

Evolution Module

6.1 Hardy-Weinberg (Revised)

Bob Gardner and Lev Yampolski

Integrative Biology and Statistics (BIOL 1810)

Fall 2007

REVIEW OF PROBABILITY RULES

Note. We start by recalling a couple of rules for the computation of probabilities. First, for the Multiplicative Rule for Probabilities, if two events E_1 and E_2 are independent (that is, the occurrence of one event does not affect the probability of the other event occurring), $P(E_1) = p_1$, and $P(E_2) = p_2$, then the probability of both E_1 and E_2 is

$$P(E_1 \text{ AND } E_2) = P(E_1) \times P(E_2).$$

Second, by the Additive Rule for Probabilities, if events E_1 and E_2 are mutually exclusive (that is, both cannot occur together), then the probability of either E_1 or E_2 occurring is

$$P(E_1 \text{ OR } E_2) = P(E_1) + P(E_2).$$

Let's Illustrate these two rules with a standard "urn" problem.

Example. Consider two urns, labeled Urn 1 and Urn 2, each filled with red and white balls. Suppose that $100 \times p\%$ of the balls in each urn is red and $100 \times q\%$ of the balls in each urn is white. Choose a ball at random from each urn.

1. What is the probability of getting two red balls?

Solution. For this to occur, we must get a red ball from Urn 1 (denote this event as R_1) and a red ball from Urn 2 (denote this event as R_2). Then, since these events are independent,

$$P(R_1 \text{ AND } R_2) = P(R_1) \times P(R_2) = p \times p = p^2.$$

2. What is the probability of getting two white balls?

Solution. With similar notation as above, we have

$$P(W_1 \text{ AND } W_2) = P(W_1) \times P(W_2) = q \times q = q^2.$$

3. What is the probability of getting one red ball and one white ball?

Solution. With the established notation, this event can occur in two different ways: $(R_1 \text{ AND } W_2)$ or $(W_1 \text{ AND } R_2)$. Since these two events are mutually exclusive,

$$P((R_1 \text{ AND } W_2) \text{ OR } (W_1 \text{ AND } R_2)) =$$

$$P(R_1 \text{ AND } W_2) + P(W_1 \text{ AND } R_2) =$$

$$P(R_1) \times P(W_2) + P(W_1) \times P(R_2) = p \times q + q \times p = 2pq.$$

ONE-LOCUS/TWO ALLELES

Note. Let's now shift our attention to genetic models. We consider a one-locus/two-alleles model. Suppose the two alleles at this locus are denoted as A and a . The *frequency* of allele A is the percentage (expressed in decimal form) of alleles at the given locus which are the A allele. Denote this frequency as p . In our model, then, the frequency of the a allele is $q = 1 - p$. Suppose that a population has these frequencies of A and a . Let's now calculate the *frequency* of the genotypes in the next generation. (Here, we are assuming nonoverlapping, discrete generations.) The only way for an offspring in the next generation to have genotype AA is to inherit an A allele from each parent. The father contributes an A allele with probability p and the mother contributes an A allele with the same probability. The probability of *both* of these events is, by the Multiplication Rule for Probability (since the two events are independent) is $p \times p = p^2$:

$$P(A \text{ from father and } A \text{ from mother}) =$$

$$P(A \text{ from father}) \times P(A \text{ from mother}) = p \times p = p^2.$$

Similarly, the probability of both parents contributing the a allele to an offspring is $q \times q = q^2 = (1 - p)^2$. It follows that

the probability of a heterozygous offspring (since this is the only other possibility) is

$$\begin{aligned} 1 - (p^2 + q^2) &= 1 - (p^2 + (1 - p)^2) = 1 - (p^2 + 1 - 2p + p^2) \\ &= 2p - 2p^2 = 2p(1 - p) = 2pq. \end{aligned}$$

Another way to calculate the probability of a heterozygous offspring is as follows. The probability that the father contributes an A allele is p and the probability that the mother contributes an a allele is $q = (1 - p)$. So the offspring can have genotype Aa in this way with probability pq :

$$P(A \text{ from father and } a \text{ from mother}) =$$

$$P(A \text{ from father}) \times P(a \text{ from mother}) = p \times q = pq.$$

However, the offspring can have the same heterozygous genotype by getting the A allele from the mother and the a allele from the father—also an event with probability pq . So, again, the probability of an Aa offspring is $2pq$ by the addition rule of probabilities (since these are disjoint events):

$$\begin{aligned} &P((A \text{ from father and } a \text{ from mother}) \text{ or } (a \text{ from father and } A \text{ from mother})) \\ &= P(A \text{ from father and } a \text{ from mother}) + P(a \text{ from father and } A \text{ from mother}) \\ &= pq + pq = 2pq. \end{aligned}$$

This can be summarized in the following table.

