Solutions Exam - Topics in Probability and Statistics (WMMA039-05) 2022/2023

Date and time: January 24, 2023, 18.15-20.15

Exercise 1

Consider the infinite sites variant of the standard Wright-Fisher model on $2N$ haploid individuals of which we analyse a segment of a chromosome in a uniform sample of size n. Assume that $n << 2N$ and that you can use Kingman's Coalescence to obtain the geneology of the sample. Let μ be the mutation probability per generation per individual (in the segment under consideration) and let $\theta = 4N\mu$.

Let S_n be the number of segregating sites in the sample. Let $h_n = \sum_{j=1}^{n-1} 1/j$ and define $\hat{\theta} = S_n/h_n$.

(a) Show that $\hat{\theta}$ is an unbiased estimator of θ. That is, show $\mathbb{E}[\hat{\theta}] = \theta$. 10

(b) Compute the variance of $\hat{\theta}$. You may use without proof that for random variables X and Y we have $Var(X) = \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y])$. 15

Solution

This is Therorem 1.22 of book, and requires proof of Theorems 1.20 and 1.21
Because we consider the infinite sites model, the number of segregating sites is equal to the number of mutations in the genealogy from the most-recent common ancestor on. We can use Kingman's coalescence and use that this is a Markov process and the time until the next "merging" if there are now k lineages (say T_k) is exponentially distributed with expectation $2/[k(k-1)]$. If there are k lineages, mutations occur according to a Poisson process at rate $k \times 2N\mu = k\theta/2$. So the expected number of mutations while there are k lineages in the genealogy is $[k\theta/2] \times 2/[k(k-1)] = \theta/(k-1)$. The expected number of mutations in the genealogy of a sample of size n is therefore

$$\mathbb{E}[S_n] = \theta/(n-1) + \theta/((n-1)-1) + \cdots + \theta/(2-1) = \theta h_n$$

and thus $\mathbb{E}[\hat{\theta}] = \theta$, which is what we should prove in part (a).

To compute the variance we denote the total number of mutations while there are k lineages in the genealogy by M_k and by independence of mutations in the different parts of the geneology

$$Var[S_n] = \sum_{k=2}^{n} Var(M_k) = \sum_{k=2}^{n} \left(\mathbb{E}[Var(M_k|T_k)] + Var(\mathbb{E}[M_k|T_k]\right),$$

Note that $Var(M_k|T_k) = (\theta/2)kT_k$ and $\mathbb{E}[M_k|T_k] = (\theta/2)kT_k$, because mutations occur according to a Poisson process with intensity $\theta/2$ on the genealogy. Furthermore, $\mathbb{E}[T_k] = 2/[k(k-1)]$ and $Var(T_k) = 4/[k^2(k-1)^2]$. This leads to

$$Var[S_n] = \sum_{k=2}^{n} \left((\theta/2)k\mathbb{E}[T_k] + (\theta/2)^2 k^2 Var(T_k)\right) = \theta \sum_{k=2}^{n} \frac{1}{k-1} + \theta^2 \sum_{k=2}^{n} \frac{1}{(k-1)^2}.$$

and $Var(\hat{\theta}) = \theta/h_n + \theta^2 \sum_{k=2}^{n} \frac{1}{(k-1)^2}/(h_n)^2$.

Exercise 2

Consider the following adaptation of the Wright-Fisher model. The $2N$ individuals in generation t (for $t \in \mathbb{Z}$) independently have a geometrically distributed number of children with mean $1/p$ (and thus variance $(1-p)/p^2$), where $p \in (0,1)$. That is, the number of children of an individual is distributed as the random variable X, with

$$\mathbb{P}(X = k) = p(1-p)^{k-1} \qquad \text{for } k = 1, 2, \cdots.$$

Then immediately upon birth all but $2N$ uniformly chosen (without replacement) children of the individuals in generation t die. The remaining $2N$ individuals survive and form generation $t+1$.

The probability that two uniformly chosen individuals (without replacement) from generation $t+1$ have the same parent converges in probability to a constant as $2N \to \infty$.

