Stockholms universitet

An introduction to mathematical modeling of the genealogical process of genes

Rikard Hellman

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2009:3
Matematisk statistik
Juni 2009

www.math.su.se/matstat

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2009:3**
http://www.math.su.se

An introduction to mathematical modeling of the genealogical process of genes

Rikard Hellman*

June 2009

Abstract

The aim with this paper is to give an introduction to the mathematical modeling of the genealogical process of genes. Humans have two copies of their genes which they inherit from their parents, one copy from each parent. A gene can be coded in different ways; the different codes are called alleles. This paper will describe the transmission of alleles between generations. The Wright-Fisher model is an urn model with replacement used to describe the variation of alleles in a population. From this model the Kingman coalescent process is derived. The Kingman coalescent process is a mathematical model for describing the line of descent for genes. It takes a sample from the population and step backwards in time to see how different lineages coalesce. The time to coalescence in this process is exponentially distributed when using a scaled time rate and letting the population size go to infinity. The coalescent process can be extended to a coalescent process including the possibility of mutation. The Ewens sampling formula is based on this extended version of the coalescent process and gives the distribution of the different alleles in a sample. Hoppe's urn model and the Chinese restaurant process can be used to simulate a sample from the Ewens sampling formula. By letting the sample size go to infinity an asymptotic estimate of the mutation rate can be derived. This estimate has quite low convergence rate which yields high variance of the estimates. Unfortunately there is no other way to estimate the mutation rate consistently with lower variance.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: r.hellman@hotmail.com . Supervisor: Andreas Nordvall Lagerås.

Contents

1 Introduction 3

2 A brief introduction to the genetic code 3

3 The Wright-Fisher Model and the Kingman Coalescent Process 4

3.1 Wright Fisher Model 5

3.2 The Kingman Coalescent Process 8

3.3 Mutations 10

4 Ewens Sampling Formula 11

4.1 Hoppe's Urn Model 12

4.2 Estimation of θ 13

4.3 Ewens Sampling Formula 15

4.4 The Chinese Restaurant Process 16

4.5 Maximum-likelihood estimation of θ 19

5 Simulation 20

1 Introduction

DNA is present in each one of our cells and carries all the genetic information that makes us human. We have thousands of genes encoded in our DNA that specifies our traits. All of us have two copies of our DNA and thus two copies of each gene; one inherited from the mother, and one from the father [1]. The purpose with this paper is to give an introduction to the mathematical modeling of the genealogical process of genes and the line of descent for them. This paper is based on chapter 1 "'Basic Models"' in Durrett, 2002 [2]. The genealogical process of genes, or the coalescent process as it will be referred to in this paper, has been developed to more and more advanced models through the years, but in this paper we will keep to the first and most basic model, the Kingman coalescent process. It can be shown that this very simple model yields the same results as the more advanced models. We will start off with explaining the Wright-Fisher model, which is a model describing the genetic variation among genes in a population. From the Wright-Fisher model the Kingman coalescent process will be derived. The coalescent process looks at a sample from the population and describes how the lines of descent are structured. From the fact that the gene focused on in the coalescent process may mutate during descent, the Kingman coalescent process will be extended to a coalescent process with mutation from which the Ewens sampling formula will be derived. The Ewens sampling formula gives the distribution of the entire sample in the coalescent process and yields some interesting results about the mutation rate and its estimate. The Ewens sampling formula will be described with two models, Hoppe's urn model and the Chinese restaurant process. With the help of either one of these two models it is easy to simulate the Ewens sampling formula, which will be done in the last section. Some interesting results from the simulations are the estimated mutation rate and the dependence between the simulated sample and the mutation rate.

2 A brief introduction to the genetic code

Before we start talking about the modeling of gene transmission it can be a good idea to gain some knowledge about the biological structure and function of the genetic material discussed in this paper.

Most living organisms have deoxyribonucleic acid (DNA) molecules which carry their genetic information about cell growth, division and function. DNA molecules are composed of two chains twisted around each other to form a double helix. Each chain consists of a sequence of four nucleotides, adenine (A), guanine (G), cytosine (C), and thymine (T). The two chains join together in a ladder-like form where adenine pairs up with thymine and guanine pairs up with cytosine. This means that the number of adenine nucleotides is equal to the number of thymine nucleotides and that the number of guanine nucleotides is equal to the number of cytosine nucleotides. It is the sequence of nucleotides on the DNA molecule that encodes the genetic information.

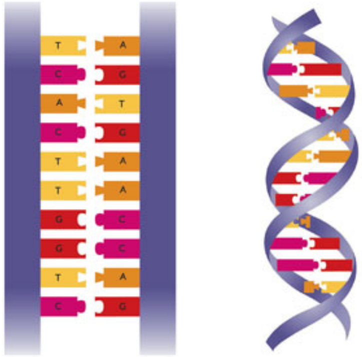


Figure 1: DNA chain [3].

A gene is a collection of nucleotides, or a segment of the DNA chain, which specifies a specific trait of an organism [4]. Each gene in the DNA chain has a special function; e.g. one gene may be responsible for your eye color while another may decide the shape of your nose. The specific sequences of nucleotides in a gene are called alleles. When we start modeling we will look at a specific gene and describe how the alleles are inherited from previous generations. Let us say that one gene is responsible for your eye color, then one allele might give you green eyes and another allele might give you blue. An organism is said to be haploid if it only has one copy of its genetic material. Humans and most higher organisms are diploid and have two copies of their genetic material; some plants can have many more copies. In diploid organisms the alleles are inherited from the parents, one from the father, and one from the mother. Thus, a diploid organism has two copies of the genes but may have two different alleles for each gene.

3 The Wright-Fisher Model and the Kingman Coalescent Process

The Wright-Fisher model is used to describe the way genes are transmitted from one generation to the next. In this section we will discuss the Wright-Fisher model and how the result from this model can be used to derive the Kingman coalescent, which is a process describing how different lineages coalesce backwards through time. Last in this section the possibility of gene mutations and how this affects the coalescent process will be discussed.

3.1 Wright Fisher Model

To describe the Wright-Fisher model we will start by explaining a strictly mathematical model and then we will discuss its application on the gene transmission process.

Assume we have $2N$ balls, all with different colors, and put them all in a big urn, urn number one. Then we draw a ball from the urn and put a new ball with the same color as the one that was drawn in a new urn, urn number two. The ball that was drawn from the first urn is put back in the first urn, i.e. it is an urn process with replacement. We repeat this process until we have $2N$ balls in urn number two. We can repeat this process several times and pick balls from urn number n and place new balls of the same colors into urn $n + 1$. We will most likely have a different distribution of balls in the new urn, so the probabilities to draw a ball with a specific color from the new urn will not be the same as before for all balls. In other words, it is likely that we from the first urn have drawn 2 or more balls of the same color.

