SPRINGER BRIEFS IN STATISTICS

Rosa Arboretti · Livio Corain
Dario Mazzaro · Luigi Salmaso

Permutation Testing for Isotonic Inference on Association Studies in Genetics

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Rosa Arboretti · Livio Corain
Dario Mazzaro · Luigi Salmaso

Permutation Testing for Isotonic Inference on Association Studies in Genetics

Springer

Rosa Arboretti
Department of Territory and Agri-Foresta
University of Padova
Agripolis-viale dell'Universita', 16
35020 Legnaro (PD)
Italy
e-mail: rosa.arboretti@unipd.it

Dario Mazzaro
Generali Group
Assicurazioni Generali s.p.a
Via Marocchesa, 14
31021 Mogliano Veneto (TV)
Italy
e-mail: mazzaro.dario@generali.it

Livio Corain
Department of Management and Engineering
University of Padova
Straoella San Nicola 3
36100 Vicenza
Italy
e-mail: livio.corain@unipd.it

Luigi Salmaso
Department of Management and Engineering
University of Padova
Stradella San Nicola 3
36100 Vicenza
Italy
e-mail: luigi.salmaso@unipd.it

ISSN 2191-544X
ISBN 978-3-642-20583-5
DOI 10.1007/978-3-642-20584-2
Springer Heidelberg Dordrecht London New York

e-ISSN 2191-5458
e-ISBN 978-3-642-20584-2

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

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

Contents

Chapter 1
Introduction

Abstract Isotonic inference concerns situations in which a set of parameters is assumed, a priori, to satisfy certain order restrictions. In the most common case, where populations are arranged in ordered groups, the parameter of interest is assumed to change monotonically with the ordering of the groups. It is then reasonable to take account of the order restrictions in making inferences about the group parameters, such as point or interval estimations or significance tests. Isotonic inference represents a statistical tool of great value in many areas of applied research, especially within the fields of genetic epidemiology and molecular genetics. The purpose of this book is to illustrate a new statistical approach to test allelic association and genotype-specific effects in the genetic study of a disease. We deal with population-based association studies via permutation testing and some likelihood-based tests, but comparisons with other methods will also be performed, analysing advantages and disadvantages of each one, particularly with regard to power properties with small sample sizes. We will focus on case-control analyses to study allelic association between marker, disease-gene and environmental factors. Permutation tests, in particular, will be extended to multivariate and more complex studies where we deal with several genes and several alleles together.

Keywords Genetic epidemiology · Isotonic inference · Linkage analysis

Isotonic inference concerns situations in which a set of parameters is assumed, a priori, to satisfy certain order restrictions. In the most common case, where populations are arranged in ordered groups, the parameter of interest is assumed to change monotonically with the ordering of the groups. It is then reasonable to take account of the order restrictions in making inferences about the group parameters, such as point or interval estimations or significance tests (Hirotsu 2005). Isotonic inference represents a statistical tool of great value in many areas of applied

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_1, © Luigi Salmaso 2011

research, especially within the fields of genetic epidemiology and molecular genetics.

The major task of genetic epidemiology and molecular genetics is to map the diseases loci, that is the identification of genes causing the pathologies. However, the more common inherited disorders are very difficult to study, because a combination of various genes and different environmental factors is often involved. Discovering the major susceptibility locus can be the starting point to advance in understanding the causes of a disease. Furthermore, the primary topic of interest has recently shifted from simple Mendelian diseases, where genotypes of a given gene cause them, to more complex diseases, where genotypes of a given set of genes together with environmental factors merely alter the probability that an individual has the disease, although individual factors are typically insufficient to cause the disease outright. To study these candidate genes and their relations we may use either linkage analysis or allelic association analysis (or linkage disequilibrium analysis).

The goal of linkage analysis is to determine whether two loci segregate independently in meiosis. Alleles of loci on different chromosomes segregate independently of each other during meiosis, as do alleles of loci on opposite ends of the same chromosome. However, when two loci are close together on the same chromosome, their alleles no longer segregate independently but are co-inherited over 50% of the time. We say these loci are linked. The closer the two loci are to each other on the chromosome, the lower the probability of recombination of their alleles. This is probability referred to as the recombination fraction θ. The genetic distance is defined to be infinity between loci on different chromosomes, and for such unlinked loci $\theta = 0.5$. For linked loci on the same chromosome, $\theta < 0.5$, and the genetic distance is an increasing function of θ. The essence of linkage analysis is to estimate the recombination fraction θ and to test whether $\theta = 0.5$.

The terms "linkage disequilibrium" and "allelic association" are sometimes used interchangeably, and sometimes different meanings are assigned to them. The most general definition of either is the condition in which alleles of two loci on a random chromosome from the population do not occur independently of one another. Linkage disequilibrium is sometimes used only when the two loci are tightly linked and not when such correlations exist between unlinked loci, as may occur, for example, as a result of population stratification.

We use the term "linkage disequilibrium" irrespective of whether or not the loci are linked. The term association (without allelic) is also used to refer to the correlation between the alleles of a locus and a phenotype. Here, therefore, we use the terms "allelic association" and "linkage disequilibrium" to refer to the correlation between alleles of two loci on haplotypes. As in linkage analysis, the goal of linkage disequilibrium analysis is to map loci relative to each other and thereby estimate the genomic position of new loci of unknown position using loci of known location. There are many measurements of linkage disequilibrium between two loci (e.g. locus 1 and locus 2), but the most commonly used is the disequilibrium coefficient $D = P_{11} - p_1 q_1$, where P_{11} is the observed frequency of the 1/1 haplotype (generally, the "1" is the most common allele present in that locus), p_1

is the frequency of the "1" allele at locus 1 in the general population and q_1 is the population frequency of the "1" allele at locus 2. The coefficient D ranges from –0.25 (linkage equilibrium) to 0.25 (linkage disequilibrium). It was shown that the rate of decay of linkage disequilibrium is dependent on the distance between loci: $D_t = D_0(1 - \theta)^t$, where t is the current generation number, D_t is the current amount of disequilibrium and D_0 is the disequilibrium at generation 0 (Liu 2010).

Linkage analysis is generally conducted on pedigrees of known structure, whereas linkage disequilibrium analysis is most often conducted on populations, which can be viewed as extremely large pedigrees with many generations of indeterminate structure. Allelic association analysis is used to locate regions of the genome shared by affected individuals more often than by a random sample of individuals from the population—it is hypothesized that affected individuals share their phenotype because they also share some disease-predisposing allele identical by descent from a common ancestor. Thus allelic association analysis is a form of linkage analysis on the largest possible hypothetical pedigree (Terwilliger and Göring 2000).

Risch and Merikangas (1996) argue that the method successfully used (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach by association studies which utilizes candidate genes has far greater power, even if every gene in the genome needs to be tested. Thus, they say that the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

Allelic association studies may be "population-based" or "family-based": the former is essentially performed over the comparison between one sample (cases) of patients and one sample (controls) of unrelated, unaffected individuals (the situation also arises where they are not unrelated); the latter is performed over a set of family units composed, at least, of one affected individual (there are many types of family-based association analysis).

The purpose of this book is to illustrate a new statistical approach to test allelic association and genotype-specific effects in the genetic study of a disease. There are some parametric and non-parametric methods available to this end. We deal with population-based association studies, but comparisons with other methods will also be performed, analysing advantages and disadvantages of each one, particularly with regard to power properties with small sample sizes. In this framework we will work out some nonparametric statistical permutation tests and likelihood-based tests to perform case-control analyses to study allelic association between marker, disease-gene and environmental factors. Permutation tests, in particular, will be extended to multivariate and more complex studies where we deal with several genes and several alleles together. Furthermore, we show simulations under different assumptions on the genetic model and analyse real data sets with the simple study of one locus with the permutation test.

The present book is addressed to practitioners and researchers working in genetics and related fields, particularly biostatisticians, with the aim of introducing them to a particularly promising new approach to deal with complex problems in the genetic association study of a disease. Prerequisites for the reader are a

background in inference and probability theory along with basic knowledge of genetics. We think these arguments could help researchers decide the best testing procedure to use, particularly with regard to complex genetic problems and unusual systems of hypotheses, and especially in the presence of small sample sizes.

References

C. Hirotsu, in *Isotonic Inference*. ed. by T. Colton, Encyclopedia of Biostatistics, (Wiley, New York, 2005)

B.-H. Liu, *Statistical Genomics: Linkage, Mapping, and QTL Analysis* (CRC Press, Boca Raton, 2010)

N. Risch, K. Merikangas, The future of genetic studies of complex human diseases. Science **273**, 1516–1517 (1996)

J.D. Terwilliger, H.H.H. Göring, Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Hum. Biol. **72**, 63–132 (2000)

Chapter 2
Association Studies in Genetics

Abstract This chapter is devoted at first to a quick review of case control studies in genetic epidemiology. The case–control method is usually applied in genetic epidemiology to elucidate the role of genetic factors and their interaction with environmental factors in the aetiology of human diseases. Case–control methodology may not be applicable in all settings and should always be integrated with family studies using genetic analytic techniques such as segregation and linkage methods. Different designs for case control studies are considered. A particular focus is given to the problem of allelic association analysis.

Keywords Allelic association analysis · Case control studies · Transmission disequilibrium test

2.1 Case–Control Studies in Genetics

The case–control method is usually applied in genetic epidemiology to elucidate the role of genetic factors and their interaction with environmental factors in the aetiology of human diseases. The human genome map will make it increasingly feasible to search for disease susceptibility genes using case–control methods in both population and family settings. The interest of epidemiology actually concerns the relation between the environment (in its more general meaning) and the occurrence of human diseases, while the interest of genetics concerns evaluating the effects of population structure and selection forces on the frequency of genetic traits. Therefore, the primary purpose of genetic epidemiology is to study genetic variation in human populations and its relation to normal and pathologic phenotypic variation. Genetic epidemiologists need to evaluate the distribution and determinants of genetic traits in human populations and address the role of genetic

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_2, © Luigi Salmaso 2011

factors and their interaction with environmental factors in the aetiology of human diseases. Both population and family methods are used to achieve this goal.

Case–control methodology may not be applicable in all settings and should always be integrated with family studies using genetic analytic techniques such as segregation and linkage methods.

In population studies, case–control approaches are used (1) to study determinants of human mutations, (2) to evaluate the role of non-specific genetic indicators (such as inbreeding and racial admixture) in the aetiology of diseases, and (3) to assess the role of specific genetic traits in the aetiology of diseases. They are particularly useful in the study of mutations because most mutations are individually rare and their ascertainment involves a combination of clinical and laboratory testing, which makes cohort studies prohibitively expensive.

An important aspect to consider in these studies is the presence of confounding factors. Confounders could be other unmeasured genetic determinants or environmental factors that could produce spurious differences in allele frequencies between cases and controls. Several confounding factors have been recognised in this type of study, making comparisons between investigations difficult (Sher 2000). Firstly, there are phenotypic differences in cases due to different studies giving different definitions of cases and controls, to subliminal differences in enrolled cases because of variations in investigators, clinical skills, and to phenotypic heterogeneity of the disease. Secondly, the genetic background of cases and controls are not identical (specifically in multiracial areas). Thirdly, there are practical difficulties due to the low number of cases, lack of specificity and methodological artefacts.

Different components can make it unclear what the correct methodological and statistical procedure to follow in clinical investigations is (Gambaro et al. 2000). (A) A complex disease is a multistage process where several genetics or environmental events mark each stage and can interact between them. (B) Interindividual variations in response to environmental factors, due to genetic heterogeneity of populations, are present in the data, so a specific polymorphism determines the type of response to exposure to a specific environmental factor. (C) There is a biological relation between the environmental factors, the particular polymorphism and one (or more) of the disease phenotypes. (D) Under special conditions, an allele may no longer be neutral.

If the first two assumptions can be checked by direct data observation, the latter two are not supported by any clear-cut evidence. Sometimes a phenotype is considered to be an indication of the disease when it is not the case. If linkage with the disease or intermediate phenotype of the disease emerges from linkage studies, this finding would strongly support the idea that the candidate gene is in some way involved with the disease. Further to the latter point, before using linkage association studies, it should be demonstrated that an allelic variant of the analysed gene is non-neutral. Clinical investigators must allow for several points in genetic studies (Gambaro et al. 2000). A gene does not have only one allelic polymorphism—there may be more of them. It is dangerous therefore to discard a candidate gene because one of its allelic variants is not found in association with the

disease. If a polymorphism is related to the disease, we cannot be sure that it is the disease allele because it could be in strong linkage disequilibrium with that allele. Because of the so-called founder effect, allelic association studies can be misleading if they are done on different populations. A positive allele-disease association can be found in a specific population and several negative studies in different populations cannot reject the first result. It is very important to rigorously select case–control subjects to guard against possible confounding effects (Gambaro et al. 2000). To minimize confounding in case–control studies, investigators need to carefully select controls from the same racial/ethnic genetic background as that from which cases are derived. Relative controls have been used in an attempt to match for genetic background. Analyses should always be stratified with regard to potentially confounding variables. For example, in Down's syndrome, advanced maternal age is the strictest factor. Therefore, in examining the potential association between a risk factor and the syndrome, possible confounding by that factor should always be considered. As such, in case–control methods the correct study design can address the issue of confounding. In the previous situation, for instance, this could be done by matching cases and controls by maternal age or by stratification in the analysis.

When we wish to evaluate the role of specific genes in the aetiology and pathogenesis of common diseases such as cancer, coronary heart disease, birth defects, etc., we are searching for correlations between specific alleles and diseases, so we need the so-called "association studies" in human genetics. They differ from "linkage studies" which seek to evidence co-segregation between a marker locus and a disease in families. In fact, Greenberg and Doneshka (1996) showed, by computer simulations, that if the disease frequency among persons with the susceptibility allele is <10 times greater than the disease frequency among persons without it, it may be quite difficult to detect linkage even in data sets consisting of 30 nuclear families with two or more affected individuals. Under these conditions, the usual linkage approaches may lack sufficient statistical power to detect linkage or may get false rejection of linkage hypotheses and suffer from the multiple testing problem. This is the primary reason for increasing the usefulness of case–control association methods to look for genetic risk factors.

Case–control evaluation of genetic traits in disease aetiology is generally guided by a "candidate" gene approach, which refers to examining allelic variation (measured either at the protein level or at the DNA level) in loci known or suspected to have some role in the pathogenesis of the disease.

In designing, analysing and interpreting case–control studies of genetic trait-disease associations, it is important to consider several methodological issues. The primary problem is that the causes of many simple and complex diseases are related to confused interactions between genetic susceptibility and environmental factors. Case–control studies provide, in this context, an efficient tool with which to search for genetic susceptibility factors along with environmental exposures. Many patterns of gene–environment interaction have been discussed—additive, multiplicative, etc. The importance of power and sample size considerations must

also be stressed for case–control studies of these gene–environment interactions (Foppa and Spiegelman 1997).

As in case–control studies of epidemiology, how we choose samples of patients and unrelated healthy individuals is very important. If we are studying a most uncommon disease, as cases we probably choose all our patients from the same region, who present the same feature with regard to environmental factors and, if necessary, confounding effects. For control subjects, selection can be tailored to the specific situations: hospital control series, random selection, friends of cases (ensuring the same environmental factors), etc. Initially we may choose more than one control group to observe the different results we obtain, but then we must decide on only one group to provide a suitable study, that is more appropriate for the characteristics of the disease, of case subjects and of the other factors (Wacholder et al. 1992).

One of the major issues in the design of case–control studies is its size. If we are bound by the frequency of the disease with regard to cases, control size is linked to the number of specific environmental and confounding factors that play a role in the influence of the disease, but most of all to the statistical power of the chosen test (Smith and Day 1984).

Gene-disease association studies that fail to examine the role of environmental exposure along with the genetic traits of interest may lead to considerable dilution in measures of association if the genetic factor confers disease susceptibility only in the presence of other genes or environmental determinants. Therefore, in designing case–control studies in genetic epidemiology, environmental risk factors should be examined along with genetic markers of interest as factors which interact with the genetic factor of interest (Gambaro et al. 2000).

In studying associations between genetic traits and disease, indirect methods are often used to assign individuals' genotypes. Such indirect measurement of the underlying genotype can lead to non-differential genotypic misclassifications, and therefore would dilute the magnitude of the relative risks found. Nevertheless, genotypic misclassification can arise when the disease itself interferes with genotypic classification. If genotypes are measured in terms of DNA, misclassification due to linkage disequilibrium can occur. Under ideal conditions, if the gene of interest has been completely sequenced, the presence and location of one or more mutations within the gene could be correlated with an altered gene product and then with case–control status in epidemiologic studies. However, in many of these studies, the researchers only have markers in the general region of the candidate gene or in a non-expressed portion of it (Sunden et al. 1996).

Unless the actual site of a deleterious mutation involved in the disease is targeted, any DNA variation between cases and controls in the region of a candidate gene could reflect DNA variation in linkage disequilibrium with the actual mutation(s) associated with the disease. Linkage disequilibrium can arise when the mutation has occurred relatively recently or if there is selective advantage for specific haplotypes, so that they are preferentially maintained in the population. Under complete equilibrium, there would be no association between any marker allele and the disease susceptibility allele (i.e. the odds ratio

should be 1 in a case–control study). Under linkage disequilibrium, a marker allele may well occur more often with the disease susceptibility allele. However, the association between the marker allele and the disease susceptibility allele may not be perfect, therefore, if the marker allele is used as a proxy for the susceptibility allele to study disease risk, a non-differential misclassification might easily occur. This would dilute the magnitude of the odds ratio between the marker allele and the disease toward the null, and would underestimate the effect of the genetic locus in the aetiology of the disease. One analytical approach with which to address the issue of linkage disequilibrium in case–control studies is to construct specific haplotypes composed of alleles at tightly linked loci within the area of the candidate gene.

Of major statistical importance in these studies is the type-I and II errors. In case–control studies involving many genetic traits and other risk factors, statistically significant associations can be down to chance. These type-I errors will be increasingly important in case–control studies involving multiple DNA markers at many candidate loci. As researchers sequence more genes and DNA polymorphisms are delineated throughout the genome, a major challenge in genetic epidemiology will be to discriminate the biologically meaningful associations from the multitude of spurious ones. The establishment of a cause–effect relation depends on many issues, including consistency of the association across studies and the presence of a biologically meaningful model underlying such associations. Finally, to address issues related to statistical power (type-II errors), investigators must ensure adequate sample sizes in designing case–control studies to search for causal genetic factors, especially to look for evidence of gene–environment interactions (Khoury and Beaty 1994).

2.2 Other Methods

The shortage of genetic case–control association studies is mainly due to the lack of large numbers of patients with a condition of interest, the lack of an adequate control group, and ethnic heterogeneity. Erroneous results may be caused by several confounding factors, which can present themselves individually or together, such as the population stratification (founder effect), multiple hypothesis testing and sub-group analysis. It is important to point out that any positive association in these studies should be reproduced in large cohorts and tested for linkage in family-based studies.

An alternative but related study design is to collect “trios”, consisting of two parents (irrespective of phenotype) and one affected offspring. The case sample consists of the alleles or haplotypes that were transmitted from the parents to an affected child, whereas the control sample consists of those alleles that were not transmitted to the affected child. The key advantage of this so-called haplotype relative risk (HRR) design is that it ensures that case and control samples come from the same genetic population. The statistical methods for analysis of both

study designs are similar for 2-point methods but may differ somewhat in multipoint analysis because the HRR design sometimes provides a means of reconstructing multilocus haplotypes, whereas case–control analysis provides only genotype information (Terwilliger and Göring 2000).

