Statistica & Applicazioni - Vol. VII, n. 1, 2009, pp. 87-103

NONPARAMETRIC DIRECTIONAL TESTS IN THE PRESENCE OF CONFOUNDING FACTORS AND CATEGORICAL DATA

Rosa Arboretti Giancristofaro*
Stefano Bonnini**

SUMMARY

In modern socio-economic systems, often the aim of a performance analysis or quality evaluation is to compare different products, different manufacturing plants or service centres, different actions or distinct treatments. The question is, ''Which is better?'' This is complicated because the considered aspects are often measured through categorical data and the results can be affected by confounding factors. To solve this problem we discuss some directional permutation tests based on the nonparametric combination of dependent permutation tests (NPC) for two-sample comparisons in the presence of ordinal categorical variables and confounding factors. In particular we present a new permutation test based on the combination of a finite number of sample moments. To reduce the confounding effects we consider the joint application of stratification and the NPC method. We also show the results of Monte Carlo simulations in order to compare permutation solutions with other nonparametric tests and to evaluate the robustness of the test based on moments.

Keywords: *confounding factors, ordered categorical variable, one-sided test, permutation test, sampling moment.*

1. INTRODUCTION

The two-sample test for categorical variables is one of the oldest and most interesting problems in the field of statistical hypothesis testing. Moreover, this kind of test for univariate ordered categorical variables is very frequently encountered in practical problems. In this paper we take stochastic dominance alternatives in two-sample testing problems into consideration. Let us assume that data are classified according to two levels of a symbolic treatment and that level j indicates group (sample) $j\,(j = 1, 2)$. X_1 and X_2 indicate the response variable for group 1 and group 2 respectively. In the alternative hypothesis the expected treatment effect is to decrease X_2 with respect to X_1 towards smaller categorical values. In other words, given two independent random samples $\mathbf{X}_j = \{X_j i, i = 1, ..., n_j\}, j = 1, 2$; we wish to test the hypothesisis

$$H_0 : \left\{ X_1 \stackrel{d}{=} X_2 \right\}$$

* Dipartimento Territorio e Sistemi Agro-Forestali - Università degli Studi di Padova - viale dell'Università, 16 - Agripolis 35020 LEGNARO (PD) (email: rosa.arboretti@unipd.it).

** Dipartimento di Matematica - Università degli Studi di Ferrara - via Machiavelli, 35 - 44100 FERRARA (e-mail: stefano.bonnini@unife.it).

against

$$H_1 : \left\{ X_1 \overset{d}{>} X_2 \right\}.$$

Thus we wish to test the hypothesis of equality in distribution against the hypothesis of stochastic dominance of X_1 with respect to X_2. This testing problem is rather difficult to cope with using parametric approaches, especially within the framework of likelihood ratio tests (Sampson and Whittaker, 1989; El Barmi and Dykstra, 1995; Dykstra *et al.*, 1995; Dardanoni and Forcina, 1998; Wang, 1996; Cohen *et al.*, 2000, 2003; Perlman and Wu, 2002; Silvapulle and Sen, 2005). One quite serious difficulty with the maximum likelihood ratio test is that its asymptotic null distribution depends on the true unknown parameters of the underlying multinomial distribution. It is therefore difficult to justify its use in practice and for this reason the use of a nonparametric test is suitable. By working within the nonparametric combination of dependent permutation tests (NPC) it is possible to find exact solutions to such problems (see Pesarin, 1994, 2001, 2004).

The aim of two-sample tests is to evaluate whether the difference between the samples is due to the treatment effect. However, in several situations, especially in observational studies, if samples are not comparable because of one or more confounding factors, the difference in responses can be caused by confounding effects. To reduce the confounding effects we propose the joint application of stratification and the NPC method.

Section 2 examines the problem of directional tests for ordered categorical variables in the one-dimensional two-sample case and presents some nonparametric methodological solutions; in Section 3 the NPC method is described and a new test based on the NPC approach and the comparison of sampling moments is presented; Section 4 considers the problem of confounding factors and describes an extension to the NPC method to reduce the confounding effects; Section 5 discusses the results of some Monte Carlo simulations to compare the power of NPC tests with that of other nonparametric tests and to assess the robustness of the tests on moments; in Section 6 an application example is described; Section 7 is devoted to our conclusions.

2. DIRECTIONAL TESTS FOR ORDERED CATEGORICAL VARIABLES: THREE NONPARAMETRIC SOLUTIONS

The testing problem described in the previous section, i.e. the so-called two-sample *goodness-of-fit* testing problem for univariate ordered categorical variables with stochastic dominance alternatives, can be defined in an equivalent way using the cumulative distribution functions (CDFs) of the two compared populations. Let us assume that the support of a univariate non-degenerate ordered categorical variable X is partitioned into a finite number $K = 2$ of ordered classes $(A_1, A_2, ..., A_K)$, and that the data are classified according to two levels of a treatment, giving rise to a typical two-sample design. Classes $A_i (i = 1, 2, ..., K)$ may represent either qualitative or

quantitative category, according to the nature of X. Thus, given two independent samples of respectively $n_j > 2, j = 1, 2$; independent and identically distributed (iid) observations, $\mathbf{X}_j = X_j i, i = 1, ..., n_j$ say, we wish to test:

$$H_0 : \left\{ X_1 \stackrel{d}{=} X_2 \right\} = \{ F_1(A_k) = F_2(A_k), k = 1, ..., K-1 \}$$

