Coalescent theory and infectious diseases

Rob van den Hengel

May 2, 2011

Studentnumber 0356999

supervisors:
Dr. M.C.J. Bootsma (MI)
Dr. W.M. van Ballegooijen (RIVM)
Ir. R.J.F. Ypma (RIVM)

Rijksinstituut voor Volksgezondheid
en Milieu
Ministerie van Volksgezondheid,
Welzijn en Sport

Universiteit Utrecht

Abstract:

The hereditary information of organisms is carried by DNA molecules. The DNA molecules can be extracted from a cell and its genetic code can be read. Within a population differences in the genetic code can occur between individuals. These differences are caused by mutations of the genes on the DNA molecules. The number of differences between to individuals is a measure for the relatedness of those two individuals. When the genetic information of a group of individuals is sampled, we can construct a kind of family tree, a so-called phylogenetic tree, using the differences between the individuals.

The coalescent theory makes a link between the phylogenetic tree and the population dynamics. It describes a model for the reproduction of the individuals to explain the phylogenetic tree. With this theory we can use the phylogenetic tree to estimate the size of the total population. This is the main reason why we want to try to apply coalescent theory to describe the spread of an infectious disease through a population.

Some problems, with the assumptions made in the classical coalescent model, arise when we use it for pathogens causing an infectious disease. In this model one individual is an infected host and the total population size, which we want to estimate, is the total number of infected hosts.

In the classical coalescent theory, among other things, there is assumed that the total population is very large and the number of sampled individuals is relative small. In the setting of infectious diseases, this assumptions can be problematic because it is possible that the total number of infected hosts is relative small, while a mayor part of the infected individuals is sampled.

In my thesis I describe two different extensions of the classical model to deal with relative large samples. From this extensions I derive two estimators for the population size and explore their performances.

Contents

This thesis is the result of my research at the department Epidemiology and Surveillance of the National Institute of Public Health and the Environment (RIVM), which I did to conlude my master study *Mathematical Sciences* at Utrecht University. I was supervised by dr. M.C.J. Bootsma from the Mathematical Institute of Utrecht University and dr. W.M. van Ballegooijen and ir. R.J.F. Ypma from the RIVM.

Preface

In case of an outbreak of an (new) infectious disease in a population, policy makers immediately want advice from specialists about the disease. When was the disease introduced in the population and how will the prevalence develop? These questions are important to determine which measures have to be taken. For example, when many individuals are already infected it is often not convenient to trace and isolate infected individuals from the rest of the population.

Research and policy making can only be done with information that is available at that moment. Often a disease has been spreading for some time before it is noticed. It is very helpful when one has good figures about the growth of the number of infections in the population until the moment of noticing. In the past years new technological developments have made it possible to get more information about the differences between the microorganisms which cause the disease. This information can be used to give estimates of the development of the disease in a population. The use of genetic diversity to infer infermation about the population dynamics is called coalescent theory. Here we use this theory for the modeling of the spread of infectious diseases in a population.

At first the aim was to look at the problems that arise when we try to use coalescent theory in a setting of infectious diseases. Gradually the focus was increasingly placed at the problem of applying coalescent theory in small, oversampled populations. Before we can formulate the goal more precisely, we introduce some terminology and explain the underlying biological processes of the spread of infectious diseases.

In Chapter 1 we shall explain some biological background and introduce the basic ideas and terminology. In Chapter 2 we describe the classic mathematical models and derive the classical estimator for the population dynamics. In Chapters 3 and 4 we give two different extensions of the classical model. In Chapter 5 we use these extensions to derive two different estimators for the number of infected hosts in a population. The performance of both estimators is explored numerically in Chapter 6. In Chapter 7 we will formulate our conclusions and give some recommendations for future work.

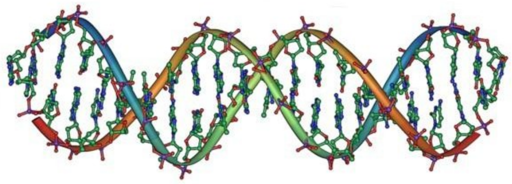


Figure 1.1: Schematic representation of a part of a DNA molecule.

1 Introduction

The hereditary information of most living organisms is carried by deoxyribonucleic acid (DNA) molecules. DNA usually consists of two complementary chains twisted around each other to form a double helix (Figure 1.1). Each chain is a sequence consisting of four nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T). Adenine pairs with thymine by means of two hydrogen bonds, while cytosine pairs with guanine by means of tree hydrogen bonds, so the strands are complementary [2].

When DNA is ready to multiply, its two strands separate, along each of the separate strands a new strand forms in the only possible way, and we wind up with two copies of the original. The precise details of the replication process are quite complicated but are not important for our study. For more information about this subject see for example Chapter 5 of B. Albers et al [1]. A gene is a collection of nucleotides that specify the amino acid sequence of a protein. When copying DNA, errors may occur. Such an error we call a mutation. The collection of all genes are called the genome. Most of the genes in our bodies reside on DNA in the nucleus of our cells and are organized into chromosomes. Lower organisms such as bacteria are haploid: they have one copy of their genetic material. Most higher organisms are diploid (i.e., have two copies).

When haploid individuals reproduce, there is one parent that passes a copy of its genetic material to its offspring. When diploid individuals reproduce, there are two parents. Each parent contributes one of each of its pairs of chromosomes. This is already a simplification of the real world since one parent's contribution may be a combination of its two chromosomes. This is because homologous pairs (e.g., the two copies of a chromosome of a diploid organism) undergo recombination, a reciprocal exchange of genetic material. Some bacteria can exchange a part of their genes (so-called plasmids) with other individuals and by this way change their genome (Chapter 9 of B. Alberts et al [1]). In this thesis we will ignore the possibility of all forms of recombination.

All species have their own unique genome. The code in DNA molecules is called the genetic information. Pathogens also have DNA molecules, although some viruses have an other kind of molecule, called RNA, but for our model this make no difference. It is possible to extract the DNA molecules from a cell and 'read' its genetic code. This process is called sampling. It is often not necessary to determine the whole genetic information of a pathogen. Only the characteristic part of the genetic information is sampled. This sampled part of the genome is a sequence of nucleotides.

By mutations in the sequence, new types of the same pathogen appears. Therefore, when we sample different infected hosts, we can find differences between the sequences we sample. We will use these differences to say something about the population dynamics. Throughout this thesis, we assume that only one variant of the pathogen can be sampled from an infected host and that all types of pathogen have the same fitness. This last assumption means that all types of a pathogen have the same ability to survive and to infect new hosts. This is also called neutrality in the mutations.

1.1 Using genetic information

If the same part of the genome is sampled for a group of individuals within the same population, one can compare the sequences and determine the nucleotide pairs in which these differ. If we assume that

all individuals in a population have one common ancestor, then all these differences have to be caused by mutations.
The number of nucleotide pairs at which two sequences are different, is an indication of how much time has gone by since the two sequences had a common ancestor. When two sequences have only one different pair of nucleotides, we assume that they are more related to each other, than two sequences with more differences. This can be depicted in a kind of 'family tree' of the pathogen, which is called a phylogenetic tree. For an example of a phylogenetic tree, see Figure 1.2. A lot of literature can be found on how to construct phylogenetic trees and of the problems which occur when constructing them. See for example [5],[9].

When we know something about the rate at which mutations occur, we can predict the number and distribution of mutations which will appear in the population. Vice versa, when we have sampled a number of sequences from a population, the number of types and their occurrence in the sample say something about how many infected individuals there are in total. This last observation is crucial for this thesis.

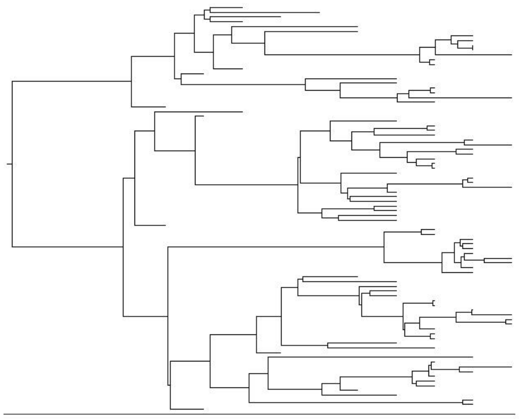


Figure 1.2: Example of a phylogenetic tree. Time is on the horizontal axis. The 'leaves' of the tree are the sampled sequences. The depth of a 'leave' corresponds to the time of sampling. The most recent samples sequences are to the outmost right of the tree. The 'root' of the tree is formed by the most recent common ancestor (MRCA)

In the years 1980-1982 J.F.C. Kingman ([6],[7],[8]) developed a theoretical background for estimations of the population size, given genetic information of sampled individuals in a population. In his model he makes many assumptions, often for mathematical reasons.

We draw our attention to the most important ones:

- The total population has a very large size N;
- The sample size n is small compared to N;
- Random mixing of all individuals in the population;
- Every mutation is neutral, i.e. each individual is equally able to produce offspring.
- Sampling from the population is done at the same moment.

Kingman's model, the 'n-coalescent', makes a link between the phylogenetic tree and the population dynamics which caused it. In literature the model has often been referred to as the classical coalescent model or just the coalescent model.

1.2 Infectious diseases

Microorganisms which cause diseases are called pathogens. Strictly speaking a virus is not a microorganism, but here we make no distinction between them. New technological developments have made it easier, cheaper and faster to extract genetic data from organisms and also from pathogens. Therefore it is nowadays feasible to sample a significant part of the genetic information of pathogens of infected hosts in a population, especially when the number of infected hosts is small. In this case, however, the first two assumptions from the list at the end of Section 1.1 are not met.

In Chapter 3 and 4 we will discuss two different extensions of the classical n-coalescent model mainly based on the articles of K. Strimmer and O.G. Pybus [15] for Chapter 3 and on that of Y.X. Fu [3] for Chapter 4. Both extensions give us methods to deal with the problems that arise when we cannot assume a large population size N and a relative small sample size n.

The reproduction rate and mutation rate of pathogens is often much larger than for higher organisms. Therefore, also the assumption that the sample is taken from the population at the same time is difficult to achieve. Often one has to take into account that sampling is done at different moments in time. In 1999 the model was extended by Rodrigo and Felsenstein [12] to genetic sequences that have been sampled at significantly different points in time (e.g., of rapidly evolving pathogen sequences or ancient DNA sequences). To simplify the model we will assume that sampling is done at one moment in time.

We assume that the phylogenetic tree is given and that this tree is correct, i.e. the tree depicts the real ancestry of the sampled sequences. In Chapter 5 we use the extensions of the classical model from Chapters 3 and 4 to derive two different estimators for the number of infected hosts in a population. Both estimators only use the features of the given phylogenetic tree and can handle relative large samples. These estimators are estimations for the development of the number of infected hosts in a population back in time. In chapter 6 we will look at the performance of both estimators.

When we have good estimations of the number of infected individuals over time, one can determine outbreak characteristics like R_0, which is the number of secondary infections caused by a single infectious individual in a completely susceptible population. Together with other outbreak characteristics, R_0 can be used to predict the speed at which the disease will spread through the population. In Figure 1.3 we give a schematic overview of the whole process, illustrating the focus of this thesis: The goal of this thesis is to derive an estimator for the number of infected hosts in time, that can deal with oversampled populations, given a phylogenetic tree of the pathogen.

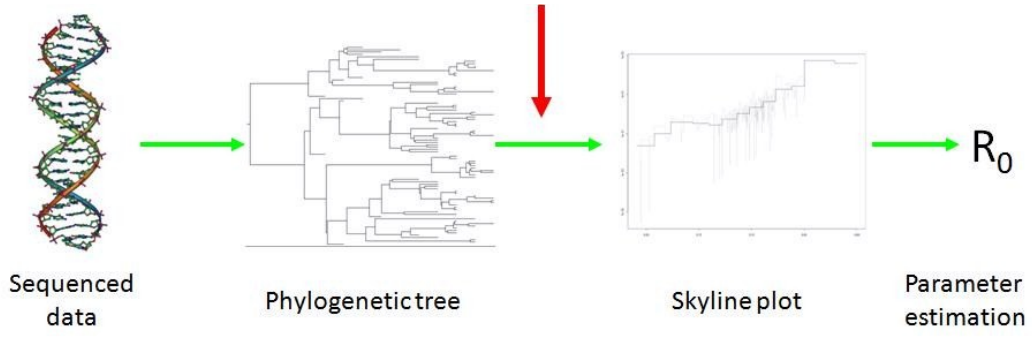


Figure 1.3: Deriving characteristics from genetics. First genetic data of the pathogen is sampled from infected hosts. From differences in these genetic sequences a phylogenetic tree is constructed. Using the phylogenetic tree one can estimate the development of the number of infected hosts back in time, depicted in a skyline plot. When the development of the infected population is known, one can estimate outbreak characteristics like R_0. In this thesis we assume that the phylogenetic tree is given and correct and from it we aim to give good estimations for the number of infected individuals.

2 The derivation of the classical n-coalescent

We first want to model the spread over time of different types of pathogens through a population of hosts. This is called a reproduction model. When we speak of the population we mean the population of infected hosts. This can be confusing since the population is not equal to the population of hosts. When we refer to the host population we will mention it explicitly. We assume that from an infected host only one type of pathogen can be sequenced.

Literature provides us with two often used models: The Wright-Fisher reproduction model and the Moran reproduction model [17]. Both models have their restrictions. In this thesis we will use the WF model, because this model can be easier applied in our theory.

2.1 Reproduction model

In this thesis we use the Wright-Fisher reproduction model (as underlying model) to describe the ancestry of individuals over time. The Wright-Fisher process describes how the proportions of different types of individuals in a population change over time. At first we assume that the population size N is constant over time. It is possible to relax this assumption and to allow for fluctuations of the total population. This results in time-varying resampling rates in the WF model but it does not change the main qualitative features, as we will show.

In the classical neutral WF model (1931) the population size N is constant over time and the population exists of two types of individuals $E = \{1, 2\}$ [19]. Time steps are discrete and generations are disjoint. This means that at every time step the whole population dies and is replaced by their offspring. We assume that every individual is equally able to produce offspring, so that the number of offspring does not depend on the type of the parent nor on the ancestry of the parent.

To explain the WF model in a biological way we imagine that before dying there is a short period in which each of the N individuals of that generation produces a very large number of offspring. Although there is an infinite number of potential individuals to form the next generation, the population size is tightly controlled so that only N of these individuals can survive to the next reproduction period. Because the offspring pool is so large, it can be assumed that it is not depleted by this sampling. So at each trial the probability that the chosen individual is the child of a specific parent remains equal [**?**].

In mathematical notation we describes the WF model as follows: If X_t denotes the number of type 1 individuals at time t, then X_n is a Markov chain with state space $\{0, \ldots, N\}$ and transition probabilities:

$$\mathbb{P}(X_{t+1} = j | X_t = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}, j = 0, \ldots, N.$$

Note that at generation $t+1$ given X_t, X_{t+1} is binomially distributed with probability $\frac{X_t}{N}$. Looking backward in time from the viewpoint of generation $t+1$, this can be interpreted as saying that each of the N individuals 'chooses' his parent at random from the population at generation t, see Figure 2.1.

