Lecture Notes in Statistics 194
Edited by P. Bickel, P.J. Diggle, S. Fienberg,
U. Gather, I. Olkin, S. Zeger

For other titles published in this series, go to
www.springer.com/series/694

Dario Basso · Fortunato Pesarin · Luigi Salmaso · Aldo Solari

Permutation Tests for Stochastic Ordering and ANOVA

Theory and Applications with R

Springer

Dario Basso
Università di Padova
Dip. to Tecnica e Gestione dei
Sistemi Industriali
Stradella San Nicola, 3
36100 Vicenza
Italy
basso@gest.unipd.it

Fortunato Pesarin
Università di Padova
Dip. to Scienze Statistiche
Via Cesare Battisti, 241/243
35121 Padova
Italy
pesarin@stat.unipd.it

Luigi Salmaso
Università di Padova
Dip. to Tecnica e Gestione dei
Sistemi Industriali
Stradella San Nicola, 3
36100 Vicenza
Italy
salmaso@gest.unipd.it

Aldo Solari
Università di Padova
Dip. to Processi Chimici
dell' Ingegneria
via Marzolo, 9
35131 Padova
Italy
aldo.solari@unipd.it

ISBN 978-0-387-85955-2 e-ISBN 978-0-387-85956-9
DOI 10.1007/978-0-387-85956-9
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: PCN applied for

© Springer Science+Business Media, LLC 2009
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is divided into two parts. The first part is devoted to some advances in testing for a stochastic ordering, and the second part is related to ANOVA procedures for nonparametric inference in experimental designs. It is worth noting that, before introducing specific arguments in the two main parts of the book, we provide an introductory first chapter on basic theory of univariate and multivariate permutation tests, with a special look at multiple-comparison and multiple testing procedures.

The concept of stochastic ordering of distributions was introduced by Lehmann (1955) and plays an important role in the theory of statistical inference. It arises in many applications in which it is believed that, given a response variable Y and an explanatory variable x, the statistical model assumes that the distribution of $Y|x$ belongs to a certain family of probability distributions that is ordered in the sense, roughly speaking, that large values of x lead to large values of the Y's.

Many types of orderings of varying degrees of strength have been defined in the literature to compare the order of magnitude of two or more distributions (see Shaked and Shanthikumar, 1994, for a review). These include likelihood ratio ordering, hazard rate ordering, and simple stochastic ordering, which are perhaps the main instances. On the one hand, these orderings make the statistical inference procedures more complicated. On the other, they contain statistical information as well, so that if properly incorporated they would be more efficient than their counterparts, wherein such constraints are ignored. These considerations emphasize the importance of statistical procedures to detect the occurrence of such orderings on the basis of random samples. Inference based on stochastic orderings for univariate distributions has been studied extensively, whereas for multivariate distributions it has received much less attention because the "curse of dimensionality" makes the statistical procedures considerably more complicated. For a review of constrained inference, we refer to the recent monograph by Silvapulle and Sen (2005).

Likelihood inference is perhaps the default methodology for many statistical problems; indeed, the overwhelming majority of work related to order-

restricted problems is based on the likelihood principle. However, there are instances when one might prefer a competitive procedure. Recently there have been debates about the suitability of different test procedures: Perlman and Chaudhuri (2004a) argue in favor of likelihood ratio tests, whereas Cohen and Sackrowitz (2004) argue in favor of the so-called class of directed tests. In multidimensional problems, it is rare that a "best" inference procedure exists. However, even in such a complex setup, following Roy's union-intersection principle (Roy, 1953), it might be possible to look upon the null hypothesis as the intersection of several component hypotheses and the alternative hypothesis as the union of the same number of component alternatives, giving rise to a multiple testing problem. A classical approach is to require that the probability of rejecting one or more true null hypotheses, the familywise error rate (Hochberg and Tamhane, 1987), not exceed a given level. Generally, it is surprising that some existing procedures seem to be satisfied to stop with a global test just dealing with the acceptance or rejection of the intersection of all null hypotheses. In the form presented, it will be difficult to interpret a statistically significant finding: The statistical significance of the individual hypotheses in multiple-endpoint or multiple-comparison problems remains very important even if global tests indicate an overall effect. Indeed, most clinical trials are conducted to compare a treatment group with a control group on multiple endpoints, and the inferential goal after establishing an overall treatment effect is to identify the individual endpoints on which the treatment is better than the control. For tests of equality of means in a one-way classification, the ANOVA F test is available, but in the case of rejection of the global null hypothesis of equality of all means, one will frequently want to know more about the means than just that they are unequal.

In the majority of the situations we shall deal with, both the hypothesis and the class of alternatives may be nonparametric, and as a result it may be difficult even to construct tests that satisfactorily control the level (exactly or asymptotically). For such situations, we will consider permutation methods that achieve this goal under fairly general assumptions. Under exchangeability of the data, the empirical distribution of the values of a given statistic recomputed over transformations of the data serves as a null distribution; this leads to exact control of the level in such models. In addition, by making effective use of resampling to implicitly estimate the dependence structure of multiple test statistics, it is possible to construct valid and efficient multiple testing procedures that strongly control the familywise error rate, as in Westfall and Young (1993).

We bring out the permutation approach for models in which there is a possibly multivariate response vector $\boldsymbol{Y}$ and an ordinal explanatory variable x taking values $\{1, \ldots, k\}$, which can be thought of as several levels of a treatment. Let $\boldsymbol{Y}_i$ denote the random vector whose distribution is the conditional distribution of $\boldsymbol{Y}$ given $x = i$. We are interested in testing $\boldsymbol{Y}_1 \overset{\mathrm{d}}{=} \ldots \overset{\mathrm{d}}{=} \boldsymbol{Y}_k$ against a stochastic ordering alternative $\boldsymbol{Y}_1 \overset{\mathrm{st}}{\leq} \ldots \overset{\mathrm{st}}{\leq} \boldsymbol{Y}_k$ with at least one $\overset{\mathrm{st}}{<}$.

In the statistical literature, there is relatively little on multivariate models for nonnormal response variables, such as ordinal response data. This is perhaps due to the mathematical intractability of reasonable models and to related computational problems. The aim is therefore to provide permutation methods that apply to multivariate discrete and continuous data. We deal with univariate and multivariate ordinal data in Chapters 2 and 3, respectively, and Chapter 4 contains results for multivariate continuous responses.

