Ochieng, Daniel

Article — Published Version

Multiple testing of interval composite null hypotheses using randomized p-values

Statistical Papers

Provided in Cooperation with:
Springer Nature

Suggested Citation: Ochieng, Daniel (2024) : Multiple testing of interval composite null hypotheses using randomized p-values, Statistical Papers, ISSN 1613-9798, Springer, Berlin, Heidelberg, Vol. 65, Iss. 8, pp. 5055-5076, https://doi.org/10.1007/s00362-024-01591-9

This Version is available at:
https://hdl.handle.net/10419/314988

Statistical Papers (2024) 65:5055–5076
https://doi.org/10.1007/s00362-024-01591-9

REGULAR ARTICLE

Multiple testing of interval composite null hypotheses using randomized *p*-values

Daniel Ochieng[1]

Received: 22 December 2023 / Revised: 7 June 2024 / Published online: 2 July 2024

Abstract
Equivalence tests are statistical hypothesis testing procedures that aim to establish practical equivalence rather than the usual statistical significant difference. These testing procedures are frequent in "bioequivalence studies," where one would wish to show that, for example, an existing drug and a new one under development have comparable therapeutic effects. In this article, we propose a two-stage randomized (RAND2) p-value that depends on a uniformly most powerful (UMP) p-value and an arbitrary tuning parameter $c \in [0, 1]$ for testing an interval composite null hypothesis. We investigate the behavior of the distribution function of the two p-values under the null hypothesis and alternative hypothesis for a fixed significance level $t \in (0, 1)$ and varying sample sizes. We evaluate the performance of the two p-values in estimating the proportion of true null hypotheses in multiple testing. We conduct a family-wise error rate control using an adaptive Bonferroni procedure with a plug-in estimator to account for the multiplicity that arises from our multiple hypotheses under consideration. The various claims in this research are verified using a simulation study and real-world data analysis.

Keywords COVID-19 · Equivalence studies · Familywise error · Randomized p-values Two One-Sided Test (TOST)

Mathematics Subject Classification 62J15

1 Introduction

Equivalence tests are testing procedures for establishing practical equivalence rather than the usual statistical significant difference. Within the frequentist framework, this test uses the fact that failing to reject a given null hypothesis of no difference is not logically equivalent to accepting the said null hypothesis (cf. (Fogarty and Small 2014)). Equivalence studies are common in the medical field, for example, where one

✉ Daniel Ochieng
dochieng@uni-bremen.de

[1] Institute for Statistics, University of Bremen, 28344 Bremen, Germany

would wish to show that an existing drug and a new one under development have comparable therapeutic effects. We refer to such studies as "bioequivalence studies." Another area of application is in genetics, where they can be used to identify non-DE (differentially expressed) genes (cf. (Qiu and Cui 2010)) or to test for the Hardy-Weinberg equilibrium (HWE) when we have multiple alleles as in Ostrovski (2020). One can also use these tests to compare the similarity between two Kaplan-Meier curves, which estimate the survival functions in two populations. See Sect. 1.3 of Wellek (2010) for an in-depth discussion of these applications.

Some studies on equivalence testing include (Romano 2005), which provides bounds for the asymptotic power of equivalence tests and constructs efficient tests that attain those bounds. The same author also gives an asymptotically UMP test based on Le Cam's notion of convergence of experiments for testing the mean of a multivariate normal. Equivalence tests can be conducted using the Two One-Sided Test (TOST) procedure. The TOST procedure is an example of an intersection–union test (cf. (Berger and Hsu 1996)) with the null hypothesis as a union of the null for the lower- and upper-sided tests and the alternative as an intersection of the rejection regions for the lower- and upper-sided tests.

Berger and Hsu (1996) consider an intersection–union test for the simultaneous assessment of equivalence on multiple endpoints. This test requires that all the $(1 - 2\alpha)100\%$ simultaneous intervals fall within the equivalence bounds for an overall α level test. This approach can be conservative depending on the correlation structure among the endpoints and the study power. Due to these difficulties, they propose a $100(1 - \alpha)\%$ confidence interval corresponding to a size α test.

Munk (1996) considered equivalence tests for Lehmann's alternative, which are unbiased for equal sample sizes within the two groups. An extension of the expected p-value of a test (EPV) to univariate equivalence tests was considered by Pflüger and Hothorn (2002). Since this procedure is independent of the distribution of the test statistic under the null hypothesis, it avoids the problem of looking for this distribution for the test statistic. Furthermore, the EPV is independent of the nominal level α.

Conducting multiple equivalence tests without multiplicity adjustments increases the probability of making false claims of equivalence (type I errors). Leday et al (2023) proposed a familywise error rate (FWER) control based on Hochberg's method. The same authors also showed that Hommel's method performs as well as Hochberg's and that an "adaptive" version of Bonferroni's method is more powerful than Hommel's for equivalence testing. Giani and Finner (1991) and Giani and Straßburger (1994) on the other hand considered simultaneous equivalence tests in the $k-$sample case and proposed tests based on the range statistic. Qiu and Cui (2010) and Qiu et al (2014) consider multiple equivalence tests based on the average equivalence criterion to identify non-DE genes. Both articles investigate the power and false discovery rate (FDR) of the TOST. Since the variance estimator in the TOST procedure can become unstable and lead to low power for small sample sizes, the latter article proposes a shrinkage variance estimator to improve the power. Huang et al (2006) also applied an average equivalence test criterion but adjusted for the multiplicity using the simultaneous confidence interval approach.

Multiple test procedures that utilize p-values are only valid if the p-value statistics follow the Uniform (0, 1) distribution under the null hypothesis. Since we use the

p-values many times, any non-uniformity in their distribution quickly accumulates and reduces the power of the overall procedure. We can decompose the equivalence hypothesis into two one-sided hypotheses, each leading to a composite null hypothesis. When dealing with a composite null hypothesis, the p- value can fail to follow the Uniform $(0, 1)$ distribution under the null hypothesis if the true parameter used is not the least favorable parameter configuration (LFC). Furthermore, we can have categorical data, for example, in genetic association studies that generate discrete data in counts, leading to test statistics with discrete distributions. Since the p-value is a deterministic transformation of the test statistic, this leads to discretely distributed p-values that are also nonuniform under the null hypothesis.

The problems of composite nulls and discrete test statistics can lead to conservative p-values, which implies that the p-value is stochastically larger than $UNI(0, 1)$ distribution under the null hypothesis. To our knowledge, no research has previously considered a two-stage randomized p-value in testing for the interval composite null hypothesis. We propose a two-stage randomized p-value for multiple testing of interval composite null hypotheses when dealing with discrete data. The two-stage procedure uses the UMP p-value in the first stage to remove the discreteness of the test statistic. The randomized p-value proposed in Hoang and Dickhaus (2022) for a continuous test statistic is then used in the second stage to deal with the composite null hypothesis.

When utilizing the non-randomized version of the Two One-Sided Test (TOST) UMP p-value in discrete models, Finner and Strassburger (2001) showed that it is possible for the power function based on a sample of size n to coincide on the entire parameter space with the corresponding power function based on size $n + i$ for small $i \in \mathbb{N}$. We illustrate that the power function of a test based on the two-stage randomized (RAND2) p-value for discrete models, just like the one for the UMP randomized p-value, is strictly increasing with an increase in the sample size. We further illustrate that for small sample sizes, it is possible that the power functions of the UMP and RAND2 p-values do not strictly increase with an increase in the sample size.

We also investigate the behavior of the distribution function for the UMP and RAND2 p-values under the null hypothesis and alternative hypothesis. Three objectives are of interest: First, to find if the power and level of conservativity of the p-values depend on the size of the equivalence limit. Second, to investigate the behavior of the CDFs of the p-values when the parameter used to compute the p-values is close to or far from the midpoint of the equivalence interval. We are interested in finding the parameter value under the null hypothesis for which the p-values are least conservative or the parameter value that maximizes the power of a test based on our p-values under the alternative hypothesis. Third, to find out if the level of conservativity of the p-values depends on the sample sizes.

Finally, we consider multiple testing of equivalence hypotheses where we assess the performance of our p-values in estimating the proportion of true null hypotheses using an empirical-CDF-based estimator. An adaptive version of the Bonferroni that utilizes the plug-in estimator of Finner and Gontscharuk (2009) is used for familywise error control.

The rest of this paper is organized as follows. General preliminaries are provided in Sect. 2. The definitions, CDFs, and investigations of the behaviors of those CDFs under the null and alternative hypothesis for the UMP and the two-stage randomized

p-values are considered in Sect. 3. We also investigate if the power function of the p-values is monotonically increasing with an increase in the sample size in the same section. Furthermore, we give the parameter value that maximizes the power of a test based on the p-values in the same section. We defer all matters concerning multiple testing until Sect. 4, where we consider a real-world data analysis and a simulation study to assess the performance of the p-values in estimating the proportion of true null hypotheses. Finally, we discuss our results and give recommendations for future research in Sect. 5.

