Statistical Methods & Applications (2021) 30:1033–1052
https://doi.org/10.1007/s10260-020-00548-0

REVIEW PAPER

A unified approach to permutation testing for equivalence

Rosa Arboretti[1] · Fortunato Pesarin[2] · Luigi Salmaso[3]

Accepted: 18 October 2020 / Published online: 10 November 2020

Abstract

The notion of testing for equivalence of two treatments is widely used in clinical trials, pharmaceutical experiments, bioequivalence and quality control. It is traditionally operated within the intersection–union principle (IU). According to this principle the null hypothesis is stated as the set of effects the differences δ of which lie outside a suitable equivalence interval and the alternative as the set of δ that lie inside it. In the literature related solutions are essentially based on likelihood techniques, which in turn are rather difficult to deal with. A recently published paper goes beyond most of likelihood limitations by using the IU approach within the permutation theory. One more paper, based on Roy's union–intersection principle (UI) within the permutation theory, goes beyond some limitations of traditional two-sided tests. Such UI approach, effectively a mirror image of IU, assumes a null hypothesis where δ lies inside the equivalence interval and an alternative where it lies outside. Since testing for equivalence can rationally be analyzed by both principles but, as the two differ in terms of the mirror-like roles assigned to the hypotheses under study, they are not strictly comparable. The present paper's main goal is to look into these problems and provide a sort of comparative analysis of both by highlighting the related requirements, properties, limitations, difficulties, and pitfalls so as to get practitioners properly acquainted with their correct use in practical contexts.

Keywords Intersection–union principle · Multi-aspect testing · Nonparametric combination · Permutation tests · Testing equivalence · Two-one-sided tests (TOST) · Union–intersection principle

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10260-020-00548-0) contains supplementary material, which is available to authorized users.

✉ Rosa Arboretti
rosa.arboretti@unipd.it

Luigi Salmaso
luigi.salmaso@unipd.it

Extended author information available on the last page of the article

1 Introduction and motivation

Testing for equivalence (*Eq*) of two treatments is widely used in clinical trials, pharmaceutical experiments, bioequivalence, quality control, etc. If we take, for example, bioequivalence, a potential risk can arise if the bioequivalence of products is not well regulated and guaranteed. This paper addresses the crucial methodological step in testing for *Eq* and provides a unified framework to nonparametric testing within the permutation approach.

In the current literature there are two different, albeit dual or mirror-like, approaches for testing for *Eq*. The first commonly adopted approach, especially in bioequivalence and pharmacostatistics (Anderson-Cook and Borror 2016; Berger 1982; Berger and Hsu 1996; D'Agostino et al. 2003; Hirotsu 2007; Hung and Wang 2009; Lakens 2017; Mehta et al. 1984; Patterson and Jones 2017; Richter and Richter 2002; Wellek 2010), is derived from the intersection–union principle (IU) and its analysis is based mainly on likelihood techniques, which in turn are rather difficult to deal with, or even unavailable outside the regular exponential family (Lehmann 1986). As far as we know, the only paper on IU based on permutation methods is Arboretti et al. (2018). The other approach (Arboretti et al. 2017; Pesarin et al. 2014, 2016) is based on Roy's (1953) union–intersection principle (UI), which is also difficult to deal with using likelihood techniques (Sen 2007; Sen and Tsai 1999). The two approaches essentially differ in terms of the roles assigned to the null and alternative hypotheses. In this paper we start with a simple description of both, before introducing the related permutation solutions. We then provide a few sampling inspection plans and an application to a bioequivalence case study (two further case studies are provided in the Supplementary Material). In the final paragraphs, after exploring the limiting behavior of permutation solutions, we discuss the most important requirements and pitfalls of both parametric and nonparametric permutation-based approaches, before drawing our conclusions. The main aim of this paper is to provide the reader with some methodological insights and suggestions in order to make the most suitable choices in relation to *Eq* testing to deal with any underlying population distribution, any sample sizes and any margins.

2 On intersection–union and union–intersection approaches

With reference to one endpoint variable X and a two-sample design, to draw inferences on the substantial *Eq* of a comparative treatment A to a new treatment B, the IU approach consists in checking if the effect δ_A of A lies in a clinically or biologically or technically unimportant interval around δ_B of B, i.e. testing the non-equivalence (*NEq*) null $H : [(\delta_A \leq \delta_B - \varepsilon_I) \bigcup (\delta_A \geq \delta_B + \varepsilon_S)]$ versus (V.s) the *Eq* alternative $K : (\delta_B - \varepsilon_I < \delta_A < \delta_B + \varepsilon_S)$, where $\varepsilon_I > 0$ and $\varepsilon_S > 0$ are the inferior (lower) and superior (upper) margins for the difference $\delta = \delta_A - \delta_B$, respectively—margins that are established by biological, clinical, pharmacological, technical or regulatory arguments and not by purely statistical considerations. Focusing on the multi-aspect nature of the problem, (Berger 1982; Berger and Hsu 1996;

Schuirmann 1981, 1987), these hypotheses can be equivalently stated as $H \equiv H_I \bigcup H_S$ and $K \equiv K_I \bigcap K_S$, where $H_I : \delta \leq -\varepsilon_I$, $K_I : \delta > -\varepsilon_I$, $H_S : \delta \geq \varepsilon_S$, and $K_S : \delta < \varepsilon_S$ are the partial one-sided sub-hypotheses into which H and K are equivalently broken down. In actual fact, H is true if one and only one of H_I and H_S is true; K is true when both sub-alternatives K_I and K_S are jointly true. Accordingly, H is retained if one and only one of two suitable partial test statistics, T_I for H_I V.s K_I and T_S for H_S V.s K_S, retains the respective sub-null. The alternative K is retained if and only if two sub-alternatives K_I and K_S are jointly retained. So, the overall (global) solution, T_G say, has to be based (Berger 1982; Schuirmann 1981) on a suitable combination of two one-sided tests (TOST).

The UI approach considers the *Eq* null $\tilde{H} : (-\varepsilon_I \leq \delta \leq \varepsilon_S)$ that δ lies inside the *Eq* interval and the alternative *NEq* hypothesis $\tilde{K} : [(\delta < -\varepsilon_I) \bigcup (\delta > \varepsilon_S)]$ that δ lies outside it. By using $\tilde{H}_I : \delta \geq -\varepsilon_I$ V.s $\tilde{K}_I : \delta < -\varepsilon_I$ and $\tilde{H}_S : \delta \leq \varepsilon_S$ V.s $\tilde{K}_S : \delta > \varepsilon_S$ to denote two one-sided sub-hypotheses into which the problem can be broken down, according to Roy (1953) we may equivalently state $\tilde{H} \equiv \tilde{H}_I \bigcap \tilde{H}_S$ and $\tilde{K} \equiv \tilde{K}_I \bigcup \tilde{K}_S$. That is, the null $\tilde{H}$ is true if both one-sided sub-null hypotheses $\tilde{H}_I$ and $\tilde{H}_S$ are jointly true and $\tilde{K}$ is true if one and only one of two sub-alternatives $\tilde{K}_I$ and $\tilde{K}_S$ is true. It is worth noting that UI, having inverted the roles of null and alternative, is effectively a mirrored formulation of IU. Of course, the global UI $\tilde{T}_G$ solution implies a suitable combination of two partial test statistics $\tilde{T}_I$ and $\tilde{T}_S$. In Arboretti et al. (2017, 2018), Pesarin et al. (2016), Sen (2007) and Wellek (2010) it is seen that both combinations of T_I and T_S for IU and $\tilde{T}_I$ and $\tilde{T}_S$ for UI are the crucial methodological points at issue for obtaining proper solutions (Pesarin 2001, 2015, 2016; Pesarin and Salmaso 2010; Sen 2007; Sen and Tsai 1999, see also the Supplementary Material).

