Mathematical and statistical necessities for population genomics

Leo Speidel

Person 1	..AAGGTGCATTGCGTAGGCTTC..
	..AAGGTGCATTCCGTAGACTTC..
Person 2	..AAGGTGCATTGCGTAGGCTTC..
	..AAGGTGCATTGCGTAGGCTTC..
Person 3	..AAGGTGCATTCCGTAGACTTC..
	..AAGGTGCATTCCGTAGGCTTC..
Person 4	..AAGGTGCATTCCGTAGACTTC..
	..AAGGTGCATTCCGTAGACTTC..
:	

How do we go from observing genetic variation to inferences about evolution?

(genetic) data

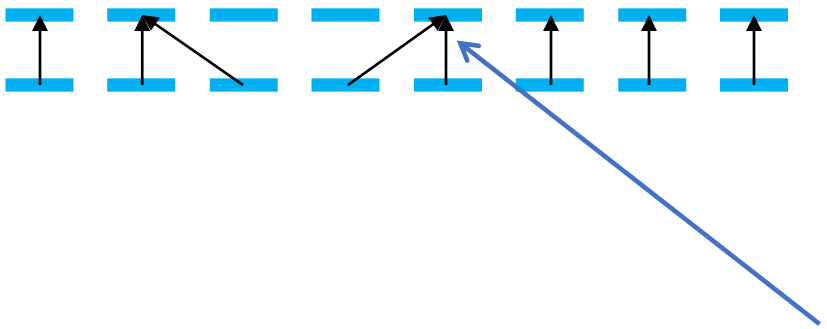
estimated parameter

The Wright-Fisher model

The **Wright-Fisher model** is able to approximate more realistic models of populations

Each member of the current generation randomly chooses one of *M* parents and inherits their DNA

Some population members have 0 children, others more than 1 child:



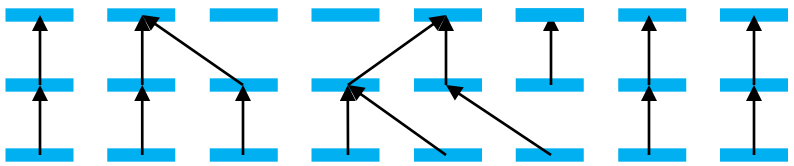
Each haplotype chooses parent in previous generation totally at random

If haplotypes **share** a parent back in time, this is called a **coalescence event**

The Wright-Fisher model

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time



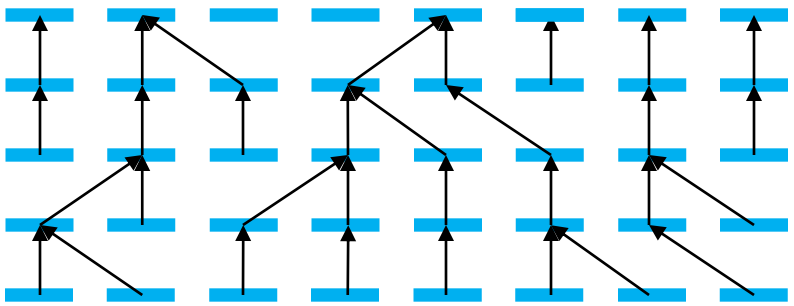
Each haplotype chooses parent in previous generation totally at random

If haplotypes **share** a parent back in time, this is called a **coalescence event**

The Wright-Fisher model

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time



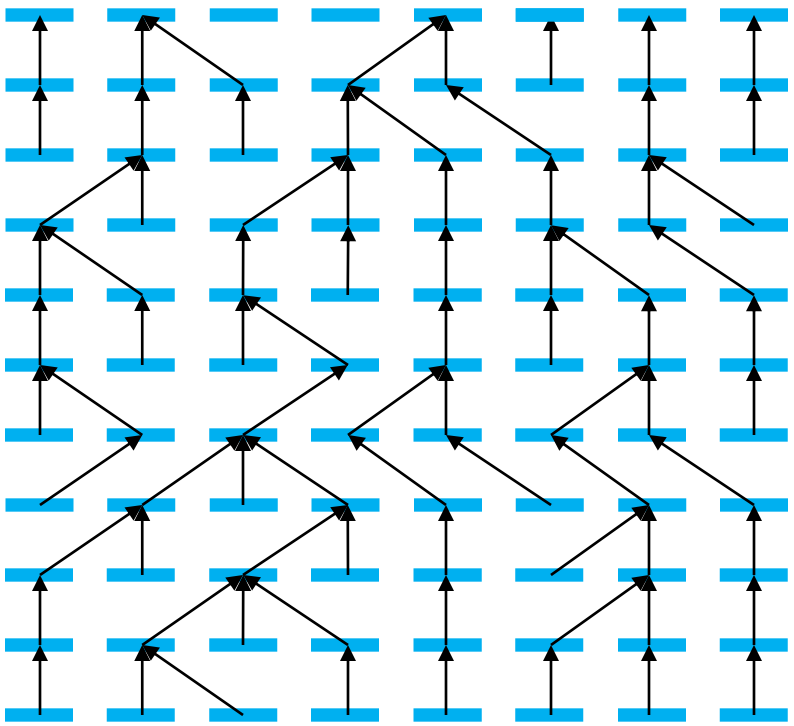
Each haplotype chooses parent in previous generation totally at random

