Chapter 6: Association analysis: Population based Studies

§ 6.1. Definition of allelic association

To map the disease genes, the linkage method usually leave a larger interval. Linkage analysis usually can map the disease genes to a interval of length from several cM to several tens cM. To further narrow down the interval, association analysis has been proved to be a useful method.

Consider two loci A and B, with alleles $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ occurring at frequencies $p_1, p_2, \ldots, p_m$ and $q_1, q_2, \ldots, q_n$ in the population, respectively. Consider haplotype with two alleles, one at locus A and the other at marker B. So, totally there are $mn$ haplotypes denoted by $A_1B_1, A_1B_2, \ldots, A_mB_1, \ldots, A_mB_n$ with the population frequencies $h_{11}, h_{12}, \ldots, h_{mn}$. If the occurrence of allele $A_i$ at locus A and the occurrence of $B_j$ at locus B in a haplotype are independent, then the frequency of haplotype $A_iB_j$ is equal to the product of the allele frequencies of $A_i$ and $B_j$, i.e.

\[ h_{ij} = p_iq_j \text{ or } P(A_iB_j) = P(A_i)P(B_j). \]

- If, for all $i$ and $j$, the above equation hold, loci A and B are said in linkage equilibrium.
- If the equation does not hold, then the two alleles $A_i$ and $B_j$ are said to be associated.
- If $h_{ij} > p_iq_j$ then $A_i$ and $B_j$ are said to be positively associated. Conversely, if $h_{ij} < p_iq_j$, then the two alleles are said to be negatively associated.

Example 1. Consider two biallelic loci A and B with alleles $A_1, A_2$ and $B_1, B_2$. Suppose that the frequencies of the four haplotypes $A_1B_1, A_1B_2, A_2B_1, A_2B_2$ are 0.4, 0.1, 0.2 and 0.3, respectively. Are there any allelic association between the two loci.

First calculate the allele frequencies of alleles $A_1, A_2, B_1$ and $B_2$

\[
\begin{align*}
P(A_1) &= P(A_1B_1) + P(A_1B_2) = 0.4 + 0.1 = 0.5 \\
P(A_2) &= 1 - P(A_1) = 0.5, \\
P(B_1) &= P(A_1B_1) + P(A_2B_1) = 0.4 + 0.2 = 0.6 \\
P(B_2) &= 1 - P(B_1) = 0.4.
\end{align*}
\]

Furthermore,

\[
\begin{align*}
P(A_1B_1) &= 0.4 > P(A_1)P(B_1) = 0.3 \\
P(A_1B_2) &= 0.1 < P(A_1B_2) = 0.2 \\
P(A_2B_1) &= 0.2 < P(A_2)P(B_1) = 0.3 \\
P(A_2B_2) &= 0.3 > P(A_2)P(B_2) = 0.2
\end{align*}
\]

So, $A_1$ and $B_1$ are positively associated; $A_1$ and $B_2$ are negatively associated; $A_2$ and $B_1$ are negatively associated; $A_2$ and $B_2$ are positively associated.
If two loci are not linkage equilibrium, the two loci are said to be in Linkage Disequilibrium (LD). Now, association and LD can be used interchangeably, that is, association also means that the two loci not in linkage equilibrium. For linkage disequilibrium, we can further define the quantities to measure how strong of the linkage disequilibrium. Let

\[ D_{ij} = h_{ij} - p_i q_j = P(A_i B_j) - P(A_i)P(B_j), \]

and

\[ D_{\text{max}} = \begin{cases} \min\{p_i q_j, (1 - p_i)(1 - q_j)\} & \text{if} \; D_{ij} < 0 \\ \min\{p_i(1 - q_j), (1 - p_i)q_j\} & \text{if} \; D_{ij} > 0. \end{cases} \]

The one of the most useful LD measure \( D' \) (Hedrick 1987) is defined by

\[ D' = \sum_{i=1}^{m} \sum_{j=1}^{n} p_i q_j |D'_{ij}|, \]

where \( D'_{ij} = \frac{D_{ij}}{D_{\text{max}}}. \)

For the case of the two biallelic loci A and B, we can easy verify that

\[ D_{11} = -D_{12} = -D_{21} = D_{22}. \]

So, the \( D' \) measure is usually defined by

\[ D = h_{11} - p_1 q_1 \\
D' = \begin{cases} \frac{D}{D_{\text{max}}} & D > 0 \\ \frac{\min\{p_1(1 - q_1), (1 - p_1)q_1\}}{\min\{p_1 q_1, (1 - p_1)(1 - q_1)\}} & D < 0. \end{cases} \]

The \( D' \in [-1, 1] \). The large value of \( |D'| \) means strong LD between the two loci. \( D' = 0 \) means that the two loci are statistically independent. Another useful LD measure is defined by

\[ \Delta^2 = \frac{D^2}{p_1 p_2 q_1 q_2}. \]

Note that, for two biallelic markers, we can define \( D \) using any one of the four haplotypes and the corresponding \( \Delta^2 \) and \( D' \).