The most used method employing trios and proving to be very powerful in showing both association and linkage is the transmission disequilibrium test (TDT), which requires DNA from an affected patient and their parents, and which examines the transmission of alleles from a heterozygous parent to the affected offspring. A significant difference from the expected Mendelian ratio of 50:50 would suggest that the allele has a role in the susceptibility to the disease in question. The TDT was proposed by Spielman et al. (1993) in response to the problem of spurious associations. This approach takes advantage of population-level associations but is not susceptible to spurious associations that result from stratification. When applied exclusively to trios, the TDT is equivalent to a valid McNemar test of linkage disequilibrium. Risch and Merikangas (1996) recommended allelic association studies as the study design of choice. Allelic association analysis can be powerful when the affected individuals in a sample share the same allele, identical by descent, at the same disease locus from a common ancestor (Terwilliger and Göring 2000).

Although very elegant, the TDT design is usually more labour intensive than a simple case–control design that uses affected individuals and unrelated controls. It may take considerable effort, or may even be impossible, to collect DNA samples from the parents of probands, particularly for late-onset diseases. It may also be difficult to collect DNA from other relatives for which TDT-like statistics have been proposed (Boehnke and Langefeld 1998; Lazzeroni and Lange 1998; Spielman and Ewens 1998). For this reason the simple case–control approach would often be an attractive study design were it not for the problem of spurious associations due to population stratification. Still, Pritchard and Rosenberg (1999) showed that the case–control design can be a valid test for association if we include an explicit test for stratification. If we use only a few marker loci, the possibility that the associations are due to population stratification cannot be eliminated. However, by typing additional unlinked markers, it is possible. The basic idea is that if stratification is present, the unlinked markers must also show association with the phenotype.

Another method, which is very useful with respect to the TDT in several diseases (Schaid and Rowland 1998), is to use the affected sib-pair approach, which has been successfully used in many research works (into for example type-I diabetes). These methods need multi-centre collection of families and large cooperative groups, but they represent the way forward in unravelling the complex genetics of polygenic diseases. Of course, this type of study can prove to be too expensive and very slow, especially if we consider the fact that results can sometimes be negative or, at least, far from what is expected. An objection might be that a study, from which no significant positive result is obtained, from a strictly statistical point of view is not “unsuccessful” or “superfluous”, but rather has as much information and importance as another study that leads to an expected

conclusion, particularly with regard to possible future studies in the same field. Nevertheless, from a more practical point of view, it is undeniable that such an employment of resources and such a strain could be considered "excessive" when facing a possible "negative" conclusion. As such, case–control studies still have a role in hypothesis testing but they must involve large numbers to provide meaningful results (Chowdhury 2000). Association methods, in many cases, have had modest results in the study of genetic polymorphisms and complex diseases, and some authors attribute this fact almost entirely to the incompetence of clinical researchers and their lack of understanding of basic genetic principles (Cheung and Kumana 2000).

In spite of the problems with case–control association studies, such as population heterogeneity, they are appealing because they do not require additional family members for cases, which can be very expensive. Devlin and Roeder (1999) developed a method that has the advantages of both case–control and family-based designs. Their method is for either single nucleotide polymorphism (SNP) association scans or tests of candidate genes. For case–control data they use the genome itself to induce controls similar to family-based studies and to determine what constitutes a significant departure from the null model of linkage disequilibrium. An advantage of dense association genomic scans is that they can detect loci having a small impact on risk to human disorders, while a disadvantage is the large number of false positive occurrences when many significance tests are conducted. Instead of a traditional Bonferroni correction, they proposed a Bayesian outlier test as a means of determining which markers exhibit significant linkage disequilibrium with the disorder, i.e. the outlier test bypasses the usual rigid assumptions required to obtain chi-square distributed random variables in favour of more flexible statistics and weaker assumptions. Another feature of their methodology is that it allows for violations in the usual model assumption, i.e. the independence of observations. When violated, this leads to extra variance in the (parametric) test statistic. Indeed, for case–control studies, affected individuals are more likely to be related than control individuals are because they share a genetic disorder and, ideally, a common genetic basis for the disorder (founder effect). For this reason, simple marker-by-marker hypothesis tests will almost certainly produce false positives, even after a Bonferroni correction. However, their simulations are not so decisive. Furthermore, the situations they considered are quite particular.

In order to briefly anticipate the advantages of the permutation approach in dealing with association studies in genetics, we emphasise that the permutation tests are essentially conditional procedures and this conditioning makes them invariant, under the null hypothesis, with respect to the underlying population distribution, which may be partially or even completely unknown (Pesarin and Salmaso 2010). Consequently, permutation tests are distribution-free and non-parametric. Moreover, as will be seen in this chapter and Chap. 3 the combination-based approach applied to permutation tests, allows us to properly break the testing problem down into a set of simpler sub-problems, each provided with a proper permutation solution.

Table 2.1 Data layout of case–control studies

Genotypes	Cases	Controls
aa	x_1	y_1
aA	x_2	y_2
AA	x_3	y_3

Table 2.2 Allele table with twice sample size

Subjects	Cases	Controls
A	$x_2 + x_3$	$y_2 + y_3$
Other	x_1	y_1

2.3 Case–Control Design Study

Allelic association may be due to pleiotropy, linkage disequilibrium, meiotic drive, selection or population stratification. Talking about association analysis, distinctions are made between model-free methods and model-based methods. The use of case–control studies to detect an association falls into the latter category. Classic case–control studies are important in genetic epidemiology, even if they can only establish an association and other designs are necessary to determine whether such associations are casual. In tabulating such data, a question arises as to whether there are one or two observations per person. One approach classifies individuals according to their genotypes, i.e. according to the pair of alleles that each individual has. Another classifies each allele. Intuitively, one might feel that, provided the alleles are independent, either approach should give a valid analysis, but Sasieni (1997) showed that this is not true.

The data appear as a standard Fisherian 3*2 or 2*2 table for which chi-squared statistics and odds ratios were developed. Table 2.1 presents the number of cases (subjects with the disease) and controls with 0 (negative), 1 (heterozygous) and 2 (homozygous) copies of the rare allele *A*.

Since each heterozygous individual has one copy of *A* and each homozygous has two copies of *A*, we can produce an allele table with twice the sample size (Table 2.2).

Table 2.3 presents the data in terms of the number of subjects with and without the rare allele *A*, ignoring the difference between homozygous and heterozygous genotypes. Such tabulation is common when it is not possible to distinguish heterozygous from homozygous individuals (a situation of perfect dominance or recessiveness).

Traditionally, odds ratios are estimated from case–control data because it is not possible to directly estimate the risk of disease in each exposure group. Instead, one relies on the identity between the ratio of odds of exposure in the diseased to that in the controls and the ratio of the odds of disease in the exposed to that in the unexposed. Provided the disease is rare, the odds ratio will be a close approximation of the relative risk. With genotype data, one can estimate the relative risk of a rare disease associated with the heterozygous genotype and with the

Table 2.3 Data layout of case–control studies ignoring the difference between homozygous and heterozygous genotypes

Allele	Cases	Controls
A	$2x_3 + x_2$	$2y_3 + y_2$
Other	$X_2 + 2x_1$	$y_2 + 2y_1$

Table 2.4 Formulas for the odds ratios to estimate the risk of disease in each exposure group

Table	Odds ratio	Formula
1: *hetero*	θ_{Hetero}	$(x_1 \cdot y_2)/(x_2 \cdot y_1)$
2: *homo*	θ_{Homo}	$(x_2 \cdot y_3)/(x_3 \cdot y_2)$
3: *allele*	θ_{Allele}	$[(2x_3 + x_2) \cdot (y_2 + 2y_1)]/$ $[(x_2 + 2x_1) \cdot (2y_3 + y_2)]$
4: *serological*	θ_{Sero}	$[(x_2 + x_3) \cdot y_1]/[x_1 \cdot (y_2 + y_3)]$

homozygous genotype, or one can combine these two groups (as is done in Table 2.3) and estimate the relative risk associated with the gene. Formulas for these estimators are given in Table 2.4.

The odds ratio from allele data is the relative odds of the allele in cases and in controls. For a rare allele, this is approximately the relative gene frequency in cases and controls. It is not, however, immediately obvious how to translate this odds ratio into a statement about the risk of disease. Whereas one can discuss the risk of disease in an individual with a given genotype, it does not make sense to talk about the risk of an allele getting the disease. The best we can do is to say that the known allele is chosen at random. By contrast, the odds ratio from the serological table does have a reasonable interpretation. For a rare disease, it will give the relative risk of disease for an individual (chosen at random from among all individuals) with at least one copy of the allele. Thus, we need not assume that homozygotes and heterozygotes have the same risk. The serological odds ratio is appropriate whenever we do not have information to distinguish homozygotes from heterozygotes.

There is, however, a special case in which the allelic odds ratio will coincide with the genotypic odds ratio. Suppose that the Hardy–Weinberg (H–W) equilibrium holds in both cases and controls, i.e. the relative proportions of the different genotypes are $p_i^2, 2pi(1-pi), (1-pi)^2$, $i = 1, 2$, where p_1 and p_2 are the allelic frequencies of the most common allele in cases and controls, respectively. Recall that the equilibrium holds under the pair of assumptions of random mating and no selection. The assumption of no selection in cases implies that the gene is not associated with the disease, but the equilibrium could also hold under weaker assumptions. Statistically, the H–W equilibrium simply states that the alleles are independent.

Sasieni (1997) showed that use of the allelic odds ratio and chi-squared statistic is not recommendable, even when it is possible to assume that the effect of different alleles at a given locus are codominant. Indeed, these statistics are not robust

Table 2.5 Contingency table for the problem of allelic association analysis

Marker	Cases	Controls	Total
aa	x_1	y_1	$S_1 = x_1 + y_1$
aA	x_2	y_2	$S_2 = x_2 + y_2$
AA	x_3	y_3	$S_3 = x_3 + y_3$
Total	$M = x_1 + x_2 + x_3$	$N = y_1 + y_2 + y_3$	$S = M + N = S_1 + S_2 + S_3$

against departures from the assumptions of the H–W equilibrium in controls and codominance between the alleles.

2.4 The Problem of Allelic Association Analysis

Suppose we have two random samples, one of M cases (individuals with disease) and one of N controls (without the disease), where each person is classified as having a particular marker allele (*a* the more common, *A* the rarer). x_1, x_2 and x_3 indicate the numbers of affected individuals who carry, respectively, zero, one or two copies of the rare allele, while y_1, y_2 and y_3 indicate the corresponding control subjects. In this way we obtain the 3×2 contingence table shown in Table 2.5.

Therefore, the odds ratio

$$\mathrm{P}(\text{disease}|aa)/\mathrm{P}(\text{no disease}|aa)/\mathrm{P}(\text{disease}|aA)/\mathrm{P}(\text{no disease}|aA),$$

and

$$\mathrm{P}(\text{disease}|aA)/\mathrm{P}(\text{no disease}|aA)/\mathrm{P}(\text{disease}|AA)/\mathrm{P}(\text{no disease}|AA),$$

or, equivalently,

$$\mathrm{P}(aa|\text{disease})/\mathrm{P}(aA|\text{disease})/\mathrm{P}(aa|\text{no disease})/\mathrm{P}(aA|\text{no disease}),$$

and

$$\mathrm{P}(aA|\text{disease})/\mathrm{P}(AA|\text{disease})/\mathrm{P}(aA|\text{no disease})/\mathrm{P}(AA|\text{no disease}),$$

are consistently estimated, respectively, by $\theta_{aA} = (x_1 \cdot y_2)/(x_2 \cdot y_1)$ and $\theta_{AA} = (x_2 \cdot y_3)/(x_3 \cdot y_2)$. Significance of the deviation of these ratios from 1 can be tested using the usual chi-square statistic with 1 degree of freedom or, for small samples, by the exact Fisher test. If the controls are obtained from a random sample of the population, rather than a sample of persons without the disease, then θ_{aA} and θ_{AA} are consistent estimations of the more meaningful *relative risks* (Elston 1998).

In genetic epidemiology of diseases of complex etiology, association studies are useful for investigating candidate disease genes. Association studies are case–control population-based studies on a comparison of unrelated affected and unaffected individuals. An allele *A* at a gene of interest is said to be associated with

the disease if it occurs at a significantly higher frequency among affected individuals compared to the control. For a bi-allelic locus with common allele *a* and rare allele *A*, individuals may carry none (subjects with genotype *aa*), one (subjects with genotype *aA*) or double (subjects with genotype *AA*) copies of the *A* allele. Conventionally, therefore, a test for allelic association is a test for the distribution of case/control genotypes using the likelihood ratio chi-square statistic (asymptotically distributed as χ^2 with 2 df) or Fisher's exact test.

However, testing only for overall effects of a gene rather than genotype-specific effects may be less powerful. For example, in a case–control study on the effect of R353Q genetic variants of factor VII (a plasma protein involved in blood coagulation) on myocardial infarction, Lacoviello et al. (1998) demonstrated great protection against myocardial infarction due to the rare genotype QQ (found in 5% of controls but in only 0.6% of cases) but only a small difference in the distribution of the RQ genotype (see Chap. 6).

It is therefore necessary to test for genotype-specific risks. However, this approach requires attention as not all models are necessarily biologically plausible, so 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 allele *A* (genotype *AA*), while the heterozygous condition (genotype *Aa*) is the same as the reference and most common condition (genotype *aa*).
2. Codominant—there is an additive effect of the *A* allele: genotype *Aa* is of risk (or of protection) in comparison with genotype *aa*, and *AA* is of risk (or of protection) in comparison with genotype *Aa*. Obviously, *AA* is of great risk (or of great protection) in comparison with genotype *aa*.
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*.

For these reasons, differences in the risk should be tested for while maximizing over the restricted parameter space that corresponds to plausible biological models: ($R_{AA} \geq R_{Aa} \geq \mathrm{R_{aa}}$) or ($R_{AA} \leq R_{Aa} \leq R_{aa}$), where R_g are the genotype-specific risks.

In case–control studies it is easy to obtain genotype-specific relative risk from odds ratios: $\theta_{AA} = R_{AA}/R_{Aa}$ and $\theta_{Aa} = R_{Aa}/R_{aa}$ (of course $\theta_{aa} = R_{aa}/R_{aa} = 1$), and the null and alternative hypotheses become:

$$H_0 : \theta_{AA} = \theta_{Aa} = 1$$
$$H_1 : \{(\theta_{AA} \geq 1) \cap (\theta_{Aa} \geq 1)\} XOR \{(\theta_{AA} \leq 1) \cap (\theta_{Aa} \leq 1)\},$$

where at least one inequality is strong.

This particular system of hypotheses was proposed for the first time by Chiano and Clayton (1998). From a statistical point of view, the alternative hypothesis is isotonic, i.e. the variables are ordered in one sense, with a further complication caused by the "*XOR*", which is an exclusive "or". This approach makes it possible to study genetic diseases for which we do not know the relative effect of the

putative allele (dominant, recessive or codominant), or if we are studying a related genetic polymorphism that may be protective or deleterious with respect to the disease.

In the following chapter we deal with this particular statistical problem using different approaches.

References

M. Boehnke, C.D. Langefeld, Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am. J. Hum. Genet. **62**, 950–961 (1998)

B.M.Y. Cheung, C.R. Kumana, Association studies of genetic polymorphisms and complex disease (correspondence). Lancet **355**, 1277 (2000)

M.N. Chiano, D.G. Clayton, Genotypic relative risks under ordered restriction. Genet. Epidemiol. **15**, 135–146 (1998)

T.A. Chowdhury, Association studies of genetic polymorphisms and complex disease (correspondence). Lancet **355**, 1277–1278 (2000)

B. Devlin, K. Roeder, Genomic control for association studies. Biometrics **55**, 997–1004 (1999)

R.C. Elston, Linkage and association. Genet. Epidemiol. **15**, 565–576 (1998)

I. Foppa, D. Spiegelman, Power and sample size calculations for case-control studies of gene-environmental interactions with a polytomous exposure variable. Am. J. Epidemiol. **146**, 596–604 (1997)

G. Gambaro, F. Angiani, A. D'Angelo, Association studies of genetic polymorphisms and complex disease (Viewpoint). Lancet **355**, 308–311 (2000)

D.A. Greenberg, P. Doneshka, Partitioned association-linkage test: distinguishing ″necessary″ from ″susceptibility″ loci. Genet. Epidemiol. **13**(3), 243–25 (1996)

L. Iacoviello, A. Di Castelnuovo, P. De Knijff, A. D'Orazio, C. Amore, R. Arboretti, C. Kluft, B.M. Donati, Polymorphisms in the coagulation factor VII gene and the risk of myocardial infarction. N. Engl. J. Med. **338**, 79–85 (1998)

M.J. Khoury, T.H. Beaty, Applications of the case-control method in genetic epidemiology. Epidemiol. Rev. **16**, 134–150 (1994)

L.C. Lazzeroni, K. Lange, A conditional inference framework for extending the transmission/disequilibrium test. Hum. Hered. **48**, 67–81 (1998)

F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software* (Wiley, Chichester, 2010)

J.K. Pritchard, N.A. Rosenberg, Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. **65**, 220–228 (1999)

N. Risch, K. Merikangas, The future of genetic studies of complex human diseases. Science **273**, 1516–1517 (1996)

P.D. Sasieni, From genotypes to genes: doubling the sample size. Biometrics **53**, 1253–1261 (1997)

J.D. Schaid, C. Rowland, Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am. J. Hum. Genet. **61**, 1492–1506 (1998)

L. Sher, Psychiatric diagnoses and inconsistent results of association studies in behavioral genetics. Med. Hypotheses **54**(2), 207–209 (2000)

P.G. Smith, N.E. Day, The design of case-control studies: the influence of confounding and interaction effects. Int. J. Epidemiol. **13**, 356–365 (1984)

R.S. Spielman, W.J. Ewens, A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62**, 450–458 (1998)

R.S. Spielman, R.E. McGinnis, W.J. Ewens, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **52**, 506–513 (1993)

S.L.F. Sunden, W.L.M. Alward, B.E. Nichols, T.R. Rokhlina, A. Nystuen, E.M. Stone, V.C. Sheffield, Fine mapping of the autosomal dominant juvenile open angle glaucoma (GLC1A) region and evaluation of candidate genes. Genome Res **6**, 862–869 (1996)

J.D. Terwilliger, H.H.H. Göring, Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Hum. Biol. **72**, 63–132 (2000)

S. Wacholder, D.T. Silverman, J.K. McLaughlin, J.S. Mandel, Selection of controls in case-control studies: II. Types of controls. Am. J. Epidemiol. **135**, 1029–1041 (1992)

Chapter 3
The Nonparametric Permutation Methodology

Abstract This chapter concerns an overview to multivariate permutation tests by nonparametric combination of dependent permutation tests. Basic theory and properties of such method are illustrated along with the most commonly used combination functions: Fisher, Tippett and Liptak.

Keywords Conditional Monte Carlo procedure · Nonparametric combination · Permutation tests

3.1 Basic Concepts of the Theory of Permutation Tests

Let us start by introducing terminology, definition and general theory of permutation tests. Permutation tests are essentially conditional procedures in which conditioning is made with respect to the permutation sample space associated with the whole data set, which is a set of sufficient statistics under the null hypothesis. It has been shown (Pesarin 2001; Pesarin and Salmaso 2010) that this conditioning makes permutation tests invariant, under the null hypothesis, with respect to the underlying population distribution, which may be partially or even completely unknown. Consequently, permutation tests are distribution-free and nonparametric.