If haplotypes **share** a parent back in time, this is called a **coalescence event**

The Wright-Fisher model

Over many generations, the population evolves

Our DNA comes from our ancestors so we look back in time



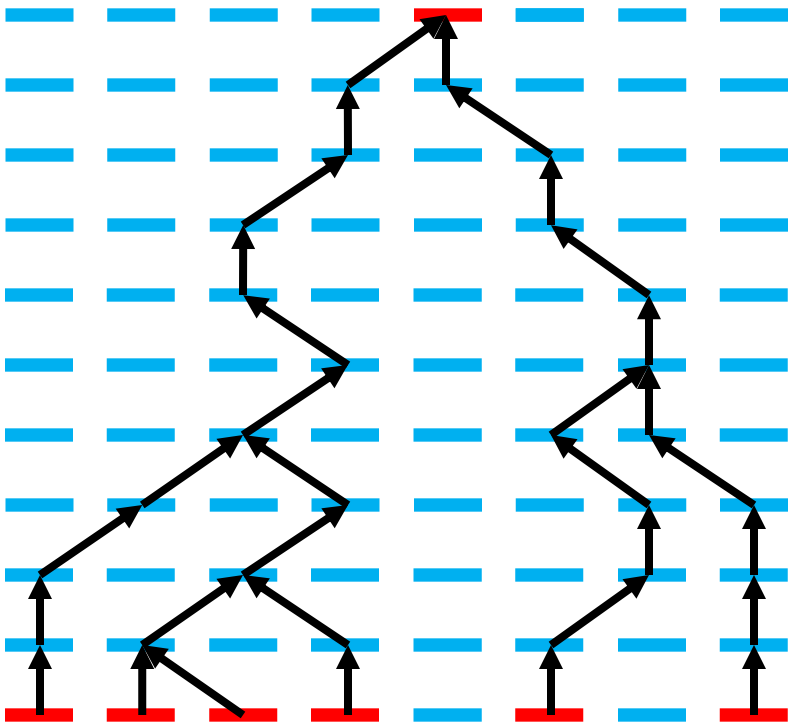
Each haplotype chooses parent in previous generation totally at random

If haplotypes **share** a parent back in time, this is called a **coalescence event**

The coalescent: history of a sample

If we take a sample from the population, we can trace their ancestry: a random tree

In this tree, the number of ancestors decreases back in time from *n* to 1



Sample of size *n*=6

Effective population size

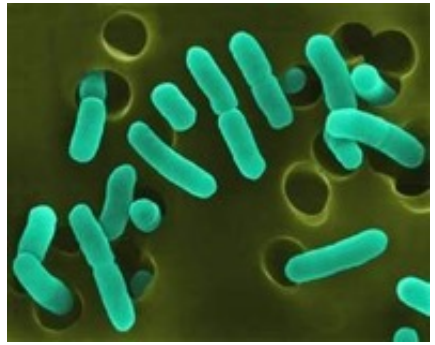
M ~10-50,000 for all human populations, highest in Africa

M varies dramatically across species
(Charlesworth, Nature Reviews Genetics 2009):

