Hardy-Weinberg Equilibrium
How do allele frequencies relate to genotype frequencies in a population?
If we have genotype frequencies, we can easily get allele 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 mutation that causes the disease. Denote the Cystic Fibrosis allele with $cf$ and the normal allele with $N$. Based on this sample, how can we estimate the allele frequencies in the population?

- We can estimate the genotype frequencies in the population based on this sample.
- \(\frac{5}{10000}\) are $cf$, $cf$
- \(\frac{442}{10000}\) are $N$, $cf$
- \(\frac{9553}{10000}\) are $N$, $N$
So we use 0.0005, 0.0442, and 0.9553 as our estimates of the genotype frequencies in the population. The only assumption we have used is that the sample is a random sample. Starting with these genotype frequencies, we can estimate the allele frequencies without making any further assumptions: Out of 20,000 alleles in the sample

- \( \frac{442 + 10}{20000} = .0226 \) are \( cf \)
- \( 1 - \frac{442 + 10}{20000} = .9774 \) are \( N \)
In contrast, going from allele frequencies to genotype frequencies requires more assumptions.

**HWE Model Assumptions**

- infinite population
- discrete generations
- random mating
- no selection
- no migration in or out of population
- no mutation
- equal initial genotype frequencies in the two sexes
Consider a locus with two alleles: \( A \) and \( a \). Assume in the first generation the alleles are not in HWE and the genotype frequency distribution is as follows:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>( u )</td>
</tr>
<tr>
<td>Aa</td>
<td>( v )</td>
</tr>
<tr>
<td>aa</td>
<td>( w )</td>
</tr>
</tbody>
</table>

where \( u + v + w = 1 \).

From the genotype frequencies, we can easily obtain allele frequencies:

\[
P(A) = \frac{u + 1}{2v} \quad P(a) = \frac{w + 1}{2v}
\]
Hardy-Weinberg Equilibrium

In the first generation: \( P(A) = u + \frac{1}{2}v \) and \( P(a) = w + \frac{1}{2}v \)

### 2nd Generation

<table>
<thead>
<tr>
<th>Mating Type</th>
<th>Mating Frequency</th>
<th>Expected Progeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>( AA \times AA )</td>
<td>( u^2 )</td>
<td>( AA )</td>
</tr>
<tr>
<td>( AA \times Aa )</td>
<td>( 2uv )</td>
<td>( \frac{1}{2} AA : \frac{1}{2} Aa )</td>
</tr>
<tr>
<td>( AA \times aa )</td>
<td>( 2uw )</td>
<td>( \frac{1}{4} AA : \frac{1}{2} Aa : \frac{1}{4} aa )</td>
</tr>
<tr>
<td>( Aa \times Aa )</td>
<td>( v^2 )</td>
<td>( \frac{1}{2} Aa : \frac{1}{2} aa )</td>
</tr>
<tr>
<td>( Aa \times aa )</td>
<td>( 2vw )</td>
<td>( aa )</td>
</tr>
<tr>
<td>( aa \times aa )</td>
<td>( w^2 )</td>
<td></td>
</tr>
</tbody>
</table>

*Check:* \( u^2 + 2uv + 2uw + v^2 + 2vw + w^2 = (u + v + w)^2 = 1 \)

- \( p \equiv P(AA) = u^2 + \frac{1}{2}(2uv) + \frac{1}{4}v^2 = (u + \frac{1}{2}v)^2 \)
- \( q \equiv P(Aa) = uv + 2uw + \frac{1}{2}v^2 + vw = 2 \left( u + \frac{1}{2}v \right) \left( \frac{1}{2}v + w \right) \)
- \( r \equiv P(aa) = \frac{1}{4}v^2 + \frac{1}{2}(2vw) + w^2 = \left( w + \frac{1}{2}v \right)^2 \)
Hardy-Weinberg Equilibrium

In the third generation:

\[ P(AA) = (p + \frac{1}{2}q)^2 = \left( \left( u + \frac{1}{2}v \right)^2 + 2 \left( u + \frac{1}{2}v \right) \left( \frac{1}{2}v + w \right) \right)^2 \]

\[ = \left( u + \frac{1}{2}v \right) \left[ \left( u + \frac{1}{2}v \right) + \left( \frac{1}{2}v + w \right) \right] \]

\[ = \left( u + \frac{1}{2}v \right) \left( u + v + w \right) \]

\[ = \left( u + \frac{1}{2}v \right) = \left( u + \frac{1}{2}v \right)^2 = p \]

Similarly, \( P(Aa) = q \) and \( P(aa) = r \) for generation 3.

