Multivariate permutation tests in genetics

R. Arboretti Giancristoforo

2003.6

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Aprile 2003

MULTIVARIATE PERMUTATION TESTS IN GENETICS

Rosa Arboretti Giancristofaro

1. INTRODUCTION

The genetic statistical problem we are going to discuss is quite common in any context related to restricted alternatives, or more generally in testing under order constraints (Hirotsu, 1986, 1998; Khoury and Beaty, 1994). In the genetic configuration introduced by Chiano and Clayton (1998), the statistical problem can be formalized in the following way. Let us assume responses are bivariate: (X_1, X_2) and that observed subjects are partitioned into two groups (according to the typical case-control study), so that data may be represented as:

$$\boldsymbol{X} = \left(X_{1ji}, X_{2ji}\right), i = 1,...,n_j, j = 1,2 \qquad (1)$$

where responses are ordered categorical such as (AA, Aa, aa). Of course, in a more general setting we may also consider real valued responses, or any kind of ordered variables, with more than two dimensions and possibly with more than two groups. The ordering relationship on responses is generally induced by the nature of the problem at hand.

The hypotheses we are interested in are:

$$H_0 : \left\{(X_{11}, X_{21}) \overset{d}{=} (X_{12}, X_{22})\right\} = \left\{(X_{11} \overset{d}{=} X_{12}) \cap (X_{21} \overset{d}{=} X_{22})\right\}, \qquad (2)$$

against the special isotonic set of alternatives:

$$H_1 : \left\{ \begin{array}{c} (X_{11} \overset{d}{\geq} X_{12}) \cap (X_{21} \overset{d}{\geq} X_{22}) \\ XOR \\ (X_{11} \overset{d}{\leq} X_{12}) \cap (X_{21} \overset{d}{\leq} X_{22}) \end{array} \right\}, \quad (3)$$

where, in each line, at least one inequality is strong. The *XOR* relation corresponds to an exclusive *OR*, so that, under H_1, one and only one of two bivariate stochastic dominance relations is true.

For convenience of interpretation, it is often useful to introduce a response model such as, for instance: $X_{hji} = \delta_{hj}(\mu_h + Z_{hji})$, where δ_{hj} is the effect on the h-th variable in the j-th group, all other notation having obvious meanings (Di Castelnuovo *et al.*, 2000). In accordance with this model, the hypotheses may be written as: $H_0 : (\delta_1 = 1) \cap (\delta_2 = 1)$ against $H_1 : [(\delta_1 \geq 1) \cap (\delta_2 \geq 1)]$ *XOR* $[(\delta_1 \leq 1) \cap (\delta_2 \leq 1)]$, where at least one inequality in each "sub-alternative" is strong. Since for each sub-hypothesis we have one partial test, in order to obtain a global test we need to combine these partial tests.

In our genetic context, this happens when a gene is associated with a given disease so that, on affected individuals (cases), at least one of the genotype frequencies with putative allele increases *XOR* decreases with respect to non-affected individuals (controls).

Of course, as under the null hypothesis, the pooled data set $\boldsymbol{X}$ is a set of sufficient statistics for the problem, partial tests to take into consideration are:

$$T_h^* = \sum_i X_{h2i}^* - \sum_i X_{h1i}^*, h = 1,2. \quad (4)$$

In the present problem, under H_1, p-values of partial tests are either stochastically smaller than α or stochastically larger than $1 - \alpha$. Thus, if responses are k-dimensional, we need to state the following assumptions:

1) all partial tests $T_h, h = 1,...,k$, are marginally unbiased and significant either for large or small values, so that their permutation distributions under H_1 are either stochastically larger or smaller than under H_0.
2) all partial tests $T_h, h = 1,...,k$, are consistent.

Furthermore, we also need to define the properties of combining functions ψ (Di Castelnuovo *et al.*, 2000) to:

a) a continuous combining function ψ must be monotonically decreasing in each argument: $\psi(...,\lambda_h,...) > \psi(...,\lambda_h',...)$, if $\lambda_h < \lambda_h', h = 1,...,k$;

b) it must attain its supremum positive value $\overline{\psi}$, possibly non finite, when at least one argument attains 0 (zero): $\psi(...,\lambda_h,...) \to \overline{\psi}$ *if* $\lambda_h \to 0$; moreover it must attain its infimum negative value $\underline{\psi}$, possibly non finite, when at least one argument attains 1: $\psi(...,\lambda_h,...) \to \underline{\psi}$ *if* $\lambda_h \to 1$;

c) $\forall \alpha > 0$, its acceptance region is bounded $\underline{\psi} < T''_{\alpha/2} < T'' < T''_{1-\alpha/2} < \overline{\psi}$.

