Permutation Approach to the Testing of Hypotheses

Fortunato Pesarin

Department of Statistical Sciences – University of Padua

pesarin@stat.unipd.it

Padova, 9-05-2023

OUTLINE

1. Introduction
2. The Data Model
3. Conditionality, Sufficiency and PT Principles
4. Main Properties of Unidimensional PTs
5. On Multidimensional PTs
6. Some Examples
7. References

1 Introduction

i	Y_B	Y_A	X	i	Y_B	Y_A	X
1	19	14	5	11	16	17	-1
2	22	23	-1	12	25	20	5
3	18	13	5	13	22	18	4
4	18	17	1	14	19	17	2
5	24	20	4	15	27	22	5
6	30	22	8	16	23	21	2
7	26	30	-4	17	24	21	3
8	28	21	7	18	18	15	3
9	15	11	4	19	28	24	4
10	30	29	1	20	27	22	5

Data regard an experiment on subjects with anxiety: observed before and after an IPAT treatment.

The example (Landenna et Al., 1988) simulates a psychological experiment on $n = 20$ subjects to see if their anxiety can be reduced after an IPAT training treatment: pairs $(Y_{Bi}, Y_{Ai}) \equiv$ anxiety before and after treatment, $X_i = Y_{Bi} - Y_{Ai}, i = 1, \ldots, n$.

The hypotheses are $H_0 : \{(Y_{Bi} \stackrel{d}{=} Y_{Ai}), i = 1, \ldots, n\}$ V.s $H_1 : Y_B \stackrel{d}{>} Y_A$.

Popular solutions by Student's $t = \bar{X}\sqrt{n}/\hat{\sigma}$ and Wilcoxon's rank (with $R_i = \mathbb{R}(|X_i|)$, $\mathbf{w}_i = 1$ if $X_i > 0$, else $\mathbf{w}_i = 0$),

$T_W = [\sum_i R_i \mathbf{w}_i - n(n+1)/4]/[n(n+1)(2n+1)/24]^{1/2} \sim \mathcal{N}(0, 1)$,

due to discrete data (integers), to too many ties and to (possibly) Non Identical Distributions (of effects), i.e. $(X_i \stackrel{d}{\neq} X_j, i \neq j)$, are not applicable here.

So, we have to find a proper solution, possibly by a different approach.

Note, under H_0, *the data within each subject are exchangeable* (i.e. permutable): $(Y_{Bi}, Y_{Ai}) \stackrel{d}{=} (Y_{Ai}, Y_{Bi})$, $i = 1, \ldots, n$. Thus, a proper solution requires to take into consideration all possible data permutations: rather a tedious job (by hand).

R.A. Fisher (1936), the main author of the permutation approach, wrote: "*the statistician does not carry out* (by hand) *this very simple and very tedious process*, *but his conclusions have no justification beyond the fact that they agree with those which could have arrived at by this elementary method*".

Apparently, Fisher seems considering the traditional parametric testing with the role of approximating the null permutation distribution.

In the last 2 decades permutation testing methods have increased both in number of applications and in solving complex and difficult multivariate problems.

When available *permutation tests* (PTs) are essentially exact *nonparametric* (NP) tools in a conditional context; *the conditioning being on the observed data that are always a set of sufficient statistics* for whatever the underlying distribution P.

The application of the *conditionality principle* (CP) provides the PT approach with nice, useful and easy to check properties under mild assumptions; so, making it effectively and easily applicable.

Except for rather simple situations, in practice the reference null distribution of most parametric tests is known only asymptotically.

There are several complex multivariate problems (common in biostatistics, clinical trials, engineering, epidemiology, experimental data, industrial statistics, marketing, pharmacology, psychology, quality control, social sciences, etc.) that are difficult, or even impossible, to solve outside the CP, especially outside the *nonparametric combination* (NPC) of dependent PTs.

Actually, parametric methods reflect a modelling approach and typically require a set of rather stringent assumptions, which are often difficult to justify.

These are often set on an *ad hoc* basis: sometimes researchers assume, without any justification (typically because it is easier to work with): multivariate normality; random sampling from a target population; homoscedasticity of responses also under the alternative, where the treatment effect might modify more than one aspect of the distribution; random effects independent of units; etc.

So consequent inferences have no real credibility.

On the contrary, NP approaches try to keep assumptions at a lower workable level, avoiding those that are difficult to justify.

Thus, they are based on more realistic foundations, are intrinsically robust and so consequent inferences are credible.

My point of view is that any statistician should have in his tool kit of methods both the parametric, including the Bayesian, and the NP, because in his life he surely meets with problems that are difficult, or even impossible, within one approach and others that are difficult, or even impossible, within the other approach.

Examples on the subject matter are in the book of mine, joint with L. Salmaso, *Permutation Tests for Complex Data, Theory, Applications and Software*; Wiley, Chichester, UK, (2010).

2 The Data Model

Let us refer to the two-sample one-dimensional layout (extensions are straightforward).

Assume that a non-degenerate variable X takes values on sample space $\mathcal{X}$, and associated with $(X, \mathcal{X})$ there are distributions P belonging to a NP family $\mathcal{P}$.

"*A family $\mathcal{P}$ of distributions is NP when it is not possible to find a finite-dimensional space* Θ (the parameters' space) *such that* (s.t.) *there is a one-to-one relationship between* Θ *and* $\mathcal{P}$, *so that each member* P *of* $\mathcal{P}$ *cannot be identified by only one member* θ *of* Θ, *and vice versa*."

Each $P \in \mathcal{P}$ gives the probability measure to events A belonging to a suitable collection (an algebra) $\mathcal{A}$ of events.

It is assumed that each family $\mathcal{P}$ admits the existence of one dominating measure $\xi_{\mathcal{P}}$ s.t. the density $f_P(X) = \mathrm{d}P(X)/\mathrm{d}\xi_{\mathcal{P}}$ is unambiguously defined.

Let $\mathbf{X}_j = \{X_{ji},\ i = 1, \ldots, n_j\} \in \mathcal{X}^{n_j}$ be the *independent and identically distributed* (IID) data sized n_j from $P_j \in \mathcal{P}$, $j = 1, 2$.

A notation for data sets with independent samples is

$$\mathbf{X} = \{X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}\} \in \mathcal{X}^n,$$

the model of which is $(\mathbf{X}, \mathcal{X}^n, \mathcal{A}^{(n)}, P^{(n)} \in \mathcal{P}^{(n)})$, with $n = n_1 + n_2$, and where $P^{(n)} = P_1^{n_1} \cdot P_2^{n_2}$.

To denote sample data in the permutation context we may also use the *unit-by-unit* representation:

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

where it is intended: first n_1 data in the list belong to first sample; the rest to the second.

Denoting by $\Pi(\mathbf{u})$ the set of permutations of unit labels $\mathbf{u} = (1, \ldots, n)$ and by $\mathbf{u}^* = (u_1^*, \ldots, u_n^*) \in \Pi(\mathbf{u})$ one of these permutations, the related permutation of $\mathbf{X}$ is:

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

hence,

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

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

denote the two permuted samples.

For notational convenience, the pooled set is also denoted by $\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2 \in \mathcal{X}^n$.

We discuss problems for stochastic dominance alternatives (one-sided) as are generated by non-negative random shift effects Δ.

The alternative assumes that treatments produce effects Δ_1 and Δ_2 with $\Delta_1 \overset{d}{>} \Delta_2$. Thus, the hypotheses under testing are

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} X \equiv P_1 = P_2, \text{ and } H_1 : (X_1 + \Delta_1) \overset{d}{>} (X_2 + \Delta_2).$$

Note: under H_0 data of two samples are *exchangeable* (units randimized to traetment).

Without loss of generality, effects under H_1 are such that:

$$\Delta_1 = \Delta \overset{d}{>} 0 \quad \text{and} \quad \Pr\{\Delta_2 = 0\} = 1.$$

The latter agrees with the notion that an *active treatment* is only assigned to units of first sample and a *placebo* to those of the second.

Moreover, we may let Δ to depend on units and on related null responses, thus pairs (X_{1i}, Δ_i), $i = 1, \ldots, n_1$, satisfy the relation $(X_{1i} + \Delta_i) \geq X_{1i}$ with *at least one strict inequality*.

