Metodološki zvezki, Vol. 4, No. 1, 2007, 21-36

Permutation Tests for Heterogeneity Comparisons in Presence of Categorical Variables with Application to University Evaluation

Rosa Arboretti Giancristofaro[1] and Stefano Bonnini[2]

Abstract

In social sciences researchers often meet the problem of determining if the distribution of a categorical variable is more concentrated in population X_1 than in population X_2. For example the effectiveness of two different PhD programs can be evaluated in terms of the heterogeneity of the set of job opportunities. The job opportunities are nominal categorical variables and populations X_1 and X_2 include all PhD holders for program 1 and program 2. We may define that a PhD program is "better than another" if it is able to offer a larger variety of job opportunities. Several other examples can be mentioned to highlight the importance of heterogeneity comparison problems in social sciences; moreover this problem occurs also very often in genetics, biology, medical studies and other sciences.

The nonparametric solution of this problem has similarities to that of permutation testing for stochastic dominance on ordered categorical variables, i.e. testing under order restrictions. If ordering of probability parameters in H_0 is unknown and it has to be estimated by sampling data, only approximate nonparametric solutions are possible within the permutation approach. Main properties of test solutions and some Monte Carlo simulations in order to evaluate the tests' behaviour under H_0 and H_1, will be presented. A real problem concerned with University evaluation is also discussed.

1 Indexes of heterogeneity

In social sciences researchers often meet the problem of establishing if the distribution of a categorical variable is more concentrated in population X_1 than in

[1] Rosa Arboretti Giancristofaro, Department of Mathematics, University of Ferrara, Via Machiavelli 35, 44100, Ferrara, Italy; rosa.arboretti@unife.it

[2] Stefano Bonnini, Department of Mathematics, University of Ferrara, Via Machiavelli 35, 44100, Ferrara, Italy; bnnsfn@unife.it

population X_2. For example the effectiveness of two different PhD programs can be evaluated in terms of the heterogeneity of the set of job opportunities. The job opportunities nominal categorical variables and populations X_1 and X_2 include all PhD holders for program 1 and program 2. We may define that a PhD program is "better than another" if it is able to offer a larger variety of job opportunities. In this case the monitoring of the first employment of PhD holders through a post-doc survey is useful. Another example is given when a company can be interested in the evaluation of the degree of heterogeneity of the customer base. Heterogeneity can concern educational qualifications, hobbies or other categorical variables. In this case the heterogeneity of the customer base of product 1 and that of product 2 can be compared to obtain useful information for the marketing strategy.

In the evaluation of an academic course the monitoring of the type of secondary school from which students come can be of interest for academic management. A heterogeneity measure of the distribution of the type of secondary school can be adopted in order to compare two different academic courses. Several other examples can be mentioned to highlight the importance of heterogeneity comparison problems in social sciences, but this problem occurs very often also in genetics, biology, medical studies and other sciences.

The concept of *heterogeneity* is mostly used in the field of descriptive statistics. *Homogeneity* notoriously means the disposition of a statistical variable X to always be manifested in the same category A_i, $i = 1,\ldots,k$, $1< k <\infty$. A set of statistical units is therefore homogeneous if all units that make it up are characterized by the same category. If this does not occur, that is if at least two categories in the set of statistical units are found, then heretogeneity is indicated by absence of homogeneity. Therefore the degree of heterogeneity obviously depends on the number of categories observed as well as on their associated frequencies. In particular the heterogeneity is at a minimum if the distribution of the observed variable is degenerate, i.e. it presents a single category with a relative frequency equal to 1 and all the others with a frequency equal to 0. On the other hand heterogeneity is at a maximum if the variable is equally distributed on all categories.

Consequently an index that syntetically translates the degree of heterogeneity of the observed phenomenon must have the following characteristics:

To assume minimum value when the phenomenon under study is manifested with a single category, i.e. in the presence of maximum homogeneity;

To assume increasingly greater values the more one moves away from the degenerate distribution and the more one approaches the equidistribution;.

To assume the maximum value in presence of equidistribution.

Heterogeneity can be associated not only with the concept of concentration but also with that of diversity, that is the attitude of a qualitative variable to assume different modalities. It is directly associated with the concept of uncertainty and that of information because in the case of minimum heterogeneity also the

uncertainty of a decision is at a minimum and the information derivable from the single observation is at a maximum. In the opposite case of maximum heterogeneity one has maximum uncertainty and minimum information derivable from the single unit. Starting from this notion, various indicators were proposed of which only the most commonly used will be mentioned.

