The Permutation Testing Approach in the Light of Conditionality and Sufficiency Principles

Fortunato Pesarin

Abstract

In recent years permutation testing methods have increased in number of applications and in solving complex multivariate problems. When available they are essentially of an exact nature in a conditional context, where the conditioning is on the pooled observed data which in general are a set of sufficient statistics in the null hypothesis. The application of the conditionality principle of inference provides this approach with important and useful properties.

1 Introduction

In recent years permutation testing methods have increased both in number of applications and in solving complex multivariate problems. Most of testing problems may also be effectively solved using traditional parametric or rank-based nonparametric (NP) methods, although in relatively mild conditions their permutation counterparts when available are asymptotically as good as the best ones (Hoeffding 1952). Permutation tests (PTs) are essentially of an exact NP nature in a conditional context, where the conditioning is on the pooled observed data which, under randomization of units to treatments, are always a set of sufficient statistics in the null hypothesis. On the one hand, the application of the conditionality principle (CP) of inference provides the PT approach with important and useful properties. On the other, the reference null distribution of most parametric tests, with the exception of rather simple situations, is only known asymptotically. Thus, for most sample sizes of practical interest, the possible lack of efficiency of PTs

F. Pesarin (✉)
Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241, 35121 Padova, Italy
e-mail: fortunato.pesarin@unipd.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_4,

may be compensated by the lack of approximation of parametric counterparts. There are many complex multivariate problems (common in biostatistics, clinical trials, experimental data, pharmacology, psychology, social sciences, etc.) which are difficult, if not impossible, to solve outside the CP and in particular outside the method of nonparametric combination (NPC) of dependent PTs (Pesarin and Salmaso 2010).

Frequently parametric methods reflect a modelling approach and generally require a set of quite stringent assumptions, which are often difficult to justify. Sometimes these assumptions are merely set on an *ad hoc* basis: too often and without any justification researchers assume multivariate normality, random sampling from a target population, homoscedasticity of responses also in the alternative, random effects independent of units, etc. In this way consequent inferences have no real credibility. On the contrary, NP approaches try to keep assumptions at a lower workable level, avoiding those which are difficult to justify. Thus, they are based on more realistic foundations, are intrinsically robust and consequent inferences credible. For instance, PT comparisons of means do not require data homoscedasticity in the alternative, provided that random effects are either negative or positive.

Our point of view, however, is that statisticians should have in their tool-kit of methods both the parametric, including the Bayesian, and the NP, because in their life they surely meet with problems which are difficult, if not impossible, within one approach and others which in turn are difficult, if not impossible, within the other. For some examples as well as for the literature on the subject matter, we refer to Pesarin and Salmaso (2010) and references therein.

Here we discuss main properties of PTs derived by direct application of sufficiency principle (SP) and CP. The outline includes: a discussion of data model which extends that commonly used by parametric approaches; a presentation of SP and CP and their involvement in the PT principle; notation, definitions, and main properties (exactness, similarity, uniform unbiasedness, consistency) of PTs; and the extension of conditional to unconditional inference.

2 The Data Model

Without loss of generality we refer to the two-sample one-dimensional design as a guide. Extensions to one-sample and multi-sample designs are straightforward. The extension to multivariate designs requires the NPC (Pesarin and Salmaso 2010).

Let us assume that a variable X takes values on sample space $\mathscr{X}$, and that associated with $(X, \mathscr{X})$ there is a parent distribution P member of an NP family $\mathscr{P}$. "A family $\mathscr{P}$ of distributions is NP when it is not possible to find a finite-dimensional space Θ (the parameter space) such that there is a one-to-one relationship between Θ and $\mathscr{P}$, in the sense that each member P of $\mathscr{P}$ cannot be identified by only one member θ of Θ, and vice versa." In practice parametric families only contain distributions defined by a well-specified finite set of parameters; whereas families of distributions in which parameters are infinitely

many or are unspecified are NP. Each $P \in \mathscr{P}$ gives the probability measure to events A member of a suitable collection $\mathscr{A}$ of events. Family $\mathscr{P}$ may consist of distributions of real (continuous, discrete, mixed) and/or categorical (nominal, ordered) type of variables. It is assumed that $\mathscr{P}$ admits the existence of a dominating measure $\xi_{\mathscr{P}}$ in which respect the density $f_P(X) = \mathrm{d}P(X)/\mathrm{d}\xi_{\mathscr{P}}$ is defined. The density on every observed sample point $X \in \mathscr{X}$ is assumed satisfying to $f_P(X) > 0$ (we do not distinguish between a variable X and its observed sample points, the context suffices to avoid misunderstandings).

