5-1-2013

A Response to Anderson's (2013) *Conceptual Distinction between the Critical p value and Type I Error Rate in Permutation Testing*

Fortunato Pesarin
University of Padova, Italy

Stefano Bonnini
University of Ferrara, Italy

A Response to Anderson's (2013) *Conceptual Distinction between the Critical p value and Type I Error Rate in Permutation Testing*

Fortunato Pesarin
University of Padova,
Padova, Italy

Stefano Bonnini
University of Ferrara,
Ferrara, Italy

Pesarin and Bonnini respond to Anderson's (2013) *Conceptual Distinction between the Critical p Value and the Type I Error Rate in Permutation Testing*.

Key words: Statistics, null hypothesis, non-parametric, permutation test, exchangeability; Type I error; p-value, discrete set, significance level, validity.

Introduction

Anderson (2013) referred to various aspects of permutation p-values. We agree with most of the notions and ideas heuristically discussed there; however, we present a more formal discussion to confirm and be more precise. The approach adopted in Pesarin (2001), Pesarin (2013) and Pesarin & Salmaso (2010a, 2010b) considered permutation tests in light of the *conditionality and sufficiency principles* of inference, and we borrow freely from those citations.

Fortunato Pesarin is a Professor Emeritus of Statistics in the Department of Statistical Sciences. Email him at: pesarin@stat.unipd.it. Stefano Bonnini is Assistant Professor of Statistics in the Department of Economics and Management. Email him at: bnnsfn@unife.it.

Without loss of generality, consider a two-sample one-dimensional design for one-sided alternatives where: $X_1 = (X_{1i}; i = 1, …, n_1)$ are data of first sample, $X_2 = (X_{2i}, i = 1, …, n_2)$ are those of the second, and $X = (X_1, X_2) = (X_i, i = 1, …, n; n_1, n_2)$ is the pooled set from which it is intended that the first n_1 are from the first sample and the remainder from the second. Denoting by $\Pi(u)$ the set of permutations of unit labels $u=(1,…,n)$ and by $u^* = (u_1^*, …, u_n^*) \in \Pi(u)$ one of these, the related permutation of X is $X^* = \{X_i^* = X(u_i^*), i = 1, …, n; n_1, n_2\}$. Thus, $X_1^* = \{X_{1i}^* = X(u_i^*), i = 1, …, n_1\}$ and $X_2^* = \{X_{2i}^* = X(u_i^*), i = n_1 + 1, …, n\}$ denote the two permuted samples; in multidimensional problems vectors associated with units are permuted (Basso, et al., 2009; Pesarin, 2001; Pesarin & Salmaso, 2010a; Pesarin & Salmaso, 2011). The hypotheses are

$$H_0 : X_1 \overset{d}{=} X_2 \quad \text{and} \quad H_1 : X_1 \overset{d}{\geq} X_2,$$ respectively stating the equality and the dominance in

distribution of two population variables. Note that H_0 implies data are exchangeable without being necessarily independent; in practice the exchangeability condition is usually attained when subjects are randomized to treatments or when related to observations on independent units.

Permutation tests are ordinary tests conditional on the observed data set X, which in H_0 is always a sufficient n-dimensional statistic for the underlying (known or unknown) distribution f in the sense that it contains all the information on f that are contained in X. Outside the regular exponential family of distributions X is generally minimal sufficient. In particular this implies that no suitable univariate statistic exists for summarizing the whole information on the aspect of interest for the analysis; therefore, no parametric or nonparametric method can aspire to be uniformly better than others.

One way to attenuate this drawback within the multi-aspect methodology (Pesarin, 2001; Pesarin & Salmaso, 2010a; Pesarin & Salmaso, 2011; Salmaso & Solari, 2005) is to use several partial test statistics, each specialized to put into evidence one aspect of interest and to nonparametrically combine their dependent results by means of a suitable function.

