Day 5 selected topics

- Gamma MLE and moment estimator
- Memoryless r.v.
- Markov Chain
- Monte-Carlo integration
- Random-effect model
- Examples in population genetics

- Although MLE is “the best” sometimes it is hard to find
  - no closed-form solution (e.g. gamma MLE), and often can only be evaluated numerically
  - computational issue due to dimensionality (to many parameters)
  - the presence of nuisance variables (e.g. mixed effect model)

Memoryless r.v.?

- I arrive at a bus stop. Nobody is there. Let T be the waiting time before the arrival of the next bus. $T \sim Exponential(\lambda)$
- $f_T(t) = \lambda e^{-\lambda t}$ is the pdf
- $F_T(t) = \Pr(T \leq t) = 1 - e^{-\lambda t}$ is the cdf
- I will be late if a bus does not arrive in the next k minutes. The probability that I will be late for work is
- $\Pr(T > k) = 1 - \Pr(T \leq k) = 1 - F_{T(k)} = e^{-\lambda k}$

- Next day, I arrive at the same bus stop. “I have been waiting here for s minutes!” a person says to me.
- I am in the same situation, that if I do not get on a bus in the next k minutes then I will be in trouble.
- $\Pr(T > s + k | T > s)$ is the probability of being late
  - extra information about the r.v. T is given by the person
  - s minutes has passed, so it is known that $T > s$
- $\Pr(T > s + k | T > s) = \frac{\Pr(T > s+k \ \& \ T > s)}{\Pr(T > s)} = \frac{\Pr(T > s+k)}{\Pr(T > s)} = \frac{e^{-\lambda(s+k)}}{e^{-\lambda s}} = e^{-\lambda k}$
- WHAT???

- Memoryless: $\Pr(X > m + n | X > m) = \Pr(X > n)$
  - no extra information given
- Another example is the discrete geometric distribution
  - the number of Bernoulli trials required before getting the first success

Markov chain

- A random process, series of r.v., $X(t), X(t+1), X(t+2), \ldots$
- There are several "states" (possible outcomes) that each $X(j)$ can take on
  - states can be discrete or continuous
- Transits from one state to another by chance over time
- The transition probabilities depend only on the current state
- The transition probability can be represented in a matrix form called Markov matrix
- Time-homogeneous Markov chain: A special case of Markov chain whose transition probabilities remain the same over time

- The four states of PonPon the rabbit

Sleeping (S)

Playing (P)

Eating (E)

Grooming (G)

- The Markov matrix the four states is

	S	P	E	G
S	0.5	0.1	0.2	0.2
P	0	0.4	0.3	0.3
E	0.3	0.1	0.6	0
G	0.25	0.25	0.25	0.25

JC69 substitution model

- Nucleotide substitution
  - mutation
  - four states: A, C, T, G

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

- Other substitution models include
  - Kimura 1980, Felsenstein 1981 etc

Wright-Fisher model as a Markov chain

- For diploids, if the effective population size is N, then the possible number of alleles (states) are $\{0, 1, 2, \dots, 2N\}$.
- Assume there are two alleles: A and B
- The dimension of the Markov matrix is $(2N + 1) * (2N + 1)$
- Genetic drift changes the allele frequency over generations
- If the frequency of allele A is $k/2N$ now, then the number of allele A in the next generation follows $binomial(2N, k/2N)$

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

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 (the elements are probabilities, of course)
- Row sum to one
- If a Markov matrix $\boldsymbol{M}$ is time-homogeneous, then the transition probabilities for T steps ahead is $\boldsymbol{M^T}$
- We can analyse the long-run behaviour of $\boldsymbol{M}$. Some Markov chains have limiting distributions, $\lim_{T\to\infty} \boldsymbol{M}^T$ exists.
- Some even have stationary distributions π, where $\pi\boldsymbol{M} = \pi$
- There are some states which you cannot leave once you have entered. They are called the absorbing states. For example, the first row and the last row of the Wright-Fisher model (Why?)

