3: Some Hypothesis Testing Problems in Genetics

§3.1. Testing hypothesis about genotype frequencies

- Testing Hardy-Weinberg Equilibrium

Consider a locus with \(k\) alleles \(A_j, j = 1, \ldots, k\). Let \(q_{ij}\) denote the genotype frequency of \(A_iA_j\) and \(p_j\) the allele frequency of \(A_j\). Let \(n_{ij}\) be the number of \(A_iA_j\) genotype. (In the existence of dominant alleles, not all \(n_{ij}\) are observable.) Need to test \(H_0\) versus \(H_1\), where

\[H_0 : q_{ij} = p_ip_j, \text{ for all } i, j = 1, \ldots, k.
\]

\[H_1 : H_0 \text{ does not hold.}
\]

**Traditional \(\chi^2\) test:**

Let \(\hat{p}_j\) be the estimate of \(p_j\). The \(\chi^2\) statistic
is given by

\[ X^2 = \sum_{u=1}^{k} \frac{(n_{uu} - n\hat{p}_u^2)^2}{n\hat{p}_u^2} + \sum_{u<v} \frac{(n_{uv} - 2n\hat{p}_u\hat{p}_v)^2}{2n\hat{p}_u\hat{p}_v}. \]

Asymptotically, \( X^2 \) has a \( \chi^2 \) distribution with d.f. \( k(k - 1)/2 \).

Note: the general form of a \( \chi^2 \) test statistic for categorical data is

\[ X^2 = \sum \frac{(\text{OBSERVED} - \text{EXPECTED})^2}{\text{EXPECTED}}. \]

The d.f. equals \( m - 1 - t \), where \( m \) is the number of categories and \( t \) is the number of parameters estimated in the expecteds.
Remark:

1. In the case of dominant alleles, the individuals are grouped by observable phenotypes, and the frequencies of the corresponding genotypes are pooled together.

2. The $\chi^2$ statistic can be inflated if the expected observations in some categories are small, which makes the test more liberal.

Likelihood ratio test (LRT)

The general LRT statistic is of the form:

$$\lambda = 2(\ln \hat{L}_1 - \ln \hat{L}_0),$$

where $\hat{L}_1$ is the maximum likelihood under $H_1$, and $\hat{L}_0$ is the maximum likelihood under $H_0$. Asymptotically, the LRT statistic has a $\chi^2$ distribution. The d.f. is the difference between the number of parameters under $H_1$ and the number of parameters under $H_0$. 

3
In the context of testing HWE, the LRT statistic is given by

\[ \lambda = 2 \sum_{i=1}^{k} n_{ii} \ln \frac{\hat{q}_{ii}}{\hat{p}_i^2} + 2 \sum_{i<j} n_{ij} \ln \frac{\hat{q}_{ij}}{2\hat{p}_i\hat{p}_j} \]

\[ \sim \chi^2_{k(k-1)/2}, \text{ asymptotically.} \]

**ABO blood group example:**

The number of ABO phenotypes of a group of duodenal ulcer patients and a group of normal people are given in the following table:

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Ulcer</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>186</td>
<td>279</td>
</tr>
<tr>
<td>B</td>
<td>38</td>
<td>69</td>
</tr>
<tr>
<td>AB</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>O</td>
<td>284</td>
<td>315</td>
</tr>
</tbody>
</table>

Considering testing HWE for the ulcer group:
\( \chi^2 \) test statistic:

\[
X^2 = \frac{[n_A - n(\hat{p}_A^2 + 2\hat{p}_A\hat{p}_O)]^2}{n(\hat{p}_A^2 + 2\hat{p}_A\hat{p}_O)} + \frac{[n_B - n(\hat{p}_B^2 + 2\hat{p}_B\hat{p}_O)]^2}{n(\hat{p}_B^2 + 2\hat{p}_B\hat{p}_O)} + \\
\frac{[n_{AB} - 2n\hat{p}_A\hat{p}_B]^2}{2n\hat{p}_A\hat{p}_B} + \frac{[n_O - n\hat{p}_O]^2}{n\hat{p}_O}
\]