The act of conditioning on X implies referring to the related conditional reference space, denoted by $\chi_{/X}$. Essentially $\chi_{/X}$ contains points of sample space χ^n that are equivalent to X in terms of information carried by the underlying – and usually unknown – likelihood function $f^{(n)}(X)$, which in turn is assumed to be positive on all observed data X. Thus, it contains all points X^* such that the likelihood ratio $f^{(n)}(X)/f^{(n)}(X^*)$ is (X, X^*)-invariant and f-independent; so it corresponds to the orbit of equivalent points associated with X. Data exchangeability in H_0 implies that the density $f^{(n)}(X)$ is invariant on permutations of its arguments, thus $f^{(n)}(X) = f^{(n)}(X^*)$ for every permutation X^* of X; $\chi_{/X}$ then contains all distinct permutations of X. That is $\chi_{/X} = \{\cup_{u^* \in \Pi(u)}[X(u_i^*), i = 1, \dots, n]\}$. Therefore, because every $X^* \in \chi_{/X}$ is sufficient for f in H_0, $\chi_{/X}$ can be considered as playing the role of a *sufficient space*.

In H_1 the set of sufficient statistics is the pair $(X_1; X_2)$ because data are only exchangeable within – but not between – groups. As a consequence of sufficiency of $\chi_{/X}$ for f under H_0 the null conditional probability of any event A, given $\chi_{/X}$, is independent of f – whatever it is – because $Pr\{X^* \in A; f|\chi_{/X}\} = Pr\{X^* \in A|\chi_{/X}\}$. Therefore, the null permutation (conditional) probability is *distribution-free* and *nonparametric* as well as the distribution of any statistic $T: \chi^n \to \mathcal{R}^1$ conditional on $\chi_{/X}$. Moreover, because for finite sample size n the number $M^{(n)} = \sum_{\chi_{/X}} 1(X^* \in \chi_{/X})$ of points in $\chi_{/X}$ is finite, in H_0 the conditional probability of any A is calculated as

$$Pr\{X^* \in A|\chi_{/X}\} = \frac{\sum_{X^* \in A} f(X^*)dX^*}{\sum_{X^* \in \chi_{/X}} f(X^*)dX^*}$$

$$= \sum_{\chi_{/X}} \frac{I(X^* \in A)}{M^{(n)}},$$

because $f(X^*)dX^* = f(X)dX$ for every $X^* \in \chi_{/X}$. Note that in carrying out the calculations for this conditional probability it is not necessary to make reference to the so-called *hypothetical repeated sampling principle*; that is, it is not necessary to examine the whole population sample space χ^n, which in turn has generally a virtual existence, and to consider all sample points that could have been realized but only the point X of which has been actually observed. In fact, as the observed data X are exclusively taken into consideration, $Pr\{A|\chi_{/X}\}$ simply requires the *complete enumeration* of $\chi_{/X}$ which, as it consists of all permutations X^* of X, has an *objective existence*.

With different arguments, a similar notion is also expressed in Anderson's article. In practice, when sample sizes (n_1, n_2) are not small, to overcome the too large cardinality of $\chi_{/X}$, the probability $Pr\{A|\chi_{/X}\}$ can be estimated at any degree of accuracy by a *conditional Monte Carlo simulation* on $\chi_{/X}$ as described by Anderson.

A first property of the permutation distribution is that in H_0 the data set X is

uniformly distributed over $\chi_{/X}$ conditionally. In fact, $\Pr\{X = x|\chi_{/X}\} = \Pr\{X^* = x|\chi_{/X}\} = 1/M^{(n)}$ if $x \in \chi_{/X}$ and 0 elsewhere. Essentially, this property states that, under H_0, the permutations in $\chi_{/X}$ are equally likely; whereas under H_1, because the set of sufficient statistics is the pair $(X_1; X_2)$, no data can be exchanged between two samples and permutations in $\chi_{/X}$ are not equally likely. This property may suggest ways for finding test statistics provided with *good* inferential properties. Further, assuming that the exchangeability condition on data X is satisfied in H_0, for any randomized test based on a statistic T for which large values are significant,

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

where $T^0 = T(X)$ is the observed value of T on data X, and

$$\gamma = [\alpha - \Pr\{T^0 > T_\alpha|\chi_{/X}\}]/\Pr\{T^0 = T_\alpha|\chi_{/X}\},$$

the α-size conditional critical value $T_\alpha = T_\alpha(X)$ which, if H_0 were true, could be determined by complete enumeration of $\chi_{/X}$. The conditional rejection probability of $\phi_R(X)$ for any given $\alpha \in (0,1)$ is then $\mathbb{E}\{\phi_R(X)|\chi_{/X}\} = \alpha$. Thus, under H_0, due to sufficiency of X, test $\phi_R(X)$ is invariant with respect to all underlying distributions f and all sample points X and so independently on how they are selected from χ^n. This property defines the so-called *uniform similarity property of randomized permutation tests*, corresponding to the stronger version of the Neymann α-structure.

