θ.

Assumption 5.1 is met in the examples presented in Sections 2 and 3. Theorem 5.2 shows that, under Assumption 5.1, it is impossible to obtain a consistent sequence of hypothesis test.

Theorem 5.2. *Under Assumption 5.1, if* $(\phi_n)_{n\in\mathbb{N}}$ *is a sequence of* $(\alpha, \beta) - size\ tests\ such\ that$ $\max(\alpha, \beta) < 1$, *then* $(\phi_n)_{n\in\mathbb{N}}$ *is not consistent. Furthermore, under the same assumption, if* $\lim_{n\to\infty} \alpha_n = 0$, $\lim_{n\to\infty} \beta_n = 0$ *and* $(\phi_n)_{n\in\mathbb{N}}$ *is a sequence of* (α_n, β_n)*-size tests, then for some* $\theta \in \Theta$, $\lim_{n\to\infty} \mathbb{P}_\theta(\phi_n = \frac{1}{2}) = 1$.

Despite Theorem 5.2, consistency can be obtained by relaxing the control over the test's errors. For instance, one might drop the requirement that the type II error probability be controlled uniformly over all points in the alternative hypothesis. That is, one might require solely that $\sup_{\theta\in H_1'} \mathbb{P}_\theta(\phi = 0) \leq \beta$, where H_1' is a subset of H_1 which is relevant for the practitioner. Example 5.3 shows that, by using such a relaxed control of the type II error, it is possible to obtain a consistent bilateral z-test.

Example 5.3. Let $X_1, \ldots, X_n$ be a i.i.d. sample with $X_i \sim N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Let $H_0 : \mu = 0$, $\alpha_n = \beta_n = \exp(-o(n))$, $a_n = \frac{-\Phi^{-1}(0.5\alpha_n)\sigma}{\sqrt{n}}$, b_n be such that $b_n \leq a_n$ and $b_n^{-1} = o(\sqrt{n})$, $\mathbf{c}_n = (-a_n, -b_n, b_n, a_n)$, and $\gamma_n = b_n + (\frac{-2\log(\sqrt{2\pi}\beta_n)}{n})^{0.5}$. The agnostic

test $\phi_{\bar{X}_n,\mathbf{c}_n}$ controls the type I error by α_n, and controls the type II error over $H_1^* : |\mu| > \gamma_n$ by β_n. Furthermore, for every $\mu \in \mathbb{R}$, $\lim_{n\to\infty} \pi_{\phi_{\bar{X}_n,\mathbf{c}_n}}(\mu) = 1$. That is, $(\phi_{\bar{X}_n,\mathbf{c}_n})_{n\in\mathbb{N}}$ is consistent.

Example 5.3 obtains consistency by not controlling the type II error in a neighborhood of H_0. The example evades the conditions of Theorem 5.2 since H_0 and H_1' are "probabilistically separated". Also, although the control of type II error is not uniform, for every $\theta \in H_1$ and $\epsilon > 0$, there exists a sample size such that the type II error over θ is controlled by θ.

In practice, one procedure to choose H_1' is to determine a desired effect size through expert knowledge elicitation. The effect size is often easier to interpret than the value of the parameter. This procedure is similar to what is done in power calculations (Neter et al. (1996)). The procedure is illustrated in Example 5.4.

Example 5.4 (Agnostic linear regression). Consider the linear regression setting in Example 2.9. Also, one wishes to test the hypothesis $H_0 : \beta_k = 0$ with the agnostic hypothesis test, ϕ_{T,c_0,c_1} (Definition 1.8), where $T = |\frac{\widehat{\beta}_k}{\sqrt{\widehat{\mathbb{V}}[\widehat{\beta}_k]}}|$. The constant c_1 is chosen so that the test's type I error is α. Also, to choose c_0 recall that, for every $\theta \in \Theta$, the probability that ϕ_{T,c_0,c_1} accepts H_0 is

$$\begin{aligned}\mathbb{P}_\theta(T \le c_0) &= \mathbb{P}_\theta\left(-c_0 \le \frac{\widehat{\beta}_k}{\sqrt{\widehat{\mathbb{V}}[\widehat{\beta}_k]}} \le c_0\right) = \mathbb{P}_\theta\left(-c_0 \le \frac{\frac{\widehat{\beta}_k-\beta_k}{\sqrt{\mathbb{V}[\widehat{\beta}_k]}} + \frac{\beta_k}{\sqrt{\mathbb{V}[\widehat{\beta}_k]}}}{\sqrt{\widehat{\mathbb{V}}[\widehat{\beta}_k]\mathbb{V}[\widehat{\beta}_k]^{-1}}} \le c_0\right) \\ &= \mathbb{P}(-c_0 \le T_{n-d-1,\delta_k} \le c_0), \end{aligned} \tag{1}$$

where $T_{p,\delta}$ has a non-central t-distribution with p degrees of freedom and non-centrality parameters δ, that is, $\delta_k = \frac{\beta_k}{\sqrt{\mathbb{V}[\widehat{\beta}_k]}} = \frac{d_k}{\sqrt{a_k}}$, a_k is the k-th element of the diagonal of the matrix $(\mathbb{X}^t\mathbb{X})^{-1}$, and $d_k = \frac{\beta_k}{\sigma}$ is the Cohen's d effect size of the k-th variable on Y (Cohen (1977)).