It is important to highlight that in order to obtain a valid global solution T_G, the IU approach requires the researcher to set its maximum type I error rate no larger than α and the maximum type II error rate β no larger than $1 - \alpha$; i.e.

$$\begin{aligned} &\text{a)} \quad \sup_{\delta \in H}\{\mathbf{E}_F(\phi_G, \delta)\} \leq \alpha, \\ &\qquad \text{and} \\ &\text{b)} \quad \inf_{\delta \in K}\{\mathbf{E}_F(\phi_G, \delta)\} = 1 - \beta \geq \alpha, \end{aligned}$$

where ϕ_G is the indicator function for the rejection region of the T_G global test and $\mathbf{E}_F(\cdot)$ the mean value of $(\cdot)$ with respect to the underlying data distribution F. Correspondingly, with clear meanings of the symbols, to obtain a valid UI global test $\tilde{T}_G$ the researcher must set its maximum type I error rate and maximum type II error rate as:

$$\begin{aligned} &\tilde{\text{a}}) \quad \sup_{\delta \in \tilde{H}}\{\mathbf{E}_F(\tilde{\phi}_G, \delta)\} \leq \alpha, \\ &\qquad \text{and} \\ &\tilde{\text{b}}) \quad \inf_{\delta \in \tilde{K}}\{\mathbf{E}_F(\tilde{\phi}_G, \delta)\} \geq \alpha. \end{aligned}$$

These conditions, that deal with inferential unbiasedness, are necessarily required by any test statistics (Lehmann 1986).

In the literature on the subject matter almost all authors apparently assume that regulatory agencies (e.g. FDA, EMEA, etc.) for testing *Eq* consider only the IU approach. For instance, the ICH-E9 glossary (Food and Drug Administration 1998) defines *Eq* of clinical trials as: "A trial with the primary objective of showing that the response to two or more treatments differs by an amount which is clinically unimportant. That is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences." This definition, however, does not contain sufficiently precise methodological indications as to which of two formulations, the IU (H, K) or the UI ($\tilde{H}, \tilde{K}$), is to be chosen, since there are circumstances where one or the other is rationally suitable for the testing problem at hand. We will see that the two share the same asymptotic behavior. This is not the case for finite sample sizes, where quite important differences are ascertained as will be seen in this paper.

Consequently, in any practical situation the researcher must choose which of (H, K) and ($\tilde{H}, \tilde{K}$) is most suitable for the proper analysis of his/her problem. We think that such an option, although not well emphasized in the literature on hypotheses testing, is common to almost all testing situations. A simple example clarifies this point: let us consider the classic problem of two simple hypotheses: $\theta_A \neq \theta_B$ and $\theta_A < \theta_B$. According to the Neyman-Pearson lemma, the best test for $H : \theta = \theta_A$ V.s $K : \theta = \theta_B$ rejects H when $T \geq T_\alpha$, where the critical value T_α is determined by the distribution of the likelihood ratio under H. On the other hand the best test for $\tilde{H} : \theta = \theta_B$ V.s $\tilde{K} : \theta = \theta_A$ rejects $\tilde{H}$ when $\tilde{T} \leq -\tilde{T}_\alpha$, where $\tilde{T}_\alpha$ is determined by the likelihood ratio distribution under $\tilde{H}$. So, the duality between two alternative formulations is evident. The researcher is therefore required to explicitly decide between (H, K) and ($\tilde{H}, \tilde{K}$); i.e. he/she has to justify which is given the role of null hypothesis, and the maximum rejection rate α when true, so as to strictly control both type I and type II inferential errors with $\beta \leq 1 - \alpha$. We believe that no researcher can escape this central necessity. Since both ways are rationally appropriate for *Eq* testing, such a notion supports our purpose to provide a sort of weak comparative (parallel) analysis of both by highlighting their respective requirements, properties, limitations, difficulties, pitfalls and inferential costs. It has to be stated, however, that two dual formulations reverse the roles of respective inferential risks: what works as type I error for (H, K) has the role, not just the related numerical value, of type II error for ($\tilde{H}, \tilde{K}$), and vice versa.

Some authors emphasize the general problem that any traditional two-sided consistent test rejects a point null hypothesis with a probability close to one for sufficiently large sample sizes, even for practically negligible violations of the null. For instance, Nunnally (1960) says: "To minimize type II errors, large samples are recommended. In psychology, practically all null hypotheses are claimed to be false for sufficiently large samples so (...) it is nonsensical to perform an experiment with the sole aim of rejecting the null hypothesis". According to Pantsulaia and Kintsurashvili (2014) the same concept is expressed by more than 200 authors. Clearly, to go beyond such a limitation of two-sided tests, this suggests considering a null hypothesis as made up of an interval of substantially equivalent points, rather than only one point. As a result the hypotheses for any traditional two-sided testing

is written as $(\tilde{H}, \tilde{K})$. Such a formulation then has its own specific merits, in spite of the fact that it is not adequately considered in the general literature (Wellek 2010, pp. 355–358, considers some likelihood-based hints). Up to now we have dedicated two papers: Arboretti et al. (2017) to the one-dimensional setting and Pesarin et al. (2016) to the multidimensional setting. It should, however, be emphasized that to find proper workable solutions requires going beyond the limitations of likelihood ratio approaches and so staying within a nonparametric approach, and specifically within the permutation theory and the nonparametric combination (NPC) of dependent permutation tests.

It could be argued that widespread use of the IU approach is due, rather than to rational analysis, to the fact that under a set of very stringent conditions (Lehmann 1986; Romano 2005), one uniformly most powerful unbiased test exists, namely T_G^{opt}, and this result is merely extended, by simple analogy, to all *Eq* problems. It will be seen that such an extension outside those conditions might give rise to several quite severe and intriguing consequences.

3 Intersection–union and union–intersection permutation tests

Without loss of generality and for the sake of simplicity, we illustrate the proposed methodology with reference to a two-sample design and one-dimensional endpoint variable X. To stay within the permutation theory and the nonparametric combination of dependent permutation tests (NPC), let us assume that a sample of n_1 IID data related to treatment A are drawn from X_1 and, independently, n_2 IID data related to treatment B are drawn from X_2. This setting can generally be obtained when n_1 units out of $n = n_1 + n_2$ are randomly assigned to A and the remaining n_2 to B. We define responses as $X_1 = X + \delta_A$ and $X_2 = X + \delta_B$, where the underlying variable X, whose distribution is F, is common to both populations. Hence, $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})$ are the data of sample A and $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})$ those of sample B. So, the pooled data set is $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) = \{X(i), i = 1, \ldots, n; n_1, n_2\}$, where in the last notation it is intended that the first n_1 data in the list are from the first sample and the rest from the second. Moreover, we assume that, possibly after suitable data transformations to obtain quasi data symmetry [e.g. as $\log(\cdot)$, $\sqrt{(\cdot)}$,$Rank(\cdot)$, etc., also point UI.3 in Sect. 7.3], variable X is provided with a finite mean value, i.e. $\mathbf{E}_F(|X|) < \infty$, so as to use consistent permutation tests based on comparison of sample means (Sen 2007; Pesarin 2015; Pesarin and Salmaso 2013).

It is assumed that two effects δ_A and δ_B are fixed and data are homoscedastic. In further research we will extend our permutation theory to random effects, that is to a condition compatible with important forms of heteroscedasticities, as are frequently met in most experimental and observational problems when a treatment, together with the mean, can also modify dispersion or even other aspects of a distribution.

In this context, both IU and UI approaches are in practice worked out by considering two partial tests for each way: one for H_I V.s K_I and one for H_S V.s K_S for IU; one for $\tilde{H}_I$ V.s $\tilde{K}_I$ and one for $\tilde{H}_S$ V.s $\tilde{K}_S$ for UI.

The two IU partial tests we consider have the (non standardized) form:

$$T_I = (\bar{X}_2 + \varepsilon_I) - \bar{X}_1 \text{ and } T_S = \bar{X}_1 - (\bar{X}_2 - \varepsilon_S),$$

and correspondingly, the two UI partial tests are:

$$\tilde{T}_I = \bar{X}_1 - (\bar{X}_2 + \varepsilon_I) \text{ and } \tilde{T}_S = (\bar{X}_2 - \varepsilon_S) - \bar{X}_1,$$

where as usual, $\bar{X}_j = \sum_{1 \le i \le n_j} X_{ji}/n_j$, $j = 1, 2$, are sample averages. It is worth noting that $T_I = -\tilde{T}_I$ and $T_S = -\tilde{T}_S$ and that large values of each test are evidence for their respective sub-alternatives. Also worth noting is that IU pair (T_I, T_S), as well as the UI pair $(\tilde{T}_I, \tilde{T}_S)$, are functions of essentially the same data $\mathbf{X}$ and so two tests in each pair are *negatively dependent* (Pesarin 2016; Pesarin et al. 2016).