\section*{§ 6.2. Maintenance of linkage disequilibrium (allelic association)}

Let’s consider two allelic markers A and B with alleles \( A_1, A_2 \) and \( B_1, B_2 \). Assume that the recombination rate between the two markers is \( \theta \). Let’s see how the linkage disequilibrium change with the evolution of the population. Suppose that at current generation the allele frequencies of \( A_1, A_2 \) and \( B_1, B_2 \) are \( p_1, p_2 \) and \( q_1, q_2 \), respectively. Let \( h_0 \) denote the frequency of the haplotype \( A_1 B_1 \) at current generation and \( h_{k} \) denote the frequencies of haplotype \( A_1 B_1 \) at \( k \)th generation from now. Under random mating, we know that \( p_i \) and \( q_i \) do not change with time. Let

\[ D_0 = h_0 - p_1 q_1 \]
is the quantity of the linkage disequilibrium at present time. At next generation ($P_i$ denote the probability in $i$th generation),

$$h_1 = P_1(A_1B_1) = (1 - \theta)h_0 + \theta P_0(A_1)P(B_1)$$

$$= (1 - \theta)h_0 + \theta p_1 q_1$$

So, the quantity of linkage disequilibrium at next generation is

$$D_1 = h_1 - p_1 q_1$$

$$= (1 - \theta)(h_0 - p_1 q_1)$$

$$= (1 - \theta)D_0.$$ 

Similarly, we can get

$$D_k = (1 - \theta)D_{k-1}$$

and so,

$$D_k = (1 - \theta)^k D_0.$$ 

We can see that the linkage disequilibrium will decay with the time. If the recombination rate is small, it will decay slowly. If the recombination rate is large, it will decay very fast.

§ 6.3. Association between two markers

The association designs are mainly two groups. One is family-based design, the other is population based-design. We consider population based design. In hereafter, all the markers used are codominant markers. Assume that there are two markers A and B with alleles $A_1, \ldots, A_m$ and $B_1, \ldots, B_n$ and frequencies $p_1, \ldots, p_m$ and $q_1, \ldots, q_n$. There are $nm$ possible haplotypes $A_iB_j$ ($1 \leq i \leq m, 1 \leq j \leq n$) with population frequencies $h_{ij}$. There are $m(m + 1)/2$ genotypes at the first marker $A_iA_j$ ($1 \leq i \leq j \leq m$) and there are $n(n + 1)/2$ genotypes at the second marker $B_kB_l$ ($1 \leq k \leq l \leq n$). So, consider the two markers together, there are totally $[m(m + 1)/2][n(n + 1)/2]$ genotypes. Denote the frequency of two marker genotype $A_iA_jB_kB_l$ by $g_{ijkl}$. Suppose we have $N$ sampled individuals with the marker genotypes data at the two markers. We can count the number of each of the two-marker genotypes. Assume that among the $N$ individuals there are $n_{ijkl}$ individuals with genotypes $A_iA_jB_kB_l$. We want to test if there is an association among the two markers.

We have three hypotheses

$H_{(0)}$: H-W equilibrium and no LD

$H_1$: H-W equilibrium and LD

$H_2$: No H-W equilibrium and LD.

If we state our test problems as

(1) $H_{(0)} \leftrightarrow H_1$: Test LD or association (under the assumption of H-W equilibrium).
(2) \( H_1 \leftrightarrow H_2 \): Test H-W (under the assumption that there is LD)

(3) \( H(0) \leftrightarrow H_2 \): Test H-W and LD simultaneously. Reject null means that either H-W is not true or there is LD.

Under \( H_2 \): The parameters are \( g_{ijkl} \). There are \([m(m + 1)/2][n(n + 1)/2] - 1\) free parameters.

Under \( H_1 \): The parameters are \( h_{ij} \) \((i = 1, \ldots, m; j = 1, \ldots, n)\). There are \(mn - 1\) parameters. There are relations

\[
\begin{align*}
g_{iikk} &= h_{ik}^2 \\
g_{ijkl} &= 2h_{ik}h_{il} \\
g_{ijkk} &= 2h_{ik}h_{jk} \\
g_{ijkl} &= 2(h_{ik}h_{jl} + h_{ij}h_{jk})
\end{align*}
\]

Under \( H(0) \): The parameters are \( p_i, q_j \) \((i = 1, \ldots, m; j = 1, \ldots, n)\). There are \(m + n - 2\) free parameters. There are relations.

\[
\begin{align*}
g_{iikk} &= h_{ik}h_{ik} = (p_iq_k)^2 \\
g_{ijkl} &= 2h_{ik}h_{il} = 2p_i^2q_kq_l \\
g_{ijkk} &= 2h_{ik}h_{jk} = 2p_ip_jq_k^2 \\
g_{ijkl} &= 2(h_{ik}h_{jl} + h_{ij}h_{jk}) = 4p_ip_jq_kq_l
\end{align*}
\]

You can construct likelihood ration test as following:

Under the full model \( H_2 \), the log-likelihood

\[
L_2(\theta_2) = \sum n_{ijkl} \log(g_{ijkl})
\]

Here the parameter \( \theta_2 = (g_{ijkl}) \) for \( 1 \leq i \leq j \leq m, 1 \leq k \leq l \leq n \). The maximum likelihood estimator of \( g_{ijkl} \) is \( \hat{g}_{ijkl} = \frac{n_{ijkl}}{n} \). So, the

\[
\max_{H_2} L_2(\theta_2) = \sum n_{ijkl} \log(\hat{g}_{ijkl})
\]