2 General preliminaries

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ denote our random data where each X_r is a real-valued, observable random variable, $1 \leq r \leq n$ with the support of $\boldsymbol{X}$ denoted by $\mathcal{X}$. We assume all X_r are stochastically independent and identically distributed (i.i.d.) with a known parametric distribution. The marginal distribution of X_1 is assumed to be P_θ, where $\theta \in \Theta \subseteq \mathbb{R}$ is the model parameter. The distribution of $\boldsymbol{X}$ under θ is as a result given by $P_\theta^{\otimes n} =: \mathbb{P}_\theta$. We will be concerned with an interval hypothesis test problem of the form

$$H : \theta \notin (\theta_1, \theta_2) \quad \text{versus} \quad K : \theta \in (\theta_1, \theta_2), \tag{1}$$

for given numbers $\theta_1, \theta_2 \in \Theta$ such that $\theta_1 < \theta_2$. When k hypotheses are of interest, then they will be expressed as $H_j : \theta_j \notin \Delta_j$ versus $K_j : \theta_j \in \Delta_j$ where Δ_j denotes the range of values in the j^{th} interval between $\theta_1^{(j)}$ and $\theta_2^{(j)}$ for $j \in \{1, \ldots, k\}$ and k is the multiplicity of the problem. Denote the resulting k p-values by $p_1, \ldots, p_k$. We consider the case $k = 1$ in Sect. 3 and defer the multiple test problem till Sect. 4. When the difference between the j^{th} true parameter $\theta^{(j)}$ and $\theta_1^{(j)}$ or $\theta_2^{(j)}$ $(j = 1, \ldots, k)$ is kept constant for all the k hypotheses, then this is referred to as the "average equivalence" criterion. We can sometimes make the interval in (1) symmetric to achieve equivariance to the permutation of groups, for example, the choice $\theta_2 = \theta_1^{-1}$ in Pflüger and Hothorn (2002) and Munk (1996).

As mentioned before, one method for testing this hypothesis is the Two One-Sided Test (TOST) procedure, where one tests for the alternatives $\theta < \theta_1$ and $\theta > \theta_2$ separately at size α and in no particular order. The TOST procedure is a special case of the intersection–union test proposed by Berger (1982) where the null hypothesis is a union of disjoint sets, and the alternative hypothesis is an intersection of the complements of those sets. For this reason, we conduct the separate individual tests at size α without a multiplicity adjustment like $\alpha/2$. Practical equivalence is declared if one rejects both tests and otherwise non-equivalence. These procedures suffer from a lack of power, and an alternative that is more powerful but too complicated has been suggested in the literature by Berger and Hsu (1996) and Brown et al (1997). Since alternative tests are difficult to implement, we use the TOST procedure in this research.

We consider test statistics $T(\boldsymbol{X})$, where $T : \mathcal{X} \to \mathbb{R}$ is a measurable mapping. Furthermore, the test statistics T_r for $r = 1, \ldots, n$ are also assumed to be mutually

independent. The marginal p-value $p(\boldsymbol{X})$ resulting from $T(\boldsymbol{X})$ is assumed to be valid, meaning that $\mathbb{P}_\theta(p(\boldsymbol{X}) \leq t) \leq t$ holds for all significance levels $t \in (0, 1)$ and for any parameter value θ in the null hypothesis. Valid p-values are stochastically larger than UNI $(0, 1)$, as investigated by, among many others, Habiger and Pena (2011) and Dickhaus et al (2012). On the same note, we call a p-value conservative if it is valid and $\mathbb{P}_\theta(p(\boldsymbol{X}) \leq t) < t$ holds for some fixed significance level $t \in (0, 1)$. Throughout the article, we refer to the cumulative distribution function (CDF) of a p-value under the alternative hypothesis as a power function because we reject the null hypothesis for small p-values. Finally, we also make use of the (generalized) inverses of certain non-decreasing functions mapping from $\mathbb{R}$ to $[0, 1]$. In this regard, we follow Appendix 1 in Reiss (1989): If F is a real-valued, non-decreasing, right-continuous function, and similarly G is a real-valued, non-decreasing, left-continuous function where we define both F and G on $\mathbb{R}$, then $F^{-1}(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}$ and $G^{-1}(y) = \sup\{x \in \mathbb{R} : G(x) \leq y\}$, respectively.

3 Interval composite hypothesis

3.1 Introduction

In this article, we are interested in the (interval) composite null hypothesis of the form in (1). We test this hypothesis using two different p-values whose definitions and the CDFs are as follows.

Definition 1 (First stage randomization) Let U be a UNI$(0, 1)$-distributed random variable independent of the data $\boldsymbol{X}$. Further assume that $T(\boldsymbol{X})$ is our test statistic whose distribution has monotone likelihood ratio (MLR), the UMP-based p-value $P^{UMP}(\boldsymbol{X}, U)$ is

$$P^{UMP}(\boldsymbol{X}, U) = \mathbb{P}_{\theta_i}(C_n < T(\boldsymbol{X}) < D_n) + U\mathbb{P}_{\theta_i}(T(\boldsymbol{X}) = C_n) + U\mathbb{P}_{\theta_i}(T(\boldsymbol{X}) = D_n), \tag{2}$$

for $i = 1, 2$ where θ_1, θ_2 such that $\theta_1 < \theta_2$ are the LFC parameters and C_n, $D_n \in \mathbb{R}$ such that $C_n \leq D_n$ are the critical constants. The CDF of $P^{UMP}(\boldsymbol{X}, U)$ is

$$\mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}) \leq t\} = \mathbb{P}_\theta(C_n < T(\boldsymbol{X}) < D_n) + \gamma_n\mathbb{P}_\theta(T(\boldsymbol{X}) = C_n) + \delta_n\mathbb{P}_\theta(T(\boldsymbol{X}) = D_n), \tag{3}$$

where θ is the chosen true parameter while γ_n and δ_n are the randomization constants. The critical constants C_n, $D_n \in \mathbb{R}$ and the randomization constants γ_n, $\delta_n \in [0, 1]$ are found by solving the equation $E_{\theta_i}[T(\boldsymbol{X})] = \alpha$ for $i = 1, 2$ where $T(\boldsymbol{X}) = \sum_{r=1}^{n} T(X_r)$. For large sample sizes, the critical and the randomization constants are $C_n = F_{\theta_1}^{-1}(1 - t)$, $D_n = F_{\theta_2}^{-1}(t)$,

$$\gamma_n = \frac{\mathbb{P}_{\theta_1}(T(\boldsymbol{X}) \leq C_n) - (1 - c)}{\mathbb{P}_{\theta_1}(T(\boldsymbol{X}) = C_n)}, \text{ and } \delta_n = \frac{c - \mathbb{P}_{\theta_2}(T(\boldsymbol{X}) \leq D_n - 1)}{\mathbb{P}_{\theta_2}(T(\boldsymbol{X}) = D_n)}.$$

We can use the p-value defined in Equation (2) with models possessing monotone likelihood ratio (MLR), for example, any one-dimensional exponential family and the location family of folded normal distribution. For continuous models, the critical constants C_n and D_n are slightly modified, for example, by introducing the variance in the case of a normal distribution. Moreover, the randomization constants in (3) are such that $\gamma_n = \delta_n = 0$ for such continuous models. Next, we give a lemma whose proof is in the Appendix to show that the UMP p-value in Definition (1) is the maximum of the p-values for a lower- and an upper-tailed test.

Lemma 1 *For a fixed but arbitrary significance level $t \in (0, 1)$ and a chosen true parameter under the null hypothesis $\theta_0 = \theta_1$ or $\theta_0 = \theta_2$, the UMP p-value in Equation (2) is the maximum of the p-values for a lower- and an upper-tailed test.*

In calculating the UMP p-value in (2), using either θ_1 or θ_2 leads to the same result for the p-value. The UMP p-value is used in the first stage of randomization to deal with the discreteness of the test statistics. We conduct a second randomization to deal with the composite null hypothesis. The second stage randomized p-value (RAND2) (cf. (Hoang and Dickhaus 2022)) is defined as follows.

