Multivariate permutation tests for ordered categorical variables

Fortunato Pesarin
Department of Statistics, University of Padova
via C. Battisti, 241,
35121 PADOVA Italy
pesarin@stat.unipd.it

1. Introduction

Let us consider ørst the goodness-of-øt problem for one-dimensional ordered categorical variable in a two-sample situation. Stochastic dominance (restricted) alternatives are considered because are rather diŒcult within the framework of likelihood ratio tests (Wang, 1996) and are frequently encountered in practical problems. The most common problems related to unrestricted alternatives, C-sample, $C > 2$, and multivariate situations are mentioned as straightforward extensions using the method of nonparametric combination of dependent tests.

Assume that the support of a univariate ordered categorical variable X is partitioned into $k \geq 2$ classes $\{A_i,\ i = 1,...,k\}$, and that the data are classiøed according to two levels of a treatment. Thus, given two independent samples $\mathbf{X}_j = \{X_{jr},\ r = 1,...,n_j\}$, $j = 1,2$, we wish to test the hypotheses $H_0 : \{X_1 \stackrel{d}{=} X_2\} = \{F_1(A_i) = F_2(A_i),\ i = 1,...,k\}$ against: $H_1 : \{X_1 \stackrel{d}{>} X_2\} = \{\bigcup_i [F_1(A_i) < F_2(A_i)]\}$, where the functions $F_j(A_i) = \Pr\{X_j \leq A_i\}$, $j = 1,2$, play the role of cdfs for X_j. Note that H_1 deønes the stochastic dominance of X_1 with respect to X_2. Observed data are usually organized in a $2 \times k$ contingency table $\{f_{ji} = \#(X_{jr} \in A_i),\ i = 1,...,k,\ j = 1,2\}$. $N_{ji} = \sum_{s \leq i} f_{js}$, $f_{\cdot i} = f_{1i} + f_{2i}$ indicate cumulative and marginal frequencies, $n_j = \sum_i f_{ji}$, and $n = n_1 + n_2$. The pooled data set $\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2$ and the set $\{n_1, n_2, f_{\cdot 1}, \ldots, f_{\cdot k}\}$ both are sets of suŒcient statistics under H_0. An optimal solution when $k = 2$ is Fisher's well-known exact probability test. Note that H_0 implies that the data of two groups are exchangeable, so that the permutation testing principle applies.

The permutation analysis of discrete distributions, especially in multidimensional situations, is easier if, in place of the usual contingency tables, data are unit-by-unit represented by listing the n individual data. For instance, in the $2 \times k$ example, pooled data are represented by the vector $\mathbf{X} = \{X(r),\ r = 1,...,n;\ n_1, n_2\}$, in which ørst n_1 responses are belonging to group 1 and the next n_2 to group 2. As a solution we may consider the permutation test

$$T_D^* = \sum_i (N_{2i}^* - N_{1i}^*) \left[4 \frac{N_{\cdot i}}{n} \left(\frac{n - N_{\cdot i}}{n} \right) \frac{n_1 n_2}{n-1} \right]^{-\frac{1}{2}},$$

where $N_{\cdot i} = N_{1i} + N_{2i} = N_{1i}^* + N_{2i}^*$, N_{1i}^* and N_{2i}^*, $i = 1,...,k-1$ are permutation cumulative frequencies. T_D essentially compares two EDFs, thus it corresponds to the discrete version of a statistic following Anderson-Darling's goodness-of-øt test for dominance alternatives. Also of interest, is the likelihood ratio test for restricted alternatives (El Barmi and Dykstra, 1995).

2. On nonparametric combination solutions

We observe that the i-th summand in T_D^* may be seen as a partial test for the sub-hypotheses: $H_{0i} : \{F_i = G_i\}$, against $H_{1i} : \{F_i < G_i\}$, corresponding to the i-th 2×2 sub-table extracted from the given $2 \times k$ complete table. Thus, T_D^* may be seen as a direct nonparametric combination (Pesarin, 1990, 2001) of $k-1$ partial tests, all expressed in standardized form. Of course, to each $k-1$ extracted sub-tables we may apply any proper marginally unbiased test, as for instance Fisher's exact probability, followed by a nonparametric combination of related

p-values λ_i. This kind of solution is discussed in Pesarin (2001). Among the many combination functions, apart for the direct, we may use Fisher's or Liptak's. These respectively give $T''_F = -\sum_i \log(\lambda_i)$ and $T''_L = \sum_i \Phi^{-1}(1-\lambda_i)$, where Φ is the standard normal cdf.