The index of heterogenity proposed by Gini (1912), for a variable X which assumes k categories with relative frequencies f_i, i=1,2,...,k, is

$$G = \sum_{s=1}^{k} f_i(1 - f_i) = 1 - \sum_{s=1}^{k} f_i^2 , \tag{1.1}$$

whose normalized version is $G^* = G(k - 1)/k$.

Shannon's index of diversity, also called index of *entropy* of a distribution, is associated with information theory (Shannon, 1948) and is

$$H = \sum_{s=1}^{k} f_i \log(1/ f_i) = -\sum_{s=1}^{k} f_i \log(f_i), \tag{1.2}$$

where log(·) are natural logarithms, even though in the original version they were in base 2, and we assume that 0 log(0) = 0. The normalized version is $H^*= H/\log(k)$.

A generalized index taken from the field of information theory, is the *generalized index of entropy* of order α proposed by Rényi (1966)

$$H_\alpha = \frac{1}{1-\alpha} \log \sum_{i=1}^{k} f_i^{\alpha} , \tag{1.3}$$

with $\alpha \neq 1$. H_α is a non-increasing function of α. As α varies, one obtains different indices of heterogeneity and some particular cases among the most frequently used are:

$$H_1 = \lim_{\alpha \to 1} \left[\frac{1}{1-\alpha} \log\left(\sum_{i=1}^{k} f_i^{\alpha} \right) \right] = -\sum_{i=1}^{k} f_i \log(f_i) = H ,$$

$$H_2 = -\log\left(\sum_{i=1}^{k} f_i^2 \right) = -\log(1 - G),$$

$$H_\infty = \lim_{\alpha \to \infty} \left[\frac{1}{1-\alpha} \log\left(\sum_{i=1}^{k} f_i^{\alpha} \right) \right] = -\log\left[\sup_{i=1,\ldots,k} (f_i) \right] .$$

2 Permutation tests for heterogeneity: The two-sample case

In the previous section we dealt with heterogeneity from the descriptive point of view, namely considering the indices that measure the degree of heterogeneity of a distribution of frequencies in a certain set of statistical units. From now on we will take into consideration the inferential problem which consists of comparing the sampling heterogeneity of a categorical variable X in two populations, in order to test the hypothesis that the heterogeneity of one population is greater than that of the other. In doing so we shall use some of the indices described in section 1. It is assumed that the values of the variable X can fall within one of the k categories A_1, A_2,..., A_k.

From a formal point of view, given two populations X_1 e X_2, if we indicate with $Het(X_j)$ the degree of heterogeneity of the population X_j (j =1,2), the problem of hypothesis testing can be expressed as follows

$$H_0 : Het(X_1) = Het(X_2)$$

against

$$H_1 : Het(X_1) > Het(X_2).$$

We take into consideration the indeces of Gini $G = \Sigma_i\, p_i(1 - p_i)$, of Shannon $H = - \Sigma_i\, p_i \log(p_i)$ and the versions of that of Rényi for $\alpha = 3$ and for $\alpha \to \infty$, corresponding to $H_3 = - 1/2 \log(\Sigma_i\, p_i^{\,3})$ and $H_\infty = - \log [\sup_i (p_i)]$, where $p_i = Pr \{X \in A_i\}$. The choice of H_3 instead of H_2, which is perhaps the index of Rényi most used in the literature, is dictated by the fact that H_2 is one-to-one related with G, and therefore the two indices imply the same inferential conclusions when applying theory and methods of permutation tests. In other words, two indices are permutationally equivalent. By indicating with p_i, i = 1,2,...k, the parameters of the underlying distribution, with $p_{(i)}$, i =1,2,...,k, we indicate the same parameters arranged in non-increasing order: $p_{(1)} \geq p_{(2)} \geq ... \geq p_{(k)}$. Four indices G, H, H_3 e H_∞ are *order invariant*, i.e. their value does not change if they are calculated with ordered parameters $p_{(i)}$ instead of with parameters p_i. If we indicate with $p_{j(i)}$, i=1,2,...,k, the ordered probabilities for population j, j = 1,2, the fact that the indices of heterogeneity are order invariant allows us to express heterogeneity through ordered parameters: *two populations such that* $\{p_{1(i)} = p_{2(i)}$, i=1,2,...,$k\}$, *i.e. with the same ordered distribution, are equally heterogeneous*. Moreover, if $\{p_{1(i)} = p_{2(i)}$, i=1,2,...,$k\}$ *data of two samples are exchangeable and so the permutation testing principle applies.* In addition, if $P_{j(i)} = \sum_{s \leq i}\, p_{j(s)}$ are the cumulative probabilities for population j referred to the ordered parameters, the null hypothesis of our problem is equivalent to:

$H_0 : P_{1(i)} = P_{2(i)}$, $i = 1,2,...,k$.

Vice versa, in the case of greater heterogeneity of population X_1 with respect to population X_2 implies that $P_{1(i)} \leq P_{2\,(i)}$ must be valid and the strict inequality is valid for some values of i. In formal terms we can write:

$H_1 : P_{1(i)} \leq P_{2(i)}$ $\forall i$ and $P_{1(i)} < P_{2(i)}$ for at least one i.

The definition of the problem by means of cumulative probabilities makes it very similar to the problem of stochastic dominance for ordered categorical variables, with the peculiarity that the order is determined according to the values of the parameters p_i and not according to the categories the variable X can assume. Therefore, our problen can be referred to as one of *dominance in heterogeneity*. We may observe that X in problems of heterogeneity can be a nominal variable because heterogeneity is a property that concerns probabilities and does not involve the categories A_i of X, whereas the problems of stochastic dominance assume that classes $A_1, A_2,..., A_k$.are ordered.

For problems of stochastic dominance the literature offers quite a long list of exact and approximate solutions. Among the many, we mention those of Agresti and Klingerberg (2005), Han *et al*. (2004), Hirotsu (1986), Loughin and Scherer (1998), Loughin (2004), Lumely (1996), Nettleton and Banerjee (2001). For the univariate case most of the methodological solutions proposed are based on the restricted maximum likelihood ratio test. Among these we mention Cohen *et al*. (2000), Silvapulle and Sen (2005), Wang (1996). In general these solutions are criticized because the distributions under the null and alternative hypotheses are asymptotically mixtures of chi-squared variables with weights essentially dependent on the unknown distribution P of the population. Nonparametric proposals are those of Troendle (2002), Brunner and Munzel (2000), Pesarin (1994 and 2001), and Pesarin and Salmaso (2006). The latters, based on the nonparametric combination of dependent permutation tests (NPC), are exact, unbiased, and consistent tests. As far as we are concerned in difference of heterogeneity, it is reasonable to set as test statistic the difference of sampling indices therefore arriving at statistics such as:

$T_G = G_1 - G_2 = \Sigma_i \, (f_{2(i)}^{\,2} - f_{1(i)}^{\,2})$ (test statistic based on Gini's index G),

$T_H = H_1 - H_2 = \Sigma_i \, [f_{2(i)}\log(f_{2(i)}) - f_{1(i)}\log(f_{1(i)})]$ (test statistic based on Shannon's entropy H),

$T_{H3} = H_{31} - H_{32} = 1/2 \, [\log(\Sigma_i \, f_{2(i)}^{\,3}) - \log(\Sigma_i \, f_{1(i)}^{\,3})]$ (test statistic based on Rényi's index H_3),

$T_{H\infty} = H_{\infty\,1} - H_{\infty\,2} = \log(\sup_i f_{2(i)}) - \log(\sup_i f_{1(i)})$ (test statistic based on Rényi's index H_∞),

where G_j, H_j, H_{3j} and $H_{\infty\,j}$ indicate the sampling values calculated for sample j, j = 1,2. Clearly the tests will be significant for large values, i.e. large values observed in the test statistic can lead to the rejection of the null hypothesis in favour of the alternative. In order to apply the tests according to the usual approach, it is necessary to know their sampling distributions subject to a proper estimate under H_0 of the vector of the marginal ordered probabilities $(p_{\bullet(1)}, p_{\bullet(2)},..., p_{\bullet(k)})'$, because the vectors of the probabilities $(p_{j1}, p_{j2},...,p_{jk})'$ as well as those of the ordered probabilities $(p_{j(1)}, p_{j(2)},..., p_{j(k)})'$, j = 1,2, are unknown. In reality this question is not easy to solve exactly, with perhaps the exception where k = 2. For this purpose, instead of the true ordering of unknown parameters $\{p_{j(1)}, p_{j(2)},..., p_{j(k)}; j=1,2\}$, we utilize their estimates based on ordereing the observed frequencies (two empirical orderings):

