Statistical population genetics

Lecture 7: Infinite alleles model

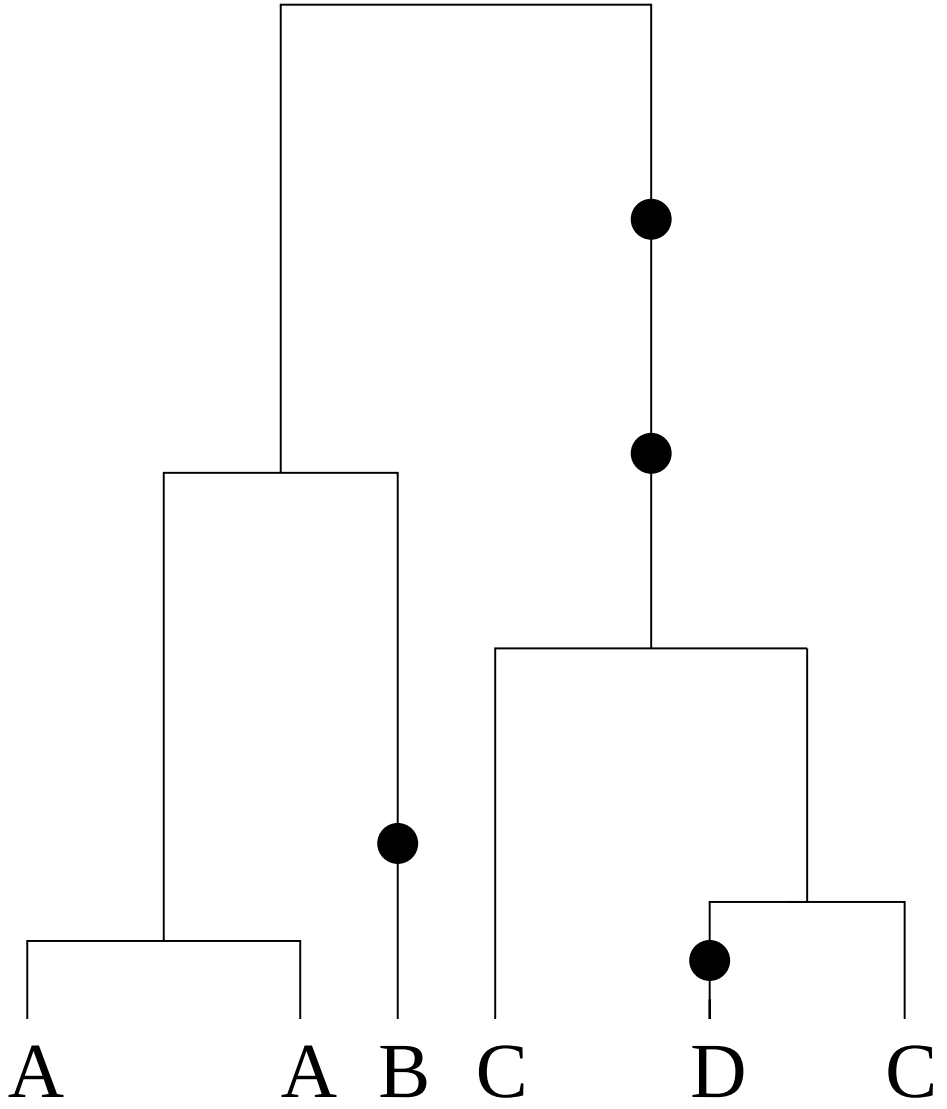
Xavier Didelot
Dept of Statistics, Univ of Oxford
didelot@stats.ox.ac.uk

Infinite alleles model

- We now discuss the **effect** of mutations.
- Kimura and Crow (1964) proposed the following mutational model:

  Definition (The infinite alleles model)**.**
  Each mutation creates a new allele.

Infinite alleles model



Infinite alleles model

- Data from the infinite alleles model can be represented as a vector $a = (a_1, ..., a_n)$ where a_i is the number of alleles for which i copies exist in the sample of size n.
- a is called the **allelic partition** of the data.
- $n = \sum_{i=1}^{n} i a_i$ and $K_n = \sum_{i=1}^{n} a_i$ is the number of allele types
- For example, in the previous slide, we have $n = 6$, $K_n = 4$ and $a = (2, 2, 0, 0, 0, 0)$.

Number of alleles

Theorem (Number of alleles)**.**

$$\mathbb{P}(K_n = k) = S(n,k)\frac{\theta^k}{\prod_{i=0}^{n-1}(\theta+i)}$$

where $S(n,k)$ is the Stirling number of the first kind.

Number of alleles

Proof.