against

$$H_1 : \left\{ X_1 \stackrel{d}{>} X_2 \right\} = \{ F_1(A_k) \leq F_2(A_k), k = 1, ..., K-1 \}$$

where at least one inequality is strict and the function $F_j(A_k) = \Pr\{X_j \leq A_k\}$ plays the role of CDF for $X_j, j = 1, 2$. By assuming that no reverse inequality such as $F_1(A_k) > F_2(A_k), k = 1, ..., K-1$ is possible, the alternative can also be written in the form of $H_1 : \left\{ \bigcup_{k=1}^{K-1} [F_1(A_k) < F_2(A_k)] \right\}$. Observed data are generally organized in a $2 \times K$ contingency table such as $\{f_j k = S_i = nj\mathbf{I}(X_j i \in A_k), k = 1, .., K; j = 1, 2$ where $\mathbf{I}(\bullet) = 1$ if event $(\bullet)$ occurs and 0 otherwise. Symbols $N_{jk} = \sum_{s \leq k} f_{js}$ indicate the cumulative frequencies for sample j, $n_j = N_{jK}$ and $f_{\bullet k} = f_{1k} + f_{2k}$ indicate marginal frequencies.

Permutation analysis is much easier if, in place of usual contingency tables, data are unit-by-unit represented by listing the $n = n_1 + n_2$ individual records. In the $2 \times K$ design, the dataset is represented by $X = \{X(i), i = 1, ..., n; n_1, n_2\}$, where it is intended that the first n_1 records belong to the first sample and the rest to the second. Sometimes $\mathbf{X}$ is also used to denote the pooled dataset. Thus, if $\left(u_1^*, ..., u_n^*\right)$ indicates a permutation of individual units $(1, ..., n)$, then $X^* = \left\{ X\left(u_i^*\right), i = 1, ..., n; n_1, n_2 \right\}$ indicates the corresponding permutation of dataset $\mathbf{X}$. $N_{jk}^* = \sum_{s \leq k} f_{js}^*$ indicate the cumulative frequencies for the permuted contingency table, i.e. for the table corresponding to the permuted dataset X^*. It is worth observing that in univariate two-sample designs, given they contain exactly the same amount of information in relation to F, the marginal frequencies $\{n_1, n_2, f_{\bullet 1}, ..., f_{\bullet K}\}$, the pooled dataset $\mathbf{X}$ as well as any of its permutations X^* are equivalent sets of sufficient statistics under H_0. Note also that the assumed iid condition implies that in H_0 the data of the two samples are exchangeable, thus the *permutation testing principle* can be applied:

''If two experiments, taking values on the same sample space Ω^n and respectively with underlying distributions F_1 and F_2, both members of $\mathbf{F}$, give the same dataset $\mathbf{X}$, then the two inferences, conditional on $\mathbf{X}$ and obtained using the same test statistic, must be the same, provided that the exchangeability of data with respect to groups is satisfied in the null hypothesis. Consequently, if two experiments, with underlying distributions F_1 and F_2, give respectively $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ and $\boldsymbol{X_1} \neq \boldsymbol{X_2}$, then the two conditional inferences may be different.'' (Pesarin, 2001)

The model for this stochastic dominance problem may be formalized with the notation $X_1 = \varphi(Y_1) \stackrel{d}{=} \varphi(Y_2 + \Delta)$, where $Y_j (j = 1, 2)$ represent underlying real-valued responses, φ is a function which transforms Y into ordered categorical data, and Δ represents a non-negative stochastic effect.

For the given testing problem there have been some interesting recent solutions: (i) the likelihood ratio test for restricted alternatives (El Barmi and Dykstra, 1995; Dardanoni and Forcina, 1998; Silvapulle and Sen, 2005); (ii) Hirotsu's solutions based on the cumulative chi-squared statistic (Hirotsu, 1982, 1986); (iii) Troendle's method, consisting in a numerical approach by which the likelihood ratio test can be calculated for the nonparametric Behrens-Fisher problem. In the following subsections we describe three further nonparametric solutions: the well-known Wilcoxon-Mann-Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947), the Brunner and Munzel test (Brunner and Munzel, 2000), and a permutation test based on the Anderson-Darling statistic (Pesarin, 2001).

2.1 *The rank sum test with correction for ties*

One of the most powerful and widely used nonparametric tests is the rank sum test proposed by Wilcoxon in 1945 and generalized and extended by Mann and Whitney in 1947. The assumptions of the method are that: 1) the two underlying distributions are continuous; 2) their shapes, from a symmetrical point of view, are equal; 3) data are not nominal categorical. Without loss of generality let us suppose that $n_1 = n_2$. $R_j i$ indicates the rank of $X_j i (i = 1, ..., n_j; j = 1, 2)$ in the pooled dataset and, where ties are present, the mean of the ranks of tied values (mid-rank) is assigned to the corresponding observations. The test statistic is

$$W = \sum_{i=1}^{n_1} R_{1i}.$$

Small values of W lead to rejecting the null hypotheses in favour of the alternative hypotheses. The standardized version of the test statistic can be obtained dividing it by

$$\sigma_W = \sqrt{\frac{n_1 n_2}{n(n-1)} \left[\frac{n^3 - n}{12} - \sum\nolimits_s \left(\frac{t_s^3 - t_s}{12} \right) \right]}$$

where $t_s (s = 1, ..., g)$ are the corrected ranks for the g tied values.