(a) Compute the limiting probability that two uniformly chosen individuals (without replacement) from generation $t+1$ have the same parent? 15

(b) How many generations should go in a time unit to obtain Kingman's coalescent as the large population limit of the genealogy of a sample of finite size? 5

Solution

This problem is closely related to claim in first 10 lines of page 126 of book.
The probability that 2 uniformly chosen individuals out of all children of generation t individuals have the same parent is the same as the probability that 2 uniformly chosen individuals out of a uniformly chosen subset of size 2N from all children of generation t individuals have the same parent.

Let X_i be the number of children (surviving or not) of the i-th individual in generation t So conditioned on $X_1, \cdots, X_{2N}$, the probability that 2 uniformly chosen individuals from generation $t+1$ have the same parent is

$$\frac{\sum_{i=1}^{2N} X_i(X_i-1)/2}{\left(\sum_{i=1}^{2N} X_i\right)\left(\left(\sum_{i=1}^{2N} X_i\right)-1\right)/2} = \frac{1}{2N}\frac{\frac{1}{2N}\sum_{i=1}^{2N} X_i(X_i-1)}{\left(\frac{1}{2N}\sum_{i=1}^{2N} X_i\right)\left(\left(\frac{1}{2N}\sum_{i=1}^{2N} X_i\right)-\frac{1}{2N}\right)}.$$

By the weak law of large numbers the numerator of the Right hand side converges in probability to

$$\mathbb{E}[X(X-1)] = Var(X) - \mathbb{E}[X] + (\mathbb{E}[X])^2 = \frac{1-p}{p^2} - \frac{1}{p} + \frac{1}{p^2} = \frac{2(1-p)}{p^2},$$

while the denominatior of the second factor of the Right hand side converges in probability to $(1/p)^2$. If you want to be completely formal, you can now use Skorohod's representation theorem to show that the answer to the exercise is $\frac{1}{2N}2(1-p)$.

This means that if there are $N/(1-p)$ generations in a time unit, each pair of lineages merges at rate 1 per time unit and Kingman's coalescent is obtained.

Exercise 3

Consider the Moran-model with selection: We consider a population of constant size of $2N$ haploid individuals. Suppose that a mutation occurs (say at time 0) which gives the mutant a selection advantage. Further assume that no other mutations occur. Non-mutants give birth to a copy of themselves independently at rate 1, while mutants give independently birth to copies op themselves at rate $1+s$, where s is strictly positive, but very small. At the moment an individual is born (mutant or non-mutant) a uniformly chosen individual from the population present just before this birth, dies.

Compute the probability of fixation. That is, compute the probability that at some time in the future all $2N$ individuals will carry the mutation. 15

Solution

This is Theorem 6.1 of book
Because we are considering the Moran model with probability 1, there are no simultanious births and the number of mutants changes at most by 1 at a time. If there are i mutants at present the rate at which a mutant gives birth multiplied by the probability that the new-born replaces a non-mutant (i.e. the rate at which the number of mutants increases by 1) is given by

$$b_i = (1+s)i \times \frac{2N-i}{2N},$$

while the rate at which a non-mutant gives birth multiplied by the probability that the new-born replaces a mutant (i.e. the rate at which the number of mutants decreases by 1) is given by

$$d_i = (2N-i) \times \frac{i}{2N-i}.$$

If we look at the embedded discrete time Markov Chain, we obtain a random walk with $1 - p_{i,i-1} = p_{i,i+1} = b_i/(b_i+d_i) = (1+s)/(2+s)$ for $i = 1, 2, \cdots, 2N-1$. Let h_i be the probability that the random walk reaches $2N$ before reaching 0 if the random walk starts in state i. Note that the answer to the exercise is h_i. Further, $h_0 = 0$ and $h_{2N} = 1$ and

$$h_i = p_{i,i+1}h_{i+1} + p_{i,i-1}h_{i-1} = \frac{1+s}{2+s}h_{i+1} + \frac{1}{2+s}h_{i-1} \qquad \text{for } i = 1, 2, \cdots, 2N-1.$$

Rearanging gives

$$h_{i+1} - h_i = \frac{1}{1+s}(h_i - h_{i-1}) = \cdots = \frac{1}{(1+s)^i}(h_1 - h_0) = \frac{1}{(1+s)^i}h_1.$$

To compute h_1, we observe

$$1 = h_{2N} = h_{2N} - h_0 = \sum_{i=0}^{2N-1}(h_{i+1} - h_i) = h_1 \sum_{i=0}^{2N-1} \frac{1}{(1+s)^i} = h_1 \frac{1-(1+s)^{-2N}}{1-(1+s)^{-1}}$$

So $h_1 = s/[1+s-(1+s)^{-(2N-1)}$.