Under model \( H_1 \), the log-likelihood function

\[
L_1(\theta_1) = \sum n_{ijkl} \log(g_{ijkl})
\]

where

\[
\begin{align*}
g_{iikk} &= h_{ik}h_{ik} \\
g_{ijkl} &= 2h_{ik}h_{il} \\
g_{ijkk} &= 2h_{ik}h_{jk} \\
g_{ijkl} &= 2(h_{ik}h_{jl} + h_{ij}h_{jk})
\end{align*}
\]
and $g_{ijkl}$ is not the free parameters, but $\theta_1 = (h_{ik})$ are parameters totally $nm - 1$ free parameter. The MLE of $h_{ij}$ can be obtained by EM algorithm which is a special case of our midterm project. Let the $h_{ij}$ is the MLE of $h_{ij}$; then,

$$\max_{H_1} L_1(\theta_1) = \sum n_{ijkl} \log(\hat{g}_{ijkl}^{(1)})$$

where

$$\hat{g}_{ikk}^{(1)} = \hat{h}_{ik} \hat{h}_{ik},$$
$$\hat{g}_{ikl}^{(1)} = 2 \hat{h}_{ik} \hat{h}_{il},$$
$$\hat{g}_{ijk}^{(1)} = 2 \hat{h}_{ik} \hat{h}_{jk},$$
$$\hat{g}_{ijkl}^{(1)} = 2 (\hat{h}_{ik} \hat{h}_{jl} + \hat{h}_{ij} \hat{h}_{jk}).$$

Under $H_0$ the log-likelihood function

$$L_0(\theta_0) = \sum n_{ijkl} \log(g_{ijkl})$$

where

$$g_{ikk} = (p_i q_k)^2,$$
$$g_{ikl} = 2 p_i^2 q_k q_l,$$
$$g_{ijk} = 2 p_i p_j q_k^2,$$
$$g_{ijkl} = 4 p_i p_j q_k q_l,$$

and $g_{ijkl}$ is not the free parameters, but $\theta_0 = (p_i, q_i)$ are parameters totally $m + n - 2$ free parameter. The MLE of $p_i$ and $q_i$ can be obtained by counting how many $A_i$ at first marker. Let $n_{1i}$ denote the number of allele $A_i$ ($i = 1, \ldots, m$) and $n_{2i}$ denote the number of $B_i$($i = 1, \ldots, n$).

Then,

$$L_0(\theta_0) = \sum_{i=1}^{m} n_{1i} \log(p_i) + \sum_{i=1}^{n} n_{2i} \log(q_i).$$

the MLE of $p_i$ is $\hat{p}_i = \frac{n_{1i}}{2N}$ and the MLE of $q_i$ is $\hat{q}_i = \frac{n_{2i}}{2N}$. Thus,

$$\max_{H_1} L_0(\theta_0) = \sum_{i=1}^{m} n_{1i} \log(\hat{p}_i) + \sum_{i=1}^{n} n_{2i} \log(\hat{q}_i).$$

In summary, the LRT for test $H_0 \leftarrow \rightarrow H_1$, that is to test LD or association when we assume H-W equilibrium, is

$$G^2_1 = 2 (\max_{H_1} L_1 - \max_{H_0} L_0)$$

which has an asymptotic distribution $\chi^2_{(n-1)(m-1)} [mn - 1 - (m + n - 2) = (m - 1)(n - 1)];$
The LRT for test $H_1 \leftrightarrow H_2$, that is, to test H-W equilibrium, is

$$G_2^2 = 2(\max_{H_2} L_2 - \max_{H_1} L_1)$$

which has an asymptotic distribution $\chi^2_d \left[ d = \left( \frac{m(m+1)}{2} n(n+1) - 1 - mn - 1 \right) = mn \left( \frac{(m+1)(n+1)}{4} - 1 \right) \right]$;

The LRT for test $H(0) \leftrightarrow H_2$, that is, to test H-W equilibrium and association simultaneously, is

$$G_2^2 = 2(\max_{H_2} L_2 - \max_{H_0} L_0)$$

which has an asymptotic distribution $\chi^2_s \left[ s = \frac{m(m+1)}{2} n(n+1) - m - n \right]$.

§ 6.4 Test for Association for complex disease: Single Marker Methods based on Case-Control Studies

§ 6.4.1 Theoretical Preparation

**Theorem 1** 6.4.1. Let $Y = (y_1, \ldots, y_M)$ follow a multinomial distribution $M_M(n, p)$, and $X = (x_1, \ldots, x_M)$ follow a multinomial distribution $M_M(m, q)$, there $p = (p_1, \ldots, p_M)$ and $q = (q_1, \ldots, q_M)$. The score test statistic to test the hypothesis

$$H_0 : p = q \text{ vs } H_1 : p \neq q$$

is given by

$$S^2 = \sum_{i=1}^M \left( \frac{\hat{p}_i - \hat{q}_i}{\frac{\hat{p}_i}{n} + \frac{\hat{q}_i}{n}} \right)^2$$

which asymptotically follow a $\chi^2$ distribution with degrees of freedom $M - 1$, where $\hat{p}_i = \frac{n_i}{n}$ and $\hat{q}_i = \frac{x_i}{m}$. When $m = n$,

$$S^2 = n \sum_{i=1}^M \left( \frac{\hat{p}_i - \hat{q}_i}{\hat{p}_i + \hat{q}_i} \right)^2$$