Definition 2 (Second stage randomization) Let U and $\tilde{U}$ be two different UNI(0, 1)-distributed random variables both stochastically independent of the data $\boldsymbol{X}$ and are also independent of each other. Assume also that we have an arbitrary constant $c \in [0, 1]$. The two-stage randomized p-value $P^{rand2}(\boldsymbol{X}, U, \tilde{U}, c)$ is

$$P^{rand2}(\boldsymbol{X}, U, \tilde{U}, c) = \tilde{U}\mathbf{1}\{P^{UMP}(\boldsymbol{X}, U) \geq c\} + P^{UMP}(\boldsymbol{X}, U)(c)^{-1}\mathbf{1}\{P^{UMP}(\boldsymbol{X}, U) < c\}. \tag{4}$$

where $P^{UMP}(X, U)$ is the UMP p-value in the first stage as defined in Equation (2). Furthermore, we define $P^{rand2}(\boldsymbol{X}, U, \tilde{U}, 0) = \tilde{U}$ and $P^{rand2}(\boldsymbol{X}, U, \tilde{U}, 1) = P^{UMP}(\boldsymbol{X})$. The CDF of $P^{rand2}(\boldsymbol{X}, U, \tilde{U}, c)$ is

$$\mathbb{P}_\theta\{P^{rand2}(\boldsymbol{X}, U, \tilde{U}, c) \leq t\} = t\mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}, U) > c\} + \mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}, U) \leq tc\}. \tag{5}$$

With our p-values so defined, we are now ready to use them to test our hypothesis. We first describe an example of a discrete model that we use to illustrate our randomized p-values in practice.

Example 1 (Binomial distribution) Assume that our (random) data is given by $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, where each X_r is a real-valued, observable random variable, $1 \leq r \leq n$, and all X_r are stochastically independent and identically distributed (i.i.d.) Bernoulli variables with parameter $\theta_i \in (0, 1)$ for $i = 1, 2$, $Bernouli(\theta_i)$ for short. A sufficient test statistic for testing the hypothesis in (1) is $T(\boldsymbol{X}) = \sum_{r=1}^{n} X_r$ which is distributed as a Binomial random variable with parameters n and θ_i, $i = 1, 2$ and we shall denote this by $Bin(n, \theta_i)$. The respective p-values with their CDFs are calculated using Equations (2), (3), (4), and (5). The critical constants C_n and D_n are

given by $C_n = F^{-1}_{Bin(n,\theta_1)}(1-t)$ and $D_n = F^{-1}_{Bin(n,\theta_2)}(t)$ for a fixed significance level $t \in (0, 1)$ where $F^{-1}(\bullet)$ denotes the quantile of a binomial random variable with parameters n and θ. The randomization constants γ_n and δ_n for large sample sizes and for arbitrary $t \in (0, 1)$ are given by $\gamma_n = \{F_{Bin(n,\theta_1)}(C_n) - (1-t)\}\{f_{Bin(n,\theta_1)}(C_n)\}^{-1}$ and $\delta_n = \{t - F_{Bin(n,\theta_2)}(D_n - 1)\}\{f_{Bin(n,\theta_2)}(D_n)\}^{-1}$, where $F_{Bin(n,\theta)}$ denotes the CDF and $f_{Bin(n,\theta)}$ the probability mass function of binomial variable with parameters n and θ.

In this section, as mentioned before, we consider the individual test problem where $k = 1$. We are interested in finding if randomization is beneficial when the equivalence limit Δ increases or decreases and if the power functions for the p-values are monotonic in sample size. Furthermore, we seek to find if the level of conservativity of the p-values depends on the sample sizes.

3.2 Sample size versus power

We expect that the power function for a test would be strictly increasing with an increase in sample size. A power function that is strictly increasing with an increase in the sample size is ideal for sample size planning since an additional observation cannot lower the power. In the case of discrete models, Finner and Strassburger (2001) showed that it is possible for the power of the (least favorable configuration) LFC-based p-value at a sample of size n to coincide over the entire parameter space with that of size $n + i$, for small $i \in \mathbb{N}$. We illustrate in the second panel of Fig. 1 and for the model in Example (1) that this paradoxical behavior can also occur for the UMP p-value and cannot be corrected even by use of randomization. The problem occurs for small samples with the chosen true parameter θ too close to the boundary of the alternative hypothesis. To generate Fig. 1, we set the tuning parameter $c = 0.5$, $t = 0.05$, $\theta_1 = 0.25$, and $\theta_2 = 0.75$ in both panels. Furthermore, we choose $\theta = 0.5$ as the true parameter under the alternative hypothesis in the left panel and $\theta = 0.4$ in the right.

On the left panel in Fig. 1, both power functions are strictly increasing with an increase in the sample size. On the right panel, both power functions are not strictly increasing with an increasing sample size. We further illustrate in Fig. 2 that this paradoxical behavior of the power function of the UMP p-value in the right panel of Fig. 1 does not occur for small equivalence limit Δ. To generate Fig. 2, we maintain the parameter settings as in the right panel of Fig. 1 but only change θ_1 to 0.35 so that the resulting Δ is decreased compared to the initial one.

From Fig. 2, the power functions for the UMP and RAND2 p-values are now strictly increasing with an increase in the sample size for most n. The problem of the power function failing to be strictly increasing with an increase in the sample size is partially dealt with, though not completely removed. Shrinking Δ from both sides, however, worsens the problem in the right panel of Fig. 1. Finally, we provide Theorem (1) with a proof in the appendix to further justify the claims in the left panel of Figure (1).

Theorem 1 *(Monotonicity of the power functions) The CDFs of the UMP and RAND2 p-values are strictly increasing with an increase in the sample size n for any fixed*

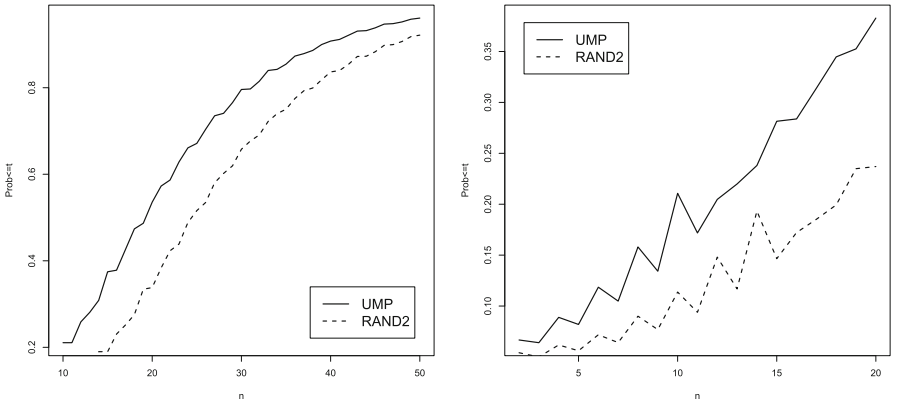

Fig. 1 The power function for the UMP and RAND2 p-values against different sample sizes for $c = 0.5$, $t = 0.05$, $\theta_1 = 0.25$, and $\theta_2 = 0.75$. Furthermore, we set $\theta = 0.5$ in the left panel and $\theta = 0.4$ in the right

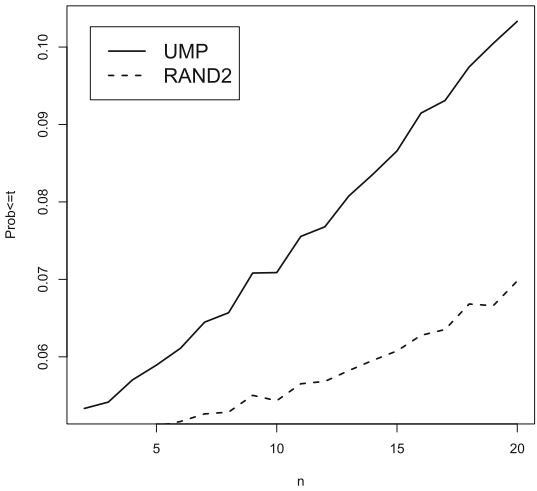

Fig. 2 The power function for the UMP and RAND2 p-values against different sample sizes for $c = 0.5$, $\theta_1 = 0.35$, and $\theta_2 = 0.75$. We maintain the true parameter under the alternative hypothesis at $\theta = 0.4$

parameter value θ under the alternative hypothesis. Consequently, for any significance level and a fixed parameter value θ under the alternative hypothesis, the power of the corresponding test is monotonically increasing with an increase in the sample size n.

