Yesterday…

- The most difficult bits of MLE…
- Confidence interval calculation
- I hope you enjoyed the reading☺

Today

- Case study
- A real problem in population genetics
- Estimating effective population size from genetic data
- Help us understand Practical 5

Genetic drift

- One of the driving forces for the change in allele frequency (what are the other forces?)
- due to random shuffling of alleles
- Infinite gamete pool from parents
- Random pairing of gametes
- Wright and Fisher both studied this process, independently
- The Wright-Fisher model for genetic drift is a *Markov Chain*

Markov chain

- A random process $X(t), X(t+1)$, … with several "states" (possible values)
- Transits from one state to another (by chance) over time
- Memoryless: The transition probability depends only on the current state (anything happened before does not matter)

- A light bulb has two states: on and off
- The transition probability can be represented in a matrix form called Markov matrix

- For diploids, if the effective population size is N, then the possible number of alleles are $\{0, 1, 2, \ldots, 2N\}$.

- Assume there are two alleles: A and B

- If the allele frequency of allele A is $k/2N$ now, then the number of the alleles in the next generation follows $binomial(2N, k/2N)$

- The Markov matrix will have the dimension $(2N + 1) * (2N + 1)$

- For instance, for $N = 2$, there are five states: $\{0, 1, 2, 3, 4\}$ representing the number of a particular allele.
- The $\{i, j\}^{th}$ element of the transition matrix is the probability from state i to state j.
- Row sums to one.

Jump to state j

from state i

```
> WF(2)
allele 0 1.00000000 0.000000 0.0000000 0.000000 0.00000000
       1 0.31640625 0.421875 0.2109375 0.046875 0.00390625
       2 0.06250000 0.250000 0.3750000 0.250000 0.06250000
       3 0.00390625 0.046875 0.2109375 0.421875 0.31640625
       4 0.00000000 0.000000 0.0000000 0.000000 1.00000000
         allele  0        1        2        3        4
```

Example

- Given the Wright-Fisher transition matrix of $N = 2$. Let $X(t)$ be the number of allele A at time t.

```
> WF(2)
allele 0 1.00000000 0.000000 0.0000000 0.000000 0.00000000
       1 0.31640625 0.421875 0.2109375 0.046875 0.00390625
       2 0.06250000 0.250000 0.3750000 0.250000 0.06250000
       3 0.00390625 0.046875 0.2109375 0.421875 0.31640625
       4 0.00000000 0.000000 0.0000000 0.000000 1.00000000
   allele  0        1        2         3        4
```

What is $\Pr(X(t+1) = 3 | X(t) = 2)$?

What is $\Pr(X(t+1) = 3 | X(t) = 0)$?

What is $\Pr(X(t+2) = 3 | X(t) = 2)$?

Some properties of Markov matrix

- Non-negative (elements are probability, of course)
- Row sum to one
- In fact we can calculate the transition probability for T steps ahead by multiplying the matrix itself T times
- There are some states which you cannot leave once you entered. They are called the absorbing state. For example, the first row and the last row of the Wright-Fisher model (Why?)

Some R code

```
WF<-function(N)
{
result<-matrix(nc=2*N+1, nr=2*N+1)
for (i in 1:nrow(result))
      {result[i,]<-dbinom(0:(2*N),
      size=2*N, prob=(i-1)/(2*N))}
return(result)
}

WF(N=2)
WF(N=8)
dim(WF(8))
```

```
# WE CAN CALCULATE THE T-STEP TRANSITION PROBABILITY
BY USING MATRIX MULTIPLICATION
# %*% IS THE COMMAND FOR MATRIX MULTIPLICATION
M<-WF(2)
M%*%M

M%*%M%*%M%*%M%*%M

# YOU MAY ALSO TRY…
M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M%*%M

# WHAT DID YOU OBSERVE?
```

