Statistical population genetics

Lecture 5: Properties of the coalescent

Xavier Didelot
Dept of Statistics, Univ of Oxford
didelot@stats.ox.ac.uk

Coalescent simulation

A coalescent tree for n genes can be **simulated** using the following algorithm:

Algorithm (Coalescent simulation)**.**

1. *Start with $k = n$ lines;*
2. *Simulate the waiting time for the next coalescence event from* $\mathrm{Exp}(k(k-1)/2)$*;*
3. *Choose without replacement a random pair of lines (i, j) amongst the* $k(k-1)/2$ *possible pairs;*
4. *Join i and j into a single line so that the number of lines k is decreased by one;*
5. *If $k > 1$, go back to Step 2, otherwise stop.*

Convolution of exponentials

Theorem (Convolution of exponentials)**.**

If $R_1, ..., R_n$ *are independent exponential distributed with parameters* $\lambda_1, ..., \lambda_n$, *then their sum is distributed as:*

$$f_{\sum_{i=1}^{n} R_i}(x) = \sum_{i=1}^{n} \lambda_i e^{-\lambda_i x} \prod_{j=1, j \neq i}^{n} \frac{\lambda_j}{\lambda_j - \lambda_i}$$

Convolution of exponentials

Proof (in the case $n = 2$)

X_1 and X_2 are exponentially distributed with parameter λ_1 and λ_2. The probability density function of $X_1 + X_2$ is therefore:

$$\begin{aligned} f_{X_1+X_2}(t) &= \int_0^t f_{X_1}(x) f_{X_2}(t-x)\mathrm{d}x \\ &= \int_0^t \lambda_1 \mathrm{exp}(-\lambda_1 x)\lambda_2 \mathrm{exp}(-\lambda_2(t-x))\mathrm{d}x \\ &= \lambda_1\lambda_2 \mathrm{exp}(-\lambda_2 t)\int_0^t \mathrm{exp}((\lambda_2-\lambda_1)x)\mathrm{d}x \\ &= \frac{\lambda_1\lambda_2}{\lambda_2-\lambda_1}(\mathrm{exp}(-\lambda_1 t) - \mathrm{exp}(-\lambda_2 t)) \end{aligned}$$

□

Height of a coalescent tree

The following theorem is due to Tavaré (1984).

Theorem (Time to the most recent common ancestor)**.**
The time to the most recent common ancestor of a sample of size n is the coalescent model has distribution:

$$f_{T_{\mathrm{MRCA}}}(t) = \sum_{i=2}^{n} \frac{i(i-1)}{2} e^{-i(i-1)t/2} \prod_{j=2, j \neq i}^{n} \frac{j(j-1)}{j(j-1) - i(i-1)}$$

and mean and variance:

$$\mathbb{E}(T_{\mathrm{MRCA}}) = 2\left(1 - \frac{1}{n}\right) \text{ and } \mathrm{var}(T_{\mathrm{MRCA}}) = 4 \sum_{i=2}^{n} \frac{1}{i^2(i-1)^2}$$

