An Analysis of Union-Intersection and Intersection-Union Tests for Equivalence and Non-Inferiority

Fortunato Pesarin[a] (pesarin@stat.unipd.it)

Joint with: *R. Arboretti*[b], *L. Salmaso*[c] and *E. Carrozzo*[c]

a: Department of Statistical Sciences; *b:* Department of Land, Environment, Agriculture and Forestry; *c:* Department of Management and Engineering, University of Padova - Italy

Adaptive Designs and Multiple Testing Workshop;

Padova, April 28-29, 2016

OUTLINE

1. Introduction

2 UI- and IU-NPC solutions to the univariate case

2.1. An algorithm for the UI-NPC

2.2. The IU-NPC algorithm

2.3. A visualization of UI- and IU-NPC

3 A parallel analysis of UI- and IU-NPC

4. Some limiting properties

5. A couple of examples

6. Some references

1 Introduction

"*To minimize type II errors, large samples are recommended. In psychology, practically all* [sharp or point] *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.*" (Nunnally, 1960. The same concept is in more than 200 papers).

Rather than of **exactly one point,** this remarkable and meaningful concept suggests to think the **null hypothesis** as **an interval of substantially equivalent points**.

In a two-sample design with: one variable X, the difference of two effects $\delta = \delta_2 - \delta_1$, two margins $(\varepsilon_I, \varepsilon_S) > 0$, the equivalence (Eq) problem consists of testing for

$$H_0 : -\varepsilon_I \leq \delta \leq \varepsilon_S \qquad \text{V.s} \qquad H_1 : [(\delta < -\varepsilon_I) \bigcup (\delta > +\varepsilon_S)].$$

Breaking-down the hypotheses, we have:

$$H_0 \equiv H_{0I} \bigcap H_{0S} \qquad \text{V.s} \qquad H_1 \equiv H_{1I} \bigcup H_{1S},$$

where sub-hypotheses are:

$$H_{0I} : \delta \geq -\varepsilon_I, \;\; H_{0S} : \delta \leq \varepsilon_S, \;\; H_{1I} : \delta < -\varepsilon_I, \text{ and } \;\; H_{1S} : \delta > \varepsilon_S.$$

The solution requires *Two One-Sided* partial *Tests* (TOST; Schuirmann, 1987), such as:

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

for respectively H_{1I} and H_{1S} (**actually, one and only one can be active!**) followed by their *suitable combination* within Roy's (1953) *Union-Intersection* (UI) principle: $T_G = UI(T_I, T_S)$.

One TOST UI solution is: $T_G = \max(T_I, T_S) \equiv \min(\lambda_I, \lambda_S)$,

(λ_I, λ_S) are p-value statistics; $H_0 : Eq$ **is accepted if both tests accept**.

When either ε_I or ε_S is **large**, measured by the T_G distribution, the problem becomes of *non-inferiority* or *non-superiority*.

Under rather stringent conditions, UMPU likelihood-based solutions for $(\varepsilon_I, \varepsilon_S) \geq 0$ do exist (Ferguson, 1967; Lehmann, 1986).

However, they are too unpractical to be used in real problems (Cox and Hinkley, 1974).

So, although not unbiased, it is common practice for $\varepsilon_I = \varepsilon_S = 0$ to only adopt tests based on divergence of sample averages, like $|\bar{X}_1 - \bar{X}_2|$.

To the best of our knowledge, the only known UI practical solution in a permutation setting is in Pesarin et al. (2016).

P.K. Sen (2007) gives a list of unsolved problems (*restraints*) for the likelihood-UI approach. He says:

"*However, computational and distributional complexities may mar the simple appeal of the UI principle 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. (...) [not so easy, anyway, since regressions on partial tests can be more complex than pair-wise linear].

In most of the complex statistical inference problems, the usual likelihood formulation stumbles into methodological as well as computational difficulties, even in asymptotic setups."

The FDA glossary defines equivalence in 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* $[\delta]$ *is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences."*

Thus, **testing for equivalence** of two treatments -agricultural and industrial experiments, clinical trials, pharmaceutical experiments, bioequivalence, quality control, etc.- **can rationally be approached by two different principles.**

One based on Roy's UI; the other, commonly adopted in the literature, based on the so-called *Intersection-Union* (IU), e.g. Wellek (2010) and references therein.

The two differ on the role assigned to the kind of null and alternative on which the inferential focus is addressed to.

With one variable X and a two-sample design, the hypotheses within the IU are

$$\tilde{H}_0 : [(\delta \leq -\varepsilon_I) \cup (\delta \geq \varepsilon_S)] \qquad \text{V.s} \qquad \tilde{H}_1 : -\varepsilon_I < \delta < \varepsilon_S;$$

that is, $\tilde{H}_0 \equiv \tilde{H}_{0I} \bigcup \tilde{H}_{0S}$ V.s $\tilde{H}_1 \equiv \tilde{H}_{1I} \bigcap \tilde{H}_{1S}$.

Note: IU *is essentially the dual formulation* (mirror-like) *w.r. to the* UI. Two partial tests are:

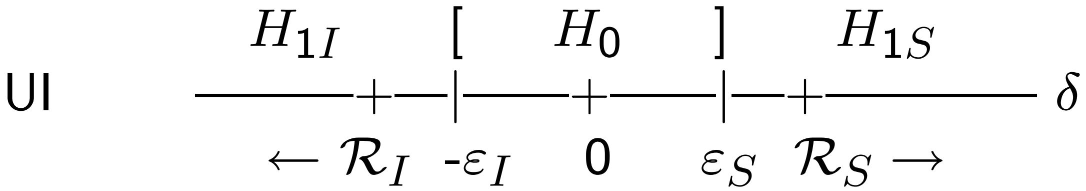
$$\tilde{T}_I = -T_I = \bar{X}_2 + \varepsilon_I - \bar{X}_1 \quad \text{and} \quad \tilde{T}_S = -T_S = \bar{X}_1 - (\bar{X}_2 - \varepsilon_S).$$

The related global test is: $\tilde{T}_G = IU(\tilde{T}_I, \tilde{T}_S)$.

One TOST IU solution is: $\tilde{T}_G = \min(\tilde{T}_I, \tilde{T}_S) \equiv \max(\tilde{\lambda}_I, \tilde{\lambda}_S)$,

$(\tilde{\lambda}_I, \tilde{\lambda}_S)$ are p-value statistics; $\tilde{H}_1 : Eq$ **is accepted if both tests reject**.

In UI **the null** is that δ **lies within** the given equivalence [closed] interval, and **the alternative** that it **lies outside**, when treatments substantially differ

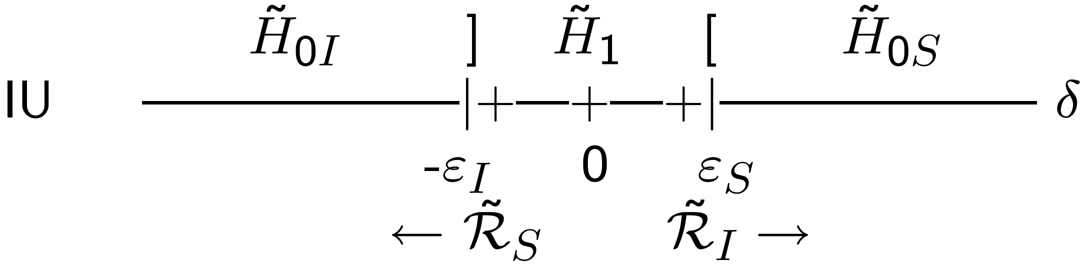


H_{1I} [H_0] H_{1S}

UI ———+—|——+——|—+——— δ

$\leftarrow \mathcal{R}_I$ -ε_I 0 ε_S $\mathcal{R}_S \rightarrow$

In IU **the alternative** is that δ **lies inside** that]open[interval, and **the null** that it **lies outside** it, when treatments are non-equivalent (N-Eq).



