Mathematical modelling of Population Genetics:

Daniel Bichener

Contents

1 Introduction

Population Genetics is the study of the change in *allele frequencies* (an *allele* is an alernative form of a gene, i.e there are different alleles which determine the gene for eye colour in humans) under the influence of four main evolutionary forces: *Natural Selection, Mutation, Genetic Drift, and Gene Flow.* In this paper we will consider only first two forces, which we will analyse in detail how they affect the allele frequencies, eventually resulting in fixation, within both a haploid(which contains only one allele per gene), and diploid population(which contains a combination of two alleles per gene). After initially showing how they would act if there was no evolutionary forces involved. Throughout the paper we will always have two distinct alleles, denoted by A and a that we will analyse in detail through the use of mathematical models. We will also assume an infinite population for each, as if we were to assume a finite population, then we would have to consider genetic drift as well.

The original explaination of *Natural Selecion* was first thought of by *Charles Darwin* in his book *On the Origin of Species* in which he defines natural selection as *"[a] principle by which each slight variation [of a trait], if useful, is preserved".*[1] A famous example of Natural Selection is the *evolution of the peppered moth* during the industrial revolution in Britain.[2]

Mutation, is a change in the structure of a gene which results in a variation of original gene that may be transmitted to the next generation. It can have four different effects on a genes fitness, the mutation is either harmful, beneficial, neutral, or nearly neutral. For the purpose of this paper we will only consider the case of a harmful mutation.

2 Haploid Genetics

In the first chapter we will consider the most basic form of organisms, haploid organisms. Haploid organisms contain only one copy of chromosomes. We will consider two Allele's, A and a with N_A and N_a being the number of each type of allele and for a total of N alleles. Each Allele will be assumed to have a fitness coefficient denoted r_A and r_a which is the contribution of offspring given that the allele survives to reproduction age. From this information we can find the number of alleles in the next generation.

$$N_A^{(i)} = r_A N_A^{(i-1)} \quad \text{and} \quad N_a^{(i)} = r_a N_a^{(i-1)} \tag{2.0.1}$$

Using (2.0.1) we can find the number of alleles in any generation given the initial number of alleles:

$$N_A^{(i)} = r_A^i N_A^{(0)} \quad \text{and} \quad N_a^{(i)} = r_a^i N_a^{(0)} \tag{2.0.2}$$

Notice that,

$$N^{(i)} = N_A^{(i)} + N_a^{(i)} = r_A N_A^{(i-1)} + r_a N_a^{(i-1)} \tag{2.0.3}$$

Where $N^{(i)}$ is the total population in the next generation. It is interesting to know proportion of A to a alleles we get this by dividing equations (2.0.2) to obtain:

$$\frac{N_A^{(i)}}{N_a^{(i)}} = \left(\frac{rA}{ra}\right)^i \frac{N_A^{(0)}}{N_a^{(0)}} \tag{2.0.4}$$

From equation (2.0.4) we can see that if $r_A > r_a$ then the proportion of A alleles to a will increase infinitely. Conversely, if $r_A < r_a$ then the proportion will decrease until it reaches zero.

2.1 Allele Frequencies

Above, we considered the numbers of each allele in the population. It is often better to compare the frequency of each allele in the population. To do this we will set p and q to be the frequency of A and a allele's respectively. From this we have;

$$p = \frac{N_A}{N} \quad \text{and} \quad q = \frac{N_a}{N} \quad \text{with} \quad p + q = 1$$

Next, we need to find the proportion of p and q in the next generation;

$$p_i = \frac{N_A^{(i)}}{N^{(i)}} \quad \text{and} \quad q_i = \frac{N_a^{(i)}}{N^{(i)}}$$

Working with p only, and using the formulas derived in equations (2.0.1) and (2.0.3) we have:

$$
\begin{aligned}
p_i &= \frac{r_A N_A^{(i-1)}}{r_A N_A^{(i-1)} + r_a N_a^{(i-1)}} \\
&= \frac{\left(\dfrac{r_A}{r_a}\right) p_{(i-1)}}{\left(\dfrac{r_A}{r_a}\right) p_{(i-1)} + q_{(i-1)}}
\end{aligned}
\qquad (2.1.1)
$$

Using $p + q = 1$ we can obtain q:

$$
q_i = \frac{q_{(i-1)}}{\left(\dfrac{r_A}{r_a}\right) p_{(i-1)} + q_i}
\qquad (2.1.2)
$$

2.2 Natural Selection in Discrete Time

Selection occurs in a population when $r_A \neq r_a$, in this section we will be considering the case for $r_A > r_a$ i.e $\dfrac{r_A}{r_a} = 1 + s$ for $s > 0$. For ease of reading we will rewrite p_i as p' and $p_{(i-1)}$ as p (similarly for q). Now using $q = 1 - p$, equation (2.1.1) becomes:

$$
p' = \frac{(1+s)p}{1+sp}
\qquad (2.2.1)
$$

We can find the equilibrium points of this equation and then analyse their stability.[3] To do this we set the next generation frequency equal to the previous generation, denote this by p^* so we have;

$$
\begin{aligned}
p^* &= \frac{(1+s)p^*}{1+sp^*} \\
p^*(p^* - 1) &= 0 \\
\Rightarrow p^* = 1 \quad &\text{or} \quad p^* = 0
\end{aligned}
$$

Next, to analyse the stability of the equilibrium points we set:

$$
p = p^* + \varepsilon \ \text{ and } \ p' = p^* + \varepsilon' \ \text{ where } \ |\varepsilon| \ll 1
$$

So;

$$p*+\varepsilon' = \frac{(1+s)(p^*+\varepsilon)}{1+s(p^*+\varepsilon)}$$

$$\varepsilon' = \frac{(1+s)p^* + (1+s)\varepsilon}{1+sp^*+s\varepsilon} - p^*$$

$$= \left(\frac{(1+s)p^*}{1+sp^*}\right)\left(1+\frac{\varepsilon}{p^*}\right)\left(\frac{1}{1+\frac{s\varepsilon}{1+sp^*}}\right) - p^*$$

Taking a Taylor expansion of order ε and ignoring higher order terms leads to:

$$\varepsilon' = \left(\frac{(1+s)p^*}{1+sp^*}\right)\left(1+\frac{\varepsilon}{p^*}\right)\left(1-\frac{s\varepsilon}{1+sp^*}\right) - p^*$$

Multiplying out, ignoring any higher order ε terms, and simplifying reduces the equation to:

$$\varepsilon' = \frac{(1+s)p^*}{1+sp^*} + \frac{(1+s)\varepsilon}{(1+sp^*)^2}) - p^*$$

$$\Rightarrow \quad \frac{\varepsilon'}{\varepsilon} = \frac{1+s}{(1+sp^*)^2} \tag{2.2.2}$$

Notice that originally if:

$$f(p) = p' = \frac{(1+s)p}{1+sp} \tag{2.2.3}$$

This a recurrence relation for p and it follows that equation (2.2.2) is $f'(p)$. The case for stability of a recurrence relation is $|f'(p^*)| < 1$, or, $\left|\frac{\varepsilon'}{\varepsilon}\right| < 1$. Analysing the equilibrium points we find that $p^* = 1$ is stable, and $p^* = 0$ is unstable.