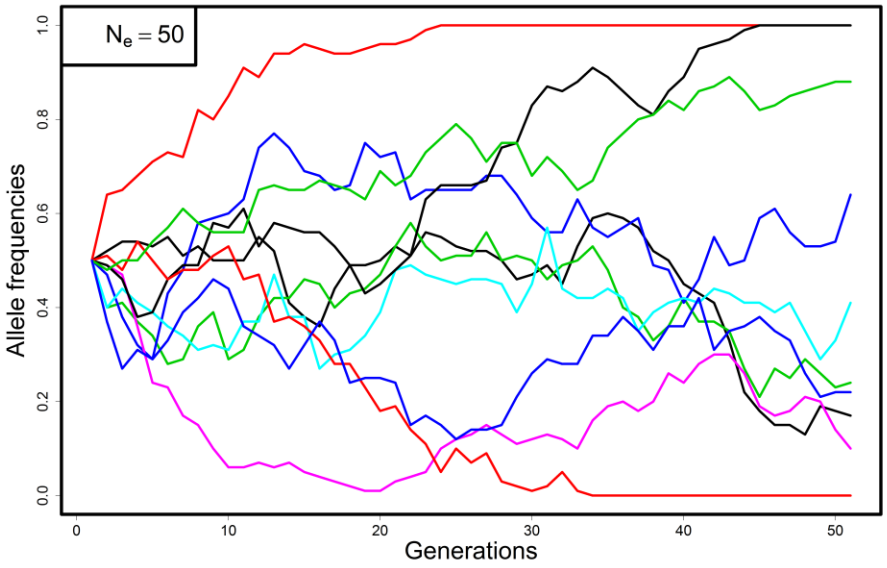
- M^{30} looks like this:

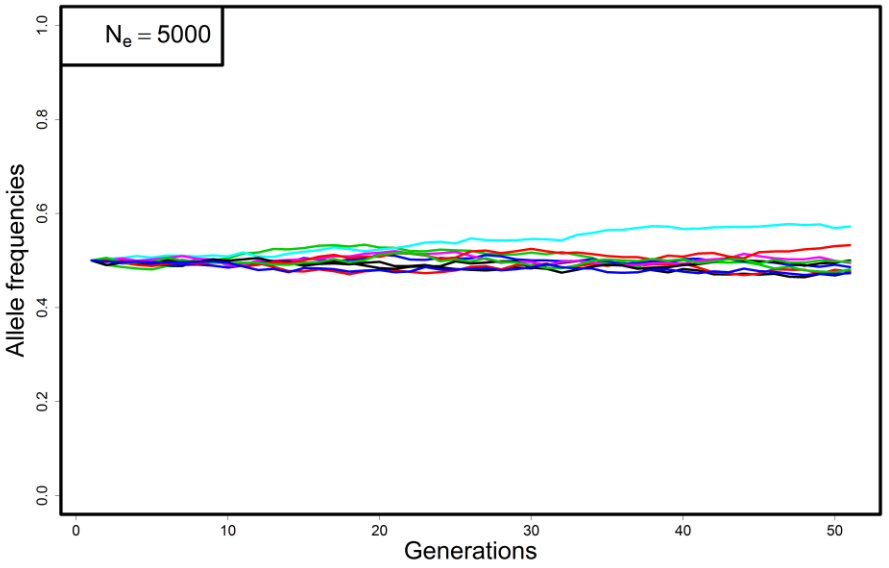
```
     [,1]         [,2]         [,3]         [,4]      [,5]
[1,] 1.0000000 0.000000e+00 0.000000e+00 0.000000e+00 0.0000000
[2,] 0.7499203 5.102345e-05 5.740139e-05 5.102345e-05 0.2499203
[3,] 0.4998937 6.803127e-05 7.653518e-05 6.803127e-05 0.4998937
[4,] 0.2499203 5.102345e-05 5.740139e-05 5.102345e-05 0.7499203
[5,] 0.0000000 0.000000e+00 0.000000e+00 0.000000e+00 1.0000000
```

- According to the WF model, all alleles go fixed/extinct in 30 generations

- Genetic drift reduces genetic variation!

Drift and population size





- extinction/fixation of alleles

- The mean allele frequency remains unchanged over time.

$$E[p_{t+1}|p_t] = p_t$$

- The variance of allele frequency increases

$$var(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N}$$

- A statistical geneticist will ask whether it is possible to infer the population size through studying the variance of allele frequency over time!

