# Not Quite Enough Selection (Or Maybe a Bit Too Much)

Alan R. Rogers

April 15, 2023

## 1  Introduction

This note summarizes basic results from the theory of natural selection. It develops the conventional algebraic model and uses this to derive a series of qualitative conclusions, which are stated below as "Results."

## 2  Basic model

Consider a locus with two alleles $A_1$ and $A_2$, whose relative frequencies in some generation are $p$ and $q = 1 - p$, respectively. Let $w_{ij}$ denote the probability that an individual of genotype $A_i A_j$ will survive to reproduce. This probability is called the fitness of genotype $A_i A_j$. The genotypic fitnesses and frequencies are shown in table 1. The frequencies before selection are in Hardy-Weinberg equilibrium. The contribution of each genotype to the gene pool of the next generation is proportional to the product of its frequency and fitness. Dividing by $\bar{w}$, the mean fitness, ensures that the genotypic frequencies after selection sum to unity.

The frequency of allele $A_1$ is $p = P_{11} + \frac{1}{2} P_{12}$, where $P_{11}$ and $P_{12}$ are the frequencies of $A_1 A_1$ and $A_1 A_2$.

$\star$ EXERCISE **1** Show that $p = P_{11} + \frac{1}{2} P_{12}$.

Applying this to the last row of table 1 gives the frequency of $A_1$ after selection.

$$p' = \frac{p^2 w_{11} + pq w_{12}}{\bar{w}} \tag{1}$$

Here, $p$ is the allele frequency among the parents and $p'$ is that among the children. Equation 1 tells us $p$'s value in one generation as a function of its value in the previous generation. Such equations are called *recurrence equations*. They provide recipes for getting from one generation to the next.

Table 1: Genotypic fitnesses and frequencies

| Genotype | $A_1 A_1$ | $A_1 A_2$ | $A_2 A_2$ |
|---|---|---|---|
| Freq. before selection | $p^2$ | $2pq$ | $q^2$ |
| Fitness | $w_{11}$ | $w_{12}$ | $w_{22}$ |
| Freq. after selection | $p^2 w_{11}/\bar{w}$ | $2pq w_{12}/\bar{w}$ | $q^2 w_{22}/\bar{w}$ |

Where $q = 1 - p$
and $\bar{w} = p^2 w_{11} + 2pq w_{12} + q^2 w_{22}$

# 3 The rate of change

How much change occurs between successive generations? To find out, we calculate $\Delta p$ (pronounced "delta pee"), which equals $p' - p$. $\Delta p$ is positive when $p$ is increasing, and negative when $p$ is decreasing. When there is no tendency to change, the system is said to be at *equilibrium*, and $\Delta p = 0$. Subtracting $p$ from equation 1 gives (after some algebra)

$$\Delta p = \frac{pq}{\bar{w}}[p(w_{11} - w_{12}) + q(w_{12} - w_{22})] \tag{2}$$

This equation provides a great deal of intuition concerning the dynamics of natural selection, which I summarize in the paragraphs that follow.

$\star$ EXERCISE **2** Derive Eqn. 2 from Eqn. 1.

**Result 1** *Response to selection is slow when either allele is rare.*

This is because $\Delta p$ is proportional to $pq$ and is therefore low when $p$ is small or when $q$ is small.

**Result 2** *Selection is especially slow when a recessive allele is rare.*

If $A_1$ is recessive, $w_{12} = w_{22}$, and Eqn. 2 becomes $\Delta p = p^2 q(w_{11} - w_{12})/\bar{w}$. When $A_1$ is rare, $p^2$ is very small, so response to selection is very slow.

**Result 3** *Selection is especially slow when a dominant allele is common.*

If $A_1$ is dominant and common, then $A_2$ is recessive and rare, so Result 2 implies that $\Delta p \approx 0$.

**Result 4** *A rare allele spreads if its heterozygote is fitter than the common homozygote.*

This result is used widely, because all mutations are initially rare, and we would like to know which ones will spread. It holds because when $A_1$ is rare, $p$ is small, the expression in square brackets is approximately $w_{12} - w_{22}$. Allele $A_1$ will increase in frequency if this quantity is positive.

