

Chapter 2

Introduction to Genetic Epidemiology

Abstract Chapter 2 introduces a background to population genetics and genetic epidemiology. It starts with basic concepts of genetics and population genetics, including genes, alleles, genotypes, phenotypes, linkage disequilibrium, Hardy-Weinberg equilibrium, and population structure. Other terminology not covered in this chapter is discussed in later chapters. Designs of genetic association studies are then introduced, including population-based and family-based designs. Testing Hardy-Weinberg equilibrium proportions is covered. Goodness-of-fit, likelihood ratio and exact tests for deviation from Hardy-Weinberg equilibrium proportions are discussed. This chapter also discusses two types of risk measures: odds ratios and relative risks. Applying a logistic regression model for case-control data is presented.

This chapter introduces a background of population genetics and genetic epidemiology. It contains two parts. First, we start with basic concepts of population genetics, including alleles, genotypes, phenotypes, and linkage disequilibrium. Other terms that are not covered here will be discussed in later chapters. Designs of genetic association studies are then introduced, including case-control and family-based designs. We will focus here on case-control designs and family-based designs will be discussed in Chap. 13.

The Hardy-Weinberg law plays an important role in population genetics and the analysis of genetic data. Hardy-Weinberg equilibrium in a population is reviewed and the implications of departure from Hardy-Weinberg equilibrium are also demonstrated. Asymptotic and exact tests for Hardy-Weinberg proportions are given with examples. Calculation of the genotype frequencies in the population with or without Hardy-Weinberg proportions is given. The impact of departure from Hardy-Weinberg proportions is reviewed. It is well known that a case-control association study may be affected by hidden population substructure. Definitions of two common population substructures are given. Methods to correct for population substructure will be discussed in Chap. 9.

We discuss two measures of genotypic risks (odds ratio and relative risk) and their inference. The logistic regression models for case-control data are reviewed. Differences in the prospective and retrospective logistic regression models are briefly discussed. The conditional logistic regression model is often used for the

analysis of matched case-control data. A discussion of conditional logistic regression is given in Chap. 4.

2.1 Basic Genetic Terminology

With a few exceptions, human beings have in each cell nucleus 23 pairs of chromosomes, among which one pair comprises the sex chromosomes, also known as the X and Y chromosomes, and the other pairs are autosomal chromosomes. Within each chromosome is a molecule of DNA, which is made up of a long sequence of four different nucleotides labeled *A*, *T*, *C*, and *G*, with a structure that allows it to replicate itself. A gene is a series of DNA sequences that contain genetic information. For the purposes of this book we shall assume that along each chromosome pair hundreds or thousands of genes are arranged in a linear order (in the case of the sex chromosomes this occurs mostly along a single chromosome—from now on we shall restrict our discussion to autosomal chromosomes). This is perhaps not the case with the latest definition of a gene, but will suffice for our purposes. Similarly we shall define a locus to be the location of a gene or any DNA sequence on a chromosome pair. When the location of a DNA sequence on a chromosome pair is known, and that sequence varies in the population, it is also called a genetic marker. An allele is an alternative DNA sequence that can occur at a particular location on a single chromosome. Since chromosomes are present in pairs, at a given locus a person's gene or marker has two alleles, one on each chromosome. In the population, however, a gene or marker could have multiple (>2) alleles. We focus on diallelic markers, which have only two alleles in the population. A single nucleotide polymorphism (SNP) is a commonly used diallelic marker that varies in individuals owing to the difference of a single nucleotide (*A*, *T*, *C*, or *G*) in the DNA sequence. Although the location of a SNP is often referred to as a locus, it is more properly referred to as a site, a locus comprising more than one site.

We denote alleles by *A* and *B* for a single marker, where *A* is referred to as a wild type or a typical allele, and *B* is the complement of *A*, or the risk allele when a disease or trait is affected by this gene. For a multiallelic marker, it is possible that more than one allele may carry risks. (Other notation may also be used for alleles. In particular, when referring to two loci, we use a different notation: *A* and *a* for the alleles at one locus, and *B* and *b* for the alleles at the other locus.) Two alleles *A* and *B* at a locus form a genotype. There are four possible genotypes: *AA*, *AB*, *BA* and *BB*, but the two orders *AB* and *BA* are not distinguished. Hence only three genotypes are possible at a diallelic locus, denoted by *AA*, *AB* or *BB*. Genotypes *AA* and *BB* are said to be homozygous, and *AB* is heterozygous. For a multiallelic marker, more genotypes may be observed. For example, if a marker has three alleles, *A*, *B* and *C*, a total of six genotypes are possible: *AA*, *AB*, *AC*, *BB*, *BC*, and *CC*. Allele frequencies are usually the relative frequencies of the alleles in the population, denoted by $\Pr(B) = p$ and $\Pr(A) = 1 - p$. Minor allele frequency (MAF) refers to the frequency of the allele with frequency no more than 0.5 in a population. Denote the three genotypes by $G_0 = AA$, $G_1 = AB$ and $G_2 = BB$.

Table 2.1 Joint distribution of marker and a functional locus

	B	b	
A	$(1 - p)(1 - q) + D$	$(1 - p)q - D$	$1 - p$
a	$p(1 - q) - D$	$pq + D$	p
	$1 - q$	q	1

The genotype frequencies are then the frequencies of the three genotypes in the population, denoted by $g_i = \Pr(G_i)$ for $i = 0, 1$ and 2 , and $g_0 + g_1 + g_2 = 1$.

A phenotype is any observable characteristic or trait of an individual. It refers to a physical expression of genotypes at many loci and/or environmental factors. The trait can be continuous, e.g., blood pressure, weight, height etc., discrete, including binary such as diseased/case and normal/control, or ordinal categories related to different stages of a disease. A discrete trait can be defined based on a continuous trait. For example, cases and controls correspond to extremely high or low values of the trait, respectively. In some study designs, cases can be obtained from the extremely high values of the trait while controls are random samples from the population.

One of the goals of genetic association studies is to detect disease susceptibility genes (or functional loci). Suppose M_1 is a functional locus for a disease. The alleles at a functional locus are not observed. Suppose M_2 is an observed marker. Assume M_1 and M_2 have alleles A, a and B, b , respectively. Suppose the allele frequencies are given by $\Pr(a) = p$ and $\Pr(b) = q$. Thus, $\Pr(A) = 1 - p$ and $\Pr(B) = 1 - q$. The joint distribution of the alleles at the two loci (the marker and the functional locus) is given in Table 2.1, where

$$D = \Pr(AB) - \Pr(A)\Pr(B),$$

is the linkage disequilibrium coefficient. When $D = 0$, the two loci are independent, and we say the two loci are in gametic phase equilibrium. When $D \neq 0$, they are correlated and in gametic phase disequilibrium. When two loci are on the same chromosome pair and close enough to each other, their alleles are not transmitted independently to each offspring and the loci are said to be linked. Gametic phase disequilibrium between two linked loci is called linkage disequilibrium (LD). Most of the discussions in this book assume that a marker is also a functional locus and that they have the same allele frequencies. We refer to this model as a single-locus model, under which either $A = B$ and $a = b$ with $p = q$ or $A = b$ and $a = B$ with $p = 1 - q$. In the former case, $\Pr(AB) = \Pr(A) = 1 - p$ and $D = p(1 - p)$. In the latter case $\Pr(AB) = 0$ and $D = -p(1 - p)$. We also call the model with $D \neq 0$ a two-locus model. A common measure for LD is the standardized LD coefficient D' , defined by

$$D' = \begin{cases} D / \min\{(1 - q)p, (1 - p)q\} & \text{if } D > 0, \\ D / \min\{(1 - p)(1 - q), pq\} & \text{if } D < 0. \end{cases} \quad (2.1)$$

Using D' , complete LD refers to $|D'| = 1$; perfect LD refers to $|D'| = 1$ together with either $A = B$ and $a = b$ with $p = q$ ($D > 0$), or $A = b$ and $a = B$ with

$p = 1 - q$ ($D < 0$). Thus, the single-locus model refers to perfect LD, while the two-locus model refers to imperfect LD.