Let X denote a response random variable whose values are points of the sample space χ. The probability distribution P on χ, associated with a symbolic random experiment characterizing X, is defined on an additive class B of subsets of χ. Sometimes, associated with P and with respect to a dominating measure ξ, we may refer to the density f of χ. Here, χ is a one-dimensional Euclidean space and B is a family of Borel sets. A random sample from X is a random experiment whose result is a sample point $X^n = \{X_1, \ldots, X_n\}$. Given a sample X of n i.i.d. observations from X, we wish to test the null hypothesis H_0 that the unknown probability distribution P on (χ, B) generating X belongs to a certain class P_0, against the

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_3, © Luigi Salmaso 2011

alternative class P_1. To be precise, we use H_0: $\{P \in P_0\}$ to denote the null hypothesis and $H_1 : \{P \in P_1\}$ the alternative, where of course $P_1 = P - P_0$. The sample point X takes values on the sample space χ^n. The most common situation is that P_0 contains only one element, and the null hypothesis in this case is said to be simple, otherwise it is said to be composite.

Let P^n indicate the probability distribution induced on χ^n by the sampling experiment. Associated with any sample point X is the orbit $(\chi^n|X)$, also called conditional sample space, containing all points of χ^n which are equivalent to the given sample point X with respect to a group of transformations characterized by suitable invariance properties. The invariance properties in question are that conditional distribution $P^n_{|X}$ on points of the conditional sample space $(\chi^n|\mathrm{X})$ is not dependent on population distribution $P\ \forall P \in \mathrm{P}_0$ (Pesarin and Salmaso 2010).

Definition Any test statistic T: $\chi^n \rightarrow \Re^1$, whose conditional c.d.f. $F_T(t|\mathbf{X})$ is induced by $P^n_{|\mathbf{X}}$ and is invariant under H_0 on $(\chi^n|\mathbf{X})$, is said to be an invariant test for testing H_0 against H_1.

However, as for any given testing problem that we may condition with respect to different sets of sufficient statistics, we may also take various groups of invariant transformations into consideration. From this point of view, on the one hand we should condition with respect to a minimal set of sufficient statistics, on the other we should take a group of maximal invariant transformations into consideration (Pesarin and Salmaso 2010).

It is important to consider that:

1. The conditional sample space $(\chi|\mathbf{X})$ always has a finite number of points, provided that sample size n is finite.
2. K denotes the cardinality of $(\chi|\mathbf{X}) : K = \#\{X^* \in (\chi|\mathbf{X})\}$, where # means the number of points satisfying condition (.);
3. On $(\chi|\mathbf{X})$ we may define an algebra of events $(B|\mathbf{X})$ containing all sub-sets of interest, so that $\{(\chi|\mathbf{X}), (B|\mathbf{X})\}$ is a conditional measurable space.
4. For every event $A \in (B|\mathbf{X})$ we have that $\Pr\{A|\mathbf{X}\} = \int_A dP_{|\mathbf{X}}$.

Another important concept is the permutation equivalence of two statistics:

Definition Two statistics T_1 and T_2, both mapping χ into $\Re^1$, are said to be permutationally equivalent when, for all points $\mathbf{X} \in \chi$ and $\mathbf{X}^* \in (\chi|\mathbf{X})$, the relationship $\{T_1(\mathbf{X}^*) \leq T_1(\mathbf{X})\}$ is true if and only if $\{T_2(\mathbf{X}^*) \leq T_2(\mathbf{X})\}$ is true, where $\mathbf{X}^*$ indicates any permutation of $\mathbf{X}$ and $(\chi|\mathbf{X})$ is the conditional sample space.

Formally, the randomized version of the permutation test Φ_R associated with $(T|\mathbf{X})$ is defined as

$$\Phi_R = \begin{cases} 1 \text{ if } T_{ob} > T_\alpha \\ \gamma \text{ if } T_{ob} = T_\alpha \\ 0 \text{ if } T_{ob} < T_\alpha \end{cases}$$

Table 3.1 Representation of the CMC-procedure

$\mathbf{X}$	$\mathbf{X}_1^*$	...	$\mathbf{X}_r^*$	...	$\mathbf{X}_B^*$
T_{ob}	T_1^*	...	T_r^*	...	T_B^*

where α is the significance level of test T, T_{ob} is the observed value of the statistic test, T_α is the critical value of the statistic and $\gamma = [\alpha - \Pr\{T_{ob} > T_\alpha|\mathbf{X}\}]/\Pr\{T^* = T_{ob}|\mathbf{X}\}$.

3.2 Sampling Inspection of Permutation Space

We first observe that, under H_0 and due to the assumed exchangeability of data with respect to symbolic treatment levels, all points of the conditional sample space ($\chi|\mathbf{X}$) are equally likely. Therefore, one way of inspecting ($\chi|\mathbf{X}$) is by means of a Monte Carlo simulation. Among the different Monte Carlo techniques, the simplest is simple random sampling.

Without loss of generality, hereafter we assume that univariate permutation test statistics T of interest are significant for large values. The permutation distribution of any test statistic T is denoted by the notation $F_T(z|\mathbf{X})$, $\forall z \in \Re^1$. A general simulation procedure for estimating the c.d.f. $F(z|\mathbf{X})$ and the associated p-value λ induced by a statistic T applied on data set $\mathbf{X}$ is described in the following steps:

1. Calculate the observed value of T:$T_{ob} = T(\mathbf{X})$.
2. Consider a data permutation $\mathbf{X}^*$ randomly selected from ($\chi|\mathbf{X}$), where all points of ($\chi|\mathbf{X}$) are equally likely, and consider the value of test statistic T on $\mathbf{X}^*$: $T^* = T(\mathbf{X}^*)$.
3. Independently repeat step 2 B times; the set of CMC-Iteration results $\{T^*, i = 1, \ldots, B\}$ is thus a random sample from the permutation distribution of T.
4. The E.D.F. $\hat{F}_B^*(z) = \sum_{i=1}^{B} I\left(T_r^* \leq z\right)/B$, $\forall z \in \Re^1$, where $I(\cdot) = 1$ if relation $(\cdot)$ is true and 0 otherwise, is a consistent estimate of the permutation distribution $F(z|\mathbf{X})$ of T; moreover, $\hat{\lambda} = \sum_{r=1}^{B} I\left(T_r^* \geq T_{ob}\right)/B$ is an unbiased and consistent estimate of the permutation p-value $\lambda = \Pr\{T^* \geq T_{ob}|\mathbf{X}\}$.
5. If, for any fixed significance level α, the result is $\hat{\lambda} < \alpha$, then reject H_0.

Table 3.1 summarizes the conditional Monte Carlo procedure (CMC-Procedure).

In statistics, the researcher usually works with complex problems that involve hypothesis systems which can be decomposed into several sub-problems with simple systems of hypotheses. The global null hypothesis can be represented as the intersection of all partial null hypotheses, while the alternative global hypothesis corresponds to the union of all partial alternative hypotheses. In these situations the use of nonparametric combination methodology for dependent tests proves to be very useful and efficient in obtaining a good solution.

3.3 The Nonparametric Combination of Dependent Tests

Let us consider a set of generic partial tests $\{T_i,\ i = 1,\ \ldots,\ k\}$, assuming that the following assumptions are satisfied.

1. All permutation partial tests T_i must be marginally unbiased and significant for large values, so that they are stochastically larger under H_1 than under H_0.

Formally, these assumptions mean that $\Pr\{T_i \geq T_{i\alpha}|\mathbf{X}, H_{1i}\} \geq \alpha, \forall\ \alpha > 0,\ i = 1, \ldots, k,$ and $\Pr\{T_i \leq z|\mathbf{X}, H_{0i}\} = Pr\{T_i \leq z|\mathbf{X}, H_{0i} \cap H_i^+\} \geq \Pr\{T_i \leq z|\mathbf{X}, H_{1i}\} = \Pr\{T_i \leq z|\mathbf{X}, H_{1i} \cap H_i^+\},\ i = 1, \ldots, k,\ \forall z \in \Re^1$, where irrelevance with respect to the complementary set of hypotheses $H_i^+\left\{\cup_{j\neq i}\left(H_{0j} \cup H_{1j}\right)\right\}$ means that it does not matter which among H_{0j} and H_{1j}, $j \neq i$, is true when testing for the i-th sub-hypothesis.

2. Partial tests T_i must be consistent, i.e. $\Pr\{T_i \geq T_{i\alpha}|H_{1i}\} \to 1,\ \forall\alpha > 0,$ $i = 1, \ldots, k$, as n tends to infinity, where $T_{i\alpha}$, assumed to be finite, is the marginal critical value T_i.

These assumptions, especially the former, imply that the set of p-values $\lambda_1, \ldots,$ λ_k, associated with the partial test statistics in $\boldsymbol{T}$, are positively dependent under the alternative, and this is irrespective of dependence relations among component variables in $\boldsymbol{X}$. They also imply that partial tests T_i, $i = 1, \ldots, k$, must be considered in such a way that their permutation distributions are monotonically related to underlying entities not implied by sub-hypotheses H_{0i} or H_{1i}, but possibly implied by H_{0j} or H_{1j}, for some $j \neq i$. In practice, when each partial test is related to a different component variable, as for instance is usual in much multidimensional testing on locations, this property is easily satisfied, provided that each partial test T_i is unbiased for the proper sub-hypothesis H_{0i} against H_{1i}, $i = 1, \ldots, k$.

Sometimes positive dependence or marginal unbiasedness is only approximately satisfied. One important example is when H_{01} and H_{02}, are, respectively, related to locations and scale coefficients in a testing problem where symbolic treatment may influence both. Therefore, for the positive dependence and marginal unbiasedness properties to be satisfied, on the one hand T_1 must be unbiased for H_{01} against H_{11}, irrespective of whether H_{02} is true or not; on the other T_2 must be unbiased for H_{02} against H_{12}, irrespective of whether H_{01} is true or not.

For the sake of simplicity and uniformity of analysis, but without loss of generality, we only refer to combining functions applied to p-values associated with partial tests. Because of assumption 1, partial tests are permutationally equivalent to their p-values: $T_i \approx \Pr\{T_i \geq T_o|\mathbf{X}\} = \lambda_i,\ i = 1, \ldots, k$. Of course, this is a direct consequence of the monotonic non-increasing behaviour with respect to t of significance level functions $L_i(t) = \Pr\left\{T_i^* \geq t|\mathbf{X}\right\}$. Thus, the nonparametric combination in a single second-order test $T'' = \psi(\lambda_1, \ldots, \lambda_k)$ is achieved by a continuous, non-increasing, univariate and non-degenerate real function $\psi : (0, 1) \to \Re^1$. Of course, ψ satisfies the measurability property as every other function does in the permutation context. In order to be suitable for test

combination, all combining functions ψ must satisfy at least the following reasonable properties:

a. Function ψ must be non-increasing in each argument: $\psi(\ldots, \lambda_i, \ldots) \geq \psi(\ldots, \lambda_i', \ldots)$ if $\lambda_i < \lambda_i', i \in \{1, \ldots, k\}$;
b. Every combining function ψ must attain its supremum value $\bar{\psi}$, possibly non finite, when at least one argument attains the zero: $\psi(\ldots, \lambda_i, \ldots) \to \bar{\psi}$ if $\lambda_i \to 0$;
c. $\forall\, \alpha > 0$, the critical value of every ψ is assumed to be finite and strictly smaller than the supremum value: $T''_\alpha < \bar{\psi}$.

These properties of combining functions are quite reasonable and intuitive, and are generally easy to justify. Property *a* is related to the unbiasedness of combined tests; *b* and *c* are related to consistency. Furthermore, these properties define a class *C* of combining functions containing the well-known combining functions of Fisher, Lancaster, Liptak and Tippett. Class *C* also contains the Mahalanobis quadratic form for invariance testing against alternatives lying at the same quadratic distance from H_0.

C contains a class of admissible combining functions of independent tests characterized by convex acceptance regions, when these are expressed in terms of *p*-value λ's. In particular, *C* includes all combining functions which nonparametrically take account of the underlying dependence structure among *p*-values λ_i, $i = 1, \ldots, k$.

Thus, one problem naturally arises: how to choose, for any given testing problem, the best combining function in class *C*. This seems to be very difficult and we believe it is unsolvable in the case of finite sample sizes and without any further restriction. At the moment, only "asymptotic optimal combinations" may sometimes be obtained. Moreover, if $D_i = \gamma_i(T_i)$, $i = 1, \ldots, k$, where γ_i are continuous monotonically increasing transformations of partial tests, then $\forall \psi \in C$, $T''_D = \psi(\lambda_{D1}, \ldots, \lambda_{Dk})$ is permutationally equivalent to $\psi(\lambda_1, \ldots, \lambda_k) = T''$, because the *p*-values are invariant under continuous monotonic increasing transformations of test statistics. Note that if partial tests are all exact permutation tests, then for every combining function $\psi \in C$, the combined test T''_ψ is an exact permutation test.

Consider a two-phase algorithm for the nonparametric combination. This algorithm is used to obtain a Monte Carlo estimate of the permutation distribution of a combined test. The first phase concerns the estimate of the *k*-variate distribution of *T*, and the second obtains the estimate of permutation distribution of combined test T''_ψ by using the same simulation results as in the first phase. Note that when it is clear from the context which combining function ψ has been adopted in place of T''_ψ, we simply use T''.

Phase I. An algorithm which simulates the first phase of a procedure estimating the *k*-variate distribution of $\boldsymbol{T}$ should include the following steps:

1. Calculate the vector of the observed values of tests $\boldsymbol{T}$: $\boldsymbol{T}_o = \boldsymbol{T}(\mathbf{X})$.

Table 3.2 Representation of a multivariate data permutation

$X_1(1)$	…	$X_1(n_1)$	$X_1(1+n_1)$	…	$X_1(n)$	→	T_{o1}
…	…	…	…	…	…		…
$X_k(1)$	…	$X_k(n_1)$	$X_k(1+n_1)$	…	$X_k(n)$		T_{ok}

$X_1(u^*_1)$	…	$X_1(u^*_{n1})$	$X_1(u^*_{1+n1})$	…	$X_1(u^*_n)$	→	T^*_1
…	…	…	…	…	…		…
$X_k(u^*_1)$	…	$X_k(u^*_{n1})$	$X_k(u^*_{1+n1})$	…	$X_k(u^*_n)$		T^*_k

Table 3.3 Representation of the CMC-procedure

$\boldsymbol{X}$	$\boldsymbol{X}^*_1$	⋯	$\boldsymbol{X}^*_1$	⋯	$\boldsymbol{X}^*_B$
T_{o1}	T^*_{11}	⋯	T^*_{r1}	⋯	T^*_{B1}
…	…	…	…	…	…
T_{ok}	T^1_{1k}	⋯	T^*_{rk}	⋯	T^*_{Bk}

2. Consider a member g^*, randomly drawn from the proper group of transformations $\boldsymbol{G}$, and the values of vector statistics $\boldsymbol{T}^* = \boldsymbol{T}(\mathbf{X}^*)$, where $\mathbf{X}^* = g^*(\mathbf{X})$. In most situations, data permutation $\mathbf{X}^*$ may be obtained by first considering a random permutation $(u^*_1, \ldots, u^*_n)$ of basic label integers (1, …, n) and then by assignment of related individual data vectors to the proper group; thus, according to the data representation given in $\mathbf{X}^* = \{\mathbf{X}(u^*_i),\ i = 1, \ldots, n;\ n_1, \ldots, n_C\}$ (see Table 3.2).
3. Independently repeat step **I.2** B times. The set of conditional Monte Carlo iteration results $\{T^*_r,\ r = 1, \ldots, B\}$ is thus a random sampling from the permutation k-variate distribution of vector test statistics $\boldsymbol{T}$.
4. The k-variate E.D.F. $\hat{F}_B(z|\mathrm{X}) = [0.5 + \sum_r I(T^*_r \le z)]/(B+1), \forall z \in \Re^k$, gives an estimate of the corresponding k-dimensional permutation distribution $F(z|\mathbf{X})$ of $\boldsymbol{T}$. Moreover, $\hat{L}_i(z|\mathrm{X}) = [0.5 + \sum_r I(T^*_{ir} \ge z)]/(B+1)$, $i = 1, \ldots, k$, gives an estimate $\forall z \in \Re^1$ of the marginal permutation significance level functions $L_i(z|\mathrm{X}) = \Pr\{T^*_i \ge z|\mathrm{X}\}$; thus $\hat{L}_t(T_o|\mathrm{X}) = \hat{\lambda}_i$ gives an estimate of the marginal p-value $\lambda_i = \Pr\{T^*_i \ge T_o|\mathrm{X}\}$, relative to test T_i. All these are unbiased and consistent estimates of corresponding true values.

Table 3.2 summarizes the observed data set and one multidimensional permutation in a two-sample problem. Table 3.3 summarizes the CMC-Procedure. In multidimensional problems, the CMC-Procedure only considers permutations of individual data vectors, so that: $\mathrm{X}^* = \{\mathrm{X}(u^*_i),\ i = 1, \ldots, n;\ n_1, \ldots, n_C\}$, as is explicitly displayed in the second part of Table 3.2, and thus all dependence relations which are present in the component variables are preserved. From this point of view, the CMC-Procedure is essentially a multivariate procedure.

With respect to standard E.D.F. estimators, 1/2 and 1 have been added, respectively, to the numerators and denominators of relationships in step **I.4**. This is done in order to obtain estimated values of c.d.f. $F(z|\mathbf{X})$ and of p-values in the open interval (0,1), so that transformations by inverse c.d.f. of continuous

Table 3.4 Nonparametric combination

T_{o1}	T^*_{11}	…	T^*_{r1}	…	T^*_{B1}
…	…	…	…	…	…
T_{o1}	T^*_{1k}	…	T^*_{rk}	…	T^*_{Bk}

$\hat{\lambda}_1$	λ^*_{11}	…	λ^*_{1r}	…	λ^*_{1B}
…	…	…	…	…	…
$\hat{\lambda}_k$	λ^*_{k1}	…	λ^*_{kr}	…	λ^*_{kB}

$T^{"}_o$	$T^{"*}_1$	…	$T^{"*}_r$	…	$T^{"*}_B$

distributions, such as $-\log(\lambda)$ or $\Phi^{-1}(\lambda)$, etc. (where Φ is the standard normal c.d.f.) are continuous. However, as B is generally large, this minor alteration is substantially irrelevant because it does not modify test behaviour or consequent inferences, neither for finite sample sizes nor asymptotically. In particular, *the* following is valid:

Proposition *As B tends to infinity*, $\hat{F}_B(z|\mathrm{X})$ *almost certainly converges to permutation c.d.f.* $F(z|\mathbf{X})$, $\forall z \in \Re^k$.

For the proof of this statement, see Pesarin and Salmaso (2010).

