May 2019

A Review of Statistical Analysis of Genetic Case-Control Data

Jin Zhang
zhang1mg@uwindsor.ca

A Review of Statistical Analysis of Genetic Case-Control Data

by

Jin Zhang

A Major Research Paper
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

A Review of Statistical Analysis of Genetic Case-Control Data

by

Jin Zhang

APPROVED BY:

M. Hlynka
Department of Mathematics and Statistics

S. Paul, Advisor
Department of Mathematics and Statistics

April 22, 2019

Declaration of Originality

I hereby certify that I am the sole author of this major paper and that no part of this major paper has been published or submitted for publication.

I certify that, to the best of my knowledge, my major paper does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my major paper, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my major paper and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my major paper, including any final revisions, as approved by my major paper committee and the Graduate Studies office, and that this major paper has not been submitted for a higher degree to any other University or Institution.

Abstract

This paper considers the analysis of genetic case-control data. We consider the allele frequency in cases and controls. Because each individual has two alleles at any autosomal locus, there will be twice as many alleles as people in the allele distribution. Simultaneously, the serological distribution is bulit by ignoring the difference between homozygous and herterozygous. We also consider the marker loci with multiple alleles. Traditional case-control studies provide a powerful and efficient method for evaluation of association between candidate gene and disease. There has been debate on how the power of tests for association changes with different allelic effect. To facilitate the design of association studies, we present power and sample size formulas for Armitage's test for trend applied to case-control studies of candidate genes.

To my loving parents

Liping Zhang and Juanfang Xing

Acknowledgments

I want to give my sincere gratitude to Dr. Paul for all his constant support and guidance during my graduate study since I started my M.Sc. program. His guidance and advice support me a lot to complete this major paper. All the support and suggestions from him made me finally succeed in my master's study.

I would also like to thank Dr. Myron Hlynka for being my department reader. He also gave me a lots of advice on statistical study.

Contents

List of Tables

Chapter 1

Introduction

Classical case-control studies are improtant in genetic epidemiology, even though they can only establish an association and other designs are necessary to determine whether such associations are causal. Chi-square tests based on simple contingency tables are useful tools for the association mapping of disease genes. These tables (Table 2.2 and Table 2.3) consist of rows representing those affected with the disease (cases) and those not affected (controls). Columns are either alleles or genotype at the genetic markers of interest. In this paper, we show the methods of testing for associations between biallelic markers and disease status, the Cochran-Armitage linear trend test and the allele test. Further, we show extension from biallelic to multiallelic trend test.

1.1 What is genotype?

Genotype: In a broad sense, the term 'genotype' refers to the genetic makeup of an organism; in other words, it describes an organism's complete set of genes. In a more narrow sense, the term can be used to refer to the alleles, or variant forms of a gene, that are carried by an organism. Humans have diploid organisms, which means that they have two alleles at each genetic position, or locus, with one allele inherited from each parent. Each pair of alleles represents the genotype of a specific gene.

1.2 How to connect statistical tests with disease genes?

Classical case-control studies are important in gentic epidemiology, even though they can only establish an association and other designs are necessary to determine whether such associations are causal. These tables (Table 2.2 and Table 2.3)consist of rows representing those affected with the disease (cases) and those not affected (controls), and columns representing either alleles or genotypes at the genetic markers of interest. Chi-square tests based on simple contingency tables are useful tools for the association mapping of disease genes.

Chapter 2

Definitions used in this paper

2.1 Definitions in Biology

Definition 2.1.1. Complete Dominance: In complete dominance, the effect of one allele in a heterozygous genotype completely masks the effect of the other. The allele that masks the other is said to be dominant to the latter, and the allele that is masked is said to be recessive to the former. Complete dominance, therefore, means that the phenotype of the heterozygote is indistinguishable from that of the dominant homozygote.

Definition 2.1.2. Incomplete Dominance: Incomplete dominance (also called partial dominance, semi-dominance or intermediate inheritance) occurs when the phenotype of the heterozygous genotype is distinct from and often intermediate to the phenotypes of the homozygous genotypes. For example, the snapdragon flower color is homozygous for either red or white. When the red homozygous flower is paired with the white homozygous flower, the result yields a pink snapdragon flower.

Definition 2.1.3. Codominance: Co-dominance occurs when the contributions of both alleles are visible in the phenotype. For example, in the ABO blood group system, the IA and IB alleles produce different modifications. Thus IAIA and IAi individuals both have type A blood, and IBIB and IBi individuals both have type B blood, but IAIB individuals have both modifications on their blood cells and thus have type AB blood, so the IA and IB alleles are said to be co-dominant.

Definition 2.1.4. Hardy-Weinberg Equilibrium: The Hardy-Weinberg equilibrium is a principle stating that the genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors. When mating is random in a large population with no disruptive circumstances, the law predicts that both genotype and allele frequencies will remain constant because they are in equilibrium. The Hardy-Weinberg equilibrium can be disturbed by a number of forces, including mutations, natural selection, nonrandom mating, genetic drift, and gene flow. For instance, mutations disrupt the equilibrium of allele frequencies by introducing new alleles into a population. Similarly, natural selection and nonrandom mating disrupt the Hardy-Weinberg equilibrium because they result in changes in gene frequencies. This occurs because certain alleles help or harm the reproductive success of the organisms that carry them. Another factor that can upset this equilibrium is genetic drift, which occurs when allele frequencies grow higher or lower by chance and typically takes place in small populations. Gene flow, which occurs when breeding between two populations transfers new alleles into a population, can also alter the Hardy-Weinberg equilibrium. Because all of these disruptive forces commonly occur in nature, the Hardy-Weinberg equilibrium rarely applies in reality. Therefore, the Hardy-Weinberg equilibrium describes an idealized state, and genetic variations

in nature can be measured as changes from this equilibrium state. There are two equations necessary to solve a HWE:

- $p + q = 1$
- $p^2 + 2pq + q^2 = 1$

where,

p : is the frequency of the dominant allele.

q : is the frequency of the recessive allele.

p^2 : is the frequency of the individuals with the homozygous dominant genotype.

$2pq$: is then frequency of individuals with the heterzygous genotype.

q^2 : is the frequency of individuals with the homozygous recessive genotype.

2.2 Cochran Armitage test for trend

Definition 2.2.1. Cochran-Armitage test for trend: is used in categorical data analysis when the aim is to assess for the presence of an association between a variable with two categories and an ordinal variable with k categories. It modifies the Pearson chi-squared test to incorporate a suspected ordering in the effects of the k categories of the second variable. For example, doses of a treatment can be ordered as 'low', 'medium', and 'high', and we may suspect that the treatment benefit cannot become smaller as the dose increases. The trend test is often used as a genotype-based test for case-control genetic association studies. In Table 2.1, $R = r_0 + r_1 + r_2$, and

$n_0 = r_0 + s_0$. The trend test statistic is:

$$T = \sum_{i=0}^{k} t_i(r_i S - s_i R),$$

where the t_i are weights, and the difference $(r_i S - s_i R)$can be seen as the difference between r_i and s_i after reweighting the rows to have the same total.

Table 2.1: Completed with the marginal totals of the two variables

	$x_0 = 0$	$x_1 = 1$	$x_2 = 2$	Sum
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Sum	n_0	n_1	n_2	N

where $x_i = i$, $i = 0, 1, 2$, denote the number of alleles.

The corresponding chi-square statistics is Armitage's test for trend, which is equivalent to the score test for the covariate effect in the logistic model. Let x_i denote a score associated with each column of the table. The general form of Armitage's test is:

$$X_T^2 = \frac{N(N \sum r_i x_i - R \sum n_i x_i)^2}{R(N-R)N \sum n_i x_i^2 - (\sum n_i x_i)^2}$$

where $x_i = i$, $i = 0, 1, 2$, denote the number of alleles M.

2.3 Distributions under different allelic effect

2.3.1 Genotype distribution

The genotype distribution table (Table 2.2)presents the number of cases (r_i) and control (s_i) for negative (which has no disease gene), heterozygous (which has one disease gene) and homozygous (which has two disease genes).

Table 2.2: Genotype distribution

	Negative	Heterozygous	Homozygous	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

2.3.2 Allele Distribution

For the Allele Test, we focus on the number of each allele. Changing the data from genotype to gene, will double the sample size. The case and control table for allele will be a 2×2 tabel as in Table 2.3.

Table 2.3: Allele distribution

	$x_0 = 0$	$x_1 = 1$	Total
Case	$r_1 + 2r_2$	$r_1 + 2r_0$	$2R$
Control	$s_1 + 2s_2$	$s_1 + 2s_0$	$2S$
Total	$n_1 + 2n_2$	$n_1 + 2n_0$	$2N$

2.4 Cochran-Mantel-Haenszel Estimator

Definition 2.4.1. Cochran-Mantel-Haenszel Estimator: To explore and adjust for confounding, use a stratified analysis in which a series of two-by-two tables are set up, one for each stratum/category of the confounding variable. Compute a weighted average of the estimates of the risk ratios or odds ratios across the strata.

$$\begin{array}{cc} a_i & b_i \\ c_i & d_i \\ \hline & n_i \end{array}$$

The weighted average provides a measure of association that is adjusted for confounding. The weighted average for odds ratio is:

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{K}\left(\frac{a_i d_i}{n_i}\right)}{\sum_{i=1}^{K}\left(\frac{b_i c_i}{ni}\right)},$$

with $K = 1, 2, 3...$

In weighted form, the weighted average for odds ratio is:

$$\hat{OR}_{MH} = \sum_{i=1}^{K} w_i \frac{OR_i}{\sum_{i=1}^{K} w_i},$$

with $OR_i = \dfrac{a_i d_i}{b_i c_i}$ and $w_i = \dfrac{b_i c_i}{n_i}$, these two formulas are exactly same by plugging in the format of w_i.

2.5 Linkage Disequilibrium (LD)

Definition 2.5.1. In population genetics, linkage disequilibrium is the non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly. Linkage disequilibrium is influenced by many factors, including selection, the rate of genetic recombination, mutation rate, genetic drift, the system of mating, population structure, and genetic linkage. As a result, the pattern of linkage disequilibrium in a genome is a powerful signal of the population genetic processes that are structuring it.

Suppose that among the gametes that are formed in a sexually reproducing population, allele A occurs with frequency p_A at one locus (p_A is the proportion of gametes with A at that locus), while at a different locus allele B occurs with frequency p_B. Similarly, let p_{AB} be the frequency with which both A and B occur together in the same gamete (p_{AB} is the frequency of the AB haplotype).

