RICE UNIVERSITY

Branching Processes with Biological Applications

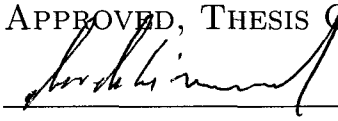
by

Xiaowei Wu

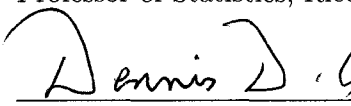
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

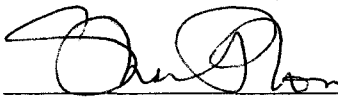
Approved, Thesis Committee:



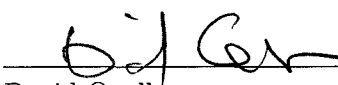
Marek Kimmel, Chair
Professor of Statistics, Rice University



Dennis D. Cox
Professor of Statistics, Rice University



Sharon E. Plon
Professor of Molecular and Human
Genetics, Baylor College of Medicine



David Queller
Harry C. and Olga K. Wiess Professor of
Natural Sciences, Rice University

Houston, Texas

October, 2009

Abstract

Branching Processes with Biological Applications

by

Xiaowei Wu

Branching processes play an important role in models of genetics, molecular biology, microbiology, ecology and evolutionary theory. This thesis explores three aspects of branching processes with biological applications. The first part of the thesis focuses on fluctuation analysis, with the main purpose to estimate mutation rates in microbial populations. We propose a novel estimator of mutation rates, and apply it to a number of Luria-Delbrück type fluctuation experiments in *Saccharomyces cerevisiae*. Second, we study the extinction of Markov branching processes, and derived theorems for the path to extinction in the critical case, as an extension to Jagers' theory. The third part of the thesis introduces infinite-allele Markov branching processes. As an important non-trivial example, the limiting frequency spectrum for the birth-death process has been derived. Potential application of modeling the proliferation and mutation of human Alu sequences is also discussed.

Acknowledgements

I would like to first express my sincere gratitude to my advisor, Dr. Marek Kimmel. Throughout my thesis-writing period, he provided encouragement, sound advice, good teaching, and lots of good ideas. I would have been lost without him.

I am highly grateful to Dr. Sharon Plon for her great help and guidance. She has provided me the opportunity to work on real genetics problems and collaborate with researchers from other disciplines.

I am also indebted to Dr. Dennis Cox, who has served in my committee and provided valuable suggestions that help to make breakthroughs in my research.

Finally, I wish to thank my family for their love and support throughout my life. I am especially thankful to my wife, Hongxiao Zhu, who has helped to proofread and revise this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The theory of branching processes started in the middle of the nineteenth century with social scientists analyzing the reasons for extinction of family lines. With time, it has been found that numerous biological processes involving reproduction can be modeled, or at least approximated, by branching processes. Branching processes play an important role in models of genetics, molecular biology, microbiology, ecology and evolutionary theory.

This thesis summarizes three topics related to branching processes with biological applications. The first topic, developed jointly with Drs. Sharon E. Plon and Erin D. Strome of Baylor College of Medicine, is fluctuation analysis with the main purpose to estimate mutation rates in microbial populations. The basis of fluctuation experiments are spontaneous mutations, i.e., mutations occurring spontaneously, not induced by mutagenic agents. This provides a chance to model the biological process

as an asexual two-type branching process. In this work, we present a novel estimator of mutation rates, which allows for unequal population sizes N_t of the parallel cultures. Simulation results show a good accuracy and robustness of this estimator compared with the commonly used median estimator and the maximum likelihood estimator. The proposed estimator is applied to 20 yeast datasets collected during three separate days of study of chromosome loss and recombination in wild-type *Saccharomyces cerevisiae* strains. In addition, we propose an alternative approach to fluctuation analysis, based on two-type Markov branching processes.

The second topic concerns the approximation of frequency of rare variants in Wright-Fisher model using a subcritical or critical branching process, and resulting investigation of transient processes leading to extinction in Markov branching processes. For the former problem, we follow Nagylaki's approach [8], which approximates frequencies of rare alleles as a subcritical or critical branching process. For the latter, we summarize the known facts concerning the *time to extinction*, *path to extinction* and *path on the verge of extinction* in subcritical Markov branching processes [16, 17, 31, 33], and extend these results to the critical case with finite variance. These results are relevant for the dynamics of extinction of disease-causing variants of genes in human populations.

The third topic focuses on the infinite-allele Markov branching process. We assume that in a Markov branching process, individuals can mutate to novel identifiable types, which we call alleles. Mutation of a new-born individual to a novel allelic type is

independent of all other members of the population. The distribution of the number of offspring is assumed to be the same for all alleles, as well as the distribution of the life-time. Based on this setting, we derived the frequency spectrum for the infinite-allele birth-death process, following Pakes' approach [32]. We then discussed the limitation and possible extension of this model in modeling the proliferation and mutation of human Alu sequences.

1.1 Background concerning Branching Processes

We first introduce the probability generating function (pgf) as an important tool in the analysis of branching process. In probability theory, the pgf of a discrete random variable is a power series representation (the generating function) of the probability mass function (pmf) of the random variable. If X is a discrete random variable taking values on some subset of the non-negative integers, with pmf $\{p_k := P(X = k), k = 0, 1, \cdots\}$, then its pgf is defined as:

$$f(s) = E[s^X] = \sum_{k=0}^{\infty} p_k s^k, |s| \leq 1.$$

Easy to see that $f(1^-) = 1$, $E[X] = f'(1^-)$ and $p_k = \frac{f^{(k)}(0)}{k!}$.

For the notation and definitions of branching processes, we will follow Athreya and Ney [3].

Definition 1.1.1. Galton-Watson branching process

A Galton-Watson branching process is a Markov chain $Z_n, n = 0, 1, \cdots$, with state

space $\{0\} \bigcup \mathcal{Z}^+$ *and transition probability*

$$P(i,j) = P(Z_{n+1} = j | Z_n = i) = \begin{cases} p_j^{*i}, & \text{if } i \geq 1, j \geq 0 \\ \delta_{0j}, & \text{if } i = 0, j \geq 0 \end{cases},$$

where δ_{ij} *is the Kronecker delta and* $\{p_k^{*i}, k = 0, 1, \cdots\}$ *is the* i*-fold convolution of probability mass function* $\{p_k, k = 0, 1, \cdots\}$.

In other words, this Markov chain is determined completely by the pmf $\{p_k, k = 0, 1, \cdots\}$, which is the offspring distribution, and 0 is the absorption state of the Markov chain.

For the Galton-Watson branching process Z_n, write its pgf as $f_n(s)$. Suppose the offspring pgf is $f(s)$. The relation between successive generations leads to an iterative rule of $f_n(s)$:

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

Assuming $Z_0 = 1$, in terms of pgf, that is $f_0(s) = s$, this iterative rule then gives a solution to the distribution of Z_n.

Definition 1.1.2. Markov branching process

A stochastic process $\{Z(t,\omega); t \geq 0\}$ *on a probability space* $(\Omega, \mathcal{F}, P)$ *is called a one dimensional continuous time Markov branching process if:*

(i) its state space is the set of non-negative integers;

(ii) it is a stationary Markov chain with respect to the σ*-fields* $\mathcal{F}_t = \sigma\{Z(s,\omega); s \leq t\}$;

(iii) the transition probabilities $P_{ij}(t)$ *satisfy*

$$\sum_{j=0}^{\infty} P_{ij}(t)s^j = \left[\sum_{j=0}^{\infty} P_{1j}(t)s^j\right]^i,$$

for all $i \geq 0$ *and* $|s| \leq 1$.

It is sometimes more convenient to use an intuitive description for branching processes. Consider the scheme of evolution and reproduction of a population of some particles. Each particle, independently of the others, lives in a life time and generates a random number of new offspring. If the life times are fixed, say to one time unit, and the splits happen only at the death of each particle, then the process is called a Galton-Watson branching process. If the life times are i.i.d. exponential, and independent of the offspring distribution, then the process is called a Markov branching process.