Phase II. The second phase of the algorithm for simulating a procedure for nonparametric combination should include the following steps:

1. The k observed p-values are estimated on data $\mathbf{X}$ by $\hat{\lambda}_i = \hat{L}_i(T_{oi}|\mathbf{X})$, where $T_{oi} = T_i(\mathbf{X})$, $i = 1, \ldots, k$, represent the observed values of partial tests and $\hat{L}_i$ are the i-th marginal significance level functions estimated by the CMC-Procedure on data set $\mathbf{X}$.
2. The combined observed value of the second-order test is again evaluated through the same conditional simulation results as the first phase, and is given by: $T^*_o = \psi\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_k\right)$.
3. The rth combined value of vector statistics is then calculated by $T''^*_r = .\psi\left(\lambda^*_{1r}, \ldots, \lambda^*_{kr}\right)$, where $\lambda^*_{1r} = \hat{L}_i\left(T^*_{ir}|\mathbf{X}\right)$, $i = 1, \ldots, k$, $r = 1, \ldots, B$.
4. Hence, the p-value of combined test T'' is estimated as: $\hat{\lambda}''_\psi = \sum_r I\left(T''^*_r \geq T''_{io}\right)/B$.
5. If $\hat{\lambda}''_\psi < \alpha$, global null hypothesis H_0 is rejected at significance level α.

Table 3.4 displays the nonparametric combination.

The CMC-procedure gives unbiased and consistent estimates of both true permutation distribution $F_\psi(t|\mathbf{X})$, $\Pr\{T''^* \leq t\}|\mathbf{X}$, $\forall t \in \mathscr{R}^1$ and true p-value $\lambda''_\psi = \Pr\{T''^* \geq T_o|\mathbf{X}\}$. In fact, $\hat{\lambda}_i \to \lambda_i$ with probability one ($i = 1, \ldots, k$), as B tends to infinity. Hence, $\hat{\lambda}''_\psi$ converges to λ_ψ with probability one, k being a fixed

finite integer and combining function ψ being continuous by assumption. This combination is a proper nonparametric method for multidimensional testing problems because it takes into consideration only the whole joint k-variate permutation (conditional) distribution $F(z|\mathbf{X})$ of $\boldsymbol{T}$, estimated by the E.D.F. $\hat{F}_{\cdot B}(z|\mathbf{X})$. In particular, it is nonparametric with respect to the latent dependence structure in population distribution P. At this point it is important to note that if proper routines for exact calculations are available, then multidimensional distribution $F(z|\mathbf{X})$, partial p-values $(\lambda_1, \ldots, \lambda_k)$, distribution of combined test $F\psi(t|\mathbf{X})$, and combined p-value λ''_ψ are all exactly evaluated.

Some examples of combining functions are now presented.

A. *Fisher's omnibus* combining function, based on the statistic:

$$T''_F = -2 \cdot \sum_i \log(\lambda_i).$$

It is well known that if the k partial test statistics are independent and continuous, then under the null hypothesis T''_F is distributed according to a central χ^2 with $2 \cdot k$ degrees of freedom. T''_F is the most popular combining function and corresponds to the so-called "multiplicative rule of combination".

B. *Liptak's* combining function, based on the statistic:

$$T''_L = \sum_i \Phi^{-1}(1 - \lambda_i),$$

where Φ is the standard normal c.d.f.. If the k test statistics are independent and continuous, then under the null hypothesis T''_L is normally distributed with mean 0 and variance k. Another version of *Liptak's* function considers logistic transformations of p-values, i.e.:

$$T''_P = \sum_i \log[(1 - \lambda_i)/\lambda_i].$$

C. *Tippett's* combining function is given by

$$T''_P = \max_{1 \le i \le k}(1 - \lambda_i),$$

where its null distribution, if the k tests are independent and continuous, behaves according to the largest of k random values from the uniform distribution in the open interval (0, 1).

D. *Lancaster's* combining solutions are based on statistics such as

$$T''_G = \sum_i \Gamma^{-1}_{r,a}(1 - \lambda_i),$$

where $\Gamma^{-1}_{r,a}$ represents the inverse c.d.f. of a central gamma distribution with known scale parameter a and r degrees of freedom. If the k partial tests are

independent, then the null distribution of T''_G is central Gamma with scale parameter a and $r \cdot k$ degrees of freedom.

References

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

F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software* (Wiley, Chichester, 2010)

Chapter 4
Statistical Problems of Allelic Association

Abstract This chapter introduces a new approach for testing for allelic association by means of nonparametric permutation tests. The genetic configuration our statistical problem can be formalized by considering bivariate responses: (X_1, X_2) where the observed subjects are partitioned into two groups (according to the typical case–control study), so that data may be represented as:

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

where responses are binary ordered categorical such as (AA, Aa, aa) or the like. An exact permutation test is then derived to efficiently solve this problem. Such novel approach is compared with competitors within parametric tests mostly used in genetics and in isotonic inference testing problems. Finally, an extension to the multivariate case is introduced and discussed.

Keywords Allelic association test · Chiano and Clayton test · Multiallelic problems

4.1 The Nonparametric Permutation Approach for the Allelic Association Test

The genetic statistical problem we are going to discuss is quite common in any context related with restricted alternatives, or more generally in testing under order constraints (see Hirotsu 1982, 1986, 1998). In the genetic configuration introduced by Chiano and Clayton (1998), our 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

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_4, © Luigi Salmaso 2011

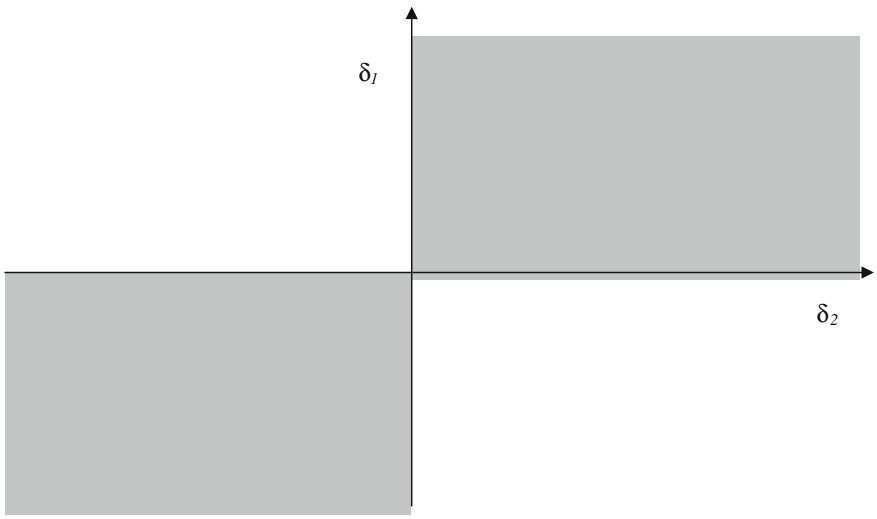

Fig. 4.1 Representation of the bivariate isotonic hypotheses

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

where responses are binary ordered categorical such as (*AA*, *Aa*, *aa*) or the like. 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 on responses is generally induced by the nature of the problem at hand. The hypotheses we are interested in are

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

against the special isotonic set of alternatives:

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

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: $X_{hji} = \delta_{hj}\ (\mu_h + Z_{hji})$, where δ_{hj} is the effect on the h-th variable in the j-th group, all other symbols having clear meanings (Pesarin and Salmaso 2010). 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. This situation is displayed in Fig. 4.1, where the two gray sections ($\delta_1 > 0$, $\delta_2 > 0$) and ($\delta_1 < 0$, $\delta_2 < 0$) represent the alternative hypothesis, the half-lines correspond to ($\delta_1 > 0$, $\delta_2 = 0$), ($\delta_1 < 0$, $\delta_2 = 0$), ($\delta_1 = 0$, $\delta_2 > 0$) and ($\delta_1 = 0$, $\delta_2 < 0$), while the null hypothesis is the single point ($\delta_1 = 0$, $\delta_2 = 0$). The points in ($\delta_1 < 0$, $\delta_2 > 0$) and ($\delta_1 > 0$, $\delta_2 < 0$) are not relevant for the analysis. These kinds of hypotheses arise when two variables are such that under the alternative at least one of them stochastically increases *XOR* decreases, whereas the other variable may remain either affected or non-affected.

In our genetic context, this happens when a gene is associated with a given disease so that in affected individuals (cases) at least one of the genotype frequencies with a 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. The partial tests to take into consideration are therefore:

$$T_h^* = \sum\nolimits_i X_{h2i}^* - \sum\nolimits_i X_{h1i}^*,\ h = 1, 2.$$

In the present problem, under H_1, p-values of partial tests are either stochastically smaller than α or stochastically larger than $1 - \alpha$. Hence, we need to modify assumptions 1 and 2 in Sect. 3.3 to:

1. All partial tests T_i, $i = 1, 2$ are marginally unbiased and significant either for large or small values, so that their permutation distribution under H_1 *is* either stochastically larger or smaller than under H_0.
2. All partial tests T_i, $i = 1, 2$ are consistent.

Furthermore, we also need to modify the properties of combining functions ψ (a, b and c in Sect. 3.3) to:

a. A continuous combining function ψ must be monotonically decreasing in each argument: $\psi(\ldots, \lambda_i, \ldots) > \psi\left(\ldots, \lambda_i', \ldots\right)$, if $\lambda_i < \lambda_i'$, $i = 1, \ldots, k$.
b. It must attain its supremum positive value $\overline{\psi}$, possibly nonfinite, when at least one argument attains 0 (zero): $\psi(\ldots, \lambda_i, \ldots) \to \overline{\psi}$ if $\lambda_i \to 0$; moreover it must attain its infimum negative value $\underline{\psi}$, possibly nonfinite, when at least one argument attains 1: $\psi(\ldots, \lambda_i, \ldots) \to \underline{\psi}$ *if* $\lambda_i \to 1$;
c. $\forall\ \alpha > 0$, its acceptance region is bounded: $\psi < T''_{\alpha/2} < T'' < T''_{1-\alpha/2} < \overline{\psi}$.

Furthermore, we also need to modify step II.5 in Sect. 3.3 to:

(II.5) if $1 - |2\hat{\lambda}''_{\psi} - 1| \leq \alpha$, then reject H_0 at significance level α.

If the exchangeability property is satisfied under H_0, the nonparametric combination methods lead to exact, unbiased and consistent permutation tests (Pesarin and Salmaso 2010).

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 compared with control individuals. For a bi-allelic locus with common allele a and rare allele

A, individuals may carry zero (subjects with genotype *aa*), one (subjects with genotype *Aa*) or two (subjects with genotype *AA*) copies of the *A* allele. Therefore, conventionally 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 $R_h = f_{h1}/f_{h2}$, $h = AA, Aa, aa$ (where f_{hj}, $j = 1, 2$ are the observed frequencies in cases and controls, respectively), 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 allele *A* (genotype *AA*), whereas the behaviour in the 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 the presence of a protective effect) *XOR* ($R_{AA} < R_{Aa} = R_{aa}$ for a deleterious effect).
2. *Codominant*. There is an ordering on effects associated with allele *A*: genotype *Aa* is of risk (or of protection) in comparison with genotype *aa*, and *AA* is of risk (or of protection) in comparison with genotype *Aa*. Obviously, *AA* is of great risk (or of great protection) in comparison with 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 allele *A* is the same in genotypes *AA* and *Aa*. 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 both directions is strong. This system of hypotheses is equivalent to the previous.

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

$$H_0 : \{(\theta_{AA} = 1) \cap (\theta_{Aa} = 1)\}$$

is equivalent to

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

Table 4.1 Data vector for the permutation test of the allelic association problem

Values	2	1	…	…	2	3	…	…	1
Position	1	2	…	…	n_{cases}	$n_{cases}+1$	…	…	n

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

The permutation solution is based on two partial statistics:

$$T_{AA} = f_{AA,\,cases} \cdot f_{Aa,\,controls} \big/ \left(f_{Aa,\,cases} \cdot f_{AA,\,controls}\right)$$

$$T_{Aa} = f_{Aa,\,cases} \cdot f_{aa,\,controls} \big/ \left(f_{aa,\,cases} \cdot f_{Aa,\,controls}\right)$$

which test the respective partial hypotheses:

$$H_{0AA} : \{\theta_{AA} = 1\} \text{ against } H_{1AA} : \{\theta_{AA} > 1 \text{ or } \theta_{AA} < 1\}$$

$$H_{0Aa} : \{\theta_{Aa} = 1\} \text{ against } H_{1Aa} : \{\theta_{Aa} > 1 \text{ or } \theta_{Aa} < 1\}.$$

Note in fact that:

$$\{\theta_{AA} = 1\} \Leftrightarrow \{f_{AA,cases} \cdot f_{Aa,controls} \overset{d}{=} f_{Aa,cases} \cdot f_{AA,controls}\},$$

so that the two relations are equivalent. To explain how the test is done, we start from the CMC method. We construct a vector of dimension n ($n = n_{cases} + n_{controls}$) and assign three different values to observations of different genotypes, for instance: 1 to all the n_{AA} subjects who stay in the cells (AA, cases) and (AA, controls), 2 to all the n_{Aa} subjects who stay in the cells (Aa, cases) and (Aa, controls), and 3 to all remaining n_{aa} subjects who stay in the cells (aa, cases) and (aa, controls).

Now, we randomly insert the $f_{AA,\,cases}$ values 1, the $f_{Aa,\,cases}$ values 2 and the $f_{aa,\,cases}$ values 3 in the first n_{cases} positions of the vector, and in the same way all the other values in second $n_{controls}$ positions of the vector. We obtain a vector such as the one in Table 4.1.

In this way we preserve all the marginal values of an association table (n_{cases}, $n_{controls}$, n_{AA}, n_{Aa}, n_{aa}). The permutation statistics T^*_{AA} and T^*_{Aa} are calculated on the same vector, after executing a random permutation of its n elements. For example, the estimation of partial p-value λ_{AA} is obtained using B CMC-iterations, such as

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

This partial p-value is distributed as $U(0, 1)$ and it allows us to reject H_{0AA} when $\hat{\lambda}_{AA} \leq \alpha/2$, or $\hat{\lambda}_{AA} \geq 1 - \alpha/2$, at a fixed significance level α. By using the same previously obtained B vectors, we also estimate the p-values $\lambda'_{AAs} = \Pr\,(T^*_{AA} \geq T^*_{AAs}.|\mathbf{X})$, where s (1, …, B):

$$\hat{\lambda}'_{AA\,s} = \frac{\#(T^*_{AA} \geq T^*_{AA\,s})}{B}.$$

With the two partial p-values and the other B type-I p-values for each of them, we use Liptak's combining function to construct the combined test which verifies the initial hypothesis system. The final p-value λ_L is estimated by:

$$\hat{\lambda}_L = \frac{\#_s^B\left\{\left[\Phi^{-1}(1-\hat{\lambda}'_{AA\,s}) + \Phi^{-1}(1-\hat{\lambda}'_{Aa\,s})\right] \geq \left[\Phi^{-1}(1-\hat{\lambda}_{AA}) + \Phi^{-1}(1-\hat{\lambda}_{Aa})\right]\right\}}{B}.$$

The final 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.

4.2 Exact Nonparametric Solution for the Genetic Problem

The previous problem can be represented by a simple case–control contingency table (Table 4.2).

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

The usual data file representation provides the structure represented in Table 4.3.

where in the first M observations (or subjects), we have X_1 AA genotypes, X_2 Aa genotypes and X_3 aa genotypes. The order among the first M subjects does not matter (and the same applies for the second N subjects) because the result in the contingency table does not change if we take two random permutations into these sub-vectors, and the frequencies X_1, X_2, X_3, Y_1, Y_2 and Y_3 remain the same. Thus, if we consider the overall permutation space associated to the data in the previous paragraph, (S!), it may be too large to explore exhaustively, even for the most modern computers (and if it were possible in some situations, the time required would be very lengthy).

We, instead, look exclusively at those specific combinations and recombinations of the permutated genotypes/haplotypes in the table which give us a particular structure of the cells.

Observe the following example to explain this concept. We have a particular permutation in the data which makes it possible to obtain the data set represented in Table 4.4.

Table 4.2 Case–control table for allelic association study

Genotype/haplotype:	Cases	Controls	Size
AA	X_1	Y_1	$S_1 = X_1 + Y_1$
Aa	X_2	Y_2	$S_2 = X_2 + Y_2$
Aa	X_3	Y_3	$S_3 = X_3 + Y_3$
Size	$M = X_1 + X_2 + X_3$	$N = Y_1 + Y_2 + Y_3$	$S = M + N = S_1 + S_2 + S_3$

Table 4.3 Data file representation

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

Table 4.4 A particular result of a permutation in the data set

	Ca.	*Co.*	
AA	x_1^*	y_1^*	$\boldsymbol{S_1}$
Aa	x_2^*	y_2^*	$\boldsymbol{S_2}$
aa	x_3^*	y_3^*	$\boldsymbol{S_3}$
	$\boldsymbol{M}$	$\boldsymbol{N}$	$\boldsymbol{S}$

The marginal sums are identical for any permutation. Only the frequencies in the cells may change. The relative data file is illustrated in Table 4.5.

Here, $\forall i,\ i'(i \neq i'),\ u_i^* = u_j^*$ *and* $u_{i'}^* = u_{j'}^*$, *where* $j \neq j'$, *and* $i, i', j, i' \in \{1, \ldots, S\}$. Furthermore, in the first M observations (or subjects), we have x_1^* *AA* genotypes, x_2^* *Aa* genotypes and x_3^* *aa* genotypes. Again, the orders of the two subvectors (first M elements and second N elements) are not important.

We see that there are not (S)! different results for the permutations, but many permutations with different numbers give a specific structure of the cells $x_1^*, x_2^*, x_3^*, y_1^*, y_2^*$ and y_3^*, which are the important parameters for our statistics. Therefore, we can construct the exact permutation distribution for the statistics, associating to the statistics their related frequencies, i.e. the times these values of the statistics appear into the (S!) permutations. In doing that we do not need expensive computer iterations; we can use the combinatorial calculus. Then, in the exploration of the total sample space, we are looking for the frequencies associated to all the different configurations of the table (Table 4.4), i.e. all the sets $\{x_1^*, x_2^*, x_3^*, y_1^*, y_2^*, y_3^*\}$ where at least one cell is different from the others.

For the data in Table 4.4, we can obtain all the different table configurations from the following algorithm:

1. $x_1^* \in [\ \max\ (0,\ S_1 - N),\ \min\ (M,\ S_1)]$;
2. $y_1^* = S_1 - x_1^*$;
3. $x_2^* \in [\ \max\ (S_2 - (N - y_1^*)),\ \min\ (M - x_1^*,\ S_2)]$;
4. $y_2^* = S_2 - x_2^*$;
5. $x_3^* = M - x_1^* - x_2^*$;
6. $y_3^* = S_3 - x_3^*$.

Table 4.5 Representation of the permutation from the data file

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

Then, for a specific set i $\{{}_ix_1^*, {}_ix_2^*, {}_ix_3^*, {}_iy_1^*, {}_iy_2^*, {}_iy_3^*\}$ we have the frequency:

$$\begin{aligned} f_i^* &= M! \cdot N! \cdot \binom{S_1}{{}_ix_1^*} \cdot \binom{S_2}{{}_ix_2^*} \cdot \binom{S_3}{{}_ix_3^*} \\ &= (M!N!S_1!S_2!S_3!)/({}_ix_1^*!{}_ix_2^*!{}_ix_3^*!{}_iy_1^*!{}_iy_2^*!{}_iy_3^*!); \end{aligned}$$

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

$$\sum_i^I f_i^* = (M+N)! = (S!);$$

where the total number of all these different configurations is

$$I = \sum_{x_1^*}^{\min(M,S_1)+1-\max(0,S_1-N)} \left[\min(M - x_1^*, S_2) + 1 - \max(0, S_2 - (N - (S_1 - x_1^*)))\right];$$

so that the relative frequencies (simpler in the computer computations) are $p_i^* = f_i^*/(S!)$. Of course, the highest relative frequency is associated to configuration where x_1^* and x_2^* are close maximally (if possible, equal) to, respectively, y_1^* and y_2^*; this coincides (in general) with the case of no association between cases and controls.