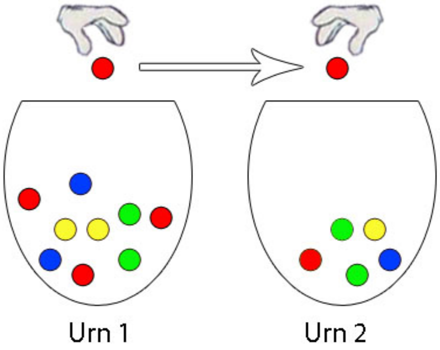


Figure 2: Urn Model.

From the urn model described above we get a probability that if there were i balls of a specific color, say red, in the nth urn there is j red balls in urn $n + 1$ which follow the binomial distribution and is calculated as,

$$p(i, j) = \binom{2N}{j} p_i^j (1 - p_i)^{2N-j}$$

where $p_i = i/2N$ is the probability to draw a red ball from an urn with i

red balls and

$$\binom{2N}{j} = \frac{(2N)!}{j!(2N-j)!}$$

is the number of ways one can choose j balls from $2N$ possible.

Now set X_n to be the number of balls in a specific color in urn number n, to be more specific say as above the number of red balls. Then we get that X_n, as discussed above, follows the binomial distribution.

$$X_n | X_{n-1} = i \sim \text{Bin}\left(2N, p_i = \frac{i}{2N}\right).$$

Expected value and absorbing states

Since X_n is binomially distributed the expected value of X_n is easily derived.

Lemma 1. $E[X_n | X_{n-1} = i] = 2N\left(\frac{i}{2N}\right) = i = X_{n-1}$.

This tells us that the expected value of X_n stays constant in time. As n goes to infinity X_n will eventually reach one of the absorbing states 0 or $2N$, i.e. eventually there will not be any red balls left or all the balls in the urn will be red. If none of the balls in the urn are red it is not possible to draw any red balls and put in the next urn and therefore $X_n = 0$ for all future urns. The same applies if all the balls are red, then all future urns will contain only red balls.

Now, let τ be the smallest amount of time it takes for X_n to reach one of the absorbing states, where time is measured in n, the number of urns.

$$\tau = \min\{n : X_n = 0 \text{ or } X_n = 2N\}$$

Then we get the following result.

Theorem 2. $P(X_\tau = 2N | X_0 = i) = \frac{i}{2N}$.

Proof. To prove Theorem 2 we first derive two equations which combined give the desired result. The first equation is

$$\begin{aligned} E[X_\tau | X_0 = i] = 0 \cdot P(X_\tau = 0 | X_0 = i) + 2N \cdot P(X_\tau = 2N | X_0 = i) = \\ = 2N \cdot P(X_\tau = 2N | X_0 = i). \end{aligned}$$

The second equation is a result of Lemma 1 and the optional stopping theorem [5]. The optional stopping theorem says that if $E\tau < \infty$ and if there exists a constant c such that $E|X_{i+1} - X_i| \leq c$ for $i = 1, 2, \ldots$ then

$EX_\tau = EX_1$. Since the expected value of X_n stays constant in time we get that

$$E[X_\tau | X_0 = i] = E[X_0 | X_0 = i] = i.$$

Combining these two equations gives

$$E[X_\tau | X_0 = i] = i = 2N \cdot P(X_\tau = 2N | X_0 = i) \Leftrightarrow P(X_\tau = 2N | X_0 = i) = \frac{i}{2N}.$$

□

Application of the Wright-Fisher model

Until now we have only described the Wright-Fisher model as a strictly mathematical process. Now we will see how this model can be used to describe how genes are transmitted between generations. Consider a population of N diploid individuals with non-overlapping generations and random mating. Non-overlapping generations mean that only one generation can exist at any one time. When we talk about random-mating we mean that each individual in the population is equally likely to mate and produce offspring. Since our interest is limited to one specific gene from a diploid organism we can consider that each individual is carrying two alleles. The possible number of different alleles is almost infinitive which will discuss in more detail in Section 4. Consider the different alleles as different colors just as the balls in the urn. We treat the diploid population of size N as a population of $2N$ haploid individuals. Now, take all the $2N$ alleles and put them in a big urn. To get the next generation of individuals we draw alleles from the urn, with replacement, in the same way we picked the colored balls before. The only difference now is that we see the colored balls as different alleles instead. In other words, the different color groups represent different families, and provided there was initially only one ball of a specific color the individuals in that color-family have at least one common ancestor.

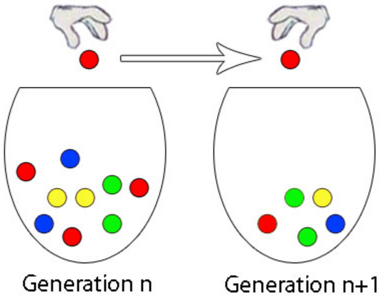


Figure 3: The Wright-Fisher Urn Model.

3.2 The Kingman Coalescent Process

When considering the Wright-Fisher model one can go backwards in time to see how different lineages coalesce. The Kingman coalescent process looks at a sample from the population and describes how the different lineages coalesce. If we look at a sample of size k from generation number n (i.e. urn n) and then go backwards in time to generation $(n-1)$ there is a probability that some of the individuals descend from the same parent. We will in this section discuss the distribution of coalescent times. We will also show that the coalescent times follow the exponential distribution if we scale the time with $2N$ and let $2N$ go to infinity.

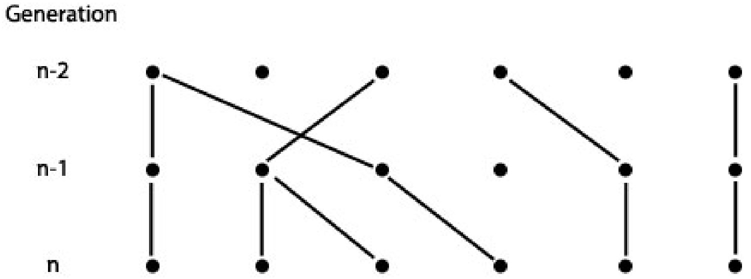


Figure 4: The Kingman coalescent process.

If we look at a sample of size k from the population, the probability that any two individuals choose the same parent is

$$\binom{k}{2} \cdot \frac{1}{2N} = \frac{k(k-1)}{2} \cdot \frac{1}{2N}$$

where $\binom{k}{2}$ is the number of ways we can choose two individuals from a sample of size k and the second term is the probability that those two choose the same parent when we have random mating. The probability that more than two individuals choose the same parent is negligible when $N \to \infty$ and will not be accounted for. We are also ignoring the probability that two or more different pairs collide. From this we get the probability of no collision as

$$\left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)$$

and the probability of no collision in the first n generations as

$$\left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)^n .$$

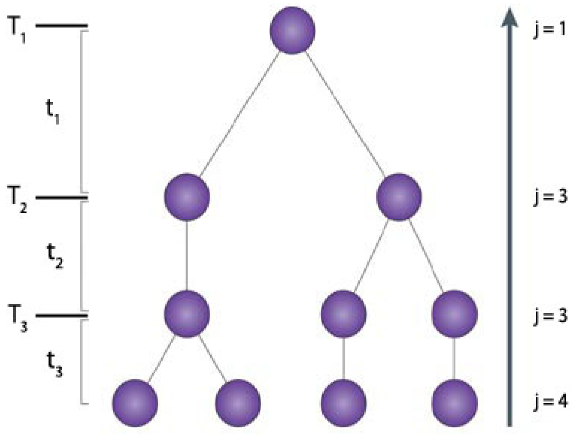
Note that $(1-x) \approx e^{-x}$ when x is small. If we apply this in the above formula we get that

$$\left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)^n \approx e^{-\frac{k(k-1)}{2} \cdot \frac{n}{2N}}$$

when N is large. From this result the following theorem is derived.

