

Review of Population Genetics Equations

1. Hardy-Weinberg Equation:

$$p^2 + 2pq + q^2 = 1$$

Derivation: Take a gene with two alleles; call them A_1 and A_2 . (Dominance doesn't matter for our purposes; this works equally well with codominance or incomplete dominance.) In a population, some members will have the $A_1 A_1$ genotype, some will have the $A_1 A_2$ genotype, and some will have $A_2 A_2$.

Now, imagine that you can somehow take all the gametes produced by the members of the population—for simplicity, we'll assume that these are eggs and sperm. Some gametes, of course, carry A_1 , and some A_2 .

$$p = \text{freq}(A_1)$$

$$q = \text{freq}(A_2)$$

NOTICE that: the frequency of an allele is equal to the probability that a randomly chosen gamete will be carrying that allele. Also notice that $p+q=1$.

What's the chance that an egg and sperm drawn randomly will both be carrying A_1 ? Obviously, it's $p \times p$, or p^2 . And the chance that both gametes will both bear A_2 is $q \times q$, or q^2 . There are two other possibilities: sperm with A_1 and egg with A_2 , or sperm with A_2 and egg with A_1 . The chance of either one happening is $p \times q$, and the total probability of producing a zygote with the $A_1 A_2$ genotype is twice that: $2pq$. All of these probabilities sum to 1. So $p^2 + 2pq + q^2 = 1$. [Since $p+q=1$, $(p+q)^2 = p^2 + 2pq + q^2 = 1^2 = 1$.]

WHO CARES? It's like this: this equation links allele frequency to genotype frequency, assuming certain conditions are met. This means that:

- 1) If you know allele frequencies, you can predict the genotype frequencies, and compare them with the actual frequencies. If they don't match, then one of your assumptions is violated—maybe there is natural selection going on, or immigration, or non-random mating. . .
- 2) If you know genotype frequencies, you can predict allele frequencies, and compare them with the actual frequencies. Again, if they don't match, then one of your assumptions is violated.
- 3) If you know phenotype frequencies, then you can estimate genotype and allele frequencies—but you can't test the underlying assumptions.

2. Fitness:

$$p^2 w_{11} + 2pq w_{12} + q^2 w_{22} = \bar{w}$$
$$p^2 \left(\frac{w_{11}}{\bar{w}} \right) + 2pq \left(\frac{w_{12}}{\bar{w}} \right) + q^2 \left(\frac{w_{22}}{\bar{w}} \right) = 1.0$$

Derivation: w in general means “relative fitness”: a measurement of the relative ability of individuals with a certain genotype to reproduce successfully. w_{11} , for instance, means the relative ability of individuals with the A_1A_1 genotype to reproduce successfully. w is always a number between 0 and 1. Adding fitness (w) to the Hardy-Weinberg equation as shown above allows you to predict the effect of selection on gene and allele frequencies in the next generation.

Take the Hardy-Weinberg equation and multiply each term (the frequency of each genotype) by the fitness of that genotype. Add those up and you get the mean fitness, \bar{w} (“w-bar”). Divide through by \bar{w} and you get the second equation. Here, each term of the equation is multiplied by the fitness of a genotype divided by the mean fitness. If a genotype is fitter than average, this quotient is greater than 1, and that genotype will increase in frequency in the next generation. If a genotype is less fit than average, the quotient is less than 1, and that genotype will decrease in frequency in the next generation.

A related term to fitness (w) that you may run across is the selection coefficient, s . The selection coefficient compares two phenotypes and provides a measure of the proportional amount that the phenotype under consideration is less fit. With no selection against a phenotype $s=0$ and if a phenotype is completely lethal $s=1$. The relationship between relative fitness (w) and the selection coefficient (s) is $s = 1-w$.

3. Mutation:

$$p_{(t+1)} = (1 - \mu)p_t + \nu q_t$$
$$\Delta p = p_t(1 - \mu - \nu) + \nu - p_t$$

Derivation: Imagine that in each generation, allele A_1 mutates to allele A_2 with a frequency of μ , and that allele A_2 “back-mutates” to A_1 with a frequency of ν . Then in each generation, q , the frequency of the A_2 allele, increases by a factor of μp (the rate of mutation of A_1 to A_2 times the frequency of A_1) and decreases by a factor of νq . These will eventually balance each other out, so that $\Delta p = 0$ (i.e. allele frequencies don’t change any further). When $\Delta p = 0$, it must be true that $\mu p = \nu q$. From this, with a little algebraic juggling, you can derive the formula

$$\hat{p} = \frac{\nu}{(\mu + \nu)}$$

where \hat{p} (“p-hat”) is the equilibrium frequency. Similar equations let you derive \hat{q} .

This isn’t all that useful an equation, however. In humans for example, mutation rates are estimated at 1.2×10^{-8} mutations/base pair/generation. Assuming an average gene size of

~ 1000 base pairs results in a mutation rate is on the order of 10^{-5} mutations/locus/generation. For example, in humans, mutations in the gene encoding the Huntington's disease protein occur spontaneously about once in every 200,000 gametes produced. This means that mutation, by itself, has very little effect on allele frequencies.

4. Genetic Drift

Genetic drift refers to chance fluctuations in allele frequency (sampling error). We have dealt with sampling error before when looking at probabilities of parents with a particular genotype having a set of offspring with a particular phenotype distribution. There we used the **Binomial Probabilities** formula:

$$Prob = \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

Now let's apply the Binomial Probabilities formula to genetic drift. Imagine we have a population of 10 diploid individuals ($N=10$) which arose from a sample of $2N (=20)$ gametes out of an essentially infinite pool of gametes. Here we can use the binomial formula to predict the probability that our population contains exactly 0, 1, 2, ... $2N$ alleles of type A_1 . This can be found by the successive terms in the expansions of $(p+q)^{2N}$ where $p = \text{freq}(A_1)$ allele and $q = \text{freq}(A_2)$ allele in the parental generation. The probability that the sample contains exactly x alleles of type A_1 is:

$$Prob = \frac{2N!}{x!(2N-x)!} p^x q^{(2N-x)}$$

Variance of allele frequencies from one generation to the next

$$Var(p_{t+1}) = \frac{p_t(1-p_t)}{2N} = \frac{p_t * q_t}{2N}$$

Where $Var(p_{t+1})$ is the **variance** in the allele frequencies after one generation, N is the population size, and p_t and q_t are the allele frequencies in the previous generation. Variance measures the uncertainty about allele frequencies in the next generation given the current allele frequency. You may have guessed the amount of uncertainty is inversely proportional to the population size. From the above equation you can also see that variance is also dependent on allele frequencies. For a given population size the greatest variance occurs when allele frequencies are equal ($p = q = 0.5$) and decrease as the difference allele frequencies increase. The square root of the variance is the **standard deviation**, which you might be more familiar with.

There are four characteristics of genetic drift that I think are particularly important for you to remember:

1. Allele frequencies tend to change from one generation to the next simply as a result of sampling error. We can specify a probability distribution for the allele frequency in the next generation, but we cannot predict the actual frequency with certainty.
2. There is no systematic bias to changes in allele frequency. The allele frequency is as likely to increase from one generation to the next as it is to decrease.
3. If the process is allowed to continue long enough without input of new genetic material

through migration or mutation, the population will eventually become fixed for only one of the alleles originally present.

4. The time to fixation on a single allele is directly proportional to population size, and the amount of uncertainty associated with allele frequencies from one generation to the next is inversely related to population size.