2.3   Natural Selection in Continuous Time

We can also analyse natural selection in continuous time. To do this we assume that $s \ll 1$, then take equation (2.2.3) and expand it in orders of s to give;

$$p' = (1+s)p(1-sp)$$
$$= p + sp - sp^2$$
$$p' - p = sp(1-p)$$

If we now consider p as a function of time, then;

$$\frac{p(t+h) - p(t)}{h} = sp(t)(1 - p(t))$$

As $h \to 0$ we get the logistical equation;

$$\frac{dp}{dt} = \tilde{s}p(1-p) \qquad (2.3.1)$$

Solving for p;

$$p(t) = \frac{Ae^{\tilde{s}t}}{1 + Ae^{\tilde{s}t}} \qquad (2.3.2)$$

We can find A in terms of p(0);

$$p(0) = \frac{A}{1 + A}$$

$$\Rightarrow A = \frac{p(0)}{1 - p(0)}$$

Using these conditions we can find $\tilde{s}$ in terms of s and see whether they differ:

$$p(1) = p(0)(1 + s - sp(0)) = \frac{A}{1 + Ae^{\tilde{s}t}}$$

Rearranging for $e^{\tilde{s}t}$;

$$e^{\tilde{s}t} = \frac{p(0)(1 + s - sp(0))}{A(1 - p(0)(1 + s - sp(0)))}$$

$$= \frac{(1 - p(0))(1 + s - sp(0))}{1 - p(0)(1 + s - sp(0))}$$

Finally solving for $\tilde{s}$:

$$\tilde{s} = \ln\left(\frac{(1-p(0))(1+s-sp(0))}{1-p(0)(1+s-sp(0))}\right)$$

$$= \ln\left(\frac{1-p(0)}{\dfrac{1}{1+s-sp(0)}-p(0)}\right)$$

$$= \ln\left(\frac{1-p(0)}{1-s+sp(0)-p(0)}\right)$$

$$= \ln\left(\frac{1}{1-s}\right)$$

$$\approx s$$

So by assuming $s \ll 1$ we have found that there is no difference between $\tilde{s}$ and s, therefore we can re-write equations (2.3.1) and (2.3.2) as:

$$\frac{dp}{dt} = sp(1-p) \qquad (2.3.3)$$

$$p(t) = \frac{Ae^{st}}{1+Ae^{st}} \qquad (2.3.4)$$

Respectively. Below there are two figures, Figure 1: represents equation (2.3.3) and Figure 2: represents equation (2.3.4).
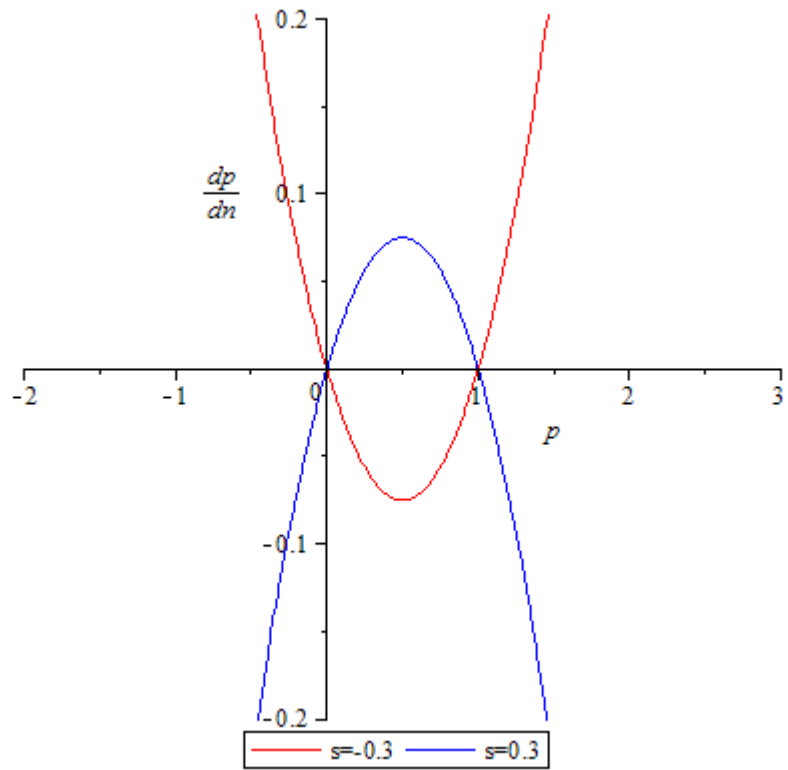
8

Figure 1: Stability analysis for Natural Selection in a Haploid Population with different fitness coefficients.

As we can see from Figure 1: a negative fitness coefficient corresponds to 0 being stable and 1 being unstable, and a positive fitness corresponds to 1 being stable and 0 unstable. From Figure 2: we see how the allele frequency becomes *fixed* for $s = 0.3$ and *lost* for $s = -0.3$.
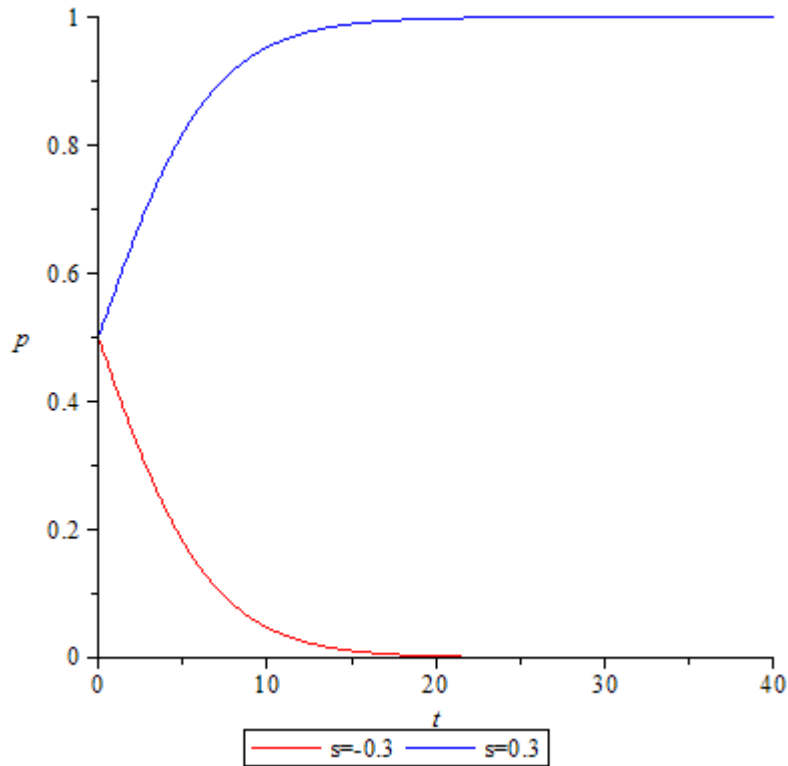
Figure 2: Natural selection in a Haploid Population: Fixation/Extinction.