3.3 Conservativity of the *p*-values

We expect that the distribution of a p-value under the null hypothesis is close to that of a $UNI(0, 1)$ distribution. A p-value can fail to meet this requirement and hence be conservative, meaning it is stochastically greater than the Uniform distribution. We compare the CDFs of RAND2 and the UMP p-values under the null and alternative hypothesis for two equivalence limits Δ_1 and Δ_2 such that $\Delta_1 < \Delta_2$ and sample sizes n_1 and n_2 such that $n_1 < n_2$. We plot Fig. 3 to compare the conservativity and

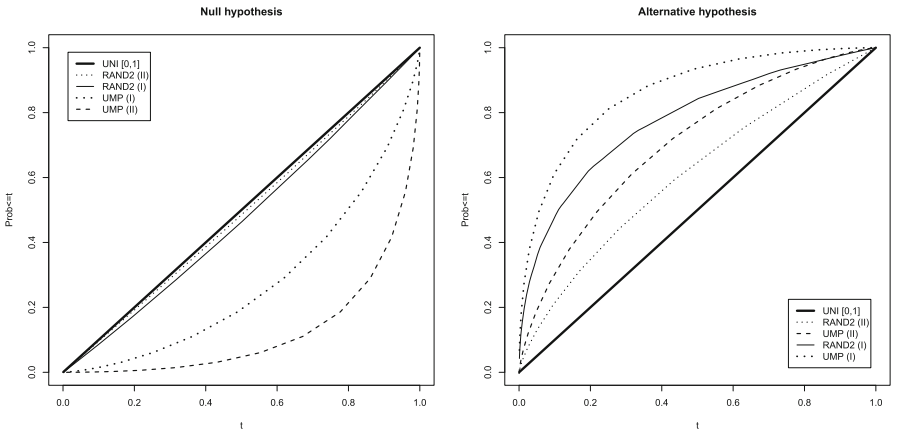

Fig. 3 The CDFs of the UMP and RAND2 p-values against t for $n = 50$ and $c = 0.5$. We choose the true parameter $\theta = 0.2$ under the null hypothesis and $\theta = 0.35$ under the alternative hypothesis. Furthermore, we set $\theta_1 = 0.25$, and $\theta_2 = 0.75$ in the first case (I) and $\theta_1 = 0.3$, and $\theta_2 = 0.75$ in the second case (II)

power functions of the two p-values for the model in Example 1 using two different equivalence limits Δ.

From Fig. 3, the CDF of the UMP p-value under the null hypothesis is far from the $UNI(0, 1)$ line compared to the one for RAND2 p-value in both cases. Therefore, the UMP p-value is more conservative than RAND2 p-value. Under the alternative hypothesis, the CDF of the UMP p-value is also far from the $UNI(0, 1)$ line compared to the one for RAND2 p-value in both cases. Therefore, as expected, the power of the UMP p-value exceeds that of RAND2 p-value.

Under the same parameter configurations and only shrinking the equivalence limit Δ, the UMP p-value becomes less powerful and more conservative. The two-stage randomized p-value also becomes less powerful, but the conservativeness of the p-value reduces even further. Next, we give Fig. 4 to illustrate the behavior of the CDFs for the two p-values under the null hypothesis using the same parameter configurations as in Fig. 3 except that the sample size n is not constant. Again, we consider two cases but with $n = 50$ in the first case (I) and $n = 100$ in the second case (II).

From Fig. 4, the CDF of the UMP p-value moves away while the one for RAND2 p-value moves closer to the $UNI(0, 1)$ line as the sample size increases. Therefore, the UMP p value becomes more conservative while the RAND2 p value becomes less conservative as the sample size increases.

3.4 Maximum power

In this section, we are interested in finding a parameter value θ_{max} under the alternative hypothesis that maximizes the power of a test based on our p-values. If such a parameter exists, choosing it as our true parameter under the alternative hypothesis will always guarantee that we have the maximum power. Also, we investigate if the value of θ_{max} is affected by the size of the equivalence limit Δ. We generate Fig. 5 to

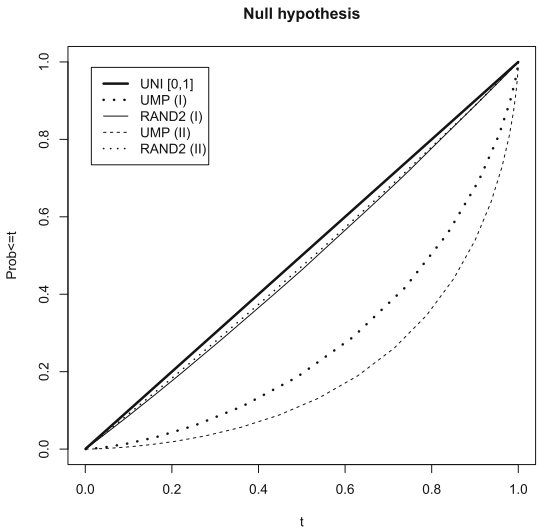

Fig. 4 The CDFs of the UMP and RAND2 p-values against different values of t with $c = 0.5$, $\theta_1 = 0.25$, $\theta_2 = 0.75$, and with $\theta = 0.2$ as the chosen parameter under the null hypothesis. Furthermore, we use $n = 50$ in the first case (I) and $n = 100$ in the second case (II)

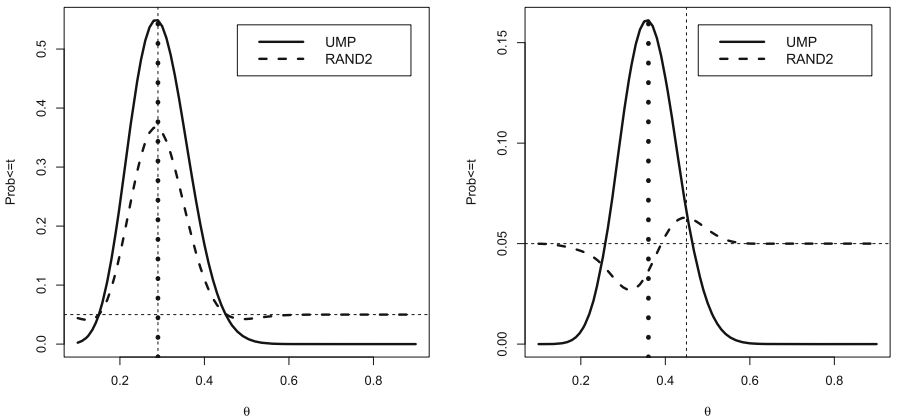

Fig. 5 The CDFs for the UMP and RAND2 p-values against the chosen parameter θ for $c = 0.5$ and $n = 50$. We set $\theta_1 = 0.15$ and $\theta_2 = 0.45$ in the left panel and $\theta_1 = 0.25$ and $\theta_2 = 0.45$ in the right one. The vertical lines intersect the respective CDF curves at their maximum and the x axis at the parameter value θ_{max} that maximizes those CDFs. The bold vertical line is for the UMP p-value while the thin dotted line is for RAND2 p-value. Furthermore, the thin dotted horizontal lines intersect the y axis at the value of α

address these questions for Example 1, where we have set $c = 0.5$ and used $n = 50$ as our sample size. Furthermore, we use $\theta_1 = 0.15$ and $\theta_2 = 0.45$ in the left panel of Fig. 5 and $\theta_1 = 0.25$ and $\theta_2 = 0.45$ in the right one.

From Fig. 5 and for a large equivalence limit Δ, the parameter θ_{max} for the two p-values always occur at the midpoint of the interval Δ. For a small Δ, the parameter θ_{max} for RAND2 p-value can occur at a point too close to θ_1 or θ_2. The one for the UMP p-value occurs at the midpoint throughout, and it does not matter how small Δ becomes. Also, for both p-values, only a single θ_{max} exists under the alternative hypothesis. Moreover, the behavior of the CDFs in the right panel further confirms that

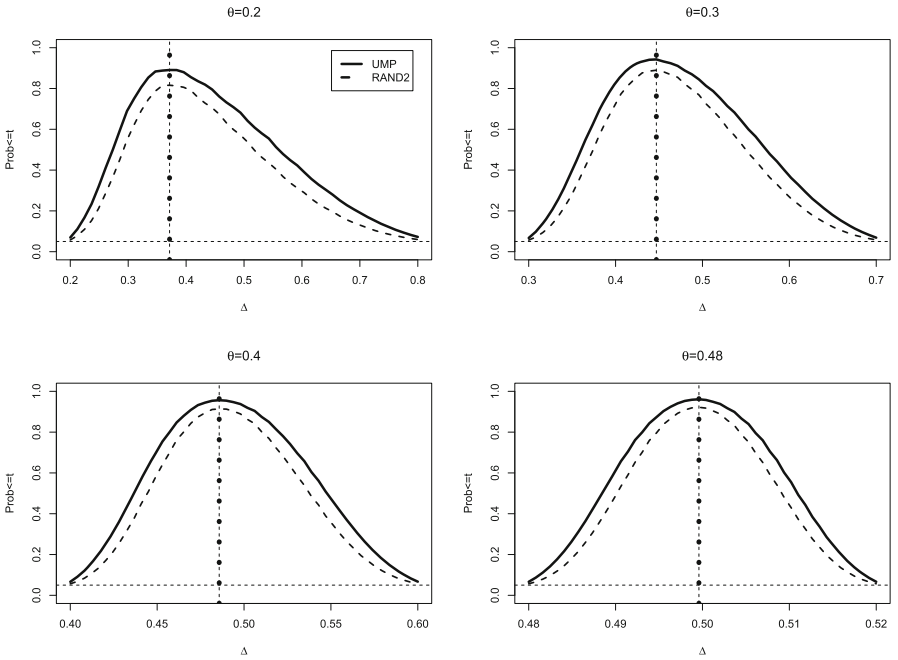

Fig. 6 The CDF under the alternative hypothesis for the UMP and RAND2 p-values against the equivalence limit Δ for $c = 0.5$ and $n = 50$. The chosen parameters under the alternative hypothesis are $\theta = 0.2, 0.3, 0.4$, and 0.48, respectively, from left to right. The vertical lines intersect the respective CDF curves at their maximum and the x axis at the Δ value, which maximizes those CDFs. The bold vertical line is for the UMP p-value while the thin dotted line is for RAND2 p-value. Furthermore, the thin dotted horizontal lines intersect the y axis at the value of α

RAND2 p-value, unlike the UMP p-value, is not unbiased. To conclude this section, we give a figure illustrating the power for the two p-values against the equivalence limit Δ. To generate Figure (6), we set $c = 0.5, n = 50$, and choose $\theta = 0.2, 0.3, 0.4$, and 0.48 as the true parameters under the alternative hypothesis. Moreover, we use different values of θ_1 and θ_2 to get different equivalence limits since $\Delta = \theta_2 - \theta_1$.