§ 6.4.2 Allele-based Method

1. $\chi^2$ test

   Consider a case-control study on genetic association between a biallelic marker $A$ with $M$ alleles $A_1, \ldots, A_m$. The data for $n$ cases and $m$ controls are given in Table 1

<table>
<thead>
<tr>
<th>Number of Allele</th>
<th>Total Number of alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$A_M$</td>
<td>$n_M$</td>
</tr>
<tr>
<td>Cases</td>
<td>$2n$</td>
</tr>
<tr>
<td>Controls</td>
<td>$2m$</td>
</tr>
</tbody>
</table>

Table 1. $2 \times M$ contingency table for case-control studies using a marker with $M$ allele.
Let \( p_i \) and \( q_i \) denote the allele frequencies of allele \( A_i \) in cases ans in controls, respectively, and \( p = (p_1, \ldots, p_M) \) and \( q = (q_1, \ldots, q_M) \). Testing association between the marker and disease locus is equivalent to testing the following hypothesis

\[
H_0 : p = q \text{ vs } H_1 : p \neq q.
\]  

(1)

Since \( (n_1, \ldots, n_M) \sim M_M(2n, p) \) and \( (m_1, \ldots, m_M) \sim M_M(2m, q) \), the score test statistic is given by (using Theorem 6.4.1)

\[
T_a = \sum_{i=1}^{M} \frac{(\hat{p}_i - \hat{q}_i)^2}{\hat{p}_i + \frac{\hat{q}_i}{2n}}
\]

which follows a Chi-square distribution with degrees of freedom \( M - 1 \), where \( \hat{p}_i = \frac{n_i}{2n} \) and \( \hat{q}_i = \frac{m_i}{2m} \).

For a biallelic marker with two alleles \( A \) and \( a \). The data can be represented in a \( 2 \times 2 \) contingency table (Table 2).

<table>
<thead>
<tr>
<th></th>
<th>Number of allele ( A )</th>
<th>Number of allele ( a )</th>
<th>Total number of alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>( n_1 )</td>
<td>( n_2 )</td>
<td>( 2n )</td>
</tr>
<tr>
<td>Controls</td>
<td>( m_1 )</td>
<td>( m_2 )</td>
<td>( 2m )</td>
</tr>
</tbody>
</table>

At this case, the statistic becomes [ Note that \( (\hat{p}_1 - \hat{q}_1)^2 = (\hat{p}_2 - \hat{q}_2)^2 \)]

\[
T_a = \frac{8(n + m)nm}{(n_1 + m_1)(n_2 + m_2)}(\hat{p}_1 - \hat{q}_1)^2.
\]

(2)

which asymptotically follows a \( \chi^2_1 \) distribution.

To see why testing association is equivalent to test the hypothesis given in (1), let’s consider the following arguments. The formulas given below are also very usefulto generate data sets for simulation studies. We only consider biallelic markers.

Assume there is a disease locus with two allele \( D \) and \( d \) with allele frequencies \( p_D \) and \( p_d \), respectively. Let’s see the allele frequency difference between cases and controls. Let

\[
f_{DD} = P(affected|DD) \\
f_{Dd} = P(affected|Dd) \\
f_{dd} = P(affected|dd)
\]

be the three penetrances. We assume the H-W equilibrium. Then the prevalence.

\[
K = P(affected) = f_{DD}p_D^2 + 2f_{Dd}p_DP_d + f_{dd}p_d^2
\]
Thus,

\[ P(DD|\text{affected}) = \frac{P(\text{affected}|DD)P(DD)}{P(\text{affected})} = \frac{f_{DD}p_D^2}{K} \]

\[ P(Dd|\text{affected}) = \frac{2f_{Dd}p_Dp_d}{K} \]

\[ P(dd|\text{affected}) = \frac{f_{dd}p_d^2}{K} \]

For the normal individuals,

\[ P(DD|\text{Normal}) = \frac{P(\text{Normal}|DD)P(DD)}{P(\text{Normal})} = \frac{(1-f_{DD})p_D^2}{1-K} \]

\[ P(Dd|\text{Normal}) = \frac{2(1-f_{Dd})p_Dp_d}{1-K} \]

\[ P(dd|\text{Normal}) = \frac{(1-f_{dd})p_d^2}{1-K} \]

Assume that D is high risk alleles, then \( f_{DD} \geq f_{Dd} \geq f_{dd} \) and \( f_{DD} = f_{Dd} = f_{dd} \) does not hold.

\[
P(DD|\text{affected}) - P(DD|\text{Normal}) = \frac{f_{DD}p_D^2 - (1-f_{DD})p_D^2}{K} - \frac{1}{1-K} \]

\[
= \frac{f_{DD}p_D^2 - Kf_{DD}p_D^2 - Kp_D^2 + Kf_{DD}p_D^2}{K(1-K)} \]

\[
= \frac{p_D^2(f_{DD} - K)}{K(1-K)} \]

Since, \( K = f_{DD}p_D^2 + 2f_{Dd}p_Dp_d + f_{dd}p_d^2 \) and \( f_{DD}(p_D^2 + 2p_Dp_d + p_d^2) = f_{DD} \). So, \( P(DD|\text{affected}) - P(DD|\text{Normal}) > 0 \).