\[
= \frac{(186 - 521 \times (.2136^2 + 2 \times .2136 \times .7363)^2}{521 \times (.2136^2 + 2 \times .2136 \times .7363)} + \\
\frac{[38 - 521 \times (.0501^2 + 2 \times .0501 \times .7363)]^2}{521 \times (.0501^2 + 2 \times .0501 \times .7363)} + \\
\frac{[13 - 521 \times 2 \times .2136 \times .0501]^2}{521 \times 2 \times .2136 \times .0501} + \frac{[284 - 521 \times .7363^2]^2}{521 \times .7363^2}
\]

\[
= 0.4059.
\]

\[P(\chi_1^2 \geq 0.4059) = 0.5240.\]

LRT statistic:

\[
2 \ln \frac{L(\hat{q})}{L(\hat{p})} = 2n_A \ln \frac{\hat{q}_A}{\hat{p}_A^2 + 2\hat{p}_A\hat{p}_O} + 2n_B \ln \frac{\hat{q}_B}{\hat{p}_B^2 + 2\hat{p}_B\hat{p}_O} \]

\[
= + 2n_{AB} \ln \frac{\hat{q}_{AB}}{2\hat{p}_A\hat{p}_B} + 2n_O \ln \frac{\hat{q}_O}{\hat{p}_O}
\]

\[
= .393.
\]

\[\hat{q}_A = \frac{186}{521}, \hat{q}_B = \frac{38}{521}, \hat{q}_{AB} = \frac{13}{521}, \hat{q}_O = \frac{284}{521},\]

\[\hat{p}_A = .2136, \hat{p}_B = .0501, \hat{p}_O = .7363.\]

\[P(\chi_1^2 \geq .393) = 0.5307.\]
Note that the $\chi^2$ test has a slightly smaller $p$-value.

**Exact tests**

If the sample size is relatively small, the asymptotic distribution of either the $ch_i^2$ statistic or the LRT statistic does not provide an accurate approximation to the test statistic. In the case that all alleles are codominant, exact tests can be carried out instead.

For a locus with $k$ alleles, the number of $A_i$-allele from a sample with genotype counts $n_{ij}$ is counted as

$$n_i = 2n_{ii} + \sum_{i \neq j} n_{ij}.$$ 

The exact test is based on the conditional distribution of the genotype counts given the alleles counts.
Two-allele case

Let $N_{AA}$ denote the random variable and $n_{AA}$ the observed value for the counts of AA. Similarly for other genotypes and alleles. Denote

$$P(N_{AA} = n_{AA}, N_{Aa} = n_{Aa}, N_{aa} = n_{aa}|N_A = n_A, N_a = n_a)$$
$$= P(n_{AA}, n_{Aa}, n_{aa}|n_A, n_a).$$

Adopt the similar notation for other probabilities. Under the hypothesis of HWE:

$$P(n_{AA}, n_{Aa}, n_{aa}|n_A, n_a)$$
$$= \frac{P(n_{AA}, n_{Aa}, n_{aa} \text{ and } n_A, n_a)}{P(n_A, n_a)}$$
$$= \frac{P(n_{AA}, n_{Aa}, n_{aa})}{P(n_A, n_a)}.$$
Note

\[ P(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} p_A^{2n_{AA}} (2p_A p_a)^{n_{Aa}} p_a^{2n_{aa}}. \]

\[ P(n_A, n_a) = \frac{(2n)!}{n_A!n_a!} p_A^{n_A} p_a^{n_a}. \]

\[ n_A = 2n_{AA} + n_{Aa}, \quad n_a = 2n_{aa} + n_{Aa}. \]

Hence

\[ P(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \frac{n!n_A!n_a!2^{n_{Aa}}}{n_{AA}!n_{Aa}!n_{aa}!(2n)!}. \]

Note that

\[ n_A + n_a = 2n, \quad n_{AA} + n_{Aa} + n_{aa} = n. \]
Let \( n_{Aa} = x \). Then
\[
P(n_{AA}, n_{Aa}, n_{aa}|n_A, n_a) = P(x|n_A) = \frac{n!n_A!(2n - n_A)!2^x}{[(n_A - x)/2]!x![n - (n_A + x)/2]!(2n)!}.
\]
If \( n_A \) is odd,
\[
x = 1, 3, \ldots, \min\{n_A, n_a\}
\]
Otherwise
\[
x = 0, 2, \ldots, \min\{n_A, n_a\}.
\]
Based on the conditional distribution above, two types of tests can be carried out.