Let $\mathbf{X}_j = \{X_{ji},\ i = 1, \ldots, n_j\} \in \mathscr{X}^{n_j}$ be the independent and identically distributed (IID) sample data of size n_j coming from $P_j \in \mathscr{P}$, $j = 1, 2$. A notation for data sets with independent samples is $\mathbf{X} = \{X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}\} \in \mathscr{X}^n$, whose related model, with clear meaning of the symbols, is $(\mathbf{X}, \mathscr{X}^n, \mathscr{A}^{(n)}, P^{(n)} \in \mathscr{P}^{(n)})$, where $n = n_1 + n_2$, and $P^{(n)} = P_1^{n_1} \cdot P_2^{n_2}$. To denote data sets in the PT context it can be useful referring to the unit-by-unit representation: $\mathbf{X} = \mathbf{X}^{(n)} = \{X(i),\ i = 1, \ldots, n;\ n_1, n_2\}$, where it is intended that first n_1 data in the list belong to first sample and the rest to the second. Indeed, 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 its members, $\mathbf{X}^* = \{X^*(i) = X(u_i^*),\ i = 1, \ldots, n;\ n_1, n_2\}$ is the related permutation of $\mathbf{X}$; so that $\mathbf{X}_1^* = \{X_{1i}^* = X(u_i^*), i = 1, \ldots, n_1\}$ and $\mathbf{X}_2^* = \{X_{2i}^* = X(u_i^*),\ i = n_1 + 1, \ldots, n\}$ are the two permuted samples, respectively. Of course, in multivariate problems data vectors associated with units are then permuted.

We discuss testing problems for stochastic dominance alternatives (one-sided) as are generated by treatments with nonnegative random shift effects Δ. In particular, the alternative assumes that two treatments produce effects Δ_1 and Δ_2, and that $\Delta_1 \overset{d}{>} \Delta_2$, where $\overset{d}{>}$ stands for stochastic (or distributional) dominance. Thus, the hypotheses are $H_0 : X_1 \overset{d}{=} X_2 \equiv P_1 = P_2$, and $H_1 : (X_1 + \Delta_1) \overset{d}{>} (X_2 + \Delta_2)$, respectively. Extensions to nonpositive and two-sided alternatives are straightforward. Note that *under H_0 data of two samples are exchangeable*, in accordance with the notion that units are randomized to treatments. Without loss of generality, we assume that effects in H_1 are such that $\Delta_1 = \Delta \overset{d}{>} 0$ and $\Pr\{\Delta_2 = 0\} = 1$. This condition 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, Δ can depend on units and on related null deviates X, so that pairs (X_{1i}, Δ_i), $i = 1, \ldots, n_1$, do satisfy $(X_{1i} + \Delta_i) \geq X_{1i}$ with at least one strict inequality. In this situation the induced stochastic dominance $(X_1 + \Delta) \overset{d}{>} X_2 = X$ is compatible with heteroscedasticities in the alternative. Thus, H_0 can also be written as $H_0 : \Delta \overset{d}{=} 0$. Other than measurability, no further distributional assumption on random effects Δ is required. It is required that null deviates X and test statistics $T : \mathscr{X}^n \to \mathscr{R}^1$ are measurable in H_0. To emphasize the roles of sample sizes and effects, we may use the notation $\mathbf{X}^{(n)}(\Delta) = \{X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}\}$ to denote data sets; and so $\mathbf{X}^{(n)}(0)$ denotes data in H_0. It is worth noting that the pooled data $\mathbf{X}^{(n)}(0)$ is always a set of sufficient statistics for P in

H_0. Indeed, since $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}) = 1$, the conditional distribution of $\mathbf{X}$ given $\mathbf{X}$ is independent of P. Furthermore, when P is NP or the number of its parameters is larger than sample size or when it lies outside the regular exponential family, $\mathbf{X}$ is *minimal sufficient*.