One major problem related to both IU and UI, that also arises when several test statistics are functions of the same data, is what to do with such a multiplicity of dependent partial tests. In this regard, a meaningful warning by Sen (2007) relating to UI says: "However, computational and distributional complexities may mar the simple appeal of the UI to a certain extent. (...) The crux of the problem is however to find the distribution theory for the maximum of these possibly correlated statistics. Unfortunately, this distribution depends on the unknown F, even under the null hypothesis. (...) An easy way to eliminate this impasse is to take recourse to the permutation distribution theory (...)". The same warning applies to the IU.

We partially disagree with this warning. The greatest obstacle to achieving suitable working solutions is finding a general method to cope with the overly complex dependence structure of two partial tests (T_I, T_S) for IU and $(\tilde{T}_I, \tilde{T}_S)$ for UI. They are negatively dependent and their dependence coefficients depend on underlying F, data $\mathbf{X}$ and margins $(\varepsilon_I, \varepsilon_S)$. Indeed, such a dependence runs from correlation $\rho = -1$, for margins $\varepsilon_I = \varepsilon_S = 0$, to almost practical independence for sufficiently large margins. Quite a general solution can be validly obtained when it is possible to deal with that dependence nonparametrically.

Moreover, in multidimensional problems, such a dependence is much more complex than pair-wise linear. So it seems impossible to deal with it by proper estimates of all associated dependence coefficients, the number and type of which are typically unknown. Thus, this dependence must be worked out nonparametrically within a well-suited theory. This requires adopting the conditionality principle of inference by conditioning on data $\mathbf{X}$ (which under the null are always sufficient), i.e. by the permutation testing principle (Pesarin 2015) and, more importantly, by the NPC of dependent permutation tests (Pesarin 1990, 1992, 2001, 2015, 2016; Pesarin and Salmaso 2010, see also the Supplementary Material).

It is worth noting that to stay within the permutation theory, i.e. by permuting the n-dimensional data $\mathbf{X}$, we have to consider permuted data associated with permutations $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)$ of unit labels $\mathbf{u} = (1, \ldots, n)$. Thus, all test statistics are calculated on corresponding data permutations $\mathbf{X}^* = \{X(u_i^*), i = 1, \ldots, n; n_1, n_2)$, where two permuted samples are $\mathbf{X}_1^* = \{X(u_i^*), i = 1, \ldots, n_1\}$ and $\mathbf{X}_2^* = \{X(u_i^*), i = n_1 + 1, \ldots, n\}$, respectively.

Our proposal is to separately test, albeit simultaneously, H_I V.s K_I and H_S V.s K_S for IU, and $\tilde{H}_I$ V.s $\tilde{K}_I$ and $\tilde{H}_S$ V.s $\tilde{K}_S$ for UI.

To test for H_I V.s K_I let us consider the statistic $T_I = \bar{X}_{I2} - \bar{X}_{I1}$, where the data $\mathbf{X}_2$ of sample B are modified to $\mathbf{X}_{I2} = \mathbf{X}_2 + \varepsilon_I$ while those of sample A are retained as they are, i.e. $\mathbf{X}_{I1} = \mathbf{X}_1$. Correspondingly, to test for H_S V.s K_S we use the statistic $T_S = \bar{X}_{S1} - \bar{X}_{S2}$, where $\mathbf{X}_{S1} = \mathbf{X}_1$ and $\mathbf{X}_{S2} = \mathbf{X}_2 - \varepsilon_S$. Thus, the global test is by one of their IU-NPC solutions, the simplest and effective of which is:

$$T_G = \min(T_I, T_S) \equiv \max(\lambda_I, \lambda_S),$$

where λ_h is the so-called p value statistic for T_h, $h = I, S$.

Correspondingly, to test for $\tilde{H}_I$ V.s $\tilde{K}_I$ and $\tilde{H}_S$ V.s $\tilde{K}_S$ we use the two statistics $\tilde{T}_I = \bar{X}_{I1} - \bar{X}_{I2} = -T_I$ and $\tilde{T}_S = \bar{X}_{S2} - \bar{X}_{S1} = -T_S$, and so $\tilde{T}_G$ is given by their UI-NPC:

$$\tilde{T}_G = \max(\tilde{T}_I, \tilde{T}_S) \equiv \min(\tilde{\lambda}_I, \tilde{\lambda}_S).$$

According to the general theory (Lehmann 1986; Romano 2005; Wellek 2010), for T_G to be unbiased with the IU-NPC, it is required that conditions a) and b) are both satisfied, thus partial critical values must be calibrated so that the global test T_G satisfies α at both extremes of K, that is

$$\alpha^c = \mathbf{E}_F(\phi_h, \delta = \varepsilon_h) \text{ s.t. } \mathbf{E}_F(\phi_G, \delta = \varepsilon_h) = \alpha, \quad \text{at } \varepsilon_h = -\varepsilon_I, \varepsilon_S, h = I, S;$$

analogously for $\tilde{T}_G$ to be unbiased with the UI-NPC (Arboretti et al. 2018), it is required that conditions ã) and $\tilde{b}$) are satisfied, thus partial critical values must be calibrated so that $\tilde{T}_G$ satisfies α at both extremes of $\tilde{H}$, that is

$$\tilde{\alpha}^c = \mathbf{E}_F(\tilde{\phi}_h, \delta = \varepsilon_h) \quad \text{s.t.} \quad \mathbf{E}_F(\tilde{\phi}_G, \delta = \varepsilon_h) = \alpha, \text{ at } \varepsilon_h = -\varepsilon_I, \varepsilon_S, \ h = I, S,$$

where ϕ_h, $\tilde{\phi}_h$, ϕ_G, $\tilde{\phi}_G$, are the indicator functions of rejection regions of concerned tests.

It is worth noting that partial critical values $C_{I\alpha}$ and $C_{S\alpha}$ of parametric tests, which depend on distribution F, sample size n and margins $(\varepsilon_I, \varepsilon_S)$, according to Lehmann (1986) and Wellek (2010) are to be numerically determined (also UI.3, Sect. 7.1). Essentially, these values can coincide only asymptotically with the standard critical values (e.g. as z_α or t_α, etc.) in use with traditional two-sided tests. Thus, in our terminology, they also must be calibrated.

In some literature, the non-calibrated IU-TOST (naive) solution $\ddot{T}_G$ is often considered (e.g. Anderson-Cook and Borror 2016; Berger and Hsu 1996; Lakens 2017; Pardo 2014; Patterson and Jones 2017; Richter and Richter 2002). This solution satisfies condition a) but not b), thus it is far from being unbiased, unless sample sizes and/or margins are sufficiently large (e.g. Sect. 5 and Supplementary Material).

When optimal likelihood solutions T_G^{opt} and $\tilde{T}_G^{opt}$ are available then for divergent sample sizes, under their conditions, we have $T_G \to T_G^{opt}$ and $\tilde{T}_G \to \tilde{T}_G^{opt}$ at quite a high rate (Hoeffding 1952).

Computational details and related algorithms are in Arboretti et al. (2018) for the IU-NPC and in Pesarin et al. (2016) for the UI-NPC (see also the Supplementary

Material). Of course, by using $T_G^{ob} = T_G(\mathbf{X})$ and $\tilde{T}_G^{ob}(\mathbf{X})$ to denote the observed values of test statistics T_G and $\tilde{T}_G$, respectively, if p value statistics of the IU-NPC T_G test $\lambda_{T_G} = \Pr\{T_G^* \geq T_G^{ob} | \mathbf{X}\} \leq \alpha^c$, then the *NEq* hypothesis H is rejected at significance level α (the naive IU-TOST $\ddot{T}_G$ rejects H if $\lambda_{T_G} \leq \alpha$; so its true type I error remains unknown, depending on F, data $\mathbf{X}$ and margins ε_I, ε_S). Correspondingly, if the UI-NPC test $\tilde{T}_G$ gives $\lambda_{\tilde{T}_G} = \Pr\{\tilde{T}_G^* \geq \tilde{T}_G^{ob} | \mathbf{X}\} \leq \tilde{\alpha}^c$, then the *Eq* hypothesis $\tilde{H}$ is rejected at significance level α. In practice p value statistics are estimated, at any desired confidence rate, by a conditional Monte Carlo procedure as: $\hat{\lambda}_h = \#[T_h(\mathbf{X}^*) \geq T_h^{ob} | \mathbf{X}]/R$, where T_h stands for $T_I, T_S, T_G, \tilde{T}_T, \tilde{T}_S, \tilde{T}_G$ and R is the number of random permutations.

