

HST.508 HW1

David (davidgoh@mit.edu)

Table of contents

1 Problem 1 Mutation and Drift for the Two-Allele Model (6 points)	1
1.1 Problem 1 a	1
1.2 Problem 1 b	2
1.3 Problem 1 C	3
2 Problem 2 Simulations of Genetic Drift (8 points)	6
2.1 Problem 2 a	6
2.2 Problem 2 b	6
2.3 Problem 2 c	7
2.4 Problem 2 d	8
2.5 Problem 2 e	9
3 Problem 3 Parameters and population genetics (3 points)	10
3.1 Problem 3 a	11
3.2 Problem 3 b	12
3.3 Problem 3 c	12
4 Problem 4 Effective population size. (3 points)	13
5 Problem 5 Extra Credit: Mean time to either fixation or loss (10 points)	14
5.1 Problem 5 (a)	15
5.2 Problem 5 (b)	15
5.3 Problem 5 (c)	16

Name: David David

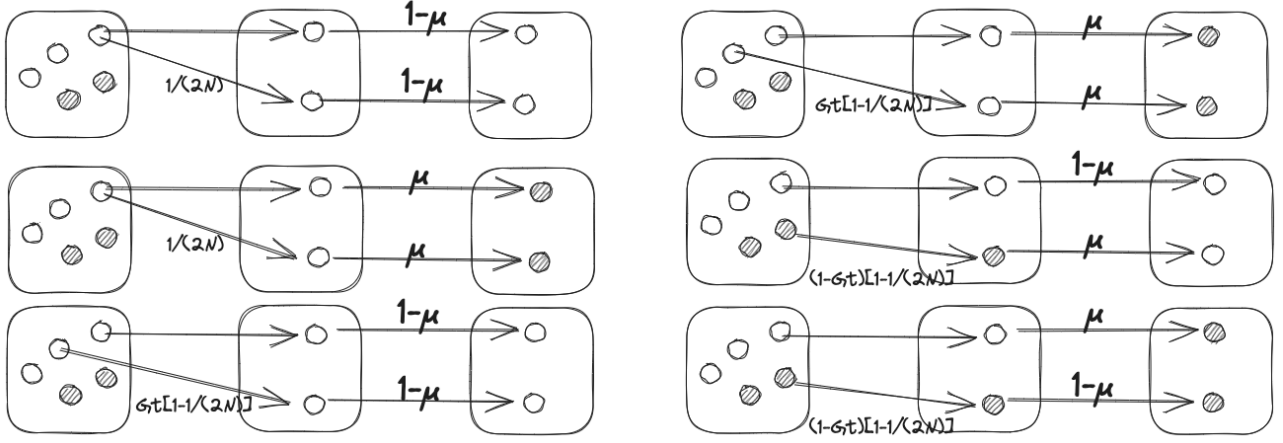
Kerberos: davidgoh

1 Problem 1 Mutation and Drift for the Two-Allele Model (6 points)

Consider the evolution of heterozygosity H_t (and homozygosity G_t) in a population subject to a random drift and mutations. Use the approach developed in class to consider a two-allele model, which has the same rate of mutation, μ , for alleles $A_2 \rightarrow A_1$ and $A_1 \rightarrow A_2$.

1.1 Problem 1 a

Obtain an expression for G_{t+1} as a function of G_t , the mutation rate, and population size. To get this expression, consider all possible ways for getting two identical alleles in the $t + 1$ generation, given that homozygosity (probability of two identical alleles) in generation t was G_t .



$$\begin{aligned}
 G_{t+1} &= \frac{1}{2N}(1-\mu)^2 \\
 &\quad + \frac{1}{2N}\mu^2 \\
 &\quad + \left(1 - \frac{1}{2N}\right) G_t(1-\mu)^2 \\
 &\quad + \left(1 - \frac{1}{2N}\right) G_t\mu^2 \\
 &\quad + \left(1 - \frac{1}{2N}\right) (1-G_t)(1-\mu)\mu \\
 &\quad + \left(1 - \frac{1}{2N}\right) (1-G_t)\mu(1-\mu)
 \end{aligned}$$

1.2 Problem 1 b

Expand the obtained expression while dropping the terms higher than the first power of or $1/N$. Obtain an expression for the steady state heterozygosity H_{ss} in the two-allele model. How does the change in heterozygosity in one generation subject to mutations and drift compare to the case due to drift only?

We drop terms higher than the first power of μ or $\frac{1}{2N}$ since $\mu \ll 1$ and $N \gg 1$, including $\frac{\mu}{N}$.

- $(1-\mu)^2 \approx 1-2\mu$
- $\mu^2 \approx 0$
- $(1-\mu)\mu \approx \mu$

$$\begin{aligned}
 G_{t+1} &= \frac{1}{2N}(1-2\mu) + \left(1 - \frac{1}{2N}\right) G_t(1-2\mu) + \left(1 - \frac{1}{2N}\right) (1-G_t)\mu + \left(1 - \frac{1}{2N}\right) (1-G_t)\mu \\
 &= \frac{1}{2N}(1-2\mu) + \left(1 - \frac{1}{2N}\right) G_t(1-2\mu) + 2\left(1 - \frac{1}{2N}\right) (1-G_t)\mu \\
 &= \left(\frac{1}{2N} - \frac{\mu}{N}\right) + \left(1 - \frac{1}{2N} - 2\mu + \frac{\mu}{N}\right) G_t + 2\left(\mu - \frac{\mu}{2N}\right) (1-G_t) \\
 &\approx \frac{1}{2N} + \left(1 - \frac{1}{2N} - 2\mu\right) G_t + 2\mu(1-G_t) \\
 &= \frac{1}{2N} + G_t - \frac{1}{2N}G_t - 2\mu G_t + 2\mu - 2\mu G_t \\
 G_{t+1} &= \frac{1}{2N} + G_t - \frac{1}{2N}G_t + 2\mu - 4\mu G_t
 \end{aligned}$$