$\tilde{H}_{0I}$] $\tilde{H}_1$ [$\tilde{H}_{0S}$

IU ——————|+—+—+|—————— δ

-ε_I 0 ε_S

$\leftarrow \tilde{\mathcal{R}}_S$ $\tilde{\mathcal{R}}_I \rightarrow$

Note: intersection of rejection regions $\tilde{\mathcal{R}}_S$ and $\tilde{\mathcal{R}}_I$ can be empty.

The IU $\tilde{T}_G$ presents some specificities and quite serious difficulties:

1) It has no solution when $\varepsilon_I = \varepsilon_S = 0$; difficulties persist for small $\varepsilon_I + \varepsilon_S$.

2) *Under very stringent assumptions one unconditionally likelihood-based optimal* (UMPUI) *test $\tilde{T}_O$ exists* (Lehmann, 1986; Romano, 2005; Wellek, 2010). To get $\tilde{T}_O$ unbiased, partial critical values must be *adjusted* so as it satisfies α at both extremes of $\tilde{H}_1$, that is

$$\tilde{\alpha} = \mathbf{E}_F(\tilde{\phi}_h, \varepsilon_h) \text{ s.t. } \mathbf{E}_F(\tilde{\phi}_O, \varepsilon_h) = \alpha,\ \varepsilon_h = -\varepsilon_I, \varepsilon_S,$$

where: $\tilde{\phi}_{(\cdot)}$ is the indicator function of rejection region $\tilde{\mathcal{R}}_{(\cdot)}$; both partial tests $\tilde{T}_h$, $h = I, S$, share the same adjusted $\tilde{\alpha}$.

3) IU mimics Lehmann's theorem that for any $\tilde{T}$ requires: **a**) $\tilde{T}$ is at most of size α : $\sup_{\delta \in \tilde{H}_0}[\mathbf{E}_F(\tilde{\phi}_{\tilde{T}}, \delta)] \leq \alpha$; **b**) $\tilde{T}$ is at least unbiased: $\inf_{\delta \in \tilde{H}_1}[\mathbf{E}_F(\tilde{\phi}_{\tilde{T}}, \delta)] \geq \alpha$.

4) $\tilde{T}_O$ is *optimal* in quite a narrow class $\tilde{\mathcal{T}}$. Typical competitors are the permutation IU $\tilde{T}_G$ and SenGupta's (2007) P^3 counterparts. However, *its optimality in class $\tilde{\mathcal{T}}$ does not imply it is good even outside.*

5) The permutation adjusted IU-TOST $\tilde{T}_G$, quickly converges to $\tilde{T}_O$.

6) Unless the invariance property works, *to get the adjusted IU-TOST $\tilde{T}_G$ available in practice, the complete knowledge of underlying distribution F of data X is required* (including all its nuisance parameters).

7) Unless $\min(n_1, n_2)$ or $\varepsilon_I + \varepsilon_S$ are very large, application of multiple testing techniques for showing which $\tilde{H}_{0h}$ is active, if not impossible, is generally *difficult* since adjusted $\tilde{\alpha} \in [\alpha, (1+\alpha)/2)$.

8) While using ranks, only within the permutation approach it seems possible to express margins in terms of the same measurement units of the data (Jannsen and Wellek, 2010; Arboretti et al., 2015).

9) The non-adjusted (naive) TOST $\ddot{T}_G$, whose type I partial errors are $\ddot{\alpha}_I = \ddot{\alpha}_S = \alpha$, satisfies Lehmann's condition **a**), but not **b**), unless $\varepsilon_I + \varepsilon_S$ and/or $\min(n_1, n_2)$ are very large [where $\tilde{T}_G$ and $\ddot{T}_G$ essentially coincide].

10) Naive TOST $\ddot{T}_G$ can be dramatically conservative [its rejection probability (RP) can be $\ll \alpha$]:

i) with $n_1 = n_2 = 12$, $\varepsilon_I = \varepsilon_S = .2$, $X \sim \mathcal{N}(0, 1)$ and $\ddot{\alpha}_I = \ddot{\alpha}_S = .05$, $\ddot{T}_G$ type I error is $\ddot{\alpha}_G \approx .000$ (with max power $\ddot{W} \approx .000$ at $\delta = 0$);

ii) $\tilde{T}_G$ type I error $.05$ requires adjusted $\tilde{\alpha} \approx .337$ (with max $\tilde{W} \approx .059$);

iii) for $\tilde{\alpha}_G = \ddot{\alpha}_I = \ddot{\alpha}_S \approx .05$, $n_1 = n_2 \approx 280$ are needed (max $\tilde{W} \approx .540)$;

iv) with $n_1 = n_2 = 12$, $\varepsilon_I = \varepsilon_S = .89$, $\tilde{\alpha}_G = \ddot{\alpha}_I = \ddot{\alpha}_S \approx .05$, max $\tilde{W} \approx .335$.

It is proved that when (n_1, n_2) and/or $\varepsilon_I + \varepsilon_S$ are not sufficiently large, the naive TOST $\ddot{T}_G$ as is commonly used in the literature can be unacceptably biased, its maximal power being $\ll \alpha$.

Paradoxically, from i): *the maximal probability for the naive TOST $\ddot{T}_G$ to find a drug equivalent to itself is about zero* (in first three decimal figures).

Indeed, when both partial rejection regions are external to the equivalence interval defined by $\tilde{H}_1$, i.e. if $(-\varepsilon_I, \varepsilon_S) \bigcap [\tilde{\mathcal{R}}_S \bigcup \tilde{\mathcal{R}}_I] = \emptyset$, there are situations where such a power is exactly zero.

By the way, in conditions i) above to get the UI T_G type I error $\alpha_G = .05 = \mathbf{E}_F(\phi_G, \varepsilon_h)$, $h = I, S$, at both extreme points of H_0, adjusted partial type I error $[\alpha^c = \mathbf{E}_F(\phi_h, \varepsilon_h)$ s.t. $\mathbf{E}_F(\phi_G, \varepsilon_h) = .05$, $h = I, S]$ is $\alpha^c \approx .046$ [the max Acceptance Probability (AP) for equivalence at $\delta = 0$ is $1 - \mathbf{E}_F(\phi_G, 0) \approx .966$].

Two UI and IU approaches, however, are not directly and meaningfully comparable.

So, a parallel analysis is to be provided.

The most difficult problem for achieving proper solutions is to find a way to cope with the too complex dependence structure on two partial tests (T_I, T_S) for the UI and $(\tilde{T}_I, \tilde{T}_S)$ for the IU. They are negatively dependent and their dependence coefficients depend on underlying F, the given data set $\mathbf{X}$ and margins $(\varepsilon_I, \varepsilon_S)$.

Indeed, in multidimensional problems, such a dependnce is much more complex than pair-wise linear. So it is impossible to deal with it by proper estimators of all related coefficients, the number and type of which are typically unknown.

Thus, *that dependence must be worked out nonparametrically within a suitable theory.* This implies taking recourse to the permutation testing principle and specifically to the **NonParametric Combination** (NPC) **of dependent Permutation Tests** (PTs) (Pesarin, 1990, 1992, 2001; Pesarin and Salmaso, 2010).

The permutation testing principle essentially requires that in the space of effects there is a point $\delta_0 \notin H_1$ s.t. data permutations are equally likely ($\approx$ exchangeable data).

