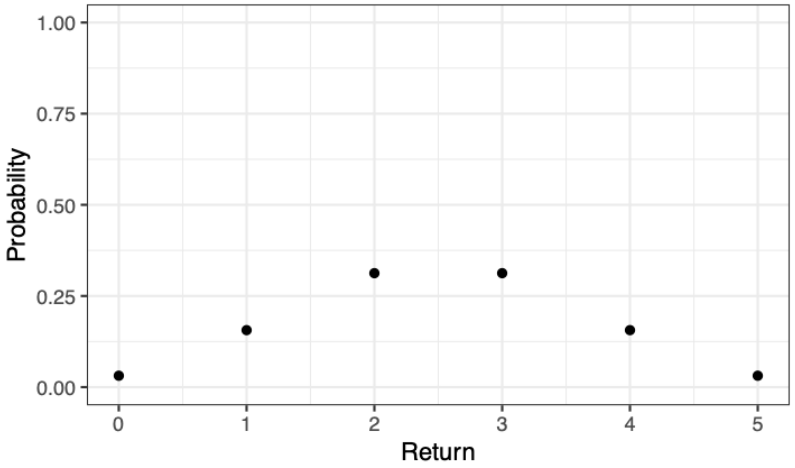
Maths & Stats primer solutions

Part 1 Coin toss, binomial distribution, geometric distribution

The coin toss is one of the most standard experiments in probability and on a fundamental level many evolutionary processes can be modelled as coin tosses.

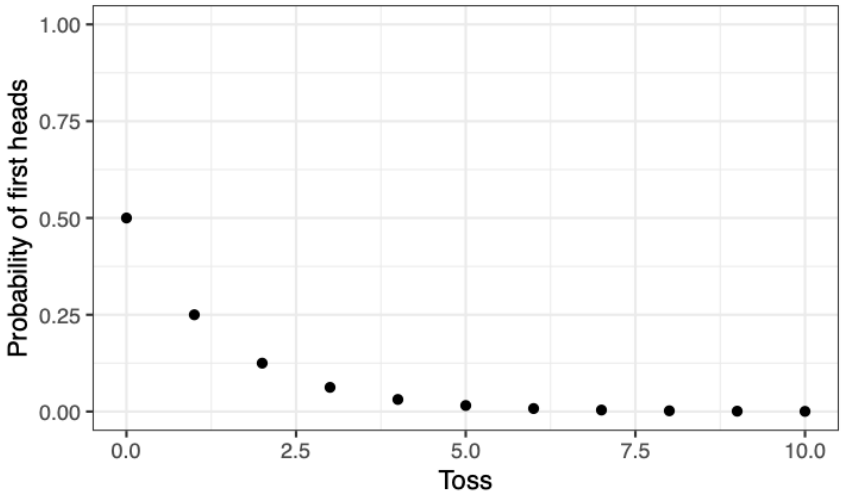
a) Assume you have a coin that shows heads with probability p and tails with probability 1-p. Heads wins you £1, tails wins you nothing. What is your expected return after 1 coin toss?
Expected return = £1 * p + £0 * (1-p) = £p

b) What is the expected return after N coin tosses? You may want to first write down an intuitive answer. For N = 5 and p = 0.5, draw a rough plot with possible returns on the x-axis and probability on the y-axis. Look up the **binomial distribution** and compare with your intuition.
Expected return after 1 coin toss is £p, therefore after N coin tosses it is £N*p.



c) Assume you have a constant number of N individuals in every generation, and offspring choose a random parent in each generation (Figure 1). This is known as the Wright-Fisher model. By thinking of the process of choosing a parent as a coin toss, calculate the expected number of offspring of a parent. What is the distribution of the number of offspring of a parent?
Assume I am a parent, then each offspring chooses "me" with probability 1/N, where N is the number of parents. This is like a coin toss with p = 1/N. There are N offspring, so N coin tosses, therefore the expected number of offspring is N*1/N = 1.
The distribution is Binom(N, 1/N).