1. **Probability test**: \( p \)-value is computed as
\[
\sum_x P(x|n_A)I\{P(x|n_A) \leq P(x_0|n_A)\},
\]
where \( x_0 \) is the observed value of \( n_{Aa} \).
2. **Likelihood ratio test**: $p$-value is computed as

$$\sum_{x} P(x|n_A)I\{\lambda(x) \geq \lambda(x_0)\},$$

where $\lambda(x)$ denote the LRT statistic computed with $n_{Aa} = x$, $n_{AA} = (n_A - x)/2$ and $n_{aa} = n - (n_A + x)/2$.

**An example:**
At the mosquito locus $Pgm$, the genotypes are obtained for 40 mosquitoes. The counts are given below:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>9</td>
</tr>
<tr>
<td>Aa</td>
<td>1</td>
</tr>
<tr>
<td>aa</td>
<td>30</td>
</tr>
</tbody>
</table>

$n_A = 19$, $n_a = 61$. 
### Exact test for HWE at $Pgm$ locus

<table>
<thead>
<tr>
<th>Possible samples</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
<th>probability</th>
<th>Cumulative probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 1 30*</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 3 29</td>
<td>.0000</td>
<td>.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 5 28</td>
<td>.0001</td>
<td>.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 7 27</td>
<td>.0023</td>
<td>.0024</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 9 26</td>
<td>.0205</td>
<td>.0229</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 19 21</td>
<td>.0594</td>
<td>.0803</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 11 25</td>
<td>.0970</td>
<td>.1793</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 17 22</td>
<td>.2308</td>
<td>.4101</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 13 24</td>
<td>.2488</td>
<td>.6589</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 15 23</td>
<td>.3411</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* is the observed data.

The $p$-value of the probability test is 0.0000. HWE is rejected.
Multi-allele case

Conditional distribution under HWE:

\[ P(\{n_{ij}\}|\{n_i\}) = \frac{n!2^H \prod_i n_i!}{(2n)! \prod_{i,j} n_{ij}!}, \]

where \( H = \sum_i \sum_{j>i} n_{ij}. \)

Derivation for the special case of three alleles:

\[ P(\{n_{ij}\}) = \frac{n!}{n_{11}! \cdots n_{33}!} p_1^{2n_{11}} p_2^{2n_{22}} p_3^{2n_{33}} \]
\[ \times (2p_1p_2)^{n_{12}} (2p_1p_3)^{n_{13}} (2p_2p_3)^{n_{23}} \]
\[ = \frac{n!}{n_{11}! \cdots n_{33}!} 2^{n_{12}+n_{13}+n_{23}} \]
\[ \times p_1^{2n_{11}+n_{12}+n_{13}} p_2^{2n_{22}+n_{12}+n_{23}} \]
\[ \times p_3^{2n_{33}+n_{13}+n_{23}}. \]

\[ P(\{n_i\}) = \frac{(2n)!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}. \]
\[
P(\{n_{ij}\}|\{n_i\}) = \frac{P(\{n_{ij}\})}{P(\{n_i\})} = \frac{n!}{n_{11}! \cdots n_{33}!} \frac{2^{n_{12} + n_{13} + n_{23}}}{(2n)!} = \frac{n!n_1!n_2!n_3!2^{n_{12} + n_{13} + n_{23}}}{(2n)!n_{11}!n_{12}!n_{13}!n_{22}!n_{23}!n_{33}!}.
\]

The \(p\)-value or the critical value of the exact test can be simulated from the above distribution. The following is a sampling scheme:

**Step 1:** Construct an array of \(n_1\) 1’s, \(n_2\) 2’s, \ldots, and \(n_k\) \(k\)’s.

**Step 2:** Randomly permute the array and take consecutive pairs to represent the genotypes.

**Step 3:** Count the number of pairs \((i, j)\) to obtain \(n_{ij}\).
Remark: The \( \{n_{ij}\} \) obtained above follows the conditional distribution.