• **Equilibrium** is reached after one generation of mating under the Hardy-Weinberg assumptions! Genotype frequencies remain the same from generation to generation.
When a population is in Hardy-Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population. We have the following relationship between genotype frequencies and allele frequencies for a population in Hardy-Weinberg equilibrium:

\[
P(AA) = P(A)P(A) \\
P(Aa) = 2P(A)P(a) \\
P(aa) = P(a)P(a)
\]
For example, consider a diallelic locus with alleles A and B with frequencies 0.85 and 0.15, respectively. If the locus is in HWE, then the genotype frequencies are:

\[
P(AA) = 0.85 \times 0.85 = 0.7225
\]

\[
P(AB) = 0.85 \times 0.15 + 0.15 \times 0.85 = 0.2550
\]

\[
P(BB) = 0.15 \times 0.15 = 0.0225
\]
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. Alleles A and B are co-dominant, and the alleles A and B are dominant to O. This leads to the following genotypes and phenotypes:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Blood Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA, AO</td>
<td>A</td>
</tr>
<tr>
<td>BB, BO</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).
From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be $P(A)=0.2877$, $P(B)=0.1065$ and $P(O)=0.6057$. If an individual has blood type B, what are the possible genotypes for this individual, what possible gametes can be produced, and what is the frequency of the genotypes and gametes if HWE is assumed?

- If a person has blood type B, then the genotype is BO or BB.
- What is $P(\text{genotype is BO}|\text{blood type is B})$?
- What is $P(\text{genotype is BB}|\text{blood type is B})$?
- What is $P(\text{B gamete}|\text{blood type is B})$?
- What is $P(\text{O gamete}|\text{blood type is B})$?
With HWE: allele frequencies $\Rightarrow$ genotype frequencies.

Violations of HWE assumption include:

- Small population sizes. Chance events can make a big difference.
- Deviations from random mating.
- Assortive mating. Mating between genotypically similar individuals increases homozygosity for the loci involved in mate choice without altering allele frequencies.
- Disassortive mating. Mating between dissimilar individuals increases heterozygosity without altering allele frequencies.
- Inbreeding. Mating between relatives increases homozygosity for the whole genome without affecting allele frequencies.

- Population sub-structure
- Mutation
- Migration
- Selection
When a locus is not in HWE, then this suggests one or more of the Hardy-Weinberg assumptions is false.

Departure from HWE has been used to infer the existence of natural selection, argue for existence of assortive (non-random) mating, and infer genotyping errors.

It is therefore of interest to test whether a population is in HWE at a locus.

We will discuss the two most popular ways of testing HWE:
- Chi-Square test
- Exact test
Chi-Square Goodness-Of-Fit Test

Compares observed genotype counts with the values expected under Hardy-Weinberg. For a locus with two alleles, we might construct a table as follows:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Observed</th>
<th>Expected under HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$n_{AA}$</td>
<td>$np_A^2$</td>
</tr>
<tr>
<td>Aa</td>
<td>$n_{Aa}$</td>
<td>$2np_A(1 - p_A)$</td>
</tr>
<tr>
<td>aa</td>
<td>$n_{aa}$</td>
<td>$n(1 - p_A)^2$</td>
</tr>
</tbody>
</table>

where $n$ is the number of individuals in the sample and $p_A$ is the probability that a random allele in the population is of type A.

