School of Mathematic and Statistics

Age of a Rare Mutation

Zhou Lu

2279337L

March 2018

Acknowledgments

This is the author's single honours statistics project, written under the supervision of Dr Alexey Lindo and Dr Vincent Macaulay at the University of Glasgow.

The author wishes to express his gratitude to Dr Alexey Lindo and Dr Vincent Macaulay for suggesting the problem and for many stimulating conversations.

Contents

Chapter 1

Introduction

1.1 Background

A gene is a sequence of DNA or RNA which codes for a molecule that has a function, and mutation can result in many different types of changes in sequences. These changes in gene encode slightly different versions of a protein, which cause different features, for a deeper discussion of process of mutation, I refer the reader to [1]. Mutations can be subdivided into germline mutations, which can be passed on to descendants, and somatic mutations (i.e. also called acquired mutations) which are not usually transmitted to descendants, see [12] for more details. Obviously, the mutation discussed in this paper is a germline mutation.

The reason for why mutation is considerable is that it has a significant effect on the process of the species evolution, because mutation is the source of novelty, creating new forms and new species, potentially instantaneously. Those mutants who are more attuned to the environment are likely to have more chance to reproductive, then the genes for these mutations are passed on to future generations, leading to the birth of new species or the extinction of less adapted species, which is called "nature selection", and mutation also leads to diversity of species, see [5] and the references given there.

However, mutation is directionless, as mutations occur randomly, which means that mutations in genes can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely, the specific explanation can be found in [21]. Hence, it is not easy to predict when and how does the mutation occur, as [22] indicates that 66 percent of cancer-causing mutations are random and 29 percent are due to the environment, even a precise equation have been deduced, the mutation may cause beneficial or harmful effect, so prediction on the mutation might be pointless. Therefore, the reverse problem might be more valuable, for instance, given the number of replications, how long ago did mutant from the first arise? There are a number of situations where this problem might arise. For example, if biologists want to know historical length of time the plant or animal has existed its present form, they can collect approximate the number of a certain specials in existence and very complete data on their rate of reproduction, Then, the formulation given in following context might be appropriate. Another source of applications would be genetics, where a geneticist might be interested in estimating how long ago a mutation he observes to took place, see for instance [15], the age of a rare mutation, leading to strong resistance against human immunodeficiency virus, was estimated. Scientists might be inspired from the original of this mutation, so as to find a cure for the disease.

This paper arises in connection with the study of rare protein variants in the American Indians populations,see [18] for more details. Research considers as a rare variant occurring only in a single South American tribe. I follow [24] in assuming that such mutation has presumably arisen since tribal differentiation (i.e. around 5500 years ago), and are probably descended from a single mutant. In addition, this rare variant has not spread to the general South American Indian population, because of the low level of inter-tribal migration. This paper uses the distribution of offspring numbers for Yanomama Indians to obtain the required distribution of mutant replicates of a mutant gene in a heterozygous parent, which is from [16].

1.2 Project Description and Aims

In this paper, the Galton-Watson branching process is a appropriate model. Let $Z_0 = 1, Z_1, Z_2, ...$ be the sizes of a population for successive generations, and the population size is assumed to behave like a Galton-Watson branching process with probability generating function $f(s) = \sum p_k s^k$, with $p_k = P(Z_1 = k)$. In detail, the population begins with the zeroth generation with a single individual who lived a unit length of time (i.e one generation), and then producing a random number offspring based on the probability distribution p_k. All offspring are assumed to behave independently of one another and of the parents , and in turn reproduce offspring in the same rule as their parents. More specific processes are discussed in [17]. Besides, this paper assumes that the probability distribution of the number of offspring of a single individual $f(s)$ is already known.

Researcher knows only population size, but does not know its age in term of generations, the whole project can be described as:

1 Model population size with Glaton-Watson branching process with linear fractional offspring distribution.

2 Infer maximum likelihood estimation (MLE) estimator for n and asymptotic confidence interval.

3 Calculate MLE estimator and confidence interval for different situation.

4 Assess the coverage of the confidence interval for finite generation sizes.

The aim of this paper is using the distribution of offspring number at present to achieve the target distribution of mutant replicates, this distribution can give us an approximate interval estimation of age of the mutation.

1.3 Data

Research is interested in the variant named Yanomama albumia variant Yan-2 , this variant has achieved polymorphic frequencies throughout the tribe, this phenomenon is usually shown in an old variant, so this paper assumes that this mutation happened at least 2000 years, this is proved by [18]. In addition, this variant is only found in this tribe, so it must have arisen since tribal differential.

The survival population usually includes more than one generation, so it is crucial to divide the whole population into different generations. In order to solve this problem, research defines the "adult generation". It means those individuals who are in reproductive age currently. In this situation, the research only includes the number of heterozygote mutant offspring surviving to adulthood from a given adult heterozygote and disregards mutants who are not surviving to adulthood in the initial generation. Extensive sampling of 47 widespread Yanomama villages provides an estimate of 875 replicates of the variant gene in the current adult population, the data is collected from [20]. Although there are some individuals homozygous for the Yan-2 variant, these sufficiently few in number for the problem of nonindependent replication to be disregarded in any preliminary analysis.

Chapter 2

Methods

2.1 Galton-Watson Branching Process

Galton-Watson Branching Process is a branching stochastic process, the concept is a population of individual (i.e. which may denote people, planets or animals) evolves in discrete time n (i.e. $n = 0, 1, 2, 3, ...$). This evolution has the following regulation:

- Each n^{th} generation individuals produces a random number (i.e. possible zero) of individuals, named offspring in the $(n+1)^{th}$ generation.
- The offspring counts $Z_\alpha, Z_\beta, Z_\gamma, ...$ for distinct individuals $\alpha, \beta, \gamma, \ldots$ are mutually independent and also independent of the offspring counts of individuals from earlier generations.
- The offsprings are identically distributed, with common distribution $\{P_k\}_{k\geq 1}$, and Z_n of the Galton-Watson process at time n is the number of individuals in the n^{th} generation.

See [10] and the references given there.

The Galton-Watson branching process performs well in describing a transmission in genetics, and the model is an appropriate choice in estimating the age of a rare mutation.

2.1.1 Basic Assumptions

In this paper, all the estimation procedures and tests of hypotheses given will be stated conditional upon the offspring not being extinct. This means all confidence intervals and significance levels are calculated given the condition that the number of offspring is more than 0, because if extinction had happened, there is no inference problem arisen, for the situation that population had happened, other method should be used.

Naturally, the branching process is continuous, it means that each generation is not independent, for instance, as the figure 2.1 is shown that one senior child of the couple living in the fist generation, and the other younger child is living in the fourth generation. This phenomena is named "overlapping generation", and it is realistic because of children are born in different time. However, using Galton-Watson branching process requests us to consider generations discretely, and the overlapping between generations might be ignored in a sufficiently large generations. Therefore, the paper considers the discrete branching process instead of continuous branching process, which is shown in the figure 2.2. Beginning with the first mutant in the original generation, and then this original mutant reproduces a random number offspring. There is no overlapping among different generations.

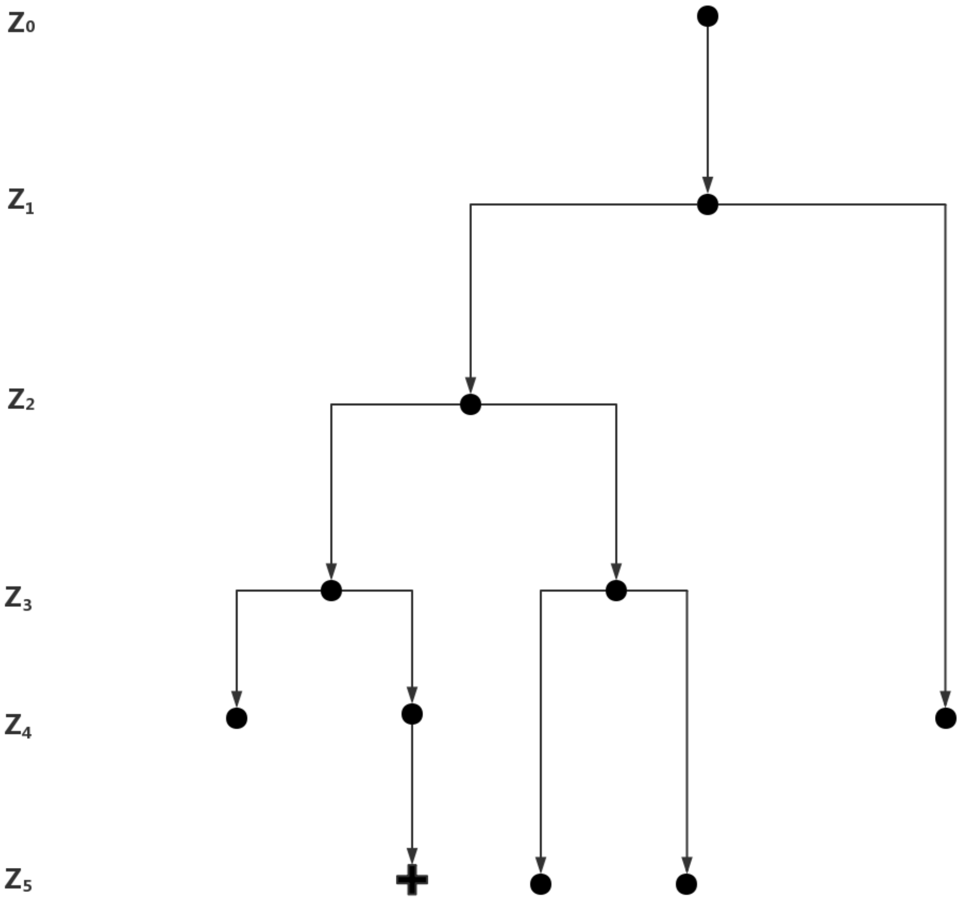

Figure 2.1: The family tree of the continue branching process; $Z_0, Z_1, Z_2, Z_3, Z_4, Z_5$ are population of the first, second, third, fourth, fifth generation; the circle means the offspring in the branching process; the cross means extinction.