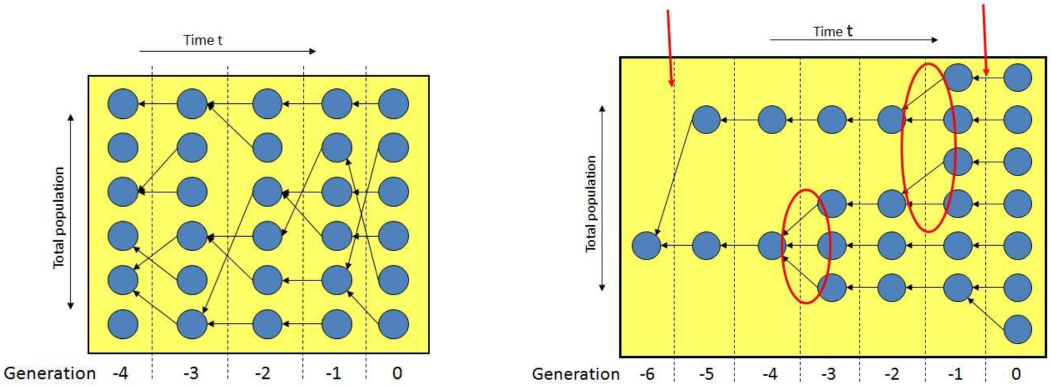
Now suppose that there are K different types of individuals $E_K = \{e_1, \ldots, e_K\}$, then the WF model is given by the Markov chain $(X_t)_{t \in \mathbb{Z}}$, with state space

$$(\beta_1, \ldots, \beta_K) \in \mathbb{N}^K \text{ and } \sum_{i=1}^{K} \beta_i = N\}$$

and transition probabilities:

$$\mathbb{P}(X_{t+1} = (\beta_1, \ldots, \beta_K) | X_t = (\alpha_1, \ldots, \alpha_K)) = \frac{N!}{\beta_1! \ldots \beta_K!} \left(\frac{\alpha_1}{N}\right)^{\beta_1} \ldots \left(\frac{\alpha_K}{N}\right)^{\beta_K}.$$

This is a multinomially distributed random variable, rather than a binomially distributed random variable. We are often interested in the number of offspring of each individual, so the number of different types is



(a) WF model. Individuals randomly 'choose' their ancestor.

(b) Ancestry of the individuals in the WF model.

Figure 2.1: Wright-Fisher reproduction model. In both pictures we see a schematic overview of the discrete time WF-model. In the left picture, each column is one generation and consists of six individuals. They all randomly choose an ancestor from the previous generation. This is depicted by the arrows. In the right picture we only kept track of individuals in the different generations which have descendants in generation 0. We ordered the arrows in such a way that they do not cross each other. The total population size is constant over time. Between the generations in which only one pair of individuals has a common ancester are indicated by a red arrow, this events are called two-coalescent events. The events at which more then two individuals have a common ancester are encircled, these are called multiple coalescent events.

equal to the population size, hence $K = N$. Because we assume that every individual is equally able to produce offspring, we get a multinomially distributed reproduction process with parameters $(N; \frac{1}{N}, \dots, \frac{1}{N})$. Using this we can compute the probability of the event A that all individuals in generation t produce exactly one offspring in generation $t+1$, or, looking backward in time we compute the probability that all individuals in generation $t+1$ choose a different ancestor in generation t.

$$\mathbb{P}(A) = \mathbb{P}(X_{t+1} = (1, \dots, 1) | X_t = (1, \dots, 1)) = \frac{N!}{1! \dots 1!} \left(\frac{1}{N}\right)^N = \frac{N!}{N^N}$$

The probability that two randomly chosen individuals from generation $t+1$ choose the same ancestor in generation t is $\frac{1}{N}$. There are $\binom{n}{2}$ different ways to form a pair of individuals. Hence, if the sample size n is relative small, the probability that there is at least one pair of individuals in generation $t+1$ with a common ancestor in generation t is $\binom{n}{2}\frac{1}{N}$. This is only true for small sample sizes. In Section 4.2 we will derive a boundary for the sample size. But for the moment we assume that the population size is very large and the sample size is small so that there is no problem.

We now introduce some notation. We denote by $\nu_{i,t-1}$ the number of offspring in generation t of individual i, which himself lives in generation $t-1$. Notice that the time index t in this notation, increases in the 'real' direction. The way of choosing ancestors by the individuals as described above means that the stochastic vector $(\nu_{1,t-1}, \nu_{2,t-1}, \dots, \nu_{N,t-1})$ of the numbers of offspring is multinomial distributed, with parameters N and, because of symmetry, $p_i = \frac{1}{N}$ for all $1 \leq i \leq N$.

2.2 Phylogenetic trees

We assume that we have sampled the sequences of the pathogen of n infected hosts. The sequences are taken at the same time. When we assume that the pathogen has a high mutation rate compared to its infectivity rate and that each mutation creates a new type of pathogen, then each host will have its own

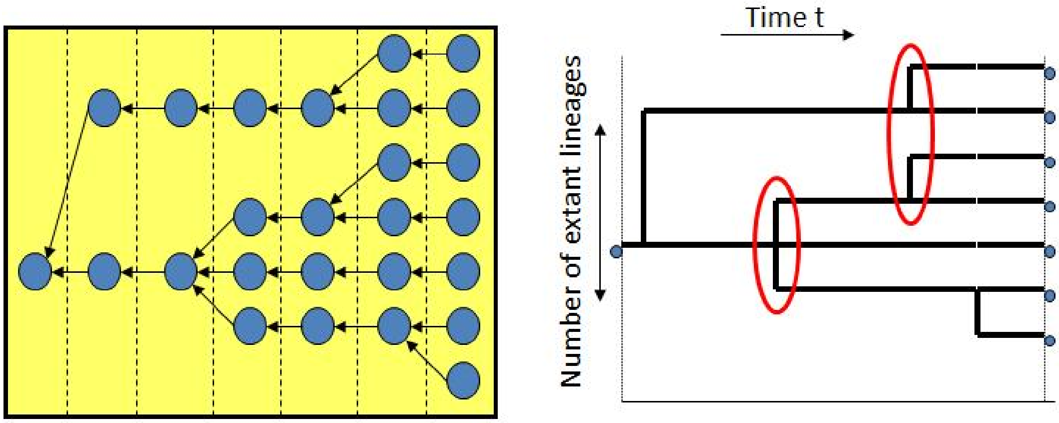


Figure 2.2: Link between WF model and phylogenetic trees. On the left side is the same schedule of a WF process as in figure 2.1. On the right is a phylogenetic tree, based on the same WF process, but with continuous time between the events. The multiple coalescent evens are encircled.

unique sequence. The number of differences between two sequences is an indication how related they are. For example when two sequences have only two pair wise differences, they are probably more related to each other then two sequences with three of more differences. The 'family lines', or lineages, of the first two sequences are likely to have been split more recently, then those of the latter, so the first sequences are likely to share a more recent commen ancester, then the latter. From this genetic data we can construct a kind of 'family tree' or 'genealogy' of the different types of the pathogen. Such a family tree is called a phylogenetic tree. We assume that the underlying reproduction model of the phylogenetic tree is a continuous time WF model. In a continious time WF model time is continious instead of discrete. The time between successive coalescent events is exponentially distributed. The convertion of the standard WF model to the continious time WF model is described in Section 2.3
We want to follow the ancestry of these n individuals back in time. When tracing back the ancestry of an individual, we speak of such a 'family line' as a lineage. When two or more individuals choose the same ancestor, we say that those lineages are merging together. This event is called a coalescent event. When in a generation exactly two lineages merge we say that a two-coalescent event has occurred. All more complex events are called multiple coalescent events, see Figure 2.1. For example, when three lineages coalesce to one one lineage, or two pairs of lineages coalesce to two lineages at the same moment, we speak of a multiple coalescent event. See for example Figure 2.2.
When we know something about the mutation rate μ, we can estimate the expected time to the most recent common ancestor (MRCA) of a number of sequences. If we don't know the mutation rate, we can only give an expression for the product of the mutation rate and time (μt). The time (t or μt) between two successive coalescent events is called the coalescent time or coalescent interval. We assume that the lengths coalescent intervals are exponentially distributed according to a continious time WF model.

Programs which construct phylogenetic trees will use the differences in the sequences sampled from a population to construct the most likely phylogenetic tree, or a set of most likely phylogenetic trees. In this thesis we consider the phylogenetic trees as given and assume that they are a correct representation of the underlying transmission process. We will use these trees to give an estimation of the population size.

2.3 Classical n-coalescent

In the model which is named after Kingman [6], we assume that we observe a randomly drawn sample of n individuals out of a larger constant population consisting of N individuals. This population is haploid and

reproduces according to the WF reproduction model.
Kingman assumes a very large population size N and a small sample size n compared to N. These assumptions make it possible to neglect the probability of a multiple coalescent event. How small the sample size should be is derived in Chapter 4. The probability of a multiple coalescent event is of order $O(N^{-2})$ and, hence, by the above assumptions these events can be neglected. This is shown explicitly in section 4. The probability that one pair of individuals chooses the same parent is $\frac{1}{N}$. So the probability of a coalescent event when observing n sequences is $P_{Coal}(n) = \binom{n}{2}\frac{1}{N}$ and the probability that no coalescence event is observed is $P_{NoCoal}(n) = 1 - P_{Coal}(n) = 1 - \frac{n(n-1)}{2N}$. Denote by τ_n the number of generations until a coalescent event occurs among n sequences. Then τ_n is known as the n-coalescent time. The distribution of τ_n is given by

$$\mathbb{P}(\tau_n = i+1) = P_{Coal}(n)P_{NoCoal}(n)^i = \left(\binom{n}{2}\frac{1}{N}\right)\left(1 - \binom{n}{2}\frac{1}{N}\right)^i, \tag{1}$$

which is a geometric distribution with expectation

$$\mathbb{E}[\tau_n] = P_{Coal}(n)^{-1} = \left(\frac{n(n-1)}{2N}\right)^{-1} = \frac{2}{n(n-1)}N.$$

Note that τ_n is decreasing with n, i.e., when the number of sequences has decreased due to the coalescent events, the expected time till the next observed coalescence event increases. Because the population size N is very large and the sample size n is small, we expect that the time until a coalescent event occurs is of order $O(N)$. In other words, during most generations no coalescent event will occur.
We want to use this model to explain the genealogy, which forms a given phylogenetic tree. Because pathogens are spreading on a continuous time scale and so the phylogenetic tree belonging to it, is also constructed in continuous time. Therefore we have to give a continuous time approximation for the waiting time. This is an exponential distribution with the same expectation. Thus we find the following density function for τ_n:

$$f(t) = \frac{n(n-1)}{2N}e^{-\frac{n(n-1)}{2N}t}.$$

We need to specify how the state of the process, i.e., the number of ancestral sequences, changes when a coalescent event occurs. Suppose a random sample of sequences is drawn from a certain generation. Suppose there are n sequences left at a generation τ, where τ is increasing backwards in time. The state of generation τ is represented by a set R_τ of all equivalent relations, which are pairs (i,j) where $i,j \in \{1,\ldots,n\}$. For a pair (i,j) holds: $(i,j) \in R_\tau$ if and only if the sequences of individuals i and j of the original sample, share a common ancestor at generation τ. At time 0, that is the time of observation of the n individuals in the sample, we are in the initial state $\Delta := \{(i,i) : 1 \le i \le n\}$.
Define $p_{\xi\eta} := \mathbb{P}(R_\tau = \eta | R_{\tau-1} = \xi)$ as the transition possibility to move from state ξ to state η, where both states are in the state space of all equivalence relations on $\{1,\ldots,n\}$.
Because under the assumptions all probabilities of multiple-coalescent events are assumed to be negligible, the transition probabilities $p_{\xi\eta}$ are almost always zero. The only type of event which changes the state that can occur is a two-coalescent event. Assume that the number of equivalence classes of state ξ is n (notation: $|\xi| = n$). That means that the remaining observable sequences in this state is n. Then

$$p_{\xi\eta} = \begin{cases} 1 - \frac{n(n-1)}{2N} & \text{if } \xi = \eta \\ \frac{1}{N} & \text{if } \xi \prec \eta \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\xi \prec \eta$ means that η is derived from ξ by merging two of its equivalence classes into one [8]. Note that the number of equivalence classes in state ξ is n, so there are $\binom{n}{2} = \frac{n(n-1)}{2}$ different ways to combine two equivalence classes.
It may be confusing that the character n is used in two different ways. When we speak of the sample size n denotes the number of sampled sequences. While, when we speak of a coalescent event n denotes the number of extant lineages before the event occurs.

2.4 Skyline plot

When estimating the population size, we assume that during a coalescent interval the population size is constant. Only at an event the estimated population size can change. The estimation depends on the length of the coalescent interval and the number of observed sequences during the interval.
From a given phylogenetic tree we can extract all the coalescent times τ_i, the number of extant lineages before and after the ith coalescent event, denoted by n_i and k_i respectively. Note that we look backward in time, that $k_i = n_{i+1}$ and that for the last coalescent interval, M, in the tree $k_M = 1$. In the classical n-coalescent model we have that $k_i = n_i - 1$, so each coalescent event in the classical n-coalescent is defined by (n_i, τ_i).
Given the assumptions made for the classical n-coalescent, the distribution of the coalescent time of τ_i is exponentially distributed with parameter $\frac{n_i(n_i-1)}{2N}$, hence we expect that the coalescent time has length

$$\mathbb{E}_K\left[\tau | n_i, N\right] = \left(\frac{n_i(n_i-1)}{2N}\right)^{-1} = \frac{2}{n_i(n_i-1)}N, \tag{3}$$

in which $\mathbb{E}_K$ means the expectation under the probability space, defined by the classical n-coalescent of Kingman. In our situation we don't know the population size N, but with the observation τ_i and the assumption that $\tau_i = \mathbb{E}_K\left[\tau | n_i, N\right]$ we can derive an estimator $\hat{N}_i^K$ for N given (n_i, τ_i):

$$\hat{N}_i^K = \tau_i \frac{n_i(n_i-1)}{2}. \tag{4}$$

In this way we get for each coalescent interval (n_i, τ_i), $1 \leq i \leq M$, an estimate $\hat{N}_i^K$. We can plot this list of estimates in a graph, with on the horizontal axis time and on the vertical axis the estimated population size $\hat{N}^K$, see for example Figure 2.3. Such a graph is called a skyline plot.

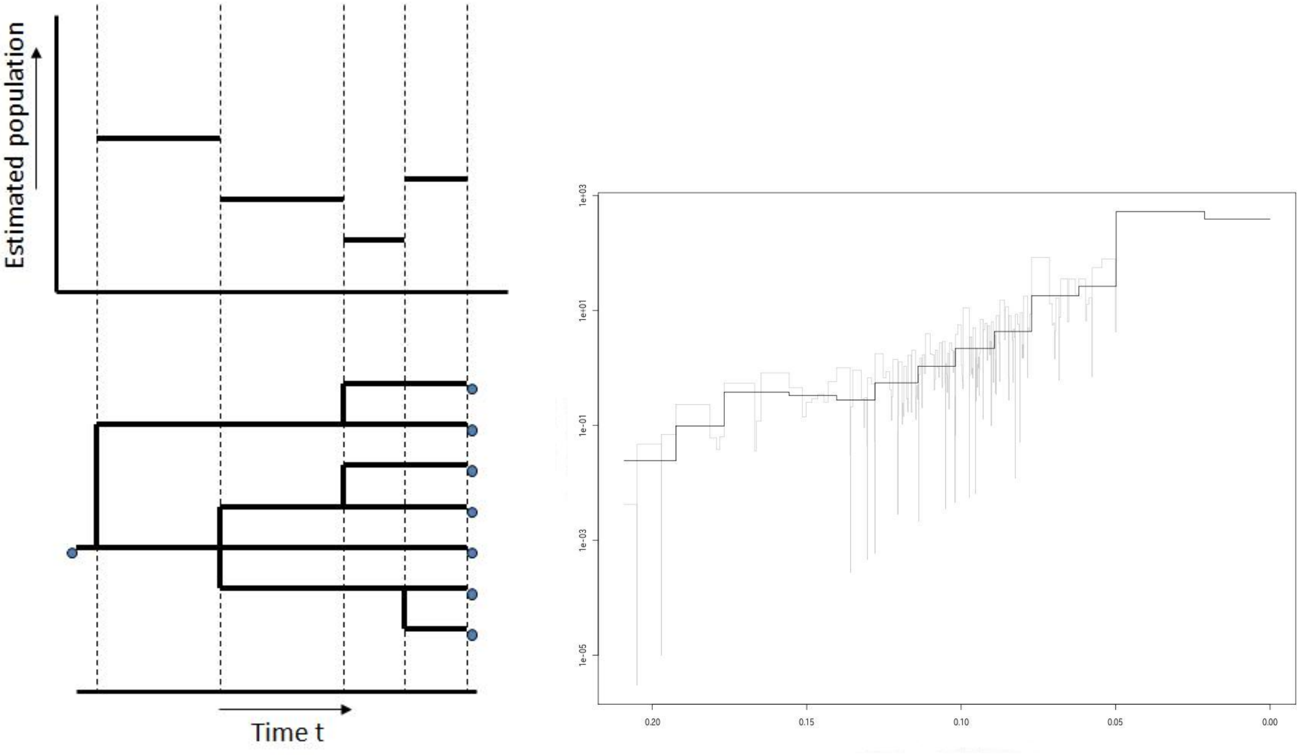


Figure 2.3: Skylineplot. The left picture shows how a skyline plot is linked to the phylogenetic tree from which it is derived. The right figure is an example of a skyline plot. On the horizontal axis is the product of the time to present in units of mutations and on the vertical axis is the product of the estimated population size and the mutation rate.