25,000,000 for *E.coli*

2,000,000 for fruit fly

D. Melanogaster



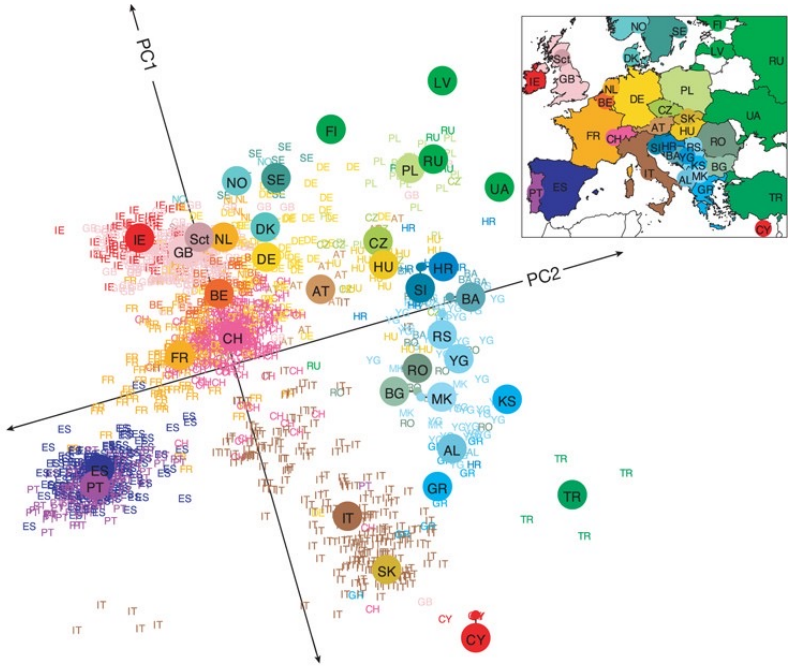
<100 for Salamanders
(Funk et al. 1999)

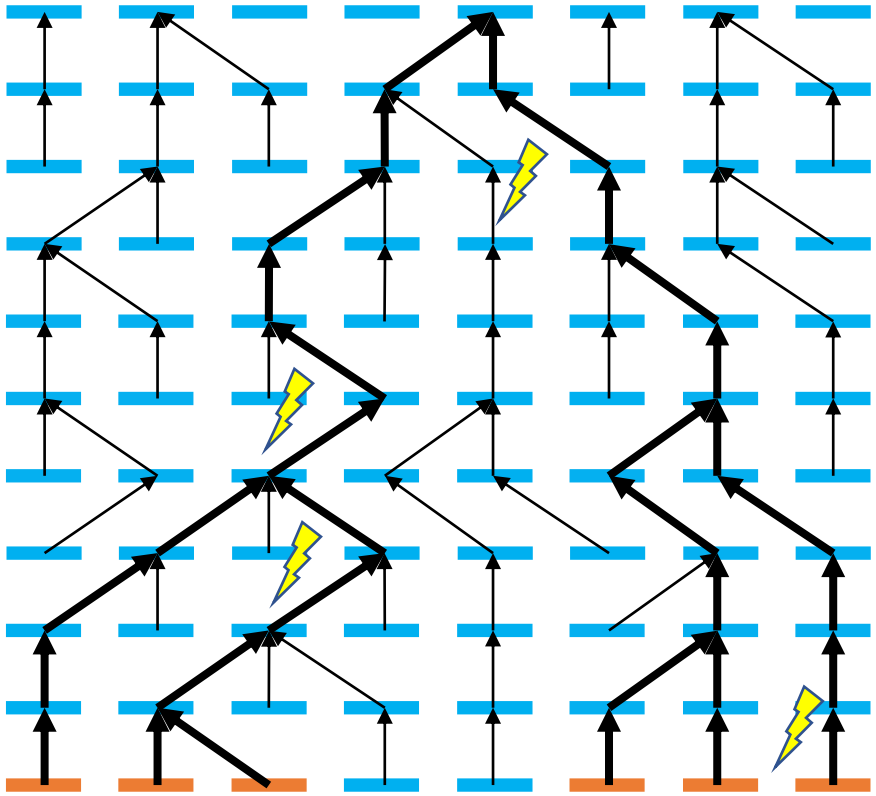
Randomness in population genetics

- Our sample is a random subset of the whole population

- Genetic inheritance is random
  - Mating choices
  - Mutation
  - Recombination

But there are patterns





Step 1: Model the data

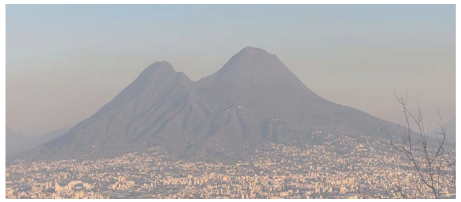
Person 1	..AAGGTGCATTGCGTAGGCTTC..
	..AAGGTGCATTCCGTAGACTTC..
Person 2	..AAGGTGCATTGCGTAGGCTTC..
	..AAGGTGCATTGCGTAGGCTTC..
Person 3	..AAGGTGCATTCCGTAGACTTC..
	..AAGGTGCATTCCGTAGGCTTC..
Person 4	..AAGGTGCATTCCGTAGACTTC..
	..AAGGTGCATTCCGTAGACTTC..
:	

(genetic) data ← Statistical model — Parameter(s)

