PM520–Homework 3

Chengliang Dong Wednesday 4th March, 2015

1 Introduction

In this homework, we are looking at Coalescent Theory. In genetics, coalescent theory is a retrospective model of population genetics. It attempts to trace all alleles of a gene shared by all members of a population to a single ancestral copy, known as the most recent common ancestor (MRCA; sometimes also termed the coancestor to emphasize the coalescent relationship). The inheritance relationships between alleles are typically represented as a gene genealogy, similar in form to a phylogenetic tree. This gene genealogy is also known as the coalescent. Understanding the statistical properties of the coalescent under different assumptions forms the basis of coalescent theory.
Starting with a sample with n individuals to create such a coalescent tree, then the time to next coalescent event follows exponential distribution $exp(n(n-1)/2)$. There is an interesting theoretical result, that is, the expected tree height is $2(1-1/n)$. Mutations may appear on lines of ancestry. In a discrete generation model, we define θ as the our mutation rate, which is two times μ, the original rate of mutation, for the simplicity of calculation. Then it follows that at k lines of ancestry, the probability of coalescence is:

$$P(coalesce) = \frac{k-1}{k-1+\theta} \tag{1}$$

the probability of mutation is:

$$P(mutate) = \frac{\theta}{k-1+\theta} \tag{2}$$

In this homework, we are assigned with four different task. The first task asks us to investigate the relationship between expected height of the tree vary as a function of the sample size. The second task asks us to investigate the distributions of number of descendants of the left ancestor. The third task asks us investigate the relationship between mean number of mutations and the mutation rate θ and sample size. The last task asks us to investigate the relationship between external branch lengths and sample size.

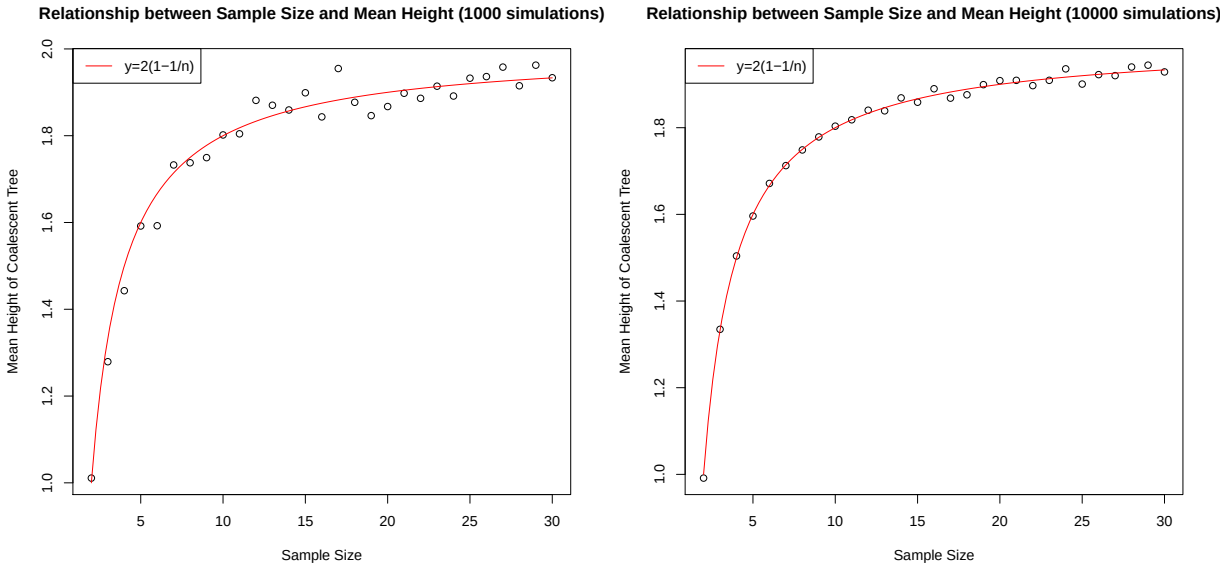
2 Method

To tackle the first task, I simulated 1000 trees and 10000 trees with sample size from 2 to 20 and plotted the mean heights of the trees for each sample size. For the second task, I also simulated 1000 trees and 10000 trees with sample size of 50 and plotted the distribution of number of descendants of the left ancestor. For the third task, I investigated the relationship between mutation rate (varies from 1 to 10) and mean number of mutations at fixed sample size of 5, with 1000 and 10000 simulations. I also investigated the relationship between sample size (varies from 2 to 20) and mean number of mutations at fixed number of mutation rate of 2. For the fourth task, I also simulated 1000 and 10000 trees with sample size varies from 2 to 10 and plotted the mean number of total length of external branches.

