Markov Chains in Biology: Moran Model

Section 1: Introduction

Mathematical Population Genetics aims to understand quantitatively the consequences of randomness in genetic inheritance. It studies how genetic variability is shaped by natural selection, demographic factors and random genetic drift.

One of the most basic models to study haploid (cells containing a single set of chromosomes) populations is due to Patrick Moran, who first proposed it in 1958. It is a birth and death model - which in the discrete sense means that at each time epoch n, a new individual is birthed with some probability, and an existing individual dies / is removed from the population.

We introduce some basic genetic terminology used in the next sections:

A gene is a basic unit of heredity and refers to a particular chromosomal locus in the genome of an organism. Alleles are the different versions of information that can be encoded at the locus. For example, the gene that controls the shape of ripe seeds in peas, has two alleles - one that makes it wrinkled and another that makes them round.

Haploid refers to the condition of a cell containing only one set of chromosomes.
A monoecious population is one where individuals are capable of producing offsprings without mating.
The Moran model describes a haploid, monoecious population.

Section 2: Setup and Results

The Moran Model describes a birth-and-death chain on a finite population of size m.

Suppose the population consists of individuals of two types - A_1 and A_2. At each time epoch n, a random individual is chosen to give birth to an individual of the same type. Simultaneously, another random individual is chosen to be removed from the population, or die. We note that the same individual can be chosen to give birth and die - in which case there is no change in the composition of the population.

Let us consider a population of size m and let $\{X_n\}_{n\geq 0}$ denote the number of individuals of type A_1 at time n. Then $\{X_n\}_{n\geq 0}$ is modelled as a Markov chain with state space $S = \{0, 1, 2, \dots, m\}$

- Initial distribution: $a \sim \text{Uniform}(S)$
- Joint distribution:
  $\forall\, n \in \mathbb{N}$ and $i_0, i_1, \dots, i_n \in S$:
  $$\mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = a_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}$$
  (where $a_{i_0} = \mathbb{P}(X_0 = i_0)$)
  $$\Rightarrow \mathbb{P}(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_n \mid X_{n-1} = i_{n-1})$$
- The transition matrix P is defined as follows:

$$p_{00} = 1 \quad \text{and} \quad p_{mm} = 1$$

For all $i \in \{1, 2, \dots, m-1\}$:

$$p_{i,i-1} = \frac{i(m-i)}{m^2}$$

$$p_{ii} = \frac{i^2 + (m-i)^2}{m^2}$$

$$p_{i,i+1} = \frac{i(m-i)}{m^2}$$

It is easy to see that the communicating classes are: $\{0\}$, $\{1, 2, \cdots, m-1\}$, $\{m\}$.
Further, it can be seen that: $\{0\}$ and $\{m\}$ are recurrent states (or absorbing states) and $\{1, 2, \cdots, m-1\}$ are transient states.

From a biological perspective, once the state $\{0\}$ is achieved, all individuals are of type A_2 and no new individuals of type A_1 can be produced. Similarly, once the state $\{m\}$ is achieved, no individuals of type A_2 can be produced. In both these cases, there is no further change in the composition of the population. When the types of individuals A_1 and A_2 are interpreted in terms of alleles of a monoecious, haploid population, the hitting probability of the state $\{m\}$ is called the fixation probability of the allele A_1.

In the Moran model, X_n eventually reaches the states $\{0\}$ or $\{m\}$. The probability of these two events is stated precisely as follows:

Lemma 1: (Fixation Probability)
In the Moran Model, the fixation probability of the allele, when there were i copies initially, is given by:

$$\mathbb{P}_i(X_n = m \text{ for some } n \geq 0) = \frac{i}{m}$$

where $\forall\ A \in \Omega$, $\mathbb{P}_i(A) := \mathbb{P}(A \mid X_0 = i)$
($\mathcal{A}_0$ is taken to be the trivial filtration)

Another important question in mathematical population genetics is: How long it takes for the genetic diversity to disappear, especially for a large population?