4 NPC limiting behavior for IU and UI

Let us assume that population mean $\mathbf{E}_F(X)$ is finite, so that $\mathbf{E}(\bar{X}^* | \mathbf{X})$ is also finite for almost all sample data $\mathbf{X}$, where $\bar{X}^*$ is the sample mean of a without replacement random sample of n_1 or n_2 elements from $\mathbf{X}$, taken as a finite population.

To find the limiting behavior of IU-NPC let us firstly consider the partial test $T_S^*(\delta) = \bar{X}_{S1}^* - \bar{X}_{S2}^*$, where its dependence on effect δ is emphasized. In Sen (2007) and Pesarin and Salmaso (2013), based on the law of large numbers for strictly stationary dependent sequences, such as are those generated by the without replacement random sampling (any random permutation is simply a without replacement random sample from $\mathbf{X}_S$), it is proved that as $\min(n_1, n_2) \to \infty$ the permutation distribution of $T_S^*(\delta)$ weakly converges to $\mathbf{E}_F(\bar{X}_{S1} - \bar{X}_{S2}) = (\varepsilon_S - \delta)$.

Thus, for any $\delta < \varepsilon_S$ the rejection rate of $T_S(\delta)$ converges to one: $\mathbf{E}_F(\phi_{T_S}, \delta) \to 1$. Moreover, for any $\delta > \varepsilon_S$ that rejection rate converges to zero. At the right extreme of H_S, $\delta = \varepsilon_S$ say, since $T_S(\varepsilon_S)$ rejects with probability α for any sample sizes (n_1, n_2), its limit rejection rate is also α.

The behavior of $T_I(\delta)$ mirrors that of $T_S(\delta)$. That is, its limiting rejection rate: i) for $\delta = -\varepsilon_I$ is α; ii) for $\delta < -\varepsilon_I$ is zero; iii) for $\delta > -\varepsilon_I$ is one.

In the global alternative $K : (-\varepsilon_I < \delta < \varepsilon_S)$, since both permutation tests T_I and T_S are jointly consistent, the global test T_G is consistent too (Pesarin 2001, 2016; Pesarin and Salmaso 2010), that is $\mathbf{E}_F(\phi_{T_G}, \delta) \to 1$. Correspondingly, for every $(\delta < -\varepsilon_I) \bigcup (\delta > \varepsilon_S)$ the limiting rejection is $\mathbf{E}_F(\phi_{T_G}, \delta) \to 0$. Moreover, in the extreme points of H, when δ is either $-\varepsilon_I$ or ε_S, as one and only one can be true if at least one differs from zero, the limiting rejection rate of T_G is α. Moreover, if $\varepsilon_I = \varepsilon_S = 0$, this rejection rate is not defined for every sample size.

To find UI-NPC's limit behavior, firstly let us analogously consider $\tilde{T}_S^*(\delta) = \bar{X}_{S2}^* - \bar{X}_{S1}^*$. As $\min(n_1, n_2) \to \infty$ implies that the permutation distribution of $T_S^*(\delta)$ weakly converges to $\mathbf{E}_F(\bar{X}_{S2} - \bar{X}_{S1}) = (\delta - \varepsilon_S)$, then for any $\delta > \varepsilon_S$ the rejection rate of $\tilde{T}_S(\delta)$ converges to one. Moreover, for any $\delta < \varepsilon_S$ its rejection rate converges to zero. At the right extreme $\delta = \varepsilon_S$, since for any sample sizes $\tilde{T}_S(\varepsilon_S)$ rejects with probability α, its limit rejection rate is also α.

The behavior of $\tilde{T}_I(\delta)$ mirrors that of $\tilde{T}_S(\delta)$. That is, the limiting rejection rate: i) for $\delta = -\varepsilon_I$ is α; ii) for $\delta > \varepsilon_I$ is zero; iii) for $\delta < -\varepsilon_I$ is one.

In the global alternative $\tilde{K} : (\delta < -\varepsilon_I) \bigcup (\delta > \varepsilon_S)$ since one and only one of $\tilde{T}_I$ and $\tilde{T}_S$ is consistent, then $\tilde{T}_G$ is consistent too (Pesarin 2001; Pesarin and Salmaso 2010; Pesarin et al. 2016).

5 A simple analysis

Calibrated values α^c and $\tilde{\alpha}^c$, so as to get global type I error rate α for the IU-NPC T_G and the UI-NPC $\tilde{T}_G$, if the underlying data distribution F is completely known, can be determined via Monte Carlo simulations as is done in Arboretti et al. (2018) for T_G and in Pesarin et al. (2016) for $\tilde{T}_G$.

Algorithms for IU-NPC and UI-NPC used to determine calibrated α^c and $\tilde{\alpha}^c$ (see also the Supplementary Material) can even be used to establish the designs $n_1 = n_2$ and $\tilde{n}_1 = \tilde{n}_2$ such that the maximum power $W_{T_G}(0; n, \varepsilon) = p$ and $W_{\tilde{T}_G}(\pm 2\varepsilon; \tilde{n}, \varepsilon) = p$ at standardized margins $\varepsilon_I = \varepsilon_S = \varepsilon$ on calibrated α^c and $\tilde{\alpha}^c$, respectively. The choice to consider designs at $\delta = 0$ for the T_G and at $\delta = \pm 2\varepsilon$ for the $\tilde{T}_G$ resides in the fact that these values are equally far away from H and $\tilde{H}$, respectively, and so their power behaviors are comparable (Wellek 2010, Chapter 11).

Assuming $X \sim N(0, 1)$ [σ unknown], $\alpha = 0.05$, $p = (0.80, 0.50)$, Table 1 contains a few designs obtained by $MC = 5000$ Monte Carlo runs, each with $R = 2500$ random permutations, for both IU-NPC and UI-NPC.

Referring to point $\varepsilon = 0.60$ as a pivot, the approximate sample sizes for $p = 0.80$, the IU-NPC designs for any intermediate margins ε' approximately agree to the empirical rule $n(\varepsilon') \approx 48.28 \cdot (0.6/\varepsilon')^2$ as obtained by interpolating simulation results. It is worth noting that these IU-NPC designs are strictly close to those obtained within the naive IU-TOST $\ddot{T}_G$ approach as reported in Lakens (2017). Such a practical coincidence is mostly due to the fact that: i) calibrated α^c coincides with non-calibrated α for interval length, adjusted with sample sizes, of about $(\varepsilon_I + \varepsilon_S)\sqrt{n_1 n_2 / n\sigma^2} > 5.4$, and ii) permutation tests are convergent at a high rate to the corresponding parametric solutions (Hoeffding 1952). On the other hand, for UI-NPC the related empirical rule for intermediate margins ε' is $\tilde{n}(\varepsilon') \approx 35.33 \cdot (0.6/\varepsilon')^2$. Similar approximate rules for $p = 0.50$ are $n(\varepsilon') \approx 30.25 \cdot (0.6/\varepsilon')^2$ and $\tilde{n}(\varepsilon') \approx 16.03 \cdot (0.6/\varepsilon')^2$, for IU-NPC and UI-NPC respectively. It is worth

Table 1 Calculations of sample sizes for IU and UI

	p	$\varepsilon \rightarrow$	1.00	0.80	0.60	0.40	0.20	0.10
IU-NPC	0.80	$n_1 = n_2$	18	28	49	109	435	1738
IU-NPC	0.50	”	–	18	30	68	270	1072
UI-NPC	0.80	$\tilde{n}_1 = \tilde{n}_2$	13	20	36	80	318	1272
UI-NPC	0.50	”	–	10	16	36	144	576

observing that to reach reasonable power the *Eq* testing process requires quite large sample sizes especially when margins are small [also point IU.2 in Sect. 7.1].

From these results we may derive a sort of relative efficiency rate of UI-NPC with respect to IU-NPC. For instance, at $\varepsilon = 0.60$ and $p = 0.80$ the rate of sample sizes is $n/\tilde{n} \approx 1.36$, for $p = 0.50$ it is $n/\tilde{n} \approx 1.88$, and for $p = 0.30$ (details not reported) it is $n/\tilde{n} \approx 2.57$. In practice, relative efficiency rates are mostly dependent on power p and are almost ε-invariant.

