On The Mutation Parameter of Ewens Sampling Formula

ON THE MUTATION PARAMETER OF EWENS SAMPLING FORMULA

BY

BENEDICT MIN-OO, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Applied Science (2016) McMaster University
(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE: On The Mutation Parameter of Ewens Sampling Formula

AUTHOR: Benedict Min-Oo
B.Sc., (Statistics)
Concordia University, Montreal, Canada

SUPERVISOR: Dr. Shui Feng

NUMBER OF PAGES: ix, 48

To my friends and family

Abstract

Ewens Sampling formula is the sampling distribution for a population assumed to follow a one parameter Poisson-Dirichlet distribution (PD(θ)), where the parameter θ is fixed. In this project this assumption will be loosened and we will look at θ as a function of the sample size n denoted $\theta_n = \alpha n^{\beta_1} (\log n)^{\beta_2}$, where $\alpha > 0, \beta_1 \geq 0, \beta_2 \geq 0$. This will result in sampling from a family of $PD(\theta_n)$ distributions. Estimators for this new construction will be tested using two different simulation methods.

Acknowledgements

I would like to thank my supervisor Dr. Shui Feng for his incredible patience, guidance, and wisdom. Furthermore, I would like to thank Dr. Hoppe and Dr. Viveros for serving on my committee.

In addition, I would like to thank my father for setting me on this path, Dr. Balakrishnan for showing me the way, my friends and classmates for keeping me sane, my friend Mu He for helping me find the finish line, and my mother for keeping me fed.

Notation and abbreviations

ESF : Ewens Sampling Formula

MLE : Maximum Likelihood Estimator

PD : Poisson-Dirichlet

RMSE : Root Mean Square Error

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

In population genetics, Warren Ewens (Ewens (1972)) discovered the sampling distribution for allele frequencies in a neutral population called the Ewens sampling formula (ESF). This famed formula has many applications. For a great overview see (Crane (2016)). Ewens sampling formula has one parameter θ, which represents the population mutation rate. Given a sample of size n, the number of distinct alleles in the sample, K_n, is approximately the magnitude of $\log(n)$ for large n and fixed θ. The focus of this project is to investigate cases where θ is not fixed. Specifically, we will be treating θ as a function of n, denoted $\theta_n = \alpha n^{\beta_1}(\log(n))^{\beta_2}$. We will also look at the maximum likelihood estimator (MLE) of θ as well as develop closed form estimators for θ and θ_n using asymptotic results. Finally, we will compare two methods of simulation as well as test the performance of our new estimators.

1.1 Motivation

Asymptotic results have been studied intensively in recent years for large θ (Feng (2010)). It was shown in (Feng (2010)) that some of these results correspond to a regime where θ and n are related in a special way. We intend to pursue further along these lines by focusing on the ESF when θ and n are related.

1.2 Layout

We start with a discussion of the Wright-Fisher model and the infinite dimensional generalization, the infinitely-many-neutral-alleles model. The ESF arises as the sampling distribution of the equilibrium distribution. Our main focus will be on the number of components K_n and the parameter θ. To gain a better understanding of θ we will go through Hoppe's Urn scheme and then look into the link between ESF and the Poisson-Dirichlet distribution. Following this, we will look at the MLE and a few closed form estimators for a fixed θ, before exploring our proposed θ_n and introducing estimators for θ_n. Then, two simulation methods will be introduced and compared. Finally, we will discuss the results and possible future work.

Chapter 2

Construction and Background

To motivate our study we start with a review of the Wright-Fisher model developed by (Wright (1931)) and (Fisher (1930)). This is followed by a discussion the infinitely-many-neutral-alleles model (Ethier and Kurtz (1981)), Hoppe's Urn model (Hoppe (1984)), Dirichlet distribution, and the Poisson-Dirichlet distribution.

2.1 Wright-Fisher Model

Consider a diploid (having a pair of each chromosome) population of finite size N, with $2N$ alleles at any given time. Let the alleles be composed of two types (C_1, C_2). Let X_t be the number of alleles of type C_1 at time t. Then, X_t can be described as a discrete time Markov chain with state space $(0, ..., 2N)$ and transition probabilities,

$$P(X_{t+1} = j | X_t = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \tag{2.1}$$

This is the Wright-Fisher Model and was introduced independently by both Wright and Fisher. Clearly, this is akin to binomial sampling with probability $p = \frac{i}{2N}$.

The Wright-Fisher model can be generalized to M allelic types by replacing the binomial sampling with multinomial sampling. For example, if we have M allelle types then,

$$P(X_{t+1} = (j_1, ..., j_M)|X_t = (i_1, ..., i_M)) = \left(\frac{2N!}{j_1!, \cdots, j_M!}\right)\left(\frac{i_1}{2N}\right)^{j_1}\cdots\left(\frac{i_M}{2N}\right)^{j_M} \tag{2.2}$$

This model does not take into account selection, mutation, population subdivision, two sexes, or any other additional effect. Note that the total number of alleles stays fixed at $2N$ and the number of different allele types is fixed at M. Now let us consider, the case where each allele has mutation rate μ and there are an infinite number of possible alleles. The key notion here is that every mutation results in a new allelic type that has yet to be seen.

So, at generation t we have X_i genes of allelic type C_i, then in generation $t + 1$ we will have Y_i genes of allelic type C_i plus Y_0 new distinct mutant genes. If our mutation rate is μ then from (Ewens (2004)) we have,

$$\text{Prob}\left\{Y_0, Y_1, Y_2, ...|X_1, X_2, ...\right\} = \frac{(2N)!}{\Pi Y_i!}\Pi {\pi_i}^{Y_i} \tag{2.3}$$

Where, $\pi_0 = \mu$ and $\pi_i = X_i(1 - \mu)/(2N)$

Now, the issue with this model is that there is no reverse mutation (mutating to a new allelic type and then returning back to the old type), so each allelic type will eventually vanish from the population. Therefore, there can exist no nontrivial stationary distribution for the frequency of any allele (Ewens (2004)). However, let us consider a delabeled configuration, where we ignore the specific type of allele and only focus on how many there are of each type. This delabeled configuration is $\{a_1, a_2, a_3, ...\}$ where a_1 is the number of genes of one type, a_2 is the number of genes of another type, and so on. The total possible number of configurations for a population N can be written down as $p(2N)$ where p is the partition function, which represents the total number of possible partitions of a natural number.

Now, using this delabled configuration (Ewens (1972)) developed an approximating partition probability formula for a sample of size n. For more on the biological construction refer to (Ewens (2004)).

When N is large and μ is small in such a way that $N\mu$ is fixed, the Wright-Fisher model can be approximated by the Wright-Fisher diffusion. The finite dimensional Wright-Fisher diffusion has an infinite dimensional approximation, the infinitely-many-neutral-alleles model, through appropriate scaling and ordering (Ethier and Kurtz (1981)).

2.2 Dirichlet and Poisson-Dirichlet Distributions

The Wright-Fisher diffusion and the infinitely-many-neutral-alleles model are reversible diffusions with respective reversible measure, the Dirichlet distribution and the one-parameter Poisson-Dirichlet distribution (Kingman (1975)).

The Dirichlet distribution has probability density function given by,

$$f\left(x_1, \cdots, x_M | \phi_1, \cdots, \phi_M\right) = \frac{1}{\mathrm{B}(\boldsymbol{\phi})} \prod_{i=1}^{M} x_i^{\phi_i - 1},$$

where $\phi_i > 0$ for any i and,

$$\mathrm{B}(\boldsymbol{\phi}) = \frac{\prod_{i=1}^{M} \Gamma(\phi_i)}{\Gamma\left(\sum_{i=1}^{M} \phi_i\right)}, \; \boldsymbol{\phi} = \{\phi_1, ..., \phi_M\}$$

which is called the beta function.

The support is, $0 < x_i < 1$ for any $1 < i < M$ and $\sum_{i=1}^{M} x_i = 1$ and although it has M variables it exists on the (M-1)-dimensional simplex. This is because if you know M-1 of the variables you know them all since $x_M = 1 - x_1 - \cdots - x_{M-1}$. For more on the Dirichlet distribution refer to (Kotz *et al.* (2000)).

Now if we set $\phi = \phi_1 = \phi_2 = \cdots = \phi_M$ we end up with the symmetric Dirichlet distribution. If we let $M \to \infty$ and $\phi \to 0$ in such a way that $\lim_{M\to\infty} M\phi = \theta$, then order x_i such that,

$$\left\{(x_1, x_2, ...) : x_1 \geq x_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} x_i = 1\right\}$$

we end up with the $PD(\theta)$.