The \( p \)-value or critical value can be simulated as follows:

(i) Simulate \( B \) samples using the sampling procedure.

(ii) For each of the simulated sample, compute the test statistic, say, the LRT statistic \( \lambda \).

(iii) Take the empirical distribution of the simulated test statistic values as the approximation to the null distribution of the test statistic.

- **Testing equality of allele frequencies of different groups under HWE**

  *ABO blood group example revisited:*

  The data is reproduced below:
Let $\mathbf{p}$ and $\mathbf{q}$ be the vectors of allele frequencies of the ulcer and normal group respectively. The null hypothesis to be tested is $H_0 : \mathbf{p} = \mathbf{q}$. Let $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ be the MLE of $\mathbf{p}$ and $\mathbf{q}$ under alternative hypothesis, $\hat{\mathbf{r}}$ the MLE under the null hypothesis. The LRT statistic is given by

$$
\lambda = 2 \ln \frac{L_u(\hat{\mathbf{p}}) L_n(\hat{\mathbf{q}})}{L_c(\hat{\mathbf{r}})} \\
= 2 \ln L_u(\hat{\mathbf{p}}) + 2 \ln L_n(\hat{\mathbf{q}}) - 2 \ln L_c(\hat{\mathbf{r}}).
$$

Based on the MLE

$$
\hat{\mathbf{p}} = (.2136, .0501, .7363), \\
\hat{\mathbf{q}} = (.2492, .0655, .6853),
$$
\[ \hat{r} = (0.2335, 0.0588, 0.7077). \]

The value of the statistic is computed as \( \lambda = 7.012 \) with \( p \)-value 0.0081.

- **The Z-test for codominant two-allele locus**

  Let \( D = q_{AA} - p_A^2 \). Then HWE \( \Leftrightarrow D = 0 \).

  Let \( \hat{D} = \hat{q}_{AA} - \hat{p}_A^2 \).

  \[
  Z = \frac{\hat{D}}{\sqrt{\text{Var}(\hat{D})}} \sim N(0, 1).
  \]

  A technique for computing the moments of frequencies:

  Define

  \[
  x_{li} = \begin{cases} 
  1, & \text{if } l\text{th allele of unit } i \text{ is } A, \\
  0, & \text{otherwise} \\
  l = 1, 2.
  \end{cases}
  \]
Then
\[ n_{AA} = \sum_{i=1}^{n} x_{1i}x_{2i}, \quad n_A = \sum_{i=1}^{n} (x_{1i} + x_{2i}). \]

Hence
\[
\begin{align*}
E(n_A) &= nE(x_{1i} + x_{2i}), \\
Var(n_A) &= n[Var(x_{1i}) + Var(x_{2i}) \\
& \quad + 2Cov(x_{1i}, x_{2i})], \\
E(n_{AA}) &= nE(x_{1i}x_{2i}), \\
Var(n_{AA}) &= nVar(x_{1i}x_{2i}).
\end{align*}
\]

\[
\begin{align*}
E(x_{1i}) &= E(x_{2i}) = p_A, \\
E(x_{1i}x_{2i}) &= q_{AA}, \\
Var(x_{1i}) &= p_A(1 - p_A), \\
Var(x_{1i}x_{2i}) &= q_{AA}(1 - q_{AA}).
\end{align*}
\]
Hence

\[
\begin{align*}
E(n_A) &= 2np_A, \\
\text{Var}(n_A) &= 2np_A(1 - p_A) + 2n(q_{AA} - p_A^2), \\
E(n_{AA}) &= nq_{AA}, \\
\text{Var}(n_{AA}) &= nq_{AA}(1 - q_{AA}).
\end{align*}
\]