$$\begin{aligned}
(1 - H_{t+1}) &= \frac{1}{2N} + (1 - H_t) - \frac{1}{2N}(1 - H_t) + 2\mu - 4\mu(1 - H_t) \\
H_{t+1} &= 1 - \frac{1}{2N} - (1 - H_t) + \frac{1}{2N}(1 - H_t) - 2\mu + 4\mu(1 - H_t) \\
&= 1 - \frac{1}{2N} - 1 + H_t + \frac{1}{2N} - \frac{H_t}{2N} - 2\mu + 4\mu - 4\mu H_t \\
&= H_t + H_t \left(-\frac{1}{2N} - 4\mu \right) + 2\mu \\
\Delta H_t &= H_t \left(-\frac{1}{2N} - 4\mu \right) + 2\mu
\end{aligned}$$

At steady state, $\Delta H_t = 0$

$$\begin{aligned}
0 &= H_{ss} \left(-\frac{1}{2N} - 4\mu \right) + 2\mu \\
H_{ss} &= \frac{2\mu}{\frac{1}{2N} + 4\mu}
\end{aligned}$$

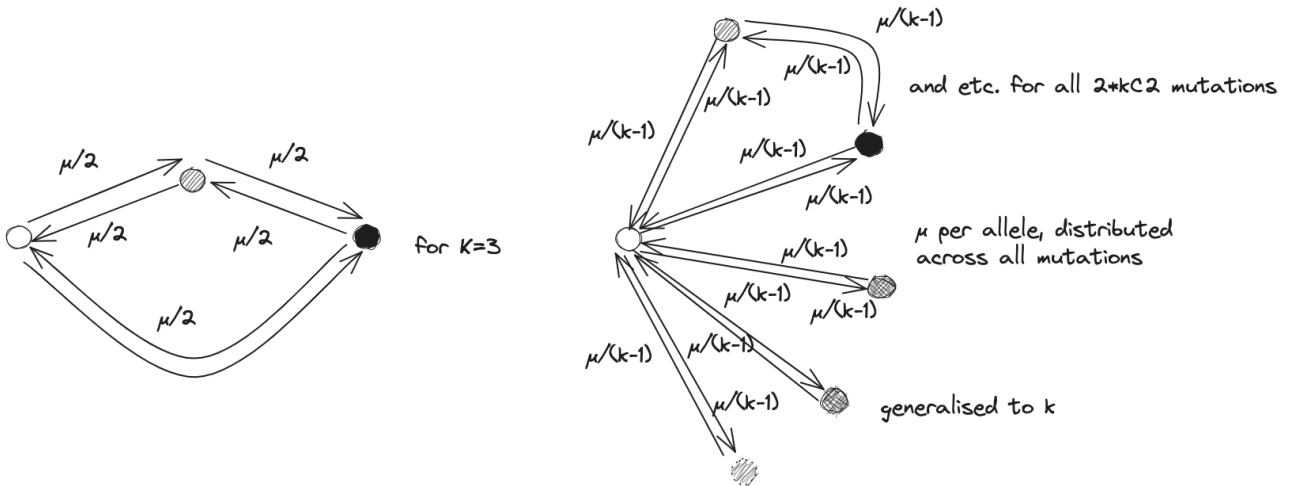
Comparing mutation and drift (M + D) with drift only (D):

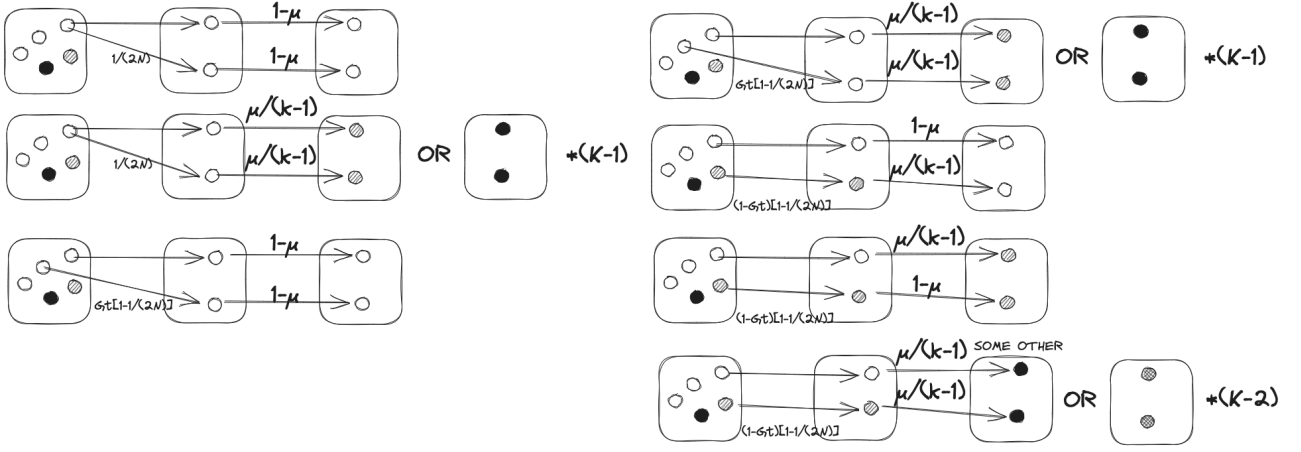
$$\begin{aligned}
\Delta H_t^D &= -\frac{H_t}{2N} \\
\Delta H_t^{M+D} &= H_t \left(-\frac{1}{2N} - 4\mu \right) + 2\mu \\
\Delta H_t^{M+D} &= \underbrace{-\frac{H_t}{2N}}_{\text{Drift}} \underbrace{- 4\mu H_t + 2\mu}_{\text{Mutation}}
\end{aligned}$$

The reversible mutation adds a $(-4\mu H_t + 2\mu)$ term to the change in heterozygosity in one generation.