3 Combining 2-coalescent intervals

To improve the standard estimator of the effective population size derived from the classical n-coalescent, K. Strimmer and O.G. Pybus noticed in their article [15], that the skyline plot has a problem with short intervals. In a short interval the classical n-coalescent estimator underestimates the real effective population size. These short intervals can appear by chance, thus are more likely to be seen in the leaves of the tree, where the number of sequences is still relatively large. It is even possible that in reality a 3-coalescent event occurred, but, because the algorithms of Kingman neglect this possibility, the event will be regarded as two 2-coalescent events which occur within a very short time interval.

Therefore Strimmer et al. suggest that short intervals should be combined with a neighboring interval, until there are no intervals left with a total length shorter than a certain ϵ. The goal of Strimmer et al. was to get rid of the 'dips' in the skyline plot, but it provides a method which can handle multiple coalescent events as well.

Remember that according to the classical n-coalescent the time between two coalescent events is exponentially distributed with parameter $\lambda^K_{N,n} = \frac{1}{N}\binom{n}{2}$, and so its expectation is $\frac{1}{\lambda^K_{N,n}} = N\binom{n}{2}^{-1}$.

The sum of two exponentially distributed random variables follows a hypo-exponential distribution which has as expectation the sum of the expectations of the underlying random variables [10]. Suppose that we combine $n-k$ different coalescent intervals and, thus, there are k sequences remaining thereafter. The expected time to observe $n-k$ coalescent events is

$$\sum_{i=k+1}^{n} {\lambda^K_{N,i}}^{-1} = \sum_{i=k+1}^{n} N\binom{i}{2}^{-1} = N\frac{2(n-k)}{nk}.$$

This is given by the following lemma:

Lemma 1. *For $1 \leq k < n \leq N$, all chosen from $\mathbb{N}$,*

$$\sum_{i=k+1}^{n} \binom{i}{2}^{-1} = \frac{2(n-k)}{nk}. \tag{5}$$

Proof. We will proof equation (5) with reversed complete induction.

Suppose $k = n-1$, then

$$\sum_{i=(n-1)+1}^{n} \binom{i}{2}^{-1} = \binom{n}{2}^{-1} = \frac{2}{n(n-1)} = \frac{2(n-(n-1))}{n(n-1)}.$$

Hence equation (5) is true for $k = n-1$. Now suppose equation (5) is true for k. When we want to prove for $k-1$ that

$$\sum_{i=(k-1)+1}^{n} \binom{i}{2}^{-1} \overset{?}{=} \frac{2(n-(k-1))}{n(k-1)}.$$

We have

$$\begin{aligned}
\sum_{i=(k-1)+1}^{n} \binom{i}{2}^{-1} &= \binom{k}{2}^{-1} + \sum_{i=k+1}^{n} \binom{i}{2}^{-1} \\
&= \frac{2}{k(k-1)} + \frac{2(n-k)}{nk} \\
&= 2\left(\frac{n}{nk(k-1)} + \frac{(n-k)(k-1)}{nk(k-1)}\right) \\
&= 2\left(\frac{n+nk-n-k^2+k}{nk(k-1)}\right) \\
&= \frac{2k(n-k+1)}{nk(k-1)} \\
&= \frac{2(n-(k-1)}{n(k-1)}.
\end{aligned}$$

□

Suppose we observe a combined coalescent interval of length τ, then we expect that $\tau = N\frac{2(n-k)}{nk}$. Hence we have an estimator for the effective population size according to Strimmer and Pybus:

$$\hat{N}^G = \tau\frac{nk}{2(n-k)} \tag{6}$$

3.1 Determining the minimal coalescent interval

We now want to determine a good choice for the minimal interval length ϵ. The goal according to Strimmer and Pybus is to get rid of the undesirable 'dips' in the skyline plot, so visual inspection of the skylineplot may do. But Strimmer and Pybus prefer a statistical approach.

As a measure of the 'fitness' of different choices for ϵ, Strimmer and Pybus give a penalty to the number of parameters, i.e., the number of coalescent intervals, of the skylineplot using the second-order extension of Akaike's information criterion (AIC_c) [15]. The adjusted log likelihood function $\log L_{AIC_c}$ that we use is corrected by AIC_c:

$$\log L_{AIC_c} = \log L - K - \frac{K(K+1)}{S-K-1}, \tag{7}$$

in which $\log L$ is the log likelihoodfunction, $S = n-1$ is the number of two-coalescent events which are needed to go from n observed sequences to one, and K is the number of composite intervals in the skyline plot, which in the literature is the number of inferred parameters. Hence K depends on the choice of ϵ.

According to Kingman the only information needed to give an estimation for the effective population size is the observed time τ_n between two coalescent events and n, the number of lineages. This time should be exponentially distributed with parameter $\binom{n}{2}\frac{1}{N}$. So the likelihood $L(N)$ of an estimated effective population size N given an observed coalescent time τ_n and n observed sequences is

$$L(N) = \binom{n}{2}\frac{1}{N}e^{-\binom{n}{2}\frac{1}{N}\tau_n}$$

and hence the log likelihood is

$$\log L(N) = \log\left(\binom{n}{2}\frac{1}{N}\right) - \binom{n}{2}\frac{1}{N}\tau_n$$

Suppose that we have a phylogenetic tree which start with n sequences, in which we observe coalescent times $\tau_n, \tau_{n-1}, \ldots, \tau_2$ with corresponding estimations for the effective populations $\hat{N}_n, \hat{N}_{n-1}, \ldots, \hat{N}_2$, then the total log likelihood of the skyline plot is

$$\begin{aligned} \log L &= \sum_{i=2}^{n} \log L(\hat{N}_i) \\ &= \sum_{i=2}^{n} \log \binom{i}{2} \frac{1}{\hat{N}_i} - \binom{i}{2} \frac{\tau_i}{\hat{N}_i} \end{aligned} \tag{8}$$

Note that when two coalescent, say j and $j+1$, $2 \leq j \leq n-1$, are combined then we find one estimation $\hat{N}$ for the whole interval. In Formula (8) we then have to put $\hat{N}_j = \hat{N}_{j+1} = \hat{N}$, so for different choices of ϵ, we still sum over $n-1$ intervals [15].
When a coalescent time $\tau_j < \epsilon$, then we merge this interval into the next interval, backward in time, with length τ_{j-1} into a new interval with length $\tau = \tau_j + \tau_{j-1}$. When the combined interval length τ is still smaller than ϵ, we add more intervals until $\tau \geq \epsilon$.
Given a phylogenetic tree, we can maximize formula (7) over $\epsilon > 0$. The argument which maximize formula (7) is our choice for the minimal interval length ϵ.

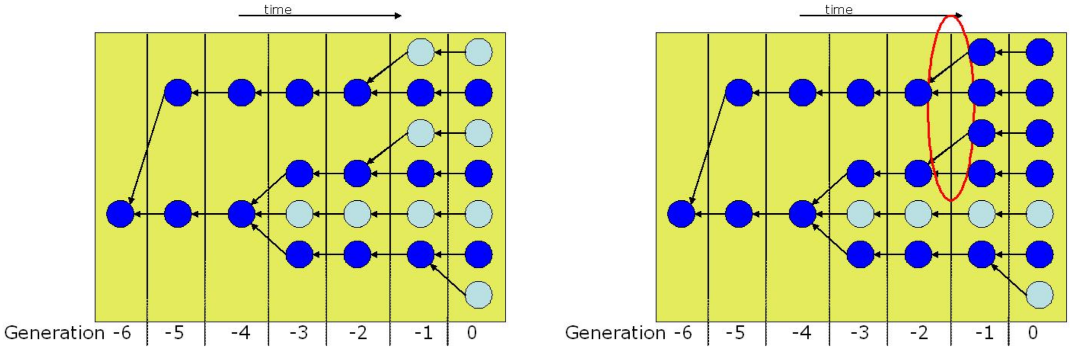


Figure 4.1: Partial observations in the WF model. In both pictures we see a WF reproduction model of which the dark dots are sampled sequences. The left figure illustrates that when a small sample is randomly drawn from the population it is very unlikely that more than two of the observed sequences will coalesce at the same time. While, as illustrated in the right figure, when the sample is large compared to the population size it is very realistic that more than two sequences will coalesce at the same time. The multiple coalescent event has been encircled.

4 Exact coalescent

In his articles, Kingman assumed that the population size N is very large compared to the sample size n, so that we can neglect the probability of multiple coalescent events per generation [6]. However, for infectious diseases this assumption is not always true. Therefore Fu extends the classical n-coalescent model by taking multiple coalescent events into account [3], deriving a theory called the exact coalescent. His theory does not need the assumption of a large N and a relative small n. We will compare his results to the results of the classical n-coalescent.

The problem with relative large samples is that we cannot ignore multiple coalescent events. When the number of observed sequences is small compared to the total population size, we could ignore the possibilities of these events, because there is only a very small probability that at a moment three or more individuals 'pick' the same ancestor. For larger samples this scenario becomes realistic. This is illustrated in Figure 4.1. The critical sample size, that is the maximum sample size for which the classical n-coalescent model of Kingman is still meaningfull, will be derived in section 4.2.

We assume that the population is constant over time. So consider a population of N sequences. From this population we take, in generation t, a sample of size n. Each individual in a generation t randomly selects an ancestor in generation $t-1$. Hence the probability that an individual in generation t is the child of a specific individual in generation $t-1$ is $\frac{1}{N}$. So the probability that a second individual chooses a different ancestor is $1-\frac{1}{N}$. Given that the first two individuals do not share the same ancestor, the probability that the third individual also selects a different ancestor is $1-\frac{2}{N}$. Applying this argument for the other $n-3$ individuals in the sample, we obtain the probability $P_{\text{No Coal}}$ that none of the n individuals in the sample shares a common ancestor in generation $t-1$. The expression 'No Coal' refers to the event that in one generation no coalescent event occurs. Hence

$$\begin{aligned} P_{\text{No Coal}} = \prod_{i=1}^{n-1}\left(1-\frac{i}{N}\right) &= \frac{1}{N^n}\prod_{i=1}^{n-1}(N-i) \\ &= \frac{N_{[n]}}{N^n}, \end{aligned} \tag{9}$$

where $N_{[n]} := N(N-1)\dots(N-n+1)$. Note that this probability depends on de sample size n at generation t and on the population size at $t-1$. So we could write N_{t-1} instead of N, which is important when we want to deal with fluctuating population sizes in time. The variable population size coalescent model are introduced by Griffiths and Tavare [4]

So the probability of a coalescent event, that is an event in which at least two individuals in generation t share an ancestor in generation $t-1$, is $\lambda_{N,n} = 1 - \frac{N_{[n]}}{N^n}$.
We call the event in which j individuals of one generation select the same ancestor an j-coalescent. When a coalescent event occurs it is well possible that different types of j-coalescent events occur and each type can occur multiple times. For instance, at one generation two 2-coalescences and one 3-coalescent can occur. In this case the number of extant lineages will be reduced by four.
The state in which the process is, when we look t generations back in time is represented by a set R_t of all equivalent relations, which are pairs (i,j) where $i,j \in \{1,\dots,n\}$. A pair $(i,j) \in R_t$ if and only if the sequences of individuals i and j of the original sample, share a common ancestor at generation t. At time 0, that is the time of observation of the n individuals in the sample, we are in the initial state $\Delta := \{(i,i) : 1 \leq i \leq n\}$.
Define $\phi_j(R_t, R_{t+1})$ as the number of j-coalescences that occur from generation t to $t+1$. Note that in this notation we can speak of a 1-coalescent, when a parent has exactly one child. Define $\phi(R_t, R_{t+1}) = \sum_{j=1}^{n} \phi_j(R_t, R_{t+1})$, hence $\phi(R_t, R_{t+1})$ is the number of extant lineages remaining at generation $t+1$. Also note that all the equations for ϕ are under the restriction that $\sum_{i=1}^{n} i\phi_i(\Delta, R_1) = n$, because the sum of all offspring of the parents has to be equal to the sample size n itself.
Denote two succeeding states of the process $(R_t)_{t\geq 0}$ by $R_t = \xi$ and $R_{t+1} = \eta$. So the process changes from state ξ to state η. Denote by $k = \phi(\eta) := \phi(\xi, \eta)$ and by $n = \phi(\xi) := \phi(R_{t-1}, \xi)$. Note that $\phi(R_t, R_{t+1})$ only depends on the state to which the process is changing, R_{t+1}, not on the state from which the process is moving, R_t. Define $p_{\xi\eta}$ as the transition probability from state ξ to state η. The number of observed sequences will change from n to k. This means that k individuals in generation $t+1$ choose an unoccupied ancestor in generation t, while the other $n-k$ individuals in generation $t+1$ choose a specific ancestor which is already occupied. Under the WF-model, the probability that k individuals all choose a different ancestor out of N is $\frac{N_{[k]}}{N^k}$ and the probability of choosing a specific individual is $\frac{1}{N}$. Thus

$$\begin{aligned} p_{\xi\eta} &= \mathbb{P}(R_{t+1} = \eta | R_t = \xi) \\ &= \frac{1}{N^{n-k}} \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right) \qquad (10) \\ &= \frac{N_{[k]}}{N^n}. \end{aligned}$$

So the transition probability $p_{\xi\eta}$ only depends on the number of observed ancestral sequences of both states. If we are only interested in the number of ancestral lines that remain, then we want an expression for the probability that the number of ancestral sequences changes from n to k:

$$\mathbb{P}(\phi(R_t, R_{t+1}) = k | \phi(R_{t-1}, R_t) = n) \quad = \quad S(n,k)\frac{N_{[k]}}{N^n}, \qquad (11)$$

where $S(n,k)$ is the Stirling number of the second kind, which is the number of ways to partition n elements into k non-empty boxes. For more information on the Stirling number of the second kind and some of its characteristics we refer to Appendix A.
Denote by $a = \{a_1, \dots, a_k\}$ the number of offspring, which are present in the genealogy of the sample, of the k individuals in the ancestral population $t+1$, which are the parents of the n observed individuals in generation t. So $\sum_{i=1}^{k} a_i = n$ and we can say that a forms a partition of n elements into k boxes.
Define κ_j, $1 \leq j \leq n$, as the number of j-coalescences. So $\kappa_j = \phi_j(R_t, R_{t+1})$. Alternatively we can formulate κ_j as $|\{a_i \in a : |a_i| = j\}|$, in which case we speak of κ_j as the multiplicity of j in the offspring-set a.
Denote by $\mathcal{A}$ a subset of the set $\{R_{t+1} : \phi(R_{t-1}, R_t) = n, \phi(R_t, R_{t+1}) = k\}$, with a specific offspring-set a for the ancestral population. So

$$\begin{aligned} \mathcal{A} &= \{R_{t+1} : \phi(R_{t-1}, R_t) = n, a = \{a_1, \dots, a_k\}, \sum_{i=1}^{k} a_i = n\} \\ &= \{R_{t+1} : \phi(R_{t-1}, R_t) = n, \phi_i(R_t, R_{t+1}) = \kappa_i, 1 \leq i \leq n\}. \end{aligned}$$

The number of ways to divide n elements into the partition a is the multinomial coefficient $\binom{n}{a_1 \dots a_k}$. But this is not the number of different states which are in the event $\mathcal{A}$, because no difference can be made between coalescents of the same kind. For example, if $|a_1| = |a_5| = 2$, then there is no difference between the event A in which individuals $i, j \in \{1, \dots, n\}$ are offspring of ancestor 1 (and so $i, j \in a_1$) and individuals $l, m \in \{1, \dots, n\}$ are offspring of ancestor 5 and event B in which $i, j \in a_5$ and $l, m \in a_1$. This is because all individuals are exchangeable, i.e. there is no order in the ancestral population. We can only differentiate between individuals with a different number of offspring.

