Biostatistics 666
Problem Set 2

1. **Mutations in the G6PD gene, which maps to the X chromosome, are associated with resistance to infection by the malaria parasite. A stretch of 2000bp surrounding the gene was sequenced in a male susceptible to malaria and in another male who appeared resistant to malaria.**

   Assume the effective population size is N = 10,000 individuals, that the mutation rate is 10^{-8} per base-pair per generation and that there is no evidence for recombination in the region.

a) Given the gene is on the X chromosome and there are 10,000 individuals in the population, how many sequences are segregating in the population? State any assumptions about the number of males and females in the population.

Assuming that there are equal numbers of males and females in the population, I estimate the number of segregating sequences is

$$N_{sequences} = 2\,N_{females} + N_{males} = 15{,}000$$

b) What is the expected time to the most recent common ancestor (MRCA) of the two sequences? Please state any assumptions you made for this calculation.

As noted in part a, the population size (in number of sequences) is ~15,000 if we assume that about half of the population is composed of males (one X chromosome each) and half is composed of females (two X chromosomes each).

Therefore, the expected coalescent time is 15,000 generations.

c) What is the expected number of differences between the two sequences?

The expected number of differences between the two sequences is

$2 * T_{MRCA} * (L * \mu) = 0.6$

Here:

T_{MRCA} = 15,000 generations

L = 2,000 bp

$\mu = 10^{-8}$ per bp per generation

d) When the two sequences were compared, 5 differences were identified. What is the probability of observing 5 or more differences between the two sequences? Could you interpret this result as evidence of natural selection at the locus?

The probability of 5 or more differences is:

$$1 - \sum_{i=0}^{4} \left(\frac{\theta}{1+\theta}\right)^{i} \left(\frac{1}{1+\theta}\right)$$

Setting $\theta = 0.6$, the probability of observing 5 or more differences turns out to be ~0.007.

This means that observing 5 differences between a pair of sequences is quite unlikely using our base model and that we expect that either model parameters are off, or that some deviation from the basic coalescent assumptions (constant population size, neutral variants, etc.) has occurred. Natural selection is one of several possible alternatives.

e) If your model allowed for recombination within this 2000 bp sequence, how might your answer to a), b) and c) above to change?

Recombination would not change a) and b) and c) [although there now might be several distinct MRCA, one of for each non-recombining portion of the sequence].

Recombination would be expected to reduce the probability of observing 5 or more differences in d).

f) In general, how do you expect patterns of genetic variation and linkage disequilibrium to compare between the X chromosome and autosomes? Do you expect to see more (or fewer) variants per base pair in one setting – or do you expect both to be about the same? Do you expect to see the same degree of linkage disequilibrium in both settings – or do you expect one to show greater linkage disequilibrium?

We expect two differences between the X and autosomes. First, we expect that the number of sequences in the population will be smaller. This will reduce the number of expected variants and increase the expected amount of linkage disequilibrium.

Further, there are fewer opportunities for recombination on the X chromosome (because recombination can only occur in females). This should further increase the amount of linkage disequilibrium.

2. Consider the following bottlenecked population model:

Historical population size N_e = 10,000 sequences,

Followed by a bottleneck with N_e = 100 sequences
Lasting for 10 generations
And ending 2000 generations ago