From Fig. 6, as is expected, the range of Δ in each panel is from the chosen parameter value θ under the alternative hypothesis to $1 - \theta$. For example, in the first panel, the parameter is $\theta = 0.2$ and Δ ranges from $\theta = 0.2$ to $1 - \theta = 0.8$. The value of θ_{max} moves closer to 0.5 as the chosen parameter under the alternative hypothesis moves closer to 0.5.

4 Estimation of the proportion of true null hypotheses

4.1 Introduction

In this section, we extend our discussions from Sect. 3 to the case when $k > 1$ hypotheses are of interest. According to Section 4.3 of Wellek (2010), it is possible to have a

case of univariate equivalence tests for a parameter of interest. Comparison of a single proportion to a fixed reference success probability was the subject of Sect. 3. We extend this idea and compare multiple proportions to a fixed reference success probability. We do this to identify the proportion of true null hypotheses (an estimation problem) and not which particular null hypotheses are true (a selection problem).

Recall that for the multiple testing problem, our hypothesis in Eq. (1) can be expressed as

$$H_j : \theta_j \notin \Delta_j \quad \text{versus} \quad K_j : \theta_j \in \Delta_j, \text{ for } j = 1, \ldots, k,$$

where Δ_j denotes the range of values in the j^{th} interval between $\theta_1^{(j)}$ and $\theta_2^{(j)}$ for $j \in \{1, \ldots, k\}$, k is the multiplicity of the problem, and $p_1, \ldots, p_k$ are the resulting k p-values. For example, assume we have a data set from $k = 1{,}000$ small regions of a country showing the number of individuals suffering from a certain disease. We are interested in testing the hypotheses that the proportion of infected individuals from all the regions lie in a certain interval $[\theta_1, \theta_2]$ when the equivalence limit is constant. We do this to find if the infection rate is at dangerously high levels in a particular region.

Conducting these hypotheses at level α increases the probability of type I errors since we do not account for the multiplicity of the problem. It is crucial to account for this multiplicity by doing, for example, a familywise error rate (FWER) control. One commonly used method for familywise error control at level α is the Bonferroni adjustment (cf. (Bonferroni 1936)). The Bonferroni procedure adjusts the raw p-values $p_1, \ldots, p_k$ by multiplying them by the number of hypotheses k. We reject the null hypothesis if an adjusted p-value is less than or equal to α. The Bonferroni procedure guarantees that the FWER is at most α regardless of the ordering or the dependence structure of the p-values.

The Bonferroni procedure can be conservative when large proportions of null hypotheses are false. The adjustment also maintains FWER at levels below $\pi_0\alpha$ instead of α where $\pi_0 = k_0/k$ is the proportion of true null hypotheses. When the number of true null hypotheses $k_0 < k$, the individual tests are conducted at a higher level α/k_0 instead of α/k, leading to a higher power for the testing procedure. We refer to this as the adaptive Bonferroni procedure, ABON for short. Since in practice we never really know the number (proportion) k_0 (π_0), we make use of ABON combined with the plug-in (ABON+plug-in) procedure of Finner and Gontscharuk (2009) to estimate π_0. The ABON+plug-in procedure, unlike closed testing procedures (like (Hommel 1988) and Hochberg (1988)), provides a theoretical guarantee to control the type I error rate at the desired level.

One classical but still commonly used estimator for k_0 is the Schweder and Spjøtvoll (1982) estimator. It is given by

$$\hat{k}_0 \equiv \hat{k}_0(\lambda) = k \cdot \frac{1 - \hat{F}_k(\lambda)}{1 - \lambda}, \tag{6}$$

where $\lambda \in [0, 1)$ is a tuning parameter and $\hat{F}_k$ is the empirical CDF (ecdf) of the k marginal p-values. It is often suggested in practice to choose $\lambda = 0.5$. One crucial pre-

requisite for the applicability of this estimator is that the marginal p-values $p_1, \ldots, p_k$ are (approximately) uniformly distributed on $(0, 1)$ under the null hypothesis; see, e. g., Dickhaus (2013); Hoang and Dickhaus (2022) and the references therein for details. The randomized p-values considered in this work are close to meeting the uniformity assumption, whereas the non-randomized p-values are over-conservative when testing two one-sided composite null hypotheses, especially in discrete models. Typically, the estimated value of k_0 becomes too large if many null p-values are conservative and the estimator from (6) is employed.

4.2 Empirical distributions

To illustrate the implication of using our proposed two-stage randomized p-value in multiple testing, we employ a graphical algorithm in computing π_0. This algorithm connects the points $(\lambda, \hat{F}_k(\lambda))$, $\lambda \in [0, 1)$ with the point $(1, 1)$. We draw a straight line to connect the two points and extend this line to intersect the y axis at the point $1 - \hat{\pi}_0$. The best p-value for use in the estimation of π_0 is that for which the resulting straight line meets the y axis at a point that is close to the actual $1 - \pi_0$. We require the empirical CDF of the p-value not to lie below the $UNI(0, 1)$ line for this to be actualized. We summarize our steps in Algorithm 1 below.

Algorithm 1 Graphical procedure for the estimation of the proportion of true null hypotheses

(1) Choose a tuning parameter $\lambda \in (0, 1)$.
(2) Compute and plot the empirical CDFs $F_k(t)$ of the p-values where $F_k(t) := \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{I}_{p_j \leq t}$ and $t \in (0, 1)$.
(3) Draw a vertical line (the dashed line in Figure 7) that intersects the x axis at the point λ. The intersection of this vertical line and the empirical CDFs of the p-values gives the point $\hat{F}_k(\lambda)$.
(4) Draw a diagonal line from the point $(1, 1)$ to the y-axis through the point $(\lambda, \hat{F}_k(\lambda))$. The intersection of this diagonal line with the y-axis gives $1 - \hat{\pi}_0$.

To generate Fig. 7, we let the number of hypotheses to be $k = 1{,}000$, the tuning parameters c and λ are both set at 0.5 and use a sample of size $n = 50$. We take the proportion of true null hypotheses to be $\pi_0 = 0.7$ and set $\theta_1 = 0.25$ and $\theta_2 = 0.75$. Furthermore, to calculate the UMP-based p-value, the parameter θ under the null and alternative hypothesis are chosen as 0.18 and 0.37, respectively.

From Fig. 7, RAND2 p-value outperforms the UMP p-value since its ECDF lies above the $UNI(0, 1)$ line for all values of $t \in (0, 1)$. Furthermore, an extension of a straight line from the points $(1, 1)$ to (λ, F_k^{RAND2}) as earlier mentioned, meets the y axis at a point which is close to $1 - \pi_0$.

4.3 Simulation study

We now conduct a simulation study based on real-world data to support the claim in Sect. 4.2 that RAND2 p-value outperforms the UMP p-value in estimating the

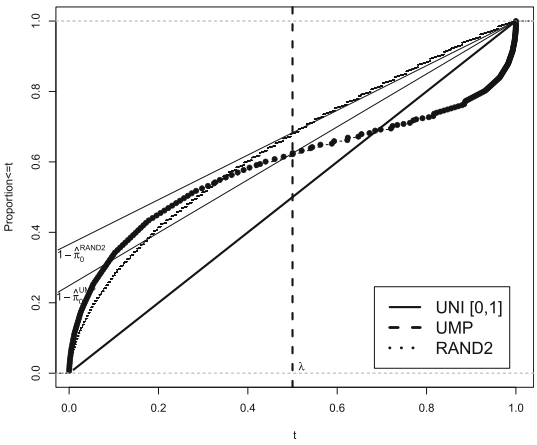

Fig. 7 Empirical CDF of the UMP p-value (black curve) and the two-stage randomized (RAND2) p-value (grey curve) for $k = 1{,}000$, $\lambda = 0.5$, $c = 0.5$, and $\pi_0 = 0.7$. We set $\theta_1 = 0.25$ and $\theta_2 = 0.75$. Furthermore, we choose the true parameter under the null as $\theta = 0.18$ and otherwise $\theta = 0.37$. The dashed vertical line intersects the x axis at λ while the thin diagonal lines intersect the y-axis at $1 - \hat{\pi}_0$

proportion of true null hypotheses in multiple testing. We use the publicly available Coronavirus Disease 2019 (COVID-19) data taken from https://github.com/CSSEGISandData/COVID-19 (cf. (Dong et al 2020)). The data set consists of confirmed COVID-19 cases and recoveries for $k = 58$ regions of the United States of America as of 12^{th} May 2020. The regions include all fifty states and eight others: American Samoa, Diamond Princess, District of Columbia, Grand Princess, Guam, Northern Mariana Islands, Puerto Rico, and the Virgin Islands.

