

[Pre-requisites](#)[Overview](#)[Definition](#)[Examples](#)

Introduction to the Wright-Fisher Model

*Joe Marcus**2016-03-29* workflowr ✓

Pre-requisites

A basic knowledge of:

- introductory probability
- genetics terminology
- discrete-time Markov chains.

Overview

A major goal of population genetics is to understand the effect of various evolutionary forces, such as genetic drift, selection, mutation and migration, on the change in allele frequencies within a population. Named after early pioneers of theoretical population genetics, Sewall Wright (https://en.wikipedia.org/wiki/Sewall_Wright) and Ronald A. Fisher (https://en.wikipedia.org/wiki/Ronald_Fisher), the Wright-Fisher model (https://en.wikipedia.org/wiki/Genetic_drift#Wright.F2.80.93Fisher_mod) describes the sampling of alleles in a population with no selection, no mutation, no migration, non-overlapping generation times and random mating. Of course, real populations in nature do not adhere to these assumptions, yet the Wright-Fisher model provides a tool for studying how introducing more complex evolutionary forces can effect a relatively simple model.

Definition

Let A and a denote two alleles segregating at a locus in a given population. The Wright-Fisher model is a discrete-time Markov chain that describes the evolution of the count of one of these alleles over time. Let X_t be the count of the A allele in a population with

N diploid individuals at generation t . The state space of this Markov chain is the set of possible counts of the A allele i.e. $X_t \in \{0, 1, \dots, 2N\}$. Each generation, a collection of alleles are sampled, with replacement, from the current population at generation t to form a new population at generation $t + 1$. This process describes the binomial sampling of alleles each generation, which allows us to write the probability transition matrix for the Markov chain as:

$$P_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

In other words the probability of transitioning from an allele count of i , at generation $t - 1$, to an allele count of j , at generation t , can be computed from binomial probability mass function with size $2N$ and success probability being equal to the frequency of the A allele at time $t - 1$:

$$X_t \mid X_{t-1} = x_{t-1} \sim \text{Binomial}(n = 2N, p = \frac{x_{t-1}}{2N})$$

Properties

Mean

One useful property to know is the expected value of X_t at any time throughout the process. This can be computed using the fact that the expectation of a binomial distribution is np and the law of total expectation:

$$\begin{aligned} E(X_t) &= E(E(X_t \mid X_{t-1})) = E(2N \frac{X_{t-1}}{2N}) \\ &= E(X_{t-1}) = E(E(X_{t-1} \mid X_{t-2})) = E(2N \frac{X_{t-2}}{2N}) \\ &= E[X_{t-2}] = \dots = X_0 \end{aligned}$$

We see that the expected value of the Wright-Fisher process at any time-point is just the count of the A allele in the first generation. If Y_t is the frequency of the A allele at time t then the above result can be extended such that:

$$E(Y_t) = Y_0$$

Variance

It is a little bit harder to find the variance of X_t but it is useful to know and will be important for later tutorials. Recall from basic probability theory the definition of the variance:

$$\text{Var}(X_t) = E(X_t^2) - E(X_t)^2$$

We have shown about that $E(X_t) = X_0$ thus:

$$\text{Var}(X_t) = E(X_t^2) - X_0^2$$

Using the law of total variance and the mean and variance of the binomial distribution:

$$\begin{aligned} E(X_t^2) &= E(E(X_t^2 \mid X_{t-1})) = E(\text{Var}(X_t \mid X_{t-1}) + E(X_t \mid X_{t-1})^2) \\ &= E(2N \frac{X_{t-1}}{2N} (1 - \frac{X_{t-1}}{2N}) + (2N \frac{X_{t-1}}{2N})^2) \\ &= E(X_{t-1} - \frac{X_{t-1}^2}{2N} + X_{t-1}^2) = E(X_{t-1}^2 - \frac{X_{t-1}^2}{2N} + X_{t-1}) \\ &= E(X_{t-1}^2)(1 - \frac{1}{2N}) + E(X_{t-1}) \end{aligned}$$

thus we can rewrite $E(X_t^2)$ as:

$$\begin{aligned} \text{Var}(X_t) + X_0^2 &= (\text{Var}(X_{t-1}) + X_0^2)(1 - \frac{1}{2N}) + X_0 \\ &= \text{Var}(X_{t-1})(1 - \frac{1}{2N}) + X_0^2 - \frac{X_0^2}{2N} + X_0 \\ \Rightarrow \text{Var}(X_t) &= \text{Var}(X_{t-1})(1 - \frac{1}{2N}) + X_0(1 - \frac{X_0}{2N}) \end{aligned}$$

solving the above recurrence results in:

$$\text{Var}(X_t) = 2N X_0 (1 - \frac{X_0}{2N}) (1 - (1 - \frac{1}{2N})^t)$$

and finally in terms of the frequency of the A allele:

$$\text{Var}(Y_t) = Y_0(1 - Y_0)(1 - (1 - \frac{1}{2N})^t)$$

Fixation probabilities

Using the above moments we can compute some interesting features of the Markov chain which also have important biological relevance. Recall that absorbing states of a Markov chain are the states, once entered, can never be exited. More formally, if X_t^* is an absorbing state then:

$$P_{\{i=X_t^*\}j} = 0$$

In the Wright-Fisher model we can intuitively see that the fixation or loss of an allele are absorbing states i.e. if all of the individuals in the population carry two copies of the A allele or a allele the allele of the other type cannot be sampled without mutation or migration. The absorbing states in the Wright-Fisher model are $X_t^* \in \{0, 2N\}$. We can compute the probability of fixation of the A allele, transitioning to the absorbing state $X_{t+1}^* = 2N$ from any other state, using the conditional expectation described above:

$$E(X_{t+1}^* \mid X_t = i) = 2NP(X_{t+1} = 2N \mid X_t = i)$$

$$\Rightarrow i = 2NP(X_{t+1} = 2N \mid X_t = i)$$

$$\Rightarrow P(X_{t+1} = 2N \mid X_t = i) = \frac{i}{2N}$$

Therefore we can reciprocally write the probability of losing the A allele (fixation of the a allele):

$$P(X_{t+1} = 0 \mid X_t = i) = 1 - P(X_{t+1} = 2N \mid X_t = i) = 1 - \frac{i}{2N}$$

We can see that the probability of fixation or loss of an allele in a pure drift Wright-Fisher model only depends on the previous count of the allele and the effective population size.

Examples

Visualizing P_{ij}

Visualization of the probability transition matrix can provide some intuition on the how the process can proceed given different starting points.

```

library(ggplot2)
library(dplyr)
library(tidyr)
library(viridis)

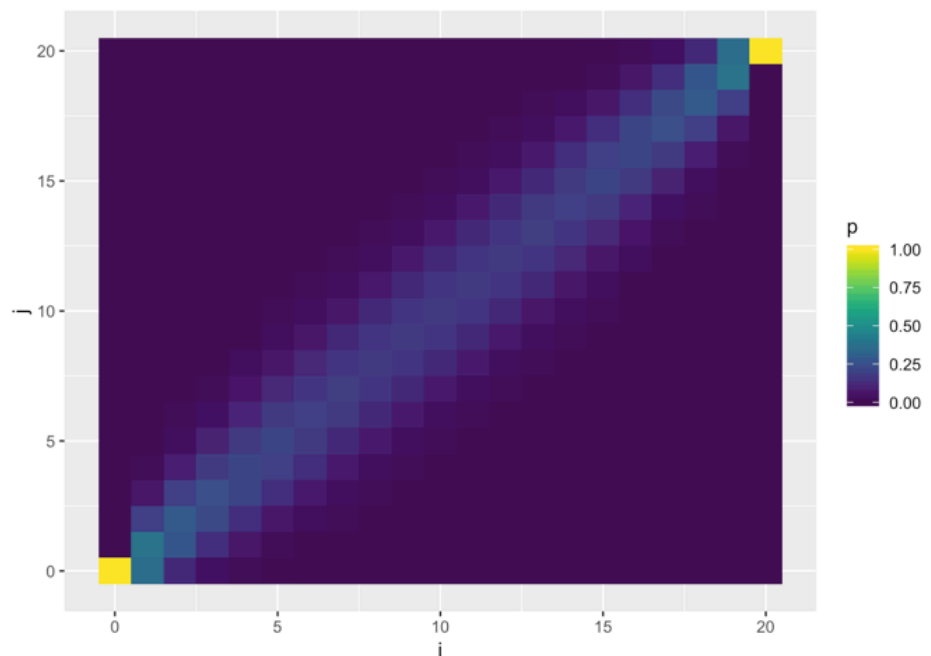
N <- 10 # effective population size

# fill up the probability transition matrix
P <- matrix(NA, ncol = 2*N + 1, 2*N + 1)
P_df <- data.frame()
for(i in 0:(2*N)){
  for(j in 0:(2*N)){
    # add to matrix for examples
    P[i+1, j+1] <- dbinom(j, 2*N, (i / (2*N)))

    # add to data.frame for viz below
    P_df <- bind_rows(P_df, data.frame(i = i, j = j
, p = P[i+1, j+1]))
  }
}

# plot it up!
p <- ggplot(P_df, aes(x = i, y = j, fill = p)) +
  geom_tile() + scale_fill_viridis(option="D")
p