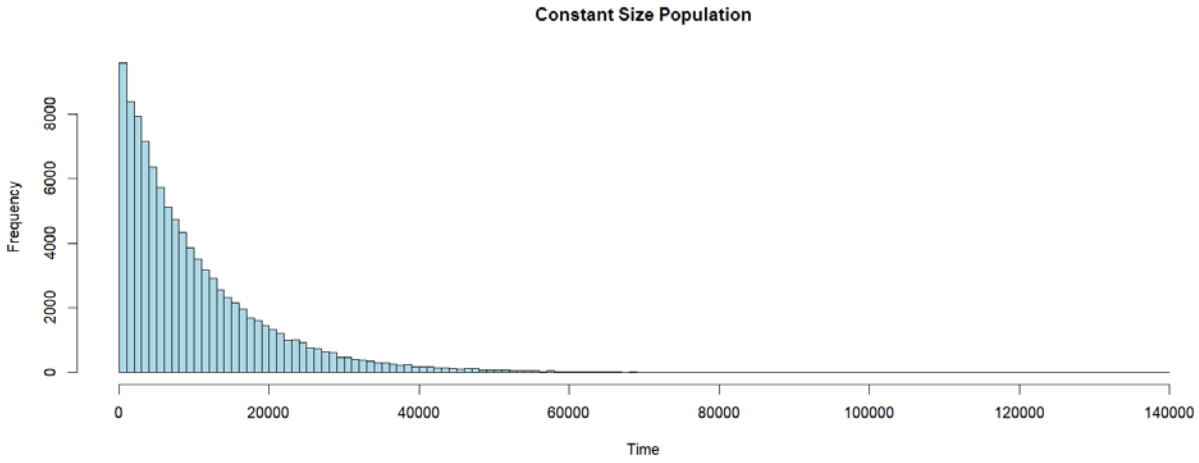
Population size after bottleneck N_e = 1,000,000 sequences

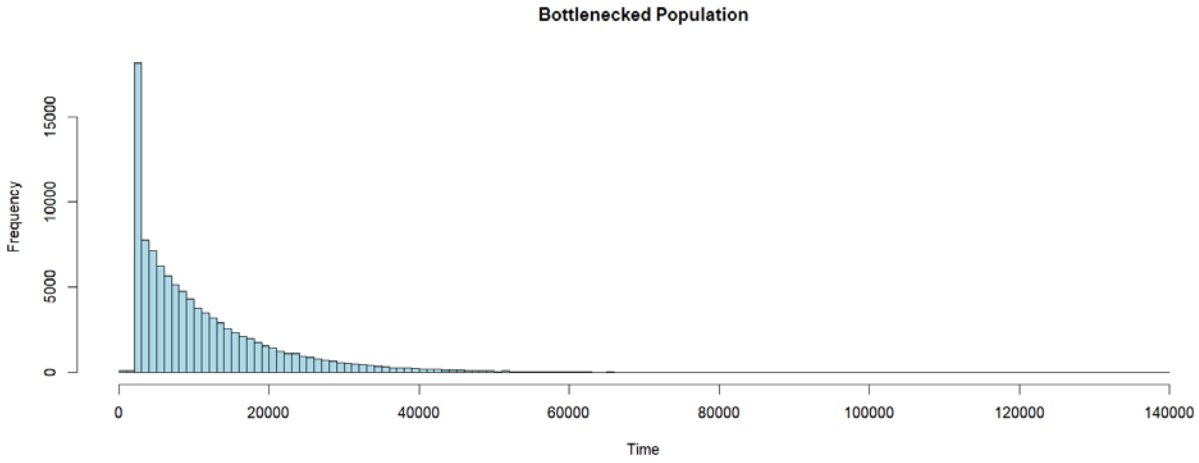
To sample a coalescent time for a pair of sequences, consider the following pseudocode:

```
SampleCoalescenceTime()
      {
      TimeToCoalescence = SampleFromExponential(Mean = 1,000,000)

      If (TimeToCoalescence < 2,000)
            Return TimeToCoalescence;

      TimeToCoalescence = 2,000 + SampleFromExponential(Mean = 100)

      If (TimeToCoalescence < 2,010)
            Return TimeToCoalescence;

      TimeToCoalescence = 2,010 + SampleFromExponential(Mean = 10,000)

      Return TimeToCoalescence;
      }
```

a) Using your favorite programming language, implement this code, sample coalescence times for 1,000 pairs of sequences and plot a histogram to summarize their distribution.