2.3 Ewens Sampling Formula

Taking a random sample of size n from a population with the frequency distribution PD(θ). For each $i = (1, .., n)$ let A_i denote the number of alleles that appear in the sample i times. The vector $\boldsymbol{A} = (A_1, .., A_n)$ is the random allelic partition of the sample.

ESF gives the distribution of $\boldsymbol{A}$ as follows

$$P_n[\mathbf{A} = \mathbf{a}] = \frac{n!}{\theta^{(n)}} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!} \mathbb{I}\left\{\sum_{i=1}^{n} i a_i = n\right\} \tag{2.4}$$

where $\mathbf{a} = (a_1, ..., a_n)$ is the given allelic partition and $\theta^{(n)} = \theta \times (\theta+1) \times \cdots \times (\theta+n-1)$ is the rising factorial.

2.4 Number of components K_n

The number of distinct alleles in the sample is the random variable K_n where $K_n = \sum_{i=1}^{n} A_i$. For any $1 \leq k \leq n$ it is known (Ewens (1972))

$$P[K_n = k] = |s(n,k)| \frac{\theta^k}{\theta^{(n)}} \tag{2.5}$$

where $|s(n,k)|$ is the unsigned Stirling number of the first kind, with value corresponding to the coefficient of θ^k in the expanded $\theta^{(n)}$. The expected value and variance are given below (Ewens (2004))

$$E[K_n] = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1} \tag{2.6}$$

$$\text{var}[K_n] = \theta \sum_{j=1}^{n} \frac{j-1}{(\theta + j - 1)^2}. \tag{2.7}$$

These results can be derived directly from (2.3), or by using Hoppe's urn below.

2.5 Hoppe's Urn

Consider an urn containing one black ball of mass θ. Select a ball from the urn, if it is black return it along with a ball of a brand new colour. If it is not the black ball, return the ball along with another ball of the same colour with mass one. Stop when you have n non-black balls and label them $1, 2, .., \tilde{K}_n$, where $\tilde{K}_n$ represents the balls colour. Now, let $\tilde{A}_i$ represent the number of colours that appear i times. Then, $\tilde{\boldsymbol{A}} = (\tilde{A}_1, ..., \tilde{A}_n)$ will have the same distribution as $\boldsymbol{A}$ above.

Proposition (Ewens (2004)): The number of distinct alleles K_n is a sufficient statistic for θ.

Proof:

$$\begin{aligned}
P(\mathbf{A} = \mathbf{a} | K_n = k) &= \frac{P(\mathbf{A} = \mathbf{a})}{P(K_n = k)} \\
&= \frac{\frac{n!}{(\theta)^n} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!}}{|s(n,k)| \frac{\theta^k}{(\theta)^n}} \\
&= \frac{n!}{|s(n,k)|} \frac{\theta^{\sum_{j=1}^{n} a_j}}{\theta^k} \prod_{i=1}^{n} \frac{1}{i^{a_i} a_i!} \\
&= \frac{n!}{|s(n,k)|} \prod_{i=1}^{n} \frac{1}{i^{a_i} a_i!} \ \square
\end{aligned}$$

This does not depend on θ. Therefore, the number of alleles K_n is a sufficient statistic for the mutation parameter θ. So, any information about θ can be inferred solely by k. Now, we return to the urn model with a focus on K_n.

Let, K_n be the number of different colored balls in the urn. After each draw, the population n will increase by 1 regardless of what colour ball is obtained. So on the

jth draw, let

$$\xi_j = \begin{cases} 1, & \text{Black ball drawn} \\ 0, & \text{Black ball not drawn} \end{cases}$$

On the first draw there is only the black ball, so $\xi_1 \equiv 1$.

For the subsequent draws $j > 1$,

$$\xi_j = \begin{cases} 1, & \frac{\theta}{\theta+j-1} \\ 0, & \frac{j-1}{\theta+j-1} \end{cases}$$

This is equivalent to saying that each ξ_j is a Bernoulli random variable with probability $\frac{\theta}{\theta+j-1}$. Now, they are not identical but they are independent since the probability of drawing the black ball on the jth draw will remain the same regardless of how many times it was drawn before.

Knowing that each time a black ball is drawn the value of K_n increases by one we can see that,

$$K_n = \xi_1 + ... + \xi_n, \tag{2.8}$$

and by taking the expected value we get,

$$\begin{aligned}E[K_n] =&E[\xi_1 + ... + \xi_n]\\ =&E[\xi_1] + ... + E[\xi_n]\\ =&1 + ... + \frac{\theta}{\theta + n - 1}\\ =&1 + \sum_{i=2}^{n} \frac{\theta}{\theta + i - 1}\\ =&\sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}\end{aligned}$$

Now, if instead of counting k we calculate x_k the frequency of each colour, so that $\sum_{k=1}^{n} x_k = 1$ and instead of stopping after n draws we just keep going, eventually we end up with an infinite number of frequencies $\{x_1, x_2, ...\}$. Now, if we order these points in descending order they can be described by the one parameter Poisson-Dirichlet distribution.

Chapter 3

Estimating θ

3.1 Maximum Likelihood Estimator

The MLE is found by maximizing the log-likelihood as a function of its parameter. In our case, since we know that K_n is a sufficient statistic for θ we will find the MLE using the probability mass function of K_n.

$$P[K_n = k] = |s(n,k)| \frac{\theta^k}{\theta^{(n)}}$$

The Likelihood of θ is,

$$L(\theta|k) = |s(n,k)| \frac{\theta^k}{\theta^{(n)}}$$

The log likelihood is, (here and below log is the natural logarithm)

$$l(\theta|k) = \log|s(n,k)| + k\log(\theta) - \log(\theta^{(n)})$$

Taking the derivative with respect to θ we get,

$$\frac{dl}{d\theta} = \frac{k}{\theta} - \frac{1}{\theta^{(n)}}\frac{d\theta^{(n)}}{d\theta}$$

where, $\frac{d\theta^{(n)}}{d\theta} = \theta^{(n)}\left(\psi(\theta+n) - \psi(\theta)\right)$ and $\left(\psi(\theta+n) - \psi(\theta)\right) = \sum_{j=0}^{n-1}\frac{1}{\theta+j}$ (see A.3) Setting equal to zero we get,

$$k = \sum_{j=1}^{n}\frac{\tilde{\theta}}{\tilde{\theta}+j-1}$$

The solution to this is the MLE for θ. We validate this by showing that the information at the MLE is greater than 0.

$$\begin{aligned}
-\frac{d^2l}{d\tilde{\theta}^2} &= \frac{k}{\tilde{\theta}^2} - \sum_{i=1}^{n}\frac{1}{(\tilde{\theta}+i-1)^2} \\
&= \frac{\sum_{i=1}^{n}\frac{\tilde{\theta}}{\tilde{\theta}+i-1}}{\tilde{\theta}^2} - \sum_{i=1}^{n}\frac{1}{(\tilde{\theta}+i-1)^2} \\
&= \sum_{i=1}^{n}\frac{1}{\tilde{\theta}(\tilde{\theta}+i-1)} - \sum_{i=1}^{n}\frac{1}{(\tilde{\theta}+i-1)^2} \\
&> 0
\end{aligned}$$

This is valid since for any $i > 1$, the first term is larger than the second and at $i = 1$ the two terms are equal.

The mean square error (MSE) can be approximated for the MLE (Ewens (2004)).

