Stockholms universitet

An analysis of beneficially mutated alleles

Sepehr Zolfeghari

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2022:10
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2022:10**
http://www.math.su.se

An analysis of beneficially mutated alleles

Sepehr Zolfeghari[*]

June 2022

Abstract

In this thesis the aim is to get a mathematical understanding of how individuals of varying reproductive fitness propagate through time. We analyze this using the Wright-Fisher model where every generation is of the same population size. Using this model we assume at first that the whole population is equally fit and obtain some interesting results. Thereafter we assume that the population can be divided into two sets where one is more fit in a reproductive sense than the other. In other words the more fit individuals carry mutated alleles and are therefore better adapted to their environment. From this we get the mathematical results needed to find the expected number of mutants for any given generation. At last we briefly discuss a growing population size.

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sepzoli@yahoo.com. Supervisor: Pieter Trapman.

Acknowledgments

I would like to give my sincerest gratitude to my supervisor Pieter Trapman for all his patience and guidance. His wealth of knowledge and understanding for the subject at hand has really enlightened me. I would also like to give my deepest appreciation to my fellow classmates and dear friends Tom Pedersen and Ossian Hätterstrand for their continued support and encouragement.

Contents

1 Introduction

The aim of this thesis is to get an understanding of the mathematics of how a population of individuals propagate through time. It is of interest to see how this population grows given that there are different individuals of varying degrees of reproductive fitness. For the size and time given for this thesis we will only consider two types of individuals, where one class of them are more fit in a reproductive sense, i.e. the mutants, and will have more of their children represented every generation.

In Section 1.1 and 1.2 we introduce the basics of genetics and the mathematics of reproductive fitness. We then go on to define the Wright-Fisher model in Section 2 which is the model we will use to explain how a population grows through time. In Section 2 it is assumed that we have a stable population for every generation. Before we go on and explore the model using two classes of individuals of different reproductive fitness we assume that they are equally fit. Doing so reveals some interesting results such as the coalescent theory, which is how one can calculate the different expected time of when two genetic lineages converge in generational time.

We then go on in Section 4 to add the assumption that there are individuals of different reproductive fitness in our population. We derive the mathematics of assuming this which will be used going on to Section 5, where we answer at what generation we can expect that the more fit individuals have out-competed their less fit counterparts.

At last we go briefly in Section 6 into what happens if we assume a growing population size for every generation.

1.1 Basic genetics

In order to study how species propagate throughout time we need to have a basic understanding of genetics.

Four molecules called nucleotides are the building blocks of DNA. These molecules are adenine, guanine, cytosine, and thymine. Sequences of DNA are called genes. Every organism has genomes which are a complete set of genetic information (Pierce 2012, p. 4). In a genome of an organism the genetic locus is a location on the said genome (Durrett 2008, p. 5). The genes that code for phenotypic traits are called alleles. An example of this is the genes that code for coat colors in cats (Pierce 2012, p. 11). The coat color can come in different colors such as black, orange or white. Eye colors in humans are also an example of different types of alleles.

It is important to distinguish between traits and genes. It is the genes that are passed on from generation to generation and the genes are responsible for certain traits that are formed by environmental factors (Pierce 2012, pp. 11-12). The phenotype of an organism is a trait encoded in the allele.

There are diploid and haploid organisms. Haploid organisms have only one copy of their genetic material while diploid organisms have two copies. Each copy in a diploid organism is given by each parent respectively (Durrett 2008, p. 4). The copies are given by chromosomes, which are a bundle of genes. Each copy of the chromosome is usually alike in structure and size and fill the same function. What differs in the two chromosomes are their alleles. One chromosome might for example code for black hair color while the other codes for blonde hair color. However they fill the same function, in this case hair color (Pierce 2012, p. 19).

1.2 Reproductive fitness

In order to study genetic fitness there needs to be a distinction between individuals. It is of interest to consider one individual more fit than another. Usually the fitness of an individual is determined by the differences in phenotypic traits. For example a white coated polar bear has a better survival advantage in a snow ridden climate than a brown coated polar bear, and is therefore more fit. As stated above these phenotypic differences are encoded in the allele.

To make this more general in this thesis we will consider two types of alleles. Allele of type A and a. One allele is more advantageous in a reproductive sense, in this thesis it is always assumed to be an allele of type a. Mathematically we can assume that an individual X that carries an allele of type A has a Poisson amount of children, such that

$$X \sim Poisson(\lambda),$$

where $\lambda \geq 1$ in order to have an expected stable population. In other words X is the distribution of offspring. An individual Y that carries an allele of type a is then said to have

$$Y \sim Poisson(\lambda s),$$

where $s > 1$ is the beneficial mutation factor. The expected amount of children that individual Y has is $E[Y] = \lambda s$ which is larger that the expected amount of children X has which is just λ (Alm and Britton 2008, p. 87).

This means that individual Y is likely to give birth to more children than X. This in of itself does not translate necessarily to reproductive fitness so an equivalent way of thinking about this is that individual Y will have more children that survive than individual X. The fact that the probability that more children of allele type a are represented in a population is an indication that that specific phenotype has an evolutionary advantage over the phenotype that is expressed by allele of type A.

2 The Wright-Fisher model

In the Wright-Fisher model it is assumed that we have an environment for which there are always $2N$ individuals in each generation where N is a natural number. In our specific case each generation n are made up of m individuals of allele type a, where $n \geq 1$ and $2N \geq m \geq 1$. Given this there are $2N - m$ individuals of allele type A, such that the sum of all individuals is $2N$.