A natural generalization of the continuous time Markov branching process is to allow life times to be i.i.d. random variables with arbitrary distribution $G(t)$. If the life times are independent of the offspring distribution, then this process is called an age-dependent (Bellman-Harris) process. In contrast to Galton-Watson and continuous time Markov branching processes, this process is not Markovian. For details about age-dependent processes, see [3, 15].

If we further assume that individuals give birth to their offspring not necessarily exactly at their death, but at randomly chosen instants during their lives, and the offspring distribution depends on the life time the general Crump-Mode-Jagers process. For details about general branching processes, see [15, 36].

Depending on the mean m of the offspring distribution, branching processes behave differently. If $m > 1$, the process is called supercritical. $m = 1$ and $m < 1$ correspond to critical and subcritical processes. In these two cases, the process will eventually die out. More details about the extinction probability and limit laws can be found in Athreya and Ney [3].

1.2 Outline of the Dissertation

The dissertation is organized as follows: Chapter 2 explores the methodology and application of mutation rate estimation in cell population; Chapter 3 summarizes and extends the results concerning extinction of subcritical/critical Markov branching processes; Chapter 4 introduces the infinite-allele Markov branching process, and provides derivation for the limiting frequency spectrum for the birth-death process.

Research described in Chapter 2 was published in *Genetics* [38] and *Mutation Research* [41]. Material of Chapter 3 is accepted by *Statistics & Probability Letters*. Material of Chapter 4 is being prepared for submission.

Chapter 2

Modeling Clonal Growth and Mutation in Cell Population

This chapter studies the model of reproduction and mutation in cell population. Section 2.1 explores the methodology of fluctuation analysis in depth, and proposes a novel mutation rate estimator, which allows for large spread of the parallel culture population sizes. This section also includes a study of chromosome loss and recombination in *Saccharomyces cerevisiae* diploid yeast strains which vary in their genetic stability. In Section 2.2, we introduce an alternative model based on two-type Markov branching processes to describe dynamics of cell population. Based on this model, we present methods for the estimation of mutation probability.

2.1 Fluctuation Analysis

Fluctuation experiments were first introduced in 1943 by Salvador Luria and Max Delbrück to show that mutations in bacteria arise spontaneously [25]. The principle of fluctuation experiments considered in this section is as follows: A number of parallel (replicate) cultures, grown from independent cells from the same strain, are separately tested for a mutation phenotype, i.e., resistance to a lethal agent, by plating on a medium containing the agent; the total number of spontaneously arising mutants and viable cells per culture are determined and these data are used for statistical analysis. The distribution of the number of mutants per independently grown culture has been called the Luria-Delbrück (LD) distribution. In the past half a century, at least four major approaches have been used to study this distribution. These include asymptotics, computational methods, probability generating functions and moments [43].

The pattern of the LD distribution is determined by the mutation rate, i.e., the frequency with which mutations appear in the population. Therefore, mutation rates can be estimated based on empirical counts of mutants and population sizes of parallel cultures. The widely used estimators of mutation rates are: the P_0 estimator [25], the median estimator [23], the Lea-Coulson estimator [23], and the maximum likelihood estimator [18, 44], as well as Bayesian estimators [2]. However, most of these estimators, such as P_0, mean and median, do not (and can not) consider variations in population sizes of the parallel cultures but assume the same population size

for each parallel culture. This assumption improves computational accessibility but may cause bias in estimation. On the other hand, the maximum likelihood estimator can handle the population size variations but the optimization algorithm itself is computationally intensive, particularly as the number of mutants becomes large. In the present paper, we propose a new estimator, the modified median estimator that accounts for variations of the parallel culture population sizes N_t. We compare the modified estimator to other mutation rate estimators through simulations, and formulate recommendations concerning the number of parallel cultures needed to achieve accurate estimates. Further, we apply our estimator to data derived from experiments using both *S. cerevisiae* diploid yeast strains and *Escherichia coli.* By analyzing the estimation variability of the two different data sets, we identify that the assumption of the independence of mutation rate on population size in the LD model may be inaccurate under certain conditions of relatively high mutation rate and small population size.

2.1.1 The Luria-Delbrück Distribution

The Luria-Delbrück distribution provides a mathematical tool to describe the expected number of mutants in a clonally growing population of cells. Mutant cells are the result of a mutational event in the original population which here we define as a stable change in the genetic material which confers a change in phenotype (e.g. drug resistance) that can be measured. Traditional methods for computing the LD

distribution require several necessary conditions: (1) Clonal growth is modeled as a continuous process over time; (2) Cell death is ignored. Therefore, when a cell divides, it is replaced by two cells capable of further proliferation. (3) The process starts at time $t = 0$ with one wild-type cell. Mutations then occur at a rate μ per time unit (generation) per single cell (or alternatively, at a rate proportional to the population size, in the entire cell population). (4) Backward mutation is ignored. (5) Wild-type and mutant cells have the same growth rate.

When these assumptions are satisfied, the probability mass function of the LD distribution can be characterized in a number of ways [26, 35]. In this paper we follow the formulation of [44], which is based on the recursive algorithm derived by Ma et al. [26]:

$$\begin{aligned} p_0 &= e^{-m}, \\ p_k &= \frac{m}{k}\sum_{j=1}^{k}\phi^{j-1}(1-\frac{j\phi}{j+1})p_{k-j} \quad (k \geq 1). \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} m &= \frac{\mu}{\beta_1}(N_t - N_0) \approx \frac{\mu}{\beta_1}N_t, \\ \phi &= 1 - e^{-\beta_1 t} = 1 - \frac{N_0}{N_t}. \end{aligned} \tag{2.2}$$

Here, p_i is the probability of getting i mutants in the population of size N_t at time t, m is the expected number of mutations by time t, μ is the mutation rate per unit time, β_1 is the growth rate of cells, N_0 is the number of wild-type cells used as seeds. Note that the expected cumulative number of mutations m is different from the number

k of mutants present at time t. m is usually smaller than k since each division of a mutant cell increments k by 1, by Assumption (3) above, whereas m is incremented only by mutation events.

The main objective of estimation is to determine the composite parameter $\mu_\beta = \frac{\mu}{\beta_1}$, called mutation rate per cell division. Estimating m is an alternative used in some literatures [23, 44, 35], since m is directly connected with $\frac{\mu}{\beta_1}$ through Equation (2.2). However, the use of m is not satisfactory in general since, even when the mutation rate μ is constant, m can vary from experiment to experiment because of the variability of the population size N_t. For this reason, we decided to estimate μ in this paper.

2.1.2 Estimation of Mutation Rate

Estimation of mutation rates is the main purpose of fluctuation analysis. Point and interval estimators of mutation rates allow quantification of the information regarding the genome stability of the target strain. We introduce three estimators in this section: the median estimator, the maximum likelihood estimator and the modified median estimator. The P_0 estimator will not be discussed due to its known drawbacks [44]. Because of the long-tail property of the LD distribution, the mean estimator (method of moments) tends to have a high variance and asymmetric confidence interval, particularly with a limited sample and hence is considered less applicable either.

(1) The Median Estimator

A general way to find the median estimator is by equating the median of the

distribution to the empirical median number of mutants based on all cultures in the batch,

$$\hat{\mu} : P(K \leq k_0 | N_t, \hat{\mu}) = \frac{1}{2},$$

where k_0 is median number of mutants based on all cultures of the batch.

Numerical calculation of the median estimator relies on the cumulative distribution function of the LD distribution, which involves iterative computation of the probabilities (see Equation (2.1)). Therefore a large value of k_0 (usually for $k_0 > 5000$) will lead to computational problems.

The Lea-Coulson estimator is a good approximation in the case of large k_0 [23]. It satisfies the empirical relation (Equation (37) of [23]),

$$\frac{k_0}{\hat{m}} - \log(\hat{m}) = 1.24,$$

where m is the expected number of mutations as defined above.