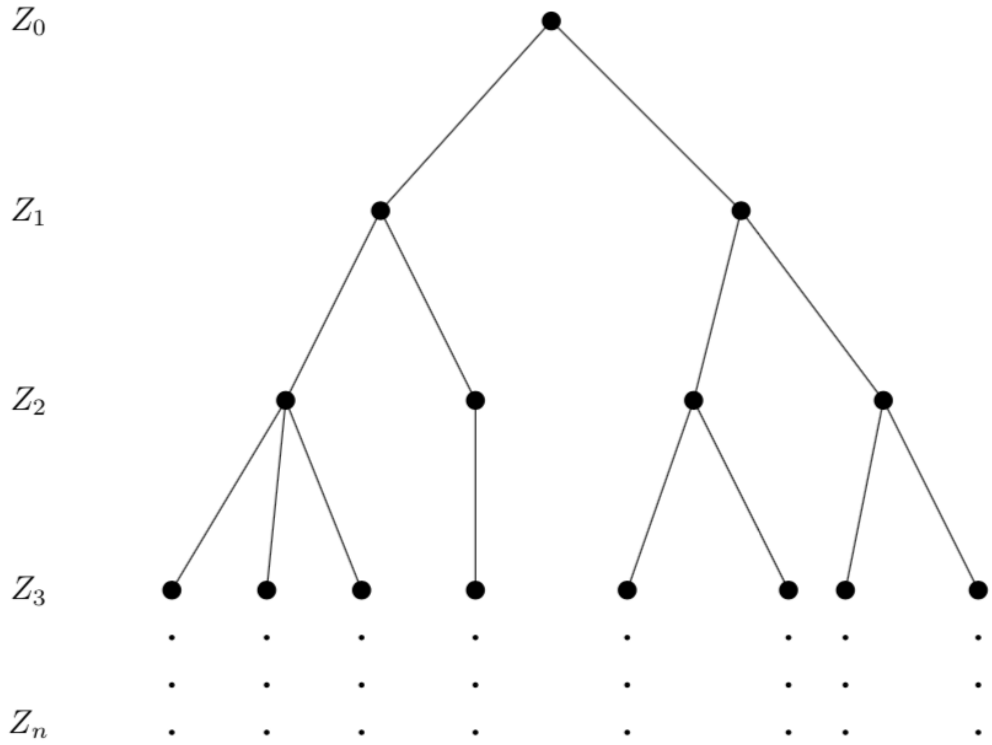

Figure 2.2: The family tree of the discrete branching process; $Z_0, Z_1, Z_2, Z_3, Z_4, Z_5$ are population of the first, second, third, fourth, nth generation;the circle means the offspring in the branching process.

2.1.2 Generating Function

An important tool in the analysis of the process is the generating function, which could be written as:

$$f(s) = \sum_{i=0}^{\infty} p_i s^i = p_0 + p_1 s + p_2 s^2 + \cdots + p_i s^i + \ldots, |s| \leq 1, \tag{2.1}$$

and it iterates:

$$\begin{aligned} f_0(s) &= s, \\ f_1(s) &= f(s), \\ f_n(s) &= f[f_{n-1}(s)], \end{aligned} \tag{2.2}$$

where s is complex in general, but will be assumed real in this paper.

Letting $P_n(i, j)$ be the n-step transition probabilities, and observe that

$$f(s) = \sum_j P(1, j) s^j;$$

$$[f(s)]^i = \sum_j P(i,j)^j, i \geq 1. \tag{2.3}$$

Using the Chapman- Kolmogorov equation, it is easily to get:

$$\begin{aligned}\sum_j P_{n+1}(1,j)s^j &= \sum_j \sum_k P_n(1,k)P(k,j)s^j \\ &= \sum_k P_n(1,k) \sum_j P(k,j)s^j \\ &= \sum_k P_n(1,k)[f(s)]^k.\end{aligned}$$

Thus letting $\sum P_n(1,j)s^j = f_{(n)}(s)$, then it is shown that

$$f_{(n+1)}(s) = f_{(n)}[f(s)].$$

Hence, it follows by induction that

$$f_{(n)} = f_n(s), \tag{2.4}$$

a crucial formula. From equation (2.3) and (2.4):

$$\sum_{j=0}^{\infty} P_n(i,j)s^j = [f_n(s)]^i. \tag{2.5}$$

Therefore, if the generating function (2.1) is known, it seems that every probabilities in every generations could be calculated by the equation (2.4). However, iterating this equation in this form is inefficient, the process would be written as:

$$f_2(s) = p_0 + p_1 f(s) + p_2 f(s)^2 + \cdots + p_i f(s)^i + \ldots;$$

$$f_n(s) = p_0 + p_1 f_{n-1}(s) + p_2 f_{n-1}(s)^2 + \ldots. + p_i f_{n-1}(s)^i + \ldots;$$

It is clear that $f_n(s)$ can not been explicitly computed in the original form, and it could be written in the form of linear fractional, this was proved by [6].

Linear Fraction Case

Intuitively, there is a relationship exists between the generations, so the project assumes the geometric distribution with parameters b and c, where b and c are positive and b is less than $(1-c)$, so rewritten the p_k and p_0:

$$p_k = bc^{k-1}, k > 0; \tag{2.6}$$

$$p_0 = 1 - \sum_{i=1}^{\infty} p_i = 1 - \frac{b}{1-c}. \tag{2.7}$$

Substitute the (2.6) and (2.7) into the 2.1, get the linear fraction, as

$$f(s) = 1 - \frac{b}{1-c} + b\sum_{i=1}^{\infty} c^{i-1}s^i. \tag{2.8}$$

Since using the geometric theorem needs the sum from zero to infinity, the equation (2.8) could be transformed as:

$$f(s) = 1 - \frac{b}{1-c} + bs\sum_{j=0}^{\infty} (cs)^j. \tag{2.9}$$

Then,

$$f(s) = 1 - \frac{b}{1-c} + \frac{bs}{1-cs} \tag{2.10}$$

Mean and Variance

Let the m and σ^2 denote the mean and variance of the offspring number distribution, which could be written as:

$$m = E(Z) = \sum_{k=1}^{\infty} kp_k; \tag{2.11}$$

$$\begin{aligned}\sigma^2 = Var(Z) &= E(Z-m)^2 \\ &= \sum_{k=0}^{\infty} (k-m)^2 p_k.\end{aligned}$$

Writing these two parameters m and σ^2 in linear fraction case:

$$m = f^{'}(1) = \sum_{k=1}^{\infty} kp_k = b/(1-c)^2;$$

$$\sigma^2 = b(1-b-c^2)/(1-c)^4.$$

Classification of the Galton-Watson Process

The Galton-Watson branching process divides the model into three cases which are based on m:

- $m > 1$, the supercritial case;
- $m = 1$, the critical case;
- $m < 1$, the subcritcal case.

Mean m denotes the average number of offspring produced by a single individual and this paper will discuss three case separately, the more detailed explanation is in the following context.

Extinct Probability

Another important parameter π named the extinction probability which denotes the probability of a population eventually extincts, so

$$\pi = P(Z_\infty = 0).$$

As the figure 2.3 shown that the extinction probability of the $\{Z_n\}$ process is the smallest non-negative root π of the equation $t = f(t)$. It is 1 if $m \leq 1$ and less than 1 if $m > 1$.