3 Result

3.1 Task One

Figure 1 demonstrated the relationship between Sample Size and Mean Height of the Coalescent Tree. We can see that from 1000 runs of simulations and 10000 runs of simulations, the Mean Heights of the Tree (y) is a function of Sample Size (n), which is $y = \frac{2(n-1)}{n}$. The more simulations are performed, the more obvious is this trend. Moreover, we can slso see that when Sample Size gets very large, the Mean Heights of Coalescent Tree approaches to 2.



(a) 1000 Simulations

(b) 10000 Simulations

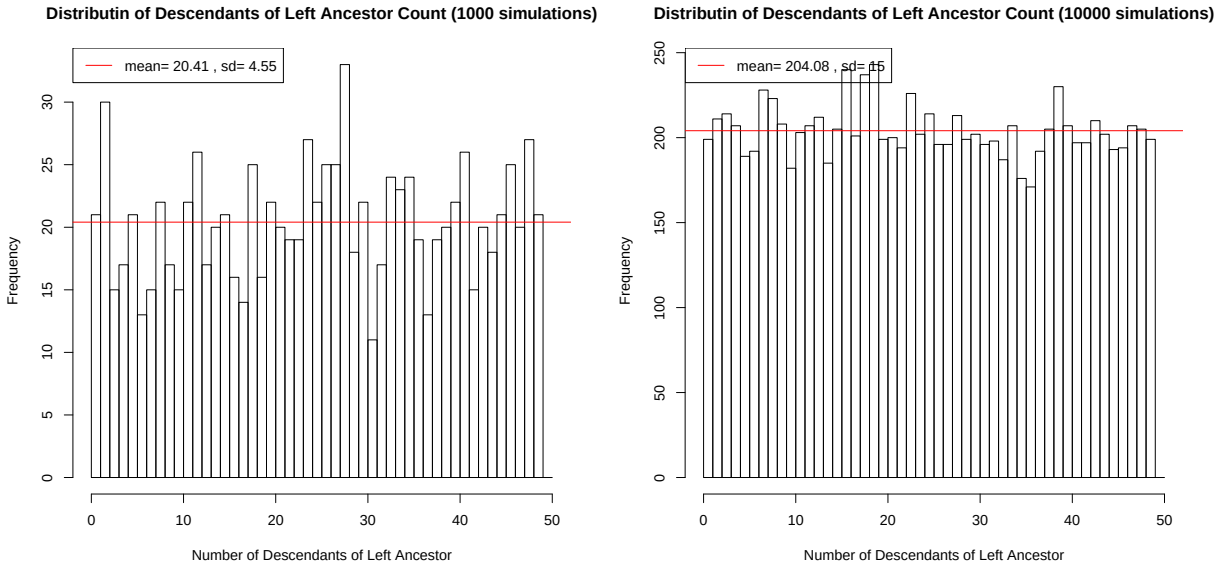
Figure 1: Relationship between Sample Size and Mean Height of the Coalescent Tree

3.2 Task Two

Figure 2 demonstrated the distribution of Number of Descendants of the Left Ancestor at the Sample Size of 50 with 1000 and 10000 runs of simulations. From this figure, we can see, that the Number of Descendants of Left Ancestor is subjected to uniform distribution and that the more simulations we run, the more obvious the trend is. The possible explaination behind it may be the symetry of Left Descendants and Right Descendants of an Ancestor.

