Wright-Fisher model and Moran model

A group project, done as a part of the Stochastic Processes course
in M.Stat 1st Year (Fall 2020), Indian Statistical Institute, Kolkata
by Ranojoy Dutta (BS1608), Soham Das (BS1710), and Somak Laha (BS1720),
under the guidance of Dr. Soumendu Sundar Mukherjee

Date: November 24, 2021

Abstract

Here we discuss about two basic models in stochastic processes used in biology for finite populations. The models, namely, Wright-Fisher Model and Moran Model/Process, successfully capture the concept of genetic drift to a preliminary level. We at first discuss about the fundamentals of the models and then move on to their corresponding genetic drift, heterozygosity and asymptotic behaviour. Also we try to find similarity between the two models under their respective assumptions. Finally, simulations are done to empirically validate the core results learned while studying the models.

1 Hardy-Weinberg Equilibrium

1.1 Introduction

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. H. Hardy and W. Weinberg, who independently discovered them in 1908.

Assumption 1.1. *We make the following assumptions:*

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

1.2 Equilibrium

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

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

Furthermore, the frequency of allele A_i in this generation is

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

.
This 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 Motivation and assumptions

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 diploid adults, $2N$ chomosomes
3. Autosomal locus segregating alleles A and a. Consider two alleles only.
4. No selection or mutation or migration
5. Random mating: 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 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:

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

Note: In this report, by genotype frequency, allele frequency of some allele, we mean their proportion in the population.

2.2 Short-term behaviour

Conditional distribution and Markov property

Let X_t = number of copies of allele A in generation t and we also define $p_t = X_t/2N$. X_t is a random variable that takes values in the set $\{0, 1, \cdots, 2N\}$. Then $X_{t+1}|X_t \sim$ Binomial$(2N, p_t)$, i.e.,

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

From here we also get

$$\mathbb{E}(X_{t+1}|X_t) = X_t \qquad \text{and} \qquad \text{Var}(X_{t+1}|X_t) = X_t\left(1 - \frac{X_t}{2N}\right)$$

The Wright-Fisher model is an example of a **discrete-time Markov chain** (DTMC). From the description of the model it is clear that the future of the allele frequency depends only on the present generation and not on generations in the past. More formally, for any integers $i, j, x_0, \cdots, x_{t-1} \in \{1, \cdots, N\}$, we have

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

which tells us that given the present, the future is conditionally independent of the past.

Genetic drift and heterozygosity

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.

Definition 2.1. *To explore the effects of genetic drift on genetic variation, we will define the **heterozygosity** of the population in generation t to be the quantity*

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

Note: Heterozygosity is basically the probability of chossing two different allele (A and a) when two alleles are sample randomly without replacement from a population of $2N$. Notice that whenever we have almost equal frequency of both alleles then H_t^0 is high and when one allele has frequency close to 1 then H_t^0 is low. Hence we can say that the expression H_t^0 of heterozygosity truly captures the genetic drift. If this value changes towards a particular direction (towards 0 or 1) with time then we say that we have genetic drift in allele frequencies.

Theorem 2.2. *Under the Wright-Fisher model, the expected heterozygosity $h(t) =$*

$\mathbb{E}[H_t^0]$ *decreases geometrically at rate* $(1-1/2N)$*:*

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

Proof. First note that,

$$\begin{aligned}\mathbb{E}(X_t^2) &= \mathbb{E}[\mathbb{E}(X_t^2|X_{t-1})] = \mathbb{E}\left[\text{Var}(X_t|X_{t-1})) + \mathbb{E}(X_t|X_{t-1})^2\right] \\ &= \mathbb{E}\left[X_{t-1}\left(1 - \frac{X_{t-1}}{2N}\right) + X_{t-1}^2\right] = \left(1 - \frac{1}{2N}\right)\mathbb{E}(X_{t-1}^2) + \mathbb{E}(X_{t-1}).\end{aligned}$$

Then it follows that,

$$\begin{aligned}\mathbb{E}(2NX_t - X_t^2) &= 2N\mathbb{E}(X_t) - \mathbb{E}(X_t^2) = 2N\mathbb{E}[\mathbb{E}(X_t|X_{t-1})] - (1 - \frac{1}{2N})\mathbb{E}(X_{t-1}^2) - \mathbb{E}(X_{t-1}) \\ &= 2N\mathbb{E}(X_{t-1}) - (1 - \frac{1}{2N})\mathbb{E}(X_{t-1}^2) - \mathbb{E}(X_{t-1}) = (1 - \frac{1}{2N})\mathbb{E}(2NX_{t-1} - X_{t-1}^2)\end{aligned}$$

$$\Rightarrow h(t) = \mathbb{E}(H_t^0) = \frac{\mathbb{E}(2NX_t - X_t^2)}{N(2N-1)} = (1 - \frac{1}{2N})\frac{\mathbb{E}(2NX_{t-1} - X_{t-1}^2)}{N(2N-1)} = (1 - \frac{1}{2N})h(t-1)$$