In their paper, Lea and Coulson also concluded in a semi-empirical manner, that the *pivotal quantity*

$$\Big(\frac{11.6}{k_0/m - \log(m) + 4.5} - 2.02\Big)$$

has an approximate $N(0,1)$ distribution. Therefore, they use this quantity to obtain the median estimator, and its confidence interval (CI).

(2) The Maximum Likelihood Estimator

The first explicit and practical algorithm for computing the maximum likelihood estimator (MLE) of m was published by Jones *et al.* [18]. To evaluate the reliability of the MLE, Stewart then provided a systematic method for constructing CIs

[37]. Inspired by their work, Zheng perfected the MLE derivation and proposed a computationally feasible method for calculation [44].

The MLE of m is given by:

$$\hat{m} = \arg\max_{m} L(m|k_1, ..., k_n, N_{t1}, ..., N_{tn}),$$

where $L(m|k_1, ..., k_n, N_{t1}, ..., N_{tn}) = \sum_{i=1}^{n} \log P(k_i|m, \phi)$.

From the asymptotic distribution of the maximum likelihood estimator, Zheng further developed a Wald-type $100(1-\alpha)\%$ interval estimation of mutation rate as

$$\frac{\hat{m}}{N_t} \pm \frac{z_{\alpha/2}}{N_t\sqrt{nI(\hat{m})}},$$

where $I(m)$ is the Fisher information defined by:

$$I(m) = E_K\Big[\frac{\partial \log P(K|m, \phi)}{\partial m}\Big]^2.$$

Like the median estimator, the MLE also has computational problems when k_0 is large.

In general, the MLE is elegant, and easy to compute, but its CI depends on the asymptotic distribution, which is usually not realizable under experimental conditions, therefore it has an obvious disadvantage when the sample size (number of parallel cultures) is not large enough. Furthermore, the MLE uses all the information in the data to find the mode of the likelihood function, and it may not be robust with respect to outliers.

(3) The Modified Median Estimator

The median estimator is robust by its nature. However, it only uses partial information carried by the data, i.e., the number of mutants in the median culture of the batch, and discards the information provided by the other cultures. Moreover, depending on the experimental design, there may exist a serious spread of Nt in the batch of parallel cultures. Because the number of mutants depends on population size, the culture with the median number of mutants does not necessarily reflect the median mutation rate in such circumstances. This motivates us to find a generalized version of the median estimator.

The modified median estimator is defined as follows: first, using the individual k and N_t, we estimate mutation rates for each parallel culture, based on the method of median (treating the single mutant colony count as the median of the size 1 sample); then we choose the median of those estimated rates as our modified median estimator. This estimator also allows detecting estimation variability and therefore can be used for exploratory and diagnostic analysis as well as for computing final estimated mutations rates (see further on, particularly Section 2.1.4 and Discussion). Mathematically the estimator is expressed as follows

$$\hat{\mu} = median(\hat{\mu}_i),$$

where $\hat{\mu}_i : P(K_i \leq k_i | N_{ti}, \hat{\mu}_i) = \frac{1}{2}$ is the estimate of the i-th culture. Detailed derivations of the modified median estimator are given in the Appendix A.

The widely used median estimator is a special case of the modified median estimator under the condition of equal population size of all the parallel cultures in

the experiment. We may also generalize other estimators, such as the means-method estimator, and so forth, by applying them to the first step of the modified median estimator.

2.1.3 Simulation Study

The quality of a point estimator is generally judged by its MSE, and CI for the estimated parameter. Based on Monte Carlo simulations, we evaluate the performance of three estimators: the median estimator, the modified median estimator and the maximum likelihood estimator, using the MSE criterion. To evaluate their CIs, we use coverage rate, i.e., the probability that the CI will contain the true value of the parameter.

The simulation procedure of the LD distribution is based on [44]. Using the simulated data, we then evaluate the performance of the three estimators. In agreement with the experimental process, we simulate 15 parallel cultures for every predetermined μ. The procedure for computing the mean squared error (MSE) is as follows:

(i) For predetermined $\mu = 10^{-5}$, simulate n pairs of (k, N_t), n from 1 to 15, where N_t follows lognormal distribution with mean $m = 10^6$ and standard deviation $s = 5 \times 10^5$, and k follows $\mathrm{LD}(N_t, \mu)$ distribution. Here the values of m and s are chosen to match typical values in our experimental data described in Section 2.1.4.

(ii) Compute point estimates using the median, maximum likelihood and modified median method.

(iii) Repeat 1000 times and calculate the MSE for the three methods.

Figure 2.1 Part A shows the MSE (in log scale) of the three estimators based on 1000 simulations when the number of parallel cultures varies from 1 to 15. We see from this figure that the modified median estimator performs better than the median estimator, and that the maximum likelihood estimator shows the lowest MSE among the three. Through another 1000 simulations, the coverage rates of the 95% CI of the three estimators when the number of parallel cultures varies from 1 to 15 are shown in Figure 2.1 Part B. Clearly, an adequate estimator should provide a 95% CI with coverage rate close to 95%, and coverage rate larger or smaller than 95% introduces additional type II or type I errors. The modified median estimator and the maximum likelihood estimator have coverage rates stabilizing at 95% as the number of parallel cultures increases. However, the coverage rates of the median estimator are always close to 1 under different number of parallel cultures. This shows that the 95% CI calculated from the method of median is wider than that expected and thus is questionable.

We also investigated the robustness of the point estimators to outliers in the number of mutants. To evaluate the robustness of estimators, we design outliers in data simulated in step (i) by assuming the normal data are from LD distribution with $\mu = 10^{-5}$ and the outliers are selected from the tails (below the 2.5 and above the 97.5 percentile). Figure 2.2 shows that when the number of cultures (out of 15) containing outliers increases from 1 to 7, the MSE (based on 1000 simulations, in log