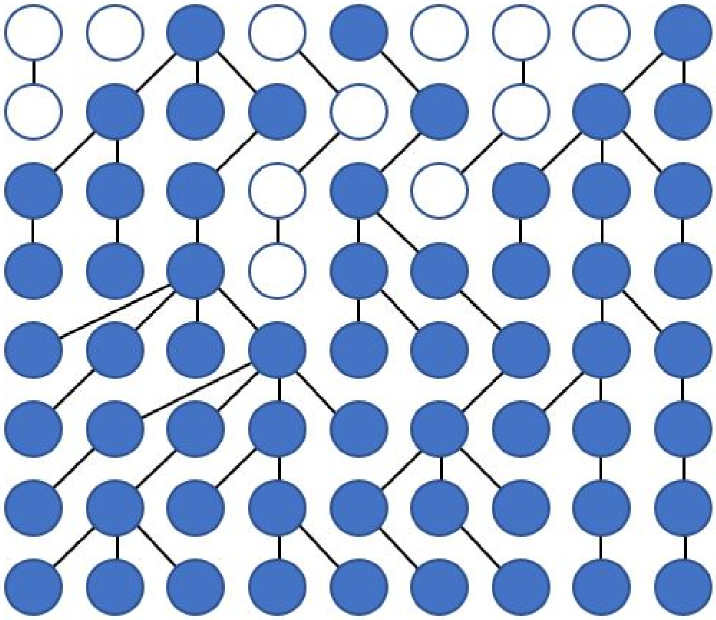

Figure 1. A demonstration of the Wright-Fisher model.

2.1 Clustering allele types

Every individual i where $1 \leq i \leq 2N$ in generation n can give birth to a stochastic amount of children such that

$$X_{n,i} \sim Poisson(\lambda_{n,i}). \tag{1}$$

In what we are modeling in this thesis every $\lambda_{n,i}$ assumes either λ or λs.

One of the neat properties of the Poisson distribution is that the sum of independent Poisson distributed variables is also a Poisson distributed variable.

Therefore in generation n all individuals i of allele type a will in total give birth to a Poisson amount of individuals such that

$$\sum_{i=1}^{m} X_i = X_a \sim Poisson(\lambda sm). \tag{2}$$

In the same way all individuals of allele type A will in total give birth to random amount of children such that

$$\sum_{i=m+1}^{2N} X_i = X_A \sim Poisson(\lambda(2N-m)) \tag{3}$$

(Held and Bové 2020, p. 360). Another useful result is that the total amount of children that can be born is

$$\begin{aligned} X_a + X_A \sim & Poisson(\lambda(2N-m) + \lambda sm) \\ = & Poisson(\lambda(2N-m+sm)). \end{aligned} \tag{4}$$

Using these results will greatly simplify a lot of calculations.

2.2 The auxiliary step

As stated in (4) the total amount of children that will be born is $Poisson(\lambda(2N-m+sm))$ distributed with an expected amount being $\lambda(2N-m+sm)$ (Alm and Britton 2008, p. 87). However in the Wright-Fisher model we condition on there being $2N$ individuals every generation. This will lead to the assumption that

$$E[X_a + X_A] = \lambda(2N-m+sm) \geq 2N \tag{5}$$

for any given generation n. Therefore it is expected that λ has been chosen which would make (5) true. The reason for this is that we can not have the probability of having less children born be smaller than $2N$, for then the Wright-Fisher model would no longer hold.

The second point is that the amount of children born is $X_a + X_A$, which could very well be larger than $2N$. In fact as stated by (5) it is in expectation. In-between every generation n and $n+1$ there is an auxiliary step n_a where there are $X_a + X_A$ individuals. These individuals are the actual children born from generation n. We then choose $2N$ uniformly at random from n_a which then will become the individuals that make up generation $n+1$. In practice this corresponds to there being some amount of children born in n_a but because of environmental factors some of them die off leaving only $2N$ to survive.

3 Equal fitness

Assume that individuals with allele type A and a have the same reproductive fitness, meaning that $s = 1$ and all individuals i in generation n such that

$1 \leq i \leq 2N$ and $1 \leq n$ can give birth to a stochastic amount of children such that

$$X_{n,i} \sim Poisson(\lambda).$$

Every individual has therefore the same expected amount of children, $E[X_{n,i}] = \lambda$. Using result (5) also gives us that

$$E[X_a + X_A] = \lambda(2N - m + m) = \lambda 2N.$$

3.1 Children choosing their parents

In this specific case we can circumvent the process of going from the auxiliary step n_a to $n+1$ by having the individuals from generation $n+1$ choose uniformly at random their parents from generation n. The reason they do not choose from n_a and instead choose directly from n is because all the individuals in generation n are equally and independently distributed.

Another result of thinking this way is that the probability that an individual i in generation n is the parent of k individuals in generation $n+1$ is given by

$$P(X_{n,i} = k) = \binom{2N}{k}\Big(\frac{1}{2N}\Big)^k\Big(1 - \frac{1}{2N}\Big)^{2N-k}. \tag{6}$$

This is just the probability of individual i choosing k individuals in generation $n+1$ where the probability of success, a child choosing the one right parent is $\frac{1}{2N}$, and there are $\binom{2N}{k}$ ways of choosing k children from generation n. Yet here again we circumvent n_a. Note that the number of children of individual i is $Bin(2N, \frac{1}{2N})$ distributed. This is not the same as the amount of children that individual i will give birth to which is $Poisson(\lambda)$.

