Evolutionary Dynamics: Homework 05

Guido Putignano, Lorenzo Tarricone, Gavriel Hannuna,
Athanasia Sapountzi

February 22, 2024

Problem 1: Pathways of Carcinogenesis

Consider three independent mutations $\{1, 2, 3\}$. Each mutation occurs after an exponentially distributed waiting time $T_i \sim \exp(\lambda_i)$, $i = 1, 2, 3$.

(a) What is the probability for the path $P = 3 \to 1 \to 2$? (1 point)

Solution

Since the mutations are independent the probability of the pathway $P = 3 \to 1 \to 2$ is:

$$\begin{aligned} Prob(P) &= \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \cdot \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{\lambda_2}{\lambda_2} \\ &= \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \cdot \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned}$$

(b) Assume cancer arises if any two of the three genes are mutated. How many possible genotypes are there? How many pathways? Compute the expected waiting time until any two out of three genes are mutated. (1 point)

Solution

The number of possible genotypes with two mutated genes can be calculated using the binomial coefficient, which represents combinations (we want to compute combinations since for genotypes we do not take into consideration the order).

Thus we calculate:

$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$

So there are 3 possible genotypes.

Due to the fact that either of the two mutations may occur first, any of these genotypes are attainable through two distinct pathways. So, there are $3 \cdot 2 = 6$ pathways.

The expected waiting time, until any two out of three genes are mutated, is:

$$
\begin{aligned}
E(\tau_2) &= \sum_P \text{Prob}(P) \cdot E(\tau_P) \\
&= \frac{\lambda_1 \cdot \lambda_2}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3}\right) \\
&\quad + \frac{\lambda_1 \cdot \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3}\right) \\
&\quad + \frac{\lambda_2 \cdot \lambda_1}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3}\right) \\
&\quad + \frac{\lambda_2 \cdot \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3}\right) \\
&\quad + \frac{\lambda_3 \cdot \lambda_1}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_2)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right) \\
&\quad + \frac{\lambda_3 \cdot \lambda_2}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_2)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right) \\
&= \frac{\lambda_1 \cdot \lambda_2 + \lambda_1 \cdot \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3}\right) \\
&\quad + \frac{\lambda_2 \cdot \lambda_1 + \lambda_2 \cdot \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3}\right) \\
&\quad + \frac{\lambda_3 \cdot \lambda_1 + \lambda_3 \cdot \lambda_2}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_2)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right) \\
&= \frac{\lambda_1 \cdot (\lambda_2 + \lambda_3)}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3}\right) \\
&\quad + \frac{\lambda_2 \cdot (\lambda_1 + \lambda_3)}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3}\right) \\
&\quad + \frac{\lambda_3 \cdot (\lambda_1 + \lambda_2)}{(\lambda_1+\lambda_2+\lambda_3)\cdot(\lambda_1+\lambda_2)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right) \\
&= \frac{\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_2+\lambda_3}\right) \\
&\quad + \frac{\lambda_2}{(\lambda_1+\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_3}\right) \\
&\quad + \frac{\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)} \cdot \left(\frac{1}{\lambda_1+\lambda_2+\lambda_3} + \frac{1}{\lambda_1+\lambda_2}\right)
\end{aligned}
$$

(c) Now consider d independent mutations. How many paths exist leading to the genotype $\{1, \ldots, d\}$ with all mutations present? If cancer already arises after any k mutations, how many different paths are there?

Solution

We have d independent mutations, there are $d! = d \cdot ((d-1) \cdot \ldots \cdot 1)$ paths that lead to the genotype $\{1, \ldots, d\}$ with all mutations present, since at each step $t = 1, \ldots, d$ there are $(d - t + 1)$ ways to choose the next mutation.

If cancer already arises after any k mutations, there are $\binom{d}{k} = \frac{d!(d-k)!}{k!}$ possible

genotypes.

As we previously stated, there are there are $k! = k \cdot ((k-1) \cdot \ldots \cdot 1)$ paths that lead to the genotype $\{1, \ldots, k\}$.

Thus we calculate: $\binom{d}{k} \cdot k! = \frac{d!}{k!(d-k)!} \cdot k! = \frac{d!}{(d-k)!}$ paths.