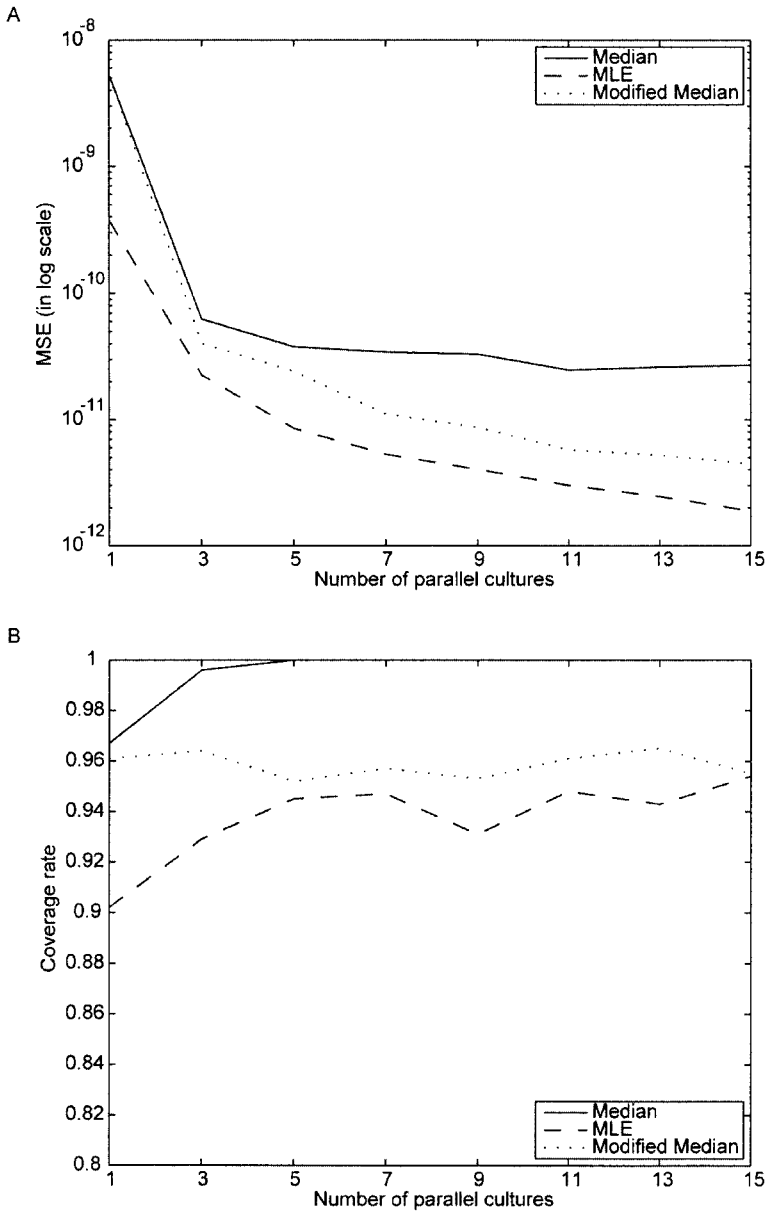


Figure 2.1: (A) Mean squared error (MSE) in log scale versus number of parallel cultures; (B) Coverage rate calculated based on 95% confidence interval versus number of parallel cultures. Parameter setting: $\beta_1 = 0.5, N_0 = 1, N_t \sim$ lognormal with mean $m = 10^6$, standard deviation $s = 5 \times 10^5, \mu = 10^{-5}$, number of simulations = 1000. Solid line: median estimator, dotted line: maximum likelihood estimator, dashed line: modified median estimator.

scale) of the three estimators also increases. Moreover, in such "noisy" environments, the maximum likelihood method gives biased estimates of the mutation rate whereas the median and modified median estimators are robust to the outliers. The modified median estimator has the best performance. This result suggests the superiority of the modified median estimator in real experimental conditions where outliers are unavoidable.

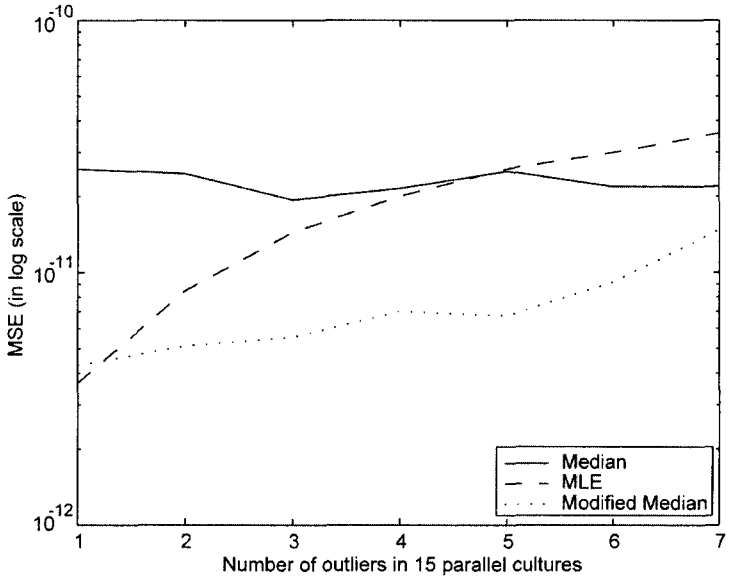


Figure 2.2: MSE in log scale versus number of outliers in 15 cultures. Parameter setting is the same as in Figure 2.1. Outliers (k-values) were randomly chosen from below the 2.5% percentile and from above the 97.5% percentile of the true LD distribution. Solid line: median estimator, dotted line: maximum likelihood estimator, dashed line: modified median estimator.

Simulation study also provides an empirical way to determine the appropriate number of parallel cultures for fluctuation experiments. Figure 2.3 shows the relative MSE, ratio of the square root of the MSE to the mean of the estimates, in our

simulation study. Under the setting of $m = 4 \times 10^6, s = 2 \times 10^6, \mu = 10^{-5}$ (shown in solid line), the relative MSE decreases with slower rates when the number of parallel cultures increases. Therefore, we can choose the optimal number of parallel cultures as a compromise between estimation accuracy and experimental expenses. Other settings of the parameters lead to similar relative MSE pattern, as shown in dashed and dotted lines. From Figure 2.3, the reasonable number of parallel cultures is between 10 and 20, and we chose 15 in our following fluctuation experiments.

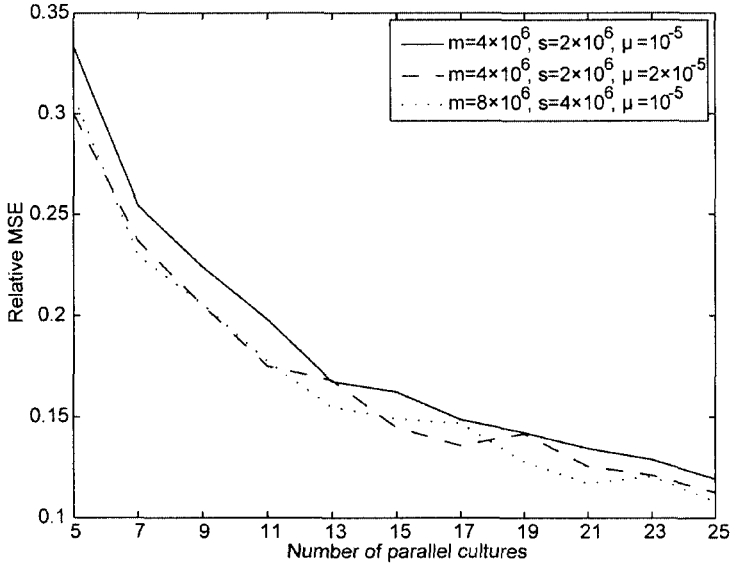


Figure 2.3: Relative MSE versus number of parallel cultures. $N_t \sim$ lognormal with mean m, standard deviation s. Number of simulations = 1000. Solid line: $m = 4 \times 10^6, s = 2 \times 10^6, \mu = 10^{-5}$, dashed line: $m = 4 \times 10^6, s = 2 \times 10^6, \mu = 2 \times 10^{-5}$, dotted line: $m = 8 \times 10^6, s = 4 \times 10^6, \mu = 10^{-5}$.

Repeated experiments or simulations provide the empirical distribution of the mutation rate estimates. The uncertainty in the estimates, here we use CI, is affected

by many factors. To answer the question "how much uncertainty in the estimates is to be expected merely on the basis of the stochastic variability inherent in the sampling process", Stewart provided some suggestions based on simulations to check the bias and standard deviation of the Lea-Coulson estimator and the maximum likelihood estimator [37]. However, these suggestions cannot be applied directly to the modified median estimator because its distribution is of no explicit form. We performed a simulation study to check the CI of the modified median estimator in relation to the settings of N_t and μ.