As previously said, the second part of the book is dedicated to nonparametric ANOVA within the permutation framework. Experimental designs are useful research tools that are applied in almost all scientific fields. In factorial experiments, processes of various natures whose behavior depends on several factors are studied. In this context, a factor is any characteristic of the experimental condition that might influence the results of the experiment. Every factor takes on different values, called levels, that can be either quantitative (dose) or qualitative (category). When several factors are observed in an experiment, every possible combination of their levels is called a treatment. The analysis of factorial designs through linear models allows us to study (and assess) the effect of the experimental factors on the response, where factors are under the control of the experimenter. They also allow for evaluating the joint effect of two or more factors (also named main factors), which are known as interaction factors. The statistical analysis is usually carried on by assuming a linear model to fit the data. Here, the model to fit the response is an additive model, where the effect of main factors and interactions are represented by unknown parameters. In addition, a stochastic error component is considered in order to represent the inner variability of the response. Usually, errors are assumed to be i.i.d. homoscedastic random variables with zero mean. This model requires some further assumptions in order to be applied. Some of them, such as independence among experimental units or the identical distribution, are reasonable and supported by experience. Other assumptions, such as normality of the experimental errors, are not always adequate. Generally it is possible to check the assumption of normality only after the analysis has been made, through diagnostic tools such as the $Q-Q$ plot (Daniel, 1959). Nevertheless, these tools are mainly descriptive; therefore the conclusions they may lead to are essentially subjective. If the normality of errors is not satisfied or cannot be justified, then the usual test statistics (such as the Student t test or the F test) are approximate. It is therefore worthwhile to reduce some assumptions, either to avoid the use of approximate tests or to extend the applicability of the methods applied.

Permutation tests represent the ideal instrument in the experimental design field since they do not require assumptions on the distribution of errors and, if normality can be assumed, they give results almost as powerful as their parametric counterpart. There are other reasons to use permutation tests; for instance, in the $I \times J$ replicated designs, even if data are normally distributed, the two-way ANOVA test statistics are positively correlated. This means that the inference on one factor may be influenced by other factors. There are

other situations where parametric tests cannot be applied at all: In unreplicated full factorial designs, the number of observations equals the number of parameters to estimate in the model; therefore there are no degrees of freedom left to estimate the error variance. Permutation tests deal with the notion of exchangeability of the responses: The exchangeability is satisfied if the probability of the observed data is invariant with respect to random permutations of the indexes. The exchangeability of the responses is a sufficient condition to obtain an exact inference. In factorial design, the responses are generally not exchangeable since units assigned to different treatments have different expectations. Thus, either a restricted kind of permutation is needed or approximate solutions must be taken into account in order to obtain separate inferences on the main factor/interaction effects.

Chapter 5 is an introduction to ANOVA in a nonparametric view. Therefore, the general layout is introduced with minimal assumptions, with some particular care about the exchangeability of errors. Some of the solutions from the literature are introduced and discussed. The kinds of errors that may arise (individual and family wise errors) in such a context are introduced, and some preliminary methods to control them are suggested. The final part of the chapter leads with direct applications of the existing methods from the literature to practical examples.

In Chapter 6 a nonparametric solution to test for effects in replicated designs is introduced. This part is dedicated to extending the solution proposed by Pesarin (2001) and Salmaso (2003) for a 2×2 balanced replicated factorial design with n units per treatment. Since the responses are not exchangeable, the solution is based on a particular kind of permutations, named synchronized permutations. In particular, by exchanging units within the same level of a factor and by assuming the standard side conditions on the constraints, it is possible to obtain a test statistic for main factors and interactions that only depends on the effects under testing and on a combination of exchangeable errors. The proposed tests are uncorrelated with each other, and they are shown to be almost as powerful as the two-way ANOVA test statistics when errors are normally distributed. After introducing the test statistics, two algorithms are proposed to obtain Monte Carlo synchronized permutations. If we desire a post hoc comparison, simultaneous confidence intervals on all pair wise comparisons can be obtained by similarly applying synchronized permutations. The tests proposed are then compared with the classical parametric analysis.

Chapter 7 is devoted to the problem of the unreplicated full factorial design analysis. Again, the problem of exchangeability of the responses arises and, given the peculiarity of the problem, it does not seem possible to obtain exact permutation tests for all factors unless testing for the global null hypothesis that there are no treatment effects. The paired permutation test introduced by Pesarin and Salmaso (2002) is exact, but it is only applicable to the first M largest effects. A further approximate solution is then proposed. Such a solution is based on the decomposition of the total response variance

under the full model and under some restricted models that are obtained in accordance with the null hypothesis under testing. The test statistic is a ratio of uncorrelated random variables, that allows us to evaluate the increase of explained variance in the full model due to the main effect under testing. The proposed test statistic allows the individual error rate to be controlled under the effect sparsity assumption. It does not control the experimental error rate, and its power is a decreasing function of the number of active effects and their sizes (the bigger the size of one effect, the bigger the noncentrality parameter in the denominator of the test statistic). To allow of control the experiment-wise error rate and in order to gain power, another version of the statistical procedure is introduced, a step-up procedure based on the comparison among noncentrality parameters of the estimates of factor effects. This test needs a calibration, which requires the central limit theorem, in order to control the experiment-wise error rate. The calibration can be obtained by either providing some critical p-values for each step of the procedure in accordance with a Bonferroni (or Bonferroni-Holm) correction or by obtaining a single critical p-value based on the distribution of the minP from simulated data under the global null hypothesis. This test is shown to be very powerful, as it can detect active factors even when there is no effect sparsity assumption (except on the smallest estimated effect, which cannot be tested). Note that a similar calibration can be provided in order to control the individual error rate at level α by choosing the critical α-quantile from the simulated null distribution of the sequential p-values. A power comparison with Loughin and Noble's test (1997) and an application from Montgomery (1991) are finally reported and discussed.

Each chapter of the book contains `R` code to develop the proposed theory. All `R` codes and related functions are available online at www.gest.unipd.it/~salmaso/web/springerbook. This Website will be maintained and updated by the authors, also providing errata and corrigenda of the code and possible mistakes in the book.

The authors wish to thank John Kimmel of Springer-Verlag and the referees for their valuable comments and publishing suggestions. In addition, they would like to acknowledge the University of Padova and the Italian Ministry for University and Scientific and Technological Research (MIUR - PRIN 2006) for providing the financial support for the necessary research and developing part of the `R` codes.

Dario Basso, Fortunato Pesarin, Luigi Salmaso, and Aldo Solari
Padova, December 2008

corresponding author: Luigi Salmaso, email: salmaso@gest.unipd.it
book Website: www.gest.unipd.it/~salmaso/web/springerbook.htm

This work has been supported by the Italian Ministry of University and Research (PRIN 2006133284_003) and by the University of Padova (CPDA088513). Research projects' coordinator: Prof. Luigi Salmaso.

Contents

1

Permutation Tests

1.1 Introduction

This book deals with the permutation approach to a variety of univariate and multivariate problems of hypothesis testing in a nonparametric framework. The great majority of univariate problems may be usefully and effectively solved within standard parametric or nonparametric methods as well, although in relatively mild conditions their permutation counterparts are generally asymptotically as good as the best parametric ones. Moreover, it should be noted that permutation methods are essentially of a nonparametrically exact nature in a conditional context. In addition, there are a number of parametric tests the distributional behavior of which is only known asymptotically. Thus, for most sample sizes of practical interest, the relative lack of efficiency of permutation solutions may sometimes be compensated by the lack of approximation of parametric asymptotic counterparts. Moreover, when responses are normally distributed and there are too many nuisance parameters to estimate and remove, due to the fact that each estimate implies a reduction of the degrees of freedom in the overall analysis, it is possible for the permutation solution to become better than its parametric counterpart (see, for example, Chapter 6). In addition, assumptions regarding the validity of parametric methods (such as normality and random sampling) are rarely satisfied in practice, so that consequent inferences, when not improper, are necessarily approximated, and their approximations are often difficult to assess.