A more extensive way of looking at it is by regarding at generation n a subset S of $2N$ such that $|S| = L$ and $1 \leq L \leq 2N$. The probability that all individuals in the subset S gives birth to k children in total is given by

$$P\Big(\sum_{i \in L} X_{n,i} = k\Big) = \binom{2N}{k}\Big(\frac{L}{2N}\Big)^k\Big(1 - \frac{L}{2N}\Big)^{2N-k}. \tag{7}$$

Note that in the subset S some may birth no children and the others might birth all the K children in total. This is analogous to (6) where the difference is that the probability of success is $\frac{L}{2N}$, given by the fact that a child chooses uniformly at random from a population of $2N$. The probability of choosing one right parent from the subset S is just $\frac{L}{2N}$. Therefore the amount of children is $Bin(2N, \frac{L}{2N})$ distributed.

However in our model we have only two alleles present and so we can forgo the analysis of individual parents and look only at how the two alleles A and a

propagate. In this case the probability of having j alleles of type A in the next generation when there are i present is said to be

$$p(i,j) = \binom{2N}{j} {p_i}^j (1-p_i)^{2N-j}, \tag{8}$$

where $p_i = \frac{i}{2N}$ (Durrett 2008, p. 6). It can be stated that we have a Markov Chain as each successive generation is only dependent on the previous state. In these terms it is not hard to see that we have two absorbing states, $X_T = 2N$ or $X_T = 0$ (Durrett 2008, p. 6).

3.2 Large population size

Using (7) and having a population $2N$ be very large, which is in practical terms a realistic assumption, reveals that as $2N$ goes to infinity we get that $Bin(2N, \frac{L}{2N}) \to Poisson(L)$ in distribution (Alm and Britton 2008, p. 171). The expected amount is therefore L children.

This result shows therefore that the expected amount of children that survives from one individual i, where $L = 1$ is 1.

In the same way (8) will give us that as $2N$ goes to infinity

$$p(i,j) \sim Poisson(i).$$

3.3 The coalescent theory

We will now under the assumption of equal fitness between individuals of allele type A and a study when two lineages coalesce in time, in other words at what generation does two individuals share a common common ancestor?

3.3.1 Two individuals

The probability that two individuals share the same parent one generation ago is

$$\frac{1}{2N},$$

meaning that there is a one hundred percent guarantee that allele 1 has a parent and the chances that the second allele shares the same parent is $\frac{1}{2N}$. Given this we can calculate the probability of when two alleles coalesce t generations back by

$$\left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right),$$

which we observe as a geometric distribution (Nordborg 2000, pp. 5-6). It can be explained by the fact that for each generation, the probability that the two alleles in question do not share a parent is

$$\left(1 - \frac{1}{2N}\right).$$

If they coalesced t generations back then they do not share the same parent for $t-1$ generations hence

$$\left(1 - \frac{1}{2N}\right)^{t-1}.$$

At last when they finally do share the same parent we multiply with the last factor

$$\left(\frac{1}{2N}\right).$$

Taking the expected value of an geometric distribution yields us $2N$, which is the expected number of generations back in which two alleles might coalesce (Held and Bové 2020, p. 359).

Note that we will not study when $k > 2$ individuals coalesce. The probability that k individuals share the same parent one generation ago is

$$\left(\frac{1}{2N}\right)^{k} = \frac{1}{(2N)^k}$$

which if $2N$ is considered very large will give an probability that is increasingly too small.

3.3.2 Two individuals from a sample of k individuals

So far we have examined the coalescence of sampling at random two individuals from the total $2N$ population. Let us say we want to instead find the coalescence of a sample of k and see how two individuals from our sample of k coalesce. We know from our previous results that the probability of coalescence one generation back for an arbitrary sample of two individuals is

$$\frac{1}{2N}.$$

Choosing two individuals from a sample of k can be done in

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

ways. This means that for each way there is a $\frac{1}{2N}$ chance that those two chosen individuals might coalesce one generation back. Hence the probability of two individuals from a sample of k coalescing one generation back is

$$\frac{\binom{k}{2}}{2N}.$$

In the same way the probability of them not coalescing t generations back and then finally coalescing is

$$\left(1-\frac{\binom{k}{2}}{2N}\right)^{t-1}\left(\frac{\binom{k}{2}}{2N}\right).$$

The expected time here as we yet again have a geometric distribution is

$$\frac{2N}{\binom{k}{2}} \tag{9}$$

(Held and Bové 2020, p. 359). Given this result we have the expected time of when two individuals coalesce from a sample of k. Once they do coalesce we have a new sample of $k-1$ left. We choose two from this population and repeat the process and get the expected time of coalesces using (9) to be

$$\frac{2N}{\binom{k-1}{2}}.$$

3.3.3 Most recent common ancestor

We can keep reapplying (9) and continue this process until we find the most recent common ancestor (Durrett 2008, p. 9). This gives us the relative proportions of how the expected time of different lineages coalescing looks like, where the branches are shorter signifying that they diverged not too long ago from each other and successively getting longer signifying that the expected time of divergence is further back in time.

