OLM11.1. Predicting the variance increase due to drift

This OLM presents the standard derivation of differentiation of populations because of their limited size based on the Fisher-Wright model.

The Fisher-Wright model uses the binomial distribution to represent genetic drift, because it considers two alleles (a and A) and assumes the gamete population to be sufficiently large so that zygote production does not change the relative frequencies of the gamete types in the gamete pool. The inclusion (or exclusion) of a gamete into (or out of) the reproducing population can be seen as the result of a binary experiment repeated 2*N* times under random mating, *N* being the size of the reproducing population (cf. Bernoulli experiment, Ch2Treat, p.29). The frequency of neutral alleles in the gamete pool is equal to that in the diploid parent population (Figure 11.10). Therefore, with the relative frequency of the A allele in the parent population denoted by p_0 and that of the a allele by q_0=1-p_0, the probability of A being present in exactly *k* copies of the 2*N* gametes producing the reproducing population is

$$P(k) = \binom{2N}{k} p_0^k (1-p_0)^{2N-k} = \binom{2N}{k} p_0^k q_0^{2N-k} \qquad (11.1.1)$$

The expectation and the variance of the number of A alleles in the reproducing offspring population are

$$E(k) = 2Np_0 \qquad (11.1.2)$$

$$V(k) = 2Np_0(1-p_0) = 2Np_0q_0 \qquad (11.1.3)$$

The relative frequency of the A allele in the next generation is

$$p_1 = k/2N \qquad (11.1.4)$$

The expected relative frequency of the A allele and its variance in the offspring population are

$$E_1 = \frac{E(k)}{2N} = p_0 \qquad (11.1.5)$$

$$V_1 = \frac{V(k)}{4N^2} = \frac{p_0q_0}{2N} \qquad (11.1.6)$$

That is, the expected allele frequency of the offspring population is the same as that of the parent population regardless of population size, but its variance is inversely proportional to *N*.

As the sampling processes are independent in each generation the changes in allel variance from one generation to the next add up. Increase in the variance in a single step is proportional to p_tq_t in each population. However, the allel frequencies of the populations get different thus the expected value of p_tq_t has to be applied in 11.1.6:

$$V_{t+1} = V_t + \frac{E(p_tq_t)}{2N} \qquad (11.1.7)$$

The expected value of p_tq_t and its variance can be calculated from the allele frequencies. Let δ denote the difference from the allele frequency of the parent population; then $p_t=p_0+\delta$ and $q_t=q_0-\delta$, therefore

$$E(p_tq_t) = E[(p_0+\delta)(q_0-\delta)] = E[p_0q_0 + \delta(q_0-p_0) - \delta^2] =$$
$$= p_0q_0 + (q_0-p_0)E(\delta) - E(\delta^2) \qquad (11.1.8)$$

As $E(\delta)$=0 and $E(\delta^2) = V_t$ by definition,

$$V_{t+1} = V_t - \frac{V_t - p_0q_0}{2N}. \qquad (11.1.9)$$

Substracting p_0q_0 from both sides we get

$$V_{t+1} - p_0q_0 = V_t - p_0q_0 - \frac{V_t - p_0q_0}{2N} \qquad (11.1.10)$$

$$V_{t+1} - p_0q_0 = (V_t - p_0q_0)\left(1 - \frac{1}{2N}\right) \qquad (11.1.11)$$

As V_0=0, it follows that

$$V_t = p_0q_0\left(1 - \left(1 - \frac{1}{2N}\right)^t\right) \approx p_0q_0\left(1 - e^{-\frac{t}{2N}}\right) \qquad (11.1.12)$$

That is, the variance of the allele frequencies increases in time according to a saturating function. Since $e^{-\frac{t}{2N}}$ approaches zero in the limit, the maximum of the variance is determined by the initial allele frequencies. All the populations have been fixed one of the alleles in the limit, the expected proportion of fixing A or a is equal to their initial frequencies.