		Maternal Allele (probability)	
		A (p)	a (q)
Paternal Allele (probability)	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

Alternatively, we can accomplish the same computation by squaring $p + q$: $(p + q)^2 = p^2 + 2pq + q^2$. The conclusion is that, regardless of the distribution of genotypes in the first generation, after one generation of random mating (and subsequently), the genotypes will be distributed according to the frequencies given above.

Note. The biological assumptions here are that the allele frequencies do not change and, therefore, the population experiences no mutation, migration, drift, or selection with respect to alleles A and a . We don't usually think of mate selection as something done randomly, but all that matters is that mating be done randomly with respect to alleles A and a . For example, you probably do not take blood type into consideration when setting up a date!

STATEMENT OF HARDY-WEINBERG

Note. We can summarize these observations in the Hardy-Weinberg Law:

Hardy Weinberg Law. *Consider a population which experiences no mutation, migration, drift, or selection with respect to a locus which contains two possible alleles, A and a . Also assume discrete (nonoverlapping) generations. If the frequency of allele A is p (in both sexes), then after one generation of random mating, the genotypes and frequencies will be AA with frequency p^2 , Aa with frequency $2pq$, and aa with frequency q^2 .*

ABO BLOOD TYPE

Note. Blood type is mentioned above. This is a trait determined by three alleles at a single locus. The alleles are commonly denoted A , B , O . These alleles combine to give the following phenotypic blood types: AA and AO (type A), BB and BO (type B), AB (type AB), and OO (type O). Denote the frequencies of alleles A , B , O as p , q , r respectively. Under the assumptions of Hardy-Weinberg, we would expect the genotypic frequencies: AA with frequency p^2 , AB with frequency $2pq$, AO with frequency $2pr$, BB with frequency q^2 , BO with frequency $2qr$, and OO with frequency r^2 . Notice that, again, these frequencies can be calculated by squaring the appropriate multinomial. This time it is $(p + q + r)^2 = p^2 + 2pq + 2pr + q^2 + 2qr + r^2$.

Example. Suppose a population is in Hardy-Weinberg equilibrium (that is, with genotypic frequencies as predicted by the Hardy-Weinberg Law) with respect to blood type. Suppose the population has 56% of individuals with blood type A and 25% of individuals with blood type O. Find the frequencies of each of the alleles A , B , and O , and the frequencies of blood types B and AB.

Solution. With the notation above, the frequency of individuals with blood type O is r^2 and so $r^2 = 25\% = 0.25$. Hence $r = 0.50$. Next, phenotypic blood type A is determined by genotypes AA and AO and so has frequency $p^2 + 2pr = p^2 + 2p(0.50) = p^2 + p$. So set this equal to 56% = 0.56: $p^2 + p = 0.56$ or $p^2 + p - 0.56 = 0$. By the quadratic equation,

$$p = \frac{-1 \pm \sqrt{(1)^2 - 4(1)(-0.56)}}{2(1)} =$$

$$-\frac{1}{2} \pm \frac{\sqrt{3.24}}{2} = -0.50 \pm 0.90.$$

So $p = -1.40$ or $p = 0.40$. Since p is between 0 and 1, it must be that $p = 0.40$. Therefore, since $p + q + r = 1$, $q = 1 - p - r = 1 - 0.40 - 0.50 = 0.10$. Notice that the

frequency of type B individuals is

$$q^2 + 2qr = (0.10)^2 + 2(0.10)(0.50) = 0.11 = 11\%$$

and the frequency of AB individuals is $2pq = 2(0.40)(0.10) = 0.08 = 8\%$. We then have:

$$\begin{aligned} \text{freq}(\text{type A}) + \text{freq}(\text{type B}) + \text{freq}(\text{type AB}) + \text{freq}(\text{type O}) \\ = 56\% + 11\% + 8\% + 25\% = 100\%, \end{aligned}$$

as expected.

THE CHI-SQUARED TEST STATISTIC

Note. We encountered the χ^2 test statistic in Section 4.5. Recall that this is a measure of the difference between observed data and the expected value of data based on some hypothesis. If there are k categories of data with O_i and E_i as the observed and expected values of the data appearing in the i th category, respectively, then the test statistic is $\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$. Under a hypothesis of Hardy-Weinberg equilibrium and known allele frequencies, we can find the predicted numbers of individuals of each genotype. The number of *degrees of freedom* is computed as the number of categories (k in our notation) minus the number of parameters estimated from the data (usually allele frequencies for a test of Hardy-Weinberg) minus 1.