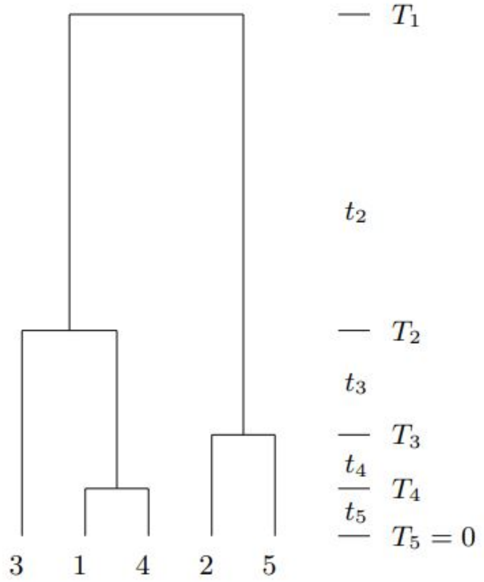

Figure 2. A realization of the coalescent for a sample of size 5 (Durrett 2008, p. 9).

To demonstrate this given a population of $2N$ and a sample size of 5 we can find the different expected generational times of when two individuals coalescence in generational time t_k, where k signifies how many individuals we have. The results are

$$E[t_5] = \frac{2N}{10}, E[t_4] = \frac{2N}{6}, E[t_3] = \frac{2N}{3}, E[t_2] = 2N$$

(Durrett 2008, p. 9).

If we sample n individuals then T_1 is the amount of time needed to get to the most common ancestor (Durrett page 9). In other words

$$T_1 = t_n + ... + t_2.$$

Thanks to the properties of expected values we can find the expected time needed to find the most recent common ancestor as finding the expected time of the sum $t_n + ... + t_2$ (Alm and Britton 2008, p. 120), we get that

$$E[T_1] = \frac{2N}{\binom{k}{2}} = 2N \cdot 2 \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 4N \cdot \left(1 - \frac{1}{n} \right)$$

(Durrett 2008, p. 9). We know that $E[t_2] = 2N$ implies that when n is sufficiently large $E[T_1]$ converges to $4N$. This means that half the expected time to find how a sample of k individuals coalesce is spent in the last coalescence t_2.

4 Mutations

Now we will re-implement a beneficial mutations into our model, i.e. when the beneficial mutation factor $s > 1$, and see how the individuals with the mutation will propagate throughout time.

4.1 The auxiliary step with mutations

As written in Section 2.2, result (5) shows that the expected amount of children born from generation n into the auxiliary step n_a is

$$\lambda(2N - m + sm).$$

We now have to find a way to uniformly choose from this auxiliary step so that the chosen individuals make up generation $n + 1$. The solution will utilize the Poisson process.

4.1.1 Part 1

Definition 1
Allan Gut states one of the definitions of a Poisson process as the following (Gut 2009, p. 222):

a) the increments $\{X(t_k) - X(t_{k-1}), 1 \leq k \leq n\}$ are independent random variable for all $0 \leq t_0 \leq t_1 \leq t_2 \leq ... \leq t_{n-1} \leq t_n$ and all n;

b) $X(0) = 0$ and there exists $\lambda > 0$ such that

$$X(t) - X(s) \in Po(\lambda(t - s)), \ for \ 0 \leq s < t.$$

The constant λ is called the intensity of the process.

For a given generation n and using (1) defined in Section 2.1, the stochastic amount of children individual i will give birth to is

$$X_i \sim Poisson(\lambda_i).$$

In our special case λ_i assumes either only λ or λs but there is no reasons not to have more than two types of λ_i corresponding to different levels of reproductive fitness. Note that

$$X_1, X_2, ..., X_{2N}$$

are all independent. Using Definition 1 we can find t_k such that

$$X_i = X(t_i) - X(t_{i-1}) \sim Poisson(\lambda(t_i - t_{i-1})) = Poisson(\lambda_i),$$

for $0 \leq t_i \leq 2N$.

What this implies is a Poisson process of intensity λ with a line divided into $2N$ parts corresponding to the $2N$ individuals. In this thesis the individual segment lengths $t_i - t_{i-1}$ of the line are either equal to s or 1. Each increment is an independent random variable which is Poisson distributed. This means that the arrival times in each increment is Poisson distributed with λ_i where the expected amount of arrivals is λ_i. (Alm and Britton 2008, p. 87).

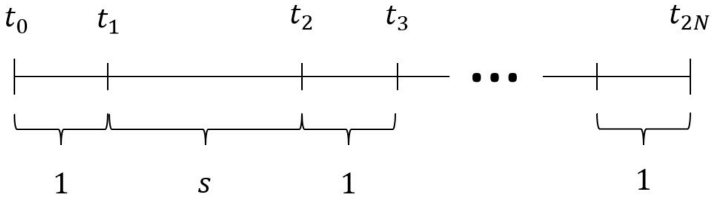

Figure 3. The Poisson process with $2N$ increments.

4.1.2 Part 2

Theorem 1
Sheldon M.Ross (p.327-328) defines the amount of events over an interval $(0, t)$ in a Poisson process as $N(t)$ (Ross 2019, pp. 327-328). Given that $N(t) = n$, the n arrival times of the n events $S_1, ..., S_n$ are uniformly distributed over the interval $(0, t)$. Note that the events are considered as unordered random variables.

