Problem set 3: Population genetics

Erik Norlin & Mattias Wiklund

March, 2023

A)

The Poisson distributed number of mutations can be written as a conditional probability distribution $P(S_n = j|T_2, \dots, T_n)$ according to

$$P(S_n = j|T_2, \dots, T_n) = \frac{(\mu T_c)^j}{j!} e^{-\mu T_c}, \tag{1}$$

where the total branch length of coalescence times T_c is given by

$$T_c = \sum_{k=2}^{n} k T_k.$$

The joint probability distribution $P(S_n = j, T_2, \dots, T_n)$ is given by

$$P(S_n = j, T_2, \dots, T_n) = P(S_n = j|T_2, \dots, T_n) P(T_2, \dots, T_n).$$

Note that the times T_j are independent random variables, which results in

$$P(T_2, \dots, T_n) = P(T_2) \dots P(T_n).$$

The probability for T_k have an exponential distribution according to

$$P(T_k) = \frac{\binom{k}{2}}{N} e^{-\frac{\binom{k}{2}}{N} T_k}.$$

The marginal probability distribution for $P(S_n = j)$, the probability that there are j SNPs in a sample of size n, can now be obtained by integrating the joint probability distribution over all T_k.

$$P(S_n = j) = \int_0^\infty \mathrm{d}T_2 \cdots \int_0^\infty \mathrm{d}T_n P(S_n = j, T_2, \dots, T_n) =$$

$$= \int_0^\infty \mathrm{d}T_2 \cdots \int_0^\infty \mathrm{d}T_n P(S_n = j|T_2, \dots, T_n) P(T_2) \dots P(T_n). \tag{2}$$

In this case, we are interested in $P(S_n = 0)$, the probability to not have any SNPs in a sample of size n. Equation (1) with $j = 0$ gives

$$P(S_n = 0|T_2, \dots, T_n) = e^{-\mu T_c} = e^{-\mu \sum_{k=2}^{n} j T_j} = \prod_{k=2}^{n} e^{-\mu k T_k}.$$

Equation (2) now gives

$$P(S_n = 0) = \int_0^\infty \mathrm{d}T_2 \cdots \int_0^\infty \mathrm{d}T_n \prod_{k=2}^{n} e^{-\mu k T_k} \frac{\binom{k}{2}}{N} e^{-\frac{\binom{k}{2}}{N} T_k} =$$

$$= \prod_{k=2}^{n} \int_0^\infty e^{-\mu k T_k} \frac{\binom{k}{2}}{N} e^{-\frac{\binom{k}{2}}{N} T_k} \mathrm{d}T_k$$

Using the standard integral

$$\int_0^\infty e^{-kx} \mathrm{d}x = \frac{1}{k} \; (k > 0),$$

the integral above can be evaluated as

$$\frac{\binom{k}{2}}{N} \int_0^\infty \exp\left[-T_k\left(\mu k + \frac{\binom{k}{2}}{N}\right)\right] \mathrm{d}T_k = \frac{\binom{k}{2}}{N} \frac{1}{\mu k + \frac{\binom{k}{2}}{N}} = \frac{\binom{k}{2}}{N} \frac{N}{kN\mu + \binom{k}{2}}$$

$$= \frac{k(k-1)}{2} \frac{1}{kN\mu + \frac{k(k-1)}{2}} = \frac{k(k-1)}{2} \frac{2}{2kN\mu + k(k-1)} =$$

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

Finally, this gives

$$P(S_n = 0) = \prod_{k=2}^{n} \int_0^\infty e^{-\mu k T_k} \frac{\binom{k}{2}}{N} e^{-\frac{\binom{k}{2}}{N} T_k} \mathrm{d}T_k$$

$$= \prod_{k=2}^{n} \frac{k-1}{k-1+\theta} = \frac{(n-1)!}{(1+\theta)(2+\theta)\dots(n-1+\theta)}.$$

To conclude, the the probability to not have any SNPs in a sample of size n is

$$P(S_n = 0) = \frac{(n-1)!}{(1+\theta)(2+\theta)\dots(n-1+\theta)}.$$

B)

In this case, we are interested in $P(S_2 = j)$, the probability to have j SNPs in a sample of size 2. Equation (1) with $n = 2$ gives

$$P(S_2 = j|T_2) = \frac{(\mu T_c)^j}{j!} e^{-\mu T_c} = \frac{(2\mu T_2)^j}{j!} e^{-2\mu T_2}.$$

Using the standard integral

$$\int_0^\infty x^j e^{-kx} \mathrm{d}x = \frac{j!}{k^{j+1}} \; (j \in \mathbb{N}, k > 0),$$

Equation (2) now gives

$$P(S_2 = j) = \int_0^\infty \frac{(2\mu T_2)^j}{j!} e^{-2\mu T_2} \frac{1}{N} e^{-\frac{T_2}{N}} \mathrm{d}T_2 = \frac{(2\mu)^j}{Nj!} \int_0^\infty T_2^j \exp\left[-T_2\left(2\mu + \frac{1}{N}\right)\right] \mathrm{d}T_2 =$$

$$= \frac{(2\mu)^j}{Nj!} \frac{j!}{\left(2\mu + \frac{1}{N}\right)^{j+1}} = \frac{(2\mu)^j}{N} \frac{N^{j+1}}{(2N\mu + 1)^{j+1}} = \frac{(2N\mu)^j}{(1+2N\mu)^{j+1}} = \frac{\theta^j}{(1+\theta)^{j+1}} =$$

$$= \frac{1}{1+\theta} \left(\frac{\theta}{1+\theta}\right)^j.$$

To conclude, the probability to have j SNPs in a sample of size 2 is

$$P(S_2 = j) = \frac{1}{1+\theta} \left(\frac{\theta}{1+\theta}\right)^j.$$