Instead, we can see that the sampling distribution has a bell shape (where the parameters are: the mean of the cell configurations, i.e. in general equal to the configuration which has the maximum relative frequency; the variance between the cell configurations). However, note that this distribution is not continuous because the data are discrete.

We can therefore repeat the same previous test with nonparametric combination to have an exact p-value associated to the hypothesis system. We can call this type of procedure CEP (conditional exact procedure) to distinguish it from the CMC-Procedure (conditional Monte Carlo procedure) shown previously.

4.3 Chiano and Clayton's Parametric Approach

Chiano and Clayton (1998) started from this specific system of hypotheses with the odds ratio that considers the admissible genetic model in the definition of the model. After, for convenience, they used log transformations of the odds: $\beta_{AA} = \log\ \theta_{AA}$ and $\beta_{Aa} = \log\ \theta_{Aa}$; so that the parametric space under the null hypothesis is:

$$\Omega_1 : \{(\beta_{AA} \geq 0) \cap (\beta_{Aa} \geq 0)\} XOR \{(\beta_{AA} \leq 0) \cap (\beta_{Aa} \leq 0)\},$$

while the null space is Ω_0, the origin of the axes $\{(\beta_{AA} = 0) \cap (\beta_{Aa} = 0)\}$, which is equivalent to the one shown in Fig. 4.1 (Sect. 4.1).

We use β to indicate the vector $[\beta_A, \beta_A]'$. By using the standard theory of estimation, they obtain $\lim_{n\to\infty} \sqrt{n}(\hat{\beta}_1 - \hat{\beta}_0) \to N(0, \sum_{\beta_0})$, where $\sum_{\beta_0}$ is the variance–covariance matrix of β evaluated into the null hypothesis. Then, if all regularity conditions hold, inference would be made by referring the likelihood ratio chi-squared statistic of Wilks:

$$\Lambda = 2 \sum nP(\hat{\beta}) \log \left[\frac{P(\hat{\beta})}{P(\beta_0)} \right],$$

to a standard χ^2 on 2 df.

Unfortunately, under such order restrictions, Wilks' regularity assumptions are not met and the null point (origin) is on the boundary. The likelihood is therefore maximized subject to order constraints as follows. First, we obtain the unrestricted maximum likelihood estimate $\hat{\beta}$ of β:

I. If $\hat{\beta} \in \{(\beta_{AA} \leq 0) \cap (\beta_{Aa} \geq 0)\} \cup \{(\beta_{AA} \geq 0) \cap (\beta_{Aa} \leq 0)\}$(the unshaded region in Fig. 4.1 of Sect. 4.1), $\hat{\beta}$ is remaximized subject to the constraint that $\beta_{Aa} = 0$ or $\beta_{AA} = 0$ (whichever is maximum) and
II. If $\hat{\beta}$ is in the shared region (Fig. 4.1 of Sect. 4.1), it is left as it is.

In other words, when $\hat{\beta}$ falls into the unshaded region, it is projected onto the line $\beta_{Aa} = 0$ or $\beta_{AA} = 0$ and the contribution to the overall distribution is χ^2_1. However, when $\hat{\beta}$ falls into the shaded region, the contribution is χ^2_2, with probability $\lambda(\beta)$ proportional to the area of the shaded region (for details, see Chiano and Clayton, 1998). It then turns out that the distribution of the likelihood ratio chi-square statistic can be represented as a mixture of 2 chi-square distributions:

$$\Lambda = \lambda(\beta)\chi^2_2 + (1 - \lambda(\beta))\chi^2_1,$$

where $\lambda(\beta)$, the mixing probability, can be approximated to

$$\lambda(\beta) = \cos^{-1}\left(\frac{I_{12}}{\sqrt{I_{11}I_{22}}} \right) / \pi,$$

I_{ij} being the components (i, j) of the variance–covariance matrix $\sum_{\beta_0}$ evaluated at the null.

We use α to denote the usual required significance level. Therefore, we have to find x such that:

$$P(\Lambda \geq x_\alpha) = \lambda(\beta)P(\chi^2_2 \geq x) + (1 - \lambda(\beta))P(\chi^2_1 \geq x).$$

4.4 The Maximum Likelihood Solution

This solution was developed by El Barmi and Dykstra (1999), referring to the previous work of Dykstra et al. (1995). The case–control table can be interpreted as two independent vectors of data. The random sample of $\boldsymbol{M}$ cases is taken from a multinomial distribution with probability vector $\boldsymbol{p} = (p_1, p_2, p_3)$ where it refers to the observed values ($p_1 = X_1/\boldsymbol{M}, p_2 = X_2/\boldsymbol{M}, p_3 = X_3/\boldsymbol{M}$), while the $\boldsymbol{N}$ controls are taken from a multinomial (independent from the other) distribution $\boldsymbol{q} = (q_1, q_2, q_3)$, where $q_1 = Y_1/\boldsymbol{N}$, $q_2 = Y_2/\boldsymbol{N}$, $q_3 = Y_3/\boldsymbol{N}$. We can derive the nonparametric maximum likelihood estimators (MLEs) of probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ under the two hypothesis systems:

$$H_0 : \{\mathbf{p} = \mathbf{q}\} \text{ VS } H_1 : \{\mathbf{p} \overset{LR}{>} \mathbf{q}\},$$

And

$$H_0 : \{\mathbf{p} = \mathbf{q}\} \text{ VS } H_1 : \{\mathbf{p} \overset{LR}{<} \mathbf{q}\},$$

and then use these estimates to construct a likelihood ratio test.

The symbols $\overset{LR}{>}$ or $\overset{LR}{<}$ mean there is a likelihood ratio ordering between the distributions of two vectors: $(X \overset{LR}{>} Y) \Rightarrow \forall a, b(a < b)$. The conditional distribution of X given $X \in (a, b)$ is stochastically greater than that of Y given $Y \in (a, b)$, or, equivalently, $[f_x(t)/f_y(t)]$ is nondecreasing in t, with f_x and f_y the density functions of X and Y.

We now need to express the likelihood function of (p, q), vectors of parameters (p_1, p_2, p_3) and (q_1, q_2, q_3), as $L \propto \prod_{i=1}^{3} p_i^{X_i} q_i^{Y_i}$. We reparameterize by letting: $\theta_i = M \cdot p_i/(M \cdot p_i + N \cdot q_i)$ and $\Phi_i = M \cdot p_i + N \cdot q_i$, to obtain $p_i = \theta_i \Phi_i / M$ and $q_i = \Phi_i(1 - \theta_i)/N$, $i = 1, 2, 3$. With some passages (see Dykstra et al. 1995), we obtain the MLEs of p and q under H_0 and H_1. Under H_0, i.e. $p \overset{LR}{=} q$, we have $p_i = q_i = (X_i + Y_i)/(M + N)$. Instead, under H_1, the MLEs are

$$\begin{aligned} \text{If } p \overset{LR}{>} q, p_i^* &= [(X_i + Y_i)/M] \cdot E_{(X+Y)}[\mathbf{X}/(\mathbf{X} + \mathbf{Y}) | \{(\theta_0, \theta_1, \theta_2) : \theta_0 \le \theta_1 \le \theta_2\}]_i \\ \text{and } q_i^* &= [(X_i + Y_i)/N] \cdot E_{(X+Y)}[\mathbf{Y}/(\mathbf{X} + \mathbf{Y}) | \{(\theta_0, \theta_1, \theta_2) : \theta_0 \ge \theta_1 \ge \theta_2\}]_i, \end{aligned}$$

$$\begin{aligned} \text{If } p \overset{LR}{<} q, p_i^* &= [(X_i + Y_i)/M] \cdot E_{(X+Y)}[\mathbf{X}/(\mathbf{X} + \mathbf{Y}) | \{(\theta_0, \theta_1, \theta_2) : \theta_0 \ge \theta_1 \ge \theta_2\}]_i \\ \text{and } q_i^* &= [(X_i + Y_i)/N] \cdot E_{(X+Y)}[\mathbf{Y}/(\mathbf{X} + \mathbf{Y}) | \{(\theta_0, \theta_1, \theta_2) : \theta_0 \le \theta_1 \le \theta_2\}]_i, \end{aligned}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the data vectors (x_1, x_2, x_3) and (y_1, y_2, y_3), while the θ_i's must satisfy some conic restrictions (see Dykstra et al. 1995).

These MLEs are consistent in the sense that the associated CDFs converge pointwise to the true CDFs when $M, N \to \infty$ and the likelihood ratio order holds. The likelihood ratio test is the statistic:

$$\Psi = \frac{\sup_{(p,q)\in H_0} L((p.q))}{\sup_{(p,q)\in H_1} L((p.q))} = \frac{L(p^0.q^0)}{L(p^*.q^*)} = \ldots = \prod_{i=1}^{3} \left(\frac{\theta_i^0}{\theta_i^*}\right)^{X_i} \left(\frac{1-\theta_i^0}{1-\theta_i^*}\right)^{Y_i},$$

because $\Phi_i^0 = \Phi_i^*$. The test rejects H_0 for large values of $T = -2 \ln \Psi$, i.e. for large values of

$$T = 2\sum_{i=1}^{3} \{X_i \ln \theta_i^* + Y_i \ln(1-\theta_i^*) - X_i \ln \theta_i^0 - Y_i \ln(1-\theta_i^0)\}.$$

T, under H_0, has, asymptotically, a mixed distribution of χ_1 and χ_2, but it is not simple to write.

We now consider a special algorithm to find the maximum likelihood estimates. We use $(m_1, m_2, \ldots, m_k)$ and $(n_1, n_2, \ldots, n_k)$ to denote the frequencies corresponding to two independent multinomials with parameters $(m, (p_1, p_2, \ldots, p_k))$ and $(n, (q_1, q_2, \ldots, q_k))$, respectively. Next we consider maximizing

$$\prod_{i=1}^{3} p_i^{m_i} \prod_{i=1}^{3} q_i^{n_i} \tag{4.1}$$

subject to

$$\frac{p_1 q_2}{p_2 q_1} \geq \frac{p_2 q_3}{p_3 q_2} \geq 1 \tag{4.2}$$

or

$$\frac{p_1 q_2}{p_2 q_1} \leq \frac{p_2 q_3}{p_3 q_2} \leq 1 \tag{4.3}$$

and

$$\sum_{i=1}^{3} p_i = 1, \quad \sum_{i=1}^{3} q_i = 1. \tag{4.4}$$

To solve this problem, we use the algorithm developed in El Barmi and Dykstra (1998). For completeness, we firstly describe the algorithm and then show how to apply it to (4.1).

Consider the problem of maximizing

$$\prod_{i=1}^{k} p_i^{n_i} \tag{4.5}$$

subject to $\mathbf{p} \in P$ and

$$\mathbf{p} \in K_1 \tag{4.6}$$

$$\ln\ \mathbf{p} \in K_2 \tag{4.7}$$

where $\ln \mathbf{p} = (\ln p_1, \ln p_2, \ldots, \ln p_k)$, K_1 and K_2 are two cones (a cone is defined as a subset of $\mathscr{R}^k$ that satisfies $\alpha\mathbf{x}$ in the cone whenever $\mathbf{x}$ is in the cone for any $\alpha \geq 0$). Examples of cones in $\mathscr{R}^k$ include any linear space and a nonnegative orthant. We assume here that K_2 contains constant vectors and note that when $K_1 = \mathscr{R}^k$ and K_2 is a linear space, this optimization problem corresponds to fitting a log-linear model to the data $(n_1, n_2, \ldots, n_k)$.

For a given cone K, its dual (polar) cone is defined **as**

$$K^* = \left\{ \mathbf{y}, \sum_{i=1}^{K} x_i y_i \leq 0, \forall \mathbf{x} \in K \right\}. \tag{4.8}$$

It is easy to see that when K is actually a linear space (and hence a cone), then K^* is its orthogonal space. El Barmi and Dykstra (1998) show that if $\mathbf{y}^* \in K_1^*$ and $\mathbf{z}^* \in K_2^*$ solve

$$\min \sum_{i=1}^{k} (\hat{p}_i - z_i) \ln \left[\frac{\hat{p}_i - z_i}{1 + y_i} \right] \tag{4.9}$$

subject to

$$\mathbf{y}^* \in K_1^* \tag{4.10}$$

$$\mathbf{z}^* \in K_2^* \tag{4.11}$$

then the vector whose i-th component is given by

$$p_i^* = \frac{\hat{p}_i - z_i^*}{1 + y_i^*}, i = 1, 2, \ldots, k, \tag{4.12}$$

solves (4.1) subject to (4.2) and (4.3).

To find $\mathbf{y}^*$ and $\mathbf{z}^*$ they developed an iterative algorithm which is guaranteed to converge to the true solution when the constraint region is not empty. To apply the algorithm, proceed as follows. Set $\mathbf{z}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \mathbf{0}$ and $v = 1$. At the vth step of the algorithm (step 1), calculate $\mathbf{z}^{(v)}$ which solves

$$\max_{z \in K_2^*} \sum_{i=1}^{k} (\hat{p}_i - z_i) \ln \left[\frac{\hat{p}_i - z_i}{1 + y_i^{(v-1)}} \right]. \tag{4.13}$$

The second step of the algorithm amounts to finding $\mathbf{y}^{(v)}$ that solves

$$\max_{y \in K_1^*} \sum_{i=1}^{k} (\hat{p}_i - z_i^{(v)}) \ln(1 + y_i). \tag{4.14}$$

At the end of the vth cycle the estimate of $\mathbf{p}$ is given by the vector whose i-th component is

$$p_i^{(v)} = \frac{\hat{p}_i - z_i^{(v)}}{1 + y_i^{(v)}}, i = 1, 2, \ldots, k. \tag{4.15}$$

This two-step procedure is repeated until sufficient accuracy is attained.

To apply the algorithm to (4.1) we reparametrize the problem as follows. Consider

$$\theta_i = \begin{cases} p_i/2, & i = 1, 2, 3 \\ q_{i-3}/2, & i = 4, 5, 6 \end{cases}$$

and

$$r_i = \begin{cases} m_i, & i = 1, 2, 3 \\ m_{i-3}, & i = 4, 5, 6 \end{cases}$$

and note that maximizing (4.1) is equivalent to maximizing

$$\prod_{i=1}^{6} \theta_i^{r_i} \tag{4.16}$$

subject to

$$\frac{\theta_1\theta_5}{\theta_2\theta_4} \geq \frac{\theta_2\theta_6}{\theta_3\theta_5} \geq 1 \tag{4.17}$$

$$\frac{\theta_1\theta_5}{\theta_2\theta_4} \leq \frac{\theta_2\theta_6}{\theta_3\theta_5} \leq 1 \tag{4.18}$$

and

$$\sum_{i=1}^{3} \theta_i - \sum_{i=4}^{6} \theta_i = 0, \tag{4.19}$$

$$\sum_{i=1}^{6} \theta_i = 1, \tag{4.20}$$

in the sense that if θ_i^*, $i = 1, 2, \ldots, 6$, solve (4.16), then $(p_i^*, q_i^*) = (2\theta_i^*, 2\theta_{i+3}^*)$, $i = 1, 2, 3$, solve (4.1).

Let $\mathbf{z}_1^{(1)} = (-1, 2, -1, 1, -2, 1)$, $\mathbf{z}_2^{(1)} = (0, -1, 1, 0, 1, -1)$ and $\mathbf{z}_1^{(2)} = (1, -2, 1, -1, 2, -1)$, $\mathbf{z}_2^{(2)} = (0, 1, -1, 0, -1, 1)$, then (4.17) and (4.18) can be expressed as

$$\sum_{j=1}^{6} z_{ij}^{(1)}\theta_j \le 0, i = 1, 2,$$

or

$$\sum_{j=1}^{6} z_{ij}^{(2)}\theta_j \le 0, i = 1, 2,$$

or equivalently, $(\theta_1, \theta_2, \ldots, \theta_6) \in K_2^{(1)}$ or $K_2^{(2)}$ where $K_2^{(1)}$ and $K_2^{(2)}$ are two cones whose respective duals are given by

$$K_2^{(l*)} = \left\{(\beta_1 z_{11}^{(l)} + \beta_2 z_{21}^{(l)}, \ldots, \beta_1 z_{16}^{(l)} + \beta_2 z_{26}^{(l)}),\ \beta_1 \ge 0, \beta_2 \ge 0\right\},$$

$l = 1, 2$. Let $K_1 = \{\theta, \Sigma_{i=1}^{3}\theta_i - \Sigma_{i=4}^{6}\theta_i = 0\}$ and note that K_1^{*}={$(\alpha y_1, \alpha y_2, \ldots, \alpha y_6)$, $\alpha \in \mathscr{R}$} where $y_i = 1$, $i = 1, 2, 3$ and $y_i = -1$, $i = 4, 5, 6$.

Our problem is then equivalent to

$$\max\left\{\max(\theta \in K_1 \cap K_2^{(1)})\prod_{i=1}^{6}\theta_i^{r_i},\ \max(\theta \in K_1 \cap K_2^{(2)})\prod_{i=1}^{6}\theta_i^{r_i}\right\} \qquad (**)$$

and θ a probability vector. The algorithm described above can now be applied to solve

$$\max(\theta \in K_1 \cap K_2^{(1)})\prod_{i=1}^{6}\theta_i^{r_i} \qquad (4.21)$$

$$\max(\theta \in K_1 \cap K_2^{(2)})\prod_{i=1}^{6}\theta_i^{r_i} \qquad (4.22)$$

and θ a probability vector, individually and hence find the overall maximum.

For simplicity, we only consider how to implement the algorithm to (**). The dual problem in this case is given by

$$\min\sum_{i=1}^{6}(\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)})\ln\left[\frac{\hat{p}_i - \beta_1^{(v)} z_{1i}^{(1)} - \beta_2^{(v)} z_{2i}^{(1)}}{1 + \alpha y_i}\right] \qquad (4.23)$$

subject to $\beta_1 \ge 0$, $\beta_2 \ge 0$ and $\alpha \in \mathscr{R}$, where $\hat{\theta}_i = r_{ij}/(m+n), i = 1, 2, \ldots, 6$.

To apply the algorithm, set $\beta_i = 0$, $i = 1, 2$, and $\alpha = 0$. At the step of the algorithm (step 1), we calculate $\beta_1^{(v)}$, $i = 1, 2$, which solve

$$\min\sum_{i=1}^{6}(\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)})\ln\left[\frac{\hat{\theta}_i - \beta_1^{(v)} z_{1i}^{(1)} - \beta_2^{(v)} z_{2i}^{(1)}}{1 + \alpha^{(v-1)} y_i}\right] \qquad (4.24)$$

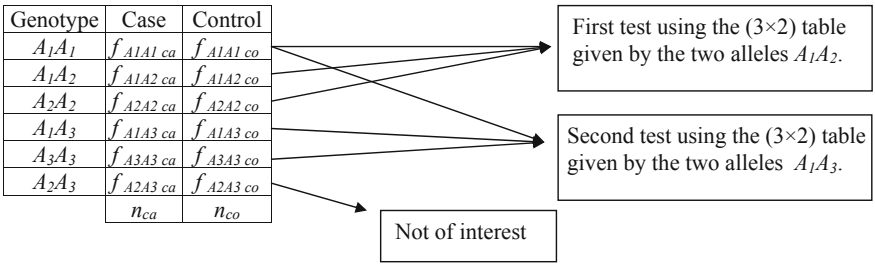

Fig. 4.2 Multiallelic problem

subject to

$$\beta_i \geq 0, i = 1, 2.$$

The second step of the algorithm amounts to finding $\alpha^{(v)}$ that solves

$$\max \sum_{i=1}^{6} (\hat{\theta}_i - \beta_1 z_{1i}^{(1)} - \beta_2 z_{2i}^{(1)}) \ln\{1 + \alpha y_i\} \tag{4.25}$$

subject to $\alpha \in \mathscr{R}$. The estimate at the v-th cycle of θ_{ij} is given by

$$\theta_i^{(v)} = \frac{\hat{\theta}_i - \beta_1^{(v)} z_{1i}^{(1)} - \beta_2^{(v)} z_{2i}^{(1)}}{1 + \alpha^{(v)} y_i}.$$

This iterative procedure is continued until sufficient accuracy is attained. What is remarkable here is that, at each step of the algorithm, a Newton–Raphson technique can be used on each variable and it requires a very small number of iterations to converge. Let $\theta^{(1)'}$ and $\theta^{(2)'}$ be the solutions corresponding to (4.21) and (4.22) respectively. Then the overall maximum must be achieved at one of them and hence is the solution to (4.18).