The number of states we are overcounting if we simply take the number of partitions depends on the multiplicity of the different i-coalescents. If, for example, the multiplicity of the 2-coalescents κ_2 is three, then we have to devide the multinomial coefficient by 3!. This we have to do for all κ_i's. Note that if for a fixed i, no i-coalescent occur, then $\kappa_i = 0$, but by convention $0! = 1$, so we are not dividing by zero.

Hence the number of states which are in the event $\mathcal{A}$ is

$$\begin{aligned}
|\mathcal{A}| &= \frac{\binom{n}{a_1 \dots a_k}}{\kappa_1! \dots \kappa_n!} \\
&= \frac{n!}{a_1! \dots a_k!} \frac{1}{\kappa_1! \dots \kappa_n!} \\
&= \frac{n!}{(1!)^{\kappa_1} \dots (n!)^{\kappa_n} \kappa_1! \dots \kappa_n!} \\
&= \frac{n!}{\prod_{i=1}^{n} (i!)^{\kappa_i} i!}
\end{aligned}$$

Note that if, for some i, $\kappa_i = 0$, then $(i!)^{\kappa_i} = 1$.

Now we have an expression for the probability of event $\mathcal{A}$ given that $\phi(R_{t-1}, R_t) = n$:

$$\mathbb{P}(\mathcal{A} | \phi(R_{t-1}, R_t) = n) = \frac{n!}{\prod_{i=1}^{n} (i!)^{\kappa_i} i!} \frac{N_{[k]}}{N^n}, \tag{12}$$

in which k can be written as $\sum_{i=1}^{n} \kappa_i$. So the probability is completely defined by the number of elements in the starting state n, and the number of observed i-coalescences κ_i, $1 \leq i \leq n$.

Suppose that we have n observed sequences in generation t, and given that a coalescent event occurs, then the probability of event $\mathcal{A}$ is

$$\begin{aligned}
\alpha_n(\mathcal{A}) &= \mathbb{P}(\mathcal{A} | \phi(R_{t-1}, R_t) = n, k \leq n) \\
&= \frac{\mathbb{P}(\mathcal{A} | \phi(R_{t-1}, R_t) = n)}{\left(1 - \frac{N_{[n]}}{N^n}\right)}.
\end{aligned}$$

By this we have defined the process state space and its transition probabilities, without the assumption of a large population size and a relative small sample size.

4.1 Comparing different coalescent probabilities

Remember from formula (11) that the probability to go from state R_t, with $\phi(R_{t-1}, R_t) = n$, to an other given state R_{t+1}, with $\phi(R_t, R_{t+1}) = k$, is $\frac{N_{[k]}}{N^n}$, which is a monotonically increasing function of k. This means that the probability of seeing a coalescent event with only one 2-coalescent, i.e. $k = n - 1$, is larger than the probability that any other coalescent event occurs. But this does not mean that the probability of observing a coalescent event with only one 2-coalescent is larger than any other coalescent event. To give a further explanation of this statement we compare the probabilities of different states below.

To make a shorter notation for $\mathcal{A}$ we only mention those κ_i for which $i > 1$ and which are non-zero. For the remainder of this section we will assume that the starting number of extant lineages is n. So if for

example three 2-coalescences, two 3-coalescences and one 5-coalescent occur, then we denote the event $\mathcal{A}$ by $\{\kappa_2 = 3, \kappa_3 = 2, \kappa_5 = 1\}$. For a given n, we know that $k = n-11$ and so that $\kappa_1 = n-3*2-2*3-1*5 = n-17$. So the complete event is described by the set $\{\kappa_2 = 3, \kappa_3 = 2, \kappa_5 = 1\}$.
We denote with $\{R_t|\phi(R_{t-1}, R_t) = n\}$ the collection of all states R_t in which there are n observable sequences. Now we want to compare some probabilities of the classical n-coalescent model and the exact coalescent model.

Lemma 2. *For every number of reduced sequences $i > 1$ after a coalescent event there exists population sizes N and large enough sample size $n < N$, such that the probability of seeing an event in which the number of observed sequences is reduced from n to $n-i$ is larger then the probability of a classical two-coalescent event. Or in mathematical notation:*
$\forall i > 1\ \exists N, n > 0$, with $n \leq N$, such that $\mathbb{P}(\{R_t|\phi(R_t, R_{t+1}) = n-i\}|\phi(R_{t-1}, R_t) = n) > \mathbb{P}(\{\kappa_2 = 1\}|\phi(R_{t-1}, R_t) = n)$

Proof: First assume that only 2-coalescent events occur, then the only difference between the classical n-coalescent model and the exact model is that per generation only one 2-coalescent can occur in the first model, while multiple 2-coalescents can occur in the second. In our new notation this means that we want to compare the probability of one two-coalescent event $\mathbb{P}_F(\{\kappa_2 = 1\})$ with the probability of i two-coalescent events $\mathbb{P}_F(\{\kappa_2 = i\})$. First we define

$$q_j = \frac{\mathbb{P}(\{\kappa_2 = j+1\})}{\mathbb{P}(\{\kappa_2 = j\})}.$$

In the case that $\kappa_2 = j$, this means that $k = n-j$ and $\kappa_1 = n-2j$. Now using formula (12) we get

$$\begin{aligned}\mathbb{P}(\{\kappa_2 = j\}) &= \frac{n!}{1^{n-2j}(n-2j)!2^j j!}\frac{N_{[n-j]}}{N^n} \\ &= \frac{n!}{(n-2j)!j!2^j}\frac{N(N-1)\dots(N-n+j+2)(N-n+j+1)}{N^n}.\end{aligned} \tag{13}$$

And in the same way for $\kappa_2 = j+1$ we have that $k = n-j-1$, $\kappa_1 = n-2(j+1)$ and

$$\mathbb{P}(\{\kappa_2 = j+1\}) = \frac{n!}{(n-2j-2))!(j+1)!2^{j+1}}\frac{N(N-1)\dots(N-n+j+2)}{N^n}.$$

So we have

$$q_j = \frac{(n-2j)!i!2^j}{(n-2j-2)!(j+1)!2^{j+1}}\frac{1}{N-n+j+1} = \frac{(n-2j)(n-2j-1)}{2(j+1)}\frac{1}{N}\left(1-\frac{n-j-1}{N}\right)^{-1}.$$

Note that $q_j > 1$ for sufficient large N and n. Hence we have

$$\frac{\mathbb{P}(\{\kappa_2 = i\})}{\mathbb{P}(\{\kappa_2 = 1\})} = q_1 \cdot q_2 \cdot \dots \cdot q_{i-1} > 1,$$

for n large enough. Hence we can conclude that if the sample size n is large in relation to the total population size N, the probability of multiple two-coalescents in the exact model is larger then the probability of only one 2-coalescent.
Because

$$\{\phi(R_t, R_{t+1}) = n-i|\phi(R_{t-1}, R_t) = n, \phi_2(R_t, R_{t+1}) = i\} \subset \{\phi(R_t, R_{t+1}) = n-i|\phi(R_{t-1}, R_t) = n\},$$

we can conclude that for every number of reduced sequences $i > 0$ after a coalescent event there exists population sizes N and large enough sample size $n < N$, such that the probability of seeing an event in which the number of observed sequences is reduced from n to $n-i$ is larger then the probability of a classical two-coalescent event. □

In the same way we can compare the probabilities of the event with two 2-coalescents $\{\kappa_2 = 2\}$, with the event of one 3-coalescent. Both events have the same $k = n-2$, while the first event has $\kappa_1 = n-4$ and the second has $\kappa_1 = n-3$, so if we use (12) we get

$$\frac{\mathbb{P}(\{\kappa_2 = 2\})}{\mathbb{P}(\{\kappa_3 = 1\})} = \frac{(n-3)! \cdot 2 \cdot 3}{(n-4)! \cdot 2 \cdot 2^2} = \frac{3}{4}(n-3)$$

So if the number of remaining sequences becomes small, then the probability of a 3-coalescent is of the same order as the probability of two 2-coalescents.

4.2 Critical sample size

Remember that the probability that no coalescent event occur in a generation is $\frac{N_{[n]}}{N^n}$, with:

$$N_{[n]} = \prod_{i=0}^{n-1} N - i = N^n - \sum_{i=1}^{n-1} iN^{n-1} + \sum_{i=1}^{n-2} i \sum_{j=i+1}^{n-1} jN^{n-2} + O(N^{n-3})$$

Hence the probability of a coalescent event $\lambda_{N,n}$ for the exact model is

$$\begin{aligned} 1 - \frac{N_{[n]}}{N^n} &= 1 - \frac{N^n - \sum_{i=1}^{n-1} iN^{n-1} + \sum_{i=1}^{n-2} i \sum_{j=i+1}^{n-1} jN^{n-2} + O(N^{n-3})}{N^n} \\ &= \frac{1}{N}\sum_{i=1}^{n-1} i + \frac{!}{N^2}\sum_{i=1}^{n-2} i \sum_{j=i+1}^{n-1} j + O(\frac{1}{N^3}) \\ &= \frac{n(n-1)}{2N} - \frac{2n(n-1)(n-2)}{3N^2} + O(\frac{1}{N^3}) \end{aligned} \tag{14}$$

If the sample size n is relatively small compared to the population size N, then we could state that

$$1 - \frac{N_{[n]}}{N^n} \approx \frac{n(n-1)}{2N}$$

In the next lemma we show for what values of N and n the above approximation is reasonable.

Lemma 3. *For all $N > 2$ and for all $2 < n \leq N$:*

$$1 - \frac{N_{[n]}}{N^n} \quad < \quad \frac{n(n-1)}{2N} \tag{15}$$

Proof. By induction. Suppose $N > 2$ and $n = 3$, then $1 - \frac{N_{[3]}}{N^3} = \frac{N^3-(N^3-3N+2)}{N^3} = \frac{3}{N} - \frac{2}{N^2} < \frac{3}{N} = \frac{3(3-1)}{2N}$
Now suppose that the equation (15) is true for n, then we have to prove it's true for $n+1$:

$$\begin{aligned} 1 - \frac{N_{[n+1]}}{N^{n+1}} &= 1 - \frac{N_{[n]}(N-n)}{N^{n+1}} \\ &= 1 - \frac{N \cdot N_{[n]}}{N^{n+1}} + \frac{n \cdot N_{[n]}}{N^{n+1}} \\ &= 1 - \frac{N_{[n]}}{N^n} + \frac{n}{N} \cdot \frac{N_{[n]}}{N^n} \\ &< \frac{n(n-1)}{2N} + \frac{2n}{2N} \\ &= \frac{n(n+1)}{2N} \\ &= \frac{(n+1)((n+1)-1)}{2N}. \end{aligned}$$

So equation (15) holds for all $N > 2$ and for all $2 < n \leq N$. □

Remember that the LHS of equation (15) is the probability that an coalescent event occurs in the exact model, while the RHS is the coalescent probability for the classical n-coalescent model. So the coalescent probability in the exact model is always smaller than in the classical n-coalescent model. Furthermore, because the time to a coalescent event occurred is geometrically distributed, the expected time to the next coalescent event is one over the coalescent probability. So the expected time length to the next coalescent in the exact model is longer then in the classical n-coalescent model.

Another note on the result (15) is that we don't have to worry that our estimate is too small. We only have to take care that it isn't too large. To achieve the estimate isn't too large, the approximation has to be at least smaller than one, so

$$\begin{aligned}\frac{n(n-1)}{2N} < 1 \quad &\Leftrightarrow \quad n^2 - n - 2N < 0 \\ &\Leftrightarrow \quad n < \frac{1}{2} + \frac{1}{2}\sqrt{1+8N} \qquad (16)\end{aligned}$$

To give a more elegant notation, notice that $\frac{1}{2} + \frac{1}{2}\sqrt{1+8N} \approx \frac{1}{2}\sqrt{8N} = \sqrt{2N}$ for large N, so when the sample size $n \leq \sqrt{2N}$, then we know for sure that $\frac{n(n-1)}{2N} < 1$. We will speak of $n_b(N) = \sqrt{2N}$ as the sample size boundary.

Note that the sample size boundary is chosen such that the classical n-coalescent makes sense. If the restriction is violated, i.e. $n > n_b(N) = \sqrt{2N}$, then the probability that a two-coalescent event occurs is larger than one. The sample size being smaller then $n_b(N)$ does not imply that the probability of a multiple coalescent event can be neglected!

For the probability of the event that only i two-coalescents occur at an event we already found an expression in formula (13). Hence the probability that only one two-coalescent occurs at an event, $\{\kappa_2 = 1\}$, is

$$\begin{aligned}\mathbb{P}(\{\kappa_2 = 1\}) &= \frac{n!}{(1!)^{n-2}(n-2)!(2!)^1 1!}\frac{N_{[n-1]}}{N^n} \\ &= \binom{n}{2}\frac{1}{N-n+1}\frac{N_{[n]}}{N^n}. \qquad (17)\end{aligned}$$

The probability that no coalescent event occur we already expressed in formula (9). In our new notation the event that no coalescent event occurred is denoted by $\{\kappa_1 = n\}$ or to be more consistent it could be denoted by $\emptyset$. Hence

$$\mathbb{P}(\emptyset) = \mathbb{P}(\{\kappa_1 = n\}) = P_{\text{No Coal}} = \frac{N_{[n]}}{N^n}.$$

From formulas (9) and (17) we can give an expression for the probability of a multiple coalescent event as the complementary probability:

$$\begin{aligned}\mathbb{P}(\{\text{Multiple coalescent}\}) &= 1 - \mathbb{P}(\emptyset) - \mathbb{P}(\{\kappa_2 = 1\}) \\ &= 1 - \frac{N_{[n]}}{N^n} - \binom{n}{2}\frac{1}{N-n+1}\frac{N_{[n]}}{N^n} \qquad (18)\end{aligned}$$

The probability of a multiple coalescent event can also be derived in another way. A multiple coalescent event is the collection of all events in which the number of observed sequences is reduced from n to $n-2$ or less. The formula for the probability that at a generation the number of sequences is reduced from n to k we derived in formula (11), so from this we get

$$\begin{aligned}\mathbb{P}(\{\text{Multiple coalescent}\}) &= \sum_{k=1}^{n-2} S(n,k)\frac{N_{[k]}}{N^n} \\ &= \sum_{k=1}^{n} S(n,k)\frac{N_{[k]}}{N^n} - S(n,n)\frac{N_{[n]}}{N^n} - S(n,n-1)\frac{N_{[n-1]}}{N^n} \\ &= 1 - \frac{N_{[n]}}{N^n} - \binom{n}{2}\frac{1}{N-n+1}\frac{N_{[n]}}{N^n},\end{aligned}$$

N	$\lfloor n_b(N) \rfloor$	$\mathbb{P}(\{\text{Multiple coalescent}\})$
50	10	0.1993
100	14	0.2119
1000	44	0.2386
10000	141	0.2577
100000	447	0.2625
1000000	1414	0.2637

Table 4.1: Probability of a multiple coalescent event near the critical sample size. In this table we show the probabilities of a multiple coalescent event for different values of N and for sample sizes chosen just below the sample size boundary n_b. By $\lfloor n_b(N) \rfloor$ we denote the floor function of $n_b(N)$, i.e. the first integer below $n_b(N) = \sqrt{2N}$.