First, we define the following notation:
Two positive sequences $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are denoted by $a_n \simeq b_n$ iff $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$

Theorem 1: (Mean Time to Fixation for Large Population)
Fix $i \in \{1, 2, \cdots, m-1\}$.
Let $T = \min(n \geq 0 : X_n \in \{0, m\})$.
Then the mean time to fixation, when there are i copies initially in the population is given by $\mathbb{E}_i(T)$, which satisfies:

$$\lim_{m\to\infty} \frac{\mathbb{E}_i(T)}{-m^2\left\{\left(1-\frac{i}{m}\right)\log\left(1-\frac{i}{m}\right) + \frac{i}{m}\log\frac{i}{m}\right\}} = 1$$

Take $p = i/m$. Then $p \in (0,1)$. Then the theorem can be stated succinctly as:

$$\mathbb{E}_{pm}(T) \simeq -m^2\{(1-p)\log(1-p) + p\log p\}$$

We note that in Lemma 1, we were looking at the fixation of the allele A_1 and A_2 separately, but in the Theorem 1 we look at the fixation of both the alleles together.

The term $(1-p)\log(1-p) + p\log p$ is the entropy for a Bernoulli(p) random variable.
The entropy of a discrete random variable X with pmf $P(X)$ is defined as:

$$\underset{\text{(entropy)}}{H(X)} = -\sum_{x \in \text{Range}(X)} P(x)\log P(x)$$

It is a measure of the randomness or "disorder" of a variable.

Section 3: Proof of the Results

Proof of Lemma 1:

We define the following stopping times:

$\tau_0 := \min\{n \geq 0 : X_n = 0\}$

$\tau_m := \min\{n \geq 0 : X_n = m\}$

Then $T = \tau_0 \wedge \tau_m$

$$\mathbb{P}_i(\tau_0 < \tau_m) + \mathbb{P}_i(\tau_m < \tau_0) = 1 \quad \text{---(1·1)}$$

Next we show that $\{X_n\}_{n\geq 0}$ is a Martingale:

$\mathbb{E}(|X_n|)$ is finite since the state space of $\{X_n\}_{n\geq 0}$ is finite.

Take $(x_{n-1}, x_{n-2}, \cdots, x_1) \in S^{n-1}$ s.t.
$\mathbb{P}(X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_1 = x_1) > 0$

Suppose $x_{n-1} \in \{0, m\}$. Then $X_n = x_{n-1}$ w.p. 1.
(Since $\{0\}$ and $\{m\}$ are closed states)
$\therefore \mathbb{E}(X_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_1 = x_1) = x_{n-1}$

Suppose $x_{n-1} \in \{1, 2, \cdots, m-1\}$, possible values of X_n are - $\{x_{n-1}-1, x_{n-1}, x_{n-1}+1\}$.

$$\mathbb{E}(X_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_1 = x_1)$$

$$= x_{n-1}\, \mathbb{P}(X_n = x_{n-1} \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_1 = x_1)$$
$$+ (x_{n-1}+1)\, \mathbb{P}(X_n = x_{n-1}+1 \mid X_{n-1} = x_{n-1}, \cdots, X_1 = x_1)$$
$$+ (x_{n-1}-1)\, \mathbb{P}(X_n = x_{n-1}-1 \mid X_{n-1} = x_{n-1}, \cdots, X_1 = x_1)$$

$$= x_{n-1}\, p_{x_{n-1}, x_{n-1}} + (x_{n-1}+1)\, p_{x_{n-1}, x_{n-1}+1}$$
$$+ (x_{n-1}-1)\, p_{x_{n-1}, x_{n-1}-1}$$

(by the following property of Markov Chains:
$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \cdots, X_1 = x_1) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})$)

$$= x_{n-1}\left[\frac{x_{n-1}^2 + (m - x_{n-1})^2}{m^2}\right] + (x_{n-1}+1 + x_{n-1}-1)\left[\frac{x_{n-1}(m - x_{n-1})}{m^2}\right]$$

$$= \frac{x_{n-1}}{m^2}\left[x_{n-1}^2 + (m - x_{n-1})^2 + 2x_{n-1}(m - x_{n-1})\right]$$

$$= \frac{x_{n-1}}{m^2}\left[x_{n-1} + m - x_{n-1}\right]^2 = x_{n-1}$$

This completes the proof that $\{X_n\}_{n\geq 0}$ is a Martingale.

Before we proceed, we will show that $\{0\}$ and $\{m\}$ are recurrent states and all other states are transient.