$$f_{1(1)} \geq f_{1(2)} \geq ... \geq f_{1(k)} \equiv \hat{p}_{1(1)} \geq \hat{p}_{1(2)} \geq ... \geq \hat{p}_{1(k)}$$

and

$$f_{2(1)} \geq f_{2(2)} \geq ... \geq f_{2(k)} \equiv \hat{p}_{2(1)} \geq \hat{p}_{2(2)} \geq ... \geq \hat{p}_{2(k)},$$

thus, obtaining the following ordered table:

Table 1: Estimation of the probabilities ordered by relative frequencies.

	Classes				Sample sizes
Population	(1)	(2)	...	(k)	
P_1	$f_{1(1)}$	$f_{1(2)}$	...	$f_{1(k)}$	n_1
P_2	$f_{2\,(1)}$	$f_{2\,(2)}$	...	$f_{2\,(k)}$	n_2
	$f_{\cdot\,(1)}$	$f_{\cdot\,(2)}$	...	$f_{\cdot\,(k)}$	N

We note that the order is realized separately for each sample and as it is based on relative frequencies rather than on classes, it could be that the i-th column of Table 1 refers to two diverse classes for the two samples. In other words class (i) corresponds to the class whose observed relative frequency occupies the i-th position in the ordered sequence and can be different for the two samples. Obviously the order imposed by the frequencies presents a random component and may vary depending upon sampling variations. Therefore under H_0 data are not exactly exchangeable as it would be if the true order of population parameters were known and used. The exchangeability property can only be obtained

asymptotically. Therefore, permutation solutions are approximate for finite sample sizes and exact only asymptotically.

Using the data in Table 1, the observed values of the test statistics T_G^o, T_H^o, $T_{H_3}^o$ and $T_{H_\infty}^o$ are calculated. For each permutation of the dataset one obtains a new permuted table (as in Table 2), with different values from those of the observed table but with fixed marginal frequencies.

Table 2: Absolute frequencies after a permutation of data.

	Classes				Sample sizes
Population	(1)	(2)	...	(*k*)	
P_1	$n^*_{1(1)}$	$n^*_{1(2)}$	...	$n^*_{1(k)}$	n_1
P_2	$n^*_{2\,(1)}$	$n^*_{2\,(2)}$	...	$n^*_{2\,(k)}$	n_2
	$n._{(1)}$	$n._{(2)}$	...	$n._{(k)}$	N

Using the data of the permutated table in the calculations of test statistics, one obtains the permutation values T_G^*, T_H^*, $T_{H_3}^*$ e $T_{H_\infty}^*$. Calculating the values that can be obtained making all the possible permutations, one obtains the permutation distribution of each test statistic. Alternatively it is possible to extract from the set of all the possible permutations a random sample, thus obtaining conditional Monte Carlo estimates. In this way, for each of the four tests it is possible to calculate the *p-value*, that, if *B* is the number of considered permutations, is given by

$$\square_G = \#\left(T_G^* \geq T_G^o \mid \mathbf{X}\right)/B,$$

$$\square_H = \#\left(T_H^* \geq T_H^o \mid \mathbf{X}\right)/B,$$

$$\square_{H_3} = \#\left(T_{H_3}^* \geq T_{H_3}^o \mid \mathbf{X}\right)/B,$$

$$\square_{H\infty} = \#\left(T_{H\infty}^* \geq T_{H\infty}^o \mid \mathbf{X}\right)/B,$$

respectively for the test based on *G*, *H*, H_3 and H_∞, where $\#\left(T_G^* \geq T_G^o \mid \mathbf{X}\right)$ indicates the number of times permutation values are not lower than the observed ones, conditionally on the dataset $\mathbf{X}=\{X_{ji},\ i=1,\ldots,n_j;\ j=1,2\}$. Therefore, according to the general deciding rule, if the *p-value* is less than or equal to a fixed significance level, the null hypothesis is rejected in favour of the alternative, otherwise the null hypothesis cannot be rejected. In order to summarize our procedure, with reference