For most problems of hypothesis testing, the observed data set $\mathbf{y} = \{y_1, \ldots, y_n\}$ is usually obtained by a symbolic experiment performed n times on a population variable Y, which takes values in the sample space $\mathcal{Y}$. We often add the adjective *symbolic* to names such as experiments, treatments, treatment effects, etc., in order to refer to experimental, pseudo-experimental, and observational contexts. For the purposes of analysis, the data set $\mathbf{y}$ is generally partitioned into *groups* or *samples*, according to the so-called *treatment levels* of the experiment. In the context of this chapter, we use capital letters

D. Basso et al., *Permutation Tests for Stochastic Ordering and ANOVA*, Lecture Notes in Statistics, 194, DOI 10.1007/978-0-387-85956-9_1,
© Springer Science+Business Media, LLC 2009

for random variables and lower case letters for the observed data set. In some sections, we shall dispense with this distinction because the context is always sufficiently clear. Of course, when a data set is observed at its $\mathbf{y}$ value, it is presumed that a sampling experiment on a given underlying population has already been performed, so that the resulting sampling distribution is related to that of the parent population, which is usually denoted by P.

For any general testing problem, in the null hypothesis (H_0), which usually assumes that data come from only one (with respect to groups) unknown population distribution P, the whole set of observed data $\mathbf{y}$ is considered to be a random sample, taking values on sample space $\mathcal{Y}^n$, where $\mathbf{y}$ is one observation of the n-dimensional sampling variable $\mathbf{Y}^{(n)}$ and where this random sample does not necessarily have independent and identically distributed (i.i.d.) components. We note that the observed data set $\mathbf{y}$ is always a set of sufficient statistics in H_0 for any underlying distribution. In order to see this in a simple way, let us assume that H_0 is true and all members of a nonparametric family $\mathcal{P}$ of nondegenerate and distinct distributions are dominated by one *dominating* measure ξ; moreover, let us denote by f_P the density of P with respect to ξ, by $f_P^{(n)}(\mathbf{y})$ the density of the sampling variable $\mathbf{Y}^{(n)}$, and by $\mathbf{y}$ the data set. As the identity $f_P^{(n)}(\mathbf{y}) = f_P^{(n)}(\mathbf{y}) \cdot 1$ is true for all $\mathbf{y} \in \mathcal{Y}^n$, except for points such that $f_P^{(n)}(\mathbf{y}) = 0$, due to the well-known factorization theorem, any data set $\mathbf{y}$ is therefore a sufficient set of statistics for whatever $P \in \mathcal{P}$.

Note that a family of distributions $\mathcal{P}$ is said to behave nonparametrically when we are not able to find a parameter θ, belonging to a known finite-dimensional parameter space Θ, such that there is a one-to-one relationship between Θ and $\mathcal{P}$ in the sense that each member of $\mathcal{P}$ cannot be identified by only one member of Θ and vice versa.

By the *sufficiency*, *likelihood*, and *conditionality principles of inference* for a review, see Cox and Hinkley, 1974, Chapter 2), given a sample point $\mathbf{y}$, if $\mathbf{y}^* \in \mathcal{Y}^n$ is such that the likelihood ratio $f_P^{(n)}(\mathbf{y})/f_P^{(n)}(\mathbf{y}^*) = \rho(\mathbf{y}, \mathbf{y}^*)$ is not dependent on f_P for whatever $P \in \mathcal{P}$, then $\mathbf{y}$ and $\mathbf{y}^*$ are said to *contain essentially the same amount of information with respect to* P, so that they are equivalent for inferential purposes. The set of points that are equivalent to $\mathbf{y}$, with respect to the information contained, is called *the coset of* $\mathbf{y}$ or *the orbit associated with* $\mathbf{y}$, and is denoted by $\mathcal{Y}^n_{/\mathbf{y}}$, so that $\mathcal{Y}^n_{/\mathbf{y}} = \{\mathbf{y}^* : \rho(\mathbf{y}, \mathbf{y}^*)$ *is* f_P*-independent*$\}$. It should be noted that, when data are obtained by random sampling with i.i.d. observations, so that $f_P^{(n)}(\mathbf{y}) = \prod_{1 \le i \le n} f_P(y_i)$, the orbit $\mathcal{Y}^n_{/\mathbf{y}}$ associated with $\mathbf{y}$ contains all permutations of $\mathbf{y}$ and, in this framework, the likelihood ratio satisfies the equation $\rho(\mathbf{y}, \mathbf{y}^*) = 1$. Also note that, as in Chapter 6, orbits of f_P-invariant points may be constructed without permuting the whole data set.

The same conclusion is obtained if $f_P^{(n)}(\mathbf{y})$ is assumed to be invariant with respect to permutations of the arguments of $\mathbf{y}$; i.e., the elements $(y_1, \ldots, y_n)$. This happens when the assumption of independence for observable data is

replaced by that of *exchangeability*, $f_P^{(n)}(y_1,\ldots,y_n) = f_P^{(n)}(y_{u_1^*},\ldots,y_{u_n^*})$, where $(u_1^*,\ldots,u_n^*)$ is any permutation of $(1,\ldots,n)$. Note that, in the context of permutation tests, this concept of exchangeability is often referred to as the *exchangeability of the observed data with respect to groups*. Orbits $\mathcal{Y}^n_{/\mathbf{y}}$ are also called *permutation sample spaces*. It is important to note that orbits $\mathcal{Y}^n_{/\mathbf{y}}$ associated with data sets $\mathbf{y} \in \mathcal{Y}^n$ always contain a finite number of points, as n is finite.

Roughly speaking, permutation tests are conditional statistical procedures, where conditioning is with respect to the *orbit* $\mathcal{Y}^n_{/\mathbf{y}}$ *associated with the observed data set* $\mathbf{y}$. We will sometimes use use the notation $\Pr\{\cdot|\mathbf{y}\}$ instead of $\Pr\{\cdot|\mathcal{Y}_{/\mathbf{y}}\}$ to denote the conditioning with respect to the orbit associated with data set $\mathbf{y}$ even though the two notations are not necessarily equivalent. Thus, $\mathcal{Y}^n_{/\mathbf{y}}$ plays the role of *reference set for the conditional inference* (see Lehmann and Romano, 2005). In this way, in the null hypothesis and assuming exchangeability, the conditional probability distribution of a generic point $\mathbf{y}' \in \mathcal{Y}^n_{/\mathbf{y}}$, for any underlying population distribution $P \in \mathcal{P}$, is

$$\Pr\{\mathbf{y}^* = \mathbf{y}'|\mathcal{Y}^n_{/\mathbf{y}}\} = \frac{\sum_{\mathbf{y}^*=\mathbf{y}'} f_P^{(n)}(\mathbf{y}^*)\cdot d\xi^n}{\sum_{\mathbf{y}^*\in\mathcal{Y}^n_{/\mathbf{y}}} f_P^{(n)}(\mathbf{y}^*)\cdot d\xi^n} = \frac{\#[\mathbf{y}^* = \mathbf{y}',\ \mathbf{y}^* \in \mathcal{Y}^n_{/\mathbf{y}}]}{\#[\mathbf{y}^* \in \mathcal{Y}^n_{/\mathbf{y}}]},$$

which is P-independent. Of course, if there is only one point in $\mathcal{Y}^n_{/\mathbf{y}}$ whose coordinates coincide with those of $\mathbf{y}'$, (i.e., if there are no ties in the data set), and if permutations correspond to permutations of the arguments, then this conditional probability becomes $1/n!$. Thus, $\Pr\{\mathbf{y}^* = \mathbf{y}'|\mathcal{Y}^n_{/\mathbf{y}}\}$ is uniform on $\mathcal{Y}^n_{/\mathbf{y}}$ for all $P \in \mathcal{P}$.