A practitioner can determine a desired Cohen's effect size value, d_k^* and a $\beta \in (0, 1)$, and use Equation (1) to choose c_0 such that the type II error is β when the effect size is d_k^*. Since, when $\delta > \delta'$, $T_{p,\delta}$ stochastically dominates $T_{p,\delta'}$ this procedure guarantees that

$$\sup_{\theta\in H_1'} \mathbb{P}_\theta(\phi = 0) = \beta,$$

where $H_1' = \{\theta \in \Theta : \delta_k\sqrt{a_k} \ge d_k^*\}$. That is, type II error probabilities are controlled by β for every parameter value with effect size greater or equal to d_k^*. Note that, when $d_k^* = 0$, the test which is obtained is the standard (α, β)-level test for $H_0^=$ in Example 2.9 when $k = (0, \ldots, 0, 1, 0, \ldots, 0)$ and $c = 0$.

The next section applies the test in Example 5.4 to real data.

6 An application of agnostic tests

The Swiss Fertility and Socioeconomic Indicators (1888) Data (Mosteller and Tukey (1977)) contains socio-economic indicators for 47 French-speaking provinces of Switzerland. In order to test the effect of the available socio-economic indicators over infant mortality rate, we apply the agnostic test in Example 5.4 using $\alpha = 0.05$, $\beta = 0.2$ and $d_k^* = 0.25$, for every k. We also compare the results with Bayes factors (Kass and Raftery (1995)) for the covariate effects, which were obtained using the package in Morey and Rouder (2018) with standard arguments. A more extensive analysis of agnostic Bayesian hypothesis tests can be found in Esteves et al. (2016), Stern et al. (2017).

Table 1 *Standard t-value, p-value and Bayes factors for a regression analysis over the Swiss dataset (Section 6) and the hypothesis $H_0 : \beta_k = 0$. The "Decision" column is the outcome of the agnostic test in Example 5.4*

	t-value	p-value	Bayes factor	Decision
Fertility	2.822	0.007	0.12	Reject
Agriculture	−0.418	0.678	2.04	Accept
Examination	0.385	0.702	2.07	Agnostic
Education	0.719	0.476	1.78	Agnostic
Catholic	0.005	0.996	2.20	Accept

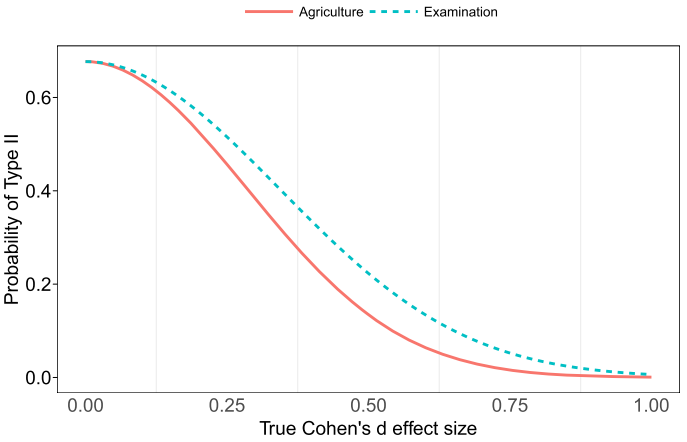

Figure 5 *Probability of type II error as a function the Cohen's effect size in the Swiss dataset for the test in Example 5.4 using the cutoff $c_0 = 1$.*

Table 1 summarizes the results obtained in this applications. One can observe that standard measures such as the t-value, p-value and Bayes factor have the same qualitative behavior, ranking the variables from least associated to most associated to infant mortality rate as: Chatolic, Examination, Agriculture, Education and Fertility. Also, when testing whether each covariate is not associated to the response variable, the agnostic test in Example 5.4 accepts the null hypothesis for the agriculture index of a province and for the percentage of catholics on it, rejects the null hypothesis for fertility and remains agnostic about the remaining variables. As a result, with the exception of the Agriculture and Examination variables, the ranking obtained by standard measures is followed by the decision of the agnostic test.

Figure 5 explains the exception above. For a fixed cutoff c_0, the probability of type II error for Agriculture decreases at a faster rate in function of Cohen's effect size than Examination. Therefore, in order to control the type II error by 0.2, the test in Example 5.4 adopts a critical region for Examination that is more conservative than the one for Agriculture. The difference in these critical regions lead the test to accept Agriculture and remain agnostic about Examination, contrary to the ranking of the statistics in Table 1.

7 Final remarks

Since agnostic tests control the type I and II error probabilities, their outcomes are more interpretable than the ones obtained using standard hypothesis tests. This paper provides several procedures to construct agnostic tests. In several statistical models, (unbiased) uniformly most powerful agnostic tests are obtained. When such tests are unavailable, an alternative that is based on standard p-values is presented. The paper also provides several links between region estimators and agnostic tests, which shows in particular that (α, β)-level tests

can be fully coherent from a logical perspective. Finally, we have shown that although one cannot obtain consistency in agnostic tests that control type I and type II error probabilities uniformly, this goal can be achieved by relaxing the control of the type II error probabilities.

An R package that implements several of the agnostic tests developed here is available at https://github.com/vcoscrato/agnostic.

Appendix: Demonstrations

Definition A.1. Let $g_0(x) = \mathbb{I}(x = 0)$ and $g_1(x) = \mathbb{I}(x = 1)$.

