

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.

- Let's assume we have a coin that shows heads with probability 0.6 and tails with probability 0.4. Heads wins you €1, tails wins you nothing. What is your expected return after 1 coin toss?
- Now assume you make 5 coin tosses. What is the probability of winning €0 and €5, respectively? What is the expected return? Draw a rough plot with possible returns on the x-axis and probability on the y-axis. In R, calculate the probabilities using the function **`dbinom(x, size = 5, p = 0.6)`** and compare with your intuition.
- 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 offspring 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?
- 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 $p = 0.1$? In R, use the function **`rgeom(n=100, prob = 0.5)`** to sample 100 times from the geometric distribution and compute the arithmetic mean. Compare with your intuition.
- 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?

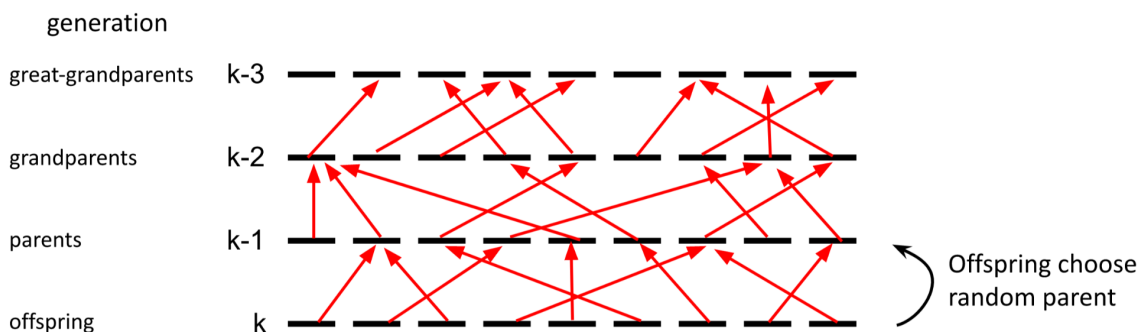


Figure 1. The Wright-Fisher model. Each generation has an equal number of individuals (N), and offspring choose a random parent in each generation.

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).

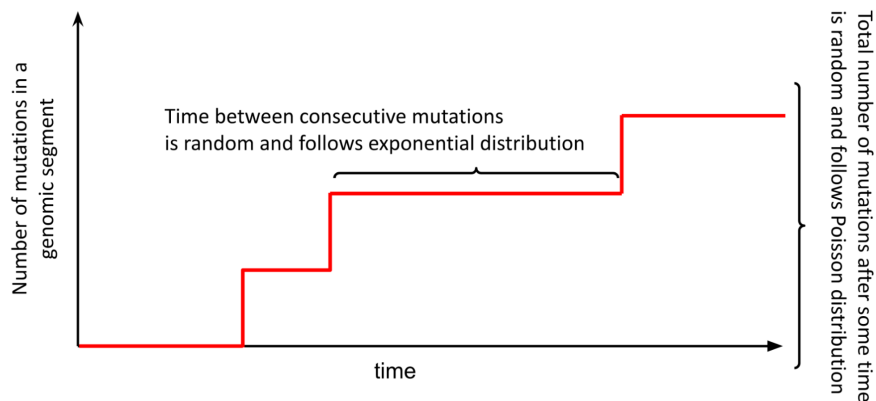


Figure 2. Illustration of how exponential distribution and Poisson distribution are related to each other. The time between consecutive events is **exponentially distributed**, the total number of events up to some time is **Poisson distributed**. This process of accumulating events in this way is also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

- 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.
- Assuming mutation occur at a rate of 1.25×10^{-8} mutations per base per generation. Can you calculate the expected time between two mutations occurring at a single base-pair.
- Let's assume that the divergence time between humans and chimps is 300,000 generations (~8M years). How many mutations do you expect in this time at a single base-pair position? (Note that mutations will accumulate in ancestors of both humans and chimps).
- 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).
- In a randomly mating population, we saw in 1e) that the expected time to the most recent common ancestor is given by the effective population size. Can you come up with a way to estimate the effective population size from the number of pairwise differences between individuals?

Part 3 Maximum likelihood estimation, Bayesian statistics

To make statistical inferences about our evolutionary past, one strategy is to come up with a probabilistic model that emulates evolution. We can then calculate the probability of observing our data given parameters. By varying parameters in our model, we can ask what the parameter that maximises the probability of observing our data. We call this approach **maximum likelihood estimation**.

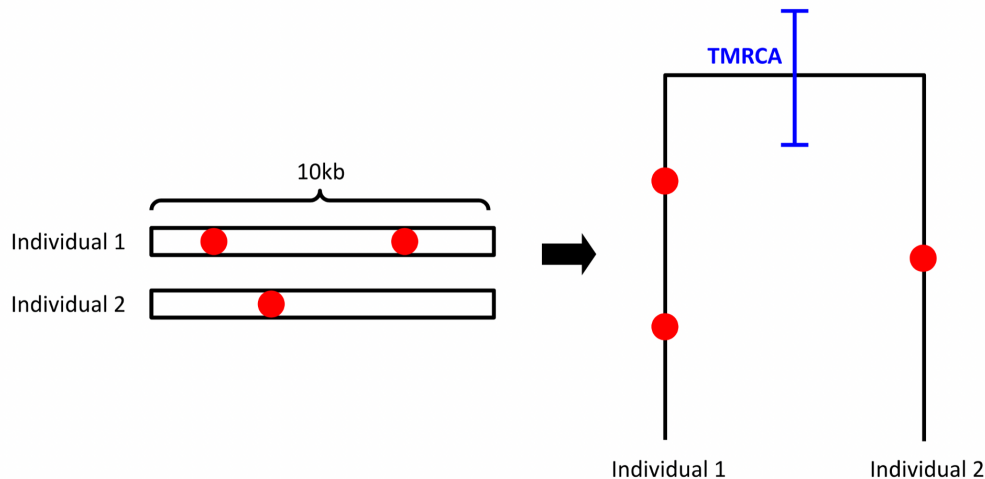


Figure 3. In this example, we want to estimate the TMRCA from the number of observed pairwise differences. Red circles indicate mutations.

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

$$2 \times \text{mutation rate} \times \text{genome length} \times \text{TMRCA}.$$

- 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 $P(x | \text{TMRCA})$ 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?
- Use the fact that the rate parameter of a Poisson is its expected value to come up with a formula for the TMRCA. Compare your answer to a).
- Often, we like to have an idea of the **uncertainty** in our estimate. We can compare the 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?
- 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**:

$$P(\text{TMRCA} | x) = \frac{P(x | \text{TMRCA}) P(\text{TMRCA})}{P(x)}.$$

- 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 $P(x | TMRCA)$ **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 assume that $P(TMRCA)$ has an exponential distribution with rate parameter $1/20,000 = 0.00005$.
- 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** $P(TMRCA | x)$ has a **gamma distribution** with parameters $\alpha = 1 + x$, $\beta = 0.00025 + 0.00005$. Note that the term 0.00025 equals $2 \times \text{mutation rate} \times \text{genome length}$ 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 .

- e) 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?
- f) 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?

Useful functions in R

| | | |
|-----------------------------|--|--|
| d*: density of distribution | Probability/density at some value x | <code>dbinom(x = 2, size = 4, prob = 0.5)</code> |
| p*: distribution function | Probability of drawing a value less or equal to value q | <code>pbinom(q = 2, size = 4, prob = 0.5)</code> |
| q*: quantile function | Value for which the probability of drawing less or equal to this value is p. | <code>qbinom(p = 0.5, size = 4, prob = 0.5)</code> |