3.3 Task Three

Figure 3 demonstrated the relationship between Mutation Rate, Sample Size and Mean Number of Mutations of the Coalescent Tree. We can see from top two figures that from 1000 runs of simulations and 10000 runs of simulations at Sample Size of 5, there is a linear relationship between Mutation Rate and the Mean Heights of the Tree. The more runs of the simulations, the more obvious is the trend, and the higher the Mutation Rate, the higher the Mean Number of Mutations. On the other hand, from bottom two figures with 1000 runs of simulations and 10000 runs of simulations at Mutation Rate of 2, there is



(a) 1000 Simulations

(b) 10000 Simulations

Figure 2: Distribution of Number of Descendants of the Left Ancestor at the Sample Size of 50

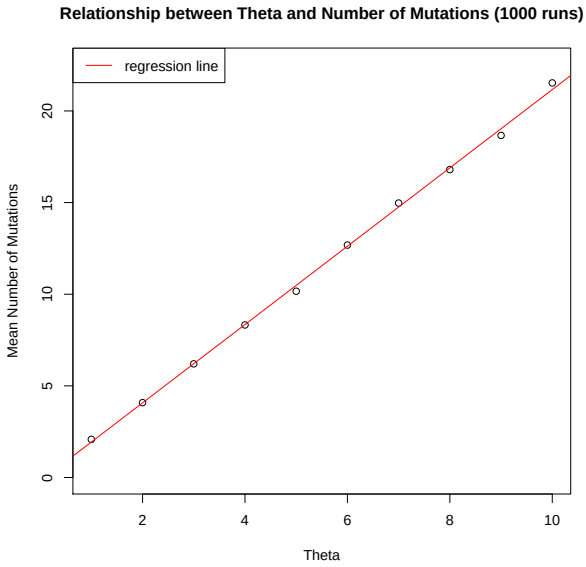
no obvious linear relationship between Sample Size and the Mean Heights of the Tree. Yet, the higher the Sample Size, the higher the Mean Number of Mutations.

3.4 Task Four

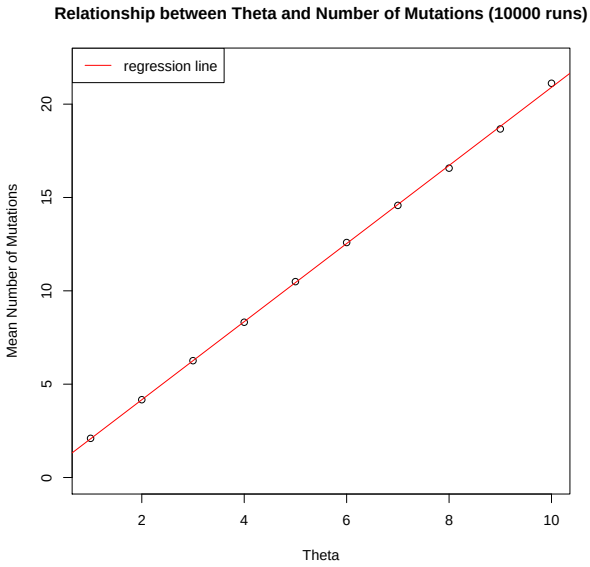
Figure 4 demonstrated the relationship between Sample Size and Total Length of External Branches of the Coalescent Tree. We can see that from 1000 runs of simulations and 10000 runs of simulations with Mutation Rate set to 0 (top two figures), the Mean Total Length of External Branch of the tree is the same regardless of Sample Size. The more simulations are performed, the more obvious is this trend (smaller variance). Moreover, we can see that from 1000 runs of simulations and 10000 runs of simulations with Mutation Rate set to 5 (bottom two figures), the Mean Total Length of External Branch of the tree, is not the same any more, but varies around a fixed value regardless of Sample Size.