After cleaning the data by removing all the missing values, we have $k = 47$ regions for our analysis. We select an interval of recovery rates θ_1 and θ_2 and conduct a TOST to find if the true rates from the data set belong to these intervals. We use a Monte Carlo simulation to assess the (average) performance of the UMP and RAND2 p-values in estimating k_0. We set the constant c and the tuning parameter λ in (6) to 0.5 for all the simulations. The recovery rates from the data set are assumed to be the true proportions. The recovery rates are defined as $\theta_i = r_i/n_i$ where r_i are the number of recoveries out of the n_i infected individuals from the i^{th} region for $i \in \{1, \ldots, 47\}$.

Using these rates and the number of confirmed cases, we generate a new data set on the computer for calculating the p-values. For simulation purposes, we choose different values for θ_1 and θ_2 in our null hypothesis leading to different equivalence limits Δ. For each equivalence limit Δ, define k_0 as the number of true proportions, that is, k_0 is the number of proportions $\theta_i, i \in \{1, \ldots, k\}$ such that $\theta_i \leq \theta_1$ or $\theta_i \geq \theta_2$. For example, with the $k = 47$ regions, k_0 can take any random value between 0 and 47. Since we are utilizing randomized p-values in (6), we average the estimated value of k_0 over the 10,000 Monte Carlo repetitions.

For exemplary purposes, we present ten choices of θ_1 and θ_2 in Table 1. A detailed description of the simulation is provided in Algorithm 2. The results from our simulation study based on the Algorithm 2 are presented in Table 1. From Table 1, for whatever value of Δ, RAND2 p-value outperforms the UMP p-value by giving estimates which are on average close to the actual value of k_0. Also, as is expected, the number of true null hypotheses k_0 decreases with an increase in the interval Δ.

Algorithm 2 Computation of the proportion of true null hypotheses

(1) For each of the $k = 47$ regions in the COVID-19 data set, find the proportions of recoveries θ_i, $i = \{1, \ldots, 47\}$ and use these as the assumed true proportions (i. e., as the assumed ground truth) in the simulations.
(2) For each θ_i from step 1.) and for each of the sample sizes n_i for $i = \{1, \ldots, 47\}$, simulate a single data point x_i from $Bin(n_i, \theta_i)$.
(3) Select two proportions $\theta_1, \theta_2 \in [0, 1]$ such that $\theta_1 < \theta_2$ as the null values to be tested against. Take k_0 as the number of values $i \in \{1, \ldots, k\}$ fulfilling that $\theta_i \leq \theta_1$ or $\theta_i \geq \theta_2$. We use the selected θ_1, θ_2 as well as the numbers x_i, and n_i from step 2.) in the computation of the p-values, where $i \in \{1, \ldots, k\}$. This step generates a pair of p-values for the UMP p-value since we decompose the null hypothesis in (1) into a lower- and an upper-sided test. Denote these p-values by p_l and p_u, respectively. In each case, pick max (p_l, p_u) which is the maximum of the two p-values.
(4) Compute the statistic in Equation (6) $r = 10{,}000$ times for the UMP and RAND2 p-values and take the mean.

Table 1 Estimates of the number of true null hypotheses

θ_1	θ_2	Δ	k_0	$\hat{k}_0^{UMP}$	$\hat{k}_0^{RAND2}$
0.4791	0.5413	0.0622	45	90.0050	44.3586
0.4509	0.5681	0.1173	43	86.0006	43.1554
0.4444	0.5946	0.1502	40	80.0034	39.4800
0.4066	0.6800	0.2734	34	67.9996	33.5392
0.3389	0.7219	0.3830	31	60.0002	33.7460
0.3188	0.7478	0.4290	29	55.9958	28.2846
0.3076	0.7566	0.4491	28	55.9958	28.6418
0.2963	0.9029	0.6065	16	32.0070	15.2562
0.2725	0.9319	0.6594	12	26.0192	12.9908
0.2456	0.9399	0.6942	12	24.6468	13.9496

4.4 Role of the tuning parameter λ

In this section, we investigate the role of the tuning parameter λ in the estimator given in (6) when using the two p-values. Proper choice of this parameter is crucial since a smaller λ will lead to high bias and low variance while a larger one leads to low bias and high variance of the proportion estimator. Based on this bias-variance trade-off, we take the optimal λ to be the one that minimizes the mean square error (MSE).

Other research in this direction includes the use of change-point concepts for choosing λ in the Storey (2002) estimator. In this approach, first approximate the p-value plot by a piecewise linear function that has a single change-point. Select the p-value at this change-point location as the value of λ. Another approach is to choose $\lambda = \alpha$. Hoang and Dickhaus (2022) noted that the default choice of $\lambda = 0.5$ works well with randomized p-values since the sensitivity of the estimator in (6) with respect to λ is least pronounced for the case of randomized p-values. The default common choice of $\lambda = 0.5$ is unstable, especially when dealing with dependent p-values. We plot Fig. 8 to illustrate how the estimator based on the UMP and RAND2 p-value is affected by different choices of λ. In this plot, we have used the same COVID-19 data and set $c = 0.5$, $\theta_1 = 0.2963$, and $\theta_2 = 0.7566$.

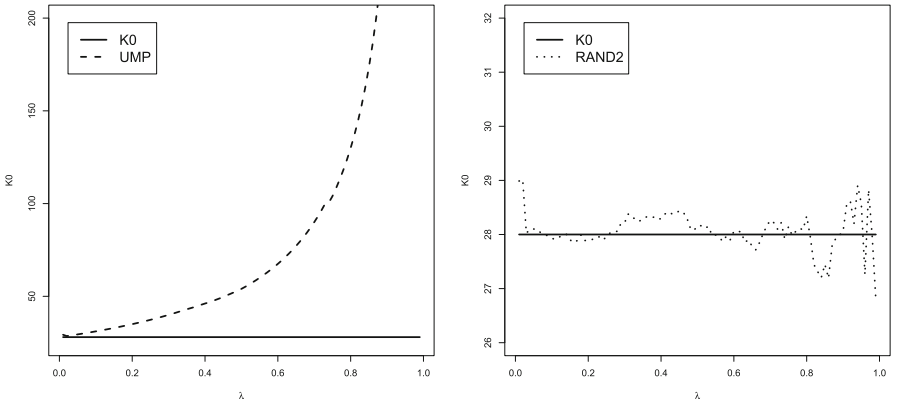

Fig. 8 An illustration of the number of true null hypotheses versus λ for the UMP and the two-stage randomized (RAND2) p-values for $\theta_1 = 0.2963$, $\theta_2 = 0.7566$, and $c = 0.5$

From Fig. 8, the estimate of k_0 based on the UMP p-value moves away from the number of true null hypotheses k_0 as the value of λ increases. The estimate based on RAND2 p-value stays close to k_0 and only oscillates wildly around k_0 when λ is close to 1.

5 Discussion

In this research, we have considered the use of UMP and randomized p-values (RAND2) in the problem of interval composite null hypothesis. Using large sample sizes, we have illustrated that the power functions for the UMP and RAND2 p-values are both monotonically increasing with an increase in the sample size. We have also found that it is possible for the power function of the UMP and the two-stage randomized p-value for a sample of size n and that of $n + i$ for small $i \in \mathbb{N}$ to coincide on the entire parameter space. This problem occurs when dealing with relatively small samples and the equivalence limit Δ is too wide while the chosen parameter θ under the alternative is too far from θ_1 or θ_2. This problem does not occur when Δ is too narrow while the chosen parameter θ is too close to θ_1 or θ_2 under the alternative hypothesis (see Fig. 2). This problem only occurs if the interval Δ gets smaller from one end while the other one is kept constant, for example, by holding θ_2 constant and increasing θ_1.

The problem of nonmonotonicity of the power functions gets worse if the equivalence limit decreases from both ends. A similar observation in Qiu and Cui (2010) is that when the equivalence limit is too narrow, the ROC curve of the TOST procedure is nonmonotonic for small sample sizes. A complete characterization of this paradox will be considered in future research following the ideas in Finner and Strassburger (2001) and Finner and Roters (1993). Of course, in practical problems, the equivalence limits are determined before the data collection and remain fixed throughout the

experiment. The adjustments made here are to illustrate the behavior of the p-values and their CDFs under different equivalence limits.