Hence, $(X_1 + \Delta) \overset{d}{>} X_2 = X$, due to treatment effect, is compatible with non-homoscedasticities under H_1. The null hypothesis can also be written as $H_0 : \Delta \overset{d}{=} 0$.

Other than measurability, no further distributional assumption on effects Δ is required.

In particular it is neither required that its mean value is finite, i.e. $\mathbf{E}(|\Delta|) < \infty$, nor that its distribution is symmetric.

To emphasize the roles of sample sizes and effects, we use

$$\mathbf{X}(\Delta) = \{X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}\}$$

to denote the actual sample data; thus, $\mathbf{X}(0)$ denotes data under H_0.

Note that the pooled data $\mathbf{X}$, if $f_P^{(n)}(\mathbf{X}) > 0$, is always a set of sufficient statistics for P, since $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}) = 1$; so the conditional distribution of $\mathbf{X}$ given $\mathbf{X}$ is independent of P (formally, $\mathbf{X}$ is also a set of ancillary (i.e. complementary) statistics; the notion of ancillarity used in NP is based on the fact that once a statistic $T(\mathbf{X})$ is considered, some useful complementary and possibly confounded information on P is still remaining in $[\mathbf{X}|T(\mathbf{X})]$).

When P is NP or the number of its parameters is larger than sample size or in most cases in which it lies outside the regular exponential family, $\mathbf{X}$ is *minimal sufficient*.

PT lie within the conditional method of inference, the conditioning is on the actually observed data $\mathbf{X}$. The related conditional reference space is denote by $\mathcal{X}^n_{/\mathbf{X}}$.

Essentially $\mathcal{X}^n_{/\mathbf{X}}$, or simply $\mathcal{X}_{/\mathbf{X}}$, contains points of $\mathcal{X}^n$ which are equivalent to $\mathbf{X}$ in terms of information carried by the underlying likelihood. Thus, it contains all points $\mathbf{X}^*$ such that the likelihood ratio $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}^*)$ is P-independent; so it corresponds to the *orbit* of equivalent points associated with $\mathbf{X}$.

Since under H_0 the density $f_P^{(n)}(\mathbf{X}) = \prod_{ji} f_P(X_{ji})$ is exchangeable by assumption, and so $f_P^{(n)}(\mathbf{X}) = f_P^{(n)}(\mathbf{X}^*)$ for every permutation $\mathbf{X}^*$ of $\mathbf{X}$, then $\mathcal{X}^n_{/\mathbf{X}}$ contains all distinct permutations of $\mathbf{X}$.

Note: the data exchangeability condition is required only under H_0.

The conditional reference space is defined as

$$\mathcal{X}_{/\mathbf{X}} = \{\cup_{\mathbf{u}^* \in \Pi(\mathbf{u})}[X(u_i^*), i = 1, \ldots, n]\}.$$

Since for $\forall \mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$ it is $\mathcal{X}_{/\mathbf{X}^*} = \mathcal{X}_{/\mathbf{X}}$, therefore $\mathcal{X}_{/\mathbf{X}}$ is a sufficient space.

Moreover, since $\forall A \in \mathcal{A}$ the conditional probability $\mathsf{Pr}(A, P|\mathbf{X}) = \mathsf{Pr}(A|\mathcal{X}_{/\mathbf{X}})$ is P-independent, the data $\mathbf{X}$ can also be considered as playing the role of a set of ancillary statistics for the problem.

Of course, when the n-$\mathcal{D}im$ $\mathbf{X}$ is *minimal sufficient* it is also *maximal ancillary*.

The fact that random effects $\mathbf{\Delta}$ may depend on null errors X is an important improvement with respect to traditional parametric approaches.

On the one hand, this leads to assumptions that are much more flexible and much closer to reality. E.g., this occurs with data obtained by physical instruments of measurement based on nonlinear monotonic transformations of underlying variables (indirect measurements).

Actually, let $X = \varphi(Y)$ be a (regular) monotonic non-decreasing transformation, so

$$X(\delta) = \varphi(Y + \delta) = \varphi(Y) + \delta\varphi'(Y + \delta h), \text{ with } 0 < h < 1,$$

to show that random effects $\Delta = \delta\varphi'(Y + \delta h)$, unless φ is linear, are dependent on *null deviates* $Y = \varphi^{-1}(X)$ and are non-homoscedastic even for fixed δ.

On the other, it is to be noted that in permutation analysis the separate estimate of variance components is generally not required.

Hence, the statistical modeling may better fit researcher's requirements.

In the NP framework, more than on parameters, the inferential interest is generally focused on functionals, i.e. functions of all parameters such as: the effect Δ; the divergence of means $\mathbb{E}(X_1 - X_2)$; etc.

PTs do not require to separate the role of parameters of interest from the nuisance ones. Such a separability property is essentially required by parametric methods.

Also note that when the n-dimensional $\mathbf{X}$ is minimal sufficient, then univariate statistics capable of summarizing the whole information do not exist (actually, looking for optimal solutions with finite sample sizes becomes impossible).

So, on the one hand, no parametric or nonparametrric method can claim to be uniformly better than others. However, by conditioning on $\mathcal{X}_{/\mathbf{X}}$, every PT counterpart of any unbiased test statistic improves its power behavior (via Rao-Blackwell).

On the other, in order to reduce the loss of information associated with using only one single overall statistic $T(\mathbf{X})$ and in connection with the notion of NP ancillarity related to $[\mathbf{X}|T(\mathbf{X})]$, it is possible to take account of a list of complementary test statistics, each able to summarizing information on a specific aspect of interest for the analysis, and so to find solutions within the so-called *multi-aspect methodology* based on the NPC of several dependent PTs.

3 Sufficiency, Conditionality and PT Principles

The SP essentially states that:

"Suppose that we are working with the model $f_X(x, \theta)$ for the random variable X, according to which the data set $\mathbf{X}$ is observed, and also suppose that the statistic S is minimal sufficient for $\theta \in \Theta$.

Then, according to the SP, so long as we accept the adequacy of the model $f_X(x, \theta)$, identical conclusions should be drawn from data $\mathbf{X}_1$ and $\mathbf{X}_2$ with the same value of S."

The CP states that:

"*Suppose that C is an ancillary statistic for the problem, then any conclusion about the parameter or the functional of interest is to be drawn as if C were fixed at its observed value.*"

The rationale for these principles in statistical inference considers problems like the following: suppose that data $\mathbf{X}$ can be obtained by means of one of two different measuring instruments, M_1 and M_2, and suppose the associated normally distributed models are respectively $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, with $\sigma_1 \ll \sigma_2$.

If it is known which instrument has generated $\mathbf{X}$ it is unavoidable to condition on the related (ancillary) model in any inference regarding μ, the value of σ being either known or unknown.

Moreover, in accordance with the SP, the statistical estimator of unknown μ should be based on a, possibly *minimal complete*, sufficient statistic for it. In addition, if the nuisance parameter σ is unknown, it is wise to stay at least on invariant statistics or on the notion of invariance of null rejection probability (according to the notion of similarity) with respect to it and so to condition on a possibly minimal sufficient statistic (via Rao-Blackwell).

Indeed, by acting outside these principles related inferential conclusions can be biased, misleading and maybe difficult to be correctly interpreted.

Thus, when P is unknown, or it is too complex to deal with or its parameters are unspecified or are infinitely many, it is wise to condition on its minimal sufficient statistic, i.e. the n-dimensional data $\mathbf{X}$, which is always sufficient for whatever $P \in \mathcal{P}$ and ancillary for the inferential problem.

This kind of conditioning implies referring to the PT principle:

"*If two experiments, taking values on the same sample space* $\mathcal{X}$ *with underlying distributions* P_1 *and* P_2 *give the same data* $\mathbf{X}$, *then two inferences conditional on* $\mathbf{X}$ *and obtained by using the same statistic* T *must be the same*, *provided that the exchangeability of the data is satisfied under* H_0."

Since both principles are satisfied, it should be emphasized that the PT principle works in accordance with CP and SP.

4 Main Properties of Unidimensional PTs

Suppose that large values of test statistics $T : \mathcal{X}^n \to \mathcal{R}^1$ are evidence against H_0.