Theorem 3. *Let the population size $N \to \infty$, and let $t = n/(2N)$ be the time scale used. Then the time to the first collision follows an exponential distribution with parameter $k(k-1)/2$.*

Let T_j be the exact time at which the sample coalesce to j lineages and let t_j be the amount of time there are exactly j lineages. It then follows from the theorem above that t_j is approximately exponentially distributed with parameter $j(j-1)/2$, i.e. $t_j \sim \text{Exp}(j(j-1)/2)$.



Nature Reviews | Genetics

Figure 5: Coalescent process with exponentially distributed coalescent times [6].

An interesting result from all this is that the expected time it takes for all lineages in the sample to coalesce to a single lineage is

$$\begin{aligned} E[T_1] = E\left[\sum_{j=2}^{k} t_j\right] &= \sum_{j=2}^{k} \frac{2}{j(j-1)} = 2\sum_{j=2}^{k}\left(\frac{1}{j-1} - \frac{1}{j}\right) = \\ &= 2\left(\left(1-\frac{1}{2}\right) + \left(\frac{1}{2}-\frac{1}{3}\right) + \cdots + \left(\frac{1}{k-1}-\frac{1}{k}\right)\right) = \\ &= 2\left(1 - \frac{1}{k}\right). \end{aligned}$$

Note that T_1 converges to two as the sample size k goes to infinity, but t_2 which is the amount of time when there are only two lineages have an expected value equal to one, i.e. $E[t_2] = 1$. This means that the expected time waiting for the last collision is the same amount of time it takes for all lineages before the last to coalesce, as the sample size goes to infinity.

3.3 Mutations

So far we have only discussed the possibility of a coalescent event. But in fact there are two possible events that may occur, either a coalescent or a mutation of the allele. Say we pick a red allele from the urn at time n, then there is a possibility that the red allele will mutate to another color as we put it down in the new urn.

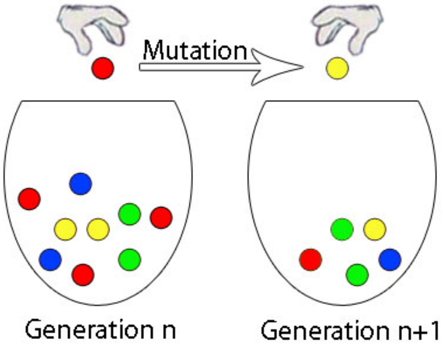


Figure 6: Wright-Fisher Urn Model with mutations.

For the coalescent process this means that if a mutation has occurred it is no longer possible to track that lineage further back in time. If we try to track an allele, say a green allele, backwards through the coalescent process, it is possible to track it as long as it stays green, but if the allele mutates to a red one it is no longer possible to track it further back in time.

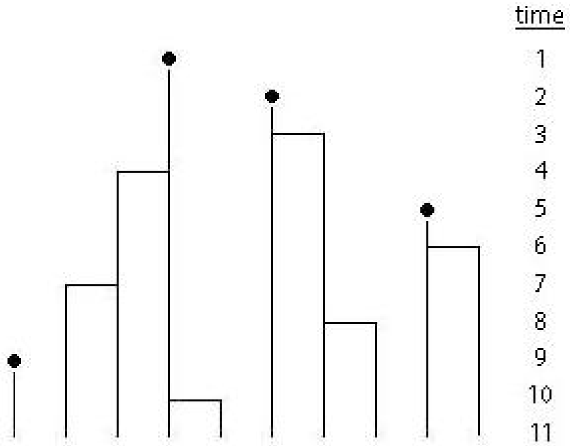


Figure 7: Coalescent process, the dots represents a mutation event.

The figure above illustrates the genealogical gene tree for a coalescent process with mutation where time is measured in coalescent and mutation events. As indicated by the dots the trace of an allele disappears when a mutation occurs.

4 Ewens Sampling Formula

As mentioned earlier there can be an almost infinite number of different alleles. Say a gene consists of 500 nucleotides, then the number of possible DNA sequences on that gene is $4^{500} = 10^{301}$. In other words, when an allele mutates it will almost certainly mutate to a new type never seen before, and therefore it is reasonable to assume that the number of alleles is infinite. In this section we will describe the coalescent process with mutation. This will be described with two different models, namely Hoppe's urn model and the Chinese restaurant process. From Hoppe's urn model we will derive the Ewens sampling formula, which gives the distribution for the number of different alleles in a sample. An important result in this section is the estimation of the mutation rate μ or equivalently the scaled mutation rate θ.

4.1 Hoppe's Urn Model

If we have a sample from the population containing k lineages the probability of a coalescent event in one time step is

$$\approx \binom{k}{2} \cdot \frac{1}{2N} = \frac{k(k-1)}{2} \cdot \frac{1}{2N}.$$

The probability to see more than one mutation at any one time step is negligible and therefore the probability that a mutation is seen instead of a coalescent is

$$\approx k\mu$$

where μ is the probability that an individual mutates. To make sure that coalescence and mutation occur with the same rate we speed up the system by running time at a rate of $2N$. The rate for coalescence is then $k(k-1)/2$ and the rate for mutation is $k\theta/2$, where $\theta = 4N\mu$. θ is called the scaled mutation rate.

Hoppe's urn model can be explained as follows. At time zero ($t = 0$) we have an urn containing a black ball with mass θ. At $t = 1$ we pick a ball from the urn which of course is the black ball and then we put back the black ball together with a colored ball with mass 1 in the urn. Also at time $t = 2$ we pick a ball from the urn, but now the probability to pick the black ball is

$$\frac{\theta}{\theta + 1}$$

and the probability to pick the colored ball is

$$\frac{1}{\theta + 1}.$$

If the black ball is picked up we again put back the black ball together with a new ball of a different color, but if the colored ball is picked we put a new ball with the same color in the urn so that we have two balls of the same color. All colored balls have mass 1 and it is assumed to be an infinite number of colors available. At time $t = k$ there are k colored balls in the urn plus the black one, so the probability to pick the black ball is

$$\frac{\theta}{\theta + k}$$

and the probability to pick a colored ball is

$$\frac{k}{\theta + k}.$$

In this model the event that a black ball is picked represents a mutation and the event that a colored ball is picked represents a coalescence. Thus we get the same coalescent process as illustrated in Figure 7 above.