When $D \neq 0$, there are nine possible combinations of genotypes for the two loci:

$$\begin{aligned} (AA, BB), & \quad (AA, Bb), & \quad (AA, bb), \\ (Aa, BB), & \quad (Aa, Bb), & \quad (Aa, bb), \\ (aa, BB), & \quad (aa, Bb), & \quad (aa, bb). \end{aligned}$$

Given genotypes (AA, BB) , it is certain that alleles on both chromosomes are A and B . In this situation, phase is known. Given genotypes (Aa, Bb) , however, it is not certain which two alleles are on the same chromosome; they can be AB and ab or Ab and aB . In this situation, phase is unknown. A different two-locus model is used in gene-gene interactions (Chap. 8), where the two loci refer to two disease susceptibility genes. It can also be modified to a two-marker model, in which both M_1 and M_2 are markers and both are in LD with a functional locus. It can be further extended to multiple markers. A disease can be affected by more than two markers in the form of a haplotype, which will be introduced and discussed in Chap. 7. Discussion of $D \neq 0$, when one locus is a marker and the other is a functional locus, will also be given for some topics in Chap. 11. Penetrance is defined as the probability of having a disease given a specific genotype at the marker, denoted by $f_i = \Pr(\text{case} | G_i)$ for genotype G_i , $i = 0, 1, 2$. Here we assume perfect LD. When there is no association, $f_0 = f_1 = f_2 = \Pr(\text{case})$. We denote $\Pr(\text{case}) = k$ as the prevalence of the disease.

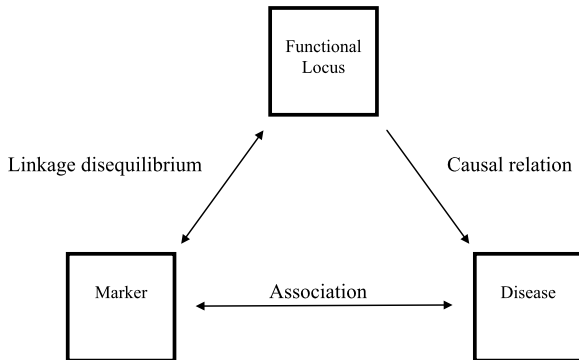
2.2 Genetic Association Studies

In this section, we first show the relationship between LD and association under the imperfect LD model by varying the parameter D defined in Table 2.1, in which one locus is a marker with alleles A/a and the other is a functional locus with alleles B/b . We will discuss two types of designs: population based and family based. In the population-based design, we focus on the retrospective case-control study. We discuss case-control designs and analyses from the epidemiological perspective. Other relevant designs are also mentioned.

2.2.1 Linkage Disequilibrium and Association Studies

Because the functional locus (or disease locus) that has a causal relationship with a disease is unknown, a marker is genotyped and tested for association with the disease. If the marker is in LD with the disease locus, an association between the marker and the disease can be identified through testing for association. Figure 2.1 shows a diagram of the association between the marker and the disease, the causal

Fig. 2.1 Diagram of LD, association and causal relationship among the marker, functional locus and disease



relationship between the disease locus and the disease, and the LD between the two loci.

To demonstrate the relationship between the LD and association, we use the notation defined in Table 2.1. In addition, we assume the penetrances at the disease locus are $f_i^* = \Pr(\text{case} | G_i^*)$, where $(G_0^*, G_1^*, G_2^*) = (BB, Bb, bb)$. The penetrances at the marker are still denoted by $f_i = \Pr(\text{case} | G_i)$, where $(G_0, G_1, G_2) = (AA, Aa, aa)$. Denote

$$F_1 = 1 - q + D/(1 - p), \quad F_2 = 1 - q - D/p,$$

$$F_3 = q - D/(1 - p), \quad F_4 = q + D/p,$$

where p , q and D are given in Table 2.1. Denote $\lambda_1^* = f_1^*/f_0^*$ and $\lambda_2^* = f_2^*/f_0^*$, which are referred to as genotype relative risks (GRRs). More discussion of GRRs will be provided in Chap. 3. Then (Problem 2.2),

$$f_0 = f_0^*(F_1^2 + 2F_1F_3\lambda_1^* + F_3^2\lambda_2^*), \quad (2.2)$$

$$f_1 = f_0^*(F_1F_2 + F_1F_4\lambda_1^* + F_2F_3\lambda_1^* + F_3F_4\lambda_2^*), \quad (2.3)$$

$$f_2 = f_0^*(F_2^2 + 2F_2F_4\lambda_1^* + F_4^2\lambda_2^*). \quad (2.4)$$

When $D = 0$, i.e., under linkage equilibrium, (2.2) to (2.4) reduce to

$$f_0 = f_1 = f_2 = f_0^*\{(1 - q)^2 + 2q(1 - q)\lambda_1^* + q^2\lambda_2^*\} = k,$$

regardless of values of λ_1^* and λ_2^* . Hence, there is no association between the marker and the disease under linkage equilibrium. From Problem 2.3, when $D \neq 0$, the penetrances (f_0, f_1, f_2) are not equal, which leads to unequal distributions for genotype counts in cases and controls. Therefore, a standard chi-squared test can be applied to detect association between the genotypes of the marker and the disease. Association, however, can arise from other factors, e.g., population substructure (see Sect. 2.4). Associations not due to LD, or more generally to gametic phase disequilibrium, are called spurious associations. Table 2.2 reports the GRRs at the marker $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_1$ given those at the disease locus $\lambda_1^* = f_1^*/f_0^*$, $\lambda_2^* = f_2^*/f_1^*$, and values of p , q , D' and disease prevalence k . Table 2.2 shows that the values of λ_2 are

Table 2.2 GRRs at the marker (λ_1, λ_2) given those at the disease locus $(\lambda_1^*, \lambda_2^*)$ with prevalence $k = 0.1$ and values of $p = \Pr(a)$ (marker allele frequency), $q = \Pr(b)$ (disease locus allele frequency) and LD parameter D'

λ_1^*	λ_2^*	p	q	D'	λ_1	λ_2
1.00	1.50	0.1	0.1	0.9	1.005	1.414
				0.8	1.008	1.336
			0.3	0.9	1.078	1.396
				0.8	1.072	1.332
	1.30	0.3	0.1	0.9	1.268	1.567
				0.8	1.237	1.474
			0.3	0.9	1.185	1.369
				0.8	1.164	1.327
1.50	1.50	0.3	0.1	0.9	1.441	1.481
				0.8	1.384	1.455
			0.3	0.9	1.224	1.244
				0.8	1.196	1.232

smaller than those of λ_2^* when $|D'| < 1$, and λ_2 decreases with $|D'|$. This indicates that association becomes weaker with a weaker LD. A similar phenomenon is observed for λ_1 except for $\lambda_1^* = 1$, which corresponds to a recessive disease (for the definitions of genetic models, see Chap. 3).

2.2.2 Population-Based Designs

A typical population-based design is the case-control study. In this design, individuals are genetically unrelated. A retrospective case-control design is cost-effective and commonly used in genetic studies. In this book, we focus on the retrospective case-control design, in which a random sample of cases (controls) is drawn from the case (control) population. The numbers of cases and controls in practice are determined in the design stage based on considerations of power, cost, and the disease prevalence. Given the total sample size, a design with equal numbers of cases and controls is more powerful than one with unequal numbers. In epidemiology, the retrospective case-control design is particularly useful to study a rare disease. For each individual, the genotype of the marker of interest is obtained. The goal of this design is to test whether or not the disease is associated with the marker. The retrospective case-control design is also used in large-scale association studies, in particular for genome-wide association studies (GWAS) using 500,000 to more than a million SNPs.

Another type of population-based design is a prospective case-control (cohort) study. In this design, individuals entering the study are drawn randomly from the study population without the disease. Following a period of time after, say, a treatment or an intervention, individuals who develop the disease are called cases and

those who do not are controls. This design, however, is not efficient for rare diseases. The outcome of this design is not restricted to a binary trait. It can be a quantitative trait, e.g., comparing the change of weight from baseline among three genotypes or between two alleles after a diet intervention.

In general, case-control designs are cost-effective for large-scale association studies. One potential concern of case-control studies is population substructure, which would lead to spurious association if not properly controlled.

2.2.3 Family-Based Designs

One simple family-based design for association studies is the case-parents trio design. In this design, an affected offspring is first ascertained and genotyped. The genotypes of the parents are also obtained. One approach to detect association in this design is based on a statistic comparing the number of marker alleles transmitted from parents to the offspring with the number not transmitted. In this case, only heterozygous parents are considered in the analysis. A typical method to analyze the trio data is called the transmission disequilibrium test (TDT). Because the untransmitted alleles can be regarded as controls for those transmitted, concern of population stratification, which can inflate the Type I error rate in population-based designs, is not relevant in the analysis of the trio design. This simple design has been extended to include multiple affected offspring (affected sibpairs), or disease-free siblings. For more details of this design and analysis, refer to Chap. 13.

