The Hardy-Weinberg Principle and estimating allele frequencies in populations

Introduction

The genetic composition of a population consists of three components when we confine ourselves to the study of variation at a single locus:

1. The number of alleles at a locus.
2. The frequency of alleles at the locus.
3. The frequency of genotypes at the locus.

It may not be immediately obvious why we need both (2) and (3) to describe the genetic composition of a population, so let me illustrate with two hypothetical populations:

	A_1A_1	A_1A_2	A_2A_2
Population 1	50	0	50
Population 2	25	50	25

It's easy to see that the frequency of A_1 is 0.5 in both populations,[1] but the genotype frequencies are very different. In point of fact, we don't need both genotype and allele frequencies. We can always calculate allele frequencies from genotype frequencies, but we can't do the reverse unless ...

[1] $p_1 = 2(50)/200 = 0.5$, $p_2 = (2(25) + 50)/200 = 0.5$.

Derivation of the Hardy-Weinberg principle

We saw last time using the data from *Zoarces viviparus* that we can describe empirically and algebraically how genotype frequencies in one generation are related to genotype frequencies in the next. Let's explore that a bit further. To do so we're going to use a technique that is broadly useful in population genetics, i.e., we're going to construct a mating table.

			Offsrping genotype		
	Mating	Frequency*	A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times$	A_1A_1	x_{11}^2	1	0	0
	A_1A_2	$x_{11}x_{12}$	$\frac{1}{2}$†	$\frac{1}{2}$	0
	A_2A_2	$x_{11}x_{22}$	0	1	0
$A_1A_2 \times$	A_1A_1	$x_{12}x_{11}$	$\frac{1}{2}$	$\frac{1}{2}$	0
	A_1A_2	x_{12}^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	A_1A_2	$x_{12}x_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
$A_2A_2 \times$	A_1A_1	$x_{22}x_{11}$	0	1	0
	A_1A_2	$x_{22}x_{12}$	0	$\frac{1}{2}$	$\frac{1}{2}$
	A_2A_2	x_{22}^2	0		0

Believe it or not, in constructing this table we've already made three assumptions about the transmission of genetic variation from one generation to the next:

Assumption #1 Genotype frequencies are the same in males and females, e.g., x_{11} is the frequency of the A_1A_1 genotype in both males and females.*

Assumption #2 Genotypes mate at random *with respect to their genotype at this particular locus.**

Assumption #3 Meiosis is fair. More specifically, we assume that there is no segregation distortion, no gamete competition, no differences in the developmental ability of eggs, or the fertilization ability of sperm.†

Now that we have this table we can use it to calculate the frequency of each genotype in newly formed zygotes, provided that we're willing to make three additional assumptions:

Assumption #4 There is no input of new genetic material, i.e., gametes are produced without mutation, and all offspring are produced from the union of gametes within this population.

Assumption #5 The population is of infinite size so that the actual frequency of matings is equal to their expected frequency and the actual frequency of offspring from each mating is equal to the Mendelian expectations.

Assumption #6 All matings produce the same number of offspring, on average.

Taking these three assumptions together allows us to conclude that the frequency of a particular genotype in the pool of newly formed zygotes is

$$\sum(\text{frequency of mating})(\text{frequency of genotype produce from mating}) \quad .$$

So

$$\begin{aligned}
\text{freq.}(A_1A_1 \text{ in zygotes}) &= x_{11}^2 + \frac{1}{2}x_{11}x_{12} + \frac{1}{2}x_{12}x_{11} + \frac{1}{4}x_{12}^2 \\
&= x_{11}^2 + x_{11}x_{12} + \frac{1}{4}x_{12}^2 \\
&= (x_{11} + x_{12}/2)^2 \\
&= p^2 \\
\text{freq.}(A_1A_2 \text{ in zygotes}) &= 2pq \\
\text{freq.}(A_2A_2 \text{ in zygotes}) &= q^2
\end{aligned}$$

Those frequencies probably look pretty familiar to you. They are, of course, the familiar Hardy-Weinberg proportions. But we're not done yet. In order to say that these proportions will also be the genotype proportions of adults in the progeny generation, we have to make two more assumptions:

Assumption #7 Generations do not overlap.

Assumption #8 There are no differences among genotypes in the probability of survival.

The Hardy-Weinberg principle

After a single generation in which *all* eight of the above assumptions are satisfied

$$\begin{aligned}
\text{freq.}(A_1A_1 \text{ in zygotes}) &= p^2 \quad (1) \\
\text{freq.}(A_1A_2 \text{ in zygotes}) &= 2pq \quad (2) \\
\text{freq.}(A_2A_2 \text{ in zygotes}) &= q^2 \quad (3)
\end{aligned}$$

It's vital to understand the logic here.