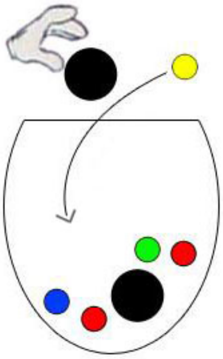


Figure 8: Hoppe's urn model.

We can track an individual backwards in time to find out where he originated from, but if a mutation has taken place we are not able to look further back, i.e. a mutation can be looked at as a completely new individual in the model. If we go backwards from time $k+1$ to time k in Hoppe's urn model we get a probability to loose track of an individual because of mutation equal to

$$\frac{\theta}{\theta + k}$$

and a probability to have a coalescent event equal to

$$\frac{k}{\theta + k}.$$

This is the same probabilities as we get when running the coalescent process with scaled time. Suppose there are $k+1$ lineages in the coalescent process, then the rate at which we will see a mutation is $(k+1)\theta/2$ and the rate for a coalescence event is $(k+1)k/2$. These event rates and the fact that all coalescent events have equal probability gives that

Theorem 4. *The genealogical relationship between k individuals in the coalescent process can be simulated by running Hoppe's urn model k time steps.*

4.2 Estimation of θ

Let K_n be the number of different alleles in a sample of size n. Further let $\eta_i = 1$ if the ith ball in Hoppe's urn model is a new color. Then it follows from Hoppe's urn model that $K_n = \eta_1 + \cdots + \eta_n$ and that

$$\eta_1, \ldots, \eta_n \text{ are independent with } P(\eta_i = 1) = \frac{\theta}{\theta + i - 1}.$$

Due to this independence we can compute the asymptotic behavior of the expected value and variance of K_n, which is then used to estimate θ.

Theorem 5. *Suppose θ is fixed in all generations. Then, as the sample size n goes to infinity*

$$EK_n \sim \theta \ln(n) \text{ and } Var(K_n) \sim \theta \ln(n)$$
$$\text{where } a_n \sim b_n \text{ means that } \frac{a_n}{b_n} \to 1 \text{ when } n \to \infty.$$

Proof. We start with proving the expected value.

$$EK_n = \sum_{i=1}^{n} E\left[\eta_i\right] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}.$$

Remember the Riemann sum approximation of an integral which gives

$$\sum_{i=1}^{n} \frac{1}{\theta + (i-1)} \approx \int_{\theta}^{n+\theta} \frac{1}{x} dx = \ln(n+\theta) - \ln(\theta) \sim \ln(n)$$

when $n \to \infty$. This gives that

$$EK_n = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \ln(n) \text{ as } n \to \infty.$$

Since all η_i are independent we get the variance of K_n as

$$\text{Var}(K_n) = \sum_{i=1}^{n} \text{Var}(\eta_i) = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}\left(1 - \frac{\theta}{\theta + i - 1}\right) = \sum_{i=1}^{n} \frac{\theta(i-1)}{(\theta + i - 1)^2}.$$

Now, we see that

$$\frac{i-1}{\theta + i - 1} = \frac{1}{\theta/(i-1) + 1} \to 1.$$

With this result and the Riemann sum approximation we get that

$$\text{Var}(K_n) \sim \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \ln(n).$$

□

Corollary 6. *By using the results from Theorem 5 we can calculate an asymptotically normal estimator of the mutation rate θ as*

$$\frac{K_n}{\ln(n)} \approx \theta.$$

The asymptotic standard deviation of the estimate is of order $1/\sqrt{\ln n}$ which is quite large. This means that if the true value of $\theta = 1$ and we want an estimate of θ with a standard error of 0.1 we need a sample of size $e^{100} \approx 2.688 \cdot 10^{43}$. However, there is no other way to estimate θ. K_n actually is a sufficient statistic for θ, i.e. K_n contains all useful information from the sample needed to estimate θ. This we will show with the help of the Chinese restaurant process later on.

4.3 Ewens Sampling Formula

In the previous section we talked about the asymptotic behavior of the number of different alleles. Now we will describe the distribution of the entire sample, which is given by the Ewens sampling formula.

Theorem 7 (Ewens sampling formula)**.** *Let a_i be the number of alleles present i times in the sample. When the scaled mutation rate is $\theta = 4N\mu$ the sample distribution is given by*

$$P_\theta(a_1, a_2, ..., a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}$$

where $\theta_{(n)} = \theta(\theta+1)(\theta+2)\cdots(\theta+n-1)$.

Proof. According to Theorem 4 it suffices to show that the distribution of the colors in Hoppe's urn at time n is given by the Ewens sampling formula. This is shown with induction. For $n = 1$, $(a_1, a_2, ..., a_n)$ equals $(1, 0, 0, ..., 0)$ and

$$P_\theta(a_1, a_2, \ldots, a_n) = \frac{1!}{\theta}\frac{\theta^1}{1!} = 1.$$

If we only have one allele in the sample the probability that the sample will have the distribution $(1, 0, \ldots, 0)$ is one. At time n we have the distribution $(a_1, a_2, \ldots, a_n)$. Let $a = (a_1, \ldots, a_n)$, and let $\bar{a} = (\bar{a}_1, \ldots, \bar{a}_n)$ be the distribution at the previous time step, then we get two different cases.

Case 1. $\bar{a}_1 = a_1 - 1$, i.e. the black ball is chosen and a new color is added into the urn. The number of one-colored groups increases with one. The transition probability for getting from $\bar{a}$ to a is,

$$p(\bar{a}, a) = \frac{\theta}{\theta + n - 1}.$$

We also have that

$$\frac{P_\theta(a)}{P_\theta(\bar{a})} = \frac{\frac{n!}{\theta_{(n)}}\prod_{j=1}^{n}\frac{(\theta/j)^{a_j}}{a_j!}}{\frac{(n-1)!}{\theta_{(n-1)}}\prod_{j=1}^{n}\frac{(\theta/j)^{\bar{a}_j}}{\bar{a}_j!}} = $$
$$= \{\bar{a}_1 = a_1 - 1\} = \frac{n}{\theta+n-1}\cdot\frac{\theta}{a_1}.$$

Case 2. One of the colored balls is chosen and a new one of the same color is added into the urn. In other words this means that for some $1 \le j < n$ we have that $a_j = \bar{a}_j - 1$ and $a_{j+1} = \bar{a}_{j+1} + 1$. The transition probability this time is

$$p(\bar{a}, a) = \frac{j\bar{a}_j}{\theta+n-1}.$$

We also have that

$$\frac{P_\theta(a)}{P_\theta(\bar{a})} = \frac{n}{\theta+n-1}\cdot\frac{\prod_{j=1}^{n}\frac{(\theta/j)^{a_j}}{a_j!}}{\prod_{j=1}^{n}\frac{(\theta/j)^{\bar{a}_j}}{\bar{a}_j!}} =$$
$$= \left\{\begin{array}{c}\bar{a}_j = a_j + 1\\ \bar{a}_{j+1} = a_{j+1} - 1\end{array}\right\} = \frac{n}{\theta+n-1}\cdot\frac{j\bar{a}_j}{(j+1)a_{j+1}}.$$

