The Wright-Fisher Model and Genetic Drift

January 22, 2015

1 Hardy-Weinberg Equilibrium

Our goal is to understand the dynamics of allele and genotype frequencies in an infinite, randomly-mating population satisfying Mendel's first law. Our results are due to G. Hardy and W. Weinberg, who independently discovered them in 1908. We make the following assumptions:

1. Non-overlapping generations
2. Infinite diploid population
3. Autosomal locus segregating alleles $A_1, \cdots, A_k$
4. Monoecious individuals
5. Random mating
6. Fair meiosis (no meiotic drive)
7. No selection or mutation.

Let p_{ij} be the frequency of genotype A_iA_j in the current generation. Then the frequency of allele A_i in that generation is equal to

$$p_i = p_{ii} + \frac{1}{2}\sum_{j \neq i} p_{ij}.$$

Because the population is infinite and randomly mating, it follows that the frequency of genotype A_iA_j in the next generation is equal to

$$p'_{ij} = \begin{cases} p_i^2 & \text{if } j = i \\ 2p_ip_j & \text{if } j \neq i \end{cases}$$

Furthermore, the frequency of allele A_i in this generation is

$$\begin{aligned} p'_i &= p'_{ii} + \frac{1}{2}\sum_{j \neq i} p'_{ij} \\ &= p_i^2 + \sum_{j \neq i} p_ip_j \\ &= p_i, \end{aligned}$$

which shows that the allele frequencies do not change between generations. Since the genotype frequencies in the next generation are calculated in exactly the same way, it follows that these also will remain unchanged in all future generations. This is why this result is described as an equilibrium.

2 The Wright-Fisher Model

2.1 Description and Motivation

In fact, allele and genotype frequencies in real populations do change over time and are affected by several processes that were neglected in our derivation of HWE. Our focus in this section is on the effects of demographic stochasticity in finite populations. To this end, we introduce the Wright-Fisher model, first for a haploid population. The assumptions of this model are as follows:

1. Non-overlapping generations
2. Constant population size: N haploid adults
3. Autosomal locus segregating alleles A and a
4. No selection or mutation
5. Each individual in generation $t+1$ chooses its parent uniformly at random and with replacement from the N adults alive in generation t.

The last assumption ('Wright-Fisher sampling') is made for convenience, but can also be justified biologically if we assume that each adult in generation t gives birth to a large number of offspring (say M), but that the environment only contains enough resources for N of these to survive to adulthood.

We can also modify this model so that it applies to a randomly mating population containing N monoecious, diploid adults. In this case, the fifth assumption is replaced by:

- Each individual in generation $t+1$ chooses two chromosomes uniformly at random and with replacement from the $2N$ chromosomes present in generation t.

This mechanism can also be justified by supposing that each adult sheds a large number of gametes, which then combine at random to give rise to N diploid offspring that survive to adulthood. In particular, selfing can occur in this model and the probability that an individual has just one parent is $1/N$.

Although simplistic, the Wright-Fisher model captures many of the key features of genetic drift and has played an important role in the development of population genetics. Here we are interested in two aspects of its behavior:

- Short-term fluctuations in allele frequencies;
- Asymptotic properties of the model.

2.2 Short-term Behavior

To describe the short-term fluctuations in the allele frequencies, let us write X_t for the number of copies of allele A in generation t. Then X_t is a random variable that takes values in the set $\{0, \cdots, 2N\}$ and, conditional on $X_t = i$, the distribution of X_{t+1} is binomial with parameters $2N$ and $i/2N$, i.e.,

$$\mathbb{P}\left(X_{t+1} = j | X_t = i\right) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

Since we are usually interested in allele frequencies rather than in counts, we will also define $p_t = X_t/2N$ to be the frequency of A in generation t. Using the properties of the binomial distribution, we see that

$$\begin{aligned} \mathbb{E}\left[p_{t+1} | p_t = p\right] &= p \\ \mathrm{Var}\left(p_{t+1} | p_t = p\right) &= \frac{p(1-p)}{2N}. \end{aligned}$$

In other words, while the average allele frequencies will remain constant from generation to generation (reflecting the absence of selection and mutation), the actually allele frequencies will change at a rate that is inversely proportional to population size. These unbiased, random fluctuations in the genetic composition of the population are known as **genetic drift**, which the Wright-Fisher model suggests is stronger in smaller populations.