These statements allow permutation inferences to be invariant with respect to P in H_0. Some authors, emphasizing this invariance property of permutation distribution in H_0, prefer to give them the name of *invariant tests*. However, due to this invariance property, permutation tests are distribution-free and nonparametric.

As a consequence, in the alternative hypothesis H_1, conditional probability shows quite different behavior and in particular may depend on P. To achieve this in a simple way, let us consider, for instance, a two-sample problem where $f_{P_1}^{(n_1)}$ and $f_{P_2}^{(n_2)}$ are the densities, relative to the same dominating measure ξ, of two sampling distributions related to two populations, P_1 and P_2, that are assumed to differ at least in a set of positive probability. Suppose also that $\mathbf{y}_1$ and $\mathbf{y}_2$ are the two separate and independent data sets with sample sizes n_1 and n_2, respectively. Therefore, as the likelihood associated with the pooled data set is $f_P^{(n)}(\mathbf{y}) = f_{P_1}^{(n_1)}(\mathbf{y}_1)\cdot f_{P_2}^{(n_2)}(\mathbf{y}_2)$, from the sufficiency principle it follows that the data set partitioned into two groups, $(\mathbf{y}_1;\mathbf{y}_2)$, is now the set of sufficient statistics. Indeed, by joint invariance of the likelihood ratio with respect to both f_{P_1} and f_{P_2}, the coset of $\mathbf{y}$ is $(\mathcal{Y}^{n_1}_{/\mathbf{y}_1}, \mathcal{Y}^{n_2}_{/\mathbf{y}_2})$, where $\mathcal{Y}^{n_1}_{/\mathbf{y}_1}$ and $\mathcal{Y}^{n_2}_{/\mathbf{y}_2}$ are partial orbits associated with $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. This implies

that, conditionally, no datum from $\mathbf{y}_1$ can be exchanged with any other from $\mathbf{y}_2$ because in H_1 permutations are permitted only within groups, separately.

Consequently, when we are able to find statistics that are sensitive to the diversity of two distributions, we may have a procedure for constructing permutation tests. Of course, when constructing permutation tests, one should also take into consideration the physical meaning of treatment effects, so that the resulting inferential conclusions have clear interpretations.

Although the concept of conditioning for permutation tests is properly related to the formal conditioning with respect to orbit $\mathcal{Y}^n_{/\mathbf{y}}$, henceforth we shall generally adopt a simplified expression for this concept by stating that *permutation tests are inferential procedures that are conditional with respect to the observed data set* $\mathbf{y}$. Indeed, once $\mathbf{y}$ is known and the exchangeability condition is assumed in H_0, $\mathcal{Y}^n_{/\mathbf{y}}$ remains completely determined by $\mathbf{y}$.

1.2 Basic Construction

In this section, we provide examples on the construction of a permutation test. We will do this by considering a two-sample design. Let $\mathbf{y}_1$ and $\mathbf{y}_2$ be two independent samples of size n_1 and n_2 from two population distributions P_1 and P_2, respectively. In addition, let $P_1(y) = P_2(y - \delta)$. That is, the population distributions differ only in location. A common testing problem is to assess whether $P_1 \stackrel{d}{=} P_2$ or not, where the symbol $\stackrel{d}{=}$ means equality in distribution. In a location problem, there are several ways to specify the underlying model generating the observed data; for instance, let

$$\begin{array}{ll} Y_{i1} = \mu_1 + \varepsilon_{i1}, & i = 1, \ldots, n_1, \\ Y_{j2} = \mu_1 + \delta + \varepsilon_{j2}, & j = 1, \ldots, n_2, \end{array} \tag{1.1}$$

be the models describing a generic observation from the first and second samples, respectively. Here $\delta = \mu_2 - \mu_1$, μ_1 and μ_2 are population constants, and ε_{i1} and ε_{j2} are identically distributed random variables with zero mean and variance $\sigma^2 < +\infty$ (the so-called experimental errors), not necessarily independent within the observations. The null hypothesis $P_1 \stackrel{d}{=} P_2$ can be written in terms of $\delta = 0$ against the alternative hypothesis $\delta \neq 0$. If H_0 is true, then Y_{i1} and Y_{j2} are identically distributued random variables. In addition, if ε_{i1} and ε_{j2} are exchangeable random variables, in the sense that $\Pr(\boldsymbol{\varepsilon}) = \Pr(\boldsymbol{\varepsilon}^*)$, where $\boldsymbol{\varepsilon} = [\varepsilon_{11}, \varepsilon_{21}, \ldots, \varepsilon_{n_1 1}, \varepsilon_{12}, \varepsilon_{22}, \ldots, \varepsilon_{n_2 2}]'$ and $\boldsymbol{\varepsilon}^*$ is a permutation of $\boldsymbol{\varepsilon}$, then also Y_{i1} and Y_{j2} are exchangeable in the sense that $\Pr(\mathbf{Y}) = \Pr(\mathbf{Y}^*)$, where $\mathbf{Y} = [\mathbf{Y_1}, \mathbf{Y_2}]'$ and $\mathbf{Y}^*$ is the corresponding permutation of $\mathbf{Y}$. As a simple example, consider the common case where the observations are independent. The likelihood can be written as

$$\begin{aligned} L(\delta; \mathbf{y}) &= f_{P_1}^{(n_1)}(y_{11}, y_{21}, \ldots, y_{n_1 1}) f_{P_2}^{(n_2)}(y_{12}, y_{22}, \ldots, y_{n_2 2}) \\ &= \prod_{i=1}^{n_1} f_{P_1}(y_{i1}) \prod_{j=1}^{n_2} f_{P_2}(y_{i2}), \end{aligned}$$

where $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]'$ and $f_{P_j}(y)$ is the density of Y_j, $j = 1, 2$. If H_0 is true, $f_{P_1}(y) = f_{P_2}(y)$, so $L_{H_0}(\delta; \mathbf{y}) = L_{H_0}(\delta; \mathbf{y}^*)$. Roughly speaking, this means that (conditionally), under H_0, $\mathbf{y}_1$ and $\mathbf{y}_2$ are two independent samples from the same population distribution P, or equivalently that $\mathbf{y}$ is a random sample of size $n = n_1 + n_2$ from P.

