# BIOB480/BIOE548 notes 9/03/2024

**Corbin et al. 1974**

- see `02_slides.pdf` for background and group questions

**Quantifying genetic diversity**

Alleic variation can be used to measure the genetic diversity in a sample of individuals:

**Proportion of polymorphic loci:**

$$P = \frac{p}{n} = \frac{\text{no. polymorphic loci}}{\text{no. loci}}$$

**Average heterozygosity:**

$$H = \frac{1}{n}\sum_{i=1}^{n} H_i$$

where $H_i$ is the average heterozygosity (=proportion of individuals that are heterozygous) at locus $i$ and $n$ is the total number of loci. (This is just a fancy way to write an arithmetic mean.)

**Allelic diversity:**

$$A = \frac{1}{n}\sum_{i=1}^{n} A_i$$

where $A_i$ is the number of alleles at locus $i$ and $n$ is the total number of loci.

**Hardy-Weinberg Principle (or equilibirum, or proportions. . . )**

Given a diallelic locus where allele $A_1$ occurs at frequency $f(A_1) = p$ and allele $A_2$ occurs at frequency $f(A_2) = q$, we expect the following genotype frequencies following one generation of random mating:

$$p * p + p * q + q * p + q * q = p^2 + 2pq + q^2$$

In addition to random mating, HWP assumes a complete absence of the four evolutionary mechanisms capable of changing allele frequencies from generation to generation:

1) No natural selection;
2) No mutation;
3) No migration;
4) No genetic drift (infinite population sizes).

For example, if $f(A_1) = 0.8$ and $f(A_2) = 0.2$, we expect genotype frequencies of $f(A_1A_1) = 0.8 * 0.8 = 0.64$, $f(A_1A_2) = 0.8 * 0.2 + 0.2 * 0.8 = 2 * 0.8 * 0.2 = 0.32$, and $f(A_2A_2) = 0.2 * 0.2 = 0.04$.

Hardy-Weinberg Proportions are important because they are *a null model for evolution*—what we expect genotype frequencies to be in the absence of evolutionary mechanisms.

We can test whether observed deviations from Hardy-Weinberg Proportions are statistically significant with a Chi-Squared test:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The statistic measures the squared difference between the number of individuals actually observed with each possible genotype and the expected number of individuals with that genotype based on Hardy-Weinberg proportions and assumptions, divded by the expected number of genotypes. For a diallelic locus, the sum indicates you repeat the operation on the right-hand side three times (for $A_1A_1$, $A_1A_2$, and $A_2A_2$). This statistic is then compared to a table where statistical signficance is inferred given a particular value for the degrees of freedom (number of free parameters). Importantly, it only works with *count* data—not frequences!

For example, if we imagine 100 chicks are born with genotype counts of $f(A_1A_1) = 20$, $f(A_1A_2) = 20$, and $f(A_2A_2) = 60$, we first determine $f(A_1) = \frac{2*20+20}{200} = 0.3$ and $f(A_2) = 1 - f(A_1) = 0.7$. Based on these values, we would expect $\#A_1A_1 = 100 * p^2 = 100 * 0.3^2 = 9$, $\#A_1A_2 = 100 * 2pq = 100 * 2 * 0.3 * 0.7 = 42$, and $\#A_2A_2 = 100 * q^2 = 100 * 0.7^2 = 49$

Our Chi-Squared statistic is then:

$$\chi^2 = \frac{(20-9)^2}{9} + \frac{(20-42)^2}{42} + \frac{(60-49)^2}{49} = 24.92$$

The value of 24.92 is our "test statistic". Since we are working with a diallelic locus where $p + q = 1$, we only have a single degree of freedom—the value $p$ depends on $q$, and vice versa. We thus look at the row $df = 1$ in a table like this one and find our value. 24.92 is much greater than the test statistic value of 3.841 required to reach statistical significance at $p = 0.05$, so we can conclude the differences between the counts of observed and expected genotypes are unlikely to be due to random sampling error.

(More on the chi-squared distribution here-in the example above, we are looking at where on the line for $k = df = 1$ our statistic falls, which is far past the right-hand side of the plot, meaning the vast majority of the probability distribution is weighted towards less extreme differences.)

The idea of expected heterozygosity under Hardy-Weinberg proportions is an important one. We can more broadly define $H_e$ for $n$ loci as:

$$H_e = 1 - \sum_{i=1}^{n} p_i^2$$

In other words, the expected frequency of heterozygotes is what you have left over (i.e. the *complement*) after accounting for the expected frequency of all homozygotes. In a diallelic system, this is $1 - p^2 - q^2 (= 2pq)$; in a triallelic system, this is $1 - p^2 - q^2 - r^2 (= 2pq + 2pr + 2qr)$, etc.

Dan Bolnick has a useful app for visualizing Hardy-Weinberg proportions: https://bolnicklab.shinyapps.io/HWEdemo/