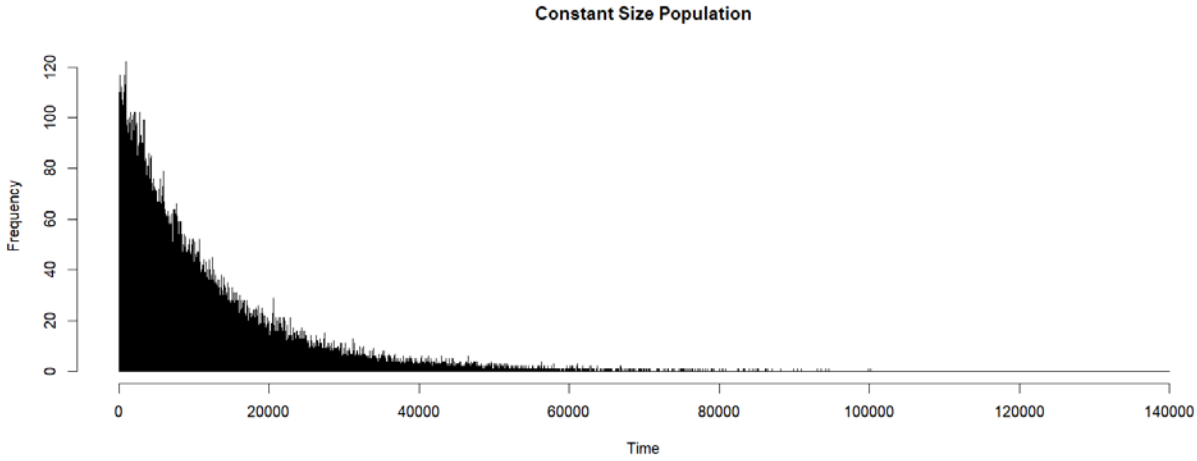
To make the patterns more obvious, I simulated 100,000 pairs of sequences, instead of 10,000… Here are the coalescence times, binned in bands of 1,000 generations, for a constant sized population and for our bottlenecked population:

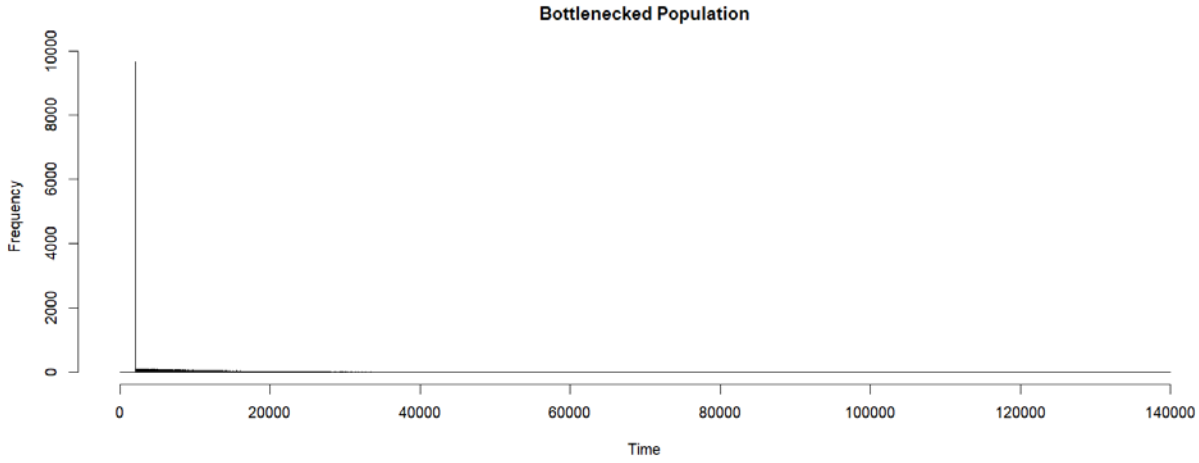




The most noticeable difference is in the first two bins (the most recent 2,000 generations) where there are now very few coalescent events. Further, in the bin that includes the interval from 2000 – 2010 generations, there a lot more coalescent events in the bottlenecked model (about 9.6% more).

The patterns are more obvious with smaller bins. Here are the histograms with 10 generation bins:





The spike between 2000 and 2010 generations is now very obvious!

b) How does this distribution of coalescence times compare to what you would expect with a constant population size N_e = 10,000 sequences?

The mean coalescent time has increased to just over 11,000 generations. Further, there are very few recent coalescent events (within the most recent 2000 generations) and a lot of coalescent events between 2000 and 2010 generations ago.

c) If you were to count pairwise differences between sequences in a population like this one, what would you expect? How would this result differ from that for a constant sized population with N_e = 10,000 sequences?

I would expect an increase in the overall number of differences between pairs of sequences and, perhaps, a decrease in the number of sequences with zero differences.