d) Going back to the coin toss, if p = 0.5, can you make a guess on the expected number of tosses until you get the first heads? How would it change if p = 0.1? Draw a rough plot with number of tosses until the first heads on the x-axis and probability on the y-axis. Look up the **geometric distribution** and compare with your intuition.
Observation 1: it gets less and less likely that the first heads is after the second, third etc toss.
Observation 2: Suppose you make a large number of coin tosses. If p = 0.5, about half of them will show heads. If p = 0.1, about 1 in 10 of them will show heads. What is the average number of tosses between two heads? If they were equally spaced out it would be 2 and 10, and this hold on average with a large number of coin tosses.
In general, the expected number of tosses is 1/p.



e) For any two individuals, the probability to choose the same parent in the previous generation is $1/N$ (the first individual chooses a parent, the second individual chooses the same parent with probability $1/N$). How many generations will it take on average for two individuals to coalesce, i.e. for their ancestors to choose the same parent? In humans, $N \approx 20,000$. Assuming a generation time of 28 years, what is the expected number of years to the most recent common ancestor?

In each generation, we do a coin toss with p = 1/N, where heads corresponds to choosing the same parent. As we have seen in d), the expected number of tosses is 1/p = N.

With N = 20000, we therefore have 20000*28 years = 560,000 years!

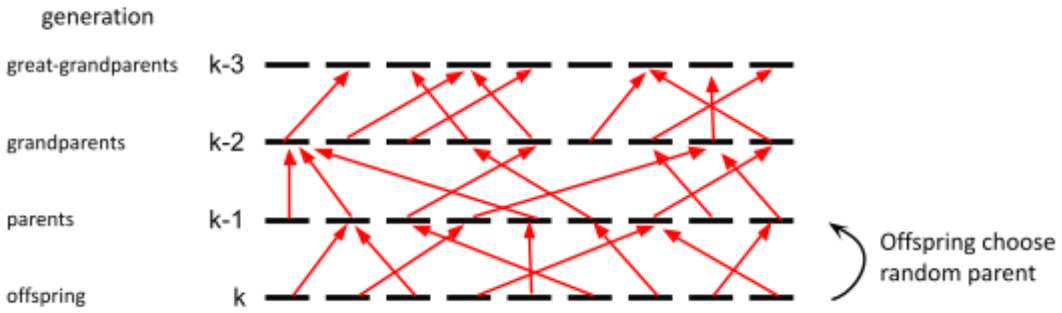


Figure 1. The Wright-Fisher model. Each generation has an equal number of individuals (N), and offspring choose a random parent in each generation.

Part 2 Exponential distribution, Poisson distribution

In part 1, we saw that the total number of heads after N coin tosses was binomially distributed and the time between two consecutive heads was geometrically distributed. Time here was modelled in discrete generations.

However, sometimes we prefer to model time as a continuous value. Events (corresponding to heads in the discrete case) are now allowed to happen at any moment with some chance, defined by a rate parameter. As before, the time between two consecutive events is random as is the total number of events up to some time (Figure 2).