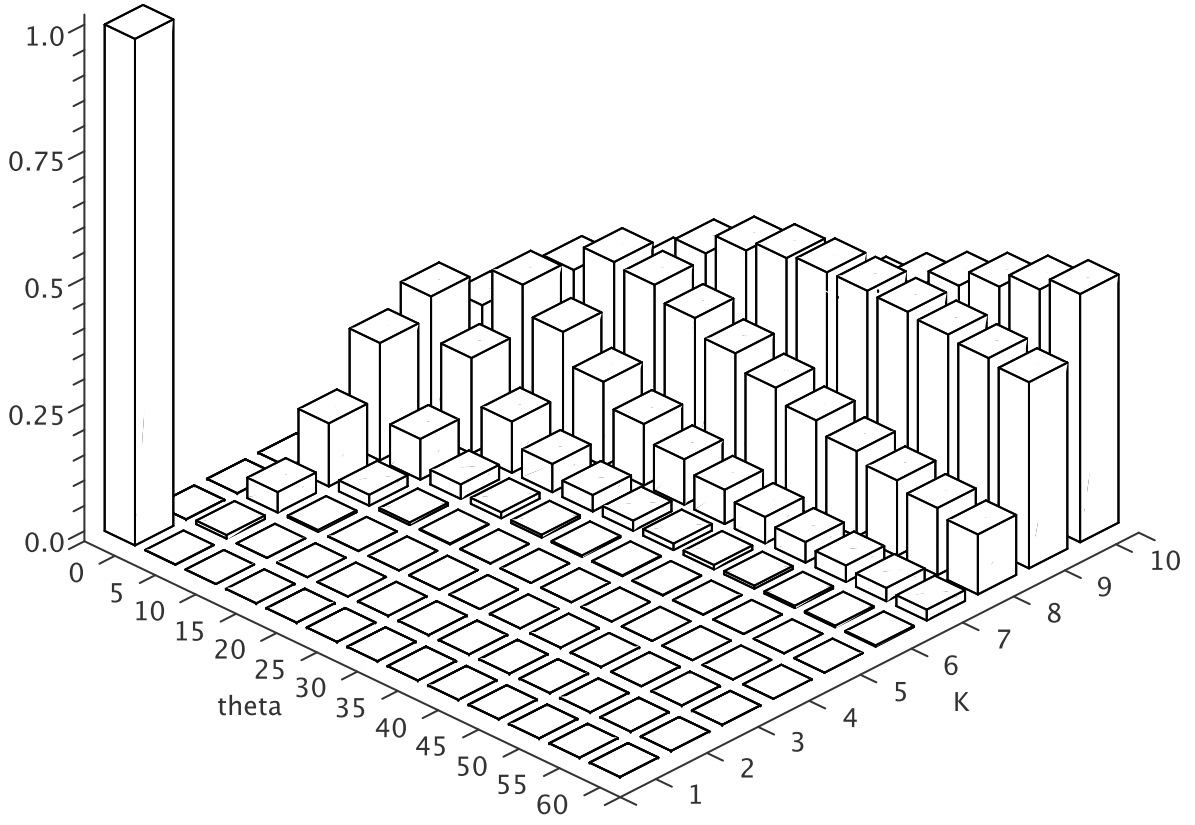
- If the last event was a coalescent, then just before that we had $n-1$ lineages and k distinct alleles.
- If the last event was a mutation, then the mutating lineage is a unique allele, and the $n-1$ other lineages contained $k-1$ distinct alleles.
- It follows that:

$$\mathbb{P}(K_n = k) = \frac{\theta}{n-1+\theta}\mathbb{P}(K_{n-1} = k-1) + \frac{n-1}{n-1+\theta}\mathbb{P}(K_{n-1} = k)$$

  with initial condition $\mathbb{P}(K_1 = 1) = 1$.

- Solving this recursive equation gives the result. □

Number of alleles



Ewens' sampling formula

Theorem (Ewens' sampling formula)**.**
The probability of an allelic partition a in a sample of size n is equal to:

$$P_n(a) = \frac{n!}{\prod_{i=0}^{n-1}(\theta+i)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}$$

- This formula is called **Ewens' sampling formula (ESF)** because it was discovered by Ewens (1972).
- The ESF has since been found to have many applications, and is thus an important result in theoretical probability.

Ewens' sampling formula

Proof. Let e_i be the vector of size n filled with zeros except for a one at the i-th position. We decompose $P_n(a)$ according to whether the last event was a coalescence (C) or a mutation (M):

$$\begin{aligned} P_n(a) &= \mathbb{P}(a|C)\mathbb{P}(C) + \mathbb{P}(a|M)\mathbb{P}(M) \\ &= \frac{n-1}{n-1+\theta}\mathbb{P}(a|C) + \frac{\theta}{n-1+\theta}\mathbb{P}(a|M) \end{aligned}$$

If the last event was a mutation, then the mutating lineage has a unique allelic type and the $n-1$ other lineages need to generate the rest of the profile, ie. $a - e_1$ so that $\mathbb{P}(a|M) = P_{n-1}(a - e_1)$. If $a_1 = 0$ then this probability is of course equal to zero.