There are $\frac{d!}{(d-k)!}$ different paths.

Problem 2: Neutral Wright-Fisher Process

Consider the neutral Wright-Fisher process for a system of N cells of two different types $\{A, B\}$. Let $X(t)$ denote the number of A-cells at time t. The process has the transition matrix

$$P_{i,j} = \text{Prob}[X(t) = j | X(t-1) = i] = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j},$$

that is, $X(t)|X(t-1) = i$ is binomially distributed with parameter $p = \frac{i}{N}$.

(a) Compute the conditional expectation $E[X(t)|X(0) = i]$ (tutorial discussion).

(b) Compute the conditional variance $Var[X(t)|X(0) = i]$ (1 point). Hint: Show that

$$Var[X(t)|X(0) = i] = V1 + \left(1 - \frac{1}{N}\right) Var[X(t-1)|X(0) = i],$$

where $V1 = Var[X(1)|X(0) = i]$. You can then use the expression above to derive the final result (no explicit calculation is needed for this last step).

Solution

We can use the recursive formula that we found above to calculate the $Var[X(2)|X(0) = i]$ as follows:

$$Var[X(2)|X(0) = i] = V1 + \left(1 - \frac{1}{N}\right) V1 = V1 \left(1 + 1 - \frac{1}{N}\right)$$

Let us remember that V1 can be calculated knowing that X(1) is a binomial distribution.

$$V1 = Npq = N\frac{i}{N}(1 - \frac{1}{N}) = i\left(1 - \frac{i}{N}\right)$$

So the variance at X(2) will be

$$Var[X(2)|X(0) = i] = i\left(1 - \frac{i}{N}\right)\left(1 + 1 - \frac{1}{N}\right)$$

$$Var[X(2)|X(0) = i] = Ni\,(1-p)\left(\frac{2}{N} - \frac{1}{N^2}\right)$$

$$Var[X(2)|X(0) = i] = Ni\,(1-p)\left(1 - (1 - \frac{1}{N})^2\right)$$

Now let us calculate the variance at X(3) and check if it follows the same pattern.

$$Var[X(3)|X(0)=i] = V1 + \left(1-\frac{1}{N}\right)V2$$

$$Var[X(3)|X(0)=i] = V1\left(1+\left(1-\frac{1}{N}\right)\left(2-\frac{1}{N}\right)\right)$$

$$Var[X(3)|X(0)=i] = i\,(1-p)\left(3-\frac{3}{N}-\frac{1}{N^2}\right)$$

$$Var[X(3)|X(0)=i] = Ni\,(1-p)\left(1-(1-\frac{1}{N})^3\right)$$

We found the same pattern also at t = 3, a pattern that will repeat itself and can be described by the following general formula

$$Var[X(t)|X(0)=i] = Ni\,(1-p)\left(1-(1-\frac{1}{N})^t\right)$$

(c) Derive an approximation for $Var[X(t)|X(0)=i]$ for large population size N. Compare the variance of the Wright-Fisher process to the variance of the Moran process, and explain the difference(s).

Solution

For large population size N, $\left(1-\frac{1}{N}\right)^t = 1-\frac{t}{N}$

$$\mathrm{Var}[X(t)\mid X(0)=i] = Ni\,(1-p)\left(1-(1-\frac{1}{N})^t\right) = Ni(1-p)\frac{t}{N} = V_1 t$$

In the Moran process, $\mathrm{Var}[X(t)\mid X(0)=i] \approx \frac{2}{N}ti\left(1-\frac{i}{N}\right)$, which is smaller than that in Wright-Fisher process. Because in one generation of the Wright-Fisher process, we actually process N generations of the Moran process.