1. If Assumptions #1–#8 are true, then equations 1–3 **must** be true.
2. If genotypes are in Hardy-Weinberg proportions, one or more of Assumptions #1–#8 may still be violated.
3. If genotypes are *not* in Hardy-Weinberg proportions, one or more of Assumptions #1–#8 **must** be false.

Point (3) is why Hardy-Weinberg is so important. There isn't a population of anything anywhere in the world that satisfies all 8 assumptions, even for a single generation.[2] But *all* possible evolutionary forces within populations cause a violation of at least one of these assumptions. Departures from Hardy-Weinberg are one way in which we can detect those forces and estimate their magnitude.[3]

Estimating allele frequencies

Before we can determine whether genotypes in a population are in Hardy-Weinberg proportions, we need to be able to estimate the frequency of both genotypes and alleles. This is easy when you can identify all of the alleles within genotypes, but suppose that we're trying to estimate allele frequencies in the ABO blood group system in humans. Then we have a situation that looks like this:

Phenotype	A	AB	B	O
Genotype(s)	aa ao	ab	bb bo	oo
No. in sample	n_A	N_{AB}	N_B	N_O

Now we can't directly count the number of a, b, and o alleles. What do we do? Well, if we knew p_a, p_b, and p_o, we could figure out how many individuals with the A phenotype have the aa genotype and how many have the ao phenotype, namely

$$\begin{aligned} N_{aa} &= n_A \left(\frac{p_a^2}{p_a^2 + 2p_a p_o} \right) \\ N_{ao} &= n_A \left(\frac{2p_a p_o}{p_a^2 + 2p_a p_o} \right) \quad . \end{aligned}$$

[2]There may be some that come reasonably close, but none that fulfill them *exactly*.

[3]Actually, there's a ninth assumption that I didn't mention. Everything I said here depends on the assumption that the locus we're dealing with is autosomal.

Obviously we could do the same thing for the B phenotype:

$$\begin{aligned} N_{bb} &= n_B \left(\frac{p_b^2}{p_b^2 + 2p_b p_o} \right) \\ N_{bo} &= n_B \left(\frac{2p_b p_o}{p_b^2 + 2p_b p_o} \right) \quad . \end{aligned}$$

Notice that $N_{ab} = N_{AB}$ and $N_{oo} = N_O$ (lowercase subscripts refer to genotypes, uppercase to phenotypes). If we knew all this, then we could calculate p_a, p_b, and p_o from

$$\begin{aligned} p_a &= \frac{2N_{aa} + N_{ao} + N_{ab}}{2N} \\ p_b &= \frac{2N_{bb} + N_{bo} + N_{ab}}{2N} \\ p_o &= \frac{2N_{oo} + N_{ao} + N_{bo}}{2N} \quad , \end{aligned}$$

where N is the total sample size.

Surprisingly enough we can actually estimate the allele frequencies by using this trick. Just take a guess at the allele frequencies. Any guess will do. Then calculate N_{aa}, N_{ao}, N_{bb}, N_{bo}, N_{ab}, and N_{oo} as described in the preceding paragraph.[4] That's the **E**xpectation part of what's called the EM algorithm. Now take the values for N_{aa}, N_{ao}, N_{bb}, N_{bo}, N_{ab}, and N_{oo} that you've calculated and use them to calculate new values for the allele frequencies. That's the **M**aximization part of the EM algorithm. Chances are your new values for p_a, p_b, and p_o won't match your initial guesses, but[5] if you take these new values and start the process over and repeat the whole sequence several times, eventually the allele frequencies you get out at the end match those you started with. These are maximum-likelihood estimates of the allele frequencies.[6]

Consider the following example:[7]

Phenotype	A	AB	AB	O
No. in sample	25	50	25	15

[4]Chances are N_{aa}, N_{ao}, N_{bb}, and N_{bo} won't be integers. That's OK. Pretend that there really are fractional animals or plants in your sample and proceed.

[5]Yes, truth *is* sometimes stranger than fiction.

[6]I should point out that this method *assumes* that genotypes are found in Hardy-Weinberg proportions.

[7]This is the default example available in the Java applet at http://darwin.eeb.uconn.edu/simulations/em-abo.html.