Then \( \delta \)-method yields

\[
\begin{align*}
E(\hat{D}) &= D + O(1/n), \\
\text{Var}(\hat{D}) &= \frac{1}{n}p_A^2(1 - p_A)^2 \\
&\quad + (1 - 2p_A)D - D^2] + O(1/n^2).
\end{align*}
\]

Under HWE, \( D = 0 \), hence

\[
Z = \frac{\sqrt{n}\hat{D}}{p_A(1 - p_A)} \sim N(0, 1).
\]
§3.2. Testing Linkage Equilibrium

• Haplotype Data

Let \( n \) independent haplotypes defined on \( m \) loci be observed. Let distinguish haplotypes be denoted by

\[ i = (i_1, \ldots, i_m). \]

\( p_i \): frequency of haplotype \( i \).

\( p_{jk} \): frequency of allele \( k \) at the \( j \)th locus.

\( n_i \): counts of haplotype \( i \).

\( n_{jk} \): counts of allele \( k \) at locus \( j \).

\[ n_{jk} = \sum_{i} n_i I\{i_j = k\}. \]

\( \{n_i\} \) follow the multinomial distribution

\[ \left( \begin{array}{c} n \\ \{n_i\} \end{array} \right) \prod_{i} p_{i}^{n_i}. \]
• Likelihood ratio tests

Distribution under $H_0$ (Linkage Equilibrium).

\[
\left( \frac{n}{\{n_i\}} \right) \prod_{i} p_i^{n_i} = \left( \frac{n}{\{n_i\}} \right) \prod_{j=1}^{m} \prod_{k} p_{jk}^{n_{jk}}.
\]

**MLE and LRT statistic**
\[
\hat{p}_i = \frac{n_i}{n}, \quad \hat{p}_{jk} = \frac{n_{jk}}{n},
\]
\[
\ln L(\{\hat{p}_i\}) = \sum_{i} n_i \ln n_i - n \ln n,
\]
\[
\ln L(\{\hat{p}_{jk}\}) = \sum_{j=1}^{m} \sum_{k} n_{jk} \ln n_{jk} - mn \ln n,
\]
\[
2 \ln \frac{L(\{\hat{p}_i\})}{L(\{\hat{p}_{jk}\})} = 2 \sum_{i} n_i \ln n_i
-2 \sum_{j=1}^{m} \sum_{k} n_{jk} \ln n_{jk} - 2(m - 1)n \ln n
\]
\[
\rightarrow \chi_r^2.
\]
• Exact tests

**Fisher-Yates Distribution**

\[
P(\{n_i\}|\{n_{jk}\}) = \frac{\binom{n}{\{n_i\}} \prod_i p_i^{n_i}}{\prod_j \binom{n}{\{n_{jk}\}} \prod_k p_{jk}^{n_{jk}}} \\
= \frac{\binom{n}{\{n_i\}}}{\prod_j \binom{n}{\{n_{jk}\}}}.
\]

**Sampling from Fisher-Yates distribution.**

**Step 1.** Arrange the \(mn\) alleles in the \(n\) observed haplotypes into an \(m \times n\) matrix whose rows correspond to loci and whose columns correspond to haplotypes. Then randomly permute each row of the matrix.

**Step 2.** In the permuted matrix, identify the number of columns with \(i = (i_1, \ldots, i_m)^T\) to obtain \(n_{i^*}\).

The resultant \(\{n_{i^*}\}\) follow the Fisher-Yates distribution.
**Simulated Exact test.**

The exact test can be based on any reasonable test statistic. The test can be simulated as follows: (i) Using the sampling scheme above to simulate a large number of \( \{ n_{i}^{\ast} \} \). (ii) For each \( \{ n_{i}^{\ast} \} \), compute the value of the test statistic. (iii) The \( p \)-value of the test is approximated by the proportion of the simulated test statistic values that exceed the observed test statistic value.

The test statistic can be taken as the LRT statistic, or the negative of the Fisher-Yates probability.