We use the following fact: $i \in S$ is recurrent $\Leftrightarrow$ $\mathbb{P}_i(\tau_i < \infty) = 1 \Leftrightarrow \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$

Look at the states $\{0\}$ and $\{m\}$:

$$p_{00}^{(n)} = p_{mm}^{(n)} = 1 \quad \forall \quad n \geq 1.$$
$$\Rightarrow \sum_{n=1}^{\infty} p_{00}^{(n)} = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} p_{mm}^{(n)} = \infty$$

Hence they are recurrent states.

Now suppose $i \in \{1, 2, \ldots, m-1\}$.
Define $\tau_i = \min\{n \geq 1 : X_n = i\}$

We note that $p_{i0}^{(i)} = \prod_{j=i}^{1} \frac{j(m-j)}{m^2} > 0$

In this case the chain stays at 0 and never returns to the state i.
Since $\{\tau_i = \infty\} \supseteq \{X_0 = i, X_1 = i-1, \ldots, X_i = 0, X_{i+1} = 0, \ldots\}$
$\Rightarrow \mathbb{P}_i(\tau_i = \infty) \geq p_{i0}^{(i)} > 0$
$\Rightarrow \mathbb{P}_i(\tau_i < \infty) < 1.$
Hence, $\{1, 2, \ldots, m-1\}$ are transient states.

Now we have sufficient conditions for applying the Optional Stopping Theorem since:

- $\mathbb{P}(T < \infty) = 1$ (Since $\{0\}$ and $\{m\}$ are the only recurrent states)
- $\{X_n\}_{n\geq 1}$ is a bounded martingale.

$\Rightarrow \mathbb{E}(|X_T|) < \infty$ and $\mathbb{P}(X_n \mid T > n)\,\mathbb{P}(T > n) \to 0$ as $n \to \infty$ (see lecture on 23rd March)

Applying the OST gives us:

$$\mathbb{E}(X_T) = \mathbb{E}(X_0) \qquad \text{---(1.2)}$$

$$\Rightarrow \mathbb{E}_i(X_T) = \mathbb{E}_i(X_0) = i$$

$$\Rightarrow m\,\mathbb{P}_i(X_T = m) = i$$

$$\Rightarrow m\,\mathbb{P}_i(\tau_m < \tau_0) = i$$

$$\Rightarrow \mathbb{P}_i(X_n = m \text{ for some } n \geq 0) = \frac{i}{m}$$

Using equation (1.1), the fixation probability of the allele A_2 can also be found:

$$\mathbb{P}_i(X_n = 0 \text{ for some } n \geq 0) = \frac{m-i}{m}$$

∎.

Proof of Theorem 1:

We want to find: $\mathbb{E}_i(T)$
In standard hitting time notation, we want to find: $k_i = \mathbb{E}_i(T)$

We first fix $j \in \{1, 2, \ldots, m-1\}$ and look at the mean time spent at j, starting at i, before absorption. For this we define a new r.v. S_j that counts the amount of time spent by the Markov chain in the state j:

$$S_j = \sum_{n=0}^{\infty} \mathbb{1}(X_n = j)$$

Let k_i^j be the mean time spent at j before absorption. Then:

$$k_i^j = \mathbb{E}_i(S_j)$$

Lemma 0.1: k_i^j is finite $\forall\ i, j \in S$.
The proof is deferred till section 4.

Claim 1: k_i^j can be described in terms of the following recurrence relation:

For $i \in \{1, 2, \cdots, m-1\}$:

$$k_i^j = \delta_{ij} + p_{i,i-1} k_{i-1}^j + p_{ii} k_i^j + p_{i,i+1} k_{i+1}^j$$

where δ_{ij} is the Kronecker-Delta function.

and $k_0^j = k_m^j = 0$

We have a recurrence relation in k_i^j.