A plot of the CDFs for the UMP and RAND2 p-values under the null and alternative hypothesis illustrates that the UMP p-value is more conservative and more powerful compared to RAND2 p-value. The conservativeness of the UMP p-value increases while the one for RAND2 reduces with a further decrease in the equivalence limit Δ. Furthermore, the power functions for the p-values decrease with a decrease in Δ. A similar trend for the CDFs occurs when Δ is kept constant, and the chosen parameter under the null (alternative) is too far from (close to) the boundary of the null (alternative).

Increasing both the parameters θ_1 and θ_2 by ϵ_1 and θ by ϵ_2 leads to an increase in power for both the p-values, a decrease in conservativity of the UMP p-value, and no change in the level of conservativity of RAND2 p-value. A similar trend occurs for a large equivalence limit, provided $\epsilon_2 > \epsilon_1$. The power increases for a large equivalence limit since θ moves closer to the midpoint of Δ, which is the parameter that gives the maximum power for both UMP and RAND2 p-values under this condition. We were also interested in finding the parameter value that maximizes the CDFs under the alternative hypothesis for the two p-values. We found that for large Δ, the parameter value that maximizes the CDFs of both p-values occurs at the midpoint of Δ. For small Δ, however, this parameter value can be too close to θ_1 or θ_2 for RAND2 p-value while the one for the UMP p-value is always at the midpoint.

Concerning the level of conservativity of the p-values to the sample size, we found that the CDF for the UMP p-value moves further away while the one for RAND2 p-value moves closer to the $UNI(0, 1)$ line with an increase in the sample size. Therefore, the UMP p-value becomes more conservative while the level of conservativity for RAND2 p-value remains the same with an increase in the sample size. Furthermore, Munk (1996) and Wellek (2010) Sect. 1.2 (p. 5) argues that equivalence tests require much larger sample sizes to achieve a reasonable power compared to the one- or two-sided tests; unless Δ is chosen too wide that even "nonequivalent" hypotheses would be declared "equivalent." Therefore, it would be better to consider RAND2 p-value for multiple equivalence tests since they are less conservative even for large sample sizes.

A plot of the empirical CDFs of the p-values evaluates their performance when used with the estimator in (6). The ECDF of RAND2 p-value, unlike the one for the UMP p-value, is above the $UNI(0, 1)$ throughout. Furthermore, the slope between the points $(1, 1)$ and (λ, F_k^{RAND2}) for RAND2 p-value is close to π_0 compared to the one for the UMP p-value. Therefore, RAND2 p-value outperforms the UMP p-value in the estimation of the proportion of true null hypotheses. To further justify this claim, we have given a real example and provided a simulation study showing that RAND2 p-value outperforms the UMP p-value for all values of Δ by giving estimates that are closer on average to the true proportions.

The choice of the tuning parameter λ for the estimator in (6) has also been of great concern in the recent literature. The sensitivity of this estimator to λ is more pronounced for conservative p-values than for non-conservative ones. Since the UMP p-value is more conservative than RAND2 p-value, the choice of λ is critical for obtaining estimates that are close to the actual number of true null hypotheses when

using the UMP p-value. Based on the results from our simulation study, we recommend a small value of λ when using the UMP p-value. Assuming we are using a small α, this choice is similar to the recommended choice of $\lambda = \alpha$ in the previous literature. When using RAND2 p-value, any choice of λ which is not close to one is recommended. We recommend this choice since the estimate of k_0 based on RAND2 p-value oscillates wildly around the value of k_0 as $\lambda \longrightarrow 1$. The recommendation in Dickhaus (2013); Habiger and Pena (2011), and Habiger (2015) that randomized p-values are nonsensical for a single hypothesis also applies to our RAND2 p-value and in that case the usage of the UMP p-value is advocated for. Furthermore, we caution the practitioner against using randomized p-values in bioequivalence studies. Some general extensions of this research include using randomized test procedures to achieve unbiased tests for Lehmann's alternative. Also, one could extend these procedures to consider multiple endpoints while accounting for the correlations among those endpoints. Finally, randomized p-values can be extended to stepwise regression since we use the p-values in these procedures several times, leading to multiple test problems.

Acknowledgements The author would like to acknowledge the editor and an anonymous reviewer for the careful reading of this manuscript and for their comments and suggestions, which improved the presentation of this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix A Appendix: Mathematical proofs

Proof of Lemma 1 Recall that our (random) data is given by $\boldsymbol{X}$, U is a $UNI(0, 1)$-distributed random variable which is independent of our data, and $t \in (0, 1)$ is an arbitrary significance level. Define $\phi(X, U; t)$ to be a decision function for a test procedure such that we reject the null when $\phi(X, U; t) = 1$ and otherwise fail to reject when $\phi(X, U; t) = 0$. A p-value based on this decision function is

$$P(X, U) = \inf\{t \in (0, 1) : \phi(X, U; t) = 1\}. \tag{A1}$$

Consider a test of $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$ where θ_0 is a prespecified constant. The size of this test for an arbitrary $t \in (0, 1)$ is

$$E_{\theta_0}\{\phi_u(X, U; t)\} = \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) > C_n) + \gamma_n \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = C_n) = t, \tag{A2}$$

where $C_n = F^{-1}_{\theta_0}(1-t)$ and $\gamma_n = \{\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) \leq C_n) - (1-t)\}\{\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = C_n)\}^{-1}$ are the critical and randomization constants, respectively as given in Definition 1. The p-value for this test based on the definition in Equation (A1) is

$$\begin{aligned} P_u(X,U) &= \inf\{t \in (0,1) : t \geq \mathbb{P}_{\theta_0}(T(\boldsymbol{x}) > c_n) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = c_n)\}, \\ &= \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) > C_n) + U\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = C_n). \end{aligned}$$

Similarly, consider a test of the form $H : \theta \geq \theta_0$ versus $K : \theta < \theta_0$ where again θ_0 is a prespecified constant. The size of this test for an arbitrary $t \in (0,1)$ is

$$E_{\theta_0}\{\phi_l(X,U;t)\} = \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) \leq D_n - 1) + \delta_n \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = D_n) = t, \quad \text{(A3)}$$

where again $D_n = F^{-1}_{\theta_0}(t)$ and $\delta_n = \{t - \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) \leq D_n - 1)\}\{\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = D_n)\}^{-1}$ are the critical and randomization constants, respectively as given in Definition 1. The p-value for this test based on the definition in Equation (A1) is

$$\begin{aligned} P_l(X,U) &= \inf\{t \in (0,1) : t \geq \mathbb{P}_{\theta_0}(T(\boldsymbol{x}) \leq d_n - 1) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = d_n)\}, \\ &= \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) \leq D_n - 1) + U\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = D_n). \end{aligned}$$

Assume now that the hypothesis is as given in Equation (1), then the overall test statistic for this problem is

$$\begin{aligned} \phi_m(X,U;t) &= \min\{\phi_l(X,U;t), \phi_u(X,U;t)\}, \\ &= \phi_l(X,U;t) \bigcap \phi_u(X,U;t). \end{aligned}$$

The overall p-value is

$$P_m(X,U) = \max\{P_l(X,U), P_u(X,U)\}, \quad \text{(A4)}$$

since

$$\begin{aligned} \{t : \phi_m(X,U;t) = 1\} &= \{t : \min[\phi_l(X,U;t), \phi_u(X,U;t)] = 1\}, \\ &= \{[t : \phi_l(X,U;t) = 1] \bigcap [t : \phi_u(X,U;t) = 1]\}, \\ &= \{t : t \geq \mathbb{P}_{\theta_0}(T(\boldsymbol{x}) \leq d_n - 1) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = d_n)\} \\ &\quad \bigcap \{t : t \geq \mathbb{P}_{\theta_0}(T(\boldsymbol{x}) > c_n) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = c_n)\}, \\ &= \{t : t \geq P_l(X,U)\} \bigcap \{t : t \geq P_u(X,U)\}, \\ &= \{t : t \geq \max[P_l(X,U), P_u(X,U)]\}, \end{aligned}$$

which gives the overall p-value in (A4) using the definition in (A1). We can express this further as

$$\{t : \phi_m(X,U;t) = 1\}$$

$$= \{t : t \geq \mathbb{P}_{\theta_0}(c_n < T(\boldsymbol{x}) < d_n) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = c_n) + u\mathbb{P}_{\theta_0}(T(\boldsymbol{x}) = d_2)\},$$

which gives the p-value

$$P_m(X, U) = \mathbb{P}_{\theta_0}(C_n < T(\boldsymbol{X}) < D_n) + U\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = C_n) + U\mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = D_2)\},$$

again based on the definition of a p-value in Equation (A1). However, this is equivalent to the UMP p-value in Definition 1 since θ_0 is the LFC parameter, which can be θ_1 or θ_2. □