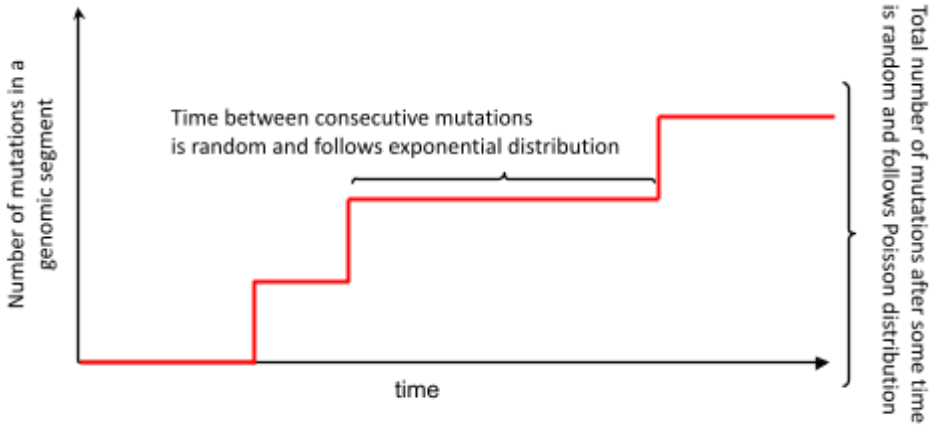
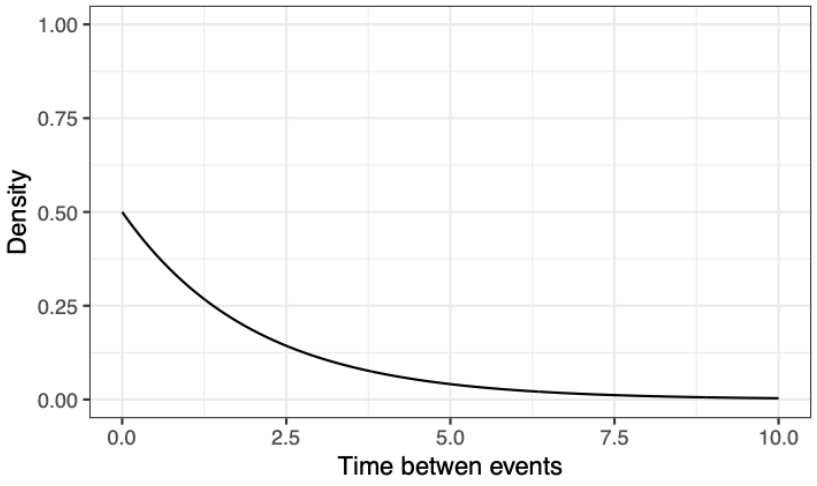


Figure 2. Illustration of how exponential distribution and Poisson distribution are related to each other. The time between consecutive events is ***exponentially distributed****, the total number of events up to some time is* ***Poisson distributed****. This process of accumulating events in this way is also called a* ***Poisson process****.*

a) The time between consecutive events is now **exponentially distributed** (Figure 2). Look up the relationship between the rate and expectation in an exponential distribution and compare it to the geometric distribution in Part 1d). Discuss the similarities and differences between a geometric and exponential distribution.
We can think of an exponential distribution as doing a coin toss at shorter and shorter time intervals. The expected value is 1/rate.



b) Assuming a mutation rate of 1.25×10^{-8} mutations per base per generation, calculate the expected time between two mutations occuring within a genomic region of 1Mb and confirm that this has the correct unit.

The rate parameter will have units **mutations/generation.**

1.25e-8 mutations/(base * generation) * 1e6 bases = 1.25e-2 mutations/generation.

Expected value is therefore 1/1.25e-2 generations/mutations = 80 generations/mutations.

c) Intuitively, using a) how many mutations would you expect between two individuals in the 1Mb region if the time to their most recent common ancestor (TMRCA) was 10,000 generations? (Hint: Note that mutations will happen in ancestors of both individuals!)

We have 80 generations/mutations.

In 10,000 generations, we therefore accumulate 10,000/80 mutations = 125 mutations.

Since mutations will happen in both individuals, we accumulate 250 mutations.

d) The random number of mutations between two individuals for which you calculated the expectation in c) has a **Poisson distribution** (Figure 2). Using c), write down a formula for how the rate parameter of the Poisson distribution is related to the mutation rate, TMRCA, and genome length (Hint: The rate parameter of a Poisson is its expected value). Using 1e) can you come up with a relationship between population size and the number of pairwise differences between individuals?

