Lecture 18 : Ewens' sampling formula

MATH285K - Spring 2010 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 1.3].

Previous class

In the previous lecture, we introduced Kingman's coalescent as a limit of the Wright-Fisher model for the lineages—backwards in time—of a small sample in a large population. One of the main advantages of the coalescent is its robustness. The coalescent also emphasizes that through their joint genealogical process the samples are correlated. That correlation structure plays an important role in the design of good statistical estimators as we will see in this lecture.

Recall that the coalescent and the Wright-Fisher models in themselves are not particularly interesting because all genetic variation is lost through fixation—a phenomenon known as genetic drift. We now introduce mutations.

1 Infinite-alleles model

We begin with a simple model of mutation known as the *infinite-alleles model*. Imagine that we are looking at a gene which has several variants called alleles. We ignore the details of the differences between the various alleles (that is, we do not know their sequence) and we assume that each time a mutation occurs it creates a new allele. Let $\theta/2$ be the rate at which mutations occur in each individual (in the rescaled time of the coalescent). The infinite-allele model can be described formally in two equivalent, but equally useful, ways.

DEF 18.1 (Infinite-alleles model: First definition) *First generate a coalescent on n samples. Conditioned on the tree obtained, generate an independent Poisson point process on the tree with rate $\theta/2$. Each sample inherits the last allele created on the path from the root (or the state at the root if no mutation occured). (Recall that such a Poisson point process can be obtained by picking an infinite sequence of independent uniform random points of the tree and keeping only the first Z where Z is Poisson with mean $T_{\mathrm{tot}}\theta/2$ with T_{tot} the total length of the tree.)*

Recall the following facts about exponential random variables.

LEM 18.2 (Minimum of exponentials) *Let $X_1, \ldots, X_k$ be k independent exponentials with parameters $\lambda_1, \ldots, \lambda_k$. Let*

$$Z = \min\{X_1, \ldots, X_k\}.$$

Then

1. *Z is exponential with parameter $\lambda_1 + \cdots + \lambda_k$.*
2. *$\mathbb{P}[Z = X_i] = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}$.*

Proof: Note that

$$\mathbb{P}[Z > z] = \prod_{i=1}^{k} \mathbb{P}[X_i > z] = \exp\left(-z \sum_{i=1}^{k} \lambda_i\right).$$

Similarly

$$\begin{aligned}
\mathbb{P}[Z = X_i] &= \mathbb{E}[\mathbb{P}[Z = X_i \,|\, X_i]] \\
&= \mathbb{E}\left[\exp\left(-X_i \sum_{j \neq i} \lambda_j\right)\right] \\
&= \int_0^\infty \mathrm{d}x \lambda_i e^{-\lambda_i x} \exp\left(-x \sum_{j \neq i} \lambda_j\right).
\end{aligned}$$

■

LEM 18.3 (Memoryless property) *If Z is exponential and $s < t$, then*

$$\mathbb{P}[Z > t \,|\, Z > s] = \mathbb{P}[Z > t - s].$$

Proof: Immediate from definition of conditional probability. ■

This leads to an equivalent definition of the infinite-alleles model where coalescence and mutations are generated simultaneously.

DEF 18.4 (Infinite-alleles model: Second definition) *For a partition Π on n samples, we let $|\Pi|$ be the number of sets in Π. Consider the following algorithm with n and θ as input:*

- *Set $\Pi = \{\{1\}, \ldots, \{n\}\}$.*

- *Repeat until* $|\Pi| = 1$*:*
  - *Set* $k := |\Pi|$.
  - *After an exponential time with parameter* $\binom{k}{2} + k\frac{\theta}{2}$ *(going backwards in time):*
    - *With probability*

$$\frac{k\frac{\theta}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{\theta}{\theta + k - 1},$$

generate a mutation in a uniformly random lineage.

    - *With probability*

$$\frac{\binom{k}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{k-1}{\theta + k - 1},$$

merge two uniformly random lineages.

2 Homozygosity

The infinite-alleles model has one parameter, the mutation rate θ. We discuss statistical estimators. A quantity that is naturally related to the mutation rate is the *homozygosity*, that is, the probability that two uniformly chosen samples have the same allele

$$\hat{F}_n = \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \delta_{i,j},$$

where the sum is over distinct pairs in $\{1, \ldots, n\}$ and $\delta_{i,j}$ is 1 if samples i and j have the same allele and 0 otherwise.