Height of a coalescent tree

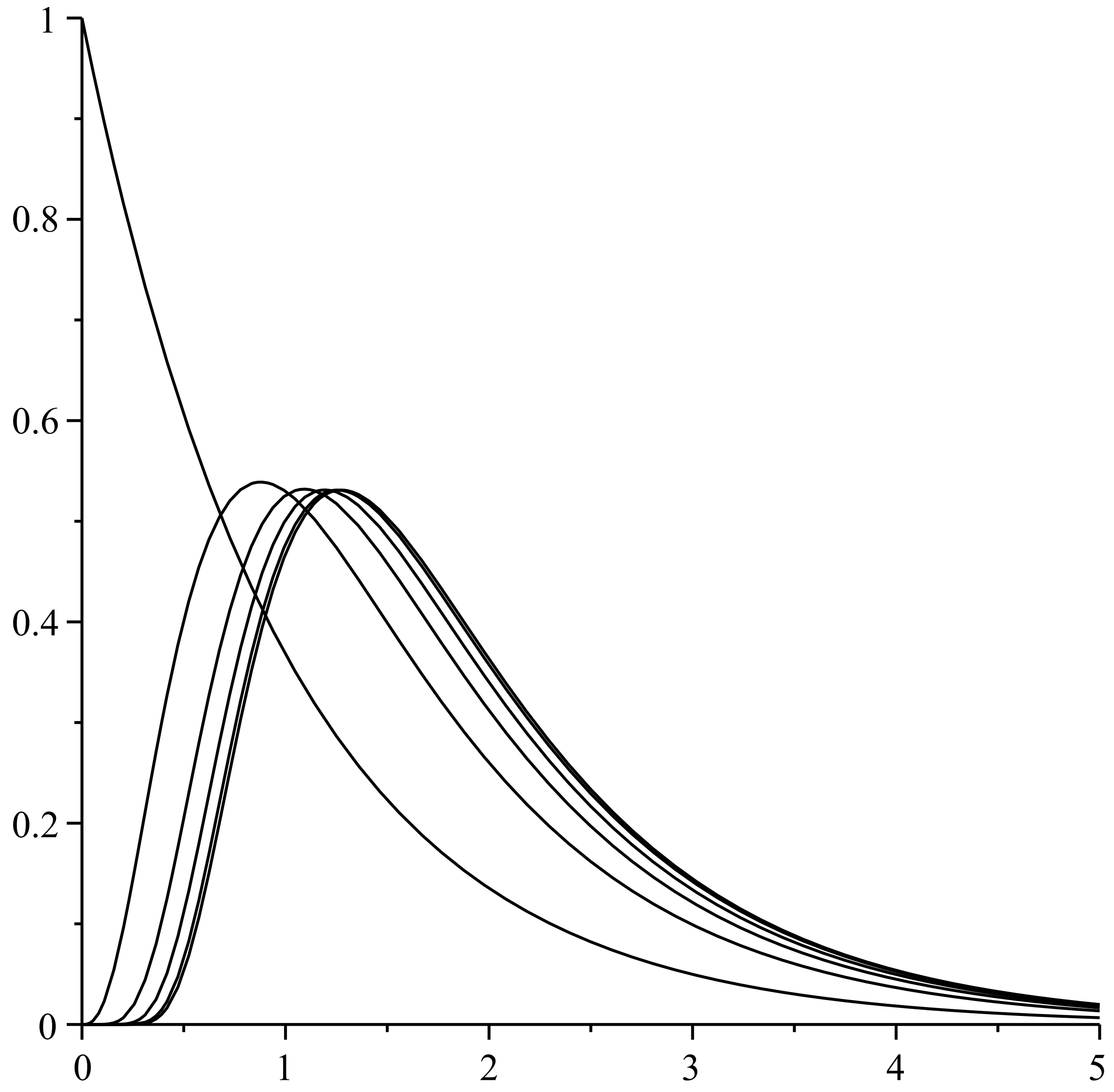
Proof. In a sample of k genes, the time T_k before the first coalescent event is Exponential with parameter $\frac{k(k-1)}{2}$. The time to the most recent common ancestor is then equal to $T_{\mathrm{MRCA}} = T_2 + T_3 + ... + T_n$. This is a sum of independently but non-identically distributed exponential variables. Thus we can use a convolution to find its probability density function. $\square$

Height of a coalescent tree

The expectation and variance of T_{MRCA} can be derived from the probability density function above, or more directly as follows:

$$\begin{aligned}\mathbb{E}(T_{\mathrm{MRCA}}) &= \sum_{i=2}^{n} \mathbb{E}(T_i) = \sum_{i=2}^{n} \frac{2}{i(i-1)} \\ &= 2\sum_{i=2}^{n}\left(\frac{1}{i-1} - \frac{1}{i}\right) = 2\left(1 - \frac{1}{n}\right) \\ \mathrm{var}(T_{\mathrm{MRCA}}) &= \sum_{i=2}^{n} \mathrm{var}(T_i) = 4\sum_{i=2}^{n} \frac{1}{i^2(i-1)^2}\end{aligned}$$

Height of a coalescent tree



Human effective population size

- Anthropological evidence shows that *Homo sapiens* appeared approximately 200,000 years ago in Africa.
- All humans must therefore share a most recent common ancestor (MRCA) less than 200,000 years ago.
- This is equal to 10,000 generations if we assume that each human generation lasted approximately 20 years.
- In coalescent theory, the expected height of a genealogical tree is equal to 2 coalescent units for a large sample size n.
- Since a coalescent unit of time is equal to M_e generations, this gives an estimate of 5,000 for the human effective population size M_e.

Total branch lengths

The following theorem is due to Tavaré (1984).