If we consider the allele frequencies, the allele frequency of allele D in the affected individuals is given by

\[
P(D|\text{affected}) = \frac{P(D, \text{affected})}{K} = \frac{P(DD, \text{Affected}) + P(Dd, \text{Affected})}{K} = \frac{f_{DD}p_D^2 + 2f_{Dd}p_D(1-p_D)}{K} \]
and the allele frequency of D in normal individuals is given by

\[
P(D|Normal) = \frac{P(D, Normal)}{K} = \frac{P(DD, Normal) + P(Dd, Normal)}{K} = \frac{(1 - f_{DD})p_D^2 + 2(1 - f_{Dd})p_D(1 - p_D)}{1 - K}
\]

So,

\[
P(D|affected) - P(D|Normal) = \left[ f_{DD}p_D^2 + 2f_{Dd}p_D(1 - p_D) \right] (1 - K) - \left[ (1 - f_{DD})p_D^2 + 2(1 - f_{Dd})p_D(1 - p_D) \right] K
\]

\[
= \left[ f_{DD}p_D^2 + 2f_{Dd}p_D(1 - p_D) - K(p_D^2 + 2p_D(1 - p_D)) \right] K
\]

\[
= \left[ (K - f_{dd}p_d) - K(1 - p_D^2) \right] K(1 - K)
\]

\[
= \left( \frac{K - f_{dd}p_d}{K(1 - K)} \right) \geq 0.
\]

and \(\frac{(K - f_{dd}p_d)}{K(1 - K)} = 0\) if and only if \(f_{DD} = f_{Dd} = f_{dd}\), that is,

\[
P(D|affected) - P(D|Normal) = 0 \text{ iff } f_{DD} = f_{Dd} = f_{dd} \quad (3)
\]

Consider a marker B with two possible alleles \(B_1\) and \(B_2\) and frequencies \(q_1, q_2\), denote \(h_{ij}\) is the frequency of \(B_iD_j\) (here \(D_1 = D\) and \(D_2 = d\)), then

\[
P(Affected|B_1B_1) = \frac{P(affected, B_1B_1)}{q_1^2} = \frac{P(affected, B_1D/B_1D) + P(affected, B_1D/B_1d) + P(affected, B_1d/B_1d)}{q_1^2}
\]

\[
= \frac{f_{DD}h_{11}^2 + 2f_{Dd}h_{11}h_{12} + f_{dd}h_{12}^2}{q_1^2}
\]

\[
P(Affected|B_1B_2) = \frac{P(affected, B_1B_2)}{2q_1q_2}
\]

\[
= \frac{P(aff, B_1D/B_2D) + P(aff, B_1D/B_2d) + P(aff, B_1d/B_2D) + P(aff, B_1d/B_2d)}{2q_1q_2}
\]

\[
= \frac{f_{DD}h_{11}h_{21} + f_{Dd}(h_{11}h_{22} + h_{12}h_{21}) + f_{dd}h_{12}h_{22}}{q_1q_2}
\]

\[
P(Affected|B_2B_2) = \frac{f_{DD}h_{21}^2 + 2f_{Dd}h_{21}h_{22} + f_{dd}h_{22}^2}{q_2^2}
\]
Let
\[ D = P(B_1D_1) - pDq_1 = h_{11} - pDq_1, \]  
then
\[ D = -(h_{12} - q_1p_d) = -(h_{21} - q_2p_d) = h_{22} - q_2p_d \]  
(5)

To prove that testing association is equivalent to testing the hypothesis given in (1), we only need to prove
\[ D = 0 \iff P(B_1|\text{affected}) = P(B_1|\text{normal}). \]

**Home Work:** (Due data: November 5th)

1. Calculate \( P(\text{Normal}|B_iB_j) \) and then \( P(B_iB_j|\text{affected}), P(B_iB_j|\text{normal}), P(B_i|\text{affected}) \) and \( P(B_i|\text{normal}) \). Show that \( P(B_i|\text{affected}) \) and \( P(B_i|\text{normal}) \) are different if there is an association between Marker and disease locus \( (D \neq 0) \), and \( P(B_i|\text{affected}) = P(B_i|\text{normal}) \) if no association \( (D = 0) \).

[Hint: Let
\[ f_{11} = P(\text{Affected}|B_1B_1) \]
\[ f_{12} = P(\text{Affected}|B_1B_2) \]
\[ f_{22} = P(\text{Affected}|B_1B_2) \]

Similar to the proof of equation (3), we can prove that \( P(B_1|\text{affected}) = P(B_1|\text{normal}) \) if and only if \( f_{11} = f_{12} = f_{22} \). Thus, we only need to prove
\[ D = 0 \iff f_{11} = f_{12} = f_{22}. \]

From equation (4) and (5), after some simplification, we have,
\[ f_{11} - f_{12} = \frac{D}{q_1q_2}[\bigl(f_{DD} - f_{Dd}\bigr)h_{11} + \bigl(f_{Dd} - f_{dd}\bigr)h_{12}]. \]

Thus, \( D = 0 \) if and only if \( f_{11} = f_{12} \), provide \( h_{11} \neq 0 \) and \( h_{12} \neq 0 \].