4.5 Some Extensions of the Nonparametric Solution to Multivariate Problems

We consider the immediate extensions of the nonparametric solution illustrated in Sect. 4.1. 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 (by CMC or CEP) separately for locus (A_1, A_2) and (A_1, A_3) because the interest lies 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, …). It is sufficient to repeat the simple test for both possible associations: rare1-common, rare2-common (see Fig. 4.2).

Table 4.6 Multiloci extension

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}

Table 4.7 The possible configurations

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	...	...

More complicated is the situation where the association study involves more than two loci, such as (a, A), where A is the rarest, and (b, B), where B is the rarest. We suppose the interest lies in knowing the specific effect of all the possible multiple configurations (Table 4.6).

The main aim in this situation is to reconstruct the possible effect that one locus may have, given a specific configuration of the other loci. Then we use six different (3×2) contingency tables, one for any specific configuration (Table 4.7).

For example, in the first table we carry out the nonparametric test (by CMC-Procedure or CEP) to study the association at locus (b, B) conditional to the genotype *aa* (more common) in the other loci. This procedure may seem a simple extension of the simple case, but if we consider the case of two loci each with three alleles, then we obtain two 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 latter situation, before executing the specific test for each configuration, it is helpful to make a single test to study if there is any type of significant association in at least one of all 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 odds ratios, as:

$$H_0 : \left\{ \bigcap_{r=1}^{k} [(\theta_{Aa\,r} = 1) \cap (\theta_{aa\,r} = 1)] \right\},$$

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

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

where, of course, at least one inequality in each of the $2 \times k$ lines is strict. The interpretation of this alternative is that there exists at least one gene 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 if we can admit that some genes may be of risk and others of protection, the remaining are neutral.

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

a. Data are organized in a unit-by-unit representation: $\{Y_{jir},\ r = 1,\ \ldots,\ k,\ i = 1, \ldots, n_j, j = case, control\}$, where Y_{jir} is the genotype of the r-th gene on the ith subject of the j-th group (i.e. Y_{jir} may assume one of the values: aa, Aa, AA);
b. Permutations exchange units (by CMC or CEP) between groups, so that k-dimensional vectors are exchanged;
c. For each gene r, $r = 1, \ldots, k$, calculate partial tests as $T^*_{r\ Aa} = f^*_{r\ AA\ case} f_{r\ Aa\ control} / (f^*_{r\ Aa\ case} f^*_{r\ AA\ control})$ and $T^*_{r\ aa} = f^*_{r\ Aa\ case} f^*_{r\ aa,\ control} / (f^*_{r\ aa\ case} f_{r\ Aa\ control})$, $r = 1, \ldots, k$, that are all significant for either large or small values;
d. Within each gene calculate second order combined test and related p-value $\hat{\lambda}''_r$, in accordance with the method previously discussed;
e. According to the nonparametric combination theory, 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 $\hat{\lambda}''' \leq \alpha$, then reject the overall null hypothesis.

A second kind of alternative of interest is

$$H'_1 : \left\{ \begin{array}{c} \cup_{1 \leq r \leq k} [(\theta_{Aa\,r} \geq 1) \cap (\theta_{aa\,r} \geq 1)] \\ XOR \\ \cup_{1 \leq r \leq k} [(\theta_{Aa\,r} \leq 1) \cap (\theta_{aa\,r} \leq 1)] \end{array} \right\},$$

where again at least one inequality in each line is strict. It 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 should 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 $\hat{\lambda}'''_1$;

f'. If $1 - |2\hat{\lambda}'''_1 - 1| \leq \alpha$, then reject the overall null hypothesis.

The third order combined tests and their p-values are always obtained by the same conditional simulation (by CMC-Procedure or CEP) results used for obtaining distributions of partial tests T^{*}_{rh} and p-values $\hat{\lambda}_{r\,h}$ and $\hat{\lambda}''_{r}, h = aa, Aa, AA, r = 1, \ldots, k$.

4.6 Allelic Association Studies with Confounding Effects

Suppose we now have a confounding factor that may have some effects on a particular pathology (for example, in individuals who live in contact with different levels of exposure to a specific agent). The study can involve one genotype (but the extension to more than one is immediate) that seems to be associated to that disease. We can represent the data as in Table 4.8.

As our first situation we consider only two levels in the confounding factor. By using the usual odds ratios, we obtain: $\theta_{aa} = (f_3 f_2)/(f_4 f_1)$ that are the odds we use as reference, $\theta_{Aa} = (f_7 f_6)/(f_8 f_5)$ and $\theta_{AA} = (f_{11} f_{10})/(f_{12} f_9)$. and These odds are estimated points of the level of association that is present in any genotype configuration between the different groups of exposure.

We may then observe, if the single odds are not one, where an association between the exposure levels and the genotype associated (or not) to the disease is present. Of course, this corresponds to a usual stratification on a variable of stratum (in case–control association studies this is very common, such as for factors like age, sex, …).

We obtained three tests of hypothesis to test if $\theta_x = 1$ or $\theta_x \neq 1$, where $x = aa$, Aa, AA. In taking into consideration the problem of isotonic inference, we can continue in the analysis and obtain the second ratios by using the first odds: $K_{aa} = \theta_{aa}/\theta_{aa} = 1$, $K_{Aa} = \theta_{Aa}/\theta_{aa}$ and $K_{AA} = \theta_{aa}/\theta_{AA}$. Then the hypotheses of interest are

$$H_0 : \{K_{aa} = K_{Aa} = K_{AA} = 1\}$$

and

$$H_1 : \{(K_{aa} \leq K_{Aa} \leq K_{AA} \leq 1)\ XOR\ (K_{aa} \geq K_{Aa} \geq K_{AA} \geq 1)\},$$

where at least one inequality is strong. The alternative is constructed in such a way that it corresponds to the possible effects that the confounding factor may have jointly with the putative allele on the affected individuals with respect to the cases. In fact, if a higher level of exposure is such that it increases the effects of the putative rare allele (or it in some way influences other factors correlated to it), the risk should in any case increase in homozygous rare subjects exposed to the factor, compared to exposed heterozygous individuals. Thus, we do not admit the points that fall in the areas such as $\{K_{aa} \geq K_{Aa} \leq K_{AA} \leq 1\}$.

Table 4.8 Confounding factor with two levels

Genotype	Confounding factor				
	Low exposure (−)		High exposure (+)		
	Case	Control	Case	Control	
aa	f_1	f_2	f_3	f_4	n_{aa}
Aa	f_5	f_6	f_7	f_8	n_{Aa}
AA	f_9	f_{10}	f_{11}	f_{12}	n_{AA}
	$n_{Ca,-}$	$n_{Co,-}$	$n_{Ca,+}$	$n_{Co,+}$	

Table 4.9 Three levels of confounding

Genotype	Confounding factor						
	Exposure level 1 (−)		Exposure level 2 (+)		Exposure level 3 (++)		
	Case	Control	Case	Control	Case	Control	
aa	f_1	f_2	f_3	f_4	f_5	F_6	n_{aa}
Aa	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n_{Aa}
AA	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	n_{AA}
	$n_{Ca,-}$	$n_{Co,-}$	$n_{Ca,+}$	$n_{Co,+}$	$n_{Ca,++}$	$n_{Co,++}$	

If we consider a more general environmental factor or a variable of stratum, where there are three different levels to order in some way, the data is as shown in Table 4.9.

In this case, we need six partial odds ratios: $\theta_{aa} = f_3 f_2/(f_4 f_1)$, $\theta_{Aa} = f_9 f_8/(f_{10} f_7)$, $\theta_{AA} = f_{15} f_{14}/(f_{16} f_{13})$ and $\theta'_{aa} = f_5 f_4/(f_6 f_3)$, $\theta'_{Aa} = f_{11} f_{10}/(f_{12} f_9)$, $\theta'_{AA} = f_{17} f_{16}/(f_{18} f_{15})$, that may be tested in the same way as previously.

The related second ratios become: $K_{aa} = \theta_{aa}/\theta_{aa} = 1$, $K_{Aa} = \theta_{aa}/\theta_{Aa}$, $K_{AA} = \theta_{aa}/\theta_{AA}$ and, for the second group, $K'_{aa} = \theta'_{aa}/\theta'_{aa} = 1$, $K'_{Aa} = \theta'_{aa}/\theta'_{Aa}$, $K'_{AA} = \theta'_{aa}/\theta'_{AA}$. So there are two sets of isotonic hypotheses:

$$H_0 : \{K_{aa} = K_{Aa} = K_{AA} = 1\}$$

against

$$H_1 : \{(K_{aa} \leq K_{Aa} \leq K_{AA} \leq 1)\ XOR\ (K_{aa} \geq K_{Aa} \geq K_{AA} \geq 1)\},$$

and

$$H_0 : \{K'_{aa} = K'_{Aa} = K'_{AA} = 1\}$$

against

$$H_1 : \{(K'_{aa} \leq K'_{Aa} \leq K'_{AA} \leq 1)\ XOR\ (K'_{aa} \geq K'_{Aa} \geq K'_{AA} \geq 1)\},$$

where all equations have a clear meaning.

Table 4.10 Two confounding factors with three levels

Gaaenotypes	V_1						V_2						V_3					
	U_1		U_2		U_3		U_1		U_2		U_3		U_1		U_2		U_3	
	ca	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*	*ca*	*co*
aa	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ab	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
bb	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ac	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
cc	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
bc	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 4.11 Data table for studying the Hardy–Weinberg equilibrium

Genotypes	Cases	Controls
AA	X_1	Y_1
Aa	X_2	Y_2
Aa	X_3	Y_3
Total	M	N

Furthermore, we may have an overall parameter that estimates the general association in the contingency table:

$$K^{FIN} = \frac{K'_{AA}K_{Aa}}{K'_{Aa}K_{AA}}.$$

If this value is equal to one, it means that there is no association in the data. If it is not one, one type of association is present (it is not important which type).

To increase the difficulty (Table 4.10), we may consider the case of two stratification or confounding factors with three levels each and the genetic locus under study is three-allelic (*a* being the more common gene, *b*, *c*).

In this case we can use the following algorithm, in basic language, to obtain the *p*-values associated to the tests on factor *U* conditioned to the levels of factor *V*:

```
for s=1 to 3
  for j=1 to 6
     for k=1 to 2
       Test1U(s,j,k)=MATRIX(j,s,k,2)
MATRIX(j,s,k+1,1)/(MATRIX(j,s,k,1)MATRIX(j,s,k+1,2))
     next k
  next j
   for k=1 to 2
      Test2Uab(s,k) = Test1U(s,1,k) / Test1U(s,2,k)
      Test2Ubb(s,k) = Test1U(s,1,k) / Test1U(s,3,k)
      Test2Uac(s,k) = Test1U(s,1,k) / Test1U(s,4,k)
      Test2Ucc(s,k) = Test1U(s,1,k) / Test1U(s,5,k)
     next k
   Test3b = Test2Uab(s,2) Test2Ubb(s,1) / (Test2Uab(s,1) Test2Ubb(s,2))
   Test3c = Test2Uac(s,2) Test2Ucc(s,1) / (Test2Uac(s,1) Test2Ucc(s,2))
next
```

where in row 5 we obtain all the observed difference tests θ_{ij}, $i = 1$ or 2 because we have three levels and so we compare level U_1 against level U_2 and U_2 against U_3, for all the genotypes of interest ($j = aa, ab, bb, ac, cc$). The 4 statements following the loop for in line 10 are the observed tests $K_{i1} = \theta_{i1}/\theta_{i2}$ and $K_{i2} = \theta_{i1}/\theta_{i3}$ for $i = 1, 2$. Finally, tests in rows 11 and 12 are K^{FIN} associated rare allele *b* and rare allele *c*. These results are repeated for all levels of V. In the same way we operate to test V|U.

4.7 The Hardy–Weinberg Equilibrium

In classic population case–control studies the researcher is often led to substitute the unrelated observed controls with the controls obtained after assuming the

Table 4.12 Expected frequencies under the H–W equilibrium

Genotypes	Cases	Controls
AA	X_1	$Y_1^1 = N[(2Y_1 + Y_2)/(2N)]^2$
Aa	X_2	$Y_2^1 = 2N[(2Y_1 + Y_2)/(2N)][(2Y_3 + Y_2)/(2N)]$
aa	X_3	$Y_3{}^1 = N[(2Y_3 + Y_2)/(2N)]^2$

Table 4.13 Calculation for expected controls under the null hypothesis

Genotypes	Cases	Controls
AA	X_1	$Y_1^2 = N[(2X_1 + X_2 + 2Y_1 + Y_2)/(2M + 2N)]^2$
Aa	X_2	$Y_2^2 = 2N[(2X_1 + X_2 + 2Y_1 + Y_2)(2X_3 + X_2 + 2Y_3 + Y_2)/(2M + 2N)^2]$
aa	X_3	$Y_3{}^2 = N[(2X_3 + X_2 + 2Y_3 + Y_2)/(2M + 2N)]^2$

Hardy–Weinberg (H–W) equilibrium. In brief, the H–W law states that if we have a bi-allelic gene with a rare allele *A* frequency of p (obviously $q = 1 - p$ is the frequency of the more common allele *a*), then the expected genotype frequencies in the population (by assuming typical conditions: diploid population, sexual reproduction, random mating, discrete generations, no mutation, no migration, very large population, no selection) are $f_{AA} = p^2$, $f_{Aa} = 2pq$ and $f_{aa} = (1 - p)^2$. Therefore, let us suppose we have the data table shown in Table 4.11.

The observed controls can be changed with the expectative frequencies under the H–W equilibrium (Table 4.12) and the relative risks can be estimated by using the odds ratios obtained from this table. This is very frequently done in literature and has been shown to produce more powerful results.

Furthermore, the parametric association studies using the case–control tables may also consider another type of expected frequency under the H–W equilibrium. As is done by Lathrop (1983) and Chiano and Clayton (1998), if we consider the null hypothesis of equal distributions in cases and controls, so that the odds ratios are both equal to one, we can assume the H–W equilibrium exists in all the table's data, both in controls and cases. The expected controls under the null hypothesis are then constructed as displayed in Table 4.13.

In this way the likelihood ratio test may be constructed by considering the ratio between the frequencies under H_1, which are the frequencies observed from the second table, and the frequencies under H_1, which are the expected frequencies we obtain from the third table. This solution may be more powerful and robust with respect to spurious association (Lathrop 1983, Chiano and Clayton 1998).

In the permutation solution we consider the adjustment for controls in the H–W equilibrium based on the correction for only the population of control individuals. This is done because the configuration of the data which involves distribution equality between cases and controls, and therefore also the adjustment for controls in H–W based on the global population, is only one (may be more than one) special permutation we obtain by changing the data in the table, therefore it is included in the analysis. Instead, at the moment, we have not completed the

likelihood-based solution with this adjustment because it needs to change the asymptotical distribution of the statistic.

References

M.N. Chiano, D.G. Clayton, Genotypic relative risks under ordered restriction. Genet. Epidemiol. **15**, 135–146 (1998)

R.L. Dykstra, S. Kochar, T. Robertson, Inference for likelihood ratio ordering in the two-sample problem. J. Amer. Statist. Assoc. **90**, 1034–1040 (1995)

H. El Barmi R. Dykstra, Maximum likelihood estimates via duality for log-convex models when cell probabilities are subject to convex constraints. Ann. Statist. **26**(5), 1878–1893 (1998)

H. El Barmi, R. Dykstra, Likelihood ratio test against a set of inequality constraints. J. Nonparametr. Statist. **11**, 233–250 (1999)

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

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

C. Hirotsu, A class of estimable contrasts in an age-period-cohort model. Ann. Inst. Statist. Math. **40**, 451–465 (1998)

G.M. Lathrop, Estimating genotype relative risks. Tissue Antigens **22**, 160–166 (1983)

B.-H. Liu, *Statistical Genomics: Linkage, Mapping, and QTL Analysis* (CRC-Press, Boca Raton, 2010)

F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software* (Wiley, Chichester, 2010)

Chapter 5
Power and Sample Size Simulations

Abstract In this chapter we perform an exhaustive simulation study focused on the nonparametric approach. At first we present some simulations for the nonparametric permutation solution by considering different types of population parameters and genetic models. The likelihood solution is not studied in depth because its asymptotical distribution is not yet known very well. Next we extend the considerations to comparisons between the nonparametric population-based methods and the sibship transmission disequilibrium test (S-TDT).

Keywords Codominant model · Dominant model · Recessive model

5.1 General Remarks

Many comparisons between population-based and family-based case–control studies have been done in the literature. In some cases the differences between parametric and nonparametric solutions have also been considered. Generally, the results are not definitive because the relations do not show a strong dominancy of one solution over another. However, one solution sometimes performs better than the other, depending on the considered sample size, the frequency of the rare allele in the population, and the disease's genetic model (dominant, codominant, recessive). In particular, the first type of comparison may be very difficult because association studies which take related controls (usually TDT or the S-TDT) into consideration must make more assumptions.

In any case, firstly we present some simulations for the nonparametric permutation solution illustrated in Sect. 4.1 by considering different types of population parameters and genetic models. The likelihood solution presented in Sect. 4.4 is not studied in depth because its asymptotical distribution is not known very well.

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_5, © Luigi Salmaso 2011

Next we extend the considerations to comparisons between the nonparametric population-based method (Sect. 4.1) and the sibship transmission disequilibrium test (S-TDT).

5.2 Simulations for Nonparametric Population-Based Solutions

We performed many simulations by 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 always set at 1,000 and the number of conditional Monte Carlo iterations (CMC-Iterations) is also 1,000. Simulations are performed by using one single locus with two alleles, one common and one rare, as in the problem presented in Sects. 2.4 and 4.1; $\alpha = 0.05$. In Fig. 5.1 we show the power simulations for the nonparametric solutions with cases = controls = 50 and rare allele frequency of 0.05.

In Fig. 5.2 we show the power simulations for the nonparametric solutions with cases = controls = 100 and rare allele frequency of 0.05.

In Fig. 5.3 we show the power simulations for the nonparametric solutions with cases = controls = 500 and rare allele frequency of 0.05.

In Fig. 5.4 we show the power simulations for the nonparametric solutions with cases = controls = 50 and rare allele frequency of 0.10.

In Fig. 5.5 we show the power simulations for the nonparametric solutions with cases = controls = 100 and rare allele frequency of 0.10.