So, we have recursive relation $h(t) = \left(1 - \frac{1}{2N}\right)h(t-1)$. Solving it we get the desired result. □

2.3 Asymptotic behaviour

Absorbing states and fixation

0 and $2N$ are two **absorbing states** for X_t. Consequently 0 and 1 are absorbing states for p_t.Let us define **time to fixation** as below.

$$\tau = \{t \geq 0 : X_t \in \{0, 2N\}\}$$

Lemma 2.3. *τ is finite w.p. 1.*

Proof. Let $\alpha = \min_{0 \leq i \leq 2N} \mathbb{P}(X_{t+1} \in \{0, 2N\}|X_t = i) \geq \left(\frac{1}{2N}\right)^{2N} > 0$.
Then can have bounds for the following probabilities:

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

So, $\sum_{t=0}^{\infty} \mathbb{P}(\tau > t) = \frac{1}{\alpha} < \infty \qquad \Rightarrow \mathbb{E}(\tau) < \infty$.
This proves $\mathbb{P}(\tau < \infty) = 1$. □

Fixation probability

For allele A, $X_t =$ no. of allele A at time t. Let $T_j = \min\{t \geq 0 : X_t = j\}$. Then **fixation probability** is defined as: $p_A = \mathbb{P}(T_{2N} < T_0 | X_0)$.

Note: Starting frequency and fixation probability are allele specific quantities. Here we have used these terms for allele A throughout the report.

Lemma 2.4. $\mathbb{P}(T_b < T_a | X_0 = i) = \dfrac{i-a}{b-a}$*, for any state i between states a and b.*

Proof. Let τ denote the time when X_t first hit a or b starting from the state $X_0 = i$. Then X_τ takes values a and b.
We know that

$$\begin{aligned} i = \mathbb{E}(X_\tau | X_0 = i) &= a\mathbb{P}(T_b > T_a | X_0 = i) + b\mathbb{P}(T_b < T_a | X_0 = i) \\ \text{and } 1 &= \mathbb{P}(T_b > T_a | X_0 = i) + \mathbb{P}(T_b < T_a | X_0 = i) \end{aligned}$$

This two equations solve for
$\mathbb{P}(T_b < T_a | X_0 = i) = \dfrac{i-a}{b-a}$ and $\mathbb{P}(T_b > T_a | X_0 = i) = \dfrac{b-i}{b-a}$. □

Theorem 2.5. *Under the Wright-Fisher model, the fixation probability of an allele is equal to its initial frequency.*

Proof. Lets prove this for allele A. Let initial frequency of A is p_0 and X_t be no, of allele A at time t. Then according to the last lemma, fixation probability of A,
$\mathbb{P}(T_{2N} < T_0 | X_0 = 2Np_0) = \dfrac{2Np_0 - 0}{2N - 0} = p_0$. □

2.4 Simulations

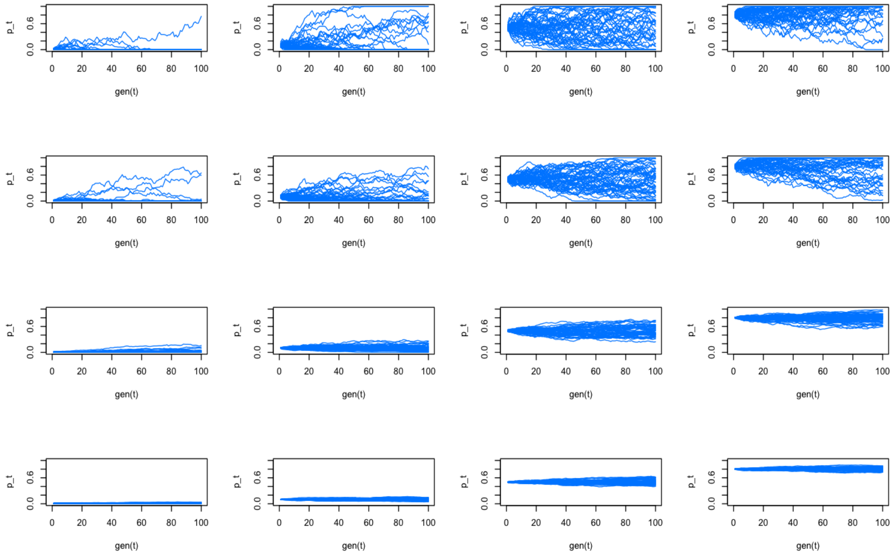

Figure 1: Plot of p_t vs t for different starting frequencies $p_0 = 0.01, 0.1, 0.5, 0.8$ and different population sizes $2N = 50, 100, 1000, 5000$. Each row corresponds to one population size(increases as we move from top to bottom) and each row corresponds to one starting frequency(increases as we move from left to right). Each plot shows 50 simulations for short time span (100 generations). Note that as population size increases relative spread of p_t decreases.