PT lie within the conditional method of inference, the conditioning being on the observed data set $\mathbf{X}$. The related conditional reference space is denoted by $\mathscr{X}^n_{/\mathbf{X}}$. Essentially $\mathscr{X}^n_{/\mathbf{X}}$ is the set of points of sample space $\mathscr{X}^n$ which are equivalent to $\mathbf{X}$ in terms of information carried by the associated 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, and so it corresponds to the *orbit* of equivalent points associated with $\mathbf{X}$. Given that, under H_0, the density $f_P^{(n)}(\mathbf{X}) = \prod_{ji} f_P(X_{ji})$ is by assumption exchangeable in its arguments, because $f_P^{(n)}(\mathbf{X}) = f_P^{(n)}(\mathbf{X}^*)$ for every permutation $\mathbf{X}^*$ of $\mathbf{X}$, then $\mathscr{X}^n_{/\mathbf{X}}$, or $\mathscr{X}_{/\mathbf{X}}$ by suppressing superscript n, contains all distinct permutations of $\mathbf{X}$. That is $\mathscr{X}_{/\mathbf{X}} = \{\bigcup_{\mathbf{u}^* \in \Pi(\mathbf{u})}[X(u_i^*), i = 1, \ldots, n]\}$. Therefore, since every element $\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ is a set of sufficient statistics for P in H_0, $\mathscr{X}_{/\mathbf{X}}$ is a *sufficient space*. Conditional reference spaces $\mathscr{X}_{/\mathbf{X}}$ are also called *permutation sample spaces*. Moreover, since $\forall A \in \mathscr{A}$ the conditional probability $\Pr(A|\mathbf{X}) = \Pr(A|\mathscr{X}_{/\mathbf{X}})$ in H_0 is P-independent (**P.1** in Sect. 4), the pooled data set $\mathbf{X}$ can be considered as playing the role of ancillary statistics for the problem. And so, when $\mathbf{X}$ is *minimal sufficient* it is also *maximal ancillary* and unique, except for a permutation.

In paired-data designs what is essential is that in H_0 the distribution of X is symmetric with respect to 0 (Pesarin and Salmaso 2010). This condition can be achieved in two main instances: (a) when data are exchangeable within each unit, i.e. when $Y_{1i} \overset{d}{=} Y_{2i}\ \forall i = 1, \ldots, n$, the Ys being paired responses, in which the difference of any two individual observations in $H_0^A : Y_1 \overset{d}{=} Y_2$ is symmetrically distributed around 0, and the set of differences $\mathbf{X} = \{X_i = Y_{1i} - Y_{2i}, i = 1, \ldots, n\}$ is sufficient for P; (b) when Y_{1i} is symmetric around μ_{1i} and Y_{2i} around μ_{2i} without being homoscedastic (and so not exchangeable), then their difference $Y_{1i} - Y_{2i}$ is symmetric around 0 in $H_0^B : (\mu_{1i} - \mu_{2i} = 0, i = 1, \ldots, n)$. In both instances, however, $\mathscr{X}_{/\mathbf{X}} = \{\bigcup_{\mathbf{S}^* \in [-1,+1]^n}[X_i S_i^*, i = 1, \ldots, n]\}$ contains all points obtained by assigning signs $S = (+1,-1)$ to differences in all possible ways. By the way, paired-data designs show that the exchangeability property is sufficient but not necessary for the PT approach.

The fact that random effects Δ may depend on null deviates X can be viewed as an improvement with respect to traditional parametric approaches, though this may imply evident difficulties for estimation and prediction. On the one hand, this leads to assumptions that are much more flexible and closer to reality. There are indeed many real problems in which the assumption of independence of effects on null deviates cannot be justified, as, for instance, when data are obtained by measurement instruments based on nonlinear monotonic transformations φ of underlying deviates Y. Indeed, with clear meaning of the symbols $\Delta\varphi'(Y + c\Delta) = X(\Delta) - \varphi(Y)$, which depends on $Y = \varphi^{-1}(X)$ and Δ. On the other hand, it is noticeable that in

PTs the separate estimate of variance components is not required. Consequently, the modeling may better fit physical requirements, results of analyses are more credible and their interpretation more clear. In addition, it is to be emphasized that in the NP framework, more than on parameters, the inferential interest is on functionals, i.e. on functions of all parameters such as the so-called treatment effect Δ. And so it can be impossible to separate the role of parameters of interest from the nuisance ones since they could be confounded in Δ.

