2 The Wright-Fisher model

2.1 Definition

The Wright-Fisher model (haploid, neutral, monoecious version): A haploid, monoecious population of \(N\) genes evolves in discrete generations. Generation \(k + 1\) is formed from generation \(k\) by choosing \(N\) genes at random with replacement, i.e. each gene in generation \(k + 1\) chooses its parent independently, uniformly at random from those present in generation \(k\).

A succinct way to represent this is as follows. Suppose \(Z_k^{(i)}\) is the number of offspring of individual \(i\) in generation \(k\). Then

\[
(Z_k^{(1)}, Z_k^{(2)}, \ldots, Z_k^{(N)}) \sim \text{Multinomial} \left( N, \frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N} \right),
\]

and \(Z_k^{(i)}, Z_l^{(j)}\) are independent for \(k \neq l\). (\(Z_k^{(i)}\) and \(Z_k^{(j)}\) are not independent!)

Note: Because each individual carries exactly one copy of a gene, we can use “individuals” and “genes” interchangeably here. In a diploid model we would have to be more careful with our terminology. It turns out that in a diploid version of this model we simply replace \(N\) with \(2N\).

What we would like is a diploid and dioecious model, which can also be handled but is slightly more complicated and so we omit it; see Wakeley (2008, Section 6.1).

3 Primer on discrete-time Markov chains

A stochastic process is just a collection of random variables, \(\{X_t : t \in T\}\), where we interpret the index set \(T\) as time. We’ll work in discrete time, \(T = \mathbb{N}\). Denote by \(\mathcal{F}_t\) the history of the stochastic process up to time \(t\); the collection \((\mathcal{F}_t)_{t \in T}\) is called the filtration.

A stochastic process possesses the Markov property if its future evolution depends only on its current state, and not on its past history. In other words, it may depend on its current value, but not on how it got to that value. The models we shall look at concern stochastic processes taking values on a countable set \(\Omega\), e.g. for the Wright-Fisher model, \(\Omega = \{0, 1, \ldots, N\}\). In discrete time, we can therefore write the Markov property as:

**Definition** (Discrete time Markov chain). Let \(X = \{X_0, X_1, \ldots\}\) be a collection of random variables taking values in some countable set \(\Omega\). Then \(X\) is said to be a discrete time Markov chain if it satisfies the Markov property:

\[
P[X_k = x_k | \mathcal{F}_{k-1}] = P[X_t = x_k | X_{k-1} = x_{k-1}],
\]

that is:

\[
P[X_k = x_k | X_{k-1} = x_{k-1}, X_{k-2} = x_{k-2}, \ldots, X_0 = x_0] = P[X_k = x_k | X_{k-1} = x_{k-1}].
\]

(1)
3.1 Transition probabilities and transition matrices

Let \( X = \{X_k : k \in T\} \) be a discrete-time Markov chain taking values on a countable set \( \Omega \), with \( T = \{0, 1, \ldots \} \). By the Markov property (1), the random evolution of \( X \) is fully specified by

(i) The (distribution of the) initial state \( X_0 \) (may be deterministic or random);

(ii) the collection of transition probabilities

\[
p(k, i, j) = \mathbb{P}[X_k = j \mid X_{k-1} = i], \quad k \in T, \ i, j \in \Omega
\]

Indeed (1) implies

\[
\mathbb{P}[X_0 = x_0, X_1 = x_1, \ldots , X_n = x_n] = \mathbb{P}[X_0 = x_0] \prod_{k=1}^{n} \mathbb{P}[X_k = x_k \mid X_{k-1} = x_{k-1}] = \prod_{k=1}^{n} p(k, x_{k-1}, x_k).
\]

for every \( n \in T \) and every \((x_1, \ldots , x_n) \in \Omega^n\).

**Definition** A discrete-time Markov chain whose transition probabilities \( p(k, i, j) \) do not depend on the time \( k \) is called a *time-homogeneous Markov chain*.