1.3 Problem 1 C

Extra credit [+2]: Now consider a k-allele model with the same rate for all possible mutations $A_i \rightarrow A_j$. How does the number of alleles affect the steady state heterozygosity? Consider limits of $k = 2$ and $k \rightarrow \infty$.





$$\begin{aligned}
G_{t+1} &= \frac{1}{2N}(1-\mu)^2 \\
&+ \frac{1}{2N} \left(\frac{\mu}{k-1} \right)^2 (k-1) \\
&+ \left(1 - \frac{1}{2N} \right) G_t (1-\mu)^2 \\
&+ \left(1 - \frac{1}{2N} \right) G_t \left(\frac{\mu}{k-1} \right)^2 (k-1) \\
&+ \left(1 - \frac{1}{2N} \right) (1-G_t)(1-\mu) \left(\frac{\mu}{k-1} \right) \\
&+ \left(1 - \frac{1}{2N} \right) (1-G_t) \left(\frac{\mu}{k-1} \right) (1-\mu) \\
&+ \left(1 - \frac{1}{2N} \right) (1-G_t) \left(\frac{\mu}{k-1} \right)^2 (k-2)
\end{aligned}$$

We drop terms higher than the first power of μ or $\frac{1}{2N}$ since $\mu \ll 1$ and $N \gg 1$.

- $(1-\mu)^2 \approx 1-2\mu$
- $\left(\frac{\mu}{k-1} \right)^2 \approx 0$
- $(1-\mu) \left(\frac{\mu}{k-1} \right) \approx \left(\frac{\mu}{k-1} \right)$

$$\begin{aligned}
G_{t+1} &= \frac{1}{2N}(1-2\mu) \\
&+ \left(1 - \frac{1}{2N} \right) G_t (1-2\mu) \\
&+ \left(1 - \frac{1}{2N} \right) (1-G_t) \left(\frac{\mu}{k-1} \right) \\
&+ \left(1 - \frac{1}{2N} \right) (1-G_t) \left(\frac{\mu}{k-1} \right)
\end{aligned}$$

Ignoring $\frac{\mu}{N}$ terms:

$$\begin{aligned}
1 - H_{t+1} &= \frac{1}{2N} + \left(1 - \frac{1}{2N} - 2\mu\right) (1 - H_t) + \frac{2\mu}{k-1} H_t \\
&= \frac{1}{2N} + 1 - \frac{1}{2N} - 2\mu - H_t + \frac{1}{2N} H_t + 2\mu H_t + \frac{2\mu}{k-1} H_t \\
1 - H_{t+1} &= 1 - 2\mu - H_t + H_t \left(\frac{1}{2N} + 2\mu + \frac{2\mu}{k-1} \right) \\
H_{t+1} &= 2\mu + H_t - H_t \left(\frac{1}{2N} + \frac{2\mu k}{k-1} \right) \\
\Delta H_t &= 2\mu + \left(-\frac{1}{2N} - \frac{2\mu k}{k-1} \right) H_t \\
0 &= 2\mu + \left(-\frac{1}{2N} - \frac{2\mu k}{k-1} \right) H_{ss} \\
H_{ss} &= \frac{2\mu}{\frac{1}{2N} + \frac{2\mu k}{k-1}}
\end{aligned}$$

For $k = 2$,

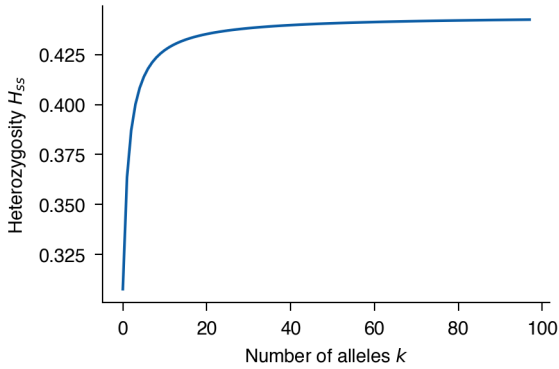
$$\begin{aligned}
H_{ss} &= \frac{2\mu}{\frac{1}{2N} + 4\mu} \\
H_{ss} &= \frac{4\mu N}{1 + 8\mu N}
\end{aligned}$$

consistent with the expression in 2b.

For $k \rightarrow \infty$,

$$\begin{aligned}
H_{ss} &= \frac{2\mu}{\frac{1}{2N} + \frac{2\mu k}{k-1}} \\
&= \frac{2\mu}{\frac{1}{2N} + \frac{2\mu}{1-\frac{1}{k}}} \\
&= \frac{2\mu}{\frac{1}{2N} + 2\mu} \\
H_{ss} &= \frac{4\mu N}{1 + 4\mu N}
\end{aligned}$$

The number of alleles k increases the steady state heterozygosity asymptotically.

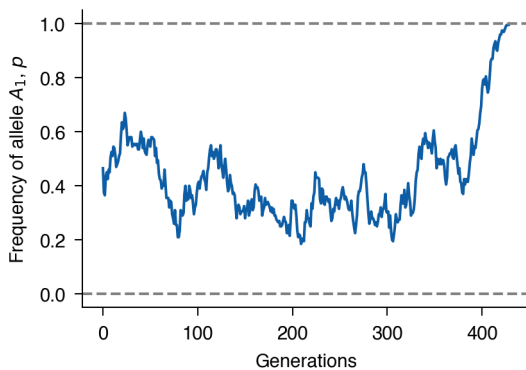


2 Problem 2 Simulations of Genetic Drift (8 points)

Start with a population of $N = 100$ diploid individuals, each of which has two homologous chromosomes, giving you a total of $2N$ chromosomes to track. Each chromosome has a single polymorphic locus (SNP) that can be in one of two states. To code this, create an array of integers, of length $2N$, and set the value of each element to either 1 or 2 depending on the allele of this chromosome. In the initial population, alleles 1 and 2 are equally abundant, i.e. $p = q = 0.5$. Please take a look at the included starter code if you need guidance (`pset1_sample_code.py`).