to permutation tests, after transformation of the original categories A_1, A_2, …, A_k into the ordered categories (1), (2), …, (*k*), where, for each sample, (1) is the category with the highest frequency, (2) is the category with the second ordered frequency,…, (k) is the category with the lowest frequency, we apply a two sample permutation test for categorical variables. Observed data are generally organized in a $2 \times K$ contingency table (see Table 1). Permutation analysis is probably 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, it is intended that in the dataset first n_1 records belong to the first sample and the rest to the second. It is worth observing that in univariate two-sample designs, since they contain exactly the same amount of information on the unknown distribution of data, the marginal frequencies the pooled data set $\mathbf{X}$ as well as any of its permutations $\mathbf{X}^*$ are equivalent sets of sufficient statistics under H_0. Under H_0 data are exchangeable, hence it is possible to change the position (the row) of some statistical units in $\mathbf{X}$ assigning them to a different sample respect to the observed dataset. In other words it is possible to make permutations of the data set. Considering all the possible $\binom{N}{n_1}$ permutations, or a random sample of them, and calculating the value of the test statistic for each permutation, it is possible to determine the permutation distribution and to calculate the p-values of the test. After each permutation, the number of units with associated category (*j*) is equal to $n_{\cdot(j)}$ (for each *j*) and it does not change but number of these units associated to the first n_1 rows (sample 1) and to the other n_2 (sample 2) can change. For this reason the column marginals and the row marginals of the contingency table remain constant but the frequencies $n_{j(i)}$ and $f_{j(i)}$ may vary.

3 Simulation study

In order to assess the properties of the test we consider a simulation study in which data are generated according to the following model:

$$X \sim 1 + Int\,[\, K \cdot U^{\,\delta}\,]$$

where $\delta \in R$, $U \sim U(0,1)$ and $Int\,[\,\cdot\,]$ denotes the integer part of $[\,\cdot\,]$.

The random variable *X* is therefore discrete and its domain consists of the first *K* positive integers. The situation of maximum heterogeneity can be simulated making $\delta = 1$, because in this case

$$X \sim U(1,2,...,K) \text{ and } f_i = \#\,(X = i) / n \cong 1 / K,$$

where n is the sample size. Increasing δ the distribution of X moves further away from maximum heterogeneity, approaching that of maximum homogeneity where frequencies tend to be concentrated on the first category. The choice of this model, as an alternative to the generation of data hypothesizing some distributions of probabilities for the two populations, mainly depends on the fact that studying the power of a nonparametric test, the variety of proposable alternatives for simulations is so vast that it is almost impossible to consider them all (Lehmann, 1953). In this way we can generate discrete distributions with different degrees of heterogeneity using a single parameter instead of K, as we would if data were generated from a completely specified distribution, such as: p_1, p_2,..,p_k. Generating the data as described, for some sample sizes, some couples of parameters δ_1 and δ_2 for respectively population X_1 and X_2 and some values of nominal significance level, we calculated the rejection rates of four tests in order to evaluate their degree of approximation as well as their power.

Table 3: Rejection rates of four tests of heterogeneity under the null hypothesis (K=8).

δ_1	δ_2	N_1	n_2	alfa nominal			Test
				0.01	**0.05**	**0.10**	
1	1	20	10	0.0020	0.0220	0.0670	T_H
				0.0005	0.0150	0.0510	T_G
				0.0020	0.0310	0.0815	T_{H3}
				0.0045	0.0130	0.0405	$T_{H\infty}$
2	2	20	10	0.0095	0.0485	0.0955	T_H
				0.0085	0.0480	0.0920	T_G
				0.0135	0.0620	0.1105	T_{H3}
				0.0070	0.0265	0.0500	$T_{H\infty}$
		40	30	0.0115	0.0455	0.1000	T_H
				0.0130	0.0515	0.1035	T_G
				0.0180	0.0625	0.1130	T_{H3}
				0.0095	0.0375	0.0685	$T_{H\infty}$
		60	30	0.0110	0.0505	0.0965	T_H
				0.0115	0.0550	0.1015	T_G
				0.0135	0.0590	0.1090	T_{H3}
				0.0075	0.0290	0.0670	$T_{H\infty}$
3	3	20	10	0.0120	0.0545	0.1030	T_H
				0.0120	0.0530	0.0995	T_G
				0.0130	0.0605	0.1165	T_{H3}
				0.0060	0.0265	0.0480	$T_{H\infty}$

Table 4: Rejection rates of four tests of heterogeneity under the null hypothesis (K=16).