**Example** Random Walk. Consider a sequence of coin flips \( Y_1, Y_2, \ldots \) where \( Y_n = 1 \) if the \( n \)-th coin flip returns Heads and \( Y_n = -1 \) otherwise. Assume that the probability of Heads of each coin is \( p \). Set \( X_0 = 0 \) and \( X_n := \sum_{i=1}^{n} Y_i, \ n = 1, 2, \ldots \). Then \( X = \{X_k : k \in T\} \) is a time-homogeneous MC with state space \( \mathbb{Z} \) and (time-homogeneous) transition probabilities

\[
p(i, j) = \begin{cases} 
  p & j = i + 1 \\
  1 - p & j = i - 1 \\
  0 & |i - j| \neq 1.
\end{cases}
\]

3.1.1 Transition matrices and the Chapman-Kolmogorov property.

**Transition matrices** In a Markov chain, the collection of transition probabilities can be organised in a matrix \( \mathbf{P} := (p(i, j) : i, j \in \Omega) \), giving, for every row \( i \), all the probability for all possible one-step transitions from \( i \). The matrix \( \mathbf{P} \) is called, unsurprisingly, the *(one-step)* transition matrix of \( X \). Thus a discrete-time MC is fully characterised by its initial distribution and its transition matrix (well-defined also in the time-inhomogeneous case).

How do we calculate probability of an event \( n \) steps in the future, when we know the present state? As usual by marginalising over all the intermediate steps. For Markov chains, this leads to the following crucial property. If \( \mathbf{P} \) is the transition matrix of a Markov chain \( X \), then the \( n \)-step transition matrix formed by the probabilities

\[
p^{(n)}(i, j) = \mathbb{P}[X_n = j \mid X_0 = i], \quad n \in T, \ i, j \in \Omega
\]
satisfies
\[ P^{(n)} = P^{(n-1)}P = PP^{(n-1)} = P \cdots P. \] (2)

[Exercise: prove it by induction]

More generally, for every \( m, n \in \mathbb{N} \),
\[ P^{(n+m)} = P^{(n)}P^{(m)} = P^{(m)}P^{(n)}. \] (3)

In other words
\[ P[X_{n+m} = j \mid X_0 = i] = \sum_{h \in \Omega} p^{(n)}(i, h)p^{(m)}(h, j) = \sum_{h \in \Omega} p^{(m)}(i, h)p^{(n)}(h, j). \] (4)

**Definition** The equivalent properties (2), (3) and (4) are referred to as the **Chapman-Kolmogorov property** of any Markov transition matrix.

### 3.1.2 Classification of states

It is clear why the word “chain” is used in association with the Markov property. This (also leads t) and the corresponding matrix interpretation reveals much of the structure of this chain: in one step, you can visit any state \( j \) only from those states corresponding to non-zero entries of the \( j \)-th column of \( P \), i.e. only from those \( i \in \Omega \) for which \( p(i, j) > 0 \). If there is at least one such \( i \), we say that \( j \) is accessible.

Similarly, any state \( j \) can be reached from the initial state \( i \) in finitely many steps only if there exist an \( n > 0 \) for which the \((i, j)\)-th entry of the matrix \( p^{(n)}(i, j) \) is non-zero. This is equivalent to saying that there exist at least a connected path from \( i \) to \( j \) of finite length that the chain \( X \) can perform with positive probability. In this case we say that \( i \) leads to \( j \).

If \( i \) leads to \( j \) and \( j \) leads to \( i \), we say that \( i \) and \( j \) intercommunicate.

The relation “\( i \sim j \) if and only if \( i \) and \( j \) intercommunicate or \( i = j \)” introduces an equivalence relation amongst the states of \( \Omega \) induced by \( P \).

**Definition** In a Markov chain with transition matrix \( P \), a state \( j \) is called an **absorbing state** if the \( j \)-th element in the diagonal of \( P \) is equal to 1. In other words, once the chain reaches state \( j \), it will stay in \( j \) for ever thereafter, with probability 1.