Lemma A.2. *For every agnostic test, ϕ,*

1. *$g_0(\phi)$ and $g_1(\phi)$ are standard tests.*
2. *for every $\theta \in \Theta$, $\mathbb{P}_\theta(g_0(\phi) = 1) = \mathbb{P}_\theta(\phi = 0)$ and $\mathbb{P}_\theta(g_1(\phi) = 1) = \mathbb{P}_\theta(\phi = 1)$.*
3. *If ϕ is unbiased, then $g_1(\phi)$ is unbiased for H_0 and $g_0(\phi)$ is unbiased for $H_0^* = H_1$.*

Proof of Lemma A.2. The first two items follow directly from Definition 1.3 and the definitions of g_0 and g_1. Next, if ϕ is unbiased, then $\alpha_\phi + \beta_\phi \leq 1$. Also,

$$\sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}\big(g_0(\phi) = 1\big) = \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi = 0) = \beta_\phi$$

$$\mathbb{P}_{\theta_0}\big(g_0(\phi) = 1\big) = \mathbb{P}_{\theta_0}(\phi = 0) \geq \beta_\phi \quad \text{for every } \theta_0 \in H_0$$

That is, $g_0(\phi)$ is unbiased for H_0^*. Similarly, $g_1(\phi)$ is unbiased for H_0. □

Lemma A.3. *Let $c, c_0, c_1 \in \mathbb{R}$, $c_0 \leq c_1$ and ϕ be an agnostic test. Also, define $H_0 : \theta_0 \leq \theta^*$ and $H_1 : \theta > \theta^*$, and let $\theta_0 \in H_0$ and $\theta_1 \in H_1$. Under Assumption 2.1,*

1. *If $\mathbb{P}_{\theta^*}(\phi_{T,c_0,c_1} = 1) \geq \mathbb{P}_{\theta^*}(\phi = 1)$, then $\mathbb{P}_{\theta_1}(\phi_{T,c_0,c_1} = 1) \geq \mathbb{P}_{\theta_1}(\phi = 1)$.*
2. *If $\beta_{\phi_{T,c_0,c_1}} \geq \beta_\phi$, then $\mathbb{P}_{\theta_0}(\phi_{T,c_0,c_1} = 0) \geq \mathbb{P}_{\theta_0}(\phi = 0)$.*
3. *If $\mathbb{P}_{\theta^*}(\phi_{T,c,c} = 1) = \mathbb{P}_{\theta^*}(\phi = 1)$, then $\phi_{T,c,c} \succeq \phi$.*

Proof.

1. Let $\theta_1 \in H_1$. Note that $g_1(\phi_{T,c_0,c_1}) = \phi_{T,c_1,c_1}$. Furthermore, it follows from Lemma A.2 that $\mathbb{P}_{\theta^*}(g_1(\phi_{T,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta^*}(g_1(\phi) = 1)$. Therefore, by defining $H_0^* : \theta = \theta^*$ and $H_1^* : \theta = \theta_1$, it follows from Assumption 2.1.2 and the Neyman-Pearson lemma that $\mathbb{P}_{\theta_1}(g_1(\phi_{T,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta_1}(g_1(\phi) = 1)$. The inequality $\mathbb{P}_{\theta_1}(\phi_{T,c_0,c_1} = 1) \geq \mathbb{P}_{\theta_1}(\phi = 1)$ follows from Lemma A.2.
2. Let $\theta_0 \in H_0$. Note that $g_0(\phi_{T,c_0,c_1}) = 1 - \phi_{T,c_0,c_0}$. Furthermore, it follows from Lemma A.2 that $\sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(g_0(\phi_{T,c_0,c_1}) = 1) \geq \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(g_0(\phi) = 1)$ Therefore, by taking $H_0^* = H_1$ and $H_1^* = H_0$, it follows from Assumption 2.1.2 and the Karlin-Rubin theorem that $\mathbb{P}_{\theta_0}(g_1(\phi_{T,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta_0}(g_1(\phi) = 1)$. It follows from Lemma A.2 that $\mathbb{P}_{\theta_0}(\phi_{T,c_0,c_1} = 0) \geq \mathbb{P}_{\theta_0}(\phi = 0)$.
3. It follows from Lemma A.3.1 that, for every $\theta_1 \in H_1$, $\mathbb{P}_{\theta_1}(\phi_{T,c,c} = 1) \geq \mathbb{P}_{\theta_1}(\phi = 1)$. Next, obtain from $\mathbb{P}_{\theta^*}(\phi_{T,c,c} = 1) = \mathbb{P}_{\theta^*}(\phi = 1)$ and $\phi_{T,c,c}$ being a standard test, that $\mathbb{P}_{\theta^*}(\phi_{T,c,c} = 0) \geq \mathbb{P}_{\theta^*}(\phi = 0)$. It follows from Lemma A.2 that $\mathbb{P}_{\theta^*}(g_0(\phi_{T,c,c}) = 1) \geq \mathbb{P}_{\theta^*}(g_0(\phi) = 1)$. By taking $H_0 : \theta = \theta^*$ and $H_1 : \theta = \theta_0$, it follows from Assumption 2.1.2 and the Neyman-Pearson lemma that $\mathbb{P}_{\theta_0}(g_1(\phi_{T,c,c}) = 1) \geq \mathbb{P}_{\theta_0}(g_1(\phi) = 1)$. Obtain from Lemma A.2 that $\mathbb{P}_{\theta_0}(\phi_{T,c,c} = 0) \geq \mathbb{P}_{\theta_1}(\phi = 0)$. Conclude that $\phi_{T,c,c} \succeq \phi$. □

Proof of Theorem 2.2. Let ϕ be an arbitrary (α, β)-size agnostic test.