In order to obtain a statistical test, we need to define a proper test statistic and obtain its null distribution. How do we find the "best" test statistic for a given inferential problem? There is no specific answer to this question when the population distributions are unknown. One reasonable criterion is, for instance, to let the unconditional expectation of a chosen test statistic depend only on the parameter of interest. For instance, since unconditionally $\mathrm{E}[\bar{y}_1] = \mu_1$ and $\mathrm{E}[\bar{y}_2] = \mu_2$, a suitable test statistic could be defined as $T(\mathbf{y}) = \bar{y}_1 - \bar{y}_2$. Another reasonable choice is to look at the parametric counter-part: In a two-sample location problem, the well-known t statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2\right]^{\frac{1}{2}}}$$

can also be considered. We will see that the t statistic and $T(\mathbf{y}) = \bar{y}_1 - \bar{y}_2$ are equivalent within a permutation framework. By equivalent test statistics we mean test statistics that lead to the same rejection region in the permutation sample space $\mathcal{Y}^n/_{\mathbf{y}}$, so they also lead to the same inference for any given set $\mathbf{y} \in \mathcal{Y}$.

Within a permutation framework, a test statistic $T : \mathcal{Y}/_{\mathbf{y}} \to \mathcal{T}$ is a real function of all the observed data that takes values on the support $\mathcal{T} = T(\mathcal{Y}/_{\mathbf{y}}) \subseteq I\!R^1$. It is worth noting that the support $\mathcal{T}$ depends on $\mathbf{y}$ in the sense that whenever $\mathbf{y} \neq \mathbf{y}'$ we may have $\mathcal{T}_{\mathbf{y}} \neq \mathcal{T}_{\mathbf{y}'}$. Moreover, if T is such that $T(\mathbf{y}^{*\prime}) \neq T(\mathbf{y}^{*\prime\prime})$ for any two distinct points of $\mathcal{Y}/_{\mathbf{y}}$, in the null hypothesis the distribution of T over $\mathcal{T}$ is uniform; that, is all points are equally likely.

The null distribution of $T(\mathbf{y})$ is given by the elements of the space $\mathcal{T}$. We will use the notation $T(\mathbf{y})$, T^o, or simply T to emphasize the observed value of the test statistic (the one obtained from the observed data), whereas T^* indicates a value of the permutation distribution of the test statistic. Note that $T^* = T^o$ if the identity permutation is applied to $\mathbf{y}$.

To perform a statistical test, we only need to define a distance function on $\mathcal{T}$ in order to specify which elements of $\mathcal{Y}/_{\mathbf{y}}$ are "far" from H_0. That is, we need a rule to determine the critical region of the test. To this end, let us explore the space $\mathcal{T}$ through the two-sample location problem example. Let $T(\mathbf{y}) = \bar{y}_1 - \bar{y}_2$ be the test statistic and $T^* = \bar{y}_1^* - \bar{y}_2^*$ be the generic element of $\mathcal{T}$. Conditionally, the expectation and variance of observations in $\mathbf{y_1}$ and $\mathbf{y_2}$ are, respectively

$$\mathrm{E}(y_{i1}|\mathbf{y}) = \bar{y}_1, \qquad \mathrm{E}(y_{i1}^2|\mathbf{y}) = \frac{1}{n_1}\sum_i y_{i1}^2,$$
$$\mathrm{E}(y_{j2}|\mathbf{y}) = \bar{y}_2, \qquad \mathrm{E}(y_{j2}^2|\mathbf{y}) = \frac{1}{n_2}\sum_j y_{y2}^2.$$

Now let y_{i1}^* be a generic observation in $\mathbf{y_1^*}$. Conditionally, $\Pr[y_{i1}^* \in \mathbf{y_1}|\mathbf{y}] = n_1/n$ and $\Pr[y_{i1}^* \in \mathbf{y_2}|\mathbf{y}] = n_2/n$. The conditional expected value of y_{i1}^* is therefore

$$\mathrm{E}[y_{i1}^*|\mathbf{y}] = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2 = \bar{y}.$$

Similarly, $\mathrm{E}[y_{j2}^*|\mathbf{y}] = \bar{y}$. Consequently, $\mathrm{E}[T^*|\mathbf{y}] = 0$, and therefore the null distribution of T^* is centered, although it is not necessarily symmetric, in the sense that $F_{T^*}(t^*) = 1 - F_{T^*}(-t^*)$, $t^* \in \mathfrak{T}$. It is symmetric, for instance, in the balanced case where $n_1 = n_2$. As regards the variance

$$\mathrm{Var}(y_{i1}^*|\mathbf{y}) = \mathrm{E}[{y_{i1}^*}^2|\mathbf{y}] - \mathrm{E}[y_{i1}^*|\mathbf{y}]^2 = \frac{1}{n}\sum_{l=1}^{2}\sum_{k=1}^{n_l} y_{kl}^2 - \bar{y}^2 = \hat{\sigma}_0^2,$$

where $\hat{\sigma}_0^2$ is the maximum likelihood estimate of the variance under H_0 when data are normally distributed. Note that $\hat{\sigma}_0^2$ is constant, in a conditional framework. Note also that the y_{i1}^*'s are not independent. By the finite population theory,

$$\mathrm{Var}(\bar{y}_1^*|\mathbf{y}) = \frac{\hat{\sigma}_0^2}{n_1}\left(\frac{n-n_1}{n-1}\right) = \frac{\hat{\sigma}_0^2}{n-1}\frac{n_2}{n_1}.$$

Now consider the relationship $n_1\bar{y}_1^* + n_2\bar{y}_2^* = Y$, where Y is the total of observations, which is permutationally invariant. Then

$$\begin{aligned}\mathrm{Var}(T^*|\mathbf{y}) &= \mathrm{Var}\left(\bar{y}_1^* - \frac{Y}{n_2} + \frac{n_1\bar{y}_1^*}{n_2}|\mathbf{y}\right) = \mathrm{Var}\left(\frac{n}{n_2}\bar{y}_1^*|\mathbf{y}\right) = \frac{n^2}{n_2^2}\mathrm{Var}(\bar{y}_1^*|\mathbf{y}) \\ &= \frac{n\hat{\sigma}_0^2}{n-1}\frac{n}{n_1 n_2} = \frac{n\hat{\sigma}_0^2}{n-1}\left(\frac{n_1+n_2}{n_1 n_2}\right) = s_0^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right),\end{aligned}$$

which is like the denominator of the t test, despite the estimate of the population variance. Note that s_0^2 is the unbiased estimate of $\mathrm{Var}(Y)$ when H_0 is true. Therefore, we can define a test statistic as

$$T^* = \frac{n_1 n_2}{n}\frac{(\bar{y}_1^* - \bar{y}_2^*)^2}{s_0^2}, \tag{1.2}$$

where the emphasis is on the fact that T^* is a random variable defined on $\mathcal{Y}/_{\mathbf{y}}$. Large values of (1.2) are significant against the null hypothesis. Since n_1, n_2 and s_0^2 are constant, (1.2) is permutationally equivalent to $T^{*\prime} = (\bar{y}_1^* - \bar{y}_2^*)^2$ and to $T^{*\prime\prime} = |\bar{y}_1^* - \bar{y}_2^*|$.