In particular, PTs take benefits from the **conditional and unconditional uniform monotonicity property** (Pesarin and Salmaso, 2010, p. 88): *testing for $H_0' : \delta \leq \delta_0$* V.s $H_1' : \delta > \delta_0$ *by the unbiased PT T, with rejection region indicator ϕ_T, such a property states that for any $\delta' < \delta_0 < \delta$, any data $\mathbf{X}$, any sample sizes (n_1, n_2), and any underlying F, the following relations respectively hold:*

$$\lambda(\mathbf{X}(\delta')) \overset{d}{\geq} \lambda(\mathbf{X}(\delta_0)) \overset{d}{\geq} \lambda(\mathbf{X}(\delta)) \ ,$$

and

$$\mathbf{E}_F(\phi_T, \delta') \leq \mathbf{E}_F(\phi_T, \delta_0) = \alpha \leq \mathbf{E}_F(\phi_T, \delta) \ .$$

Our main goal is to **realize the UI-NPC and IU-NPC approaches for the equivalence and non-inferiority problems** and to propose a first parallel analysis.

2 UI- and IU-NPC solutions to the univariate case

Suppose: two-sample IID data are $\mathbf{X}_j$, with $n_j \geq 2$, $j = 1, 2$; ε_I and ε_S the equivalence margins for $\delta = \delta_2 - \delta_1$; and X is such that $\mathbf{E}_F(X)$ is finite [if not finite: *multi-aspect* methods or rank tests can be used].

Also: $\mathbf{X}_1$ belongs to the reference experiment and $\mathbf{X}_2$ to the competitor.

For testing the one-sided $H_{0I} : \delta \geq -\varepsilon_I$ V.s $H_{1I} : \delta < -\varepsilon_I$ consider $\mathbf{X}_{I1} = \mathbf{X}_1$ and $\mathbf{X}_{I2} = \mathbf{X}_2 + \varepsilon_I$,

and for $H_{0S} : \delta \leq \varepsilon_S$ V.s $H_{1S} : \delta > \varepsilon_S$ the $\mathbf{X}_{S1} = \mathbf{X}_1$ and $\mathbf{X}_{S2} = \mathbf{X}_2 - \varepsilon_S$.

So, a univariate problem is transformed into a special bivariate one, where two vectors $\mathbf{X}_I = \mathbf{X}_{I1} \biguplus \mathbf{X}_{I2}$ and $\mathbf{X}_S = \mathbf{X}_{S1} \biguplus \mathbf{X}_{S2}$ are deterministically related.

Two partial tests are: $T_I = \bar{X}_{I1} - \bar{X}_{I2}$ and $T_S = \bar{X}_{S2} - \bar{X}_{S1}$.

Note: large values of both partial tests are significant.

Moreover: $X \overset{P}{>} 0$ implies $\bar{X}_{h1} - \bar{X}_{h2} \equiv \bar{X}_{h1}/\bar{X}_{h2}$, $h = I, S$. So: *for positive variables difference intervals and ratio intervals have the same handling within the permutation setting*.

When H_{0I} is true, data in $\mathbf{X}_I$ are exchangeable at point $\delta = -\varepsilon_I$. Thus, the rejection probability (RP) of test T_I at $\delta = -\varepsilon_I$ is: $\mathbf{E}_F(\phi_I, -\varepsilon_I) = \alpha$ (attainable).

Since $\delta < \delta'$ implies $\mathbf{E}_F(\phi_I, \delta) \geq \mathbf{E}_F(\phi_I, \delta')$ [note *monotonicity in* δ], RP is not smaller than α at $\delta < -\varepsilon_I$, and not larger than α at $\delta > -\varepsilon_I$. And this uniformly for all sample data and all F.

Correspondingly, when H_{0S} is true, data in $\mathbf{X}_S$ are exchangeable at $\delta = \varepsilon_S$ and so the RP of T_S is not smaller than α at $\delta > \varepsilon_S$, is not larger than α at $\delta < \varepsilon_S$ and equals α at $\delta = \varepsilon_S$.

Two tests T_I and T_S are negatively related, since when one tries to reject the other tries to accept (e.g., $\varepsilon_I = \varepsilon_S = 0$ implies $T_I^* + T_S^* = 0$).

The related permutation process defines the bivariate distribution of (T_I^*, T_S^*).

As in H_1 **either** H_{1I} **or** H_{1S} is true, *one way* for defining the global test is:

$$T_G^o = \mathsf{min}(\lambda_I^o, \lambda_S^o),$$

where $\lambda_h^o = \mathsf{Pr}\{T_h^*(\delta) \geq T_h^o(\delta) | \mathbf{X}_h(\delta)\}$ is the permutation p-value-like statistic of test T_h, $h = I, S$ [also suitable can be *Tippett's* and *Fisher's* rules].

Unless the cardinality of permutation space $\Pi(\mathbf{X})$ is relatively small, we estimate at any desired degree of accuracy the bivariate distribution of (T_I^*, T_S^*) by means of a conditional Monte Carlo procedure, consisting of a random sample of R elements from $\Pi(\mathbf{X})$. Commonly, R is at least of 1000.

Two p-value statistics $\hat{\lambda}_h^o$, $h = I, S$, are then estimated as

$$\hat{\lambda}_h^o = \textstyle\sum_{r=1}^R \mathbf{I}\{T_{hr}^*(\delta) \geq T_h^o(\delta)|\mathbf{X}_h(\delta)\}/R,$$

where $T_{hr}^* = T[\mathbf{X}_h(\mathbf{u}_r^*)]$ is T_h statistic at the rth permutation of $(1, \ldots, n)$.

Negative relation between T_I^* and T_S^* implies that T_G is not unbiased for all δ, all sample sizes and all $(\varepsilon_I, \varepsilon_S)$ and consistency is not for all combining functions.

For $\varepsilon_I = \varepsilon_S = 0$ and type I error of both partial PTs $\alpha^c = \alpha$, type I error $\alpha_G = 2\alpha$ (as for the traditional two-sided testing).

When the length $\varepsilon_I + \varepsilon_S$, measured w.r. to the distribution of T^*_G, is moderately large all three tests T_I, T_S and T_G share type I error at α : $\alpha^c = \alpha_G = \alpha$.

Thus, application of multiple testing techniques becomes easy. E.g., if for sufficiently large sample sizes T_G is rejected at type I error α, the branch h, $h = I, S$, s.t. $\hat{\lambda}^o_h = \min(\hat{\lambda}^o_I, \hat{\lambda}^o_S)$ is declared active at type I error not larger than α. In general, the exact errors α^c lie in the (closed) range $[\alpha/2,\ \alpha]$.

With $n_1 = n_2 = 12$, $\varepsilon_I = \varepsilon_S = .2$ and $X \sim \mathcal{N}(0, 1)$, UI-NPC uses $\alpha^c \approx .046$ for T_G type I error $\alpha_G = .05$.

So, if $\lambda^o_G \leq .05$, the related branch h were declared significant at type I error α^c in the interval [.025, .05].

To appreciate how far from α is type I error of T_G in some typical situations, a simple simulation study is reported in the following table.

$\varepsilon_I, \varepsilon_S$	$n_1 = n_2$	$\alpha^c = \alpha$	α_G
.00	10	.01| .05| .10	.0200| .100| .200
.10	"	"	.0130| .067| .141
.25	"	"	.0103| .053| .109
.50	"	"	.0101| .051| .101
.75	"	"	.0100| .050| .100
.10	20	"	.0102| .063| .128
.25	"	"	.0101| .052| .103
.50	"	"	.0100| .050| .100
.10	40	"	.0101| .053| .116
.25	"	"	.0100| .050| .102
.50	"	"	.0100| .050| .100

Table 1: Values of α_G of T_G when partial tests are at $\alpha^c = .01, .05, .10$, with $n_1 = n_2 = 10, 20, 40$ and $\varepsilon_I = \varepsilon_S = .1, .25, .5, .75$ for $X \overset{d}{=} \mathcal{N}(0, 1)$, based on $R = 10000$ permutations and $MC = 20000$ Monte Carlo runs.