§3.3. Other multinomial problems in genetics

- \( Z_{\text{max}} \) tests
To test the alternative that there might be clustering in some categories in the multinomial distribution, the $Z_{\text{max}}$-test is more appropriate.

**Test statistic**

$$Z_{\text{max}} = \max_{1 \leq i \leq m} \frac{N_i - np_i}{\sqrt{np_i(1 - p_i)}},$$

where $N_i$ is the counts and $p_i$ is the probability in category $i$.

**p-value:**

Let $z_{\text{max}}$ be the observed value of $Z_{\text{max}}$. The $p$-value can be approximated by

$$m[1 - \Phi(z_{\text{max}})] - \frac{m(m - 1)}{2}[1 - \Phi(z_{\text{max}})]^2 \leq P(Z_{\text{max}} \geq z_{\text{max}}) \leq m[1 - \Phi(z_{\text{max}})],$$

where $\Phi$ is the CDF of $N(0, 1)$.  

23
**Example:** In Situ Hybridization.

\[ Z_{\text{max}} \text{-test for the ZYF Probe} \]

in Macropus eugenii

<table>
<thead>
<tr>
<th>Segment</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>( Z_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p</td>
<td>0.042</td>
<td>24</td>
<td>3.666</td>
</tr>
<tr>
<td>1q</td>
<td>0.189</td>
<td>37</td>
<td>-2.406</td>
</tr>
<tr>
<td>2p</td>
<td>0.019</td>
<td>4</td>
<td>-0.571</td>
</tr>
<tr>
<td>2q</td>
<td>0.136</td>
<td>25</td>
<td>-2.261</td>
</tr>
<tr>
<td>3/4p</td>
<td>0.104</td>
<td>35</td>
<td>1.174</td>
</tr>
<tr>
<td>3/4q</td>
<td>0.178</td>
<td>44</td>
<td>-0.886</td>
</tr>
<tr>
<td>5p</td>
<td>0.031</td>
<td>29</td>
<td>7.030</td>
</tr>
<tr>
<td>5q</td>
<td>0.097</td>
<td>28</td>
<td>0.190</td>
</tr>
<tr>
<td>6p</td>
<td>0.048</td>
<td>11</td>
<td>-0.670</td>
</tr>
<tr>
<td>6q</td>
<td>0.062</td>
<td>11</td>
<td>-1.564</td>
</tr>
<tr>
<td>7</td>
<td>0.053</td>
<td>19</td>
<td>1.126</td>
</tr>
<tr>
<td>Xp</td>
<td>0.011</td>
<td>4</td>
<td>0.534</td>
</tr>
<tr>
<td>Xq</td>
<td>0.018</td>
<td>3</td>
<td>-0.911</td>
</tr>
<tr>
<td>Y</td>
<td>0.012</td>
<td>5</td>
<td>0.908</td>
</tr>
</tbody>
</table>
$H_0$: No presence of ZYF at any of the 14 regions.

$z_{\text{max}} = 7.030$.

$P(Z_{\text{max}} \geq 7.03) \approx 0$.

The $p$-value confirms the presence of ZYF at the p arm of chromosome 5.

Repeated conditional $Z_{\text{max}}$-test.

Once a category is claimed significant, the test is carried out conditioning on the remaining categories. The procedure repeats until no category is tested significant.
• **$W_d$ tests**

*Test statistic*

$$W_d = \sum_{i=1}^{m} I\{N_i \geq d\}.$$  

*Recursive formula for the computation of $p$-values*

$t_{j,k,l}$: the probability that $W_d \leq j$ given $k$ trials and $l$ categories.

The $l$ given categories refers to the first $l$ of the overall $m$ categories, and the $i$th of these $l$ categories is assigned the conditional probability $p_i/(p_1 + \cdots + p_l)$.