Example. Consider a sample from a population with the following numbers of the three possible genotypes in a one locus/two alleles setting: The number of AA individuals in the sample is 46, the number of Aa individuals in the sample is 40, and the number of aa individuals in the sample is 14. Use a χ^2 test statistic to see if the population is in Hardy-Weinberg equilibrium.

Solution. First, we use the sample to estimate the frequency, p , of allele A (this estimation decreases the number of degrees of freedom by 1). The number of individuals in the sample is $46 + 40 + 14 = 100$, and so the sample contains 200 alleles. The genotype AA individuals contain a total of $2 \times 46 = 92$ A alleles and the genotype Aa individuals contain 40 A alleles. So our estimation of p is $p = (92 + 40)/200 = 0.66$. Therefore we calculate q as $q = 1 - p = 1 - 0.66 = 0.34$. Assuming Hardy-Weinberg equilibrium (that is, our null hypothesis is H_0 : “The population is in Hardy-Weinberg equilibrium”), we calculate the expected number of AA individuals as:

$$\text{Expected Number of } AA \text{ Individuals} = p^2 \times (\text{Population Size}).$$

So we have that the expected number of AA genotype individuals is $p^2 \times 100 = 0.66^2 \times 100 = 43.56$. Similarly, the

expected number of Aa genotype individuals is $2pq \times 100 = 2(0.66)(0.34) \times 100 = 44.88$ and the expected number of aa genotype individuals is $q^2 = (0.34)^2 \times 100 = 11.56$. Therefore for the three categories of genotypes, we have the following observed and expected values:

	AA	Aa	aa
observed (O_i)	46	40	14
expected (E_i)	43.56	44.88	11.56

Next, we calculate the test statistic as:

$$\begin{aligned} \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} &= \frac{(46 - 43.56)^2}{43.56} + \frac{(40 - 44.88)^2}{44.88} + \frac{(14 - 11.56)^2}{11.56} \\ &= 0.1367 + 0.5306 + 0.5150 = 1.1877. \end{aligned}$$

Now we have 3 categories and have estimated 1 parameter. So the number of degrees of freedom is $3 - 1 - 1 = 1$. Recall from Section 4.5, that the χ^2 distribution with one degree of freedom yields $\chi_{.100}^2 = 2.70554$ and $\chi_{.050}^2 = 3.84146$. Since our test statistic is smaller than both of these, we fail to reject the null hypothesis that the population is in Hardy-Weinberg equilibrium.

AN OBSERVATION ABOUT SAMPLE SIZE

Note. The following example illustrates how sample size can affect a hypothesis test.

Example. Consider a sample from a population with the following numbers of the three possible genotypes in a one locus/two alleles setting: The number of AA individuals in the sample is 230, the number of Aa individuals in the sample is 200, and the number of aa individuals in the sample is 70. Use a χ^2 test statistic to see if the population is in Hardy-Weinberg equilibrium.

Solution. This sample is directly proportional to the one in the previous example, only it is five times larger. Again, we find $p = 0.66$ and $q = 0.34$. Since the sample size is 500, we get the following expected values:

	AA	Aa	aa
observed (O_i)	230	200	70
expected (E_i)	217.8	224.4	57.8

The test statistic is:

$$\sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} = \frac{(230 - 217.8)^2}{217.8} + \frac{(200 - 224.4)^2}{224.4} + \frac{(70 - 57.8)^2}{57.8}$$

$$= 0.6834 + 2.6531 + 2.5751 = 5.9116.$$

Since we now have the test statistic larger than $\chi^2_{.050}$, we see that we can reject the null hypothesis that the population is in Hardy-Weinberg equilibrium with $1 - .050 = 0.950 = 95\%$ confidence.

Note. Why does the second example have such a dramatically different conclusion from the first? Afterall, the first and second samples are directly proportional in terms of the number of individuals in each category! The answer lies in the confidence we can put in large samples versus the confidence we can put in small samples. Notice that in both examples, the expected values differ somewhat from the observed values. Since a small sample can differ from the the population proportions with a higher probability than a large sample, we can put more confidence in the larger sample. That is, we are more confident that the large sample reflects the true population parameters and hence that the differences between expected and observed numbers are actually present in the population.