δ_1	δ_2	n_1	n_2	alfa nominal			Test
				0.01	**0.05**	**0.10**	
1.5	1.5	20	10	0.0060	0.0510	0.1075	T_H
				0.0030	0.0290	0.0730	T_G
				0.0125	0.0585	0.1185	T_{H3}
				0.0120	0.0285	0.0470	$T_{H\infty}$
		40	30	0.0080	0.0395	0.0810	T_H
				0.0060	0.0340	0.0700	T_G
				0.0095	0.0540	0.0965	T_{H3}
				0.0105	0.0345	0.0780	$T_{H\infty}$
		60	30	0.0070	0.0420	0.0910	T_H
				0.0045	0.0390	0.0840	T_G
				0.0115	0.0560	0.1155	T_{H3}
				0.0080	0.0305	0.0725	$T_{H\infty}$
2	2	20	10	0.0110	0.0635	0.1245	T_H
				0.0080	0.0490	0.0955	T_G
				0.0185	0.0755	0.1340	T_{H3}
				0.0100	0.0355	0.0560	$T_{H\infty}$
		40	30	0.0105	0.0570	0.1100	T_H
				0.0100	0.0525	0.1020	T_G
				0.0155	0.0665	0.1215	T_{H3}
				0.0065	0.0350	0.0730	$T_{H\infty}$
		60	30	0.0085	0.0470	0.1025	T_H
				0.0090	0.0465	0.1070	T_G
				0.0125	0.0605	0.1220	T_{H3}
				0.0060	0.0250	0.0670	$T_{H\infty}$
2.5	2.5	20	10	0.0125	0.0610	0.1200	T_H
				0.0125	0.0505	0.1050	T_G
				0.0200	0.0690	0.1355	T_{H3}
				0.0055	0.0265	0.0500	$T_{H\infty}$
		40	30	0.0145	0.0575	0.1145	T_H
				0.0095	0.0445	0.1120	T_G
				0.0135	0.0545	0.1195	T_{H3}
				0.0050	0.0270	0.0600	$T_{H\infty}$
		60	30	0.0115	0.0605	0.1120	T_H
				0.0120	0.0570	0.1010	T_G
				0.0150	0.0640	0.1160	T_{H3}
				0.0070	0.0320	0.0720	$T_{H\infty}$

Table 3 reports the rejection rates under the null hypothesis for a discrete variable with K=8 categories, for some degrees of heterogeneity with δ_1 and δ_2 ranging from 1 to 3. 2000 dataset were generated each with 2000 permutations in order to approximate the related permutation distribution. Reported results show that, in general, the most conservative test is that based on the index H_∞ of Rényi. For the other three tests the performances are very similar, even though that based on H_3 shows itself to be slightly anticonservative but in any case we can conclude

that the tests are substantially well approximated. In general, by increasing of δ_1, δ_2 and δ_2-δ_1, the rejection rates tend to increase.

Table 4 shows the results of analogous simulations for K = 16 categories. Also in this case, in the presence of maximum heterogeneity, the rejection rates are clearly below the nominal significance levels. Again the two tests based on Rényi's statistics stand out: $T_{H\infty}$ for its lower than nominal rejection rates and $\boldsymbol{T_{H3}}$, vice versa, for its tendency towards rejection rates slightly higher than the nominal levels. In any case T_{H3} behaves not so far away from the tests based on Shannon's and Gini's statistics.

Table 5: Power of nonparametric tests of heterogeneity (K=16 classes).