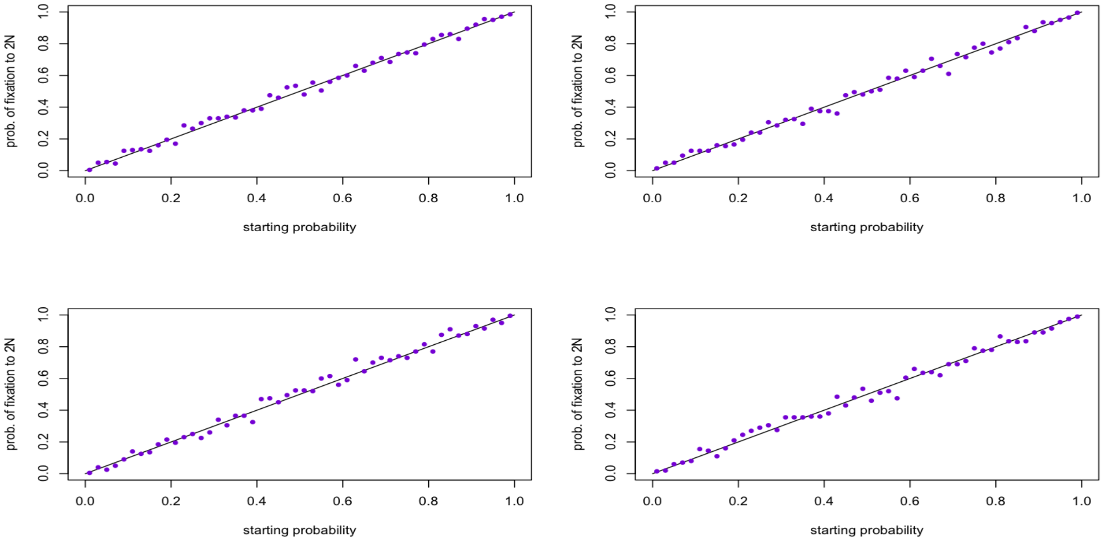

Figure 2: Plot of p_A (fiaxtion probability of A) vs p_0 (starting frequency of A) for $2N = 100, 200, 1000, 2000$. This confirms that fixation probability of allele A is same as its starting probability for different population size.

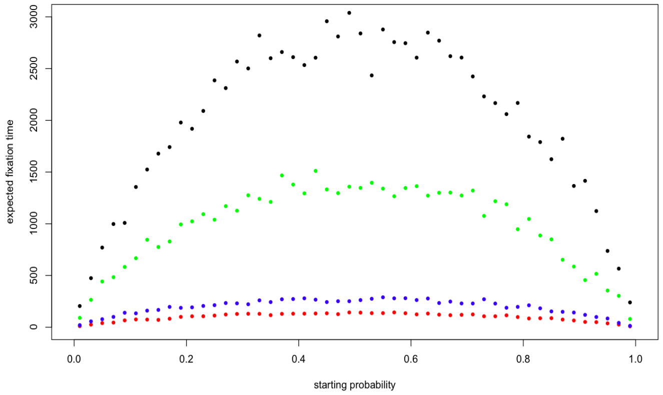

Figure 3: Plot of expected fixation time $\mathbb{E}(\tau)$ of an allele vs starting probability p_0 of allele A for $2N = 100, 200, 1000, 2000$. Note that when p_0 is close to $\frac{1}{2}$, $\mathbb{E}(\tau)$ is large. Also if population size increases, allele-fixation takes more time.