To illustrate the two equivalent definition of the model, we compute the expectation of $\hat{F}_n$ under both:

THM 18.5 (Homozygosity) *We have*

$$\mathbb{E}\left[\hat{F}_n\right] = \frac{1}{1+\theta}.$$

Proof:(Proof 1) By linearity of expectations, it is enough to consider a 2-coalescent. The probability that there are no mutations on the path between 1 and 2 is

$$\mathbb{E}[\delta_{1,2}] = \mathbb{P}[\delta_{1,2} = 1] = \mathbb{E}[\mathbb{P}[\delta_{1,2} = 1 \,|\, T_{\text{tot}}]] = \mathbb{E}[e^{-2T_{\text{tot}}(\theta/2)}] = \int_0^\infty \mathrm{d}t e^{-t} e^{-t\theta}.$$

■

Proof:(Proof 2) The probability that a coalescence occurs before the first mutation is

$$\frac{2-1}{\theta+2-1} = \frac{1}{1+\theta}.$$

■

However, because of the correlation in the samples, the variance of $\hat{F}_n$ does not converge to 0.

THM 18.6 *As $n \to \infty$, we have*

$$\text{Var}\left[\hat{F}_n\right] \to \frac{2\theta}{(1+\theta)^2(2+\theta)(3+\theta)}.$$

Proof: We begin by computing the second moment. Note that

$$\mathbb{E}\left[\hat{F}_n^2\right] = \mathbb{E}\left[\binom{n}{2}^{-2} \sum_{\substack{\{i_1,j_1\}\\\{i_2,j_2\}}} \delta_{i_1,j_1}\delta_{i_2,j_2}\right].$$

Because of the $O(n^{-4})$ in front, in the limit only the terms with $|\{i_1, j_1, i_2, j_2\}| = 4$ contribute to the second moment. Note that there are $\binom{n}{2}\binom{n-2}{2}$ such terms so that

$$\mathbb{E}\left[\hat{F}_n^2\right] \to \mathbb{E}[\delta_{1,2}\delta_{3,4}].$$

Ewens' sampling formula derived below will give the result. ■

3 Ewens' sampling formula

For $i = 1, \ldots, n$, let a_i be the number of alleles that appear i times in the sample and let $q(\mathbf{a})$ be the distribution of $\mathbf{a} = (a_1, \ldots, a_n)$. Ewens showed that q satisfies the following recursion:

THM 18.7 *Letting $\mathbf{e}_i$ be the unit vector in direction i, we have*

$$q(\mathbf{a}) = \frac{\theta}{\theta+n-1}\left[\frac{a_1}{n}q(\mathbf{a}) + \sum_{j=2}^{n}\frac{j(a_j+1)}{n}q(\mathbf{a}-\mathbf{e}_1-\mathbf{e}_{j-1}+\mathbf{e}_j)\right]$$
$$+ \frac{n-1}{\theta+n-1}\left[\sum_{j=1}^{n}\frac{j(c_j+1)}{n}q(\mathbf{a}+\mathbf{e}_j-\mathbf{e}_{j-1})\right].$$

Proof: The proof is left to the reader. Hint: condition on the first event in the second definition of the infinite-alleles model. ■

By working out basic cases, Ewens conjectured that:

THM 18.8 (Ewens' sampling formula) *Letting* $\|\mathbf{a}\| = \sum_{i=1}^{n} ia_i$*, in a sample of size* n *we have*

$$q(\mathbf{a}) = \mathbb{1}\{\|a\| = n\} \frac{n!}{\theta_{(n)}} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!},$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

Note that for $n = 2$,

$$q((2,0)) = \frac{2!}{\theta(\theta+1)} \left(\frac{\theta}{1}\right)^2 \frac{1}{2!} = \frac{\theta}{1+\theta},$$

and

$$q((0,1)) = \frac{2!}{\theta(\theta+1)} \left(\frac{\theta}{2}\right)^1 \frac{1}{1!} = \frac{1}{1+\theta},$$

confirming our previous calculations on homozygosity.

We will give Kingman's proof of Ewens' sampling formula in the next lecture.

Further reading

The material in this section was taken from Section 1.3 of the excellent monograph [Dur08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.