δ_1	δ_2	n_1	n_2	alfa nominal			Test
				0.01	**0.05**	**0.10**	
2	2.5	90	30	0.0810	0.2050	0.3190	T_H
				0.0830	0.2210	0.3420	T_G
				0.0970	0.2290	0.3660	T_{H3}
				0.0460	0.1600	0.2520	$T_{H\infty}$
2	3			0.2110	0.4360	0.5790	T_H
				0.2375	0.4690	0.6095	T_G
				0.2620	0.4920	0.6270	T_{H3}
				0.1565	0.3780	0.5110	$T_{H\infty}$
2	3.5			0.4180	0.6630	0.7860	T_H
				0.4330	0.6970	0.8090	T_G
				0.4510	0.7090	0.8180	T_{H3}
				0.3320	0.5880	0.7260	$T_{H\infty}$
2	3	60	30	0.1990	0.4170	0.5710	T_H
				0.2070	0.4280	0.5910	T_G
				0.2440	0.4580	0.6130	T_{H3}
				0.1400	0.3290	0.4790	$T_{H\infty}$
2	3.5			0.3540	0.6150	0.7330	T_H
				0.3710	0.6310	0.7540	T_G
				0.3930	0.6470	0.7620	T_{H3}
				0.2690	0.5200	0.6570	$T_{H\infty}$
2	4			0.5480	0.7790	0.8750	T_H
				0.5620	0.8090	0.8910	T_G
				0.5900	0.8240	0.8960	T_{H3}
				0.4500	0.7190	0.8320	$T_{H\infty}$

To evaluate the power of four tests we considered some situations in which the heterogeneity of the two populations were different. In this case 1000 dataset were generated and, for each of these, 1000 permutations. Obviously, the power of the tests increases with the increase in the difference of heterogeneity parameters δ_i, i=1,2 (Table 5). Comparing the performances of the four tests it emerges that the

test $T_{H\infty}$ seems slightly worse than the others, whereas a preference is shown for the test T_{H3}, based on Rényi's entropy index of order 3.

4 An example: University Evaluation

In this section we deal with an application problem in the context of the University Evaluation. The data refer to the type of secondary school (TSS) attended by graduates of the Faculties of Engineering and Economics at the University of Ferrara in 2005 (see Table 6). The problem consists of the comparision of the two faculties from the point of view of TSS attended by graduates. The aspect of interest is the heterogeneity of TSS. High heterogeneity means that the undergraduate degree can be obtained by students coming from a large set of secondary schools. The goal of the study is to answer this question. “Is the heterogeneity of TSS of graduates in Economics greater than that of TSS of graduates in Engineering?”. In order to answer this question we applied a two-sample heterogeneity test for categorical data and one-sided alternative hypothesis.

Table 6: Type of secondary school attended by graduates at the University of Ferrara in 2005 (relative frequencies). Data from Almalaurea (www.almalaurea.it).

Type of Secondary School	**Economics**	**Engineering**
Scientific	38.2	43.3
Technical	48.5	50.2
Humanistic	8.2	4.5
For elementary school teacher	1.7	0.0
Linguistic	1.7	0.0
Vocational	1.7	2.0
Artistic	0.0	0.0
	100	100

The system of hypotheses can be formulated as follows:

$$H_0: Het(\text{Economics}) = Het(\text{Engineering})$$

against

$$H_1: Het(\text{Economics}) > Het(\text{Engineering}).$$

Table 7: Ordered table of the relative frequencies.

	Ordered frequencies		**Cumulative ordered frequencies**	
Ordered Class	**Economics**	**Engineering**	**Economics**	**Engineering**
(1)	48.5	50.2	48.5	50.2
(2)	38.2	43.3	86.7	93.5
(3)	8.2	4.5	94.9	98
(4)	1.7	2.0	96.6	100
(5)	1.7	0.0	98.3	100
(6)	1.7	0.0	100	100
(7)	0.0	0.0	100	100

Looking at the plots in figure 1 we can say that, from a descriptive point of view, TSS of Economics dominates that of Engineering in heterogeneity.