2.2 *The Brunner and Munzel test*

Brunner and Munzel (2000) proposed a rank test for the Behrens-Fisher two-sample problem in a nonparametric model where the assumption of continuous distribution functions is relaxed. For the Brunner and Munzel rank test W_{BM}, arbitrary distribution functions are admitted including the case where ties occur by observing ordered categorical data. The test statistic proposed by Brunnel and Munzel is

$$W_{BM} = \frac{\bar{R}_{2\cdot} - \bar{R}_{1\cdot}}{\sqrt{n\hat{\sigma}^2}}$$

where $\bar{R}_{j\cdot} = \sum_{i=1}^{n_j} R_{ji}/n_j$ is the mean of the ranks R_{ji} in the jth sample and $\hat{\sigma}^2$ is a consistent estimator of variance: $\hat{\sigma}^2 = n \cdot \left(\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2\right)$, where $\hat{\sigma}_j^2 = \left[\left(\frac{1}{n_j - 1}\right)\sum_{i=1}^{n_j} \left(R_{ji} - R_{ji}^{(j)} - \bar{R}_{j\cdot} + \left(n_j + 1\right)/2\right)^2\right]/\left(n - n_j\right)^2$ and $R_{ji}^{(j)}$ denotes the (within) rank of $X_j i$ among the n_j observations within the jth sample $\mathbf{X}_j = \left\{X_{ji}, i = 1, ..., n_j\right\}$.

2.3 *The permutation test based on the Anderson-Darling statistic*

As a solution to the two-sample one-dimensional testing problem, we may consider the permutation test statistic:

$$T_{AD}(\mathbf{X}^*) = T_{AD}^* = \sum_{k=1}^{K-1} \left(\hat{F}_{2k}^* - \hat{F}_{1k}^*\right)\left[\bar{F}_{\bullet k}(1 - \bar{F}_{\bullet k})\frac{4n_1}{n_2(n-1)}\right]^{-1/2}$$

where $\hat{F}_{jk}^* = N_{jk}^* \Big/ n_j, j = 1, 2;$ are the empirical distribution functions (EDFs) corresponding to the permuted contingency table and $\bar{F}_{\bullet k} = N_{\bullet k}/n$ are the marginal EDFs. Note that large values for T_{AD} are significant. Statistic T_{AD} essentially compares two EDFs and corresponds to the discrete version of a statistic following Cramér-von Mises' two-sample goodness-of-fit test statistic for stochastic dominance alternatives, adjusted according to Anderson-Darling. The p-value is defined as $\lambda_{AD} = \Pr\left\{ T_{AD}^* \geq T_{AD}^o | \mathbf{X}\right\}$, where $T_{AD}^o = T_{AD}(\mathbf{X})$ represents the observed value of T_{AD}. Thus, according to the general testing rule, if $\lambda_{AD} \leq \alpha$, the null hypothesis is rejected at significance level $\alpha > 0$. The exact determination of permutation distribution of any statistic can clearly be obtained by complete enumeration of all its permutation values. Of course, this way becomes unsuitable when sample sizes are not very small and when problems are complex. Alternatively, the permutation distribution can be estimated to the desired degree of accuracy by a Conditional Monte Carlo Method (CMCM) consisting of a simple random sampling from the set of all permutations. This solution is especially recommended for NPC methods and in general for complex problems (see Pesarin, 2001).

3. SCORE TRANSFORMATION AND TEST ON MOMENTS

Testing analysis using NPC methods requires that problems be broken down into a set of simpler sub-problems for each of which a permutation partial test is available and that these partial tests can be jointly processed. From a formal point of view, the null hypothesis $H_0 : \bigcap_k H_{0k}$ is true if every null sub-hypothesis H_{0k} is true; the alternative hypothesis $H_1 : \bigcup_k H_{1k}$ is true if at least one alternative sub-hypothesis H_{1k} is true. Hence, in order to obtain an overall solution, one way is to properly combine all related partial results. The partial tests and associated p-values are dependent in a way that in general is extremely difficult to take explicitly into account. Consequently,

when considering their combination, we shall nonparametrically take account of their underlying dependence relations; hence we shall work within the NPC approach. Theory and methods for these kinds of solutions are fully discussed in Pesarin (2001). In the NPC approach we need to combine the p-values λ_k associated with partial tests by a non-degenerate and measurable combining function ψ (for example, Fisher's: $T''_F = -\sum_k \log(\lambda_k)$ and Tippett's: $T''_T = \max_k(1-\lambda_k)$). A further frequently used combining function is the so-called direct combination consisting in a function of partial test statistics instead of related p-values (for example $T''_D = \sum_k T_k$, where T_k denotes the partial test). Thus, T_{AD} may be seen as a direct nonparametric combination of Fisher's exact probability tests and so it enjoys all NPC properties.

If permutation partial tests satisfy some mild conditions and the combining function satisfies some properties, which are generally easy to check and justify, it is possible to obtain a combined test which is at least exact, unbiased and consistent (for details see Pesarin, 2001).