P.1. *Sufficiency of $\mathcal{X}_{/\mathbf{X}}$ for P implies that the conditional probability of every event $A \in \mathcal{A}$, given $\mathcal{X}_{/\mathbf{X}}$, is independent of P; i.e.*

$$\mathsf{Pr}\{\mathbf{X}^* \in A; P|\mathcal{X}_{/\mathbf{X}}\} = \mathsf{Pr}\{\mathbf{X}^* \in A|\mathcal{X}_{/\mathbf{X}}\}.$$

Note: the plain (basic) definition of sufficiency is applied.

Thus, since for finite n the number $M = M^{(n)} = \sum_{\mathcal{X}_{/\mathbf{X}}} \mathbb{I}(\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}})$ of points in $\mathcal{X}_{/\mathbf{X}}$ is finite, the conditional probability of any $A \in \mathcal{A}$ is

$$\Pr\{\mathbf{X}^* \in A | \mathcal{X}_{/\mathbf{X}}\} = \frac{\sum_{\mathbf{X}^* \in A} f_P(\mathbf{X}^*) \mathrm{d}\mathbf{X}^*}{\sum_{\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}} f_P(\mathbf{X}^*) \mathrm{d}\mathbf{X}^*} = \sum_{\mathcal{X}_{/\mathbf{X}}} \frac{\mathbb{I}(\mathbf{X}^* \in A)}{M},$$

because $\forall \mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$ it is $f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* = f_P(\mathbf{X})\mathrm{d}\mathbf{X}$ (and > 0).

Note:
a) it is not necessary to take recourse to the *hypothetical repeated sampling principle* (no frequentist interpretation).
b) PTs only require *the existence of a likelihood* f_P (not its calculability!).
c) *The data set* $\mathbf{X}$ *is uniformly distributed over* $\mathcal{X}_{/\mathbf{X}}$ *conditionally*, i.e. *all permutations are equally likely over* $\mathcal{X}_{/\mathbf{X}}$ (under H_0 and H_1).

P.2. *If* $T : \mathcal{X}^{(n)} \to \mathcal{R}^1$ *is any measurable statistic* (e.g. a test) *then its permutation distribution* (given $\mathcal{X}_{/\mathbf{X}}$) *is independent of* P; *i.e.*

$$\mathsf{Pr}\{T(\mathbf{X}^*) \leq t, P|\mathcal{X}_{/\mathbf{X}}\} = \mathsf{Pr}\{T(\mathbf{X}^*) \leq t|\mathcal{X}_{/\mathbf{X}}\} = \sum_{\mathcal{X}_{/\mathbf{X}}} \frac{\mathbb{I}(T(\mathbf{X}^*) \leq t)}{M}.$$

Consequence C1): the permutation (conditional) distribution of $T(\mathbf{X}^*)$ under H_0 may differ from that under H_1 (see later).

Consequence C2): *Let* $\mathbf{T} = (T_1, \ldots, T_K)$ *be* $K \geq 1$ *real statistics* (e.g., tests) *and let* $\psi : \mathcal{R}^K \to \mathcal{R}^1$ *be any measurable function, then the conditional –permutation– distribution of* ψ *is independent of the underlying population distribution* P; *specifically, it is independent of all nuisance parameters underlying* $\mathbf{T}$, *including in particular those related to any kind of dependence relations on* $\mathbf{T}$ (Sen, 2007). *Indeed* :

$$\mathsf{Pr}\{\psi(T_1^*,\ldots,T_K^*) \leq t; P|\mathcal{X}_{/\mathbf{X}}\} = \mathsf{Pr}\{\psi(T_1^*,\ldots,T_K^*) \leq t; |\mathcal{X}_{/\mathbf{X}}\} =$$

$$= \mathsf{Pr}[\psi_{\mathbf{T}}^{-1}(t)|\mathcal{X}_{/\mathbf{X}}] = \frac{\sum_{\mathcal{X}_{/\mathbf{X}}} \mathbb{I}[\mathbf{X}^* \in \psi_{\mathbf{T}}^{-1}(t)]}{M} \quad ,$$

since, by measurability of ψ *the inverse image* $\psi_{\mathbf{T}}^{-1}(t)$ *of* $(-\infty, t]$ *is, for every* $t \in \mathcal{R}^1$, *an event belonging to* $\mathcal{A}_{/\mathbf{X}}$, *the measure of which is the count provided by P.1*.

Consequence C2) is essentially the main formal result useful for multivariate and multi-aspect PT.

The figure provides a sketch for the NPC

$\mathbf{X}$	$\mathbf{X}_1^*$	$\cdots$	$\mathbf{X}_r^*$	$\cdots$	$\mathbf{X}_R^*$
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}^*
$T_\varphi^o(\mathbf{X})$	$T_{\varphi 1}^*$	$\cdots$	$T_{\varphi r}^*$	$\cdots$	$T_{\varphi R}^*$

Note:
- the conditional probability has always an *objective existence*;
- when $\mathbf{X}$ is minimal sufficient, it makes no sense to work outside the PT P;
- PTs are nonparametric, distribution-free and intrinsically robust;
- P2, C2) is the central property for the NPC: all dependences are worked out;
- *some* functions φ are useful for testing *multi-aspect* hypotheses.

P.3. (Uniform similarity) *Assume that the exchangeability condition on data* $\mathbf{X}$ *is satisfied under* H_0*, then the conditional rejection probability* $\mathbb{E}\{\phi_R(\mathbf{X})|\mathcal{X}_{/\mathbf{X}}\}$ *of randomized test*

$$\phi_R(\mathbf{X}) = \begin{cases} 1 & \text{if} \quad T^o > T_\alpha \\ \gamma & \text{"} \quad T^o = T_\alpha \\ 0 & \text{"} \quad T^o < T_\alpha \end{cases} ,$$

is such that, $\forall \alpha \in (0,1)$*, it is* $\mathbb{E}\{\phi_R(\mathbf{X})|\mathcal{X}_{/\mathbf{X}}\} = \alpha$*; thus* ϕ_R *is* $\mathbf{X}$*-*P*-invariant for all* $\mathbf{X} \in \mathcal{X}^n$ *and all* $P \in \mathcal{P}$*, where:* $T^o = T(\mathbf{X})$ *is the observed value value of* T *on data* $\mathbf{X}$*,* $T_\alpha = T_\alpha(\mathbf{X}(0))$ *is the* α*-size conditional critical value which could be determined by complete enumeration of* $\mathcal{X}_{/\mathbf{X}}$ *when it is known that* $\Delta = 0$ *and*

$$\gamma = \left[\alpha - \mathsf{Pr}\left\{T^o > T_\alpha|\mathcal{X}_{/\mathbf{X}}\right\}\right] \Big/ \mathsf{Pr}\left\{T^o = T_\alpha|\mathcal{X}_{/\mathbf{X}}\right\}.$$

The p-*value statistic* $\lambda = \lambda_T(\mathbf{X}) = \mathsf{Pr}\{T^* \geq T^o | \mathcal{X}_{/\mathbf{X}}\}$ is a non-increasing function of T^o and is one-to-one related with the attainable α-value of T, i.e. $\lambda_T(\mathbf{X}) > \alpha$ implies $T^o < T_\alpha$, and vice versa.

Hence, the non-randomized version is stated as

$$\phi = \begin{cases} 1 & \text{if} \quad \lambda_T(\mathbf{X}) \leq \alpha \\ 0 & \text{"} \quad \lambda_T(\mathbf{X}) > \alpha \end{cases} ,$$

thus, under H_0 it is: $\mathbb{E}\{\phi(\mathbf{X}) | \mathcal{X}_{/\mathbf{X}}\} = \mathsf{Pr}\{\lambda_T(\mathbf{X}) \leq \alpha | \mathcal{X}_{/\mathbf{X}}\} = \alpha$ for every *attainable* $\alpha \in (0, 1)$.

Note: λ were a *genuine* p-*value* only if H_0 were true ($\Delta = 0$).

P.4. *Based on* **P.1**, *if* X *is a continuous variable and* T *is a continuous non-degenerate function, then* p*-value statistic* $\lambda_T(\mathbf{X})$ *under* H_0 *is uniformly distributed over its attainable support*.