If $1 - \left|2\hat{\lambda}^*_{\psi} - 1\right| \leq \alpha$, then reject H_0 at significance level α.

If the exchangeability property is satisfied under H_0, the nonparametric combination method leads to exact, unbiased and consistent permutation tests (Pesarin, 2001).

An allele A at a gene of interest is said to be associated with the disease if it occurs at a significantly higher or smaller frequency among affected individuals compared with control individuals. For a bi-allelic locus with common allele *a* and rare allele A, individuals may carry zero copies of allele A (subjects with genotype *aa*), one (subjects with genotype Aa) or two (subjects with genotype AA).

Therefore, conventional testing for allelic association implies testing for the joint equality in distribution of genotype frequencies against an alternative of *XOR* dominance of cases with respect to controls by using a proper test statistic. In doing this, it should be taken into consideration that, by referring to genotype-specific risks (Lathrop, 1983) $R_h = f_{h1}/f_{h2}$, $h = AA, Aa, aa$, (where f_{hj}, $j = 1,2$, are respectively the observed frequencies in cases and controls) the effect of an allele can be expressed in only one of the following ways:

1. *Recessive:* there is an effect only in the presence of two copies of A allele (genotype AA), whereas the behaviour in heterozygous condition (genotype Aa) is the same as the reference and most common condition (genotype *aa*), so that: ($R_{AA} > R_{Aa} = R_{aa}$, in presence of a protective effect) *XOR* ($R_{AA} < R_{Aa} = R_{aa}$ for a deleterious effect).
2. *Co-dominant:* there is an ordering on effects associated with the A allele: genotype Aa is of risk (or protection) in comparison with the genotype *aa*, and AA is of risk (or protection) in comparison with the genotype Aa. Obviously, AA is of great risk (or of great protection) in comparison with the genotype *aa*, so that ($R_{AA} > R_{Aa} > R_{aa}$, for a protective effect) *XOR* ($R_{AA} < R_{Aa} < R_{aa}$ for a deleterious effect).
3. *Dominant:* the effect of the A allele is the same in the AA and Aa genotype. In this situation, there is no relative risk (or protection) between AA and

Aa, but only between AA (or Aa) and aa, so that: ($R_{AA} = R_{Aa} > R_{aa}$, protection) XOR ($R_{AA} = R_{Aa} < R_{aa}$, risk).

For these reasons, differences in risk should be tested for over the restricted parameter space, which properly fits the plausible biological models, defined as: $(R_{AA} \geq R_{Aa} \geq R_{aa})\ XOR\ (R_{AA} \leq R_{Aa} \leq R_{aa})$.

Following Chiano and Clayton (1998), in order to reduce the analysis from three to two dimensions, because in a 2×3 contingency table there are only 2 degrees of freedom, we may consider odds ratios of genotype-specific relative risks, which contain all relevant information and are defined as $\theta_{AA} = R_{AA}/R_{Aa}$ and $\theta_{Aa} = R_{Aa}/R_{aa}$, respectively. Thus, the hypotheses under testing may be equivalently expressed as: $H_0 : \theta_{AA} = \theta_{Aa} = 1$, against $H_1 : [(\theta_{AA} \geq 1) \cap (\theta_{Aa} \geq 1)]\, XOR\, [(\theta_{AA} \leq 1) \cap (\theta_{Aa} \leq 1)]$, where at least one inequality in each direction is strong. This system of hypotheses is equivalent to the previous one.

In order to solve the problem within the permutation approach, it should be noted that relation defining the null hypothesis:

$$H_0 : [(\theta_{AA} = 1) \cap (\theta_{Aa} = 1)] \tag{5}$$

is equivalent to:

$$H_0 : \left\{ \left(f_{AA,cases} \cdot f_{Aa,controls} \overset{d}{=} f_{Aa,cases} \cdot f_{AA,controls} \right) \cap \left(f_{Aa,cases} \cdot f_{aa,controls} \overset{d}{=} f_{aa,cases} \cdot f_{Aa,controls} \right) \right\}, \tag{6}$$

which is easier for computations because it is expressed in terms of products of frequencies.