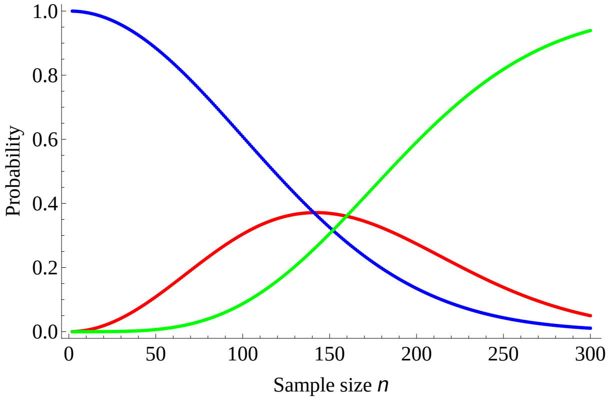


Figure 4.2: Probabilities of different types of events. The probabilities of $\mathbb{P}(\emptyset)$ (blue), $\mathbb{P}(\{\kappa_2 = 1\})$ (red) and $\mathbb{P}(\{\text{Muliple coalescent}\})$ (green) for $N = 10000$. On the horizontal axis is the sample size, and on the vertical axis the probability. The critical sample size $\lfloor n_b(N) \rfloor = 141$ for $N = 10000$.

because $S(n, n) = 1$ and $S(n, n-1) = \binom{n}{2}$ (see Appendix A). Which is indeed equal to formula (18).
In Table 4.1 we show that the probability of a multiple coalescent event can be reasonable large when the sample size n is near the sample size boundary n_b. One can see that for large N the probability of a multiple coalescent event is around 0.26. So, despite the fact that the criterion of $n < \sqrt{2N}$ is not violated, in situations in which n is close to n_b, the probability of a multiple coalescent event cannot be neglected.

An impression of the relative sizes of the three probabilities $\mathbb{P}(\emptyset)$, $\mathbb{P}(\{\kappa_2 = 1\})$ and $\mathbb{P}(\{\text{Multiple coalescent}\})$ is depicted in Figure 4.2.

4.3 Most recent common ancester

One of the most interesting features of a set of sequences is the expected time to the most recent common ancestor (MRCA). We want to compare the results of both models. Therefore we look at the expected fraction $s(n)$ of the number of lineages which remain of the n extant lineages after one unit of time.

$$s(n) = \frac{\mathbb{E}[\phi(R_t, R_{t+1}) | \phi(R_{t-1}, R_t) = n]}{n}$$

If, for the classical n-coalescent model, $\frac{2N}{n(n-1)} < 1$ than the expected time to a coalescent event is smaller than one. (So $\frac{n(n-1)}{2N} > 1$ and hence $n > n_b(N) = \sqrt{2N}$.) The expected time to see i successive classical two-coalescent events is:

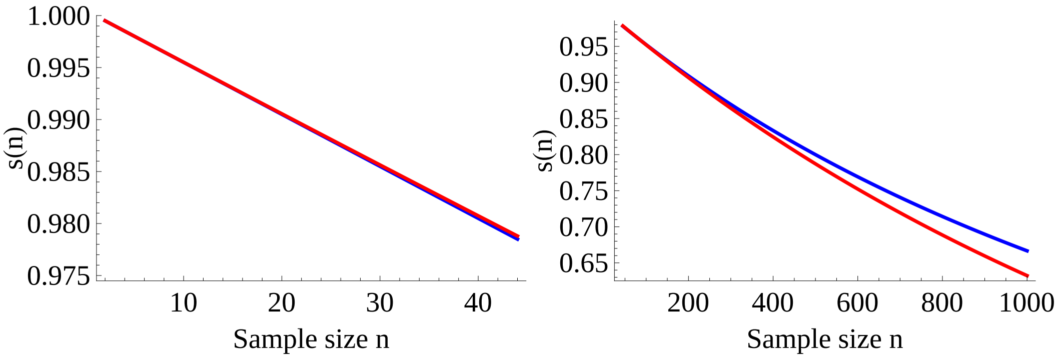
$$\begin{aligned}\frac{2N}{n(n-1)} + \frac{2N}{(n-1)(n-2)} + \ldots + \frac{2N}{(n-i+1)(n-i)} &= 2N\sum_{j=1}^{i}\frac{1}{(n-j+1)(n-j)} \\ &= 2N\left(\frac{i}{(n-i)n}\right) \\ &= 2N\left(\frac{1}{n-i} - \frac{1}{n}\right),\end{aligned}$$

where the second equation can be found by induction and the third equation by decomposing into partial fractions. If we put the above equation equal to one and solve it for i, then we get an expression for the expected number of coalescent events per time unit.

$$2N\left(\frac{1}{n-i} - \frac{1}{n}\right) = 1 \Leftrightarrow i = n\left(1 - \frac{2}{2 + \frac{n}{N}}\right)$$

So if $\frac{2N}{n(n-1)} < 1$, then we expect that after one time unit $n - n\left(1 - \frac{2}{2+\frac{2}{N}}\right)$ lineages remain, hence $s(n) = \frac{2}{2+\frac{n}{N}}$. If $\frac{2N}{n(n-1)} \geq 1$, then there are on average $\frac{n(n-1)}{2N} \leq 1$ coalescent events per time unit, so the expected number of lineages remaining after one time unit is $ns(n) = n - \frac{n(n-1)}{2N} = n\left(1 - \frac{n-1}{2N}\right)$ and hence $s(n) = \frac{1}{2}\left(2 - \frac{n-1}{N}\right)$ To calculate $s(n)$ in case of the exact model, first remember that $\mathbb{P}(\phi(R_t, R_{t+1}) = k | \phi(R_{t-1}, R_t) = n) = S(n,k)\frac{N_{[k]}}{N^n}$. If we then fix n and take the sum over all possible k's, then this sum should be one, hence:

$$\sum_{k=1}^{n} S(n,k)\frac{N_{[k]}}{N^n} = 1 \Leftrightarrow \sum_{k=1}^{n} S(n,k)N_{[k]} = N^n$$



(a) $n \leq n_b(N) = 44$ **(b)** $n > n_b(N) = 44$

Figure 4.3: The fraction of extant lineages after one time unit for the classical n-coalescent model (blue) and the exact model (red) for $N = 1000$.

and so

$$\begin{aligned}
ns(n) &= \mathbb{E}[\phi(R_t, R_{t+1}) | \phi(R_{t-1}, R_t) = n] \\
&= \sum_{k=1}^{n} k\mathbb{P}(\phi(R_t, R_{t+1}) = k | \phi(R_{t-1}, R_t) = n) \\
&= \sum_{i=1}^{n} kS(n,k) \frac{N_{[k]}}{N^n} \\
&= \frac{1}{N^n} \sum_{i=1}^{n} kS(n,k) \cdot N(N-1)\ldots(N-k+1) \\
&= \frac{1}{N^n} \sum_{i=1}^{n} S(n,k) \cdot N(N-1)\ldots(N-k+1) \cdot (N-(N-k)) \\
&= \frac{1}{N^n} \sum_{i=1}^{n} NS(n,k) \cdot ((N(N-1)\ldots(N-k+1)) - ((N-1)(N-2)\ldots(N-k+1)(N-k))) \\
&= \frac{1}{N^n} \left(\sum_{i=1}^{n} NS(n,k)N_{[k]} - \sum_{i=1}^{n} NS(n,k)(N-1)_{[k]} \right) \\
&= \frac{N}{N^n} \left(N^n - (N-1)^n\right) \\
&= N \left(1 - \frac{(N-1)^n}{N^n}\right) \\
&= N \left(1 - \left(1 - \frac{1}{N}\right)^n\right) \\
s(n) &= \left(\frac{n}{N}\right)^{-1} \left(1 - \left(1 - \frac{1}{N}\right)^n\right) \qquad (19)
\end{aligned}$$

The graphs in Figure 4.3 illustrate that the fraction of extant lineages after one time unit $s(n)$ is slightly smaller for the exact model. This implies that the exact model reaches its MRCA faster, but the difference is quite small [3].

5 Estimations for multiple coalescent events

In this section we will derive two different estimators for the effective population size, where we assume that the real underlying reproduction process is that of the exact coalescent algorithm, described in Chapter 4.

5.1 Generalized skylineplot estimator

If we want to use $\hat{N}^G$ for estimations of the effective population size on trees constructed according to the exact coalescent algorithm, we have to apply formula (7) on those trees. The likelihood function is defined for trees with only two-coalescent events in it, but for the exact coalescent model also multiple coalescent events can occur. Despite this, we can still use formula (8) for the log likelihood function. We just assume that every multiple coalescent event, in which the number of observed sequences is reduced by i, is the result of merging i classical two-coalescent intervals together.

It wasn't the purpose of Strimmer and Pybus to give an estimator for the effective population size outside the domain defined by the constraints given by Kingman, but only to get rid of the 'dips' in the skyline plot. But, by their efforts, they derived an estimator which can also be applied to phylogenetic trees which are composed according the exact coalescent theory of Fu. The observation of 'dips' can be explained as follows: when three of more identical sequences are sampled, then, according to most algorithms to construct phylogenetic trees, the time until they merge will be very small or even zero. Strimmer and Pybus noticed that this were in fact multiple coalescent events. Therefore they group the smaller intervals, which they consider artifacts of the tree-constructing algorithm.

Phylogenetic trees which are constructed according to the exact coalescent theory, already allow for multiple coalescent events. Therefore it isn't correct to group intervals of that kind of trees, since, by construction, those intervals were different coalescent events. So, when applying $\hat{N}^G$, to 'exact' phylogenetic trees we do not group intervals further.

5.2 Exact coalescent estimator

We want to use the exact coalescent model derived by Fu to construct a good estimator for the population size N, which depends on all characteristics of a phylogenetic tree. We assume that the population size N is piecewise constant and that the phylogenetic tree is given.

Given the number of extant lineages after each coalescent event $n_1, \dots, n_M$ and waiting times $\tau_1, \dots, \tau_M$. Remember that $k_i = n_{i+1}$. The waiting time τ_i is the time to wait between coalescent event i and $i+1$. Now each coalescent event is defined by the triple (n_i, k_i, τ_i).

The coalescent interval τ_i depends on the observed number of extant lineages n_i during the interval and the population size N.

In Chapter 4 we have derived for the discrete model the probability that no coalescent event occurs (Formula (9)). In Chapter 2 we derived a continuous time distribution for the waiting time until the next coalescent event (Formula (1)), an exponential distribution. For the exact coalescent model this can be done in the same way. Let t_n now be the waiting time until the next coalescent event in the exact model. Then t_n is the number of generations we have to wait until we see a coalescent event among the n observable sequences. Note that t_n has a geometric distribution: If $t_n = l + 1$, then the first l generations nothing happened while on the $l + 1$-th generation a coalescent event has to occur. Hence we have that:

$$\mathbb{P}(t_n = l + 1) = \left(\frac{N_{[n]}}{N^n}\right)^l \left(1 - \frac{N_{[n]}}{N^n}\right).$$

Similar to the derivation in Chapter 2, a continuous approximation for t_n is an exponentially distributed random variable with parameter $\lambda_{N,n} = 1 - \frac{N_{[n]}}{N^n}$, which is the rate at which coalescent events occur.

The probability $\mathbb{P}(k_i, \tau_i | n_i, N)$ that we have to wait a time τ_i until the next coalescent event and that the number of observable sequences, given population size N and starting number sequences n, after the event is k_i, is given by

$$\mathbb{P}(k_i,\tau_i|n_i,N) = S(n_i,k_i)\frac{N_{[k_i]}}{N^{n_i}}\frac{1}{\lambda_{N,n_i}}\lambda_i e^{-\lambda_{N,n_i}\tau_i} = S(n_i,k_i)\frac{N_{[k_i]}}{N^{n_i}}e^{-\lambda_{N,n_i}\tau_i},$$

with $\lambda_{N,n_i} = 1-\frac{N_{[n_i]}}{N^{n_i}}$ the rate at which a coalescent event occurs. So for each such triple we can define the likelihood function

$$L_i(N) = L(N|n_i,k_i,\tau_i) = S(n_i,k_i)\frac{N_{[k_i]}}{N^{n_i}}e^{-\lambda_{N,n_i}\tau_i}.$$

Since we want to maximize $L_i(N)$ over N we can neglect $S(n_i,k_i)$ in the equation, because it doesn't depend on N. For each i we obtain a maximum likelihood estimate $\hat{N}_i$ of the population size. However, one has to be careful with the interpretation of these estimates. For instance, even asymptotically the average $\frac{1}{M}\sum_{i=1}^{M}\hat{N}_i$ fails to converge to the real population size. This is caused by the skewness of the distribution of the remaining observable sequences k_i. See the article of E.M. Volz [16].

If we can assume that the population size is constant over time, we could take a product of the likelihood functions and maximize over this combined likelihood function. So we derive the following estimator for the population size

$$\hat{N} := \arg\max_{N\geq n}\prod_{i=1}^{M}\frac{N_{[k_i]}}{N^{n_i}}e^{-\lambda_{N,n_i}\tau_i}.$$

When we cannot assume a constant population size over time, we have to use the exact coalescent estimator $\hat{N}^F$, which is derived at each interval separately. Hence $\hat{N}^F$ is defined by

$$\hat{N}^F_i \quad = \quad \arg\max_{N\geq n}\frac{N_{[k_i]}}{N^{n_i}}e^{-\lambda_{N,n_i}\tau_i}. \tag{20}$$

If we plot all these estimations $\hat{N}^F_i$ for one phylogenetic tree in a graph we get a skyline plot. Remember that we assume a constant population during a coalescent interval, so the plot consists of a step function with M steps.

5.3 Likelihood region

We want to quantify how good the exact coalescent estimator $\hat{N}^F$ is. Therefore we want some kind of confidence intervals for the maximum likelihood estimations, called likilihood regions. By the exponential term in the likelihood function $L_i(N)$, we know that the integral over this function is finite. A typical graph for this function can be seen in Figure 6.5c.

We will use a much used method to approximate the likelihood regions for each coalescent event. We first have to introduce some theory and notations.

We define the *logarithmic relative likelihood function* as

$$r_i(N) = r(N|n_i,k_i,\tau_i) := \log L(N|n_i,k_i,\tau_i) - \log L(\hat{N}^F_i|n_i,k_i,\tau_i).$$

Note that $r_i(N)\in(-\infty,0]$, with zero as its maximum. This function is the logarithm of the *likelihood ratio*

$$R(N|n_i,k_i,\tau_i) := \frac{L(N|n_i,k_i,\tau_i)}{L(\hat{N}^F_i|n_i,k_i,\tau_i)}$$

S.S. Wilks (1937 [18]) has proved that $-2r_i(N)$ is $\chi^2(1)$-distributed, in which the '(1)' denote that the distribution has one degree of freedom, when the sample is large. He showed that $-2r_i(N)$ is $\chi^2(1)$-distributed up to terms of order $\frac{1}{\sqrt{h}}$, where h is the number of trails.