Ewens' sampling formula

If the last event was a coalescence, we decompose $\mathbb{P}(a|C)$ according to all the profiles of size $n-1$ that could be observed just before the coalescence:

$$\mathbb{P}(a|C) = \sum_{a'} P_{n-1}(a')\mathbb{P}(a|C,a')$$

The coalescence may have happened between any two genes that share the same allele in a. Let j denote the number of copies in a of the allele of the genes that coalesced. Given j, we have $a' = a - e_j + e_{j-1}$. Thus:

$$\mathbb{P}(a|C) = \sum_{j=2}^{n} P_{n-1}(a - e_j + e_{j-1})\mathbb{P}(a|C, a - e_j + e_{j-1})$$

Ewens' sampling formula

The last term is the probability that a coalescence event happens to one of the $(j-1)(a_{j-1}+1)$ genes for which there are a_{j-1} copies in $a - e_j - e_{j-1}$. Since there are $n-1$ genes in $a - e_j - e_{j-1}$, we have:

$$\mathbb{P}(a|C, a - e_j + e_{j-1}) = \frac{(j-1)(a_{j-1}+1)}{n-1}$$

Putting this altogether we get:

$$P_n(a) = \frac{\theta}{n-1+\theta} P_{n-1}(a - e_1) + \frac{n-1}{n-1+\theta} \sum_{j=2}^{n} \frac{(j-1)(a_{j-1}+1)}{n-1} P_{n-1}(a - e_j + e_{j-1})$$

with boundary condition $P_1(1) = 1$ and $P_n(a) = 0$ if any of the $a_j < 0$. Solving this recursion equation leads to the ESF. □

Example

For a sample of size $n = 3$, there are three possible allelic profiles: $(3, 0, 0)$, $(1, 1, 0)$ and $(0, 0, 1)$ with respective probabilities:

$$
\begin{aligned}
P_3(3,0,0) &= \frac{\theta}{\theta+2} P_2(2,0) = \frac{\theta^2}{(\theta+1)(\theta+2)} \\
P_3(1,1,0) &= \frac{\theta}{\theta+2} P_2(0,1) + \frac{2}{\theta+2} P_2(2,0) \\
&= \frac{\theta}{\theta+2}\frac{1}{\theta+1} + \frac{2}{\theta+2}\frac{\theta}{\theta+1} \\
&= \frac{3\theta}{(\theta+1)(\theta+2)} \\
P_3(0,0,1) &= \frac{2}{\theta+2} P_2(0,1) = \frac{2}{(\theta+1)(\theta+2)}
\end{aligned}
$$

Sufficiency of number of alleles

Definition (Sufficiency of a statistic)**.**
A statistic $T(X)$ is sufficient for underlying parameter θ if the conditional distribution of the data X given the statistic $T(X)$ is independent of θ, ie:

$$\mathbb{P}(X|T(X),\theta) = \mathbb{P}(X|T(X))$$

Theorem (Sufficiency of the number of alleles)**.**
The number of alleles is a sufficient statistic for parameter θ.

Sufficiency of number of alleles

Proof. Since the number of alleles K_n is completely determined by the allelic profile a, the distribution of a given K_n reduces to:

$$\mathbb{P}(a|K_n = k, \theta) = \frac{P_n(a)}{\mathbb{P}(K_n = k)} = \frac{n!}{S(n,k)} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}$$

This distribution does not depend on θ, therefore K_n is sufficient for parameter θ. □

Example

- Coyne (1976) studied the xanthine dehydrogenase gene (Xdh) of Drosophila persimilis by **electrophoresis**.
- This method reveals whether two genes are identical, but not how closely related they are.
- The infinite alleles model is therefore particularly well suited for the analysis of such data.
- They found $K_n = 23$ alleles in a sample of $n = 60$ individuals with the following allelic profile:

$$a_1 = 18,\ a_2 = 3\ , a_4 = 1,\ a_{32} = 1$$

- What is the maximum likelihood estimator of θ based on this data?

Example

- Since K_n is sufficient for θ, we estimate θ based on K_n only.
- The likelihood of θ is:

$$\mathbb{L}(\theta) = \mathbb{P}(K_{60} = 23|\theta) = S(60, 23)\frac{\theta^{23}}{\prod_{i=0}^{59}(\theta + i)}$$

- Taking the logarithm and deriving by θ gives:

$$\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta} = \frac{23}{\theta} - \sum_{i=0}^{59}\frac{1}{\theta + i}$$

- This is equal to zero when:

$$23 = \sum_{i=0}^{59}\frac{\theta}{\theta + i}$$

- Solving gives a maximum likelihood estimator for θ of 13.17.

Summary

- In the **infinite alleles model**, each mutation creates a new allele
- The **Ewens' sampling formula** gives the probability of a dataset occurring under this mutational model
- We derived an equation for the **number of alleles**
- The **number of alleles** is a sufficient statistic in this model, making it very useful to draw **inference** from genetic data
- The infinite alleles model is particularly well suited to Analise data from **electrophoresis**