Claim 2: The solution of the above recurrence relation is given by:

$$k_i^j = \begin{cases} i/j \; m & \text{for } i \leq j \\ \frac{m-i}{m-j} \; m & \text{for } i > j \end{cases}$$

$$k_i = \sum_{j=1}^{m-1} k_i^j = m \left\{ \sum_{j=1}^{i} \frac{m-i}{m-j} + \sum_{j=i+1}^{m-1} \frac{i}{j} \right\}$$

We want to look at the case where m is large.
$p := i/m$ and $p \in (0, 1)$.
Replacing i by pm gives us:

$$k_{pm} = m \left\{ \sum_{j=1}^{pm} \frac{m-pm}{m-j} + \sum_{j=pm+1}^{m-1} \frac{pm}{j} \right\}$$

$$= m^2 \left\{ (1-p) \sum_{j=1}^{pm} \frac{1}{m-j} + p \sum_{j=pm+1}^{m-1} \frac{1}{j} \right\}$$

We now state Lemma 0.2:

$$\sum_{j=1}^{i} \frac{1}{m-j} \approx -\log\left(1 - \frac{i}{m}\right)$$

and $$\sum_{j=i+1}^{m-1} \frac{1}{j} \approx -\log \frac{i}{m}$$

In this lemma, using $p = \frac{i}{m}$ gives us:

$$\sum_{j=1}^{pm} \frac{1}{m-j} \approx -\log(1-p)$$

and $$\sum_{j=pm+1}^{m-1} \frac{1}{j} \approx -\log p$$

$$\Rightarrow k_{pm} = \mathbb{E}_{pm}(T) \simeq -m^2 \{(1-p)\log(1-p) + p \log p\}$$

This completes the proof the theorem, barring the claims and Lemma 0.1 and 0.2. The claims are proved below, while the proofs of the lemmas are deferred till Section 4.

Proof of Claim 1:

As stated above, the random variable S_j counts the amount of time spent in state j:

$$S_j = \sum_{n=0}^{\infty} \mathbb{1}(X_n = j)$$

k_i^j, the mean time spent at j, starting from i, before absorption:

$$k_i^j = \mathbb{E}(S_j \mid X_0 = i)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i)$$

(Since each indicator function is a non-negative r.v., the order of summation can be changed)

$$k_i^j = \delta_{ij} + \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i)$$

$$= \delta_{ij} + \sum_{n=1}^{\infty} \sum_{\ell=0}^{m} \mathbb{E}(\mathbb{1}(X_n = j), X_1 = \ell \mid X_0 = i)$$

$$= \delta_{ij} + \sum_{n=1}^{\infty} \sum_{\ell=0}^{m} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i, X_1 = \ell) \cdot \mathbb{P}(X_1 = \ell \mid X_0 = i)$$

$$= \delta_{ij} + \sum_{n=1}^{\infty} \sum_{\ell=0}^{m} p_{i\ell}\, \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i, X_1 = \ell)$$

(property of Markov Chains)

Again we note that the summation consists of non-negative terms. Hence order of summation can be changed.

$$k_i^j = \delta_{ij} + \sum_{\ell=0}^{m} p_{i\ell} \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i, X_1 = \ell)$$

$$= \delta_{ij} + \sum_{\ell=0}^{m} p_{i\ell} \left(\delta_{j\ell} + \sum_{n=2}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_0 = i, X_1 = \ell) \right)$$

$$= \delta_{ij} + \sum_{\ell=0}^{m} p_{i\ell} \left(\delta_{j\ell} + \sum_{n=2}^{\infty} \mathbb{E}(\mathbb{1}(X_n = j) \mid X_1 = \ell) \right)$$

(by the following property of Markov Chains: $\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})$)

$$= \delta_{ij} + \sum_{\ell=0}^{m} p_{i\ell}\, k_\ell^j$$

Fix, $i, j \in \{1, 2, \cdots, m-1\}$

$$k_i^j = \delta_{ij} + k_{i-1}^j p_{i,i-1} + k_i^j p_{i,i} + k_{i+1}^j p_{i,i+1}$$

and $k_0^j = k_m^j = 0$

This completes the proof of claim 1.