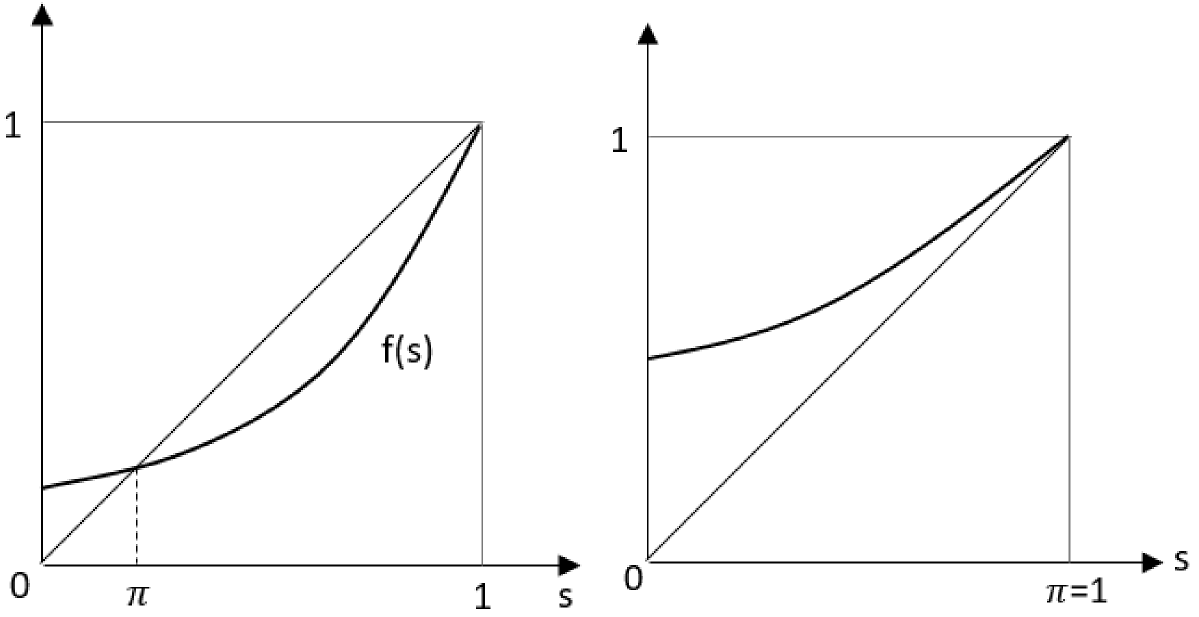

Figure 2.3: The plot of linear fraction generating function when $m > 1$ (left) and the plot of linear fraction generating function when $m \leq 1$ (right). In the left figure, the curve is the graph of function $f(s)$, the straight line is graph of function $f(x) = x$, and they have two intersections π and 1; in the right figure, the curve is the graph of function $f(s)$ when $m \leq 1$, and the straight line is graph of function $f(x) = x$, they have am intersections 1 (the plots are adapted from [6]).

The figure 2.4 is shown that for any value t belongs to [0,1], $f_n(t) = \pi$ as $n \to \infty$, this phenomenon is hold in the supercritical, critical and subcritical case.

If $f(s)$ is a fractional linear generating function, the equation $f(s) = s$ could be written as:

$$1 - \frac{b}{1-c} + \frac{bs}{1-cs} = s;$$

Therefore,

$$\pi = \min\Big\{1, \frac{1-b-c}{c(1-c)}\Big\}. \tag{2.12}$$

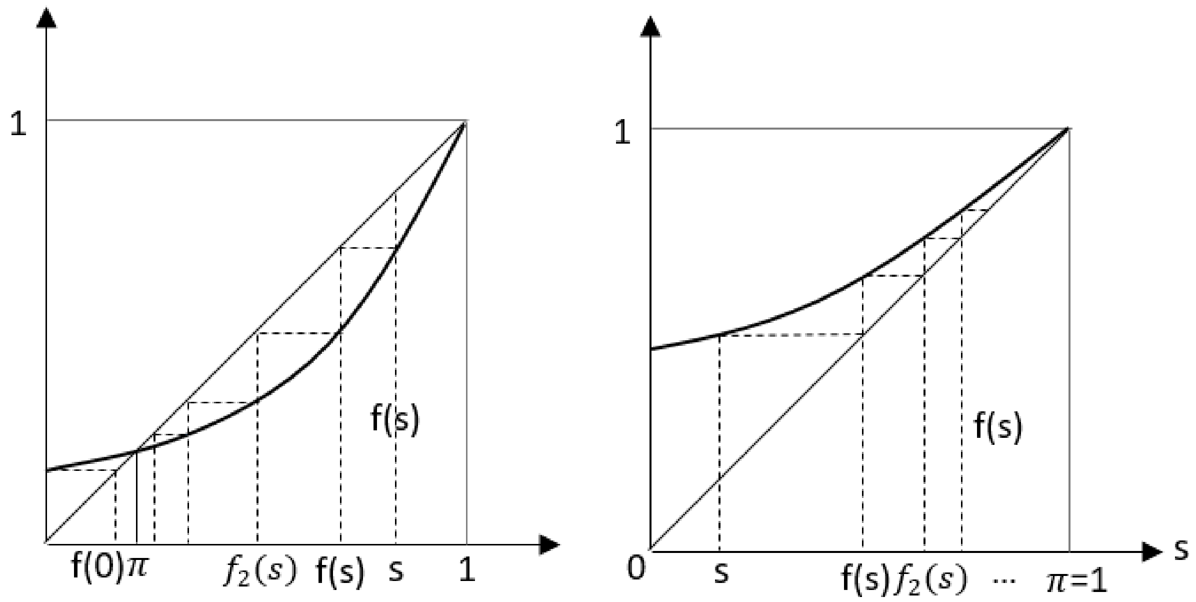

Figure 2.4: The plot of iteration of linear fraction generating function when $m > 1$ (left) and the plot of iteration of linear fraction generating function when $m \leq 1$ (right). In the left figure, the curve is the graph of function $f(s)$ when $m > 1$, and the straight line is graph of function $f(x) = x$, for any positive value $s \in [0, 1]$, the iteration finally reaches π; in the right figure, The curve is the graph of function $f(s)$ when $m \leq 1$, and the straight line is graph of function $f(x) = x$, for any positive value $s \in [0, 1]$, the iteration finally reaches π (the plots are adapted from [6]).

The Linear Fraction for the n^{th} Generating Function

The equation (2.10) is the generating function for the first generation, and further generating function could deduces by [6]:

for any two points u,v

$$\frac{f(s) - f(u)}{f(s) - f(v)} = \frac{s - u}{s - v}\frac{1 - cv}{1 - cu}. \tag{2.13}$$

According to previous extinction probability part, the equation $f(s) = s$ has root π and 1. If suppose $u = \pi$ and $v = 1$, then for $m \neq 1$, the above formula becomes

$$\frac{1 - cv}{1 - cu} = \frac{1 - c}{1 - c\pi} = \lim_{s \to 1} \Big(\frac{f(s) - \pi}{s - \pi}\Big)\Big(\frac{f(s) - 1}{s - 1}\Big)^{-1} = \frac{1}{m};$$

and hence equation (2.13) becomes

$$\frac{f(s) - \pi}{f(s) - 1} = \frac{1}{m}\frac{s - \pi}{s - 1}.$$

Iterating this, getting

$$\frac{f_n(s) - \pi}{f_n(s) - 1} = \frac{1}{m^n}\frac{s - \pi}{s - 1},$$

which can be solved explicitly for $f_n(s)$, and answer is

$$f_n(s) = 1 - m^n\Big(\frac{1-\pi}{m^n-\pi}\Big) + \frac{m^n(\frac{1-\pi}{m^n-\pi})^2 s}{1-(\frac{m^n-1}{m^n-\pi})} \quad if \quad m \neq 1. \tag{2.14}$$

The equation (2.14) is same as

$$f_n(s) = 1 - \frac{b(n)}{1-c(n)} + \frac{b(n)s}{1-c(n)s}, \tag{2.15}$$

where

$$b(n) = m^n\Big(\frac{1-\pi}{m^n-\pi}\Big)^2,$$

$$c(n) = \Big(\frac{m^n-1}{m^n-\pi}\Big);$$

This form is very similar to equation 2.10.

If m=1, then $b = (1-c)^2$ and $\pi = 1$. Then

$$f(s) = \frac{c-(2c-1)s}{1-cs},$$

which can be iterated to yield

$$f_n(s) = \frac{nc-(nc+c-1)s}{1-c+nc-ncs}.$$

2.2 Maximum Likelihood Estimation

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations, see more details in [2]. In this paper, probability function could be generated from $f_n(s)$, and then get the MLE estimator $\hat{n}$.

Supercritical Case

Basically, supercriticla case (i.e $m > 1$) is the most common situation in the population structure, so this paper firstly discuss the problem of estimating n for supercritical case (i.e. $m > 1$). The expected population size in nth generation is m^n, so when conditional on non-excitinction, the population grows exponentially in the supercritical case, which is showed in the figure 2.5. Suppose P_n denotes the probability of n is the actually age in generation of the population under the conditional distribution of Z given $Z > 0$, the reason why the probability should be given a condition is that the present inference problem would not have

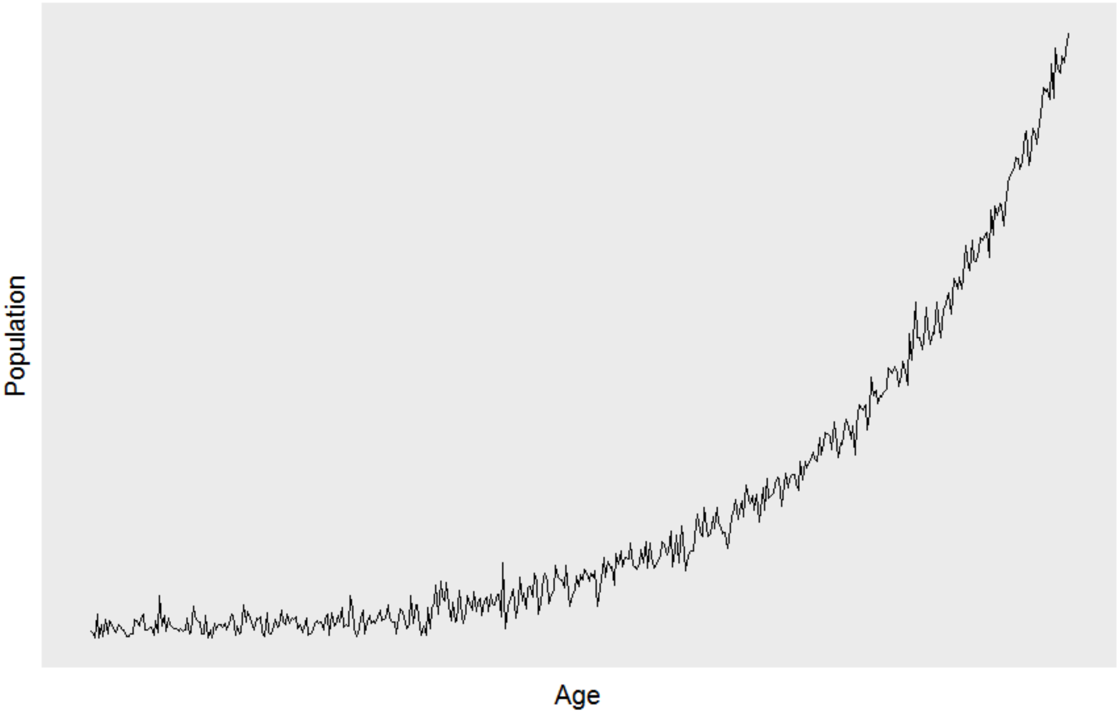

Figure 2.5: The exponential growth of population in supercritical case when conditional on non-extinction.