Let us denote $f(x) = \sum_{i=1}^{n} \frac{x}{x+i-1}$, $f'(x) = \sum_{j=1}^{n} \frac{i-1}{(x+i-1)^2}$. Then,

$$K_n - E[K_n] = f(\tilde{\theta}) - f(\theta)$$

Using first-order Taylor approximations for the right hand we get,

$$\begin{aligned} f(\tilde{\theta}) - f(\theta) & \\ &\approx f'(a)(\tilde{\theta} - a) - f'(a)(\theta - a) \\ &= (\tilde{\theta} - \theta) f'(a) \end{aligned}$$

letting $a = \theta$ we get $K_n - E[K_n] \approx (\tilde{\theta} - \theta) f'(\theta)$. Now,

$$\begin{aligned} E[(\tilde{\theta} - \theta)^2] &\approx \frac{E\left[(K_n - E[K_n])^2\right]}{f'(\theta)^2} \\ \to MSE(\tilde{\theta}) &\approx \frac{\text{var}[K_n]}{f'(\theta)^2} \end{aligned}$$

From (2.3) we have that,

$$\frac{\text{var}[K_n]}{f'(\theta)^2} = \frac{\theta \sum_{i=1}^{n} \frac{i-1}{(\theta+i-1)^2}}{\left(\sum_{i=1}^{n} \frac{i-1}{(\theta+i-1)^2}\right)^2} = \frac{\theta}{\sum_{i=1}^{n} \frac{i-1}{(\theta+i-1)^2}}$$

Therefore,

$$MSE(\tilde{\theta}) \approx \frac{\theta}{\sum_{i=1}^{n} \frac{i-1}{(i+\theta-1)^2}}. \tag{3.1}$$

Please refer to (Ewens (2004)) for more details.

3.2 Closed Form Estimators

The MLE has no closed form solution. However, some closed form estimators can be constructed using asymptotic approximations of E$[K_n]$.

$E[K_n]$ can be approximated by the integrals,

$$\begin{aligned}\sum_{i=1}^{n} \frac{\theta}{\theta+i-1} \geq& \theta \int_0^{n-1} \frac{1}{\theta+x} dx \\ =& \theta \int_{\theta}^{n+\theta-1} \frac{1}{y} dy \\ =& \theta \left(\log(n+\theta-1) - \log(\theta)\right) \\ \approx& \theta \left(\log(1+\frac{n}{\theta})\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n} \frac{\theta}{\theta+i-1} \leq& 1+\theta \int_{\theta}^{n+\theta-1} \frac{1}{y} dy \\ \leq& 1+\theta \left(\log(1+\frac{n}{\theta})\right)\end{aligned}$$

Now, for n large, both integrals will resemble $\theta\left(\log(n)\right)$.

This results in the simple consistent estimator,

$$\hat{\theta} = \frac{k}{\log(n)} \tag{3.2}$$

The problem is that if k is significantly larger than $\log(n)$, $\hat{\theta}$ is far from the MLE.

For extreme cases where k is very close to n let us look at the following,

$$\begin{aligned}
E\left[1-\frac{k}{n}\right] &= 1-\frac{1}{n}E[k] \\
&= 1-\frac{1}{n}\sum_{j=1}^{n}\frac{\theta}{\theta+j-1} \\
&\leq 1-\frac{\theta}{n}\left(\log(1+\frac{n}{\theta})\right) \\
&= 1-\frac{\theta}{n}\left(\frac{n}{\theta}-\frac{n^2}{2\theta^2}+o(\frac{n^3}{\theta^3})\right) \\
&= \frac{n}{2\theta}+o(\frac{n^2}{\theta^2}) \\
&\approx \frac{n}{2\theta} \text{ for } \theta > n
\end{aligned}$$

Remark: The Above calculation will work for any $\theta > n$, however the bigger θ is in comparison to n the faster it will converge.

So, I propose the estimator.

$$\hat{\hat{\theta}} = \frac{n^2}{2(n-k)} \tag{3.3}$$

Remark: This estimator requires k to be very close to n but not equal to n.

Another estimator, which tries to combine $\hat{\theta}$ and $\hat{\hat{\theta}}$ is,

$$\theta^* = \frac{nk}{(n-k)\log(n)} \tag{3.4}$$

This estimator will look like $k/\log(n)$ for small k while it will look like a tempered version of $\hat{\hat{\theta}}$ for k close to n.

I plotted this estimator vs the MLE below.