2.1 An algorithm for the UI-NPC:

1. Read data $\mathbf{X} = (X_i, i = 1, \ldots, n; n_1, n_2)$ and margins $\varepsilon_I, \varepsilon_S > 0$.

2. Define two vectors $\mathbf{X}_I = (\mathbf{X}_{I1} \uplus \mathbf{X}_{I2}) = (X_{I1i} = X_{1i}, i = 1, \ldots, n_1; X_{I2i} = X_{2i} + \varepsilon_I, i = 1, \ldots, n_2)$ and $\mathbf{X}_S = (\mathbf{X}_{S1} \uplus \mathbf{X}_{S2}) = (X_{S1i} = X_{1i}, i = 1, \ldots, n_1; X_{S2i} = X_{2i} - \varepsilon_S, i = 1, \ldots, n_2)$.

3. Compute $T_I^o = \bar{X}_{I1} - \bar{X}_{I2}$ and $T_S^o = \bar{X}_{S2} - \bar{X}_{S1}$ and take memory.

4. Take a random permutation $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)$ of unit labels $\mathbf{u} = (1, \ldots, n)$.

5. Define the two permuted samples: $\mathbf{X}_I^* = [X_I(u_i^*), i = 1, \ldots, n; n_1, n_2]$ and $\mathbf{X}_S^* = [X_S(u_i^*), i = 1, \ldots, n; n_1, n_2]$, both defined on the same $\mathbf{u}^*$.

6. Compute $T_I^* = \bar{X}_{I1}^* - \bar{X}_{I2}^*$ and $T_S^* = \bar{X}_{S2}^* - \bar{X}_{S1}^*$ and take memory.

7. Independently repeat R times steps 4 to 6; so $[(T_{Ir}^*, T_{Sr}^*),\ r = 1, \ldots, R]$ simulates the bivariate permutation distribution of two partial tests (T_I, T_S).

8. Estimate two partial p-value statistics $\hat{\lambda}_h^o = \sum_{r=1}^{R} \mathbf{I}(T_{hr}^* \geq T_h^o)/R$, $h = I, S$, and the UI-global test statistic as $\hat{T}_G^o = \min(\hat{\lambda}_I^o, \hat{\lambda}_S^o)$.

9. If $\hat{T}_G^o \leq \alpha^c$, reject the global null hypothesis H_0 at type I error α.

2.2 The IU-NPC algorithm

Everything else being as for the UI-NPC, the algorithm for IU-NPC modifies steps 3 and 6 to 9 into:

$\tilde{3}$. Compute two observed tests: $\tilde{T}_I^o = \bar{X}_{I2} - \bar{X}_{I1}$ and $\tilde{T}_S^o = \bar{X}_{S1} - \bar{X}_{S2}$ and take memory.

$\tilde{6}$. Compute two permuted tests: $\tilde{T}_I^* = \bar{X}_{I2}^* - \bar{X}_{I1}^*$ and $\tilde{T}_S^* = \bar{X}_{S1}^* - \bar{X}_{S2}^*$ and take memory.

$\tilde{7}$. Independently repeat R times steps 4 to 6; $[(\tilde{T}_{Ir}^*, \tilde{T}_{Sr}^*), r = 1, \ldots, R]$ simulates the bivariate permutation distribution of two partial tests $(\tilde{T}_I, \tilde{T}_S)$.

$\tilde{8}$. Estimate two partial p-value statistics: $\tilde{\lambda}_h^o = \sum_{r=1}^R \mathbf{I}(\tilde{T}_{hr}^* \geq \tilde{T}_h^o)/R$, $h = I, S$, and the IU-global test as $\tilde{T}_G^o = \max(\tilde{\lambda}_I^o, \tilde{\lambda}_S^o)$.

$\tilde{9}$. If $\tilde{T}_G^o \leq \tilde{\alpha}$, reject the global null hypothesis $\tilde{H}_0$ at type I error α.

2.3 A visualization of UI- and IU-NPC

$\mathbf{X}$	$\mathbf{X}_1^*$	$\cdots$	$\mathbf{X}_r^*$	$\cdots$	$\mathbf{X}_R^*$
T_I^o	T_{I1}^*	$\cdots$	T_{Ir}^*	$\cdots$	T_{IR}^*
T_S^o	T_{S1}^*	$\cdots$	T_{Sr}^*	$\cdots$	T_{SR}^*
T_G^o	T_{G1}^*	$\cdots$	T_{Gr}^*	$\cdots$	T_{GR}^*
$\tilde{T}_I^o$	$\tilde{T}_{I1}^*$	$\cdots$	$\tilde{T}_{Ir}^*$	$\cdots$	$\tilde{T}_{IR}^*$
$\tilde{T}_S^o$	$\tilde{T}_{S1}^*$	$\cdots$	$\tilde{T}_{Sr}^*$	$\cdots$	$\tilde{T}_{SR}^*$
$\tilde{T}_G^o$	$\tilde{T}_{G1}^*$	$\cdots$	$\tilde{T}_{Gr}^*$	$\cdots$	$\tilde{T}_{GR}^*$

where: T_G^* and $\tilde{T}_G^*$ are obtained according to an *adaptive weighted rule*, with weights: $w_h = 1$ if $h = \arg\max_k(T_k^o)$ and $\tilde{w}_h = 1$ if $\tilde{h} = \arg\min_k(\tilde{T}_k^o)$, and 0 elsewhere.

So, the observed values of two global tests are:

$$T_G^o = w_I T_I^o + w_S T_S^o \quad \text{and} \quad \tilde{T}_G^o = \tilde{w}_I \tilde{T}_I^o + \tilde{w}_S \tilde{T}_S^o;$$

the permutation distributions of which, for $r = 1, \ldots, R$, are:

$$T_{Gr}^* = w_I T_{Ir}^* + w_S T_{Sr}^* \ \text{ and } \ \tilde{T}_{Gr}^* = \tilde{w}_I \tilde{T}_{Ir}^* + \tilde{w}_S \tilde{T}_{Sr}^*, \text{ respectively.}$$

Note: the reference p-values for T_G and $\tilde{T}_G$ are the adjusted ones: for $\lambda_G = \text{Pr}\{T_G^* \geq T_G^o | \mathbf{X}, (\varepsilon_I, \varepsilon_S)\}$ it is α^c (not α); for $\tilde{\lambda}_G = \text{Pr}\{\tilde{T}_G^* \geq \tilde{T}_G^o | \mathbf{X}, (\varepsilon_I, \varepsilon_S)\}$ it is $\tilde{\alpha}$ (not α).