2.4 Mutation-Selection in Discrete Time

In the above section, we we notice that if the fitness of allele A is greater than the fitness of allele a then eventually the population will consist only of A alleles. In this chapter we will consider what happens when we introduce *mutation* which is the probability u that each new A allele will defect into an a allele which is typically assumed to have a lower fitness than allele A. For this section we will take allele A as the comparison for fitness so allele a has a relative fitness of $1 - s$. Since we are assuming that natural selection also takes place, each generation a proportion up of A alleles are lost so we have using the simplified notation established previously:

$$p' = \frac{(1 - u)p}{p + (1 - s)p}$$

10

The denominator in this case is different to that of equation (2.2.1) because we have set p as the comparison for fitness. So in terms of p only;

$$p' = \frac{(1-u)p}{1-s(1-p)} \qquad (2.4.1)$$

And for q;

$$q' = \frac{(1-s)q+up}{p+(1-s)q}$$

Again, we will proceed to find the equilibrium points of (2.4.1). We do this by setting $p^* = p' = p$;

$$p^* = \frac{(1-u)p^*}{1-s(1-p^*)}$$
$$p^*(-s(1-p^*)+u) = 0$$

This gives:

$$p^* = 0 \quad \text{or} \quad p^* = 1 - \frac{u}{s}$$

Now to analyse these points for stability using the same tactics for the case without mutation:

$$p = p^* + \varepsilon \text{ and } p' = p^* + \varepsilon' \text{ where } |\varepsilon| \ll 1$$

Substituting into equation (2.4.1) gives:

$$p*+\varepsilon = \frac{(1-u)(p*+\varepsilon)}{1-s(1-(p*+\varepsilon))}$$

$$\varepsilon' = \left(\frac{(1-u)p^*}{1-s(1-p^*)}\right)\left(1+\frac{\varepsilon}{p^*}\right)\left(\frac{1}{1+\dfrac{s\varepsilon}{1-s(1-p^*)}}\right) - p^*$$

$$= \left(\frac{(1-u)p^*}{1-s(1-p^*)}\right)\left(1+\frac{\varepsilon}{p^*}-\frac{s\varepsilon}{1-s(1-p^*)}\right) - p^*$$

$$= \left(\frac{(1-u)p^*}{1-s(1-p^*)}\right)\left(1+\frac{(1-s)\varepsilon}{p^*(1-s(1-p^*))}\right) - p^*$$

$$\Rightarrow \frac{\varepsilon'}{\varepsilon} = \frac{(1-u)(1-s)}{(1-s(1-p^*))^2}$$

11

Substituiting the equilibrium points in we get;

$$f'(0) = \frac{\varepsilon'}{\varepsilon} = \frac{1-u}{1-s} \quad \text{or;}$$

$$f'(1 - \frac{u}{s}) = \frac{\varepsilon'}{\varepsilon} = \frac{1-s}{1-u}$$

The case for stability is again that $f'(p) < 1$. Analysing the equilibrium points we find that $p^* = 0$ is stable for $s < u$ i.e. if the relative fitness is weaker than the force of mutation. $p^* = 1 - \frac{u}{s}$ is stable for $s > u$ i.e. if the relative fitness is greater than the force of mutation.

2.5 Mutation-Selection in Continuous Time

Next we should consider what happens if we assume continuous time. The method involved is analogous to that in Section 2.3. Assume $u, s \ll 1$, then we can approximate equation (2.4.1) with a taylor expansion in s:

$$p' = (1-u)(1 + s(1-p))$$
$$= p + sp - sp^2 - up - usp + usp^2$$

Ignoring terms containing us;

$$p' - p = p(s - u - sp)$$

This can be represented in terms of the differential equation;

$$\frac{dp}{dt} = p(\tilde{s} - \tilde{u} + \tilde{s}p) \tag{2.5.1}$$

Seperating the variables and using partial fractions we get;

$$\int \left(\frac{1}{(\tilde{s} - \tilde{u})p} + \frac{\tilde{s}}{(\tilde{s} - \tilde{u})(\tilde{s} - \tilde{u} + \tilde{s}p)} \right) dp = \int 1.dt$$

We can rewrite this as;

$$\int \left(\frac{1}{p} + \frac{\tilde{s}}{d - \tilde{s}p} \right) dp = \int d.dt \quad \text{Where } d = \tilde{s} - \tilde{u}$$

Solving for p;

$$p(t) = \frac{Ade^{dt}}{1 + \tilde{s}Ae^{dt}} \tag{2.5.2}$$

Where

$$A = \frac{p(0)}{d - \tilde{s}p(0)}$$

If we analyse the limit of equation (2.5.2) we see that for $d > 0$, i.e. $\tilde{s} > \tilde{u}$

$$\lim_{t \to \infty} p(t) = 1 - \frac{\tilde{u}}{\tilde{s}}$$

Which means that the population will never always consist of just A alleles as some will always mutate after selection has taken place. However if $d < 0$ then $\lim_{t \to \infty} p(t) = 0$, eventually allele A will become extinct. Below we have two figures; Figure 3: and Figure 4: which represent equations (2.5.1) and (2.5.2) respectively. In each case $p(0) = 0.4$ and $s = 0.3$.
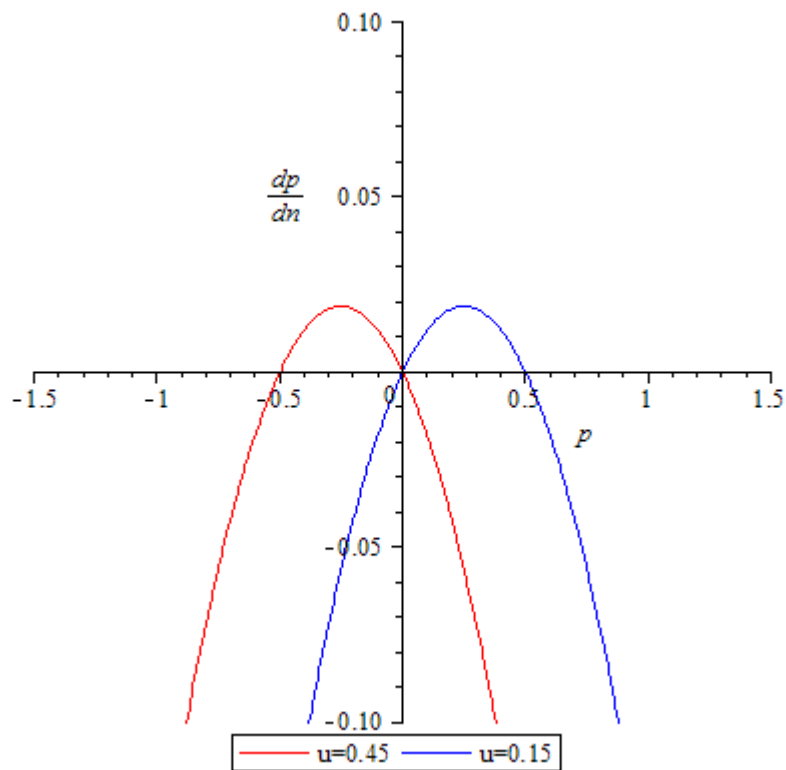