arisen is $Z = 0$, and if the equation (2.15) means the probability generating function of the unconditional distribution of Z, p_n could be written as:

$$\begin{aligned} P_n(Z = k) &= P(Z = k|Z > 0) \\ &= \frac{f_n^{(k)}(0)}{k!\{1 - f_n(0)\}}, (k = 1, 2, ...); \end{aligned} \tag{2.16}$$

where $f_n^{(k)}$ denotes the k^{th} derivative of $f_n(s)$, then substituting the equation (2.15) into the equation (2.16), rewritten as:

$$P_n(Z = k) = P(Z = k|Z > 0) \tag{2.17}$$

$$= \{1 - c(n)\}c(n)^{k-1}, (k = 1, 2, ...) \tag{2.18}$$

$$P(Z = 0) = f_n(0) = 1 - b(n)/\{1 - c(n)\}.$$

Intuitively, MLE might be used here to estimate the value of n. If the value of k is given, then maximizing the equation (2.17), getting the estimated value $\hat{n}$. One thing to note is the $\hat{n}$ is not an integer thought in fact n must be integer; in the MLE, this paper treats n as a parameter varying continuously from zero to infinity. If desired, the $\hat{n}$ could be rounded off.

In order to simplify the derivation, one common way is to use logarithm, then

$$log\{P_n(Z = k)\} = \log(1 - \pi) + (k - 1)\log(m^n - 1) - k\log(m^n - \pi). \tag{2.19}$$

Differentiating the equation (2.19):

$$\frac{k-1}{m^n-1}m^n\log(m)-\frac{k}{m^n-\pi}m^n\log(m)=0; \tag{2.20}$$

Therefore

$$\hat{n}=\frac{\log\{(1-\pi)Z_n+\pi\}}{\log(m)}; \tag{2.21}$$

Then, the second derivation of the equation (2.19) is:

$$\log m(1-k)\frac{m^n}{(m^n-1)^2}+((\log m)^2)(k-1)\frac{m^n}{m^n-1}+ \\ (\log m)^2k\frac{m^n}{(m^n-\pi)^2}-(\log m)^2k\frac{m^n}{m^n-\pi}. \tag{2.22}$$

Since the $m>1$, so the equation (2.22) is negative, which can infer that $\hat{n}$ gives the maximum value for equation (2.19). The MLE is unbiased estimator, but

$$\lim_{n\to\infty}P(|n-\hat{n}|>\epsilon)\neq 0 \text{ for any } \epsilon>0.$$

Therefore, $\hat{n}$ is not consistent.

α-Consistence

Although the $\hat{n}$ do not have the property of consistency, it could be okay when the $\hat{n}$ is a relatively consistent estimate of n for large n, because research only collects the size of the present population and ignores any past random fluctuations, there is no necessary for strict consistence. Then, another lower requirement consistence would be selected, called α-consistence which was found in [8], it is defined that estimate $\hat{\theta}$ of θ is α-consistent if

$$\lim_{n\to\infty}\frac{\hat{\theta}-\theta}{\theta^\alpha}\xrightarrow{p}0.$$

Therefore,

$$\begin{aligned}\hat{n}-n&=[\log\{(1-\pi)Z_n+\pi\}-n\log m]/\log m\\&=\log\{(1-\pi)m^{-n}Z_n+\pi m^{-n}\}/\log m.\end{aligned}$$

It is clear that $m^{-n}\to 0$ when $n\to\infty$, and the most confused part is $m^{-n}Z_n$. In the case of supercritical, $m>1$ and $\sum k(\log k)p_k<\infty$, and all probability equations are conditional on $Z_n>0$. Therefore, supposing M is a random variable with a continuous density function on $(0,\infty)$, the $m^{-n}Z_n$ is almost surely convergence to M. Hence, $\hat{n}-n$ is almost surely convergence to $\log\{(1-\pi)M\}/\log m$, and for any $\theta>0$, θ^α is a constant, so

$$\lim_{n\to\infty}\frac{\hat{n}-n}{n^\alpha}\xrightarrow{a.s.}0.$$

Following the definition of α-consistence, $\hat{n}$ is approximately consist to n. For a further discussion of proof, I refer the reader to [23].

Confidence Interval

The reason for why proofing the $\hat{n}$ is α-Consistence is to calculate the confidence interval for $\hat{n}$. Define the distribution function for the random variable M, where

$$F(z) = P(M \leq z).$$

Then draws two parameters a_1 and a_2 from the distribution $F(z)$, let

$$\begin{aligned} a_1 &= (1-\pi)F^{-1}(\alpha_1) \\ a_2 &= (1-\pi)F^{-1}(\alpha_2), \end{aligned}$$

Where α_1 and α_2 are two positive variables which satisfies $\alpha_1+\alpha_2 = \alpha$, and α is the confidence level. Thus, the n has an approximate (1-α) confidence interval:

$$\left[\hat{n} - \frac{\log a_2}{\log m}, \hat{n} - \frac{\log a_1}{\log m}\right]. \tag{2.23}$$

This confidence interval is an approximate estimate, approximate means that the probability of confidence interval includes the true might not equal to $(1-\alpha)$ when n is small, but as n increases, the probability will convergence to the $(1-\alpha)$.

However, distribution of $F(z)$ is not easy to find, one solution is the Laplace-Stieltjes transform of $F(z)$ which is found in [10], $f(s) = E\{\exp(-sM)\}$ is the unique root of the functional equation:

$$f(ms)(1-\pi) + \pi = \psi\{f(s)(1-\pi) + \pi\}, \tag{2.24}$$

satisfying

$$f^{'}(0) = -E(M) = -(1-\pi)^{-1}.$$

The equation (2.24) could be solved when ψ is fractional linear generating function. The result is shown that M is an exponential distribution with the density function $(1-\pi)\exp\{-(1-\pi)u\}$ for $u > 0$, the parameter π is solved by the equation (2.12). In this case, an approximate (1-α) confidence interval could be calculated for n from (2.23) with

$$\begin{aligned} a_1 &= -\log(1-\alpha_1) \\ a_2 &= -\log(\alpha_2). \end{aligned}$$

Considering the value of α_1 and α_2 could be any positive value satisfied $\alpha_1 + \alpha_2 = \alpha$, it will result in many different confidence intervals. Basically, the interval with minimum length is likely to use, and the length of the interval could be calculated as:

$$L = \log(a_2/a_1)/\log m.$$

An easy differentiation shows that minimum length interval will be obtained by choosing α_1, as the unique positive root of the equation:

$$(\alpha - \alpha_1)\log(\alpha - \alpha_1) = (1-\alpha_1)\log(1-\alpha_1).$$

The table 2.1 shows that the shortest length L of (1-α) confidence interval for n.

α	α_1	$\log \alpha_1$	$\log \alpha_2$	$L\ \log m$
0.001	0.0009	-7.0101	2.2229	9.2330
0.005	0.0044	-5.4221	2.0058	7.4279
0.010	0.0187	-4.7105	1.8935	6.6340
0.020	0.0171	-4.0601	1.7652	5.8253
0.050	0.0415	-3.1615	1.5613	4.7228
0.100	0.0804	-2.4791	1.3692	3.8483

Table 2.1: Lengths of shortest $(1-\alpha)$ confidence interval for n.

Critical Case

For the critical case in the Galton-Watson process, the increase rate of offspring m equals to 1, it means that population in each generation is similar, the figure 2.6 shows that if the population conditional on the not being extinct, the population will grow linearly. Hence, estimating the generation n is different from the supercritcial case.