Proof of Theorem 1 To verify that the CDFs of the UMP and the two-stage randomized p-values are point-wise monotonically increasing with an increase in the sample size for any parameter value θ belonging to the alternative hypothesis, it suffices to prove that these CDFs for a sample of size $n + 1$ are greater than those for size n. Recall from Equation (5) that

$$\begin{aligned}\mathbb{P}_\theta\{P^{rand2}(\boldsymbol{X}, U, \tilde{U}, c) \leq t\} &= t\mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}, U) > c\} + \mathbb{P}_\theta\{P_T^{UMP}(\boldsymbol{X}, U) \leq tc\},\\ &= t - t\mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}, U) \leq c\} + \mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}, U) \leq tc\}.\end{aligned}$$

For an arbitrary, but fixed $c \in [0, 1]$, further recall that $C_n = F^{-1}_{Bin(n,\theta_1)}(1 - c)$ and $D_n = F^{-1}_{Bin(n,\theta_2)}(c)$ denotes the quantile of a binomial random variable with parameters n and θ_i for $i = 1, 2$. Again recall that the randomization constants are given by

$$\gamma_n = \frac{\mathbb{P}_{\theta_1}(T(\boldsymbol{X}) \leq C_n) - (1 - c)}{\mathbb{P}_{\theta_1}(T(\boldsymbol{X}) = C_n)} \text{ and } \delta_n = \frac{c - \mathbb{P}_{\theta_2}(T(\boldsymbol{X}) \leq D_n - 1)}{\mathbb{P}_{\theta_2}(T(\boldsymbol{X}) = D_n)}.$$

Define

$$\begin{aligned}\beta_n \equiv \beta_n(c, \theta, \theta_1, \theta_2) &= \mathbb{P}_\theta\{P^{UMP}(\boldsymbol{X}) \leq c\},\\ &= \mathbb{P}_\theta(C_n < T(\boldsymbol{X}) < D_n) + \gamma_n\mathbb{P}_\theta(T(\boldsymbol{X}) = C_n) + \delta_n\mathbb{P}_\theta(T(\boldsymbol{X}) = D_n). \quad \text{(A5)}\end{aligned}$$

Since X is a random variable that follows a binomial distribution with parameters n and θ, the above power function can be expressed as

$$\beta_n = (1 + \varrho)^{-n}\left\{\sum_{x=c_n+1}^{n}\binom{n}{x}\varrho^x + \gamma_n\binom{n}{c_n}\varrho^{c_n} - \sum_{x=d_n}^{n}\binom{n}{x}\varrho^x + \delta_n\binom{n}{d_n}\varrho^{d_n}\right\}, \quad \text{(A6)}$$

where $\varrho = \left(\frac{\theta}{1-\theta}\right)$ and it is such that $\varrho \in (0, \infty)$. For a sample of size $n+1$, Equation (A6) becomes

$$\beta_{n+1} = (1 + \varrho)^{-n-1}\left\{\sum_{x=c_{n+1}+1}^{n+1}\binom{n+1}{x}\varrho^x + \gamma_{n+1}\binom{n+1}{c_{n+1}}\varrho^{c_{n+1}}\right.$$

$$- \sum_{x=d_{n+1}}^{n+1} \binom{n+1}{x} \varrho^x + \delta_{n+1} \binom{n+1}{d_{n+1}} \varrho^{d_{n+1}} \Bigg\}. \tag{A7}$$

Since $\gamma_n, \gamma_{n+1}, \delta_n, \delta_{n+1} \in (0, 1)$, to verify that $\beta_{n+1} > \beta_n$, we compare the coefficients of ϱ^x in Equations (A6) and (A7). To do this for the coefficients of ϱ^x in the first terms in Equations (A6) and (A7), we have

$$(1+\varrho)\Bigg[\sum_{x=c_n+1}^{n} \binom{n}{x} \varrho^x\Bigg] < \Bigg[\sum_{x=c_{n+1}+1}^{n+1} \binom{n+1}{x} \varrho^x\Bigg],$$

provided $c_n = c_{n+1}$. The proof for the other case when $c_n + 1 = c_{n+1}$ can be shown similarly. Next, comparing the other coefficients of ϱ^x in Equations (A6) and (A7), we have

$$(1+\varrho)\Bigg[\sum_{x=d_n}^{n} \binom{n}{x} \varrho^x\Bigg] > \Bigg[\sum_{x=d_{n+1}}^{n+1} \binom{n+1}{x} \varrho^x\Bigg],$$

provided $d_n + 1 = d_{n+1}$. Again proving the other case when $d_n = d_{n+1}$ will follow similar steps. With this, it is evident that $\beta_{n+1} > \beta_n$. The proof for $\mathbb{P}_\theta\{P_T^{rand}(\boldsymbol{X}, U) \leq tc\}$, which yields the same result, can be carried out similarly. With this, the CDF for the two-stage randomized p-value $RAND2$ is, under the stated conditions, monotonically increasing with an increase in n, which we needed to prove. Repeating the above calculations for β_n with $t \in (0, 1)$ in place of c provides the proof that the CDF of the UMP p-value is monotonically increasing with an increase in the sample size (under the stated conditions). □

References

Berger RL (1982) Multiparameter hypothesis testing and acceptance sampling. Technometrics 24(4):295–300

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. Stat Sci 1:283–302

Bonferroni C (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubbl R Ist Super Sci Econ Commer Firenze 8:3–62

Brown LD, Hwang JG, Munk A (1997) An unbiased test for the bioequivalence problem. Ann Stat 25(6):2345–2367

Dickhaus T (2013) Randomized p-values for multiple testing of composite null hypotheses. J Stat Plan Inference 143(11):1968–1979

Dickhaus T, Strassburger K, Schunk D et al (2012) How to analyze many contingency tables simultaneously in genetic association studies. Stat Appl Genet Mol Biol 11(4):12

Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20(5):533–534

Finner H, Gontscharuk V (2009) Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. J R Stat Soc Ser B Stat Methodol 71(5):1031–1048

Finner H, Roters M (1993) On the behaviour of expectations and power functions in one-parameter exponential families. Stat Risk Model 11(3):237–250

Finner H, Strassburger K (2001) Increasing sample sizes do not always increase the power of UMPU-tests for 2× 2 tables. Metrika 54(1):77–91
Fogarty CB, Small DS (2014) Equivalence testing for functional data with an application to comparing pulmonary function devices. Ann Appl Stat 1:2002–2026
Giani G, Finner H (1991) Some general results on least favorable parameter configurations with special reference to equivalence testing and the range statistic. J Stat Plan Inference 28(1):33–47
Giani G, Straßburger K (1994) Testing and selecting for equivalence with respect to a control. J Am Stat Assoc 89(425):320–329
Habiger JD (2015) Multiple test functions and adjusted p-values for test statistics with discrete distributions. J Stat Plan Inference 167:1–13
Habiger JD, Pena EA (2011) Randomised P-values and nonparametric procedures in multiple testing. J Nonparametr Stat 23(3):583–604
Hoang AT, Dickhaus T (2022) On the usage of randomized p-values in the Schweder–Spjøtvoll estimator. Ann Inst Stat Math 74(2):289–319
Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75(4):800–802
Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75(2):383–386
Huang Y, Hsu JC, Peruggia M et al (2006) Statistical selection of maintenance genes for normalization of gene expressions. Stat Appl Genet Mol Biol 5:1
Leday GG, Hemerik J, Engel J et al (2023) Improved family-wise error rate control in multiple equivalence testing. Food Chem Toxicol 178:113928
Munk A (1996) Equivalence and interval testing for Lehmann's alternative. J Am Stat Assoc 91(435):1187–1196
Ostrovski V (2020) New equivalence tests for Hardy–Weinberg equilibrium and multiple alleles. Stats 3(1):34–39
Pflüger R, Hothorn T (2002) Assessing equivalence tests with respect to their expected p-value. Biometr J 44(8):1015–1027
Qiu J, Cui X (2010) Evaluation of a statistical equivalence test applied to microarray data. J Biopharm Stat 20(2):240–266
Qiu J, Qi Y, Cui X (2014) Applying shrinkage variance estimators to the tost test in high dimensional settings. Stat Appl Genet Mol Biol 13(3):323–341
Reiss RD (1989) Approximate distributions of order statistics. With applications to nonparametric statistics. Springer Series Statistics. Springer, New York
Romano JP (2005) Optimal testing of equivalence hypotheses. Ann Stat 33(3):1036–1047
Schweder T, Spjøtvoll E (1982) Plots of P-values to evaluate many tests simultaneously. Biometrika 69:493–502
Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol 64(3):479–498
Wellek S (2010) Testing statistical hypotheses of equivalence and noninferiority. CRC Press, New York