The permutation solution is based on two partial test statistics:

$$\begin{aligned} T_{AA} &= f_{AA,cases} \cdot f_{Aa,controls} / f_{Aa,cases} \cdot f_{AA,controls}, \\ T_{AA} &= f_{Aa,cases} \cdot f_{aa,controls} / f_{aa,cases} \cdot f_{Aa,controls}, \end{aligned} \tag{7}$$

suitable for testing for the following two system of hypotheses:

$$\begin{aligned} &H_{0AA} : [\theta_{AA} = 1] \text{ against } H_{1AA} : [\theta_{AA} > 1 \text{ or } \theta_{AA} < 1] \\ &H_{0Aa} : [\theta_{Aa} = 1] \text{ against } H_{1Aa} : [\theta_{Aa} > 1 \text{ or } \theta_{Aa} < 1] \end{aligned} \tag{8}$$

Note, in fact, that:

$$[\theta_{AA}=1] \Leftrightarrow \left[f_{AA,cases} \cdot f_{Aa,controls} \overset{d}{=} f_{Aa,cases} \cdot f_{AA,controls} \right]. \quad (9)$$

The permutation tests T^*_{AA} and T^*_{Aa} are calculated by using a Conditional Monte Carlo (CMC) procedure (Pesarin, 2001). For example, the estimation of partial *p*-value λ_{AA} is obtained using *B* CMC-iterations:

$$\hat{\lambda}_{AA} = \frac{\#(T^*_{AA} \geq T^{oss}_{AA})}{B}. \quad (10)$$

This partial *p*-value is distributed as *U*(0,1) and leads to reject H_{0AA} if $\hat{\lambda}_{AA} \leq \alpha/2$, or $\hat{\lambda}_{AA} \geq 1-\alpha/2$, at a significance level α. By using the same *B* CMC-iterations, we also estimate:

$$\hat{\lambda}_{AAs} = \frac{\#(T^*_{AA} \geq T^*_{AAs})}{B}, s = 1,\ldots,B. \quad (11)$$

Now, we adapt Liptak's combining function to construct the combined test for the system of hypotheses in (3). The global *p*-value λ_L is estimated by:

$$\hat{\lambda}_L = \frac{\#^B_s\{[\Phi^{-1}(1-\hat{\lambda}_{AAs})+\Phi^{-1}(1-\hat{\lambda}_{Aas})] \geq [\Phi^{-1}(1-\hat{\lambda}_{AA})+\Phi^{-1}(1-\hat{\lambda}_{Aa})]\}}{B}. \quad (12)$$

This *p*-value also follows a distribution U(0,1). Furthermore, if $\hat{\lambda}_L \leq \alpha/2$, we consider the rare allele to be of risk, whereas if $\hat{\lambda}_L \geq 1-\alpha/2$ we consider it to be of protection.

2. EXACT EXPLORATION OF THE PERMUTATION SPACE

We can represent the previous problem by a simple case-control contingency table (table 1).

It should be noted that in all these types of studies, the data may be represented in a contingency table (in this case 3×2) with fixed marginal values. The total cases, M, and the total controls, N, are given and are obtained from experimental observations. At the same time, the number of genotypes AA, in cases and controls together, S_1, is also given, and the same holds also for S_2 and S_3.

TABLE 1

Case-control contingency table for allelic association study

Genotype/ haplotype:	*Cases*	*Controls*	
AA	g_1	l_1	$S_1 = g_1 + l_1$
Aa	g_2	l_2	$S_2 = g_2 + l_2$
Aa	g_3	l_3	$S_3 = g_3 + l_3$
	$M = g_1 + g_2 + g_3$	$N = l_1 + l_2 + l_3$	$S = M + N = S_1 + S_2 + S_3$

With the usual representation of the data file, we have the following structure:

Observation	1	2	3	4	...	M	$M+1$	...	$S = M + N$
Genotype	Aa	AA	AA	aa	...	Aa	Aa	...	AA
Permutation order	u_1	u_2	u_3	u_4	...	u_M	u_{M+1}	...	u_S

Figure 1 - Data representation

where, in the first M observations (or subjects), we have g_1 genotypes AA, g_2 genotypes Aa and g_3 genotypes aa. It does not matter what order we have among the first M subjects (or in the second N subjects), because the contingency table does not change if we take two random permutations into these sub-vectors, and frequencies g_1, g_2, g_3, l_1, l_2 and l_3 remain the same.

Thus, if we consider the overall permutation space associated to data (in the previous paragraph), its cardinality is S! and it may be too large to explore exhaustively.

Les us look at those specific combinations and recombinations of the permutated genotypes/haplotypes in the table which gives the cells a particular structure displayed in table 2.

The marginal sums are identical for any permutation, only the frequencies in the cells may change. The relative data file is illustrated in figure 2, where $\forall i, i' (i \neq i'), u_i^* = u_j^* \text{ and } u_{i'}^* = u_{j'}^*, \text{where } j \neq j', \text{and } i, i', j, j' \in \{1,...,S\}$. Furthermore, in the first M observations (or subjects) we have g_1^* genotypes AA, g_2^* genotypes Aa and g_3^* genotypes aa. Again, the order of the elements in the two sub-vectors (firstly M elements and secondly N elements) is not important.