In our specific case $t = \sum_{i=1}^{2N} \lambda_i$, which means our interval is the sum of all λ_i. The reason why this result is so useful is that instead of having an auxiliary generation n_a, the $2N$ individuals from generation $n+1$ can uniformly choose their parents at random from a line of length t. In this line each segment is proportional to the expected amount of children each individual i from generation n will have. Therefore we can circumvent the process of uniformly excluding individuals at random in the auxiliary step. Note that the expected amount of children for each parent is still λ_i even though they might have less or more children than that.

Assume that the sum of children that generation n has given birth to is

$$\sum_{i=1}^{2N} x_i = X_c,\ x_i \geq 0.$$

Mathematically we want to find

$$P(X_1 = x_1, X_2 = x_2, ..., X_{2N} = x_{2N} | X_c = 2N), \tag{10}$$

given that $X_i \sim Poisson(\lambda_i)$. This probability is equal to a multinomial distribution of character

$$Multinomial\left(2N, \frac{\lambda_1}{\sum_{i=1}^{2N}\lambda_i}, \frac{\lambda_2}{\sum_{i=1}^{2N}\lambda_i}, ..., \frac{\lambda_{2N}}{\sum_{i=1}^{2N}\lambda_i}\right),$$

(Agresti 2014, pp. 7-8). Note that it is possible for one or more individuals to not have any children survive to generation $n+1$. In that case their unique λ_i does not get represented the next generation, resulting in fewer overall different types of λ_i.

4.1.3 Part 3

As stated in Section 2.1 in (2) and (3) the total amount of children of allele type A born into the auxiliary step is X_A and the total amount of children born with allele type a into the auxiliary step is X_a. Because we are interested of how the beneficial mutation propagates, i.e. individuals of allele type a, leads us therefore to regard all individuals of the same allele types as one large organism and see how many children that organism will have represented into the next generation $n+1$. The only λ_i we therefore need to focus on is given by (2) and (3) as λsm and $\lambda(2N-m)$. Visually this would correspond to a Poisson process of intensity λ with the interval $(0,(2N-m)+sm)$ divided into two parts of length $2N-m$ and sm.

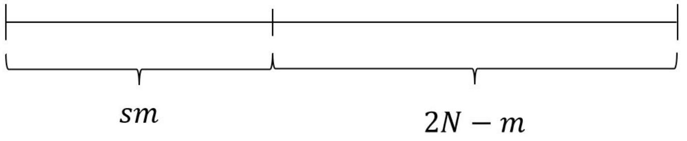

Figure 4. The Poisson process with 2 increments.

We can now therefore imagine having $2N$ points representing the individuals in generation $n+1$ and uniformly at random place them on this line. Depending on what interval they are placed into will constitute what allele they will carry.

Using (10) and applying it to our specific case yields

$$P(X_a = x_a, X_A = x_A | x_a + x_A = 2N) \tag{11}$$
$$\sim Multinomial\left(2N, \frac{\lambda sm}{\lambda sm + \lambda(2N-m)}, \frac{\lambda(2N-m)}{\lambda sm + \lambda(2N-m)}\right).$$

The probability density function of (11) is

$$\frac{2N}{x_a!x_A!}\left(\frac{\lambda sm}{\lambda sm + \lambda(2N-m)}\right)^{x_a}\left(\frac{\lambda(2N-m)}{\lambda sm + \lambda(2N-m)}\right)^{x_A}$$

(Alm and Britton 2008, p. 135). However because $x_a + x_A = 2N$ implies $x_A = 2N - x_a$ we get that

$$\frac{2N}{x_a!(2N-a)!}\left(\frac{\lambda sm}{\lambda sm + \lambda(2N-m)}\right)^{x_a}\left(1 - \frac{\lambda sm}{\lambda sm + \lambda(2N-m)}\right)^{2N-x_a} \tag{12}$$

$$= \binom{2N}{x_a}\left(\frac{sm}{sm + (2N-m)}\right)^{x_a}\left(1 - \frac{sm}{sm + (2N-m)}\right)^{2N-x_a}$$

which we see is the probability density function of an binomial distribution variable with $n = 2N$ and $p = \frac{sm}{sm+(2N-m)}$ (Alm and Britton 2008, p. 77).

Note that (12) implies that (11) is not dependent on the intensity λ of the underlying Poisson process. We can therefore see the λ as an scaling factor which can be arbitrarily chosen.

Define α_n as the number of individual with the beneficial mutated allele type a in generation n, such that $0 \leq \alpha_n \leq 2N$ for any $1 \leq n$. In conclusion the amount of individuals in generation $n+1$ with the beneficial allele type a given that there were m individuals of allele type a in generation n is given by

$$\alpha_{n+1}|\alpha_n = m \sim Bin\left(2N, \frac{sm}{sm + (2N-m)}\right). \tag{13}$$

4.2 The expected number of mutants

Using the results found in Section 4.1.3 and result (13) we get

$$E[\alpha_{n+1}|\alpha_n = m] = \frac{2Nsm}{sm + (2N-m)}, \tag{14}$$

which is the expected amount of individual with allele type a in generation $n+1$ given that there were m such individuals the previous generation (Alm and Britton 2008, p. 77). However if we assume that we do not know what value α_n takes then (14) becomes

$$E[\alpha_{n+1}|\alpha_n] = \frac{2Ns\alpha_n}{s\alpha_n + (2N-\alpha_n)}, \tag{15}$$

which is a function of the stochastic variable α_n. What we want to determine is $E[\alpha_{n+1}]$, which is the expected amount of individuals in generation $n+1$ with an allele type a unconditioned on generation n. It is also known that