With reference to the $2 \times K$ design, another application of the NPC methodology relates to the joint analysis of tests on sampling moments. To this end, let us assign ranks W to ordered classes, i.e. let us transform A_k into k, and consider the following rule: two discrete distributions defined on the same support, with a finite number K of distinct real values, are equal if and only if their first $K-1$ moments are equal, because their characteristic functions, as well as their probability generating functions, depend only on these few moments. Consequently, we are allowed to write the global hypotheses as

$$H_0 : \left\{X_1 \overset{d}{=} X_2\right\} = \left\{ \bigcap_{r=1}^{K-1} \mathbf{E}\left(W_1^r\right) = \mathbf{E}\left(W_2^r\right)\right\}$$

and

$$H_1 : \left\{X_1 \overset{d}{>} X_2\right\} = \left\{ \bigcup_{r=1}^{K-1} \mathbf{E}\left(W_1^r\right) > \mathbf{E}\left(W_2^r\right)\right\}.$$

Let us observe that the hypothesis $H_1^{\cdot}\left\{X_1 \overset{d}{>} X_2\right\} \equiv \left\{W_1 \overset{d}{>} W_2\right\}$ is not equivalent to the hypothesis $H_1 : \left\{ \bigcup_{r=1}^{K-1} \mathbf{E}\left(W_1^r\right) > \mathbf{E}\left(W_2^r\right)\right\}$ because, from a mathematical point of view, the equivalence is true only in the two-sided test, that is $\left\{X_1 \overset{d}{\neq} X_2\right\} \equiv \left\{ \bigcup_{r=1}^{K-1} \mathbf{E}\left(W_1^r\right) \neq \mathbf{E}\left(W_2^r\right)\right\}$, but the moment inequalities in H_1'' do not imply stochastic dominance in H_1'. In order to prove this, let us consider the data in Table 1, where for the two compared populations the hypothesis H_1'' is true. In fact $E(W_1) = 2.5 > 2.4 = E(W_2)$, $\mathbf{E}(W_1^2) = 7.3 > 6.4 = E(W_2^2)$ and $E(W_1^3) = = 23.5 > 18.6 = E(W_2^3)$.

If we consider the CDFs, we can see that W_1 do not dominate W_2 (see Figure 1), so we are under H_1'' but not under H_1' and inequalities on moments do not imply stochastic dominance.

TABLE 1. - *Distribution of two ordered categorical variables with four categories*

	categories				
	1	*2*	*3*	*4*	
W_1	0.2	0.3	0.3	0.2	1.00
W_2	0.1	0.5	0.3	0.1	1.00

The test on moments with a directional alternative can be applied when the rejection of the null hypothesis implies that the stochastic dominance hypothesis is true, i.e. when just one direction is possible under the alternative hypothesis. For example, in several medical studies the objective is to show the superiority of a new experimental or investigative treatment over the active control (superiority trial) or to establish that the effect of the experimental treatment, when compared to the active control, is not below a given non-inferiority margin (non-inferiority trial). In the first case, if it is proved that the treatment effect cannot be lower than the control effect, the alternative hypothesis must be directional and only one choice is possible for the direction of the alternative hypothesis. A sponsor of an experimental treatment may logically decide to conduct a non-inferiority trial when he believes that the active control effectiveness cannot be exceeded. Examples are: new drugs that may have fewer side effects, a new product's possible lower costs, the sponsor's better access to the market, etc. If the direction of the alternative hypothesis is not known, the test on moments cannot be performed.

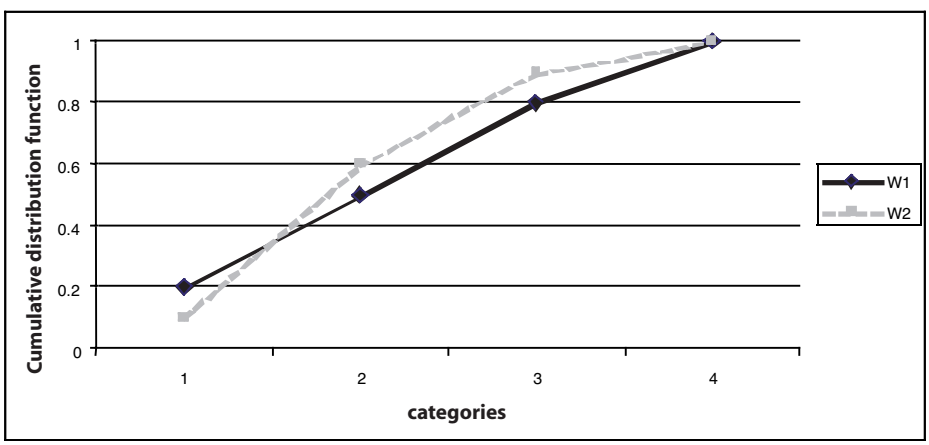

FIGURE 1. - *CDFs of distributions in Table 1*

For each $r, r = 1, ..., K-1$; let us now consider the permutation partial test statistic based on the difference between the rth sampling moments or on a permutationally equivalent statistic, such as $T^*_{Wr} = \sum_{k \le K} k^r f^*_{1k} / n_1$. We note that all these partial

tests are exact, unbiased and consistent. An NPC test associated with a combining function ψ is $T''_{W\psi} = \psi(\lambda_{W1}, ..., \lambda_{WK-1})$ and therefore, for any ψ, combined tests $T''_{W\psi}$ are at least exact, unbiased and consistent (for details see Pesarin, 2001). Of course the same kind of solution works even if, instead of ranks, distinct bounded real scores $w_1 < w_2 < ... < w_K$ are assigned to classes. In section 5, results of Monte Carlo simulations are shown to evaluate how much the results of the test change according to the set of chosen scores.