1. Conclude from Assumption 2.1 that

$$\begin{aligned} P_{\theta^*}(\phi_{T,c_0,c_1} = 1) = \alpha \geq \alpha_\phi \geq \mathbb{P}_{\theta^*}(\phi = 1) \\ \beta_{\phi_{T,c_0,c_1}} = \beta \geq \beta_\phi \end{aligned} \tag{A.1}$$

   It follows from eq. (A.1) and Lemma A.3 that $\phi_{T,c_0,c_1} \succeq \phi$. Since ϕ was arbitrary, conclude that ϕ_{T,c_0,c_1} is an UMP (α, β)-level agnostic test.
2. Either there exists $c \in [c_1, c_0]$ such that $\mathbb{P}_{\theta^*}(\phi_{T,c,c} = 1) = \mathbb{P}_{\theta^*}(\phi = 1)$ or there exists no such c. If there exists such a c, then it follows from Lemma A.3 that $\phi_{T,c,c} \succeq \phi$. Next, assume there exists no such c. Note that ϕ has size (α, β) and, therefore,

$$\mathbb{P}_{\theta^*}(\phi_{T,c_0,c_0} = 1) = \alpha \geq \mathbb{P}_{\theta^*}(\phi = 1).$$

   Since $\mathbb{P}_{\theta^*}(\phi_{T,c,c} = 1)$ decreases continuously over c, conclude that

$$\begin{aligned} \mathbb{P}_{\theta^*}(\phi_{T,c_1,c_1} = 1) \geq \mathbb{P}_{\theta^*}(\phi = 1) \\ \beta_{\phi_{T,c_1,c_1}} = \beta \geq \beta_\phi \end{aligned} \tag{A.2}$$

   Conclude from eq. (A.2) and Lemma A.3 that $\phi_{T,c_1,c_1} \succeq \phi$. □

Lemma A.4. *Let $c, c_0, c_1 \in \mathbb{R}$, $c_0 \leq c_1$ and ϕ be an unbiased test. Define $H_0 : \theta(1) \leq \theta^*$ and $H_1 : \theta(1) > \theta^*$ and let $\bar{\theta} \in \Theta$ be such that $\theta(1) = \theta^*$. Under Assumption 2.4,*

1. *If $\mathbb{P}_{\bar{\theta}}(\phi_{V,c_0,c_1} = 1) \geq \mathbb{P}_{\bar{\theta}}(\phi = 1)$, then $\forall \theta_1 \in H_1$, $\pi_{\phi_{V,c_0,c_1}}(\theta_1) \geq \pi_\phi(\theta_1)$.*
2. *If $\mathbb{P}_{\bar{\theta}}(\phi_{V,c_0,c_1} = 0) \geq \beta_\phi$, then $\forall \theta_0 \in H_0$, $\pi_{\phi_{V,c_0,c_1}}(\theta_0) \geq \pi_\phi(\theta_0)$.*

Proof.

1. Let $\theta_1 \in H_1$. We wish to show that $\mathbb{P}_{\theta_1}(g_1(\phi_{V,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta_1}(g_1(\phi) = 1)$, Since $g_1(\phi_{V,c_0,c_1})$ and $g_1(\phi)$ are standard tests, our strategy is to obtain the inequality from Lehmann and Romano (2006), p. 151. In order to obtain this result, Assumption 2.4 is used to show that $g_1(\phi_{V,c_0,c_1})$ satisfies the required conditions.

   Let $\Theta^* = \{\theta \in \Theta : \theta(1) \geq \bar{\theta}\}$. Note that $g_1(\phi_{V,c_0,c_1}) = \phi_{V,c_1,c_1}$. Also, it follows from Lemma A.2 that $g_1(\phi)$ is unbiased for H_0 under Θ. Since H_0 is more restrictive under Θ^*, $g_1(\phi)$ is also unbiased for H_0 under Θ^*. Moreover, it follows from Lemma A.2 that $\mathbb{P}_{\bar{\theta}}(g_1(\phi_{V,c_0,c_1}) = 1) \geq \mathbb{P}_{\bar{\theta}}(g_1(\phi) = 1)$. It follows from Assumption 2.4 that, under Θ^*, $\alpha_{g_1(\phi_{V,c_0,c_1})} \geq \alpha_{g_1(\phi)}$. Putting all of the above conditions together, conclude that $\mathbb{P}_{\theta_1}(g_1(\phi_{V,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta_1}(g_1(\phi) = 1)$ by applying Lehmann and Romano (2006), p. 151, in Θ^*. It follows directly from Lemma A.2 that $\mathbb{P}_{\theta_1}(\phi_{V,c_0,c_1} = 1) \geq \mathbb{P}_{\theta_1}(\phi = 1)$, which is equivalent to, $\pi_{\phi_{V,c_0,c_1}}(\theta_1) \geq \pi_\phi(\theta_1)$.
2. Let $\theta_0 \in \{\theta \in \Theta : \theta(1) < \theta^*\}$. Note that $g_0(\phi_{V_c 0,c_1}) = 1 - \phi_{V,c_0,c_0}$. Also, it follows from Lemma A.2 that $g_0(\phi)$ is unbiased for $H_0^* = H_1$. Also, obtain from Lemma A.2 and Assumption 2.4.2 that $\mathbb{P}_{\bar{\theta}}(g_0(\phi_{V,c_0,c_1}) = 1) \geq \sup_{\theta_1 \in \Theta_1 \cup \{\theta\}} \mathbb{P}_{\theta_1}(g_0(\phi) = 1)$. Therefore, by taking $H_0^* : \theta(1) \geq \theta^*$, it follows from from Assumption 2.4 and Lehmann and Romano (2006), p. 151, that $\mathbb{P}_{\theta_0}(g_0(\phi_{V,c_0,c_1}) = 1) \geq \mathbb{P}_{\theta_0}(g_0(\phi) = 1)$. Conclude from Lemma A.2 that $\mathbb{P}_{\theta_0}(\phi_{V,c_0,c_1} = 0) \geq \mathbb{P}_{\theta_0}(\phi = 0)$. Since θ_0 was arbitrary in H_1^*, conclude from Assumption 2.4.2 that, for every $\theta_0 \in \overline{H_1^*} = H_0$, $\mathbb{P}_{\theta_0}(\phi_{V,c_0,c_1} = 0) \geq \mathbb{P}_{\theta_0}(\phi = 0)$, that is, $\pi_{\phi_{V,c_0,c_1}}(\theta_0) \geq \pi_\phi(\theta_0)$. □