$$E[\alpha_{n+1}] = E[E[\alpha_{n+1}|\alpha_n]] \tag{16}$$

(Gut 2009, p. 34). Therefore with the definition of expected value for a function we yield the following (Alm and Britton 2008, p. 59):

$$E[\alpha_n] = E[E[\alpha_{n+1}|\alpha_n]] = E\left[\frac{2Ns\alpha_n}{s\alpha_n + (2N - \alpha_n)}\right]$$

$$= \sum_{k=0}^{2N} \frac{2Nsk}{sk + (2N-k)} \binom{2N}{k} \left(\frac{s\alpha_{n-1}}{s\alpha_{n-1} + (2N - \alpha_{n-1})}\right)^k \left(1 - \frac{s\alpha_{n-1}}{s\alpha_{n-1} + (2N - \alpha_{n-1})}\right)^{2N-k}.$$

The sum above is therefore a function of α_{n-1}. The reason for this is that $\frac{2Ns\alpha_n}{s\alpha_n+(2N-\alpha_n)}$ is

$$Bin\left(2N, \frac{2Ns\alpha_{n-1}}{s\alpha_{n-1} + (2N - \alpha_{n-1})}\right)$$

distributed. Unless we know what value α_{n-1} has taken the expected value $E[\alpha_n]$ will be a function of α_{n-1}.

5 Expected convergence of mutants

It is now of relevance to find a numeric way of calculating at what generation the population is made up entirely of individuals carrying the beneficial mutated alleles. In other words at what n can we expect $\alpha_n = 2N$.

5.1 The initial stages

What we like to determine is $E[\alpha_n]$, which to clarify is the expected amount of alleles with a beneficial mutation in any generation. We would like to calculate it without having to condition on any previous generations. Using (15) from Section 4.2 we yield

$$E[\alpha_{n+1}|\alpha_n] = \frac{2Ns\alpha_n}{s\alpha_n + 2N - \alpha_n} = \frac{s\alpha_n}{\frac{s\alpha_n}{2N} + 1 - \frac{\alpha_n}{2N}}. \tag{17}$$

If $2N$ is considered really large which is a fair assumption to make as in practical applications the population size in question is very large, then we get the following result using (17):

$$\lim_{2N\to\infty} E[\alpha_{n+1}|\alpha_n] = s\alpha_n.$$

Note however in order for this to be true α_n has to be of a much smaller order then $2N$ in order for the quotient $\frac{\alpha_n}{2N}$ to converge to 0. If that is true we can approximate the expected value as

$$E[\alpha_{n+1}|\alpha_n] \approx s\alpha_n. \tag{18}$$

Applying (18) to (16) gives us

$$\begin{aligned} E[\alpha_{n+1}] =& E[E[\alpha_{n+1}|\alpha_n]] \\ \approx& E[s\alpha_n] \\ =& sE[\alpha_n]. \end{aligned} \tag{19}$$

Reapplying (19) to itself gives us

$$E[\alpha_{n+1}] \approx s^2 E[\alpha_{n-1}].$$

Continuing this process until the first generation , given that $s^n\alpha_1 \leq 2N$ we get that

$$E[\alpha_{n+1}] \approx s^n E[\alpha_1] = s^n \alpha_1, \tag{20}$$

because α_1 is the amount of individuals with the beneficial mutation in generation 1, this is not a stochastic variable and therefore known (Allen p.34). The expected amount of individuals with mutations will therefore grow exponentially for each successive generation with a factor of s, which is the beneficial mutation factor.

There are two cases where (20) will not hold. The first is that the mutant population in the early stages have their highest probability of dying off in the first initial generations. This is especially true if α_1 makes up a very small percentage of the population size, i.e. $\frac{\alpha_1}{2N} \approx 0$. When this happens we enter one of the absorbing states discussed in Section 3.1. Once the mutants makes up a more substantial percentage of the population the chances of them dying out decreases and (20) can be used. Therefore (20) is best used for the initial first generations.

The second case is when the the quotient $\frac{\alpha_n}{2N}$ can no longer be approximated to 0. Then (20) will be unsuitable to use as we have assumed the quotient to be approximated to 0.

5.2 The later stages

We will now handle the case where the approximation used in (20) does not hold. In order to solve this issue we will introduce a new random variable

$$\frac{\alpha_n}{2N}.$$

We therefore want to analyze the random variable $\frac{\alpha_n}{2N}$ and see for what n

$$E\left[\frac{\alpha_n}{2N}\right] = 1 \tag{21}$$

which is equivalent to at what generation the whole population is said to be mutant.

5.2.1 The variance

Assume that for every α_n it holds that

$$\alpha_n = 2N \cdot c_n,$$

where $0 \leq c_n \leq 1$. The following variance given that $\frac{\alpha_n}{2N} = c_n$ is

$$\begin{aligned} & V\left(\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N} = c_n\right) \\ =& E\left(\left(\frac{\alpha_{n+1}}{2N} - E\left(\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N} = c_n\right)\right)^2\middle|\frac{\alpha_n}{2N} = c_n\right). \end{aligned} \tag{22}$$