The association between the alleles A and B can be regarded as completely random—which is known in statistics as independence—when the occurrence of one does not affect the occurrence of the other, in which case the probability that both A and B occur together is given by the product $p_A p_B$ of the probabilities. There is said to be a linkage disequilibrium between the two alleles whenever p_{AB} differs from $p_A p_B$ for any reason.

The level of linkage disequilibrium between A and B can be quantified by the

coefficient of linkage disequilibrium D_{AB}, which is defined as:

$$D_{AB} = p_{AB} - p_A p_B,$$

provided that both p_A and p_B are greater than zero. Linkage disequilibrium corresponds to $D_{AB} \neq 0$.

Chapter 3

Extension from genotype to genes

In different genetic markers of interest, there are three tables (Genotype distribution, Allele distribution and Serological distribution). The chi-square test statistics and the odds ratios were developed in this chapter. We consider which tabulation is the most appropriate in different allelic effects.

3.1 Tests under Complete Dominance

3.1.1 Genotype distribution under Complete Dominance

Table 3.1 presents the number of cases and control for negative (which has no marker allele), heterozygous (which has one marker allele) and homozygous (which has two marker allele).

Table 3.1: Genotype distribution

	Negative	Heterozygous	Homozygous	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

In genotype distribution, the trend statistic X_G^2 for the 2×3 genotype tabel (Table 2.3) is:

$$X_G^2 = \frac{N(N(r_1+2r_2)-R(n_1+2n_2))^2}{R(N-R)[N(n_1+4n_2)-(n_1+2n_2)^2]}$$

3.1.2 Allele distribution under Complete Dominance

Since each heterozygous person has one copy of the marker allele and each homozygous person has two copies, one can produce an allele table(e.g, Apple et al., 1994, 1995; Odunsi et al., 1995) with twice the sample size (Table 3.2).

The test statistic for the 2×2 allele table, is given by:

Table 3.2: Allele distribution

	Marker allele	Other	Total
Case	$r_1 + 2r_2$	$r_1 + 2r_0$	$2R$
Control	$s_1 + 2s_2$	$s_1 + 2s_0$	$2S$
Total	$n_1 + 2n_2$	$n_1 + 2n_0$	$2N$

$$X_A^2 = \frac{2N(2N(r_1 + 2r_2) - 2R(n_1 + 2n_2))^2}{2R(2N - 2R)[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2]}.$$

3.1.3 Serological distribution under Complete Dominance

We built the serological disrtibution table by ignoing the difference between homozygous and heterozygous genotypes. Such a tabulation was common when the disease type was done by serology, so that it was not possible to distinguish between someone who was homozygous for the allele of interest and someone who was heterozygous. Intuitively, this table will be appropriate whenever the allele of interest is dominant. Table 3.3 present the data in terms of the number of patients with and without the disease allele.

The Chi-square test statistics X_S^2 for 2×2 serological table obtained by pooling heterozygous and homozygous individuals, is equivalent to the trend test applied to the 2×3 genotypic table using $x_0 = 0$ and $x_1 = x_2$. The test is efficient when the gene is dominant. It is inefficient for testing whether a codominant gene is associated

Table 3.3: Serological distribution

	Patients with marker allele	Other	Total
Case	$r_1 + r_2$	r_0	R
Control	$s_1 + s_2$	s_0	S
Total	$n_1 + n_2$	n_0	N

with disease. The test statistic X_S^2 is:

$$X_S^2 = \frac{N[N(r_1 + r_2) - R(n_1 + n_2)]^2}{R(N - R)[N(n_1 + n_2) - (n_1 + n_2)^2]}.$$

3.2 Tests under Codominance

3.2.1 Genotype distribution under Codominance

Under the Codominance genotype, A represents the marker allele, B represents others. Table 3.4 represents the number of cases and controls for three types of genotype: Negative, Homozygous and Heterozygous. The phenotype of heterozygous will always have the disease. The test statistic will be same as:
$X_G^2 = \frac{N(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{R(N - R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$ in Chapter 2.4.

3.2.2 Allele distribution under Codominance

The table for Allele distribution under Codominance (Table 3.5) summarizes the number of marker allele (A) and the number of the other gene (B), where $x_1 = 1$ is dis-

Table 3.4: Genotype distribution

	Negative(BB)	Heterozygous(AB)	Homozygous(AA)	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

ease gene A, $x_0 = 0$ is the other gene B. Since human is diploid organism, the sample size will be doubled. The test statistic $X_A^2 = \frac{2N(2N(r_1+2r_2)-2R(n_1+2n_2))^2}{2R(2N-2R)[2N(n_1+2n_2)-(n_1+2n_2)^2]}$ is same as the test statistic in Chapter 2.5.

Table 3.5: Allele distribution

	Marker allele(A)	Other(B)	Total
Case	r_1+2r_2	r_1+2r_0	$2R$
Control	s_1+2s_2	s_1+2s_0	$2S$
Total	n_1+2n_2	n_1+2n_0	$2N$

3.2.3 Relation between Trend Test and Allele Test

Calculating the ratio of two test statistics to check the relation.

$$
\begin{aligned}
Ratio =& \frac{X_A^2}{X_G^2} \\
=& \frac{2N(2N(r_1+2r_2)-2R(n_1+2n_2))^2}{2R(2N-2R)[2N(n_1+2n_2)-(n_1+2n_2)^2]} \times \frac{R(N-R)[N(n_1+4n_2)-(n_1+2n_2)^2]}{N(N(r_1+2r_2)-R(n_1+2n_2))^2} \\
=& [\frac{2N(r_1+2r_2)-2R(n_1+2n_2)}{N(r_1+2r_2)-R(n_1+2n_2)}]^2 \times \frac{N(n_1+4n_2)-(n_1+2n_2)^2}{2[2N(n_1+2n_2)-(n_1+2n_2)^2]} \\
=& \frac{2N(n_1+4n_2)-2(n_1+2n_2)^2}{2N(n_1+2n_2)-(n_1+2n_2)^2} \\
=& \frac{2(n_0+n_1+n_2)(n_1+4n_2)-2(n_1+2n_2)^2}{2(n_0+n_1+n_2)(n_1+2n_2)-(n_1+2n_2)^2}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{2(n_1^2+5n_1n_2+4n_2^2+n_1n_0+4n_2n_0)-2(n_1^2+4n_2^2+4n_1n_2)}{2(n_1^2+3n_1n_2+2n_2^2+n_1n_0+2n_2n_0)-(n_1^2+4n_2^2+4n_1n_2)}\\
=&\frac{2n_1n_2+2n_0n_1+8n_0n_2}{n_1^2+2n_1n_2+2n_0n_1+4n_0n_2}\\
=&\frac{2n_1n_2+2n_0n_1+8n_0n_2}{(n_1+2n_0)(n_1+2n_2)}\\
=&\frac{2n_1n_2+2n_0n_1+8n_0n_2+n_1^2-n_1^2}{(n_1+2n_0)(n_1+2n_2)}\\
=&\frac{(n_1+2n_0)(n_1+2n_2)+4n_0n_2-n_1^2}{(n_1+2n_0)(n_1+2n_2)}\\
=&1+\frac{4n_0n_2-n_1^2}{(n_1+2n_0)(n_1+2n_2)}
\end{aligned}
$$

Let $B = \frac{4n_0n_2-n_1^2}{(n_1+2n_0)(n_1+2n_2)}$. The equation of B only has $n_i's$, which are the combined sample sizes with $n_i = r_i + s_i$. So B depends only on the counts in the combined sample. Then, the ratio of X_A^2 and X_G^2 is $1+\frac{4n_0n_2-n_1^2}{(n_1+2n_0)(n_1+2n_2)} = 1+B$, which depends only on the combined sample sizes.

If $B = 0$, the ratio is one, which means that $X_A^2 = X_G^2$.

$$B = 0 \implies 4n_0n_2 = n_1^2$$

The HWE holds in data if the genotype distribution and the allele distribution are same, which means the test statistics X_G^2 and X_A^2 are equal. Clearly, the two test statistics are equal only when HWE holds ($B = 0$) in the combined sample which is $4n_0n_2 = n_1^2$. Otherwise, the alleles statistic is larger than the valid trend test statistic if there is an excess of homozygotes and smaller when there is an excess of heterozygotes, meaning the test will be conservative for the excess of heterozygote.

The binomial distribution that the allele test is based on does not hold due to dependence between alleles within individuals when there is departure from HWE.

3.2.4 Summary

Association between disease and gene under three distributions (Genotype distribution, Allele distribution and Serological distribution) and the power of test statistics under different allelic effect are concluded in the following Proposition 3.1.

Proposition 3.1. Tests under different distributions:

1. Under the null hypothesis of no association between the disease and the gene:
   (a) both X_G^2 and X_S^2 are asymptotically distributed as chi-squared with one degree of freedom;
   (b) X_A^2 is also asymptotically chi-squared provided the population from which the cases and controls are sampled is in Hardy-Weinbery equilibrium;
   (c) X_A^2 will be anticonservative if there is an excess of homozygotes relative to the Hardy-Weinbery equilibrium.
2. Concerning power:
   (a) X_G^2 is locally most powerful if and only if the allele effect is exactly co-dominant(i.e., if the homozygous odds ratio is the square of the heterozygous one).

(b) X_S^2 is locally most powerful if and only if the allele effect is dominant.

(c) Provided the population is in Hardy-Weinberg equilibrium, X_A^2 is locally most powerful if and only if the allele effect is (exactly) codominant.

3.3 Odds Ratio

3.3.1 Odds Ratio for general cases

Definition 3.3.1. The formulas for odds ratio for general cases: heterozygous, homozygous, allele and serological are summrized in Table3.6.

Table 3.6: Odds Ratios

Distribution	Odds Ratio
Heterozygous	$\psi_{hetero} = \frac{r_1 s_0}{r_0 s_1}$
Homozygous	$\psi_{homo} = \frac{r_2 s_0}{r_0 s_2}$
Allele	$\psi_{allele} = \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)}$
Serological	$\psi_{sero} = \frac{(r_1 + r_2)s_0}{r_0(s_1 + s_2)}$

Theorem 3.1. Suppose that the heterozygous odds ratio is not equal to one, zero or infinity. Then any two of the following four conditions imply the other two.