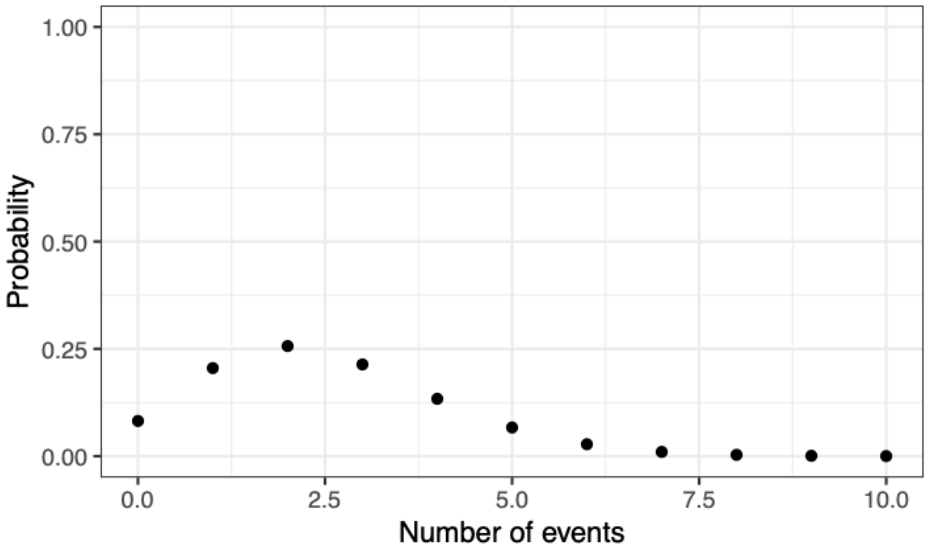
We want to generalise c). There we had

250 = 2*10,000/80 = 2*10,000/(1/1.25e-2) = 10,000 * 1.25e-8 * 1e6.

Generalising, this is **2 * TMRCA * mutation_rate * genome_length**.

From 1e), we know that E[TMRCA] = N, the population size. Therefore

N ≈ number of pairwise differences / (2*mutation_rate*genome_length).



Part 3 Maximum likelihood estimation, Bayesian statistics

To make statistical inferences about our evolutionary past given some data that we collected, one strategy is to come up with a probabilistic model that emulates evolution, which we use to calculate the probability of observing our data. By varying parameters in our model, we can ask what the **most likely** set of parameters are that would give rise to the observed data. We call this approach **maximum likelihood estimation**.

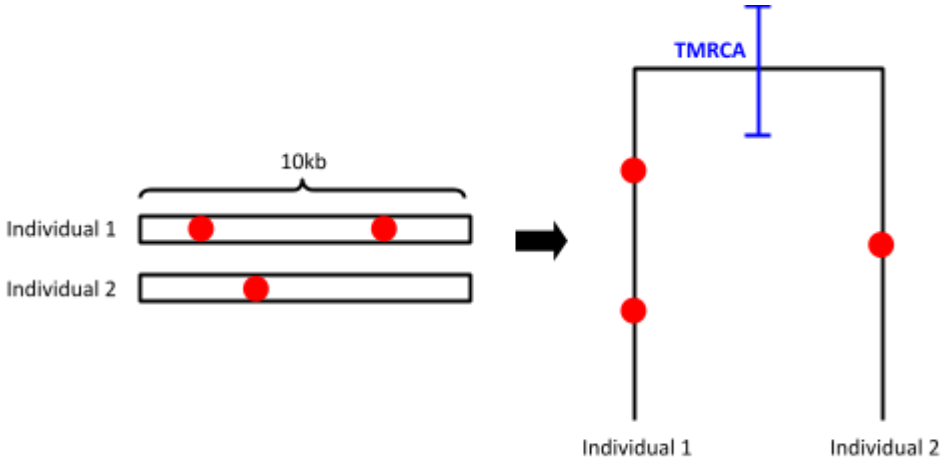


Figure 3. In this example, we want to estimate the TMRCA from the number of observed pairwise differences. Red circles indicate mutations.