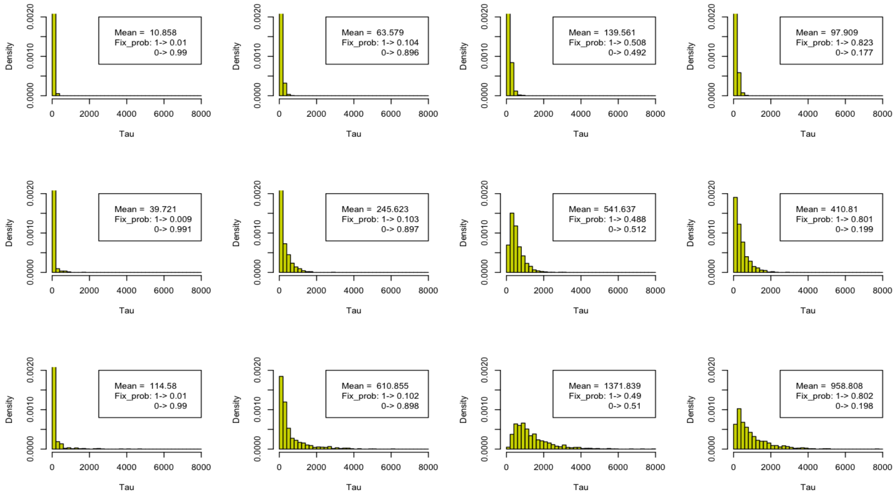

Figure 4: Histograms of observed values of τ for different starting frequencies of allele A $p_0 = 0.01, 0.1, 0.5, 0.8$ and different population sizes $2N = 100, 400, 1000$. In each plot, mean of observed τ, fixation probabilities of allele A to 0 and 1 are shown. Note that fixation probabilities to state 1 is close to starting frequency of A.

3 The Moran model

3.1 Description

A population of $2N$ genes (labelled $1, 2, \ldots, 2N$) evolves according to the **Moran model** if at exponential rate $\binom{2N}{2}$ a pair of genes are sampled (with replacement) and from the population, one dies and the other splits in two.

To visualize, suppose there are $2N$ individuals present at certain time. Each of the $\binom{2N}{2}$ pairs of individuals has an exponential clock with rate 1 ticking. When a clock goes off, with equal probabilities either one of them splits, and the other one dies.

3.2 Assumptions

The basic comparison of assumptions between Wright-Fisher model and Moran model are as follows.

- The population is always constant in both the models, which is $2N$.
- Wright - Fisher model described a Discrete Time Markov Chain. Moran model describes a Continuous Time Markov Chain.
- W-F model evolves in discrete generations. But Moran model allows overlapping generations.

3.3 CTMC

Like before, we assume that the population of size $2N$ consists of two alleles - A and a. It is enough to consider the Continuous time Markov Chain $(X_t)_{t\geq 0}$ which observes the number of allele A only. The state space is $\{0, 1, 2, \ldots, 2N\}$. Define $p_t = X_t/2N$ to be frequency of A at time t.

The generator matrix $Q = ((q_{ij}))$ is formulated as below.

$$q_{i,j} = \begin{cases} -4Np_i(1-p_i) & j = i, \\ 2Np_i(1-p_i) & j = i+1, \\ 2Np_i(1-p_i) & j = i-1, \\ 0 & \text{otherwise.} \end{cases}$$

3.4 Short term behaviour

First we state and prove a useful result that we need to prove further results.

Lemma 3.1. $(X_t : t \geq 0)$ *be a Markov Chain with values in a finite state space* $E = \{1, \ldots, n\}$ *and rate matrix* Q. *Let* $f : E \to \mathbb{R}$ *be a real-valued function and for each* $i \in E$ *and* $t \geq 0$ *define* $u_i(t)$ *to be the following.*

$$u_i(t) = \mathbb{E}[f(X_t) \mid X_0 = i].$$

Then the functions $u_1(t), \ldots, u_n(t)$ *satisfy the following differential equations:*

$$\frac{d}{dt}u_i(t) = \sum_k p_{ik}(t) \left(\sum_j q_{kj}(f(j) - f(k)) \right), \quad i \in E,$$

where $p_{ik}(t) = \mathbb{P}(X_t = k \mid X_0 = i)$ *is the transition probability of* X.

Proof. We first recall that the transition probabilities $p_{ij}(t)$ satisfy the Kolmogorov forward equations, which take the form

$$\frac{d}{dt} p_{ij}(t) = \sum_k p_{ik}(t) q_{kj}$$

Consequently,

$$\begin{aligned}
\frac{d}{dt} u_i(t) &= \frac{d}{dt} \sum_j p_{ij}(t) f(j) \\
&= \sum_j \frac{d}{dt} p_{ij}(t) f(j) \\
&= \sum_j \sum_k p_{ik}(t) q_{kj} f(j) \\
&= \sum_k p_{ik}(t) \sum_j q_{kj} f(j) \\
&= \sum_k p_{ik}(t) \left(\sum_{j \neq k} q_{kj} f(j) + q_{kk} f(k) \right) \\
&= \sum_k p_{ik}(t) \left(\sum_{j \neq k} q_{kj} (f(j) - f(k)) \right)
\end{aligned}$$

where we have used the fact that $q_{kk} = -\sum_{j \neq k} q_{jk}$ to pass to the final line. □