4. CONFOUNDING EFFECTS: A SOLUTION BASED ON STRATIFICATION AND THE NPC METHOD

In a two-sample test with independent samples the observed difference between the compared groups may be caused not only by the treatment effect but also by confounding factors. Thus the confounding effects may cause a bias in the observed effect which consequently does not reflect the real treatment effect. Often some of the expected confounding factors are known and observed. In these cases it is possible to apply suitable statistical methods to adjust the observed effects with respect to confounding effects. In stratification methods statistical units are grouped into strata defined according to the values of confounding factors. The aim of the method is to remove the confounding effects comparing units within the strata. Thus the results obtained for each stratum are summarized in a single measure related to the whole sample.

The application of this procedure to our testing problem implies performing partial tests for each stratum and the combination of these partial tests to obtain a global overall test. The NPC method described in the previous sections can be used to this end.

Let us again consider the test on moments and suppose that the units are grouped into S strata, so that $S \times (K-1)$ partial tests are performed, where K is the number of categories of the response variable. The overall hypotheses can be written in the following way and correspondingly analysed within strata:

$$H_0 : \bigcap_{s=1}^{S} \left\{ \bigcap_{r=1}^{K-1} \left[\mathbf{E}\left(W_{1s}^{r}\right) = \mathbf{E}\left(W_{2s}^{r}\right) \right] \right\} = \bigcap_{s=1}^{S} H_{0s}$$

against

$$H_1 : \bigcup_{s=1}^{S} \left\{ \bigcup_{r=1}^{K-1} \left[\mathbf{E}\left(W_{1s}^{r}\right) > \mathbf{E}\left(W_{2s}^{r}\right) \right] \right\} = \bigcup_{s=1}^{S} H_{0s}$$

where a breakdown into a set of suitable sub-hypotheses is emphasized and $W_j s, j = 1, 2;$ denotes the score variable W for the jth population of the sth stratum. The solution to the multistrata testing problem consists of three steps: 1) perform $S \times (K-1) T_{sr}$ partial tests on moments; 2) for each stratum s, combine the tests

$T_s1, ..., TsK-1$ (within strata combination) obtaining partial combined tests $T'_1, ..., T'_S$; 3) combine $T'_1, ..., T'_S$ to obtain the overall multistrata test on moments T''. This procedure is displayed in Figure 2.

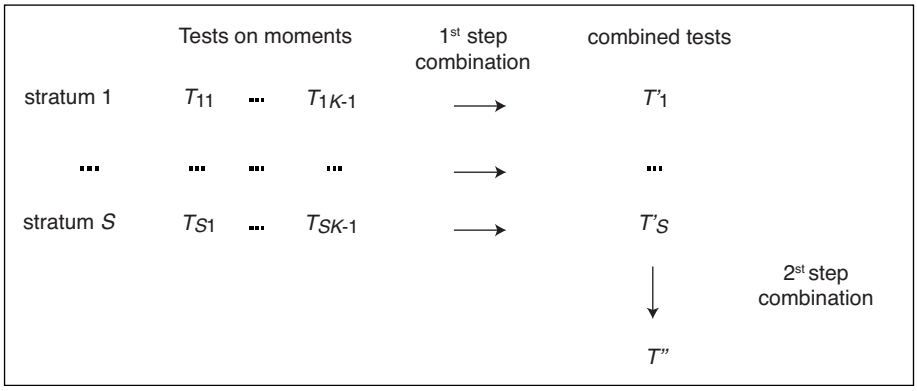

FIGURE 2. - *Multistrata NPC test on moments*

5. AN OBSERVATIONAL STUDY ON THE PROFESSIONAL PLACING OF POST-DOCS

Let us consider data from Arboretti *et al.* (2005) regarding an observational study carried out in 2004 at the University of Ferrara on the professional placing of Post-Docs. We report results by two doctorate fields - economic-legal (EL) and biological-medical (BM) - regarding Post-Docs' ''satisfaction with courses and seminars attended when they were PhD students''. All variables are ordered categorical (scores from 1 to 4: not at all, not very, quite, very satisfied, i.e. $1 = k = K = 4$). It is of interest to highlight differences in the interviewed Post-Docs' satisfaction profile and in particular we wish to answer the question, ''Are EL Post Docs more satisfied then BM Post Docs?'' From an inferential point of view this gives rise to a testing problem with a stochastic dominance alternative.

Table 2 shows the composition of observed sample by gender and doctorate field. It is clear that the number of males in the BM field is very small compared to the EL field where the number of males is quite similar to the number of females. Indeed the percentage of males in the BM field is 14.8% against 41.9% in the EL field.

TABLE 2. - *Sample composition by gender and doctorate field*

	field		
Gender	*BM*	*EL*	*Total*
M	4	13	17
F	23	18	41
Total	27	31	58

Gender is a typical confounding factor in observational studies derived from surveys on university evaluation therefore the multistrata test on moments can be suitable in this case. The hypotheses are

$$H_0 : \left(X_{mEL} \overset{d}{=} X_{mBM}\right) \bigcap \left(X_{fEL} \overset{d}{=} X_{fBM}\right)$$

against

$$H_1 : \left(X_{mEL} \overset{d}{>} X_{mBM}\right) \bigcup \left(X_{fEL} \overset{d}{>} X_{fBM}\right)$$

where X_sEL and X_sBM, $s = m,f$; denote the within stratum s response variable (Post Docs' satisfaction) for the *EL* and *BM* groups respectively. The observed relative frequencies for each sample and each stratum are listed in Table 3.

TABLE 3. - *Relative frequencies for each sample (BM and EL) and each stratum (gender)*

	BM		*EL*	
category	*M*	*F*	*M*	*F*
1	0.25	0.04	0.00	0.06
2	0.25	0.22	0.08	0.06
3	0.50	0.57	0.46	0.67
4	0.00	0.17	0.46	0.22
Total	1.00	1.00	1.00	1.00

Figure 3 shows the results for each stratum of the multistrata test on moments with Fisher's combining function T''_{WF}.