2.1 Problem 2 a

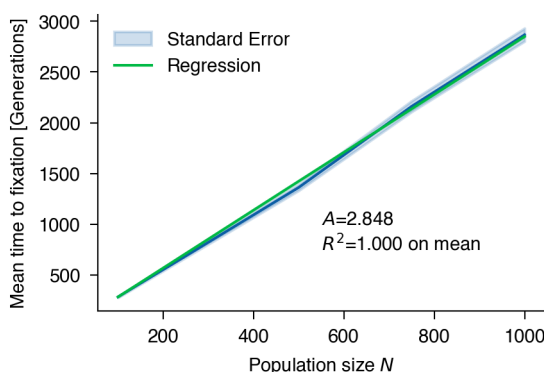
Simulate drift in one population. Generate the next generation by drawing individuals at random from the current generation with replacement. Compute heterozygosity for each generation. Continue until all individuals become of one type – this is called fixation. Show a plot of heterozygosity (p) over time for 1 trajectory.



2.2 Problem 2 b

Simulate 1000 trajectories, keeping a record of the fixation time for each simulation. Compute the mean time to fixation and compare it to N . You can run your simulations at 3-5 different values of N to establish this dependence better.

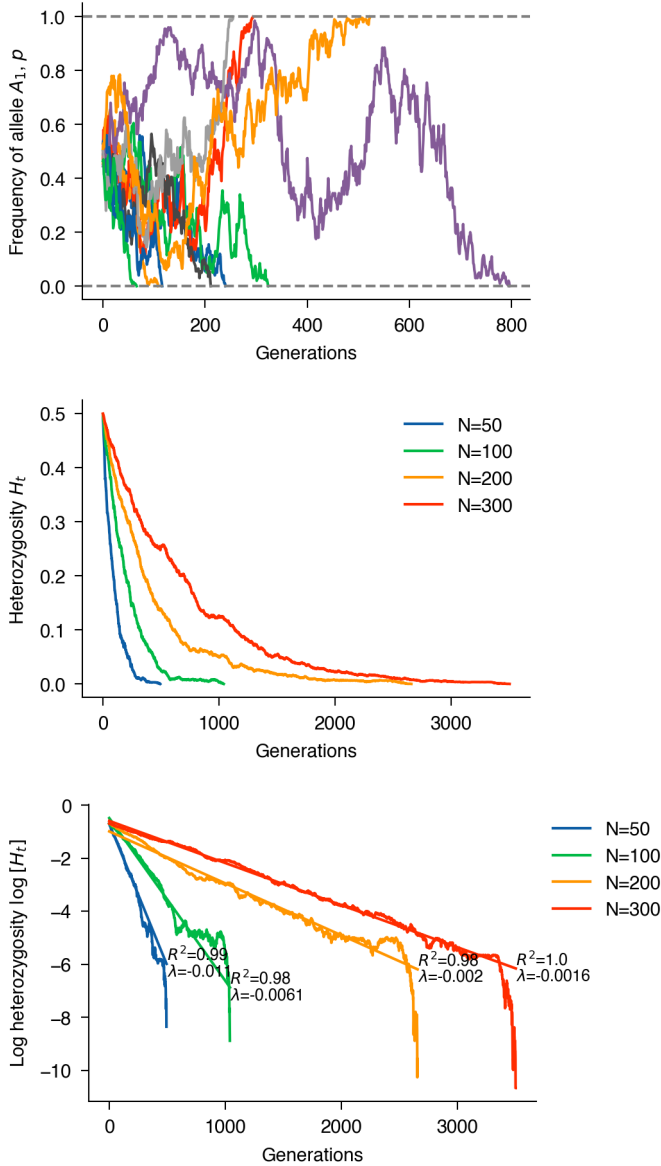
$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t \approx H_0 \exp\left[-\frac{t}{2N}\right]$$



The mean time to fixation is proportional to the population size. This result aligns with our estimates from the characteristic time $\tau \sim 2N$ from the exponential decay of the heterozygosity.

2.3 Problem 2 c

Plot 10 individual trajectories of heterozygosity vs. time on a single graph. In addition, for each time point, compute the average heterozygosity by averaging over all 1000 trajectories and plot the average heterozygosity vs time. To better see the exponential decay of heterozygosity you can plot the log of heterozygosity vs time. Measure the rate of decay of heterozygosity (the slope on the log-linear plot). Compare it to N .

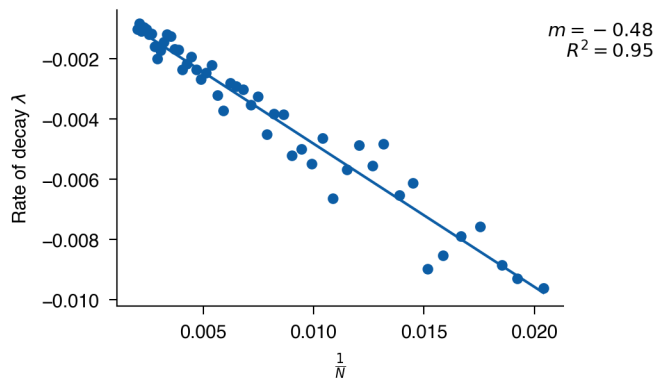


The log heterozygosity is largely linear with negative gradient and implies an exponential decay of the heterozygosity. Let the rate of decay of heterozygosity be λ . For random mating, fixed population sizes, no mutation, no selection, we have:

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

$$\log H_t = \log H_0 - \frac{1}{2N}t$$

$$\log H_t = \log H_0 + \lambda t$$

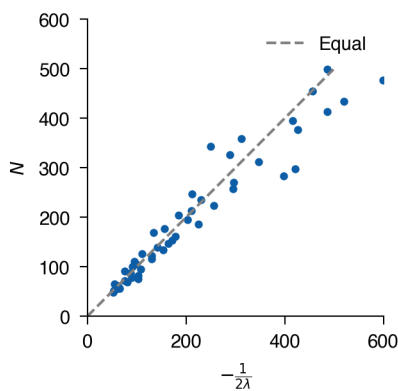


Plotting λ against $\frac{1}{N}$, we observe the gradient of $-\frac{1}{2}$:

$$\lambda = -\frac{1}{2} \frac{1}{N}$$

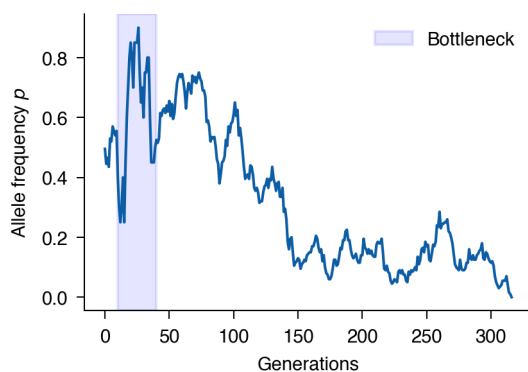
Practically speaking, we can estimate the population size from the rate of decay of heterozygosity λ :

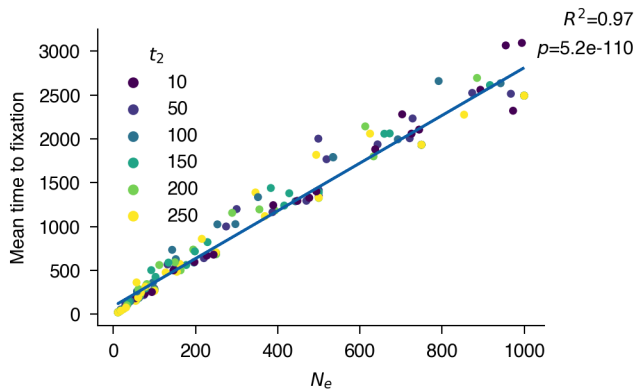
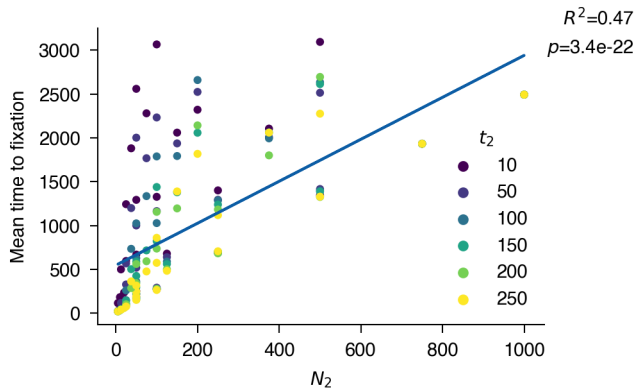
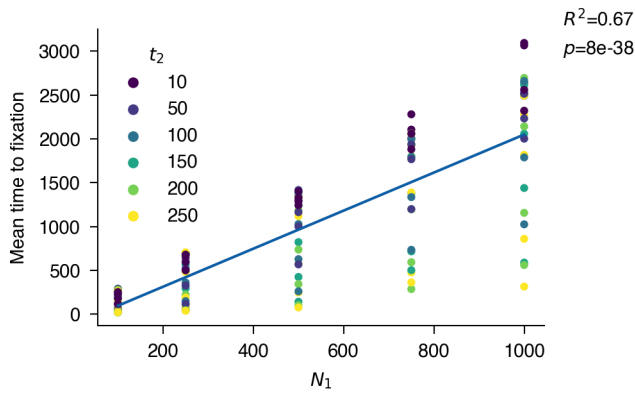
$$\lambda = -\frac{1}{2} \frac{1}{N} \Rightarrow N = -\frac{1}{2\lambda}$$



2.4 Problem 2 d

Consider an effect of changing population size on the time to fixation. Run simulations first at some $N = N_1$ and then change N to $N = N_2$ for t_2 number of generations then back to N_1 . Try $N_2 \ll N_1$, e.g. $N_2 = N_1/10$. Make 1000 runs and compute the mean time to fixation. Try several N_1 and N_2 . How does decrease in the population size change the mean time to fixation? Compute N_{eff} for each parameters (N_1 , N_2 , t_2), and correlated mean time to fixation with N_1 , N_2 , and N_{eff} . Interpret your findings.



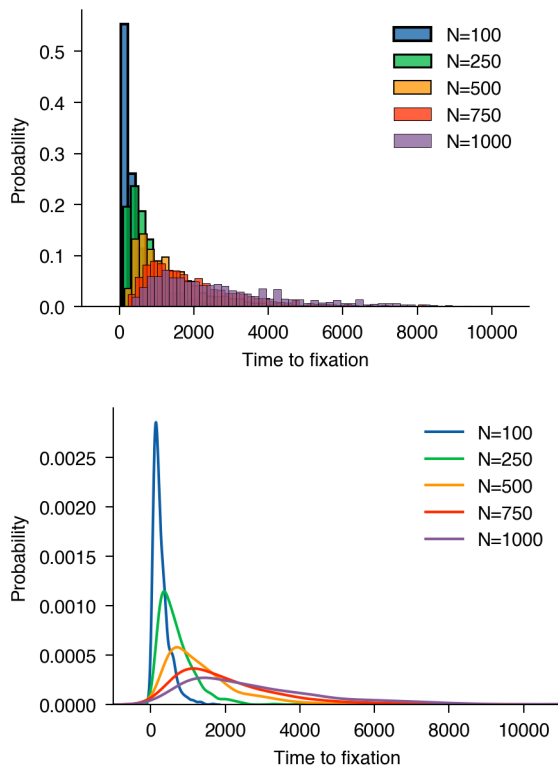


The mean time to fixation is correlated with all three of N_1 , N_2 , and N_e , but N_e is the best linear predictor of the mean time to fixation. N_e best explains the variation in mean time to fixation because it incorporates the number of generations experiencing the bottleneck, t_2 , through the harmonic mean.