	S	P	E	G
S	0.5	0.1	0.2	0.2
P	0	0.4	0.3	0.3
E	0.3	0.1	0.6	0
G	0.25	0.25	0.25	0.25

R Console

```
> m
     [,1] [,2] [,3] [,4]
[1,] 0.50 0.10 0.20 0.20
[2,] 0.00 0.40 0.30 0.30
[3,] 0.30 0.10 0.60 0.00
[4,] 0.25 0.25 0.25 0.25
> m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m%*%m
     [,1]  [,2]  [,3] [,4]
[1,]  0.3 0.175 0.375 0.15
[2,]  0.3 0.175 0.375 0.15
[3,]  0.3 0.175 0.375 0.15
[4,]  0.3 0.175 0.375 0.15
> 
```

Stationary distribution $\pi = (0.3, 0.175, 0.375, 0.15)$

- For WF model 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 if $N = 2$

- Genetic drift reduces genetic variation!

Other Markov processes

- Moran model
  - alternative to Wright-Fisher model
  - allows overlapping generations

- Birth and Death process
  - continuous-time Markov process
  - birth: state+1; death: state-1
  - infinitely many states

Monte Carlo integration

- To evaluate the following integral: $I = \int_0^1 \sqrt{1 - x^2} dx$
- Calculate by hand… ☹
- Numerical methods such as quadrature rule
  - counting areas of rectangles / trapeziums
  - `integrate()` in R

```
integrate(function(x){sqrt(1-x^2)}, lower=0, upper=1)
```

Monte Carlo integration

- $I = \int_0^1 \sqrt{1 - x^2} dx$

- Sample $\{x_1, x_2, \ldots, x_n\}$ from $uniform(0, 1)$ distribution

- Compute $I_n = \frac{1}{n} \sum_{i=1}^{n} \sqrt{1 - x_i^2}$

- I_n is an approximation to the integral I, if n is reasonably large

Justification

- $I = \int_a^b g(x)dx = \int_a^b \frac{g(x)}{f(x)} f(x)\, dx = E_X[\frac{g(X)}{f(X)}]$
  - where X is a r.v. with pdf f and support (a, b). The integral becomes the expectation of the transformed r.v. $\frac{g(X)}{f(X)}$.
- Remember, expectation is the population mean, the average of infinitely many trials, which can be "replicated" by computer
- Draw $\{x_1, \ldots, x_n\}$ from f, $I_n = \frac{1}{n}\sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)}$ is a good approximation to I for sufficiently large n

- $\int_0^\infty xe^{-2x}dx$

- $\int_0^\infty xe^{-2x}dx$
- Let $X \sim Exponential(\lambda = 1)$, $f_X(x) = e^{-x}$

$$\int_0^\infty \frac{xe^{-2x}}{e^{-x}} e^{-x} dx$$

$$= \int_0^\infty xe^{-x}e^{-x}dx$$

$$= E_X[Xe^{-X}]$$

- Sample $\{x_1, x_2, \ldots, x_n\}$ from $Exponential(1)$ distribution for some large n

- $I_n = \frac{1}{n}\sum_{i=1}^{n} x_i e^{-x_i}$ is an approximation to the original integral

```
x<-rexp(1e6, 1)
mean(x*exp(-x))
```

- Stochastic simulation -> no deterministic answers (unlike `integrate()` in R)
  - "random" answers

- Able to work with $\pm\infty$ bounds
  - as long as the chosen r.v. is defined in those bounds (e.g. normal)

Exercise

- Evaluate $\int_{-\infty}^{+\infty} e^{-x^2} dx$ via MC integration
  - which r.v. should you use? Those with $\pm\infty$ bounds perhaps?
  - some choices of r.v. are better than the others

- The intrinsic variance of MC integration is unavoidable
  - slow convergence
  - variance $\propto \frac{1}{n}$ as each draw is independent
  - need large n
- Multivariate integrals -> Multivariate distributions
  - requires samples from a multivariate distribution
  - requires even more sampling points
- There are ways to reduce variance (beyond this course)
  - e.g. importance sampling
- It is also possible to use dependent samples
  - Markov Chain Monte Carlo (MCMC)

MCMC integration

- $\int_a^b g(x)dx \approx \frac{1}{n}\sum \frac{g(x_i)}{f(x_i)}$
- In pure MC integration we assume $\{x_1, \dots, x_n\}$ are independent
- In fact the above approximation still holds for correlated $\{x_1, \dots, x_n\}$
- Sometimes it is easier to generate a series of dependent $\{x_1, \dots, x_n\}$
- One can (smartly) construct a Markov chain, whose stationary distribution is f
  - Gibbs sampling
  - Metropolis-Hastings (MH) algorithm