Proof of Theorem 2.5. Since $\alpha + \beta \leq 1$, obtain $c_0 \leq c_1$. It follows from Assumption 2.4 that ϕ_{V,c_0,c_1} is a (α, β)-level test. Let ϕ be an unbiased (α, β)-size test. Therefore, note that $\mathbb{P}_{\bar{\theta}}(\phi_{V,c_0,c_1} = 1) = \alpha \geq \alpha_\phi$ and $\mathbb{P}_{\bar{\theta}}(\phi_{V,c_0,c_1} = 0) = \beta \geq \beta_\phi$. Conclude from Lemma A.4 that $\phi_{V,c_0,c_1} \succeq \phi$. □

Proof of Theorem 2.7. Since $\alpha + \beta \leq 1$, obtain $c_{1,l} \leq c_{0,l} \leq c_{0,r} \leq c_{1,r}$. Let ϕ be an unbiased (α, β)-size test and $\theta_1 \in H_1$. Since $\alpha_{g_1(\phi_{V,\mathbf{c}})} = \alpha_{\phi_{V,\mathbf{c}}} \geq \alpha_\phi = \alpha_{g_1(\phi)}$, it follows from Assumption 2.6 and Lehmann and Romano (2006), p. 151, that one can obtain $\mathbb{P}_{\theta_1}(g_1(\phi_{V,\mathbf{c}}) = 1) \geq \mathbb{P}_{\theta_1}(g_1(\phi) = 1)$. Conclude from Lemma A.2 that $\mathbb{P}_{\theta_1}(\phi_{V,\mathbf{c}} = 1) \geq \mathbb{P}_{\theta_1}(\phi = 1)$, which is equivalent to, $\pi_{\phi_{V,\mathbf{c}}}(\theta_1) \geq \pi_\phi(\theta_1)$. Next, let $\theta_0 \in H_0$. Since ϕ is an (α, β)-size test, for every $\theta_1 \in H_1$, $\mathbb{P}_{\theta_1}(\phi = 0) \leq \beta$. It follows from Assumption 2.6 that $\mathbb{P}_{\theta_0}(\phi = 0) \leq \beta$, that is, $\pi_\phi(\theta_0) \leq \beta$. Since $\pi_{\phi_{V,\mathbf{c}}}(\theta_0) = \beta$, obtain $\pi_{\phi_{V,\mathbf{c}}}(\theta_0) \geq \pi_\phi(\theta_0)$. □

Definition A.5. A statistic, $T \in \mathbb{R}$, is unbiased for H_0 if, for every $t \in \mathbb{R}$, $\theta_0 \in H_0$ and $\theta_1 \in H_1$, $\mathbb{P}_{\theta_0}(T \leq t) \geq \mathbb{P}_{\theta_1}(T \leq t)$.

Assumption A.6.

1. Θ is a connected space.
2. T is an unbiased statistic for H_0.
3. For every $t \in \mathbb{R}$, $\mathbb{P}_\theta(T \geq t)$ is a continuous function over θ.

Lemma A.7. *Under Assumption A.6, for every $t \in \mathbb{R}$,*

$$\sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(T > t) = 1 - \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(T \leq t)$$

Proof. Let ∂H_0 and ∂H_1 denote the boundaries of H_0 and H_1. Since $H_1 = H_0^c$, $\partial H_0 = \partial H_1$. Also, since Θ is connected, $\partial H_0 \neq \varnothing$. Therefore,

$$\begin{aligned}
\sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(T > t) &= 1 - \inf_{\theta_0 \in H_0} \mathbb{P}_{\theta_1}(T \leq t) \\
&\geq 1 - \inf_{\theta_0 \in \partial H_0} \mathbb{P}_{\theta_0}(T \leq t) \quad \text{Assumption A.6.3} \\
&= 1 - \inf_{\theta_1 \in \partial H_1} \mathbb{P}_{\theta_1}(T \leq t) \\
&\geq 1 - \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(T \leq t) \quad \text{Assumption A.6.3} \qquad \text{(A.3)}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(T > t) &= 1 - \inf_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(T \leq t) \\
&\leq 1 - \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(T \leq t) \quad \text{Assumption A.6.2} \qquad \text{(A.4)}
\end{aligned}$$