3. Some extensions

a) In a two-sample problem with unrestricted or non-dominance alternatives the hypotheses are $H_0 : \{X_1 \stackrel{d}{=} X_2\}$ against $H_1 : \{X_1 \stackrel{d}{\neq} X_2\} = \{\bigcup_i [F_1(A_i) \neq F_2(A_i)]\}$. Thus, T_D becomes

$$T^{*2}_D = \sum_i \left(N^*_{2i} - N^*_{1i}\right)^2 \left[N_{\cdot i} \cdot (n - N_{\cdot i})\right]^{-1},$$

which corresponds to a two-sample Anderson-Darling test adjusted for discrete variables.

b) The extension of hypotheses and test statistics to $C > 2$ groups for unrestricted alternatives and ordered variables is straightforward. Denoting by $F^*_{ji} = N^*_{ji}/n_j$ and $\bar{F}_i = N_{\cdot i}/n$, where $N_{\cdot i} = \sum_j N_{ji}$, the partial and pooled EDFs, we clearly have

$$T^{*2}_{CD} = \sum_{j=1}^{C} \sum_i \left(F^*_{ji} - \bar{F}_i\right)^2 \left[\bar{F}_i \cdot (1 - \bar{F}_i) \cdot (n - n_j)/n_j\right]^{-1}.$$

c) For an extension to the multivariate version of the two-sample problem with dominance alternatives, let us assume that the response variable is q-dimensional $\mathbf{X} = (X_1, ..., X_q)$, whose related numbers of ordered classes are $\mathbf{k} = (k_1, ..., k_q)$ and that units n_1 and n_2 are independently observed from $\mathbf{X}_1$ and $\mathbf{X}_2$. The hypotheses we wish to test are $H_0 : \{\mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2\} = \{\bigcap_{h=1}^{q} \bigcap_i (F_{hi} = G_{hi})\}$ and $H_1 : \{\mathbf{X}_1 \stackrel{d}{>} \mathbf{X}_2\} = \{\bigcup_{h=1}^{q} \bigcup_i (F_{hi} < G_{hi})\}$, where F_{hi} and G_{hi} play the role of cdfs for the h-th variable.

This problem (Wang, 1996) is considered as extremely diŒcult when approached within the likelihood ratio, whereas within the nonparametric combination method its solution is straightforward. For instance, the direct combination on standardized partial tests, which shows the Anderson-Darling structure, is

$$T''^*_{MD} = \sum_h \sum_i \left(N^*_{h2i} - N^*_{h1i}\right) \left[4\tfrac{N_{h\cdot i}}{n}\left(\tfrac{n - N_{h\cdot i}}{n}\right)\tfrac{n_1 n_2}{n-1}\right]^{-1/2},$$

where it is worth noting that the component $T^*_{Dh} = \sum_i \left(N^*_{h2i} - N^*_{h1i}\right) \left[4\tfrac{N_{h\cdot i}}{n}\left(\tfrac{n - N_{h\cdot i}}{n}\right)\tfrac{n_1 n_2}{n-1}\right]^{-1/2}$ is the partial permutation test related to the hth variable, $h = 1, \ldots, q$.

REFERENCES

El Barmi, H. and Dykstra, R. (1995) Testing for and against a set of linear inequality constraints in a multidimensional setting. The Canadian J. of Statist., 23, 131-143.

Pesarin, F. (1990) Tecniche di ricampionamento e veriøca multidimensionale delle ipotesi. Statistica, L, 4, 483-501.

Pesarin, F. (2001) Multivariate Permutation Tests With Applications to Biostatistics. Wiley, Chichester.

Wang. Y. (1996) A likelihood ratio test against stochastic ordering in several populations. J. of the Am. Statist. Ass., 91, 1676-1683.

RESUME

Cette communication trate des tests de permutation pour variables categoriales multidimensionnelles sous des alternatives de dominance stochastique. La solution est obtenue par la methode de combinaison non-parametrique de tests partiels dependants.