**Result 5** *If the heterozygote has intermediate fitness, then one allele will increase to fixation.*

To see why, note that if $w_{11} > w_{12} > w_{22}$, then the quantity in square brackets in Eqn. 2 is always positive, and $A_1$ will increase to fixation. On the other hand, $A_2$ will fix if $w_{11} < w_{12} < w_{22}$.

**Result 6** *If the heterozygote has the highest fitness, then $A_1$ evolves toward an intermediate equilibrium.*

This is the case of "fitness overdominance" or "heterozygote superiority." In this case, $w_{11} - w_{12} < 0$ and $w_{12} - w_{22} > 0$, so $\Delta p$ is positive when $p$ is small but negative when $p$ is large.

**Result 7** *When the heterozygote has highest fitness, the allele that is most common at the stable equilibrium is the one whose homozygote has highest fitness.*
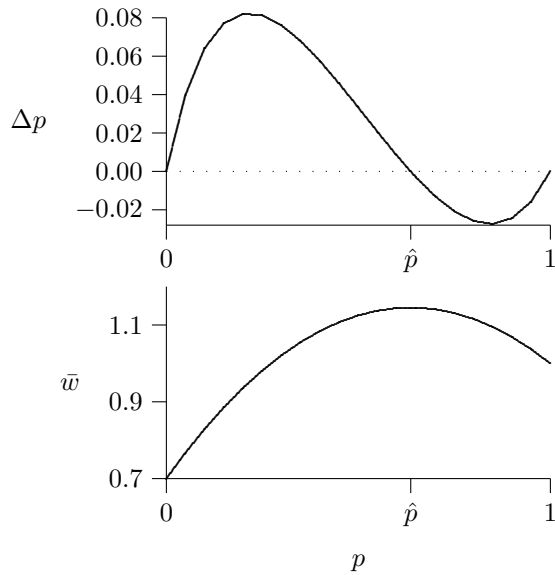
Figure 1: Model of heterozygote advantage. Upper panel: the rate, $\Delta p$, of change in $p$. Equilibria occur where $\Delta p = 0$. The equilibria at $p = 0$ and $p = 1$ are unstable. That at $p \approx 0.64$ is stable. Lower panel: mean fitness, $\bar{w}$, as a function of $p$. The equilibrium value, $\hat{p}$, maximizes $\bar{w}$ and coincides the point at which $\Delta p = 0$.
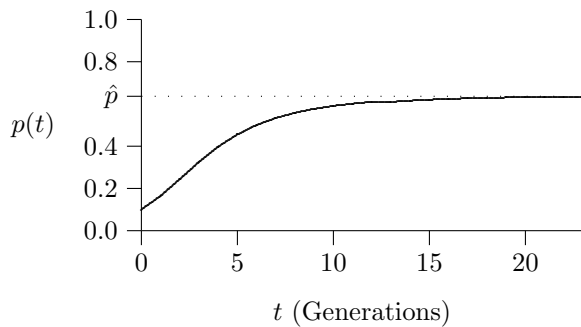


Figure 2: Time path of allele frequency using the same model as Fig. 1. The dotted line indicates the equilibrium value, $\hat{p} \approx 0.64$.

To derive this result, set to zero the quantity in square brackets in Eqn. 2 and rearrange to obtain

$$\frac{\hat{p}}{\hat{q}} = \frac{w_{12} - w_{22}}{w_{12} - w_{11}}$$

where $\hat{p}$ and $\hat{q}$ are the equilibrium values of $p$ and $q$. On the right side of this equation, the numerator and denominator are both positive, because we are concerned with the case of heterozygote superiority. If $w_{11} > w_{22}$, then the denominator will be smaller than the numerator, and this implies that $\hat{p}$ (the frequency of $A_1$) exceeds $\hat{q}$ (that of $A_2$). For example, suppose that $w_{22} = 0.1$, $w_{11} = 0.2$, and $w_{12} = 0.3$. Then

$$\frac{\hat{p}}{\hat{q}} = \frac{0.3 - 0.1}{0.3 - 0.2} = 2$$

and $\hat{p}$ is twice as large as $\hat{q}$.

**Result 8** *If the heterozygote has lowest fitness, then there are two stable equilibria, one at which $A_1$ is fixed and another at which it is lost.*

In this case, $w_{11} - w_{12} > 0$ and $w_{12} - w_{22} < 0$, so $\Delta p$ is negative when $p$ is small but positive when $p$ is large.