2. Prove Theorem 6.4.1.

2. Likelihood Ratio Test

First, we consider one marker with \( M \) alleles \( A_1, \ldots, A_M \). The data are summarized in Table 1. Test the hypothesis \( H_0 : p = q \) vs \( H_1 : p \neq q \).

The log-likelihood function under \( H_0 \) and \( H_1 \) are
\[ L_0(p) = \sum_{i=1}^{m} (n_i + m_i) \log p_i \]
\[ L_1(p, q) = \sum_{i=1}^{m} n_i \log p_i + \sum_{i=1}^{m} m_i \log q_i \]
The MLE of $p_i$ under $H_0$ is $\hat{p}_i = \frac{(n_i+m_i)}{2(n+m)}$. The MLE of $p_i$ and $q_i$ under $H_1 \cup H_1$ are $\hat{p}_i = \frac{n_i}{2n}$ and $\hat{q}_i = \frac{m_i}{2m}$. The likelihood ratio test statistic is 

$$G^2 = 2(L_1(\hat{p}, \hat{q}) - L_0(\hat{p}_0))$$

which approximately has $\chi^2$ distribution with degree freedom $M-1$, where $\hat{p}_0 = (\hat{p}_{10}, \ldots, \hat{p}_{M0})$, $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_M)$, and $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_M)$.

§ 6.4.3 The False-positive of the $\chi^2$ Test Due to Population Stratification

The $\chi^2$ test statistic given in equation (2) can be also written as

$$T_a = \frac{(\hat{p}_1 - \hat{q}_1)^2}{\hat{\sigma}^2}.$$ 

where $\hat{\sigma}^2$ is an estimator of the variance of $\hat{p}_1 - \hat{q}_1$. It of a valid test under the assumption that the population is homogeneous, that is, if the underlying population is homogeneous and there is no association between the trait and the marker, $E(\hat{q}_d - \hat{q}_c) = 0$ and the test statistic $T_p$ has a asymptotic chi-square distribution with one degree of freedom. However, if the underlying population is not homogeneous, e.g., there is population stratification, the expectation of $\hat{q}_d - \hat{q}_c$ may not be zero and statistical inference based on the test statistic $T_p$ may be biased. For example, let us assume that the sampled individuals come from two different subpopulations and there is no association between marker $A$ and disease locus within each subpopulation. Let $q_1$ and $q_2$ denote the allele $A$ frequency within subpopulations 1 and 2, respectively, $f_1$ denote the probability that a sampled affected individual is from the first subpopulation, and $g_1$ denote the probability that a sampled normal individual is from the first subpopulation. Then we have

$$E(\hat{q}_d - \hat{q}_c) = q_1f_1 + q_2(1 - f_1) - (q_1g_1 + q_2(1 - g_1))$$

$$= (q_1 - q_2)(f_1 - g_1).$$

Therefore, if the allele frequencies differ between the two subpopulations and $f_1 \neq g_1$ (i.e., the disease prevalence is different between the two subpopulations), then $E(\hat{q}_d - \hat{q}_c) \neq 0$. Therefore, the simple chi-square test statistic for the 2 × 2 table does not have a chi-square distribution with one degree of freedom even when there is no disease-marker association in each subpopulation. Such statistical tests that ignore population heterogeneity may lead to erroneous conclusions.

As an example, we can consider an extreme cases. There are two subpopulations. Within each subpopulation there are no association between marker and disease locus. In the first subpopulation, the allele frequency is $q_1 = 0.7$ while in the second subpopulation, the allele frequency is $q_2 = 0.3$. If we sample 100 cases from the first population and 100 controls from the second population, then $\hat{q}_a \approx q_1 = 0.7, \hat{q}_c \approx q_2 = 0.3$ and

$$Var(\hat{q}_a - \hat{q}_c) = \frac{0.7 \times 0.3 + 0.7 + 0.3}{\sqrt{200}} = 0.029$$
and
\[ X^2 = \frac{(0.7 - 0.3)^2}{0.029} = \frac{0.16}{0.029} = 5.5 \]
the p-value is 0.019. Then your conclusion is that there is an association between the markers and the disease. This is called false-positive due to population stratification. So, when you do the sampling it is very important to match the cases and controls.

