Heterozygosity

- One measure of the diversity of a population is its **heterozygosity**.

**Definition** (Heterozygosity).

*Heterozygosity is the probability that two genes chosen at random from the population have different alleles.*

- In a biallelic WF model, the heterozygosity is equal to:

\[
H_t = 2 \frac{X_t}{M} \left(1 - \frac{X_t}{M}\right)
\]

- How does this evolve with time in the WF?
Heterozygosity in the WF

**Theorem** (Heterozygosity under the biallelic WF model).
*Under the biallelic WF model, the expected heterozygosity decays approximately at rate $1/M$ when $M$ is large.*
Heterozygosity in the WF

Proof.

\[ \mathbb{E}(H_{t+1}) = \frac{2}{M^2} \mathbb{E}(X_{t+1}(M - X_{t+1})) \]

\[ = \frac{2}{M^2} \left\{ M \mathbb{E}(X_{t+1}) - \mathbb{E}(X_{t+1}^2) \right\} \]

\[ = \frac{2}{M^2} \left\{ M \mathbb{E}(X_{t+1}) - \text{var}(X_{t+1}) - \mathbb{E}(X_{t+1})^2 \right\} \]

\[ = \frac{2}{M^2} \left\{ MX_t - X_t + \frac{X_t^2}{M} - X_t^2 \right\} \]

\[ = H_t \left( 1 - \frac{1}{M} \right) \]

By induction on \( t \) we get that:

\[ \mathbb{E}(H_t) = H_0 \left( 1 - \frac{1}{M} \right)^t \]

\[ \approx H_0 e^{-t/M} \]
Heterozygosity

• The decay of the **heterozygosity** illustrates how **genetic drift** tends to remove genetic variation from populations.

• Smaller populations loose variation faster than larger populations.

• The rate at which heterozygosity decays can be used to estimate the **effective population size**.
Once again, the data of Buri (1956) does not fit our expectation when $M = 32$ but behaves as if $M = 18$. 
Fixation

- $X_t = 0$ and $X_t = M$ are absorbing states of the biallelic WF process.
- Genetic drift leads to either $A$ or $a$ being lost from the population.
- When this happens, the surviving allele is said to be fixed in the population, and the lost allele is said to be extinct.
- What is the probability that $A$ will reach fixation rather than $a$ given its initial frequency?
Fixation

**Theorem** (Probability of fixation).

The probability that an allele will reach fixation given its initial frequency is equal to its initial frequency.
Fixation

Proof.

• The result is implied by the fact that $\mathbb{E}(X_t)$ remains constant and equal to $X_0$: If fixation is reached at time $t$, then:

$$\mathbb{E}(X_t) = \mathbb{P}(A \text{ fixed}) \times M + \mathbb{P}(a \text{ fixed}) \times 0$$

so that:

$$\mathbb{P}(A \text{ fixed}) = \mathbb{E}(X_t)/M = X_0/M$$

• Genealogical approach: eventually all genes in the population will be descended from one unique gene in generation 0, and this gene has probability $X_0/M$ to be of allele $A$.

• Markov Chain approach: let $q_i$ be the probability of fixation of $A$ given $X_t = i$, solve:

$$q_i = \sum_{j=0}^{M} q_j P_{i,j}$$
Examples

- In the Buri (1956) experiment, 58 of the 107 populations reached fixation: 28 for allele $bw^{75}$ and 30 for the other allele.

- The probability that a new allele appearing in a population through mutation will eventually become fixed is equal to $1/M$ provided no further mutation occurs.

- What is the expected time before fixation?
Time before fixation

**Theorem** (Time before fixation).

Let $\tau(p)$ be the expected time before fixation given that $X_0 = pM$. Then:

$$\tau(p) \approx -2M(p \log(p) + (1 - p)\log(1 - p))$$

with the approximation being valid for large populations.
**Time before fixation**

**Proof.** If \( p = 0 \) or \( p = 1 \), fixation is reached so that \( \tau(0) = 0 \) and \( \tau(1) = 0 \). Otherwise, \( \tau(p) \) is equal to one plus the fixation time in the next step. By summing over all possibilities for the next step, we get:

\[
\tau(p) = 1 + \sum_{j=0}^{M} P_{pM,j} \tau(j/M)
\]

This expresses \( \tau \) as the solution of a linear equation. Unfortunately, this equation becomes increasingly difficult to solve as \( M \) increases. We therefore use an approximation.
Time before fixation

Let $p_t = X_t/M$. Recall that the variance of $p_{t+1}$ about $p_t$ is of order $1/M$. Thus when $M$ is large, the terms in the sum for which $\text{abs}(pM - j)$ is “large” can be ignored. This suggests a continuous approximation. Let us assume that $p$ is a continuous function in $[0, 1]$.

Then we can rewrite as:

$$\tau(p) = 1 + \int\epsilon \mathbb{P}(p \rightarrow p + \epsilon) \tau(p + \epsilon) d\epsilon$$
Time before fixation

Since $\epsilon$ is small, we can expand $\tau(p + \epsilon)$ as a Taylor series:

$$
\tau(p) \approx 1 + \int_{\epsilon} \mathbb{P}(p \rightarrow p + \epsilon) \left( \tau(p) + \epsilon \tau'(p) + \epsilon^2 \tau''(p)/2 \right) d\epsilon
$$

$$
= 1 + \tau(p) + \tau'(p) \int_{\epsilon} \mathbb{P}(p \rightarrow p + \epsilon) \epsilon d\epsilon
$$

$$
+ (\tau''(p)/2) \int_{\epsilon} \mathbb{P}(p \rightarrow p + \epsilon) \epsilon^2 d\epsilon
$$

$$
= 1 + \tau(p) + \tau'(p) \mathbb{E}(\epsilon) + (\tau''(p)/2) \mathbb{E}(\epsilon^2)
$$
Time before fixation

Since $\mathbb{E}(\epsilon) = \mathbb{E}(p_{t+1} - p_t) = 0$ and
$\mathbb{E}(\epsilon^2) = \text{var}(\epsilon) = \text{var}(p_{t+1}) = p(1 - p)/M$, we have:

$$\tau(p) = 1 + \tau(p) + \tau''(p)p(1 - p)/(2M)$$

or

$$\tau''(p) = \frac{-2M}{p(1 - p)}$$

This can be solved with boundary conditions $\tau(0) = 0$ and $\tau(1) = 0$ to give
the required result. □.
Time before fixation

• Thus, for the Wright-Fisher model, the expected time to fixation is of order $O(M)$.

• This is the so-called diffusion approximation to the mean absorption time, although we have not used diffusion theory explicitly here.

• For example, in the case of a newly appeared mutation, we have $p = 1/M$ and

$$\tau(p) \approx 2 + 2\log(M)$$

• In the case where $p = 1/2$, we have

$$\tau(p) \approx 1.38M$$
Summary

• The pure Wright-Fisher model results in a **decay of genetic variation**.
• This is the effect of **genetic drift**, which is compensated by **mutation**.
• It is straightforward to **extend** the WF model to incorporate mutations.
• Exact calculations are impossible so that we need to use **diffusion approximations** as we did to find the time before fixation.
• This approach was championed in the 50s and 60s by **Kimura**.
• This is one of the most sophisticated branches of applied probability.
• We will avoid these complications!