Because asymptotically minus two times the logarithmic likelihood ratio

$$-2r_i(N) = 2(\log L_i(\hat{N}^F_i) - \log L_i(N)) = 2\log\left(\frac{L_i(\hat{N}^F_i)}{L_i(N)}\right),$$

is $\chi^2(1)$-distributed, we can estimate the likelihood region.
Note that both events beneath are the same

$$\{N \geq n| - 2r_i(N) < \chi^2_{1-\alpha}(1)\} \quad \Leftrightarrow \quad \{N \geq n | L_i(N) > L_i\left(\hat{N}_i^F\right) e^{-\frac{1}{2}\chi^2_{1-\alpha}(1)}\}, \tag{21}$$

where $\chi^2_{1-\alpha}(1)$ is the value such that $\mathbb{P}(-2r_i(N) < \chi^2_{1-\alpha}(1)) = 1 - \alpha$, which can be found in various tables. For example when we allow a 5 percent uncertainty, then $\alpha = 0.05$ and hence $\chi^2_{1-\alpha}(1) = \chi^2_{0.95}(1) = 3.84$ [11]. This is because

$$\begin{aligned} -2r_i(N) = 2\log\frac{L_i(\hat{N}_i^F)}{L_i(N)} < \chi^2_{1-\alpha}[1] \quad &\Leftrightarrow \quad \log\frac{L_i(\hat{N}_i^F)}{L_i(N)} < \frac{1}{2}\chi^2_{1-\alpha}[1] \\ &\Leftrightarrow \quad \frac{L_i(\hat{N}_i^F)}{L_i(N)} < e^{\frac{1}{2}\chi^2_{1-\alpha}[1]} \\ &\Leftrightarrow \quad L_i(N) > L_i(\hat{N}_i^F)e^{-\frac{1}{2}\chi^2_{1-\alpha}[1]}. \end{aligned}$$

Hence for all values $N \geq n$ for which the likelihood is larger than $L_i\left(\hat{N}_i^F\right)e^{-\frac{1}{2}\chi^2_{1-\alpha}[1]}$, also $-2r_i(N) < \chi^2_{1-\alpha}[1]$, and so the likelihood region given by these N is an approximate $(1-\alpha)$-level confidence region for N. But we want to use this logarithmic likelihood ratio test on single coalescent events, so the number of trials $h = 1$. Hence the $\chi^2(1)$-distribution is not a good approximation for $-2r_i(N)$. Therefore we will not make further use of likelihood regions to give a measure for the variance of the estimations of $\hat{N}^F$.

6 Results

In this section we will show the characteristics of the different estimators, which we derived in the previous chapters. One way to compare the estimatars is by calculating the expectation and the variation of the estimators. But for many estimators and probability spaces this can not be done analytically, therefore we often have to use simulated data.
We will assume that the real underlying process is the one described in Chapter 4. Hence we assume that the underlying probability space is that of the exact coalescent and the data is also simulated according to that model. In this chapter, to shorten notation, we denote with $\mathbb{E}_F$ the conditional expectation under the exact coalescent model given population size N and sample size n, unless we mention otherwise. The same applies to the variance Var_F and the standard deviation $\sigma_F = \sqrt{\mathrm{Var}_F}$.
Because all the described estimators, i.e., the classical n-coalescent estimator $\hat{N}^K$, the generalized skyline plot estimator $\hat{N}^G$ and the 'exact' estimator $\hat{N}^F$ do not make use of the exact form of the phylogenetic tree, but only of the coalescent times and of the number of extant lineages, these are the only features we have to simulate. So if we want to simulate phylogenetic trees from a total population N and sample size $n < N$, according to exact coalescent algorithm, we have to make a list $(n_0, \dots, n_M)$ in which we choose the number of extant lineages left, given the last number of sequences, where $n_0 = n$ according to the distribution described below until $k = 1$. So $n_M = 1$. Then we appoint to each $n_i > 1$ an exponentially distributed coalescent time τ_i with parameter λ_{N,n_i}, which results in a list $(\tau_0, \dots, \tau_{M-1})$.
Furthermore we have to adapt some of the notation to the continuous time situation. In discrete time we described the state of the process at time t with R_t. Define $t_i = \sum_{j=0}^{i-1} \tau_j$ for $i \geq 1$ and $t_0 = 0$, and denote by R_{t_i} the state of the process after the i-th coalescent event occurs. Then the continuous time state process $(R_t)_{t\geq 0}$ is defined as $R_t = R_{t_i}$ $\forall t \in [t_i, t_{i+1}[$ and $i \geq 0$. Similar to Chapter 4 we denote with $\phi(R_{t_{i-1}}, R_{t_i})$ the number of extant lineages left after the i-th coalescent event. So $\phi(R_{t_{i-1}}, R_{t_i}) = n_i$ and we set $\phi(R_{t_{-1}}, R_{t_0}) = n_0 = n$.
The number of extant lineages is declining according to the distribution as in Formula (11), given a coalescent event occurs

$$\mathbb{P}(\phi(R_{t_i}, R_{t_{i+1}}) = k_i | N, \phi(R_{t_{i-1}}, R_{t_i}) = n_i) \quad = \quad S(n_i, k_i) \frac{N_{[k_i]}}{N^{n_i}} \lambda_{N,n_i}^{-1}, \tag{22}$$

for $1 \leq k_i < n_i$. While the coalescent times, as described in Section 5.2, are exponentially distributed with parameter

$$\lambda_{N,n_i} = 1 - \frac{N_{[n_i]}}{N^{n_i}}. \tag{23}$$

6.1 Performance of the classical n-coalescent estimator with large samples.

The classical n-coalescent was originally used to give estimates for how closely different species of organisms (mammals, birds, plants, etc.) were related. A measure for this is when they had their most recent common ancestor (MRCA). It has also been applied to estimate the fluctuations of the population size in time for animals like, mammoths [13]. When the classical n-coalescent is applied in these kind of circumstances, the assumptions of a large population and a relatively small sample size are easily met. Also the assumption that the samples are taken all at the same moment can be defended when the theory is used to look at time periods of thousands of years or even longer. But with modern methods it is becoming easier and less expensive to sample sequences and so the datasets are becoming larger, hence the assumptions of a large population and a relatively small sample size can not always be made. The consequences for the performance of $\hat{N}^K$ are showed in Figure 6.1. We see that the expected values of the estimator are growing quadratically with the sample size. So for large samples the estimator $\hat{N}^K$ gives a huge overestimation of the real population size N.

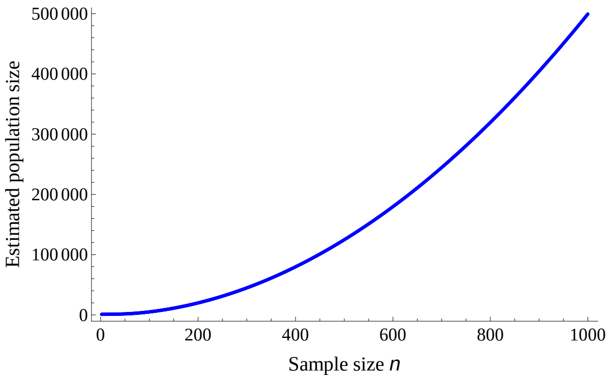


Figure 6.1: Expected estimations of $\hat{N}^K$ for $N = 1000$ and sample size varying from 2 to 1000.

6.2 Estimations with the generalized skyline plot

For the generalized skyline plot estimator $\hat{N}^G$, we can compute the expectation under the exact coalescent probability space analytically. We derived the formula for $\hat{N}^G$ in (6), which was

$$\hat{N}^G = \frac{nk}{2(n-k)}\tau.$$

So the expectation $\mathbb{E}_f$ under the exact coalescent probability space is

$$\begin{aligned} \mathbb{E}_F[\hat{N}^G] &= \mathbb{E}_F\left[\frac{nk}{2(n-k)}\tau\right] \\ &= \mathbb{E}_F\left[\frac{nk}{2(n-k)}\right]\mathbb{E}[\tau] \end{aligned}$$

Define $g_n(k) := \frac{nk}{2(n-k)}$, then

$$\begin{aligned} \mathbb{E}_F[g_n(k)] &= \sum_{i=1}^{n-1} g_n(i)\mathbb{P}(\phi(R_{t_{j-1}}, R_{t_j}) = i|N, \phi(R_{t_{j-2}}, R_{t_{j-1}}) = n) \\ &= \sum_{i=1}^{n-1} \frac{ni}{2(n-i)} S(n,i) \frac{N_{[i]}}{N^n} \lambda_{N,n}^{-1} \end{aligned}$$

And so

$$\mathbb{E}_F[\hat{N}^G] = \lambda_{N,n}^{-2} \sum_{i=1}^{n-1} \frac{ni}{2(n-i)} S(n,i) \frac{N_{[i]}}{N^n}, \tag{24}$$

which is not a very elegant formula, but can be computed. In Figure 6.2 we plot the values of $\mathbb{E}_F[g_n(k)]$ for $N = 1000$ and n varying from two to thousand. We did the same for different population sizes, all these results look the same. So, when we assume that the underlying probability space is the one defined by the exact coalescent, then in expectation the generalized skyline plot estimator performs quite well, compared to the classical n-coalescent estimator $\hat{N}^K$. Despite the fact that the estimator is not really taking into account the probabilities of multiple coalescent events, its expected estimation is of the same order as the real population size, even for large samples.

In the same way we also want to compute the variance of the estimator $\hat{N}^G$:

$$\mathrm{Var}_F(\hat{N}^G) = \mathbb{E}_F[\hat{N}^{G^2}] - \mathbb{E}_F^2[\hat{N}^G]. \tag{25}$$

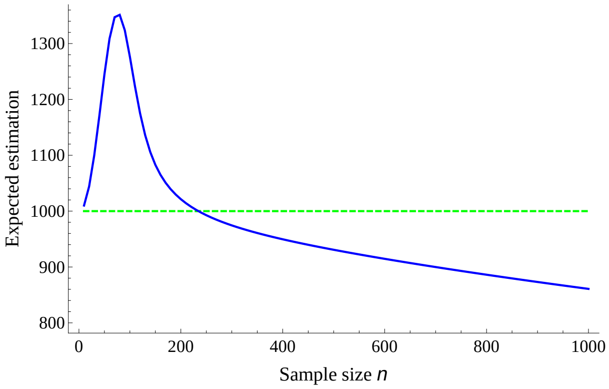


Figure 6.2: The expected estimation of the generalized skylineplot estimator $\hat{N}^G$ for $N = 1000$ and varying sample sizes n under the exact coalescent model probability space.

the last term in the above equation we derive from formula 24. The first term can be calculated as follows

$$\begin{aligned}
\mathbb{E}_F[\hat{N}^{G\,2}] &= \mathbb{E}_F[\tau^2]\,\mathbb{E}_F\left[\left(\frac{nk}{2(n-k)}\right)\right] \\
&= \frac{1}{\lambda_{N,n}^2}\,\mathbb{E}_f[g_n^2(k)] \\
&= \frac{1}{\lambda_{N,n}^2}\sum_{i=1}^{n-1} g_n^2(i)\mathbb{P}(\phi(R_t,R_{t+1}) = i|N, \phi(R_{t-1},R_t) = n) \\
&= \frac{1}{\lambda_{N,n}^2}\sum_{i=1}^{n-1}\left(\frac{ni}{2(n-i)}\right)^2 S(n,i)\frac{N_{[i]}}{N^n}\frac{1}{\lambda_{N,n}}. \qquad (26)
\end{aligned}$$

Now we are able to compute the variance of $\hat{N}^G$, for each N and n by combining Formulas (24) and (26) in formula (25). The results of (25) for $N = 1000$ and n varying from 2 to 1000 can be seen in Figure 6.3a. We see that the variance of the estimator under the exact coalescent event is quite large for all sample sizes. The problem which can arise from this high variance in the estimations is shown in Figure 6.3b. In this graph we see a cloud of estimations of the population size obtained from one hundred phylogenetic trees, all simulated with $N = 1000$ and $n = 1000$. On the horizontal axis are the sample sizes of all coalescent events which occurred and on the vertical axis are the estimations of $\hat{N}^G$. One can see the 'cloud' of estimations is very wide for each sample size. When we have only a single phylogenetic tree on which we have to estimate the development of the effective population size in time, we have a huge uncertainty.

6.3 Estimations with the exact coalescent estimator

For the exact coalescent estimator $\hat{N}^F$, we couldn't find a closed form formula for the expectation. Therefore we cannot do better then compute a sample mean and a sample variance. We have simulated 800 pylogenetic trees with a constant total population of 1000 and a starting sample size of 1000. By evaluating the exact coalescent estimations for the effective population size over all the trees, we can say something about the accuracy of the estimator for different sample sizes.

When we speak about the expected estimation of $\mathbb{E}_F\,\hat{N}_n^F$ for a certain sample size n we mean the approximation by the sample mean $\bar{N}_n^F = \frac{1}{m}\sum_{i=1}^m \hat{N}_{n,i}^F$, where m is the number of realizations $\hat{N}_{n,i}^F$ of phylogenetic trees, which pass a state with n observable sequences. Note that $0 \leq m \leq 800$ and that there are many sample sizes n for which we have no observation and therefore no approximation for the expected estimation. The same applies to the variance $\text{Var}(\hat{N}_n^F)$, which is the sample variance $\frac{1}{m-1}\sum_{i=1}^m(\hat{N}_{n,i}^F - \bar{N}_n^F)^2$. In Figure

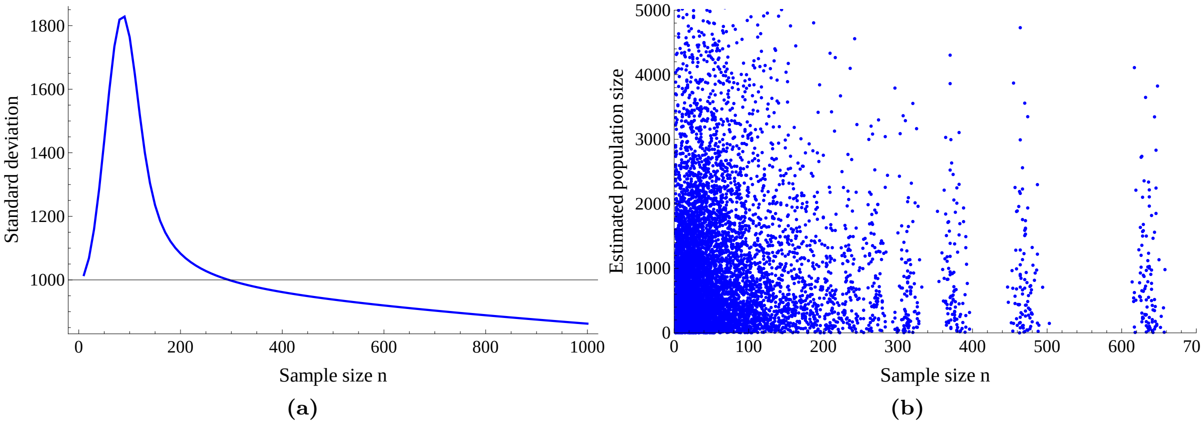


Figure 6.3: Variation of estimates of $\hat{N}^G$. In 6.3a we see the standard deviation of $\hat{N}^G$ for $N = 1000$ and n varying from 2 to 1000. In 6.3b each dot is an estimation of $\hat{N}^G$. The results are from one hundred simulated trees.

6.4 we depict the number of observations for each sample size $2 \leq n \leq 1000$ from the 800 simulated trees.

If we want the expected estimation and its variance of $\hat{N}^F$ for a specific sample size n, we should simulate a lot of remaining sequences k_i's and coalescent intervals τ_i's given sample size n and population size N, and compute the sample mean and variance. But, since our objective is to give estimations for the population size from a given phylogenetic tree or a bunch of trees, we want to give the estimations without extra simulated data. Besides, for the smaller sample sizes we have enough observations, while for the very large sample sizes the behaviour of the estimators is clear from the estimations we have from the view area's in that domain, for which we have data.