The decreases in the population size decreases the mean time to fixation by increasing the variance of the steps and increasing the probability to fixation or loss. The longer the decrease (the greater the t_2) the lower the mean time to fixation for a given N_2 .

2.5 Problem 2 e

[Extra credit: +2 points] Plot the distribution of the fixation time for each N . How broad is the distribution? Think of various ways to quantify this. Study the mean time to fixation as a function of the number K of alleles (types of individuals), starting with $1/K$ initial frequency.

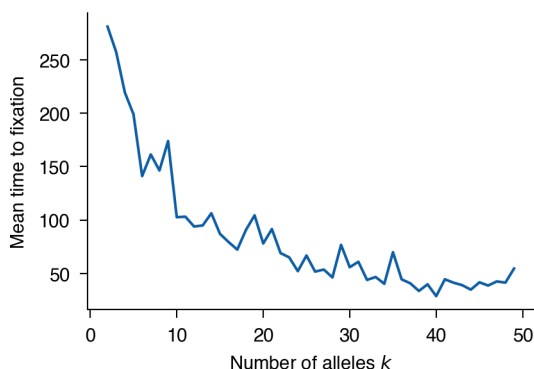


The mean and variance of the distribution increases with N . We can quantify the broadness through the coefficient of variance.

$N = 1000$, $CV = 0.7536712324154138$
 $N = 500$, $CV = 0.7510492725102286$
 $N = 100$, $CV = 0.7179094134418806$

For a given population size N , the mean time to fixation or loss decreases as the number K alleles increases. Here, we implement a new drift function with k alleles.

```
Text(0, 0.5, 'Mean time to fixation')
```



3 Problem 3 Parameters and population genetics (3 points)

A researcher is trying to understand the genetic mutations that lead to resistance to a new antibiotic they have discovered, called techacillin. To do this, they decide to grow many replicate populations of *E. coli* in a morbidostat (an experimental device that keeps

a constant population size and growth rate, while increasing antibiotic concentration as bacteria become more resistant) until they have reached a high level of techacillin resistance. They will then sequence the final and intermediate mutants and identify the mutations that emerged and their order.

3.1 Problem 3 a

The researcher wants to observe the full diversity of possible resistance mutations, including those with relatively minor, but significant, effects on resistance ($>10\%$ increase in growth rate in presence of low levels of drug). To do so, they will need to minimize clonal interference between mutants. They estimate that during every replication, 10^{-5} new mutations emerge that provide significant resistance (as by the aforementioned criteria). What is the maximum population size they can use without interference occurring?

We can estimate the maximum population size directly based on the rate of emergence of beneficial mutations. The rate of beneficial mutations that provide resistance emerging is:

$$\mu_b = 10^{-5} \text{ per generation per locus}$$

The rate of acquiring new beneficial mutations per generation is:

$$n_b = N\mu_b$$

Clonal interference occurs when more than one beneficial mutation occurs per generation:

$$n_b = N\mu_b > 1$$

The maximum population size to avoid clonal interference is:

$$\begin{aligned} N\mu_{\text{res}} &= 1 \\ N &= \frac{1}{\mu_{\text{res}}} \\ N &= 10^5 \end{aligned}$$

Another way of estimating the maximum population size is to compare the time to fixation and loss. The time to fixation for strongly beneficial mutations in a haploid is:

$$T_{\text{fixation}} = \frac{\ln N}{s}$$

The probability for a mutation to emerge and fix is:

$$P_{\text{occur}} = 2N\mu_b s$$

The time for a mutation to emerge and fix:

$$T_{\text{occur}} = \frac{1}{2N\mu_b s}$$

To avoid clonal interference, we want beneficial mutations to arise (occur) and then have enough time to become fixed before another beneficial mutation emerges.

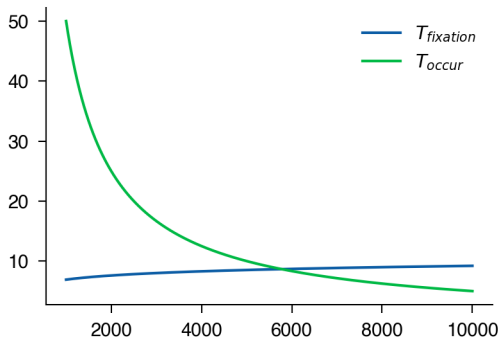
Clonal interference occurs when a beneficial mutation that emerges does not have enough time to become fixed before another one emerges, that is, the time to occurrence is less than the time to fixation:

$$T_{\text{occur}} < T_{\text{fixation}}$$

We want to operate in the regime of no clonal interference, where new beneficial mutations fix fast enough before another emerges:

$$\begin{aligned}
T_{\text{occur}} &> T_{\text{fixation}} \\
T_{\text{occur}} - T_{\text{fixation}} &> 0 \\
\frac{1}{2N\mu_b s} - \frac{\ln N}{s} &> 0 \\
\frac{1}{s} \left(\frac{1}{2N\mu_b} - \ln N \right) &> 0 \\
\frac{1}{2N\mu_b} - \ln N &> 0 \\
N &> 5773
\end{aligned}$$

5773.037773037773



This corresponds to $N < 5773$. This is a more conservative estimate of the maximum population size than $N < 10000$.

3.2 Problem 3 b

Why would the researcher want to use the highest possible population size within this regime?