```



Past versions of p_matrix-1.png

Simulating the Wright-Fisher Model

Lets simulate a grid of parameter values to explore the concepts introduced above:

```
# data.frame to be filled
wf_df <- data.frame()

# effective population sizes
sizes <- c(50, 100, 1000, 5000)

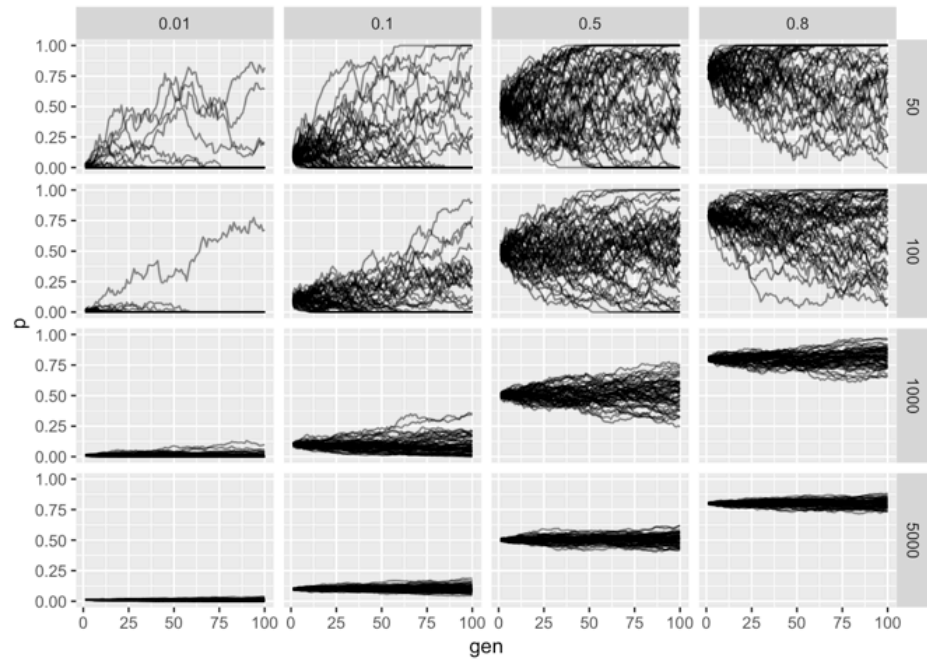
# starting allele frequencies
starting_p <- c(.01, .1, .5, .8)

# number of generations
n_gen <- 100

# number of replicates per simulation
n_reps <- 50


# run the simulations
for(N in sizes){
  for(p in starting_p){
    p0 <- p
    for(j in 1:n_gen){
      X <- rbinom(n_reps, 2*N, p)
      p <- X / (2*N)
      rows <- data.frame(replicate = 1:n_reps, N =
rep(N, n_reps),
                        gen = rep(j, n_reps), p0 =
rep(p0, n_reps),
                        p = p)
      wf_df <- bind_rows(wf_df, rows)
    }
  }
}

# plot it up!
p <- ggplot(wf_df, aes(x = gen, y = p, group = repl
icate)) +
  geom_path(alpha = .5) + facet_grid(N ~ p0) + g
uides(colour=FALSE)
p
```



Past versions of wf_simulation-1.png

The horizontal facet is the starting frequency and the vertical facet is the effective population size. We see that as we increase the population size the less noisy the allele frequency trajectory as expected from our derivation. Additionally we see that probability of fixation or loss is more likely when lower and higher starting allele frequencies (again as expected).

 Session information

This site was created with R Markdown
(<http://rmarkdown.rstudio.com>)

