Allele Frequencies, Genotype Frequencies, and Hardy-Weinberg Equilibrium

Establishing the genetics of the ABO blood group system was one of the first breakthroughs in Mendelian genetics. The locus corresponding to the ABO blood group has three alleles, A, B and O and is located on chromosome 9q34. The alleles A and B are dominant to O. This leads to the following genotypes and phenotypes (blood types):

<table>
<thead>
<tr>
<th>Genotype(s)</th>
<th>Blood Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO, AA</td>
<td>A</td>
</tr>
<tr>
<td>BO, BB</td>
<td>B</td>
</tr>
<tr>
<td>AB</td>
<td>AB</td>
</tr>
<tr>
<td>OO</td>
<td>O</td>
</tr>
</tbody>
</table>

Mendel’s first law allows us to quantify the types of gametes an individual can produce. For example, an individual with type AB produces gametes A and B with equal probability (1/2).

Exercise 1.1(estimating genotype frequencies from allele frequencies): From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be Pr(A)=0.2877, Pr(B)=0.1065 and Pr(O)=0.6057. (Vogel and Motulsky). If an individual has blood type B, what gametes can be produced and with what frequency? What assumptions are made in your calculation?
Exercise 1.2 (estimating allele frequencies from genotype frequencies): Cystic Fibrosis is caused by a recessive allele. The locus for the allele is in region 7q31. Of 10,000 Caucasian births, 5 were found to have Cystic Fibrosis and 442 were found to be heterozygous carriers of the disease causing mutation. Denoting Cf to be the Cystic Fibrosis gene and N to be the normal gene, estimate the allele frequencies. What assumptions are made in your calculation?
Hardy-Weinberg Equilibrium

When a population is in Hardy-Weinberg Equilibrium (HWE), the alleles that compose a genotype can be thought of as having been chosen at random from the alleles in a population. That is, for a locus with alleles A and B, the locus is in Hardy-Weinberg Equilibrium if

\[
\begin{align*}
P(AA) &= p_A^2 \\
P(AB) &= 2p_A(1-p_A) \\
P(BB) &= (1-p_A)^2
\end{align*}
\]

where \( p_A = \text{Pr}(A) \) is the allele frequency.

Example 1.2: Consider a diallelic locus with alleles A and B having frequencies 0.85 and 0.15. If the locus is in HWE, what are the expected genotype frequencies?

\[
\begin{align*}
\text{Pr}(AA) &= \\
\text{Pr}(AB) &= \\
\text{Pr}(BB) &=
\end{align*}
\]

Now suppose 50 individuals are sampled and genotyped. The data are \( n_{AA}=43 \), \( n_{AB}=5 \) and \( n_{BB}=2 \). Using these data, estimate the genotype frequencies.

\[
\begin{align*}
P_{AA} &= \\
P_{AB} &= \\
P_{BB} &=
\end{align*}
\]

The assumption that a population is in HWE is often used in genetics. Thus it is important to be able to test this assumption. One method for testing for a departure from HWE is to use the likelihood ratio test statistic:

\[
LRTS = -2\ln\left[ \frac{n^{n_A} (n_A)^{n_{AA}} (n_B)^{n_{BB}} 2^{n_{AB}}}{(2n)^{2n} (n_{AA})^{n_{AA}} (n_{AB})^{n_{AB}} (n_{BB})^{n_{BB}}} \right]
\]

Under the null hypothesis of HWE, the LRTS is asymptotically distributed as a \( \chi^2 \) random variable with 1 degree of freedom. Here, \( n \) is the total number of individuals and \( n_X \) is the count of the number of alleles or genotypes depending on what the subscript \( X \) represents.

Q: What does “asymptotically distributed” mean?
Where does this test statistic come from and why does it make sense?