Similar to supercritial case, this process also begins with the fractional linear generating function 2.10 with $b = (1-c)^2$, so it could rewritten as:

$$f(s) = c + \frac{(1-c)^2 s}{1-cs}.$$

Then, the $f_n(s)$ is still a fractional linear generating function, and

$$P(Z = k|Z > 0) = \{1 - c(n)\}c(n)^{k-1} \quad (k = 1, 2, ...), \tag{2.25}$$

where $c(n) = nc/(1 - c + nc)$.
In order to get the estimator $\hat{n}$, applying the maximum likelihood estimate method for the equation (2.25), so the first derivation of log-likelihood is:

$$\frac{d}{dn}\{\log P(Z = k|Z > 0)\} = \frac{k-1}{n} - \frac{kc}{1-c+nc}. \tag{2.26}$$

It could be easy to verify that the second derivation is greater than 0, so suppose the equation 2.26 equals to zero, and solve the equation for n,

$$\hat{n} = \frac{1-c}{c}(Z-1)n. \tag{2.27}$$

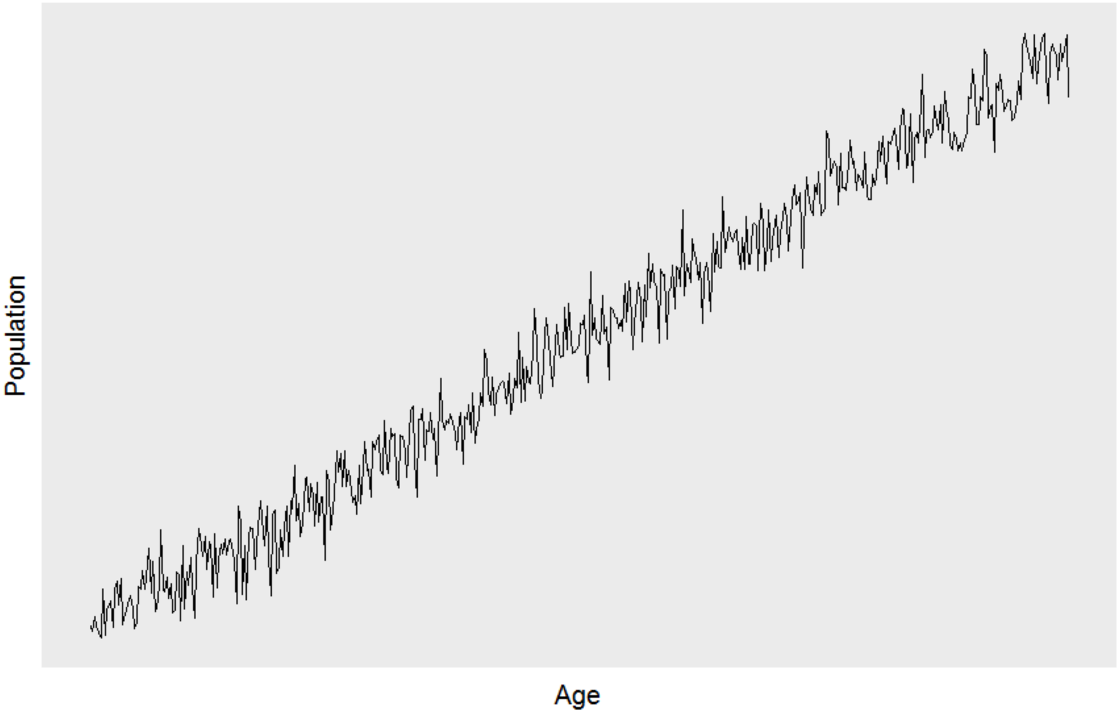

Figure 2.6: The linear growth of population in critical case when conditional on non-extinction.

Confidence Interval

The Kesten, Ney and Spitzer [13] have proved that:

$$\lim_{n\to\infty} n[1 - f_n(0)] \to \frac{2}{\sigma^2}$$

and given $Z > 0$ the conditional distribution of $\hat{n}/n$ convergences to an exponential distriution with parameter one. Since $E(Z|Z > 0) = [1 - f_n(0)]^{-1}$, it follows that $\hat{n}$ is an asymptotically unbiased estimated of n in the sense that $E(\hat{n}/n|Z > 0) \to 1$ as $n \to \infty$. The approximate $(1 - \alpha)$ confidence interval $S(\hat{n})$ for n can be written as:

$$S(\hat{n}) = \left[\frac{\hat{n}}{a_2}, \frac{\hat{n}}{a_1}\right]. \tag{2.28}$$

The coverage $P[n \in S(\hat{n})] \to 1 - \alpha$ as $n \to \infty$, where $a_1 = -\log(1 - \alpha_1)$ and $a_2 = -\log \alpha_2$ (e.g. $\alpha_1 + \alpha_2 = \alpha$).

Subcritical Case

As for the subcritical case, since the increase rate of offspring m less than 1, the population will eventually die out with probability equals to 1, which is similar to the figure 2.7. but Yaglom [10] indicates that there is another situation in the subcritical case. He has proved that

$$\lim_{n\to\infty} P(Z = k|Z > 0) = d_k,$$

and $d_1 + d_2 + .. = 1$. Therefore, for identify population size k, it is impossible to estimate the generation n.

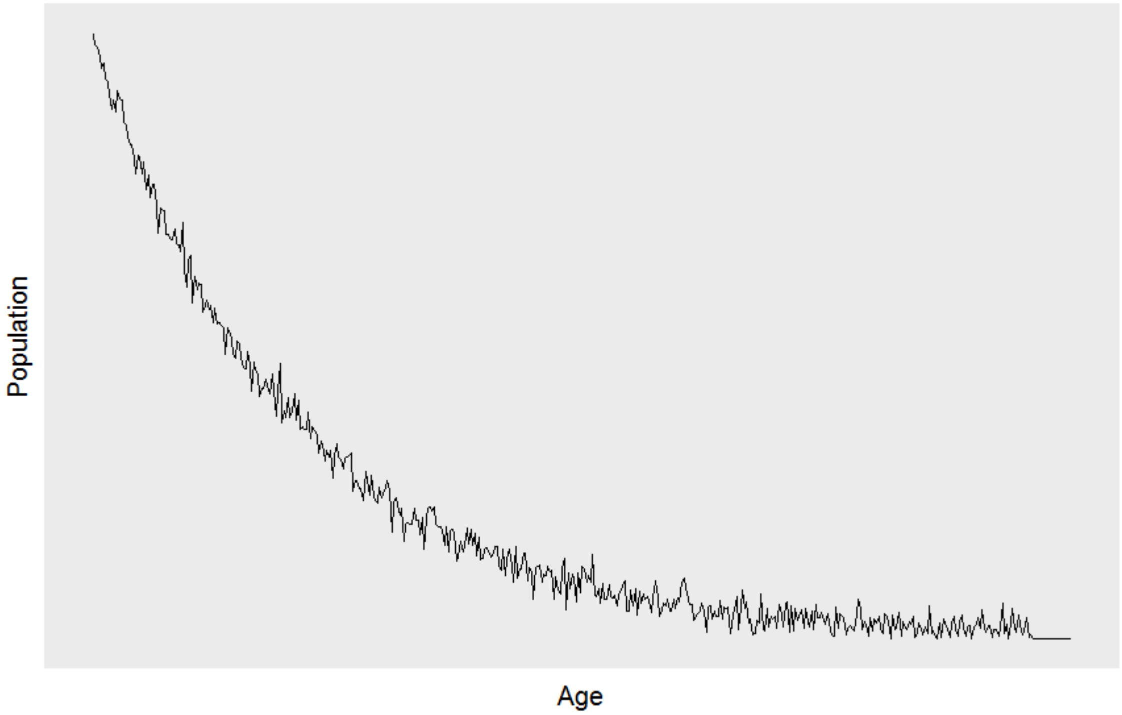

Figure 2.7: The extinction of population begins with a large number of individuals in subcritical case.

Chapter 3

Result

3.1 Maximum Likelihood Estimator

Parameter π and m

The equation (2.21) already shows the way to calculate the $\hat{n}$, but first problem is how to estimate π and m from the data. Neel and Chagnon [19] study the distribution of offspring numbers for Yanomama Indians and find that the adult generation and surviving children have a good fit for the generalized geometric distribution with $c = 0.4$ and m equals to half the offspring number. Kojima and Kelleher [14] used the negative binomial distribution instead of geometric distribution, but the result showed that the geometric distribution fits better. Therefore, it makes sense for the assumption of a geometric distribution. From the equation (2.10) and (2.11), it is easy to infer that

$$1 - \pi = (m - 1)(1 - c)/c.$$

Substituting $c = 0.4$:

$$\pi = 1.5m + 2.5.$$

Since research already gets the value of c from others' study. So if the value of m is estimated, the value of π could be solved.

Supercritical Case

Neel and Weiss [19] found the evidence that the value of m in each generation is slightly greater than one in the Yanomama population, and they calculate the value of m from the family size means, but using the survival of offspring to adulthood is more sensible. Therefore, the m calculated by them is slightly greater than the m estimated from the survival of offspring to adulthood.

In their paper, the mean m in Yanomama was around 1.02 during a long period, and the mean m increases slightly in recent generations. The increase of mean is around between 0.005 to 0.01 annually or the mean m is range from 1.1 to 1.2. Considering the situation when $c = 0.4$, the average rate of increase m should be taken the range from 0.95 to 1.1, and in the supercritical case, this paper chooses three values when m equals to 1.02, 1.05 and 1.1.

According to the equation (2.21) and (2.23), when m equals to 1.02 (i.e. similar to annual growth of 0.001 in population), the MLE estimator is around 167 generations, with 88 and 326 generations being the lower and upper boundary of confidence interval. When m equals to 1.05, the MLE estimator is around 86 generations, with 54 and 151 generations being the lower and upper boundary of confidence interval, and when m equals to 1.1, the maximum likelihood estimator is around 51 generations, with 35 and 84 generations being the lower and upper boundary of confidence interval.