Figure 3: Stability analysis for Natural Selection with Mutation in a Haploid Population with different fitness coefficients.

13

As we can see from Figure 3: when $u > s$, which is represented by the red line, $p = 0$ is stable. Whereas when $u < s$, $p = 0$ is unstable and $p = 1 - \frac{u}{s}$ is stable.
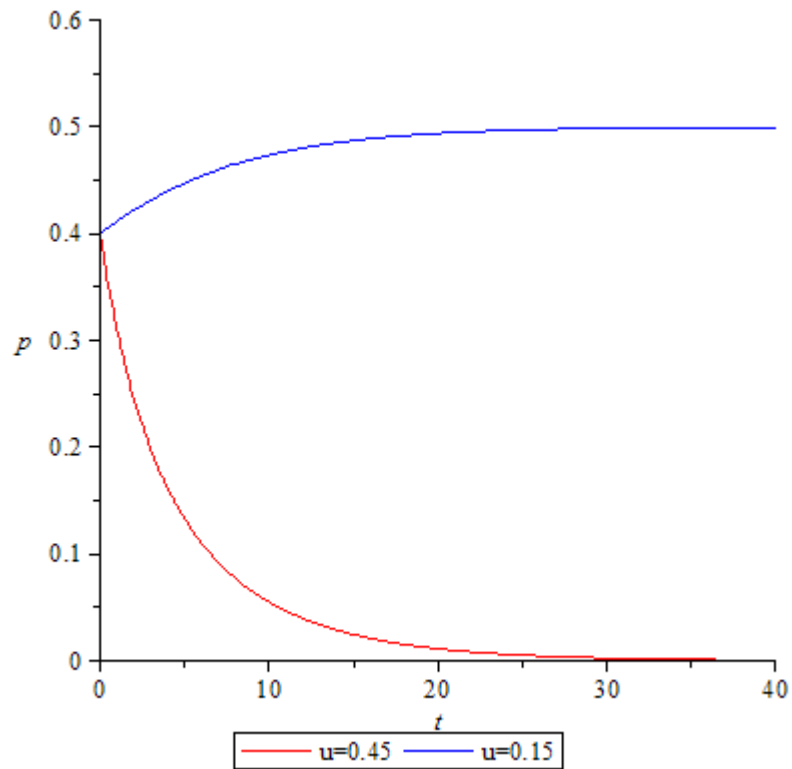

Figure 4: Natural selection with Mutation in a Haploid Population: Fixation/Extinction.

Notice from Figure 4: we can see how the red line which represents the case for when $d < 0$ and the allele is lost, whereas the blue line represents $d > 0$ and the allele eventually tends to a frequency of $\frac{1}{2}$.

3 Diploid Genetics

In this chapter we will consider the other form of organisms which are diploids. Diploids are organisms that contain two copies of chromosomes, one from each parent. Again we will assume two distinct allele's, A and a and we will proceed through the chapter in a similar way as we did for our haploid models.

3.1 Allele Frequencies

There are two different ways in which a diploid organism can reproduce that we will consider in this section. The two methods we will introduce are self-fertilisation or *selfing* and *random mating*. In each case we will find the frequency of each allele in the ith, $i^{th}+1$, and n^{th} generations.

For a diploid organism with two distinct alleles we have three possible combinations of genotypes. We have the homozygous combinations AA and aa or the heterozygous combination Aa. For each of the combinations we define their frequency in the population as:

$$P = \frac{N_{AA}}{N}, \quad Q = \frac{N_{Aa}}{N}, \quad R = \frac{N_{aa}}{N} \tag{3.1.1}$$

From the equations in (3.1.1) we can find the proportion of A and a alleles in the population;

$$\begin{aligned} p &= \frac{n_A}{n} \\ &= \frac{2P+Q}{2N} \\ &= P + \frac{1}{2}Q \end{aligned} \tag{3.1.2}$$

Similarly, the proportion of q alleles is;

$$q = R + \frac{1}{2}Q \tag{3.1.3}$$

Since we are dealing with proportions, we also have the two constraints:

$$p + q = 1 \quad \text{and} \quad P + Q + R = 1 \tag{3.1.4}$$

For finding the allele frequencies in a diploid population we will first begin by making six assumptions, and we will prove that how under these conditions the allele frequencies in both random mating and selfing populations remains constant. The assumptions we make are as follows:

1. Large (in theory infinite) population

2. No selective reproduction

3. No cross-overs or mating between different generations

4. No natural selection

5. No mutation

6. No genetic drift

3.1.1 Random Mating

To model sexual reproduction in a diploid population we will use two methods of *random mating* and show how they in fact produce the same results. Random Mating assumes that each organism in the population has a chance of mating with the opposite sex.

From this definition we can introduce the two different methods of modelling; the *Gene pool approach* and the *Mating table Method.*

Gene Pool Approach

First assume that all males and females release their gametes into two different pools each pool containing either all male or all female gametes. The offspring for the next generation is then determined by selecting one gamete from each pool. Then the probability of getting P, Q or R genotypes in the next generation is as follows:

$$\text{Probability of offspring P'} = p(AA) = p^2$$
$$\text{Probability of offspring Q'} = p(Aa \text{ or } aA) = 2pq$$
$$\text{Probability of offspring R'} = p(aa) = R^2$$

If we sum up the probability of getting each genotype in the next generation we have:

$$P' + Q' + R' = 1$$
$$\Rightarrow p^2 + +2pq + q^2 = 1$$
$$\Rightarrow (p + q)^2 = 1 \tag{3.1.5}$$

This result is known as the *Hardy-Weinberg equation*[4] and we will now show how we can get the same result from the mating table method.