The proof follows from eqs. (A.3) and (A.4). □

Lemma A.8. *If Φ is a nested family of standard tests for H_0 such that, for every $\phi \in \Phi$, ϕ is unbiased for H_0, then $1 - p_{H_0,\Phi}$ is an unbiased statistic for H_0.*

Proof. For each $t \in [0, 1]$, let $\phi_t^* \in \Phi$ be such that $\alpha_{\phi_t^*} = t$.

$$\begin{aligned}
\mathbb{P}_{\theta_0}(1 - p_{H_0} \leq t) &= 1 - \mathbb{P}_{\theta_0}(p_{H_0} < 1 - t) \\
&= 1 - \mathbb{P}_{\theta_0}\left(\phi_{1-t}^* = 1\right) \\
&\geq 1 - \alpha_{\phi_{1-t}^*} \\
&\geq 1 - \mathbb{P}_{\theta_1}\left(\phi_{1-t}^* = 1\right) \\
&= 1 - \mathbb{P}_{\theta_1}(p_{H_0} < 1 - t) = \mathbb{P}_{\theta_1}(1 - p_{H_0} \leq t)
\end{aligned}$$

□

Proof of Theorem 3.1.

$$\begin{aligned}
\alpha_{\phi_{\alpha,\beta}} &= \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}\big(1 - p_{H_0}(X) > 1-\alpha\big) \\
&= \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}\big(p_{H_0}(X) < \alpha\big) = \alpha \\
\beta_{\phi_{\alpha,\beta}} &= \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}\big(1 - p_{H_0}(X) \leq \beta\big) \\
&= \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}\big(p_{H_0}(X) \geq 1-\beta\big) \\
&= \sup_{\theta_0 \in H_0} 1 - \mathbb{P}_{\theta_0}\big(p_{H_0}(X) < 1-\beta\big) = \beta \quad \text{Lemmas A.7 and A.8}
\end{aligned}$$

□

Proof of Theorem 4.2. Since $R(x)$ has confidence $1-\alpha$, $\mathbb{P}_\theta(\theta \notin R(x)) \geq \alpha$, for every $\theta \in \Theta$. Therefore,

$$\begin{aligned}
\alpha_{\phi_{R,H_0}} &= \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(\phi_{R,H_0} = 1) = \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}\big(R(X) \subseteq H_0^c\big) \\
&\leq \sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}\big(\theta_0 \notin R(X)\big) \leq \alpha \\
\beta_{\phi_{R,H_0}} &= \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi_{R,H_0} = 0) = \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}\big(R(X) \subseteq H_0\big) \\
&\leq \sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}\big(\theta_1 \notin R(X)\big) \leq \alpha
\end{aligned}$$

□

Proof of Theorem 4.5. Let $\theta^* \in \mathbb{R}$ and $R_1(x)$ be a set. We write $\theta^* < R_1(x)$ if, for every $\theta(1) \in R_1(x)$, $\theta^* < \theta(1)$. Also, $\theta^* > R_1(x)$ if, for every $\theta(1) \in R_1(x)$, $\theta^* > \theta(1)$.

For each $x \in \mathcal{X}$, let $R_1(x) = \{\theta(1) : \phi_{H_{0,\theta(1)}}(x) = \frac{1}{2}\}$. If $\phi_{H_{0,\theta^*}}(x) = 1$, then conclude from Assumption 4.3 that for every $\theta(1) \leq \theta^*$, $\phi_{H_{0,\theta(1)}}(x) = 1$. Therefore, if $\phi_{H_{0,\theta^*}}(x) = 1$, $\theta^* < R_1(x)$. Similarly, if $\phi_{H_{0,\theta^*}}(x) = 0$, then it follows from Assumption 4.3 that $\theta^* > R_1(x)$. Since $\phi_{H_{0,\theta^*}}(x) \in \{0, \frac{1}{2}, 1\}$, conclude that $\phi_{H_{0,\theta^*}}(x) = 1$ if and only if $\theta^* < R_1(x)$ and $\phi_{H_{0,\theta^*}}(x) = 0$ if and only if $\theta^* > R_1(x)$. That is, for every θ^*, $\phi_{H_{0,\theta^*}}$ is based on $R(x) := R_1(x) \times \mathbb{R} \times \cdots \times \mathbb{R}$ for H_{0,θ^*}.

Finally, if Assumption 4.4 holds, then for every $\theta \in \Theta$,

$$\mathbb{P}_\theta\big(\theta \in R(X)\big) = \mathbb{P}_\theta\big(\theta(1) \in R_1(X)\big) = \mathbb{P}_\theta\left(\phi_{H_{0,\theta(1)}} = \frac{1}{2}\right) \geq 1 - 2\alpha$$

That is, $R_1(X)$ has confidence $1-2\alpha$ for $\theta(1)$ and $R(X)$ has confidence $1-2\alpha$ for θ. □