3 A parallel analysis of IU- and UI-NPC

Simulation results for: $X \sim \mathcal{N}(0,1)$, $\alpha = 5\%$, $MC = 5000$, $R = 2500$, $n_1 = n_2$, $\varepsilon_I = \varepsilon_S$, max power at $\delta = 0$ for naive TOST-IU $W\ddot{T}^*_G$, adjusted partial $\tilde{\alpha}_h$, $W\tilde{T}^*_G$ adjusted, $W\tilde{T}_O$ optimal (Wellek, 2010, page 122), UI adjusted α^c, max acceptances for UI AT^*_G :

n	IU					UI	
10	ε	$W\ddot{T}^*_G$	$\tilde{\alpha}$	$W\tilde{T}^*_G$	$W\tilde{T}_O$	α^c	AT^*_G
	1.0	.392	.054	.426	.453	.05	1.00
	.75	.085	.078	.190	.198	.05	.999
	.50	.001	.154	.091	.093	.049	.994
	.25	.000	.310	.054	.058	.048	.971
	.10	.000	.434	.050	——	.036	.956

n			IU			UI	
15	ε	$W\ddot{T}^*_G$	$\tilde{\alpha}$	$W\tilde{T}^*_G$	$W\tilde{T}_O$	α^c	AT^*_G
	1.0	.704	.05	.704	.714	.05	1.00
	.75	.040	.059	.348	.355	.05	1.00
	.50	.008	.113	.123	.127	.05	.998
	.25	.000	.271	.061	.063	.047	.979
	.10	.000	.417	.053	——	.039	.956

n			IU			UI	
20	ε	$W\ddot{T}^*_G$	$\tilde{\alpha}$	$W\tilde{T}^*_G$	$W\tilde{T}_O$	α^c	AT^*_G
	1.0	.846	.05	.846	.859	.05	1.00
	.75	.513	.052	.527	.533	.05	1.00
	.50	.032	.084	.163	.171	.05	.999
	.25	.000	.237	.065	.068	.048	.984
	.10	.000	.402	.056	——	.040	.963

We report: **I**) the IU-adjusted $\tilde{\alpha}$ so as $\tilde{\alpha}_G \approx .05$ for $n_1 = n_2 = 12$, $X \sim \mathcal{N}(0, 1)$, max IU power $W\tilde{T}^*_G$ at $\delta = 0$ and $\tilde{\alpha}_G \approx .05$, max IU power $W\ddot{T}^*_G$ of naive IU-TOST at $\ddot{\alpha} = .05$; and **II**) max UI-TOST acceptances AT^*_G at $\delta = 0$ and UI-adjusted α^c so as $\alpha_G \approx .05$ (by simulation with $R = 5000$ and $MC = 10000$)

I)

$\varepsilon_I = \varepsilon_S$	$\tilde{\alpha}$	$W\tilde{T}^*_G$	$W\ddot{T}^*_G$
0.80	0.060	0.301	0.235
0.40	0.185	0.076	0.001
0.333	0.225	0.066	0.000
0.20	0.337	0.059	0.000
0.10	0.428	0.052	0.000
0.02	0.504	0.051	0.000
0.01	0.513	0.0505	0.000
0,001	0.523	0.0502	0.000

II)

AT^*_G	α^c
0.999	0.050
0.991	0.049
0.985	0.048
0.966	0,046
0.954	0.040
0.952	0.029
0.951	0.027
0.950	0.025

.

Results confirm optimality of $\tilde{T}_O$. Moreover, $\tilde{T}_O$ and IU-NPC $\tilde{T}_G$ seem comparable.

Also confirmed is that adjusted $\tilde{\alpha} \in [\alpha,\ (1+\alpha)/2)$ and $\alpha^c \in [\alpha/2,\ \alpha]$.

The AP of equivalence generally results better for UI-NPC T_G than for IU-NPC $\tilde{T}_G$; and so T_G appears as a *more efficient* testing way.

However, **this conclusion is not completely correct**.

We have to ask "*who pays costs of inferential errors*?"

As inferential errors are essentially referred to two opposite "subjects" (one pays for H_0 and one for $\tilde{H}_0$), in our opinion there is no clear sense for their meaningful comparison.

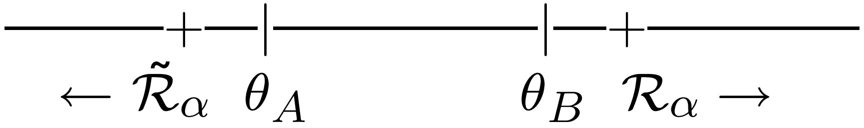
Similarly, for instance, to the very standard Neyman-Pearson lemma with $\theta_A < \theta_B$ where for

$$H_0 : \theta = \theta_A \ \text{ V.s } \ H_1 : \theta = \theta_B,$$

there is an optimal test T, whose rejection region $\mathcal{R}_\alpha$ is mostly determined by θ_A;

whereas for $\tilde{H}_0 : \theta = \theta_B$ V.s $\tilde{H}_1 : \theta = \theta_A$,

there is quite a different optimal test $\tilde{T}$, whose rejection region $\tilde{\mathcal{R}}_\alpha$ is mostly determined by θ_B;


$\leftarrow \tilde{\mathcal{R}}_\alpha$
θ_A
θ_B
$\mathcal{R}_\alpha \rightarrow$

.

Thus, the choice between two testing strategies cannot be obtained by only considering their rejection and acceptance corresponding probabilities, as for instance Pr$\{T \in \mathcal{R}_\alpha|\theta_B\}$ and Pr$\{\tilde{T} \notin \tilde{\mathcal{R}}_\alpha|\theta_B\}$, because related to different subjects on differently defined rejection regions: $\mathcal{R}_\alpha \neq \mathcal{X}-\tilde{\mathcal{R}}_\alpha$.

Of course, this parallel analysis has to be completed and extended to one-sample, multi-sample and multivariate situations.

Including, in particular, testing equivalence on $V_E \geq 1$ variables and non-superiority on $V_{NS} \geq 0$ variables (restricted alternatives).

Such extensions are possible within the NPC.

4 Some limiting properties

Assume that $\mathbf{E}_F(X)$ is finite, so that also $\mathbf{E}(\bar{X}^*|\mathbf{X})$ is finite for almost all $\mathbf{X} \in\mathcal{X}^n$ ($\bar{X}^*$ is the sample mean of a WRRS of n_1 or n_2 elements from the pooled set $\mathbf{X}$ taken as a finite population).

For the UI-NPC way firstly consider the $T_S^*(\delta) = \bar{X}_{S2}^* - \bar{X}_{S1}^*$, where its dependence on δ is emphasized.

From the results in Pesarin and Salmaso (2013), based on the law of large numbers for strongly stationary dependent sequences as those generated by the WRRS process, it can be proved that, as $\min(n_1, n_2) \to \infty$, the PT $T_S^*(\delta)$ weakly converges to $\mathbf{E}_F(\bar{X}_{S2} - \bar{X}_{S1}) = (\delta - \varepsilon_S)$.

Thus:

i) at $\delta < \varepsilon_S$, as $T_S^*(\delta) \xrightarrow{P} (\delta - \varepsilon_S) \in H_{0S}$ then its RP converges to zero;

ii) at $\delta > \varepsilon_S$, as $(\delta - \varepsilon_S) \in H_{1S}$ the RP converges to one;

iii) at $\delta = \varepsilon_S$, as for sufficiently large sample sizes the RP of $T_S(\delta)$ is α, its limit is α too.

The behavior of T_I is s.t. $T_I^*(\delta) \xrightarrow{P} -(\delta + \varepsilon_I)$. Thus, its limiting RP: i) for $\delta > -\varepsilon_I$ is zero; ii) for $\delta < -\varepsilon_I$ is one; iii) for $\delta = -\varepsilon_I$ is α.