For example, in a \( N \)-Wright-Fisher population, the states \( \{0\} \) and \( \{N\} \) are absorbing states (prove it).

**Proposition** Let \( X \) be a Markov chain with transition matrix \( P \) on a finite state space \( \Omega \). If \( P \) is full rank and the initial state \( x_0 \) of \( X \) leads to at least one absorbing state, then \( X \) will be absorbed with probability 1.

**proof.** Exercise. hint: Try to calculate the probability that the chain is not absorbed by time \( n \).

### 3.1.3 First-step analysis

From the Chapman-Kolmogorov property follows a very useful tool to evaluate probabilities of potentially complex events, just by conditioning on “what can happen on day 1” and averaging over
all the possibilities. This is called first-step analysis. If the chain is time-homogeneous, then

\[ P[X_{n+1} = j \mid X_0 = i, X_1 = l] = P[X_{n+1} = j \mid X_1 = l] = P[X_n = j \mid X_0 = l] = p^{(n)}(l, j), \]

where the first equality follows from the Markov property and the second from time-homogeneity. Thus from (4) one has

\[ P[X_{n+1} = j \mid X_0 = i] = \sum_l P[X_1 = l \mid X_0 = i]P[X_{n+1} = j \mid X_0 = i, X_1 = l] = \sum_l p(i, l)p^{(n)}(l, j) \tag{5} \]

and the conditional probability of an event to occur at time \( n + 1 \) steps in the future, can be written as a linear combination of \( n \)-step transition probabilities. This, along with boundary conditions yields a linear system to work out several important quantities.

For example, first step analysis gives immediate information on what is the probability that, in a random walk with up-step probability \( p \), the chain returns to zero at time \( 2n \) (try it).

4 Some path properties of the Wright-Fisher model.

4.1 Probability of fixation

Let \( X_n \) denote the number (not the proportion) of individuals of a given allelic type (‘red type’) in a two-type WF population of size \( N \). We would like to compute \( a_x := P(X_\infty = N \mid X_0 = x) \) (which is an expectation of an indicator function), and denote \( P_{xy} = P(X_1 = y \mid X_0 = x) \). Condition on the first transition of the Markov chain:

\[ a_x = \sum_{y=0}^{N} P_{xy}P(X_\infty = N \mid X_1 = y) = \sum_{y=0}^{N} P_{xy}a_y, \]

where this last equality follows from the time homogeneity of the model. This is a system of equations in \( (a_x)_{x=0,..,N} \) with boundary conditions \( a_0 = 0, \ a_N = 1 \). This system is quite easy to solve, by noting that the binomial expectation “\( np \)” for \( X_1 \) is

\[ \mathbb{E}[X_1 \mid X_0 = x] = N \times \frac{x}{N} = x = \sum_{y=0}^{N} yP_{xy}, \]

where the last equality is simply the definition of an expectation. Notice that

\[ x = \sum_{y=0}^{N} yP_{xy}, \]

\[ a_x = \sum_{y=0}^{N} a_yP_{xy} \]

are of exactly the same form, which shows that \( a_x = C x \) is one solution to the latter. The boundary conditions tell us that \( C = \frac{1}{N} \), so:

\[ P(X_\infty = N \mid X_0 = x) = \frac{x}{N}. \]
4.2 Time to absorption

A similar argument to the previous example leads to a system for \( m_x = \mathbb{E}[T \mid X_0 = x] \), where \( T = \inf\{k \geq 0 : X_k \in \{0, N\}\} \):

\[
m_x = \sum_{y=0}^{N} P_{xy}(1 + m_y) = 1 + \sum_{y=0}^{N} P_{xy}m_y,
\]