Mating Table Method

For this approach we will consider all the different mating possibilities between the different genotypes. We have 3 different genotypes P, Q, R and we have 2 different combinations where order doesn't matter, so we have 6 different combinations; AAxAA, AAxAa (twice), AAxaa (twice), AaxAa, Aaxaa (twice), and aaxaa. We will now put these mating combinations into a table along side their frequencies and the probability of offspring:

Mating	Frequency	P'	Q'	R'
AAxAA	P^2	$1 * P^2$	–	–
AAxAa	$2PQ$	$\frac{1}{2} * 2PQ$	$\frac{1}{2} * 2PQ$	–
AAxaa	$2PR$	–	$1 * 2PR$	–
AaxAa	Q^2	$\frac{1}{4} * Q^2$	$\frac{1}{2} * Q^2$	$\frac{1}{4} * Q^2$
Aaxaa	$2QR$	–	$\frac{1}{2} * 2QR$	$\frac{1}{2} * 2QR$
aaxaa	R^2	–	–	$1 * R^2$

Table 1: Random Mating table with two distinct alleles

If we now add up the offspring columns we get:

$$P' = P^2 + PQ + \frac{1}{4}Q^2 = \left(P + \frac{1}{2}Q\right)^2 = p^2 \qquad (3.1.6)$$

$$Q' = PQ + 2PR + \frac{1}{2}Q^2 + QR = 2\left(\frac{1}{2}PQ + PR + \frac{1}{4}Q^2 + \frac{1}{2}QR\right)$$

$$= 2\left(P + \frac{1}{2}Q\right)\left(R + \frac{1}{2}Q\right) = 2pq \quad (3.1.7)$$

$$R' = R^2 + PQ + \frac{1}{4}Q^2 = \left(R + \frac{1}{2}Q\right)^2 = q^2 \qquad (3.1.8)$$

Adding up equations (3.1.6), (3.1.7) and (3.1.8) we get the same result as equation (3.1.5) using the gene pool approach. Hence both methods produce the same answer. We can use the Hardy-Weinberg Equation to show how

allele frequencies remain constant in each generation. Using equation (3.1.2) for the next generation we have;

$$p' = P' + \frac{1}{2}Q'$$

Substituting our results from equations (3.1.6) and (3.1.7);

$$
\begin{aligned}
p' &= p^2 + \frac{1}{2} * 2pq \\
&= p(p + q) \\
&= p(p + (1 - p)) \\
&= p
\end{aligned}
$$

This result is analogous for q, therefore proving that allele frequencies remain constant in a diploid population modelling random mating under our 6 assumptions.

Another interesting result coming from these assumptions is the genotype frequencies P, Q, and R remain constant after the second generation. After one generation of mating we have the genotype frequencies P', Q' and R' described above. For the generation after that we will have:

$$
\begin{aligned}
P'' &= P'^2 + P'Q' + \frac{1}{4}Q'^2 \\
&= \left(P' + \frac{1}{2}Q'\right)^2
\end{aligned}
$$

substituting P' and Q' from equations (3.1.6) and (3.1.7);

$$P'' = \left(P^2 + PQ + \frac{1}{4}Q^2 + \left(P + \frac{1}{2}Q\right)\left(R + \frac{1}{2}Q\right)\right)^2$$

Using R=1-P-Q and expanding;

$$
\begin{aligned}
P'' &= \left(P^2 + \frac{1}{4}Q^2 + PQ + P(1 - P - Q) + \frac{1}{2}Q(P + 1 - P - Q) + \frac{1}{4}Q^2\right)^2 \\
&= \left(P + \frac{1}{2}Q\right)^2 = P'
\end{aligned}
$$

Showing that it only takes one generation before the genotype frequencies are mixed in the population. The figure below will show this graphically:
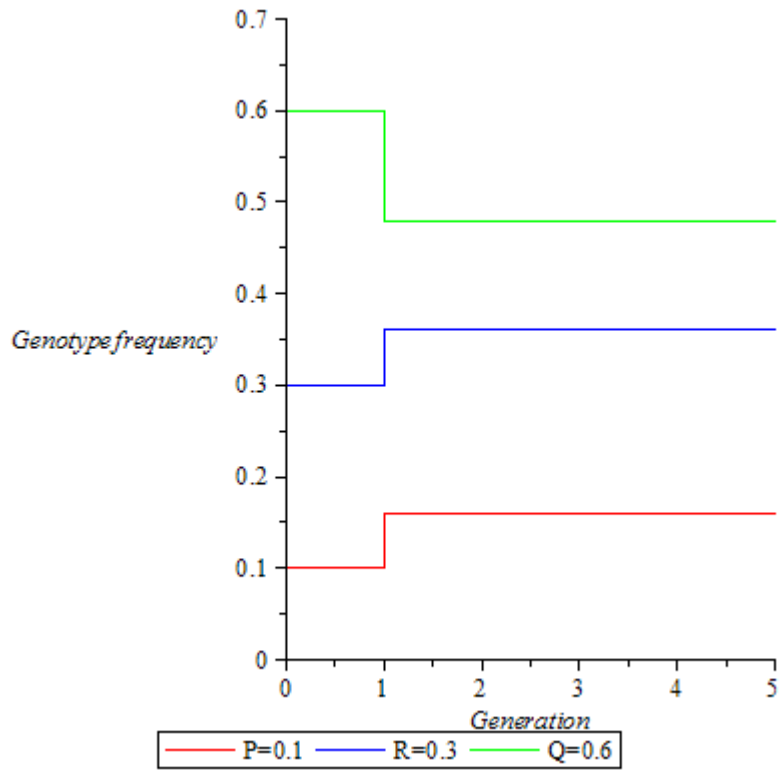
Figure 5: Random Mating with two distinct alleles

We see that the initially we have $P(0) = 0.1$ $Q(0) = 0.6$ and $R(0) = 0.3$ then after one generation the genotypes remain fixed at $P = 0.16$, $Q = 0.48$, and $R = 0.36$

3.1.2 Selfing

The other way a diploid organism can reproduce is by self reproduction or selfing. Selfing is when the organism involved provides both the male and female gametes. We can use a mating table similar to table 1 to model selfing:

Parent	Frequency	P'	Q'	R'
AA	P	$P * 1$	–	–
Aa	Q	$\frac{1}{4}Q$	$\frac{1}{2}Q$	$\frac{1}{4}Q$
aa	R	–	–	$R * 1$

Table 2: Self-fertisation table with two distinct alleles.