In Figure 6.5a we see that the estimator gives an overestimation of the population size on an interval between 40 and 160 observable sequences. This is remarkable because we assume that the underlying process is the exact coalescent model and $\hat{N}^F$ is derived from that model. The overestimation can be explained by looking at the likelihood function in case of a two coalescent event, i.e. $k = n - 1$. See Figure 6.5c. This Figure illustrates that the likelihoodfunction of a two-coalescent event is much less steep or peaked, then the likelihoodfunctions of multiple coalescent events. The likelihood funcion in the case of a two-coalescent event is only pushed down by the negative exponential factor added by the coalescent time. The two-coalescent event alone, given that a coalescent event occurs, gives no information about the most likely value of the population size, so every population size $N > n$ will do. Thus, when a two-coalescent event occurs, it will probably gives an overestimation of the population size.

For small sample sizes, when most occuring events are two-coalescent events, every estimation of a two-coalescent event contains less information then the estimation of a multiple coalescent event. But since we take the sample mean we do not taken this into account. Therefore we derive an overestimation for the population size for small sample sizes.

When it becomes very unlikely that a single two-coalescent occurs, then the exact coalescent estimator $\hat{N}^F$ performs very well.

The estimator $\hat{N}^F$ is sensitive to the kind of coalescent event that is happening, while the estimation shows only minor drifts caused by the observed coalescent time. For this see Figures 6.5c and 6.5d.

In the same way we compute the sample variance of the estimator. We see that for small samples the variance of the estimator is more or less the same as for the generalized skyline plot estimator $\hat{N}^G$, but for larger samples the variance quickly drops to zero. See Figure 6.6.

When we look at the individual estimations of the first hundred simulated phylogenetic trees, (the same trees as in the section 6.2) we see indeed that the estimations for larger samples have less fluctuations (Figure 6.7a). In this graph there seems to be 'lines' of estimations. To have a better view of these 'lines' we plot

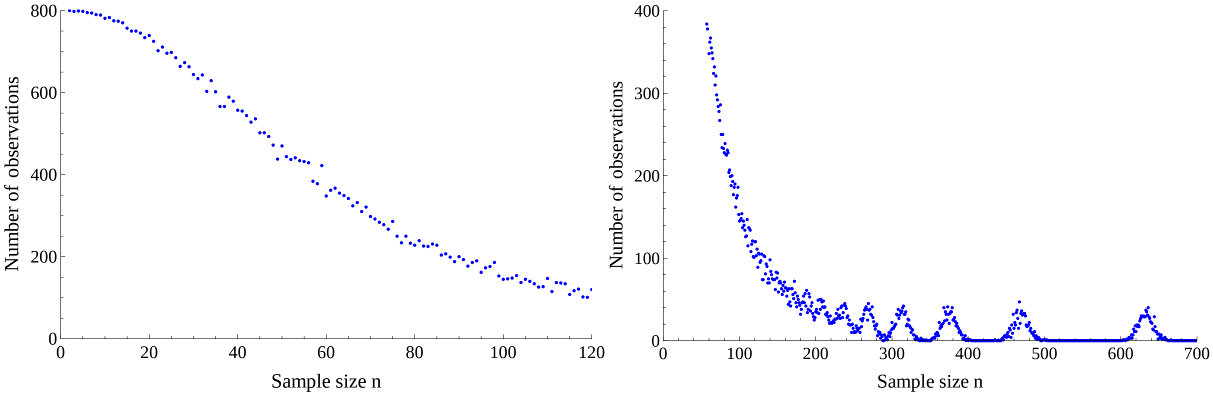


Figure 6.4: Number of observations. In both graphs we see on the horizontal axis the sample size n and on the vertical axis the number of observations for that sample size from the 800 simulated trees. In the left graph one can see that for relative small sample sizes we have a large number of observations, so here the sample mean and the sample variance would give good estimations. In the right graph we see that the number of observations becomes very small or even zero for large sample sizes. The individual peaks on the right side are caused by first few coalescent events.

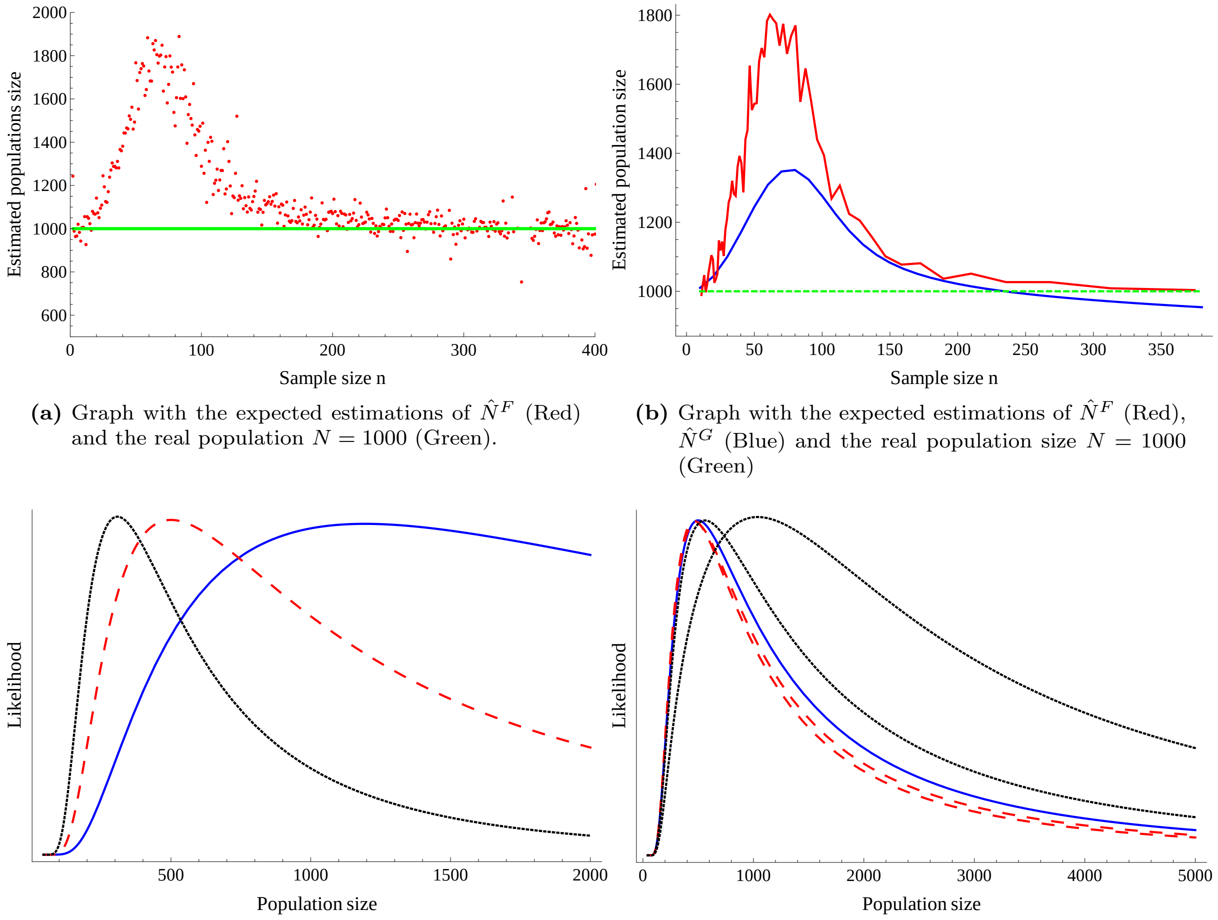
the same estimations on a logarithmic scale (Figure 6.7b) and we color the different type of coalescent event. An estimation of $\hat{N}^F$ derived from a two-coalescent event is colored blue, a three-coalescent event is colored red and the other events are colored green. We see that the 'lines' in the graph are caused by the different type of events. This can be explained by the fact that the estimator $\hat{N}^F$ is more sensitive to the type of event then to the length of the coalescent interval and that these two random variables are independent of each other.

When we compare the graphs of the expected estimation of $\hat{N}^F$ (Figure 6.5a) and the expected variance (Figure 6.6), with the lines in Figure 6.7a, it seems that when the first line stops, i.e. no two-coalescent events occur anymore, then the expected estimation of $\hat{N}^F$ becomes better and its variance drops dramatically. Therefore we want to determine the sample size at which it becomes unlikely that any two-coalescent event occurs. In Table 4.1 we have computed the probability of a multiple coalescent event near the critical sample size $n_b(N) = \sqrt{2N}$. It seems that this probability depends on the ratio $a = \frac{n}{\sqrt{2N}}$ and not directly on the population size N. To check this statement we plot the graphs of $\mathbb{P}(\emptyset)$, $\mathbb{P}(\{\kappa_2 = 1\})$ and $\mathbb{P}(\{\text{Multiple coalescent}\})$ in Figure 6.8 for $N = 1000$, 10000 and 100000, $n = \lfloor a\sqrt{2N} \rfloor$ and $a \in [0.01, 3]$. Here we see that all three pictures are the same.

So to determine for which sample size n it becomes unlikely for a two-coalescent event to happen, we have to determine for which a $\mathbb{P}(\{\kappa_2 = 1\}) < 0.01$, where 0.01 is an arbitrary cut-off-value. Therefore we want to solve

$$\min_{a>0} |\mathbb{P}(\{\kappa_2 = 1\}) - 0.01| = \min_{a>0} \left| \binom{n}{2} \frac{1}{N-n+1} \frac{N_{[n]}}{N^n} - 0.01 \right|,$$

for different choices of N and with $n = \lfloor a\sqrt{2N} \rfloor$. The results for this optimization problem can be seen in Table 6.1. The 'boundary' n_{2b} we found here is the smallest sample size for which $\mathbb{P}(\{\kappa_2 = 1\}) < 0.01$ and is defined as $n_{2b} = \lceil a\sqrt{2N} \rceil$. When we look at the performance of $\hat{N}^F$ for $N = 1000$ and $n = n_{2b} = 114$, then the expected estimation $\mathbb{E}[\hat{N}^F_{114}] \approx 1177$ and the standard deviation $\sigma(\hat{N}^F_{114}) = 822$.



(a) Graph with the expected estimations of $\hat{N}^F$ (Red) and the real population $N = 1000$ (Green).

(b) Graph with the expected estimations of $\hat{N}^F$ (Red), $\hat{N}^G$ (Blue) and the real population size $N = 1000$ (Green)

(c) Plot with the likelihoodfunctions of a two-coalescent event (Blue), a three-coalescent event (Red, dashed), and a four-coalescent event (Black), all with the same coalescent time $\tau = \lambda_{N,n}^{-1}$, with $N = 1000$ and $n = 44$. The likelihoodfunction of the events are rescaled such that all maximums are the same. Notice that the likelihoodfunctions of the multiple coalescent events are more peaked.

(d) Shifting of the likelihoodfunction for different coalescent times and $N = 1000$, $n = 44$ and $k = 42$. Plotted for $\tau = \lambda_{N,n}^{-1}$ (Blue), $\tau = \frac{1}{2}\lambda_{N,n}^{-1}$, $\tau = \frac{1}{5}\lambda_{N,n}^{-1}$ (Both red, dashed) and $\tau = 2\lambda_{N,n}^{-1}$, $\tau = 5\lambda_{N,n}^{-1}$ (Both black). Notice that there seems to be some kind of minimum estimation for the type of event notwithstanding the observed coalescent time.

Figure 6.5: Exact coalescent features.

N	500	1000	5000	10000
a	2.56	2.54	2.54	2.54
$n_{2b} = \lceil a\sqrt{2N} \rceil$	81	114	255	360

Table 6.1: Right bound

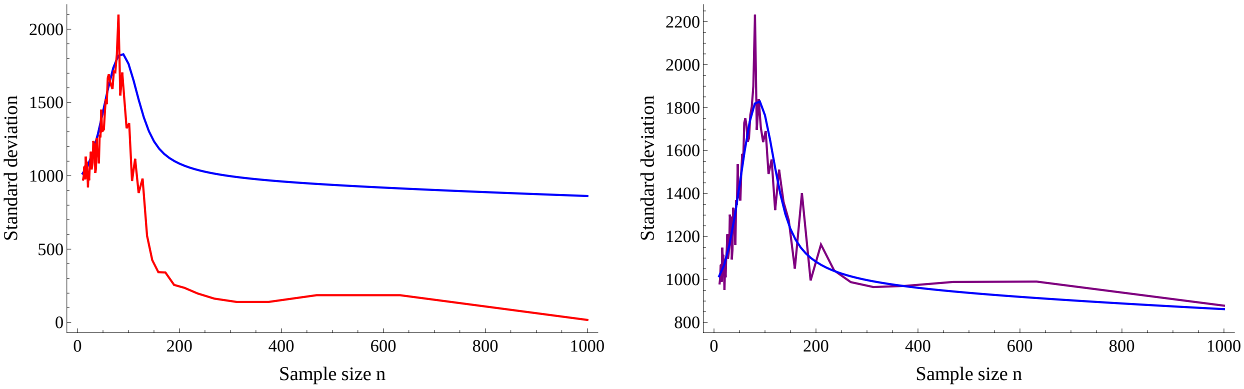
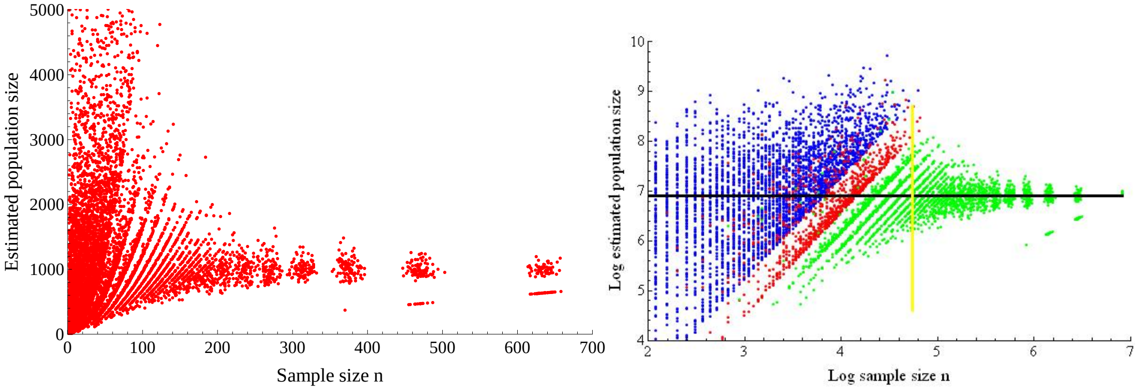


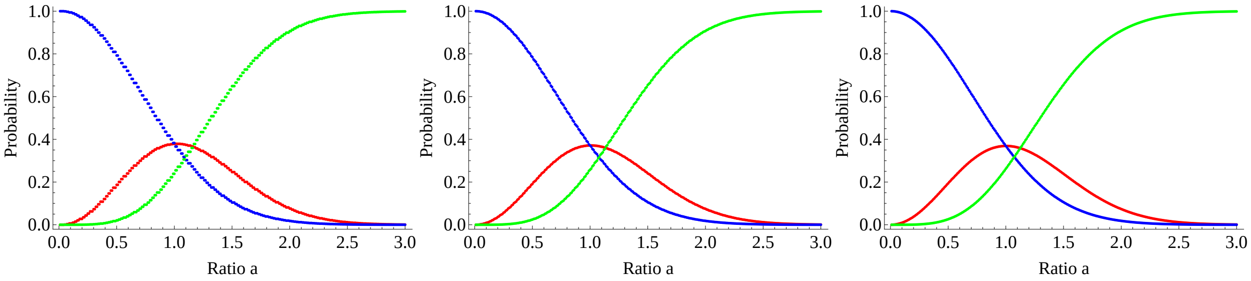
Figure 6.6: Variation of estimates of $\hat{N}^G$ and $\hat{N}^F$. In both graphs on the horizontal axis is the sample size varying from 2 to 1000. The real population size N is 1000. On the vertical axis is the standard deviation $\sigma = \sqrt{\text{Var}}$. In the left graph we see the standard deviation for both estimators $\hat{N}^F$ (in red) and $\hat{N}^G$ (in blue). In the right picture we see the sample standard deviation and the real standard deviation of $\hat{N}^G$, in which we see that the sample variance closely approaches the real variance.