with \( m_0 = m_N = 0 \). But this is much harder to solve, so we turn to an approximation. Introduce the allele frequency \( p_k := X_k/N \), which is also a Markov chain. Denote the expected time to absorption by \( t(p) := m_{Np} \). Since \( (X_{k+1}, X_k = x) \sim \text{Binomial}(N, p_k) \), we have that the change in the number of copies of allele \( A \) over one generation, \( \delta X = X_{k+1} - X_k \), satisfies

\[
\mathbb{E}[\delta X \mid X_k = x] = \mathbb{E}[X_{k+1} \mid X_k = x] - x = x - x = 0,
\]

\[
\text{Var}[\delta X \mid X_k = x] = \text{Var}[X_{k+1} \mid X_k = x] = Np(1 - p),
\]

and hence the change in allele frequency, \( \delta p = \frac{\delta X}{N} \), satisfies

\[
\mathbb{E}[\delta p \mid p_k = p] = 0,
\]

\[
\text{Var}[\delta p \mid p_k = p] = \frac{1}{N^2} \text{Var}[\delta X \mid X_k = x] = \frac{1}{N}p(1 - p).
\]

Since \( N \) is large, this describes a distribution most of whose probability mass is close to 0. In other words, we expect \( \delta p \) to be small. We can now approximate \( t(p) \) via a first-step analysis and a Taylor expansion of \( t(p + \delta p) \) around \( t(p) \), provided \( t(p) \) is twice differentiable, which we will assume:

\[
t(p) = \sum_{\delta p} P(p' = p + \delta p | p)[t(p + \delta p) + 1],
\]

\[
\approx 1 + \sum_{\delta p} P(p' = p + \delta p | p) \left( t(p) + (\delta p)t'(p) + \frac{(\delta p)^2}{2}t''(p) \right),
\]

\[
= 1 + t(p) + t'(p)\mathbb{E}[\delta p | p] + \frac{1}{2}t''(p)\mathbb{E}[(\delta p)^2 | p],
\]

\[
= 1 + t(p) + t'(p) \cdot 0 + \frac{1}{2}t''(p) \cdot \frac{1}{N}p(1 - p).
\]

Hence we have the ODE

\[
p(1 - p)t''(p) = -2N,
\]

with \( t(0) = t(1) = 0 \), whose solution is (exercise)

\[
t(p) = -2N[p \log p + (1 - p) \log(1 - p)].
\]

4.3 Decay in heterozygosity

Definition. The heterozygosity at a locus in a population is the probability that if we sample two chromosomes independently from the population then they form a heterozygous pair at this locus. (This definition is also relevant for diploid populations.)

Denote heterozygosity by \( H_k := 2p_k(1 - p_k) \). What happens to \( \mathbb{E}[H_k \mid p_0 = p] \) as \( k \to \infty \)?

We saw above that \( \mathbb{E}[p_1 \mid p_0 = p] = p \) and \( \text{Var}[p_1 \mid p_0 = p] = \frac{1}{N}p(1 - p) \). Hence, by a first step analysis:

\[
\mathbb{E}[H_1 \mid p_0 = p] = 2 \left( \mathbb{E}[p_1 \mid p_0 = p] - \mathbb{E}[p_1 \mid p_0 = p]^2 - \text{Var}[p_1 \mid p_0 = p] \right),
\]
\[= 2 \left( p - p^2 - \frac{1}{N} p (1 - p) \right),\]
\[= H_0 \left( 1 - \frac{1}{N} \right).\]

Iterating, we find \(E[H_k | H_0] = H_0 (1 - \frac{1}{N})^k\). Thus, heterozygosity decays geometrically at rate \(1/N\).

5 Approximating the Wright-Fisher model

5.1 The limit of heterozygosity

Rescale time by introducing a continuous-time parameter \(t\) such that in the \(k\)th generation of the Wright-Fisher model we have \(k = \lfloor Nt \rfloor\). To get a feel for the effect of the scaling, think first of keeping \(k\) fixed and increasing \(N\). Then \(t\) must decrease, so the continuous time point at which the \(k\)th generation takes place is approaching 0. On the other hand, keeping \(t\) fixed and increasing \(N\) requires \(k\) to increase, meaning more and more generations are squeezed into the time interval \([0, t]\).