More complicated family-based designs use data on large pedigrees with two or more generations. Some genotypes may not be available, especially for late onset diseases. Both binary traits and quantitative traits can be analyzed using family-based designs. Different kinds of family-based association tests can be used to analyze large family data in which correlations of traits among family members and their genetic relationships are incorporated into the analysis. A well-known example of this design is the Framingham Heart Study, which is a community-based family design. The original study began in 1948 and was designed to study cardiovascular disease and its risk factors in Framingham, Massachusetts. Now data from three generations have been obtained. Recent genetic studies, including linkage studies and GWAS, have been extensively reported.

Unlike population-based designs, family-based designs can eliminate the effects of population stratification using TDT statistics, without the need of the kinds of methods briefly described later. But these designs are typically not as efficient as population-based designs, especially for late onset diseases.

2.2.4 Other Designs

In addition to purely population-based and family-based designs, there are other designs for genetic studies. These designs use multi-stage samples. For example, in

stage 1, family data are obtained for linkage studies, from which candidate-genes are identified. In the second stage, association studies (either population-based or family-based) are conducted. In another population-based design, controls may be shared by association studies for different diseases. In the Wellcome Trust Case-Control Consortium (WTCCC) study, 3,000 controls drawn from the British population were shared among association studies of seven diseases. Data from population-based and family-based designs can also be combined to enhance the power to detect true associations. The community-based Framingham Heart Study has been used to supply controls for association studies of diabetes. These designs arise because a genetic study often uses multi-stage samples and thousands of individuals to detect small to moderate genetic effects.

In testing gene-environment interaction for a rare disease when a genetic susceptibility and an environmental factor are independent in the population, a case-only design can be employed because the odds ratio relating the gene and environment to a disease is approximately the odds ratio relating the environmental factor to the genetic factor among cases. The case-only design is often more powerful to detect gene-environment interaction than a case-control design using a logistic regression model.

Many hybrid designs have been proposed for cost-effectiveness, including combining case-control and family-based designs to test for genetic associations, and combining case-control and case-only designs to test for gene-environment interactions. Many genetic studies have been conducted and data from various study designs are available. Thus, hybrid designs based on data sharing are becoming more important and popular.

2.3 Hardy-Weinberg Principle

The Hardy-Weinberg principle, also known as the Hardy-Weinberg law or Hardy-Weinberg equilibrium (HWE), is a well-known model in population genetics. It states that under random mating both allele and genotype frequencies in a population remain constant or stable if no disturbing factors are introduced. We first introduce HWE followed by testing for departure from Hardy-Weinberg proportions (usually erroneously called “testing for HWE”). Both asymptotic chi-squared tests and an exact test will be discussed. The impact of departure from Hardy-Weinberg proportions will also be discussed.

2.3.1 *What Is Hardy-Weinberg Equilibrium?*

Autosomal Chromosomes

Consider a diallelic locus with alleles A and B with population frequencies q and p , respectively. Assume males and females have the same allele frequencies.

Table 2.3 Genotype frequencies under random mating given male and female gametes

			Male gametes	
			<i>A</i>	<i>B</i>
			<i>q</i>	<i>p</i>
Female gametes	<i>A</i>	<i>q</i>	g_0	$g_1/2$
	<i>B</i>	<i>p</i>	$g_1/2$	g_2

Table 2.4 Mating types (MTs) with frequencies and conditional probabilities of zygotes given mating types

MTs	Freq.	Freq. of zygotes		
		<i>AA</i>	<i>AB</i>	<i>BB</i>
MT ₁ : <i>AA</i> × <i>AA</i>	g_0^2	1	0	0
MT ₂ : <i>AA</i> × <i>AB</i>	$2g_0g_1$	1/2	1/2	0
MT ₃ : <i>AA</i> × <i>BB</i>	$2g_0g_2$	0	1	0
MT ₄ : <i>AB</i> × <i>AB</i>	g_1^2	1/4	1/2	1/4
MT ₅ : <i>AB</i> × <i>BB</i>	$2g_1g_2$	0	1/2	1/2
MT ₆ : <i>BB</i> × <i>BB</i>	g_2^2	0	0	1

Then, under random mating, Table 2.3 gives the genotype frequencies $g_0 = \Pr(AA)$, $g_1 = \Pr(AB)$ and $g_2 = \Pr(BB)$ together with the male and female gametes and their frequencies.

When HWE holds in the population,

$$g_0 = q^2, \qquad g_1 = 2pq, \qquad g_2 = p^2.$$

(2.5)

Equations (2.5), known as HWE proportions or simply Hardy-Weinberg proportions, can be obtained assuming random mating, under which alleles of male and female gametes are independent. More assumptions, however, are required for HWE. In addition to random mating, it also requires that the population size is infinite, males and females have identical allele frequencies, there is no effect of migration or mutation, and there is no natural selection.

Assume that HWE does not hold at the current generation in the population. Under random mating, we will show that for one locus the proportions (2.5) hold in the population after one generation. The genotype frequencies are denoted by g_0 , g_1 , and g_2 as before. Then the allele frequencies of *A* and *B* in the population given the genotype frequencies can be written as $q = g_0 + g_1/2$ and $p = g_2 + g_1/2$, respectively. Table 2.4 shows six mating types MT_{*j*} for $j = 1, \dots, 6$, their corresponding frequencies, and the conditional probabilities of their zygotes given the mating types.

In the next generation, the genotype frequencies can be obtained from

$$\Pr(G_i) = \sum_{j=1}^6 \Pr(G_i | MT_j) \Pr(MT_j), \quad (2.6)$$

where $(G_0, G_1, G_2) = (AA, AB, BB)$. It can be shown that (Problem 2.4), using (2.6) and Table 2.4,

$$\begin{aligned} \Pr(G_0) &= (g_0 + g_1/2)^2 = q^2, \\ \Pr(G_1) &= 2(g_0 + g_1/2)(g_2 + g_1/2) = 2pq, \\ \Pr(G_2) &= (g_2 + g_1/2)^2 = p^2, \end{aligned} \quad (2.7)$$

where p and q are the frequencies of alleles B and A , respectively. Hence, at the next generation, the Hardy-Weinberg proportions hold. Note carefully that we have shown that one round of random mating results in these proportions, *not* that these proportions imply equilibrium. It is possible for these proportions to hold at every generation and yet the allele frequencies change from generation to generation.

The above results can be extended to multiallelic loci. In general, assume a locus has m alleles, denoted by A_j , $j = 1, \dots, m$, with population frequencies $p_j = \Pr(A_j)$. Under HWE, the genotype frequencies are given by $\Pr(A_i A_j) = 2p_i p_j$ for $i \neq j$ and $\Pr(A_j A_j) = p_j^2$.

The X Chromosome

For sex-linked loci, females have two copies of the X chromosome, while males have one copy of the X chromosome and one copy of the Y chromosome. We focus on the X chromosomes. Then the Hardy-Weinberg proportions, as defined for autosomal chromosomes, can be applied to females. For males, we assume that the allele frequency is identical to that of females. Hence, with Hardy-Weinberg proportions at the X chromosome, we have

$$\Pr(B|\text{male}) = \Pr(B|\text{female}) = p, \quad \Pr(A|\text{male}) = \Pr(A|\text{female}) = q, \quad (2.8)$$

$$\Pr(G_0|\text{female}) = q^2, \quad \Pr(G_1|\text{female}) = 2pq, \quad \Pr(G_2|\text{female}) = p^2. \quad (2.9)$$

If (2.8) holds and (2.9) does not hold, it takes one generation to reach Hardy-Weinberg proportions. If (2.8) does not hold but (2.9) holds, it takes infinitely many generations to reach the proportions.

Table 2.5 Testing Hardy-Weinberg proportions: the observed genotype counts, the expected genotype counts under HWE, and estimates of the expected genotype counts

	AA	AB	BB
Observed	n_0	n_1	n_2
Expected	nq^2	$2npq$	np^2
Estimated	$n\hat{q}^2$	$2n\hat{p}\hat{q}$	$n\hat{p}^2$

2.3.2 Testing Hardy-Weinberg Equilibrium Proportions

We discuss asymptotic tests and an exact test for Hardy-Weinberg proportions in the population for autosomal chromosomes. Results for the X chromosome are given next.

A Simple Chi-Squared Test

Suppose a random sample of size n is drawn from the population and the genotype counts of the n individuals are obtained. Denote the genotype counts by (n_0, n_1, n_2) for $(G_0, G_1, G_2) = (AA, AB, BB)$, and $n_0 + n_1 + n_2 = n$. The allele counts are $2n_0 + n_1$ for A and $2n_2 + n_1$ for B among a total of $2n$ alleles. Let $p = \Pr(B)$. An estimate of p is given by $\hat{p} = (2n_2 + n_1)/(2n)$. Likewise, an estimate of $q = \Pr(A)$ is given by $\hat{q} = (2n_0 + n_1)/(2n)$. Table 2.5 shows the observed genotype counts, the expected genotype counts under Hardy-Weinberg proportions (the null hypothesis H_0), and estimates of the expected genotype counts under H_0 .