TABLE 2

A particular result of a permutation in the dataset

	Cases	Controls	
AA	g^*_1	l^*_1	S_1
Aa	g^*_2	l^*_2	S_2
aa	g^*_3	l^*_3	S_3
	M	N	S

We see that there are no S! different results for each permutation, but many permutations give a specific structure of the cells $g^*_1, g^*_2, g^*_3, l^*_1, l^*_2$ and l^*_3.

Observation	1	2	3	4	...	M	$M+1$	...	$S = M+N$
Genotype	*aa*	*Aa*	*Aa*	*AA*	...	*aa*	*Aa*	...	*Aa*
Permutation order	u^*_1	U^*_2	u^*_3	u^*_4	...	u^*_M	u^*_{M+1}	...	u^*_S

Figure 2 - Representation of the permuted data file

Thus, we can construct the exact permutation distribution for the test statistics, associating the related frequencies to the statistics, i.e. the number of times these values of the statistics appear in the S! permutations. In the exploration of the overall permutation space we are looking for the frequencies associated to all possible different configurations in table 2, i.e. all possible sets $\{ g^*_1, g^*_2, g^*_3, l^*_1, l^*_2, l^*_3 \}$ where at least one cell is different from the others.

For data in table 2, we can obtain all possible different table configurations, that are:

1) $g^*_1 \in [\max(0, S_1 - N), \min(M, S_1)]$;
2) $l^*_1 = S_1 - g^*_1$;
3) $g^*_2 \in [\max(S_2 - (N - l^*_1)), \min(M - g^*_1, S_2)]$;
4) $l^*_2 = S_2 - g^*_2$;
5) $g^*_3 = M - g^*_1 - g^*_2$;
6) $l^*_3 = S_3 - g^*_3$.

Then, for a specific set $i\{ {}_i g^*_1, {}_i g^*_2, {}_i g^*_3, {}_i l^*_1, {}_i l^*_2, {}_i l^*_3 \}$ we have the frequency:

$$f^*_i = M!N! \binom{S_1}{{}_i g^*_1} \cdot \binom{S_2}{{}_i g^*_2} \cdot \binom{S_3}{{}_i g^*_3} = \tag{13}$$

$$(M!N!S_1!S_2!S_3!)/({}_i g^*_1!{}_i g^*_2!{}_i g^*_3!{}_i l^*_1!{}_i l^*_2!{}_i l^*_3!);$$

and, of course, the sum of all the frequencies is:

$$\sum_{i}^{I} f_i^* = (M+N)! = (S!); \tag{14}$$

where the total number of all these different configurations is:

$$I = \sum_{g_1^*}^{\min(M,S_1)+1-\max(0,S_1-N)} [\min(M-g_1^*, S_2)+1-\max(0, S_2-(N-(S_1-g_1^*)))]; \tag{15}$$

so that the relative frequencies are $p_i^* = f_i^*/(S!)$.

Of course, the highest relative frequency is associated to the configuration where g_1^* and g_2^* are maximally close (if possible, equal) to l_1^* and l_2^* respectively, which, in general, coincides with the case of no association between cases and controls.

3. EXTENSION OF THE PERMUTATION SOLUTION TO MULTIVARIATE PROBLEMS

In this paragraph we consider an extension of the previous solution to multivariate genetic testing problems (Cappuccio *et al.*, 2000; Cheung and Kumana, 2000; Chowdhury, 2000; Gambaro *et al.*, 2000).

Of course, we may have multiallelic loci such as (A_1, A_2, A_3), where loci A_2 and A_3 can both be rare. In this case we can construct the previous nonparametric tests separately for locus (A_1, A_2) and (A_1, A_3), because the interest is in making comparisons between the rare alleles and the more common ones. We are not interested in knowing the association between two rare alleles (maybe one is of risk and the other of protection or one is neutral and the other of risk, etc.). It is then possible to repeat the previous test for both possible associations: rare1-common, rare2-common (figure 3), where abbreviations CA and CO stands for cases and controls respectively).

The situation is more complicated when the association study involves more than two loci, such as (a,A) and (b,B) where A and B are the rarest alleles.

We suppose interest lies in knowing the specific effect of all the multiple possible configurations (figure 4).

The main topic in this situation is to reconstruct the possible effect that one locus may have, given a specific configuration of the other locus. Then we use six different (3×2) contingency tables, one for each specific configuration (figure 5). Note that in figure 5, for simplicity, cell frequencies are not reported.