Under this time scaling, \(E[H_k | H_0] = H_0 \left( 1 - \frac{1}{N} \right)^k \rightarrow H_0 e^{-t}\), as \(N \rightarrow \infty\).

How do we know that the rounding of \(Nt\) in the exponent does not cause problems when taking this limit? To be more precise, we should write

\[H_0 \left( 1 - \frac{1}{N} \right)^{Nt} \leq E[H_k | H_0] \leq H_0 \left( 1 - \frac{1}{N} \right)^{Nt-1}\]

with

\[H_0 \left( 1 - \frac{1}{N} \right)^{Nt-1} = H_0 \left( 1 - \frac{1}{N} \right)^{-1} \left( 1 - \frac{1}{N} \right)^{Nt} \rightarrow 1 \times H_0 e^{-t},\]

\[H_0 \left( 1 - \frac{1}{N} \right)^{Nt} \rightarrow H_0 e^{-t},\]

so that we must have \(E[H_k | H_0] \rightarrow H_0 e^{-t}\) by a squeezing argument.

5.2 Probabilities in the Wright-Fisher model

Let \(T_2^{(N)}\) denote the time until a common ancestor is found between any two of the \(n\) samples from a Wright-Fisher model of population size \(N\). Let’s consider \(T_2^{(N)}\) first. We have that

\[\mathbb{P}(T_2^{(N)} = k) = \left( 1 - \frac{1}{N} \right)^{k-1} \frac{1}{N}, \quad k = 1, 2, \ldots.\]

In other words, \(T_2^{(N)} \sim \text{Geometric}(1/N)\). What about \(T_n^{(N)}\)? We first need to compute the probability, \(p_{nj}\), that a sample of size \(n\) has \(j\) distinct ancestors in the previous generation. If we
identify parents with boxes and offspring with balls, this is the probability that dropping \( n \) balls at random into \( N \) boxes results in \( j \) occupied boxes. Let \( \mathcal{S}_n^{(j)} \) be the Stirling number of the second kind, which counts the number of ways of partitioning a set of \( n \) elements into \( j \) nonempty subsets. Then

\[
p_{nj} = \mathcal{S}_n^{(j)} \frac{N(N-1) \ldots (N-j+1)}{N^n},
\]

To see why this is correct, note that \( N(N-1) \ldots (N-j+1) \) counts the number of ways to pick \( j \) distinct parents, \( \mathcal{S}_n^{(j)} \) is the number of ways of assigning \( n \) individuals to these \( j \) parents, and \( N^n \) is the total number of ways of assigning \( n \) individuals to their parents. An explicit formula for \( \mathcal{S}_n^{(j)} \) is

\[
\mathcal{S}_n^{(j)} = \frac{1}{j!} \sum_{i=0}^{j} (-1)^{j-i} \binom{j}{i} i^n,
\]

but this is not very informative and you certainly don’t need to remember it.

Hence,

\[
P(T_n(N) = k) = p_{kn}^{k-1}(1 - p_{nn}).
\]

### 5.3 Approximation of coalescence times

A brief introduction to big-O and little-o notation. A sequence \( f(N) \) is \( \mathcal{O}(g(N)) \) for some other sequence \( g(N) \), where \( N \in \mathbb{N} \), if

\[
\limsup_{N \to \infty} \left| \frac{f(N)}{g(N)} \right| < \infty.
\]

Think of this as \( f \) growing no faster than \( g \), e.g. \( f(N) = N^2 + N \) and \( g(N) = N^2 \). Similarly, a sequence \( f(N) \) is \( o(g(N)) \) if

\[
\limsup_{N \to \infty} \left| \frac{f(N)}{g(N)} \right| = 0,
\]

or \( f \) grows strictly more slowly than \( g \), e.g. \( f = N^2 + 1 \) and \( g(N) = N^2 \log N \). A collective name for big-O and little-o notation is Landau notation.