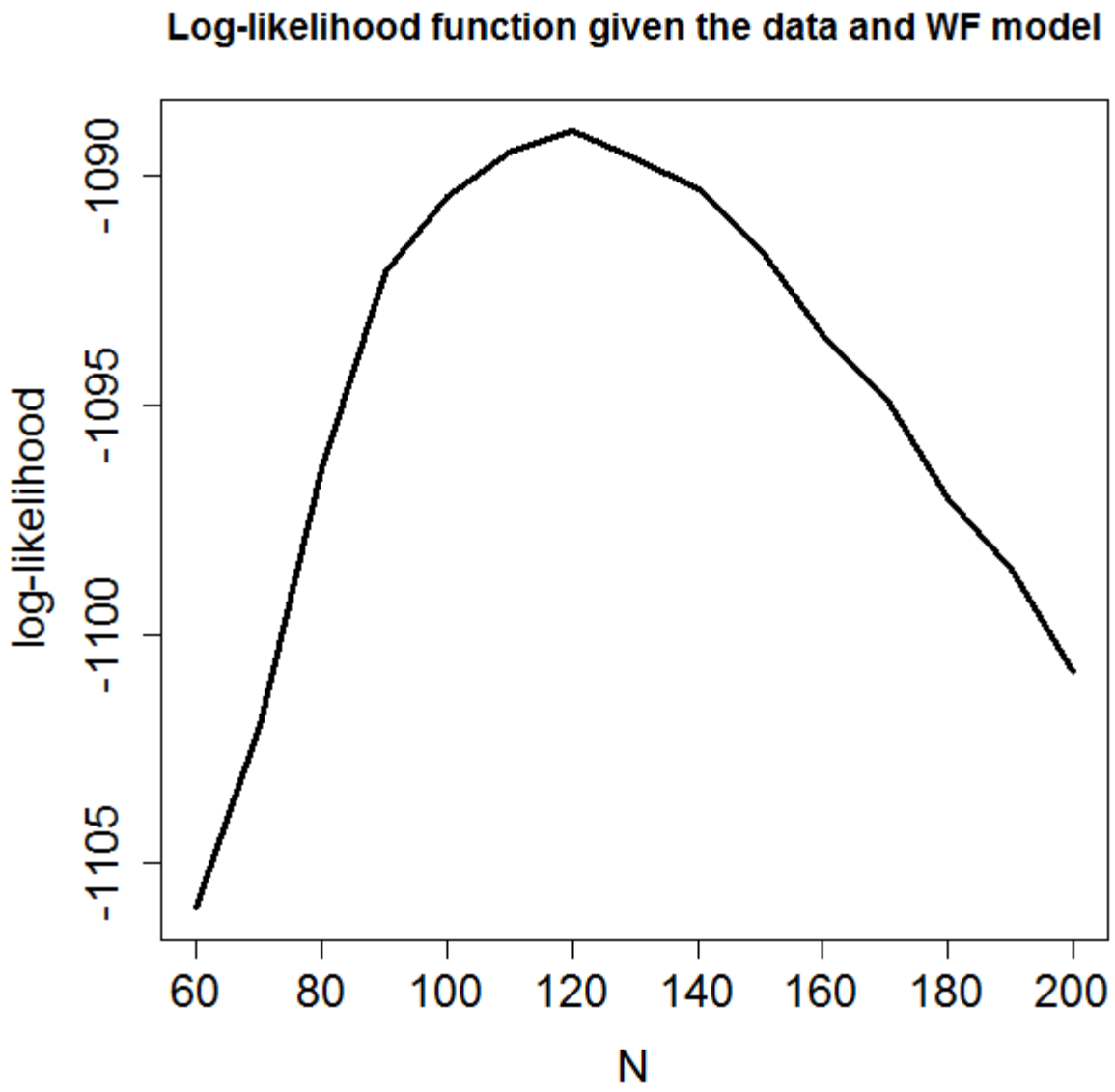
- Model: Wright-Fisher model
- Parameter of interest: N, the population size
- Data: I have plenty if you want

- So why not MLE???

- For every N, we can compute a corresponding WF matrix, and calculate the transition probability $p_t|p_0$ for each locus

- We sum all the log transition probabilities across all loci, and this is our log-likelihood value

- Plot the log-likelihood function against N


Log-likelihood function given the data and WF model
log-likelihood
-1090
-1095
-1100
-1105
60
80
100
120
140
160
180
200
N

The real scenario is more complex

- There are sampling error. The true allele frequency cannot be observed. (State-space model)

True allele freq $p_0 \rightarrow p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_{t-1} \rightarrow p_t$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

Observed $x_0 \quad x_1 \quad x_2 \quad \dots \quad x_{t-1} \quad x_t$

- The down arrows represent sampling error (noise!)
- Hidden Markov model. Exactly when ML suffers from computational issues

- The WF matrix only models the $\{p_t\}$ part. We need to take sampling error into account

- $L(N) = \sum_{all\ p_t} \sum_{all\ p_0} f(x_t|p_t) f(p_t|p_0, N_e) f(x_0|p_0) f(p_0)$

- "sum over all the possible values of the underlying true allele frequencies at two time points"

- Things can get very complicated when there are more than one populations (as in Practical 5)

- Williamson & Slatkin (1999) Using ML to estimate population size from temporal changes in allele frequencies.

- Wang & Whitlock (2003) Estimating effective population size and migration rates from genetic samples over space and time.