In Fig. 5.6 we show the power simulations for the nonparametric solutions with cases = controls = 500 and rare allele frequency of 0.10.

As can be observed from the previous figures, the nonparametric solution's power is also very good for small sample sizes, and for low rare allele frequency. Of course, the situation where the rare allele is recessive is the worst, and generally the number of subjects needed to perform an analysis of association is very large.

In Fig. 5.7 we consider a simulation study for the permutation test with a rare allele frequency of 0.10 and significance level α fixed at 0.05. The rare allele is codominant and the odds are 2.

In Fig. 5.8 we consider simulations with a rare allele frequency of 0.10 and significance level $\alpha = 0.05$ for the case where the rare allele is dominant and the odds are 2.

As we can note from the figures, the nonparametric permutation solution displays very good power behaviour.

We cannot a priori know what the best type of test for a specific case is because it may depend on the specific data table. Furthermore, in a real data set the samples of cases and controls are not generally equal which means the solutions may be quite different with respect their power.

In any case, for small sample sizes, nonparametric tests should generally be preferred over their parametric counterparts.

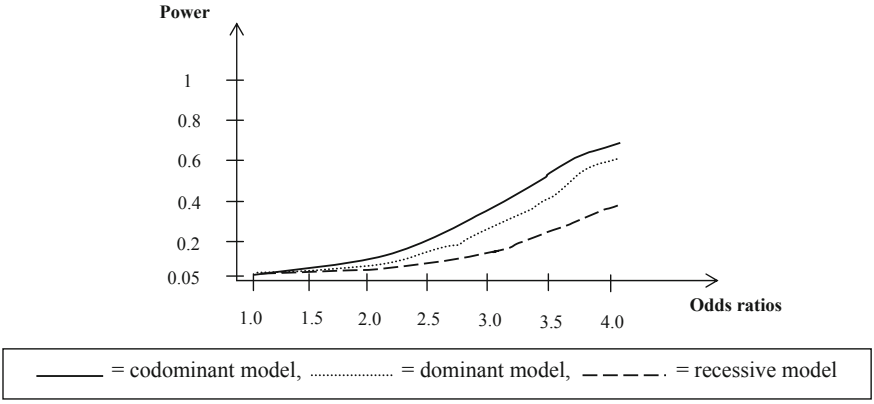

Fig. 5.1 Cases = controls = 50, $f = 0.05$

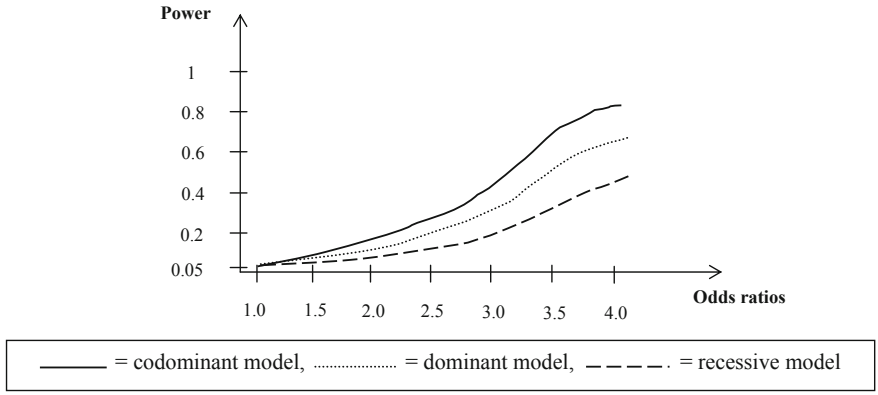

Fig. 5.2 Cases = controls = 100, $f = 0.05$

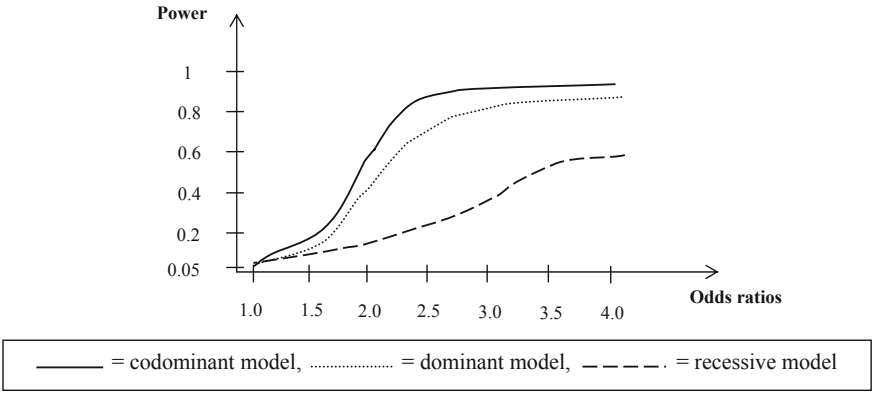

Fig. 5.3 Cases = controls = 500, $f = 0.05$

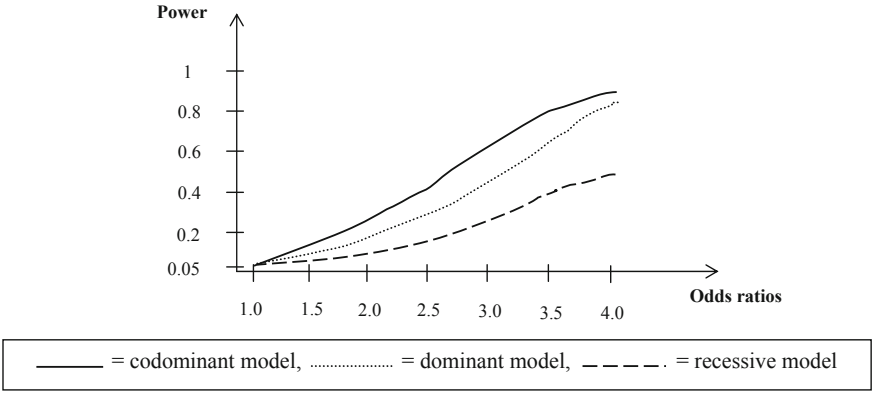

Fig. 5.4 Cases = controls = 50, $f = 0.10$

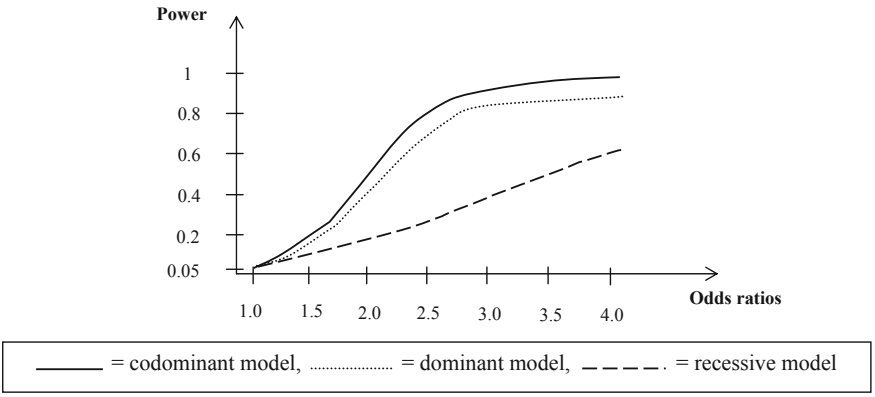

Fig. 5.5 Cases = controls = 100, $f = 0.10$

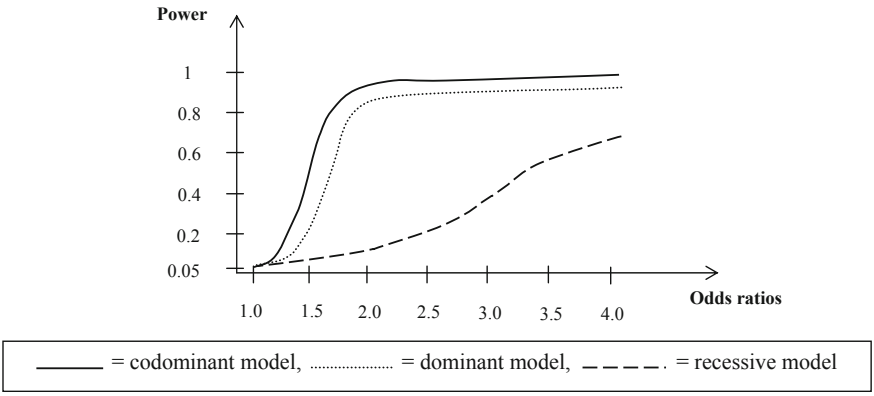

Fig. 5.6 Cases = controls = 500, $f = 0.10$

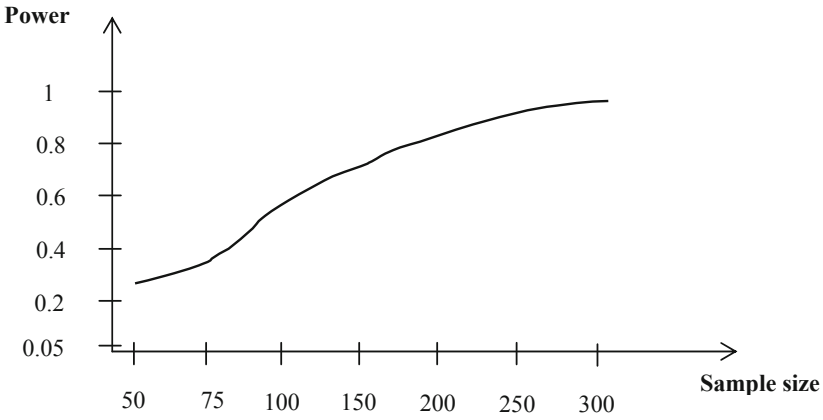

Fig. 5.7 Codominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$

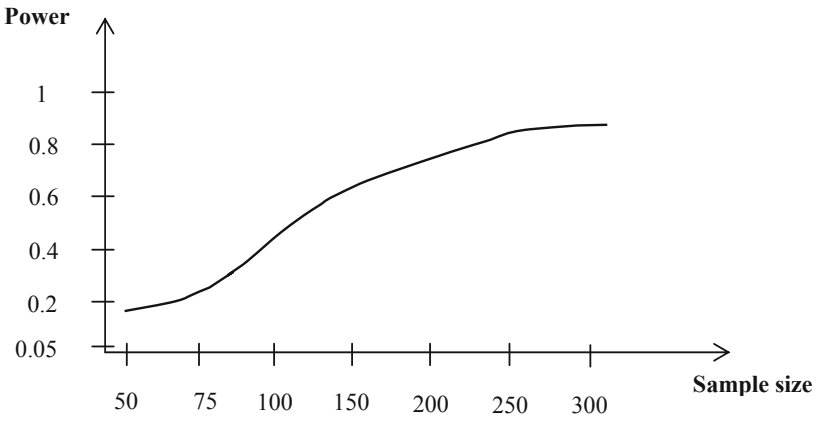

Fig. 5.8 Dominant model with odds ratios = 2, $f = 0.10$, $\alpha = 0.05$

5.3 Comparison Between S-TDT and Population-Based Methods

As reported by Schaid and Rowland (1998), Table 5.1 displays the number of sibships required to have 80% power, under an assumption of 5% false-positive rate, to detect association with a genetic autosomal dominant locus disease.

We performed the simulations with the permutation test (1,000 permutations) using the same parameters for allele frequencies, odds ratios, significance level and assuming the dominant model. The results are shown in Table 5.2.

Table 5.1 Number of sibships required to have 80% power (Schaid and Rowland 1998)

Rare allele frequency (*f*)	Odds ratios	Number of sibships for single locus test with one sib
0.1	2	257
0.1	4	59
0.2	2	212
0.2	4	57

Table 5.2 Permutation test results

Rare allele frequency (*f*)	Odds ratios	Number of cases = number of controls	Power of permutation test for single locus test
0.1	2	257	0.829
0.1	4	59	0.821
0.2	2	212	0.811
0.2	4	57	0.831

From this study we can conclude that population-based association studies may be more powerful with same sample sizes. Furthermore, in using related controls, there is the big problem of finding a sufficient number of individuals to achieve a power equivalent to that of methods with unrelated controls. This could be very expensive.

Reference

J.D. Schaid, C. Rowland, Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am. J. Hum. Genet. **61**, 1492–1506 (1998)

Chapter 6
Case Study

Abstract This chapter reviews a well-known case study from the international biomedical literature. It is well-known that high blood levels of coagulation factor VII represent a risk factor for ischaemic cardiovascular disease. This is supported by many studies from the past decade. Here we refer to a couple of papers which investigates whether the risk of myocardial infarction is associated with polymorphisms in the factor VII gene and whether these polymorphisms are associated with factor VII levels. The results obtained using nonparametric combination of dependent permutation tests, which can refer to complex hypothesis systems and isotonic alternatives, support and confirm the results from the literature, allowing the practitioner to use a more powerful and insightful tool for statistical inference.

Keywords Ischaemic Cardiovascular Disease · Factor VII Gene · Permutation Test

6.1 Background

High blood levels of coagulation factor VII represent a risk factor for ischaemic cardiovascular disease. This is supported by many studies from the past decade. The Northwick Park Heart Study group reported that high levels of factor VII were independently associated with an increase in the risk of coronary events in middle-aged men (Meade et al. 1986) and (1993) showed that the level of factor VII was predictive of the risk of fatal but not nonfatal myocardial infarction. The same trend was observed in the Prospective Cardiovascular Münster Study (Heinrich et al. 1994). Other investigations, however, failed to find such associations (Cortellaro et al. 1992; Vaziri et al. 1992). Factor VII blood levels are influenced by both environmental and genetic factors such as age, body-mass index, use of oral contraceptive, etc. (see Iacoviello et al. 1998). Here we refer to works by

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_6, © Luigi Salmaso 2011

Iacoviello et al. (1998) and Di Castelnuovo et al. (2000) who investigate whether the risk of myocardial infarction is associated with polymorphisms in the factor VII gene and whether these polymorphisms are associated with factor VII levels (see Iacoviello et al. 1998). They also consider the role of the decanucleotide insertion/deletion functional polymorphism in the promoter region of the factor VII gene and of possible interactions of the promoter with the HVR4 intron polymorphism (see Di Castelnuovo et al. 2000).

6.2 Study Population

Iacoviello et al. (1998) and Di Castelnuovo et al. (2000) studied patients with a family history of thrombosis. Case patients were persons over 45 years who had a myocardial infarction and who were reported to have at least one first-degree relative who had a myocardial infarction or a stroke (or both) before the age of 65. The patients were selected from the trial population of the "Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico" (GISSI) on the basis of an interview regarding their family history of thrombosis.

Controls were consecutive patients over the age of 45 who were hospitalized for any clinical reason other than myocardial infarction, stable or unstable angina, stroke, or transient ischaemic attacks. Furthermore, they could not report a personal or family history of thrombosis or have definite defects of the haemostatic system. For more details, see Iacoviello et al. (1998).

6.3 Laboratory Measurements and Techniques

The cases were assessed 5–7 months after their most recent ischaemic event. Blood was collected from the patients between 8 and 10 am after an overnight fast and after a 20 min rest in the supine position. Blood was not collected from patients who were receiving oral anticoagulant drugs. Blood samples for DNA and biochemical analyses were available from 165 of 171 cases and 225 of 272 controls.

Three polymorphisms for the factor VII gene were studied:

1. Hypervariable region 4 of intron 7 of the factor VII gene was amplified (see Iacoviello et al. 1998) and three alleles were identified: a common allele (H6) of 443 bp with six monomers, a less frequent allele (H7) of 480 bp with seven monomers of 37 bp, and a very rare allele (H5) of 406 bp with five monomers.
2. Two fragments were detected for the R353Q polymorphism: the most common (allele R) of 205 bp and the rarest (allele Q) of 272 bp.
3. For the promoter region they considered the normal allele (1) and the rare allele (2) with the deletion.

Table 6.1 Case–control table for R353Q

R353Q	Cases	Controls	Total
QQ	1	10	11
RQ	49	76	125
RR	114	138	252
Total	164	224	388

Table 6.2 Case–control table for the promoter region

Promoter region	Cases	Controls	Total
22	1	4	5
12	42	75	117
11	119	140	259
Total	162	219	381

For more details, see Iacoviello et al. (1998) and Di Castelnuovo et al. (2000).

6.4 Results of Statistical Analyses

From the data, we obtained the following case–control tables for the polymorphisms of factor VII loci (Tables 6.1, 6.2 and 6.3).

Table 6.1 shows the data for the R353Q polymorphism, where R is the most common allele: here there is one missing in both cases and controls. Table 6.2 shows the data for the promoter region polymorphism of the factor VII gene: in this case the more common allele is allele 1 and there are three missing data in the cases and six in the controls. Table 6.3 shows the data for hypervariable region 4, where H6 is the most common allele while H7 and H5 are the rarest alleles.

Only the data for Table 6.2 are not in H–W equilibrium, whereas in Tables 6.1 and 6.3, they follow the H–W law; furthermore, it has been proven that R353Q and hypervariable region 4 polymorphisms are in linkage disequilibrium (Iacoviello et al. 1998).

Consider the odds ratios for myocardial infarction with each of the three tables below (see Tables 6.4, 6.5 and 6.6) by imposing too "perfect" H–W equilibrium in the controls.

If we now perform the simple permutation test illustrated in Sect. 4.1 by looking at the odds ratios in Tables 6.4 and 6.5, we obtain (by using 10,000 CMC-iterations and Liptak's nonparametric combination) the p-values represented in Table 6.7. In the same way, Table 6.7 presents the results with the multiallelic permutation test from Sect. 4.5.

Results are significant for the Q allele of the R353Q polymorphism at the level $\alpha = 0.05$. We can say that the Q allele has a protective effect (Odd < 1) against myocardial infarction. Furthermore, by looking at the two estimated relative risks, we may consider this a codominant effect. The equivalence of odds ratios between

Table 6.3 Case–control table for hypervariable region 4

Hypervariable Region 4	Cases	Controls	Total
H7H7	12	31	43
H6H7	60	97	157
H6H6	84	94	118
H7H5	4	1	5
H6H5	5	2	7
Total	165	225	390

Table 6.4 Odds ratios for R353Q by imposing H–W equilibrium in the controls

R353Q	Estimated relative risks
Odd QQ/RQ	0.155
Odd QQ/RQ in H–W	0.153
Odd RQ/RR	0.780
Odd RQ/RR in H–W	0.797

Table 6.5 Odds ratios for the promoter by imposing H–W equilibrium in the controls

Promoter	Estimated relative risks
Odd 22/12	0.446
Odd 22/12 in H–W	0.199
Odd 12/11	0.659
Odd 12/11 in H–W	0.759

Table 6.6 Odds ratios for hypervariable region 4 by imposing H–W equilibrium in the controls

Hypervariable region 4	Estimated relative risks
Odd 77/67	0.626
Odd 77/67 in H–W	0.729
Odd 67/66	0.692
Odd 67/66 in H–W	0.644
Odd 55/65	Infinity
Odd 55/65 in H–W	Infinity
Odd 65/66	2.798
Odd 65/66 in H–W	2.738

observed and H–W adjusted controls supports the above observation that controls are in H–W equilibrium.

The *p*-value for the protective region polymorphism is not significant for the observed data, but is 0.036 with the controls adjusted by H–W equilibrium. However, we cannot infer (the allele 2) produces a protective effect. Finally, we can note the strong protective effect of the allelic form 7 (allele 7) in the hypervariable region 4 polymorphism of the factor VII region. This allele is also protective against myocardial infarction and the effect seems to be codominant. Allele 5 does not show any particular effect.