Theorem (Total branch lengths)**.**
The sum of branch lengths of a coalescent tree for a sample of size n has distribution:

$$f_{T_{\text{total}}}(t) = \sum_{i=2}^{n} \frac{i-1}{2} \exp\left(-\frac{i-1}{2}t\right) \prod_{\substack{j=2 \\ j\neq i}}^{n} \frac{j-1}{j-i}$$

and mean and variance:

$$\mathbb{E}(T_{\text{total}}) = 2\sum_{i=1}^{n-1} \frac{1}{i} \text{ and } \operatorname{var}(T_{\text{total}}) = 4\sum_{i=1}^{n-1} \frac{1}{i^2}$$

Total branch lengths

Proof. Let T_{total} denote the sum of branch lengths. Then: $T_{\text{total}} = \sum_{i=2}^{n} iT_i$. Notice that iT_i is Exponentially distributed:

$$\mathbb{P}(iT_i < t) = \mathbb{P}(T_i < t/i) = 1 - \exp\left(-\frac{i(i-1)}{2}\frac{t}{i}\right)$$

which is the cumulative distribution function of an exponential with parameter $(i-1)/2$. We can therefore use a convolution to find the probability distribution of T_{total}. □

Total branch lengths

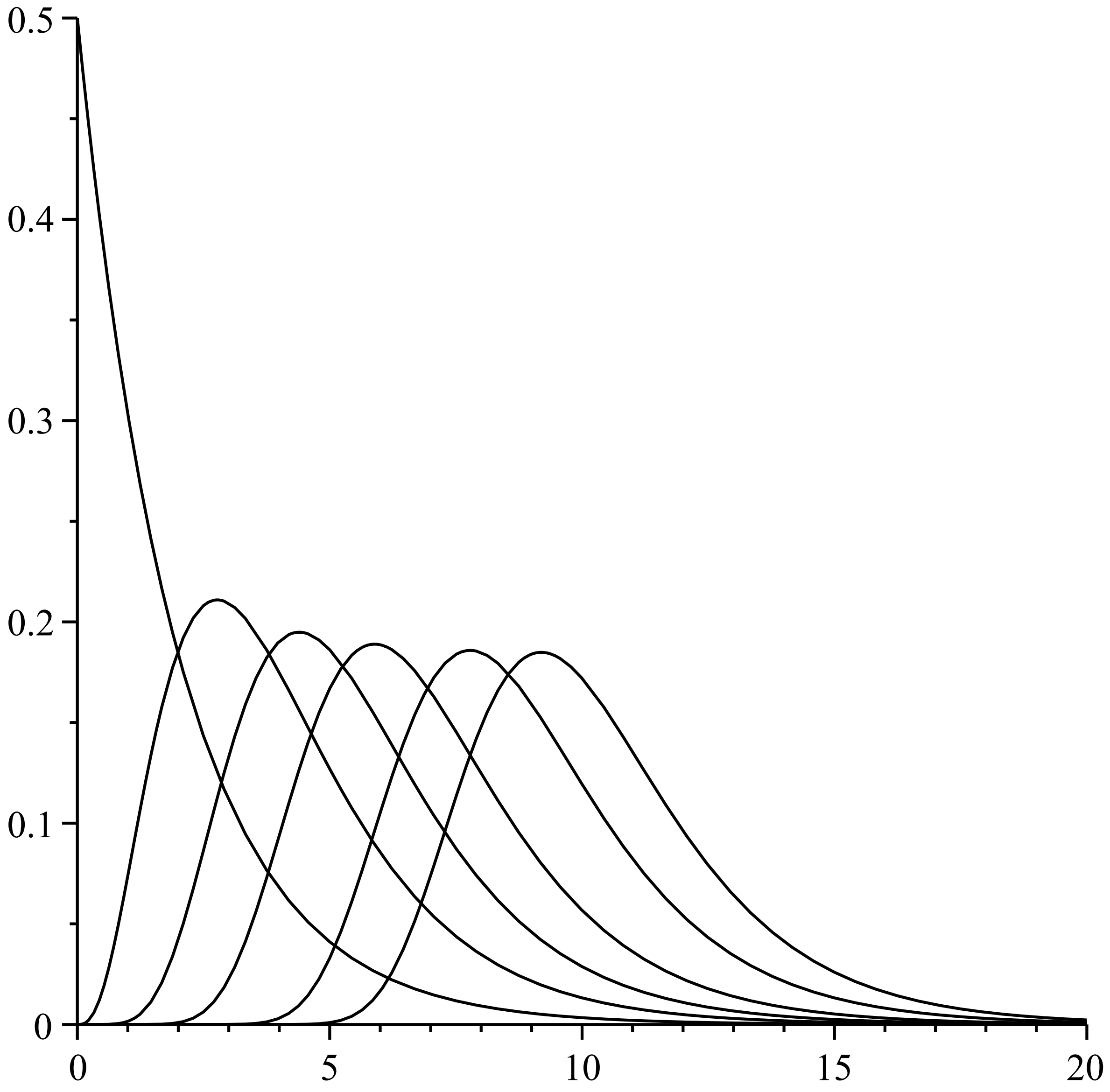
The expectation and variance can be derived from the above, or more directly:

$$\mathbb{E}(T_{\text{total}}) = \sum_{i=2}^{n} \mathbb{E}(iT_i) = \sum_{i=2}^{n} \frac{2i}{i(i-1)} = 2\sum_{i=1}^{n-1} \frac{1}{i}$$

$$\text{var}(T_{\text{total}}) = \sum_{i=2}^{n} \text{var}(iT_i) = \sum_{i=2}^{n} \frac{4i^2}{i^2(i-1)^2} = 4\sum_{i=1}^{n-1} \frac{1}{i^2}$$

□

Total branch lengths

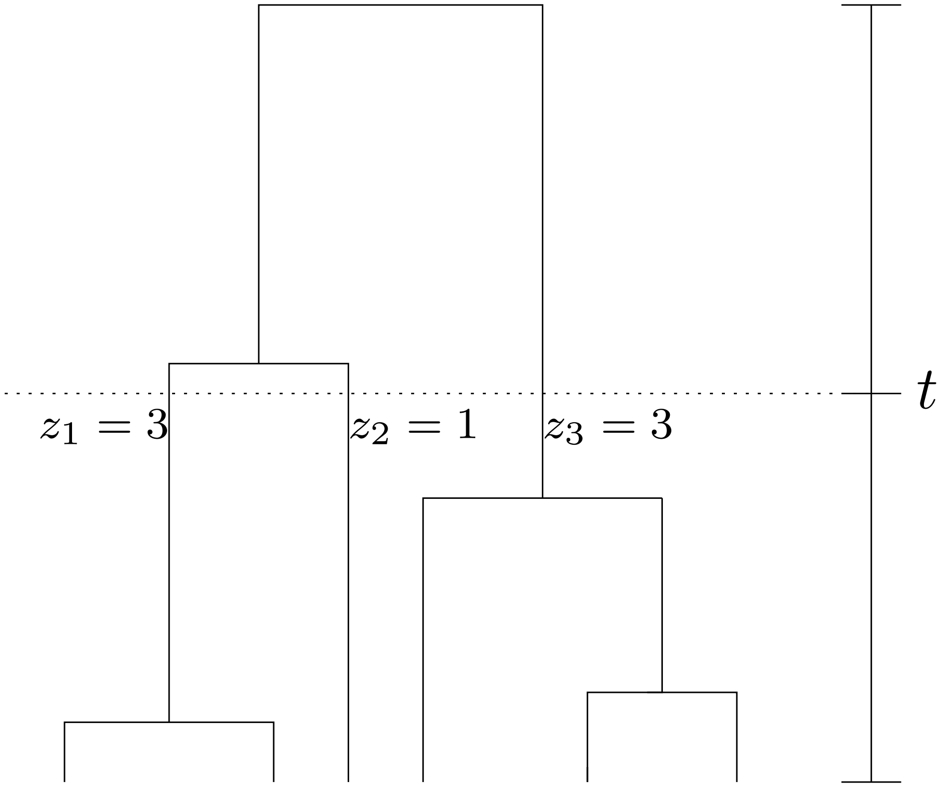


Hanging configuration

Theorem (Probability of hanging configuration)**.**
Consider a coalescent tree for which at time t back in time, there were k ancestral lineages. Let $Z = (z_1, ..., z_k)$ be the number of descendants of these k lineages such that $\sum z_i = n$. We have:

$$\mathbb{P}(Z = z|n, k) = \begin{pmatrix} n - 1 \\ k - 1 \end{pmatrix}^{-1}$$

Hanging configuration



Hanging configuration

Proof.