To complete the proof we see that

$$\sum_{\bar{a}}\frac{P_\theta(\bar{a})}{P_\theta(a)}p(\bar{a},a) = \frac{\theta+n-1}{n}\cdot\frac{a_1}{\theta}\cdot\frac{\theta}{\theta+n-1} +$$
$$+\sum_{j=1}^{n-1}\frac{\theta+n-1}{n}\cdot\frac{(j+1)a_{j+1}}{j\bar{a}_j}\cdot\frac{j\bar{a}_j}{\theta+n-1} =$$
$$= \frac{a_1}{n} + \sum_{j=1}^{n-1}\frac{(j+1)a_{j+1}}{n} =$$
$$= \frac{1}{n}(a_1 + 2a_2 + \cdots + na_n) = \frac{1}{n}\cdot n = 1.$$

Rearranging the above gives that

$$\sum_{\bar{a}} P_\theta(\bar{a})\cdot p(\bar{a},a) = P_\theta(a).$$

Since the distribution of Hoppe's urn also satisfies this recursion with the same initial condition the two must be equal according to Theorem 4. □

4.4 The Chinese Restaurant Process

The same coalescent process as described with Hoppe's urn can also be described with the Chinese restaurant process. Consider a restaurant with an infinite number of tables labeled 1,2,3,... . The first person arriving to the restaurant will sit down at the first table, table 1. The second person that arrives will choose to sit down at the first unoccupied table with probability

$$\frac{\theta}{\theta+1}$$

and at the occupied table with probability

$$\frac{1}{\theta+1}.$$

When the nth person arrives at the restaurant he will choose to sit at an unoccupied table with probability

$$\frac{\theta}{\theta+n-1}$$

and at occupied table number i with probability

$$\frac{c_i}{\theta+n-1}$$

where c_i ='the number of persons at table i'. That a person chooses to sit down at a new unoccupied table corresponds to a mutation and that he chooses to sit at one of the occupied tables corresponds to a coalescent event.

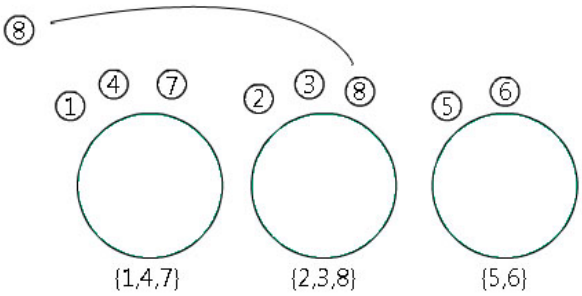


Figure 9: The Chinese restaurant process.

One of the differences between Hoppe's urn model and the Chinese restaurant process is that in the later one knows the exact permutation of the sample. Let Π_n be the permutation when there are n individuals. A property when using this notation is that given a permutation, the path to it is unique, i.e. in the example given in the figure above we know just by knowing the permutation that person 2 is a mutation from person one and that person 3 is a descendant from person 2 and so on. The following is a result from this property.

Theorem 8. *If π is a permutation with k cycles (i.e. k tables) then*

$$P_\theta(\Pi_n = \pi) = \frac{\theta^k}{\theta_{(n)}}.$$

Note that when $\theta = 1$, $P_\theta(\Pi_n = \pi) = 1/n!$. That is, all permutations are equally likely.

Proof. In the example given in the figure above we have, if the probabilities are as given, that

$$\begin{aligned}
&P_\theta(\Pi_8 = \{(741)(832)(65)\}) = \\
&= \frac{\theta}{\theta} \cdot \frac{\theta}{\theta+1} \cdot \frac{1}{\theta+2} \cdot \frac{1}{\theta+3} \cdot \frac{\theta}{\theta+4} \frac{1}{\theta+5} \cdot \frac{1}{\theta+6} \cdot \frac{1}{\theta+7} = \\
&= \frac{\theta^3}{\theta_8}.
\end{aligned}$$

Now, if a permutation of $\{1, 2, \dots, n\}$ has k cycles the numerator is always θ^k and the denominator is always $\theta_{(n)}$. □

Now, let $|S_n^k|$ be the number of permutations from {1,2,. . . ,n} with k cycles, i.e. $|S_n^k|$ is the number of ways we can get k groups from a sample of size n. $|S_n^k|$ is called the Stirling numbers of the second kind. The Stirling numbers of the second kind satisfy the relationship

$$|S_n^k| = (n-1)|S_{n-1}^k| + |S_{n-1}^{k-1}|.$$

In words this means that we can construct a $\pi \in |S_n^k|$ from a member of $|S_{n-1}^{k-1}|$ by adding $\{n\}$ as a new cycle, or from a $\sigma \in |S_{n-1}^k|$ by picking an integer $1 \le j \le n-1$ and setting $\pi(j) = n$ and $\pi(n) = \sigma(j)$.

Furthermore, let K_n just as before be the number of different alleles (number of groups or tables) in a sample of size n. Now, it follows that

Lemma 9. $P_\theta(K_n = k) = \frac{\theta^k}{\theta_{(n)}} \cdot |S_n^k|$.

It can be shown by calculating the Ewens sampling formula and conditioning on K_n that

Theorem 10. *K_n is a sufficient statistic for estimating θ.*

Proof.

$$\begin{aligned} P_\theta(a_1, a_2, \ldots, a_n) &= \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!} = \\ &= \frac{n!}{\theta_{(n)}} \cdot \theta^k \prod_{j=1}^{n} \frac{(1/j)^{a_j}}{a_j!} = \\ &= n! \frac{P_\theta(K_n = k)}{|S_n^k|} \prod_{j=1}^{n} \frac{(1/j)^{a_j}}{a_j!}. \end{aligned}$$

Conditioning on K_n gives

$$P_\theta(a_1, a_2, \ldots, a_n | K_n = k) = \frac{P_\theta(a_1, a_2, \ldots, a_n)}{P_\theta(K_n = k)} = \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \frac{(1/j)^{a_j}}{a_j!}.$$

Since this conditional distribution does not depend on θ, K_n is a sufficient statistic for estimating θ. □

4.5 Maximum-likelihood estimation of θ

Earlier we gave an asymptotic estimate of θ. In this section we will derive the maximum-likelihood estimator for θ based on K_n. The two estimates are not the same but they are asymptotically equal when the sample size n goes to infinity.

$$L_n(\theta, k) = \frac{\theta^k}{\theta_{(n)}} \cdot |S_n^k|.$$

This is the likelihood of observing k when the true value is θ. To find the value for θ that maximizes the probability to observe k we have to take the derivative of the likelihood with respect to θ.

$$\frac{\partial}{\partial\theta} L_n(\theta, k) = |S_n^k| \frac{k\theta^{k-1}\theta_{(n)} - \theta^k \theta'_{(n)}}{(\theta_{(n)})^2} = \frac{\theta^k |S_n^k|}{\theta_{(n)}} \left(\frac{k}{\theta} - \frac{\theta'_{(n)}}{\theta_{(n)}} \right).$$