(Gut 2009, p. 36). Now using Theorem 2.2 *a*) in An intermediate Course in probability, we get that (22) becomes

$$\begin{aligned} & V\left(\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N} = c_n\right) \\ =& E\left(\left(\frac{\alpha_{n+1}}{2N} - \frac{1}{2N}E(\alpha_{n+1}|\alpha_n = 2Nc_n)\right)^2\middle|\alpha_n = 2Nc_n\right) \\ =& E\left(\frac{1}{(2N)^2}\Big(\alpha_{n+1} - E(\alpha_{n+1}|\alpha_n = 2Nc_n)\Big)^2\middle|\alpha_n = 2Nc_n\right) \\ =& \frac{1}{(2N)^2}E\left(\Big(\alpha_{n+1} - E(\alpha_{n+1}|\alpha_n = 2Nc_n)\Big)^2\middle|\alpha_n = 2Nc_n\right) \\ =& \frac{1}{(2N)^2}V(\alpha_{n+1}|\alpha_n = 2Nc_n) \end{aligned} \tag{23}$$

(Gut 2009, p. 36). Using the distribution given by (13) from Section 4.1.3 and (23) gives us that

$$\begin{aligned} & V\left(\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N} = c_n\right) \\ =& \frac{1}{(2N)^2}2N\frac{s2Nc_n}{s2Nc_n + 2N - 2Nc_n}\left(1 - \frac{s2Nc_n}{s2Nc_n + 2N - 2Nc_n}\right) \\ =& \frac{1}{2N}\frac{sc_n}{1 + sc_n - c_n}\frac{1 - c_n}{1 + sc_n - c_n} \\ =& \frac{1}{2N}\frac{sc_n - sc_n^2}{(1 + sc_n - c_n)^2} \end{aligned} \tag{24}$$

(Alm and Britton 2008, p. 77). The following results gives us an expression for the variance of $\frac{\alpha_{n+1}}{2N}\Big|\frac{\alpha_n}{2N} = c_n$.

5.2.2 Chebyshev's inequality

Let us substitute the random variable $\frac{\alpha_{n+1}}{2N}\big|\frac{\alpha_n}{2N}=c_n$ as X and $E\big[\frac{\alpha_{n+1}}{2N}\big|\frac{\alpha_n}{2N}=c_n\big]$ as μ. Using Chebyshev's inequality we get that

$$\begin{aligned}
&P(|X-\mu|\geq a)\leq\frac{\sigma^2}{a^2}\\
\Rightarrow&1-P(|X-\mu|\geq a)\geq 1-\frac{\sigma^2}{a^2}\\
\Rightarrow&P(|X-\mu|<a)\geq 1-\frac{\sigma^2}{a^2}
\end{aligned}$$

(Alm and Britton 2008, p. 68). Now we can substitute a for $\frac{1}{(2N)^{\frac{1}{4}}}$ as the requirements for a is that it is strictly larger than 0. We also substitute σ^2 for $\frac{1}{2N}\frac{sc_n-sc_n^2}{(1+sc_n-c_n)^2}$ using (24) and get

$$\begin{aligned}
&P\bigg(|X-\mu|<\frac{1}{(2N)^{\frac{1}{4}}}\bigg)\geq 1-\frac{1}{2N}\frac{1}{\sqrt{2N}}\frac{sc_n-sc_n^2}{(1+sc_n-c_n)^2} \qquad (25)\\
\Rightarrow&P\bigg(|X-\mu|<\frac{1}{(2N)^{\frac{1}{4}}}\bigg)\geq 1-\frac{1}{2N^{1.5}}\frac{sc_n-sc_n^2}{(1+sc_n-c_n)^2}
\end{aligned}$$

Now assume that the population $2N$ is very large. Having used this assumption (25) yields

$$\begin{aligned}
\lim_{2N\to\infty}P\bigg(|X-\mu|<\frac{1}{(2N)^{\frac{1}{4}}}\bigg)\geq&1-\frac{1}{2N^{1.5}}\frac{sc_n-sc_n^2}{(1+sc_n-c_n)^2}\\
=&P(|X-\mu|<0)\geq 1
\end{aligned}$$

but because a probability is never larger than 1 we get that

$$P(|X-\mu|<0)=1$$

given that the population size $2N$ goes to infinity. However if $2N$ is sufficiently large enough which in practical sense is a realistic assumption to make, the probability is

$$P(|X-\mu|<0)\approx 1.$$

What this means is that the stochastic variable X and its expected value μ are roughly always the same such that we can assume that for any outcome of X we get that

$$\frac{\alpha_{n+1}}{2N}\bigg|\frac{\alpha_n}{2N}=c_n\approx E\bigg[\frac{\alpha_{n+1}}{2N}\bigg|\frac{\alpha_n}{2N}=c_n\bigg]. \qquad (26)$$

5.2.3 The expected value

Note that using (26) and (16) from Section 4.2 gives us that

$$\begin{aligned} E\left[\frac{\alpha_{n+1}}{2N}\right] =& E\left[E\left[\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N}\right]\right] \\ \approx& E\left[\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N}\right]. \end{aligned}$$

At last we can use the results found in Section 5.2.1 and 5.2.2 and (14) from Section 4.2 and get that

$$\begin{aligned} & E\left[\frac{\alpha_{n+1}}{2N}\middle|\frac{\alpha_n}{2N}\right] \\ \approx& E\left[\frac{\alpha_{n+1}}{2N}\right] \\ =& E\left[E\left[\frac{\alpha_{n+1}}{2N}\middle|\left(\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right)\right]\right] \\ =& E\left[\frac{1}{2N}E\left[\alpha_{n+1}\middle|\left(\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right)\right]\right] \\ =& \frac{1}{2N}E\left[\frac{2Ns\left(\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right)}{2N-(s-1)\left(\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right)}\right] \\ \approx& E\left[\frac{sE\left[\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right]}{2N-(s-1)E\left[\frac{\alpha_n}{2N}\middle|\frac{\alpha_{n-1}}{2N}\right]}\right]. \end{aligned} \tag{27}$$

In conclusion we get a recursive formula where we reapply (27) until we get to generation 1. With very tedious calculations it is possible to find at what n (21) roughly holds true.

