Some estimators of molecular polymorphism and their asymptotic behavior *

Samara F. Kiihl[1], Hildete P. Pinheiro[1], Aluísio Pinheiro[1] and Sérgio F. dos Reis[2]
[1] *Department of Statistics* University of Campinas, Brazil
[2] *Departament of Parasitology* University of Campinas, Brazil

Abstract

Important aspects of population evolution have been investigated using nucleotide sequences. The quantity θ, representing the mean number of mutations per gene per generation, is an essential parameter in population evolution studies, since it determines the degree of polymorphism in a locus. Inferences about the evolution of a population are measured by the accuracy of estimation of this parameter. In this article we present various methods of estimation of θ, analytical studies of their asymptotic distributions as well as comparisons of the distribution's behavior of these estimators through simulations.

Keywords and phrases: asymptotic theory, discrete distribution, maximum likelihood estimate, molecular polymorphism, population evolution.

1 Introduction

The mean number of mutations per gene per generation, represented by θ, is an essential parameter in population evolution studies, since it determines the degree of polymorphism in a locus. Inferences about the evolution of a population are measured by the accuracy of estimation of this parameter. Hypotheses tests about the evolution of natural populations can be done using an estimator of θ and its distribution. One can think of many potential applications of tests like this. Ramos-Arroyo et al. (2005) investigated the incidence and mutation

*Corresponding author: *E-mail address*: hildete@ime.unicamp.br (H. Pinheiro)

rates of Huntington's disease in Spain and concluded that the incidence and mutation rate of the disease are 2-3 times higher than previously reported. In human immunodeficiency virus (HIV) studies, understanding the evolution of HIV is crucial for reconstructing its origin. This can help develop new ways to treat or vaccinate against HIV (Rambaut et al., 2004). Also, in studies of association between disease and genotypes (Whittemore, 2004), it is important to estimate the degree of polymorphism to see how it can affect a certain disease. Comparisons of groups with respect to evolution and/or mutation rates can be made by the comparisons of one of these estimators of θ for each group. Smith et al. (2002) show significant variation in substitution rates within the non-coding part of the human genome using human-chimpanzee pairwise comparisons. Crow (1997) discusses about the higher human mutation rate for base substitutions in males compared to females. Huang et al. (1997) study sex differences in mutation rate in primates.

There are three estimators of θ most used in the literature: $\mathcal{T}_1$, based on the number of segregating sites (S), which will be discussed in Section 2, $\mathcal{T}_2$, the mean number of pairwise nucleotide differences discussed in Section 3 and $\mathcal{T}_3$, based on the number of *singletons* ($S^\star$), presented in Section 4. All of these estimators are unbiased or asymptotically unbiased for θ under the assumption of the Wright-Fisher model, i.e., constant effective population size and large number of sites in the sequences, in such a way that every mutation occurs in a new site and all mutations are selectively neutral (Hartl and Clark, 2007). The estimation of θ using these estimators is computationally simple, but they may have a big variance. These estimators are largely used in the literature for inferences about population evolution. Many authors (e.g., Tajima, 1989; Fu and Li, 1993a) developed statistical tests of the neutral mutation hypothesis based on DNA polymorphism using as test statistics linear combination of these estimators ($\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$). The distributions of these test statistics for neutral mutation are based on heuristic motivation and the critical values are obtained by computer simulations. In view of this, we study analytically the asymptotic behavior of each of these estimators of θ and perform some simulation studies as well.

A great attention will be given here to the maximum likelihood estimator (MLE) of θ. The MLE has been mentioned in the literature only to obtain the Cramer-Rao Lower Bound (CRLB) for its variance (Fu and Li, 1993b) to compare the efficiency of the estimators ($\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$). In Section 5 we present the MLE of θ, its method of computation and we showed that its asymptotic distribution is normal. The best asymptotic behavior of the MLE is illustrated

by the simulation studies in Section 6. We found that $\mathcal{T}_1$ and the MLE of θ have the same asymptotic distribution, i.e., they are both normal with the same mean and variance.

2 The number of segregating sites

Suppose we have a sample of five DNA sequences, from which 500 sites were sequenced. Table 1 presents 16 polymorphic or *segregating* sites, i.e., sites where nucleotides differ. Among the polymorphic sites, we have sites 3, 5, 6, 7, 8, 11, 12, 13 and 15 which are *singletons*, since there is only one nucleotide different from the others.

Table 1: Polymorphic sites in a sample of five genes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
b	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
c	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
d	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
e	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A
*	(6)	(6)	(4)	(7)	(4)	(4)	(4)	(4)	(6)	(6)	(4)	(4)	(4)	(6)	(4)	(6)

* : number of pairwise differences

For Table 1, the number of segregating sites: $S = 16$; the number of singletons: $S^\star = 9$ and; the number of pairwise differences in each site is shown on the last row of the table, with the mean number of pairwise differences: $\mathcal{T}_2 = (6 + 6 + 4 + 7 + \cdots + 4 + 6)/\binom{5}{2} = 79/10 = 7.9$.

Consider the sites of each of $2N$ gametes in a population. If a sample of size n of i gametes is chosen randomly, S is *the number of segregating sites in the sample*, i.e., the polymorphic sites in which nucleotides are different. According to Watterson (1975) and Tajima (1989), the moments of S can be found assuming that i) recombination due to crossing over is rare; ii) infinite number of sites; iii) mutations can occur in any site; iv) mutations occurring in different gametes are mutually independent; v) a gamete inherit all sites from its paternal gamete; vi) two mutations never occur on the same site; vii) for a population of effective size N (i.e., N diploid individuals), there are $2N$ gametes in the population. If ν is the rate of mutation in a gene and it is very small, assuming that the number of mutations follows a Poisson distribution with parameter ν and by using Taylor series expansions, we get that $E(S) = \theta a_n$, where $\theta = 4N\nu$ is our parameter of interest and a_n is given in (2.2). Here, θ is interpreted as the expected number of mutations in the population per gene per generation and Watterson (1975) introduced an estimator of θ, $\mathcal{T}_1$, and its

variance

$$\mathcal{T}_1 = \frac{S}{a_n} \quad \text{and} \quad Var(\mathcal{T}_1) = \frac{\theta}{a_n} + \frac{\theta^2 b_n}{a_n^2}, \tag{2.1}$$

where

$$a_n = \sum_{j=1}^{n-1} \frac{1}{j} \quad \text{and} \quad b_n = \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{2.2}$$

Here we will study analytically the asymptotic behavior of $\mathcal{T}_1$.