Example 1.1: Estimating haplotype frequencies and LD

	B	**b**	
A	p_{AB}	p_{Ab}	
a	p_{aB}	p_{ab}	
			1

- Two-locus, two-allele setting
- Four haplotypes: AB, Ab, aB, ab
  - with true haplotype frequencies $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$
  - sum of four haplotype frequencies =1

- If haplotypic data is obtained then the MLE is
  - $\widehat{p_{AB}} = \frac{\#\ of\ AB\ haplotype\ observed}{total\ haplotype\ sample\ size}$
  - same for all four haplotype frequencies

- $\widehat{r^2} = \frac{(\widehat{p_{AB}}\widehat{p_{ab}} - \widehat{p_{Ab}}\widehat{p_{aB}})^2}{\widehat{p_A}.(1-\widehat{p_A}.)\widehat{p_B}.(1-\widehat{p_B}.)}$
  - according to the invariant principle, $\widehat{r^2}$ is also the MLE for r^2, the standardised LD coefficient
  - heavily biased for small sample size

Table 1 Expected genotypic frequencies under HWE

	BB	*Bb*	*bb*
AA	$f_1 = p_{AB}^2$	$f_2 = 2p_{AB}p_{Ab}$	$f_3 = p_{Ab}^2$
Aa	$f_4 = 2p_{AB}p_{aB}$	$f_5 = 2(p_{AB}p_{ab} + p_{Ab}p_{aB})$	$f_6 = 2p_{Ab}p_{ab}$
aa	$f_7 = p_{aB}^2$	$f_8 = 2p_{aB}p_{ab}$	$f_9 = p_{ab}^2$

The expected frequency of genotypes given the haplotype frequencies under HWE [2]. All the expected frequencies $f_1, f_2, \ldots, f_9$ add up to one

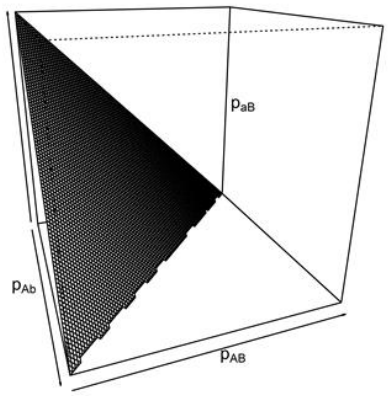
- Sometimes only genotypic information is obtained.

- Nine genotypes, each has an expected frequency under HWE.
- The genotype counts $\{n_1, n_2, \ldots, n_9\}$ are assumed to follow a multinomial distribution with size n and expected frequencies $\{f_1, f_2, \ldots, f_9\}$
- The log-likelihood function is

$$l(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) = constant + \sum_{i=1}^{9} n_i \log(f_i)$$

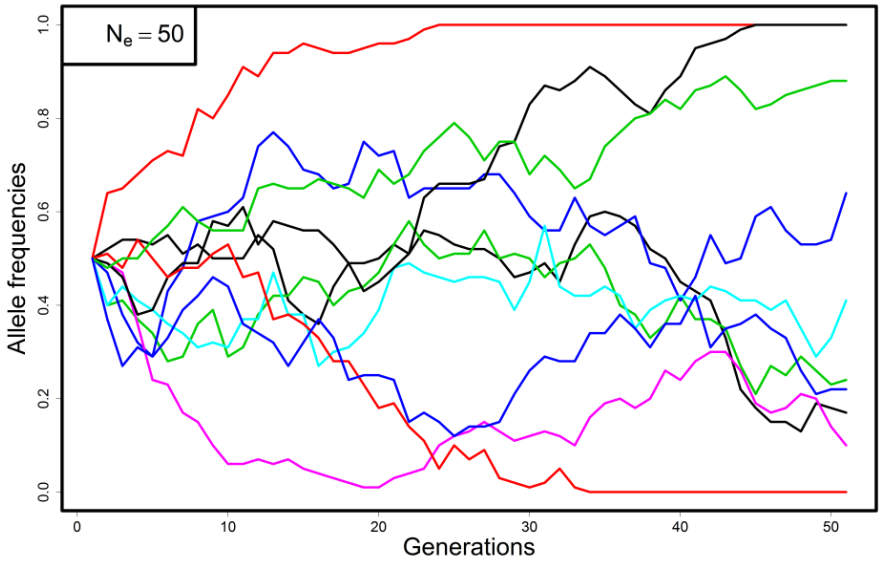
- In theory, we can maximise the above log-likelihood to find the MLE for the four haplotype frequencies, but…