The likelihood of the data if Hardy-Weinberg Equilibrium is not assumed is:

\[
L(p_{AA}, p_{AB}) = \frac{N!}{N_{AA}!N_{AB}!N_{BB}!} p_{AA}^{N_{AA}} p_{AB}^{N_{AB}} p_{BB}^{N_{BB}}
\]

On the other hand, if Hardy-Weinberg Equilibrium holds, then we have a relationship between the genotype frequencies and the allele frequencies. The likelihood of the data is then:

\[
L(p_A) = \frac{N!}{N_{AA}!N_{AB}!N_{BB}!} (p_A^2)^{N_{AA}} (2p_A(1-p_A))^{N_{AB}} ((1-p_A)^2)^{N_{BB}}
\]

The likelihood ratio test statistic comes from taking the log of the ratio of two likelihoods.

Null Hypothesis $H_0$: the population is in HWE so that $P(AA) = P(A)^2$ and $P(AB) = 2P(A)P(B)$ and $P(BB) = (1-P(A))^2$.

Alternative Hypothesis $H_1$: the population is not in HWE so that $P(AA) \neq P(A)^2$ and/or $P(AB) \neq 2P(A)P(B)$ and/or $P(BB) \neq (1-P(A))^2$.

The degrees of freedom of the test is the number of free parameters under the alternative minus the number of free parameters under the null. In this case:

\[
df = ...
\]
An Aside: Maximum Likelihood Estimates

A maximum likelihood estimate of a parameter $\theta$ is the estimate of $\theta$ that maximizes the likelihood function. This provides an estimate of $\theta$ that “best explains” the observed data in some sense.

For our example, we want to find the $p_{AA}$ and $p_{BB}$ that maximizes $L(p_{AA}, p_{AB})$. Sometimes it is possible to determine this through a closed-form solution i.e. an explicit formula. In fact, when testing HWE for a diallelic locus, the maximum likelihood estimates for $p_{AA}$ and $p_{AB}$ under the alternative hypothesis are just the genotype frequencies. Under the null hypothesis, the maximum likelihood estimate of $p_A$ is the allele frequency. For other situations, maximum likelihood estimates are not so easy to calculate. It may be necessary to mathematically traverse the likelihood space in order to find the peak.

The likelihood ratio test statistic consists of maximizing the likelihood under the null hypothesis, maximizing the likelihood under the alternative hypothesis and forming the statistic:

$$LRTS = -2 \ln \left( \frac{\text{maximum likelihood under the null}}{\text{maximum likelihood under the alternative}} \right)$$

If the null hypothesis is nested within the alternative hypothesis, then given the null hypothesis is true the LRTS has an asymptotic $\chi^2$ distribution with degrees of freedom equal to the difference in the number of parameters for the two hypotheses.
Is our null hypothesis nested within the alternative hypothesis?
Here, the null and alternative hypotheses are defined:

Ho: HWE or \([P_{AA} = p_A^2, P_{AB} = 2p_A(1-p_A) , P_{BB} = (1-p_A)^2]\]

Ha: HWD or \([P_{AA} \neq p_A^2, P_{AB} \neq 2p_A(1-p_A), P_{BB} \neq (1-p_A)^2]\]

d.f. = # free parameters under Ha - # free parameters under Ho

Another option is the use of a goodness-of-fit chi-square test:

\[
X^2 = \sum_{\text{genotypes}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}
\]

Here, the expected counts are calculated based on the allele frequency estimate and the Hardy-Weinberg proportions. The sampling distribution of the test statistic can be approximated with a \(\chi^2\) distribution with 1 degree of freedom.
Hardy-Weinberg Equilibrium – Why must we be skeptical of this assumption?

1. Small population sizes. Chance events can make a big difference.
2. Deviations from random mating.
   a. Assortive mating. Mating between genotypically similar individuals increases homozygosity for the loci involved in mate choice without altering allele frequencies.
   b. Disassortive mating. Mating between dissimilar individuals increases heterozygosity without altering allele frequencies.
   c. Inbreeding. Mating between close relatives increases homozygosity for the whole genome without affecting allele frequencies.
3. Mutation.
4. Migration.
5. Selection.