Table 2 reports, for standard normal data (σ unknown) with $n_1 = n_2 = 12$ and $\varepsilon = (4/5, 3/5, 2/5, 1/3, 1/5, 1/10)$, calibrated α^c, $\tilde{\alpha}^c$, rejection rates of H at $\delta = 0$ and $\delta = \pm 2\varepsilon$ for the IU-NPC T_G and the naive IU-TOST $\ddot{T}_G$, and of $\tilde{H}$ for the UI-NPC $\tilde{T}_G$, all obtained with $MC = 5000$ and $R = 2500$.

In order to clarify how to read Table 2, let us consider line $\varepsilon = 0.40$: calibrated $\alpha^c = 0.185$, $W_{T_G}(0) = 0.076, W_{T_G}(\pm 0.8) = 0.987$; $W_{\ddot{T}_G}(0) = 0.001$, $W_{\ddot{T}_G}(\pm 0.8) = 1.000; \tilde{\alpha}^c = 0.049$; $W_{\tilde{T}_G}(0) = 0.991$, and $W_{\tilde{T}_G}(\pm 0.8) = 0.249$, and so on. In particular, the naive IU-TOST $\ddot{T}_G$ appears to be dramatically conservative since its maximum power of $W_{\ddot{T}_G}(0) = 0.001$ is much smaller than $\alpha = 0.05$. Thus, since its power is much smaller than α, naive $\ddot{T}_G$ cannot be seriously considered as a practical way to test for *Eq*. $W_{T_G}(0) = 0.076$ in contrast with $W_{\tilde{T}_G}(\pm 0.8) = 0.249$, when a comparison can be stated, manifests that the UI-NPC is considerably more efficient than the IU-NPC in detecting the respective comparable alternative.

From these results we can see that IU-NPC appears to be mostly focused on *NEq* as the main assertion under testing, i.e. the one to be falsified if not true, so exhibiting an intrinsic propensity to retain H even when it is not true. Thus, its applications are mostly with problems where rejection of true *Eq* has relatively smaller costs than its acceptance when *NEq* is true while taking under strict control related global errors. This is typically the case in the areas of bioequivalence and pharmacostatistics, where it is considered ethical to retain A (the "old drug") unless there is empirical evidence that B (the "competitor") is *Eq* to it. On the other hand, UI-NPC appears to be mostly focused on *Eq*, so exhibiting a relatively larger propensity to retain $\tilde{H}$ when it is true. Thus, its applications are mostly with problems where rejection of true *Eq* has relatively greater costs than acceptance of a false *NEq* while taking under strict control related global errors. This generally occurs with testing aims to go beyond traditional two-sided procedures, as for

Table 2 Power behavior of IU and UI with NPC versus TOST

ε	IU			UI	
	α^c	T_G	$\ddot{T}_G$	$\tilde{\alpha}^c$	$\tilde{T}_G$
0.80	0.060	**0.301**/0.996	0.235/1.000	0.050	0.999/**0.603**
0.60	0.099	**0.144**/0.994	0.025/1.000	0.050	0.998/**0.402**
0.40	0.185	**0.076**/0.987	0.001/1.000	0.049	0.991/**0.249**
0.333	0.225	**0.066**/0.984	0.000/1.000	0.048	0.985/**0.191**
0.20	0.337	**0.059**/0.963	0.000/1.000	0,046	0.966/**0.109**
0.10	0.428	**0.052**/0.959	0.000/1.000	0.037	0.954/**0.065**

instance with quality control, etc. It is also important to emphasize that for $\varepsilon \leq 0.333$ the maximum probability for the naive IU-TOST $\ddot{T}_G$ to retain *Eq*, when it is true, is zero [see also points ÏÜ.3, ÏÜ.5, ÏÜ.6 in Sect. 7.2], so resulting in pure costs without any inferential benefits.

Table 3 reports, for data from $N(0, 1)$ (σ unknown), the minimal sample sizes $\ddot{n}_1 = \ddot{n}_2$, in terms of $\varepsilon = \varepsilon_I = \varepsilon_S$, for naive IU-TOST $\ddot{T}_G$ when conditions a) and b) are satisfied, i.e. to be unbiased at $\ddot{\alpha}_G = .05$, and the maximum probability (i.e. the power) to accept *Eq* $[\ddot{W}(Eq)]$ at $\delta = 0$.

If σ were known we have essentially the same results. It is proved that when (n_1, n_2) and/or $\varepsilon_I + \varepsilon_S$ are not sufficiently large (Wellek, 2010, p. 5), the naive IU-TOST $\ddot{T}_G$, as is frequently used in the literature, can be unacceptably biased (Sect. 7.2).

6 A bioequivalence application

Let us consider data from (Hirotsu (2017), p. 108) on the end-point variable Log $C_{\max}$ (Log of maximum blood concentration of a drug), related to $n_1 = 20$ Japanese subjects and $n_2 = 13$ Caucasians, after prescribing a standard dose of a drug. Data concern a bridging study conducted to investigate for bioequivalence between two populations. So, the test is to see if two populations can be considered bioequivalent with respect to that variable. Data are reported in Table 4.

The basic statistics are: $\bar{X}_{Jap} = 1.518$; $\hat{\sigma}_{Jap} = 0.0813$; $\bar{X}_{Cau} = 1.457$; $\hat{\sigma}_{Cau} = 0.0951$; pooled $\hat{\sigma} = 0,0869$. By firstly using the permutation test $T' = |\bar{X}_J - \bar{X}_C|$ for the point null hypothesis $H' : X_J \overset{d}{=} X_C$ versus the two-sided alternative $K' : X_J \overset{d}{\neq} X_C$, with $R = 100\,000$ we obtain the p value statistic $\hat{\lambda}' = 0.0535$. There is no evidence for non-equality between two data sets at $\alpha = 0.05$, although $\bar{X}_{Jap}$ appears to be slightly larger than $\bar{X}_{Cau}$ (Student's $t = 1.991$, 31 df, $\lambda'_t > 0.05$).

Let us consider the IU-NPC T_G and the UI-NPC $\tilde{T}_G$ with the list of margins $\varepsilon_I = \varepsilon_S = (0.058,\ 0.071,\ \ 0.109,\ \ 0.125)$, corresponding to studentized values (in terms of $\hat{\sigma}$) of $(2/3,\ \ 0.82,\ \ 1.25,\ 1.44\ )$, respectively.

The results, with $R = 100\,000$ on data X, are in Table 5.

The results in brackets are related to mid-rank data transformations. In our opinion, due to some apparent irregular data concentrations and some ties, mid-rank results can be slightly more reliable than those on plain data X (IU.3 Sect. 7.1, and UI.1 Sect. 7.3). At ε such that $\ddot{\alpha}(\varepsilon) = \alpha_G = 0.05$, i.e. $\varepsilon \approx 0.071$ (corresponding to $\approx 0.82\hat{\sigma}$), type I error rates of naive IU-TOST $\ddot{T}_G(\varepsilon)$ and of IU-NPC T_G approximately coincide. Of course, this coincidence also remains for larger sample sizes and margins. With the data of the example, the *Eq* of two data sets is retained

Table 3 Minimal sample sizes for unbiasedness of naive IU-TOST

$\varepsilon_I = \varepsilon_S$	0.10	0.25	0.50	0.75
$\ddot{n}_1 = \ddot{n}_2$	1210	194	49	22
$\ddot{W}(Eq)$	0.58	0.58	0.58	0.58

Table 4 Data from Hirotsu (2007)

Jap									
1.567	1.515	1.500	1.591	1.624	1.691	1.531	1.456	1.351	1.478
1.461	1.571	1.565	1.586	1.406	1.488	1.500	1.577	1.500	1.407
Cau									
1.455	1.375	1.474	1.650	1.464	1.375	1.479	1.413	1.423	1.389
1.441	1.650	1.348							

Table 5 Analysis with IU and UI NPC

$\varepsilon_I = \varepsilon_S$	IU			UI		
	α^c	$\hat{\lambda}_G$	Inference	$\tilde{\alpha}^c$	$\hat{\lambda}_{\tilde{G}}$	Inference
0.058	0.068	0.545 (0.730)	$H: NEq$	0.050	0.455 (0.277)	$\tilde{H}: Eq$
0.071	0.050	0.382 (0.607)	$H: NEq$	0.050	0.618 (0.400)	$\tilde{H}: Eq$
0.109	0.050	0.071 (0.155)	$H: NEq$	0.050	0.929 (0.849)	$\tilde{H}: Eq$
0.125	0.050	0.025 (0.040)	$K: Eq$	0.050	0.975 (0.962)	$\tilde{H}: Eq$

by the IU-NPC T_G for margins $\varepsilon_I = \varepsilon_S \gtrsim 0.12 \approx 1.38\hat{\sigma}$. For margins $\varepsilon_I = \varepsilon_S < 0.12$, we could state that there is not enough information to conclude that the two populations are equivalent. Anyway, it also well known in the literature that the IU-NPC approach suffers from a lack of power (e.g., Berger and Hsu 1996; Wellek 2010).