The rate of acquiring new beneficial mutations per generation is:

$$n_b = N\mu_{\text{res}}$$

Increasing the population size results in a greater flux of new mutations per generation into the population. To maximise the probability of appearance of new mutations, the researcher should maximise population size at the limit of the regime of no clonal interference ($N < 10^5$).

3.3 Problem 3 c

How will the types of observed mutations (in terms of fitness effect) change if they use a higher population size?

Using a higher population size ($N > 10000$ or $N > 5773$), clonal interference will occur. In this regime, only the mutation with the highest fitness will win and be observed. The diversity of mutations decreases because only one type of mutation is observed, at least before a mutation with even higher fitness appears.

4 Problem 4 Effective population size. (3 points)

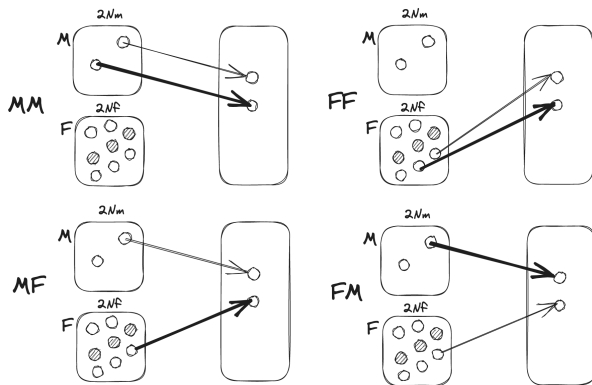
Derive a formula for the effective population size, N_e , if the number of males (N_m) and number of females (N_f) are not equal. (Hint: use the same method we used to derive N_e for the case of a population that is fluctuating in size.)

To fix a concrete case, consider the following example. Imagine a zoo population of primates with 20 males and 20 females. Due to the dominance hierarchy, only one of the males actually breeds. What is the relevant population size that informs us about the strength of drift in this system? Is it 40? 21? To get the answer, compute the probability that two genes drawn at random are alike (identical by descent) in this new situation, depending on random draws of genes from both the males (actually just one male) and 20 females

There are four permutations of taking alleles from the male and female pool. Assume a large population size, such that the distribution of alleles in male and female populations are the same (homogeneous).

The probability of selecting the same kind of allele from the same sex pool (P_{MM} or P_{FF}) should have the same expression as the single population case, just with N_m or N_f .

The probability of selecting the same kind of allele from the opposite sex pool is simply the homozygosity.



Assume hermaphrodite diploid individuals, random mating, no selection, and no migration.

$$\begin{aligned}
G_{t+1} &= P_{MM} + P_{MF} + P_{FM} + P_{FF} \\
&= \frac{1}{4} \left[\frac{1}{2N_m} + \left(1 - \frac{1}{2N_m}\right) G_t \right] \\
&\quad + \frac{1}{4}(1)G_t \\
&\quad + \frac{1}{4}(1)G_t \\
&\quad + \frac{1}{4} \left[\frac{1}{2N_f} + \left(1 - \frac{1}{2N_f}\right) G_t \right] \\
G_{t+1} &= \frac{1}{4} \left(\frac{1}{2N_m} + \frac{1}{2N_f} \right) + \frac{1}{4} \left(4 - \frac{1}{2N_m} - \frac{1}{2N_f} \right) G_t \\
G_{t+1} &= \frac{2N_m + 2N_f}{16N_m N_f} + \left(1 - \frac{2N_m + 2N_f}{16N_m N_f} \right) G_t \\
1 - H_{t+1} &= \frac{2N_m + 2N_f}{16N_m N_f} + \left(1 - \frac{2N_m + 2N_f}{16N_m N_f} \right) (1 - H_t) \\
1 - H_{t+1} &= \frac{2N_m + 2N_f}{16N_m N_f} + 1 - H_t - \frac{2N_m + 2N_f}{16N_m N_f} + \frac{2N_m + 2N_f}{16N_m N_f} H_t \\
-H_{t+1} &= -H_t + \frac{2N_m + 2N_f}{16N_m N_f} H_t \\
H_{t+1} &= H_t \left(1 - \frac{2N_m + 2N_f}{16N_m N_f} \right)
\end{aligned}$$

Using the effective population size N_e as an approximation:

$$\begin{aligned}
H_{t+1} &\approx H_t \left(1 - \frac{1}{2N_e} \right) \\
2N_e &= \frac{16N_m N_f}{2N_m + 2N_f} \\
N_e &= \frac{8N_m N_f}{2N_m + 2N_f} \\
N_e &= \frac{4N_m N_f}{N_m + N_f}
\end{aligned}$$

For $N_m = 1$ and $N_f = 20$,

$$N_e = \frac{4N_m N_f}{N_m + N_f} = \frac{4(1)(20)}{1 + 20} = 3.81$$

5 Problem 5 Extra Credit: Mean time to either fixation or loss (10 points)

The mean time of either fixation or loss $t(p)$ can be derived similar to the way probability of fixation was derived in the class. Consider a two-allele model, an allele with initial frequency p in a population of N diploids without selection or mutations. Derive a recursive relationship for the mean time to fixation-or-loss, which only slightly differs from the recursive relationship from the probability of fixation

The probability of fixation is the mean probability over all possible branches in the next time step.

$$\Pi(p) = \sum_{\Delta p} \Pi(p + \Delta p) \text{Prob}(\Delta p)$$

Likewise, the time to fixation is the mean time over all branches in the next step, but we need to account for stepping into the future by adding 1.

$$t(p) = 1 + \sum_{\Delta p} t(p + \Delta p) \text{Prob}(\Delta p)$$

5.1 Problem 5 (a)

To obtain ODE for $t(p)$ do Taylor expansion while keeping only two leading terms.