In our case, $Y_1, Y_2, \ldots$ are independent random variables and $S = V_n = Y_1 + \ldots Y_{n-1}$. Here, V_n is the number of segregating sites in a sample of n gametes and Y_j, $j = 1, 2, \ldots n-1$, is the number of new mutations occurring in the previews generations when there were exactly $j+1$ distinct ancestors present. Using the fact that

$$P(Y_j = n) \quad = \quad \left(\frac{1}{1+\frac{\theta}{j}}\right)\left(1 - \frac{1}{1+\frac{\theta}{j}}\right)^n, \quad n = 0, 1, 2, 3, \ldots, \tag{2.3}$$

Note that the characteristic function of Y_j is

$$\varphi_{Y_j}(t) = E(e^{itY_j}) = \frac{j}{j - \theta(e^{it} - 1)} \tag{2.4}$$

and

$$E(Y_j) = \frac{\theta}{j} \;\; ; \quad E(Y_j^2) = \frac{\theta}{j} + \frac{2\theta^2}{j^2}\,. \tag{2.5}$$

As $Y_1, Y_2, \ldots, Y_{n-1}$ are independent r.v. and a_n is given in (2.2), we can study the characteristic function of $\sqrt{a_n}(\mathcal{T}_1 - \theta)$, i.e.,

$$\varphi_{\sqrt{a_n}(\mathcal{T}_1-\theta)}(t) = e^{-it\sqrt{a_n}\theta} \prod_{j=1}^{n-1} \frac{j}{j + \theta(1 - e^{ita_n^{-1/2}})}.$$

By Taylor series expansions, we get

$$\begin{aligned} log(\varphi_{\sqrt{a_n}(\mathcal{T}_1-\theta)}(t)) \quad &\approx \quad -it\sqrt{a_n}\theta - \theta a_n \left[-ita_n^{-1/2} - \frac{i^2t^2a_n^{-1}}{2} - \frac{i^3t^3a_n^{-3/2}}{3!} - \ldots\right] \\ &= \quad -\frac{t^2\theta}{2} - \frac{it^3\theta}{6\sqrt{a_n}} + \frac{t^4\theta}{24a_n} + \ldots \qquad (2.6) \\ &\approx \quad -\frac{t^2}{2}\,\theta \;\text{ as }\; n \to \infty. \qquad (2.7) \end{aligned}$$

Therefore, as $n \to \infty$,

$$\sqrt{a_n}(\mathcal{T}_1 - \theta) \xrightarrow{D} N(0, \theta).$$

3 The number of pairwise differences

Note that the mean number of pairwise differences ($\mathcal{T}_2$) can be defined in terms of U-statistics (Pinheiro et al., 2008). Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})'$ and $\mathbf{X}_j = (X_{j1}, \ldots, X_{jm})'$ be random vectors representing DNA sequences i and j. So, X_{il} can assume values in the set $\{A, C, T, G\}$, where A represents Adenine, C, Cytosine, T, Thymine and G, Guanine.

$$\mathcal{T}_2 = \left(\sum_{1 \leq i < j \leq n} K_{ij} \right) \Big/ \binom{n}{2} = \left(\sum_{i<j}^{n} \sum_{l=1}^{m} \mathbb{I}(X_{il} \neq X_{jl}) \right) \Big/ \binom{n}{2} \ , \tag{3.8}$$

where K_{ij} is the number of pairwise differences between genes i and j, in sequences of length m; X_{il} is a random variable representing the nucleotide present at the i-th sequence in the l-th site and $\mathbb{I}(X_{il} \neq X_{jl}) = 1$, if $X_{il} \neq X_{jl}$ and 0 otherwise.

On the other hand, if one is interested in estimating a parameter, say $\mathcal{H}$, representing the proportion of sites where nucleotides differ, then we can write

$$\mathcal{H} = \sum_{i<j} \frac{1}{m} \sum_{l=1}^{m} P(X_{il} \neq X_{jl}) \Big/ \binom{n}{2} = 1 - \frac{1}{m} \sum_{l=1}^{m} \sum_{c=1}^{4} \Pi_{cl}^2,$$

where $\Pi_{cl} = P(X_{il} = c)$ is the probability of having category (nucleotide) c at site l ($c = 1, \ldots, 4$) and $\mathcal{H}$ is called in the literature the Hamming distance (Pinheiro et al., 2005). In this case an unbiased estimator of $\mathcal{H}$ is $\bar{D} = \mathcal{T}_2/m$ and it is a U-statistic of degree 2. Pinheiro et al. (2008) showed that $\bar{D}$ is asymptotically normal distributed, i.e.,

$$\sqrt{nm} \frac{(\bar{D} - \mathcal{H})}{2\sigma_1} \xrightarrow{D} N(0,1) \text{ as } n \to \infty \text{ (and either } m \text{ limited or } m \to \infty), \tag{3.9}$$

where

$$\begin{aligned} \sigma_1^2 \ \equiv \ & \frac{1}{m^2} \left\{ \sum_{l=1}^{m} \sum_{c=1}^{C} \Pi_{cl}(1 - \Pi_{cl})^3 \right. \\ & + \sum_{l \neq k=1}^{m} \sum_{c,d=1}^{C} (\Pi_{cl,dk} - \Pi_{cl}\Pi_{dk})(1 - \Pi_{cl})(1 - \Pi_{dk}) \\ & \left. - \sum_{l=1}^{m} \sum_{c \neq d=1}^{C} \Pi_{cl}\Pi_{dl}(1 - \Pi_{cl})(1 - \Pi_{dl}) \right\}. \end{aligned} \tag{3.10}$$

In view of this, we can say that the estimator $\mathcal{T}_2 = m\bar{D}$ is a function of $\bar{D}$ and has also an asymptotic normal distribution for large n. Also, as $\mathrm{E}(\mathcal{T}_2) = \theta$

(Tajima, 1989), $\theta = m\mathcal{H}$. Then,

$$\sqrt{n}\frac{(\mathcal{T}_2 - \theta)}{2\sigma_1^\star} \xrightarrow{D} N(0,1) \text{ as } n \to \infty\,, \tag{3.11}$$

where $\sigma_1^\star = \sqrt{m}\sigma_1$.

4 The number of singletons

The number of singletons, $\mathcal{S}^\star$, represents the number of sites in which appears a single individual (sequence) with a nucleotide different from the others and its moments were computed by Fu and Li (1993) using coalescence theory (Hein et al., 2005).

$$E(\mathcal{S}^\star) = \frac{n}{n-1}\,\theta \text{ and } Var(\mathcal{S}^\star) = \frac{n}{n-1}\theta + d_n\theta^2. \tag{4.12}$$

where

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right),$$
$$c_2 = 1, \quad c_n = 2[na_n - 2(n-1)]/[(n-1)(n-2)], \quad \text{for } n > 2, \tag{4.13}$$

and $a_{n+1} = \sum_{j=1}^{n} 1/j$.