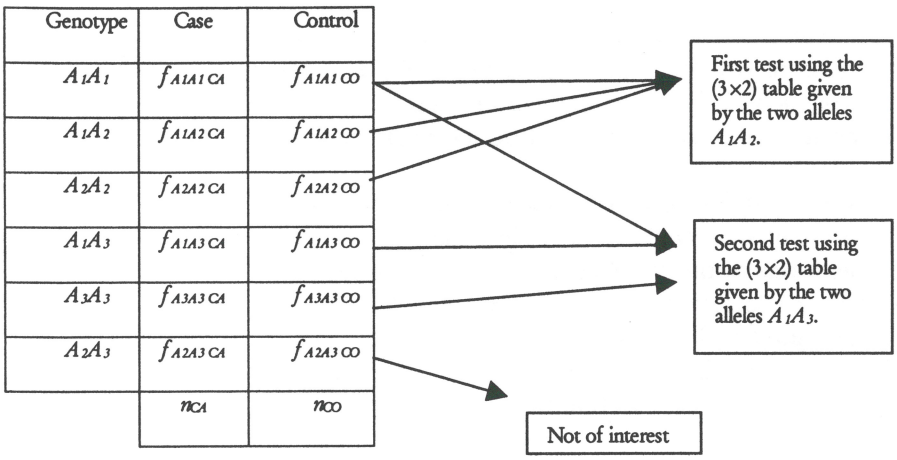

Figure 3 - Multiallelic problem

Genotype	Case	Control
Aa, bb	$f_{aa,bb\,CA}$	$f_{aa,bb\,CO}$
Aa, Bb	$f_{aa,Bb\,CA}$	$f_{aa,Bb\,CO}$
Aa, BB	$f_{aa,BB\,CA}$	$f_{aa,BB\,CO}$
Aa, bb	$f_{Aa,bb\,CA}$	$f_{Aa,bb\,CO}$
Aa, Bb	$f_{Aa,Bb\,CA}$	$f_{Aa,Bb\,CO}$
Aa, BB	$f_{Aa,BB\,CA}$	$f_{Aa,BB\,CO}$
AA, bb	$f_{AA,bb\,CA}$	$f_{AA,bb\,CO}$
AA, Bb	$f_{AA,Bb\,CA}$	$f_{AA,Bb\,CO}$
AA, BB	$f_{AA,BB\,CA}$	$f_{AA,BB\,CO}$
	n_{CA}	n_{CO}

Figure 4 - Multiloci extension

1) *aa*	CA	CO
bb	...	...
Bb	...	...
BB	...	...

2) *Aa*	CA	CO
bb	...	...
Bb	...	...
BB	...	...

3) *AA*	CA	CO
bb	...	...
Bb	...	...
BB	...	...

4) *bb*	CA	CO
aa	...	...
Aa	...	...
AA	...	...

5) *Bb*	CA	CO
aa	...	...
Aa	...	...
AA	...	...

6) *BB*	CA	CO
aa	...	...
Aa	...	...
AA	...	...

Figure 5 - Possible configurations

For example, in the first table, we carry out a permutation test to investigate the association at locus (*b*,*B*) conditional to the genotype *aa* (more common) in the other locus. This procedure is an extension of the bivariate case (presented in section 1). If we consider the case of two loci with three alleles each, then we obtain two contingency tables for each of the six configurations; and if we have more than two loci together, the analysis of each type of association may be very difficult.

In the last situation, before carrying out the specific test for each configuration, it is helpful to carry out an overall test to study if there is any type of significant association in at least one of the configurations (it does not matter, for the moment, if it is of risk or protection). Then, we may suppose that k polymorphic genes are jointly examined and that (with the usual notation) $\{(aa)_r, (Aa)_r, (AA)_r, r = 1,\ldots,k\}$ is the set of related genotypes. In this situation we express the null hypothesis in terms of odd ratios as follows:

$$H_0 : \left\{ \bigcap_{r=1}^{k} [(\theta_{Aar} = 1) \cap (\theta_{aar} = 1)] \right\}, \tag{16}$$

which means that all k genes are jointly irrelevant for discrimination. The alternative of interest may assume two different expressions. The first is

$$H_1 : \left\{ \bigcup_{r=1}^{k} \begin{bmatrix} (\theta_{Aar} \geq 1) \cap (\theta_{aar} \geq 1) \\ XOR \\ (\theta_{Aar} \leq 1) \cap (\theta_{aar} \leq 1) \end{bmatrix} \right\}, \tag{17}$$

where, of course, at least one inequality in each of the $2 \times k$ lines is strict. The interpretation of this alternative is that at least one gene exists which is relevant for discriminating cases with respect to controls. The aim of this alternative is not to know if all genes are of risk (*XOR* protection), but to know if we can admit that some genes may be of risk, some of protection, and the remaining neutral.