P.5. *A PT* T *is exact if its null distribution essentially depends on exchangeable null error deviates* $\mathbf{X}$ *only.*

P.6. (Uniform unbiasedness) *PT for random shift alternatives* $(\boldsymbol{\Delta} \overset{d}{\geq} 0)$ *based on divergence of symmetric statistics of non-degenerate measurable non-decreasing transformations of the data, i.e.* $T^*(\boldsymbol{\Delta}) = S_1[\mathbf{X}_1^*(\boldsymbol{\Delta})] - S_2[\mathbf{X}_2^*(\boldsymbol{\Delta})]$, *where* $S_j(\cdot)$, $j = 1, 2$, *are symmetric functions of their entry arguments* $(\cdot)$, *are conditionally unbiased for every attainable* α, *every population distribution* P, *and uniformly for all data sets* $\mathbf{X} \in \mathcal{X}^n$. *In particular*

$$\Pr\{\lambda(\mathbf{X}(\boldsymbol{\Delta})) \leq \alpha | \mathcal{X}_{/\mathbf{X}(\boldsymbol{\Delta})}\} \geq \Pr\{\lambda(\mathbf{X}(0)) \leq \alpha | \mathcal{X}_{/\mathbf{X}(0)}\} = \alpha;$$

thus, the p-value statistic under H_1 is stochastically dominated by that under H_0; actually, it is $\lambda(\mathbf{X}(\boldsymbol{\Delta})) \overset{d}{\leq} \lambda(\mathbf{X}(0))$.

Of course, if $\boldsymbol{\Delta}' \overset{d}{>} \boldsymbol{\Delta} \overset{d}{>} 0$ then $\lambda(\mathbf{X}(\boldsymbol{\Delta}')) \overset{d}{\leq} \lambda(\mathbf{X}(\boldsymbol{\Delta})) \overset{d}{\leq} \lambda(\mathbf{X}(0))$.

Observe that uniform similarity (**P.3**) and uniform unbiasedness (**P.6**) require the exchangeability under H_0; thus, random sampling from a population is not required. Actually, they properly work also with the *selection-bias samples*, that are much more frequently met in practice.

P.7. *For each permutation* $\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$, *the Empirical Probability Measure (EPM) of any* $A \in \mathcal{A}$ *is defined as* $\hat{P}_{\mathbf{X}^*}(A) = \sum_{i \leq n} \mathbb{I}(X_i^* \in A)/n$ *which, since* $\forall \mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$ *it is* $\sum_{i \leq n} \mathbb{I}(X_i^* \in A) = \sum_{i \leq n} \mathbb{I}(X_i \in A)$, *is a permutation invariant function over* $\mathcal{X}_{/\mathbf{X}}$: $\hat{P}_{\mathbf{X}^*}(A) = \hat{P}_{\mathbf{X}}(A)$.

The latter implies that conditioning on $\mathcal{X}_{/\mathbf{X}}$ is equivalent to conditioning on the EPM $\hat{P}_{\mathbf{X}}$, which then is a *sufficient function*.

P.8. *The* (unconditional or population) *power of a PT* T, *as a function of* Δ, α, T, P, *and* n, *is defined as*

$$W(\Delta, \alpha, T, P, n) = \mathbb{E}_{P^n}[\Pr\{\lambda_T(\mathbf{X}(\Delta)) \leq \alpha | \mathcal{X}^n_{/\mathbf{X}}\}].$$

Of course, $W(\Delta, \alpha, T, P, n) \geq W(0, \alpha, T, P, n) = \alpha$, $\forall \alpha > 0$, *since, by* **P.6,** *the integrand is* $\geq \alpha$ *for all* $\mathbf{X} \in \mathcal{X}^n$, *all* $P \in \mathcal{P}$ *and all* n.

Clearly, unconditional unbiasedness in **P.8,** requiring the whole model $(\mathcal{X}, P^n)$, imply recourse to the *hypothetical repeated sampling principle*, and so it is in accordance with a frequentist interpretation.

To state the weak consistency of a test T, i.e."*if* $\Delta \overset{d}{>} 0$, *as* $\min[n_1, n_2] \to \infty$ *the rejection probability of test* T *tends to one for all* $\alpha > 0$", let us consider sequences of data where first n_1 IID values are from $X_1(\Delta) = X + \Delta$ and the other n_2 from $X_2 = X(0) = X$.

Denote such sequences as

$$\{\mathbf{X}^{(n)}(\Delta)\}_{n\in\mathbb{N}} = \{[X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}]\}_{(n_1,n_2)\in\mathbb{N}}.$$

Of course, $\{\mathbf{X}^{(n)}(0) = \mathbf{X}^{(n)}\}_{n\in\mathbb{N}}$ represents sequences under H_0.

Besides, we assume that $n \to \infty$ implies $\min[n_1, n_2] \to \infty$.

P.9. *Let X be any population variable and suppose that $\{\mathbf{X}^{(n)}(\Delta)\}_{n\in\mathbb{N}}$ is a sequence of data the first n_1 IID from $(X_1(\Delta), \mathcal{X})$ and independently the other n_2 IID from $(X, \mathcal{X})$. Suppose that the null distribution of X is $P \in \mathcal{P}$, and let $\varphi : \mathcal{X} \to \mathcal{R}^1$ be any non-decreasing and non-degenerate measurable function.*

Also, *suppose that*:

a) *the φ-mean $\mathbb{E}_P[\varphi(X)] = \mathbb{E}_P[\varphi(X(0))]$ is s.t. $\mathbb{E}_P[|\varphi(X)|] < +\infty$;*

b) *the finite φ-mean under H_1 is s.t. $\mathbb{E}_P[\varphi(X(\Delta))] > \mathbb{E}_P[\varphi(X(0))]$ for every $\Delta \overset{d}{>} 0$ and $\Delta' \overset{d}{>} \Delta$ implies $\mathbb{E}_P[\varphi(X(\Delta'))] > \mathbb{E}_P[\varphi(X(\Delta))]$;*

c) *the PT is based on $T^* = \sum_{i\leq n_1} \varphi(X_i^*)/n_1$ or on equivalent statistics.*

Then, for every $\alpha > 0$, **a), b)**, *and* **c)** *imply that the rejection probability of the PT* ϕ, *associated with* T^*, *converges weakly to one as* $n \to \infty$.

Note that population variable X can be either real or ordered categorical, and that its transformation $\varphi(X)$ is real, i.e. continuous, discrete, or mixed.

4.1 An algorithm for two-sample permutation tests

A *conditional Monte Carlo* (CMC) algorithm for evaluating the p-value-like λ of a test statistic T on two-sample data $\mathbf{X}(\boldsymbol{\Delta}) = \{X(i; \boldsymbol{\Delta}_i), i = 1, \ldots, n;\, n_1, n_2\}$ is based on the steps:

1) Calculate, on the data $X(\boldsymbol{\Delta})$, the observed value $T^o(\boldsymbol{\Delta}) = T[X(\boldsymbol{\Delta})]$, and take note.
2) Generate a random permutation $X^*(\boldsymbol{\Delta})$ of $X(\boldsymbol{\Delta})$.
3) Calculate $T^*(\boldsymbol{\Delta}) = T(X^*(\boldsymbol{\Delta}))$ and take note.
4) Independently repeat R times steps 2 and 3.
5) The estimated p-value-like is $\hat{\lambda}_T[\mathbf{X}(\boldsymbol{\Delta})] = \sum_{1 \le r \le R} \mathbb{I}[T_r^*(\boldsymbol{\Delta}) \ge T^o(\boldsymbol{\Delta})]/R$.

For testing on paired observations replace step 2) with: *for each of the n differences in $\mathbf{X}$, consider a random assignment of signs S^*, obtaining the data permutation* $\mathbf{X}^* = \{X_i \cdot S_i^*, i = 1, \ldots, n\}$.

A simple power comparison of Student's t with its permutation counterpart on one-sample normally distributed data, at $\alpha = 0.05$, with: $X \sim N(\delta, \sigma^2)$, $\sigma^2 = 1$ assumed unknown, $R = 2500$ CMC and $MC = 5000$ Monte Carlo runs gave:

n	$\delta \rightarrow$	0.1	0.2	0.4	0.8
10	t / T^*	0.096 / 0.087	0.163 / 0.148	0.351 / 0.324	0.778 / 0.750
20	t / T^*	0.115 / 0.108	0.217 / 0.207	0559 / 0.543	0.967 / 0.965

.

Results that show the very small lack of power in favour of the UMPU invariant –similar– Student's t, which is conditional on the "minimal" sufficient statistic for σ under H_0, over its permutation counterpart T^*, which instead is conditional on the "maximal" sufficient statistic for P.