By the method of moments, an unbiased estimator of θ is

$$\mathcal{T}_3 = \frac{(n-1)\mathcal{S}^\star}{n}\,. \tag{4.14}$$

Then,

$$Var(\mathcal{T}_3) = \left(\frac{n-1}{n}\right)\theta + \left(\frac{n-1}{n}\right)^2 d_n\theta^2\,. \tag{4.15}$$

Note that the number of singletons can be written as

$$\mathcal{S}^\star = \sum_{l=1}^{m} \mathbb{I}\left\{\sum_{1\le i<j\le n} \mathbb{I}(X_{il} \neq X_{jl}) = n-1\right\},$$

where m is the number of sites.

$$E(\mathcal{S}^\star) = \sum_{l=1}^{m} \underbrace{\mathrm{P}\left(\sum_{i<j} \mathbb{I}(X_{il} \neq X_{jl}) = n-1\right)}_{\mathcal{P}} = m\mathcal{P} \tag{4.16}$$

The probability of a site be singleton, $\mathcal{P}$, can be expressed as

$$\begin{aligned}\mathcal{P} &= n\{\pi_A[(\pi_C)^{n-1} + (\pi_G)^{n-1} + (\pi_T)^{n-1}] + \pi_C[(\pi_A)^{n-1} + (\pi_G)^{n-1} + (\pi_T)^{n-1}] \\ &+ \pi_G[(\pi_A)^{n-1} + (\pi_C)^{n-1} + (\pi_T)^{n-1}] + \pi_T[(\pi_A)^{n-1} + (\pi_G)^{n-1} + (\pi_C)^{n-1}]\},\end{aligned}$$

where $\pi_c = P(X_{il} = c)$ is the probability of having nucleotide c at site l ($c \in \{A, C, T, G\}$).

By (4.12) and (4.16),

$$\begin{aligned} E(\mathcal{S}^\star) &= m\mathcal{P} = \frac{n}{n-1}\,\theta \Rightarrow \mathcal{P} = \left(\frac{n}{n-1}\right)\frac{\theta}{m}\,. \\ \mathcal{P} &\to \frac{\theta}{m} \quad , \text{ when } \quad n \to \infty\,. \end{aligned}$$

$$P(\mathcal{S}^\star = k) = \binom{m}{k}\mathcal{P}^k(1-\mathcal{P})^{m-k} \to \binom{m}{k}\left(\frac{\theta}{m}\right)^k\left(1-\frac{\theta}{m}\right)^{m-k}\,, \text{ when } n \to \infty\,.$$

But, note that

$$\left(1-\frac{\theta}{m}\right)^m \approx e^{-\theta} \text{ when } m \to \infty\,,$$

$$\left(1-\frac{\theta}{m}\right)^{m-k} \approx \exp\left[(m-k)\left(-\frac{\theta}{m}\right)\right] \approx e^{-\theta} \text{ when } m \to \infty$$

and

$$\binom{m}{k} \approx \frac{m^k}{k!} \text{ when } m \to \infty\,.$$

So, when $m \to \infty$,

$$P(\mathcal{S}^\star = k) = \frac{\theta^k}{k!}e^{-\theta}\,.$$

In summary:

- The distribution of $\mathcal{S}^\star$ is Binomial$(m, \mathcal{P})$.
- When $n \to \infty$, the distribution of $\mathcal{S}^\star$ is Binomial$(m,\, \theta/m)$.
- When $n \to \infty$ and $m \to \infty$, in such a way that $n >> m$, the distribution of $\mathcal{S}^\star$ is Poisson(θ).

5 The maximum likelihood estimator of θ

Suppose that a sample of n genes is drawn from a population with random mating and under the assumption of neutral mutations, the process of evolution of the sequences sampled is entirely determined by the value of θ. This process is known as coalescence process (Wakeley, 2009).

Here it is assumed the Wright-Fisher model for the population and for convenience the time in which the sample is taken (external nodes) is considered the

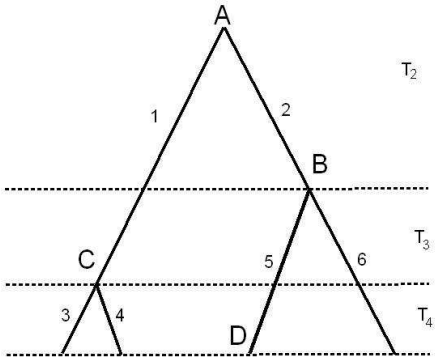

Figure 1: An example of topology of four sequences. The 1st, 2nd, 3rd and 4th branching events are A, B, C and D respectively.

n-th ramification event (Fu, 1994; Figure 1). As $T_i \sim \text{Exp}\left(\frac{i(i-1)}{4N}\right)$ represents the coalescence time between the $(i-1)$-th and i-th node,

$$f_{T_i}(t_i) = \frac{i(i-1)}{4N}\, e^{-\frac{i(i-1)}{4N} t_i}, \quad t_i = 0,\, 1, \ldots. \quad i = 2,\, 3, \ldots$$

Considering the genealogy of a random sample from a population with random mating, there are $n-1$ internal nodes in the tree enumerated from 2 to $n-1$ according to their order of occurrence (B and C in Figure 1). Therefore, between the $(i-1)$-th and i-th node there are exactly i branches, enumerated from 1 to i. The time since the $(i-1)$-th node until the i-th is the coalescence time T_i. Define Y_{ij} as the number of mutations in the j-th branch within the i existing branches during the coalescence time T_i. We assume that $Y_{ij} \sim \text{Poisson}(\nu t_i)$.

Let $Y_{i1}, Y_{i2}, \ldots Y_{ii}$ conditioned on t_i be a random sample with distribution Poisson(νt_i). Then, we know that $Y_i = \sum_{j=1}^{i} Y_{ij}$ is a sufficient statistic for νt_i and $Y_i \mid t_i \sim \text{Poisson}(\nu i t_i)$. Y_i is the total number of mutations occurred between the $(i-1)$-th and i-th node, i.e., during the coalescence time T_i.

It is reasonable to assume that the spatial distribution of a certain number of mutations between sites of a gene is independent of θ, even though the number of mutations depends on it. For example, one can assume that the spatial distribution is uniform among sites, or any other distribution which does not

depend on θ. Then, all relevant information about the value of θ in a sample of n genes is on the vector $\mathbf{\Psi} = \{T_i, Y_i : i = 2, ..., n\}$, whose elements are nonnegative.