6 Growing population size

So far the Wright-fisher model has only allowed for a constant population of size $2N$ to take place in every generation. It is of interest to study a changing population size for every generation.

Assume we know a function $f(n)$ which is a function of the generation n. The function $f(n)$ is not a stochastic variable and is therefore known beforehand. This means that we already know how many individuals the environment in question can sustain and keep alive for every generation. For every n we are looking at $f(n)$ is a contestant no different than $2N$. We can therefore substitute $2N$ for $f(n)$ and have the result

$$\alpha_{n+1}|\alpha_n = m \sim Bin\bigg(f(n), \frac{sm}{sm + f(n) - m}\bigg). \tag{28}$$

hold true given by result (13) from Section 4.1.3.

Instead of the $2N$ children choosing uniformly at random their parents in a Poisson process we instead have that $f(n)$ children choose their parents uniformly at random from generation $n-1$. Note however that (28) builds upon (11) from Section 4.1.3 where in generation n

$$\begin{aligned} &P(X_a = x_a, X_A = x_A | x_a + x_A = f(n)) \\ \sim &Bin\Bigg(f(n), \frac{sx_a}{sx_a + f(n) - x_a}\Bigg). \end{aligned}$$

The assumption we are making however is that $x_a + x_A = f(n)$. Therefore the probability in the auxiliary generation n_a that $X_a + X_A \geq f(n)$ has to be fairly large to insure a growing population. Otherwise if it is too low the individuals can not give birth to sufficient amount of children to fill up the next generation of capacity $f(n)$.

7 Discussion

7.1 Results

In this section we will present the main conclusions gathered in this thesis.

The auxiliary step posed a problem presented in Section 2.2. However we have found a way to circumvent this in a population of equal fitness in Section 3.1 and later with individuals having a mutation in Section 4.1. What is interesting in both of these cases is that the children themselves choose their parents from the previous generation, instead of uniformly at random excluding children from the auxiliary step. This method of uniformly choosing at random is also used in Section 3.1 discussing the coalescent theory.

Using the results gathered in Section 4 we answered the question of the expected amount of mutants in a given generation n. The reason that this is of special interests is that we can find at what generation it is expected that we only have mutants. In Section 5.1 we explain the problem of having too small of a percentage of mutants but later go on to discuss the first up-jump of mutants. In the first few generations we can use the results presented in Section 5.1 to calculate the expected amount of mutants in those first generations. However it is shown that these results do not hold once the mutants start to grow more significantly and make up a larger percentage of the population. It is therefore in Section 5.2 where this problem is dealt with and we find a way to calculate the expected number of mutants in every generation. We can therefore conclude that using both this methods in conjugation with each other can guarantee a relatively good method of calculating the expected number of mutants in every generation.

7.2 Improvements

In this section we will discuss what improvements and ideas could be explored given more time.

What could be interesting is to delve into what happens if there is a population where there are more than two types of alleles. These alleles correspond to different types of reproductive fitness. Seeing how they will propagate and at what generational time one allele is fully represented would be interesting to explore. It is also interesting to go further into the coalescent theory and see how that would relate to a population of unequal fitness.

Another point to mention here is what would happen if a mutation could randomly appear in any given generation. Also some simulations added to explore numerically the results presented in this thesis, especially in relation to mutations would have been beneficial to understanding these concepts better. At last the most interesting idea is to explore a changing population. In this case it could be strictly growing, decreasing or be of a random stochastic nature. I think this reflects reality the best as the environment can shift and is not always assumed to be known in advance.

8 References

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution.* 2nd ed. [ebook]. Available at: `https://services.math.duke.edu/~rtd/Gbook/PM4DNA_0317.pdf` [Accessed: 30 may 2022].

Pierce, B.A. (2012). *Genetics, A Conceptual Approach.* 4th ed. [ebook] New York:W. H. Freeman and Company Available at: `https://generalgenetics.files.wordpress.com/2015/09/pierce-genetics-conceptual-approach-4th-txtbk.pdf` [Accessed: 30 may 2022].

Nordborg, M. (2000). *Coalescent Theory.* Available at: `https://cseweb.ucsd.edu/classes/sp05/cse291-a/doc/nordborg_coalescent.pdf#:~:text=Magnus%20Nordborg%E2%88%97%20Department%20of%20Genetics%2C%20Lund%20University%E2%80%A0%20March,the%20genealogical%20and%20mutational%20history%20of%20these%20copies.` [Accessed: 30 may 2022].

Held, L and Sabanés Bové, D. (2020). *Likelihood and Bayesian Inference With Applications in Biology and Medicine* . Springer

Alm, S.E. and Britton, T. (2008). *Stokastik : sannolikhetsteori och statistikteori med tillampningar.* Stockholm: Liber.

Gut, A. (2009). *An intermediate course in probability.* New York: Springer.

Ross, S.M. (2019). *Introduction To Probability Models.* Academic Press.

Agresti, A. (2018). *Categorical Data Analysis.* Hoboken Wiley.