Moreover, when the data set $\mathbf{X}$ is minimal sufficient in H_0, even if the parent likelihood model depends on a finite set of parameters only one of which is of interest, univariate statistics capable of summarizing the contained information do not exist. So no parametric method can claim to be uniformly better than others. Indeed, conditioning on $\mathscr{X}_{/\mathbf{X}}$, i.e. by considering PT counterparts, improves the power behavior of any unbiased test statistic (via Rao-Blackwell). However, to reduce the loss of information associated with using one single statistic, it is possible to find solutions within the so-called multi-aspect methodology and based on the NPC of several dependent PTs, each capable of summarizing information on a specific aspect of interest for the analysis (Pesarin and Salmaso 2010). A procedure which may improve efficiency and interpretability of results. For instance, when of two unbiased partial PTs only one is consistent, their NPC is consistent.

3 Conditionality, Sufficiency, and Permutation Testing Principles

Let us briefly recall the CP and the SP, as are used in parametric inference (Cox and Hinkley 1974). We consider these principles as key guides also for the NP approach and relate them to the PT principle.

The SP 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, 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."

Basically, the rationale for adopting these principles in statistical inference considers typical examples as the following: suppose that data $\mathbf{X}$ can be obtained by means of one of two different measuring instruments, I_1 and I_2, and suppose the associated normally distributed models are, respectively, $X_1 \sim \mathscr{N}(_1, \sigma_1)$ and $X_2 \sim \mathscr{N}(_2, \sigma_2)$, with $\sigma_1 \ll \sigma_2$. If it is known which instrument has generated $\mathbf{X}$ it seems unavoidable to condition on the related (ancillary) model in any inference regarding μ, the value of σ being 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 invariance of null rejection probability (according to the notion of similarity) and so to condition on a possibly minimal sufficient statistic for it. Indeed, by acting outside these principles related inferential conclusions can be biased, misleading and maybe impossible to be correctly interpreted.

Thus, in the general situation it is wise to condition on its minimal sufficient statistic in H_0, i.e. to condition on the pooled observed data $\mathbf{X}$ which is always sufficient for whatever $P \in \mathscr{P}$ and ancillary for the inferential problem. It is to be recognized that in the literature there is general agreement on the SP; whereas the CP, especially when the ancillary statistic C is not unique, gives rise to known questions and so it is somewhat doubtful (Frosini 1991). These doubts, however, do not apply to the PT approach when $\mathbf{X}$ is minimal sufficient and so maximal ancillary and unique.

This kind of conditioning implies referring to the PT principle: "If two experiments, taking values on the same sample space $\mathscr{X}$ 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 in H_0." Of course, it is intended that in order to obtain reliable inferences there must be a form of stochastic dominance of $(T|H_1)$ with respect to $(T|H_0)$.

On the one hand it should be emphasized that the PT principle works in accordance with both CP and SP since it satisfies both. On the other hand, the related conditional inference can be extended from the set of really observed units to the family of all populations whose associated distributions P satisfy the condition $f_P^{(n)}(\mathbf{X}) > 0$, so as to also include most of the problems in which the sample data are obtained by selection-bias procedures from a target population. However, it should be noted that, due to conditioning on sufficient statistics for all nuisance entities, the extension to a family of distributions is also typical of all parametric conditional inferences in the presence of nuisance parameters (Sect. 5). For instance, this feature is clearly enjoyed by Student's t whose inference can be extended from the observed data set $\mathbf{X}$ to all normal populations which assign positive density to the variance estimate $\hat{\sigma}^2$; thus, its inference can be extended to a family of distributions more than to only the target one.

4 Main Properties of PTs

In this section we briefly outline main terminology, definitions, and general theory of PTs for some one-dimensional problems. Emphasis is again on two-sample one-sided designs in which large values of test statistics $T : \mathscr{X}^n \to \mathscr{R}^1$ are evidence against H_0.