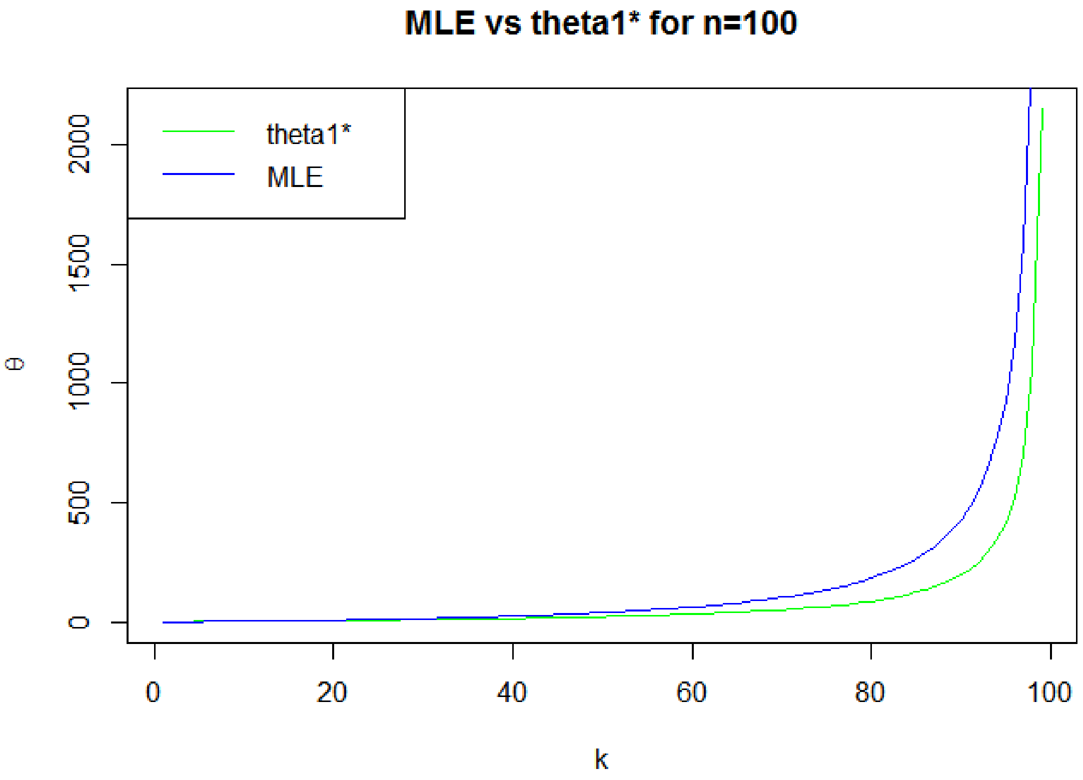

Figure 3.1: MLE vs θ^*

Based on this graph I believe that there should be another term, so I propose,

$$\theta^{**} = \frac{nk}{(n-k)\left(\log(n) - \log(\log(n))\right)} \tag{3.5}$$

which is plotted against the MLE below in Figure 3.2,

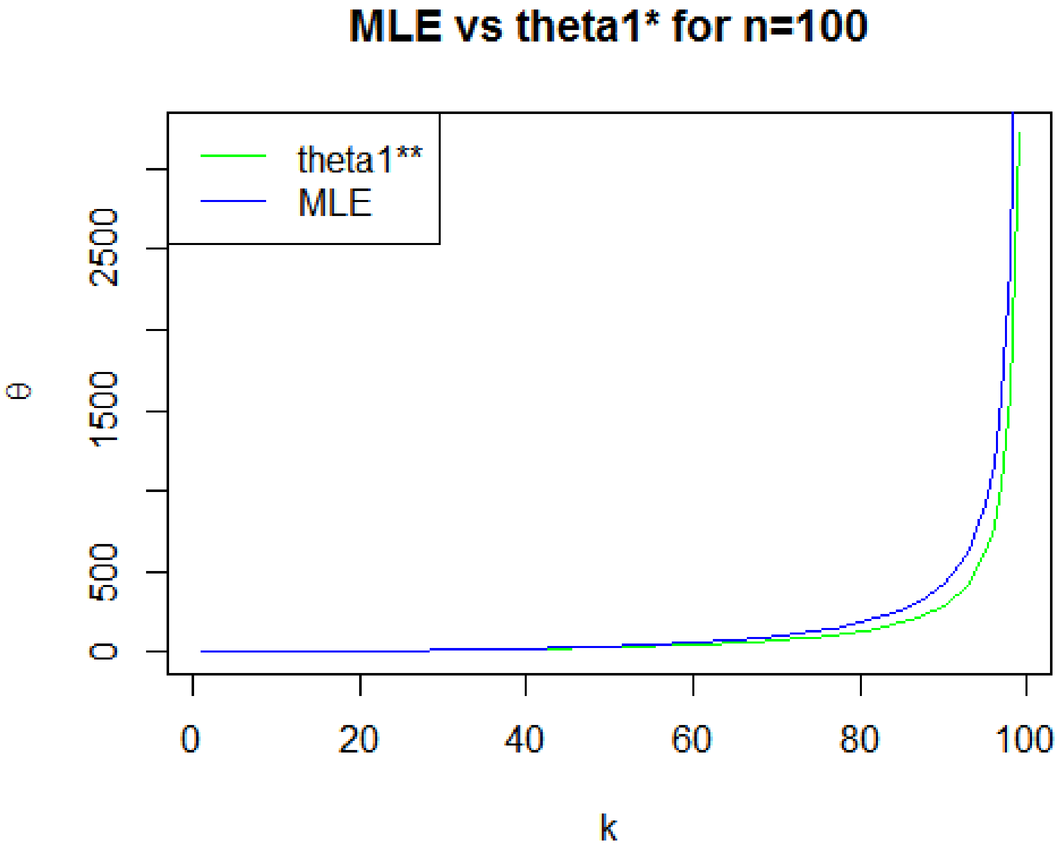

Figure 3.2: MLE vs θ^{**}

Chapter 4

θ as a funciton of n

If we consider θ as fixed then we know that $K_n \approx \log(n)$. However, if we see a sample where this is not the case, then the assumption that θ is fixed may not be valid. For these cases I propose treating θ as a function of n in the following way. $\theta_n = \alpha n^{\beta_1}(\log(n))^{\beta_2}$.

4.1 Construction

We will look at the two following cases,

Case 1: $\beta_1 = 1 - \beta_2$, $\beta_2 > 0$

$$\theta_n = \alpha n^{\beta_1}(\log(n))^{1-\beta_1} \tag{4.1}$$

Case 2: $\beta_2 = 0$

$$\theta_n = \alpha n^{\beta_1} \tag{4.2}$$

We will now approximate $E[K_n]$ given $\theta = \theta_n$.

For $\beta_1 < 1$, $\beta_2 < \frac{(1-\beta_1)\log(n)-\log(\alpha)}{\log(\log(n))}$

$$\begin{aligned}
E[K_n] &= \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1} \\
&\geq \theta \left(\log(1 + \frac{n}{\theta}) \right) \\
&= \theta \log \left(\frac{n}{\theta} \right) + \theta \left(\log(1 + \frac{\theta}{n}) \right) \\
&= \alpha n^{\beta_1} (\log(n))^{\beta_2} \left[\log \left(\frac{n}{\alpha n^{\beta_1} (\log(n))^{\beta_2}} \right) + \log \left(1 + \frac{\alpha n^{\beta_1} (\log(n))^{\beta_2}}{n} \right) \right] \\
&= \alpha n^{\beta_1} (\log(n))^{\beta_2} \left[\log \left(\frac{n^{1-\beta_1}}{\alpha (\log(n))^{\beta_2}} \right) + \log \left(1 + \frac{\alpha (\log(n))^{\beta_2}}{n^{1-\beta_1}} \right) \right] \\
&= \alpha n^{\beta_1} (\log(n))^{\beta_2} \left[(1 - \beta_1) \log(n) - \beta_2 \log(\log(n)) \right] + O(n^{\beta_1 - 1})
\end{aligned}$$

when n is large and similarly,

$$\begin{aligned}
E[K_n] &= \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1} \\
&\leq 1 + \theta \left(\log(1 + \frac{n}{\theta}) \right) \\
&= \alpha n^{\beta_1} (\log(n))^{\beta_2} \left[(1 - \beta_1) \log(n) - \beta_2 \log(\log(n)) \right] + O(n^{\beta_1 - 1})
\end{aligned}$$

this result is obtained using the asymptotic expansion of $\log(1 + x)$ at $x = 0$ (see A.4).

For $\beta_1 = 1,\ \beta_2 = 0$

$$
\begin{aligned}
E\left[\frac{K_n}{n}\right] &= \frac{1}{n}\sum_{j=1}^{n}\frac{\theta}{\theta+j-1}\\
&= \sum_{j=0}^{n-1}\frac{\alpha}{\alpha n + j}
\end{aligned}
$$

using integral approximations and assuming n large (see A.5)

$$
\begin{aligned}
&\approx \alpha[\log(\alpha n + n) - \log(\alpha n)]\\
&= \alpha[\log(\alpha+1) + \log(n) - \log(\alpha) - \log(n)]\\
&= \alpha\,[\log(\alpha+1) - \log(\alpha)]\\
&= \alpha\log\left(1+\frac{1}{\alpha}\right)
\end{aligned}
$$

4.2 Estimators

Let k_n denote the observed value of K_n in a sample of size n.

For $0 \le \beta_1 < 1,\ \beta_2 < \frac{(1-\beta_1)\log(n)-\log(\alpha)}{\log(\log(n))}$ I propose the estimator,

$$\hat{\theta}_n = \frac{k_n}{(1-\beta_1)\log(n) - \beta_2\log(\log(n))} \tag{4.3}$$

For $\beta_1 = 1, \beta_2 = 0$ I propose the solution to

$$
\begin{aligned}
\frac{k_n}{n} &= \hat{\hat{\alpha}}\log\left(1+\frac{1}{\hat{\hat{\alpha}}}\right)\\
\hat{\hat{\theta}}_n &= \hat{\hat{\alpha}}n
\end{aligned}
$$

This is not a closed form solution, so let us use the asymptotic expansion of $\log(1+x)$

once again. For $(\alpha > 1)$,

$$\begin{aligned}
\frac{k_n}{n} &\approx \hat{\hat{\alpha}} \log\left(1 + \frac{1}{\hat{\hat{\alpha}}}\right) \\
&\approx \hat{\hat{\alpha}}\left(\frac{1}{\hat{\hat{\alpha}}} - \frac{1}{2(\hat{\hat{\alpha}})^2} + \frac{1}{3(\hat{\hat{\alpha}})^3} - o(\hat{\hat{\alpha}}^4)\right) \\
&\approx 1 - \frac{1}{2(\hat{\hat{\alpha}})} + \frac{1}{3(\hat{\hat{\alpha}})^2} \\
&\Rightarrow 6\left(1 - \frac{k_n}{n}\right)(\hat{\hat{\alpha}})^2 - 3\hat{\hat{\alpha}} + 2 = 0 \\
&\Rightarrow \hat{\hat{\alpha}} = \frac{3 \pm \sqrt{9 - 48(1 - \frac{k_n}{n})}}{12(1 - \frac{k_n}{n})} \\
&= \frac{3 \pm \sqrt{48\frac{k_n}{n} - 39}}{12(1 - \frac{k_n}{n})}
\end{aligned}$$

Now, for the solution to be real, $k_n \geq \frac{13n}{16}$. For positive solutions there are two cases.

Case 1

$$\begin{aligned}
3 - \sqrt{48\frac{k_n}{n} - 39} &\geq 0 \\
\rightarrow \sqrt{48\frac{k_n}{n} - 39} &\leq 3 \\
\rightarrow k_n &\leq n
\end{aligned}$$

which will always be true and,

Case 2

$$3 + \sqrt{48\frac{k_n}{n} - 39} \geq 0$$

which is always true for real solutions.

Therefore,

$$\hat{\hat{\alpha}} = \begin{cases} \text{No real solution} & k_n < \frac{13n}{16} \\ \frac{3 \pm \sqrt{48\frac{k_n}{n} - 39}}{12(1 - \frac{k_n}{n})} & \frac{13n}{16} \leq k_n < n \\ \infty & k_n = n \end{cases} \tag{4.4}$$

and,

$$\hat{\hat{\theta}}_n = \hat{\hat{\alpha}} \log(n) \tag{4.5}$$

For the case when $\beta_1 > 1$, and $\beta_2 > 0$ it is suitable to use the estimator $\hat{\hat{\theta}}$ since θ will be larger than n.

4.3 Approximating α, β_1

Using R, we can find a solution to the MLE for a given k and n. We can then examine the shape of $\tilde{\theta}$ as n grows for different values of k. This is shown in the plots below.