If we set this to zero and solve it for k we get that

$$k = \theta \cdot \frac{\theta'_{(n)}}{\theta_{(n)}} = \theta \cdot \frac{d}{d\theta} \ln(\theta_{(n)}).$$

Remember that $\theta_{(n)} = \theta(\theta + 1) \cdots (\theta + n - 1)$, so

$$\frac{d}{d\theta} \ln(\theta_{(n)}) = \frac{d}{d\theta} \sum_{i=1}^{n} \ln(\theta + i - 1) = \sum_{i=1}^{n} \frac{1}{\theta + i - 1}$$

and we get that

$$k = \theta\left(\frac{1}{\theta} + \frac{1}{\theta+1} + \cdots + \frac{1}{\theta+n-1}\right) = EK_n.$$

This means that the maximum likelihood estimator is the θ that makes the expected number of different alleles equal to the observed number, i.e. the θ that solves $EK_n = k$. When n goes to infinity this is the same asymptotic estimate as we got from Corollary 6.

$$\hat{\theta} \approx \frac{k}{\ln(n)} \quad \text{, when } n \text{ is large.}$$

According to the theory of maximum likelihood estimation, $E\hat{\theta}$ is asymptotically equal to θ, and $\mathrm{Var}(\hat{\theta}) = 1/I(\hat{\theta})$ where $I(\theta)$ is the Fisher information.

$$I(\theta) = E\left(\frac{\partial}{\partial\theta}\ln(L_n(\theta,k))\right)^2$$

In our case we get that

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln(L_n(\theta,k)) &= \frac{\partial}{\partial\theta}\ln\left(\frac{\theta^k|S_n^k|}{\theta_{(n)}}\right)\\ &= \frac{k}{\theta} - \frac{\partial}{\partial\theta}\ln(\theta_{(n)}) = \frac{1}{\theta}\left(k - \sum_{i=1}^{n}\frac{\theta}{\theta+i-1}\right).\end{aligned}$$

Since $EK_n = \sum_{i=1}^{n}\frac{\theta}{\theta+i-1}$ it follows from the definition of the variance that

$$I(\theta) = \frac{1}{\theta^2}E[k - EK_n]^2 = \frac{1}{\theta^2}\mathrm{Var}(K_n).$$

Combining this with previous result about the asymptotic behavior of the variance we get that

$$\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{\mathrm{Var}(K_n)} \to 0, \text{ as } n \to \infty.$$

This asymptotic result tell us that $\hat{\theta}$ is a consistent estimator of θ, but it converge rather slow since $\mathrm{Var}(\hat{\theta}) \sim \theta/\ln(n)$.

5 Simulation

In this section we will simulate data from the Ewens sampling formula (Theorem 7) with the help of Hoppe's urn model. It is possible to simulate data with the Chinese restaurant process as well but it demands more computer power and compared to Hoppe's urn it does not give any extra information that is useful for us. When we make the simulations the aim is to estimate the scaled mutation rate θ and see how accurate the estimation is. As we previously have mentioned and will see in this section the convergence rate of $EK_n/\ln(n) \rightarrow \theta$ is quite slow. This means that n has to be *very* large for good estimation of θ. Unfortunately, as discussed in the previous section, the only way to estimate θ is with $K_n/\ln(n)$. An interesting thing that will be looked at is how the number of alleles (groups) depends on the mutation rate θ. We will also see how the estimation of θ depends on the real θ and the sample size n.

Remember the relationship between the mutation rate μ and the scaled mutation rate θ

$$\mu = \frac{\theta}{4N}.$$

The mutation rate in human mitochondrial DNA is estimated to approximately $2.7 \cdot 10^{-5}$ [7]. Instead of using the actual population size when making the calculations an effective population size is used. This is due to its more accurate representation of the genetic variation among humans. The effective population size for humans has in many studies been shown to be $\sim 10\,000$ [8]. So to give an example, if we have an effective population size of 10 000, the scaled mutation rate $\theta = 0.27 \cdot 4 = 1.08$.

Before we start simulating let us look at the convergence rate for the asymptotic results. As proved in the previous section $EK_n \sim \theta \cdot \ln(n)$ when $n \rightarrow \infty$ or equivalent

$$\frac{EK_n}{\theta \cdot \ln(n)} \rightarrow 1 \quad \text{and} \quad \frac{EK_n}{\ln(n)} \rightarrow \theta$$

when $n \rightarrow \infty$. The convergence rate will be illustrated in the figure below where we let n go from 1 to 100 000 in

$$\frac{EK_n}{\theta \cdot \ln(n)} = \frac{1}{\ln(n)} \cdot \sum_{i=1}^{n} \frac{1}{\theta + i - 1}$$

for different values of θ. As n increase this should converge to one.

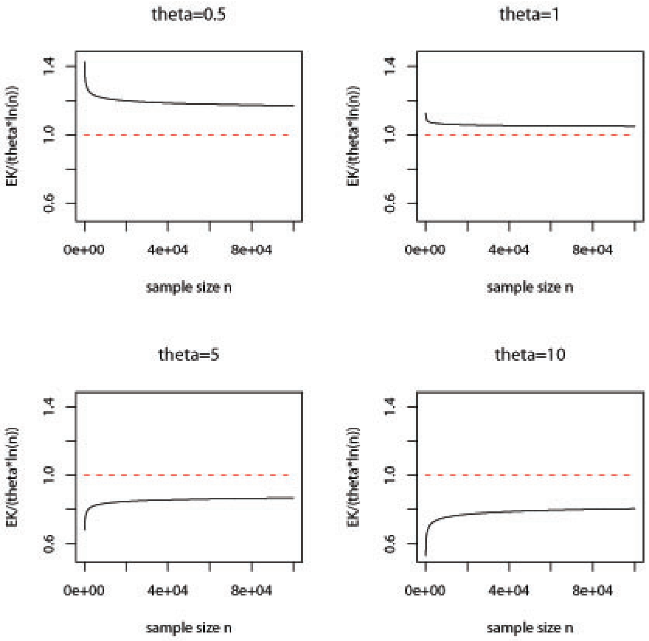


Figure 10: Plots of $EK_n/(\theta \ln(n))$ for $\theta = 0.5, 1, 5, 10$.

From the plots we see that $EK_n/(\theta \ln(n))$ converges quickly for small n:s but more slowly as n increases. The plots also show that $\theta = 1$ gives the best convergence rate and that the convergence rate decreases as θ increases (for $0 < \theta < 1$ the convergence rate increase as θ increases). The expected values when $n = 100\,000$ are given in the following table.