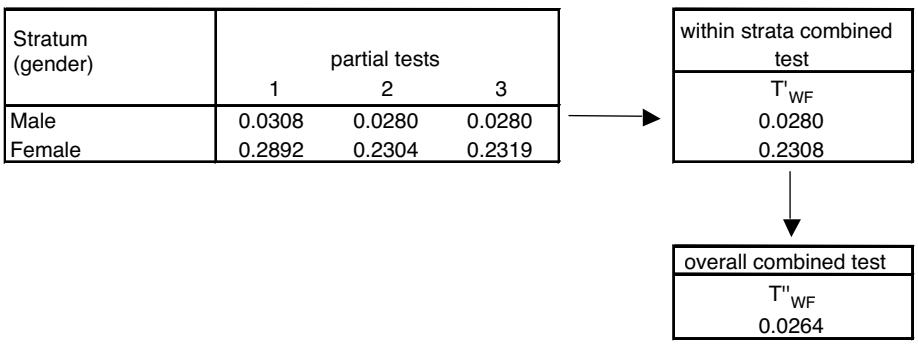

FIGURE 3. - *Results of multistrata permutation tests on moments*

The p-value of the global test is 0.0264 (Fisher's combination) indicating a global difference between the two doctorate fields.

6. MONTE CARLO SIMULATIONS

The performance of the test on moments (with Fisher's or Tippett's combining function) for the univariate case discussed in section 3 with respect to the two competitors proposed in the literature, described in section 2, has been studied using Monte Carlo simulations.

For the present simulation study, data were generated from ordinal multinomial distributions with four categories. By assuming the 1st sample came from the control population and the 2nd sample from the treatment population, we supposed that the treatment effect reduced the frequency of upper categories. The restricted alternative hypothesis is defined as

$$H_1 : \left\{ X_1 \overset{d}{>} X_2 \right\} = \{F_1(A_k) \leq F_2(A_k), k = 1, ..., 3\}$$

where at least one inequality is strict. The associated distribution of the 2nd sample presented an absolute frequency reduction of 40% for category 4. In settings (a) and (b) shown in Table 4, this reduction shifted entirely with respect to category 3 and 1, while in setting (c) the reduction split over categories 2 and 3.

For each configuration we performed 1000 Monte Carlo simulations and 1000 Conditional Monte Carlo iterations to evaluate the permutation distribution. For each independent pair of samples, we considered sample sizes of $n_1 = n_2 = 30$, $n_1 = n_2 = 60$, $n_1 = 30$, $n_2 = 20$ and $n_1 = 60, n_2 = 40$. The results in Table 5 show good behaviour of all solutions under the null hypothesis. In particular, for nominal level $\alpha = 1\%$, the simulated type-I error rates for T''_{WF} ranged from 0.8% $(n_1 = 30, n_2 = 20)$ to 1.2% $(n_1 = n_2 = 60)$ and for W_BM they ranged from 0.8% $(n_1 = 60, n_2 = 40)$ to 1.5% $(n_1 = n_2 = 30)$; for nominal $\alpha = 5\%$, the rejection rates for T''_{WF} ranged from 4.5% (n_1=60, n_2 =40) to 5.1% (n_1= n_2 = 30) and for W_{BM} they ranged from 4.0% (n_1=60, n_2 =40) to 5.8% (n_1= n_2 =30); for nominal α =10%, the rejection rates for T''_{WF} ranged from 8.6% (n_1=60, n_2 =40) to 10.7% (n_1= n_2 = 30) and for W_{BM} they ranged from 8.4% (n_1=60, n_2 =40) to 11.9% (n_1= n_2 =30). In general we can say that simulated type-I error rates of tests on moments and the Wilcoxon test are very similar. The rejection rates under H_0 of the Brunnel and Munzel test are somewhat higher than those of the others.

TABLE 4. - *Frequency distributions (%) for data generation*

Frequency distribution (%)	*Category*			
	1	*2*	*3*	*4*
1st sample	5	10	15	70
2nd sample (a)	5	10	55	30
2nd sample (b)	45	10	15	30
2nd sample (c)	5	30	35	30

Under the alternative hypothesis (see Table 6), the tests on moments present the highest empirical powers for almost all the probability configurations described in Table 4 and for all sample sizes. In Table 6, results for $n_1 = n_2 = 30$ and $n_1 = 30$ and $n_2 = 20$ are shown. Thus we can say that tests on moments using Fisher's combining function or Tippett's combining function seem be the most powerful.

TABLE 5. - *Achieved significance levels ($n_1 = n_2 = 30$)*

Nominal α	T''_{WF}	T''_{WT}	W	W_{BM}
0.010	0.010	0.011	0.010	0.015
0.025	0.018	0.018	0.021	0.027
0.050	0.051	0.052	0.053	0.058
0.100	0.107	0.105	0.104	0.119
0.200	0.196	0.198	0.206	0.220
0.300	0.299	0.296	0.298	0.308
0.400	0.408	0.405	0.394	0.410
0.500	0.494	0.495	0.497	0.515
0.600	0.584	0.584	0.578	0.593
0.700	0.673	0.671	0.679	0.691
0.800	0.793	0.790	0.788	0.807
0.900	0.897	0.896	0.898	0.900
1.000	1.000	1.000	1.000	1.000

To evaluate the robustness of the test on moments in relation to different score transformations, we performed Monte Carlo simulations and plotted the rejection rates corresponding to different possible transformations. The considered transformations are illustrated in Table 7. In transformation 1 the ranks of categories are assigned to classes; transformations 2 and 3 are asymmetric transformations which assign very large values to the best judgements and very small values to the worst judgements; transformation 4 is symmetric but the distance between adjacent scores is not constant.