Suppose we observe the number of pairwise differences between two individuals, denoted by x. Given this data, we aim to estimate their TMRCA. We use our probabilistic model from Part 2, which says that the number of pairwise differences that accumulate up to some time is Poisson distributed, where the rate parameter equals

$$2 \times mutation\,rate \times genome\,length \times TMRCA.$$

a) Assume a mutation rate of 1.25×10^{-8} and genome length of 10,000. Using the R function "dpois" or otherwise, calculate the probability of observing $x = 3$ differences, if the TMRCA equals 5,000, 10,000, or 20,000. By trying different values, can you find the **maximum-likelihood** TMRCA?

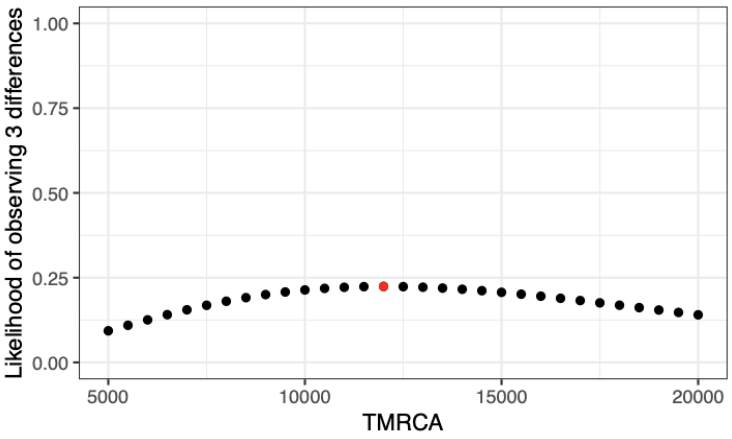
The TMRCA is the parameter which we want to infer.
Our data is that we observe 3 genetic differences (mutations).

The number of differences is Poisson distributed, with parameter
2 * mutation_rate * genome_length * TMRCA = 2 * 1.25e-8 * 10,000 * TMRCA.

So the probability of observing 3 differences when TMRCA = 5000 is given by
dpois(x = 3, lambda = 2*1.25e-8*10000*5000) = 0.09326328

Similarly, dpois(x = 3, lambda = 2*1.25e-8*10000*10000) = 0.213763
dpois(x = 3, lambda = 2*1.25e-8*10000*20000) = 0.1403739
So this suggests that the maximum-likelihood TMRCA is somewhere between 5000 and 20000.

Trying out different values gives



b) Use the fact that the rate parameter of a Poisson is its expected value to come up with a formula for the TMRCA. Can you intuitively describe this formula? Compare your answer to a).
2*mutation_rate*genome_length*TMRCA = "expected number of differences"

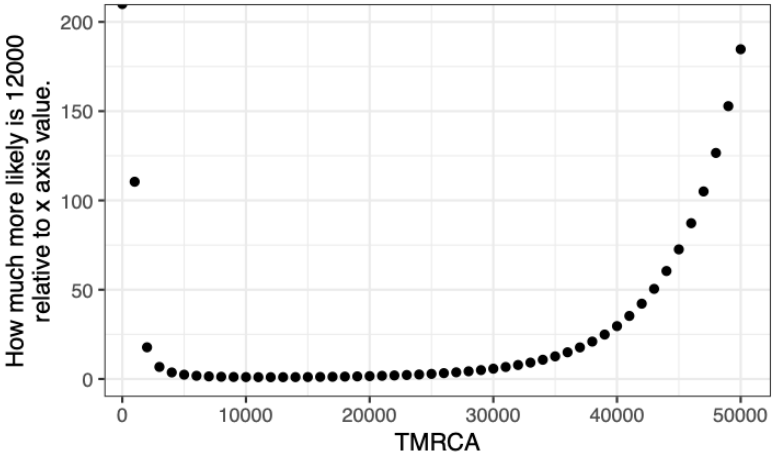
Solving for TMRCA, we get
TMRCA = "expected number of differences"/(2*mutation_rate*genome_length)

We know all variables on the right hand side, plugging these in gives
TMRCA = 3/(2*1.25e-8*10000) = 12,000.