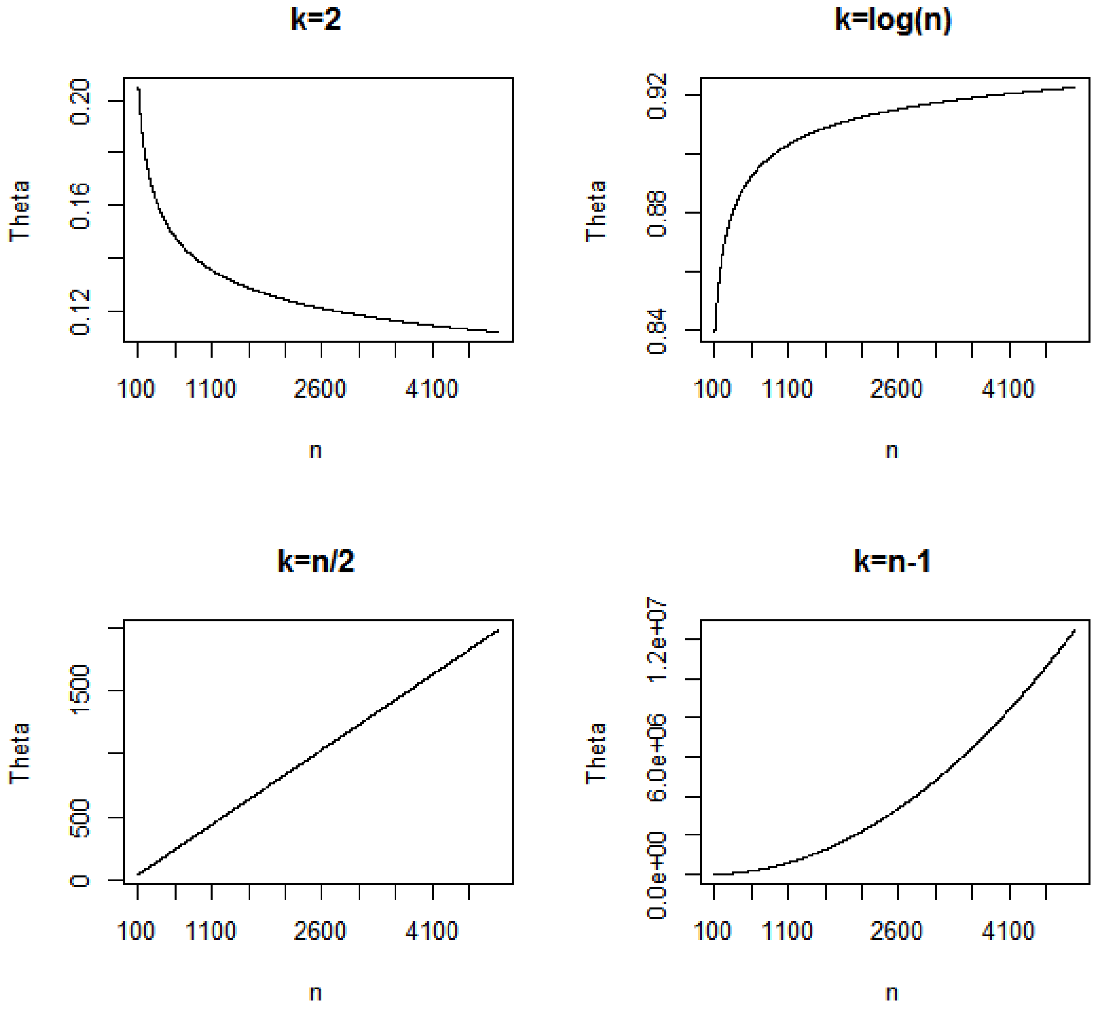

Figure 4.1: Shape of θ for different k_n

Clearly, we can see that for different k the shape of θ changes. Now, using least squares regression I will fit $\theta_n = \alpha n^{\beta_1}(\log(n))^{1-\beta_1}$,

$$\theta_n = \alpha n^{\beta_1}(\log(n))^{1-\beta_1}$$

$$\Rightarrow \log(\theta_n) = \beta_1(\log(n) - \log(\log(n))) + \log(\alpha) + \log(\log(n))$$

Let $y = \log(\theta_n)$, $\beta_0 = \log(\alpha) + \log(n)$ and, $x_n = \log(n) - \log(\log(n))$

$$\Rightarrow y = \beta_1 x_n + \beta_0$$

Remark: We see that $\alpha = \frac{e^{\beta_0}}{\log(n)}$ is a function of $log(n)$. So, we can rewrite θ_n as,

$$\theta_n = e^{\beta_0}\left(\frac{n}{\log(n)}\right)^{\beta_1}$$

so I will record the β_1's and β_o's in the following table.

Table 4.1: Regression Results for Case 1

	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n-1$
β_1	-0.16	0.02	0.46	1.17	2.33
β_0	-1.16	-0.21	-0.46	0.15	1.46
R^2	0.99	0.95	1	1	1

The results in this table were generated using the `lm` function in R.

Here are the plots showing the MLE vs the fitted θ_n's,

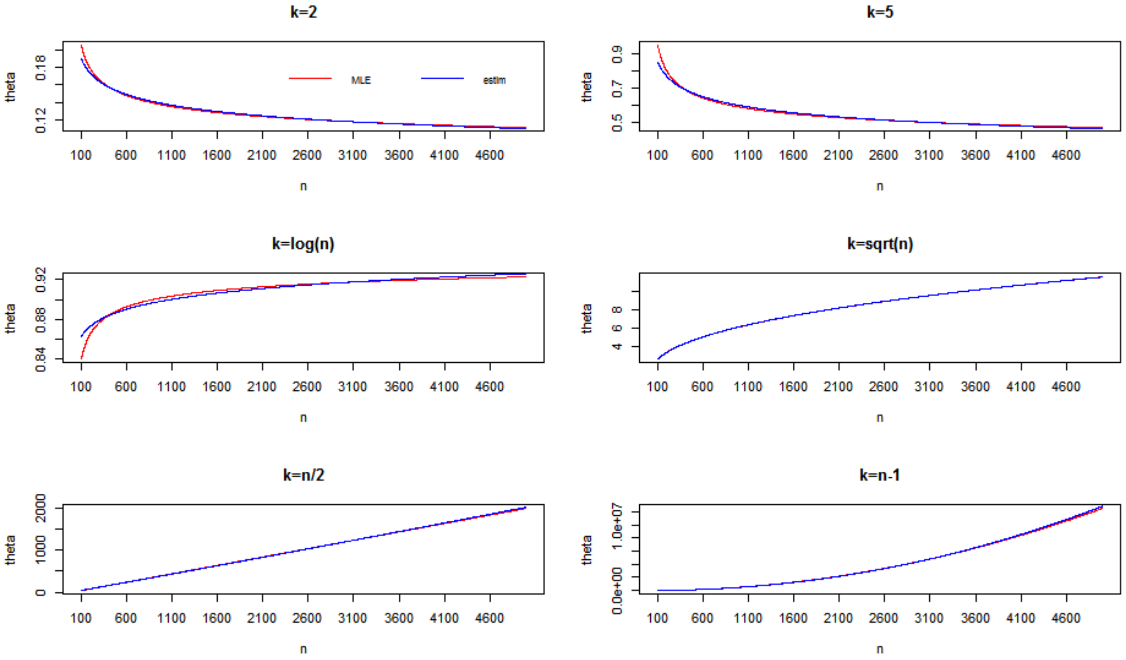

Figure 4.2: Least square approximations for θ_n

4.3.1 $\theta_n = \alpha n^\beta$

For our second case we have,

$$\log(\theta) = \log(\alpha) + \beta \log(n)$$

$$\text{Let } \log(\theta) = y, \text{ and } \log(n) = x_n \text{ and } \log(\alpha) = \beta_0$$

$$\Rightarrow y = \beta x_n + \beta_0$$

This time we can directly record the $\alpha' s$

Table 4.2: Regression Results for Case 2

	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
β	-0.13	0.018	0.39	1.00	2.00
α	0.36	0.79	0.41	0.39	0.48
R^2	0.99	0.96	1	1	1

Notice that as expected $\beta \approx 0$ when $k_n = \log(n)$. However, this is the only case where this occurs, which validates our proposition that θ should be a function of n. Another interesting thing to note, is that when $k_n = n/2$ we get $\beta = 1$, this implies that if $k_n = n/2$, θ has a linear relationship with n.

Chapter 5

Simulations and Results

I simulated an observed k_n given n and a true fixed value θ

5.1 Method 1: Hoppe's Urn method

As seen in Chapter 2 (Equation 2.8), K_n can be constructed with the sum of n Bernoulli random variables. This is a very straightforward method as R can generate these random variables directly. This method generates k_n, to generate the full allelic partition a second method must be considered.