1. The homozygous odds ratio is the square of the heterozygous one,

$$\psi_{homo} = \frac{r_2 s_0}{r_0 s_2} = (\frac{r_1 s_0}{r_0 s_1})^2 = \psi^2_{hetero}.$$

2. The allelic odds ratio is equal to the heterozygous one,

$$\psi_{allele} = \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} = \frac{r_1 s_0}{r_0 s_1} = \psi_{hetero}.$$

3. The Hardy-Weinberg equilibrium holds in the control population,

$$4s_0 s_2 = s_1^2.$$

4. The Hardy-Weinberg equilibrium holds in the case population,

$$4r_0 r_2 = r_1^2.$$

The stipulation that the heterozygous odds ratio is not equal to one is only required to derive (3) or (4) from (1) and (2).

Proof. Assume that the common odds ratio $\psi = \frac{r_1 s_0}{r_0 s_1} = \frac{r_2 s_1}{r_1 s_2}$.

- (1) and (2) $\implies$ (3)
  From (1) one have $r_2 = \frac{\psi r_1 s_2}{s_1}$, plug this equation into (2). We have :

$$(r_1 s_1 + 2\psi r_1 s_2)(2s_0 + s_1) = \psi s_1 (2r_0 + r_1)(s_1 + 2s_2)$$

$$\begin{aligned}
&\implies s_2(2\psi r_1(2s_0+s_1)-2\psi s_1(2r_0+r_1)) = \psi s_1^2(2r_0+r_1) - r_1 s_1(2s_0+s_1)\\
&\implies s_2 = \frac{2s_1(\psi r_0 s_1 - r_1 s_0) + r_1 s_1^2(\psi-1)}{4\psi(r_1 s_0 - r_0 s_1)}\\
&\implies s_2 = \frac{r_1 s_1^2(r_1 s_0 - r_0 s_1)}{4r_1 s_0(r_1 s_0 - r_0 s_1)}\\
&\implies s_2 = \frac{s_1^2}{4s_0}\\
&\implies 4s_0 s_2 = s_1^2,
\end{aligned}$$

which is (3).

- (1) and (2) $\implies$ (4)

  Similar to proof 1, plug $s_2 = \frac{s_1 r_2}{\psi r_1}$ into the equstion(2):

$$\begin{aligned}
&\implies r_2 = \frac{r_1^2}{4r_0}\\
&\implies 4r_0 r_2 = r_1^2.
\end{aligned}$$

- (1) and (3) $\implies$ (4)

  (1) implies that:

$$\begin{aligned}
\frac{r_2 s_0}{r_0 s_2} &= (\frac{r_1 s_0}{r_0 s_1})^2\\
\implies r_2 &= \frac{r_1^2 s_0 s1}{r_0 s_1^2}.
\end{aligned}$$

  Then, plug $s_1^2 = 4s_0 s_1$ into r_2,

$$\implies r_2 = \frac{r_1^2 s_0 s_2}{4r_0 s_0 s_2}$$

$$\implies r_2 = \frac{r_1^2}{4r_0}$$
$$\implies 4r_0r_2 = r_1^2,$$

which is (4).

- (2) and (4) $\implies$ (3):
  Substituting for $r_2 = \frac{r_1^2}{4r_0}$ from(4) into $\frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} = \frac{r_1 s_0}{r_0 s_1}$

$$\implies (2r_1^2 + 4r_0r_1)(2s_0 + s_1)r_0s_1 = 4r_0r_1s_0(2s_2 + s_1)(2r_0 + r_1)$$
$$\implies s_2 = \frac{2r_1s_1(r_1 + 2r_0)(s_1 + 2s_0) - 4r_1s_0s_1(2r_0 + r_1)}{8r_1s_0(2r_0 + r_1)}$$
$$\implies s_2 = \frac{s_2^3 + 2s_0s_1 - 2s_0s_1}{4s_0}$$
$$\implies s_2 = \frac{s_1^2}{4s_0}$$
$$\implies 4s_0s_2 = s_1^2,$$

which is (3).

- (3) and (4) $\implies$ (1):
  From (3) and (4), we have $s_1^2 = 4s_0s_2$ and $r_1^2 = 4r_0r_2$.
  The square of the odds ratio for heterozygous is $\psi_{hetero}^2 = (\frac{r_1s_0}{r_0s_1})^2$, using (3) and (4) we have:

$$\implies \psi_{hetero}^2 = (\frac{r_1s_0}{r_0s_1})^2 = \frac{r_1^2s_0^2}{r_0^2s_1^2}$$
$$\implies \psi_{hetero}^2 = \frac{4r_0r_2s_0^2}{4s_0s_2r_0^2} = \frac{r_2s_0}{r_0s_2} = \psi_{homo}.$$

- (3) and (4) $\implies$ (2):

  From (3) and (4), we have $s_2 = \dfrac{s_1^2}{4s_0}$ and $r_2 = \dfrac{r_1^2}{4r_0}$. Use (3) and (4) to write s_2 snd r_2.

  The odds ratio for allele is

$$\begin{aligned}
\psi_{allele} &= \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} = \frac{(2r_1^2/4r_0 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_1^2/4s_0 + s_1)} \\
\implies \psi_{allele} &= \frac{(r_1^2 s_0 + 2r_0 r_1 s_0)(2s_0 + s_1)}{(2r_0 + r_1)(2s_1^2 r_0 + 2r_0 s_0 s_1)} \\
\implies \psi_{allele} &= \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} = \frac{r_1 s_0}{r_0 s_1} = \psi_{hetero}.
\end{aligned}$$

□

Proposition 3.2. Odds ratios

1. Suppose that the Hardy-Weinberg equilibrium holds in the control population. Then the allelic odds ratio is greater (less) than the heterozygous odds ratio if and only if the homozygous odds ratio is greater (less) than the square of the heterzygous one.

   Proof. The proof in the left to right direction follows that used to prove (1) and (3) $\implies$ (4) and (3) and (4) $\implies$ (2) in Theorem 1. □

2. Suppose that only the homozygotes are not viable, i.e., there are no homozygotes in the population. Then the allelic odds ratio always lies between the heterozygous odds ratio and 1. (Suppose $r_2 = s_2 = 0$, ψ_{allele} always between ψ_{hetero} and 1.)

   Proof. Assume $\psi_{hetero} < 1$,

$$
\begin{aligned}
&\implies \frac{r_1 s_0}{r_0 s_1} < 1 \\
&\implies \frac{r_1 s_0}{r_0 s_1} \times \frac{r_1 s_1}{r_0 s_1} < \frac{r_1 s_1}{r_0 s_1} \\
&\implies \frac{2r_1 s_0}{r_0 s_1} + \frac{r_1 s_0}{r_0 s_1} \times \frac{r_1 s_1}{r_0 s_1} < \frac{2r_1 s_0}{r_0 s_1} + \frac{r_1 s_1}{r_0 s_1} \\
&\implies \frac{r_1 s_0}{r_0 s_1} \times (2 + \frac{r_1 s_1}{r_0 s_1}) < \frac{2r_1 s_0}{r_0 s_1} + \frac{r_1 s_1}{r_0 s_1} \\
&\implies \frac{r_1 s_0}{r_0 s_1} < \frac{\frac{2r_1 s_0}{r_0 s_1} + \frac{r_1 s_1}{r_0 s_1}}{2 + \frac{r_1 s_0}{r_0 s_1}} < 1 \\
&\implies \frac{r_1 s_0}{r_0 s_1} < \frac{2r_1 s_0 + r_1 s_1}{2r_0 s_1 + r_1 s_1} < 1 \\
&\implies \frac{r_1 s_0}{r_0 s_1} < \frac{r_1(2s_0 + s_1)}{s_1(2r_0 r_1)} < 1 \\
&\implies \psi_{hetero} < \psi_{allele} < 1.
\end{aligned}
$$

   Similarly for $\psi_{hetero} > 1$. □

3. Suppose only that the gene is dominant so that the homozygous odds ratio is equal to the heterzygous odds ratio. Then the allelic odds ratio lies between the heterozygous odds ratio and 1. ($\psi_{homo} = \psi_{hetero} \implies \psi_{allele}$ lies between ψ_{hetero} and 1.)

   Proof. Assume $\psi_{hetero} < 1$

From $\psi_{homo} = \psi_{hetero}$ one has $r_2 = \dfrac{r_1 s_1}{s_1}$.
Let $\psi_{hetero} = \dfrac{r_1 s_0}{r_0 s_1} = x$, and $\dfrac{r_1 s_1}{2 r_0 s_1} = y$. Then one has $x < 1$.

$$
\begin{aligned}
&\Longrightarrow xy < 1 \\
&\Longrightarrow x + xy < x + y \\
&\Longrightarrow x(1+y) < x + y \\
&\Longrightarrow x < \frac{x+y}{1+y} < 1, (x < 1) \\
&\Longrightarrow \frac{r_1 s_0}{r_0 s_1} < \frac{\frac{r_1 s_0}{r_0 s_1} + \frac{r_1 s_1}{2 r_0 s_1}}{1 + \frac{r_1 s_1}{2 r_0 s_1}} = \frac{r_1(2 s_0 + s_1)}{s_1(2 r_0 + r_1)}.
\end{aligned}
$$

Similarly for $\psi_{hetero} > 1$. □

4. Suppose only that the gene is recessive, so that the heterozygous odds ratio is equal to 1. Then the allelic odds ratio lies between homozygous odds ratio and 1. ($\psi_{hetero} = 1 \implies \psi_{allele}$ lies between ψ_{homo} and 1.)