Critical Case

For the critical case, rate of increase m equals to 1, and according to the equation (2.27) and (2.28) the MLE estimator of the age is approximately 1313 generations. The lower and upper boundary confidence interval are 275 generations and 30930 generations respectively. The length of confidence interval is quiet large, for the upper boundary is 30930. Assuming each generation interval is 25 years which is quite small, the occurrence of mutant happened about 773250 years ago, this is impossible since the human species existed a half million years ago. At the other extreme, an estimate of approximately 275 generations; Thompson [24] found that this estimator would put the origin of the mutation at a point in time midway between the crossing of the isthmus and he present, and very likely subsequent to tribal differentiation, so it is impossible.

3.2 Coverage Probability

In statistics, the coverage probability of a technique for calculating a confidence interval is the proportion of the time that the interval contains the true value of interest [7]. In particular, it is the simulation results of the proportion of the true generations which are assumed in the simulation locate in the confidence interval. The reason for doing the simulation is to make sure of the asymptotic confidence interval (2.23) is appropriate. Since research chooses the significance level as $\alpha = 0.05$, so the coverage probability is expected to converge to 0.95. Here, research choose three values $m = 1.02$, 1.05 and 1.1 to do simulation, all of them are in supercritical case. The previous result of critical case shows the population is unlikely in the critical case, so it is no necessary for simulation.

Mean m Equals to 1.02

The figure 3.1 shows the result of simulation for $m = 1.02$, coverage probability is nearly 1 for the first 30 generations, and it seems to be convergence after around 100 generations. The MLE estimator is 167 generations when $m = 1.02$, which indicates that the coverage probability is already convergence.

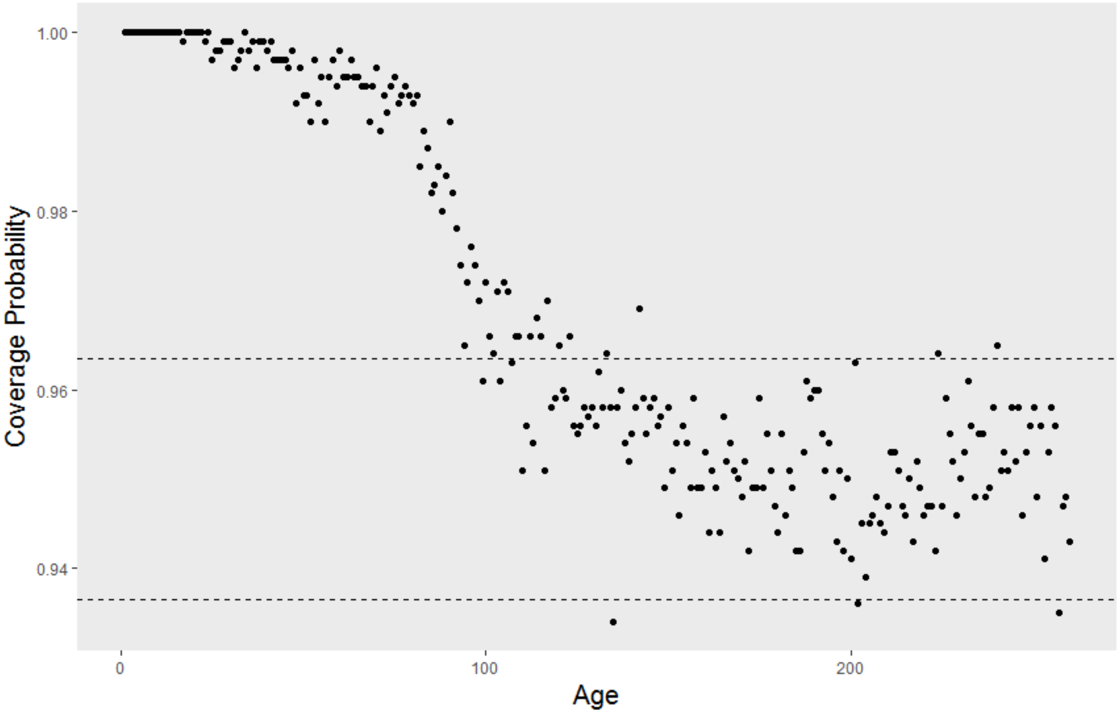

Figure 3.1: The plot of coverage probability versus generations for m=1.02, the dashed lines are binomial confidence interval for 95% confidence level.

Mean m Equals to 1.05

The figure 3.2 shows the result of simulation for $m = 1.05$, situation is similar, since coverage probability begins with 1 and converges to 0.95 after 80 generations and the MLE estimator is 86 generations.

Mean m Equals to 1.1

The figure 3.2 shows the result of simulation for $m = 1.1$, coverage probability reaches convergence around 50 generations, and the MLE estimator for $m = 1.1$ is exactly 51 generations.

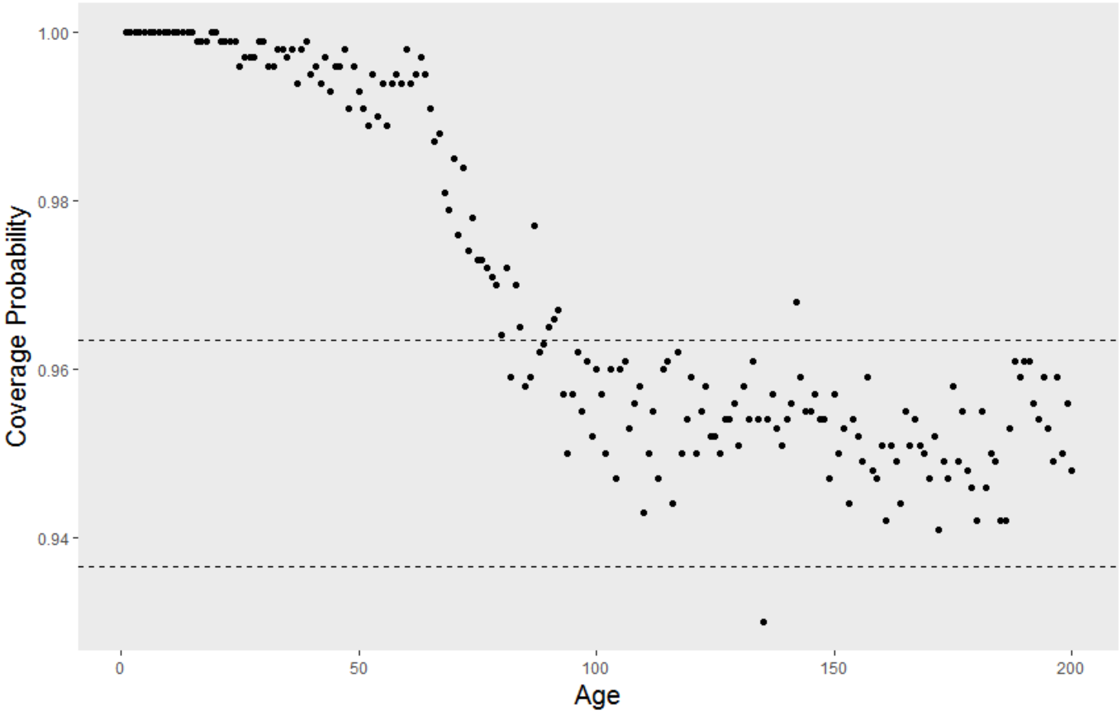

Figure 3.2: The plot of coverage probability versus generations for $m = 1.05$, the dashed lines are binomial confidence interval for 95% confidence level.

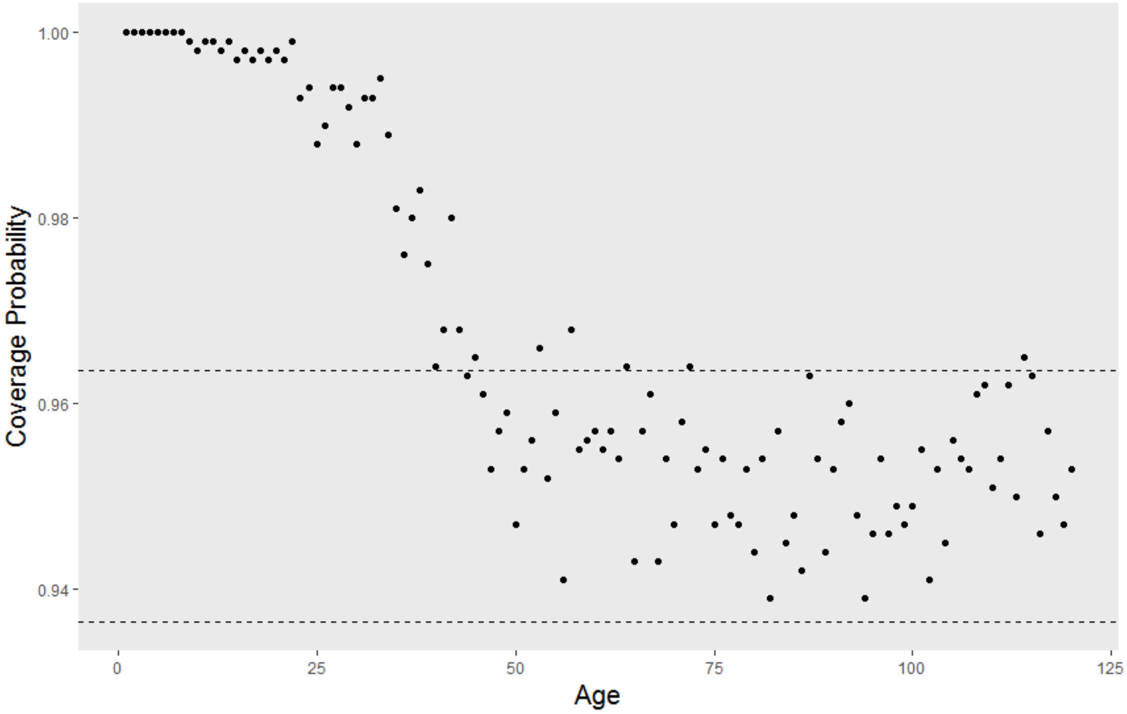

Figure 3.3: The plot of coverage probability versus generations for $m = 1.1$, the dashed lines are binomial confidence interval for 95% confidence level.