On the other hand, the UI-NPC $\tilde{T}_G$ retains Eq for all margins, including $\varepsilon_I = \varepsilon_S = 0$ (in such a case $\tilde{\alpha}^c = \alpha/2$ and the related p value statistic is $\hat{\lambda}_{\tilde{G}} = \hat{\lambda}' = 0.0535 > 0.025$).

7 Warnings and good practices for NPC equivalence

7.1 The IU-NPC solution

The IU-NPC approach, as well as the likelihood-based IU, presents some pitfalls, as is evident from previous results (see also the Supplementary Material). The most important requirements and pitfalls are:

- IU.1) It does not admit any solution when $\varepsilon_I = \varepsilon_S = 0$, i.e. when the null hypotheses is $H: [(\delta \leq 0) \bigcup (\delta \geq 0)]$ in which case the alternative K becomes logically impossible as it is empty, $K = \varnothing$ say.
- IU.2) When the T_G measure of $\varepsilon_I + \varepsilon_S$ is small, there still remain difficulties in retaining Eq when it is true. This difficulty is well recognized in the literature.

For instance, (Wellek (2010), p. 5) says "...the sample sizes required in an equivalence test in order to achieve a reasonable power typically tend to be considerably larger than in an ordinary one- or two-sided testing procedure ...unless the range of tolerance deviations ...is chosen so wide that even distributions exhibiting pronounced dissimilarities would be declared 'equivalent' ...". The same difficulties are confirmed by simulation results (Arboretti et al. 2018).

- IU.3) Using Monte Carlo to achieve the IU-NPC calibrated α^c requires complete knowledge of underlying distribution F of variable X, *including all its nuisance parameters*. When a central limit theorem is working for partial test distributions, calibrated α^c can be approximately assessed by a simulation algorithm as in (Arboretti et al. (2018), see also the Supplementary Material) where the unknown standard deviation σ_X is replaced by its sampling estimate $\hat{\sigma}_X$. Thus, assessing α^c endures two sources of approximations. This difficulty is particularly intriguing when sample sizes (n_1, n_2) and/or Eq interval length $\varepsilon_I + \varepsilon_S$ are small. However, if data X are provided with finite second moment, IU-NPC is little influenced by mis-specification of data distribution F [also UI.3, Sect. 7.3)]. When no assumption on underlying F is undertaken, then mid-rank transformation of the numeric data $\mathbf{X}$ and margins $(\varepsilon_I, \varepsilon_S)$ may provide for reliable evaluation of calibrated α^c, provided that normal approximation for Wilcoxon-Mann-Whitney statistics takes place, i.e. for sample sizes of about 10 or larger.
- IU.4) According to results in Arboretti et al. (2018) and based on limiting behavior of the permutation test as stated in (Hoeffding (1952), IU-NPC test $T_G = \min(T_I, T_S)$ quickly converges to T_G^{opt} in the conditions for the latter.
- IU.5) Unless $\min(n_1, n_2)$ or $\varepsilon_I + \varepsilon_S$ are very large, once Eq is rejected, the application of a Bonferroni-like rule for establishing which H_h, $h = I, S$, is active, if not always impossible, is generally difficult since calibrated α^c lie in the half-open interval $[\alpha, (1 + \alpha)/2)$.
- IU.6) In practice, to analyze a given data set $(\mathbf{X}_1, \mathbf{X}_2)$, with (n_1, n_2) sample sizes, margins $(\varepsilon_I, \varepsilon_S)$, at significance level α, one has to firstly establish or to estimate α^c via Monte Carlo as in point IU.3; then, one can proceed with the IU-NPC analysis. This implies using two computing algorithms.
- IU.7) While using any kind of ranks, only within the IU-NPC permutation approach is it possible to express margins in terms of the same physical measurement units of variable X. Rank solutions as discussed in Wellek (2010) and Janssen and Wellek (2010) express margins in terms of rank transformations so as to mimic solutions based on normal settings. However, this implies considering something similar to random margins, the meaning of which become doubtful or at least questionable and too difficult to justify (Arboretti et al. 2015; Hirotsu 2007). This same difficulty is also met when monotonic data transformations, such as $X = \varphi(Y)$, are necessary and margins are expressed in terms of transformed values X. In any case, provided that margins are clearly justified, IU-NPC can be correctly applied if these are expressed either in terms of original data Y or in terms of transformed data X.

- IU.8) The multidimensional extension of the IU approach by likelihood methods is far from satisfactory, especially outside normal distributions. We think this extension can easily be done under the NPC and we intend to do it in future research.
- IU.9) Calibrated reference values under the parametric likelihood ratio approach are obtained by numerical calculations (Wellek 2010; Lehmann 1986) only for population distributions lying within the regular exponential family if the invariance property for nuisance parameters (if any) works. Outside, only approximated solutions can be obtained [IU.3]. So the IU parametric approach is extremely demanding. Moreover, whenever the minimal sufficient statistics in the null hypothesis is the whole n-dimensional data set $\mathbf{X}$, only nonparametric permutation solutions can be set up correctly (Pesarin 2015, 2016; Pesarin and Salmaso 2010).

7.2 The naive IU-TOST solution

The naive IU-TOST solution, $\ddot{T}_G = \min(T_I, T_S)$ say, as is frequently considered in the literature (Anderson-Cook and Borror 2016; Berger 1982; Berger and Hsu 1996; Pardo 2014; Patterson and Jones 2017; Richter and Richter 2002; Schuirmann 1987; Wellek 2010), corresponds to the non-calibrated version that rejects the global H at type I error rate α when both partial tests reject each other at the same rate α in place of calibrated α^c, i.e. when $\ddot{\alpha}_I = \ddot{\alpha}_S = \alpha$. This naive $\ddot{T}_G$ solution has several further specific pitfalls:

- ÏÜ.1) It satisfies condition a) but not b) in Sect. 2; however, it trivially satisfies Theorem 1 in Berger (1982) and Berger and Hsu (1996).
- ÏÜ.2) When the T_G measure of $\varepsilon_I + \varepsilon_S$ is very large, the non-calibrated naive $\ddot{T}_G$, whose partial type I error rates are $\ddot{\alpha}_I = \ddot{\alpha}_S = \alpha^c = \alpha$, and the calibrated IU-NPC T_G coincide, and so they are both consistent (Sect. 4).
- ÏÜ.3) The naive IU-TOST $\ddot{T}_G$ can be dramatically conservative and its maximum rejection probability can be much smaller than α, even exactly zero (Arboretti et al. 2018), see ÏÜ.5 and results in Sect. 5 (see also the Supplementary Material).
- ÏÜ.4) In Theorem 2 in Berger (1982) and Berger and Hsu (1996), essentially states that margins $(\varepsilon_I, \varepsilon_S)$ exist such that the power under K of naive test $\ddot{T}_G$ is not smaller than α. Since $\ddot{T}_G$ is consistent, as the standardized length of the Eq interval diverges at the rate $[n_1 n_2/(n_1 + n_2)]^{1/2}$, if $\min(n_1, n_2)$ diverges, such an existence corresponds to consistency. However, it is important to underline that such a condition is not constructive and so is not beneficial to finding practical solutions. Indeed, in any real problem, based on technical or biological or regulatory considerations, margins are established before the experiment for data collection is conducted. So, since it is unknown if the $\ddot{T}_G$ measure of $(\varepsilon_I + \varepsilon_S)$ with actual sample data is sufficiently large so that $\ddot{\alpha} = \alpha^c = \alpha$, naive $\ddot{T}_G$ solutions do not guarantee minimal requirements in order to be considered valid test statistics.