Proof. Assume $\psi_{homo} < 1, \psi_{hetero} = \dfrac{r_1 s_0}{r_0 s_1} = 1 \implies s_1 = \dfrac{r_1 s_0}{r_0}$.
Let $x = \dfrac{r_2 s_0}{r_0 s_2}, y = \dfrac{r_1 s_0}{2 r_0 s_2}$

$$
\begin{aligned}
&\Longrightarrow x < 1 \\
&\Longrightarrow xy < y < 1 \\
&\Longrightarrow x + xy < x + y < 1 \\
&\Longrightarrow x < \frac{x+y}{1+y} < 1 \\
&\Longrightarrow \frac{r_2 s_0}{r_0 s_2} < \frac{\dfrac{r_2 s_0}{r_0 s_2} + \dfrac{r_1 s_0}{2 r_0 s_2}}{1 + \dfrac{r_1 s_0}{2 r_0 s_2}}
\end{aligned}
$$

$$\begin{aligned}
&\implies \frac{r_2 s_0}{r_0 s_2} < \frac{s_0(2r_2 + r_1)}{r_0(2s_2 + s_1)} = \frac{\frac{s_0}{r_0}(2r_2 + r_1)(2r_0 + r_1)}{(2s_2 + s_1)(2r_0 + r_1)} \\
&\implies \frac{r_2 s_0}{r_0 s_2} < \frac{(2r_2 + r_1)(\frac{r_1 s_0}{r_0} + 2s_0)}{(2r_0 + r_1)(2s_2 + s_1)} \\
&\implies \frac{r_2 s_0}{r_0 s_2} < \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} \\
&\implies \psi_{homo} = \frac{r_2 s_0}{r_0 s_2} < \frac{(2r_2 + r_1)(2s_0 + s_1)}{(2r_0 + r_1)(2s_2 + s_1)} = \psi_{allele} < 1 \\
&\implies \psi_{homo} < \psi_{allele} < 1.
\end{aligned}$$

Similarly for $\psi_{homo} > 1$. □

5. Provided r_0 and s_0 are both nonzero, the serological odds ratio is greater (less) than the heterozygous odds ratio if and only if the homozygous odds ratio is greater (less) than the heterozygous one. Note that in symbols,

$$\psi_{homo} < \psi_{hetero} \iff \psi_{sero} < \psi_{hetero}.$$

Conversely,

$$\psi_{homo} > \psi_{hetero} \iff \psi_{sero} > \psi_{hetero}.$$

Proof.

$$\begin{aligned}
&\psi_{homo} > \psi_{hetero} \iff \frac{r_2 s_0}{r_0 s_2} > \frac{r_1 s_0}{r_0 s_1} \\
&\iff \frac{r_2}{s_2} > \frac{r_1}{s_1} \\
&\iff r_2 s_1 > r_1 s_2
\end{aligned}$$

$$\begin{aligned}
&\iff r_2 s_1 + r_1 s_1 > r_1 s_2 + r_1 s_1 \\
&\iff s_1(r_2 + r_1) > r_1(s_2 + s_1) \\
&\iff \frac{r_2 + r_1}{s_2 + s_1} > \frac{r_1}{s_1} \\
&\iff \frac{s_0}{r_0}\left(\frac{r_2 + r_1}{s_2 + s_1}\right) > \frac{r_1 s_0}{s_1 r_0} \\
&\iff \psi_{sero} > \psi_{hetero}.
\end{aligned}$$

Similarly for $\psi_{sero} < \psi_{hetero} \iff \psi_{homo} < \psi_{hetero}$. □

6. The serological odds ratio always lies between the heterozygous and homozygous odds ratios. (ψ_{sero} always lies between ψ_{hetero} and ψ_{homo}.)

Proof. Assume $\psi_{hetero} < \psi_{homo}$

$$\begin{aligned}
&\implies \frac{r_1 s_0}{r_0 s_1} < \frac{r_2 s_0}{r_0 s_2} \\
&\implies \frac{r_1}{s_1} < \frac{r_2}{s_2} \\
&\implies \frac{r_1 s_2}{r_2 s_1 < 1} \\
&\implies r_1 s_2 < r_2 s_1.
\end{aligned}$$

Part 1: multiplying $r_1 s_2 < r_2 s_1$ on both sides by s_2:

$$\begin{aligned}
&\implies r_1 s_2 s_2 < r_2 s_1 s_2 \\
&\implies r_1 s_1 s_2 + r_1 s_2 s_2 < r_1 s_1 s_2 + r_2 s_1 s_2 \\
&\implies r_1 s_2 (s_1 + s_2) < s_1 s_2 (r_1 + r_2)
\end{aligned}$$

$$\implies \frac{1}{s_1 s_2(s_1+s_2)} r_1 s_2(s_1+s_2) < \frac{1}{s_1 s_2(s_1+s_2)} s_1 s_2(r_1+r_2)$$
$$\implies \frac{r_1 s_2(s_1+s_2)}{s_1 s_2(s_1+s_2)} < \frac{s_1 s_2(r_1+r_2)}{s_1 s_2(s_1+s_2)}.$$

Part 2: multiply by s_1 for both sides:

$$\implies r_1 s_1 s_2 < r_2 s_1 s_1$$

$$\implies r_1 s_1 s_2 + r_1 s_1 s_2 < r_1 s_1 s_2 + r_2 s_1 s_1$$
$$\implies \frac{1}{s_1 s_2(s_1+s_2)} s_1 s_2(r_1+r_2) < \frac{1}{s_1 s_2(s_1+s_2)} r_2 s_1(s_1+s_2)$$
$$\implies \frac{s_1 s_2(r_1+r_2)}{s_1 s_2(s_1+s_2)} < \frac{r_2 s_1(s_1+s_2)}{s_1 s_2(s_1+s_2)}.$$

Combining Part 1 and Part 2 together:

$$\implies \frac{r_1 s_2(s_1+s_2)}{s_1 s_2(s_1+s_2)} < \frac{s_1 s_2(r_1+r_2)}{s_1 s_2(s_1+s_2)} < \frac{r_2 s_1(s_1+s_2)}{s_1 s_2(s_1+s_2)}$$
$$\implies \frac{r_1}{s_1} < \frac{r_1+r_2}{s_1+s_2} < \frac{r_2}{s_2}$$
$$\implies \frac{s_0}{r_0}\frac{r_1}{s_1} < \frac{s_0}{r_0}\frac{r_1+r_2}{s_1+s_2} < \frac{s_0}{r_0}\frac{r_2}{s_2}$$
$$\implies \frac{r_1 s_0}{r_0 s_1} < \frac{s_0(r_1+r_2)}{r_0(s_1+s_2)} < \frac{r_2 s_0}{r_0 s_2}$$

$$\implies \psi_{hetero} < \psi_{sero} < \psi_{homo}.$$

Similarly for $\psi_{hetero} > \psi_{homo}$. □

Chapter 4

Extension from Biallelic to Multiallelic

The extension from biallelic to multiallelic markers is not as straightforward as the contingency-table-based test. But multiallelic trend test has the same power advantage of the allele test over the genotype test due to fewer degrees of freedom and remains valid when Hardy-Weinberg Disequilibrium (HWD) exists in the sample. First, we rewrite the biallelic test statistics into matrix forms. Then we extend x from a 3×1 matrix into X, a $m(m+1)/2 \times m$ matrix for the multiallelic trend test. In this chapter we also consider the biallelic test and the multiallelic test in proportion form.

4.1 Biallelic Trend Test

4.1.1 Biallelic Trend Test in matrix term

To extend Sasieni's findings to multiallelic markers, we start by looking at different representations of the biallelic trend test statistic. Slager and Schaid (2001) present the trend test statistic in the form $\frac{u^2}{Var(u)}$ where $u = x'[(1-\phi)r - \phi s]$, with $\phi = \frac{R}{N}$ and

$$x = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \quad r = \begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} \quad s = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \end{pmatrix} \quad n = \begin{pmatrix} n_0 \\ n_1 \\ n_2 \end{pmatrix}.$$

4.1.2 Algebraic Simplification

A small amount of algebra shows the equality of the trend test statistic in this form to the one given in Chapter 2.4.1,
$X_G^2 = \frac{N(N(r_1+2r_2) - R(n_1+2n_2))^2}{R(N-R)[N(n_1+4n_2) - (n_1+2n_2)^2]}$.

1. Part 1, the u term:

$$\begin{aligned} u &= x'[(1-\phi)r - \phi s] \\ &= (0\ 1\ 2)\left[(1-\frac{R}{N})\begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} - \frac{R}{N}\begin{pmatrix} s_0 \\ s_1 \\ s_2 \end{pmatrix}\right] \end{aligned}$$

$$
\begin{aligned}
=&(0\ 1\ 2)\left[\frac{N-R}{N}\begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} - \frac{R}{N}\begin{pmatrix} s_0 \\ s_1 \\ s_2 \end{pmatrix}\right] \\
=&(0\ 1\ 2)\begin{pmatrix} \frac{N-R}{N}r_0 - \frac{R}{N}s_0 \\ \frac{N-R}{N}r_1 - \frac{R}{N}s_1 \\ \frac{N-R}{N}r_2 - \frac{R}{N}s_2 \end{pmatrix} \\
=&0 \times \left(\frac{N-R}{N}r_0 - \frac{R}{N}s_0\right) + 1 \times \left(\frac{N-R}{N}r_1 - \frac{R}{N}s_1\right) + 2 \times \left(\frac{N-R}{N}r_2 - \frac{R}{N}s_2\right) \\
=&\frac{1}{N}[(N-R)(r_1+2r_2) - R(s_1+2s_2)].
\end{aligned}
$$

2. Part 2, the $Var(u)$ term:

$$
\begin{aligned}
Var(u) =&Var(x'[(1-\phi)r - \phi s]) \\
=&Var\left[(1-\phi)x'\sum_{i=0}^{2} r_i - \phi x'\sum_{j=0}^{2} s_j\right] \\
=&Var\left[(1-\phi)x'\sum_{i=0}^{2} r_i\right] + Var\left[\phi x'\sum_{j=0}^{2} s_j\right] - 2Cov\left[(1-\phi)x'\sum_{i=0}^{2} r_i, \phi x'\sum_{j=0}^{2} s_j\right] \\
=&(1-\phi)^2 x' Var\left(\sum_{i=0}^{2} r_i\right)x + \phi^2 x' Var\left(\sum_{j=0}^{2} s_j\right)x \\
&- 2\phi(1-\phi)x' Cov\left(\sum_{i=0}^{2} r_i, \sum_{j=0}^{2} s_j\right)x \\
=&(1-\phi)^2 x'\left[Var(\sum_{i=0}^{2} r_i) + 2\sum\sum_{i\neq j} Cov(r_i, r_j)\right]x
\end{aligned}
$$

$$+ \phi^2 x' \left[Var(\sum_{j=0}^{2} s_j) + 2 \sum \sum_{i \neq j} Cov(s_i, s_j) \right] x$$
$$- 2\phi(1-\phi) x' \left[\sum_i \sum_j Cov(r_i, s_j) \right] x.$$

3. Variance and Covariance (Given disease status, the distribution of genotypes is multinomial with parameter vector $p = (p_0, p_1, p_2)$ for cases and $q = (q_0, q_1, q_2)$ for controls):

$$Var(r_i) = \frac{R(N-R)n_i(N-n_i)}{N^2(N-1)} = Rp_i(1-p_i);$$
$$Var(s_i) = \frac{S(N-S)n_i(N-n_i)}{N^2(N-1)} = Sq_i(1-q_i);$$
$$Cov(r_i, r_j) = -\frac{R(N-R)n_i n_j}{N^2(N-1)} = -Rp_i p_j;$$
$$Cov(s_i, s_j) = -\frac{S(N-S)n_i n_j}{N^2(N-1)} = -Sq_i q_j;$$
$$Cov(r_i, s_j) = \frac{RSn_i n_j}{N^2(N-1)}.$$

Note that the three covariance terms in the last step account for the biological relationships among the subjects: $Cov(r_i, r_j)$, for correlation between the ith and jth cases; $Cov(s_i, s_j)$, for correlation between the ith and jth controls; and $Cov(r_i, s_j)$, for correlation between the ith case and jth control.