θ	0.5	1	5	10
$EK_n/\ln(n) = \theta^*$	0.8670559	1.050137	4.345921	8.044235
$EK_n/(\theta \ln(n))$	1.170548	1.050137	0.8691842	0.8044235

The table shows that the estimated values of θ differ from the true values quite much when $n = 100\,000$. From the table we also see that when the true θ increases the expected estimation of θ, $EK_n/\ln(n)$, is further from the true value. This is due to the slower convergence rate for high values of θ and thus a larger sample size n is needed for more accurate estimation when θ is large. The larger sample size we have the better the estimate will be but as seen in the plots θ^* close up on the true θ slower and slower as n increases. To give an example, if $\theta = 1$ and $n = 200000$ we get $\theta^* = 1.047289$ which yields a difference of only 0.002847574 compared to θ^* at $n = 100000$.

With this slow convergence in mind we will make simulations from the Ewens sampling formula with the help of Hoppe's urn model. From these simulations we will estimate θ. The estimate from the sample is given by

$$\hat{\theta} = \frac{K_n}{\ln(n)}.$$

To get a good estimate of θ we know that n has to be large. But just to illustrate the variance within a sample, i.e. the variation among groups we simulate a very small sample. If we are not interested in estimating θ the small sample size does not cause any trouble, and it is hard to get an overview of the variation if the sample size is too large. So let us simulate a sample of size 20 from Hoppe's urn model where $\theta = 2$, and illustrate the variance among groups with the following genealogical gene tree.

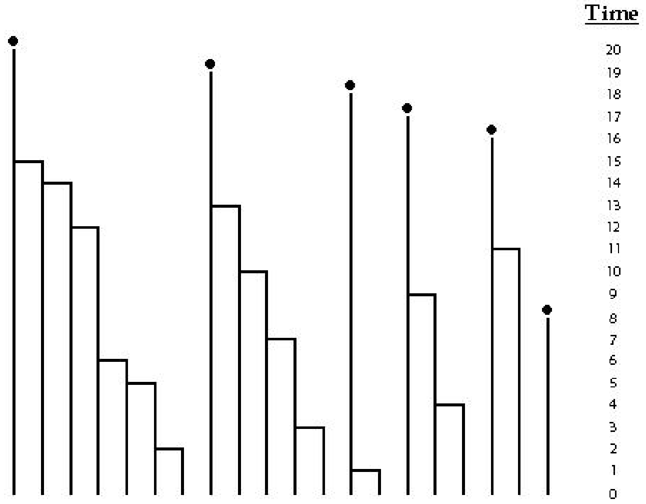


Figure 11: Genealogical gene tree of simulations. $\theta = 2$, $n = 20$ and time is measured in events.

As the genealogical tree shows the simulation gives us six groups, and five mutation events take place before the first collision. The expected number of groups is $\sum_{i=1}^{n} \theta/(\theta + i - 1) \approx 5.3$. If we estimate θ from our simulation we get $\hat{\theta} = 6/\ln(20) = 2.0028$ which seems to be a pretty good estimate considering that the true θ equals two. But the standard deviation for $\hat{\theta}$ is approximately $2/\ln(20) \approx 0.82$. So the estimate has in this case quite high variance and the sample size is to small for us too trust the estimation.

We will see how the number of different alleles in the simulated sample depend on the scaled mutation rate θ. The number of alleles in the sample, or the number of different colors in Hoppe's urn, is given by K_n. According to Hoppe's urn model K_n should increase when θ does, which comes from that $\theta/(\theta + k)$ is the probability for a new allele. Let us examine what happens when K_n is a function of θ.

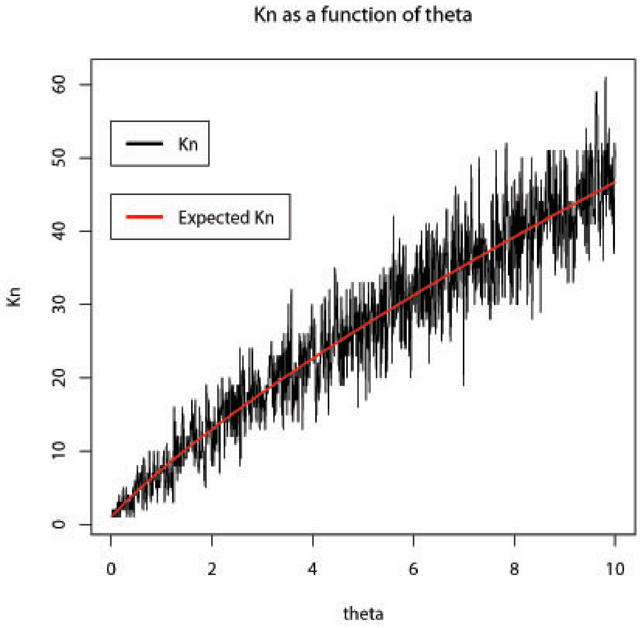


Figure 12: Number of alleles K_n as a function of θ, sample size $n = 1\,000$.

The plot clearly indicates that the number of alleles depend on θ, it shows that K_n increase with θ. The expected value of K_n, EK_n, which is plotted in red, follow the simulated values well and therefore gives an indication that the simulations are accurate. The plot also give example of the high variance in the simulations which seems to increase with θ just as expected $(\mathrm{Var}(K_n) \sim \theta \cdot \ln(n))$.

To get an idea of how $\hat{\theta}$ differs from the real θ we will make 1000 simulations, calculate $\hat{\theta}$ for each one of the 1 000 simulations and draw an histogram of the result. $\hat{\theta}$ will differ from the true θ more or less depending on what the true value is so we will see what happens when we change θ, the sample size n is set to 10 000 in the simulations.

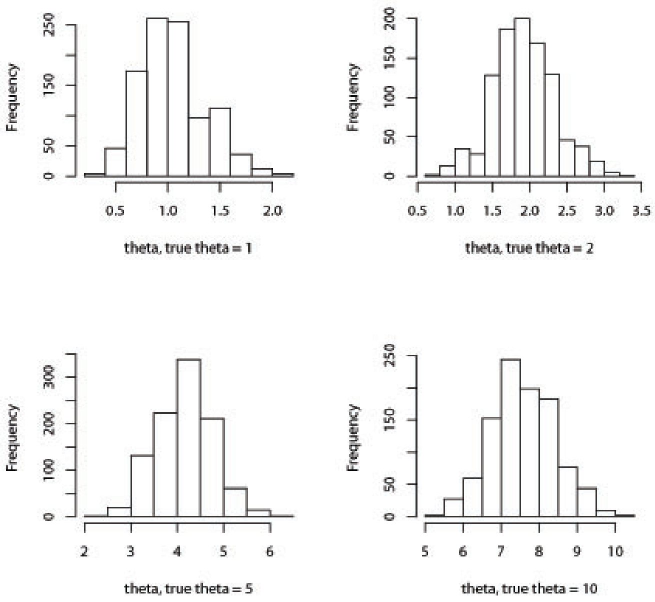


Figure 13: Histograms of $\hat{\theta}$ for $\theta = 1, 2, 5, 10$.