- We estimate $p_A$ with $\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n}$
Chi-Square Goodness-Of-Fit Test

Test statistic is for Allelic Association is:

\[ X^2 = \sum \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}} \text{ genotypes} \]

\[ X^2 = \frac{(n_{AA} - n\hat{p}_a^2)^2}{n\hat{p}_a^2} + \frac{(n_{Aa} - 2n\hat{p}_a(1 - \hat{p}_a))^2}{2n\hat{p}_a(1 - \hat{p}_a)} + \frac{(n_{aa} - n(1 - \hat{p}_a)^2)^2}{n(1 - \hat{p}_a)^2} \]

- Under \( H_0 \), the \( X^2 \) test statistic has an approximate \( \chi^2 \) distribution with 1 degree of freedom
- Recall the rule of thumb for such \( \chi^2 \) tests: the expected count should be at least 5 in every cell. If allele frequencies are low, and/or sample size is small, and/or there are many alleles at a locus, this may be a problem.
The Hardy-Weinberg exact test is based on calculating probabilities

\[ P(\text{genotype counts} \mid \text{allele counts}) \] under HWE.
Suppose we have a sample of 5 people and we observe genotypes AA, AA, AA, aa, and aa.

If five individuals have among them 6 A alleles and 4 a alleles, what genotype configurations are possible?
## HWE Exact Test Example

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>theoretical probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0.048</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0.571</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0.381</td>
</tr>
</tbody>
</table>

**Hardy-Weinberg Equilibrium**
Now suppose we have a sample of 100 individuals and we observe 21 "a" alleles and 179 "A" alleles, what genotype configurations are possible?
Note that specifying the number of heterozygotes determines the number of AA and aa genotypes.

<table>
<thead>
<tr>
<th>aa</th>
<th>Aa</th>
<th>AA</th>
<th>theoretical probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>≪ .000001</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>≪ .000001</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>&lt; .000001</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>.000001</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>.000047</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>.000870</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>.009375</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td>.059283</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td>.214465</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td>.406355</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td>.309604</td>
</tr>
</tbody>
</table>

Wiggington, Cutler, Abecasis (AJHG, 2005)
The formula is:

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

If we had actually observed 13 heterozygotes in our sample, then the exact test p-value would be

\[
\approx 0.009375 + 0.000870 + 0.000047 + 0.000001 = 0.010293
\]

(To get the p-value, we sum the probabilities of all configurations with probability equal to or less that the observed configuration.)
The next slide is Figure 1 from Wigginton et al (AJHG 2005). The upper curves give the type I error rate of the chi-square test; the bottom curves give the type I error rate from the exact test. The exact test is always conservative; the chi-square test can be either conservative or anti-conservative.
HWE TYPE I ERROR

A. Sample size = 100, $\alpha = 0.05$

B. Sample size = 100, $\alpha = 0.01$

C. Sample size = 100, $\alpha = 0.001$

D. Sample size = 1,000, $\alpha = 0.05$

E. Sample size = 1,000, $\alpha = 0.01$

F. Sample size = 1,000, $\alpha = 0.001$
The Exact Test should be preferred for smaller sample sizes and/or multiallelic loci, since the $\chi^2$ test is not valid in these cases (rule of thumb: must expect at least 5 in each cell).

The coarseness of Exact Test means it is conservative. In Example 4, we reject the null hypothesis that HWE holds if 13 or fewer heterozygotes are observed. But the observed p-value is actually 0.010293. Thus to reject at the 0.05 level, we actual have to see a p-value as small as 0.010293.
The $\chi^2$ test can have inflated type I error rates. Suppose we have 100 genes for which HWE holds. We conduct 100 $\chi^2$ tests at level 0.05. We expect to reject the null hypothesis that HWE holds in 5 of the tests. However, the results of Wiggington et al (AJHG, 2005) say, on average, it can be more than 5 depending on the minor allele count. Although it is not desirable for a test to be conservative (Exact Test), an anti-conservative test is considered unacceptable.

Wiggington et al (AJHG, 2005) give an extreme example with a sample of 1000 individuals. At a nominal $a=0.001$, the true type I error rate for the $\chi^2$ test exceeds 0.06.
The $\chi^2$ test is a two-sided test. In contrast, the Exact Test can be made one-sided, if appropriate. Specifically, one can test for a deficit of heterozygotes (if one suspects inbreeding or population stratification); test for an excess of heterozygotes (which indicate genotyping errors for some genotyping technologies).

Exact test is more computationally intensive