GENETIC DRIFT – THE WRIGHT-FISHER MODEL

I. Introduction

1. Our goal is to consider a large number of replicate populations, all starting with the same population size and the same gene frequency, and ask, what happens to the distribution of gene frequencies in these populations over time.
2. One reason for asking this is that we can test the predictions of theoretical models of this process by examining how replicate experimental populations diverge in gene frequency.
3. In addition, by describing the gene frequency distribution over an ensemble of replicate populations, this is equivalent to describing the probability that a particular population will have any given gene frequency over time.

II. The Wright-Fisher model

A. The so-called Wright-Fisher model (Named after Sewell Wright and R. F. Fisher, who both independently developed this model) is really very simple. Developing it consists of two steps:

1. Step 1: Develop an equation that predicts the probability that a population will change from gene frequency p at time t to gene frequency $p + 1$ at time t+1
2. Step 2: Use this equation, in matrix form, to calculate distributions of gene frequencies in successive generations.

B. Step 1

1. Suppose that we have a population of N individuals, which means that there are $2N$ copies of the A gene.
2. Suppose also that at the A locus, there are two alleles, A_1 and A_2 .
3. Finally, suppose that initially there are i copies of the A_1 allele, so that $p_1 = \frac{i}{2N}$.
4. We want to calculate the probability that a population that starts with i copies of the A_1 allele ends up with j copies after one generation in a finite population in which drift alone is acting.
5. This probability is given by:

$$T_{ij} = \binom{2N}{j} (\tfrac{i}{2N})^j (1 - \tfrac{i}{2N})^{2N-j}$$

6. To see where this comes from, first think of the alleles in a given generation as being sampled from a very large pool of gametes with frequency p_1.

   a. Then, the probability that an A_1 allele will be drawn is p_1, or $\frac{i}{2N}$.

   b. and, the probability that an A_2 allele will be drawn is $1 - p_1$, or $1 - \frac{i}{2N}$.

   c. Now, one way of getting j A_1 alleles is to draw alleles in the following order:

$$\begin{array}{cccccccccc} A_1 & A_1 & A_1 \ldots & A_1 & A_2 & A_2 & A_2 & A_2 & \ldots & A_2 \\ 1 & 2 & 3 \ldots & j & 1 & 2 & 3 & \ldots & & 2N-j \;. \end{array}$$

   d. The probability of this happening is

$$\text{Prob} = p_1 \times p_1 \times p_1 \times \;\ldots\; \times p_1 \;\times (1-p_1) \times (1-p_1) \times (1-p_1) \times \ldots \times (1-p_1)$$

$$= (p_1)^j (1-p_1)^{2N-j}$$

$$= (\tfrac{i}{2N})^j \,(1 - \tfrac{i}{2N})^{2N-j}$$

   e. Now, this is just one way of obtaining j A_1 alleles and $2N - j$ A_2 alleles.

   f. Another way is drawing the alleles in the following order:

$$\begin{array}{ccccccccccc} A_1 & A_1 & A_1 \ldots & A_1 & A_2 & A_2 & A_2 & A_2 & \ldots & A_2 & A_1 \\ 1 & 2 & 3 \ldots & j-1 & l & 2 & 3 & \ldots & & 2N-j & 1 \;. \end{array}$$

   g. But the probability of this is likewise

$$\text{Prob} = p_1 \times p_1 \times p_1 \times \;\ldots\; \times p_1 \;\times (1-p_1) \times (1-p_1) \times (1-p_1) \times \ldots \times (1-p_1) \;\times p_1$$

$$= (p_1)^j (1-p_1)^{2N-j}$$

$$= (\tfrac{i}{2N})^j \,(1 - \tfrac{i}{2N})^{2N-j}$$

   h. In fact there are $\binom{2N}{j} = \frac{2N!}{j!\,(2N-j)}$ different ways to get j A_1 alleles and $2N - j$ A_2 alleles.

   i. Thus, the total probability of obtaining this number of A_1 and A_2 alleles is just

$$T_{ij} = \binom{2N}{j} (\tfrac{i}{2N})^j \,(1 - \tfrac{i}{2N})^{2N-j} \quad \text{, as claimed.}$$

C. Step 2.

1. Now we want to calculate how a distribution of gene frequencies changes over time.
2. We must first define what we mean by the distribution of gene frequencies.
3. Suppose we have a large number of populations, each with the same number of individuals.
4. Then the distribution of gene frequncies is essentially a graph, on which the x-axis is the number of A_1 alleles in the population, whereas the y-axis is the proportion of populations having that many A_1 alleles.
5. This distribution can be represented as a vector of $2N + 1$ elements, with each element corresponding to a different number of A_1 alleles in the population.
6. For example, if there are 2 individuals in the population, there can be 0, 1, 2, 3, or 4 copies of the A_1 allele in the population.
7. Then if fractions θ_0, θ_1, θ_2, θ_3, and θ_4 represent the proportions of populations having 0, 1, 2, 3, and 4 A_1 alleles, respectively, we can represent this population by the vector