Chapter 4

Conclusion and Discussion

4.1 Conclusion

The problem considered is that of estimating the age or rate of increase of a variant on the basis of the present number of replicates observed in a population. The research starts with diffusion equation analysis of age probability distributions, the linear fractional generating function is shown good properties in the iteration process. Then inferring the likelihood for the age which is studied on the basis of a discrete branching process model, and maximizing likelihood to get the MLE estimator for generations. According to the research of Helgason and Agnar [11], the generation intervals for the matriarchal society and patrilineal society is 28.12 years and 31.13 year respectively. Since there is no direct evidence to show whether the Yanomama tribe is matriarchal society or patrilineal society, this paper uses average value to estimate the generation intervals, which is 29.625 years. According to result, estimated generation is around 167 when m equals 1.02, so this variant happened 4947.375 years ago, this estimator and range of likely values conform extraordinary well with the hypothesis that this mutation should be quite old but arising after tribal differentiation (i.e. around 5500 years ago). As for m equals 1.05, estimated generation is around 86, and 2547.75 years ago; in the situation of m equals 1.1, estimated generation is around 51 generations, and 1510.875 years ago, but there is no evidence showed that the variant is less than 2000 years old. The figure 4.1 is shown that with increase of mean in supercritical case, the age will decrease. Because of MLE estimator for $m = 1.1$ is already in two thousand years, so there is no need to invokes larger m values. For critical case (i.e. $m = 1$), result shows that the estimated generation is 38898 or around 1152353.25 years ago. However, this estimator is not realistic since the period is remarkable earlier than the American Indian crossed the isthmus into South America.

In conclusion, the previous result shows that the length of confidence interval is very wide, so it is impossible to provide a reliable point estimate of the age of a specified variant in the process of gene evolution in natural populations, although the likelihood analysis provides a confidence interval which may place useful boundaries on the period in which a variant originated, and the observed distribution of numbers of several variants may also provide useful information.

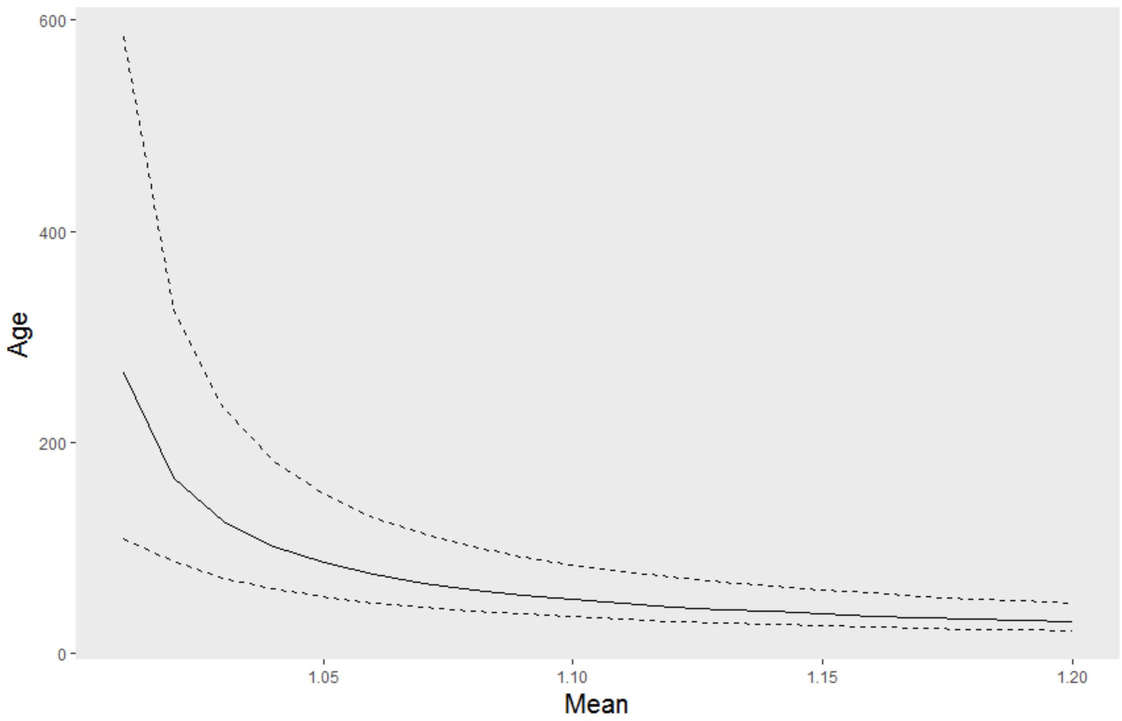

Figure 4.1: The plot of age versus mean when present number of replicates observed is 875, the dashed lines are 95% approximately confidence interval.

4.2 Discussion for Further Research

Value of Mean m

In this paper, research tries different values of m to find a reasonable estimation, and m is assumed to be fixed, but in realistic the population structure will change. Since m denotes the rate of increase for population, this value should not be considered as a fixed value. In addition, the figure 4.1 is shown m is sensitive to estimator, we should consider it carefully, so how to choose the value of m still need further exploration.

Using Birth Order to Estimate Family Sizes

According to method, it is fairly clear to estimate the offspring the distribution consistently if the entire family tree is observed, conditional on non-extinction. That leads to the question, for instance, if more parameters in the generation size case need to be estimated conditionally consistently, it means that the entire family tree cannot be observed. One such intermediate scheme is to randomly sample individuals in the process, and determine each sampled individuals family size. Obviously, this scheme yields no information about the probability of zero offspring.

Burks [4] discusses the demographic problem of obtaining data on the distribution of family sizes. To deal with this problem, she suggests taking a random sample of individuals from the population, and ascertain their birth order rather than their family size. One way this might be achieved is by sampling birth certificate, which in the past (e.g. in the United States) usually required that mothers report whether a particular birth was their first, second, third and so on. The number of individuals in the sample with birth order k should approximately $n\sum_{j\geq k} p_j$, where n is the sample size, since given that the eventually family size is j, the probability is $1/j$ for each particular birth order entering the sample, and a family of size j has probability jp_j to enter the sample. This assumes that the population behaves in a stationary fashion, for a treatment of a more general case, I refer the reader to [9].

Young Variants

The variant research considered is a quiet old which has at least 2000 years, and research also wants to consider some young variant. However, a variant present in the low number may, however, be of considerable age. One case which is found by Neel [20] is that of a Makiritare albumin variant which was found in four villages in South America. It has been found in 13 individuals, of whom four are adults. Research assumes that these are the only current adult representatives of the only current adult representatives of the variant which is it is truly localized. The maximum likelihood estimate of the age is 4 generations for m between 1.02 and 1.1 and increases only to 5 for m in the range from 1 to 1.05. As for upper confidence interval, the upper boundary will reach 106 generations, there is no essentially no information concerning the age. At m=1.02, 1.05 and 1.1 the upper boundary of confidence intervals are 164, 69 and 37 generations respectively, so an age of several hundred is well within the bounds of possibility. Hence, for variants in very low numbers, it is unlikely that a useful estimate will be obtained through this method and how to deal with this case still need further research.

Continuous Time Markov Branching Processes

In this paper, the research is only concerned with the Galton-Waltson process, which is a discrete time Markov process. For the further study, research will consider more general models. In particular, all of the result can be applied immediately to continuous time Markov branching process. This is because of the fact that if $\{Y(t), t \geq 0\}$ is a continuous time Markov branching process, then for ant t_0 the imbedded process $\{Z_n\}$ given by $Z_n = Y(nt_0)$ is a Galton-Watson process [10]. More generally, all of our technique can be carried over to supercritical age-dependent branching process by appealing to the results of Athreya [3].

Bibliography

[1] I. Adeniran. *Modelling the Short Qt Syndrome Gene Mutations: And Their Role in Cardiac Arrhythmogenesis*. Springer, 2014.

[2] J. Aldrich. "RA Fisher and the making of maximum likelihood 1912-1922". In: *Statistical science* (1997), pp. 162–176.

[3] K. B. Athreya. "On the supercritical one dimensional age dependent branching processes". In: *The Annals of Mathematical Statistics* (1969), pp. 743–763.

[4] B. S. Bukks. "A statistical method for estimating the distribution of sizes of completed fraternities in a population represented by a random sampling of individuals". In: *Journal of the American statistical association* 28.184 (1933), pp. 388–394.

[5] C. Darwin. *The origin of species*. Dent, 1909.

[6] P. A. M. Dirac. *Branching Processes*. Springer, 1972. ISBN: 978-3-642-65371-1.

[7] Y. Dodge. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand, 2006.