4. Adding equations together, the trend test statistic is:

$$\begin{aligned}\frac{u^2}{Var(u)} =& \frac{1}{N^2}[(N-R)(r_1+2r_2)-R(s_1+2s_2)]^2 \times \frac{1}{Var(u)} \\ =& \frac{N(N(r_1+2r_2)-R(n_1+2n_2))^2}{R(N-R)[N(n_1+4n_2)-(n_1+2n_2)^2]} \\ =& X_G^2.\end{aligned}$$

4.1.3 Biallelic Trend Test in proportion form

If the proportion of marker alleles for case and the proportion of marker alleles for control are equal, there is no marker-disease association. The null hypothesis is $\hat{p_R} - \hat{p_S} = 0$, and the alternative hypothesis is $\hat{p_R} - \hat{p_S} \neq 0$, where $\hat{p_R}$ is the proportion of marker alleles that appears in cases and $\hat{p_S}$ is the proportion of marker alleles that appears in controls. Under H_0, $\hat{p_R} = \hat{p_S} = \hat{\bar{p}}$.

$$\begin{aligned}&\hat{p_R} = \frac{2r_2 + r_1 + 0r_0}{2R}; \\ &\hat{p_S} = \frac{2s_2 + s_1 + 0s_0}{2S}; \\ &\hat{\bar{p}} = \frac{2n_2 + n_1 + 0n_0}{2N}; \\ &Var(\hat{p_R} - \hat{p_S}) = Var(\hat{p_R}) + Var(\hat{p_S}) \\ &= \frac{\hat{p_R}(1-\hat{p_R})}{NR} + \frac{\hat{p_S}(1-\hat{p_S})}{NS} = \frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{2R} + \frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{2S}.\end{aligned}$$

In matrix form,

$$x = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \quad r = \begin{pmatrix} r_0 \\ r_1 \\ r_2 \end{pmatrix} \quad s = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \end{pmatrix} \quad n = \begin{pmatrix} n_0 \\ n_1 \\ n_2 \end{pmatrix}.$$

$$\hat{p_{R_i}} = \frac{1}{2R} X'r, \ \hat{p_{S_i}} = \frac{1}{2S} X's$$

$$Var(\hat{p_R} - \hat{p_S})$$

$$= \begin{bmatrix} Var(p_{R_1} - p_{S_1}) & Cov(p_{R_1} - p_{S_1}, p_{R_2} - p_{S_2}) & Cov(p_{R_1} - p_{S_1}, p_{R_3} - p_{S_3}) \\ Cov(p_{R_2} - p_{S_2}, p_{R_1} - p_{S_1}) & Var(p_{R_2} - p_{S_2}) & Cov(p_{R_2} - p_{S_2}, p_{R_3} - p_{S_3}) \\ Cov(p_{R_3} - p_{S_3}, p_{R_1} - p_{S_1}) & Cov(p_{R_3} - p_{S_3}, p_{R_2} - p_{S_2}) & Var(p_{R_3} - p_{S_3}) \end{bmatrix}$$

$$= Var(\hat{p_R}) + Var(\hat{p_S})$$

$$= \frac{\hat{p_R}(1 - \hat{p_R})}{NR} + \frac{\hat{p_S}(1 - \hat{p_S})}{NS} = \frac{\hat{\bar{p}}(1 - \hat{\bar{p}})}{2R} + \frac{\hat{\bar{p}}(1 - \hat{\bar{p}})}{2S}.$$

Since in biallelic case, we only have one value of p_R and one value of p_S.

The Biallelic Trend Test statistic in proportion form:

$$\frac{u^2}{Var(u)} = (\hat{p_R} - \hat{p_S})^2 \times \frac{1}{Var(u)}.$$

4.2 Multiallelic Trend Test

4.2.1 Variables used in Multiallelic Trend Test

To construct such a statistic for m alleles.

1. We create a matrix X such that:

   (a) The jth column in X corresponds to the jth allele, $(j = 1, 2, ..., m)$.

   (b) An element in the jth column is the number of type j alleles.

   (c) The matrix X has $\frac{m(m+1)}{2}$ rows, which is the total number of possible genotypes.

   (d) X is a $\frac{m(m+1)}{2} \times m$ matrix.

2. r's, s's and n's:

$$r = \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ r_{m(m+1)/2} \end{pmatrix} \quad s = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ \vdots \\ s_{m(m+1)/2} \end{pmatrix} \quad n = \begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ \vdots \\ n_{m(m+1)/2} \end{pmatrix}.$$

3. The vector of trend test statistic U is:

$$U = X'[(1 - \phi)r - \phi s]$$

where the vector r and s are defined before but with length equal to $\frac{m(m+1)}{2}$, the number of distinct genotypes.

4. The $Var(U)$ is easily derived by replacing the vector x by the matrix X in the previous variance function for u.

$$\begin{aligned} Var(U) =&(1-\phi)^2 X' \left[Var(\sum_{i=0}^{2} r_i) + 2 \sum \sum_{i \neq j} Cov(r_i, r_j) \right] X \\ &+ \phi^2 X' \left[Var(\sum_{j=0}^{2} s_j) + 2 \sum \sum_{i \neq j} Cov(s_i, s_j) \right] X \\ &- 2\phi(1-\phi) X' \left[\sum_i \sum_j Cov(r_i, s_j) \right] X \end{aligned}$$

4.2.2 Multiallelic Trend Test in proportion form

The allele test statistic X_A^2 is not conducive to a multiallelic extension. We need a method of calculating the allelic test statistic that can be translated from a scalar to matrix format as with the trend test statistic. Under the null hypothesis of no marker-disease association we assume that the alleles in cases and controls come from independent multinomial samples, each with a probability that can be estimated with $\bar{p}$ (the sample frequency of the corresponding allele).

In matrix notation the chi-square statistic can be expressed as:

$$(\hat{p}_R - \hat{p}_S)'[Var(\hat{p}_R - \hat{p}_S)]^{-1}(\hat{p}_R - \hat{p}_S)$$

with $U = \hat{p}_R - \hat{p}_S$, $\hat{p}_R = \frac{1}{2R}X'r$, $\hat{p}_S = \frac{1}{2S}X's = \frac{1}{2N-2R}X'(n-r)$ and $\hat{p} = \frac{1}{2N}X'n$.

$$
\begin{aligned}
&Var(\hat{p}_{R_i} - \hat{p}_{S_i}) = Var(\hat{p}_{R_i}) + Var(\hat{p}_{S_i}) \\
&= \frac{\hat{p}_{R_i}(1-\hat{p}_{R_i})}{NR} + \frac{\hat{p}_{S_i}(1-\hat{p}_{S_i})}{NS} = \frac{\hat{p}_i(1-\hat{p}_i)}{2R} + \frac{\hat{p}_i(1-\hat{p}_i)}{2S} \\
&= \left(\frac{1}{2R} + \frac{1}{2S}\right)(\hat{p}_i(1-\hat{p}_i));
\end{aligned}
$$

$$
Cov(\hat{p}_{R_i} - \hat{p}_{S_i}, \hat{p}_{R_j} - \hat{p}_{S_j}) = \left(\frac{1}{2R} + \frac{1}{2S}\right)(-\hat{p}_i\hat{p}_j);
$$

$$
\begin{aligned}
&Var(\hat{p}_R - \hat{p}_S) \\
&= \begin{bmatrix}
Var(p_{\hat{R}_1} - p_{\hat{S}_1}) & Cov(p_{\hat{R}_1} - p_{\hat{S}_1}, p_{\hat{R}_2} - p_{\hat{S}_2}) & \dots \\
Cov(p_{\hat{R}_2} - p_{\hat{S}_2}, p_{\hat{R}_1} - p_{\hat{S}_1}) & Var(p_{\hat{R}_2} - p_{\hat{S}_2}) & \dots \\
\vdots & \ddots & \vdots \\
Cov(p_{\hat{R}_m} - p_{\hat{S}_m}, p_{\hat{R}_1} - p_{\hat{S}_1}) & \dots & Var(p_{\hat{R}_m} - p_{\hat{S}_m})
\end{bmatrix} \\
&= \left(\frac{1}{2R} + \frac{1}{2S}\right)\begin{bmatrix}
\hat{p_1}(1-\hat{p_1}) & -\hat{p_1}\hat{p_2} & \dots \\
-\hat{p_1}\hat{p_2} & \hat{p_2}(1-\hat{p_2}) & \dots \\
\vdots & \ddots & \vdots \\
-\hat{p_1}\hat{p_m} & \dots & \hat{p_m}(1-\hat{p_m})
\end{bmatrix} \\
&= \left(\frac{1}{2R} + \frac{1}{2N-2R}\right) \cdot \left[\frac{1}{(2N)^2}\left(2N diag(X'n) - X'nn'X\right)\right] \\
&= \frac{2N}{2R(2N-2R)} \cdot \left[\frac{1}{(2N)^2}\left(2N diag(X'n) - X'nn'X\right)\right]
\end{aligned}
$$

$$= \frac{1}{2R(2N-2R)(2N)} \cdot (2N diag(X'n) - X'nn'X)$$

with the vector n of length $\frac{m(m+1)}{2}$ containing the overall sample counts for each marker genotype and p_i representing the sample frequency of the ith marker allele.