A similar proof applies to the classic t statistic: Let t^{*2} be the (squared) value of the t statistic obtained from a random permutation of $\mathbf{y}^*$,

$$t^{*2} = \frac{n_1 n_2}{n} \frac{(\bar{y}_1^* - \bar{y}_2^*)^2}{s^{*2}}.$$

It can be easily proved (see Section 5.2) that this is a special case of one-way ANOVA framework (when $C = 2$). Therefore, t^{*2} is a monotone nondecreasing function of $T^{*\prime}$, and since permutation tests are based on the ordered values of $\mathcal{T}$ (see Section 1.3), t^{*2} is permutationally equivalent to T^* as well.

The exact p-value of the test is

$$p = \frac{1}{C} \sum_{T^* \in \mathcal{T}} I(T^* \geq T^o) = \frac{\#[T^* \geq T^o]}{C},$$

where $T^o = T(\mathbf{y})$, $I(\cdot)$ is the indicator function, and C is the cardinality of $\mathcal{T}$. If Y is a continuous random variable (i.e., the probability of having ties is zero), then

$$C = \binom{n}{n_1}.$$

Clearly C increases very rapidly with n, so in practice the c.d.f. of T^* is approximated by a Monte Carlo sampling from $\mathcal{T}$. Let B be the number of Monte Carlo permutations. Then the c.d.f. of T^* is estimated by

$$\hat{F}_{T^*}(t) = \frac{\#[T^* \leq t]}{B} \qquad t \in I\!R.$$

1.3 Properties

In this section, we investigate some properties of the permutation tests, such as exactness and unbiasedness; for consistency we refer to Hoeffding (1952). Let Y be a random variable such that $\mathrm{E}[Y] = \mu$ and $\mathrm{Var}[Y]$ exists. Let $H_0 : \mu \leq \mu_0$ be the null hypothesis to be assessed and $T = T(\mathbf{Y})$ a suitable test statistic for H_0 (in the sense that large values of T are significant against H_0). Then, a (nonrandomized) test ϕ of size α is a function of the test statistic $T = T(Y)$ such as

$$\phi(T) = \begin{cases} 1 & \text{if} \quad T \geq T^{1-\alpha} \\ 0 & \text{if} \quad T < T^{1-\alpha}, \end{cases}$$

where $T^{1-\alpha}$ is the $1-\alpha$ quantile of the null distribution of T, i.e. $\Pr[T \geq T^{1-\alpha} | \mathcal{Y}_{/\mathbf{y}}] = \alpha$. The α-values that satisfy $\Pr[T \geq T^{1-\alpha} | \mathcal{Y}_{/\mathbf{y}}] = \alpha$ are called attainable α-values. The set of attainable α-values is a proper subset of $(0, 1]$. Thus, if $H_0 : \mu = \mu_0$, permutation tests are exact for all attainable α-values, whereas if $H_0 : \mu \leq \mu_0$, they are conservative.

If the distribution of T is symmetric, one can define a test for two-sided alternatives by replacing T with $|T|$ in the definition of ϕ, or T^α with $T^{1-\alpha}$ if the alternative hypothesis is $H_1 : \mu < \mu_0$. Clearly, the expected value of ϕ is

$$\mathrm{E}[\phi] = 1 \cdot \Pr[T \geq T^{\alpha}] + 0 \cdot \Pr[T < T^{\alpha}] = \alpha.$$

That is why ϕ is usually called a test of size α.

Permutation tests are conditional procedures; therefore the definitions of the usual properties of *exactness* and *unbiasedness*, and *consistency* require an ad hoc notation: From now on, we denote by $\mathbf{y}(\delta)$ the set of data when the alternative hypothesis is true and by $\mathbf{y}(\mathbf{0})$ the set of data when the null hypothesis is true.

The test ϕ of size α is said to be exact if $\forall\ 0 < \alpha < 1$:

$$\Pr[\phi = 1 | \mathbf{y}(\mathbf{0})] = \alpha.$$

The test ϕ is said to be unbiased if

$$\Pr[\phi = 1 | \mathbf{y}(\mathbf{0})] \leq \alpha \leq \Pr[\phi = 1 | \mathbf{y}(\delta)].$$

The test ϕ is said to be consistent if

$$\lim_{n \to +\infty} \Pr[\phi = 1 | \mathbf{y}(\delta)] = 1.$$

To prove the properties of permutation tests, we will still refer to a univariate two-sample problem. In the previous section, we have given an informal definition of a permutation test.

Formally, let $\mathcal{Y}^n/\mathbf{y}$ be the orbit associated with the observed vector of data $\mathbf{y}$. The points of $\mathcal{Y}^n/\mathbf{y}$ can also be defined as $\mathbf{y}^* : \mathbf{y}^* = \pi\mathbf{y}$ where π is a random permutation of indexes $1, 2, \ldots, n$. Define a suitable test statistic T on $\mathcal{Y}^n/\mathbf{y}$ for which large values are significant for a right-handed one-sided alternative: The image of $\mathcal{Y}^n/\mathbf{y}$ through T is the set $\mathcal{T}$ that consists of C elements (if there are no ties in the given data). Let

$$T^*_{(1)} \leq T^*_{(2)} \leq \cdots \leq T^*_{(C)}$$

be the ordered values of $\mathcal{T}$. Let T^o be the observed value of the test statistic, $T^o = T(\mathbf{y})$. For a chosen attainable significance level $\alpha \in \{1/C, 2/C, \ldots, (C-1)/C\}$, let $k = C(1-\alpha)$. Define a permutation test for a one-sided alternative the function $\phi^* = \phi(T^*)$

$$\phi^*(T) = \begin{cases} 1 & \text{if} \quad T^o \geq T^*_{(k)} \\ 0 & \text{if} \quad T^o < T^*_{(k)} \end{cases}.$$

Since the critical values of the distribution of T^* depend on the observed data, one can provide a more general definition of a permutation test based on the p-values, whose distribution depends on sample size n:

$$\phi^*(T) = \begin{cases} 1 & \text{if} \quad \Pr[T^* \geq T^o | \mathbf{y}] \leq \alpha \\ 0 & \text{if} \quad \Pr[T^* \geq T^o | \mathbf{y}] > \alpha \end{cases}.$$

The equivalence of the two definitions is ensured by the relationship

$$\begin{aligned}\Pr\{\Pr[T^* \geq T^o|\mathbf{y}] \leq \alpha\} &= \Pr\{\Pr[T^* \leq T^o|\mathbf{y}] \geq 1-\alpha\} \\ &= \Pr\{F_{T^*}(T^o) \geq 1-\alpha\} \\ &= \Pr\{F_{T^*}^{-1}(F_{T^*}(T^o)) \geq F_{T^*}^{-1}(1-\alpha)\} \\ &= \Pr\{T^o \geq T^*_{(k)}\}.\end{aligned}$$

To prove exactness, suppose H_0 is true. Then the elements of $\mathcal{T}$ are equally likely under the null hypothesis. This means that

$$\Pr\{T^* = T^o|\mathbf{y}(\mathbf{0})\} = \frac{1}{C} \quad \Rightarrow \quad \Pr\{T^o \in \mathcal{A}|\mathbf{y}(\mathbf{0})\} = \frac{\#[T^* \in \mathcal{A}]}{C},$$

where $\mathcal{A}$ is one element of the Borel set defined on $\mathcal{T}$. Hence, for any attainable significance level α

$$\begin{aligned}\Pr\{\phi^*(T) = 1|\mathbf{y}(\mathbf{0})\} &= \Pr\{T^o \geq T^*_{(k)}|\mathbf{y}(\mathbf{0})\} \\ &= \frac{\#[T^* \geq T^*_{(k)}]}{C} = \frac{C\alpha}{C} = \alpha.\end{aligned}$$

Note that, since permutation tests are conditionally exact, they are unconditionally exact as well.