We'll start with the guess that $p_a = 0.33$, $p_b = 0.33$, and $p_o = 0.34$. With that assumption we would calculate that $25(0.33^2/(0.33^2 + 2(0.33)(0.34))) = 8.168$ of the A phenotypes in the sample have genotype *aa*, and the remaining 16.832 have genotype *ao*. Similarly, we can calculate that 8.168 of the B phenotypes in the population sample have genotype *bb*, and the remaining 16.823 have genotype *bo*. Now that we have a guess about how many individuals of each genotype we have we can calculate a new guess for the allele frequencies, namely $p_a = 0.362$, $p_b = 0.362$, and $p_o = 0.277$. By the time we've repeated this process four more times, the allele frequencies aren't changing anymore. So the maximum likelihood estimate of the allele frequencies is $p_a = 0.372$, $p_b = 0.372$, and $p_o = 0.256$.

What is a maximum-likelihood estimate?

I just told you that the method I described produces "maximum-likelihood estimates" for the allele frequencies, but I haven't told you what a maximum-likelihood estimate is. The good news is that you've been using maximum-likelihood estimates for as long as you've been estimating anything, without even knowing it. Although it will take me awhile to explain it, the idea is actually pretty simple.

Suppose we had a sock drawer with two colors of socks, red and green. And suppose we were interested in estimating the proportion of red socks in the drawer. One way of approaching the problem would be to mix the socks well, close our eyes, take one sock from the drawer, record its color and replace it. Suppose we do this N times and we found k red socks. If we knew p, the proportion of red socks in the drawer, we could calculate the probability of getting k red socks in a sample of size N. It would be

$$\binom{N}{k} p^k (1-p)^{(N-k)} \quad . \tag{4}$$

This is the *binomial probability distribution.*

But we don't know p. That's what we're trying to estimate. What gives? Well, suppose we reverse the question to which equation 4 is an answer and call the expression in 4 the "likelihood of the data." Suppose further that we find the value of p that makes the likelihood bigger than any other value we could pick. Then $\hat{p}$ is the maximum-likelihood estimate of p.[8]

In the case of the ABO blood group that we just talked about, the likelihood is a bit more complicated

[8] You'll be relieved to know that in this case, $\hat{p} = k/N$.

$$\binom{N}{N_A N_{AB} N_B N_O} p_a^{N_A} p_{ab}^{N_{AB}} p_b^{N_B} p_o^{N_O} \tag{5}$$

This is a *multinomial probability distribution*.[9]

Testing Hardy-Weinberg

Testing the hypothesis that genotypes are in Hardy-Weinberg proportions is quite simple. We can simply do a χ^2 or G-test for goodness of fit between observed and predicted genotype (or phenotype) frequencies, where the predicted genotype frequencies are derived from our estimates of the allele frequencies in the population. There's only one problem. To do either of these tests we have to know how many degrees of freedom are associated with the test. How do we figure that out? In general, the formula is

$$\begin{aligned} \text{d.f.} = & \quad (\text{\# of categories in the data -1 }) \\ - & \quad (\text{\# number of parameters estimated from the data}) \end{aligned}$$

For this problem we have

$$\begin{aligned} \text{d.f.} = & \quad (\text{\# of phenotype categories in the data - 1}) \\ - & \quad (\text{\# of allele frequencies estimated from the data}) \end{aligned}$$

In the ABO blood group we have 4 phenotype categories, and 3 allele frequencies. That means that a test of whether a particular data set has genotypes in Hardy-Weinberg proportions will have $(4-1)-(3-1)=1$ degrees of freedom for the test. Notice that this also means that if you have completely dominant markers, like RAPDs or AFLPs, you can't determine whether genotypes are in Hardy-Weinberg proportions because you have 0 degrees of freedom available for the test.

An example

Here's data from the ABO blood group:[10]

Phenotype	A	AB	B	O	Total
Observed	862	131	365	702	2060

[9] In the notes at http://darwin.eeb.uconn.edu/notes/frequencies.pdf you'll find a description of another way to use likelihoods to make inferences about parameters, but I'll save that discussion until we've had a chance to look at WinBUGS and Bayesian inference in lab.

[10] Yet again!

The maximum-likelihood estimate of allele frequencies, assuming Hardy-Weinberg, is:

$$\begin{aligned} p_a &= 0.281 \\ p_b &= 0.129 \\ p_o &= 0.590 \quad , \end{aligned}$$

giving expected numbers of 846, 150, 348, and 716 for the four phenotypes. $\chi^2 = 3.8$, $0.05 < p < 0.1$.