Corollary 3.2. *We take the identity function $f(i) = i$ and use the above lemma to obtain that*

$$\frac{d}{dt} \mathbb{E}[X_t \,|\, X_0 = i] = 0,$$

for every $i \in \{0, 1, \ldots, 2N\}$. This shows that the expected number of copies of allele A is constant over time and therefore

$$\mathbb{E}[X_t \,|\, X_0 = i] = \mathbb{E}[X_0 \,|\, X_0 = i] = i.$$

3.5 Heterozygosity

Definition 3.3. *The probability that at time point t, we randomly sample two chromosomes and get both alleles A and a, is known as the heterozygosity $H(t)$.*

$$H(t) = \frac{2X_t(2N - X_t)}{(2N)^2}.$$

We have this following result regarding expected heterozygosity.

Theorem 3.4. *Define* $\overline{H}_i(t) := \mathbb{E}[H(t) \,|\, X_0 = i]$*. Then,*

$$\overline{H}_i(t) = e^{-t/N}\overline{H}_i(0).$$

Proof. A little algebra shows that,

$$\begin{aligned}
&\sum_j q_{kj}(f(j) - f(k)) \\
&= 2Np_k(1-p_k)\left[2\left(p_k + \frac{1}{2N}\right)\left(1 - p_k - \frac{1}{2N}\right) + 2\left(p_k - \frac{1}{2N}\right)\left(1 - p_k + \frac{1}{2N}\right) - 4p_k(1-p_k)\right] \\
&= 2Np_k(1-p_k)\cdot\left(-\frac{1}{N^2}\right)
\end{aligned}$$

if $1 \le k \le 2N-1$, while the corresponding expression vanishes if $k = 0$ or $k = 2N$. It then follows from the lemma that the expected heterozygosity satisfies the following equation

$$\begin{aligned}
\frac{d}{dt}H_i(t) &= \sum_k p_{ik}(t)\cdot\left(-\frac{1}{N}2p_k(1-p_k)\right) \\
&= -\frac{1}{N}H_i(t),
\end{aligned}$$

which has solution

$$H_i(t) = e^{-\frac{t}{N}}H_i(0),$$

where $\overline{H}_i(0) = 2p_i(1-p_i)$ □

Note: For Wright - Fisher model, we had $\overline{H}_i(t) = \left(1 - \frac{1}{2N}\right)^t \overline{H}_i(0) \approx e^{-t/2N}\overline{H}_i(0)$.
Hence, we may say that a population of size $2N$ governed by the Moran model looses its variation twice as rapidly as a population of same size governed by the Wright - Fisher model.

3.6 Asymptotic Behaviour

Fixation and fixation probability

The absorbing states are 0 and $2N$ in Moran model, and let

$$\tau = \min\{t \ge 0 : \, X_t \in \{0, 2N\}\}$$

be the time of fixation of one or the other allele. τ is finite with probability 1 since the embedded chain is absorbed with probability 1.
The next result shows that the fixation probability of allele A is the same as the initial proportion of A, just like Wright - Fisher model.

Theorem 3.5. *Let $u(i) = \mathbb{P}(X_\tau = 2N \,|\, X_0 = i)$ be the probability that A is eventually fixed in a population of size $2N$ that initially contains i copies of A. Under the Moran model,*

$$u(i) = \frac{i}{2N}.$$

Proof. Since τ is almost surely finite, and $\mathbb{E}[X_t \,|\, X_0 = i] = i$ holds for all fixed time $t \geq 0$,

$$i = \mathbb{E}[X_\tau \,|\, X_0 = i] = 2N\,\mathbb{P}(X_\tau = 2N \,|\, X_0 = i) = 2Nu(i).$$

Hence, $u(i) = \dfrac{i}{2N}$. □

Fixation Time

It is quite interesting to investigate the distribution of the fixation time τ for different population sizes and initial proportions of alleles. However here we present a result regarding only the unconditional and conditional (given A gets fixed) expected fixation time as function of population size and starting allele frequencies. Consider the following two notations.

$$\mathbb{E}_i[\tau] = \mathbb{E}[\tau \,|\, X_0 = i],$$
$$\overline{\mathbb{E}}_i[\tau] = \mathbb{E}[\tau \,|\, X_0 = i, X_\tau = 2N].$$