Initial probabilities:

$$t_{0,k,1} = \begin{cases} 
1, & \text{if } k < d, \\
0, & \text{if } k \geq d.
\end{cases}$$

$$t_{j,k,1} = 1, \text{ for } j > 0.$$
Recursive formulae:

\[ t_{0,k,l} = \min\{d-1,k\} \sum_{i=0}^{\min\{d-1,k\}} \binom{k}{i} \left[ \frac{p_l}{p_1 + \cdots + p_l} \right]^i \left[ 1 - \frac{p_l}{p_1 + \cdots + p_l} \right]^{k-i} \times t_{0,k-i,l-1}, \]

\[ t_{j,k,l} = \min\{d-1,k\} \sum_{i=0}^{\min\{d-1,k\}} \binom{k}{i} \left[ \frac{p_l}{p_1 + \cdots + p_l} \right]^i \left[ 1 - \frac{p_l}{p_1 + \cdots + p_l} \right]^{k-i} \times t_{j,k-i,l-1} \]

\[ + \sum_{i=d}^{k} \binom{k}{i} \left[ \frac{p_l}{p_1 + \cdots + p_l} \right]^i \left[ 1 - \frac{p_l}{p_1 + \cdots + p_l} \right]^{k-i} \times t_{j-1,k-i,l-1}. \]

**Example:**

66 Mutations have been observed in 141 amino acids of the homoglobin α chain. Of these 141 amino acids, 16 show two or more mutations. \(W_d\)-test is used to test whether or not the mutations occur randomly. The \(p\)-value of the
test with \( d = 2 \) is 0.028, which suggests non-randomness.

§3.4. Testing for detecting disease genes

• **Case-control association tests**

  **Goal**

  To test whether or not a disease is associated with a locus.

  **Setting of the problem**

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ( \cdots ) j ( \cdots ) m</td>
<td>( c_1. ) ( c_2. ) ( n )</td>
</tr>
<tr>
<td>Control</td>
<td>( c_{11} \cdots c_{1j} \cdots c_{1m} )</td>
</tr>
<tr>
<td>Case</td>
<td>( c_{21} \cdots c_{2j} \cdots c_{2m} )</td>
</tr>
<tr>
<td>Subtotal</td>
<td>( c_{.1} \cdots c_{.j} \cdots c_{.m} )</td>
</tr>
</tbody>
</table>
Under the null hypothesis of no association between disease status and genotypes, $E(c_{1j}) = E(c_{2j})$.

**Test Statistic**

$$Z_{\max} = \max_{i,j} Z_{ij} = \max_{j} |Z_{1j}|,$$

where

$$Z_{ij} = \frac{c_{ij} - E(c_{ij})}{\sqrt{\text{Var}(c_{ij})}},$$

$$E(c_{ij}) = \frac{c_{1j}\cdot}{n},$$

$$\text{Var}(c_{ij}) = \frac{c_{1}(c_{1} - 1)c_{j}(c_{j} - 1)}{n(n - 1)}$$

$$+ E(c_{1j}) - [E(c_{1j})]^2.$$

**Note**

$$c_{1j} - \frac{c_{1j}\cdot}{n} = \frac{(c_{1.} + c_{2.})c_{1j} - c_{1j}\cdot}{n}$$

$$= \frac{c_{2}(c_{1j} + c_{2j}) - (c_{1.} + c_{2.})c_{2j}}{n}$$

$$= -(c_{2j} - E(c_{2j})).$$
The conditional moments of $c_{1j}$ given the marginals

Let $x_i = 1$ if the $i$th person is from the control group and has genotype $j$, $x_i = 0$ otherwise. Then

$$c_{1j} = \sum_{i=1}^{n} x_i.$$  

$$E(c_{1j}) = E[\sum_{i=1}^{n} x_i] = nE(x_i) = \frac{c_1 c_j}{n}.$$  

$$E([c_{1j}]^2) = E \left[ \sum_{i=1}^{n} \sum_{k=1}^{n} x_ix_k \right] = E[\sum_{i=1}^{n} x_i + \sum_{i \neq k} x_ix_k] = E(c_{1j}) + n(n-1)E(x_ix_k) = E(c_{1j}) + n(n-1)\frac{c_1(c_1 - 1)c_j(c_j - 1)}{[n(n-1)]^2}.$$  

Simulating p-value by permutation

Randomly permute all the individuals, and
randomly divide the permuted individuals into $c_1$. controls and $c_2$. cases. Then compute the test statistic (LRT or $Z_{ij}$) with the permuted data. The proportion of the computed values which exceed the observed value of the test statistic is taken as the approximation of the $p$-value.