Our simulation study deals with the case of 15 parallel cultures. The uncertainty in the estimates is indicated by confidence intervals with ends being the 0.025 and 0.975 percentile of the estimates in 1000 simulations. We show the result of these percentiles in Figure 2.4 under a typical setting of N_t with mean $m = 10^6$, coefficient of variation (i.e., the ratio of standard deviation over mean) $\rho = 0.4$, and mutation rate changing from 10^{-6} to 1.5×10^{-5}. This percentile range can be summarized by two regression lines, as shown by the upper and lower solid lines in Figure 2.4. Table 2.1 lists the empirical regression coefficients under other settings of N_t (means: $10^6, 2\times10^6, 4\times10^6$, and 8×10^6, coefficients of variation: $0, 0.2, 0.4, 0.6, 0.8$, and 1). We see that the variability of the estimates is a function of the mean and standard deviation of N_t, as well as of the mutation rate. To provide an example of application of this table, let us consider that in 15 parallel cultures N_t has mean 10^6 and standard deviation 4×10^5 (the corresponding coefficient of variation is 0.4), and suppose the estimate

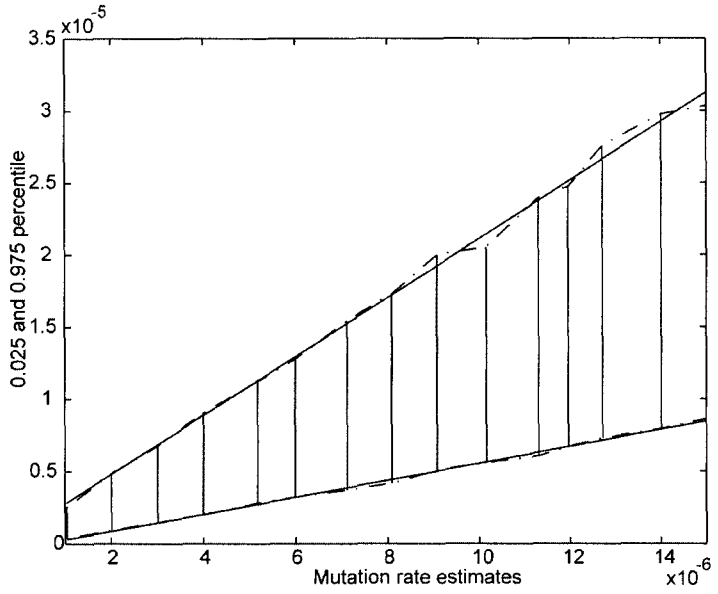


Figure 2.4: The 0.025 and 0.975 percentile of the estimates based on simulations. $N_t \sim$ lognormal with mean $m = 10^6$, coefficient of variation $\rho = 0.2$. Number of simulations = 1000. Dashed line: lower and upper percentiles with mutation rate estimates are from $1.02 \times 10^{-6} to 5.08 \times 10^{-6}$. Solid line: regression lines.

of the mutation rate is $\hat{\mu}$, then with 95% probability $\hat{\mu}$ should fall into the interval $-3.008\times10^{-7}+0.5821\times\hat{\mu}$ to $0.609\times10^{-6}+2.0677\times\hat{\mu}$. In this way Table 2.1 provides a simple empirical rule to compute the confidence intervals of the estimates in repeated experiments. Monotonicity of the percentiles as the mean/standard deviation of N_t changes can be seen from the pattern of the slope coefficients. Let us note that the intercept coefficients play a less important role since they are usually 10 times smaller than the product of the slope coefficient and $\hat{\mu}$.

Another application of Table 2.1 is to investigate whether the dilution procedure in measuring N_t can lead to unexpected large variability in the estimates. In fluctuation experiments, N_t is always very large and can only be approximately measured through serial dilutions. Suppose the true N_t has mean m and standard deviation s, denote the coefficient of variation by $\rho = \frac{s}{m}$. After r dilutions, each with dilution rate $p_i, i = 1, \cdots, r$, this coefficient of variation can be shown using conditional expectation as

$$\rho' = \sqrt{\rho^2 + \frac{1 + p_r + p_r p_{r-1} + \cdots + p_r p_{r-1} \cdots p_2}{p_r p_{r-1} \cdots p_1 \cdot m}}.$$

In our yeast assay, $r = 4$ and $p_i = 0.1$ for $i = 1, \cdots, 4$, m is of scale 10^6 and usually we observe ρ about 0.5, so we can see that the dilution procedure only increases the coefficient of variation of N_t by no more than 1.1%. Using Table 2.1, we see the change of the CI in the estimates is negligible. This eliminates a possible reason of the excess variability in the estimates described below (see Section 2.1.5).

Table 2.1: Practical computation of 95% confidence intervals: Regression coefficients for the 0.025 and 0.975 percentile of the estimates based on 1000 simulations. (A) 0.025 percentile; (B) 0.975 percentile. Rows represent different coefficient of variation of N_t settings ranging from 0 to 1, columns represent different mean settings of N_t ranging from 10^6 to 8×10^6.

A	$m = 10^6$		$m = 2 \times 10^6$		$m = 4 \times 10^6$		$m = 8 \times 10^6$	
	Intercept	slope	Intercept	slope	Intercept	slope	Intercept	slope
$\rho = 0.0$	-3.682×10^{-7}	0.7624	-2.546×10^{-7}	0.7784	-1.964×10^{-7}	0.8034	-1.839×10^{-7}	0.8229
$\rho = 0.2$	-1.620×10^{-7}	0.6447	-1.656×10^{-7}	0.6668	-1.570×10^{-7}	0.6906	-1.432×10^{-7}	0.7060
$\rho = 0.4$	-3.008×10^{-7}	0.5821	-2.681×10^{-7}	0.6176	-1.672×10^{-7}	0.6290	-2.201×10^{-7}	0.6658
$\rho = 0.6$	-3.264×10^{-7}	0.5498	-3.026×10^{-7}	0.5897	-2.912×10^{-7}	0.6240	-2.342×10^{-7}	0.6442
$\rho = 0.8$	-3.442×10^{-7}	0.5248	-4.228×10^{-7}	0.5747	-3.248×10^{-7}	0.6067	-2.751×10^{-7}	0.6395
$\rho = 1.0$	-3.230×10^{-7}	0.5115	-4.120×10^{-7}	0.5705	-2.761×10^{-7}	0.5899	-3.504×10^{-7}	0.6374
B	$m = 10^6$		$m = 2 \times 10^6$		$m = 4 \times 10^6$		$m = 8 \times 10^6$	
	Intercept	slope	Intercept	slope	Intercept	slope	Intercept	slope
$\rho = 0.0$	0.608×10^{-6}	1.4210	0.619×10^{-6}	1.3742	0.416×10^{-6}	1.3508	0.392×10^{-6}	1.3121
$\rho = 0.2$	0.701×10^{-6}	1.6230	0.445×10^{-6}	1.5962	0.495×10^{-6}	1.5627	0.216×10^{-6}	1.5540
$\rho = 0.4$	0.609×10^{-6}	2.0677	0.550×10^{-6}	2.0104	0.549×10^{-6}	1.9602	0.318×10^{-6}	1.9311
$\rho = 0.6$	0.893×10^{-6}	2.4638	0.605×10^{-6}	2.4142	0.458×10^{-6}	2.3440	0.531×10^{-6}	2.3236
$\rho = 0.8$	0.812×10^{-6}	2.9124	1.286×10^{-6}	2.7281	1.211×10^{-6}	2.6121	1.408×10^{-6}	2.5598
$\rho = 1.0$	2.101×10^{-6}	3.1858	1.188×10^{-6}	3.0996	2.384×10^{-6}	2.8878	1.558×10^{-6}	2.8259

2.1.4 Estimation of Mutation Rates in Yeast Strains

Our experiments include two parts, one for budding yeast, *Saccharomyces cerevisiae*, and the other for *Escherichia coli*. The yeast experiments were carried out in the following manner [38], which is a modification of the method described by Klein [22] to measure chromosome V instability in a large number of different strain backgrounds. Wild-type *S. cerevisiae* strains were struck-out on YPD, rich non-selective media, so that colonies arose from single cells and allowed to grow for 3 days at 30^oC. The colonies represent the parallel cultures in the fluctuation experiment. Twenty-four separate colonies per strain were then chosen and dispersed in $200\mu l$ of water each in 96 well plates. Culture size was estimated by measuring absorbance using the Tecan Spectroflour Plus at $620nm$ and the 15 out of 24 colonies with the closest optical densities were carried forward in the experiment. Tenfold serial dilutions up to 10^{-4} in water were made. In one set of experiments (noted as day 3 in Figure 2.5, Part A) the 15 colonies were randomly selected without clustering by optical density. $100\mu l$ of the 10^{-1} dilution was then plated onto a SC-Arg plus canavanine at $60\mu g/ml$ (to determine the number of mutant colonies) and $100\mu l$ of either 10^{-4} or 10^{-5} dilution was plated onto a YPD plate (to determine the number of viable cells) and spread with glass beads. Plates were grown for 3 days at 30^oC and then colonies were counted using the aCOLyte SuperCount Colony Counter. The total number of viable cells N_t and the mutant cells k were used for fluctuation analysis.