[8] D. Feldman and M. Fox. "Estimation of the parameter n in the binomial distribution". In: *Journal of the American Statistical Association* 63.321 (1968), pp. 150–158.

[9] P. Guttorp. *Statistical inference for branching processes*. Vol. 122. Wiley-Interscience, 1991.

[10] T. E. Harris. *The theory of branching processes*. Courier Corporation, 2002.

[11] A. Helgason et al. "A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes". In: *The American Journal of Human Genetics* 72.6 (2003), pp. 1370–1388.

[12] M. Hollstein et al. "Database of p53 gene somatic mutations in human tumors and cell lines." In: *Nucleic acids research* 22.17 (1994), p. 3551.

[13] H. Kesten, P. Ney, and F. Spitzer. "The Galton-Watson process with mean one and finite variance". In: *Theory of Probability & Its Applications* 11.4 (1966), pp. 513–540.

[14] K.-I. Kojima and T. M. Kelleher. "Survival of mutant genes". In: *The American Naturalist* 96.891 (1962), pp. 329–346.

[15] H. Kuipers et al. "An HIV-1-infected individual homozygous for the CCR-5 D32 allele and the SDF-1 3 A allele". In: *Aids* 13.3 (1999), p. 433.

[16] J. W. MacCluer, J. V. Neel, and N. A. Chagnon. “Demographic structure of a primitive population: a simulation”. In: *American Journal of Physical Anthropology* 35.2 (1971), pp. 193–207.

[17] T. Maruyama. “The age of a rare mutant gene in a large population.” In: *American journal of human genetics* 26.6 (1974), p. 669.

[18] J. V. Neel. ““Private” genetic variants and the frequency of mutation among South American Indians”. In: *Proceedings of the National Academy of Sciences* 70.12 (1973), pp. 3311–3315.

[19] J. V. Neel and N. A. Chagnon. “The demography of two tribes of primitive, relatively unacculturated American Indians”. In: *Proceedings of the National Academy of Sciences* 59.3 (1968), pp. 680–689.

[20] J. V. Neel and K. M. Weiss. “The genetic structure of a tribal population, the Yanomama Indians. XII. Biodemographic studies”. In: *American Journal of Physical Anthropology* 42.1 (1975), pp. 25–51.

[21] S. A. Sawyer et al. “Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila”. In: *Proceedings of the National Academy of Sciences* 104.16 (2007), pp. 6504–6510.

[22] M. J. Stensrud and M. Valberg. “Inequality in genetic cancer risk suggests bad genes rather than bad luck”. In: *Nature Communications* 8.1 (2017), p. 1165.

[23] S. M. Stigler. “Estimating the age of a Galton—Waston branching process”. In: *Biometrika* 57.3 (1970), pp. 505–512.

[24] E. Thompson. “Estimation of age and rate of increase of rare variants.” In: *American journal of human genetics* 28.5 (1976), p. 442.

Appendices

Appendices

R code

```
library(ggplot2)
#######################define the parameter
c<-0.4
alpha1<- 0.0415
alpha2<- 0.05-alpha1
##################
########################situation for m=1.02
m<-1.02 #mean
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
n.hat<-log((1-pi)*875+pi)/log(m) ######MLE estimator
a1<- -log(1-alpha1)
a2<- -log(alpha2)
#get the confidence interval
cilow<-n.hat-log(a2)/log(m)  ##########low boundary of confidence interval
ci.high<-n.hat-log(a1)/log(m) ########high boundary of confidence interval
########################situation for m=1.05
m<-1.05
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
n.hat<-log((1-pi)*875+pi)/log(m)
a1<- -log(1-alpha1)
a2<- -log(alpha2)
#get the confidence interval
cilow<-n.hat-log(a2)/log(m)  ##########low boundary of confidence interval
cihigh<-n.hat-log(a1)/log(m)  ########high boundary of confidence interval
########################situation for m=1.1
m<-1.1
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
n.hat<-log((1-pi)*875+pi)/log(m)
a1<- -log(1-alpha1)
```

```r
a2<- -log(alpha2)
#get the confidence interval
cilow<-n.hat-log(a2)/log(m)   ##########low boundary of confidence interval
cihigh<-n.hat-log(a1)/log(m)  ########high boundary of confidence interval
############################situation for m=1
n.hat<- (1-c)/c*(875-1)
a1<- -log(1-alpha1)
a2<- -log(alpha2)
cilow<- n.hat/a2
cihigh<- n.hat/a1
###################################
m<-seq(101,120,1)/100
n.hat<-numeric(20)
cilow<- numeric(20)
cihigh<- numeric(20)
for (i in 1:20) {
  b<-m[i]*(1-c)^2
  pi<-(1-b-c)/(c*(1-c))
  n.hat[i]<-log((1-pi)*4+pi)/log(m)
  a1<- -log(1-alpha1)
  a2<- -log(alpha2)
  #get the confidence interval
  cilow[i]<-n.hat[i]-log(a2)/log(m[i])
  cihigh[i]<-n.hat[i]-log(a1)/log(m[i])
}
##############get the coverage plot
m<-1.02 ##########change this value to get the coverage prabability for m=
1.05, m=1.1
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
#####################################################
###############define the number of generarion
coverage<-numeric(500)
for (n in 1:500) {


  bn<-m^n*((1-pi)/(m^n-pi))^2
  cn<-(m^n-1)/(m^n-pi)
```

```
 pn<-numeric(1000)

 for (i in 1:1000) {
  pn[i]<-(1-cn)*cn^(i-1)
  if(pn[i]<0.000000001)
   break
 }
 pn<-pn[pn>0]
 nopn<-length(pn)#############the number of probability
 result10<- numeric(1000)
 for (times in 1:1000) {
  result10[times]<-sample(1:nopn,1,prob = pn)
 }
##############################
n.hat<-log((1-pi)*result10+pi)/log(m)
a1<- -log(1-alpha1)#define the value of a1
a2<- -log(alpha2)#define the value of a2
#get the confidence interval
 cilow<-n.hat-log(a2)/log(m)
 cihigh<-n.hat-log(a1)/log(m)
 coverage[n]<-sum(cilow<n&cihigh>n)/1000
}
df<-data.frame(x=c(1:length(coverage)),y=coverage)
st<-sqrt(0.95*0.05/1000) #binomial standard error
h<- 0.95+1.96*st  #high boundaruy of confidence interval
l<- 0.95-1.96*st  #low boundary of confidence interval
ggplot()+geom_point(data = df,aes(df[,1],df[,2]))+geom_hline(yintercept =
h,lty="dashed")+
 geom_hline(yintercept = l,lty="dashed")+
 ylab("Coverage Probability")+theme(axis.title.y = element_text(size = 1
5))+
 xlab("Age")+theme(axis.title.x = element_text(size = 15))+
 theme(panel.grid.major = element_blank(), panel.grid.minor = element_blan
k())

#####the length of the confidence interval
L<-log(a2/a1)/log(m)
```

```
L*log(m)
########## get The plot of age versus mean
m<-1.01
n.hat<-numeric(20)

cilow<- numeric(20)
cihigh<- numeric(20)
for (i in 1:20) {
  b<-m*(1-c)^2
  pi<-(1-b-c)/(c*(1-c))
  n.hat[i]<-log((1-pi)*875+pi)/log(m)
  a1<- -log(1-alpha1)
  a2<- -log(alpha2)
  #get the confidence interval
  cilow[i]<-n.hat[i]-log(a2)/log(m)
  cihigh[i]<-n.hat[i]-log(a1)/log(m)
  m<-1.01+0.01*i
}
mm<-seq(1.01,1.20,0.01)
plot(mm,n.hat)
library(ggplot2)
df<-data.frame(Mean=mm,Age=n.hat,l=cilow,h=cihigh)
sp<-ggplot()+geom_line(data=df,mapping=aes(df$Mean,df$Age))+geom_line(data
= df,mapping = aes(Mean,l),lty="dashed")+
  geom_line(data = df,mapping = aes(Mean,h),lty="dashed")+xlim(c(1.01,1.2))
sp + theme(panel.grid.major = element_blank(), panel.grid.minor = element_b
lank())+
  xlab("Mean")+ theme(axis.title.x = element_text(size = 15))+
  ylab("Age")+ theme(axis.title.y = element_text(size = 15))
###########################few k=4 in the situaion of young variant

c<-0.4
alpha1<- 0.0415
alpha2<- 0.05-alpha1
###########m=1.02
m<-1.02
b<-m*(1-c)^2
```

```
pi<-(1-b-c)/(c*(1-c))

n.hat<-log((1-pi)*4+pi)/log(m)
a1<- -log(1-alpha1)
a2<- -log(alpha2)
#get the confidence interval
cilow<-n.hat-log(a2)/log(m)

cihigh<-n.hat-log(a1)/log(m)
##########m=1.05
m<-1.05
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
n.hat<-log((1-pi)*4+pi)/log(m)
a1<- -log(1-alpha1)
a2<- -log(alpha2)
#get the confidence interval
cilow<-n.hat-log(a2)/log(m)
cihigh<-n.hat-log(a1)/log(m)
############m=1.1
m<-1.1
b<-m*(1-c)^2
pi<-(1-b-c)/(c*(1-c))
n.hat<-log((1-pi)*4+pi)/log(m)
a1<- -log(1-alpha1)
a2<- -log(alpha2)
#get the confidence interval

cilow<-n.hat-log(a2)/log(m)
cihigh<-n.hat-log(a1)/log(m)
```