In order to solve this specific problem, let us suppose that:

a) data are organized in a unit-by-unit representation: $\left(Y_{jir}, r = 1,\ldots,k, i = 1,\ldots,n_j, j = case, control\right)$, where Y_{jir} is the genotype of the r-th gene on the i-th subject of the j-th group (i.e. Y_{jir} may assume one of the values: aa, Aa, AA);

b) permutations exchange units between groups, so that k-dimensional vectors are exchanged;

c) for each gene r, $r = 1,...,k$, calculate partial tests as
$T^*_{r\,Aa} = f^*_{r\,AA\ cases} \cdot f^*_{r\,Aa\ controls} / f^*_{r\,Aa\ cases} \cdot f^*_{r\,AA\ controls}$ and
$T^*_{r\,aa} = f^*_{r\,Aa\ cases} \cdot f^*_{r\,aa\ controls} / f^*_{r\,aa\ cases} \cdot f^*_{r\,Aa\ controls}, r = 1,...,k$, and all tests are significant for either large or small values;

d) within each gene calculate a second order combined test and related p-value $\hat{\lambda}''_r$, in accordance with the method previously discussed in section 1;

e) according to the nonparametric combination theory (Pesarin, 2001), we combine k second order transformed p-values $1-|2\hat{\lambda}''_r - 1|$ through any combining function ψ to obtain a third order overall combined test and related p-value $\hat{\lambda}'''$;

f) if $\hat{\lambda}''' \le \alpha$, then reject the overall null hypothesis.

A second type of alternative of interest is:

$$H_1' : \left\{ \begin{array}{c} \bigcup_{1\le r\le k} [(\theta_{Aa\,r} \ge 1) \cap (\theta_{aa\,r} \ge 1)] \\ XOR \\ \bigcup_{1\le r\le k} [(\theta_{Aa\,r} \le 1) \cap (\theta_{aa\,r} \le 1)] \end{array} \right\}, \qquad (18)$$

where again at least one inequality in each line is strict. This means that there is at least one gene which is of protection (XOR risk), whereas others are neutral.

Again, in order to solve the problem, we must modify steps e) and f) respectively into:

e') according to the nonparametric combination theory, combine k second order p-values $\hat{\lambda}''_r$ through any suitable combining function to obtain a proper third order overall combined test and related p-value $\hat{\lambda}'''$;

f) if $1-|2\hat{\lambda}''' - 1| \le \alpha$, then reject the overall null hypothesis.

The third order combined tests and their p-values are always obtained by the CMC procedure used for obtaining distributions of partial tests $T^*_{r\,h}$ and p-values $\hat{\lambda}_{r\,h}$ and $\hat{\lambda}''_r$, $h = aa, Aa, AA$, $r = 1,...,k$.

4. POWER AND SAMPLE SIZE SIMULATIONS

We present some simulations for the nonparametric permutation solution illustrated in section 1 by considering different types of population parameters and genetic models. We perform a set of power simulations considering different parameter types (allelic frequency in the population, the three genetic models for the allele effect, several values of the odds ratios) for the permutation solution. The number of simulations is 1000 and the number of CMC-iterations is again 1000.

Simulations are performed by using a single locus with two alleles, one more common and one rare, and the significance level α=0.05. In figures 6-11 we show the power simulations for the permutation solution pointing out the sample size for cases and controls and the frequency of the rare alleles.

As we can observe from the previous figures, the power of the nonparametric solution is very good for small sample sizes as well, and also for a low frequency of rare alleles. Of course, the situation where the rare allele is recessive is the worst, and generally, in this case, the number of subjects is large.

In figure 12, we consider a simulation study for the permutation test with a frequency of 0.10 for the rare allele and significance level α=0.05 for the case where the rare allele is co-dominant and the odds-ratios are equal to 2.