- **P.1.** *Sufficiency of $\mathscr{X}_{/\mathbf{X}}$ for P under H_0 implies that the null conditional probability of every event $A \in \mathscr{A}$, given $\mathscr{X}_{/\mathbf{X}}$, is independent of P; that is, with clear meaning of the symbols,* $\Pr\{\mathbf{X}^* \in A; P|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\mathbf{X}^* \in A|\mathscr{X}_{/\mathbf{X}}\}$.

Thus, the permutation distribution induced by any test statistic $T : \mathscr{X}^n \to \mathscr{R}^1$, namely $F_T(t|\mathscr{X}_{/\mathbf{X}}) = F_T^*(t) = \Pr\{T^* = T(\mathbf{X}^*) \leq t|\mathscr{X}_{/\mathbf{X}})\}$, is P-invariant. Hence, any related conditional inference is distribution-free and NP. Moreover, since for finite sample sizes the number $M = M^{(n)} = \sum_{\mathscr{X}_{/\mathbf{X}}} \mathbb{I}(\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}})$ of points in $\mathscr{X}_{/\mathbf{X}}$ is finite, a relevant consequence of both independence of P and finiteness of M is that in H_0 the permutation probability on every $A \in \mathscr{A}$ is calculated as

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

because $\forall \mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ it is $f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* = f_P(\mathbf{X})\mathrm{d}\mathbf{X}$. It is worth noting here that for calculating the conditional probability distribution it is not necessary to make reference to the so-called *hypothetical repeated sampling principle*. Actually, $\Pr\{\mathbf{X}^* \in A|\mathscr{X}_{/\mathbf{X}}\}$ is *objectively determined* by complete enumeration of $\mathscr{X}_{/\mathbf{X}}$ which once data are observed has a physical existence, and so no hypothetical sampling experiment is referred to in its determination. Since in determining the permutation probability measure in H_0 knowledge of P, or of f_P, is not required, it is to be emphasized that *only the existence of a likelihood is required* by the PT approach (if this existence could not be assumed, no statistical problem would be on the stage). One more relevant consequence of finiteness of $\mathscr{X}_{/\mathbf{X}}$ is that in H_0 permutations $\mathbf{X}^*$ are equally likely conditionally, i.e. $\Pr\{\mathbf{X} = \mathbf{x}|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\mathbf{X}^* = \mathbf{x}|\mathscr{X}_{/\mathbf{X}}\} = 1/M$ if $\mathbf{x} \in \mathscr{X}_{/\mathbf{X}}$ and 0 elsewhere. And so:

- **P.2.** *In H_0 the data set $\mathbf{X}$ is uniformly distributed over $\mathscr{X}_{/\mathbf{X}}$ conditionally.*
- **P.3.** (Uniform similarity of randomized PTs). *Let us assume that the exchangeability condition on data $\mathbf{X}$ is satisfied in H_0, then the conditional rejection probability $\mathbb{E}\{\phi_R(\mathbf{X})|\mathscr{X}_{/\mathbf{X}}\}$ of randomized test $\phi_R = 1$ if $T^o > T_\alpha$, $= \gamma$ if $T^o = T_\alpha$,* and *$= 0$ if $T^o < T_\alpha$, is $\mathbf{X}$-P-invariant for all $\mathbf{X} \in \mathscr{X}^n$ and all $P \in \mathscr{P}$, where: $T^o = T(\mathbf{X})$ is the observed value of T on data $\mathbf{X}$, T_α is the α-sized critical value, and $\gamma = \left[\alpha - \Pr\left\{T^o > T_\alpha|\mathscr{X}_{/\mathbf{X}}\right\}\right] \Big/ \Pr\left\{T^o = T_\alpha|\mathscr{X}_{/\mathbf{X}}\right\}$.*

  For non-randomized PTs such a property is satisfied in the almost sure form for continuous variables and at least asymptotically for discrete variables.