The joint density function of $(T_2, \dots T_n, Y_2, \dots, Y_n)$ is

$$\begin{aligned} f_{\mathbf{\Psi}}(\boldsymbol{\psi}) &= \prod_{i=2}^{n} P(Y_i = y_i \mid t_i) f_{T_i}(t_i) \\ &= \exp\left[-\nu \sum_{i=2}^{n} i t_i - \frac{\nu}{\theta} \sum_{i=2}^{n} i(i-1)t_i - (n-1)\log\theta\right] \\ &\times \exp\left[\left(\sum_{i=2}^{n} y_i + n - 1\right)\log\nu + \sum_{i=2}^{n} \log\left(\frac{i-1}{y_i!}\right)\right]. \end{aligned} \quad (5.17)$$

The joint density function of $\mathbf{Y} = (Y_2, ..., Y_n)$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=2}^{n} \left(\frac{i-1}{\theta+i-1}\right)\left(\frac{\theta}{\theta+i-1}\right)^{y_i} \text{ with } y_i = 0, 1, 2, \dots. \quad (5.18)$$

An ideal situation would be one in which one can observe the whole process of evolution of a sample of genes, from the root of the tree to the present. In this situation one can count not only the number of mutations at each branch $(\{Y_i, i = 2, ...n\})$, but also the number of generations $(\{T_i, i = 2, ...n\})$ between the successive nodes. A good estimator of θ with these two sets of data will have smaller variance than all other estimators which contain less information about the process. This estimator can be obtained by the method of maximum likelihood.

We consider here a more realistic and more interesting situation, from the practical point of view. We observe only the number of mutations at each branch of the genealogy, i.e., information about $(\{Y_i, i = 2, ...n\})$, but not about the coalescence time $(\{T_i, i = 2, ...n\})$.

Denote $l_n(\theta)$ the log-likelihood function of θ and $l'_n(\theta), l''_n(\theta), \dots$ its derivatives with respect to θ. Then,

$$l_n(\theta) = \sum_{i=2}^{n} \left\{ y_i \log\left(\frac{\theta}{i-1}\right) - (y_i + 1)\log\left(\frac{\theta+i-1}{i-1}\right)\right\}$$

and the score function is

$$l'_n(\theta) = \frac{\partial l_n(\theta)}{\partial\theta} = \sum_{i=2}^{n}\left(\frac{y_i}{\theta} - \frac{y_i+1}{\theta+i-1}\right) = \sum_{i=2}^{n}\left[\frac{y_i(i-1)-\theta}{\theta(\theta+i-1)}\right].$$

Therefore, the maximum likelihood estimator of θ is the solution of the equation

$$\sum_{i=2}^{n}\left(\frac{y_i+1}{\hat{\theta}_m+i-1}\right) = \frac{\sum_{i=2}^{n} y_i}{\hat{\theta}_m} \quad (5.19)$$

and

$$l_n''(\theta) = \frac{\partial^2 l_n(\theta)}{\partial \theta} = \sum_{i=2}^{n} \left(\frac{y_i + 1}{(\theta + i - 1)^2} - \frac{y_i}{\theta^2} \right).$$

Note that, by (5.19),

$$\frac{\partial^2 l_n(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}_m} = -\sum_{i=2}^{n} \frac{(y_i + 1)(i-1)}{\hat{\theta}_m(\hat{\theta}_m + i - 1)} < 0,$$

which shows that $\hat{\theta}_m$ maximizes $L(\theta, \mathbf{y})$.

Also,

$$\begin{aligned} E(l_n'(\theta) &= 0, \\ Var(l_n'(\theta)) &= \frac{1}{\theta} \sum_{i=2}^{n} \frac{1}{(\theta + i - 1)}. \end{aligned} \tag{5.20}$$

And we can write

$$l_n'(\theta) = \sum_{j=1}^{n-1} \left[\frac{jY_j - \theta}{\theta(\theta + j)} \right].$$

Now let $a_n^\star = \sum_{j=1}^{n-1} (\theta + j)^{-1} = a_n + o(1)$, where a_n is given in (2.2).

$$\varphi_{\frac{l_n'(\theta)}{\sqrt{a_n}}}(t) = e^{-ita_n^{-1/2} a_n^\star} \prod_{j=1}^{n-1} \left(\frac{j}{j - \theta(e^{itj(\theta+j)^{-1}\theta^{-1}a_n^{-1/2}} - 1)} \right) \text{ by (2.4).}$$

Again, by Taylor series expansions, we get

$$\begin{aligned} \log(\varphi_{\frac{l_n'(\theta)}{\sqrt{a_n}}}(t)) &\approx \frac{(it)^2}{2\theta a_n} \sum_{j=1}^{n-1} \frac{j}{(\theta + j)^2} + \frac{(it)^3}{6\theta^2 a_n^{3/2}} \sum_{j=1}^{n-1} \frac{j^2}{(\theta + j)^3} + \dots \\ &= -\frac{t^2}{2\theta a_n} a_n^\star - \frac{it^3}{6\theta^2 a_n^{3/2}} a_n^\star + \frac{t^2}{2a_n} b_n^\star + \frac{it^3}{2\theta a_n^{3/2}} a_n^\star - \frac{it^3}{6a_n^{3/2}} c_n^\star \quad (5.21) \\ &\approx -\frac{t^2}{2\theta} \text{ as } n \to \infty, \end{aligned} \tag{5.22}$$

where $b_n^\star = \sum_{j=1}^{n-1} (\theta + j)^{-2} = b_n + o(1)$ and $c_n^\star = \sum_{j=1}^{n-1} (\theta + j)^{-3}$.

Therefore, as $n \to \infty$,

$$\varphi_{\frac{l_n'(\theta)}{\sqrt{a_n}}}(t) \longrightarrow e^{-t^2/(2\theta)}, \text{ i.e., } \frac{l_n'(\theta)}{\sqrt{a_n}} \xrightarrow{D} N\left(0, \frac{1}{\theta}\right). \tag{5.23}$$

An expansion of $l_n'(\hat{\theta}_m)$ about $l_n'(\theta)$ will give us

$$l_n'(\hat{\theta}_m) = l_n'(\theta) + (\hat{\theta}_m - \theta) l_n''(\theta) + \frac{1}{2}(\hat{\theta}_m - \theta)^2 l_n'''(\theta_n^\star),$$

with $\theta_n^\star \in (\theta, \hat{\theta}_m)$. But, $l_n'(\hat{\theta}_m) = 0$. Therefore,

$$\sqrt{a_n}(\hat{\theta}_m - \theta) = \frac{l_n'(\theta)/\sqrt{a_n}}{-l_n''(\theta)/a_n - \frac{1}{2a_n}(\hat{\theta}_m - \theta) l_n'''(\theta_n^\star)}\,. \tag{5.24}$$