This solution can be reached also through the Central Limit Theorem (CLT), assuming that $N \to \infty$ and there is an independent and identical distribution (i.i.d.) we can apply the CLT so that the binomial distribution can be approximated to a normal one, with the variance being

$$Var[X(t)|X(0)=i] = tnpq = tN\frac{i}{N}(1-\frac{i}{N}) = tV_1$$

The Moran process includes competition and selection, which affect allele frequencies differently, while Wright-Fisher mainly accounts for random genetic drift, and most importantly, it represents the sampling of a population. The formula for the variance in a Moran process depends on the population size (N), the initial allele frequencies, and the selection coefficient (s) for one of the alleles. The Variance for a Moran process can be calculated in this way.

$$Var[X(t)|X(0)=i] = \frac{2i}{N}\left(1-\frac{i}{N}\right)\frac{1-\left(1-\frac{2}{N^2}\right)^t}{\frac{2}{N^2}}$$

$$Var[X(t)|X(0)=i] = Ni\left(1-\frac{i}{N}\right)\left(1-\left(1-\frac{2}{N^2}\right)^t\right)$$

As we can see, the variance for a Moran process is quite similar to the variance of Wright-Fisher, other for the part dependent on t.

(d) Show that in the Wright-Fisher process, the heterozygosity H_t at time t satisfies (1 point)

$$E[H_t|X_0=i] = H_0(i)\left(1-\frac{1}{N}\right)^t$$

and hence decreases exponentially at rate $1/N$. Compare this behavior with the Moran model. Note: Heterozygosity in this context is defined as the probability that two individuals chosen at random from the population are of different types.

Solution

To show that $E[H_t|X_0=i]$ corresponds to the equation stated above we can start analyzing the exact definition of Heterozygosity $H_t = X(t)/N(1-X(t)/N)$, meaning the chance of drawing 1 individual of one population and the second of the other population. At t = 0 $H_0 = p(1-p)$. The expectation of H_t can be written as:

$$E[H_t|X_0=i] = E[X(t)/N(1-X(t)/N)|X_0=i] = \frac{E[X(t)]}{N} - \frac{E[X(t)^2]}{N^2}$$

Now we must remember two equations, the first $E[X(t)] = i$ and the second is that $Var[X(t)] = E[X^2] - E[X]^2$. We can use the last one to find the value of $E[X^2]$ based on the previous equation that described the variance of a Wright-Fisher process.

$$E[X^2] = Ni(1-p)\left(1-\left(1-\frac{1}{N}\right)^t\right) + i^2$$

$$= (Ni-i^2)\left(1-\left(1-\frac{1}{N}\right)^t\right) + i^2$$

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

So finally $E[H_t|X_0=i]$ will be

$$E[H_t] = \frac{i}{N} - \frac{1}{N^2}\left[Ni - (Ni-i^2)\left(1-\frac{1}{N}\right)^t\right]$$

$$= (i/N - (i/N)^2)\left(1-\frac{1}{N}\right)^t = p(1-p)\left(1-\frac{1}{N}\right)^t = H_0(i)\left(1-\frac{1}{N}\right)^t$$

For the Moran model, similar to the variance, the expectation of heterozygosity will decrease by a factor of $2/N^2$, instead of $1/N$

$$E[H_t] = H_0(i)\left(1 - \frac{2}{N^2}\right)^t$$

(e) Simulate the Wright-Fisher process. Compute the empirical mean and variance and compare them with your analytical results. Use $n = 100$ simulations with population sizes of $N \in \{10, 100\}$, respectively, and $X(0) = N/2$.

Solution

The following are two plots representing the simulations of the Wright-Fisher process with N = 10 and N = 100. The black dots in the figures represent the simulated trajectories.

