1 Introduction

This lecture explores the relationships among population frequencies for sets of alleles within one locus or between two loci. The question of interest is whether the joint frequency of a set of alleles equals the product of each of the separate, marginal allele frequencies. If this holds, then the allele frequencies are in *equilibrium*; if not, they are in *disequilibrium*. Usually, equilibrium is treated as the null hypothesis and tested based on statistics derived from the discrepancy of the joint frequency and the product of the marginal frequencies.

2 Hardy–Weinberg equilibrium (HWE)

The equilibrium of the alleles within one locus is called Hardy–Weinberg equilibrium (HWE). Equilibrium holds (or will be converged to) if the allele frequencies are static in the population (no mutations, no migration), the alleles are neutral (e.g., none of the genotypes is lethal), and mating is random. Formally we define HWE as a plain mathematical relation on the allele and genotype frequencies.

For simplicity, we will consider a biallelic locus, with alleles A and a. We denote the population frequencies of the alleles and the three possible (unordered) genotypes by p_A, p_a, p_{AA}, p_{Aa}, and p_{aa}. Now, HWE is said to hold if

$$p_{AA} = p_A^2\,, \quad p_{Aa} = 2p_A p_a\,, \quad p_{aa} = p_a^2\,.$$

Given a sample of n genotypes at the loci, a common task is to judge whether HWE holds or not. The most popular approach is to view HWE as a null hypotheses and the n samples independent draws from the corresponding null distribution of the genotypes; if the actual data look extreme, in some sense, with respect to the null distribution, then the null hypothesis of HWE is rejected. We next consider two implementations of this hypothesis testing approach. Both stem from the idea of comparing the observed genotype counts, denoted as n_{AA}, n_{Aa}, and n_{aa}, to the counts expected under HWE conditionally on the observed allele counts, n_A and n_a.

2.1 Fisher's exact test

Fisher's exact test evaluates the probability of genotype counts that are equally or less likely than the observed counts, called the *tail probability*. Under HWE and the data sampling assumption the probability of the genotype counts n_{AA}, n_{Aa}, and n_{aa} given the allele counts n_A and n_a is given as

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \Pr(n_{AA}, n_{Aa}, n_{aa}) / \Pr(n_A, n_a) = \frac{n! n_A! n_a! 2^{n_{Aa}}}{n_{AA}! n_{Aa}! n_{aa}! (2n)!}\,. \qquad (1)$$

We leave it as an exercise to show that the unknown allele frequencies indeed cancel out. We denote by $T(n_{AA}, n_{Aa}, n_{aa})$ the corresponding tail probability, also called the *p-value* of the test. If $T(n_{AA}, n_{Aa}, n_{aa})$ is less than α, then HWE is rejected at significance level α. A drawback of Fisher's exact test is that evaluating the p-value becomes slow when n gets large; see the exercises.

2.2 The chi-squared test

There is another popular test, called the chi-squared test, that, for large n, is significantly faster to evaluate than Fisher's exact test. The test statistic is obtained as

$$X^2 = \sum_{\text{genotypes}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

where the expectation is with repect to independence of the two alleles. This is sometimes handy to write in terms of the observed (also the maximum-likelihood) allele frequency

$$\tilde{p}_A = n_A/(2n) = (n_{AA} + \frac{1}{2}n_{Aa})/n$$

and the estimated *disequilibrium coefficient*

$$\hat{D}_A = n_{AA}/n - (n_A/(2n))^2$$

as

$$X^2 = \frac{n\hat{D}_A^2}{\tilde{p}_A^2(1-\tilde{p}_A^2)}\,.$$

When the sample size gets large, the distribution of the test statistic X^2 approaches the χ^2 distribution (with one degree of freedom). This enables easy and fast evaluation of an approximate p-value for an observed values of X^2.

3 Linkage equilibrium (LE)

The equilibrium of the alleles or genotypes between two (or more) loci is called linkage equilibrium (LE). Usually nearby loci are *not* in LE due to the fact that genetic material is transmitted from one generation to the next generation in relatively long junks or segments of the DNA, the boundaries of the segments determined by cross-overs. However, if the population is old and large, then even nearby loci can be "very close" to LE, since the dependency of the loci decays when many cross-overs happen in between the loci. For a formal treatment of LE, we make a distinction between two types of linkage equilibrium: *allelic LE* and *genotypic LE*. In the literature allelic LE is sometimes called *gametic LE*.

3.1 Allelic LE

For simplicity we consider two biallelic loci, one with alleles A and a and the other with alleles B and b. Now, allelic LE (ALE) is said to hold if the population frequency of each of the four haplotypes equals the product of the respective allele frequencies:

$$p_{ux} = p_u p_x \qquad \text{for all } u \in \{A, a\}, x \in \{B, b\}\,.$$

Under the supposition that haplotypes are sampled (that is, observed), the development of Fisher's exact test and a chi-squared test is analogous to what we did in the case of HWE. Both are based on comparing the entries of the *contingency table* (n_{ux}) to its marginals

(n_A, n_a) and (n_B, n_b). For example, the chi-squared test statistic takes the form $n\hat{\Delta}^2$, where

$$\hat{\Delta} = \frac{\hat{D}}{\sqrt{\tilde{p}_A \tilde{p}_a \tilde{p}_B \tilde{p}_b}}$$

with the estimated disequilibrium coefficient

$$\hat{D} = \tilde{p}_{AB} - \tilde{p}_A \tilde{p}_B .$$

The actual, population-related (not sample-related) quantities are defined analogously with the hats and tildes removed.