## 4   The same result in terms of genic fitnesses

We can also write equation 2 as

$$\Delta p = \frac{pq(w_1 - w_2)}{\bar{w}} \tag{3}$$

where $w_1$ and $w_2$ are the *genic* or *marginal* fitnesses of $A_1$ and $A_2$. The genic fitness of allele $A_i$ is simply the average fitness of copies of $A_i$.

In diploid organisms, genes occur in pairs. Each copy of the $A_1$ allele is paired with some other gene. What is the probability that this second gene is also a copy of $A_1$? If mating is at random, then the second gene is (in effect) drawn at random from among all the genes in the population. Since a fraction $p$ of these are copies of $A_1$, $p$ is also the probability that the second gene is an $A_1$. In this case, the original copy of $A_1$ resides in an $A_1A_1$ homozygote with fitness $w_{11}$. On the other hand, the original $A_1$ might (with probability $q$) be paired with a copy of $A_2$. If so, then the initial $A_1$ resides in a heterozygote with fitness $w_{12}$. The mean fitness of allele $A_1$ is therefore

$$w_1 = pw_{11} + qw_{12}$$

A similar argument applies if the initial gene is a copy of $A_2$ and leads to

$$w_2 = pw_{12} + qw_{22}$$

Since $pq/\bar{w}$ is always positive, equation 3 implies that

**Result 9** *Selection favors the allele whose genic fitness is highest.*

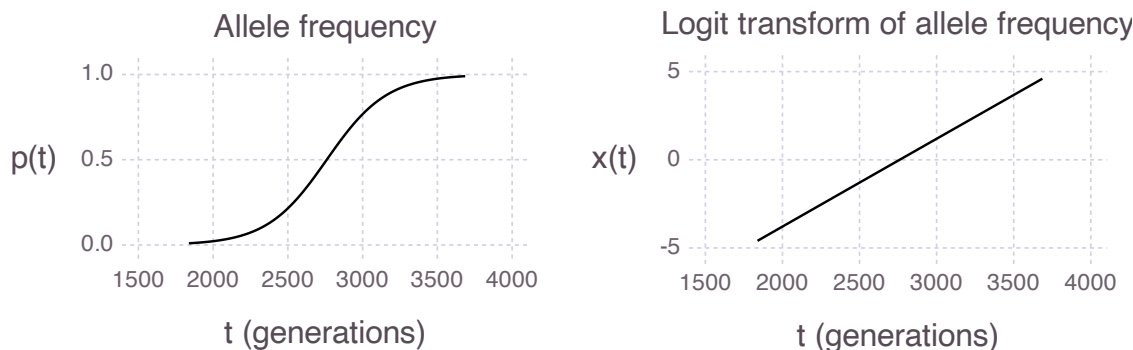$\star$ EXERCISE **3** Verify that $w_1 - w_2$ is equal to $p(w_{11} - w_{12}) + q(w_{12} - w_{22})$.

Figure 3: Time path of $p$, the frequency of an advantageous allele, $A_1$, over the range $0.01 \leq p \leq 0.99$. The model assumes that the fitness of the favored homozygote ($A_1A_1$) exceeds that of the unfavored one ($A_2A_2$) by $s = 0.01$ and that there is no dominance. Calculations used Eqn. 1.

## 5  Wright's equation

We saw above that Eqn. 2 is rich in insight about the behavior of natural selection. Most of that insight came from studying the quantity in square brackets. Sewall Wright [4] noticed that this quantity is equal to $\frac{1}{2}\frac{d\bar{w}}{dp}$, where $d\bar{w}/dp$ is the derivative (or slope) of the curve that describes $\bar{w}$ as a function of $p$. Thus,

$$\Delta p = \frac{pq}{2\bar{w}} \cdot \frac{d\bar{w}}{dp} \tag{4}$$

This equation is equivalent to 2, but provides a different kind of insight. Notice that $pq/2\bar{w}$ is always positive when $p$ is neither zero nor unity. This means that $p$ will increase when $d\bar{w}/dp > 0$ and will decrease when this derivative is negative. Selection pushes allele frequency in a direction that increases $\bar{w}$, the mean fitness. It pushes the population "uphill" on the function $\bar{w}(p)$. Consequently,

**Result 10** *Stable equilibria occur at peaks in the graph of mean fitness against allele frequency.*

The lower panel of Fig. 1 illustrates this principle.