§ 6.4.4 Genotype-based Method

1. \( \chi^2 \) Test or Goodness of Fit Test

Consider a marker with \( L \) alleles \( A_1, \ldots, A_L \). Let \( g_1, \ldots, g_M \) denote all the distinct genotype observed in the sample, where \( M \leq \frac{L(L+1)}{2} \). Let \( p_i \) and \( q_i \) denote the genotype frequencies of genotype \( g_i \) in cases and in controls, respectively, and \( p = (p_1, \ldots, p_M) \) and \( q = (q_1, \ldots, q_M) \). The data can summarized as the following table:

<table>
<thead>
<tr>
<th>Number of genotype</th>
<th>Total Number of genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1 )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( g_2 )</td>
<td>( n_2 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( g_M )</td>
<td>( n_M )</td>
</tr>
<tr>
<td>Cases</td>
<td>( n )</td>
</tr>
<tr>
<td>( m_1 )</td>
<td>( m_1 )</td>
</tr>
<tr>
<td>( m_2 )</td>
<td>( m_2 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( m_M )</td>
<td>( m_M )</td>
</tr>
<tr>
<td>Controls</td>
<td>( m )</td>
</tr>
</tbody>
</table>

Testing association between the marker and disease locus is equivalent to testing the following hypothesis:

\[ H_0 : p = q \] vs \( H_1 : p \neq q. \]

Since \( (n_1, \ldots, n_M) \sim M_M(n, p) \) and \( (m_1, \ldots, m_M) \sim M_M(m, q) \), the score test statistic is given by (using Theorem 6.4.1)

\[ T_g = \sum_{i=1}^{M} \left( \frac{\hat{p}_i - \hat{q}_i}{\frac{n_i}{n} + \frac{m_i}{m}} \right)^2 \]

which follows a Chi-square distribution with degrees of freedom \( M - 1 \), where \( \hat{p}_i = \frac{n_i}{n} \) and \( \hat{q}_i = \frac{m_i}{m} \). The score test here is called \( \chi^2 \) test or goodness of fit test.

2. Likelihood Ratio Test

Then the log-likelihood under null (no association between disease locus and the marker) genotype frequencies are same in cases and controls is given by

\[ L_0(p) = \sum_{i=1}^{m} (n_i + m_i) \log p_i \]

MLE of \( p_i \) is \( \hat{p}_{i0} = \frac{n_i}{n} \). Under alternative hypothesis, the likelihood is given by

\[ L_1(p, q) = \sum_{i=1}^{m} n_i \log p_i + \sum_{i=1}^{m} m_i \log q_i, \]
the MLE of \( p_i \) and \( q_i \) are \( \hat{p}_i = \frac{n_i}{n}, \hat{q}_i = \frac{m_i}{m} \), so the likelihood ratio statistic is
\[
G^2 = 2(L_1(\hat{p}, \hat{q}) - L_0(\hat{p}_0))
\]
which approximately has a \( \chi^2 \) distribution with degree freedom \( 2 \ast [M - 1] - [M - 1] = M - 1 \).

For more than one markers, in the same way, we can consider multimarker genotypes. The methods is exact the same.

§ 6.4.3 Logistic regression method
Consider a biallelic with alleles \( A_1 \) and \( A_2 \). We have \( n \) individuals and their genotypes \( g_1, \ldots, g_n \), where each genotype may be one of the three forms \( A_1A_1, A_1A_2 \) and \( A_2A_2 \). Let \( x_1 \) and \( x_2 \) be two variables. For \( i \)th individual, define
\[
x_{i1} = \begin{cases} 1, & A_1A_1 \\ 0, & A_1A_2 \\ -1, & A_2A_2 \end{cases}
\]
and
\[
x_{i2} = \begin{cases} 0, & A_1A_1 \\ 1, & A_1A_2 \\ 0, & A_2A_2 \end{cases}
\]
as additive and dominant genetic score. Let \( p_i = P(\text{individual } i \text{ given his/her genotype}) \). Assume the logistic linear model
\[
\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}
\]
Let \( y = 1 \) denote the disease and \( y = 0 \) denote normal. Now, the data is \((y_1, g_1), \ldots, (y_n, g_n)\) . The log-likelihood function of \( y_i \) is
\[
l(\beta) = \sum_{i=1}^{n} \left( y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)
\]
\[
= \sum_{i=1}^{n} \left( y_i \log \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}} - (1 - y_i) \log(1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}) \right)
\]
\[
= \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}})
\]
where \( p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}} \) and \( 1 - p = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}} \). To test association, we can test the null hypothesis
\[
H_0 : \beta_1 = \beta_2 = 0.
\]

Considering a more general case, suppose there are \( k \) independent variable that satisfy the logistic linear model
\[
\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.
\]
Similarly, the log-likelihood can be written as \((x_{i0} = 1)\)

\[
 l(\beta) = \sum_{i=1}^{n} y_i (\sum_{j=0}^{k} \beta_j x_{ij}) - \sum_{i=1}^{n} \log(1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}})
\]

We want to test the null hypothesis

\[
 H_0 : \beta_1 = \ldots = \beta_k = 0.
\]

1. Score test.

Let \(x_i = (x_{i0}, \ldots, x_{ik})^T\), \(\beta = (\beta_0, \ldots, \beta_k)\), and \(p(\beta, x_i) = p_i\), then

\[
 p(\beta, x_i) = p_i = \frac{e^{\sum_{j=0}^{k} \beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}}} = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}
\]

\[
 1 - p(\beta, x_i) = 1 - p_i = \frac{1}{1 + e^{x_i^T \beta}} = \frac{1}{1 + e^{x_i^T \beta}}.
\]

The log-likelihood function

\[
 l(\beta) = \sum_{i=1}^{n} y_i (\sum_{j=0}^{k} \beta_j x_{ij}) - \sum_{i=1}^{n} \log(1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}})
\]

\[
 = \sum_{i=1}^{n} y_i x_i^T \beta + \sum_{i=1}^{n} \log(1 + e^{x_i^T \beta})
\]

The score

\[
 \frac{\partial l(\beta)}{\partial \beta_l} = \sum_{i=1}^{n} y_i x_{il} - \sum_{i=1}^{n} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} x_{il}
\]

\[
 = \sum_{i=1}^{n} x_{il} (y_i - p(\beta, x_i))
\]

for \(l = 0, \ldots, k\).