However, if haplotypes are not observed, but only unordered genotypes at the two loci, then the presented approaches do not apply as such. The key difficulty here is that the haplotype pairs AB/ab and Ab/aB result in the same observed genotype $AaBb$, and so the frequencies of the two haplotypes cannot be deduced from the genotype counts.

Weier's solution is based on a variant of the disequilibrium coefficient D, called the *composite disequilbirum coefficient*, defined by

$$D^c = (p_{A*/B*} - p_A p_B) + (p_{A*/*B} - p_A p_B) .$$

Here $p_{A*/B*}$ denotes the population frequency of haplotype pairs that contain the haplotype AB and $p_{A*/*B}$ denotes the population frequency of haplotype pairs that contain A in one haplotype and B in the other, the other two alleles being "wild cards." The key observation is that while the frequencies $p_{A*/B*}$ and $p_{A*/*B}$ cannot be meaningfully estimated from a sample of (unordered) genotypes, *their sum can be be estimated*, since

$$p_{A*/B*} + p_{A*/*B} = 2p_{AABB} + p_{AABb} + p_{AaBB} + \frac{1}{2} p_{AaBb} .$$

The composite correlation coefficient generalizes the above defined Δ by

$$\Delta^c = \frac{D^c}{\sqrt{(p_A p_a + D_A)(p_B p_b + D_B)}} .$$

From a sample of genotypes, the respective estimate $\hat{\Delta}^c$ is easily obtained. Again, the distribution of $n(\hat{\Delta}^c)^2$ approaches $\chi^2(1)$ under the supposition that the genotypes at the two loci are independent. One can also show that $\hat{\Delta}^c$ is nothing but the correlation coefficient of two random variables, one for each locus, that take value 0, 1/2, or 1 if the genotype is AA, Aa, or aa, respectively; similarly for the genotypes at the other locus. Thus, Δ^c is a measure of *linear dependency* of the allele counts at the two loci.

3.2 Genotypic LE

In general, linkage equilibrium is defined by independency of the genotypes at the two loci, and not just as the lack of linear or log-linear dependency of the allele counts. Thus, such *genotypic LE* (GLE) is said to hold if the population frequency of each possible genotype combination at the two loci equals the product pof the respective genotype frequencies at the loci:

$$p_{uvxy} = p_{uv} p_{xy} \qquad \text{for all } uv \in \{AA, Aa, aa\}, xy \in \{BB, Bb, bb\} .$$

Again, the general ideas of Fisher's exact test and the chi-squared test apply. Both are based on comparing the entries of the *contingency table* (n_{uvxy}) to its marginals (n_{uv}) and(n_{xy}).

The chi-squared test statistic, for example, can be written as

$$X^2 = \sum_{uvxy} \frac{(n_{uvxy} - n_{uv}n_{xy}/n)^2}{n_{uv}n_{xy}/n} = \sum_{uvxy} \frac{(nn_{uvxy} - n_{uv}n_{xy})^2}{nn_{uv}n_{xy}}$$

For large samples, the distribution of X^2 under GLE approaches the $\chi^2(d)$ distribution, where $d = 4$ is the degree of freedom in the case of biallelic loci.

4 Estimating the p-value by permutation testing

The chi-squared test statistic is problematic to use when the sample size is small. Namely, the distribution is "close enough" to the asymptotic $\chi^2(d)$ distribution only when each genotype combination $uvxy$ is sampled sufficiently many times, say, at least five observed occurences.

There is a generic technique to overcome this problem, known as *permutation testing*. The idea is very simple: the genotype data in the second locus are shuffled at random to obtain a new, permuted genotype data set. Formally: if the g_i^1 and g_i^2 denote the ith observed genotype at the first and the second locus, respectively, then in the permuted data set these are replaced by $\tilde{g}_i^1 = g_i^1$ and $\tilde{g}_i^2 = g_{\pi(i)}^2$, where π is a random permutation distributed uniformly on the set of all the $n!$ possible permutations on the set $\{1, 2, \ldots, n\}$. Thus, the distribution of the test statistic on the permuted data follows the distribution under the null hypothesis of independence of the two loci, that is, GLE.

Suppose that T permuted data sets have been generated, and in $t \leq T$ of these the test statistic is greater or equal to its value on the original observed data, then an estimate for the p-value is obtained as

$$\text{p-value} \approx t/T\,.$$

The time complexity of a straightforward implementation of this procedure becomes $O(nT)$.

Exercises

1. Prove (1). Also, write the probability as a function of the number of heterozygous genotypes $x = n_{Aa}$ (and the n_A and n) alone.

2. Suppose n genotypes are sampled at a locus and you want to test for HWE. How fast can you evaluate (a) Fisher's exact test, assuming there are two alleles at the locus, (b) χ^2 test, assuming there are m alleles at the locus? We assume that basic arithmetical operations with two numbers take time $O(1)$.

3. Bob has got data consisting of 10,000 genotypes over 6 biallelic loci. He wants to compute the contingency tables for each of the $\binom{6}{2}$ pairs of loci as fast as possible. How would you suggest Bob to process his data?