   $$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4).$$

8. Asking how the distribution of allele frequencies changes over time is thus equivalent to asking how this vector changes over time.
9. Now, suppose we consider how a population comes to have 1 A_1 allele at generation $t+1$.
10. There are 3 ways this can happen:

    a. In the previous generation (t) it may have had 1 A_1 alleles, with no change in frequency over one generation do to genetic drift. The probability of this happening is

    $$T_{11} = \binom{4}{1}\left(\tfrac{1}{4}\right)^1 \left(1 - \tfrac{1}{4}\right)^3 = .422 \ .$$

    b. In the previous generation it may have had 2 A_1 alleles, and over one generation genetic drift changed the number of A_1 alleles to 1. The probability of this happening is

    $$T_{21} = \binom{4}{1}\left(\tfrac{1}{2}\right)^1 \left(1 - \tfrac{1}{2}\right)^3 = .250 \ .$$

    c. In the previous generation, it may have had 3 A_1 alleles, and over one generation genetic drift changed the number of A_1 alleles to 1. The probability of this happening is

    $$T_{31} = \binom{4}{1}\left(\tfrac{3}{4}\right)^1 \left(1 - \tfrac{3}{4}\right)^3 = .047 \ .$$

d. Note that if a population starts with either 0 or 4 A_1 alleles, it can not change to having 1 A_1 allele due to drift. This means that the probability of these happening are

$$T_{01} = T_{41} = 0.$$

11. We can now write down the proportion of the the entire population that comes to have 1 A_1 allele in generation $t+1$:

$$\theta_1' = \theta_0 T_{01} + \theta_1 T_{11} + \theta_2 T_{21} + \theta_3 T_{31} + \theta_4 T_{41}$$

12. In similar fashion, we can write down the proportions of the population that come to have 0, 2, 3, and 4 copies of the A_1 allele at time $t+1$:

$$\theta_0' = \theta_0 T_{00} + \theta_1 T_{10} + \theta_2 T_{20} + \theta_3 T_{30} + \theta_4 T_{40}$$

$$\theta_2' = \theta_0 T_{02} + \theta_1 T_{12} + \theta_2 T_{22} + \theta_3 T_{32} + \theta_4 T_{42}$$

$$\theta_3' = \theta_0 T_{03} + \theta_1 T_{13} + \theta_2 T_{23} + \theta_3 T_{33} + \theta_4 T_{43}$$

$$\theta_4' = \theta_0 T_{04} + \theta_1 T_{14} + \theta_2 T_{24} + \theta_3 T_{34} + \theta_4 T_{44}$$

13. But this can be rewritten in matrix notation as

$$\begin{matrix} \theta_0' \\ \theta_1' \\ \theta_2' \\ \theta_3' \\ \theta_4' \end{matrix} = \begin{bmatrix} T_{00} & T_{10} & T_{20} & T_{30} & T_{40} \\ T_{01} & T_{11} & T_{21} & T_{31} & T_{41} \\ T_{02} & T_{12} & T_{22} & T_{32} & T_{42} \\ T_{03} & T_{13} & T_{23} & T_{33} & T_{43} \\ T_{04} & T_{14} & T_{24} & T_{34} & T_{44} \end{bmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$$

or,

$$\boldsymbol{\theta}' = \boldsymbol{T\theta} .$$

14. Thus, suppose that we have our population of 2 individuals. The we can calculate the matrix $\boldsymbol{T}$ using the formula given previously for the elements as

$$T = \begin{bmatrix} 1.0 & 0.316 & 0.062 & 0.004 & 0 \\ 0 & 0.422 & 0.250 & 0.047 & 0 \\ 0 & 0.211 & 0.375 & 0.211 & 0 \\ 0 & 0.047 & 0.250 & 0.422 & 0 \\ 0 & 0.004 & 0.062 & 0.316 & 1 \end{bmatrix}$$

15. If we start out with an ensemble of populations in which all populations have 2 A_1 alleles at generation t ($p_1 = .5$), then $\theta =$ (0,0,1,0,0) and after one generation of drift the distribution of gene frequencies will be:

$$\boldsymbol{\theta'} = \boldsymbol{T\theta}$$

$$= \begin{bmatrix} 1.0 & 0.316 & 0.062 & 0.004 & 0 \\ 0 & 0.422 & 0.250 & 0.047 & 0 \\ 0 & 0.211 & 0.375 & 0.211 & 0 \\ 0 & 0.047 & 0.250 & 0.422 & 0 \\ 0 & 0.004 & 0.062 & 0.316 & 1 \end{bmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} .062 \\ .250 \\ .375 \\ .250 \\ .062 \end{pmatrix}$$

D. We can use this model to ask several questions:

1. In general, what happens to gene frequencies in finite populations?
2. What is the probability of fixation of an allele due to drift?
3. How is rate of fixation related to population size?
4. What is the equilibrium distribution of the non-fixation allele-classes?