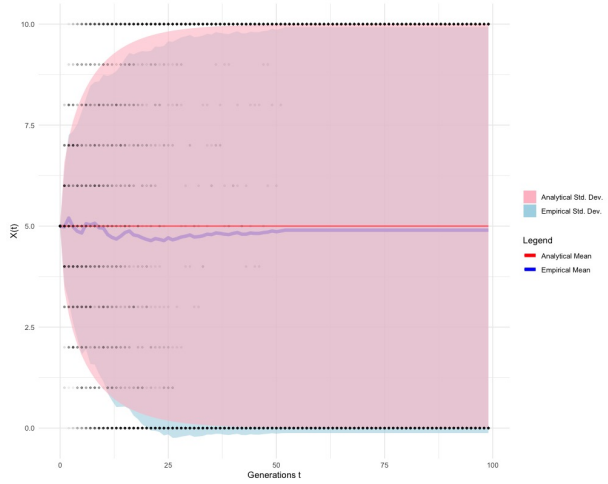

Figure 1: 100 simulations with N = 10 of the Wright-Fisher process

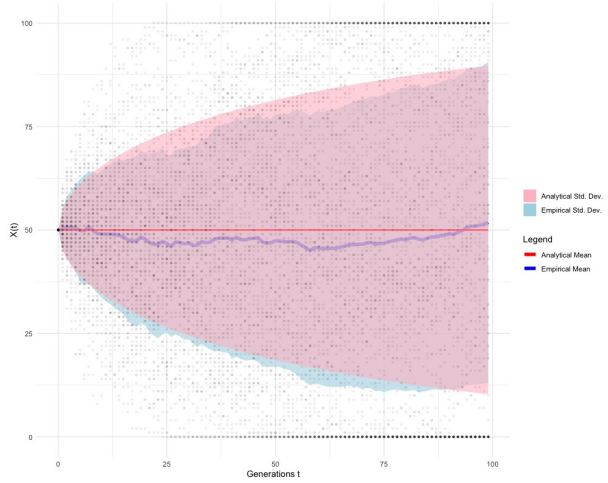

Figure 2: 100 simulations with N = 100 of the Wright-Fisher process

Problem 3: Wave Approximation

Consider the wave approximation of the Wright-Fisher model for cancer progression. Here, the growth of a clone with j mutations is given by

$$\dot{x}_j = s x_j (j - < j >)$$

For small times, the average fitness $s < j >= s \sum_j j x_j$ can be considered constant. Use this throughout your calculations.

(a) Find the analytic solution for the initial condition $x_j(0) = \frac{1}{N}$. (1 point)

Solution: Given that the factor $s < j >$ (and therefore also $s(j - < j >)$) is constant, we just need to solve a first-order linear and homogenous ODE of the kind $x' = \alpha x$ this has a well-known solution (obtainable for example by separation of variables) of $x(t) = A_0 e^{\alpha t}$ where the value A_0 needs to be determined with some initial condition. For us, the general solution is $x_j(t) = A_0 e^{s(j-<j>)t}$. We are given as initial condition $x_j(0) = \frac{1}{N}$. substituting this in our general solution yields $x_j(0) = \frac{1}{N} = A_0 e^{s(j-<j>)0} = A_0$ The final analytic solution will therefore be $x_j(t) = \frac{1}{n} e^{s(j-<j>)t}$

(b) The rate at which an additional mutation occurs is given by $u d x_j(t)$. Find the time τ when the cumulative probability exceeds $\frac{1}{N}$. (1 point)

Solution What is probably required in this exercise is to calculate when the nonhomogenous rate exceeds $\frac{1}{N}$. This is the case if:

$$\frac{1}{N} = ud \int_0^\tau x_j(t) dt = \frac{ud}{N} \int_0^\tau e^{s(j-<j>)t} dt \iff$$

$$1 = \frac{ud}{s(j - < j >)} \left[e^{s(j-<j>)t} \right]_0^\tau = \frac{ud}{s(j - < j >)} (e^{s(j-<j>)\tau} - 1) \iff$$

$$\frac{s(j - < j >)}{ud} + 1 = e^{s(j-<j>)\tau} \iff \tau = \frac{ln(\frac{s(j-<j>)}{ud})}{s(j - < j >)}$$

If we want to find a time at which the probability of any mutation occurs in cell j, then we would have $\frac{1}{N} = 1 - e^{-ud \int_0^\tau x_j(t)}$ and we would need to solve for τ...

(c) Compute the waiting time until the next mutation for a mutation rate $u = 10^{-7}$/cell generation, $d = 80$ genes, and a fitness advantage of $s = 1.15\%$ per mutation. Use that $j - < j > \approx \sqrt{\log N}$ with $N = 10^7$ cells and assume a cell generation time of 1 day.

Solution: What we calculated at point b is exactly the desired waiting time as the presence of a new mutant in a discrete population of N individuals is $\frac{1}{N}$. Now substituting the number and the approximation given we get:

$$\tau = \frac{ln(\frac{s\sqrt{lnN}}{ud})}{s\sqrt{lnN})} = \frac{ln(\frac{0.0115\sqrt{ln10^7}}{10^{-780}})}{0.0115\sqrt{ln10^7})} \approx 187.58 \; days$$