By Taylor series expansion, we have

$$\begin{aligned}
\log \varphi_{\frac{l_n''(\theta)}{a_n}}(t) &= it\frac{b_n^\star}{a_n} - \frac{it}{\theta a_n}\sum_{j=1}^{n-1}\frac{(2\theta j + j^2)}{j(\theta+j)^2} + \frac{(it)^2}{2a_n^2\theta^3}\sum_{j=1}^{n-1}\frac{(2\theta j + j^2)^2}{(\theta+j)^4} + \dots \\
&\approx -\frac{it}{a_n}b_n^\star - \frac{it}{\theta a_n}a_n^\star + \frac{it}{a_n}b_n^\star - \frac{2it^2}{\theta a_n^2}\left[b_n^\star + \theta^2 d_n^\star - 2\theta c_n^\star\right] + \dots \\
&= -\frac{it}{\theta a_n}a_n^\star - \frac{2it^2}{\theta a_n^2}b_n^\star - \frac{2it^2\theta}{a_n^2}d_n^\star + \frac{4it^2}{a_n^2}c_n^\star + \dots \qquad (5.25) \\
&\approx \frac{-it}{\theta} \text{ as } n \to \infty, \qquad (5.26)
\end{aligned}$$

where $a_n^\star = \sum_{j=1}^{n-1}(\theta+j)^{-1}$, $b_n^\star = \sum_{j=1}^{n-1}(\theta+j)^{-2}$, $c_n^\star = \sum_{j=1}^{n-1}(\theta+j)^{-3}$ and $d_n^\star = \sum_{j=1}^{n-1}(\theta+j)^{-4}$.

Therefore,

$$\frac{-l_n''(\theta)}{a_n} \xrightarrow{p} \frac{1}{\theta}\,. \tag{5.27}$$

Now, note that

$$l_n'''(\theta) = -\sum_{i=2}^{n}\left(\frac{2(Y_i+1)}{(\theta+i-1)^3} - \frac{2Y_i}{\theta^3}\right)$$

and using Taylor series expansion, one can see that

$$\frac{-l_n'''(\theta_n^\star)}{2a_n} \xrightarrow{p} \frac{1}{(\theta_n^\star)^2}\,, \text{ i.e., } \left|\frac{-l_n'''(\theta_n^\star)}{2a_n}\right| < \infty. \tag{5.28}$$

Then, using (5.24), the results given by (5.23), (5.27) and (5.28), we obtain that

$$\sqrt{a_n}(\hat{\theta}_m - \theta) \xrightarrow{D} N(0, \theta). \tag{5.29}$$

6 Simulation Results

A number of genealogies was simulated according to different values of θ and n. First, one generates a genealogy tree of DNA sequences. When there are n sequences of DNA, randomly choose two of those sequences to coalesce. Therefore, after this procedure one has $n-1$ DNA sequences. Figure 2 shows an example of this process. In the case of 5 DNA sequences (*A, B, C, D* and *E*), if *B* and *C* are chosen, one gets four new DNA sequences (*A, BC, D* and *E*). Then, after the coalescence of *D* and *E*, one gets three sequences (*A, BC* and

DE). When A and BC are chosen, one gets the genealogic relationship of five DNA sequences shown in Figure 2.

The next step is to generate the number of mutations in each branch. Let Y_{ij} be the number of mutations occurring in the j-th branch among the remaining i branches during coalescence time T_i (Figure 2), and let Y_i be the total number of mutations in i branches. Here, we assume that $T_i \sim Exp\,(i(i-1)/(4N))$, $Y_{ij} \mid t_i \sim Poisson(\nu t_i)$ and $Y_i = \sum_{j=1}^{i} Y_{ij} \mid t_i \sim Poisson(\nu i t_i)$, with $Y_i = \sum_{j=1}^{i} Y_{ij}, \quad i = 2, 3, \ldots$.

The marginal distribution of Y_i can be derived by integrating on t_i the joint distribution of (Y_i, t_i) and it is a geometric distribution as shown in (5.18).

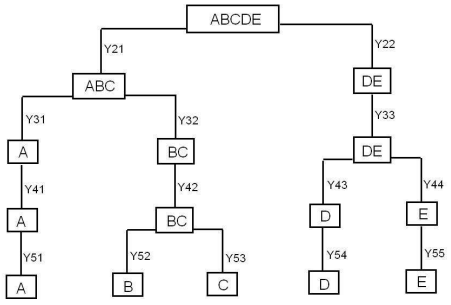

Figure 2: Evolutionary relationship among five genes

The joint distribution of Y_{i1}, Y_{i2}, ..., Y_{ii} for a given value of Y_i is a Multinomial:

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ii} = y_{ii} \mid Y_i = y_i) = \frac{y_i!}{\prod_{j=1}^{i} y_{ij}!} \left(\frac{1}{i}\right)^{y_i}. \quad (6.30)$$

First, one generates Y_i according to (5.18), then Y_{ij}'s are obtained according to (6.30).

With the tree constructed and the mutations in each branch simulated, we computed the number of segregating sites ($S = V_n = Y_2 + \ldots + Y_n$), the mean number of pairwise nucleotide differences ($\mathcal{T}_2$), the number of singletons ($S^\star$) and the maximum likelihood of θ ($\hat{\theta}_m$). Ten thousand simulations were performed, i.e., 10000 trees were generated for each value of n and θ.

Tables 2 and 3 present the values of the estimates and sample variances for each estimator. It is clear that $\hat{\theta}_m$ has always the smallest variance and it is also the less biased among all the estimators in study. $\mathcal{T}_3$, which is based on

the number of singletons, is the worst. It has the largest sample variance and the greatest bias. Comparing $\mathcal{T}_1$ and $\mathcal{T}_2$, one can say that the variance of $\mathcal{T}_1$ is smaller than the variance of $\mathcal{T}_2$. As expected, the variances of all estimators decreases as n increases.

Looking at Figure 3, which shows the behavior of the estimators when $\theta = 2$, one can see that $\mathcal{T}_3$ has a very bad behavior for all values of n. On the other hand, $\hat{\theta}_m$ seems to have the smallest variance and its distribution gets closer to symmetry for $n \geq 30$, but all the estimators have skewed distributions (median $< \theta$). As n increases, the best behavior of $\hat{\theta}_m$ is more evident. Figure 4 shows the distributions when $\theta = 20$ for various values of n. It is evident the improvement of $\mathcal{T}_3$ and the difference among the estimators is more evident as well. The distributions are still skewed to the right, but they are more symmetric than the cases of $\theta = 2$. From Tables 2 and 3, one can see that the mean of $\mathcal{T}_1$ is always very close to θ, but its distribution is still very skewed (median $< \theta$), which can be seen in Figures 3 and 4. It is evident that $\hat{\theta}_m$ still has the best behavior and as n increases, one can see that the distribution of $\hat{\theta}_m$ is more symmetric (Figures 3 and 4).