Proof of Claim 2 : Solving the recurrence relation:

$$k_i^j = \delta_{ij} + p_{i,i-1} k_{i-1}^j + p_{ii} k_i^j + p_{i,i+1} k_{i+1}^j \quad ---(2.1)$$

$$p_{i,i-1} = p_{i,i+1} = \frac{i(m-i)}{m^2}$$

Let $a_i := p_{i,i-1} > 0$

Then $p_{ii} = 1 - 2a_i$

(2.1) can be re-written as follows :

$$k_i^j = \delta_{ij} + a_i k_{i-1}^j + a_i k_{i+1}^j + (1-2a_i) k_i^j$$

$$2k_i^j = k_{i+1}^j + k_{i-1}^j + \delta_{ij}$$

$$\Rightarrow \left(k_i^j - k_{i-1}^j\right) = \left(k_{i+1}^j - k_i^j\right) + \frac{\delta_{ij}}{a_i} \quad ---(2.2)$$

Define a new variable: $y_i := k_i^j - k_{i+1}^j$

(2.2) in terms of y_i's:

$$y_i = y_{i+1} + \frac{\delta_{ij}}{a_i} \quad ---(2.3)$$

We try to write y_i's in terms of k_1^j :

$$y_1 = k_1^j - k_0^j = k_1^j$$

$$y_2 = y_1 - \frac{\delta_{1j}}{a_1} \qquad \text{(Using (2.3))}$$
$$= k_1^j - \frac{\delta_{ij}}{a_1}$$

$$y_3 = y_2 - \frac{\delta_{2j}}{a_2} = k_1^j - \frac{\delta_{1j}}{a_1} - \frac{\delta_{2j}}{a_2}$$

Similarly, $y_i = k_1^j - \sum_{l=1}^{i-1} \frac{\delta_{lj}}{a_l}$

$$= \begin{cases} k_1^j & \text{for } i \le j \\ k_1^j - \frac{1}{a_j} & \text{for } i > j \end{cases}$$

From the definition of y_i's:

$$\sum_{l=1}^{i} y_l = k_i^j$$

$$\therefore k_i^j = \begin{cases} i\, k_1^j & \text{for } i \le j \\ i\, k_1^j - \frac{i-j}{a_j} & \text{for } i > j \end{cases}$$

We know: $k_m^j = 0$ and $a_j = \frac{j(m-j)}{m^2}$

$$\therefore k_m^j = \sum_{l=1}^{m} y_l = m\, k_1^j - \frac{m-j}{j(m-j)/m^2}$$

$$\Rightarrow 0 = m\, k_1^j - \frac{m^2}{j}$$

$$\Rightarrow k_1^j = m/j$$

This gives us: $k_i^j = \begin{cases} \frac{i}{j}\, m & i \le j \\ \frac{m-i}{m-j}\, m & i > j \end{cases}$

∎

Section 4 - Proof of Lemma 0.1 and Lemma 0.2

Proof of Lemma 0.1:

Suppose $i \in \{0, m\}$. Then $k_0^j = k_m^j = 0 \quad \forall j \in S$ since the chain stays at i.
Suppose $j \in \{0, m\}$. Then $k_i^0 = k_i^m = 0 \quad \forall i \in S$ since absorption takes place as soon as the state 0 or m is hit.

Suppose $i, j \in \{1, 2, \cdots, m-1\}$.

Claim: $\mathbb{P}_i(S_j \geq \ell) = \mathbb{P}_i(S_j < \infty)^\ell$

Proof of the claim:

$$\tau_j := \min\{n \geq 0 : X_n = j\}$$

Then we note that:

$$\{S_j \geq \ell\} = \{S_j \geq \ell, \tau_j < \infty\}$$

$$= \bigcup_{k=1}^{\infty} \{S_j \geq \ell, \tau_j = k\}$$

For $k \geq 1$, if $\mathbb{P}_i(\tau_j = k) > 0$, then:

$$\mathbb{P}_i(S_j \geq \ell, \tau_j = k) = \mathbb{P}_i(S_j \geq \ell \mid \tau_j = k) \cdot \mathbb{P}_i(\tau_j = k)$$

$$= \mathbb{P}_i(S_j \geq \ell \mid X_k = j, X_{k-1} \neq j, \cdots, X_1 \neq j) \cdot \mathbb{P}_i(\tau_j = k)$$

$$= \mathbb{P}_i(S_j \geq \ell - 1)\, \mathbb{P}_i(\tau_j = k)$$

The last step follows from the fact:

$$\{S_j \geq \ell\} \cap \{X_k = j, X_{k-1} \neq j, \cdots, X_1 \neq j\}$$

$$= \left\{ \sum_{n=k+1}^{\infty} \mathbb{1}(X_n = j) \geq \ell - 1 \right\} \cap \{X_k = j, X_{k-1} \neq j, \cdots, X_1 \neq j\}$$

Now use a change of index for the Markov Chain.