We now consider the analysis of data using the multiloci extension of the permutation test illustrated in Sect. 4.5. The three following tables represent the analysis of haplotypes: hypervariable and promoter region polymorphisms

Table 6.7 *P*-Values of the permutation test

Test	*p*-Value
R353Q:{(Odd QQ/RQ < 1) ∩ (Odd RQ/RR < 1) XOR (Odd QQ/RQ > 1) ∩ (Odd RQ/RR > 1)}	0.013
R353Q in exact H–W equilibrium	0.009
Promoter:{(Odd 22/12 < 1) ∩ (Odd 12/11 < 1) XOR (Odd 22/12 > 1) ∩ (Odd 12/11 > 1)}	0.129
Promoter in exact H–W equilibrium	0.036
Hypervariable:{(Odd 77/67 < 1) ∩ (Odd 67/66 < 1) XOR (Odd 77/67 > 1) ∩ (Odd 67/66 > 1)}	0.015
Hypervariable region for 7 allelic form (allelic form 7/allele 7/7th allelic form) in exact H–W equilibrium	0.016
Hypervariable:{(Odd 55/65 < 1) ∩ (Odd 65/66 < 1) XOR (Odd 55/65 > 1) ∩ (Odd 65/66 > 1)}	0.250
Hypervariable region for allelic form 5 (allele 5) in exact H–W equilibrium	0.250

Table 6.8 Case–control table for hypervariable and promoter region polymorphism

Hypervariable and promoter	Cases	Controls	Total
66-11	75	74	149
66-12	7	18	25
66-22	0	0	0
67-11	37	55	92
67-12	21	38	59
67-22	1	1	2
77-11	2	8	10
77-12	10	19	29
77-22	0	3	3
65-11	3	2	5
65-12	2	0	2
65-22	0	0	0
55-11	0	0	0
55-12	0	0	0
55-22	0	0	0
75-11	2	1	3
75-12	2	0	2
75-22	0	0	0
Total	162	219	381

(Table 6.8), hypervariable region and R353Q polymorphism (Table 6.9), promoter region and R353Q polymorphism (Table 6.10).

We studied the association of each locus by conditioning on the other loci, and we used the same isotonic system of hypotheses as above. We performed a permutation test to obtain combined *p*-values which test each pair of odds ratios for one locus conditional on a specific genotype of the other loci. We then performed permutation tests using 10,000 CMC-iterations and Liptak's combining function

Table 6.9 Case–control table for hypervariable region and R353Q polymorphisms

Hypervariable and R353Q	Cases	Controls	Total
66-RR	77	78	155
66-RQ	6	15	21
66-QQ	0	1	1
67-RR	28	50	78
67-RQ	32	45	77
67-QQ	0	2	2
77-RR	3	7	10
77-RQ	8	16	24
77-QQ	1	7	8
65-RR	4	2	6
65-RQ	1	0	1
65-QQ	0	0	0
55-RR	0	0	0
55-RQ	0	0	0
55-QQ	0	0	0
75-RR	2	1	3
75-RQ	2	0	2
75-QQ	0	0	0
Total	164	224	388

Table 6.10 Case–control table for promoter region and R353Q polymorphisms

Promoter and R353Q	Cases	Controls	Total
11-RR	107	126	233
11-RQ	12	14	26
11-QQ	0	0	0
12-RR	6	7	13
12-RQ	35	62	97
12-QQ	1	6	7
22-RR	0	0	0
22-RQ	1	0	1
22-QQ	0	4	4
Total	162	219	381

(we used Liptak's combination function here simply because it seems to be more powerful for the present case study). In this case we did not adjust the controls for the H–W law because there is linkage disequilibrium between the two loci and we are studying their possible associated effect.

From the data in Table 6.8 we obtained six p-values (we do not report the results that include the hypervariable region 4 polymorphism with allele 5 because the relative sample size is too low):

1. The first p-Value jointly tests the two odds ratios of the promoter region polymorphism conditionally on (the genotypic form 66) of the hypervariable

region 4 polymorphism. We call it p-value (66-..) and it is equal to 0.054, not significant at level $\alpha = 0.05$, so we cannot say that the allele 2 in the promoter region polymorphism presents one particular variation of effect if it is associated to form 66 of the hypervariable region 4 polymorphism.
2. p-Value (67-..) is 0.887; we accept the null hypothesis that there is no variation of effect with allele 2 in the promoter region polymorphism conditionally on genotype 67 of the hypervariable region 4 polymorphism.
3. p-Value (77-..) is 0.984; we accept the null hypothesis that there is no variation of effect with allele 2 in the promoter region polymorphism conditionally on genotype 77 of the hypervariable region 4 polymorphism.
4. p-Value (7.-11) is 0.024; we have significant evidence that there is an increasing protective effect (the two odds are 0.37 and 0.66) with allele 7 in the hypervariable region 4 polymorphism conditionally on genotype 11 of the promoter region polymorphism.
5. p-Value (7.-12) is 0.621; we accept the null hypothesis that there is no variation of effect with allele 7 in the hypervariable region 4 polymorphism conditionally on genotype 12 of the promoter region polymorphism.
6. p-Value (7.-22) is 0.603; we accept the null hypothesis that there is no variation of effect with allele 7 in the hypervariable region 4 polymorphism conditionally on genotype 22 of the promoter region polymorphism.

From the data in Table 6.9 we obtained six p-values (we do not report the results that include the hypervariable region 4 polymorphism with allele 5 because it is not interesting and the sample size is too low):

1. p-Value (66-..) is 0.049; it is slightly significant at level $\alpha = 0.05$, so we could say that the Q allele in the R353Q polymorphism presents an increasing protective effect (odds ratios are 0 and 0.40) if it is associated to form 66 of the hypervariable region 4 polymorphism. In any case, by also looking at the small sample size of the 66-RR haplotype, we need further analyses to confirm this result.
2. p-Value (67-..) is 0.877; we accept the null hypothesis that there is no variation of effect with the Q allele in the R353Q polymorphism conditionally ongenotype 67 of the hypervariable region 4 polymorphism.
3. p-Value (77-..) is 0.383; we accept the null hypothesis that there is no variation of effect with the Q allele in the R353Q polymorphism conditionally ongenotype 77 of the hypervariable region 4 polymorphism.
4. p-Value (7.-RR) is 0.043; we have a slightly significant result that shows there is an increasing protective effect (odds ratios are 0.76 and 0.56) with allele 7 in the hypervariable region 4 polymorphism conditionally ongenotype RR of the R353Q polymorphism.
5. p-Value (7.-RQ) is 0.841; we accept the null hypothesis that there is no variation of effect with allele 7 in the hypervariable region 4 polymorphism conditionally ongenotype RQ of the R353Q polymorphism.

Table 6.11 Case–control data for the R353Q polymorphism jointly with the smoking stratum variable

R353Q	Smoking = No Cases	Smoking = No Controls	Smoking = Yes Cases	Smoking = Yes Controls
RR	27	60	87	53
RQ	17	42	32	25
QQ	0	1	1	6
Total	44	103	120	84

6. *p*-Value (7.-QQ) is 0.734; we accept the null hypothesis that there is no variation of effect with allele 7 in the hypervariable region 4 polymorphism conditionally on genotype QQ of the R353Q polymorphism.

The six *p*-values for Table 6.10 are: *p*-value (11-..) = 0.998, *p*-value (12-..) = 0.158, *p*-value (22-..) = 0.000, *p*-value (..-RR) = 0.981, *p*-value (..-RQ) = 0.723, *p*-value (..-QQ) = 0.076. Only the *p*-value that tests for a possible variation of effect with the Q allele in the R353Q polymorphism conditionally on genotype 22 of the promoter region polymorphism shows significant evidence for that, but its result is strongly conditioned on the small sample size (see lines 22-RR, 22-RQ and 22-QQ in Table 6.9). From these results we can say that allele 7 in the hypervariable region 4 polymorphism and allele Q in the R353Q polymorphism of the factor VII region have a protective (and codominant) effect against myocardial infarction. Furthermore, the protective effect of allele 7 may increase if in the haplotype of the factor VII region we have genotype 22 of the promoter region polymorphism. These results agree with those obtained in Iacoviello et al. 1998. We also performed the permutation test illustrated in Sect. 4.6 by considering the confounding factor smoking to test if there is a difference of interaction between subjects under different levels of the confounding factor: in fact people are subdivided in the data into smokers and non-smokers. If we look, therefore, at Table 6.11, we see the case–control data for the R353Q polymorphism jointly with the smoking stratum variable. We then test if there is an interaction effect between smoking and the R353Q polymorphism such that there is a different level of protection for the R353Q polymorphism between the two levels: Smoking = No and Smoking = Yes. We obtained a combined *p*-value of 0.000 with 10,000 CMC-iterations using Liptak's combining function. This significant result means that the allelic variant Q of the R353Q polymorphism has a protective effect that increases in people who are smokers with respect to non smokers.

In the same way, Table 6.12 shows the joined data for the hypervariable region 4 polymorphism and the smoking stratum variable. Here we have a combined *p*-value of 0.002 so we can conclude the protective effect of allelic variant 7 of hypervariable region 4 polymorphism has an increasing protective effect when people are smokers. Tests on allelic variant 5 are not reported because they are not of interest.

Results in Table 6.13 for the promoter region are not significant.

Table 6.12 Case–control data for the hypervariable region 4 polymorphism and the smoking stratum variable

Hypervariable	Smoking = No Cases	Smoking = No Controls	Smoking = Yes Cases	Smoking = Yes Controls
66	24	41	60	39
67	17	50	43	29
77	4	12	8	15
65	0	0	5	1
55	0	0	0	0
75	0	1	4	0
Total	45	104	120	84

Table 6.13 Case–control data for the promoter polymorphism and the smoking stratum variable

Promoter	Smoking = No Cases	Smoking = No Controls	Smoking = Yes Cases	Smoking = Yes Controls
11	33	59	86	58
12	10	41	32	23
22	0	1	1	2
Total	43	101	119	83

Table 6.14 Case–control data for the association between the presence or absence of the rare allele

R353Q	Smoking = No Cases	Smoking = No Controls	Smoking = Yes Cases	Smoking = Yes Controls
RR	27	60	87	53
RQ and QQ	17	43	33	31
Total	44	103	120	84

We can repeat the previous analyses for Table 6.11 by looking at the association between the presence or absence of the rare allele in cases and controls. The results of these analyses are shown in Table 6.14. We obtained more differences between case and control groups in smokers than in non smokers.

The *p*-values for permutation tests relative to the odds ratios of data groups for Smoking = No and Smoking = Yes are, respectively, 0.082 (odds = 0.878) and 0.001 (odds = 0.648), which suggests the protective effect of allele R353Q is active when we consider people that smoke.

These results obtained using nonparametric combination of dependent permutation tests, which can refer to complex hypothesis systems and isotonic alternatives, support and confirm the results obtained by Iacoviello and Donati (1999) and Di Castelnuovo et al. (1999), who applied a more simple parametric approach.

References

M. Cortellaro, C. Boschetti, E. Cofrancesco, C. Zanussi, M. Catalano, G. de Gaetano, L. Gabrielli, B. Lombardi, G. Specchia, L. Tavazzi, The PLAT Study: hemostatic function in relation to atherothrombotic ischemic events in vascular disease patients: principal results. Arterioscler. Thromb. **12**, 1063–1070 (1992)

A. Di Castelnuovo, A. D'Orazio, C. Amore, A. Falanga, M.B. Donati, L. Iacoviello, The decanucleotide insertion/deletion polymorphism in the promoter region of the coagulation factor VII gene and the risk of familial myocardial infarction. Thromb. Res. **98**, 9–17 (2000)

A. Di Castelnuovo, D. Mazzaro, F. Pesarin, L. Salmaso, Multidimensional permutation testing for isotonic inference: an application to genetics. Volume of Abstracts. *International Biometric Society Italian Region*, (Roma, Italy, 1999), pp. 7–9

J. Heinrich, L. Balleisen, H. Schulte, G. Assmann, J. van de Loo, Fibrinogen and factor VII in the prediction of coronary risk: results from the PROCAM study in healthy men. Arterioscler. Thromb. **14**, 54–59 (1994)

L. Iacoviello, M.B. Donati, Gene-environment interactions: implications for cardiovascular disease. Cardiologia **44**, 227–232 (1999)

L. Iacoviello, A. Di Castelnuovo, P. De Knijff, A. D'Orazio, C. Amore, R. Arboretti, C. Kluft, B.M. Donati, Polymorphisms in the coagulation factor VII gene and the risk of myocardial infarction. N. Engl. J. Med. **338**, 79–85 (1998)

T.W. Meade, S. Mellows, M. Brozovic, G.J. Miller, R.R. Chakrabarti, W.R. North, A.P. Haines, Y. Stirling, J.D. Imeson, S.G. Thompson, Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study. Lancet **2**, 533–537 (1986)

T.W. Meade, V. Ruddock, Y. Stirling, R. Chakrabarti, G.J. Miller, Fibrinolytic activity, clotting factors, and long-term incidence of ischaemic heart disease in the Northwick Park Heart Study. Lancet **342**, 1076–1079 (1993)

N.D. Vaziri, S.C. Kennedy, D. Kennedy, E. Gonzales, Coagulation, fibrinolytic, and inhibitory proteins in acute myocardial infarction and angina pectoris. Am. J. Med. **93**, 651–657 (1992)

Chapter 7
Conclusions

Abstract This final chapter concerns some discussion to emphasize the role of nonparametric combination as a flexible methodology for solving complex problems. These complex testing problems are not adequately taken into consideration in the standard literature. This is in spite of the fact that they are very frequently encountered in a great variety of practical applications. These problems emphasize the versatility and effectiveness of the nonparametric combination methodology. It should also be stressed that since permutation tests are conditional with respect to a set of sufficient statistics, the nonparametric combination, in very mild conditions, frees the researcher from the necessity to model the dependence relations among responses. Some future research guidelines using permutation tests are also presented.

Keywords Complex problems · Nonparametric combination · Permutation tests

7.1 Statistical Methodology

In conclusion, we would like to emphasize the role of nonparametric combination as a flexible methodology for solving complex problems. These complex testing problems are not adequately taken into consideration in the standard literature. This is in spite of the fact that they are very frequently encountered in a great variety of practical applications. These problems emphasize the versatility and effectiveness of the nonparametric combination methodology.

It should also be stressed that because permutation tests are conditional with respect to a set of sufficient statistics, the nonparametric combination, in 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 behaviour of combined tests is similar to that of their best parametric counterparts, in the conditions for the latter.

L. Salmaso et al., *Permutation Testing for Isotonic Inference on Association Studies in Genetics*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-20584-2_7, © Luigi Salmaso 2011

The nonparametric combination of dependent permutation partial tests is a method for the combination of significance levels or rejection probabilities. Conversely, the way generally followed by most parametric tests, based for instance on likelihood ratio behaviour, essentially corresponds to the combination of discrepancy measures usually expressed by distance of points in sample space χ. In this sense, this method appears as a substantial extension of standard parametric approaches. Further, in the presence of a stratification variable, the nonparametric combination, through a multi-phase procedure, allows for flexible solutions. For instance, we may first combine partial tests with respect to variables within each stratum and then combine the *combined* test with respect to strata. Alternatively, we may first combine partial tests related to each variable with respect to strata, and then combine the *combined* tests with respect to the variables.

As a final remark, in very mild conditions, the nonparametric combination method may be considered as a way of reducing the degree of complexity for most testing problems. We saw these characteristics in the genetic problem discussed, when we perform partial tests on the different odds ratios obtained from the same case–control data, and then calculate the combined test by using the first p-values. For case–control association studies we used these permutation methods but we also introduced other parametric likelihood methods. We did not perform power comparisons between these two types of tests because we cannot obtain a good approximation of the asymptotical distribution for the likelihood test. In literature, there are many parametric methods and often their properties may differ depending on the assumptions made. As we saw, permutation tests perform well even for small sample sizes. This is important because often for complex and late onset diseases, it is very difficult to collect a lot of data. Furthermore, by comparing the population-based nonparametric association method with the sibships-based test, we observe that the power of the two tests is quite similar: this might imply that family-based methods are rather expensive due to the need for extensive data collection. Of course, the results obtained by case–control studies, based on unrelated controls, have always been suspicious because we cannot include some particular confounding factors in the analysis, or the association may be spurious due to population stratification effects.

Despite this, we believe that techniques using population-based association studies may still help gene location of complex diseases, and they should be used in combination with traditional genetic studies based on family data. Perhaps the main advantage in using population-based association studies is that we do not need to know the exact parameters of the genetic model. Certainly a characteristic of permutation tests is that they can achieve good power even when using only a few data, as shown with the simulations performed in Chap. 5. The association study presented in Sects. 2.4 and 4.1–4.4, performed by analyzing the odds ratios, has the merit from a methodological point of view of testing the data by also taking into consideration particular biological models. This test is done on both likelihood-based and permutation methods by using case–control studies. However, the permutation tests do not need explicit expression of the likelihood

ratio. This is very important because this way it can be extended to more difficult analyses involving multiallelic and multiloci problems. Furthermore, in the study of genetic traits associated to complex diseases, confounding factors represent an important aspect of the analysis, and permutation tests can be used to perform the analyses which consider these aspects. Finally, but no less importantly, nonparametric permutation tests allow us to consider missing data in the analyses (see for example Piepho (2005) and Pesarin and Salmaso (2010), paragraph 7.5, p. 232). A software to perform permutation tests in genetic studies is available at: www.salmasowigi.it.

7.2 Future Work Prospects

The applications of nonparametric tests proposed in this book suggest there is no need to abandon population-based association studies for genetic and epidemiologic studies, using evidence obtained from family-based studies. It would be helpful to collect data for complex diseases that may be used in suitable TDT-type tests and population-based analyses, in order to compare the two solutions on the same data for both power and limit aspects. The next step of this analysis would be to use nonparametric permutation methods in family-based studies, perhaps involving complex hypothesis systems which need a nonparametric combination methodology similar to the isotonic alternative we saw in Sect. 4.1.

In actual fact, it seems weak to rely only on association-based methods. Additionally, the S-TDT, even if it does utilize some linkage information, is highly inefficient in the absence of linkage disequilibrium. Since linkage disequilibrium is so unpredictable, studies that are dependent upon it will often fail. The goal of typing multiple polymorphisms (up to three) within candidates does not guarantee LD between the QTL and the chosen markers. For this reason, the power calculations are meaningless since they assume complete LD between the marker and the QTL. In fact, we may not have power to detect even the QTLs with the largest effects because LD-based methods do not find genes in order of their relative importance in the population. It is highly likely that the underlying QTLs will be multiallelic thus rendering all LD-based approaches with low power. However, if we collect sibships, there is a large amount of linkage information to be exploited (jointly with the available LD), and this requires the use of more powerful linkage-based methods (we believe it is more appropriate to perform powerful linkage-based methods as well, such as the variance component method, which can be used for both quantitative and discrete traits).

For the multivariate case we also need to consider some comparisons among different permutation tests, in particular by looking at the Type-I error levels and power.

Acknowledgments Authors wish to thank the University of Padova (CPDA092350/09) and the Italian Ministry for University and Research (2008WKHJPK/002) for providing the financial support for this research.

References

H.P. Piepho, Permutation tests for the correlation among genetic distances and measures of heterosis. Theor. Appl. Genet. **111**, 95–99 (2005)

F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software* (Wiley, Chichester, 2010)