5 On Multidimensional PTs

According to Roy's (1953) Union-Intersection way, suppose the hypotheses H_0 and H_1 are **equivalently broken-down** into $K \geq 2$ sub-hypotheses: H_{0k} V.s H_{1k}, $k = 1, \ldots, K$, so:

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

thus, the problem implies using one test for each sub-hypotheses H_{0k} V.s H_{1k}.

Note: K sub-hypotheses can be one- and/or two-sided, simple and/or composite. Importantly, K can be not related with n, $\# V$ of variables and $\mathcal{D}_{im}(\Theta)$.

Assume that for each H_{0k} V.s H_{1k} a *separately unbiased* partial PT T_k, significant for large values, is available.

Let δ_k be the **effect detectable** by the PT T_k. So, δ_k is a *functional* dependent on all underlying parameters defining F.

T_k is separately -marginally- unbiased if $\forall \alpha > 0$, $\forall \delta_k \in H_{1k}$ and independently of all partial effects except for δ_k, i.e. $\boldsymbol{\delta}_{\setminus k} \in \bigcup_{h \neq k} \delta_h$, its p-value statistic satisfies:

$$\Pr\{\lambda_k(\mathbf{X}(\delta_k, \boldsymbol{\delta}_{\setminus k})) \leq \alpha | \mathbf{X}\} \geq \Pr\{\lambda_k(\mathbf{X}(0, \boldsymbol{\delta}_{\setminus k})) \leq \alpha | \mathbf{X}\} = \alpha,$$

consequently λ_k doesn't increase if δ_k increases.

Also assume:

a) *at least one test* T_k, $k = 1, \ldots, K$, *is consistent*;

b) T_k *is* **monotonically** *dependent with any other test* T_h, $h \neq k$;

c) *If* b) *fails, some NPCs might fail.*

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

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

To be suitable for multidimensional testing purposes and assuming the rule *large values are significant*, **non-degenerate measurable combining functions ψ have to satisfy :**

1. are non-increasing in each argument: $\psi(.., \lambda_k, ..) \leq \psi(.., \lambda'_k, ..)$ if $\lambda_k > \lambda'_k$;
2. each ψ must attain its supremum $\vec{\psi}$ if at least one argument attains 0;
3. $\alpha > 0$ implies $T_{\psi\alpha} < \vec{\psi}$ asymptotically; i.e. no concentration at $\vec{\psi}$ under H_0.

With reference to positive one-sided partial alternatives, of specific interest for the UI-NPC is the notion that the K partial tests are **positively dependent** (Lehmann, 1986).

This implies that $\forall \delta_k \in H_{1k}$ and $\forall(\boldsymbol{\delta}'_{\setminus k} > \boldsymbol{\delta}_{\setminus k})$ the following condition holds

$$\lambda_k(\mathbf{X}(\delta_k, \boldsymbol{\delta}'_{\setminus k})) \overset{d}{\leq} \lambda_k(\mathbf{X}(\delta_k, \boldsymbol{\delta}_{\setminus k})),$$

i.e. the kth p-value statistic does not increase in distribution when some other effects $\boldsymbol{\delta}_{\setminus k}$ increase in at least one component.

5.1 Main NPC combining functions

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

Admissible combining functions mostly used in practice are:

$$
\begin{array}{rcl}
T_F^* & = & -2\sum_k \log(\lambda_k^*), \text{ Fisher's [omnibus, i.e. the product rule];} \\
T_L^* & = & \sum_k \Phi^{-1}(1-\lambda_k^*), \text{ Liptak-Stouffer's [suitable if all } T_k^* \text{ are positively related];} \\
T_D^* & = & \sum_k T_k^*, \text{ the direct [suitable if all } T_k^* \text{ share the same limiting null distribution];} \\
T_T^* & = & \max_k(1-\lambda_k^*), \text{ Tippett's [the best at each permutation; often} \equiv \max_k(T_k^*)\text{];} \\
T_G^o & = & \max_k(1-\lambda_k^o), \text{ the best observed partial [suitable when only one } H_{1k} \text{ can be true} \\
 & & \text{or when some } T_k^* \text{ are negatively related; often } T_G^o \equiv \min_k(\lambda_k^o)\text{] .}
\end{array}
$$

5.2 Main NPC Properties

- **NP.1.** UI-NPC works with one-sample, multi-sample and other complex designs.

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

• **NP.3.** *If all K PTs are separately unbiased and positively dependent* (Lehmann, 1986), *T_ψ is unbiased $\forall \psi \in \mathcal{C}$.*

• **NP.4.** *If all K PTs are separately unbiased, positively dependent and at least one is consistent* (for divergent sample sizes), *T_ψ is consistent $\forall \psi \in \mathcal{C}$.* If some PTs were negatively dependent, then T_ψ-unbiasedness may not be true and consistency can be valid only for a subset of combining functions: $\psi_C \in C_C \subset C$.

• **NP.5.** *Under mild conditions UI-NPC satisfies the so-called "finite-sample consistency"* [Pesarin & Salmaso, 2009, 2010], that which occurs when K diverges while n_1 and n_2 are fixed (useful when $n < K$, with some stochastic processes, "omics", functional, shape, and image data).

- **NP.6.** *UI-NPC works even when different degrees of importance are assigned to the K sub-hypotheses.* E.g., if $w_k \geq 0$, $k = 1, \ldots, K$, and $w_k > 0$ for at least one k, Fisher's becomes $T^*_{FW} = -\sum_k w_k \cdot \log(\lambda^*_k)$. When $w_k = w > 0$ an equivalent formulation of T_F occurs; weighted rules are typically used in survival analyses.

In this context, T_G results as an *adaptive weighted Tippt't rule*, where weights are: $w_k = 1$ if $k = \arg\min_h(\lambda_h)$ and 0 elsewhere, so $T^*_G = \max_k(1 - \lambda^{*w_k}_k)$ (for general adaptive tests see O'Gorman, 2012.

• **NP.7.** *Under conditions for the permutation central limit theorem (PCLT) and partial tests* $T_k^* = \bar{X}_{1k}^* - \bar{X}_{2k}^*$, $k = 1, \ldots, K$, so that each partial test is asymptotically optimal (practically, $0 < Var(X_k) < \infty$ often suffices for PCLT), *combined test by any* $\psi \in \mathcal{C}_A$ *results as an admissible combination of asymptotically optimal tests*.

• **NP.8.** *Since PTs do not require to be expressed in standardized formulation, UI-NPC does not require existence of population inverse of covariance matrix* $\mathbf{\Sigma}^{-1}$, thus it easily deals with $n < K$ and/or constrained alternatives.

- **NP.9.** *UI-NPC does not require knowledge of dependence coefficients among partial PTs.* Thus, since $\mathbf{X}$ is sufficient for P, including all its dependence coefficients, the UI-NPC properly deals with the nonparametric combination of dependent permutation tests (Prp P5).

- **NP.10.** *UI-NPC provides a natural solution to the* **multi-aspect testing,** *that which arises when a list of different tests are used for the same hypotheses* (suitable when under specific conditions there exists a good test, but it is unknown if such conditions occur).

6 Some Examples

Example 1. A testing problem on the effectiveness of training for the reduction of anxiety (IPAT) with $n = 20$ subjects: $X = Y_B - Y_A$.

i	Y_B	Y_A	X	i	Y_B	Y_A	X
1	19	14	5	11	16	17	-1
2	22	23	-1	12	25	20	5
3	18	13	5	13	22	18	4
4	18	17	1	14	19	17	2
5	24	20	4	15	27	22	5
6	30	22	8	16	23	21	2
7	26	30	-4	17	24	21	3
8	28	21	7	18	18	15	3
9	15	11	4	19	28	24	4
10	30	29	1	20	27	22	5

Permutation test on paired data, $T^* = \sum_i X_i \cdot S_i^*$,

S_i^* are IID, s.t. $\Pr(S_i^* = 0) = \Pr(S_i^* = 1) = 1/2$.

With $R = 100000$ random permutations, $\hat{\lambda} = 0.00020$.

The McNemar test $\#(X_i > 0) = 17 \sim Bin(1/2, 20)$ gives $\lambda = 0.0013$.

Permutation test reject H_0 at $\alpha = 0.0005$, McNemar rejects at $\alpha = 0.005$.