For $\mathbb{P}_i(\tau_j = k) = 0$, $\mathbb{P}_i(S_j \geq \ell, \tau_j = k) = 0$

Hence $\mathbb{P}_i(S_j \geq \ell, \tau_j = k) = \mathbb{P}_i(S_j \geq \ell - 1)\, \mathbb{P}_i(\tau_j = k)$ follows trivially.

$$\mathbb{P}_i(S_j \geq \ell) = \sum_{k=1}^{\infty} \mathbb{P}_i(S_j \geq \ell,\ \tau_j = k)$$

$$= \sum_{k=1}^{\infty} \mathbb{P}_i(S_j \geq \ell - 1)\,\mathbb{P}_i(\tau_j = k)$$

$$= \mathbb{P}_i(S_j \geq \ell - 1) \cdot \sum_{k=1}^{\infty} \mathbb{P}_i(\tau_j = k)$$

$$= \mathbb{P}_i(S_j \geq \ell - 1)\cdot \mathbb{P}_i(\tau_j < \infty)$$

An inductive argument shows that $\forall\ \ell \geq 1$:

$$\mathbb{P}_i(S_j \geq \ell) = \left(\mathbb{P}_i(\tau_j < \infty)\right)^{\ell}$$

This completes the proof of the claim.

Since S is a discrete state space,

$$\mathbb{E}_i(S_j) = \sum_{\ell=0}^{\infty} \mathbb{P}_i(S_j \geq \ell)$$

$$= \sum_{\ell=0}^{\infty} \left(\mathbb{P}_i(\tau_j < \infty)\right)^{\ell}$$

Since j is a transient state, $\mathbb{P}_i(\tau_j < \infty) < 1$ and the geometric series converges. Alternately, for the Moran model, S is finite hence the sum is finite.

∎.

Proof of Lemma 0.2:

Fix $i \in \mathbb{N}$

Let $H_n = \sum_{i=1}^{n} \frac{1}{i}$

From basic analysis, we know that:

$$H_n - 1 < \log n < H_{n-1} \quad \text{--- (4.1)}$$

$$-\log i/m = -\log i + \log m$$

Then (4.1) gives:

$$\sum_{j=1}^{m} \frac{1}{j} - 1 - \sum_{j=1}^{i-1} \frac{1}{j} < -\log i/m < \sum_{j=1}^{m-1} \frac{1}{j} - \left(\sum_{j=1}^{i} \frac{1}{j} - 1 \right)$$

For $m > i+2$:

$$\sum_{j=i}^{m} \frac{1}{j} - 1 < -\log i/m < \sum_{j=i+1}^{m-1} \frac{1}{j} + 1$$

Dividing by $\sum_{j=i+1}^{m-1} \frac{1}{j}$, a positive number, preserves the inequality:

$$\Rightarrow \frac{\sum_{j=i}^{m} \frac{1}{j} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}} < \frac{-\log i/m}{\sum_{j=i+1}^{m-1} \frac{1}{j}} < \frac{\sum_{j=i+1}^{m-1} \frac{1}{j} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}}$$

$$\lim_{m \to \infty} \frac{\sum_{j=i+1}^{m-1} \frac{1}{j} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}} = 1$$

$$\lim_{m\to\infty} \frac{\sum_{j=i}^{m} \frac{1}{j} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}} = \lim_{m\to\infty} \frac{\sum_{j=i+1}^{m-1} \frac{1}{j} - \frac{1}{i} - \frac{1}{m} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}}$$

$$= \lim_{m\to\infty} 1 + \frac{-\frac{1}{i} - 1}{\sum_{j=i+1}^{m-1} \frac{1}{j}} - \frac{1}{m} \cdot \frac{1}{\sum_{j=i+1}^{m-1} \frac{1}{j}}$$