5.2 Method 2: Symmetric Dirichlet(K,ϕ)

A sample of size n with $\phi > 0$ from an $M-$dimensional symmetric Dirichlet(ϕ,...,ϕ) follows (Feng (2007))

$$P\left(\mathbf{A}_n = (a_1, ..., a_n)\right) = \frac{n!}{\theta^{(n)}} \frac{\phi^s \Gamma(M+1)}{\Gamma(M-k+1)} \prod_{j=1}^{n} \left(\frac{\Gamma(j+\phi)}{\Gamma(j+1)\Gamma(\phi+1)} \right)^{a_j} \frac{1}{a_j}$$

with $\theta = M\phi$ and $s = \sum_{i=1}^{n} a_i$. When $M \to \infty$ and $\phi \to 0$ in such a way that θ is fixed we get ESF. Computationally we can not let $M \to \infty$ but if we take M large enough we will get a reasonable approximation of the ESF.

To generate $\mathbf{x}= (x_1, ..., x_K)$ from a Dirichlet(K, ϕ), let $\mathbf{y}= (y_1, ..., y_K)$, where Y_i ~Gamma$(\frac{\theta}{K}, 1)$. Now, set $x_i = \frac{y_i}{\sum_{j=1}^{K} y_j}$ and we are left with the desired result. To sample from this distribution we will do the following.

- Generate $\mathbf{b}= (b_1, ..., b_n)$ where $b_i \sim$ Uniform$(0, 1)$.
- Create, $\mathbf{z}= (0, z_1, ..., z_M)$ where $z_i = \sum_{j=1}^{i} x_j$.
- Draw from $\mathbf{z}$ using $\mathbf{b}$. For example, if $0 < b_1 < z_1$ it is considered selecting allele of type 1 from the population, whereas if $z_5 < b_1 < z_6$ it is considered selecting an allele of type 6 from the population.
- Finally, count the number of distinct alleles chosen in your sample. This will be your simulated k_n.

Remark: This method can also be used to generate the allelic partition for ESF.

5.3 Tables

First, let us go over the metrics used in the following tables. Let $\bar{\theta}$ be any estimator.

I calculated the mean of $100\left(\bar{\theta}/\theta - 1\right)$, so the closer to 0 the better.

I calculated 1000 times the variance of $\left(\bar{\theta}/\theta\right)$, so again the closer to 0 the better.

Finally, I calculated 100 times the square root of the mean squared error (RMSE) of $\hat{\theta}/\theta$, given by,

$$\left(\frac{1}{m}\sum_{i=1}^{m}(X_i - 1)^2 + \frac{1}{n-1}\sum_{i=1}^{m}\left(X_i - \frac{1}{n}\sum_{i=1}^{m}X_i\right)^2\right)^{\frac{1}{2}}$$

Where m is the number of iterations and X_i is the value of $\hat{\theta}/\theta$ at iteration i. The closer the RMSE is to 0 the better the estimator.

In Method 1 I ran 1000 repetitions for each case. This ran quickly, taking less than an hour.

In Method 2, for each case I generated a Dirichlet distribution and sampled from it 100 times. I then repeated this process 10 times to get 1000 k values, the total time it took was around 13 hours.

Remark: When $k_n = n$, the value of k_n will be replaced with $k_n* = k_n - 1$. The number of times this occurs will be counted and the higher the count the more skewed the results. Since the MLE, $\theta^*, \theta^{**}, \hat{\hat{\theta}}_n$ and $\hat{\hat{\theta}}$ go to infinity when $k_n = n$.

5.3.1 Simulation Method 1

Table 5.1: n=50

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.08,-0.16,1.16,0.21)	Mean	7	130	204	141	270	NA	12360	0
$\bar{k} = 1.9$	Variance	1479	1276	2235	1562	3681	NA	5974	0
	RMSE	138	221	312	238	399	NA	14457	0
(0.21,0.02,0.98,0.85)	Mean	7	23	93	36	109	NA	3098	0
$\bar{k} = 4.1$	Variance	427	229	561	335	791	NA	1283	0
	RMSE	92	91	167	107	189	NA	3648	0
(0.16,0.46,0.54,2.04)	Mean	2	-12	151	5	61	NA	1331	0
$\bar{k} = 7.0$	Variance	229	71	573	136	322	NA	522	0
	RMSE	78	61	227	73	127	NA	1587	0
(0.30,1.17,-0.17,22.9)	Mean	4	-70	-370	-33	3	NA	140	0
$\bar{k} = 26.8$	Variance	115	1	109	35	83	NA	135	0
	RMSE	71	72	406	55	68	NA	205	0
(1.10,2.33,-1.33,1631)	Mean	-33	-99	-101	-65	-46	-100	-30	474
$\bar{k} = 48.8$	Variance	21	0	0	6	14	0	22	474
	RMSE	54	99	101	69	58	100	54	474

Table 5.2: n=100

$(\alpha,\beta_1,\beta_2,\theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n=k_n$
(0.07,-0.16,1.16,0.19)	Mean	5	119	183	125	236	NA	26517	0
$\bar{k}=1.9$	Variance	1262	1181	1965	1302	2915	NA	6904	0
	RMSE	116	173	244	181	307	NA	27581	0
(0.18,0.02,0.98,0.86)	Mean	4	19	81	26	88	NA	5989	0
$\bar{k}=4.7$	Variance	328	200	465	246	550	NA	1304	0
	RMSE	65	59	118	66	127	NA	6236	0
(0.14,0.46,0.54,2.6)	Mean	3	-16	134	-5	42	NA	2040	0
$\bar{k}=10.1$	Variance	170	54	414	85	191	NA	452	0
	RMSE	51	37	162	40	73	49	2131	0
(0.25,1.17,-0.17,42.6)	Mean	3	-74	-333	-44	-16	NA	146	0
$\bar{k}=51.9$	Variance	46	1	39	10	23	NA	54	0
	RMSE	36	74	340	48	33	NA	164	0
(0.94,2.33,-1.33,5607)	Mean	-23	-100	-100	-66	-49	-100	-21	415
$\bar{k}=98.7$	Variance	42	0	0	8	18	0	42	415
	RMSE	38	100	100	68	53	100	37	415

Table 5.3: n=500

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.05,-0.16,1.16,0.16)	Mean	1	108	154	109	196	NA	100000	0
$\bar{k} = 2.0$	Variance	974	971	1447	990	1986	NA	9560	0
	RMSE	99	146	195	148	241	NA	100000	0
(0.13,0.02,0.98,0.88)	Mean	1	12	62	14	61	NA	28504	0
$\bar{k} = 6.2$	Variance	244	171	356	180	361	NA	1738	0
	RMSE	50	43	86	45	86	NA	28518	0
(0.10,0.46,0.54,4.75)	Mean	0	-23	101	-19	14	NA	5412	0
$\bar{k} = 22.6$	Variance	61	21	146	26	51	NA	248	0
	RMSE	25	28	108	25	27	NA	5415	0
(0.19,1.17,-0.17,197.1)	Mean	1	-80	-270	-59	-42	NA	154	0
$\bar{k} = 249.6$	Variance	9	0	5	1	2	NA	11	0
	RMSE	10	80	270	59	43	NA	154	0
(0.69,2.33,-1.33,118581)	Mean	-10	-100	-100	-71	-59	-100	-10	353
$\bar{k} = 498.6$	Variance	70	0	0	7	15	0	70	353
	RMSE	29	100	100	72	60	100	28	353

Table 5.4: n=1000

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta}_n$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.05,-0.16,1.16,0.14)	Mean	-1	101	144	104	184	NA	100000	0
$\bar{k} = 2.0$	Variance	1045	1053	1509	1064	2051	NA	12689	0
	RMSE	102	146	189	147	233	NA	100000	0
(0.12,0.02,0.98,0.90)	Mean	1	10	56	11	54	NA	56126	0
$\bar{k} = 6.8$	Variance	188	137	275	141	272	NA	1685	0
	RMSE	43	38	77	39	75	NA	56126	0
(0.09,0.46,0.54,6.22)	Mean	0	-25	92	-23	7	NA	8199	0
$\bar{k} = 32.1$	Variance	37	13	85	15	28	NA	175	0
	RMSE	19	28	96	26	18	NA	8199	0
(0.17,1.17,-0.17,391.9)	Mean	0	-82	-250	-63	-49	NA	154	0
$\bar{k} = 496.7$	Variance	5	0	2	0	1	0	6	0
	RMSE	7	82	250	64	49	NA	154	0
(0.62,2.33,-1.33,466031)	Mean	-9	-100	-100	-74	63	-100	-9	339
$\bar{k} = 998.6$	Variance	73	0	0	6	12	0	74	339
	RMSE	29	100	100	74	64	100	29	339