Proof of Corollary 4.6. Follows directly from Theorems 4.1, 4.2 and 4.5. □

Proof of Corollary 4.7. Let T be such as in Assumption 2.1 and $\theta_1, \theta_2, \theta_3 \in \Theta$ be such that $\theta_1 \leq \theta_2 \leq \theta_3$. It follows from Theorem 2.2 that $\phi_{H_{0,\theta_i}} = \phi_{T,c_{0,\theta_i},c_{0,\theta_i}}$, where c_{0,θ_i} and c_{1,θ_i} are such that $\sup_{\theta_1 \in H_{1,\theta_i}} \mathbb{P}_{\theta_1}(T \leq c_{0,\theta_i}) = \alpha$ and $\sup_{\theta_0 \in H_{0,\theta_i}} \mathbb{P}_{\theta_0}(T > c_{1,\theta_i}) = \alpha$. Since $\theta_1 \leq \theta_2$, $H_{0,\theta_1} \subset H_{0,\theta_2}$. Therefore, $c_{1,\theta_1} \leq c_{1,\theta_2}$, that is, if $\phi_{T,c_{0,\theta_2},c_{1,\theta_2}}(x) = 1$, then $\phi_{T,c_{0,\theta_1},c_{1,\theta_1}}(x) = 1$. Similarly, if $\phi_{T,c_{0,\theta_3},c_{1,\theta_3}}(x) = 0$, then $\phi_{T,c_{0,\theta_2},c_{1,\theta_2}}(x) = 0$. Conclude that, if $\phi_{H_{0,\theta_2}}(x) = 0$, then $\phi_{H_{0,\theta_3}}(x) = 0$ and, if $\phi_{H_{0,\theta_2}}(x) = 1$ then $\phi_{H_{0,\theta_1}}(x) = 1$. Also, for every $\theta^* \in \Theta$, it follows from Theorem 2.2 and the continuity of $\mathbb{P}_\theta(T \leq t)$ over θ that $\mathbb{P}(\phi_{H_{0,\theta^*}} = \frac{1}{2}) = 1 - 2\alpha$. The proof follows directly from Theorem 4.5. □

Proof of Corollary 4.10. Since $g(V_\theta, \theta)$ is ancillary, there exist v_α and $v_{1-\alpha}$ such that, for every $\theta \in \Theta$, $\mathbb{P}_\theta((V_\theta, \theta) \leq v_\alpha) = \alpha$ and $\mathbb{P}_\theta(g(V_\theta, \theta) > v_{1-\alpha}) = \alpha$. Since $g(v, \theta)$ is decreasing over θ, for every $\theta \in \Theta$, $\mathbb{P}_\theta(V_\theta \leq g^{-1}(v_\alpha, \theta)) = \alpha$ and $\mathbb{P}_\theta(V_\theta > g^{-1}(v_{1-\alpha}, \theta)) = \alpha$. Conclude from Theorem 2.5 that

$$\phi_{H_{0,\theta^*}} = \phi_{V_{\theta^*}, g^{-1}(v_\alpha, \theta^*), g^{-1}(v_{1-\alpha}, \theta^*)} \tag{A.5}$$

Let $\theta_1 \leq \theta_2 \leq \theta_3$. Since $g^{-1}(v, \theta)$ is increasing over θ, conclude from eq. (A.5) that, if $\phi_{H_{0,\theta_2}}(x) = 1$, then $\phi_{H_{0,\theta_1}}(x) = 1$. Also, if $\phi_{H_{0,\theta_2}}(x) = 0$, then $\phi_{H_{0,\theta_3}}(x) = 0$. Also, it follows from Theorem 2.5 that, for every $\theta \in \Theta$ such that $\theta(1) = \theta^*$, $\mathbb{P}_\theta(\phi_{H_{0,\theta^*}} = \frac{1}{2}) = 1 - 2\alpha$. The proof follows directly from Theorem 4.5. □

Proof of Theorem 4.14. Let

$$R_1^{(1)}(x) = \{\theta^* \in \mathbb{R} : \phi_{H_{0,\theta^*}} = 0\}$$

$$R_2^{(1)}(x) = \left\{\theta^* \in \mathbb{R} : \phi_{H_{0,\theta^*}} \in \left\{0, \frac{1}{2}\right\}\right\}$$

$$R_1(x) = R_1^{(1)}(x) \times \mathbb{R} \times \cdots \times \mathbb{R}$$

$$R_2(x) = R_2^{(1)}(x) \times \mathbb{R} \times \cdots \times \mathbb{R}$$

By construction $R_1(x) \subseteq R_2(x)$, $\phi_{H_{0,\theta^*}}(x) = 0$ if and only if $\{\theta^*\} \subseteq R_1^{(1)}(x)$ (and thus $\phi_{H_{0,\theta^*}}(x) = 0$ if and only if $H_{0,\theta^*} \subseteq R_1(x)$) and $\phi_{H_{0,\theta^*}}(x) = 1$ if and only if $R_2^{(1)}(x) \subseteq \{\theta^*\}^c$ (and thus $\phi_{H_{0,\theta^*}}(x) = 1$ if and only if $R_2(x) \subseteq H_{0,\theta^*}^c$). That is, $\phi_{H_{0,\theta^*}}(x)$ is based on $R_1(x)$ and $R_2(x)$. Furthermore, for every $\theta \in \Theta$,

$$\mathbb{P}_\theta\big(\theta \in R_1(X)\big) = \mathbb{P}_\theta\big(\theta(1) \in R_1^{(1)}(X)\big) = \mathbb{P}_\theta(\phi_{H_{0,\theta(1)}} = 0) = \beta$$

$$\mathbb{P}_\theta\big(\theta \notin R_2(X)\big) = \mathbb{P}_\theta\big(\theta(1) \notin R_2^{(1)}(X)\big) = \mathbb{P}_\theta(\phi_{H_{0,\theta(1)}} = 1) = \alpha$$

Conclude that $R_1(X)$ and $R_2(X)$ are confidence regions with confidence of, respectively, β and $1 - \alpha$. □