Again adding up the columns for the next generation genotype frequencies we find that:

$$P' = P + \frac{1}{4}Q \tag{3.1.9}$$

$$Q' = \frac{1}{2}Q \tag{3.1.10}$$

$$R' = R + \frac{1}{4}Q \tag{3.1.11}$$

Now to find the allele frequencies in the next generation:

$$p' = P' + \frac{1}{2}Q'$$
$$= P + \frac{1}{4}Q + \frac{1}{2}\left(\frac{1}{2}Q\right)$$
$$= p$$

We can use a similar method to show that $q' = q$. It's worth saying that although we have consistency in allele frequencies over generations, unlike in random mating the genotype frequencies do not remain the same each time. In fact after each generation half of the Q genotypes are lost from the previous generation, and the P and R frequencies each grow by an additional quarter of Q from the previous generation. If initially at the first generation we have $P^{(0)}$, $Q^{(0)}$ and $R^{(0)}$ then in the second generation we have:

$$P^{(1)} = P^{(0)} + \frac{1}{4}Q^{(0)}, \;\; Q^{(1)} = \frac{1}{2}Q^{(0)}, \;\; R^{(1)} = R^{(0)} + \frac{1}{4}Q^{(0)}$$

For the third generation we will have;

$$P^{(2)} = P^{(1)} + \frac{1}{4}Q^{(1)}$$
$$= P^{(0)} + \frac{1}{4}(Q^{(0)} + Q^{(1)})$$
$$= P^{(0)} + \frac{1}{4}Q^{(0)}(1 + \frac{1}{2})$$

In general we have;

$$P^{(n)} = P^{(0)} + \frac{1}{4}Q^{(0)} \sum_{n=0}^{n-1}(\frac{1}{2})^n$$

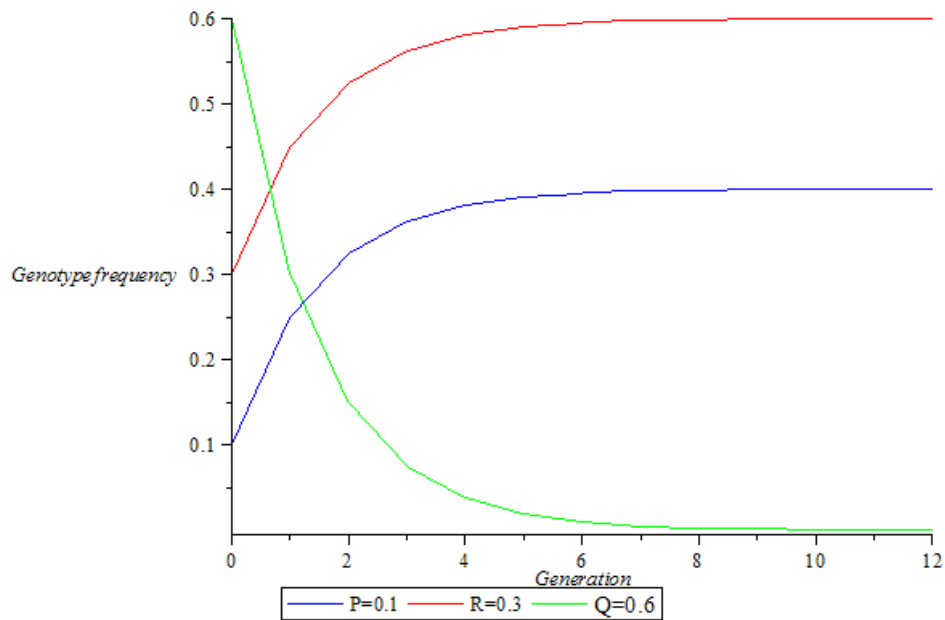Again, we will show this graphically with the figure below:

Figure 6: Self-Fertilisation with two distinct alleles

This time we notice that it takes much longer for the genotypes to eventually become fixed at $P = 0.4$, $Q = 0$, and $R = 0.6$, even though we started with the same frequencies that we had for Figure 5.

3.2 Random Mating with Natural Selection in Discrete Time

To model natural selection in a diploid population we will start by assuming that one allele has a higher fitness coefficient than the other, as we did with a haploid population. We will choose this to be allele A, so that the fitness of the genotype AA is greater than that of the genotype aa, but the genotype Aa doesn't necessarily have the same fitness as either of the two. So we need to introduce a new parameter h to allow for this condition. Next we will set genotype aa as the comparison genotype such that the relative fitness of $aa = 1$. We will summarise all this into one table, bearing in mind we are assuming random mating throughout.

Genotype	AA	Aa	aa
Freq. of offspring	p^2	$2pq$	q^2
Relative fitness	$1+s$	$1+sh$	1
Freq. after selection	$\dfrac{(1+s)p^2}{w}$	$\dfrac{(1+sh)2pq}{w}$	$\dfrac{q^2}{w}$

Table 3: Random mating with Natural selection with two distinct alleles

Where $w = (1+s)p^2 + (1+sh)2pq + q^2$. So that the sum of the frequencies in the next generation is one. Hence the frequency of allele A after selection is given by;

$$p' = P' + \frac{1}{2}Q'$$
$$= \frac{(1+s)p^2 + (1+sh)pq}{(1+s)p^2 + (1+sh)2pq + q^2}$$

Using $q = 1 - p$ and simplifying;

$$p' = \frac{(1+sh)p + (1-h)sp^2}{1 + 2sh + (1-2h)sp^2} \tag{3.2.1}$$

We will find and analyse the stability points in the next section.

3.3 Random Mating with Natural Selection in Continuous Time

To find a model for natural selection in continuous time we will use the trick we established in previous sections. We begin by rearranging equation (3.2.1) into:

$$p' = \frac{p + s(hp + (1-h)p^2)}{1 + s(2hp + (1-2h)p^2)} \qquad (3.3.1)$$

Next we will assume that $s \ll 1$ and take a Taylor expansion of equation (3.3.1) to get:

$$p' = (p + s(hp + (1-h)p^2))(1 - s(hp + (1-h)p^2)(1 + s(2hp + (1-2h)p^2) + ...) \qquad (3.3.2)$$

We can then ignore any higher terms of s and simplify equation (3.3.2) to:

$$p' - p = sp((1-3h)p + h - (1-2h)p^2) \qquad (3.3.3)$$

Equation (3.3.3) Can be represented by the differential equation:

$$\frac{dp}{dt} = sp((1-3h)p + h - (1-2h)p^2) \qquad (3.3.4)$$

We will now use this equation to find the stationary solutions of our model and analyse their stability then we will proceed to solve the differential equation. We find stationary points from $f(p) = \dfrac{dp}{dt} = 0$, there are three solutions to this equation. One obvious solution is $p = 0$, the next solution is $p = 1$ so we can factor this out of equation (3.3.4) to find the remaining solution:

$$sp(p-1)(-p - h + 2hp) = 0$$

Hence:

$$p = \frac{h}{2h - 1} \text{ for } h \neq \tfrac{1}{2}$$

Is our final solution. For the case when $h = \dfrac{1}{2}$ the cubic term in equation (3.3.4) will disappear and our only solutions will be $p = 0$ and $p = 1$. To

analyse the stability of our three points we will differentiate equation (3.3.4) to get:

$$f'(p) = s((1 - 3h)p + h - (1 - 2h)p^2) + sp(1 - 3h - (2 - 4h)p) \quad (3.3.5)$$