$$\begin{aligned}
\hat{p_R} - \hat{p_S} =& \frac{1}{2R} X'r - \frac{1}{2S} X's \\
=& \frac{1}{2R} X'r - \frac{1}{2N-2R} X'n + \frac{1}{2N-2R} X'r \\
=& \frac{2N-2R}{2R(2N-2R)} X'r + \frac{2R}{2R(2N-2R)} X'r - \frac{1}{2N-2R} X'n \\
=& \frac{N}{R} \frac{1}{2N-2R} X'r - \frac{1}{N-R} X'n \\
=& \frac{1}{2N-2R} (\phi^{-1} X'r - X'n) \\
=& \frac{1}{\phi(2N-2R)} X'(r - \phi n) \\
=& \frac{N}{2R(N-R)} X'(r - \phi n).
\end{aligned}$$

The test statistics:

$$\begin{aligned}
X^2 =& (\hat{p}_R - \hat{p}_S)' [Var(\hat{p}_R - \hat{p}_S)]^{-1} (\hat{p}_R - \hat{p}_S) \\
=& \frac{N}{2R(N-R)} \cdot (r - \phi n)' X \cdot [2R(2N)(2N-2R)] \\
& \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot \frac{N}{2R(N-R)} X'(r - \phi n) \\
=& \frac{N^2}{4R^2(N-R)^2} \cdot 8NR(N-R) \cdot (r - \phi n)' X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r - \phi n) \\
=& \frac{2N^3}{R(N-R)} \cdot (r - \phi n)' X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r - \phi n).
\end{aligned}$$

The allele test statistic for multiallelic markers is $X_A^2 = X^2$. The multiallelic trend test statistic X_G^2 can be put into a form similar to X_A^2, by replacing $2N$, $2R$ and $2S$ in X_A^2 into N, R and S.

$$
\begin{aligned}
X_A^2 &= \frac{2N^3}{R(N-R)} \cdot (r-\phi n)'X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r-\phi n) \\
&= \frac{\frac{1}{4}\cdot 4 \cdot 2N^3}{\frac{1}{4}\cdot 4R(N-R)} \cdot (r-\phi n)'X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r-\phi n) \\
&= \frac{\frac{1}{4}\cdot (2N)^3}{\frac{1}{4}\cdot 2R(2N-2R)} \cdot (r-\phi n)'X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r-\phi n) \\
&= \frac{(2N)^3}{\frac{1}{4}\cdot 2R(2N-2R)} \cdot (r-\phi n)'X \cdot [2N diag(X'n) - X'nn'X]^{-1} \cdot X'(r-\phi n) \\
X_G^2 &= \frac{N^3}{R(N-R)} \cdot (r-\phi n)'X \cdot [N(X'diag(n)X) - X'nn'X]^{-1} \cdot X'(r-\phi n).
\end{aligned}
$$

As in the biallelic case, without the assumption of HWE, the allele test statistic under the null hypothesis does not hold. The requirement for the equality of these two statistics is:

$$
\begin{aligned}
& [N(X'diag(n)X) - X'nn'X]^{-1} = 2\,[2N diag(X'n) - X'nn'X]^{-1} \\
\Longrightarrow\ & 2\,[N(X'diag(n)X) - X'nn'X] = 2N diag(X'n) - X'nn'X \\
\Longrightarrow\ & 2N(X'diag(n)X) - X'nn' = 2N diag(X'n).
\end{aligned}
$$

These two statistics are formed with the same vector U, but have different variances. The requirement for the equality can also be expressed in more familiar terms:

$$\begin{bmatrix} \hat{p_1}^2 & \hat{p_1}\hat{p_2} & \dots \\ \hat{p_1}\hat{p_2} & \hat{p_2}^2 & \dots \\ \vdots & \ddots & \vdots \\ \hat{p_1}\hat{p_m} & \dots & \hat{p_m}\hat{p_m} \end{bmatrix} = \begin{bmatrix} \hat{P_{11}} & \frac{1}{2}\hat{P_{12}} & \dots \\ \frac{1}{2}\hat{P_{21}} & \hat{P_{22}} & \dots \\ \vdots & \ddots & \vdots \\ \frac{1}{2}\hat{P}_{(m)1} & \dots & \hat{P}_{(m)(m)} \end{bmatrix}$$

where $\hat{p}_i$ represents the frequency of allele in the sample, $\hat{P}_{ij}$ represent the sample frequency of genotype with allele i and j, $\hat{P}_{ii} = \hat{p}_i\hat{p}_i$ and $\hat{P}_{ij} = \hat{p}_i\hat{p}_j$.

4.3 The Power of trend test and Sample sizes

We approximate the formula of the power of trend test and the formula of necessary sampe sizes for given power of test. First, we calculate power for the trend test in biallelic case using normal distribution without HWE. Then we use Chi-square distribution to calculate the power for the trend test for multiallelic statistic.

4.3.1 The Power of the test and sample sizes required by biallelic statistic

The candidate gene status of a particular allele in any individual subject can be defined by two classes: (A) for high risk and (a) for all other alleles. We took R

random cases and S controls, with $R+S=N$. Given disease status, the distribution of genotypes is multinomial with parameter vector $p=(p_0,p_1,p_2)$ for cases and $q=(q_0,q_1,q_2)$ for controls. The population prevalence of the disease K, γ_1 and γ_2, the relative risks of genotypes Aa and AA, respectively, to aa.

The Armitage trend test is written as $\frac{u^2}{Var(u)}$. Under null hypothesis $H_0: p_i = q_i$:

$$
\begin{aligned}
\mu_0 =& 0;\\
\sigma_0^2 =& Var(u) = Var(x'[(1-\phi)r - \phi s])\\
=& N\phi(1-\phi)^2\left[\sum_i x_i^2 p_i - (\sum_i x_i p_i)^2\right] + N\phi^2(1-\phi)\left[\sum_i x_i^2 q_i - (\sum_i x_i q_i)^2\right]\\
=& N\phi(1-\phi)\left[\sum_i x_i^2 p_i - (\sum_i x_i p_i)^2\right][(1-\phi)\phi]\\
=& N\phi(1-\phi)\left[\sum_i x_i^2 p_i - (\sum_i x_i p_i)^2\right]\\
=& N(\sigma_0^*)^2.
\end{aligned}
$$

Under alternative hypothesis $H_a: p_i \neq q_i$:

$$
\begin{aligned}
\mu_1 =& N\left[\phi(1-\phi)\sum_i x_i(p_i - q_i)\right] = N\mu_1^*;\\
\sigma_1^2 =& Var(u) = Var(x'[(1-\phi)r - \phi s])\\
=& N\phi(1-\phi)^2\left[\sum_i x_i^2 p_i - (\sum_i x_i p_i)^2\right] + N\phi^2(1-\phi)\left[\sum_i x_i^2 q_i - (\sum_i x_i q_i)^2\right]\\
=& N(\sigma_1^*)^2.
\end{aligned}
$$

The power for the test is:

$$\begin{aligned}(1-\beta) =& P\left(Z < \frac{-z_{1-\alpha/2}\sigma_0 - \mu_1}{\sigma_1}\right) + P\left(Z > \frac{z_{1-\alpha/2}\sigma_0 - \mu_1}{\sigma_1}\right) \\ \Longrightarrow N \geq& \frac{(z_{1-\alpha/2}\sigma_0^* - z_\beta \sigma_1^*)^2}{(\mu_1^*)^2}\end{aligned}$$

where $p_i = \dfrac{f_i g_i}{\sum_i f_i g_i}$, $q_i = \dfrac{(1-f_i)g_i}{\sum_i (1-f_i)g_i}$, K is the prevalence rate and p is the frequency of high-risk allele (A). g_0, g_1 and g_2 are the population genotypic probabilities, and f_0, f_1 and f_2 are the penetrances for the respective genotypes (aa), (Aa) and (AA). $g_0 = (1-p)^2, g_1 = 2(1-p)p$ and $g_2 = p^2$. $f_0 = \dfrac{K}{g_2\gamma_2 + g_1\gamma_1 + g_0}, f_1 = f_0\gamma_1$ and $f_2 = f_0\gamma_2$.

4.3.2 Linkage Disequilibrium coefficients

We assume a biallelic disease locus with alleles A_1 and A_2, and the pupolation frequency for the high-risk allele A_2 is p. And the penetrance for $(A_r A_s)$ is f_{rs}. We consider markers with alleles M_i with population frequencies $q_i, i = 1, ..., m$, m is the number of alleles at marker locus. As in Chapman and Wijsman (1998), we assume equifrequent alleles in all markers ($q_i = 1/m$ for all i) to maximize heterozygosity. Thus, for any number of alleles at a marker, there are only two distinct linkage disequilibrium (LD) coefficients. The LD coefficients D_{ri} between A_r and M_i have

values:

$$D_{11} = -D_A,\ D_{21} = D_A$$

$$D_{1i} = -D_B,\ D_{2i} = D_B$$

where $D_A = \frac{(m-1)p}{m}$, $D_B = \frac{-p}{m}$. The usual bounds on linkage disequilibria imply that $p \leq \frac{1}{m}$ for this fomulation.