(a) Estimations of $\hat{N}^F$

(b) Estimations of $\hat{N}^F$ on a logarithmic scale. The estimations derived by a two-coalescent event are colored blue, that of a three coalescent event red, and the others green. The yellow line depicts the n_{2b} boundary.

Figure 6.7: Variation of estimates of $\hat{N}^F$ depending on the type of event. In the left graph we see each estimation of $\hat{N}^F$ in one hundred simulated phylogenetic trees. On the horizontal axis is the sample size varying from 2 to 1000 and on the vertical axis the estimated values of the population size. The real population size $N = 1000$. In the right graph the same estimations but now both axes are on a logarithmic scale and we have colored the type of coalescent events. The groups of low estimations for the large sample sizes are caused by a defect in the numerical algorithm to maximize the likelihoodfunction. Sometimes the algorithm stops immediate at the start point n and returns n as maximum of the likelihoodfuncion.



(a) N=1000 **(b)** N=10000 **(c)** N=100000

Figure 6.8: Probabilities of different types of events for different population sizes. The probabilies of $\mathbb{P}(\emptyset)$ (blue), $\mathbb{P}(\{\kappa_2 = 1\})$ (red) and $\mathbb{P}(\{\text{Multiple coalescent}\})$ (green) for $N = 1000$, 10000 and 100000, and ratio $a = \frac{n}{\sqrt{2N}}$ varrying from 0.01 to 3.

$n = 44$	$\hat{N}^F$	$\hat{N}^G$
$\mathbb{E}_F$	1466	1200
σ_F	1215	1343

(a) $n = n_b = 44$

$n = 114$	$\hat{N}^F$	$\hat{N}^G$
$\mathbb{E}_F$	1177	1203
σ_F	822	1597

(b) $n = n_{2b} = 114$

$n = 210$	$\hat{N}^F$	$\hat{N}^G$
$\mathbb{E}_F$	1068	1014
σ_F	287	1068

(c) $n = 210$

$n = 223$	$\hat{N}^F$	$\hat{N}^G$
$\mathbb{E}_F$	1054	1006
σ_F	219	1038

(d) $n = 223$

$n = 238$	$\hat{N}^F$	$\hat{N}^G$
$\mathbb{E}_F$	1004	998
σ_F	200	1038

(e) $n = 238$

Table 6.2: The expected estimations and standard deviations of $\hat{N}^F$ and $\hat{N}^G$ for $N = 1000$ and different choices of n.

In Table 6.2 we compare the expected estimations and standard deviations of $\hat{N}^F$ and $\hat{N}^G$ for $n = n_b = 44$, $n = n_{2b} = 114$, $n = 210$, $n = 223$ and $n = 238$. We see that for all sample sizes $n < n_{2b}$ the variances of $\hat{N}^F$ and $\hat{N}^G$ are both large while the expected estimation of $\hat{N}^G$ is better than that of $\hat{N}^F$. But when $n > n_{2b}$ $\hat{N}^F$ performs better, and probably would perform even better when we simulate more data points for the larger sample sizes. Furthermore the variance of $\hat{N}^F$ becomes very small, unlike the variance of $\hat{N}^G$. So, as long as we cannot neglect the probability of a two-coalescent event (i.e. $n < n_{2b} = 2.54\sqrt{2N}$), we should use the generalized skyline plot estimator $\hat{N}^G$. When the probability of a two-coalescent event is very small $n < n_{2b}$ we should choose the exact coalescent estimator $\hat{N}^F$, since it has a much smaller variance and for most instances has a better expectation.

7 Discussion

We want to give estimations for the population size, given a phylogenetic tree. We have seen two different methods to describe the genealogy of phylogenetic trees. The first was the classical method in which only two-coalescent events are allowed. It is reasonable to neglect the other types of coalescent events when the population size N is very large and the sample size n is relative small. But, in the case of infectious diseases, the number of infected hosts can be quite small, while from a significant part of the infected hosts the pathogen is sampled. Therefore we derived the exact coalescent model, for small, oversampled populations, which can deal with multiple coalescent events. We have seen that the time to the MRCA for both methods is of the same order.

From the classical n-coalescent model we derived two estimators for the population size: the classical n-coalescent estimator $\hat{N}^K$ and the generalized skyline plot estimator $\hat{N}^G$. From the exact model we derived the exact estimator $\hat{N}^F$. We investigated how these estimators perform on phylogenetic trees constructed according to the exact coalescent model.

The critical sample size which is derived in the classical n-coalescent model is chosen such that as long as $n \leq \sqrt{2N}$ the probabilities of a coalescent event of no event are nicely between zero and one. But this does not mean that the probability of a multiple coalescent event is negligibly small. So even when one is sure that the sample size is smaller then the critical sample size, the classical n-coalescent estimator $\hat{N}^K$ can give an considerable overestimation of the population size.
There is another major problem when we use $\hat{N}^K$ for estimations of the population size. Because the expected estimation grows quadratically with the sample size n, the estimations will generally not violate the critical sample size boundary $n \leq \sqrt{2N}$. Therefore it is extremely difficult to check for this condition. So actually we should never opt for $\hat{N}^K$ as our estimator, unless we have independent reasons which guarantee that the assumptions are met.

The generalized skyline plot estimator $\hat{N}^G$ was derived to get rid of the dips in the estimations. By merging the coalescent intervals from two-coalescent events with each other, the derived estimator can be applied onto phylogenetic trees which allow for multiple coalescent events to happen. However the method treats such an event as being a combination of successive two-coalescent events. Furthermore, the estimator has a large variance in its estimations for all sample sizes.
The exact coalescent estimator $\hat{N}^F$ takes into account the multiple coalescent events. A disadvantage is that there does not exist a closed form formula for it. We can only compute its value numerically. So we also have to approximate the expected value and variance numerically. However for relative large sample sizes its variance is small compared with the variance of $\hat{N}^G$ and it has a good expectation.
Especially the two-coalescent events produce an overestimation of the population size. When the number of merging sequences in the event becomes larger, the curve in the likelihood function becomes more peaked. Such kind of multiple coalescent events happen more often when the sample size becomes large relative to the population size. A peaked likelihood function results in a small variance of the estimator and in a small likelihood regio. Therefore the estimations become trustworthy.
As a kind of rule of thumb we could state that when the sample size n is smaller than $n_{2b} = 2.54\sqrt{2N}$ we should opt for the estimator $\hat{N}^G$, because it is expected to give a better estimation of the population size, the variance of both estimators on that domain is more or less the same and its computational time is less. For sample sizes larger than n_{2b}, the exact coalescent estimator $\hat{N}^F$ should be chosen, since in this domain the expected value is better and its variance is much smaller than that of $\hat{N}^G$.
To decide whether or not the sample size $n > n_{2b}$, one should first use $\hat{N}^G$ to have an impression of the population size. This is because $\hat{N}^G$ has a much smaller computational time. If $n < \sqrt{2\hat{N}^G}$ it is very unlikely that $\hat{N}^F$ gives an better estimation of the population size. But if $n > \sqrt{2\hat{N}^G}$ it is reasonable to invest the extra computational time to compute $\hat{N}^F$ and to check for the boundary conditions with this estimations.

To obtain a better understanding of the properties of both estimators, research should be done to explain why the generalized skyline plot estimator $\hat{N}^G$ gives a smaller overestimation of the population size than the exact estimator $\hat{N}^F$. Maybe it is caused by the fact that $\hat{N}^G$ multiple coalescent events treats as a collecttion of successive two-coalescent events and by that the method neglects the events that three or more individuals choose the same ancestor.
Furthermore we want to know whether or not the performance of $\hat{N}^F$ is the same, when we apply it to phylogenetic trees constructed according to the classical n-coalescent model and we group intervals in the same way as Strimmer and Pybus did with the second order Akaike's information criterion. When, after grouping of the intervals, we derive the same tree as we should have got, when we used the exact model, than $\hat{N}^F$ should still perform better for large sample sizes. But, since the expected coalescent times in both models differ, we are not sure that we derive the same tree. It is possible, for instance, that after grouping, less intervals are left. In that case it is possible that $\hat{N}^G$ also performs better for large samples.

From the start we assume that the mutation rate is large compared to the speed at which new hosts are infected. As a consequence we could assume that every sampled individual has its own unique pathogen sample. When we have to deal with pathogens with a smaller mutation rate it becomes likely that we sample the same sequences at different hosts. When we assume that all sampling is done on the same moment, a problem arise when constructing a phylogenetic tree from data with equal sequences. Those sequences have distance zero to each other and therefore will al merge immediately. But when we assume that the sampling is done at different time points and we now the moment on which every sample is taken, this information can be used in constructing phylogenetic trees in which equal sequences do not immediately merge.
About the population size we only assume that it is constant during a coalescent interval. The estimates becomes probably better, when we can make further assumptions about the population size. For instance, if we can assume that there is a dependency between successive estimates in time of the population size, we could at further constrains on our estimation model and in this way getting better estimates. Additional information which may be available in case of an outbreak could be that the disease is in an endemic state in the population and so the total size of infected individuals is constant over time. Or that the disease is just introduced in the population and so the number of infected individuals is growing exponential.

Most algorithms to construct phylogenetic trees do not provide a single tree, but a collection of trees that are likely and their probability that they are the real one. The expected stochastic properties of these trees are similar, so using this collection of trees, one can take a weighted average at every moment in time to reduce the uncertainty of the estimates. Hence, when we have a collection of phylogenetic trees, we can provide better estimates of the population size.

In the setting of infectious diseases it is possible to have a situation in which there is a relative small, oversampled infected population. We have shown in this thesis that, when we want to apply coalescent theory this scenario, it is very important to take multiple coalescent events into account. The estimator $\hat{N}^F$, which we derived from the exact coalescent model, performs much better than the estimators based on the classical model. Furthermore we have given a method to decide which estimator should be used.

Appendices

A Stirling Numbers

The Stirling number of the second kind $S(n,k)$, defined on the natural numbers, is the number of different ways to divide n distinct elements over k nonempty subsets.
It is possible to express the Stirling numbers of the second kind as a recurrence equation. This can be done by the following reasoning:
Suppose we want to compute $S(n,k)$ by looking at all Stirling numbers of the second kind of sets with $n-1$ elements and adding one element to them. The only Stirling numbers of the second kind with $n-1$ elements which can be used are $S(n-1,k)$ and $S(n-1,k-1)$, because partitions with more then k subsets can not be reduced to k subsets by adding one element and in the same way partitions with $k-2$ or less subsets can not increase to k by adding one element.
If we are in the case of $n-1$ elements devided over k non-empty subsets (this can be done in $S(n-1,k)$ different ways), then we can add one element in k diffent ways by adding this element to one of the k subsets. While if we have $n-1$ elements devided over $k-1$ nonempty subsets, which is possible in $S(n-1,k-1)$ different ways we have to add the new element into a new subset to reach the desired number of k subsets. This is the only way. Now we can add the number of possibilities of the two cases to obtain $S(n,k)$;

$$S(n,k) \quad = \quad S(n-1,k-1) + kS(n-1,k). \tag{27}$$

There is a closed form formula for the Stirling numbers of the second kind, which is

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{n}(-1)^j\binom{k}{j}(k-j)^n.$$

[14] Note that $S(n,0)=0$, $S(n,1)=1$ and $S(n,n)=1$ for all $n \geq 1$. Furthermore $S(n,k)=0$ for all $k>n$. When we want ot devide n elements over $n-1$ different nonempty subsets ($S(n,n-1)$), we have to 'put' $n-1$ elements in the $n-1$ subsets and then it doesn't matter where we put the last element. There are $\binom{n}{n-1}=\binom{n}{2}$ different ways to choose $n-1$ elements from n. Hence $S(n,n-1)=\binom{n}{2}$.
We want to prove the next lemma

Lemma 4. *For all $N \geq 2$, $2 \leq n \leq N$ and $1 \leq k < n$*

$$\sum_{k=1}^{n} S(n,k)N_{[k]} = N^n. \tag{28}$$

Proof. We do this by induction so first suppose that $n=1$ and $N \geq n$, then equation (28) states that

$$S(1,1)N_{[1]} = 1N = N^1.$$

Now the induction step: Suppose that equation (28) is true for n, now we want to prove for $n+1$ that

$$\sum_{k=1}^{n+1} S(n+1,k)N_{[k]} \overset{?}{=} N^{n+1}.$$

Therefore we are going to rewrite the LHS of the above equation, by using the recurrence equation (27).

$$\begin{aligned}
\sum_{k=1}^{n+1} S(n+1,k)N_{[k]} &= \sum_{k=1}^{n+1} (S(n,k-1)+kS(n,k))N_{[k]} \\
&= \sum_{k=1}^{n+1} S(n,k-1)N_{[k]} + \sum_{k=1}^{n+1} kS(n,k)N_{[k]} \\
&= \sum_{k=0}^{n} S(n,k)N_{[k+1]} + \sum_{k=1}^{n} kS(n,k)N_{[k]} \\
&= \sum_{k=1}^{n} S(n,k)N_{[k]}(N-k) + \sum_{k=1}^{n} kS(n,k)N_{[k]} \\
&= \sum_{k=1}^{n} S(n,k)N_{[k]}(N-k+k) \\
&= N\sum_{k=1}^{n} S(n,k)N_{[k]} \\
&= N^{n+1}.
\end{aligned}$$

Hence equation (28) is true. □

References

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology.* Garland Science, second edition, 2004.

[2] R. Durrett. *Probability Models for DNA Sequence Evolution.* Springer, 2002.

[3] Y.X. Fu. Exact coalescent for the wright-fisher model. *Theoretical Populations Biology*, 69:385–394, Jun 2006.

[4] R.C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying envionment. *Phil. Tans. R. Soc. Lond.*, 344:403–410, 1994.

[5] B.M. Hallström and A. Janke. Mammalian evolution may not be strictly bifurcating. *Mol Biol Evol*, 27(12):2804–2816, 2010.

[6] J.F.C. Kingman. *Mathematics of Genetic Diversity.* Society for Industrial an Applied Mathematics, 1980.

[7] J.F.C. Kingman. Exchangeability and the evolution of large populations. *Exchangeability in probability and statistics*, pages 97–112, 1982.

[8] J.F.C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[9] W.-H. Li. Simple method for constructing phylogenetic trees from distance matrices. *Proc. Natl. Acad. Sci. USA*, 78(2):1085–1089, 1981.

[10] S.I. Resnick. *Adventures in Stochastic Processes.* Birkhäuser, 1992.

[11] J.A. Rice. *Mathematical Statistics and Data Analysis.* Thomson Brooks/Cole, 2007.

[12] A.G. Rodrigo and J. Felsenstein. Coalescent approaches to hiv population genetics. In K.A. Crandall, editor, *The evolution of HIV.* Johns Hopkins Univ. Press, Baltimore, 1999.

[13] B. Shapiro, A.J. Drummond, A. Rambout, et al. Rise and fall of the beringian steppe bison. *Science*, 306:1561–1565, 2004.

[14] H. Sharp. Cardinality of finite topologies. *J Combinatorial Theory*, 5:82–86, 1968.

[15] K. Strimmer and O.G. Pybus. Exploring the demographic history of dna sequences using the generalized skyline plot. *Mol. Biol. Evol.*, 18(12):2298–2305, 2001.

[16] E.M. Volz, S.L. Kosakovsky Pond, M.J. Ward, et al. Phylogdynamics of infectious disease epidemics. *Genetics*, 183:1421–1430, 2009.

[17] J. Wakeley. *Coalescent Theory.* Roberts & Company Publishers, 2009.

[18] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

[19] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.