Example 2. Testing symmetry with a sample of 24 washers:

i	X_0	X	i	X_0	X
1	0	1.6	13	0	2.1
2	0	1	14	0	-1
3	0	-0.8	15	0	3.5
4	0	-1.3	16	0	0.6
5	0	1.4	17	0	-0.2
6	0	-0.1	18	0	0.5
7	0	1.1	19	0	0.5
8	0	-1	20	0	4
9	0	-0.1	21	0	1.9
10	0	-0.6	22	0	-0.4
11	0	0.7	23	0	0.4
12	0	-0.6	24	0	1.4

Symmetry of X with respect to X_0 implies $H_0 : \Pr(X_0 - x) = \Pr(X_0 + X)$.

It is required to test against positive asymmetry.

Hence $H_1 : \Pr(X_0 + X) > \Pr(X_0 - X)$ (note the analogy with the IPAT data)

Test: $T^* = \sum_i X_i \cdot S_i^*$.

With $R = 100000$, $\lambda = 0.01856$

The null hypothesis is rejected at $\alpha = 0.05$

Example 3. Job satisfaction of 20 workers (fictitious data).

12 were classified as anxious X_1, and 8 classified as normal X_2.

It is required to test for $H_0 : X_1 \stackrel{d}{=} X_2$ against $H_1 : X_1 \stackrel{d}{>} X_2$.

X_1 :	66	57	81	62	61	60	73	59	80	55	67	70
X_2 :	64	58	45	43	37	56	44	42				

Using NPC on $T^* = \bar{X}_1^* - \bar{X}_2^*$, with $R = 100000$, $\hat{\lambda} = 0.000415$.

H_0 is then rejected at $\alpha = 0.001$.

The Wilcoxon-Mann-Whitney rank solution:

$$T_W = (\sum_i R_{1i} - n_1(n+1)/2)/[n_1 n_2(n+1)/12]^{1/2}$$

where $R_{ji} = \sum_{1 \leq k,h \leq n} \mathbb{I}(X_{kh} \leq X_{ji})$, $j = 1, 2$,

gives $T_W = 4.146$, with standard normal null distribution: $T_W(0) \sim \mathcal{N}(0, 1)$,

then with p-value $\lambda_W < 0.001$, leading to rejection of H_0.

Example 4. Multivariate Fisher's exact probability test

Consider the table

Gr.	1°		2°	
$X \backslash Y$	0	1	0	1
0	27	1	15	0
1	13	7	2	0

X	1°	2°
0	28	15
1	20	2

Y	1°	2°
0	40	17
1	8	0

the data are related to Hamilton's Anxiety (X) and Depression (Y) on 65 workers in an Oil Company (Abbate et al., 2001), employed for 12-15 years.

48 units (Gr.1°) were exposed to *stress conditions*, 17 (Gr.2°) were exposed to *normal conditions*.

It is asked to test if stress may induce anxiety and/or depression.

Thus, $H_1 : (X_1 \overset{d}{>} X_2) \cup (Y_1 \overset{d}{>} Y_2)$ V.s $H_0 : (X_1 \overset{d}{=} X_2) \cap (Y_1 \overset{d}{=} Y_2)$

The UI-NPC test, with $R = 100000$ CMC gives: $\hat{\lambda}_1 = 0.0221$, $\hat{\lambda}_2 = 0.0743$;

(the exact probability values on two marginals are $\lambda_1 = 0.02209$ and $\lambda_2 = 0.07476$, respectively; note, the essential closeness).

Fisher's combination gives $\hat{\lambda}_F = 0.00789$, significant at 0.01;

Note: with $\alpha = 001$ the combined is significant, partials are not, (application of Bonferroni's type analyses may become problematic).

Example 5. Shrinkage of synthetic fibers (Johnson & Leone, 1964).

Group 1 with $n_1 = 12$ items at 120 C^o, group 2 of $n_2 = 10$ at 140 C^o

1^o			2^o		
3.45	3.62	3.60	3.72	4.01	3.54
3.49	3.64	3.56	3.67	4.03	3.40
3.52	3.53	3.57	3.96	3.60	3.76
3.44	3.56	3.43	3.91		

Test for $H_0 : X_1 \stackrel{d}{=} X_2$ against $H_1 : X_1 \stackrel{d}{<} X_2$. Using NPC on $T_1^* = \bar{X}_2^* - \bar{X}_1^*$ and $T_2^* = \overline{(X_2^*)^2} - \overline{(X_1^*)^2}$; with $R = 100000$, we have: $\hat{\lambda}_1 = 0.00178$, $\hat{\lambda}_2 = 0.00164$, $\hat{\lambda}_F = 0.00162$; so, H_0 is rejected at $\alpha = 0.002$ (Aspin-Welch test: $\bar{X}_2 = 3.76$, $\bar{X}_1 = 3.534$, $\hat{\sigma}_2^2 = 0.4569$, $\hat{\sigma}_1^2 = 0.0495$, $t_{\approx 11} = 3.20$, significant at $\alpha = 0.01$).

Example 6. Blood testosterone on 11 women.

Blood testosterone is observed on 11 women at 8.00, 8.30, 9.00, 10.00, and 15.00.

These observations were aimed at evaluating whether testosterone is subject to change during the day.

The hypotheses are:

$$H_0 : \{\bigcap_{i=1}^{n}(X_{1i} \stackrel{d}{=} ... \stackrel{d}{=} X_{ki})\} = \{\delta_1 = ... = \delta_k\},$$

against $H_1 : \{H_0$ *is not true*$\}$.

The data in $(p/ml) \times 1000$ are:.

$i \setminus t$	8.00	8.30	9.00	10.00	15.00
1	320	278	236	222	232
2	478	513	415	359	292
3	921	701	645	526	458
4	213	230	261	253	199
5	273	338	323	332	222
6	392	302	289	305	172
7	469	443	292	235	233
8	422	389	359	331	185
9	613	649	626	588	636
10	395	318	298	269	328
11	462	400	360	247	284

Permutation test is: $T_P^* = \sum_{j=1}^{k} (\bar{X}_{j\cdot}^* - \bar{X}_{\cdot\cdot})^2 / \sum_{ji} (X_{ji}^* - \bar{X}_{\cdot i} - \bar{X}_{j\cdot}^* + \bar{X}_{\cdot\cdot})^2$

and Friedman's rank test $T_F = \sum_{j=1}^{k} \left[\bar{R}_j - \frac{k+1}{2}\right]^2 / \frac{12k(k+1)}{12n}$.

Considering $R = 2000$ permutations T_P^* gives $\hat{\lambda}_P = 0.0003$, which leads to the rejection of H_0 at $\alpha = 0.001$.

This result fits that of Friedman's rank test $T_F = 19.709$, whose null distribution is approximated by a central χ^2 with 4 d.f. significant at $\alpha = 0.001$.

Example 6. Dominance on ordered categorical data

Data are from Troendle (2002), also discussed by Lumely (1996) and Brunner & Langer. (2000).

Data concern "a trial of shoulder tip pain, and the observed variables are pain scores following laparoscopic surgery.

Pain scores take integer values from one (low) to five (high), i.e. $1 \leq k \leq K = 5$.

Treatment A (active drug) and P (placebo) were assigned randomly to 25 female patients with 14 receiving A and 11 receiving P.

Pain scores recorded on the third day following surgery are:

A : 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 4, 1, 1

P : 3, 3, 4, 3, 1, 2, 3, 1, 1, 5, 4

It is of interest to test if treatment with active drug A produces stochastically better responses than placebo P.

This gives rise to a dominance alternative $H_1 : \{\bigcup_k F_{Ak} > F_{Pk}\}$,

and, as subjects were randomly assigned to treatments, $H_0 : \{\bigcap_k F_{Ak} = F_{Pk}\}$

where F_j are cumulative relative frequencies.

Based on $R = 100\,000$ permutations Anderson–Darling's test statistic

$$T^*_{AD} = \sum_{k=1}^{K-1} \left(\hat{F}^*_{2k} - \hat{F}^*_{1k}\right) \left[\bar{F}_{\cdot k}(1 - \bar{F}_{\cdot k})\right]^{-\frac{1}{2}},$$

where $\hat{F}^*_{jk} = \#(X^*_{ji} \leq C_k)/n_j$, $j = 1, 2$, are permutation EDFs at class C_k and $\bar{F}_{\cdot k} = (\hat{F}^*_{2k} n_2 + \hat{F}^*_{1k} n_1)/n$, $1 \leq k \leq K$,

gives a p-value $\hat{\lambda}_{AD} = 0.00455$,

the t-test on ranks of classes as scores (Wilcoxon-Mann-Whitney) gives $\hat{\lambda}_t = 0.00520$,

both leading to the rejection of H_0 at $\alpha = 0.01$.