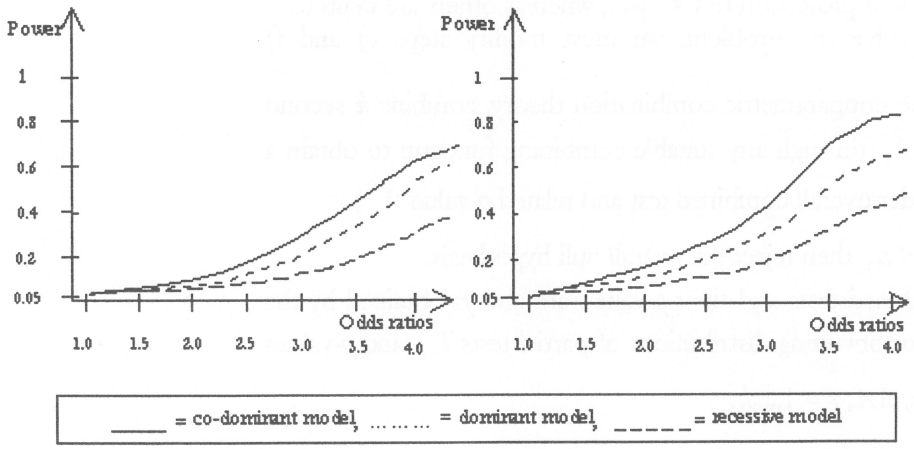

Figure 6 - Cases = controls = 50, f = 0.05

Figure 7 - Cases = controls = 100, f = 0.05

In figure 13, we consider simulations with a frequency of 0.10 for the rare allele and a significance level $\alpha = 0.05$ for the case where the rare allele is dominant and the odds-ratios are equal to 2.

As we can see from the figures, even in these cases, the nonparametric permutation solution has a very good power behaviour.