$\star$ EXERCISE **4** Verify that equations 2 and 4 are equivalent.

$\star$ EXERCISE **5** Define $w_{11} = 1$, $w_{12} = 1.1$, and $w_{22} = 0.8$. Find the equilibria in this system by graphing $\Delta p$ as a function of $p$.

## 6  The time required for a selective sweep

How long does it take for an allele to increase from one value to another, under the influence of natural selection? This question has been addressed by Haldane [3] and several later authors [2, sec. 5.3; 1, sec. 3.1.iii.a]. The model is simplest if there is no dominance, so that alleles contribute additively to fitness. Let's suppose that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ have relative fitnesses $w_{11} = 1 + s$, $w_{12} = 1 + s/2$, and $w_{22} = 1$. The left panel of Fig. 3 shows the time path of allele $A_1$ over the range $0.01 \leq p \leq 0.99$. As the graph shows, this selective sweep takes more than 1500 generations.

The graph on the right shows the time path of the transformed variable,

$$x = \ln \frac{p}{1 - p}$$

This is called the "logit transform" and is often used in statistics. As you can see, the graph of $x$ is a straight line, or very nearly so. Appendix A shows that the slope of this line is approximately $s/2$. Consequently, it is easy to write $x$ as a function of $t$:

$$x(t) = x_0 + ts/2$$

where $x_0$ is the initial value of $x$. Solving for $t$ gives [2, Eqn. 5.3.13; 1, Eqn. 3.3]

$$\begin{aligned} t &= \frac{2}{s}(x(t) - x_0) \\ &= \frac{2}{s} \ln \frac{p(t)q_0}{q(t)p_0} \end{aligned} \tag{5}$$

where $p_0$ and $q_0$ are the initial values of $p$ and $q = 1 - p$. This result makes it easy to figure out how long it would take for selection to raise the frequency of an allele from one value to another. It is an approximation for weak selection, but it works remarkably well for values of $s$ up to about 0.1.

Let us use it to ask about the time required for a selective sweep. Equation 5 is deterministic and consequently does not inform us about the initial generations, when the advantageous allele is very rare. Neither does it inform us about the final generations, when the alternate allele is rare. Those dynamics are stochastic and are not captured by our deterministic model. We therefore ignore the beginning and end of the selective sweep and ask how long it takes for the advantageous allele to increase from 1% to 99%. For this case, Eqn. 5 becomes $t \approx 18/s$.

**Result 11** *It takes roughly $18/s$ generations for an advantageous allele to increase in frequency from 0.01 to 0.99, if there is no dominance.*

For example, in Fig. 3, $s = 1/100$, so the selective sweep lasts about 1800 generations. If generations are 25 years in length, this corresponds to about 45 thousand years. If there is dominance, the time required for a selective sweep is longer [2, p. 193; 1, Box 3.2].

$\star$ EXERCISE **6** How many generations are required for selection to increase the frequency of an allele from 1% to 99%, if the coefficient of selection is $s = 0.001$, and there is no dominance?

$\star$ EXERCISE **7** Repeat exercise 6, but this time figure out the time required for the allele to increase in frequency from 25% to 75%.

## 7 Summary

This note introduces the conventional algebraic model of natural selection. Equilibria occur where $\Delta p = 0$. There are always equilibria at $p = 0$ and $p = 1$. There may also be an intermediate equilibrium at which $0 < p < 1$. Stable equilibria occur at peaks in the graph of mean fitness against $p$.

When the heterozygote has intermediate fitness, selection favors the allele whose homozygote has the higher fitness. A rare allele will spread if the fitness of its heterozygote is higher than that of the common homozygote. When the heterozygote's fitness is higher than that of either homozygote, the system evolves toward a stable equilibrium at which $p$ is greater than zero but less than unity. When the heterozygote has lowest fitness, the system evolves away from an unstable intermediate equilibrium and toward either $p = 0$ and $p = 1$.

When gene effects are additive, the selective sweep of an advantageous allele takes roughly $18/s$ generations.