5.3.2 Simulation Method 2

Table 5.5: n=50

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.08,-0.16,1.16,0.21)	Mean	-6	119	190	128	251	NA	12335	0
$\bar{k} = 1.8$	Variance	223	212	372	252	593	NA	962	0
	RMSE	81	132	203	142	266	NA	12400	0
(0.21,0.02,0.98,0.85)	Mean	-10	10	73	21	85	NA	3067	0
$\bar{k} = 3.7$	Variance	92	55	135	77	182	NA	295	0
	RMSE	55	42	91	51	104	NA	3084	0
(0.16,0.46,0.54,2.04)	Mean	-10	-19	131	-5	46	NA	1312	0
$\bar{k} = 6.5$	Variance	70	22	180	42	99	NA	161	0
	RMSE	45	32	140	35	61	NA	1320	0
(0.30,1.17,-0.17,22.9)	Mean	-2	-71	-364	-36	-3	NA	133	0
$\bar{k} = 26.1$	Variance	68	1	73	21	50	NA	81	0
	RMSE	32	71	366	40	28	NA	138	0
(1.10,2.33,-1.33,1631)	Mean	-34	-99	-101	-65	-47	-100	-31	482
$\bar{k} = 48.8$	Variance	23	0	0	6	15	0	24	482
	RMSE	38	99	101	66	48	100	35	482

Table 5.6: n=100

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.07,-0.16,1.16,0.19)	Mean	-32	84	138	88	181	NA	26433	0
$\bar{k} = 1.6$	Variance	148	151	251	163	366	NA	867	0
	RMSE	67	93	147	97	191	NA	26433	0
(0.18,0.02,0.98,0.86)	Mean	-2	13	72	19	78	NA	5975	0
$\bar{k} = 4.5$	Variance	69	42	97	51	115	NA	273	0
	RMSE	55	45	92	51	98	NA	5975	0
(0.14,0.46,0.54,2.6)	Mean	13	-11	148	1	52	NA	2056	0
$\bar{k} = 10.7$	Variance	48	14	107	23	51	NA	121	0
	RMSE	44	27	152	31	60	NA	2056	0
(0.25,1.17,-0.17,42.6)	Mean	1	-74	-331	-45	-18	NA	144	0
$\bar{k} = 51.5$	Variance	37	0	31	8	18	NA	43	0
	RMSE	22	74	332	46	23	NA	146	0
(0.94,2.33,-1.33,5607)	Mean	-24	-100	-100	-66	-50	-100	-22	405
$\bar{k} = 98.7$	Variance	42	0	0	8	18	0	43	405
	RMSE	31	100	100	67	51	100	30	405

Table 5.7: n=500

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.05,-0.16,1.16,0.16)	Mean	-26	81	121	82	158	NA	100000	0
$\bar{k} = 1.8$	Variance	38	39	57	39	79	NA	379	0
	RMSE	63	86	125	87	161	NA	100000	0
(0.13,0.02,0.98,0.88)	Mean	31	37	98	39	97	NA	28583	0
$\bar{k} = 7.5$	Variance	42	27	56	29	58	NA	277	0
	RMSE	53	48	103	50	103	NA	28677	0
(0.10,0.46,0.54,4.75)	Mean	-2	-25	97	-21	11	NA	5406	0
$\bar{k} = 22.2$	Variance	10	3	24	4	8	NA	40	0
	RMSE	22	27	100	24	23	NA	5425	0
(0.19,1.17,-0.17,197.1)	Mean	-1	-80	-269	-60	-43	NA	152	0
$\bar{k} = 250.8$	Variance	6	0	3	1	1	NA	7	0
	RMSE	12	80	269	60	43	NA	154	0
(0.69,2.33,-1.33,118581)	Mean	-38	-100	-100	-80	-72	-100	-37	106
$\bar{k} = 497.6$	Variance	113	0	0	12	24	0	114	106
	RMSE	51	100	100	81	73	100	51	106

Table 5.8: n=1000

$(\alpha, \beta_1, \beta_2, \theta)$	Criteria	$\tilde{\theta}$	$\hat{\theta}$	$\hat{\theta_n}$	θ^*	θ^{**}	$\hat{\hat{\theta}}_n$	$\hat{\hat{\theta}}$	$n = k_n$
(0.05,-0.16,1.16,0.14)	Mean	-12	92	130	93	167	NA	100000	0
$\bar{k} = 1.9$	Variance	97	97	139	98	188	NA	1166	0
	RMSE	106	103	139	104	176	NA	100000	0
(0.12,0.02,0.98,0.90)	Mean	5	14	62	15	60	NA	56139	0
$\bar{k} = 7.1$	Variance	24	17	34	18	34	NA	211	0
	RMSE	37	33	68	34	66	NA	56372	0
(0.09,0.46,0.54,6.22)	Mean	-3	-27	87	-24	4	NA	8192	0
$\bar{k} = 31.3$	Variance	7	3	17	3	6	NA	35	0
	RMSE	19	29	90	27	17	NA	8227	0
(0.17,1.17,-0.17,391.9)	Mean	-1	-82	-249	-64	-50	NA	153	0
$\bar{k} = 494.6$	Variance	3	0	1	0	1	NA	4	0
	RMSE	11	82	250	64	50	NA	155	0
(0.62,2.33,-1.33,466031)	Mean	-78	-100	-100	-94	-91	-100	-77	2
$\bar{k} = 994.1$	Variance	23	0	0	2	4	0	23	2
	RMSE	79	100	100	94	91	100	79	2

Chapter 6

Discussion

6.1 Simulation Results

6.1.1 Simulation Method Comparison

In tables 5.1,5.2,5.3 and 5.4 I used the Hoppe's Urn simulation method, while in tables 5.5,5.6,5.7 and 5.8 I used the Dirichlet simulation method. To compare the two methods we will examine the returned value k and the variance of K_n. The percentage error will be used to determine how close the simulated values of k_n are to the theoretical values. Let R be the percentage error calculated by,

$$R = 100 \times \left| \frac{\text{simulated value} - \text{theoretical value}}{\text{theoretical value}} \right|$$

The value of R is given by the following table

Table 6.1: Simulation Method Comparison using k_n

Method 1	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n-1$
n=50	5.26	4.58	1.02	6.72	0.41
n=100	5.26	2.02	0.99	3.66	0.30
n=500	0.00	0.24	1.06	0.16	0.08
n=1000	0.00	1.58	1.49	0.66	0.04
Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n-1$
n=50	11.11	5.73	8.79	4.21	0.41
n=100	25.00	2.34	6.54	2.91	0.30
n=500	11.11	17.14	0.72	0.32	0.28
n=1000	5.26	2.71	1.03	1.09	0.49

In Table 6.1 above, we can see that for method 1 the simulation remains within 7% of the theoretical values across the board. While in method 2 there continues to be errors over 10% until we reach $n = 1000$. This indicates that the mean of our simulated k_n lies closer to the theoretical value when using method 1.

We will now look at the variance of K_n under both methods. Now, I did not directly calculate the variance of K_n. However, by looking at 1000 times the variance of $\frac{\hat{\theta}}{\theta}$ they can be inferred. Remember that $\hat{\theta} = \frac{k_n}{\log(n)}$ and θ is given. So, var$[\hat{\theta}/\theta]$ =var$[K_n]/\theta\log(n)$.

Table 6.2: Variance of $\hat{\theta}/\theta$ times 1000 under both Simulation Methods

Method 1	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	1276	229	71	1	0
n=100	1181	200	54	1	0
n=500	971	171	21	0	0
n=1000	1053	137	13	0	0

Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	212	55	22	1	0
n=100	151	42	14	0	0
n=500	39	27	3	0	0
n=1000	97	17	3	0	0

In Table 6.2 above, we can clearly see that the variance in method 2 is significantly lower than in method 1. However, despite the larger variance, method 1 does provide closer results in under 1/10th the time. Therefore, method 1 is recommended when simulating k_n.

6.1.2 Estimator Comparison

Since simulation method 1 was shown to be the more effective method, the estimators will be judged based solely on the results of the method 1 simulations.