Proof of Theorem 5.2. Since Θ is connected and $H_0 \notin \{\varnothing, \Theta\}$, $\partial H_0 \neq \varnothing$. Let $\theta^* \in \partial H_0$. If ϕ_n has size (α_n, β_n), $\sup_{\theta_0 \in H_0} \mathbb{P}_{\theta_0}(\phi_n = 1) \leq \alpha_n$ and $\sup_{\theta_1 \in H_1} \mathbb{P}_{\theta_1}(\phi_n = 0) \leq \beta_n$. It follows from the continuity of $\mathbb{P}_\theta(\phi_n = i)$ that $\mathbb{P}_{\theta^*}(\phi_n = 1) \leq \alpha_n$ and $\mathbb{P}_{\theta^*}(\phi_n = 0) \leq \beta_n$. Therefore, for the first part of the theorem, $\pi_{\phi_n}(\theta^*) \leq \max(\alpha, \beta) < 1$. That is, $\lim_{n\to\infty} \pi_{\phi_n}(\theta^*) \neq 1$ and $(\phi_n)_{n\in\mathbb{N}}$ is not consistent. For the second part of the theorem, Since $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \beta_n = 0$, one obtains that $\lim_{n\to\infty} \mathbb{P}_{\theta^*}(\phi_n = \frac{1}{2}) = 1$. □

Acknowledgments

This research was partially funded by Fundação de Amparo à Pesquisa do Estado de São Paulo (2017/03363-8 and 2019/11321-9) and CNPq (306943/2017-4).

References

Berg, N. (2004). No-decision classification: An alternative to testing for statistical significance. *The Journal of Socio-Economics* **33**, 631–650. https://doi.org/10.1016/j.socec.2004.09.036

Bickel, P. J. and Doksum, K. A. (2015). *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I*, Vol. 117. CRC Press. MR3445928

Casella, G. and Berger, R. L. (2002). *Statistical Inference, Vol. 2*. Pacific Grove, CA: Duxbury. MR1051420

Cohen, J. (1977). Chapter 9—F tests of variance proportions in multiple regression/correlation analysis. In *Statistical Power Analysis for the Behavioral Sciences*, Revised ed. (J. Cohen, ed.) 407–453. Academic Press.

DeGroot, M. H. and Schervish, M. J. (2002). *Probability and Statistics*. Addison-Wesley. MR0373075

Esteves, L. G., Izbicki, R., Stern, J. M. and Stern, R. B. (2016). The logical consistency of simultaneous agnostic hypothesis tests. *Entropy* **18**, 256. MR3550265 https://doi.org/10.3390/e18070256

Geisser, S. and Johnson, W. O. (2006). *Modes of Parametric Statistical Inference, Vol. 529*. John Wiley & Sons. MR2272721

Izbicki, R. and Esteves, L. G. (2015). Logical consistency in simultaneous statistical test procedures. *Logic Journal of the IGPL* **23**, 732–758. MR3403972 https://doi.org/10.1093/jigpal/jzv027

Jeske, D. R., Linehan, J. A., Wilson, T. G., Kawachi, M. H., Wittig, K., Lamparska, K., Amparo, C., Mejia, R., Lai, F., Georganopoulou, D. and Steven, S. S. (2017). Two-stage classifiers that minimize pca3 and the psa proteolytic activity testing in the prediction of prostate cancer recurrence after radical prostatectomy. *The Canadian Journal of Urology* **24**, 9089–9097.

Jeske, D. R. and Smith, S. (2017). Maximizing the usefulness of statistical classifiers for two populations with illustrative applications. In *Statistical Methods in Medical Research*. MR3825911 https://doi.org/10.1177/0962280216680244

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer. MR2135927

Lei, J. (2014). Classification with confidence. *Biometrika* **101**, 755–769. MR3286915 https://doi.org/10.1093/biomet/asu038

Levine, T. R., Weber, R., Park, H. S. and Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research* **34**, 188–209. https://doi.org/10.1111/j.1468-2958.2008.00318.x

Monahan, J. F. (2008). *A Primer on Linear Models*. CRC Press. MR2402599

Morey, R. D. and Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models, Vol. 4*. Chicago: Irwin.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics Theory and Methods* **5**, 737–751. MR0458772 https://doi.org/10.1080/03610927608827392

Sadinle, M., Lei, J. and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*. **114**, 223–234. https://doi.org/10.1080/01621459.2017.1395341

Shao, J. (2003). *Mathematical Statistics, Vol. 2*. Springer. MR2002723 https://doi.org/10.1007/b97553

Stern, J. M. (2011). Symmetry, invariance and ontology in physics and statistics. *Symmetry* **3**, 611–635. MR2845301 https://doi.org/10.3390/sym3030611

Stern, J. M. (2017). Continuous versions of haack's puzzles: Equilibria, eigen-states and ontologies. *Logic Journal of the IGPL* **25**, 604–631. MR3685097 https://doi.org/10.1093/jigpal/jzx017

Stern, J. M., Esteves, L. G., Izbicki, R. and Stern, R. B. (2017). Logically-consistent hypothesis testing and the hexagon of oppositions. *Logic Journal of the IGPL* **25**, 741–757. https://doi.org/10.1093/jigpal/jzx024

Departament of Statistics
Federal University of São Carlos
São Carlos, São Paulo, 13565-905
Brazil
E-mail: vcoscrato@gmail.com
rafaelizbicki@gmail.com
rbstern@gmail.com