As $\hat{\theta}_m$ seems to be the best estimator, in order to emphasize its asymptotic symmetric behavior, we show the distribution of $\hat{\theta}_m$ for different values of θ and n in Figure 5. For any value of θ (2, 5, 10, 15 or 20), its distribution is more symmetric as n increases. In terms of symmetry, $\hat{\theta}_m$ seems reasonable even for $n = 30$. Also, its behavior is uniform in θ and the tails of the distributions are very light compared with the normal distribution. The skewness of the distribution of $\hat{\theta}_m$ is illustrated by the Quantile-quantile plots shown in Figure 6. We show here the Q-Q plots for the extreme cases ($\theta = 2$ with $n = 5$ and $n = 1000$, and $\theta = 20$ with $n = 5$ and $n = 1000$). It is clear that as the sample size increases, the distribution gets more symmetric and closer to normality. This is true for all values of θ and n.

Table 2: Summary Statistics of the estimators with $\theta = 2$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\hat{\theta}_m$
5	Estimate	2.01	2.07	1.79	2.00
	Sample variance	2.28	2.64	3.57	2.10
10	Estimate	1.99	2.04	1.81	2.01
	Sample variance	1.44	1.98	3.27	1.34
30	Estimate	2.02	2.07	1.91	2.00
	Sample variance	0.93	1.77	2.78	0.79
50	Estimate	2.00	2.05	1.92	2.00
	Sample variance	0.76	1.68	2.40	0.65
100	Estimate	1.99	2.03	1.94	2.00
	Sample variance	0.62	1.59	2.27	0.56

Table 3: Summary Statistics of the estimators with $\theta = 20$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\hat{\theta}_m$
5	Estimate	20.06	20.62	17.81	20.01
	Sample variance	138.62	163.28	225.99	112.37
10	Estimate	19.89	20.39	18.21	20.02
	Sample variance	81.63	122.85	179.40	55.40
30	Estimate	20.06	20.49	19.18	20.05
	Sample variance	46.99	107.32	108.45	22.64
50	Estimate	20.04	20.49	19.36	19.96
	Sample variance	37.59	104.50	81.58	15.96
100	Estimate	20.01	20.49	19.57	20.00
	Sample variance	28.41	102.35	61.60	11.31

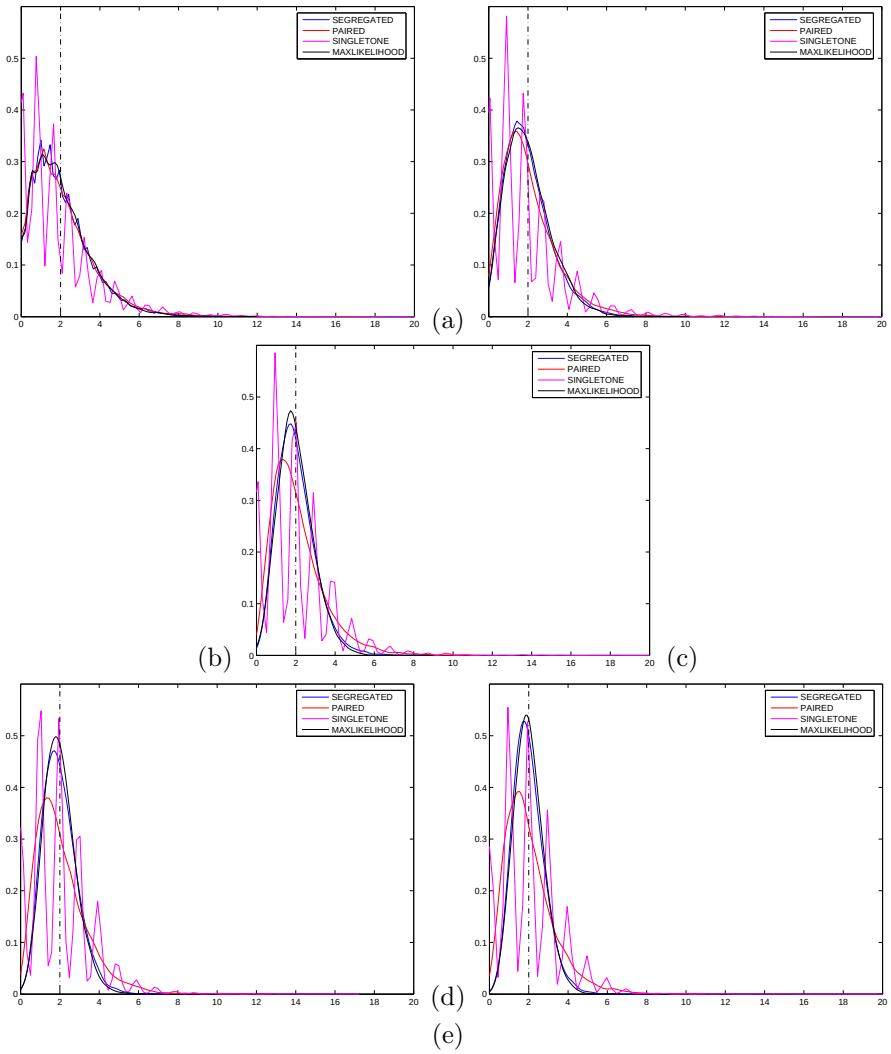

Figure 3: Smoothed Simulated Distribution of the estimators $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and $\hat{\theta}_m$, with $\theta = 2$ and sample sizes: (a) 5, (b) 10, (c) 30, (d) 50 and (e) 100. Kernel Smoothing parameter $h = 0.19$