A typical chi-squared test has the form

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

Applying the above test to the data in Table 2.5 with the expected counts being replaced by the estimated ones, we have

$$\chi^2 = \frac{(n_0 - n\hat{q}^2)^2}{n\hat{q}^2} + \frac{(n_1 - 2n\hat{p}\hat{q})^2}{2n\hat{p}\hat{q}} + \frac{(n_2 - n\hat{p}^2)^2}{n\hat{p}^2}, \quad (2.10)$$

which has an asymptotic χ_1^2 distribution under H_0 . Using a chi-squared distribution for the statistic based on discrete genotype data, a bias correction of $1/2$ may be used

$$\chi^2 = \sum \frac{(|\text{observed} - \text{expected}| - 1/2)^2}{\text{expected}}.$$

To apply the chi-squared test in (2.10) requires a large n and the expected counts in each of the three cells not too small. This may not be true for alleles with small MAFs, or a small sample size.

Test Based on Hardy-Weinberg Disequilibrium

An alternative derivation of the above chi-squared test is based on the Hardy-Weinberg disequilibrium (HWD) coefficient, defined by

$$\Delta = \Pr(BB) - \{\Pr(B)\}^2 = \Pr(BB) - \{\Pr(BB) + \Pr(AB)/2\}^2.$$

Under H_0 , $\Delta = 0$. Hence, to test Hardy-Weinberg proportions, we can test $H_0 : \Delta = 0$. A test statistic can be constructed based on $\hat{\Delta}$, given by

$$\hat{\Delta} = \hat{\Pr}(BB) - \{\hat{\Pr}(BB) + \hat{\Pr}(AB)/2\}^2 = \frac{n_2}{n} - \left(\frac{2n_2 + n_1}{2n} \right)^2.$$

Denote the mean and variance of $\hat{\Delta}$ by $\mu = E(\hat{\Delta})$ and $\sigma^2 = \text{Var}(\hat{\Delta})$, where, ignoring terms with orders higher than $1/n$,

$$\begin{aligned} \mu &= \Delta - \{p - 2p^2 + \Pr(BB)\}/(2n), \\ \sigma^2 &= \{p^2(1-p)^2 + (1-2p)^2\Delta - \Delta^2\}/n. \end{aligned}$$

Under $H_0 : \Delta = 0$, after ignoring the terms with order $1/n$ in μ ,

$$\mu = 0 \quad \text{and} \quad \sigma^2 = \frac{1}{n}p^2(1-p)^2.$$

Hence, asymptotically,

$$\sqrt{n}\hat{\Delta} \sim N(0, p^2(1-p)^2) \quad \text{under } H_0,$$

which leads to an asymptotic chi-squared test

$$\chi^2 = \frac{n\hat{\Delta}^2}{\hat{p}^2(1-\hat{p})^2} \sim \chi_1^2, \quad (2.11)$$

where $\hat{p} = n_2/n + n_1/(2n)$ is same as in (2.10).

Using data on the *MN* blood groups in a British population, $n = 1000$ with $n_0 = 298$, $n_1 = 489$ and $n_2 = 213$, the estimate of p is $\hat{p} = (2 \times 213 + 489)/(2000) = 0.4575$, and the estimate of Δ is $\hat{\Delta} = 213/1000 - \hat{p}^2 = 0.003694$. From (2.11),

$$\chi^2 = \frac{1000 \times 0.003694^2}{0.4575^2(1-0.4575)^2} \approx 0.22152.$$

If we use (2.10), the estimates of the expected genotype counts corresponding to (n_0, n_1, n_2) are (294.306, 496.388, 209.306). Hence,

$$\chi^2 = \frac{(298 - 294.306)^2}{294.306} + \frac{(489 - 496.388)^2}{496.388} + \frac{(213 - 209.306)^2}{209.306} \approx 0.22152,$$

which is identical to the previous chi-squared statistic (Problem 2.5). The p-value for $\chi^2 = 0.222$ is 0.67, so there is no strong evidence to indicate deviation from Hardy-Weinberg proportions.

Likelihood Ratio Test

The LRT can be used to test Hardy-Weinberg proportions. The genotype counts (n_0, n_1, n_2) follow the multinomial distribution $Mul(n; p_0, p_1, p_2)$, where $p_i = \Pr(G_i)$ for $i = 0, 1, 2$. Then the likelihood function can be written as

$$L(p_0, p_1, p_2) = \frac{n!}{n_0!n_1!n_2!} p_0^{n_0} p_1^{n_1} p_2^{n_2}.$$

The MLE for p_i is $\hat{p}_i = n_i/n$. Thus, the maximum of the likelihood function is

$$L(\hat{p}_0, \hat{p}_1, \hat{p}_2) = \frac{n!}{n_0!n_1!n_2!} \frac{n_0^{n_0} n_1^{n_1} n_2^{n_2}}{n^n}.$$

Under the null hypothesis H_0 , the likelihood function is

$$L_0(p, q) = \frac{n!}{n_0!n_2!n_2!} 2^{n_1} p^{n_1+2n_2} q^{n_1+2n_0}.$$

The MLEs are $\hat{p} = (2n_2 + n_1)/(2n)$ and $\hat{q} = (2n_0 + n_1)/(2n)$. Thus,

$$L_0(\hat{p}, \hat{q}) = \frac{n!}{n_0!n_2!n_2!} \frac{2^{n_1} (n_1 + 2n_2)^{n_1+2n_2} (n_1 + 2n_0)^{n_1+2n_0}}{(2n)^{2n}}.$$

Hence, the LRT can be written as

$$\begin{aligned} \text{LRT} &= 2 \log \frac{L(\hat{p}_0, \hat{p}_1, \hat{p}_2)}{L_0(\hat{p}, \hat{q})} \\ &= 2 \log \frac{(2n)^{2n} n_0^{n_0} n_1^{n_1} n_2^{n_2}}{2^{n_1} n^n (n_1 + 2n_2)^{n_1+2n_2} (n_1 + 2n_0)^{n_1+2n_0}} \sim \chi_1^2 \quad \text{under } H_0. \end{aligned}$$

Applying the LRT to the above data, we obtain

$$\begin{aligned} \text{LRT} &= 2 \sum_{i=0}^2 n_i \log n_i + 4n \log(2n) - 2n \log n - 2n_1 \log 2 \\ &\quad - 2(n_1 + 2n_2) \log(n_1 + 2n_2) - 2(n_1 + 2n_0) \log(n_1 + 2n_2) \approx 0.22147, \end{aligned}$$

which is essentially the same p-value as we obtained before.

The asymptotic test given in (2.10) can be easily modified for testing Hardy-Weinberg proportions for a multiallelic locus. Suppose the following genotype counts are observed for $m(m-1)/2$ genotypes with m alleles:

$$\{n_{ij} : i, j = 1, \dots, m, i \leq j\}.$$

Let $n = \sum_{i \leq j} n_{ij}$ and $\hat{p}_j = (2n_{jj} + \sum_{i < j} n_{ij} + \sum_{k > j} n_{jk})/(2n)$. Then

$$\chi^2 = \sum_{j=1}^m \frac{(n_{jj} - n\hat{p}_j^2)^2}{n\hat{p}_j^2} + \sum_{j=1}^m \sum_{i < j} \frac{(n_{ij} - 2n\hat{p}_i\hat{p}_j)^2}{2n\hat{p}_i\hat{p}_j} + \sum_{j=1}^m \sum_{k > j} \frac{(n_{jk} - 2n\hat{p}_k\hat{p}_j)^2}{2n\hat{p}_k\hat{p}_j}.$$

Under H_0 , $\chi^2 \sim \chi_{m(m-1)/2}^2$. However, a more powerful test, based on χ_1^2 , is obtained by modeling the genotype frequencies as

$$\Pr(A_i A_j) = 2(1 - F)p_i p_j \quad \text{for } i \neq j,$$

$$\Pr(A_i A_i) = (1 - F)p_i^2 + Fp_i,$$

and testing the null hypothesis $H_0 : F = 0$, where F is Wright's inbreeding coefficient.

Exact Test

The performance of asymptotic chi-squared tests depends on approximations of the distributions of test statistics under H_0 . Because the genotype counts are discrete data, the approximations are not always accurate. In this case, an exact test may be preferred. Note that HWE is a model for calculating genotype frequencies using allele frequencies. Therefore, the exact test for Hardy-Weinberg proportions is based on the probability distribution of all possible genotype counts under HWE conditional on the observed allele counts.