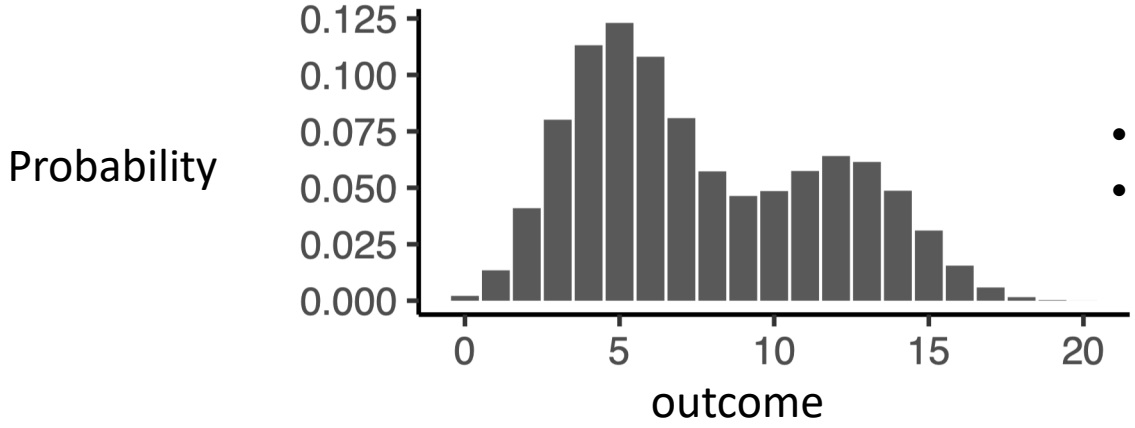
P(data | parameter)

The likelihood

What are probability distributions?



Mathematical function describing how likely each outcome is



- Mean (expected value)
- Standard deviation

Discrete probability distributions

- Outcome takes discrete values e.g., (0,1,2,3,…,20)
- Probability mass function $P(X = k)$

$$\sum_{\text{all outcomes } k} P(X = k) = 1$$

Continuous probability distributions

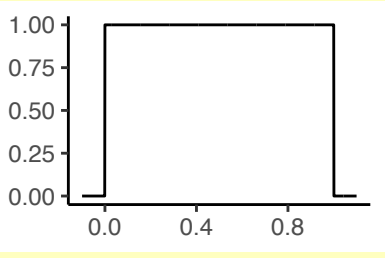
- Outcome takes continuous values e.g., any number between 0 and 20
- Probability density function f(X = t)

$$\int_{\text{all outcomes } t} f(X = t)\, dt = 1$$

Common probability distributions

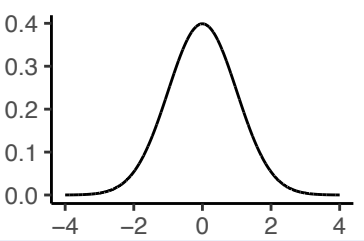
Uniform distribution

"no information / p-values under the null"



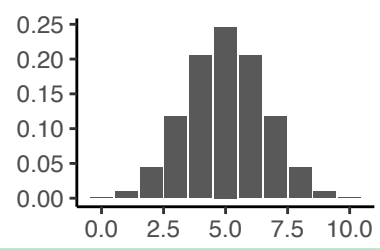
Normal distribution

"**everything** & if in doubt"



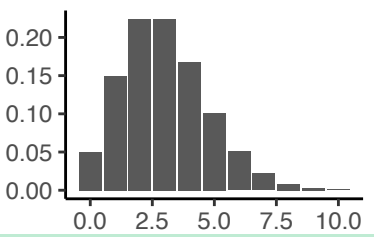
Binomial distribution

"**counting** the number of a type"



Poisson distribution

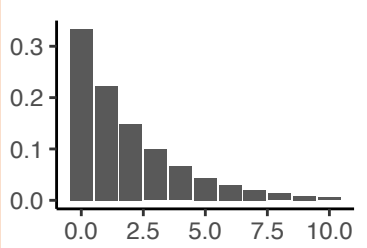
"**counting** the number of events"

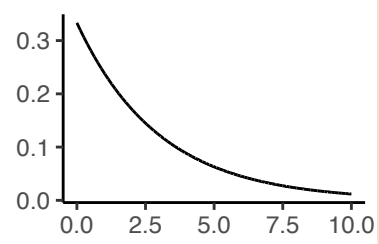


Geometric distribution

Exponential distribution

"**time** until first failure / next event"