One consequence, but not the most important, of conditioning on X is that the permutation critical value $T_\alpha = T_\alpha(X)$ does not need to be determined in practice. Unless it is known that H_0 is true – in which case knowing T_α would be essentially useless – its determination would require the ordering of $\{T(X^*), X^* \in \chi_{/X}\}$, that is, that of all permutation values of T, thus it necessarily depends on actual data X and on some unknown treatment effect δ (if any). One efficient way to overcome this apparent drawback is by making reference to the *significance level transformation* of T, that is, by considering the so-called observed p-value-like score. The observed p-value-like score of test ϕ_R, which is defined as $\lambda_R^0 = \lambda_R(X, T) = \Pr\{T^* > T^0|\chi_{/X}\}$, is a non-increasing function of T^0 and is one-to-one related with the α-value of T. Because $T^0 < T_\alpha$ implies $\lambda_R^0 > \alpha$ and $T^0 > T_\alpha$ implies $\lambda_R^0 < \alpha$:

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

is the general form of all permutation tests. Of course, simple or non-randomized or ordinary test $\phi(X)=1$ if $T \geq T_\alpha$ and 0 elsewhere, the p-value score $\lambda^0 = \Pr\{T^* \geq T^0|\chi_{/X}\}$ is defined in the discrete set of *attainable values*, $\Lambda(X, T) \in (0,1)$.

The statistic λ^0 coincides with the p-value of test T if H_0 were true, then it works as a p-value score. Thus, we agree with Anderson's notion that λ^0 is not a true p-value, unless H_0 were true. We partially disagree, however, because α is simply the critical value (i.e., the type I error rate α) of any randomized permutation test (for non-randomized tests, the resulting type I error rate is $\alpha' \leq \alpha$ so giving rise to conservative inferences); exactly as for parametric tests.

In order to establish whether the p-value score is a test statistic furnished with good inferential properties it is useful to observe that:

i. in H_0 the distribution of λ^0 is uniform on its attainable set $\Lambda(X, T)$;

ii. in $\quad H_1: \left(X_1 \overset{d}{\geq} X_2\right) = \left(X + \delta \overset{d}{\geq} X_2 = X\right) \equiv$

$(\delta > 0)$, where the unknown treatment effect δ can be either fixed or random and, in the latter case, it could depend on latent unobservable values X. The permutation distribution of the p-value score is such that $\lambda^0(X, \delta) \overset{d}{\leq} \lambda^0(X, 0) \overset{d}{=} \mathcal{U}[\Lambda(X, T)]$, and thus it

is uniformly dominated by the uniform

distribution over $\Lambda(\boldsymbol{X}, T)$; moreover, $\delta' > \delta > 0$ implies that

$$\lambda^0(\boldsymbol{X}, \delta') \overset{d}{\leq} \lambda^0(\boldsymbol{X}, \delta) \overset{d}{\leq} \lambda^0(\boldsymbol{X}, 0),$$ so the

dominance in distribution is monotone with respect to δ for every $\boldsymbol{X}$.

Provided that the exchangeability property is satisfied in H_0, these properties imply that the test statistic based on λ^0 is uniformly unbiased for all data sets $\boldsymbol{X}$, all distributions f, all sample sizes (n_1, n_2) and independently on how subjects are selected from a population (i.e., possibly under selection-bias sampling if they are randomized to treatments). These properties guarantee that tests based on the permutation p-value score are valid for making inference whenever the permutation testing principle applies (Pesarin, 2001; Pesarin & Salmaso, 2010a; Pesarin & Salmaso, 2010b). In this we agree with Anderson's view-point that the permutation p-value does not impugn the validity of related tests.

With respect to the consistency property, if in H_0 the population mean value of X is finite, that is, $\mathbb{E}(|X|) < \infty$, then the permutation test based on difference of two sample means, $T = \bar{X}_1 - \bar{X}_2$ for example, is consistent (Pesarin & Salmaso, 2013), that is, its rejection probability converges to 1 as $min(n_1, n_2)$ diverges. Thus, not all test statistics are "valid irrespective of the shapes of the population distributions," in the sense that there exist population f for which there are permutation tests T that are not consistent. A way to obtain valid consistent test statistics is described in (Pesarin & Salmaso, 2010a; Pesarin & Salmaso, 2010b; Salmaso & Solari, 2005) within the so-called multi-aspect testing.