Let (n_0, n_1, n_2) be the genotype counts for (AA, AB, BB) and (n_A, n_B) be the allele counts for (A, B) . Then $n_A = 2n_0 + n_1$, $n_B = 2n_2 + n_1$, and $n_A + n_B = 2n$. The genotype counts follow the multinomial distribution: $(n_0, n_1, n_2) \sim \text{Mul}(n; q^2, 2pq, p^2)$ under HWE, where $p = \Pr(B)$, i.e.,

$$\Pr(n_0, n_1, n_2) = \frac{n!}{n_0!n_1!n_2!}(q^2)^{n_0}(2pq)^{n_1}(p^2)^{n_2} = \frac{n!}{n_0!n_1!n_2!}2^{n_1}p^{n_B}q^{n_A}.$$

The allele counts (n_A, n_B) have the binomial distribution given by

$$\Pr(n_A, n_B) = \frac{(2n)!}{n_A!n_B!}p^{n_B}q^{n_A}.$$

Since the allele counts are determined by the genotype counts, we have

$$\Pr(n_0, n_1, n_2, n_A, n_B) = \Pr(n_0, n_1, n_2).$$

Hence,

$$\Pr(n_0, n_1, n_2 | n_A, n_B) = \frac{\Pr(n_0, n_1, n_2)}{\Pr(n_A, n_B)} = \frac{n!n_A!n_B!2^{n_1}}{n_0!n_1!n_2!(2n)!}. \quad (2.12)$$

Substituting $n_A = 2n - n_B$, $n_0 = (n_A - n_1)/2 = n - (n_B + n_1)/2$ and $n_2 = (n_B - n_1)/2$ into (2.12), we obtain

$$\Pr(n_1 | n_B) = \frac{n!(2n - n_B)!n_B!2^{n_1}}{[n - (n_B + n_1)/2]!n_1![(n_B - n_1)/2]!(2n)!}, \quad (2.13)$$

which only depends on n_1 given n_B and n (n_A is determined by n_B and n).

Table 2.6 Conditional probabilities of n_1 given $n_B = 8$ and $n = 30$

n_1	$n - \frac{n_B + n_1}{2}$	$\frac{n_B - n_1}{2}$	$\Pr(n_1 n_B)$
0	26	4	0.000011
2	25	3	0.0022
4	24	2	0.0557
6	23	1	0.3565
8	22	0	0.5856

Choose all valid values of n_1 given n_B and n , including the observed number with genotype AB . The valid value of n_1 has to be bounded by $0 \leq n_1 \leq \min(n_B, 2n - n_B)$, and $(n_B - n_1)/2$ is an integer. This implies that n_1 is even (odd) if n_B is. Calculate the probability in (2.13) for each valid n_1 . The exact p-value is the sum of probabilities with valid n_1 smaller than or equal to the observed number with genotype AB . In practice, to apply (2.13), the allele with smaller allele count is denoted by B and its corresponding allele count by n_1 , which will reduce the computation burden of (2.13).

Applying the exact test for Hardy-Weinberg proportions to the genotype counts $(n_0, n_1, n_2) = (24, 4, 2)$ with $n = 30$, the allele counts are $(n_A, n_B) = (52, 8)$. Note that $n_B < n_A$ and n_B is even. Hence, the only valid values for n_1 are 0, 2, 4, 6 and 8. The corresponding probabilities of (2.13) are reported in Table 2.6. Then the sum of probabilities that are smaller than or equal to 4 is the exact p-value. Using Table 2.6, we have $p = 0.000011 + 0.0022 + 0.0557 = 0.0579$, not significant at the 5% level.

Test Hardy-Weinberg Proportions for the X Chromosome

If the allele frequency in males is identical to that in females, Hardy-Weinberg proportions can be tested among females using χ^2 as given in (2.10) or (2.11) and the exact test. The male allele frequency may not be equal to that of females owing to many reasons, including genotyping errors. Therefore one may also test whether or not the male allele frequency is equal to that of females.

Let $p_m = \Pr(B|\text{male})$ and $p_f = \Pr(B|\text{female})$. Let (n_0^f, n_1^f, n_2^f) be the genotype counts in females and (n_A^m, n_B^m) be the allele counts in males. Let Δ_f be the HWD coefficient for females. The null hypothesis consists of $H_{0a} : p_M = p_F$ and $H_{0b} : \Delta_f = 0$, i.e., Hardy-Weinberg proportions hold in females. Let $n^m = n_A^m + n_B^m$ and $n^f = n_0^f + n_1^f + n_2^f$. Using the data, estimates of p_m and p_f are given by $\hat{p}_m = n_B^m/n^m$ and $\hat{p}_f = (2n_2^f + n_1^f)/(2n^f)$. Note that $\text{Var}(\hat{p}_m) = p_m(1 - p_m)/n^m$ and $\text{Var}(\hat{p}_f) = p_f(1 - p_f)/(2n^f)$, which can be estimated by $\widehat{\text{Var}}(\hat{p}_m) = \hat{p}_m(1 - \hat{p}_m)/n^m$ and $\widehat{\text{Var}}(\hat{p}_f) = \hat{p}_f(1 - \hat{p}_f)/(2n^f)$. A test statistic for H_{0a} can be written as

$$Z = \frac{\hat{p}_m - \hat{p}_f}{\sqrt{\widehat{\text{Var}}(\hat{p}_m) + \widehat{\text{Var}}(\hat{p}_f)}} \sim N(0, 1) \quad \text{under } H_{0a}.$$

In Problem 2.6, it is shown that Z and χ^2 are asymptotically uncorrelated under H_{0a} and H_{0b} . Thus, each test can be applied at the $\alpha/2$ level to control for multiple testing.

2.3.3 Impact of Hardy-Weinberg Equilibrium or Disequilibrium

In population genetics, HWE is used as a reference model to compare with other models. It is built on many assumptions which may not hold true. On the other hand, for genetic association studies, testing Hardy-Weinberg proportions has been used as a tool to detect genotyping errors. In research articles, p-values from testing Hardy-Weinberg proportions are often reported together with p-values of the association tests. Others, however, argue that a typical chi-squared test is not sensitive to departure from Hardy-Weinberg proportions (see Bibliographical Comments in Sect. 2.7). Hence the power to detect genotyping errors is low. Random mating is a necessary condition for HWE, which is also a requirement for applying the allele-based association test (Chap. 3). But failure to reject the null hypothesis that Hardy-Weinberg proportions hold does not mean the null hypothesis is true. Deviation from Hardy-Weinberg proportions in case-control data may also imply inbreeding or population stratification.

Testing Hardy-Weinberg proportions is based on samples drawn from the population. When case-control data are used, testing Hardy-Weinberg proportions is usually based on controls when studying a rare disease, because then the population and control genotypic distributions are similar. On the other hand, deviation from Hardy-Weinberg proportions in cases may indicate association when it holds in the population (or, for a rare disease, controls). Association tests incorporating HWD have been proposed to improve efficiency and power to detect true associations.

When Hardy-Weinberg proportions do not hold, the inbreeding coefficient, F , has been used above. Using F , given the allele frequencies, the genotype frequencies can be written as

$$g_0 = q^2 + pqF, \quad g_1 = 2pq(1 - F), \quad g_2 = p^2 + pqF.$$

Hence, Hardy-Weinberg proportions hold if and only if $F = 0$.

2.4 Population Substructure

The case-control design for genetic association studies may be affected by population substructure. Two types of substructure are considered in the literature. One is population stratification (PS) and the other is cryptic relatedness (CR). Case-control samples may be affected by PS or CR or both. We give brief introductions to PS and CR in this section. More details and corrections for these two substructures will be deferred to Chap. 9.

2.4.1 Population Stratification

Population stratification often refers to the situation that the allele (or genotype) frequency of the marker changes across the subpopulations. However, when testing association between a marker and a disease, the hidden PS influences the test result only if the following two conditions are both satisfied:

- I. *The allele (or genotype) frequency of the marker varies across the subpopulations,*
- II. *The disease prevalence varies across the subpopulations.*

Suppose there are two subpopulations, denoted by Z_1 and Z_2 , with allele frequencies of a marker of interest $P_j = \Pr(B|Z_j)$ for $j = 1, 2$. In each subpopulation, penetrances are all equal to the disease prevalence $f_{0j} = f_{1j} = f_{2j} = k_j = \Pr(\text{case} | Z_j)$, $j = 1, 2$. Hence, there is no association between the marker and the disease in each subpopulation.