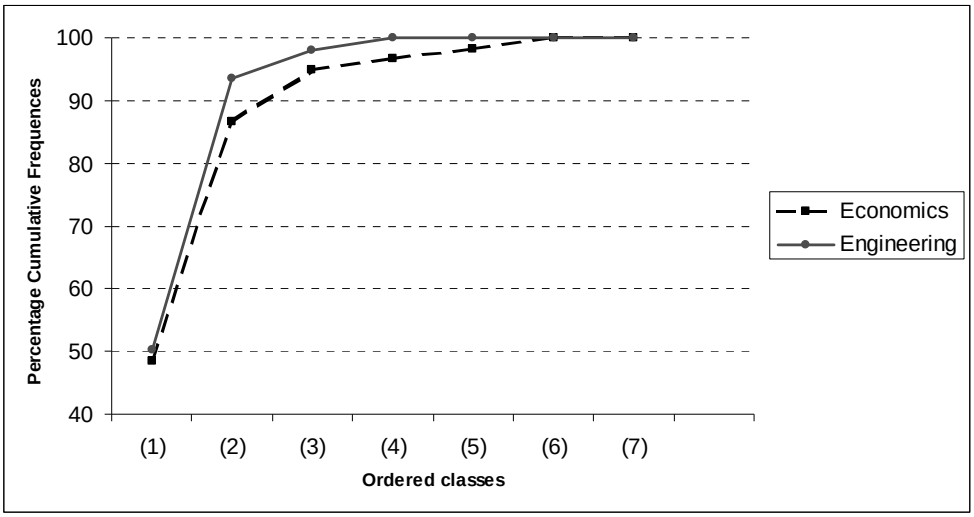

Figure 1: Cumulative relative frequencies for the ordered classes.

Using a simple Monte Carlo random sampling of B = 50.000 permutations from the set of all permutations we obtain the *p-values* reported in Table 8.

Table 8: p-value of the nonparametric tests of heterogeneity.

Test	*p-value*
T_H	0.05924
T_G	0.10362
T_{H3}	0.14158
$T_{H\infty}$	0.38760

5 Concluding remarks

This work presents an inferential nonparametric procedure that allows us for a solution to the problem of hypothesis testing, in which the objective is comparing the heterogeneity of two populations on the basis of sampling data, i.e. to test the hypothesis that the heterogeneity of one population is greater than that of another population.

The proposed test statistic consists of the comparison of the sampling indices of heterogeneity calculated for the two samples and it can vary according to the index of heterogeneity considered. Therefore we propose a general method to solve a peculiar problem. We think that other possible indeces of heterogeneity, i.e. other test statistics, can be adopted within the same framework. We think that the kind of index used is one of the aspect of the solution but not the main problem. In fact from the simulation results we can say that the performances of the four tests are very similar and we cannot conclude that one test is the best. The simulation study allowed us to assess that the proposed nonparametric tests of heterogeneity show high degree of approximation under the null hypothesis and good power behaviour under the alternative. The rejection rates increase with the increase in the homogeneity of distributions. Among the test statistics considered, that based on the index of Rényi of order 3 seems to show higher rejection rates under H_0 but a slightly higher power under H_1.

Moreover the choice of a nonparametric test proves to be both practical and efficient, easy to apply and it requires few and weak nonparametric assumptions.

References

[1] Agresti, A. and Klingenber, G.B. (2005): Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics*.

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

[3] Cohen, A., Kemperman, J.H.B., Madigan, D., and Sakrowitz, H.B. (2000): Effective directed tests for models with ordered categorical data. *Australian and New Zealand Journal of Statistics*, **45**, 285-300.

[4] Gini, C. (1912): Variabilità e mutabilità, in Studi economico-giuridici della Facoltà di Giurisprudenza, Università di Cagliari.

[5] Han, K.E., Catalano, P.J., Senchaudhuri, P., and Metha, C. (2004): Exact analysis of dose-response for multiple correlated binary outcomes. *Biometrics*, **60**, 216-24.

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

[7] Lehmann, E.L. (1953): The power of rank tests, *The Annals of Mathematical Statistics*. **24** (1), 23-43.

[8] Loughin, T.M. (2004): A Systematic Comparison of Methods for Combining P-Values from Independent Tests. *Computational Statistics and Data Analysis*, **47**, 467–485.

[9] Loughin, T.M. and Scherer, P.N. (1998): Testing for Association in Contingency Tables with Multiple Column Responses. *Biometrics*, **54**, 630–637.

[10] Lumley, T. (1996): Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, **52**, 354-361.

[11] Nettleton, D. and Banerjee T. (2000): Testing the equality of distributions of random vectors with categorical components. *Computational Statistics and Data Analysis*, **37**, 195-208.

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

[13] Pesarin F. (2001): *Multivariate Permutation Test With Application to Biostatistics*. Chichester: Wiley.

[14] Pesarn, F. and Salmaso L. (2006): Permutation tests for univariate and multivariate ordered categorical data. *Austrian Journal of Statistics*, **35**, 315-324.

[15] Rényi, A. (1996): Calculus des probabilitès, Dunod, Paris.

[16] Shannon, C.E. (1948): A mathematical theory of communication. *Bell System Technological Journal*, **27**, 623-656.

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

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

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