The Wright-Fisher model is an example of a **discrete-time Markov chain**. Indeed, from the description of the model it is clear that if we know the allele frequencies in the current generation, then those in the future will depend only on the present and not on generations further in the past. More formally, for any integers $i, j, x_0, \cdots, x_{t-1} \in \{0, \cdots, N\}$, we have

$$\mathbb{P}\left(X_{t+1} = j | X_t = i, X_{t-1} = x_{t-1}, \cdots, X_0 = x_0\right) = \mathbb{P}\left(X_{t+1} = j | X_t = i\right),$$

which tells us that given the present, the future is conditionally independent of the past. This property, which characterizes Markov chains in general, is key to the analysis of the Wright-Fisher model.

To explore the effects of genetic drift on genetic variation, we will define the **heterozygosity** (or allelic diversity) of the population in generation t to be the quantity

$$H_t^0 = \frac{2X_t(2N - X_t)}{2N(2N-1)} = \left(\frac{2N}{2N-1}\right) 2p_t(1-p_t).$$

As is clear from the first definition, the heterozygosity is equal to the probability of sampling both alleles when we sample two chromosomes at random and without replacement from the population. Also, by exploiting the Markov property of the Wright-Fisher model and the moment equations given above, we obtain the following classical result.

Theorem 1. *Under the Wright-Fisher model, the expected heterozygosity $h(t) = \mathbb{E}[H_t^0]$ decreases geometrically at rate $(1 - 1/2N)$:*

$$h(t) = \left(1 - \frac{1}{2N}\right)^t h(0).$$

While this result can be deduced directly from the moment equations, we provide an alternative proof that foreshadows the intimate relationship between the Wright-Fisher model and the coalescent.

Proof. By definition, $h(t)$ is the probability that two chromosomes, sampled at random and without replacement from the population in generation t, carry different alleles. Since we have not incorporated mutation into the model, this event will only occur if the two chromosomes have distinct ancestors in generation $t-1$ and if these ancestral chromosomes carried different alleles. Under the Wright-Fisher model, the probability that the two chromosomes have distinct ancestors in the previous generation is $1 - 1/2N$, in which case, because the ancestral chromosomes are chosen uniformly at random and without replacement from generation $t-1$, the probability that they carry different alleles is $h(t-1)$. Combining these two results, we obtain the recursion

$$h(t) = \left(1 - \frac{1}{2N}\right) \cdot h(t-1),$$

and the main result follows by continuing this argument back to generation 0. □

This shows that, on average, genetic variation tends to be reduced by genetic drift and that the rate of loss of variation due to drift is greater in smaller populations than in larger populations. Furthermore, notice that as $t \to \infty$, the expected heterozygosity decreases to 0. This suggests that in the long-term all genetic variation will be lost under the Wright-Fisher model.

2.3 Asymptotic Behavior

To characterize the asymptotic behavior of the Wright-Fisher model, define the **time to fixation** τ to be the random variable

$$\tau = \min\{t \geq 0 : p_t = 0 \text{ and } p_t = 1\}.$$

Notice that if $p_t = 1$, then the population only contains A alleles, while if $p_t = 0$, then the population only contains a alleles. In either case, the population has lost all genetic variation and we say that the surviving allele has been **fixed** in the population. Because our model lacks a mechanism to reintroduce alleles that have been lost from the population (i.e., no mutation or migration), once $p_t = 0$ or $p_t = 1$, the process remains in this state at all future times. For this reason, 0 and 1 are said to be **absorbing states** for the Wright-Fisher model. We would like to know whether this is certain to occur, i.e., whether $\mathbb{P}(\tau < \infty) = 1$, and, if so, how the probability that A is fixed in the population depends on its initial frequency.

To show that fixation is certain to occur in finite time, notice that no matter how many copies of A are present in the population in the current generation, there is always a positive probability that either A or a will be fixed in the next population. Let α be the minimum of these probabilities, i.e.,

$$\alpha = \min_{0 \leq i \leq 2N} \mathbb{P}\left(X_{t+1} \in \{0, 2N\} | X_t = i\right) > 0$$

and notice that α does not depend on t. In particular, for any initial frequency p, we have $\mathbb{P}_p(\tau > 1) \leq (1 - \alpha)$, since the probability that we have fixation within the first generation is at least α. Furthermore, by the Markov property,

$$\begin{aligned} \mathbb{P}_p(\tau > 2) &= \mathbb{P}(\tau > 2|\tau > 1) \cdot \mathbb{P}(\tau > 1) \\ &\leq (1-\alpha) \cdot (1-\alpha) \\ &= (1-\alpha)^2, \end{aligned}$$

since the probability that fixation occurs during the second generation given that it didn't occur during the first generation is also at least $(1 - \alpha)$. In general, for any $t > 0$ and any initial frequency p, we have

$$\mathbb{P}_p(\tau > t) \leq (1-\alpha)^t \to 0 \text{ as } t \to \infty.$$

However, since $\{\tau = \infty\} \subset \{\tau > t\}$ for every $t > 0$, this shows that $\mathbb{P}_p(\tau = \infty) = 0$ and so we can conclude that the time to fixation is almost surely finite: $\mathbb{P}_p(\tau < \infty) = 1$.