The fluctuation experiment is designed to avoid jackpots due to having mutant

cells in the starting culture. Each colony (the parallel culture in question) is started from a single cell. If that cell is mutant then $k = N_t$. Even if one assumes that the colony starts from 2-3 cells, one of which is mutant, this would result in extremely large k's which were not seen on any regular basis.

For the *E. coli* experiments, we use the same data and experimental procedures as described in Hastings et al. [14]. In this experimental protocol, 25 parallel cultures were assayed. Each parallel culture was initiated by a single colony which was then inoculated into liquid media and grown further. Mutant cells were identified at the end of the experiment by their ability to grow in the presence of valine.

Because of its robustness and computational accessibility (compared to the maximum likelihood method) and favorable coverage and accuracy (compared to the method of median), the modified median estimator was chosen to analyze the experimental data obtained from a set of *S. cerevisiae* (budding yeast) strains. In the experiments involving these strains, the "mutation event" is defined as the instability (chromosome loss or mitotic recombination) of Chromosome V during cell division, as measured by conversion of a sensitive strain to a strain resistant to the drug *canavanine*. Statistical inference for Chromosome V instability rates (μ_{CV}) of the wild-type strains, including estimation and hypothesis testing, is based on 15 parallel cultures in every experiment. We use box plot for graphical depiction of the estimates of μ_{CV} derived by the modified median method in the yeast fluctuation experiments. Figure 2.5, Part A shows the result of this method for the chromosome V instability assay

in a wild-type *S. cerevisiae* strain. In this figure, each box summarizes the distribution of the μ_{CV} estimates based on data from the 15 parallel cultures assayed in one experiment. The middle line in the box represents the modified median estimator of the corresponding strain. As noticed in the beginning of this section, we initially performed a total of 15 repeated experiments, each on 15 parallel cultures, on two separate days using the same wild-type strain. We then followed up with additional 5 repeats carried out over a year later, by a different investigator. The corresponding μ_{CV} estimates are shown by the 20 boxes in Figure 2.5, Part A. In comparison, the μ_{CV} derived using the standard median estimator for each experiment is depicted by an asterisk. The differences in the estimates of μ_{CV} between the modified and standard median estimators for each experiment reflect the effect of unequal N_t of the parallel cultures. In other words, the modified median estimator is taking into account variability in N_t to predict the mutation rate.

2.1.5 Variability of Estimates and Dependence of Mutation Rates on Cell Population Size

In the previous section we studied the estimation of μ_{CV} in yeast fluctuation experiments. To examine reproducibility, each experiment was replicated a number of times on the same day and/or different days. Theoretically, under the same experimental conditions and using the same LD model, the estimates based on independent repeats of the experiment involving the same strain should show limited variability.

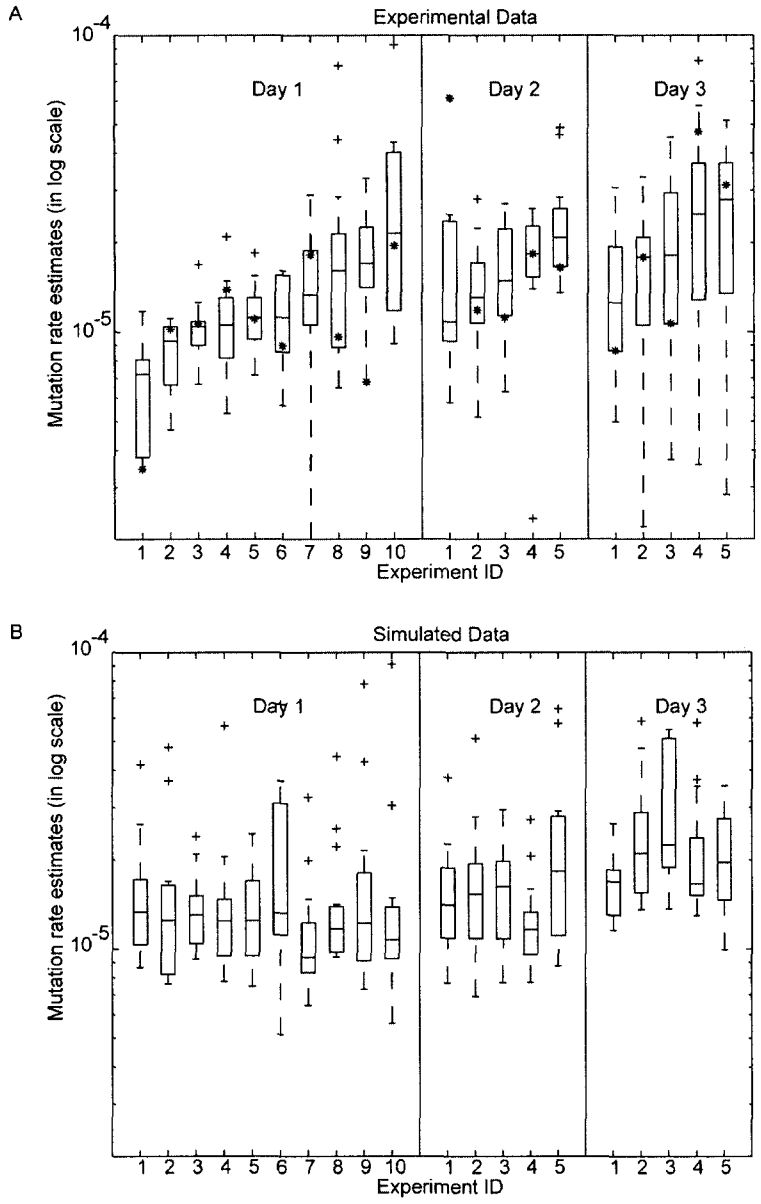


Figure 2.5: Box plot of mutation rate estimates. Each box represents summary statistics of 15 mutation rate estimates (using the modified median estimator where we treat the mutant colony count per culture as the median of the size 1 sample) in the parallel cultures of each strain. Within each box, the middle line represents the modified median estimator; the asterisk represents the median estimator. (A) Box plot of mutation rate estimates in yeast wild-type strains in 20 replicate experiments on 3 separate days, ordered by the point estimates. Day 3 experiments were done separated in time from Days 1 and 2 by an independent investigator. In Day 3 experiments the 15 cultures (colonies) were not first clustered by size at the beginning of the experiment (see Section 2.1.4). (B) Box plot of mutation rate estimates in simulated data, ordered by the point estimates in Panel (A).

However, it is observed that the variability of the estimates in replicate experiments exceeds the simulated one due to unknown effects. This effect is visualized in Figure 2.5, which compares the estimates of 20 replicate experiments carried out on three separate days (Figure 2.5, part A) with their simulated counterparts (Figure 2.5, Part B). In both parts of Figure 2.5, the left panel depicts 10 replicate experiments on Day 1, the center panel depicts 5 replicate experiments on Day 2, whereas the right panel depicts 5 additional experiments performed by a different investigator on "Day 3" over a year after Days 1 and 2. As already remarked, the design of our estimator allows visualizing the within-replicate variability in a direct way, by plotting estimates of mutation rates based on individual replicates. Comparison of the Day $1, 2$ and 3 data demonstrates a remarkable reproducibility of the magnitude of estimated mutation rates and also of the among-replicate variability.