4 Conclusion

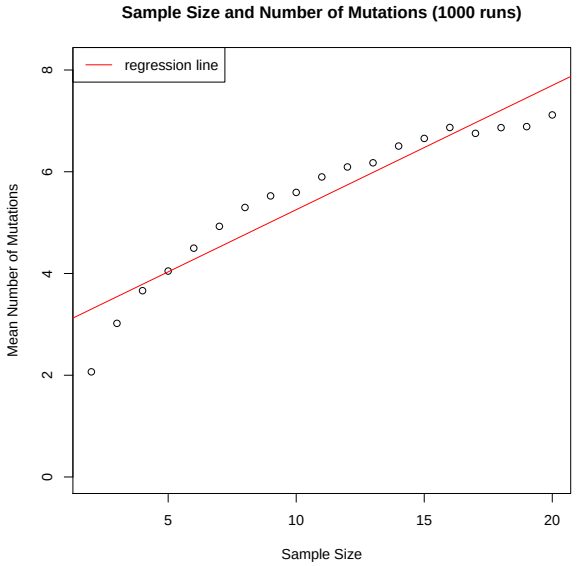
Here we investigated different variables in Coalescent Theory with four different task. The first task asks us to investigate the relationship between expected height of the tree vary as a function of the sample size. The second task asks us to investigate the distributions of number of descendants of the left ancestor. The third task asks us investigate the relationship between mean number of mutations and the mutation rate θ and sample size. The last task asks us to investigate the relationship between external branch lengths and sample size. From our analysis, we can see that first, the Mean Heights of the Tree (y) is a function of Sample Size (n), which is $y = \frac{2(n-1)}{n}$; second, the Number of Descendants of Left Ancestor is subjected to uniform distribution; third, there is a linear relationship between Mutation Rate and the Mean Heights of the Tree at fixed Sample Size and that there is no obvious linear relationship between Sample Size and



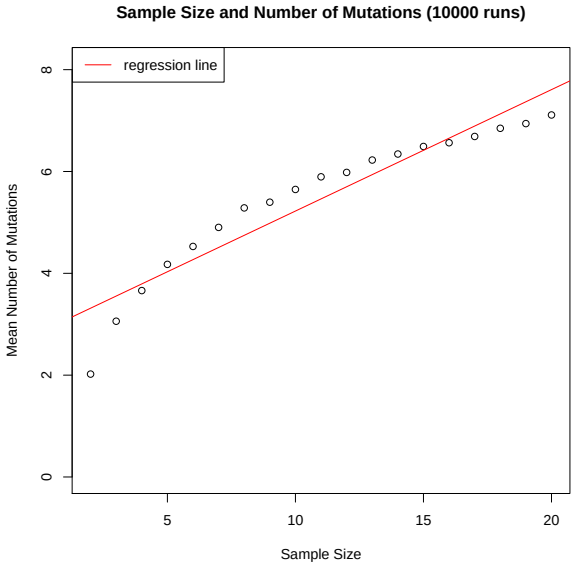
(a) Relationship between Mutation Rate and Mean Number of Mutations of the Coalescent Tree at Sample Size of 5 (1000 Simulations)



(b) Relationship between Mutation Rate and Mean Number of Mutations of the Coalescent Tree at Sample Size of 5 (10000 Simulations)



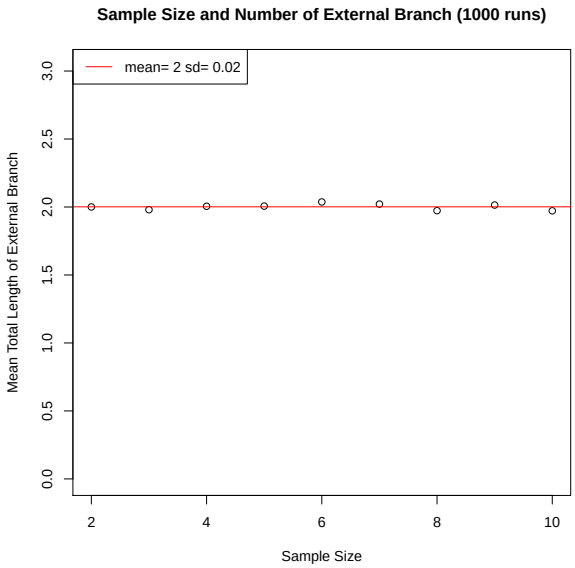
(c) Relationship between Sample Size and Mean Number of Mutations of the Coalescent Tree at Mutation Rate of 2 (1000 Simulations)