Theorem 3.6. *Under Moran model with two alleles, A and a, the expected time to fixation of either allele is approximately*

$$\mathbb{E}_i[\tau] \approx -2N\{p\log(p) + (1-p)\log(1-p)\},$$

where $p = \frac{i}{2N}$ is the initial proportion of allele A and time is measured in units of generations. Similarly, the expected time to fixation conditional on eventual fixation of A is given by

$$\overline{\mathbb{E}}_i[\tau] = \frac{-2N(1-p)}{p}\log(1-p).$$

Proof. If S_j is the amount of time that the process X spends in state j before time τ, then $\tau = S_1 + \cdots + S_{2N-1}$ and so the unconditioned expected time to fixation is equal to

$$\mathbb{E}_i[\tau] = \sum_{j=1}^{2N-1} \mathbb{E}_i\left[S_j\right]$$

Furthermore, if N_j is the number of times that X visits state j before time τ and we let $q_j = 2j(2N-j)/2N$ be the rate at which the process leaves state j, then because each visits lasts on average $1/q_j$ generations, we have

$$\mathbb{E}_i\left[S_j\right] = \frac{1}{q_j}\mathbb{E}_i\left[N_j\right]$$

To calculate the expected value of N_j, we can reason as follows. Let T_k be the hitting time of k as defined in Proposition 1 and notice that N_j will be greater than 0 if and only if the process hits state j before either allele is fixed in the population. If i is between 0 and j, then by Proposition 1 this probability is

$$\mathbb{P}_i\left(N_j \geq 1\right) = \mathbb{P}_i\left(T_j < T_0\right) = \frac{i}{j}$$

Similarly, if i is between j and $2N$, then this probability is

$$\mathbb{P}_i\left(N_j \geq 1\right) = \mathbb{P}_i\left(T_j < T_{2N}\right) = 1 - \mathbb{P}_i\left(T_{2N} < T_j\right) = 1 - \frac{i-j}{2N-j} = \frac{2N-i}{2N-j}$$

Suppose that $N_j \geq 1$. In this case, following the first visit from i to j, the process is at state j and its history prior to that time is independent of the number of subsequent visits to state j. In fact, if we let π_j denote the probability that the process never returns to state j when it starts in state j, then

$$\mathbb{P}_i\left(N_j \geq n+1 \mid N_j \geq n\right) = 1 - \pi_j$$

for every $n \geq 1$. This shows that conditional on $N_j \geq 1, N_j$ is geometrically-distributed with success parameter π_j and so

$$\mathbb{E}_i\left[N_j\right] = \mathbb{P}_i\left(N_j \geq 1\right) \cdot \frac{1}{\pi_j}$$

To calculate π_j, observe that because the process is equally likely to move to state $j-1$ or to state $j+1$ when it leaves state j, we have

$$\begin{aligned}
\pi_j &= \frac{1}{2} \cdot \mathbb{P}_{j-1}\left(T_0 < T_j\right) + \frac{1}{2} \cdot \mathbb{P}_{j+1}\left(T_{2N} < T_j\right) \\
&= \frac{1}{2} \cdot \frac{1}{j} + \frac{1}{2} \cdot \frac{1}{2N-j} \\
&= \frac{2N}{2j(2N-j)}
\end{aligned}$$

This shows that the expected number of visits to state j prior to fixation is

$$\mathbb{E}_i\left[N_j\right] = \begin{cases} \frac{i}{j} \cdot \frac{2j(2N-j)}{2N} & \text{if } i \leq j \\ \frac{2N-i}{2N-j} \cdot \frac{2j(2N-j)}{2N} & \text{if } i \geq j \end{cases}$$

and therefore

$$\mathbb{E}_i\left[S_j\right] = \begin{cases} \frac{i}{j} & \text{if } i \leq j \\ \frac{2N-i}{2N-j} & \text{if } i \geq j \end{cases}$$

since (fortuitously) $q_j = 1/\pi_j$. Summing then gives

$$\begin{aligned}
\mathbb{E}_i[\tau] &= \sum_{j=1}^{i-1} \frac{2N-i}{2N-j} + \sum_{j=i}^{2N-1} \frac{i}{j} \\
&= 2N \cdot \left\{ \left(1 - \frac{i}{2N}\right) \cdot \frac{1}{2N} \sum_{j=1}^{i-1} \frac{1}{1-j/2N} + \left(\frac{i}{2N}\right) \cdot \frac{1}{2N} \sum_{j=i}^{2N-1} \frac{1}{j/2N} \right\} \\
&\sim 2N \left\{ (1-p) \int_0^p \frac{1}{1-q} dq + p \int_p^1 \frac{1}{q} dq \right\} \\
&= -2N\{p\log(p) + (1-p)\log(1-p)\}
\end{aligned}$$

where $p = i/2N$ is the initial frequency. To calculate the conditional mean fixation time, observe that the transition probabilities of the Moran model conditional on eventual fixation of A are