Remark: The same procedure can be considered for allele frequencies if the allele frequencies can be counted.
**ABO example (revisited)**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>O</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>279</td>
<td>69</td>
<td>17</td>
<td>315</td>
<td>680</td>
</tr>
<tr>
<td>Ulcer</td>
<td>186</td>
<td>38</td>
<td>13</td>
<td>284</td>
<td>521</td>
</tr>
<tr>
<td>Subtotal</td>
<td>465</td>
<td>107</td>
<td>30</td>
<td>599</td>
<td>1201</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|cccc|c}
 c_{1j} & 279 & 69 & 17 & 315 \\
 E(c_{1j}) & 263.281 & 60.583 & 16.986 & 339.151 \\
 \text{Var}(c_{1j}) & 70.050 & 23.960 & 7.190 & 73.808 \\
 Z_{1j} & 1.878 & 1.720 & 0.005 & -2.811 \\
\end{array}
\]

\[Z_{\text{max}} = 2.811.\]

With \( B = 10,000 \), approximate \( p \)-value:

\[Z_{\text{max}}: 0.0169 \pm 0.0026,\]

Fisher’s Exact: \( 0.0335 \pm 0.0036,\)

LRT: \( 0.0295.\)
• Transmission/disequilibrium tests (TDT)

*Rationale of TDT*

Significant association detected by Association test might be caused by other reasons rather than linkage.

TDT circumvents misleading association by exploring internal controls provided by parents.

The rationale of TDT is that if the disease is not linked with the locus, the alleles of the parents at the locus must segregate according to the Mendelian law.

*Data for TDT*

Raw data: Genotypes of the parents of affected children and the affecteds themselves.
From the raw data, the alleles of heterozygous parents which are transmitted and not transmitted are summarized as

<table>
<thead>
<tr>
<th>Transmission Pattern</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmitted</td>
<td>$c_{11} \ c_{12} \ \cdots \ c_{1m}$</td>
</tr>
<tr>
<td>Not Transmitted</td>
<td>$c_{21} \ c_{22} \ \cdots \ c_{2m}$</td>
</tr>
<tr>
<td></td>
<td>$h_{1} \ h_{2} \ \cdots \ h_{m}$</td>
</tr>
</tbody>
</table>

Example: Raw data

\{(Aa, AA), Aa\}, \{(Aa, Aa), Aa\},
\{(Aa, Aa), AA\}, \{(Aa, aa), Aa\}

<table>
<thead>
<tr>
<th>Transmission Pattern</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmitted</td>
<td>A \ a</td>
</tr>
<tr>
<td>Not Transmitted</td>
<td>4 \ 2</td>
</tr>
</tbody>
</table>

34
Test Statistic

\[ Z_{ij} = \frac{c_{ij} - \frac{h_j}{2}}{\sqrt{\frac{h_j}{4}}} \cdot \]

\[ Z_{\text{max}} = \max_{ij} Z_{ij}. \]

\[ \chi^2 = \sum_{ij} Z_{ij}^2. \]

Note: the \( \chi^2 \) is not the usual \( \chi^2 \)-statistic.

Simulating the distribution of test statistics

Simulate the genotypes of the affecteds according to random segregation from their parents. For resultant new table, compute the test statistic. Repeat this procedure for a large
$B$ times. The empirical distribution of the simulated values of the test statistic provides the approximation to the null distribution of the statistic.

**Example:** Ataxia-telangiectasia (AT) in Costa Rica.

<table>
<thead>
<tr>
<th>Transmission Pattern</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>11</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmitted</td>
<td>3</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Not Transmitted</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

$\chi^2 = 92.91$, $Z_{\text{max}} = 4.69$. 