## Acknowledgements

## A  The slope of the straight line in Fig. 3

In sec. 6, we assumed that genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$ had relative fitnesses $w_{11} = 1 + s$, $w_{12} = 1 + s/2$, and $w_{22} = 1$. With these definitions, Eqn. 2 becomes $\Delta p = pqs/2\bar{w}$, which is approximately $pqs/2$ if selection is weak. Under weak selection, the change in each generation will be small, and we can treat $p$ as a continuous function of time, and $\Delta p$ as a derivative:

$$dp/dt = pqs/2$$

As we saw above in Fig. 3, the time path is much simpler if we transform $p$ into

$$x = \log \frac{p}{q} = \log p - \log q$$

where $q = 1 - p$. The derivative of $x$ is

$$
\begin{aligned}
dx/dt &= \frac{1}{p}\frac{dp}{dt} + \frac{1}{q}\frac{dp}{dt} = \frac{pqs}{2}\left(\frac{1}{p} + \frac{1}{q}\right) = \frac{pqs}{2}\left(\frac{q}{pq} + \frac{p}{pq}\right) \\
&= s/2
\end{aligned}
$$

Because $s$ is a constant, this result says that the graph of $x$ against $t$ is a straight line with slope $s/2$.

## B  Answers to Exercises

$\star$ EXERCISE **1** The population contains $NP_{11}$ individuals of genotype $A_1A_1$ and $NP_{12}$ of genotype $A_1A_2$. Each of the former carries two copies of $A_1$ and each of the latter carries one copy. The total number of copies of $A_1$ is thus $2NP_{11} + NP_{12}$. Dividing by $2N$ (the total number of genes) gives the relative frequency of $A_1$ within the population: $p = P_{11} + \frac{1}{2}P_{12}$.

$\star$ EXERCISE **2**

$$
\begin{aligned}
\Delta p &= \frac{p^2 w_{11} + pq w_{12}}{\bar{w}} - p \\
&= \frac{p^2 w_{11} + pq w_{12} - p^3 w_{11} - 2p^2 q w_{12} - pq^2 w_{22}}{\bar{w}} \\
&= \frac{p^2 q w_{11} + pq(q - p) w_{12} - pq^2 w_{22}}{\bar{w}} \\
&= \frac{pq}{\bar{w}}[p w_{11} + (q - p) w_{12} - q w_{22}] \\
&= \frac{pq}{\bar{w}}[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]
\end{aligned}
$$

$\star$ EXERCISE **3** The calculation is very simple:

$$
\begin{aligned}
w_1 - w_2 &= p w_{11} + q w_{12} - q w_{22} - p w_{12} \\
&= p(w_{11} - w_{12}) + q(w_{12} - w_{22})
\end{aligned}
$$

$\star$ EXERCISE **4** Mean fitness is

$$
\begin{aligned}
\bar{w} &= p^2 w_{11} + 2pq w_{12} + q^2 w_{22} \\
&= p^2 w_{11} + 2(p - p^2) w_{12} + (1 - 2p + p^2) w_{22}
\end{aligned}
$$

Taking the derivative,

$$
\frac{d\bar{w}}{dp} = 2p w_{11} + (2 - 4p) w_{12} + (-2 + 2p) w_{22}
$$

$$
\begin{aligned}
\frac{1}{2}\frac{d\bar{w}}{dp} &= p w_{11} + (1 - 2p) w_{12} - (1 - p) w_{22} \\
&= p w_{11} + (q - p) w_{12} - q w_{22} \\
&= p(w_{11} - w_{12}) + q(w_{12} - w_{22})
\end{aligned}
$$

which equals the quantity in square brackets in Eqn. 2.

$\star$ EXERCISE **5** There are unstable equilibria at 0 and 1, and a stable equilibrium at 3/4.

$\star$ EXERCISE **6** 18,380 generations

$\star$ EXERCISE **7** 4394 generations

# References

[1] Brian Charlesworth and Deborah Charlesworth. *Elements of Evolutionary Genetics.* Roberts, 2010.

[2] James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory.* Harper and Row, New York, 1970.

[3] J. B. S. Haldane. *The Causes of Evolution.* Cornell University Press, Ithaca, 1932.

[4] Sewall Wright. *The Theory of Gene Frequencies*, volume 2 of *Evolution and the Genetics of Populations.* University of Chicago Press, Chicago, 1969.