Suppose the subpopulations are known so that r_j cases and s_j controls can be drawn from the j th subpopulation. One example is that the subpopulations are defined by geographical regions or ethnicities. The total numbers of cases and controls are $r = r_1 + r_2$ and $s = s_1 + s_2$, respectively. Denote the estimates of allele frequencies using cases and controls from the j th subpopulation by $\hat{p}_{1j} = n_{Aj}/r_j$ and $\hat{p}_{0j} = m_{Aj}/s_j$, respectively, where n_{Aj} and m_{Aj} are the numbers of A alleles among cases and controls in the j th subpopulation. Then,

$$E(\hat{p}_{1j}) = E(\hat{p}_{0j}) = p_j, \quad j = 1, 2.$$

That is, the estimates using cases and controls have the same expectation in each subpopulation. When cases and controls from the two subpopulations are pooled (ignoring the existence of subpopulations), we have r cases and s controls. We estimate allele frequency from the r cases, denoted by $\hat{p}_1 = (n_{A1} + n_{A2})/r$, and from the s controls, denoted by $\hat{p}_0 = (m_{A1} + m_{A2})/s$. Their expectations, conditional on $\{r_i, s_i; i = 1, 2\}$, are given by

$$E(\hat{p}_1) = (r_1 p_1 + r_2 p_2)/r,$$

$$E(\hat{p}_0) = (s_1 p_1 + s_2 p_2)/s.$$

If $E(\hat{p}_1) \neq E(\hat{p}_0)$, then there is association between the disease and the marker at the total population level, even though there is no association in each subpopulation. This is spurious association caused by PS, which is ignored in the above calculations.

A sufficient condition for $E(\hat{p}_1) = E(\hat{p}_0)$ is $s_j = m r_j$ for all j , where m does not change with j . This condition is satisfied under the matched case-control design (Chap. 4). When $m = 1$, the design is called a matched-pair design, in which equal numbers of cases and controls are drawn from each subpopulation. Another sufficient condition for $E(\hat{p}_1) = E(\hat{p}_0)$ is $p_1 = p_2$, i.e., the allele frequency does not change across the subpopulations.

To take account of the known subpopulations, stratified estimates should be used, which are given by

$$\hat{p}_1 = \sum_{j=1}^2 w_j \frac{n_{Aj}}{r_j}, \quad \hat{p}_0 = \sum_{j=1}^2 w_j \frac{m_{Aj}}{s_j},$$

with weights $w_j = (r_j + s_j) / \sum_j (r_j + s_j)$, proportional to the size of the j th subpopulation. It follows that

$$E(\hat{p}_1) = E(\hat{p}_0) = \sum_j w_j p_j.$$

For PS to have an effect, it is necessary that the disease prevalence k_j varies across the subpopulations. The disease prevalence is not used in the above arguments because the subpopulations are known, so that cases and controls can be drawn separately from each subpopulation and then pooled. In practice, PS is latent (i.e., the subpopulations are not known), and definitions of subpopulations by geographical regions are not perfect (see discussion in Chap. 9). In this case, suppose r cases and s controls are drawn from the case population and control population, respectively. Then r_j cases (s_j controls) belong to the j th subpopulation for $j = 1, 2$, which are random and unknown. Note that

$$E(r_j/r) = \Pr(Z_j | \text{case}) = \frac{\Pr(Z_j)k_j}{\Pr(\text{case})},$$

$$E(s_j/s) = \Pr(Z_j | \text{control}) = \frac{\Pr(Z_j)(1 - k_j)}{1 - \Pr(\text{case})}.$$

If $k_1 = k_2$, there is no change in disease prevalence across the subpopulations, and $k_1 = k_2 = \Pr(\text{case})$. Hence, $E(r_j/r) = E(s_j/s)$, equivalent to a matched case-control design. Under this sampling design,

$$E(\hat{p}_1) = E\{E(\hat{p}_1 | r_1, r_2)\} = E(r_1/r)p_1 + E(r_2/r)p_2,$$

$$E(\hat{p}_0) = E\{E(\hat{p}_0 | s_1, s_2)\} = E(s_1/s)p_1 + E(s_2/s)p_2.$$

Hence $E(\hat{p}_1) = E(\hat{p}_0)$ if either $k_1 = k_2$ or $p_1 = p_2$.

2.4.2 Cryptic Relatedness

A simple model for cryptic relatedness in a population is that HWE does not hold due to unknown relatedness among individuals in the population. For this simple CR model, we assume the population does not contain subpopulations with varying allele frequencies. However, HWE does not hold because of unknown relatedness among individuals. This CR model is also studied in Chap. 9. In Sect. 2.4.1, we

considered the bias of estimates of allele frequencies using cases and controls in the presence of PS. The variance of the estimates would be affected in the presence of CR, which will be also discussed in Chap. 9.

One can also consider a more general model of CR with several subpopulations. In each subpopulation, HWE does not hold, but individuals across the subpopulations are not genetically related. In this generalized model, how the allele frequencies and disease prevalences change across the subpopulations is not specified. If, in addition to relatedness among individuals in each subpopulation, the allele frequencies and disease prevalences also change across the subpopulations, then the PS and CR can be studied simultaneously.

2.5 Odds Ratio and Relative Risk

2.5.1 Odds Ratios

Definitions

The odds ratio (OR) is commonly used to measure association in epidemiology. For a prospective case-control study, the odds of being a case versus a control for a given risk factor $R = E+$ (exposed) or $R = E-$ (not exposed) is defined as

$$\frac{\Pr(d = 1|R)}{\Pr(d = 0|R)}, \quad (2.14)$$

where $d = 1$ is for a case and $d = 0$ is for a control. The OR with respect to two levels of R is defined as

$$\text{OR}_{d=1:d=0} = \frac{\Pr(d = 1|E+)}{\Pr(d = 0|E+)} \bigg/ \frac{\Pr(d = 1|E-)}{\Pr(d = 0|E-)}.$$

For a retrospective case-control study, the odds of being $E+$ versus being $E-$ in cases ($d = 1$) or controls ($d = 0$) is defined as

$$\frac{\Pr(E+|d)}{\Pr(E-|d)}. \quad (2.15)$$

The OR with respect to case and control groups is

$$\text{OR}_{R=E+:R=E-} = \frac{\Pr(E+|d=1)}{\Pr(E-|d=1)} \bigg/ \frac{\Pr(E+|d=0)}{\Pr(E-|d=0)}.$$

It can be shown that

$$\frac{\Pr(E+|d=1) \Pr(E-|d=0)}{\Pr(E-|d=1) \Pr(E+|d=0)} = \frac{\Pr(d=1|E+) \Pr(d=0|E-)}{\Pr(d=1|E-) \Pr(d=0|E+)}.$$

Thus, the ORs for the retrospective and prospective case-control studies are identical. This property, along with its close relation with relative risk as mentioned below, makes the OR a widely used measure of association in epidemiology.

Table 2.7 A 2×2 table with case and control status and levels of a risk factor for n subjects

	$E+$	$E-$	
Case	a	b	$a + b$
Control	c	d	$c + d$
	$a + c$	$b + d$	n

Inference

For a general 2×2 table as given in Table 2.7, the estimate of the OR is given by

$$\widehat{\text{OR}} = \frac{ad}{bc}, \quad \text{or} \quad \log \widehat{\text{OR}} = \log \left(\frac{ad}{bc} \right).$$

Note that if any entry in Table 2.7 is 0, a constant $1/2$ is often added to each cell in the table. When there is no association in the 2×2 table, $\widehat{\text{OR}} \approx 1$. If the exposure to the risk factor ($E+$) increases the risk of having the disease, $\widehat{\text{OR}} > 1$.

A consistent estimate of the variance of the log OR can be written as (Problem 2.7)

$$\widehat{\text{Var}}(\log \widehat{\text{OR}}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}, \quad (2.16)$$

which is referred to as Woolf's estimate of the variance of the log OR. The confidence interval for the log OR can be obtained from

$$\frac{\log \widehat{\text{OR}} - \log \text{OR}}{\sqrt{\widehat{\text{Var}}(\log \widehat{\text{OR}})}} \rightarrow N(0, 1). \quad (2.17)$$

That is, $\log \widehat{\text{OR}} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\log \widehat{\text{OR}})}$. Denote $z = z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\log \widehat{\text{OR}})}$. Then, the $100(1 - \alpha)\%$ confidence interval for the OR is $(\widehat{\text{OR}}/e^z, \widehat{\text{OR}}e^z)$. Under the null hypothesis of no association H_0 , $\log \text{OR} = 0$. Thus, after substituting $\log \text{OR} = 0$, the left hand side of (2.17) can be used as a test statistic for association.