θ	1	2	5	10
mean($\hat{\theta}$)	1.058593	1.919473	4.165427	7.579199

The mean values of $\hat{\theta}$ is the center of mass in the histograms and shows approximately how big the difference is between the estimated and the true θ. As we see the estimated value of θ differs more and more from the true value as θ increases (when $\theta \geq 1$). This is due to that the variance of $\hat{\theta}$ increases with θ, $\mathrm{Var}(\hat{\theta}) \sim \theta / \ln(n)$. For fixed variance the sample size n increases exponentially with θ, e.g. say the true $\theta = 10$ and we want to estimate θ with a standard deviation of 0.5, then we would need a sample size equal to $e^{10/0.5^2} = e^{40} \approx 2.35 \cdot 10^{17}$. If $\theta = 1$ a sample size of $e^4 \approx 55$ had resulted in a standard deviation equal to 0.5. As we have seen earlier the convergence rate is slower at larger values of θ and therefore we need a larger sample size to get the same variance.

To more clearly see the relationship between $\hat{\theta}$ and the true θ the simulated $\hat{\theta}$ are plotted against the true θ.

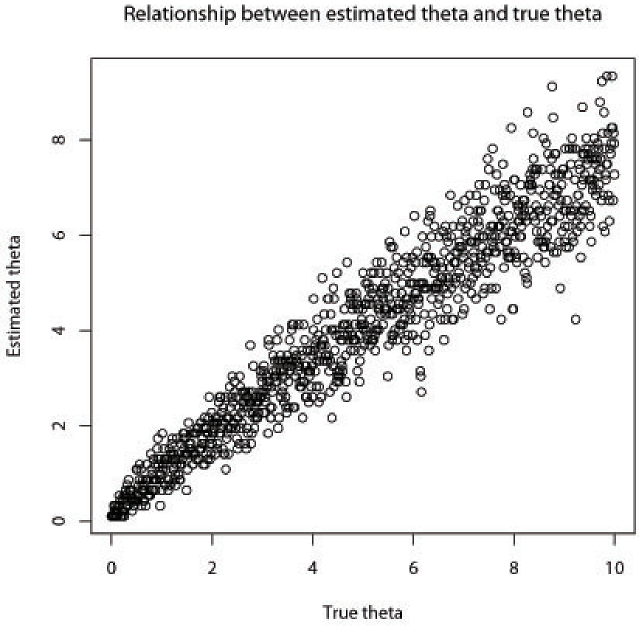


Figure 14: $\hat{\theta}$ from simulations with $n = 10\,000$ as a function of the true θ.

This scatter plot suggests, with its cone like structure, that the variation of $\hat{\theta}$ increase when θ do, just as we have seen before. The figure also show that $\hat{\theta}$ differs more and more from the true value as θ increase.

We have earlier discussed how $\hat{\theta}$ depend on n. Now we will illustrate how this dependence between $\hat{\theta}$ and n looks like. We will simulate samples of different sizes for a fixed value of θ and calculate $\hat{\theta}$ for them. It would according to the structure of $\hat{\theta}$ be expected to see that $\hat{\theta} \to \theta$ as n increases. With the following figure we see what the simulations say about this.

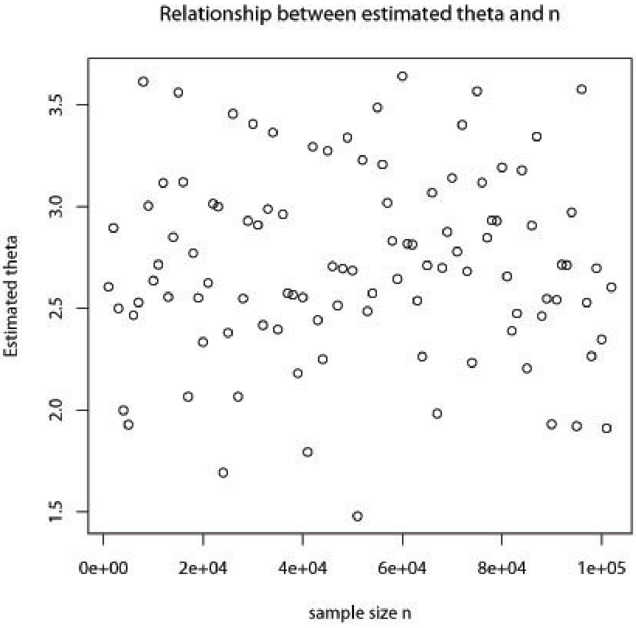


Figure 15: $\hat{\theta}$ as a function of the sample size n, true $\theta = 3$.

It is not clear in the figure that $\hat{\theta} \to \theta$ as n increases as we could expect to see. The reason for this is that we need exponentially higher sample size to see any noticeable difference in variance. In our case we use a sample size from 1 000 to 100 000 and a $\theta = 3$ which yields a variance from $3/\ln(100\,000)$ to $3/\ln(1\,000)$.

To round off we conclude that $K_n/\ln(n)$ is a poor estimate of θ but yet a function of the sufficient statistic K_n. The convergence rate is very slow which makes the estimate of θ smaller than the real θ. The sample size n is exponentially dependent on the variance and the scaled mutation rate θ, i.e. $e^{\theta/\mathrm{Var}(\hat{\theta})} \sim n$. $\theta = 1$ gives the best convergence rate and thus the simulated estimate with the lowest mean squared error for any given n. The increased uncertainty in the estimate when θ increases can be explained mostly by the increasing bias, even if, as we have seen, the variance increases slightly as well. The maximum-likelihood estimate of θ is probably approximately unbiased, and would probably yield a better estimate than $K_n/\ln(n)$. But it is more difficult to calculate the maximum-likelihood estimate due to the nonlinear likelihood equations.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Andreas Nordvall Lagerås at Stockholm University, Division of Mathematical Statistics, for his guidance and support which have been of great value for my work.

References

[1] Genetic Inheritance - United Leokodystrophy Foundation. **Genetic Inheritance**. Page last updated 8 February 2007.
http://www.ulf.org/patients/inheritance.html.

[2] Durrett, R., 2002. **Probability models for DNA sequence Evolution** Springer-Verlag, New York. ISBN: 0-387-95435-X.

[3] Cancer Research UK. **Double Helix** (picture). Page last updated 23 August 2007.
http://info.cancerresearchuk.org/images/gpimages/ys_DNA_4and5.

[4] Biology Online Dictionary. **Gene**. Page last updated 17 June 2008.
http://www.biology-online.org/dictionary/Gene.

[5] http://www.wikipedia.org. **Optional Stopping Theorem**. Page last updated 16 April 2009.
http://en.wikipedia.org/wiki/Optional_stopping_theorem.

[6] Charlesworth, B., March 2009. **Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation**. Nature Reviews Genetics 10, 195-205, doi:10.1038/nrg2526.

[7] http://www.wikipedia.org. **Mutation rate**. [Online] November 2008.
http://en.wikipedia.org/wiki/Mutation_rate.

[8] Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins et al., 2000. **DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22**. Proc. Natl. Acad. Sci. USA 97: 11354-11358.