Since the sign of $f'(p)$ depends on our choice of h, we will consider four different cases, either $h = 0$, $0 < h < \frac{1}{2}$, $\frac{1}{2} < h < 1$, or $h > 1$. These four different cases can be interpreted as follows; $h = 0$ means there is no difference between the fitness of the homozygous genotype aa and the heterozygous genotype Aa this means that allele A would be recessive. $0 < h < \frac{1}{2}$ implies a partial dominance of allele A whereas $\frac{1}{2} < h < 1$ implies a strong dominance. $h > 1$ models a phenomena known as *heterozygous dominance* where the heterozygous genotype has the highest fitness, a few examples in humans are sickle cell anaemia and thalassemia.Two mutations of the haemoglobin cell that when paired with a normal haemoglobin gene offer a natural immunity to malaria. [5]

For $h = 0$;

$$f'(p) = sp(1 - 3p)$$

So;

$$f'(p) = \begin{cases} 0, \text{ if } p = 0 \text{ or } p = \dfrac{h}{2h - 1} \text{ hence is a saddle point.} \\ -2, \text{if } p = 1 \text{ hence is a stable equilibrium point.} \end{cases}$$

Next for $h \neq 0$;

$$f'(p) = 2sp - 6shp + sh - 3sp^2 + 6shp^2$$

For $p = 0$ we have $f'(p) = sh > 0$ for all $h > 0$ so $p = 0$ is an unstable equilibrium point. Now instead of computing the value for $f'(p)$ at the other points we will do it by analysis. If $0 < h < \dfrac{1}{2}$ we have a negative cubic, hence we have two stable equilibria points and one unstable equilibrium point, this means that the other two points are stable equilibria. Next, for $h > \dfrac{1}{2}$ we have a positive cubic, hence two unstable equilibria and one stable equilibria. If $\dfrac{1}{2} < h < 1$ then $p = \dfrac{h}{2h - 1} > 1$ meaning that $p = 1$ is stable and $p = \dfrac{h}{2h - 1}$ is unstable, and if $h > 1$ then $p = 1$ is unstable and $p = \dfrac{h}{2h - 1}$ is stable. All of this information will be summarised in the figure below:
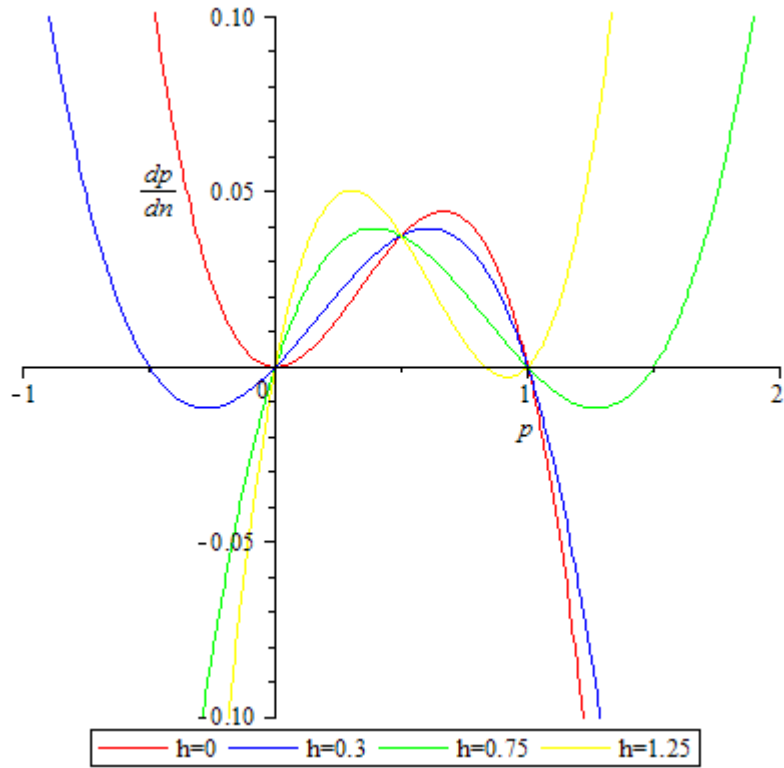
Figure 7: Natural selection with Random mating for different h parameters with $s = 0.3$

From Figure 5 we have each of the four different cases represented by a line on the figure. The red line represents $h = 0$ i.e a recessive A allele, the yellow line represents heterozygous dominance, the blue and green line represent partial and strong dominance of allele A respectively.

Now that we have found the equilibrium points and determined the stability of the allele frequency we will attempt to solve the differential equation (3.3.4). First we will factorise the equation, and separate the variables to get the following:

$$\int \frac{dp}{p(p-1)(p - \frac{h}{2h-1})} = \int s.dt$$

Next, using partial fractions to simplify the left hand side we get:

$$\int \left(\frac{2h-1}{hp} + \frac{2h-1}{(h-1)(p-1)} + \frac{(2h-1)^3}{h(1-h)((2h-1)p-h)} \right) .dp = st + c$$

Integrating;

$$\frac{2h-1}{h} \ln{(p)} + \frac{2h-1}{h-1} \ln{(p-1)} + \frac{(2h-1)^2}{h(1-h)} \ln{((2h-1)p-h)} = st + c$$

Even with the help of mathematical software, this equation is difficult to solve for p, so we will consider a specific example. If we take a population of diploid organisms where the dominant allele is recessive such that $h = 0$ and $p_0 = p(0) \ll 1$ then we can rewrite (3.3.4) as:

$$\frac{dp}{dt} = sp^2$$

solving for p;

$$p = \frac{p_0}{1 - sp_0 t}$$

taking a taylor expansion;