As regards unbiasedness, we will refer to the two-sample problem of the previous section. Let's suppose that data of the two samples are generated under the model (1.1), and let $H_0 : \mu_2 - \mu_1 \leq 0$ be the null hypothesis to assess. Define the test statistic as $T = \bar{y}_2 - \bar{y}_1$. Let $T^o(0)$ and $T^o(\delta)$ be respectively the observed value of T when data are $\mathbf{y}(\mathbf{0})$ and $\mathbf{y}(\delta)$, respectively,

$$\begin{aligned}T^o(0) = T^*(\mathbf{y}(\mathbf{0})) &: \bar{y}_2 - \bar{y}_1 = \bar{\varepsilon}_2 - \bar{\varepsilon}_1, \\ T^o(\delta) = T^*(\mathbf{y}(\delta)) &: \bar{y}_2 - \bar{y}_1 = \delta + \bar{\varepsilon}_2 - \bar{\varepsilon}_1,\end{aligned}$$

where $\bar{\varepsilon}_2$ and $\bar{\varepsilon}_1$ are sampling averages of n_2 and n_1 exchangeable errors, respectively. Since the event $\Pr[T^* \geq T^o|\mathcal{Y}/_{\mathbf{y}}] \leq \alpha$ implies the event $\{T^o \geq T^*_{(k)}|\mathcal{Y}/_{\mathbf{y}}\}$, we may write

$$\Pr[T^o(0) \geq T^*_{(k)}|\mathbf{y}(\mathbf{0})] = \Pr[\bar{\varepsilon}_2 - \bar{\varepsilon}_1 \geq T^*_{(k)}]$$

and

$$\Pr[T^o(\delta) \geq T^*_{(k)}|\mathbf{y}(\delta)] = \Pr[\bar{\varepsilon}_2 - \bar{\varepsilon}_1 \geq T^*_{(k)} - \delta].$$

Now, without loss of generality, let $\delta \geq 0$ and $T^*_{(k)} \geq 0$. Then, from the exactness of ϕ^*, we have:

$$\Pr[\phi^* = 1|\mathbf{y}(\delta)] = \Pr[T^o \geq T^*_{(k)}|\mathbf{y}(\delta)] \geq \Pr[T^o \geq T^*_{(k)}|\mathbf{y}(\mathbf{0})] = \Pr[\phi^* = 1|\mathbf{y}(\mathbf{0})],$$

which proves unbiasedness.

1.4 Multivariate Permutation Tests

There are some problems where the complexity requires a further approach. Consider, for instance, a multivariate problem where q (possibly dependent) variables are considered, or a multiaspect problem (such as the Beherens-Fisher problem), or a stratified analysis. The difficulties arise because of the underlying dependence structure among variables (or aspects), which is generally unknown. Moreover, a global answer involving several dependent variables (aspects) is often required, so the question is how to combine the information related to the q variables (aspects) into one global test.

Let us consider a one-sample multivariate problem with q dependent variables: Here the data set $\mathbf{Y}$ is an $n \times q$ matrix, where n is the sample size. What we are generally interested in is to test the null hypothsis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$ against the alternative hypothesis $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$, where $\boldsymbol{\mu}$ is a $q \times 1$ vector of population means and $\boldsymbol{\mu_0} = [\mu_{01}, \mu_{02}, \ldots, \mu_{0q}]$ is a target vector. Assuming $\mathbf{Y}_i$ $i = 1, \ldots, n$ is a multivariate normal random variable, a parametric solution is Hotelling's T^2 test. In a bivariate problem, we may specify it as

$$\begin{aligned} T^2 &= n[\bar{\mathbf{y}} - \boldsymbol{\mu_0}]' \boldsymbol{\Sigma}^{-1} [\bar{\mathbf{y}} - \boldsymbol{\mu_0}] \\ &= \frac{n[\bar{y}_1 - \mu_1]^2}{s_1^2(1-\hat{\rho}_{12}^2)} + \frac{n[\bar{y}_2 - \mu_2]^2}{s_2^2(1-\hat{\rho}_{12}^2)} - 2\frac{n\hat{\rho}_{12}[\bar{y}_1 - \mu_1][\bar{y}_2 - \mu_2]}{s_1^2 s_2^2(1-\hat{\rho}_{12}^2)} \\ &= T(\mathbf{x_1}, \mu_1 | \hat{\rho}_{12}) + T(\mathbf{x_2}, \mu_2 | \hat{\rho}_{12}) - 2T'(\mathbf{x_1}, \mathbf{x_2}, \mu_1, \mu_2), \end{aligned}$$

where $T(\cdot)$ and $T'(\cdot)$ are test statistics, $\hat{\rho}_{12}$ is the estimate of the correlation between Y_1 and Y_2, and s_1^2 and s_2^2 are unbiased estimates of population variances. Note that Hotelling's T^2 is a combination of marginal tests on μ_1 and μ_2 accounting for the dependence between Y_1 and Y_2. Hotelling's T^2 depends on the estimated variance-covariance matrix $\boldsymbol{\Sigma}$, which has rank $n - q$, and it is appropriate only for two-sided alternatives. This means that either when $n \leq q$ or alternatives are one-sided, the Hotelling T^2 test cannot be applied. If Y_1 and Y_2 are independent, Hotelling's T^2 reduces to

$$T^2 = \frac{n[\bar{y}_1 - \mu_1]^2}{s_1^2} + \frac{n[\bar{y}_2 - \mu_2]^2}{s_2^2} = T(\mathbf{y_1}, \mu_1 | \rho_{12} = 0) + T(\mathbf{y_2}, \mu_2 | \rho_{12} = 0).$$

Within a conditional approach, there are no assumptions on the dependence structure among the q variables. Let us consider the matrix of observations partitioned into n q-dimensional arrays; that is,

$$\mathbf{Y_{n \times q}} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}.$$

Each row of $\mathbf{Y}$ is a determination of the multivariate variable $[Y_1, Y_2, \ldots, Y_q]$, which has distribution P with unknown dependence structure. But, being

determinations of the same random variable, the rows of $\mathbf{Y}$ (i.e., the data related to the statistical units) have an intrinsic dependence structure, which does not need to be modelled in order to do a permutation test if the permutations involve the rows of $\mathbf{Y}$. Note that this is true even if the vectors of the observations are repeated measures, or functions of the same data (e.g., the first r powers of a random variable Y).