TABLE 6. - *Empirical power for the three distribution configurations*

nominal alpha	*0.010*	*0.025*	*0.050*	*0.100*
	Situation: a; sizes: $n_1 = n_2 = 30$			
Rank Sum Test	0.355	0.523	0.627	0.747
Test on Mom. Fisher	0.572	0.704	0.787	0.875
Test on Mom. Tippett	0.330	0.470	0.592	0.691
Brunner-Munzel Test	0.449	0.584	0.666	0.777
	Situation: b; sizes: $n_1 = n_2 = 30$			
Rank Sum Test	0.847	0.921	0.958	0.985
Test on Mom. Fisher	0.856	0.921	0.957	0.983
Test on Mom. Tippett	0.884	0.936	0.972	0.988
Brunner-Munzel Test	0.909	0.946	0.977	0.989
	Situation: c; sizes: $n_1 = n_2 = 30$			
Rank Sum Test	0.540	0.698	0.789	0.886
Test on Mom. Fisher	0.674	0.783	0.868	0.928
Test on Mom. Tippett	0.577	0.714	0.803	0.885
Brunner-Munzel Test	0.634	0.756	0.841	0.905
	Situation: a; sizes: $n_1 = 30$, $n_2 = 20$			
Rank Sum Test	0.255	0.388	0.542	0.669
Test on Mom. Fisher	0.455	0.596	0.697	0.792
Test on Mom. Tippett	0.242	0.364	0.489	0.621
Brunner-Munzel Test	0.365	0.511	0.625	0.720
	Situation: b; sizes: $n_1 = 30$, $n_2 = 20$			
Rank Sum Test	0.710	0.809	0.879	0.933
Test on Mom. Fisher	0.735	0.820	0.882	0.934
Test on Mom. Tippett	0.778	0.864	0.920	0.964
Brunner-Munzel Test	0.768	0.854	0.903	0.953
	Situation: c; sizes: $n_1 = 30$, $n_2 = 20$			
Rank Sum Test	0.377	0.532	0.662	0.778
Test on Mom. Fisher	0.520	0.682	0.770	0.856
Test on Mom. Tippett	0.443	0.578	0.682	0.786
Brunner-Munzel Test	0.496	0.623	0.737	0.814

TABLE 7. - *Score transformations*

Frequency distribution (%)	*Category*			
	1	*2*		*4*
Transformation 1	1	2	3	4
Transformation 2	1	2	4	8
Transformation 3	1	5	7	8
Transformation 4	2	5	7	10

In Figure 4 the rejection rates of the test on moments with Fisher's combining function are plotted for different score transformations in the balanced case ($n_1 = n_2 = 30$). For the unbalanced case the plots are similar. In general the results are not transformation invariant. In situations (a) and (c), where the difference between the distributions is linked only to a couple of parameters (probabilities), transformation 3 gives bad results because its power seems to be very low. In situation (b), the least powerful test corresponds to transformation 2. The other transformations show very similar performances. In general we can conclude that effects induced by score transformation on the performance of the test on moments depends on the real unknown distribution. However, the test performs well and the results are transformation invariant if a symmetric or almost symmetric transformation is applied.

7. CONCLUSIONS

The nonparametric combination method is suitable and effective for many complex testing problems which, in a parametric framework, are very difficult or even impossible to solve. One major feature of the nonparametric combination of dependent tests, provided that the permutation principle applies, is that one must pay attention to a set of partial tests, each appropriate for the related sub-hypothesis. The researcher is only required to make sure all partial tests are marginally unbiased, a sufficient condition which is generally easy to check.

We propose a two-sample permutation test for directional alternatives and categorical variables based on the transformation of the ordered categorical responses into a numerical variable, using an extension of the NPC method through stratification, in order to take the bias induced by confounding effects into account.

The Monte Carlo experiments shown in this contribution prove that the test on moments presents general good behaviour. Its power is higher than that of other widespread nonparametric tests and the results are transformation invariant if a symmetric or almost symmetric score transformation is applied.