- We prove the result by induction on n.
- For $n = 2$ the hypothesis is true.
- Let us now assume that it is true for $n - 1$ and show that it is true for n.
- If $k = n$ then there is only one possibility $z = (1, ..., 1)$ and the hypothesis is true.
- If $k < n$, consider the class in which the last coalescent event happened. The probability that it is class i is $(z_i - 1)/(n - 1)$ since any branch is equally likely to split.
- Thus:

$$\begin{aligned}\mathbb{P}(Z = z|n, k) &= \sum_{i=1}^{k} \frac{z_i - 1}{n - 1} \mathbb{P}(Z = z - 1_i|n - 1, k) \\ &= \frac{n - k}{n - 1} \begin{pmatrix} n - 2 \\ k - 1 \end{pmatrix}^{-1} = \begin{pmatrix} n - 1 \\ k - 1 \end{pmatrix}^{-1}\end{aligned}$$

□

Hanging configuration

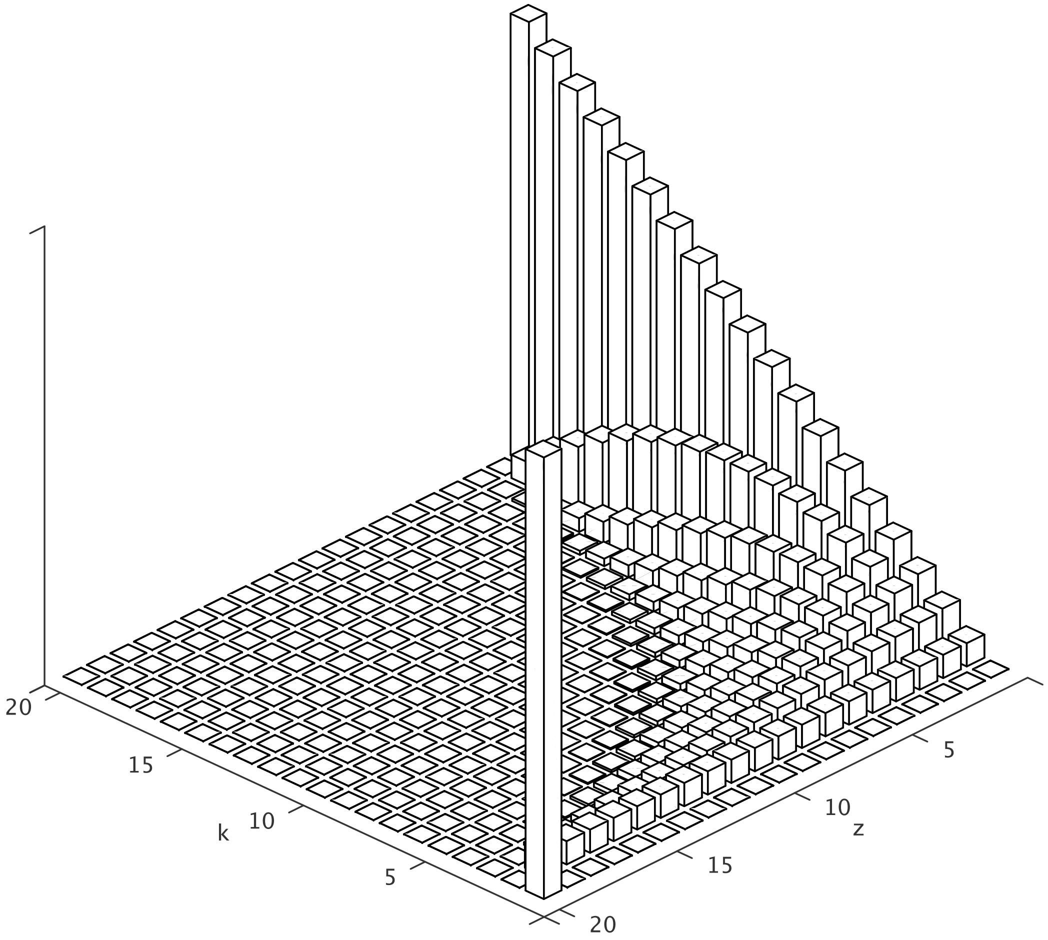
Lemma 1. In a coalescent tree for n genes, we consider a branch at a time back in time when there were k ancestral lineages. The distribution of the number Y of genes that are derived from that branch is:

$$\mathbb{P}(Y = y|n,k) = \frac{\begin{pmatrix} n-1-y \\ k-2 \end{pmatrix}}{\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}}$$

Lemma 2. The number of descendants on one side of the root of a coalescent tree is uniformly distributed.

Descendants of a branch

With $n = 20$:



Summary

- The coalescent model is **easy to simulate**
- We derived the distribution of the **time to the most recent common ancestor** for a sample
- We also derived the **sum of branch lengths** of a coalescent tree
- The probability distribution of **hanging configurations** is uniform