We would like to also underline some few other points:

1. The permutation test is unbiased and consistent when the effect δ is such that the distribution of $X_1(\delta)$ in the alternative is not only shifted, but also when the dispersion increases or decreases provided that $X_1(\delta) \overset{d}{\geq} X_2$. For example, the effect on

high blood pressure of a suitable drug is to reduce it towards smaller values by diminishing both mean and dispersion. That is, it is valid under homoscedasticity in H_0 (when the exchangeability is satisfied) and for possible non-homoscedasticity in H_1. When $\mu_1 = \mu_2$ and homoscedasticity is not satisfied in H_0, so the exchangeability is also not satisfied, it falls within the so-called Behrens-Fisher problem for which no exact parametric and non-parametric non-randomized solutions exist. In Pesarin (2001, Chapter 10) an almost exact (but not exact) univariate and multivariate permutation solution is provided for symmetric distributions. Thus, in our opinion, some of the related arguments in Anderson's paper should be accordingly modified.

2. Traditional parametric tests are said to be generally valid under their conditions. This, however, is not always true. For example, in a simulation under multivariate normality we found that as the number q of processed variables increases, while keeping sample sizes fixed, the power of Hotelling's T^2 increased up to a maximum at approximately $q \cong (n-2)/2$, then it decreased and was not calculable when $q \geq (n-1)$, whereas that of a permutation test based on $T_p = \sum_{h \leq q} (\bar{X}_{1h} - \bar{X}_{2h})^2$ always increased up to 1 for divergent q. Moreover, depending on the correlation matrix, there is a value q' such that the permutation test T_p is more powerful than T^2. To this end we would like to report a brief simulation study from Pesarin and Salmaso (2010b). The results in Table 1, where $\alpha = 0.01$, **0.05** (bold face), B = 1000 random permutations and MC = 1000 Monte Carlo simulations from q-dimensional normal distributions $X_1 \overset{d}{=} \mathcal{N}_q(\boldsymbol{\delta}, \boldsymbol{1})$ and

$X_2 \overset{d}{=} \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{1})$ were considered, show that:

(i) as q increases, the power of Hotelling's T^2 increases up to a maximum and then decreases to a minimum for $q = n–2$, after

that it cannot be calculated due to singularity of estimated covariance matrix; (ii) power of T_p increases monotonically with q; (iii) power of T_p is not invariant with respect to alternatives lying at Mahalanobis distance from H_0 and so in some circumstances and even for small q it can be more powerful than T^2, which in stringent conditions is the uniformly most powerful unbiased similar invariant (whereas T_p is simply unbiased). A similar result, however less general, is also in Blair, Higgins, Karniski and Kromrey (1994).

Table 1: Simulations under
$H_1: n_1 = n_2 = 10,\ \mu = 0,\ \delta = 0.40$

q	T^2	T_p
4	.079/**.219**	.081/**.237**
8	.063/**.234**	.126/**.347**
12	.037/**.186**	.176/**.436**
18	.013/**.067**	.253/**.543**
19		.244/**.544**
22		.340/**.618**
25		.365/**.656**

3. Under selection-bias sampling, any parametric inference cannot be extended to the target population, even asymptotically, unless a suitable treatment independent selection model is well-defined, identifiable and estimable. Conversely, under mild conditions and based on the uniform similarity and the uniform unbiasedness, permutation inference can be extended (Pesarin, 2001; Pesarin, 2002; Pesarin & Salmaso, 2010a, Pesarin & Salmaso, 2010b) to the possibly unknown population from which selected subjects belong. Therefore, the result of the inferential conclusion is not only valid to the actual subjects (e.g., *treatment is effective on actual subjects*), as suggested by Edgington and Onghena (2007) and by Lehmann (2009), it is also valid to the latent population (e.g., *treatment is effective*), at least in a weak sense. This in

part contradicts one of Anderson viewpoints.