  Determining the critical values T_α of a test statistic T, given the observed data $\mathbf{X}$, in practice presents obvious difficulties. Therefore, it is common to make reference to the associated p-value. This is defined as $\lambda = \lambda_T(\mathbf{X}) = \Pr\{T^* \geq T^o|\mathscr{X}_{/\mathbf{X}}\}$, the determination of which can be obtained by complete enumeration of $\mathscr{X}_{/\mathbf{X}}$ or estimated, to the desired degree of accuracy, by a conditional Monte Carlo algorithm based on a random sampling from $\mathscr{X}_{/\mathbf{X}}$ (Pesarin and Salmaso 2010). For quite simple problems it can be evaluated by efficient computing routines such as those in Mehta and Patel (1983); moreover, according to Mielke and Berry (2007) it can be approximately evaluated by using a suitable approximating distribution, e.g. as within Pearson's system of distributions,

sharing the same few moments of the exact permutation distribution, when these are known in closed form in terms of data $\mathbf{X}$.

The p-value λ is a non-increasing function of T^o and is one-to-one related with the attainable α-value of a test, in the sense that $\lambda_T(\mathbf{X}) > \alpha$ implies $T^o < T_\alpha$, and vice versa. Hence, the non-randomized version can be stated as $\phi = 1$ if $\lambda_T(\mathbf{X}) \le \alpha$, and $= 0$ if $\lambda_T(\mathbf{X}) > \alpha$, for which in H_0 it is $\mathbb{E}\{\phi(\mathbf{X})|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\lambda_T(\mathbf{X}) \le \alpha|\mathscr{X}_{/\mathbf{X}}\} = \alpha$ for every *attainable* α. Thus, attainable α-values play the role of critical values, and in this sense $\lambda_T(\mathbf{X})$ itself is a test statistic.

- **P.4.** (Uniform null distribution of p-values). *Based on* **P.1**, *if X is a continuous variable and T is a continuous non-degenerate function, then p-value $\lambda_T(\mathbf{X})$ in H_0 is uniformly distributed over its attainable support*.
- **P.5.** (Exactness of permutation tests). *A PT T is exact if its null distribution essentially only depends on exchangeable null deviates* $\mathbf{X}(0)$.
- **P.6.** (Uniform unbiasedness of test statistic T). *PTs for random shift alternatives* $(\Delta \overset{d}{\ge} 0)$ *based on divergence of symmetric statistics of non-degenerate measurable non-decreasing transformations of the data, i.e.* $T^*(\Delta) = S_1[\mathbf{X}_1^*(\Delta)] - S_2[\mathbf{X}_2^*(\Delta)]$, *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* $\mathbf{X} \in \mathscr{X}^n$. *In particular*: $\Pr\{\lambda(\mathbf{X}(\Delta)) \le \alpha|\mathscr{X}_{/\mathbf{X}(\Delta)}\} \ge \Pr\{\lambda(\mathbf{X}(0)) \le \alpha|\mathscr{X}_{/\mathbf{X}(0)}\} = \alpha$, *thus p-value in H_1 is stochastically dominated by that in H_0*: $\lambda(\mathbf{X}(\Delta)) \overset{d}{\le} \lambda(\mathbf{X}(0))$.

  An immediate consequence of **P.6** is that, if $\Delta' \overset{d}{>} \Delta$ and so $\lambda(\mathbf{X}(\Delta')) \overset{d}{\le} \lambda(\mathbf{X}(\Delta)) \overset{d}{\le} \lambda(\mathbf{X}(0))$, the permutation p-values of any T are stochastically decreasingly ordered with respect to effect Δ. Without further assumptions, uniform unbiasedness cannot be extended to two-sided alternatives.

  It is worth observing that uniform similarity **P.3** and uniform unbiasedness **P.6** since are at least satisfied for almost all data sets $\mathbf{X}$ under exchangeability in H_0 do not require random sampling from a population. Thus, they also work for selection-bias sampling.
- **P.7.** (The empirical probability measure, EPM). *For each permutation* $\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$, *the EPM of any* $A \in \mathscr{A}$ *is defined as* $\hat{P}_{\mathbf{X}^*}(A) = \sum_{i \le n} \mathbb{I}(X_i^* \in A)/n$ *which, since* $\forall \mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ *it is* $\sum_{i \le n} \mathbb{I}(X_i^* \in A)/n = \sum_{i \le n} \mathbb{I}(X_i \in A)/n = \hat{P}_{\mathbf{X}}(A)$, *is a permutation invariant function over* $\mathscr{X}_{/\mathbf{X}}$.