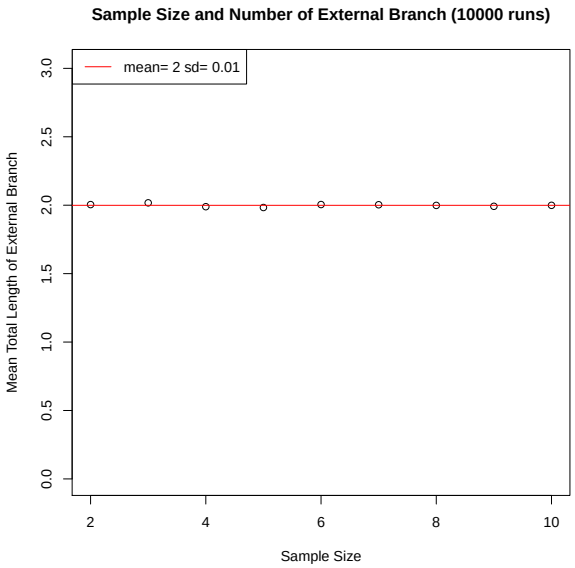
(d) Relationship between Sample Size and Mean Number of Mutations of the Coalescent Tree at Mutation Rate of 2 (10000 Simulations)

Figure 3: Relationship between Mutation Rate, Sample Size and Mean Number of Mutations of the Coalescent Tree

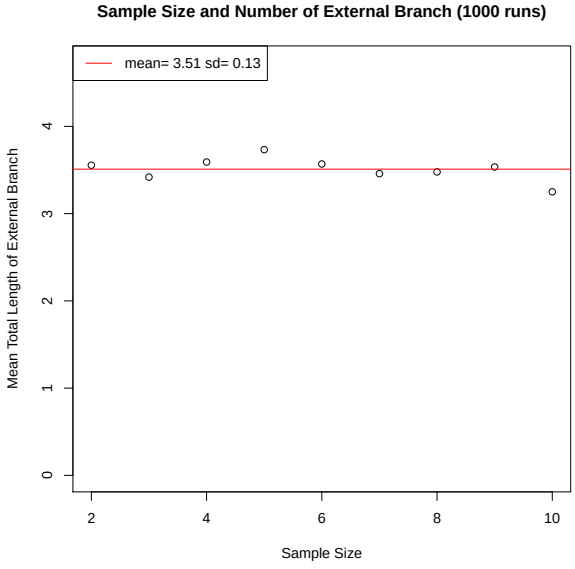
the Mean Heights of the Tree at fixed Mutation Rate; four, the Mean Total Length of External Branch of the tree is the same regardless of Sample Size at Mutation Rate of 0 and that the Mean Total Length of External Branch of the tree varies around a fixed value regardless of Sample Size at non-zero Mutation Rate.



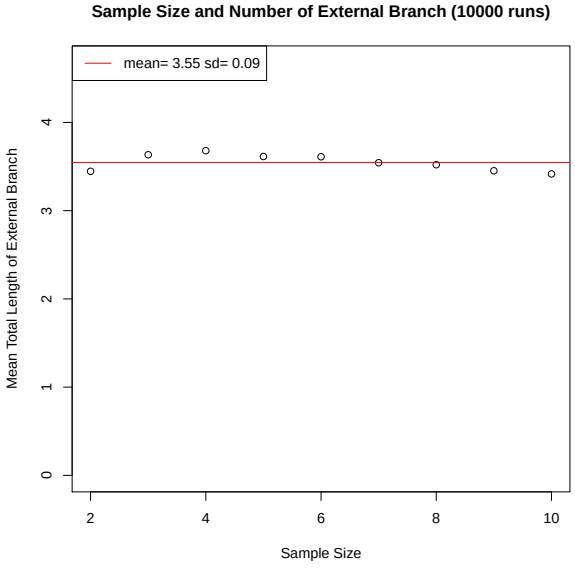
(a) Mutation Rate = 0 (1000 Simulations)



(b) Mutation Rate = 0 (10000 Simulations)



(c) Mutation Rate = 5 (1000 Simulations)



(d) Mutation Rate = 5 (10000 Simulations)

Figure 4: Relationship between Sample Size and Total Length of External Branches of the Coalescent Tree