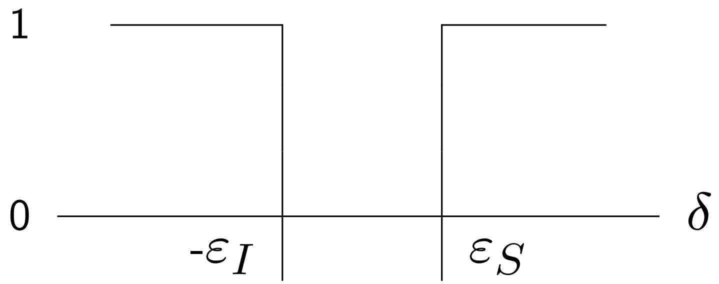
In the global alternative $H_1 : (\delta < -\varepsilon_I) \bigcup (\delta > \varepsilon_S)$ one and only one between T_I and T_S is consistent, then T_G is consistent too.

Indeed, since either ($\lambda_I \overset{P}{\rightarrow}$ 0 and $\lambda_S \overset{P}{\rightarrow}$ 1) OR ($\lambda_I \overset{P}{\rightarrow}$ 1 and $\lambda_S \overset{P}{\rightarrow}$ 0) for respectively $\delta \in H_{1I}$ OR $\delta \in H_{1S}$. Thus, $\forall \delta \in H_1$, min$(\lambda_I, \lambda_S) \overset{P}{\rightarrow} 0 < T_{G\alpha}$, $\forall \alpha > 0$. Then consistency.

Moreover, in the extreme points of H_0 when δ is either $-\varepsilon_I$ OR ε_S, as one and only one can be true if at least one is positive the RP of T_G is α (if both ε_I and ε_S are 0, RP is 2α); when $-\varepsilon_I < \delta < \varepsilon_S$, i.e. in the open equivalence interval, the limiting RP is zero being such for both partial tests.

By permuting 0 and 1, it is proved that the IU procedures exactly enjoy the same limits.

Limits of RP of H_0 for the UI-NPC and of AP of $\tilde{H}_0$ for the IU-NPC



It is remarkable to note that:

1- If there is a point δ_0 s.t. pooled data $\mathbf{X}(\delta_0)$ are exchangeable, then both UI-NPC T_G and IU-NPC $\tilde{T}_G$ are admissible tests since both rejection regions are convex (Birnbaum, 1954, 1955).

2- If $0 < Var_F(X) < \infty$, that often suffices for the PCLT, then both UI-NPC T_G and IU-NPC $\tilde{T}_G$ are admissible combinations of asymptotically optimal tests, as each partial test is asymptotically optimal.

5 A couple of examples

1. As a first example, consider (from Hirotsu, 2004) $Log\ C_{\mathsf{max}}$ for 20 Japanese subjects and 13 Caucasians.

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

With $R = 100000$ for $H_0 : X_J \overset{d}{=} X_C$ V.s $H_1 : X_J \overset{d}{\neq} X_C$, test $T^* = \bar{X}^*_J - \bar{X}^*_C$ obtains a two-sided p-value $\hat{\lambda} = 0.0535$ (one-sided is $\hat{\lambda} = 0.0268$).

Both p-values agree with a substantial equivalence (Eq), although $\bar{X}_J = 1.518$ is lightly larger than $\bar{X}_C = 1.457$.

Consider UI-NPC and IU-NPC for testing equivalence with margins $\varepsilon_I = \varepsilon_S =$ (0.017, 0.022, 0.058, 0.071, 0.109, 0.120), approximately equal to (1/5, 1/4, 2/3, 0.82, 1.25, 1.38) the s.d. $\hat{\sigma} = 0.0869$, respectively.

The UI-NPC results are:

$$\begin{array}{l}
\lambda_G(0.017) = \min(0.079; 0.992) \Longrightarrow \text{acceptance of } H_0 : \text{Eq} \\
\lambda_G(0.022) = \min(0.098; 0.994) \Longrightarrow \text{acceptance of } H_0 : \text{Eq} \\
\lambda_G(0.058) = \min(0.455; 0.999) \Longrightarrow \text{acceptance of } H_0 : \text{Eq} \\
\lambda_G(0.071) = \min(0.618; 1.000) \Longrightarrow \text{acceptance of } H_0 : \text{Eq} \\
\lambda_G(0.109) = \min(0.929; 1.000) \Longrightarrow \text{acceptance of } H_0 : \text{Eq} \\
\lambda_G(0.120) = \min(0.961; 1.000) \Longrightarrow \text{acceptance of } H_0 : \text{Eq}
\end{array}$$

The IU-NPC resuls are:

$$
\begin{aligned}
\tilde{\lambda}_G(0.017) &= \max(0.921; 0.008) \Longrightarrow \text{acceptance of } \tilde{H}_0 : \text{N-Eq } (\tilde{\alpha} \approx 0.312)\\
\tilde{\lambda}_G(0.022) &= \max(0.902; 0.006) \Longrightarrow \text{acceptance of } \tilde{H}_0 : \text{N-Eq } (\tilde{\alpha} \approx 0.264)\\
\tilde{\lambda}_G(0.058) &= \max(0.545; 0.001) \Longrightarrow \text{acceptance of } \tilde{H}_0 : \text{N-Eq } (\tilde{\alpha} \approx 0.068)\\
\tilde{\lambda}_G(0.071) &= \max(0.382; 0.000) \Longrightarrow \text{acceptance of } \tilde{H}_0 : \text{N-Eq } (\tilde{\alpha} \approx 0.05)\\
\tilde{\lambda}_G(0.109) &= \max(0.071; 0.000) \Longrightarrow \text{acceptance of } \tilde{H}_0 : \text{N-Eq } (\tilde{\alpha} \approx 0.05)\\
\tilde{\lambda}_G(0.120) &= \max(0.039; 0.000) \Longrightarrow \text{acceptance of } \tilde{H}_1 : \text{Eq } \quad (\tilde{\alpha} \approx 0.05)
\end{aligned}
$$

These results confirm that UI-NPC properly detects Eq; whereas IU-NPC manifests severe difficulties to detect a substantial Eq when it really exists.

By a normal approximation, naif $\ddot{T}_G(\varepsilon)$ rejects at $\varepsilon[s.t.\ \ddot{\alpha}(\varepsilon) = \tilde{\alpha}_G = 0.05] \approx 0.071$ (corresponding to $\approx 0.82\ \hat{\sigma}$). With the data, Eq is accepted if $\varepsilon_I = \varepsilon_S \gtrsim 1.38\ \hat{\sigma}$ (basically, this is an extremely poor result).

2. A psychological experiment on job satisfaction: data from Pesarin and Salmaso (2010), page 24.

Data are:

Extroverted:	$\mathbf{X}_1 =$	(66, 57, 81, 62, 61, 60, 73, 59, 80, 55, 67, 70)	$n_1 = 12$
Introverted:	$\mathbf{X}_2 =$	(64, 58, 45, 43, 37, 56, 44, 42)	$n_2 = 8$

Basic statistics are: $\bar{X}_1 = 65.92$; $\bar{X}_2 = 48.63$; $\hat{\sigma} = 8.93$.

For the one-sided (sharp) hypotheses $H_0 : X_1 \overset{d}{=} X_2$ V.s $H_1 : X_1 \overset{d}{>} X_2$, with $R = 100000$, PT $T = |\bar{X}_1 - \bar{X}_2|$ leads to $\hat{\lambda} = 0.00086$ for a substantial N-Eq.

The UI- and IU-NPC for equivalence using $\varepsilon_I = \varepsilon_S$, respectively give:

UI

$\varepsilon_I = \varepsilon_S$	$\hat{\lambda}_G$		Inference
8	0.019	$\Longrightarrow$	H_1 : N-Eq
10	0.048	$\Longrightarrow$	H_1 : N-Eq
11	0.074	$\Longrightarrow$	H_0 : Eq

IU

$\varepsilon_I = \varepsilon_S$	$\tilde{\lambda}_G$		Inference
22	0.136	$\Longrightarrow$	$\tilde{H}_0$: N-Eq
24	0.062	$\Longrightarrow$	$\tilde{H}_0$: N-Eq
25	0.035	$\Longrightarrow$	$\tilde{H}_1$: Eq

UI-NPC results permit declaring N-Eq for margins $\varepsilon \leq 10$ and Eq for larger.values.