  The latter implies that conditioning on $\mathscr{X}_{/\mathbf{X}}$ is equivalent to conditioning on the EPM $\hat{P}_{\mathbf{X}}(A)$, which then is sufficient too.
- **P.8.** (The power of test T). *The (unconditional or population) power of a PT T as a function of Δ, α, T, P, and n is defined as* $W(\Delta, \alpha, T, P, n) = \mathbb{E}_{P^n}[\Pr\{\lambda_T(\mathbf{X}(\Delta)) \le \alpha|\mathscr{X}_{/\mathbf{X}}^n\}]$. *Of course,* $W(\Delta, \alpha, T, P, n) \ge W(0, \alpha, T, P, n) = \alpha$, $\forall \alpha > 0$, *since, in force of* **P.6** *the integrand is* $\ge \alpha$ *for all* $\mathbf{X} \in \mathscr{X}_{/\mathbf{X}}^n$, *all* $P \in \mathscr{P}$, *and all* n.

It is worth noting that **P.8** implies unconditional unbiasedness. It is also to be noted that the power determination of T implies referring to the hypothetical repeated sampling principle.

To introduce the weak consistency property of PTs, stating that "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 first consider sequences of related data sets where first n_1 IID values are from $X_1(\Delta) = X + \Delta$ and the other n_2 from $X_2 = X(0) = X$. Such sequences are denoted by $\{\mathbf{X}^{(n)}(\Delta)\}_{n\in\mathbb{N}} = \{[X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}]\}_{(n_1,n_2)\in\mathbb{N}}$. Of course, $\{\mathbf{X}^{(n)}(0) = \mathbf{X}^{(n)}\}_{n\in\mathbb{N}}$ represents sequences in H_0. Besides, we assume that $n \to \infty$ implies $\min[n_1, n_2] \to \infty$.

- **P.9.** (Weak Consistency). *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 $(X_1(\Delta), \mathscr{X})$ and independently the other n_2 IID from $(X, \mathscr{X})$. Suppose that the null distribution of X is $P \in \mathscr{P}$, and let $\varphi : \mathscr{X} \to \mathscr{R}^1$ be any non-decreasing and non-degenerate measurable function. Suppose also that*: (a) *the φ-mean $\mathbb{E}_P[\varphi(X)] = \mathbb{E}_P[\varphi(X(0))]$ is finite, i.e. $\mathbb{E}_P[|\varphi(X)|] < +\infty$*; (b) *the φ-mean in H_1 is such that $\mathbb{E}_P[\varphi(X(\Delta))] > \mathbb{E}_P[\varphi(X(0))]$ for every $\Delta \overset{d}{>} 0$*; (c) *the PT is based on $T^* = \frac{1}{n_1}\sum_{i\le n_1}\varphi(X_i^*)$, or on permutationally equivalent statistics. Then, for every $\alpha > 0$*, (a)–(c) *imply that the rejection probability of the PT ϕ, associated with T^*, converges weakly to one as $n \to \infty$.*

It is worth noting 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. As an application of **P.9** we see details for proving consistency of a test based on well-known Cramér–von Mises statistic for one-sided alternatives. Indeed: (1) with $\Delta \overset{d}{>} 0$, $T^*_{CM} = \sum_{i=1}^{n}[\hat{F}_2^*(X_i) - \hat{F}_1^*(X_i)]$, where $\hat{F}_j^*(z) = \sum_i^{n_j}\mathbb{I}(X_{ji}^* \le z)/n_j$, $j = 1, 2$, is permutationally equivalent to $-\sum_{i\le n}\hat{F}_1^*(X_i)/n$, since $\hat{F}_{\mathbf{X}^{(n)}}(t) = [n_2\hat{F}_2^*(t) + n_1\hat{F}_1^*(t)]/n$ is a permutation invariant function; (2) as F_P is bounded, $\mathbb{E}_P(F_P(X))$ is finite; (3) as $\hat{F}_1^*$ is a sample mean, we have that $\Pr\{|\hat{F}_1^*(z) - \hat{F}_{\mathbf{X}^{(n)}}(z)| < \varepsilon|\hat{F}_{\mathbf{X}^{(n)}}\} \longrightarrow 1$, $\forall z \in \mathscr{R}^1$ and $\varepsilon > 0$; (4) $\Delta \overset{d}{>} 0$ implies $\mathbb{E}_P[F_P(X(\Delta))] < \mathbb{E}_P[F_P(X(0))]$. Therefore, since conditions (a)–(c) are satisfied, T^*_{CM} is weakly consistent.