$$= \underline{1}$$

By the Sandwich Theorem,

$$\lim_{m\to\infty} \frac{-\log i/m}{\sum_{j=i+1}^{m-1} \frac{1}{j}} = 1$$

In other words, $\sum_{j=i+1}^{m-1} \frac{1}{j} \approx -\log i/m$

We use the same procedure for

$$-\log\left(1 - \frac{i}{m}\right) = -\log(m-i) + \log m$$

(4.1) gives us:

$$\sum_{j=1}^{m} \frac{1}{j} - 1 - \sum_{j=1}^{m-i-1} \frac{1}{j} < -\log\left(1-\frac{i}{m}\right) < \sum_{j=1}^{m-1} \frac{1}{j} - \left(\sum_{j=1}^{m-i} \frac{1}{j} - 1\right)$$

For $i > 2$:

$$\Rightarrow \sum_{j=m-i}^{m} \frac{1}{j} - 1 < -\log\left(1-\frac{i}{m}\right) < \sum_{j=m-i+1}^{m-1} \frac{1}{j} - 1$$

Dividing by $\sum_{j=1}^{i} \frac{1}{m-j}$ preserves the inequality.

$$\lim_{m\to\infty} \frac{\sum_{j=m-i+1}^{m-1} \frac{1}{j} - 1}{\sum_{j=1}^{i} \frac{1}{m-j}} = \lim_{m\to\infty} 1 + \frac{\frac{-1}{m-i} - 1}{\sum_{j=1}^{i} \frac{1}{m-j}}$$

$$= 1$$

$$\lim_{m\to\infty} \frac{\sum_{j=m-i}^{m} \frac{1}{j} - 1}{\sum_{j=1}^{i} \frac{1}{m-j}} = \lim_{m\to\infty} 1 + \frac{\frac{1}{m} - 1}{\sum_{j=1}^{i} \frac{1}{m-j}}$$

$$= 1$$

By the Sandwich Theorem,

$$\lim_{m\to\infty} \frac{-\log\left(1-\frac{i}{m}\right)}{\sum_{j=1}^{i} \frac{1}{m-j}} = 1$$

In other words,

$$-\log\left(1-\frac{i}{m}\right) \approx \sum_{j=1}^{i} \frac{1}{m-j}$$

▨

Section 5: Genetic Interpretation and Some Generalizations

The Moran model described above is interpreted as the birth and death process of a monoecious, haploid population. We have studied the inheritance of a particular gene with two alleles - A_1 and A_2.

K-allele Moran Model:

In real life, many genes have more than two alleles - like the ABO blood group system. In case there are three alleles: A_1, A_2, A_3, the behaviour of the population cannot be studied fully by looking at the number of A_1 alleles, we also have to look at the number of A_2 alleles. In general, for a k-allele model, a set of Markov chains: $\{X_n^1, X_n^2, \ldots, X_n^{k-1}\}_{n \geq 1}$ is studied.

We note that for a random selection model as described above, the transition matrix for each $\{X_n^i\}_{n \geq 1}$ is the same. This is because in each step, the change in the number of a particular allele can change at most by 1.

Selection:

If individuals of different allele types have a selective advantage of $\alpha_1, \alpha_2 > 0$ for reproduction, then the probability of choosing allele A_1 for reproduction when $X_n = i$ is given is:

$$R_i^1 = \frac{\alpha_1 \cdot i}{\alpha_1 i + \alpha_2 (m-i)}$$

Similarly, probability of choosing A_2 for reproduction is:

$$R_i^2 = \frac{\alpha_2 (m-i)}{\alpha_1 i + \alpha_2 (m-i)}$$

Similarly, probabity of choosing A_1 and A_2 for death is given by:

$$D_i^1 = \frac{\beta_1 i}{\beta_1 i + \beta_2 (m-i)} \quad , \quad D_i^2 = \frac{\beta_2 (m-i)}{\beta_1 i + \beta_2 (m-i)}$$

where $\beta_1, \beta_2 > 0$ are the relative selection biases.