4.3.3 Genotype frequencies

HWD at the marker locus, is denoted d_{ij} for genotype M_iM_j; HWD at the disease locus, is denoted d_{rs} for genotype A_rA_s. The digenic gametic disequilibrium (LD) is defined above as D_{ri}, and the digenic nongametic disequilibrium, is denoted $D_{r/i}$ for allele A_r at the disease locus and M_i at the marker locus. And f_{rs} is the penetrance of (ArA_s).The marker genotype frequencies in affected individuals in case is:

$$\begin{aligned}
Pr(M_iM_i|Aff.) =& \frac{1}{K}\sum_{r,s} f_{rs}[(p_rq_i + D_{ri})(p_sq_i + D_{si}) + p_rp_sd_{ii} \\
&+q_i^2d_{rs} + d_{ii}d_{rs} + p_rq_iD_{s/i} + p_sq_iD_{r/i} + D_{r/i}D_{s/i}] \\
=& q_i^2 + \frac{1}{K}(2q_i\delta_i^c + \delta_{ii}^c) \\
=& Pr(M_iM_i) + \frac{1}{K}(2q_i\delta_i^c + \delta_{ii}^c); \\
Pr(M_iM_j|Aff.) =& \frac{1}{K}\sum_{r,s} f_{rs}[(p_rq_i + D_{ri})(p_sq_j + D_{sj}) + p_rp_sd_{ij} \\
&+q_iq_jd_{rs} + d_{ij}d_{rs} + p_rq_jD_{s/i} + p_sq_iD_{r/j} + D_{r/j}D_{s/i}
\end{aligned}$$

$$
\begin{aligned}
&+(p_r q_j + D_{rj})(p_s q_i + D_{si}) + p_r p_s d_{ij} \\
&+q_i q_j d_{rs} + d_{ij} d_{rs} + p_r q_i D_{s/j} + p_s q_j D_{r/i} + D_{r/i} D_{s/j}] \\
=&2q_i q_j + \frac{2}{K}(q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c) \\
=&Pr(M_i M_j) + \frac{2}{K}(q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c).
\end{aligned}
$$

Similarly, equations of genotype frequencies for the controls (unaffected individuals) are obtained by simply substituting $(1 - f_{rs})$ for f_{rs} and $(1 - K)$ for K:

$$
\begin{aligned}
Pr(M_i M_i|Unaff.) =&\frac{1}{1-K}\sum_{r,s}(1 - f_{rs})[(p_r q_i + D_{ri})(p_s q_i + D_{si}) + p_r p_s d_{ii} \\
&+q_i^2 d_{rs} + d_{ii} d_{rs} + p_r q_i D_{s/i} + p_s q_i D_{r/i} + D_{r/i} D_{s/i}] \\
=&q_i^2 + \frac{1}{1-K}(2q_i \delta_i^c + \delta_{ii}^c) \\
=&Pr(M_i M_i) + \frac{1}{1-K}(2q_i \delta_i^c + \delta_{ii}^c); \\
Pr(M_i M_j|Unaff.) =&\frac{1}{1-K}\sum_{r,s}(1 - f_{rs})[(p_r q_i + D_{ri})(p_s q_j + D_{sj}) + p_r p_s d_{ij} \\
&+q_i q_j d_{rs} + d_{ij} d_{rs} + p_r q_j D_{s/i} + p_s q_i D_{r/j} + D_{r/j} D_{s/i} \\
&+(p_r q_j + D_{rj})(p_s q_i + D_{si}) + p_r p_s d_{ij} \\
&+q_i q_j d_{rs} + d_{ij} d_{rs} + p_r q_i D_{s/j} + p_s q_j D_{r/i} + D_{r/i} D_{s/j}] \\
=&2q_i q_j + \frac{2}{1-K}(q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c) \\
=&Pr(M_i M_j) + \frac{2}{1-K}(q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c).
\end{aligned}
$$

We simplify these two equations for marker genotype frequencies in cases and controls by assuming the HWE holds in population. When HWE holds in the whole popula-

tion, we have $d_{ij} = d_{rs} = 0$, it is reasonable to assume that the digentic nongametic disequilibria are also zero. In that case:

$$\delta_i^c = D_{2i} f_{11}[-(1-p) + ((1-p) - p)\gamma_1 + p\gamma_2]$$
$$\delta_{ij}^c = D_{2i} D_{2j} f_{11}(1 - 2\gamma_1 + \gamma_2)$$

where f_{11} is calculated by $f_{11} = \dfrac{K}{g_2\gamma_2 + g_1\gamma_1 + g_0}$.

4.3.4 The Power of the test and sample sizes required by multiallelic statistic

We took R random cases and S controls, with $R + S = N$. We assume the sample sizes for case and control ar eequal, which means $R = S = \dfrac{1}{2}N$ and $\phi = \dfrac{R}{N} = 0.5$. Given disease status, the distribution of genotypes is multinomial with parameter P_R for cases and P_S for controls. K represents the population prevalence of the disease, γ_1 and γ_2, the relative risks of genotypes Aa and AA, respectively, to aa. We create the matrix X such that: The jth column in X corresponds to the jth allele, $(j = 1, 2, ..., m)$. An element in the jth column is the number of alleles of type j. The matrix X has $\dfrac{m(m+1)}{2}$ rows, which is the total number of possible genotypes.
We calculate the power for the trend test for the multiallelic statistic by using Chi-square distribution. Under the alternative hypothesis that genotypes in case and control are from two independent multinomial distributions with unequal probabilities $P_R = Pr(M_iM_j|Aff.)$ and $P_S = Pr(M_iM_j|Unaff.)$, repectively. The mean of U is $\mu_1 = N\phi(1-\phi)X'(P_R - P_S)$ and the variance of U is $\Sigma_1 = N\phi(1-\phi)X'[(1-\phi)(diag(P_R) - P_RP_R') + \phi(diag(P_S) - P_SP_S')]X$, which is the results in chapter 4.3.1.

By the definitions for a multinomial distribution, the noncentrality parameter $\lambda = \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1$. The power of the test is:

$$Power = 1 - X'^2_{df,\lambda}$$

where $X'^2_{df,\lambda}$ is the left-tail area of the noncentral Chi-square distribution with *df* degree of freedom and noncentrality parameter λ.

At any given power of test we also can compute the necessary sample sizes by combining the formulas of P_R, P_S, X, μ_1 and Σ_1:

$$\begin{aligned}
&\lambda = \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1 \\
\Longrightarrow\ &2\lambda = \mu_1'\Sigma_1^{-1}\mu_1 \\
\Longrightarrow\ &2\lambda = [\mu_1 = N\phi(1-\phi)X'(P_R - P_S)]' \\
&\quad \cdot [N\phi(1-\phi)X'[(1-\phi)(diag(P_R) - P_RP_R') + \phi(diag(P_S) - P_SP_S')]X]^{-1} \\
&\quad \cdot [\mu_1 = N\phi(1-\phi)X'(P_R - P_S)] \\
\Longrightarrow\ &2\lambda = N\phi(1-\phi)(P_R - P_S)'X \\
&\quad \cdot [X'[(1-\phi)(diag(P_R) - P_RP_R') + \phi(diag(P_S) - P_SP_S')]X]^{-1} \\
&\quad \cdot X'(P_R - P_S) \\
\Longrightarrow\ &N^{-1} = \frac{\phi(1-\phi)}{2\lambda}(P_R - P_S)'X \\
&\quad \cdot [X'[(1-\phi)(diag(P_R) - P_RP_R') + \phi(diag(P_S) - P_SP_S')]X]^{-1} \\
&\quad \cdot X'(P_R - P_S) \\
\Longrightarrow\ &N^{-1} = \frac{0.5}{2\lambda}(P_R - P_S)'X \\
&\quad \cdot [X'[(diag(P_R) - P_RP_R') + (diag(P_S) - P_SP_S')]X]^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\cdot X'(P_R - P_S) \\
\Longrightarrow N = &\frac{2\lambda}{0.5}[(P_R - P_S)'X \\
&\cdot [X'[(diag(P_R) - P_R P_R') + (diag(P_S) - P_S P_S')]X]^{-1} \\
&\cdot X'(P_R - P_S)]^{-1}
\end{aligned}
$$

where the value of $\lambda = X^2_{(m-1),\ \beta}$ depends on the power of test and the degrees of freedom $(m-1)$, X depends on the number of alleles at the marker locus, $P_R = Pr(M_iM_j|Aff.)$ and $P_S = Pr(M_iM_j|Unaff.)$.

Chapter 5

Calculations

5.1 From genotypes to genes

5.1.1 Test statistics

The data concerns HLA-DQ and HLA-DR typing and cervical intraepithelial neoplasia (CIN). The table of the data presents the number of cases (women with CIN 3) and controls with 0 (negative), 1 (heterozygous), and 2 (homozygous) copies of the allele DQ3 (DQ3 as co-domminants also called DQB1*03) at the HLA locus DQ. The odds ratio and the Chi-square test statistics for heterozygous and homozygous refer to the respective 2×2 subtables (Table 5.2 and Table 5.4) within Table 5.1. The expected values are concluded in another 2×2 table (Table 5.3 and Table 5.5). The odds ratios: $\psi_{hetero} = \frac{45 \times 273}{40 \times 100} = 3.07$, $\psi_{homo} = \frac{28 \times 273}{40 \times 43} = 4.44$ and the Chi-square test statistics: $X^2_{hetero} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 21.85$, $X^2_{homo} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 28.22$.

Since a heterozygous woman has one copy of DQ3 and each homozygous woman

has two copies, we can produce an allele table with twice the sample size (Table 5.6), where DQ3 represent the total number of the DQ3 for all observations. Similarly, one can build a table of expected values (Table 5.7).

Finally, the data in terms of the number of women with and without the allele, treat homozygous and heterozygous genotype in one group. Such a tabulation (Table 5.8) was common when HLA typing was done by serology, so that it was not possible to distinguish between someone who was homozygous or heterozygous for the allele of interest. Meanwhile, we can not test the association between disease and unknown allele. Intuitively, this table will be appropriate whenever the allele of interest is dominant.

Table 5.1: Genotype Distribution

	Negative	Heterozygous	Homozygous	Total
Case	40	45	28	113
Control	273	100	43	416
Total	313	145	71	529

Table 5.2: I: hetero (O_{ij})

	Negative	Heterozygous	Total
Case	40	45	85
Control	273	100	373
Total	313	145	458

From Table 5.1 to 5.9, one can calculate odds ratios and Chi-square test statis-

Table 5.3: I: hetero (E_{ij})

	Negative	Heterozygous	Total
Case	58	27	85
Control	255	118	43
Total	313	145	458

Table 5.4: I: homo (O_{ij})

	Negative	Homozygous	Total
Case	40	28	68
Control	273	43	316
Total	313	71	384

tics ($X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$). The row labeled "I: linear" is the maximum likelihood estimate from the model using a covariate that counts the number of DQ3 alleles that each woman has. The test statistic for "I: linear" is the sum of two independent Chi-square test statistics, concluded in Table 5.10.

5.1.2 Conclusion

From Table 5.10, we can see the allele distribution gives the largest chi-squared test statistic value. However, we can not conclude that the test for allele distribution is the most powerful test. The power of trend test for genotype distribution, the test for allele distribution and the test for serological distribution may be affected by the types of allele effect and satisfaction of HWE.