- ÏÜ.5) Paradoxically, when the *Eq* interval length $\varepsilon_I + \varepsilon_S$ is small in terms of the $\ddot{T}_G$ distribution, the maximum probability for the naive IU-TOST $\ddot{T}_G$ of finding a drug equivalent to itself can be exactly zero. This generally occurs when two partial rejection regions have no common points, i.e. when $\phi_S \bigcap \phi_I = \emptyset$ so leading to impossible events, where ϕ_I and ϕ_S are the α-rejection regions of T_I and T_S, respectively. For instance, with: $n_1 = n_2 = 12$, $\varepsilon_I = \varepsilon_S = 0.25$, $X \sim N(0,1)$ and $\ddot{\alpha}_I = \ddot{\alpha}_S = 0.05$, by a simulation with $MC = 5000$ and $R = 2500$ the type I error for $\ddot{T}_G$ is $\ddot{\alpha}_G \approx 0.000$ and, much worse, the maximum estimated power $\hat{W}\ddot{T}_G(0) \approx 0.000$. Interestingly, the calibrated IU-NPC α^c is about 0.293 (that of UI-NPC $\tilde{T}_G$ is $\tilde{\alpha}^c \approx 0.047$; see also the Supplementary Material). In this respect it is easy to see that for normal data, with known σ and $\varepsilon_I = \varepsilon_S = \varepsilon$, the maximum probability to retain *Eq* is exactly zero up to $n_1 = n_2 = \lfloor 2(z_\alpha \sigma/\varepsilon)^2 \rfloor$, with $\lfloor(\cdot)\rfloor$ the integer part of $(\cdot)$ and z_α the α-quantile of $N(0, 1)$.
- ÏÜ.6) As a consequence, naive IU-TOST $\ddot{T}_G$ tests are not members of the set of test statistics that satisfy conditions a) and b) in Sect. 2. Moreover, as the global type I error rate and power can both be considerably smaller than α for small sample sizes and/or small *Eq* interval length, we may state that the naive $\ddot{T}_G$ testing procedure is based on an incorrect methodology, meaning that it can happen that true type I error results $\alpha(\pm\varepsilon, n) \ll \alpha$ and maximal power, at $\delta = 0$, $W\ddot{T}_G(0, n, \varepsilon) \ll \alpha$, conditions that do not agree with the minimal requirements for any test (Nunnally 1960, and Sect. 2). Thus, in our opinion, unless sample sizes and/or *Eq* interval length are sufficiently large, there is no reason for taking naive $\ddot{T}_G$ into consideration in *Eq* testing. Essentially, this is our basic criticism regarding the widespread use of the naive IU-TOST method (e.g. Anderson-Cook and Borror (2016), Pesarin (1990, (1992), among the many). We think that this intrinsic defect remains hidden to most practitioners because naive IU-TOST apparently sounds non-counter-intuitive.
- ÏÜ.7) Direct consequence of former two points is that for naive IU-TOST $\ddot{T}_G$ the cumulation of inferences from independent studies could be unsuitable. For instance, if there are $m \geq 2$ analyses, each based on insufficiently large sample sizes, as is common in some meta-analyses and multicenter studies, their combination might always reject that a drug is *Eq* to itself. Indeed, for valid combination it is required that all m partial tests are unbiased (i.e. minimal power $\geq \alpha$, Sect. 2). In fact, if for study h, $h = 1, \ldots, m, \phi_{Sh} \bigcap \phi_{Ih} = \emptyset$, i.e. the joint rejection region of any two partial tests T_{Sh} and T_{Ih} is empty, the p value related to $\ddot{T}_{Gh}$ is 1 and so p value of any of its combinations is also 1, hence always providing for *NEq*, true or not.

7.3 The UI-NPC solution

The most important requirements and pitfalls of the UI-NPC are:

- UI.1) Using Monte Carlo to establish the calibrated $\tilde{\alpha}^c$ requires complete knowledge of underlying distribution F of endpoint variable X, including all its

nuisance parameters [same as IU.3, Sect. 7.1)]. When, for partial test distributions, a central limit theorem is working, calibrated $\tilde{\alpha}^c$ can be approximately determined according to Arboretti et al. (2018), since the *Eq* interval length $\varepsilon_I + \varepsilon_S$ can be measured in terms of underlying standard error $\sigma_X[n_1 n_2/(n_1 + n_2)]^{-1/2}$. As in practice σ_X is unknown, substitution by its sampling estimate $\hat{\sigma}_X$ implies that $\tilde{\alpha}^c$ can be assessed only approximately. It is worth noting, however, that the related degree of approximation is generally negligible in practice because: i) its true value lies in the closed interval $[\frac{1}{2}\alpha, \alpha]$ and so the maximum approximation error is bounded by $\alpha/2$; ii) for any given *Eq* interval, calibrated $\tilde{\alpha}^c$ quickly converges to α for increasing sample sizes, provided that the population mean $\mathbf{E}_F(X)$ is finite. When population mean is assumed not to be finite, then mid-rank transformation of the numeric data $\mathbf{X}$ and margins $(\varepsilon_I, \varepsilon_S)$, can provide for well approximated evaluations of calibrated $\tilde{\alpha}^c$, provided that normal approximation for Wilcoxon-Mann-Whitney statistics takes place.

- UI.2) Similarly to point IU.6 (Sect. 7.1) to analyze a given data set $(\mathbf{X}_1, \mathbf{X}_2)$, with (n_1, n_2) sample sizes, margins $(\varepsilon_I, \varepsilon_S)$, at significance level α, one has to firstly establish or to estimate $\tilde{\alpha}^c$ via Monte Carlo as at point UI.1; then, one can proceed with the UI-NPC analysis. This too implies using two computing algorithms, but with much less impact than with IU-NPC, because $\tilde{\alpha}^c \in [\frac{1}{2}\alpha, \alpha]$, which is a much smaller range than $[\alpha, (1 + \alpha)/2)$. Indeed, a similar five-entry table would require much smaller numbers of sample sizes and margins.
- UI.3) Once *NEq* is retained at significance level α, identifying which of two arms is mostly responsible for that result using a Bonferroni-like rule implies that the related type I error is in $[\frac{1}{2}\alpha, \alpha]$, and so the related type I error rate is not less than $\alpha/2$. Indeed, it is close to α even for moderate sample sizes and small *Eq* interval length since, in practice, UI-NPC is intrinsically robust against mis-specification of underlying F, possibly after data transformations achieving near symmetry. In this regard, with data from a Student's t distribution with 2 df (zero mean and infinite variance), $n_1 = n_2 = 12$, $\varepsilon_I = \varepsilon_S = 0.321$, corresponding to margins of about 0.25 for standard normally distributed data since $\Pr\{-0.25 \le N(0,1) \le 0.25\} = \Pr\{-0.321 \le t_2 \le 0.321\}$, we have $\alpha^c \approx 0.375$ and $\tilde{\alpha}^c \approx 0.047$. Compared to those that are active under standard normal data, as in ÏÜ.5 (Sect. 7.2), α^c proves to be much larger than 0.293, and so the IU-NPC appears not to be robust against F; instead $\tilde{\alpha}^c$ coincides at the third figure with 0.047, confirming that the UI-NPC is at least approximately invariant on F, provided that near symmetry for the data is achieved. The robustness properties of IU-NPC and UI-NPC will be considered in further research.
- UI.4) When $\varepsilon_I = \varepsilon_S = 0$, i.e. for sharp null and two-sided alternatives, unless the underlying data distribution is symmetric, it is well known that it is difficult to find unbiased tests based on comparison of sample averages (Cox and Hinkley 1974; Lehmann 1986). Within the UI-NPC, however, the test $\tilde{T}_G = \max[(\bar{X}_1 - \bar{X}_2), (\bar{X}_2 - \bar{X}_1)]$ is always at least unbiased at $\alpha/2$.
- UI.5) Similarly to IU.9 (Sect. 7.1), calibrated reference values under the parametric likelihood ratio approach are obtained by numerical calculations only

for population distributions lying within the regular exponential family if the invariance property for nuisance parameters (if any) works (Ferguson 1967). So, like the IU, the UI parametric approach is also quite demanding. On the contrary, when no parametric UI is available, approximations within UI-NPC generally suffice for most practical applications [UI.1].