Odds Ratios for Genetic Associations

For a diallelic marker with alleles A and B , and three genotypes $(G_0, G_1, G_2) = (AA, AB, BB)$, case-control data can be displayed in a 2×3 table. For Table 2.8, two ORs can be used to measure association. One is between AB and AA , denoted as OR_1 . The other is between BB and AA , denoted as OR_2 . In both ORs, genotype $G_0 = AA$ is used as the reference.

The formulas for the estimates of the two log ORs and their asymptotic variances are given by

$$\log \widehat{\text{OR}}_1 = \log \left(\frac{r_{1s_0}}{r_{0s_1}} \right), \quad \widehat{\text{Var}}(\log \widehat{\text{OR}}_1) = \frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{s_1}, \quad (2.18)$$

Table 2.8 A 2×3 table with case and control status and three genotypes

	<i>AA</i>	<i>AB</i>	<i>BB</i>
Case	r_0	r_1	r_2
Control	s_0	s_1	s_2

$$\log \widehat{\text{OR}}_2 = \log \left(\frac{r_2 s_0}{r_0 s_2} \right), \quad \widehat{\text{Var}}(\log \widehat{\text{OR}}_2) = \frac{1}{r_0} + \frac{1}{r_2} + \frac{1}{s_0} + \frac{1}{s_2}. \quad (2.19)$$

The estimates $\log \widehat{\text{OR}}_1$ and $\log \widehat{\text{OR}}_2$ are negatively correlated with covariance (Problem 2.8)

$$\widehat{\text{Cov}}(\log \widehat{\text{OR}}_1, \log \widehat{\text{OR}}_2) = -\frac{1}{r_0} - \frac{1}{s_0}. \quad (2.20)$$

If one is interested in the OR between *AA* versus genotypes with at least one allele *B* (i.e., *AB* and *BB*), the estimate of the log OR can be written as

$$\log \widehat{\text{OR}} = \log \left(\frac{s_0(r_1 + r_2)}{r_0(s_1 + s_2)} \right),$$

with an estimated asymptotic variance

$$\widehat{\text{Var}}(\log \widehat{\text{OR}}) = \frac{1}{r_0} + \frac{1}{r_1 + r_2} + \frac{1}{s_0} + \frac{1}{s_1 + s_2}.$$

On the other hand, to calculate the OR between *BB* versus genotypes with at least one allele *A* (i.e., *AA* and *AB*), one has

$$\log \widehat{\text{OR}} = \log \left(\frac{r_2(s_0 + s_1)}{s_2(r_0 + r_1)} \right),$$

$$\widehat{\text{Var}}(\log \widehat{\text{OR}}) = \frac{1}{r_0 + r_1} + \frac{1}{r_2} + \frac{1}{s_0 + s_1} + \frac{1}{s_2}.$$

As before, infinite estimates and variances can be avoided by adding 1/2 to each cell in Table 2.8.

2.5.2 Relative Risks

Definition and Relation with Odds Ratio

Define $f_1 = \Pr(d = 1|E+)$ and $f_0 = \Pr(d = 1|E-)$. Then the relative risk (RR) of the disease on being exposed versus not being exposed is

$$\text{RR} = \frac{f_1}{f_0}.$$

If the data in Table 2.7 are collected in a prospective case-control design, the estimate of the RR is given by

$$\widehat{RR} = \frac{\widehat{f}_1}{\widehat{f}_0} = \frac{a(b+d)}{b(a+c)},$$

where $\widehat{f}_0 = b/(b+d)$ and $\widehat{f}_1 = a/(a+c)$. To obtain the asymptotic variance of $\widehat{RR}$, it is easier to work with the log RR. Note that, in a prospective study, $\widehat{f}_0$ and $\widehat{f}_1$ are independent. Thus, $\text{Var}\{\log(\widehat{f}_1/\widehat{f}_0)\} = \text{Var}(\log \widehat{f}_1) + \text{Var}(\log \widehat{f}_0)$. Denote $n_0 = b+d$ and $n_1 = a+c$. Then both a and b follow binomial distributions, $a \sim B(n_1; f_1)$ and $b \sim B(n_0; f_0)$. Thus, by the Delta method,

$$\text{Var}(\log \widehat{f}_i) \approx \frac{1}{f_i^2} \frac{f_i(1-f_i)}{n_i} = \frac{1-f_i}{f_i n_i}.$$

Hence,

$$\widehat{\text{Var}}\left\{\log\left(\frac{\widehat{f}_1}{\widehat{f}_0}\right)\right\} = \frac{c}{a(a+c)} + \frac{d}{b(b+d)}.$$

From the expression for the OR, we have

$$\text{OR}_{d=1:d=0} = \text{OR}_{R=E+:R=E-} = \frac{f_1(1-f_0)}{f_0(1-f_1)} = \frac{f_1}{f_0} \left(1 + \frac{f_1-f_0}{1-f_1}\right).$$

Thus, for a rare disease ($f_1 - f_0 \approx 0$ and $f_1 \approx 0$),

$$\text{OR}_{d=1:d=0} = \text{OR}_{R=E+:R=E-} \approx \text{RR}.$$

Genotype Relative Risks

For the data presented in Table 2.8, two GRRs can be defined,

$$\text{GRR}_1 = \Pr(d=1|AB)/\Pr(d=1|AA),$$

$$\text{GRR}_2 = \Pr(d=1|BB)/\Pr(d=1|AA).$$

When there is no association between the case-control status and the marker, $\text{GRR}_1 = \text{GRR}_2 = 1$. In general, unbiased estimates for GRRs are not available using retrospective case-control data.

2.6 Logistic Regression for Case-Control Studies

2.6.1 Prospective Case-Control Design

Likelihood Function

A logistic regression model is often used for the analysis of prospective case-control data. Let $d = 1$ denote a case and $d = 0$ a control. Denote $\mathbf{X} = (X_1, \dots, X_p)^T$ a vector of covariates and $H(\mathbf{X}) = (h_1(\mathbf{X}), \dots, h_p(\mathbf{X}))^T$, where h_i is a coding function or transformation of covariates. Then using logistic regression, one has

$$P(\mathbf{x}) = \Pr(d = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\beta_0 + \beta_1^T H(\mathbf{x})\}}{1 + \exp\{\beta_0 + \beta_1^T H(\mathbf{x})\}}$$

and $\Pr(d = 0 | \mathbf{X} = \mathbf{x}) = 1 - \Pr(d = 1 | \mathbf{X} = \mathbf{x})$. The prospective likelihood function can be written as

$$\begin{aligned} L_{\text{pros}}(\beta_0, \beta_1) &= \prod_{j=1}^n \{p(\mathbf{x}_j)\}^{d_j} \{1 - p(\mathbf{x}_j)\}^{1-d_j} \\ &= \prod_{j=1}^n \frac{\exp\{\beta_0 d_j + \beta_1^T H(\mathbf{x}_j) d_j\}}{1 + \exp\{\beta_0 + \beta_1^T H(\mathbf{x}_j)\}}. \end{aligned} \quad (2.21)$$

Inference for β_1 can be made using (2.21).

Examples

For the first example, consider a single binary covariate in the logistic regression model. Hence, $H(\mathbf{X}) = H(X) = 0$ for not exposed ($E-$) and 1 for exposed ($E+$). For the second example, consider a single genetic marker G as a covariate. Denote the three genotypes as $(G_0, G_1, G_2) = (AA, AB, BB)$. There are several ways to code (or score) the genotype G . For examples, a two-dimensional scoring function is $H(G) = (h_1(G), h_2(G))^T$, where $h_1(G) = 0, 0, 1$ and $h_2(G) = 0, 1, 1$ if $G = G_0, G_1, G_2$, respectively, and a one-dimensional scoring function is $H(G) = i$ if $G = G_i$ for $i = 0, 1, 2$.

2.6.2 Retrospective Case-Control Design

The retrospective case-control design differs from the prospective case-control design. The data in the retrospective design are not drawn from the general population. They are sampled from a population with selected samples ($S = 1$) of cases and controls. The proportion of cases in the selected population is often different from the

disease prevalence in the general population. In fact, this is an important difference between the retrospective and prospective designs.