5 Extending Permutation Inference

The non-randomized permutation test ϕ associated with a given test statistic T based on divergence of symmetric functions of the data possesses both conditional unbiasedness and similarity properties, the former **P.6** satisfied by *all population distributions P and all data sets $\mathbf{X} \in \mathscr{X}^n$*, the latter **P.3** satisfied for continuous, non-degenerate variables and *almost all data sets*. These two properties jointly suffice to weakly extend conditional inferences to unconditional or population ones, i.e. for the extension of conclusions related to the specific set of actually observed units (e.g., *drug is effective on the observed units*) to conclusions related to the

population from which units have been drawn (e.g., *drug is effective*). Such an extension is done with weak control of inferential errors. With clear meaning of symbols let us observe:

- (i) for each attainable α and all sample sizes n, the similarity property implies that the power of the test under H_0 satisfies $W(0, \alpha, T, P, n) = \alpha$, because $\Pr\{\lambda(\mathbf{X}(0)) \leq \alpha | \mathscr{X}^n_{/\mathbf{X}}\} = \alpha$ for almost all samples $\mathbf{X} \in \mathscr{X}^n$ and all continuous non-degenerate distributions P, independently of how data are selected;
- (ii) the uniform conditional unbiasedness implies that the unconditional power is $W(\Delta, \alpha, T, P, n) \geq \alpha$ **(P 8)** for all distributions P and, provided that $f_P^{(n)}(\mathbf{X}) > 0$, independently of how data are selected.

As a consequence, if, for instance, the inferential conclusion related to actual data $\mathbf{X}$ is in favor of H_1, so we say that "data $\mathbf{X}$ are evidence of treatment effectiveness on actually observed units," due to (i) and (ii) we are allowed to say that this conclusion is also valid unconditionally for all populations $P \in \mathscr{P}$ such that $f_P^{(n)}(\mathbf{X}) > 0$. Thus, the extended inference becomes "treatment is likely to be effective." The condition $f_P^{(n)}(\mathbf{X}) > 0$ implies that inferential extensions must be carefully interpreted. To illustrate this aspect simply, let us consider an example of an experiment in which only males of a given population of animals are observed. Hence, based on the result actually obtained, the inferential extension from the observed units to the selected sub-population is immediate. Indeed, on the one hand, rejecting the null hypothesis with the actual data means that *data are evidence for a non-null effect of treatment*, irrespective of how data are collected, provided that they are exchangeable in the null hypothesis. On the other hand, if females of that population, due to the selection procedure, have a probability of zero of being observed, then in general we can say nothing reliable regarding them, because it may be impossible to guarantee that the test statistic used for male data satisfies conditional unbiasedness and/or similarity properties for female data as well. For instance, effect may be positive on male and negative on female. In general, *the extension* (i.e., the extrapolation or the inductive generalization) *of any inference to populations which cannot be observed can be formally done only with reference to assumptions that lie outside those that are adopted under the control of experimenters while working on actual data*. For instance, extensions to humans of inferences obtained from experiments on animals essentially require specific hypothetical assumptions.

We observe that for parametric tests, when there are nuisance entities to remove, the extension of inferences from conditional to unconditional can generally be done only if the data are obtained through well-designed sampling procedures applied to the entire target population. When selection-bias data $\mathbf{X}$ are observed and the selection mechanism is not well designed and/or modelled there is no point in staying outside the conditioning on the associated sufficient orbit $\mathscr{X}_{/\mathbf{X}}$ and the related distribution induced by the chosen statistic T. On the one hand this implies adopting the permutation testing principle; on the other, no parametric approach can be invoked to obtain reliable inferential extensions.

References

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

Frosini, V.B.: On some applications of the conditionality principle. Statistica Applicata **3**, 555–568 (1991)

Hoeffding, W.: The large-sample power of tests based on permutations of observations. Ann. Math. Stat. **23**, 169–192 (1952)

Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)

Mielke, P.W., Berry, K.J.: Permutation Methods, A Distance Function Approach, 2nd edn. Springer, New York (2007)

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