Let \(H(\beta) = (h_{ml}(\beta))_{(k+1) \times (k+1)}\) denote Hession matrix, where

\[
 h_{ml}(\beta) = \frac{\partial^2 l(\beta)}{\partial \beta_m \partial \beta_l} = -\sum_{i=1}^{n} \frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2} x_{im} x_{il}
\]

\[
 = -\sum_{i=1}^{n} x_{im} x_{il} (1 - p(\beta, x_i)) p(\beta, x_i).
\]
Let

\[ X = \begin{pmatrix}
1 & x_{11} & \cdots & x_{1k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & \cdots & x_{nk}
\end{pmatrix}_{(n+1) \times (n+1)} \]

and \( W = \text{diag}(p(\beta, x_1)(1 - p(\beta, x_1)), \ldots, p(\beta, x_k)(1 - p(\beta, x_k))) \) denote a \( n \times n \) matrix. Then \( H(\beta) = -X^T W X \). Let \( I(\beta) \) denote the Fisher information matrix, the \((m, l)\)th element of \( I(\beta) \) is \( I_{ml}(\beta) = -E(\frac{\partial^2 l(\beta)}{\partial \beta_m \partial \beta_l}) = -h_{ml} \). So, in this case \( I(\beta) = -H \).

Under null hypothesis, the likelihood function becomes

\[ l(\beta) = \sum_{i=1}^{n} y_i \left( \sum_{j=0}^{k} \beta_j x_{ij} \right) - \sum_{i=1}^{n} \log(1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}}) \]

\[ = \sum_{i=1}^{n} y_i \beta_0 + \sum_{i=1}^{n} \log(1 + e^{\beta_0}) \]

and

\[ \frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} p_0 = 0. \]

Thus, we have the MLE of \( p_0 \) is \( \hat{p}_0 = \bar{y} \), where \( p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \). The score

\[ \frac{\partial l(\beta)}{\partial \beta_i} = \sum_{i=1}^{n} y_i x_{it} - \sum_{i=1}^{n} \frac{e^{\beta_0} x_{it}}{1 + e^{\beta_0} x_{it}} = \sum_{i=1}^{n} y_i x_{it} - \sum_{i=1}^{n} \frac{e^{\beta_0} x_{it}}{1 + e^{\beta_0} x_{it}} \]

\[ = \sum_{i=1}^{n} (y_i - p_0) x_{it} \]

Thus, under null hypothesis, the score vector

\[ S(\beta)|_{\beta=\hat{\beta}} = \left( \frac{\partial l(\beta)}{\partial \beta_0}, \ldots, \frac{\partial l(\beta)}{\partial \beta_k} \right)^T |_{\beta=\hat{\beta}} = X^T (y - \bar{y} 1), \]

where \( y = (y_1, \ldots, y_n)^T \) and \( 1 = (1, \ldots, 1)^T \). The Fisher information matrix \( I = \bar{y}(1 - \bar{y}) X^T X \). Score test statistic is

\[ T_s = \frac{1}{\bar{y}(1 - \bar{y})} (y - \bar{y} 1)^T X(X^T X)^{-1} X^T (y - \bar{y} 1) \]

which has an asymptotic distribution \( \chi^2_k \).

### 2. Likelihood Ratio Test

Under null hypothesis, \( \beta_1 = \cdots = \beta_k = 0 \), and \( \hat{\beta}_0 = \log(\frac{\hat{p}_0}{1 - \hat{p}_0}) = \log(\frac{\bar{y}}{1 - \bar{y}}) \). The maximum likelihood
\[
\max_{H_0} l(\beta) = \max_{H_0} \sum_{i=1}^{n} y_i \left( \sum_{j=0}^{k} \beta_j x_{ij} \right) - \sum_{i=1}^{n} \log \left( 1 + e^{\sum_{j=0}^{k} \beta_j x_{ij}} \right)
= ny \log \left( \frac{y}{1-y} \right) + n \log(1-y).
\]

Under whole parametr space \(H_0 \cup H_1\), the MLE of the \(\beta\) can be calculated using Newton-Raphson algorithm. Let \(S(\beta) = \left( \frac{\partial l(\beta)}{\partial \beta_0}, \ldots, \frac{\partial l(\beta)}{\partial \beta_k} \right)^T\) be the score vector. \(H(\beta)\) is the Hessian matrix.

Starting with \(m = 0\) (\(\beta^{(0)}\) is an initial value of \(\beta\)), Use the following iterative equation

\[
\beta^{(m+1)} = \beta^{(m)} - H^{-1}(\beta^{(m)}) S(\beta^{(m)})
\]

till the convergence of \(\beta^{(m+1)}\).

Reference