We will begin by looking at the estimator $\hat{\hat{\theta}}_n$.

Table 6.3: RMSE of θ^{ii} times 100

Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	NA	NA	NA	NA	100
n=100	NA	NA	NA	NA	100
n=500	NA	NA	NA	NA	100
n=1000	NA	NA	NA	NA	100

The value NA indicates that there is no real solution. Remember that $\hat{\hat{\theta}}_n$ has no real values for $k_n < 13n/16$. In the simulation, if at any iteration $k_n < 13n/16$ the result would be NA. Also, remember that this estimator was designed for cases where $\beta_1 = 1$ and $\alpha > 1$. This case never occurred in our simulations, therefore the performance of $\hat{\hat{\theta}}_n$ should not be judged based on these results.

Now, let us look at $\hat{\theta}_n$.

Table 6.4: RMSE of $\hat{\theta}_n$ times 100

Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	312	167	227	406	101
n=100	244	118	162	340	100
n=500	195	86	108	270	100
n=1000	189	77	96	250	100

These results are considered poor since any estimator with an RMSE value above

100 is considered a poor estimator. However, we can see that as n increases the estimator improves across the board.

Remember that in our construction $\beta_2 = 1 - \beta_1$, so the approximation to $\mathrm{E}[K_n]$ is controlled by $\left(\frac{\log(n)}{n}\right)^{1-\beta_1}$. Therefore, the smaller the β_1 the better, hence the better performance of $k_n = \log(n)$ where $\beta_1 \approx 0$.

There is another problem, when constructing this estimator there was a $-log(\alpha)$ that we ignored since its effect becomes negligible for large n. Of course, our largest n is only 1000 and our $\alpha < 1.2$ throughout, especially for the $k_n = 2$ case where $\alpha < 0.1$. Therefore, this $\log(\alpha)$ may have a significant influence when dealing with small n. While for very large n this estimator may be valid, it still cannot compete with the MLE under these circumstances.

Now, let us examine the estimators θ^* and θ^{**}.

Table 6.5: RMSE of (θ^*, θ^{**}) times 100

Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	(238,399)	(107,189)	(**73**,127)	(**55**,**68**)	(69,58)
n=100	(181,307)	(66,127)	(**40**,73)	(48,**33**)	(68,53)
n=500	(148,241)	(**45**,86)	(**25**,27)	(59,43)	(72,60)
n=1000	(147,233)	(**39**,75)	(26, **18**)	(64,49)	(74,64)

In bold are the times where the estimator was an improvement on the MLE. We can see that when $k_n = \sqrt{n}$ one of our estimators always outperforms the MLE. For $k_n = \log(n)$, θ^* performs well for larger values of n and improves as n grows. In the

$k_n = n/2$ case, θ^* performs well for low values of n but does not improve as n grows. We can also see that in the $k_n = \sqrt{n}$ case as n grows larger, θ^{**} begins to outperform θ^*.

Now, let us look at our final estimator $\hat{\hat{\theta}}$.

Table 6.6: RMSE of $\hat{\hat{\theta}}$ times 100

Method 2	$k_n = 2$	$k_n = \log(n)$	$k_n = \sqrt{n}$	$k_n = n/2$	$k_n = n - 1$
n=50	> 1000	> 1000	> 1000	205	54
n=100	> 1000	> 1000	> 1000	164	37
n=500	> 1000	> 1000	> 1000	154	28
n=1000	> 1000	> 1000	> 1000	154	29

Here, we can see that this estimator is only reasonable in the extreme case of $k_n = n - 1$. In fact, by looking at Tables 5.1,5.2,5.3, and 5.4 we can see that $\hat{\hat{\theta}}$ and the MLE are almost identical when $k_n \approx n - 1$ for every value of n. Therefore, for this extreme case $\hat{\hat{\theta}}$ could be used in place of the MLE.

6.2 Conclusion

In this thesis, the mutation parameter of Ewens Sampling Formula was discussed. Asymptotic approximations to the MLE were explored, and as a result closed form estimators were introduced. The parameter $\theta_n = \alpha n^{\beta_1}(\log n)^{1-\beta_1}$ was proposed for cases where θ may not be fixed. α, and β_1 were approximated for several cases, by using least squares regression to fit θ_n against the MLE. It was established that in a sample where $k \geq \sqrt{n}$, θ should not be assumed to be fixed. Two simulation techniques were carried out, one that only simulated an observed value k, and one that simulated the allelic partition of ESF. These two methods were compared by calculating the percentage error between the theoretical and simulated values of k. The MLE was also compared against the new estimators using these simulations. The results indicate that method 1 achieves values closer to the theoretical values while running 10 times faster. Our two closed form estimators θ^* and θ^{**} were also found to perform as well or better than the MLE when $k \approx \sqrt{n}$.

6.3 Future Work

As for future work, one could look more closely at the case where β_1 and β_2 do not have a relationship. One could also construct and simulate ESF using truncated stick breaking methods, which was not discussed. The estimators θ^* and θ^{**} could be further explored, to identify when exactly to use each one. Other regimes of K_n could be explored. Finally, one could look at alternate constructions of θ_n.

Appendix A

Your Appendix

A.1 Gamma Function

The Gamma function is defined by

$$\begin{aligned} \Gamma(n) &= (n-1)!, \text{ for positive integer } n \\ \Gamma(t) &= \int_0^\infty x^{t-1} e^{-x}\, dx, \text{in general} \end{aligned}$$

where $n! = n \times (n-1) \times (n-2) \times \cdots \times 1$

A.2 Digamma Function

The Digamma function is defined by

$$\psi(x) = \frac{d}{dx} \ln \Big(\Gamma(x)\Big) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Also $(\psi(\theta + n) - \psi(\theta)) = \sum_{j=0}^{n-1} \frac{1}{\theta+j}$ for positive integer n.

A.3 The Rising Factorial

The rising factorial is defined by

$$\theta^{(n)} = \theta \times (\theta+1) \times \cdots \times (\theta+n-1) = \frac{\Gamma(\theta+n)}{\Gamma(\theta)}$$

and

$$\begin{aligned}\frac{d(\theta^{(n)})}{d\theta} &= \frac{\Gamma'(\theta+n)\Gamma(\theta) - \Gamma(\theta+n)\Gamma'(\theta)}{\Gamma(\theta)^2} \\ &= \frac{\Gamma(\theta+n)}{\Gamma(\theta)}\left[\frac{\Gamma'(\theta+n)}{\Gamma(\theta+n)} - \frac{\Gamma'(\theta)}{\Gamma(\theta)}\right] \\ &= \theta^{(n)}\left(\psi(\theta+n) - \psi(\theta)\right)\end{aligned}$$

A.4 Asymptotic Expansions

Using the Taylor series expansion near 0.

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - ... = O(x)$$

A.5 Integral Approximation

$$\begin{aligned}
\sum_{j=0}^{n-1} \frac{\alpha}{\alpha n + j} \geq & \alpha \int_0^{n-1} \frac{1}{\alpha n + x} dx \\
& \text{apply the transformation } y = \alpha n + x \\
= & \alpha \int_{\alpha n}^{\alpha n + n - 1} \frac{1}{y} dy \\
\approx & \alpha [\log(\alpha n + n) - \log(\alpha n)] \text{ for large n} \\
= & \alpha [\log(\alpha + 1) + \log(n) - \log(\alpha) - \log(n)] \\
= & \alpha \log \left(1 + \frac{1}{\alpha}\right)
\end{aligned}$$

Bibliography

Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.*, **31**, 1–19.

Ethier, S. and Kurtz, T. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. Appl. Probab.*, **13**, 429–452.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoret. Popn Biol*, **3**, 87–112.

Ewens, W. (2004). *Mathematical Population Genetics I. Theoretical Introduction.* Springer.

Feng, S. (2007). Large deviations associated with Poisson-Dirichlet distribution and Ewens sampling formula. *The Annals of Applied Probability*, **17**, 1570–1595.

Feng, S. (2010). *The Poisson-Dirichlet Distribution and Related Topics.* Springer.

Fisher, R. (1930). *The Genetical Theory of Natural Selection.* Oxford University Press.

Hoppe, F. (1984). Pó lya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94.

Kingman, J. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser.*, **37**, 1–22.

Kotz, S., Balakrishnan, N., and Johnson, N. (2000). *Continuous Multivariate Distributions, Models and Applications.* Wiley-Interscience.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97–159.