Exercise 4

For parts (a) and (b), consider the infinite sites model of the standard Wright-Fisher model on $2N$ haploid individuals, with a uniform sample of size $n = 5$.

We read out the nucleotides at the sites of a segment of a chromosome on which recombination is not possible. The segregating sites in our sample are denoted by letters $a, b, \cdots, i$ and the nucleotides on those sites for the different individuals are given in the following table.

Individual \ Sites	a	b	c	d	e	f	g	h	i
1	C	A	C	A	A	C	A	A	A
2	A	C	C	A	A	C	C	A	A
3	C	A	C	A	A	C	A	A	A
4	A	A	A	C	C	A	A	C	A
5	A	A	C	A	A	C	C	A	C

(a) Assume that the nucleotides on sites $a, \cdots, i$ of the most recent common ancestor of the sample are all A's. Draw a geneology of the sample. include the possible locations of the mutations in it (with their site label), that leads to the above table. **Hint:** The answer does not necessarily allow for both depiction of individuals 1 to 5 in that order and for a depiction of the genealogy without crossing lineages. 10

(b) Now assume that you do not know the nucleotides on sites $a, \cdots, i$ of the most recent common ancestor of the sample. Draw a geneology for which the most recent common ancestor of individual 1 and 3 is the same as the most recent common ancestor of the entire sample, but still leading to the above table. What are the nucleotides on sites $a, \cdots, i$ of the most recent common ancestor of the entire sample, based on this geneology? 10

(c) Assume that in another population, we take a sample of size $n = 5$, and we obtain the following table giving nucleotides at all 10 segregating sites in the sample.

Individual \ Sites	a'	b'	c'	d'	e'	f'	g'	h'	i'	j'
1	C	A	C	A	A	C	A	A	A	A
2	A	C	A	A	A	A	C	A	A	A
3	A	A	A	A	A	A	A	A	A	C
4	A	A	A	C	C	A	A	C	A	A
5	A	A	A	A	A	A	A	A	C	A

Assume that you know that this second population was either exponentially growing or exponentially decaying, but that apart from that the assumptions of the Wright-Fisher model apply to the population. Based on the nucleotides on sides $a', \cdots, j'$ is it more likely that the population is exponentially growing or exponentially decaying? Argue why. 10

Solution

(a) From site a we can deduce that 1 and 3 have a common ancestor who is not an ancestor of any of the other individuals. From site g we can deduce that individuals 2 and 5 have a common ancestor who is not an ancestor of any of the other individuals. From site c we then deduce that individuals 1,2,3,5 have a common ancestor that is not an ancestor of site 4. This gives a genealogy. Mutations a occurs in line segment from which only 1 and 3 descend. mutation b occurs in line segment from which only individual 2 descents etc.

(b) Note that individual 1 and 3 are identical, and since their most recent common ancestor is the most recent common ancestor of the entire sample, the most recent common ancestor of the population should be identical on the segregating sites to indivduals 1 and 3, because any mutation on line segments leading from this most recent common ancestor to individual 1, will cause individual 1 and 3 to differ from eachother. By site g, 2 and 5 must have common ancestor that is not an ancestor of other individuals in the sample. Similarly by site a, individuals 2,4 and 5 must have common ancestor that is not an ancestor of sites 1 and 3. Any genealogy that satisfies these conditions is a correct answer to the exercise.

(c) All segregating sites have either 1 or 4 mutations. This implies that the total length of line segments in genealogy from which 2 or 3 individuals in sample descent is small. This in turn implies that the genealogy is star-shaped which is expected in a (fast enough) growing population, where longer ago the population was smaller, and "merging" events were more likely in the past. In a declining population, merging of lineages (going backward in time) becomes less and less frequent and a star-shaped genealogy is unlikely.