Note that there is no guarantee that a) and b) give the same answer, but often it turns out they do.

c) Often, we like to have an idea of the **uncertainty** in our estimate. There are generally three options:
First, we can compare likelihoods for different TMRCA values. Using the R function "dpois" as in a), can you calculate how many times more likely a TMRCA of 12,000 is compared to a TMRCA of 30,000 or 50,000?
dpois(x = 3, lambda = 0.00025*12000)/dpois(x = 3, lambda = 0.00025*30000)=5.761096
dpois(x = 3, lambda = 0.00025*12000)/dpois(x = 3, lambda = 0.00025*50000)=184.6849

In other words, 12000 is 6 times more likely than 30000 and 185 times more likely than 50000. So perhaps we would think that 30000 is plausible, but 50000 seems quite a lot less likely than 12,000.



Second, we can calculate a "standard error" of our estimate. However, this is impossible in our example above, because we only have one data point x.

Finally, if we have a prior guess (e.g. from previous experiments or intuition), we can use a **Bayesian approach**. This can work particularly well if we only have a small amount of data.

Central to the Bayesian approach is **Bayes' rule**:

$$P(TMRCA \mid x) = \frac{P(x \mid TMRCA)\, P(TMRCA)}{P(x)}.$$

- The left hand side is the **posterior probability** of a TMRCA value, given the data x (to be precise, it is the "density" of the TMRCA, as this is a continuous number, but this is not important for our purposes).

- The right hand side contains a term familiar to us. We call $P(x \mid TMRCA)$ **the likelihood**, which we know has a Poisson distribution from a).

- We call $P(TMRCA)$ **the prior probability** of the TMRCA. This is where we can use our prior knowledge. In Part 1e), we saw that the expected TMRCA equals the population size. Therefore, if we had an idea of the population size, we could use this as prior knowledge here. Let's assume that the population size is 20,000, in other words, from Part 1e) we expect our TMRCA to be 20,000. We can model this as an exponential distribution with rate parameter 1/20,000 = 0.00005 (discuss why by drawing parallels to Part 2!).

- Finally, the denominator P(x) is the probability of the data; for our purposes this is a constant that we don't need to calculate.

Doing some maths we will find that the **posterior** $P(TMRCA \mid x)$ has a **gamma distribution** with parameters $\alpha = 1 + x$, $\beta = 0.00025 + 0.00005$. Note that the term 0.00025 equals $2 \times mutation\ rate \times genome\ length$ and comes from the likelihood, and 0.00005 comes from the prior above. For how these parameters determine the gamma distribution, you can look up the box on the right of https://en.wikipedia.org/wiki/Gamma_distribution). The expected value of a gamma distribution is given by α/β.

d) What is the posterior mean of the TMRCA? How does it compare to the maximum likelihood estimate from a)? How does it compare to our prior estimate of 20,000?
Posterior mean is alpha/beta = (1+3)/(0.00025+0.00005) = 13333.33.
Prior estimate was 20,000 and our maximum-likelihood estimate was 3/0.00025 = 12,000, so the posterior mean falls between the two. Intuitively, our data alone tells us 12,000, but our prior tells us 20,000 so Bayes' rule will give us a value in between.

e) Using the R function qgamma or otherwise, can you find the 2.5% and 97.5% percentile of this distribution and hence a 95% credible interval (Bayesian equivalent of a confidence interval) for the TMRCA?
qgamma(p = 0.025, shape = 4, rate = 0.00025+0.00005) = 3632.885
qgamma(p = 0.975, shape = 4, rate = 0.00025+0.00005) = 29224.24
So [3632.885, 29224.24] is our equal-tailed 95% credible interval for the TMRCA.