As approximately adjusted $\tilde{\alpha}_h \approx 0.05$ for margins $\varepsilon \geq 15$, the IU-NPC results permit declaring N-Eq for margins $\varepsilon \leq 24$ and Eq for larger values, when data X lie in the range $\bar{X} \pm 2.7\hat{\sigma}$, i.e 59 ± 25.

The latter appears as a too wide range for any meaningful practical purpose.

6 Some references

1. Arboretti, R., Carrozzo, E. and Caughey, D. (2015). A rank-based permutation test for equivalence and noninferiority; *Italian Journal of Applied Statistics,* **25**(1): 81-92.

2. Basso, D., Pesarin, F., Salmaso, L. and Solari, A. (2009). *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications in R.* Lecture notes N. 194, Springer, New York.

3. Berger, R.L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*; **24:** 295-300.

4. Bertoluzzo, F., Pesarin, F. and Salmaso, L. (2013). On multi-sided permutation tests. *Communications in Statistics - Simulation and Computation*; **42(**6): 1380-1390.

5. Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*; **49:** 559-574.

6. Birnbaum, A. (1954). Characterization of complete classes of tests of some multiparametric hypotheses, with application to likelihood ratio tests. *Annals of Mathematical Statistics*; **26**: 21-36.

7. Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

8. Ferguson, T.S. (1967). *Mathematical Statistics, A Decision Theoretic Approach.* Academic Press, New York.

9. Frosini, B.V. (2004). On Neyman-Pearson theory: Information content of an experiment and a fancy paradox. *Statistica*, **64:** 271-286.

10. Hirotsu, C. (2004). *Statistical analysis for medical and pharmaceutical data: from data summarization to multiple comparisons for interactions*. University of Tokyo Press, Tokyo, in Japanese.

11. Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*; **23:** 169-192.

12. Hung, H.M.J. and Wang, S.U.(2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics*; **19**: 1–11.

13. Janssen, A. and Wellek, S. (2010). Exact linear rank tests for two-sample equivalence problems with continuous data. *Statistica Neerlandica*; **64(**4): 482–504.

14. Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* 2nd Edition. Wiley: New York, USA.

15. Marozzi, M. and Salmaso, L. (2006). Multivariate bi-aspect testing for the two-sample location problem. *Communication in Statistics—Theory and Methods*; **35**(3): 477–488.

16. Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurements*; **20**: 641-650.

17. Pantsulaia, G. and Kintsurashvili, M. (2014). Why is the null hypothesis rejected for "almost every" infinite sample by some hypothesis testing of maximal reliability. *Journal of Statistics: Advances in Theory and Applications*; **11:** 45-70.

18. Pesarin, F. (1990). On a nonparametric combination method for dependent permutation tests with applications. *Psychometrics and Psychosomatics.* **54:** 172-179.

19. Pesarin, F. (1992). A resampling procedure for nonparametric combination of several dependent tests. *Journal of the Italian Statistical Society*; **1:** 87-101.

20. Pesarin F. (2001). *Multivariate Permutation Tests, with Applications in Biostatistics*. Wiley: Chichester, UK.

21. Pesarin F. and Salmaso L. (2010). *Permutation Tests for Complex Data, Theory, Applications and Software*. Wiley: Chichester, UK.

22. Pesarin, F., Salmaso, L. (2010,a). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics*; **22**(5): 669–684.

23. Pesarin, F., Salmaso, L., (2013). On the weak consistency of permutation tests. *Communications in Statistics - Simulation and Computation*; **42**: 1368-1397.

24. Pesarin, F,. Salmaso, L., Carrozzo, E. and Arboretti, R. (2014). Testing for equivalence and non-inferiority: IU and UI tests within a permutation approach. *JSM 2014 - Section on Nonparametric Statistics.*

25. Pesarin, F,. Salmaso, L., Carrozzo, E. and Arboretti, R. (2016). Union-Intersection Permutation Solution for Two-Sample Equivalence Testing. *Statistics & Computing;* **26**: 693-701, DOI 10.1007/s11222-015-9552-y

26. Romano, J.P. (2005). Optimal testing of equivalence hypotheses. *Annals of Statistics*; **33**: 1036-1047.

27. Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*; **24:** 220-238.

28. Salmaso L., Solari A. (2005). Multiple aspect testing for case-control designs. *Metrika*; **62**: 331-340.

29. Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*; **15**: 657-680.

30. Sen, P.K. (2007). Union–intersection principle and constrained statistical inference. *Journal of Statistical Planning and Inference*; **137**: 3741–3752.

31. Sen, P.K. and Tsai, M.T. (1999). Two-stage likelihood ratio and union intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix. *Journal of Multivariate Analysis*; **68**: 264 282.

32. SenGupta, A. (2007). P^3 Approach to intersection-union testing of hypotheses. *Journal of Statistical Planning and Inference*; **137**: 3753-3766.

33. Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority.* Chapman & Hall/CRC: Boca Raton, USA.

34. Westfall, P.H., Ho, S.-Y. and Prillaman, B.A. (2001). Properties of multiple intersection-union tests for multiple endpoints in combination therapy trials. *Journal of Biopharmaceutical Statistics*; **11**(3): 125-138.

35. Wiens, B.L. (2006). Randomization as a basis for inference in noninferiority trials. *Pharmaceutical Statistics;* **5:** 265-271.

7 The nonparametric combination (NPC)

Let $\mathbf{X}_j = \{X_{j1}, \ldots, X_{jn_j}\}$ IID from $\mathcal{X}_Q$, $j = 1, 2$, be two independent Q-dimensional data sets, $Q \geq 1$, $n_1, n_2 \geq 2$, $n = n_1 + n_2$.

A representation of pooled data set

$$\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2 \in \mathcal{X}_Q^n \quad \text{is}$$

$$\mathbf{X} = \mathbf{X}^{(n)} = \{X(i),\ i = 1, \ldots, n;\, n_1, n_2\},$$

where first n_1 lines are of first sample and the rest of the second.

Let $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)$ be one permutation of $\mathbf{u} = (1, \ldots, n)$, the related permutation of $\mathbf{X}$ is:

$$\mathbf{X}^* = \{X^*(i) = X(u_i^*),\, i = 1, \ldots, n;\, n_1, n_2\}, \text{ and so}$$

$$\mathbf{X}_1^* = \{X_{1i}^* = X(u_i^*),\, i = 1, \ldots, n_1\} \text{ and}$$

$$\mathbf{X}_2^* = \{X_{2i}^* = X(u_i^*),\, i = n_1 + 1, \ldots, n\}$$

are the two permuted samples.

Suppose that H_0 implies $\mathbf{X}_1 \overset{d}{=} \mathbf{X}_2$, i.e. data exchangeability (in H_0 "sharp", $\mathbf{X}$ is always sufficient for any underlying F).

Moreover, suppose the hypotheses can be broken-down into $K \geq 1$ sub-hypotheses: H_{0k} V.s H_{1k}, $k = 1, \ldots, K$, so that

$$H_0 \equiv \bigcap_{k=1}^{K} H_{0k} \qquad \text{and} \qquad H_1 \equiv \bigcup_k H_{1k}.$$

Note: K can be smaller, equal or larger than Q.

Also assume that for each sub-hypothesis H_{0k} V.s H_{1k} a "separately unbiased" partial PT T_k is available, at least one of which is consistent.