Example 7. Repeated observations with ordered categorical data

Data, from Lumley (1996) and Brunner & Langer (2000), concern a clinical trial on shoulder tip pain scores X observed on each patient at $\tau = 6$ time points after *laparoscopic surgery*.

Pain scores (ordered categories) range from $1 \equiv$ *low*, to $C = 5 \equiv$ *high*.

Two treatments were randomly assigned to $n = 41$ eligible patients:

$n_1 = 22$ received Y (active drug), $n_2 = 19$ received N (placebo).

$t\downarrow$	Y																					
1	1	3	1	1	2	1	3	2	1	2	4	1	1	3	1	4	1	1	1	3	2	1
2	1	2	1	1	1	1	1	1	1	1	4	1	1	2	2	4	3	1	2	1	1	1
3	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	4	2	1	2	1	1	1
4	1	2	1	1	1	1	1	1	1	1	4	2	2	1	1	2	1	1	2	1	1	1
5	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	2	3	1	1
6	1	1	1	1	1	1	1	2	1	1	2	1	2	1	1	1	1	1	2	3	1	1

$t\downarrow$	N																		
1	5	1	4	3	1	1	2	2	1	5	5	4	2	3	1	1	1	3	1
2	2	5	4	4	1	3	2	2	1	5	4	4	3	3	1	5	1	3	3
3	3	3	4	3	1	2	3	1	1	5	4	4	4	4	1	5	1	3	3
4	5	4	4	3	1	2	4	3	1	4	4	4	3	4	1	5	1	3	3
5	5	5	1	3	1	1	2	3	1	3	2	4	3	4	1	4	1	1	2
6	4	3	1	2	1	1	2	2	1	3	2	3	2	3	1	3	1	1	1

One related testing problem

Besides solutions from the literature [e.g., Lumley (1996), Brunner & Langer (2000), etc.], it is required to test if patients taking Y exhibit, along time, a *stochastically lower pain* than those taking N.

Pain distributions appear to be time-decreasing: $X_t \overset{d}{\geq} X_q$, $t < q \leq \tau$;

i.e., data look like a two-sample of *independent discrete non-stationary stochastic processes* with **unknown time-converging transition matrices** (maybe random).

Accordingly, testing has to be set up *without assuming time-invariant distributions*.

The hypotheses for *multivariate stochastic dominance* setting are:

$$H_0 : \mathbf{X}_Y \overset{d}{=} \mathbf{X}_N = \bigcap_{t=1}^{6} (\mathbf{X}_{Yt} \overset{d}{=} \mathbf{X}_{Nt}) \equiv \bigcap_{t=1}^{6} \bigcap_{h=1}^{4} [F_{Yt}(c_h) = F_{Nt}(c_h)],$$

V.s

$$H_1 : \mathbf{X}_Y \overset{d}{<} \mathbf{X}_N = \bigcup_{t=1}^{6} (\mathbf{X}_{Yt} \overset{d}{\leq} \mathbf{X}_{Nt}) \equiv \bigcup_{t=1}^{6} \bigcup_{h=1}^{4} [F_{Yt}(c_h) > F_{Nt}(c_h)],$$

(note: at least one strict inequality).

Also note:

- Scores $\equiv$ ordered categorical values: $X \in (c_1 = 1 \prec \ldots \prec c_5 = 5)$;
- underlying CDFs: $F_{Xt}(c_h) = \Pr\{X_t \leq c_h\}$, $h = 1, \ldots, 5$, $t = 1, \ldots, 6$;
- since $F_{Yt}(c_5) = F_{Nt}(c_5) = 1$, $\forall t$; i.e. c_5 points are not considered;
- patient's trajectories $\{X_t,\ t = 1, \ldots, 6\}$ are independently observed;
- as X_t is non-stationary under H_0, usual 2-way modelling is unsuitable;
- the problem is broken-down into 24 partial sub-problems;
- parameters' $Dim(\Theta) = 5^6 - 1 = 15624$ (quite a highly sparse table);
- [if parameters were specific to units, $Dim(\Theta) = 640584$];
- alternatives are restricted to the 24-Dim positive orthant;
- *Multi-One-Sided Tests* are required;
- when feasible, likelihood maximization requires non-standard methods;
- when $n < Dim(\Theta)$, LR methods are not available.

Setting $Y \to 1$ and $N \to 2$, the testing problem:

$$\begin{array}{l} \\ \text{V.s} \\ \\ \end{array} \left| \begin{array}{l} H_{0th} : F_{1t}(c_h) = F_{2t}(c_h) \\ \\ H_{1th} : F_{1t}(c_h) > F_{2t}(c_h) \end{array} \right| h = 1, \ldots, 4;\ t = 1, \ldots, 6,$$

requires the *joint comparison of* 24 *one-sided differences* : $\hat{F}_{1t}(c_h) - \hat{F}_{2t}(c_h)$.

Of course, *the* 24 *partial tests are dependent* in a complex way**: dependence coefficients are functions of all parameters of underlying processes**.

When feasible the related testing problem has a rather difficult solution within the LR theory (Colombi & Forcina, 2016; Perlman & Wu, 2006; Silvapulle & Sen, 2005; Zhu & Chen, 2018; etc.). With categorical data, LR solutions are also not unique (Cohen et al., 2000, 2003; Wang, 1996). Supplementary options, difficult to justify in terms of the real problem under study, are required when the orthant dimension is "not small".

This difficulty mostly consists in that the set of alternatives is restricted to lie in the 24-*Dimensional* **positive orthant** *where the likelihood cannot be maximized under* H_0 *by ordinary methods of maximization*.

Thus, we analyze the problem within the *conditional theory of inference* by **conditioning on a set of sufficient statistics, as the pooled data** $\mathbf{X}$, (Cox & Hinkley, 1974; Lehmann, 1986; etc.); i.e., by *permutation methods*.

Note: sufficiency of $\mathbf{X}$ requires generalized density $f_F(\mathbf{X}) > 0$.

Due to evident complexity, for testing analyses it seems wise to stay within Roy's (1953) *Union-Intersection approach* (UI) and the *Permutation Testing Principle* (PTP; Pesarin, 2013).

The UI-NPC analysis of medical example

According to Roy, the hypotheses $H_0 : \mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2$ V.s $H_1 : \mathbf{X}_1 \stackrel{d}{<} \mathbf{X}_2$ are broken down as

$$H_0 : \mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2 \equiv \bigcap_{t=1}^{6} \bigcap_{h=1}^{4} [F_{1t}(c_h) = F_{2t}(c_h)]$$

and

$$H_1 : \mathbf{X}_1 \stackrel{d}{<} \mathbf{X}_2 \equiv \bigcup_{t=1}^{6} \bigcup_{h=1}^{4} [F_{1t}(c_h) > F_{2t}(c_h)],$$

(note: *at least one strict inequality*).

Each partial test we use corresponds to a standardized comparison of empirical frequencies (Pesarin, 2018):

$$T^*_{th} = G(n_1, n_2) \cdot [\hat{F}^*_{1th} - \hat{F}^*_{2th}] \left[\bar{F}_{\cdot th}(1 - \bar{F}_{\cdot th})\right]^{-\frac{1}{2}},$$

$$t = 1, \ldots, 6 = \tau, \;\; h = 1, \ldots, 4 = C - 1,$$

where:
- $\hat{F}^*_{jth} = \#(X^*_{jt} \leq c_h)/n_j$, $j = 1, 2$: permutation relative frequencies;
- $\bar{F}_{\cdot th} = [\#(X_{1t} \leq c_h) + \#(X_{2t} \leq c_h)]/n$: marginal relative frequencies;
- each T^*_{th} is a reformulation of Fisher's exact probability test;
- under indep. units T^*_{th} is UMPS; if not it is unbiased, consistent and admissible;
- $G(n_1, n_2) = [n_1 n_2 (n-1)/n^2]^{1/2}$, i.e. a permutation invariable quantity;
- $\tau \times (C-1)$ partial tests T^*_{th} are positively dependent.