Notice that

\[
p_{nn} = N^{-n} N(N-1) \ldots (N-n+1),
\]

\[
= \left( 1 - \frac{1}{N} \right) \left( 1 - \frac{2}{N} \right) \ldots \left( 1 - \frac{n-1}{N} \right),
\]

\[
= 1 - \binom{n}{2} \frac{1}{N} + \mathcal{O} \left( \frac{1}{N^2} \right),
\]

\[
p_{n,n-1} = \mathcal{S}_n^{(n-1)} \frac{N(N-1) \ldots (N-n+2)}{N^n},
\]

\[
= \binom{n}{2} \frac{1}{N} + \mathcal{O} \left( \frac{1}{N^2} \right),
\]

since \( \mathcal{S}_n^{(n-1)} = \binom{n}{2} \), and for \( j < n-1 \),

\[
p_{nj} = \mathcal{S}_n^{(j)} N^{-n} N(N-1) \ldots (N-j+1),
\]

\[
= \mathcal{O} \left( \frac{1}{N^2} \right).
\]
Thus, if we ignore terms of $O(N^{-2})$, transitions are now very simple. It follows from the above probabilities that $T_n^{(N)} \sim \text{Geometric}\left(\frac{n}{2} \frac{1}{N}\right)$, and when a common ancestor is found it is between precisely two (and no more) individuals. In other words, under this approximation only binary mergers are ever seen in the genealogical tree.

Now, we make the same time-rescaling as before by introducing the continuous random variable $T_n = T_n^{(N)}/N$ and writing $k = \lfloor Nt \rfloor$ to measure time in units of $N$ generations. We find

$$
\mathbb{P}(T_n \leq t) \approx \mathbb{P}(T_n^{(N)} \leq \lfloor Nt \rfloor) = 1 - \left[1 - \left(\frac{n}{2} \frac{1}{N} + O\left(\frac{1}{N^2}\right)\right)^{\lfloor Nt \rfloor}\right]_{N \to \infty} 1 - e^{-\left(\frac{n}{2}\right)t}.
$$

In other words, $T_n \sim \text{Exponential}\left(\frac{n}{2}\right)$ in units of $N$ generations. (This can be made more precise by a squeezing argument as in Section 5.1.)

### 5.4 The ancestral process

**Definition.** Sample $n$ individuals at time 0 (the present day). The *ancestral process* of the Wright-Fisher model, $A_n^{(N)}(k)$, is the number of distinct ancestors of these $n$ individuals in generation $-k$ (that is, $k$ generations ago).

From above, we have the transition probabilities

$$
\mathbb{P}(A_n^{(N)}(k+1) = j \mid A_n^{(N)}(k) = i) = p_{ij}.
$$

The limit $A_n(t) := \lim_{N \to \infty} A_n^{(N)}(\lfloor Nt \rfloor)$ has the following properties:

- $A_n(0) = n$.
- $\{A_n(t) : t \geq 0\}$ is a (continuous-time) Markov process.
- In the language of birth-and-death processes, it is a pure death process with death rates $\mu_k = \left(\frac{k}{2}\right)$ for $k \geq 2$ and $\mu_1 = 0$. In other words:

$$
\mathbb{P}(A_n(t + dt) = j \mid A_n(t) = i) = \begin{cases}
1 - \left(\frac{i}{2}\right)dt + o(dt), & \text{if } j = i, \\
\left(\frac{i}{2}\right)dt + o(dt) & \text{if } j = i - 1, \\
o(dt) & \text{otherwise}.
\end{cases}
$$

When the process reaches 1, it stays there indefinitely.

- It is related to coalescence times by the observation

$$
A_n(t) < j \text{ if and only if } T_n + T_{n-1} + \ldots + T_j \leq t,
$$

for $2 \leq j \leq n$. 