We next consider how the fixation probability of allele A depends on its initial frequency. To this end, define

$$u(p) = \mathbb{P}_p\{p_\tau = 1\}.$$

Before stating our main result, we will first prove a useful lemma concerning expectations on events.

Lemma 1. *Let X be a bounded random variable, i.e., for some real number $M < \infty$, $\mathbb{P}(|X| \leq M) = 1$, and let A be an event. Then*

$$|\mathbb{E}[X; A]| \leq M \cdot \mathbb{P}(A).$$

Proof. Let 1_A be the indicator variable for the event A, i.e., $1_A = 1$ if A occurs and $1_A = 0$ if A does not occur. Then 1_A is a Bernoulli random variable with parameter $p = \mathbb{P}(A) = \mathbb{E}[1_A]$ and the expectation of X on the event A is defined as

$$\mathbb{E}[X; A] = \mathbb{E}[X \cdot 1_A].$$

Since $|X| \leq M$ almost surely and 1_A is non-negative, it follows that

$$\begin{aligned}
|\mathbb{E}[X; A]| &= |\mathbb{E}[X \cdot 1_A] \\
&\leq \mathbb{E}[|X \cdot 1_A|] \\
&= \mathbb{E}[|X| \cdot 1_A] \\
&\leq M \cdot \mathbb{E}[1_A] \;=\; M \cdot \mathbb{P}(A).
\end{aligned}$$

□

Theorem 2. *Under the Wright-Fisher model, the fixation probability of an allele is equal to its initial frequency, i.e., $u(p) = p$.*

Proof. To prove this result, notice that if $\tau \leq t$, then because 0 and 1 are both absorbing states for the Wright-Fisher model, $p_t = p_\tau$. Consequently, for any fixed $t < \infty$,

$$\begin{aligned}
\mathbb{E}_p\left[p_\tau\right] &= \mathbb{E}_p\left[p_\tau; \tau \leq t\right] + \mathbb{E}_p\left[p_\tau; \tau > t\right] \\
&= \mathbb{E}_p\left[p_t; \tau \leq t\right] + \mathbb{E}_p\left[p_\tau; \tau > t\right] \\
&= \mathbb{E}_p\left[p_t\right] - \mathbb{E}_p\left[p_t; \tau > t\right] + \mathbb{E}_p\left[p_\tau; \tau > t\right] \\
&= p - \mathbb{E}_p\left[p_t; \tau > t\right] + \mathbb{E}_p\left[p_\tau; \tau > t\right].
\end{aligned}$$

However, since $p_t \in [0, 1]$, the last two terms on the right-hand side can each be bounded by the probability $\mathbb{P}_p(\tau > t)$, which we know tends to 0 as $t \to \infty$. Taking this limit shows that the left-hand side is equal to p, which completes the proof. □

As we will later see, this result holds quite generally in neutral population genetical models and has even been taken as a abstract definition of neutrality.

2.4 Simulation

The Wright-Fisher model can be simulated in several ways. The most straightforward (but not the most efficient) approach is to keep track of individual genotypes in each generation and to randomly sample the parent of each individual from the previous generation. Alternatively, if we are only interested in the dynamics of the allele frequencies, then it suffices to just record these and to generate a binomially-distributed random variable $X_{t+1} \sim \text{Binomial}(2N, p_t)$ and set $p_{t+1} = X_{t+1}/2N$. However, when N is large, say $N \geq 1000$, even this approach may be too slow for many purposes. In this case, we can replace the binomial distribution by a pair of approximations based on the central limit theorem and the law of rare events. There are three cases, depending on the product $m \equiv 2Np_t$. If $m \leq 5$, then we can use the law of rare events to approximate X_{t+1} by a Poisson random variable with mean m. Notice that this random variable is guaranteed to be non-negative, as is required by the exact model. Alternatively, if $m \geq 2N - 5$ (i.e., $2N(1 - p_t) \leq 5$), then we can use the law of rare events to approximate $2N - X_{t+1}$ by a Poisson random variable with mean $2N - m$. In the third case where $5 < m < 2N - 5$, we can invoke the central limit theorem and approximate p_{t+1} by a normally-distributed random variable with mean p_t and variance $p_t(1 - p_t)/2N$. While the resulting simulations will not exactly reproduce the Wright-Fisher model, the accuracy of the approximations based on the law of rare events and the central limit theorem for sufficiently large N is quite good.