In a retrospective case-control study, the covariates $\mathbf{X}$ of a case $d = 1$ or a control $d = 0$ in the selected population $S = 1$ are obtained. Thus, the likelihood of observing $\mathbf{X} = \mathbf{x}$ is

$$\Pr(\mathbf{X} = \mathbf{x} | d, S = 1),$$

where $d = 0$ or 1 . The likelihood function can be written as

$$\begin{aligned} L_{\text{retro}}(\beta_0, \beta_1) &= \prod_{j=1}^n \{\Pr(\mathbf{X}_j = \mathbf{x}_j | d_j = 1, S_j = 1)\}^{d_j} \\ &\quad \times \{1 - \Pr(\mathbf{X}_j = \mathbf{x}_j | d_j = 0, S_j = 1)\}^{1-d_j}. \end{aligned} \quad (2.22)$$

Denote $\tilde{p}(\mathbf{x}_j) = \Pr(d_j = 1 | S_j = 1, \mathbf{X}_j = \mathbf{x}_j)$ and $\pi_i = \Pr(S_j = 1 | d_j = i, \mathbf{X}_j = \mathbf{x}_j)$, $i = 0, 1$. Then it can be shown that

$$\text{logit} \tilde{p}(\mathbf{x}_j) = \frac{\tilde{p}(\mathbf{x}_j)}{1 - \tilde{p}(\mathbf{x}_j)} = \frac{\pi_1}{\pi_0} \frac{p(\mathbf{x}_j)}{1 - p(\mathbf{x}_j)} = \exp\{\tilde{\beta}_0 + \beta_1^T H(\mathbf{x}_j)\},$$

where $\tilde{\beta}_0 = \beta_0 + \log(\pi_1/\pi_0)$. Thus, the odds of observing $\mathbf{x}$ in the selected case-control samples is proportional to the odds of observing $\mathbf{x}$ in the population. Therefore, the OR in the selected population equals that in the population. The proportion

$$\pi_1/\pi_0 = \frac{\Pr(d_j = 1 | S_j = 1, \mathbf{x}_j)}{\Pr(d_j = 0 | S_j = 1, \mathbf{x}_j)} \bigg/ \frac{\Pr(d_j = 1 | \mathbf{x}_j)}{\Pr(d_j = 0 | \mathbf{x}_j)}$$

is the ratio of probabilities of cases to controls in the selected samples with respect to that in the population.

The likelihood (2.22) can be further written as

$$L_{\text{retro}}(\beta_0, \beta_1) = \prod_{j=1}^n \{\tilde{p}(\mathbf{x}_j)\}^{d_j} \{1 - \tilde{p}(\mathbf{x}_j)\}^{1-d_j} \quad (2.23)$$

$$\times \prod_{j=1}^n \left\{ \frac{\Pr(\mathbf{x}_j | S_j = 1)}{\Pr(d_j = 1 | S_j = 1)} \right\}^{d_j} \left\{ \frac{\Pr(\mathbf{x}_j | S_j = 1)}{\Pr(d_j = 0 | S_j = 1)} \right\}^{1-d_j}. \quad (2.24)$$

If (2.24) does not depend on the coefficient β_1 which appears in (2.23), then (2.23) can be used for the analysis of the retrospective data. Thus, the prospective likelihood function L_{pros} can be used for the retrospective case-control data for inference of β_1 .

2.7 Bibliographical Comments

This book focuses on the analysis of genetic case-control association studies. Therefore, only the background in population genetics needed for case-control association studies is introduced in this chapter. More about population genetics can be found in other textbooks or Refs. [49, 71], and [117]. In Chap. 13, we will discuss linkage and association studies using family data. Some background related to the analysis of family data will be given there. Basic statistical methods for testing association and other genetic studies, including linkage analysis and family-based association studies, can be found in [12, 165, 240, 245], and [299]. Elston et al. [74] reviewed multi-stage sampling for various genetic studies, including family and/or case-control data. The TDT was proposed by Spielman et al. [255]. For the GWAS design with 3,000 common controls shared with seven different diseases, refer to the WTCCC [301].

The Hardy-Weinberg principle was independently introduced by Hardy [115] and Weinberg [297] in 1908. A triangle diagram for Hardy-Weinberg proportions was given by Edwards [67]. Testing Hardy-Weinberg proportions and interpretation of deviation from Hardy-Weinberg proportions can be found in Weir [299] and Sham [240]. The example used to test Hardy-Weinberg proportions in Sect. 2.3.2 comes from Hartl and Clark [117]. Comparison of various asymptotic tests for Hardy-Weinberg proportions can be found in Emigh [76]. The exact test for Hardy-Weinberg proportions was first proposed by Haldane [112]. Guo and Thompson [109] studied exact tests for Hardy-Weinberg proportions for multiallelic loci. A definition of HWE on the X chromosome and its properties were given in Li [165]. Testing Hardy-Weinberg proportions on the X chromosome was studied by Zheng et al. [339]. Nielsen et al. [193] and Song and Elston [251] studied the departure from Hardy-Weinberg proportions in cases and/or case-control association studies. Li [164] showed that one can have Hardy-Weinberg proportions and yet the population is not in equilibrium.

The concept of LD between two loci and the use of D' can be found in Lewontin [162] and Weir [299]. The formulas (2.2) to (2.4) are studied by Zheng et al. [340]. Similar formulas were also given in Nielsen and Weir [195]. Lewontin did not mean his coefficient D' to refer solely to linked loci (personal communication) but rather to any two loci, and hence intended D' to be the more general measure of genetic phase disequilibrium. See also discussion of this in Wang et al. [293].

Population substructure is an important issue for genetic case-control association studies [66, 283]. The definition of PS can be also found in Crow and Kimura [49] (p. 54) and Eland-Johnson [71] (p. 228). The simple definition of CR was given by Voight and Pritchard [282]. The more general definition of CR given in Sect. 2.4.2 was used by Crow and Kimura [49] (p. 64), Eland-Johnson [71] (p. 213), Devlin and Roeder [60], and Whittemore [302]. Recent discussions can be found in Astle and Balding [10] and Zheng et al. [341].

Measures of risks and their inference can be found in many epidemiological or biostatistics textbooks, e.g., Fleiss et al. (Chaps. 7, 11 and 13) [86] and Sahai and Khurshid [222]. Using the prospective logistic regression model for the retrospective

case-control data was studied by Prentice and Pyke [204]. Their results hold for both the unmatched case-control design discussed here and for the matched case-control design discussed in Chap. 4. The derivation of the likelihood functions for the retrospective data presented here can be found in Sahai and Khurshid [222].

2.8 Problems

2.1 Let F_1 , F_2 , F_3 and F_4 be defined as in Sect. 2.2.1. Show that $F_1 F_4 = F_2 F_3$ if and only if $D = 0$.

2.2 Using the notation defined in Sect. 2.2.1, show that

$$\begin{aligned} f_0 &= f_0^* (F_1^2 + 2F_1 F_3 \lambda_1^* + F_3^2 \lambda_2^*), \\ f_1 &= f_0^* (F_1 F_2 + F_1 F_4 \lambda_1^* + F_2 F_3 \lambda_1^* + F_3 F_4 \lambda_2^*), \\ f_2 &= f_0^* (F_2^2 + 2F_2 F_4 \lambda_1^* + F_4^2 \lambda_2^*). \end{aligned}$$

2.3 Using (2.2) to (2.4), show that

$$\begin{aligned} f_1 - f_0 &= \frac{Df_0^*}{p(1-p)} \{F_1(\lambda_1^* - 1) + F_3(\lambda_2^* - \lambda_1^*)\}, \\ f_2 - f_1 &= \frac{Df_0^*}{p(1-p)} \{F_2(\lambda_1^* - 1) + F_4(\lambda_2^* - \lambda_1^*)\}. \end{aligned}$$

Further, when $D \neq 0$, $f_0 = f_1 = f_2$ holds if and only if $\lambda_1^* = \lambda_2^* = 1$ holds.

2.4 Using (2.6) and Table 2.4, derive (2.7).

2.5 Show that the chi-squared tests for Hardy-Weinberg proportions in (2.10) and (2.11) are identical.

2.6 Show that, ignoring higher order terms, the test for equal allele frequencies in males and females and the test for Hardy-Weinberg proportions in females are uncorrelated under H_{0a} and H_{0b} .

2.7 Under a prospective case-control design, $a \sim B(a+c; f_1)$ and $b \sim B(b+d; f_0)$ (binomial distributions). The OR is given by $OR = f_1(1-f_0)/\{f_0(1-f_1)\}$. Derive the variance of the estimate of log OR and show its estimate can be written as (2.16).

2.8 Derive the covariance of the estimates of ORs given in (2.18) and (2.19) using multinomial distributions for the genotype counts of cases and controls and the fact that the genotype counts of cases and controls are independent.

Analysis of Genetic Association Studies
Zheng, G.; Yang, Y.; Zhu, X.; Elston, R.C.
2012, XXII, 414 p., Hardcover
ISBN: 978-1-4614-2244-0