$$\begin{aligned}
\bar{p}_t(i,j) &= \mathbb{P}\left(X_t = j \mid T_{2N} < T_0, X_0 = i\right) \\
&= \mathbb{P}\left(X_t = j \mid X_0 = i\right) \cdot \frac{\mathbb{P}\left(T_{2N} < T_0 \mid X_t = j, X_0 = i\right)}{\mathbb{P}\left(T_{2N} < T_0 \mid X_0 = i\right)} \\
&= p_t(i,j) \cdot \frac{\mathbb{P}\left(T_{2N} < T_0 \mid X_t = j\right)}{\mathbb{P}\left(T_{2N} < T_0 \mid X_0 = i\right)} \\
&= p_t(i,j) \cdot \frac{j/2N}{i/2N} \\
&= p_t(i,j) \cdot \frac{j}{i}
\end{aligned}$$

In fact. it can be shown that the Moran model conditioned on eventual fixation of A is also a continuous-time Markov chain, with transition rates $\bar{q}_{ij} = q_{ij} \cdot (j/i)$. Since $\bar{q}_{ij} > q_{ij}$ whenever $j > i$, the frequency of A is disproportionately likely to increase in the conditioned process, as might be expected. In this case, we can express the conditional mean time to fixation as the sum

$$\overline{\mathbb{E}}_i[\tau] = \sum_{j=1}^{2N-1} \overline{\mathbb{E}}_i\left[S_j\right]$$

where $\overline{\mathbb{E}}_i\left[S_j\right]$ is the mean occupancy time in state j by the conditioned process. Since

$$S_j = \int_0^\infty 1_j\left(X_t\right) dt$$

where $1_j\left(X_t\right)$ is equal to 1 if $X_t = j$ and 0 otherwise, the conditional expectation of this variable is

$$\begin{aligned}\overline{\mathbb{E}}_i\left[S_j\right] &= \overline{\mathbb{E}}\left[\int_0^\infty 1_j\left(X_t\right)dt\right] = \int_0^\infty \overline{\mathbb{E}}\left[1_j\left(X_t\right)\right]dt \\ &= \int_0^\infty \bar{p}_t(i,j)dt = \left(\frac{j}{i}\right)\int_0^\infty p_t(i,j)dt = \frac{j}{i}\cdot\mathbb{E}_i\left[S_j\right].\end{aligned}$$

Using the value of $\mathbb{E}_i\left[S_j\right]$ calculated above, we find

$$\overline{\mathbb{E}}_i\left[S_j\right] = \begin{cases} 1 & \text{if } i \le j \\ \frac{2N-i}{i}\cdot\frac{j}{2N-j} & \text{if } i \ge j \end{cases}$$

Summing from 1 to $2N-1$ then gives

$$\begin{aligned}\overline{\mathbb{E}}_i[\tau] &= \sum_{j=1}^{i-1}\frac{2N-i}{i}\cdot\frac{j}{2N-j} + \sum_{j=i}^{2N-1}1 \\ &= \left\{2N\cdot\left(\frac{1-i/2N}{i/2N}\right)\cdot\frac{1}{2N}\sum_{j=1}^{i-1}\frac{j/2N}{1-j/2N} + 2N - i\right\} \\ &= 2N\cdot\left\{\left(\frac{1-i/2N}{i/2N}\right)\cdot\frac{1}{2N}\sum_{j=1}^{i-1}\frac{j/2N}{1-j/2N} + (1-i/2N)\right\} \\ &\sim 2N\left\{\frac{(1-p)}{p}\int_0^p\frac{q}{1-q}dq + (1-p)\right\} \\ &= 2N\left\{\frac{(1-p)}{p}(-\log(1-p)-p) + (1-p)\right\} \\ &= -\frac{2N(1-p)}{p}\log(1-p)\end{aligned}$$

□

3.7 Simulations

Computational complexity

We must mention that it is historically pretty expensive to simulate Moran processes. We used a R package *GillespieSSA* to simulate the moran process. Though the package allowed performing exact simulation of Moran processes, doing it multiple times (say, for empirical distribution of τ) required multiple hours which we could not afford. Instead we used some variants of 'tau leaping algorithm' which is specifically used for simulation of Stochastic systems.