The simulated data were generated using the same population sizes as in our repeated wild-type experiments, and assuming a common μ_{CV} for the Day 1 and Day 2 plots (based on combined estimates of all 20 Day 1 and Day 2 experiments), and another μ_{CV} for the Day 3 plot (based on combined estimates of 5 Day 3 experiments). Obviously, point estimators of the 15 replicates (shown as the middle line in each box) in Figure 2.5, Part A show higher variability than those in Figure 2.5, Part B. We note that the absolute difference in these estimates in the 20 experiments shown in Figure 2.5, Part A varies less then fourfold (from a minimum of 7.21×10^{-6} to a maximum of 2.79×10^{-5}). Similarly, in most other cases of replicate experiments

reported by Strome et al. (ref. [38], not shown), the variation of estimates among different day replicates is also less than fourfold. However, the corresponding μ_{CV} difference based on simulations in Figure 2.5, Part B is only about twofold (from a minimum of 9.38×10^{-6} to a maximum of 2.24×10^{-5}). For many experimental systems where biologists are comparing strains whose mutation rate may vary by several units in the logarithmic scale, this excess variability is often overlooked, but needs to be strictly investigated from a statistical point of view since it may restrict model applicability.

Confidence intervals are also helpful in judging the variability of the estimates. Reproducibility of experiments, or low variability in the estimates, leads to a high overlap rate among CIs. To realize this, notice that, for two independent experiments concerning the same μ_{CV}, if we use 95% CI, then with probability 0.952 both intervals should contain the true parameter, and hence overlap. Accordingly, the variability difference between the estimates coming from experimental data and those coming from simulated data can be seen more clearly through the comparison of the overlapping of CIs. Figure 2.6, part A shows the CIs of the modified median estimators of the 20 experiments with the wild-type strain. Figure 2.6, Part B shows the CIs using the same simulated data as in Figure 2.5, part B. The comparison suggests that the estimates of μ_{CV} based on the experimental data have larger variability than those based on the simulated data.

This excess variability of the estimates reflects either fluctuations in the underly-

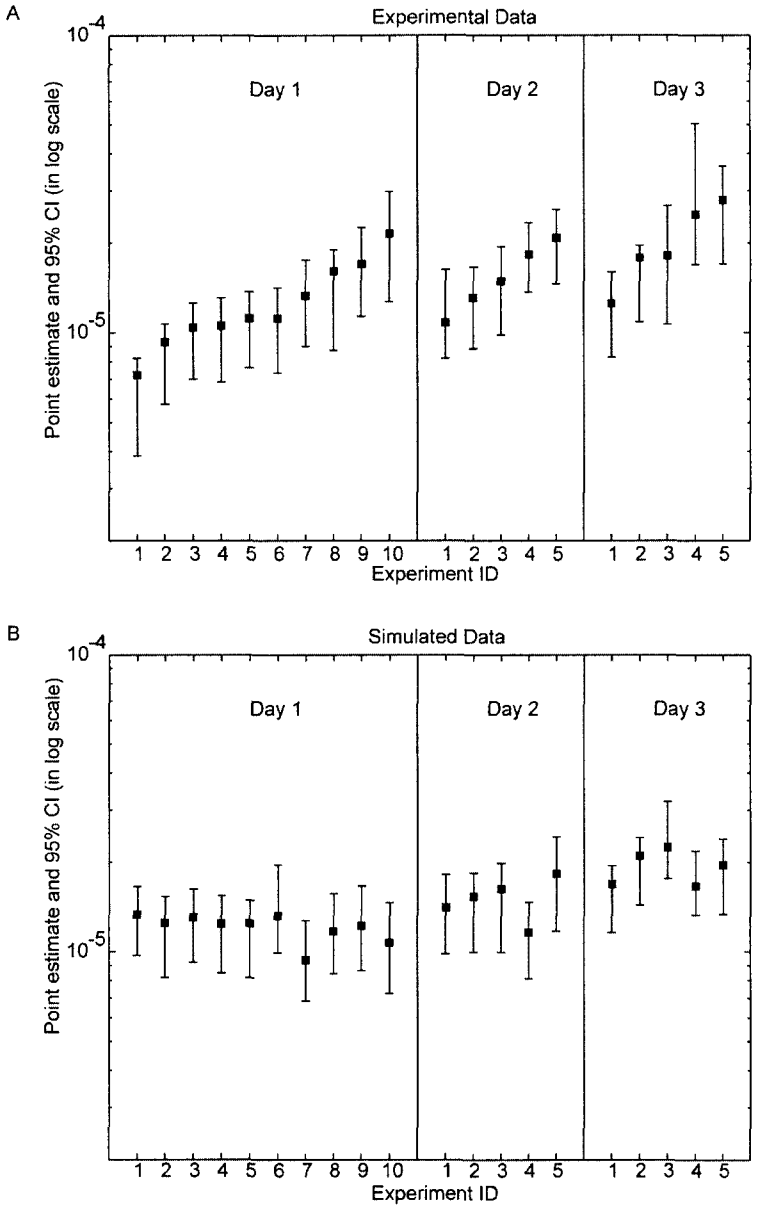


Figure 2.6: Confidence intervals of mutation rate estimation. (A) Confidence intervals of mutation rate estimation in yeast wild-type strains in 20 replicate experiments on 3 separate days, ordered by the point estimates. (B) Confidence intervals of mutation rate estimates in simulated data, ordered by the point estimates in Panel (A).

ing μ_{CV} of the repeated experiments on the same yeast strain, or variability of the estimates due to deviations from the LD distribution. In order to answer "what is the basis for this variation", we need to analyze the variability problem in detail. Suppose in the total 15 experiments performed on Day 1 and Day 2 (Day 3 experiments select colonies differently with Day 1 and Day 2 experiments, to avoid this confounding, we only consider Day 1 and Day 2 experiments), the underlying μ_{CV} is $\mu_1, \mu_2, \cdots, \mu_15$ respectively, and the corresponding estimates are $\hat{\mu}_{ij}$, $i = 1, \cdots, 15$ representing the experiment number and $j = 1, \cdots, 15$ representing the culture number in each experiment. To check whether there exist fluctuations in the underlying μ_{CV}, we need to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_15$ versus the alternative H_1 : there is at least one μ_i different from others. This is accomplished by an approach similar to analysis-of-variance (ANOVA). First we define the test statistic (F-statistic in ANOVA) to be the ratio of the "among-experiment-variance" to the "within-experiment-variance", where the "among-experiment-variance" is the sample variance of the modified median estimates, and the "within-experiment-variance" is the average of the sample variances of within each experiment. Second, we find the distribution of this statistic under null hypothesis H_0. Due to lack of explicit form, this is accomplished by permutation, i.e., we randomly permute among the experiments and calculate the value of the test statistic after permutation. Using this approach, we obtain a highly significant p-value equal to 0.0001 based on the experimental data. As a supplementary control, we determine that the p-value based

on simulated data with no "among-experiment-variance" is equal to 0.7732. This provides us sufficient evidence that there exist fluctuations in the underlying μ_{CV} in our experimental data.

To identify possible reasons for variability of the estimates, we consider our expectation of constant mutation rate. Figure 2.7, Part A shows the scatter plot of μ_{CV} estimates versus population sizes of the 300 cultures (20 replicate experiments of 15 parallel cultures each) for the wild-type *S. cerevisiae* strain (same data as in Figure 2.5, Part A and Figure 2.6, Part A). Under our experimental conditions, a significant inverse relationship can be found within certain population size range ($N_t < 8 \times 10^6$ cells/culture). This clearly contradicts the starting assumption of constant mutation rate. For comparison, the analogous scatter plot using simulated data (from the same simulations as those in Figure 2.5, Part B and Figure 2.6, Part B) is shown in 2.7, Part B. This scatter plot shows uniformity of the mutation rate estimates, as expected under the constant mutation rate used in the simulation. Similar scatter plots as 2.7, Part A can be observed if we assume an inverse relationship between N_t and μ when simulating data. This suggests that there may exist an inverse relationship between the population size N_t and the mutation rate μ under certain experimental conditions.