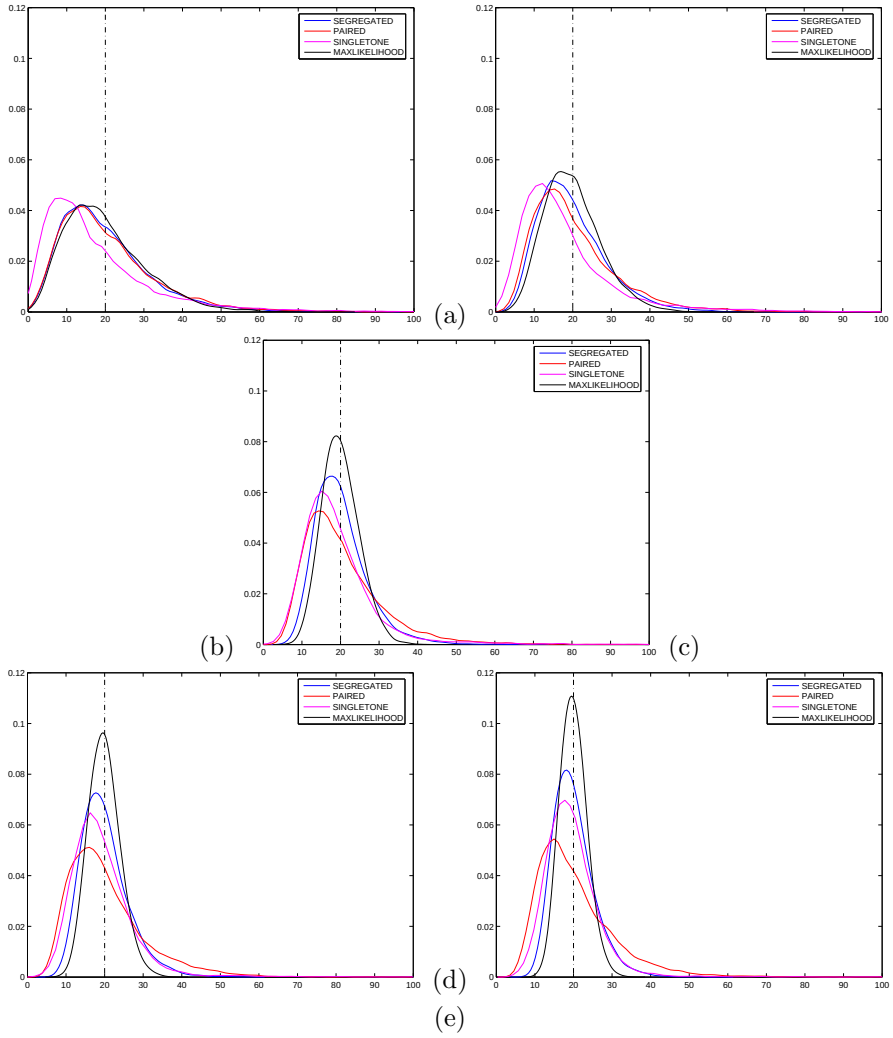

Figure 4: Smoothed Simulated Distribution of the estimators $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and $\hat{\theta}_m$, with $\theta = 20$ and sample sizes: (a) 5, (b) 10, (c) 30, (d) 50 and (e) 100. Kernel Smoothing parameter $h = 1.22$

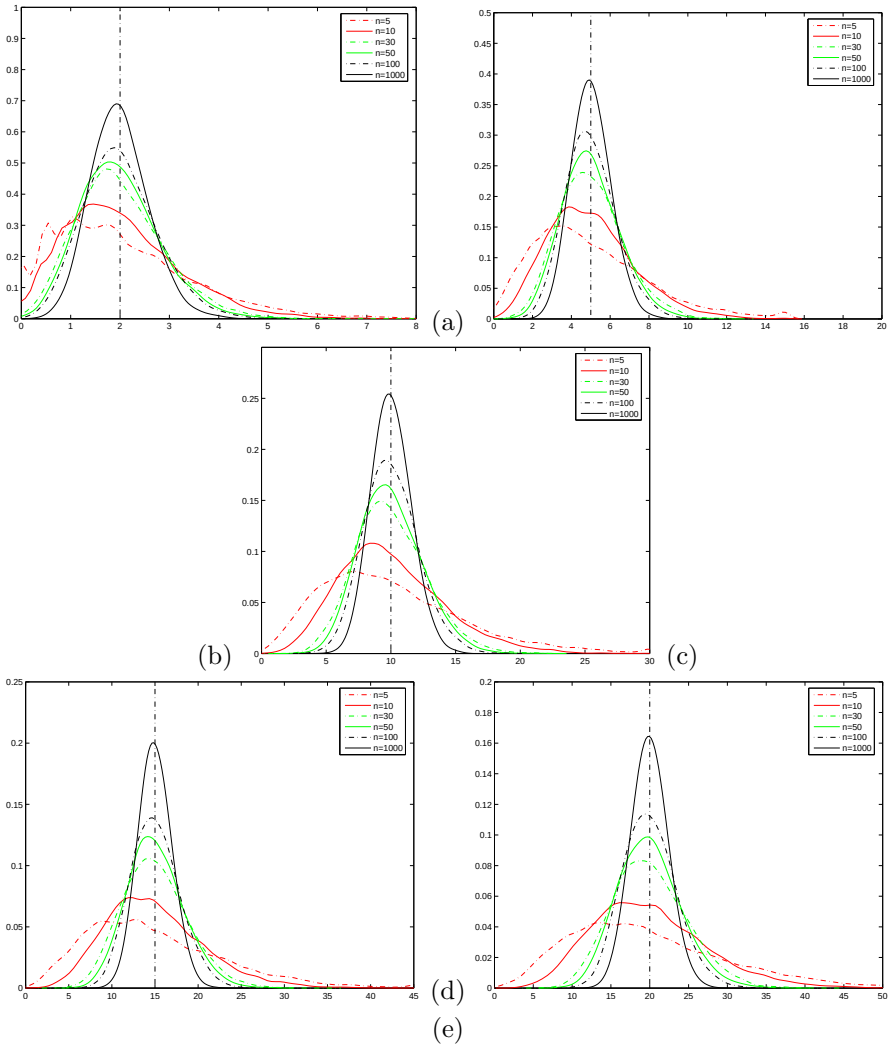

Figure 5: Smoothed Simulated Distribution of $\hat{\theta}_m$ for sample sizes: 5, 10, 30, 50, 100 and 1000 with θ: (a) 2, (b) 5, (c) 10, (d) 15 and (e) 20. Kernel Smoothing parameters are (a) 0.15, (b) 0.29, (c) 0.49, (d) 0.69, (e) 0.88