Simulation of the process for different populations

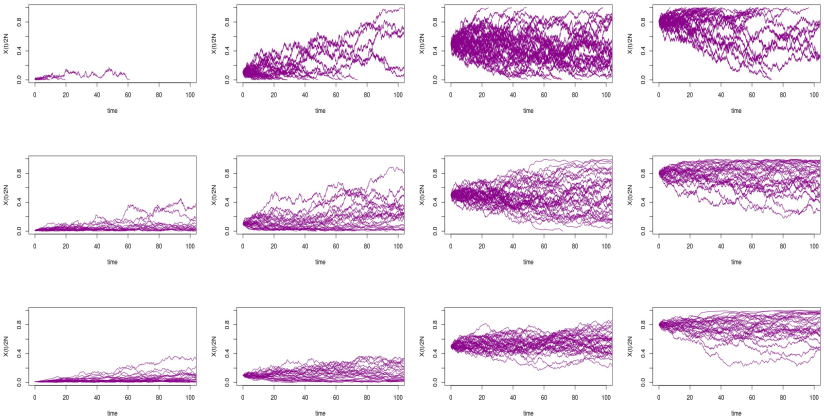

Figure 5: p vs t plots for $p_0 = 0.01, 0.1, 0.5, 0.8$ and $2N = 200, 500, 1000$.

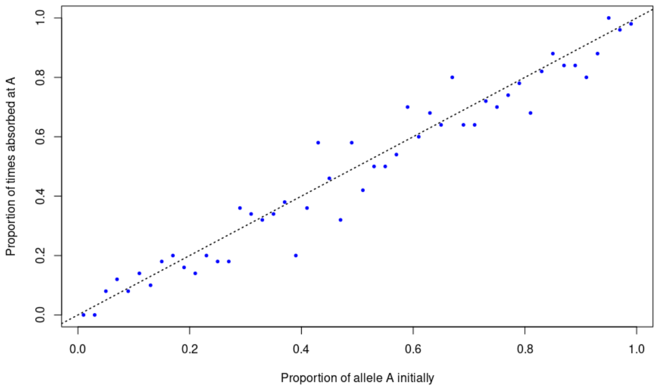

Figure 6: Simulations were done for population size $2N = 500$. It verifies the result that the fixation probability of allele A is same as the initial proportion of A.

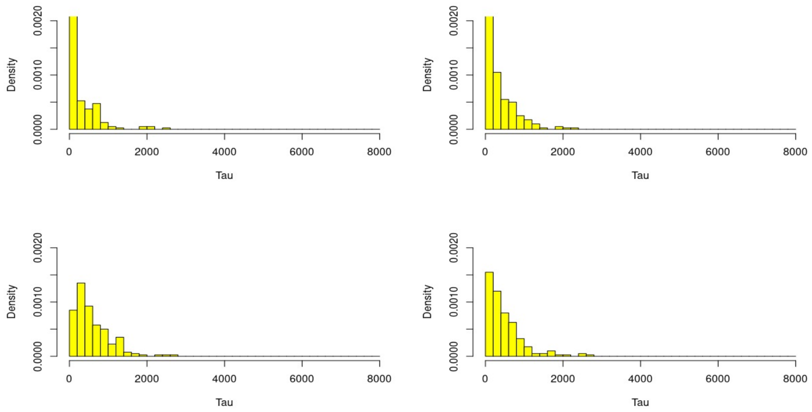

Figure 7: Histograms of Tau for $p_0 = 0.01, 0.1, 0.5, 0.8$ and $2N = 500$.

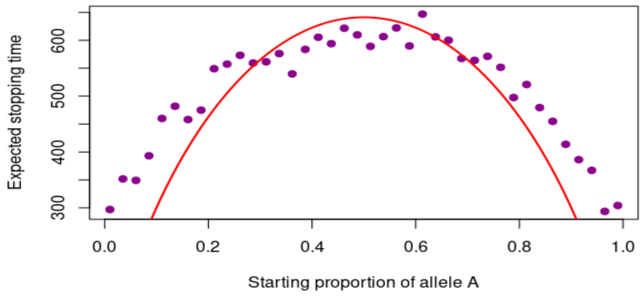

Figure 8: Expected fixation time vs $p_0 \in (0, 1)$. Population size was fixed to be 500. The red line is the graph of $f(p) = -2N\{p\log(p) + (1-p)\log(1-p)\}$. So, simulation also supports the theoretical result.

4 R codes and references

All study materials and R codes used for this project are available at this GitHub repository: https://github.com/Sohamdas-stat/Stochastic-Processes-2020

References

[1] Introduction to the Wright-Fisher Model by Joe Marcus, 2016-03-29.
Link: https://stephens999.github.io/fiveMinuteStats/ wright_fisher_model.html

[2] Wright-Fisher model and genetic drift
Link: https://math.la.asu.edu/ jtaylor/teaching/Spring2015/APM504/ lectures/Wright-Fisher_model.pdf

[3] Moran model
Link: https://math.la.asu.edu/ jtaylor/teaching/Spring2015/APM504/lectures/Moran.pdf

[4] Moran model
Link: https://en.wikipedia.org/wiki/Moran_process