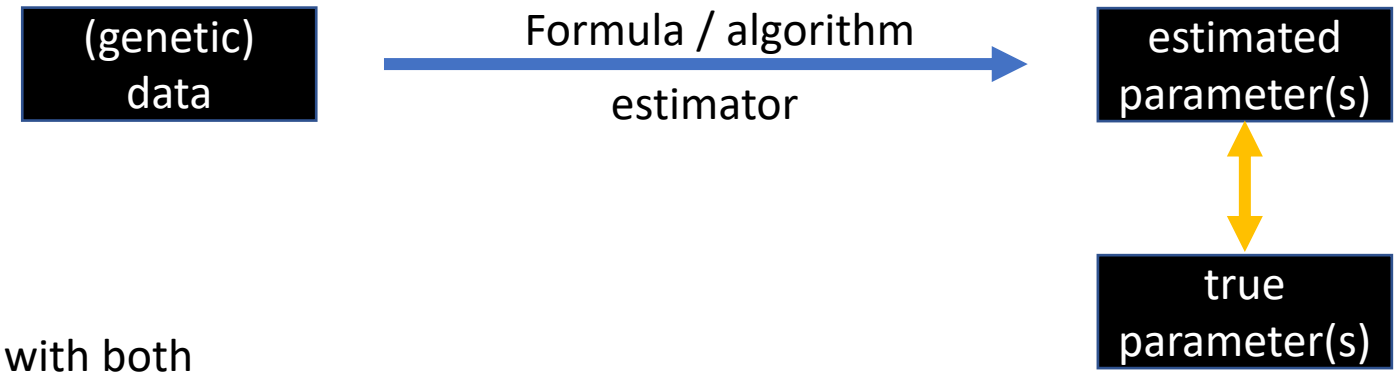
WIKIPEDIA
The Free Encyclopedia

In R:

```
runif()
rnorm()
rbinom()
rpois()
rgeom()
rexp()
```

Practical will explore these distributions in the context of population genetics!

Step 2: Fit the model (statistical inference)



Statistics provides us with both

An estimator/formula to get "the best possible guess" of the parameter of interest

some description of the relationship between estimated and true parameters

1. Link between data and parameter is **the likelihood**

   P(data | parameter)

   "What is the probability of the data given parameters?"
   "E.g. what is the probability of my genetic data assuming the mutation rate is x"

2. For any estimator, some desirable properties are:
   - **Unbiased:** On average, it should give the right answer
   - **Consistent:** The more data we get, the closer to the true parameter we get.
   - **Small variance:** How far off the true value do I expect my estimated parameter to be?

Likelihood: the link between the data and parameters

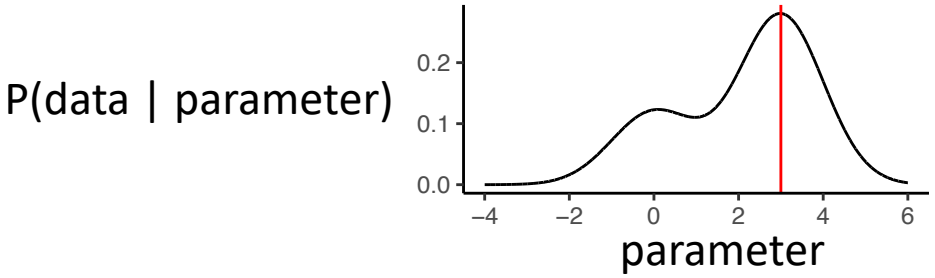
Key is to find some link between the data and our parameters of interest

Data → P(data | parameter) → estimated parameter

Two common strategies

Maximum likelihood:

Find the parameter value that maximises P(data | parameter)



Bayesian inference:

Define a prior $\mathrm{P(parameter)}$

$$\mathrm{P(parameter \mid data)} = \frac{\mathrm{P(data \mid parameter)\ P(parameter)}}{\mathrm{P(data)}}$$

Practical will explore both strategies!

How to find the “best” parameters

Data → P(data | parameter) → Best parameter

“lots of parameters”
“dependencies between data points”

e.g.,

- Individuals are genetically related through common ancestors
- Mutations nearby in the genome are linked (linkage disequilibrium)
- Two or more distinct processes generating similar patterns in the data:
  E.g., effective population size changes and natural selection

Some of the approaches we will encounter this week:

Markov-Chain Monte Carlo

Approximate Bayesian Computation

Hidden Markov models

Bootstrapping and jackknife

Machine learning

Overview of the practical

“Estimating the effective population size by hand”

“Maximum likelihood estimate of the time to the most-recent common ancestor”

- **Part 1** Coin toss, binomial distribution, geometric distribution
- **Part 2** Exponential distribution, Poisson distribution

(Solutions will be discussed ~15:00)

- **Part 3** Maximum likelihood estimation, Bayesian statistics

(Solutions will be discussed ~17:00)