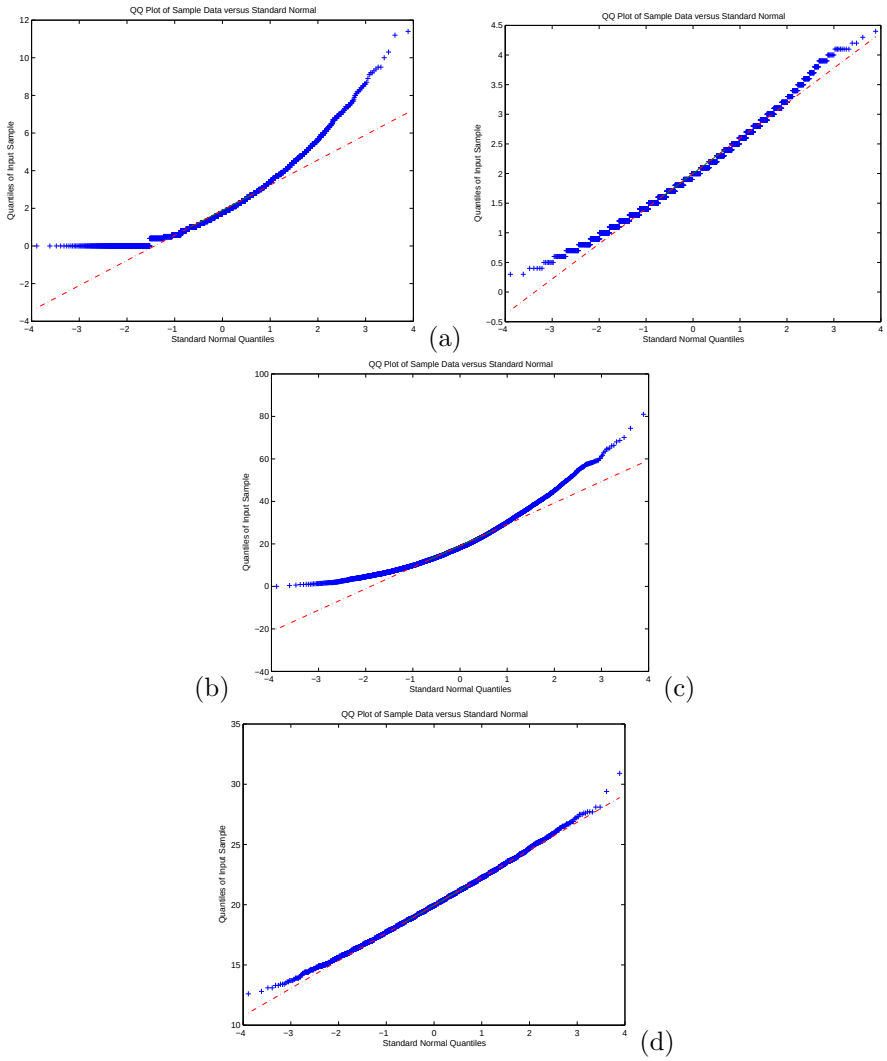

Figure 6: Q-Qplot of $\hat{\theta}_m$ for cases: (a) θ=2 and n=5, (b) θ=2 and n=1000, (c) θ=20 and n=5, (d) θ=20 and n=1000

7 Discussion

We noticed that $\mathcal{T}_1$ and $\hat{\theta}_m$ are asymptotically equivalent, but looking at the second terms in the Taylor series expansion given in (2.6) for $\mathcal{T}_1$ and in (5.21) for $l_n'(\theta)$, we can see that for $0 < \theta < 1$, $\left|\frac{it^3\theta}{6a_n^{1/2}}\right| < \left|\frac{it^3 a_n^\star}{6\theta^2 a_n^{3/2}}\right|$, i.e., $\mathcal{T}_1$ has a better rate of convergence than $\hat{\theta}_m$. On the other hand, for $\theta > 1$, this relationship is reversed and therefore, $\hat{\theta}_m$ is better than $\mathcal{T}_1$. As for $\theta = 1$, we have $\left|\frac{it^3 a_n^\star}{6a_n^{3/2}}\right| < \left|\frac{it^3}{6a_n^{1/2}}\right|$ and $\hat{\theta}_m$ is also better than $\mathcal{T}_1$. This is illustrated in the simulations by looking at Figures 3 and 4, which is evidently the best behavior of $\hat{\theta}_m$. Also, we can see from Figures 5 and 6 that n needs to be at least 1000 in order to get good asymptotic results.

An interesting thing noticed here is that the rate of convergence of normality is $\sqrt{a_n}$, i.e., $\sqrt{\log n}$ for $\mathcal{T}_1$ and $\hat{\theta}_m$. Usually, the asymptotic behavior of MLE's ($\hat{\theta}$, say) are similar to the mean, i.e, the variance of $\sqrt{n}\hat{\theta} = o(1)$, but here we found that $Var(\sqrt{a_n}\hat{\theta}_m) = o(1)$, which explains why the tails of the distribution of $\hat{\theta}_m$ are so light (Figure 6).

Acknowledgements

This research was funded in part by Fundação de Amparo à Pesquisa do Estado de São Paulo (02/03284-5), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Fundo de Apoio ao Ensino, à Pesquisa e à Extensão (858/05) and Conselho Nacional de Desenvolvimento Científico e Tecnológico .

References

Crow, J.F. (1997). The high spontaneous mutation rate: Is it a health risk? *PNAS* **94**, 8380–8386.

Fu, Y.X (1994). A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**, 685–692.

Fu, Y.X. and Li, W.H. (1993a). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Fu, Y.X. and Li, W.H. (1993b). Maximum likelihood estimation of population parameters. *Genetics* **134**, 1261–1270.

Hartl, D. and Clark, A. (2007). *Principles of Population Genetics*. Sinauer Associates, Sunderland.

Hein, J., Schierup, M.H. and Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press.

Huang, W., Chang, B.H.-J., Gu, X., Hewett-Emmett, D. and Li, W.-H. (1997). Sex Differences in Mutation Rate in Higher Primates Estimated from AMG Intron Sequences. *Journal of Molecular Evolution* **44(4)**, 463–465.

Lehmann, E.L. and Casella, G. (2003). *Theory of Point Estimation*. Springer-Verlag, New York.

Pinheiro, H.P., Pinheiro, A. and Sen, P.K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference* **130(1-2)**, 325–339.

Pinheiro, A., Pinheiro, H.P. and Kiihl, S.F. (2008). An asymptotically normal test for the selective neutrality hypothesis. *Institute of Mathematical Statistics Collections* **1**, 377–389.

Ramos-Arroyo, M.A., Moreno, S. and Valiente, A. (2005). Incidence and mutation rate of Huntington's disease in Spain: experience of 9 years of direct genetic testing. *Journal of Neurology and Psychiatry* **76**, 337–342.

Rambaut, A., Posada, D., Crandall, K.A. and Holmes, E.C. (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, **5**, 52–61.

Smith, N.G.C., Webster, M.T. and Ellegren, H. (2002). Deterministic mutation rate variation in the human genome. *Genome Res* **12(9)**, 1350–1356.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 555–595.

Wakeley, J. (2009). *Coalescent Theory - An Introduction*. Robert and Company Publishers, Greenwood Village, Colorado.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Whittemore, A.S. (2004). Miscellanea Estimating genetic association parameters from family data. *Biometrika* **91**, 219–225.