$$t(p + \Delta p) = t(p) + t'(p)\Delta p + \frac{1}{2}t''(p)\Delta p^2 + \dots$$

Ignoring all terms higher than Δp^2 :

$$\begin{aligned} t(p) &= 1 + \sum_{\Delta p} t(p + \Delta p) \text{Prob}(\Delta p) \\ &\approx 1 + \sum_{\Delta p} \left[t(p) + t'(p)\Delta p + \frac{1}{2}t''(p)\Delta p^2 \right] \text{Prob}(\Delta p) \\ &= 1 + t(p) \sum_{\Delta p} \text{Prob}(\Delta p) + t'(p) \sum_{\Delta p} \Delta p \text{Prob}(\Delta p) + \frac{1}{2}t''(p) \sum_{\Delta p} \Delta p^2 \text{Prob}(\Delta p) \\ t(p) &= 1 + t(p) + t'(p) \langle \Delta p \rangle + \frac{1}{2}t''(p) \langle \Delta p^2 \rangle \\ 0 &= 1 + t'(p) \langle \Delta p \rangle + \frac{1}{2}t''(p) \langle \Delta p^2 \rangle \end{aligned}$$

For no selection or mutation,

$$\begin{aligned} \langle \Delta p \rangle &= 0 \\ \langle \Delta p^2 \rangle &= \frac{p(1-p)}{2N} \\ 0 &= 1 + \frac{1}{2}t''(p) \frac{p(1-p)}{2N} \end{aligned}$$

5.2 Problem 5 (b)

Write down boundary conditions for $t(p)$ and solve the equation.

$$\begin{aligned} t(0) &= 0 \\ t(1) &= 0 \end{aligned}$$

$$\begin{aligned}
0 &= 1 + \frac{1}{2}t''(p)\frac{p(1-p)}{2N} \\
-1 &= t''(p)\frac{p(1-p)}{4N} \\
t''(p) &= -\frac{4N}{p(1-p)} \\
t''(p) &= -4N\left[\frac{1}{p} + \frac{1}{1-p}\right] \\
t'(p) &= -4N[\ln p - \ln(1-p)] + A \\
t(p) &= -4N[(p \ln p - p) - \{-(1-p) \ln(1-p) - p\}] + Ap + B \\
t(p) &= -4N[p \ln p + (1-p) \ln(1-p)] + Ap + B
\end{aligned}$$

$$t(0) = -4N(0) + 0 + B$$

Hence, $B = 0$

$$t(1) = -4N(0) + A + B$$

Hence, $A = 0$ The solution is:

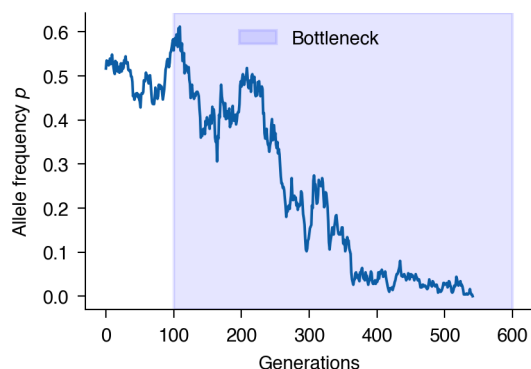
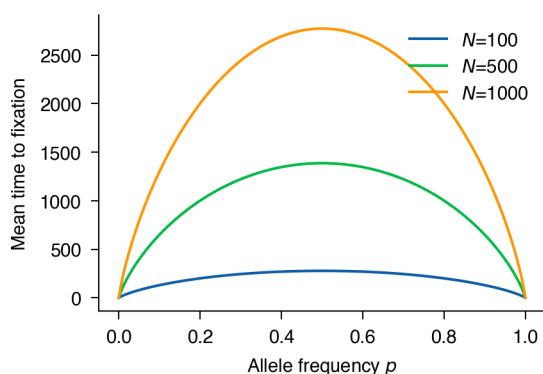
$$t(p) = -4N[p \ln p + (1-p) \ln(1-p)]$$

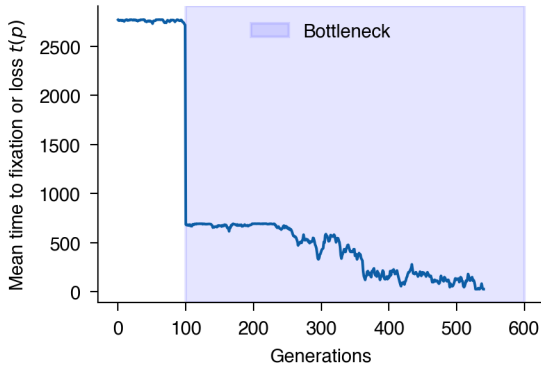
5.3 Problem 5 (c)

Discuss genetic implications of your result. Consider $t(p)$ for a new mutation. For how long does a polymorphism live a population? Discuss the role of the population size, effect of bottlenecks. Make numeric estimates.

The mean time to fixation shows a linear dependency on N , and a “parabolic” dependency on the allele frequency p .

$$t(p) = -4N[p \ln p + (1-p) \ln(1-p)]$$





The mean time to fixation $t(p)$ is directly proportional to the population size N . Hence, bottlenecks decrease the mean time to fixation and speeds up fixation or loss (increases probability of fixation or loss).

For a new mutation,

$$p = \frac{1}{2N}$$

For large population sizes,

$$p \ll 1$$

$$1 - p \approx 1$$

$$\begin{aligned} t(p) &= -4N \left[\frac{1}{2N} \ln \frac{1}{2N} + (1) \ln 1 \right] \\ &= -2 \ln \frac{1}{2N} \\ t(p) &= 2 \ln 2N \end{aligned}$$

The mean time to fixation or loss $t(p)$ has a logarithmic dependency on N . Note that the probability of fixation is $p = 1/2N$ itself, so new mutations are significantly more likely to experience loss for large N .