$$p \approx p_0(1 + sp_0 t)$$

From this we see that a recessive dominant allele increases only linearly across generations. This will be shown in the figure below.
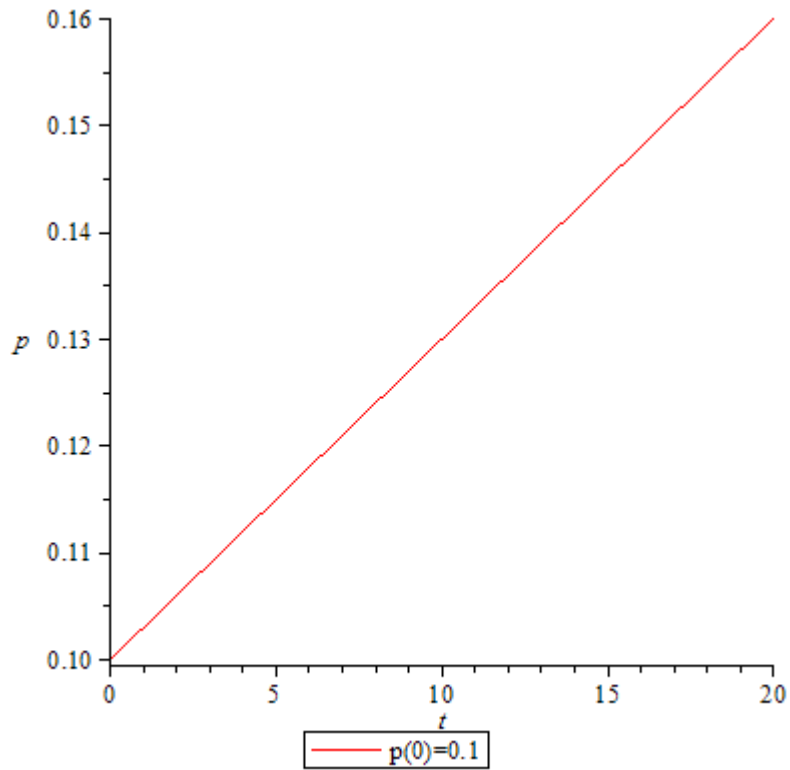
Figure 8: Natural selection in continous time with $p(0) = 0.1$

3.4   Mutation-Selection in Discrete Time

Now we will introduce mutation into our model. Again, we will use u to denote the force of mutation of allele a onto allele A, and again like we did for our mutation in a haploid population model, we will set the genotype AA as the comparison for our fitness so we can amend table 3 as follows:

Genotype	AA	Aa	aa
Freq. of offspring	p^2	$2pq$	q^2
Relative fitness	1	$1 - sh$	$1 - s$
Freq. after selection	$\dfrac{p^2}{w}$	$\dfrac{(1 - sh)2pq}{w}$	$\dfrac{(1 - s)q^2}{w}$

Table 4: Modelling Mutation-selection in Diploids

With $w = p^2 + 2(1 - sh)pq + (1 - s)q^2$. Next, however, instead of dealing with the frequency of A alleles we will consider the frequency of a alleles instead, which we have previously denoted q as it slightly simplifies our equations. First, let $\hat{p}$ and $\hat{p}$ be the frequencies of alleles A and a after selection respectively. Now we apply the force of mutation so that the frequency of allele a in the next generation is given by:

$$q' = u\hat{p} + \hat{q}$$

Substituting $\hat{p}$ and $\hat{q}$;

$$= \frac{u(p^2 + (1 - sh)pq) + (1 - sh)q^2 + (1 - sh)pq}{w}$$

Where the values for $\hat{p}$ and $\hat{q}$ are the frquencies after selection and we obtain from the formulas (3.1.2) and (3.1.3) respectively. Now, using $q = 1 - p$ we get:

$$q' = \frac{u + (1 - u - sh(1 + u))q - s(1 - h(1 + u))q^2}{1 - 2shq - s(1 - 2h)q^2}$$

Now to find the equilibrium solutions of this equation we begin by setting $q* = q = q'$ One solution is readily found, by neglecting back mutation (i.e.a mutated gene reverting back to it's original state) as $q_1 = 1$. The other two solutions are found from the equation:

$$s(1 - 2h)q^2 + sh(1 + u)q - u = 0; \quad [6]$$

30

As;

$$q_2 = \frac{h(1+u)}{2(2h-1)} \left(1 + \sqrt{1 - \frac{4(2h-1)u}{sh^2(1+u)^2}} \right) \quad \text{and;}$$

$$q_3 = \frac{h(1+u)}{2(2h-1)} \left(1 - \sqrt{1 - \frac{4(2h-1)u}{sh^2(1+u)^2}} \right)$$

For the special case when $h = \dfrac{1}{2}$ the quadratic equation becomes linear, and the other solution is given by $q = \dfrac{2u}{s(1+u)}$. The stability of these points are analysed in detail in a paper published by $J.R.Chasnov.$[7]

4 Conclusion

In conclusion, we have seen how the introduction of either evolutionary force can affect the allele frequencies over generations. This was to be expected, as we had initially shown in the beginning of each chapter how without any forces the frequencies remained constant regardless of size, and then the introduction of either force would favour one of the allele's, eventually making that allele fixed in the population, whilst the other was lost.

The interesting results came from when we considered a diploid population. It would have been possible to analyse a diploid population that experienced self-fertilisation, however, for random mating within a diploid population we seen how the results for the equilibrium points and the stability of them, really only depended on the heterozygous genotype and in particular, how dominant allele A was compared with allele a. This result also extends for the case with mutation. As was shown by the work of J.R.Chasnov, the equilibrium points and their stability also depended on the value for our h parameter. It would also be interesting to consider what would happen for a diploid population that experiences a mix of both self-fertilisation.

It be natural to follow up the results we have gathered and consider Natural Selection and Mutation for organisms with multiple alleles, not just two and see how or if, they differ. Or also relaxing some of our assumptions such as selective reproduction, or cross-overs and seeing what impact they have upon our model.

However, even though we have shown how these organisms react to these evolutionary forces under certain assumptions. We have neglected to consider the case for a finite population. In such a circumstance it is unrealistic that any one allele would become totally fixed within a population because of the effect of gene flow and genetic drift which we have yet to discuss. Extensive research into the idea of genetic drift has been discussed by mathematicians Motoo Kimura, and James Crow. It is also conceivable to combine all the evolutionary forces into one model.

In the end, the beauty of population genetics is that it is still a very new area of mathematics with very much left to be discovered, from Darwin's original theory of evolution we have came a long way in a short period of time and we know have a "Modern evolutionary synthesis" explaination of evolution, which has resulted from the works of Charles Darwin, R.A.Fisher, and Sewell Wright, to name a few.

References

[1] Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* 1859.

[2] Cook, L. M., B. S. Grant, I. J. Saccheri and J. Mallet. *Selective bird predation on the peppered moth: the last experiment of Michael Majerus.* 2012.

[3] Hartl, L. D; Clark, G. A. *Principles of Population Genetics fourth edition.* 220-222. 2007.

[4] Stern, C. *The HardyWeinberg law.* Science 97 (2510): 137-138, 1943.

[5] Bridges, K. *Malaria and the Sickle Hemoglobin Gene* 2002.

[6] CROW, J. F; KIMURA, M. *An introduction to population genetics theory.* 1970.

[7] J. R. Chasnov. *Mutation-Selection Balance, Dominance and the Maintenance of Sex* . Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 2000.