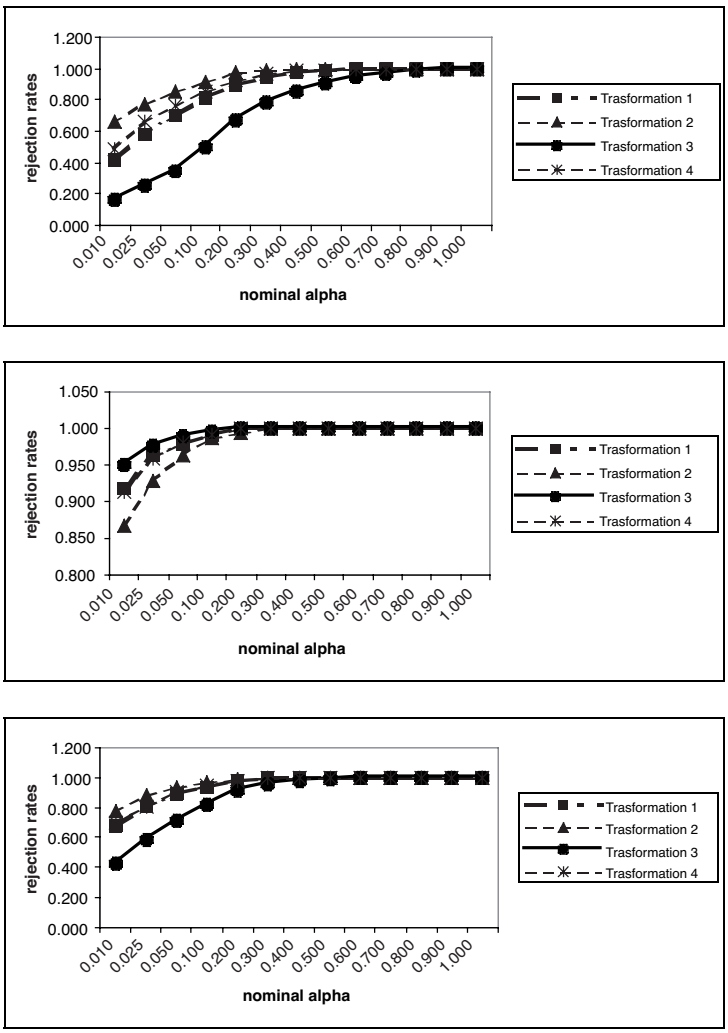

FIGURE 4. - *Rejection rates of the test on moments with Fisher's combining function for different score transformations ($n_1=n_2=30$)*

ACKNOWLEDGEMENT

The research has been supported by the University of Padova CPDA088513. (Coordinator: Prof. Luigi Salmaso)

RIASSUNTO

Spesso nei moderni sistemi socio-economici lo scopo di un'analisi di performance o di una valutazione di qualità consiste nel confronto di diversi prodotti o servizi, diversi impianti produttivi o centri di erogazione del servizio, diverse azioni, diversi trattamenti, con l'obiettivo di individuare quale

sia il migliore. Tale valutazione è complessa perché le caratteristiche da valutare sono spesso misurate con dati categoriali e per la presenza di fattori di confondimento. In questo articolo viene proposta una soluzione basata su un nuovo test di permutazione costruito a partire da numero finito di momenti campionari e per ridurre gli effetti di confondimento viene proposta l'applicazione del metodo NPC dopo un opportuna stratificazione dei campioni. I risultati delle simulazioni realizzate col metodo Monte Carlo per confrontare la soluzione di permutazione proposta con altri test non parametrici dimostra il buon comportamento in potenza del test sui momenti oltre che la sua robustezza rispetto a diverse possibili trasformazioni dei dati.

REFERENCES

Arboretti Giancristofaro R., Pesarin F., Salmaso L. (2005). Nonparametric approaches for multivariate testing with mixed variables and for ranking on ordered categorical variables with an application to the evaluation of PhD programs, in S. Sawilowsky (Ed.), *Real Data Analysis*, American Educational Research Association, Age Publishing.

Brunner E., Munzel U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and small-sample approximation. *Biometrical Journal*, **42**, 17-25.

Cohen A., Kemperman J.H.B., Madigan D., Sakrowitz H.B. (2003). Effective directed tests for models with ordered categorical data. *Australian and New Zealand Journal of Statistics*, **45**, 285-300.

Cohen A., Kemperman J.H.B., Sakrowitz H.B. (2000). Properties of likelihood inference for order restricted models. *Journal of Multivariate Analysis,* **72**, 50-77.

Dardanoni V., Forcina A. (1998). A unified approach to likelihood inference on stochastic ordering in a non-parametric context. *Journal of the American Statistical Association*, **93**, 1112-1123.

Dykstra R.L., Kochar S., Robertson T. (1995). Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association*, **90**, 1039-1040.

El Barmi H., Dykstra R. (1995). Testing for and against a set of linear inequality constraints in a multinomial setting. *The Canadian Journal of Statistics*, **23** (2), 131-143.

Hirotsu C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika*, **69**, 567-577.

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

Mann H.B., Whitney D.R. (1947). On a test of whether one of two random variables is stocastically larger than the other. *Annals of mathematical statistics*, **18**, 50-60.

Perlman M.D., Wu L. (2002). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *Journal of Statistical Planning and Inference*, **107**, 173-186.

Pesarin F. (1994). Goodness-of-fit testing for ordered discrete distributions by resampling techniques. *Metron*, **LII**, 57-71.

Pesarin F. (2001). *Multivariate Permutation Test With Application to Biostatistics*. Wiley, Chichester.

Pesarin F. (2004). Alcuni problemi di verifica delle ipotesi per variabili categoriali. *Statistica*, **LXIV** (2), 367-386.

Sampson A.R., Whitaker L.R. (1989). Estimation of multivariate distributions under stochastic ordering. *Journal of the American Statistical Association*, **84**, 541-548.

Silvapulle M.J., Sen P.K. (2005). *Constrained Statistical Inference, Inequality, Order, and Shape Restrictions*. Wiley, New York.

Troendle J.F. (2002). A likelihood ratio test for the nonparametric Behrens-Fisher problem. *Biometrical Journal*, **44** (7), 813-824.

Wang Y. (1996). A likelihood ratio test against stochastic ordering in several populations. *Journal of the American Statistical Association*, **91**, 1676-1683.

Wilcoxon F. (1945). Individual comparisons by ranking method. *Biometrics*, **1**, 80-83.