The simplest combination is the Direct; i.e. *the sum of dependent standardized partial statistics*; leading to the global test :

$$T^*_{AD} = G(n_1, n_2) \cdot \sum_{t=1}^{\tau} \sum_{h=1}^{C-1} \left(\hat{F}^*_{1th} - \hat{F}^*_{2th} \right) \left[\bar{F}_{\cdot th} (1 - \bar{F}_{\cdot th}) \right]^{-\frac{1}{2}};$$

that is, a version of *Multivariate Anderson-Darling discrete goodness-of-fit solution for Multi-One-Sided alternatives*.

Note:
- the form $0/0$ is set 0;
- $T^*_{AD} = \sum_t T^*_{ADt}$, is the sum of τ Anderson-Darling partial tests;
- under H_1 at least one summand assumes values not smaller than under H_0;
- each partial test T^*_{th} is separately unbiased, consistent and cond. optimal;
- T^*_{AD} *is unbiased, consistent and admissible*;
- also appropriate are partial tests on scores T_W, on mid-ranks T_M, or..., etc.

The analyses of data based on $R = 100\,000$ give:

t	1	2	3	4	5	6	$\hat{\lambda}$
$\hat{\lambda}_{ADt}$	0.05848	0.00044	0.00005	0.00002	0.00024	0.00622	0.000035
$\hat{\lambda}_{Wt}$	0.09036	0.00045	0.00005	0.00002	0.00026	0.00739	0.000025
$\hat{\lambda}_{Mt}$	0.10795	0.00035	0.00007	0.00003	0.00026	0.00736	0.000075

,

where scores are $w_h = h$, $h = 1, \ldots, C$.

These results exhibit quite a strong evidence for the distributional higher pain scores, i.e. *multivariate stochastic dominance*, of patients taking N comparing to those taking Y. It is worth noting that three combined tests give about the same results.

To approximately take account of observed and/or unobserved covariates and of baseline data $X_{t=1}$, we also tried a multivariate dominance analysis on *difference pain scores* $Y_{tji} = X_{tji} - X_{1ji}$, $t = 2, \ldots, 6$, $j = 1, 2$, $i = 1, \ldots, n$.

The related results, still based on $R = 100\,000$ and Fisher's combined test, are:

t	2	3	4	5	6	$\hat{\lambda}_{F,2\div6}$
$\hat{\lambda}_{Yt}$	0.03729	0.00631	0.00162	0.07842	0.47577	0.01731

Apparently, at time *6* two groups of patients almost completely recover from difference pain scores; whereas at times *2* to *5* there seems remaining a significant greater pain on patients taking *N* than those taking *Y*.

According to a one-step closed testing procedure, this further conclusion relies on a partial analysis limited to times 2 to 5, whose combined test leads to $\hat{\lambda}_{F,2\div5} = 0.00504$.

7 References

Abbate, C., Giorgianni, F., Munaò, F., Pesarin, F., Salmaso, L. (2001). Neurobehavioural evaluation in humans exposed to hydrocarbons. *Psychotherapy and Psychosomatics*; **70**, 44–49.

Arboretti, R. Pesarin, F. &. Salmaso, L. (2021). A Unified Approach to Permutation Testing for Equivalence. *Statistical Methods & Applications;* **30:**3 1033–1052; https://doi.org/10.1007/s10260-020-00548-0).

Basso D., Pesarin F., Salmaso L., Solari A. (2009). *Permutation tests for stochastic ordering and ANOVA: theory and applications in R*. Lecture Notes in Statistics, N. 194, Springer, New York.

Basso, D., Salmaso, L. (2009). A permutation test for umbrella alternatives.*Statistics and Computing*, (DOI 10.1007/s11222-009-9145-8).

Bertoluzzo F., Pesarin F., Salmaso L. (2011). Multi-sided permutation tests: an approach to random effects. *Journal of Statistical Planning and Inference* (forthcoming).

Brombin, C., Salmaso, L. (2009). Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics & Data Analysis*, **53**, 3921-3931.

Brunner, E., Langer, F. (2000). Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical Journal;* **42**, 663-675.

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

David, H.A. (2008). The beginnings of randomization Tests. *The American Statistician*, **62**, 70–72.

Edgington, E.S., Onghena, P. (2007). *Randomization Tests* (4th ed.).Chapman and Hall/CRC, London.

Finos, L., Salmaso, L. (2006). Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations. *Journal of Nonparametric Statistics*, **18**, 245-261.

Finos, L., Salmaso, L., Solari, A. (2007). Conditional inference under simultaneous stochastic ordering constraints. *Journal of Statistical Planning and Inference*, **137**, 2633-2641.

Fisher, R.A. (1936). "The coefficient of racial likeness" and the future of craniometry. *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, **66**, 57-63.

Goggin M.L. (1986). The "Too Few Cases/Too Many Variables" Problem in Implementation Research. *The Western Political Quarterly*, **39**, 328-347.

Good, P. (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (3rd ed.), Springer-Verlag, New York.

Hirotsu, C. (1986). Cumulative chi-squared statistic or a tool for testing goodness of fit. *Biometrika*, **73**, 165-173.

Hirotsu, C. (1998b). Isotonic inference. In *Encyclopedia of Biostatistics*, 2107 - 2115, Wiley, New York.

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

Janssen, A. (2005). Resampling student's t-type statistics. *Annals of the Institute of Statistical Mathematics*, **57**, 507-529.

Johnson, N.L., Leone, F.C. (1964). *Statistical ad Experimental Design in Engineering and Physical Sciences.* Wiley, New York.

Kempthorne, O. (1955). The randomization theory of experimental inference.Journal of the American Statistical Association, **50**, 964-967.

Klingenberg, B., Solari, A., Salmaso, L., Pesarin F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, **65**, 452–462.

Lachenbrook, P.A., (1976). Analysis of data with clumping at zero.*Biometrical Journal,* **18**, 351-356.

Landenna, G., Marasuni, D., Ferrari, P. (1998). *La Verifica delle Ipotesi Statistiche*. Il Mulino, Bologna.

Lehmann, E.L. (2009). Parametric versus nonparametrics: two alternative methodologies. *Journal of Nonparametric Statistics,* **21**, 397-405.

Lehmann, E.L., Romano J.P. (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer, New York.

Lehmann, E.L., Scheffé, H. (1950). Completeness similar regions, and unbiased estimation. *Sankhyā*, **10**, 305-340.

Lehmann, E.L., Scheffé, H. (1955). Completeness similar regions, and unbiased estimation - part II. *Sankhyā*, **15**, 219-236.

Ludbrook, J., Duddley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician,* **52**, 127-132.

Lumley, T., (1996). Generalized estimating equations for ordinal data: A note on working correlation structures. *Biometrics*, **52**, 354–361.

Mehta, C.R., Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association,* **78**, 427-434.

Moder K., Rasch D., Kubinger K.D. (2009). Don't use the two-sample *t*-test anymore! *Proceedings of the 6th St. Petersburg Workshop on Simulation* (Edited by S.M. Ermakov, V.B. Melas and A.N. Pepelyshev), 258-264.

Pesarin, F. (2001). *Multivariate Permutation tests: with Application in Biostatistics*. John Wiley & Sons, Chichester, UK.

Pesarin, F. (2002). Extending permutation conditional inference to unconditional one. *Statistical Methods and Applications*, **11**, 161-173.

Pesarin F. and Salmaso L. (2009). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparametric Statistics* (DOI 10.1080/10485250902807407).

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

Pesarin, F. & Salmaso, L., (2013). On the weak consistency of permutation tests. *Communications in Statistics - Simulation and Computation*; **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. *Statistics & Computing;* **26**(3): 693-701, DOI 10.1007/s11222-015-9552-y

Salmaso, L., Solari, A. (2005). Multiple Aspect Testing for case-control designs. *Metrika*, **12**, 1-10.

Salmaso, L., Solari, A. (2006). Nonparametric iterated combined tests for genetic differentiation. *Computational Statistics & Data analysis,* **50**, 1105-1112.

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

Troendle, J.F., (2002). A Likelihood Ratio Test for the Nonparametric Behrens-Fisher Problem. *Biometrical Journal*; **44:7**, 813-824.

Westfall, P. H., Young S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-values adjustment*. Wiley, New York.