A suitable nonparametric test to assess the hypothesis on marginal distributions $H_{0j} : \mu_j = \mu_{0j}, j = 1, \ldots, q$, is McNemar's test,

$$S_j = \sum_{i=1}^{n} I(y_{ij} - \mu_{0j} > 0),$$

where $I(\cdot)$ is the indicator function. If data in $\mathbf{Y_j}$ are symmetric and H_{0j} is true, then μ_{0j} represents the mean and the median of the distribution. Therefore, if μ_{0j} is true, S_j should be close to $n/2$. The null distribution of S_j is binomial with parameters n and $1/2$. Clearly, S_j is significant for small and large values, and the p-value of the test is obtained as

$$p_j = \Pr[X \leq (n - S_j)] + \Pr[X \geq S_j] \qquad \text{where} \qquad X \sim \text{Bi}(n, 1/2).$$

An equivalent version of McNemar's test is the test statistic

$$T^*(\mathbf{y_j}, \mu_{0j}) = \sum_{i=1}^{n} (y_{ij} - \mu_{0j}) \text{sgn}^*(y_{ij} - \mu_{0j}), \tag{1.3}$$

where

$$\Pr[\text{sgn}^*(y_i - \mu_0) = z] = \begin{cases} 1/2 & \text{if } z = +1 \\ 1/2 & \text{if } z = -1 \end{cases}.$$

Note that in one-sample location problems, the usual permutations do not apply since what is really informative here on the location parameter is the vector of observed signs $\mathbf{S_j} = [I(y_{1j} - \mu_{0j} > 0), I(y_{2j} - \mu_{0j} > 0), \ldots, I(y_{nj} - \mu_{0j} > 0)]$. According to McNemar's test, two points $\mathbf{y_j^*}$ and $\mathbf{y_j'}$ have the same likelihood if $\sum_{i=1}^{n} I(y_{ij}^* - \mu_{0j} > 0) = \sum_{i=1}^{n} I(y_i' - \mu_{0j} > 0)$. Here, the permutation sample space $\mathcal{Y}^{(n)}/_{\mathbf{y_j}}$ is given by

$$\mathcal{Y}^{(n)}/_{\mathbf{y_j}} = \{\mathbf{y_j^*} : \mathbf{y_j^*} = \pi^{\pm}(\mathbf{y_j} - \boldsymbol{\mu_{0j}})\},$$

where $\pi^{\pm}$ is a combination of $n \pm$ signs, $\boldsymbol{\mu_{0j}} = \mu_{0j}\mathbf{1_n}$ and $\mathbf{1_n}$ is an $n \times 1$ vector of 1's. The permutation sample space therefore has 2^n points. Note that in (1.3) we have

$$\begin{aligned} \text{E}[T^*(\mathbf{y_j}, \mu_{0j})|\mathbf{y_j}] &= 0, \\ \text{Var}[T^*(\mathbf{y_j}, \mu_{0j})|\mathbf{y_j}] &= \sum_{i=1}^{n} (y_{ij} - \mu_{0j})^2, \end{aligned}$$

so the null distribution is always centered on μ_{0j}.

Since we have the relationship

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \qquad \Longrightarrow \qquad \bigcap_{j=1}^{q} H_{0j},$$

the *global* null hypothesis H_0 can be viewed as an intersection of *partial* null hypotheses H_{0j}. Let λ_j, $j = 1, \ldots, q$, be a *partial* test statistic for the univariate hypothesis H_{0j}. By partial test we mean a test statistic to assess $H_{0j} : \mu_j = \mu_{0j}$ $j = 1, \ldots, q$. For instance, one may consider $\lambda_j = |T_j^*(\mathbf{y_j}, \mu_{0j})|$ or $\lambda_j = T_j^*(\mathbf{y_j}, \mu_{0j})^2$, which is significant for large values against $H_{0j} : \mu_j = \mu_{0j}$. The partial test statistics may also be significant for one-sided alternatives. For instance if $H_{1j} : \mu_j < \mu_{0j}$, then a test statistic is $\lambda_j = -T_j^*(\mathbf{y_j}, \mu_{0j})$. Now let

$$\psi^* = \psi(\mathbf{Y}^*, \boldsymbol{\mu_0}) = \sum_{j=1}^{q} \lambda_j \tag{1.4}$$

be a *global test statistic*. In order to account for the (possible) dependence among the q variables, the domain of ψ^* is

$$\mathcal{Y}^{(n)}/_{\mathbf{Y}^*} = \left\{ \mathbf{Y}^* : \mathbf{Y}^* = [\pi^{\pm}(\mathbf{y}_1 - \boldsymbol{\mu_{01}}), \pi^{\pm}(\mathbf{y}_2 - \boldsymbol{\mu_{02}}), \ldots, \pi^{\pm}(\mathbf{y}_q - \boldsymbol{\mu_{0q}})] \right\},$$

where $\pi^{\pm}$ is the same combination of n $\pm$ signs applied to all q vectors. If the q variables are independent, one may consider

$$\mathcal{Y}^{(n)}/_{\mathbf{Y}_\perp} = \left\{ \mathbf{Y}_\perp^* : \mathbf{Y}_\perp^* = [\pi_1^{\pm}(\mathbf{y}_1 - \boldsymbol{\mu_{01}}), \pi_2^{\pm}(\mathbf{y}_2 - \boldsymbol{\mu_{02}}), \ldots, \pi_q^{\pm}(\mathbf{y}_q - \boldsymbol{\mu_{0q}})] \right\}.$$

where the $\pi_i^{\pm}$'s are q independent combinations of n $\pm$ signs.

Note that $\mathcal{Y}^{(n)}/_{\mathbf{Y}_\perp}$ and $\mathcal{Y}^{(n)}/_{\mathbf{Y}^*}$ are different spaces. In particular, $\mathcal{Y}^{(n)}/_{\mathbf{Y}^*} \subseteq \mathcal{Y}^{(n)}/_{\mathbf{Y}_\perp}$, where $\mathcal{Y}^{(n)}/_{\mathbf{Y}_\perp}$ is the orbit associated to $\mathbf{Y}$ if the q variables are assumed to be independent, whereas in $\mathcal{Y}^{(n)}/_{\mathbf{Y}^*}$ the inner dependence among variables is maintained. The cardinality of $\mathcal{Y}^{(n)}/_{\mathbf{Y}_\perp}$ is 2^{nq}, whereas the cardinality of $\mathcal{Y}^{(n)}/_{\mathbf{Y}^*}$ is 2^n since the same combinations of signs apply to all q vectors.

If (1.5) is computed on $\mathcal{Y}^{(n)}/_{\mathbf{Y}^*}$, then $T^*(\mathbf{y_j}, \mu_{0j}) = T^*(\mathbf{y_j}, \mu_{0j}|\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the matrix of (true) variances and covariances among q variables. That is, since the test statistic is defined on a permutation sample space accounting for dependence, the partial test statistic T_j^*'s also account for dependence. If $q = 2$, then let

$$H_0^G = \begin{cases} H_{01} : \mu_1 \leq 0 \\ H_{02} : \mu_2 \leq 0 \end{cases}$$

be the global null hypothesis, which is true if the partial null hypotheses H_{01} and H_{02} are jointly true and which should be rejected whenever one of the partial null hypotheses is rejected. Define a global test to assess H_0^G as

$$\psi^* = \psi(\mathbf{Y}^*, \mathbf{0}) = T_1^*(\mathbf{y_1}, 0) + T_2^*(\mathbf{y_2}, 0), \tag{1.5}$$