4. When in a multivariate testing problem the alternatives are restricted to be one-sided and the underlying distribution is not multivariate normal with known covariance matrix, no parametric solution is known (under multivariate normality some asymptotic solution are known (Silvapulle & Sen, 2005)); whereas within the permutation testing principle these solutions are straightforward through the so-called nonparametric combination of dependent partial tests, one for each component variable (Basso, Pesarin, Salmaso & Solari, 2009; Basso & Salmaso, 2011; Finos, Salmaso & Solari, 2007; Pesarin & Salmaso, 2010a). This, together with the fact that no single test can recover all of the information on f contained in X, may solve the question regarding the "unresolved ambiguity concerning the degree to which each of the ... tests ... provides unique information about the null hypothesis" expressed in Andersons manuscript.

5. Reported results from Hayes (1996) regarding the non "distribution-free" property of permutation tests are, in our opinion, misleading because in order for the permutation test based on $\sum X_i Y_i$ for testing $H_0: \rho = 0$ to be valid, the null hypothesis *must* be $H_0: X \perp Y$, that is, *independence must be assumed* between X and Y, not only zero correlation. In other words it is not possible to test $\rho = 0$ via permutation without assuming independence due to the lack of exchangeability (remember that $\rho = 0$ characterizes independence on bivariate normal distributions). In a sense, this implies some limitations on the inferential possibilities of permutation tests.

6. One important – and very useful property – of permutation tests is when there are missing or censoring or zero-inflated data informative on the treatment effect (see Pesarin, 2001; Pesarin & Salmaso, 2010a; Pesarin & Salmaso, 2010b) where, unless a

suitable, identifiable and separately estimable model is provided, no parametric test exists (Sen, 2007).

7. Permutation testing can be used in an intriguing problem as that in which some treated subjects have negative and others positive effect, as for example with some drugs sensitive to genetic interaction, so that both $\delta < 0$ and $\delta > 0$ can be jointly true in the alternative. This problem is studied in Bertoluzzo, Pesarin & Salmaso (2013).

8. In Pesarin and Salmaso (2010a) and Pesarin and Salmaso (2010c) an extended notion of consistency is provided for permutation tests where it is assumed that sample sizes remain fixed and the number of informative variables, those which contribute with non-null effect, diverge to the infinity. A notion which can have application, for example, in many problems of modern biological analyses where the number of observed variables per subject can be practically unlimited or even in problems where data are curves or images which can be reduced to a large number of coefficients (e.g., functional principal components, Fourier and/or wavelets coefficients, area under the curve and/or number of picks over a list of thresholds, etc.).

9. It is possible to download without cost the software NPC-Test R.10 from the Wiley website http://www.wiley.com/go/npc for Pesarin and Salmaso (2010a), which may solve the large majority of unidimensional and multidimensional permutation testing problems. R coded programs are also available for download from the same website.

References

Anderson, R. B. (2013). Conceptual Distinction between the Critical p Value and the Type I Error Rate in Permutation Testing. *Journal of Modern Applied Statistical Methods*, *12*(*1*), 2-8.

Basso, D., Pesarin, F., Salmaso, L., & Solari, A. (2009). *Permutation tests for stochastic ordering and ANOVA: Theory and applications*. New York, NY: Springer.

Basso, D., & Salmaso, L. (2011). A permutation test for umbrella alternatives. *Statistics and Computing*, 21, 45-54.

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

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's t2 test in prescribed circumstances. *Multivariate Behavioral Research*, *29*, 141-163.

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

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

Pesarin, F. (2001). *Multivariate permutation tests: with application in biostatistics*. Chichester, UK: John Wiley & Sons.

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

Pesarin, F. (2013). Some elementary theory of permutation tests. *Communications in Statistics - Theory and Methods*. Submitted, R1 version.

Pesarin F., & Salmaso, L. (2010a). *Permutation tests for complex data. Theory, applications and software: Wiley series in probability and statistics*. Chichester, UK: John Wiley & Sons.

Pesarin F., & Salmaso, L. (2010b). The permutation testing approach: a review. *Statistica*, *LXX*(*4*), 1-29.

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

Pesarin, F. & Salmaso, L. (2011). A review and some new results on permutation testing for multivariate problems. *Statistics and Computing*, *22*(*2*), 639-646.

Pesarin, F., & Salmaso, L. (2013). On the weak consistency of permutation tests. *Communications in Statistics - Simulation and Computation*, 42, 1368-1379.

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

Sen, P. K. (2007). Union-intersection principle and constrained statistical inference. *Statistical Planning and Inference*, *137*, 3741-3752.

Silvapulle, J. S., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order and shape restrictions*. Hoboken, NJ: Wiley.