Table 5.5: I: homo (E_{ij})

	Negative	Homozygous	Total
Case	55	13	68
Control	258	58	316
Total	313	71	384

Table 5.6: II: allele (O_{ij})

	DQ3	Other	Total
Case	101	125	226
Control	186	646	832
Total	287	771	1058

Table 5.7: II: allele (E_{ij})

	DQ3	Other	Total
Case	61	165	226
Control	226	606	832
Total	287	771	1058

5.2 Extension from Biallelic to Multiallelic

5.2.1 The Power of the test and sample sizes required by biallelic statistic without HWE

Tables 5.11 to 5.14 conclude the necessary sample size N to achieve 80% power for Armitage's trend test under four different allelic effects (multiplicative, additive, dominant and recessive). Under a multiplicative model, $\gamma_1^2 = \gamma_2$, under an additive

Table 5.8: III: ser (O_{ij})

	DQ3	Other	Total
Case	73	40	113
Control	143	273	416
Total	216	313	529

Table 5.9: III: ser (E_{ij})

	DQ3	Other	Total
Case	46	67	113
Control	170	246	416
Total	216	313	529

Table 5.10: Odds ratios and Chi-square statistics

Table	Odds ratio	Chi-square
I: hetero	3.07	21.85
I: homo	4.44	28.22
I: linear	2.22	34.32
II: allele	2.81	44.85
III: serological	3.48	33.61

model, $\gamma_2 = 2\gamma_1 - 1$, under dominant model, $\gamma_1 = \gamma_2$, under recessive model, $\gamma_1 = 1$. For each genetic model, we assumed equal number of cases and controls, which means $\phi = 0.5$. We calculate the necessary sample sizes by $N = \frac{(z_{1-\alpha/2}\sigma_0^* - z_\beta \sigma_1^*)^2}{(\mu_1^*)^2}$.

5.2.2 Conclusion

In this case, we ignore the HWE and assume the sizes of case and the sizes of control are equal, which means $R = S = \frac{1}{2}N$. Table 5.11 − 5.14 conclude necessary sample size under different allelic effects (multiplicative, additive, dominant and recessive) with different values of population prevalence of the disease K, γ_1 and γ_2(the relative risks of genotypes Aa and AA, respectively, to aa) and the population frequencies of high-risk allele p.

First of all, the necessary sample sizes under the recessive model are much larger than others. It is easily to understand that, because the recessive disease genes has lower probability to affect people. Then, as the population prevalence of the disease K increases, the necessary sample sizes decreases. Similarly, the population frequencies of high-risk allele p increasing makes the necessary sample sizes decreasing.

Table 5.11: Sample sizes N for multiplicative model

γ_1	γ_2	K	p	$\alpha = 0.05$
2	4	0.01	0.01	599
			0.10	64
			0.50	19
		0.10	0.01	505
			0.10	54
			0.50	16
3	9	0.01	0.01	240
			0.10	26
			0.50	7
		0.10	0.01	204
			0.10	22
			0.50	6

Table 5.12: Sample sizes N for additive model

γ_1	γ_2	K	p	$\alpha = 0.05$
2	3	0.01	0.01	604
			0.10	70
			0.50	33
		0.10	0.01	510
			0.10	59
			0.50	27
3	5	0.01	0.01	243
			0.10	29
			0.50	16
		0.10	0.01	206
			0.10	24
			0.50	13

Table 5.13: Sample sizes N for dominant model

γ_1	γ_2	K	p	$\alpha = 0.05$
2	2	0.01	0.01	609
			0.10	77
			0.50	91
		0.10	0.01	514
			0.10	64
			0.50	74
3	3	0.01	0.01	244
			0.10	30
			0.50	39
		0.10	0.01	207
			0.10	26
			0.50	32

Table 5.14: Sample sizes N for recessive model

γ_1	γ_2	K	p	$\alpha = 0.05$
1	2	0.01	0.01	2,544,269
			0.10	3,424
			0.50	72
		0.10	0.01	2,104,248
			0.10	2,845
			0.50	60
1	3	0.01	0.01	653,082
			0.10	1,033
			0.50	26
		0.10	0.01	540,517
			0.10	862
			0.50	22

5.2.3 Sample sizes required for multiallelic distribution with HWE and complet LD

Tables $5.15 - 5.19$ show the sample sizes required for achieving 80% power using the multiallic trend test. Multiplicative, additive, dominant and recessive disease models were examined, defined by K, p, γ_1, γ_2 and f_{11}. As the usual bounds on linkage disequilibria imply that $p \leq \frac{1}{m}$ mentioned in section 4.3.2, we don't consider the situation of the number of alleles at the marker locus greater than 2.

Table 5.15: Necessary sample sizes for multiplicative models with HWE and complete LD

				Number of alleles at marker locus	
γ_1	γ_2	K	p	2	3
2	4	0.01	0.01	2666	4,795
			0.10	32	132
			0.50	3	
		0.10	0.01	23,306	8,125
			0.10	36	186
			0.50	3	

We can conclude that, the necessary sample sizes for recessive models is much larger than other models. The larger number of alleles at marker locus needs larger sample size to achieve 80% power of test. Then, as the population prevalence of the disease K is increasing, the necessary sample size is decreasing. Similarly, the population frequencies of high-risk allele p increasing makes the necessary sample

Table 5.16: Necessary sample sizes for additive models with HWE and complete LD

				Number of alleles at marker locus	
γ_1	γ_2	K	p	2	3
2	4	0.01	0.01	27	6,147
			0.10	35	160
			0.50	4	
		0.10	0.01	437	7,874
			0.10	24	65
			0.50	5	

sizes decreasing. We got similar results in the bialleic statistic in section 5.2.1.

Table 5.17: Necessary sample sizes for dominant models with HWE and complete LD

				Number of alleles at marker locus	
γ_1	γ_2	K	p	2	3
2	4	0.01	0.01	1,319	7,682
			0.10	45	152
			0.50	12	
		0.10	0.01	729	7,096
			0.10	31	126
			0.50	12	

Table 5.18: Necessary sample sizes for recessive models with HWE and complete LD

				Number of alleles at marker locus	
γ_1	γ_2	K	p	2	3
2	4	0.01	0.01	2,560	301,995
			0.10	2,624	1,918
			0.50	7	
		0.10	0.01	2561	301,995
			0.10	350	963
			0.50	11	

Chapter 6

Conclusion

Initially, we consider the extension from genotype to genes under complete dominance and codominance. Table $3.1 - 3.5$ shows that the extension under complete dominance and codominance are same.

We also compute the test statistics for genotype distribution and allele distribution. Besides this, we found out the relations between X_A^2 and X_G^2. X_A^2 is asymptotically chi-squared provided the population from which the cases and controls are sampled is in Hardy-Weinbery equilibrium. X_G^2 is locally most powerful if and only if the allele effect is exactly co-dominant (i.e., if the homozygous odds ratio is the square of the heterozygous one). Provided the population is in Hardy-Weinberg equilibrium, X_A^2 is locally most powerful if and only if the allele effect is (exactly) codominant.

After that, we extend the biallelic statistic to the multiallelic statistic, and derive the test statistics for trend tests in both normal form and proportion form. Furthermore, we approximate the formula for the power of trend test and the formula for necessary sample sizes at given power level for the trend test in the biallelic case using normal

distribution without HWE and Chi-square distribution to calculate the power of the trend test for the multiallelic statistic.

Finally, we calculate the test statistics for genotype distribution, allele distribution and serological distribution to compare the power of the tests. It is observed that the power of trend test for genotype distribution, the test for allele distribution and the test for serological distribution may be affected by the types of allele effect and satisfaction of HWE. We also calculate the necessary sample sizes for the biallelic statistic without HWE (Table $5.11 - 5.14$) and the necessary sample sizes for multiallelic distribution with HWE (Table $5.15 - 5.18$) under different values of K, p, γ_1, γ_2 and f_{11}. Thereby, the recessive model always needs a larger sample size to achieve 80% power compared to other models. If the population prevalence of the disease K increases, the necessary sample sizes will decrease. Likewise, increase in the population frequencies of high-risk allele p will decrease the necessary sample sizes.

Bibliography

[1] Sasieni, P. D.(1997). From genotypes to genes: Doubling the sample size. Biometrics 53, 1253-1261.

[2] Czika, W. and Weir, B. S. (2004). Properties of the multiallelic Trend Test. Biometrics, 60, 69-74.

[3] Apple, R. J., Erlich, H. A., Klitz, W., Becker, T. M. and Wheeler, C. M. (1994). HLA DR-DQ associations with cervical carcinoma show papillomavirus-type specificity. Nature Genetics 6, 157-162.

[4] Odunsi, K., Terry, G., Ho, L., Bell, J., Cuzick, J. and Ganesan, T. S. (1995). Association between HLA BQB1*03 and cervical intra-epithelial neoplasia. Molecular Medicine 1, 161-171.

[5] Wank, R. and Thomssen, C. (1991). High risk of squamous cell carcinoma of the cervix for women with HLA-DQw3. Nature 352, 723-725.

[6] Armitage, P. (1955). Test for linear trends in proportions and frequencies. Biometrics 11, 375-386.

[7] Chapman, N. H. and Wijsman, E. M. (1998). Genome screens using link-age disequilibrium tests: Optimal marker characeristics and feasibility. American Journal of Human Genetics 63, 1872-1885.

[8] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. Biometrics 55, 997-1004.

[9] Nielsen, D. M. and Weir, B. S. (1999). A classical setting for associations between markers and loci affecting quantitative traits. Genetical Research 74, 271-277.

[10] Nielsen, D. M., Ehm, M. G. and Weir, B. S. (1990). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. American Journal of Human Genetics 63, 1531-1540.

[11] Slager, S. L. and Schaid, D. J. (2001). Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. Human Heredity 52, 149-153.

[12] Slager, S. L. and Schaid, D. J. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. American Journal of Human Genetics 68, 1457-1462.

[13] Weir, B. S. (1979). Inferences about linkage disequilibrium. Biometrics 35, 235-254.

[14] Weir, B. S. (1996). Genetic Data Analysis *II*. Sunderland, MA: Sinauer Associates.

[15] Feder, J. N. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nature Publishing Group.

[16] Grover, V. K., Cole, David E. C. and Hamilton, David C. (2009). Attributing Hardy-Weinberg Disequilibrium to Population Stratification and genetic association in case-control studies. Annals of Human Genetics 74, 77-87.

Vita Auctoris

Jin Zhang was born in Hohhot, China in 1993. She graduated from University of Toronto in 2017 in Toronto, and obtained her Honour B.Sc. with specialist in Statistics and major in mathematics. She is currently a master candidate in the department of Mathematics and Statistics at the University of Windsor and hopes to graduate in May 2019.