Then the transition matrix is given by:

$$P_{i,i-1} = R_i^2 D_i^1 \quad , \quad P_{i,i+1} = R_i^1 D_i^2$$

$$P_{ii} = 1 - P_{i,i-1} - P_{i,i+1}$$

Mutation:

Suppose A_1 mutates to A_2 with probability u and A_2 mutates to A_1 with probability v. Mutation takes place at each time epoch before individuals are chosen to give birth or to die. Then the probability of choosing A_1 when $X_n = i$ is:

$$\phi_i = \frac{i(1-u) + (m-i)(1-v)}{m}$$

The transition probabilities then become:

$$p_{i,i-1} = \phi_i (1-\phi_i) \quad , \quad p_{i,i+1} = \phi_i (1-\phi_i)$$

$$p_{ii} = 1 - p_{i,i-1} - p_{i,i+1} .$$

We observe that introducing mutation to the model changes the structure of the Markov chain. The states $\{0\}$ and $\{m\}$ are no longer absorbing states and the chain becomes recurrent and irreducible.

Section 5: Wright-Fisher Model

We will now discuss the Wright-Fisher Model, which is closely related to the Moran model. It shares many properties with the Moran model, which will be discussed in the results.

- **Setup:**

  Consider a population of $m/2$ individuals, where m is an even number.
  Suppose the population is diploid, ie, each individual possesses two copies of each chromosome - one from each parent. This means that there are m chromosomes in the population at any given time. Let the alleles be of two types - A_1 and A_2. Each gene looks like one of - A_1A_1, A_1A_2 or A_2A_2.
  At each time epoch n, the types of alleles in the next generation is found by choosing randomly from the alleles in the previous generation. This can be interpreted as mating of randomly selected individuals from the previous generation to produce m new individuals. We allow both parents to be the same and impose no requirement on the parents to be of opposite sexes.

Let $\{X_n\}_{n\geq 0}$ denote the number of alleles of type A_1 in generation n. Then $\{X_n\}_{n\geq 0}$ can be modelled as a Markov chain with state space $S = \{0, 1, 2, \ldots, m\}$ as follows:

- Initial distribution : $a \sim \text{Uniform}(S)$

- Joint distribution:
  $\forall\, n \in \mathbb{N}$ and $i_0, i_1, \ldots, i_n \in S$
  $$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = a_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}$$
  (where $a_{i_0} = \mathbb{P}(X_0 = i_0)$)

- Suppose there are 'i' number of alleles in the generation at time n, ie, $X_n = i$. At the $(n+1)^{th}$ time epoch, the alleles are chosen randomly (with replacement) from the alleles in generation n.

  Hence, $X_n | X_{n-1} = i \sim \text{Bin}\left(m, \frac{i}{m}\right)$

  This gives the transition matrix P:

  $$p_{00} = 1, \quad p_{mm} = 1$$

  $$p_{ij} = \binom{m}{j} \left(\frac{i}{m}\right)^j \left(\frac{m-i}{m}\right)^{m-j} \quad \text{for } (i,j) \neq (0,0) \text{ or } (m,m)$$

- **Results**:
  The following results have been stated as facts without proofs.

- The structure of the Markov chain is same as that of the Moran model. The communicating classes are: $\{0\}$, $\{1, 2, \ldots, m-1\}$ and $\{m\}$.

- $\{0\}$ and $\{m\}$ are recurrent (absorbing states), where as $\{1, 2, \ldots, m-1\}$ are all transient.

- The hitting time probability for the state $\{m\}$ is also the same as that for the Moran model:

$$h_i := \mathbb{P}_i(X_n = m \text{ for some } n \geq 0) = \frac{i}{m}$$

- $\{X_n\}_{n \geq 0}$ is a Martingale.

- Let T be the hitting time of $\{0, m\}$. Fix a state $i \in \{1, 2, \ldots, m-1\}$. Take $p = \frac{i}{m}$, where $p \in (0, 1)$. Then the mean hitting time as $m \to \infty$ is given by:

$$\mathbb{E}_{pm}(T) \simeq -2m\{(1-p)\log(1-p) + p\log p\}$$

We note that this limiting value of the expected hitting time for the Wright-Fisher model differs by a factor of $\frac{m}{2}$ as compared to the Moran model. This is because both the above models have the same limiting distribution - called the Kingman's Coalescent, but with different time clocks.
For further information, we refer the reader to Theorem 1.30 and Theorem 4.1 in [2].

References :

1. Markov Chains - J. R. Norris (Section 5.1)

2. Probability Models for DNA Sequence Evolution - Richard Durrett