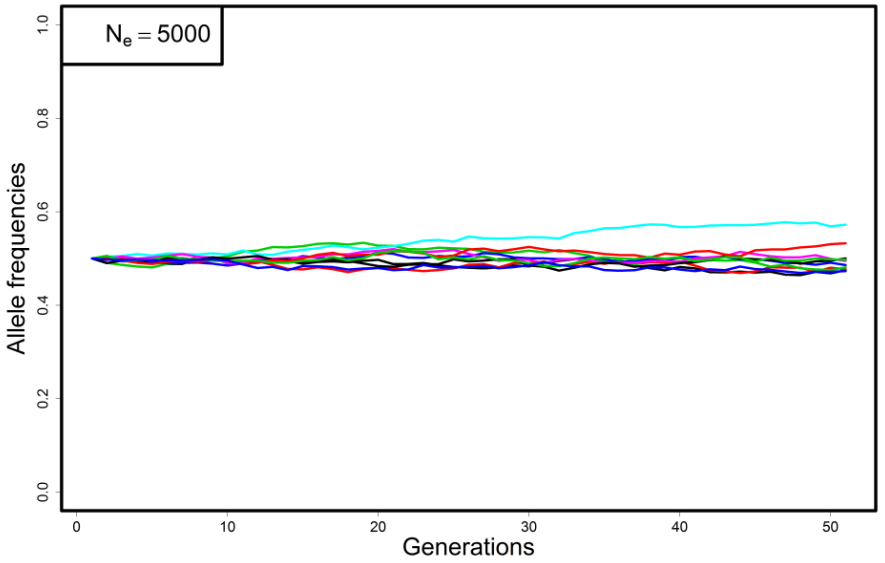
- There are challenges:
  - no closed-form solution
  - The four haplotype frequencies add up to 1; the fourth frequency is redundant. Hence the parameter space is a tetrahedron
  - sometimes there are multiple points with zero gradients (saddle / local points)?



- Possible solutions:
  - Expectation-Maximisation (EM) algorithm (Excoffier & Slatkin 1995)
  - some kind of transformation before maximising the log-likelihood? into a cube? (Hui and Burt, 2020)

Example 2: Drift and population size





- Model: Wright-Fisher model (the Markov matrix)
- Parameter of interest: N, the effective population size
- Data: The allele frequencies at two time points, across many unlinked loci.

- So why not MLE???

Transition prob, from WF matrix

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

Sampling at time t Sampling at time 0

- The two sampling $f(x_t|p_t)$ and $f(x_0|p_0)$ are modelled by binomial distributions, independently
- The transition probabilities $f(p_t|p_0, N)$ can be obtained from the WF matrix M^t

- "The likelihood of observing x_0 and x_t, given the true allele frequencies p_0 and p_t, and effective population size N"
  - then sum these likelihood values over all possible p_0 and p_t
  - state-space model
  - from 0 to $2N$

- Williamson & Slatkin (1999); Hui & Burt (2015)

Summary - MLE

- Day 1: Common r.v. and their pmf/pdf. Expectation. Moment generation functions.
- Day 2: Multivariate r.v., independence. Define likelihood functions. The triplet: model, data, parameters. Maximisation via differentiation and `optim()`.
- Day 3: Properties of MLE. Likelihood-ratio test. Logistic regression.
- Day 4: C.I. by log-likelihood. C.I. by approximate normality. Joint confidence region. Profile likelihood.
- Day 5: Examples and more examples

Beyond this course

- R functions and packages that help implement MLE
  - `mle(), confint()`
  - `{stats4}`

- Alternative optimisation routines
  - `nlm(), nlminb()`
  - `{optimx},{lbfgsb3},{BB},`… etc

- Require some ‘grammatical’ changes
- Multivariate testing

(Possible) solutions

- Approximation to the integrals (e.g. Laplace approximation), EM algorithm
- Statistical sampling (e.g. Monte Carlo, MCMC)
- Approximate Bayesian Computation (ABC)
- More Statistics!

MLE is...

- Not just a method, but THE method
- A collection of methods that share a common belief towards how “the best parameters” should be
- Many canned software and functions make use of the results from MLE (with or without acknowledging it)