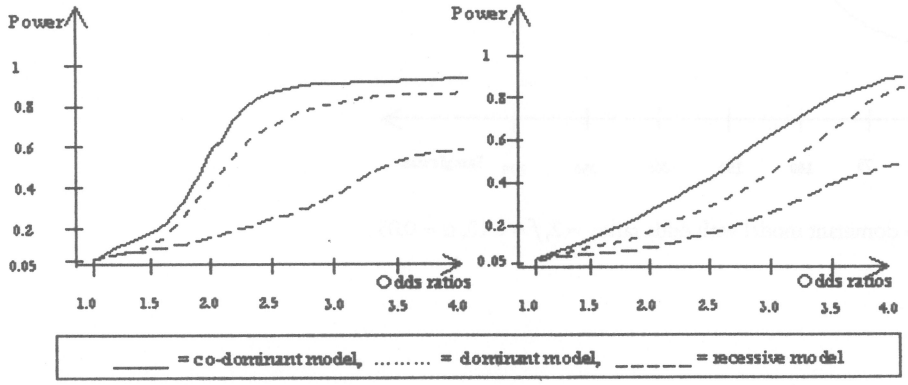

Figure 8 - Cases = controls = 500, $f = 0.05$

Figure 9 - Cases = controls = 50, $f = 0.10$

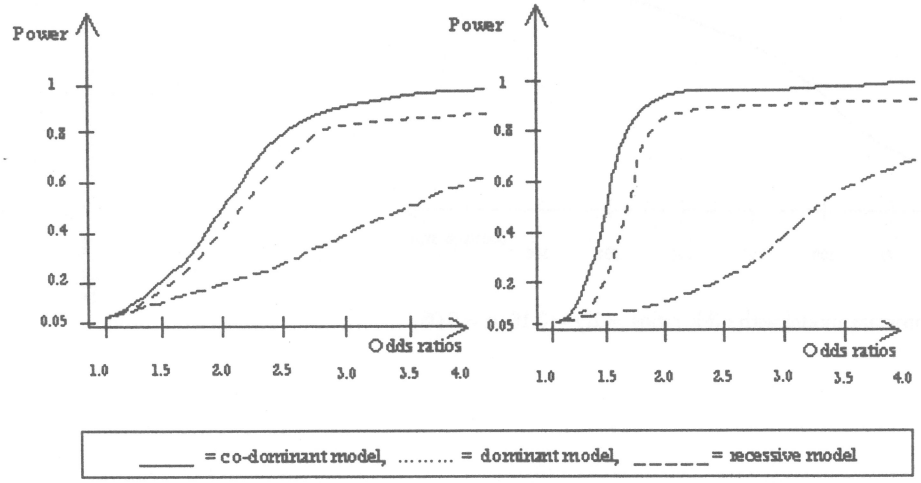

Figure 10 - Cases = controls = 100, $f = 0.10$

Figure 11 - Cases = controls = 500, $f = 0.10$

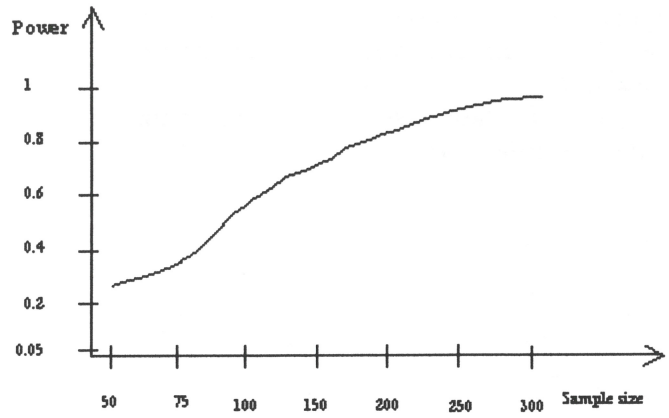

Figure 12 - Co-dominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$

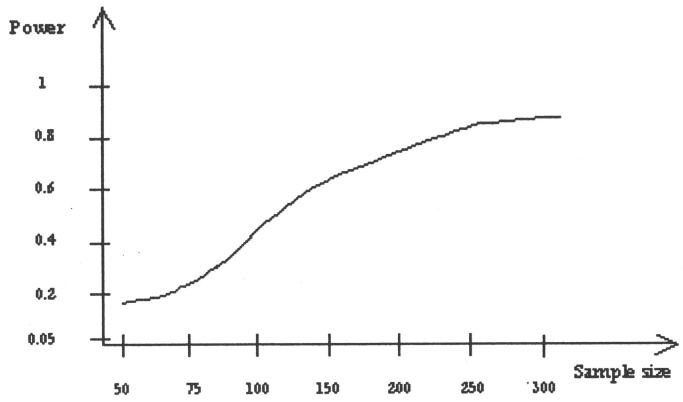

Figure 13 - Dominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$.

5. CONCLUSIONS

We would like to emphasise the role of nonparametric combination as a flexible methodology for solving complex problems.

It should also be stressed that, since permutation tests are conditional with respect to a set of sufficient statistics, the nonparametric combination, under

very mild conditions, frees the researcher from the necessity to model the dependence relations among responses. Furthermore, several Monte Carlo experiments have shown that the unconditional power of the permutation solution has a very good behaviour even for small sample sizes and for a low frequency of rare alleles.

Dipartimento di Scienze Statistiche ROSA ARBORETTI GIANCRISTOFARO
Università di Padova

ACKNOWLEDGEMENT

The author wishes to thank Dr. Dario Mazzaro for helpful comments and suggestions.

REFERENCES

F.P. CAPPUCCIO, G.A. SAGNELLA, G.A. MACGREGOR, (2000), *Association studies of genetic polymorphisms and complex disease (Correspondence)*, "The Lancet", 355, pp. 1278.

B.M.Y. CHEUNG, C.R. KUMANA, (2000), *Association studies of genetic polymorphisms and complex disease (Correspondence)*, "The Lancet", 355, pp. 1277.

M.N. CHIANO, D.G. CLAYTON, (1998), *Genotypic relative risks under ordered restriction*, "Genetic Epidemiology", 15, pp. 135-146.

T.A. CHOWDHURY, (2000), *Association studies of genetic polymorphisms and complex disease (Correspondence)*, "The Lancet", 355, pp. 1277-1278.

A. DI CASTELNUOVO, D. MAZZARO, F. PESARIN, L. SALMASO, (2000), *Test di permutazione multidimensionali in problemi di inferenza isotonica: un'applicazione alla genetica*, "Statistica", LX, 4, pp.691-700.

G. GAMBARO, F. ANGIANI, A. D'ANGELO, (2000), *Association studies of genetic polymorphisms and complex disease (Viewpoint)*, "The Lancet", 355, pp. 308-311.

C. HIROTSU, (1986), *Cumulative chi-squared statistic or a tool for testing goodness of fit*, "Biometrika", 73, pp. 165-173.

C. HIROTSU, (1998), *Isotonic inference*, "Encyclopedia of Biostatistics", Wiley, New York, pp. 2107-2115.

M.J. KHOURY, T.H. BEATY, (1994), *Applications of the case-control method in genetic epidemiology*, "Epidemiologic Reviews", 16, pp. 134-150.

G.M. LATHROP, (1983), *Estimating genotype relative risks*, "Tissue Antigens", 22, pp. 160-166.

F. PESARIN, (2001), *Multivariate permutation tests with applications to biostatistics*, Wiley, Chichester.

SUMMARY

Multivariate permutation tests in genetics

In this paper we provide some new statistical results for hypotheses testing in genetics particularly referred to multivariate allelic association studies. An extensive power simulation study is also provided on permutation solutions.

RIASSUNTO

Test di permutazione multivariati in genetica

In questo lavoro vengono proposte alcune nuove procedure di verifica di ipotesi nell'ambito di problemi in genetica particolarmente riferiti alle analisi multivariate di associazione allelica. Viene inoltre proposto un esteso studio di simulazione per alcune soluzioni di permutazione.