8 Concluding remarks

The present paper provides a sort of comparative analysis of two nonparametric permutation approaches for *Eq* testing problems. In accordance with the majority of the literature on the subject matter, one is based on the IU principle. The other is based on the UI principle. Although they entail different evaluations of inferential errors, both are rationally suitable for such testing and so they are not strictly comparable. As such, rather than a proper comparison, we have proposed a sort of weak comparative (parallel) analysis. However, we believe that neither can be considered uniformly the best to be used for all possible problems. Thus, our analysis is mostly concerned with highlighting their respective requirements, properties, difficulties, inferential costs, limitations and pitfalls.

One important point we took into consideration was that in some of the literature the IU solution is used referring to the so-called non-calibrated reference critical values. We called it the naive IU-TOST solution. In this regard, we showed (see ÏÜ.5 and ÏÜ.6, Sect. 7.2) that since its type I error rate and power for relatively small margins and/or sample sizes can be zero, thus implying rejection of *Eq*, true or not, with a probability close to one, its related testing process can become absolutely useless, resulting in pure costs without any inferential benefits. This rather erroneous feature may lead, for instance, to the unacceptable conclusion that "the probability to find that a drug is *Eq* to itself by the naive IU-TOST can be zero".

A further aspect we would like to consider is a sort of comparison between the IU and the UI with respect to the so-called point null hypotheses. A point null is equivalent to considering $\varepsilon_I = \varepsilon_S = 0$, with length of equivalence interval of zero. On the one hand, the UI way coincides with the traditional two-sided solution plus one more: once the null has been rejected, its p value $\lambda = \min(\lambda_I, \lambda_S)$ satisfies Bonferroni's rule (UI.3, Sect. 7.3) and allows us to make inference on which is the active arm: e.g. if $\lambda = \lambda_I$ then $\delta < 0$, at type I rate $\tilde{\alpha}^c = \alpha/2$ (similarly for δ_S). On the other hand, the IU way cannot have any solution, so in this formulation a point null cannot be considered as a null interval. This too shows that two formulations are essentially different.

A problem faced by any researcher is finding guidance to choose between two approaches. Our point of view is that if he/she considers that rejection of *Eq* when it is true has relatively smaller costs than its acceptance when *NEq* is true, as can typically be considered the case with bioequivalence and pharmacostatistics, then the IU-NPC is the correct choice. Correspondingly, if he/she considers that rejection of *Eq* when it is true has relatively greater costs than its acceptance when *NEq* is true, as can typically be the case with traditional two-sided testing (quality control, etc.), then the UI-NPC is the correct choice.

In the usual literature on the subject matter, both IU and UI parametric approaches are essentially worked out within likelihood techniques. These approaches, which in any case imply approximate solutions, are rather difficult to deal with since they require quite severe conditions of validity, such as population distributions lying within the regular exponential family and enjoying the invariant property for nuisance parameters, if any—conditions that are generally quite difficult to meet and/or justify. Our IU-NPC and UI-NPC permutation solutions are also approximated. However, when a parametric optimal solution exists, its NPC counterpart is asymptotically convergent to it at a high rate. When a solution within likelihood ratio is not invariant with respect to one or more nuisance parameters, it cannot be worked out unless these nuisance parameters are completely known. Our IU-NPC and UI-NPC solutions, given they are working conditionally on a set of sufficient statistics in one point of the null hypothesis, do not require any knowledge of nuisance parameters, so are sufficiently flexible to cope with most practical problems.

Acknowledgements Authors wish to thank the Editor, Associate Editor and Referees for helping to improve the manuscript.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

References

Anderson-Cook C, Borror C (2016) The difference between "equivalent" and "not different". Qual Eng 28:249–262

Arboretti R, Carrozzo E, Caughey D (2015) A rank-based permutation test for equivalence and noninferiority. Ital J Appl Stat 25:81–92

Arboretti R, Carrozzo E, Pesarin F, Salmaso L (2017) A multivariate extension of union–intersection permutation solution for two-sample testing. J Stat Theory Pract 11:436–448

Arboretti R, Carrozzo E, Pesarin F, Salmaso L (2018) Testing for equivalence: an intersection–union permutation solution. Stat. Biopharm Res 10:130–138

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

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection–union tests and equivalence confidence sets. Stat Sci 11:283–319

Cox D, Hinkley D (1974) Theoretical statistics. Chapman and Hall, London

D'Agostino RB, Massaro JM, Sullivan LM (2003) Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. Stat Med 22:169–186

Ferguson TS (1967) Mathematical statistics, a decision theoretic approach. Academic Press, New York

Food and Drug Administration (1998) Guidance for industry: E9 statistical principles for clinical trials. Food and Drug Administration, Rockville

Hirotsu C (2007) A unifying approach to non-inferiority, equivalence and superiority tests via multiple decision processes. Pharm Stat 6:193–203
Hirotsu C (2017) Advanced analysis of variance. Wiley, Hoboken
Hoeffding W (1952) The large-sample power of tests based on permutations of observations. Ann Math Stat 23:169–192
Hung H, Wang S (2009) Some controversial multiple testing problems in regulatory applications. J Biopharm Stat 19:1–11
Janssen A, Wellek S (2010) Exact linear rank tests for two-sample equivalence problems with continuous data. Stat Neerl 64:482–504
Lakens D (2017) Equivalence trials: a practical primer for t test, correlations and meta-analyses. Soc Psychol Pers Sci 8:355–362
Lehmann E (1986) Testing statistical hypotheses. Wiley, New York
Mehta CR, Patel NR, Tsiatis AA (1984) Exact significance testing to establish treatment equivalence with ordered categorical data. Biometrics 40:819–825
Nunnally J (1960) The place of statistics in psychology. Educ Psychol Meas 20:641–650
Pantsulaia G, Kintsurashvili M (2014) Why is the null hypothesis rejected for 'almost every' infinite sample by some hypothesis testing of maximal reliability. J Stat Adv Theory Appl 11:45–70
Pardo S (2014) Equivalence and noninferiority tests for quality. Manufactoring and test engineers. Chapman & Hall/CRC, Boca Raton
Patterson S, Jones B (2017) Bioequivalence and statistics in clinical pharmacology, 2nd edn. Chapman & Hall/CRC, Boca Raton
Pesarin F (1990) On a nonparametric combination method for dependent permutation tests with applications. Psychometrics Psychosomatics 54:172–179
Pesarin F (1992) A resampling procedure for nonparametric combination of several dependent tests. J Ital Stat Soc 1:87–101
Pesarin F (2001) Multivariate permutation tests, with applications in biostatistics. Wiley, Chichester
Pesarin F (2015) Some elementary theory of permutation tests. Commun Stat Theory Methods 44:4880–4892
Pesarin F (2016) Encyclopedia of statistical sciences, chap permutation test: multivariate. Wiley-StatRef, Hoboken
Pesarin F, Salmaso L (2010) Permutation tests for complex data, theory, applications and software. Wiley, Chichester
Pesarin F, Salmaso L (2013) On the weak consistency of permutation tests. Commun Stat Simul Comput 42:1368–1397
Pesarin F, Salmaso L, Carrozzo E, Arboretti R (2014) Testing for equivalence and non-inferiority: Iu and ui tests within a permutation approach. JSM 2014—section on nonparametric statistics
Pesarin F, Salmaso L, Carrozzo E, Arboretti R (2016) Union-intersection permutation solution for two-sample equivalence testing. Stat Comput 26:693–701
Richter S, Richter C (2002) A method for determining equivalence in industrial applications. Qual Eng 14:375–380
Romano J (2005) Optimal testing of equivalence hypotheses. Ann Stat 33:1036–1047
Roy S (1953) On a heuristic method of test construction and its use in multivariate analysis. Ann Math Stat 24:220–238
Schuirmann D (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15:657–680
Schuirmann DL (1981) On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. Biometrics 37:617
Sen P (2007) Union–intersection principle and constrained statistical inference. J Stat Plan Inference 137:3741–3752
Sen P, Tsai M (1999) Two-stage likelihood ratio and union intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix. J Multivar Anal 68:264–282
Wellek S (2010) Testing statistical hypotheses of equivalence and noninferiority. Chapman & Hall/CRC, Boca Raton

Affiliations

Rosa Arboretti[1] · **Fortunato Pesarin**[2] · **Luigi Salmaso**[3]

[1] Department of Civil, Environmental and Architectural Engineering, University of Padova, Padova, Italy

[2] Department of Statistical Sciences, University of Padova, Padova, Italy

[3] Department of Management and Engineering, University of Padova, Padova, Italy