The global hypotheses are then tested by combining K dependent partial PTs:

$$T_\psi = T_\psi(T_1, \ldots, T_K) \equiv \psi(\lambda_1, \ldots, \lambda_K),$$

where $\lambda_k = \Pr\{T_k^* \geq T_k^o | \mathbf{X}\}$ is the p-value-like statistic of T_k.

In accordance with the rule "large values are significant", each combining function ψ should satisfy:

- it is non-increasing in each argument: $\psi(.., \lambda_k, ..) \geq \psi(.., \lambda_k', ..)$ if $\lambda_k < \lambda_k'$;
- it must attain its supremum $\vec{\psi}$ if at least one argument attains 0;
- $\alpha > 0$ implies $T_{\psi\alpha} < \vec{\psi}$, i.e. no concentration at $\vec{\psi}$ under H_0.

A visualization for a two-sample Q-dimensional case and the NPC:

$X_1(1)$	$\cdots$	$X_1(n_1)$	$X_1(1+n_1)$	$\cdots$	$X_1(n)$		T_1^o
$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\longrightarrow$	$\vdots$
$X_Q(1)$	$\cdots$	$X_Q(n_1)$	$X_Q(1+n_1)$	$\cdots$	$X_Q(n)$		T_K^o

$X_1(u_1^*)$	$\cdots$	$X_1(u_{n_1}^*)$	$X_1(u_{1+n_1}^*)$	$\cdots$	$X_1(u_n^*)$		T_1^*
$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\longrightarrow$	$\vdots$
$X_Q(u_1^*)$	$\cdots$	$X_Q(u_{n_1}^*)$	$X_Q(u_{1+n_1}^*)$	$\cdots$	$X_Q(u_n^*)$		T_K^*

$\mathbf{X}$	$\mathbf{X}_1^*$	$\cdots$	$\mathbf{X}_r^*$	$\cdots$	$\mathbf{X}_R^*$
T_1^o	T_{11}^*	$\cdots$	T_{1r}^*	$\cdots$	T_{1R}^*
$\cdots$	$\cdots$		$\cdots$		$\cdots$
T_K^o	T_{K1}^*	$\cdots$	T_{Kr}^*	$\cdots$	T_{KR}^*

$$\downarrow$$

$\hat{\lambda}_1^o$	$\hat{L}_{11}^*$	$\cdots$	$\hat{L}_{1r}^*$	$\cdots$	$\hat{L}_{1R}^*$
$\cdots$	$\cdots$		$\cdots$		$\cdots$
$\hat{\lambda}_K^o$	$\hat{L}_{K1}^*$	$\cdots$	$\hat{L}_{Kr}^*$	$\cdots$	$\hat{L}_{KR}^*$

$$\Downarrow$$

$T_\psi^o(\mathbf{X})$	$T_{\psi 1}^*$	$\cdots$	$T_{\psi r}^*$	$\cdots$	$T_{\psi R}^*$

where $\hat{L}^*(T_{kr}^*) = \left[\frac{1}{2} + \sum_{j=1}^{R} \mathbf{I}(T_{kj}^* \geq T_{kr}^*)\right] / (R+1)$ is the ESF, an empirical measure in $(0,1)$; last line visualizes the (simulated) reference distribution of T_ψ.

7.1 Main NPC Properties

Combining functions ψ define a class $\mathcal{C}$ of possibilities. A sub-class $\mathcal{C}_A \subseteq \mathcal{C}$ contains admissible combining functions. A combining function ψ is admissible if its rejection region in the $(\lambda_1, \ldots, \lambda_K)$ representation is convex (Birnbaum, 1954, 1955).

Admissible combining functions mostly used in practice are:

$$\begin{aligned}
T_F^* &= -\textstyle\sum_k \log(L_k^*), \text{ Fisher's [the product rule];}\\
T_L^* &= \textstyle\sum_k \Phi^{-1}(1 - L_k^*), \text{ Liptak's [suitable if all } T_k^* \text{ are positively related];}\\
T_D^* &= \textstyle\sum_k T_k^*, \text{ the direct [suitable if all } T_k^* \text{ share the same limiting null distribution];}\\
T_T^* &= \max_k(1 - L_k^*), \text{ Tippett's [the best at each permutation, often } \equiv \max_k(T_k^*)\text{];}\\
T_G &= \max_k(1 - \lambda_k^o), \text{ the best observed partial [suitable when only one } H_{1k} \text{ can be true}\\
&\quad \text{or when some } T_k^* \text{ are negatively related; by reversing signs } \equiv \min_k(\lambda_k^o)\text{].}
\end{aligned}$$

P.1. NPC works with one-sample and multi-sample designs as well.

P.2. *If all K partial PTs are exact, T_ψ is exact $\forall \psi \in \mathcal{C}$.*

P.3. *If all K PTs are separately unbiased and positively dependent, T_ψ is unbiased $\forall \psi \in \mathcal{C}$.* The K partial tests are separately unbiased if $\Pr\{T_k^* \geq T_k^o, H_{1k}|\mathbf{X}\} \leq \alpha$, $k = 1, \ldots, K$.

P.4. *If all K PTs are separately unbiased, positively dependent and at least one is consistent* (for divergent sample sizes), *T_ψ is consistent $\forall \psi \in \mathcal{C}$.*

P.5. *Under mild conditions NPC satisfies the so-called "finite-sample consistency"*, that which occurs when K diverges while n_1 and n_2 are fixed (useful when $n < K$, with some stochastic processes and "omics", functional or shape data).

P.6. *NPC works even when different degrees of importance are assigned to the K sub-hypotheses*. E.g., if $w_k \geq 0$, $k = 1, \ldots, K$, with $w_k > 0$ for at least one k, Fisher's becomes $T^*_{FW} = -\sum_k w_k \cdot \log(L^*_k)$. When $w_k = w > 0$ an equivalent formulation of T_F occurs. In this context, T_G results as an *adaptive weighted Tipptt's rule*, where weights are: $w_k = 1$ if $k : \min_k(\lambda_k)$ and 0 elsewhere, so $T_G = \max_k(1 - L_k^{*w_k})$

P.7. *If* $0 < Var(X_k) < \infty$ *and* $T^*_k = \bar{X}^*_{1k} - \bar{X}^*_{2k}$, $k = 1, \ldots, K$, so that each partial test is asymptotically optimal (condition $0 < Var(X_k) < \infty$ is sufficient for PCLT), *combined test by any $\psi \in \mathcal{C}_A$ results as an admissible combination of asymptotically optimal tests*.

P.8. *NPC does not require knowledge of dependence coefficients among partial PTs*.

P.9. **NPC achieves Roy's UI** within a permutation framework.

Suppose: $n_1 = n_2 = 15$; $X \sim \mathcal{N}(0,1)$; $\varepsilon_I = \varepsilon_S = .75$; $\alpha_I = \alpha_S = .05$;

by simulation within permutation theory with $MC = 10000$ and $R = 4000$.

IU: $\tilde{\alpha}_G = .050$ ($\tilde{\alpha}_I = \tilde{\alpha}_S \approx .059$), max power (acceptances of $\tilde{H}_1$) $\approx .348$,

[Student's t max power is .355, numerically determined: Wellek, 2010, page 122],

[with $n_1 = n_2 = 10$ and $\varepsilon_I = \varepsilon_S = .25$ Student's t max power is .0584];

UI: $\alpha_G = .050$, max acceptances of $H_0 \approx 1.000$.

Within the permutation theory and the UI-NPC most restraints discussed in Sen (2007), plus several others, enjoy clear, effective and practical solutions.