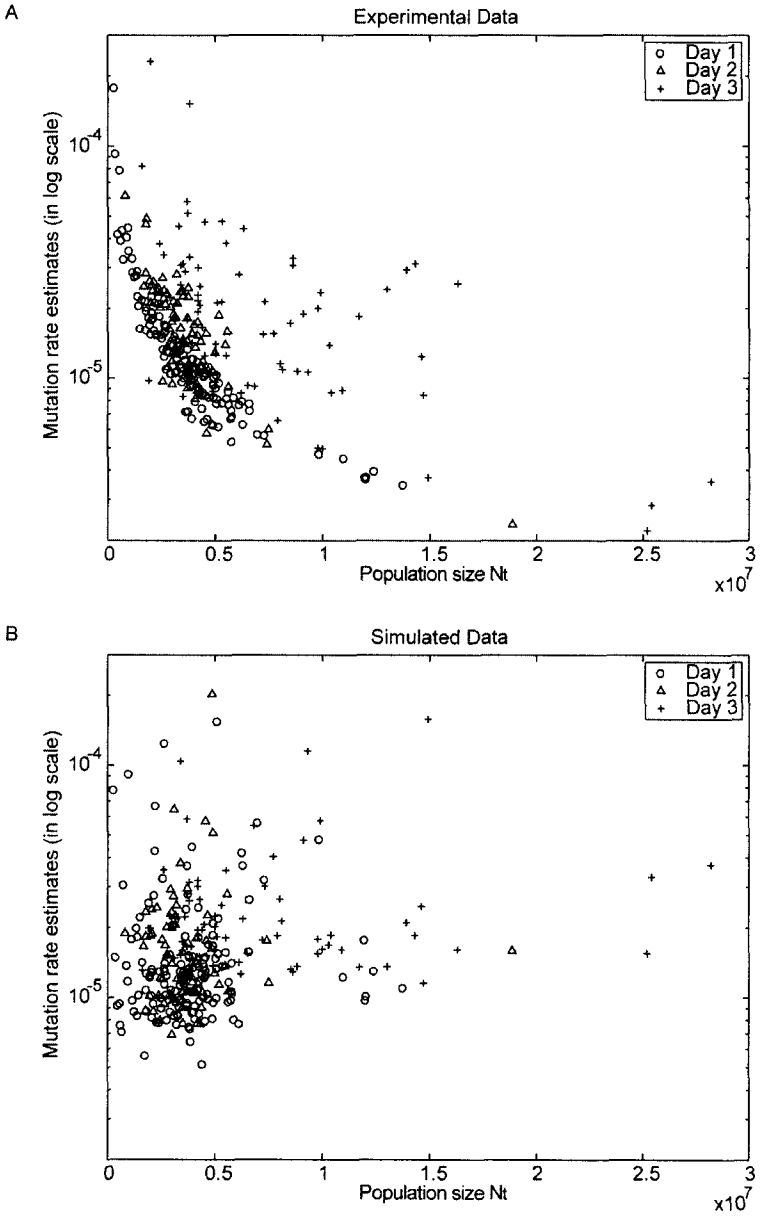


Figure 2.7: Scatter plot of mutation rate estimates versus population sizes. (A) Budding yeast data in wild-type background on 3 separate days, (B) Simulated data using constant mutation rate assumption.

2.1.6 Discussion

We report in this chapter the derivation of a modified median estimator of mutation rates, and its application to a number of Luria-Delbrück type fluctuation experiments in *S. cerevisiae*. Through simulation, we see that this estimator is accurate, reliable and robust. In addition, this estimator provides flexibility as it provides an estimate of the mutation rate from each individual parallel culture in a fluctuation analysis experiment, which helps in experiments where N_t varies and in detecting variation of mutation rates and exploring the causes of variation. Using *S. cerevisiae* data, we found excessive variation compared to that seen in the simulated data. This finding demonstrates that the half-century old fluctuation experiment deserves a new look.

We first validated the modified median estimator by comparing it to the usually used median estimator and to the maximum likelihood estimator, using an array of simulations. We also demonstrated that mutation rate estimates for strains known to differ in mutation rate, e.g. wild-type versus *rad*9 deficient were comparable to other published data [38].

As a conclusion, although the modified median estimator may be less accurate than the maximum likelihood estimator under ideal conditions, such as when all data are independently generated from one and the same LD distribution, it seems more robust to outliers in the data. Therefore, it appears to be superior under experimental conditions. The estimator exhibits good coverage properties in simulated data.

Application of the modified median estimator thus leads to increased accuracy and robustness of mutation rate estimation in fluctuation experiments. Also, as documented in Figure 2.3, the relative MSE of the mutation rate estimates using the modified median method falls below 20% if 15 cultures are used for analysis (solid line). This result provides a practical indication on "how many cultures should be used" for researchers planning to perform fluctuation analysis experiments to estimate mutation rates. Generally, the number of cultures needed to reduce the relative MSE below certain level depends on the population size and the underlying mutation rate for the experiment in question.

We applied the modified median estimator to estimate mutation rates in replicate experiments for a large number of yeast strains derived in a mutagenesis screen as described in Strome *et al.* [38]. The baseline computations on the starting wild-type yeast strain disclosed unexpected effects compared to the simulated data as described here. We detected excess variability of estimates in certain experimental conditions, which was not considered in other models. First, the twenty replicate experiments using the same wild-type yeast strain show a fourfold range of among-experiment variability of mutation rates. The variability exceeds the level consistent with the confidence intervals computed using not only our method but also the method of maximum likelihood. Although not important when comparing strains that differ in their mutation rates by several units in the logarithmic scale, this effect may be particularly important when designing experiments to compare strains whose mutation

rates may only differ slightly.

We decided to try to identify the causes of higher than expected variability of estimates. There may be several reasons for the variability of the estimates. First, it may be impossible to control the experimental conditions of replicate experiments to be exactly the same, especially when experiments are performed on different days. However, we saw similar variability among replicate experiments performed on the same day as for experiments done on several different days, by different investigators. Second, the LD model may not be flexible enough to explain the data. The growth and mutation process of a cell population may involve cell death and variation of mutation rates. As described below, the mutation rate (either per unit time or per cell division) may depend on the population size N_t at each time point. As a consequence of this or similar effects, final experimental data might not follow the simple LD distribution, but a complex distribution determined by the mutually dependent parameters (N_t, μ). From a mathematical viewpoint, to generalize the LD model, this means considering a population size-dependent process instead of a Markov branching process [3, 21].

In summary, the modified median estimator has been designed for two features, which are important for mutation rate analysis, i.e., robustness to outliers and control of variability of estimates. Particularly the latter feature is important as it may provide a lower limit for detection of differences when comparing mutation rate between different strains in large series of data (more than 100 strains, as in Strome *et al.* [38]).

As documented in Results, exploratory analysis using the budding yeast data shows an interesting inverse relationship between the population size N_t and the estimated mutation rate under certain experimental conditions. To verify whether this relationship holds more generally in clonal growth and mutation process, we also investigated data derived from measuring mutation rate of a series of strains of *Escherichia coli* using parallel liquid cultures (for details, see [14]). Figure 2.8 shows the box plot and scatter plot for this bacterial dataset and simulated data analogous to those for yeast and simulated data (Figure 2.5 and Figure 2.7). We see that in the bacterial case, although the mutation rate estimates in experimental data still have larger variability than those in simulated data (comparing Figure 2.8, Part A with C), the dependence on N_t is not present (Figure 2.8, Part B) which is similar to what we expect and observe in simulated data (Figure 2.8, Part D). This indicates that the dependence on N_t is not a unique explanation of excess variability.

Another factor may be the biological process being measured itself. As an example for the yeast experiments described here, other investigators have reported that once diploid yeast cells lose one copy of Chromosome V, they grow slower and then rapidly reduplicate the mutant chromosome [39, 42]. Thus, growth rate of mutant cells may vary during the growth of the culture which is not considered in the model. Another example of these phenomena is the work of Boesen *et al.* [4]. They demonstrate in an assay of mutation rate in a mouse lymphoma cell line that the mutation rate varies over a tenfold range during the growth of the culture only when they allow

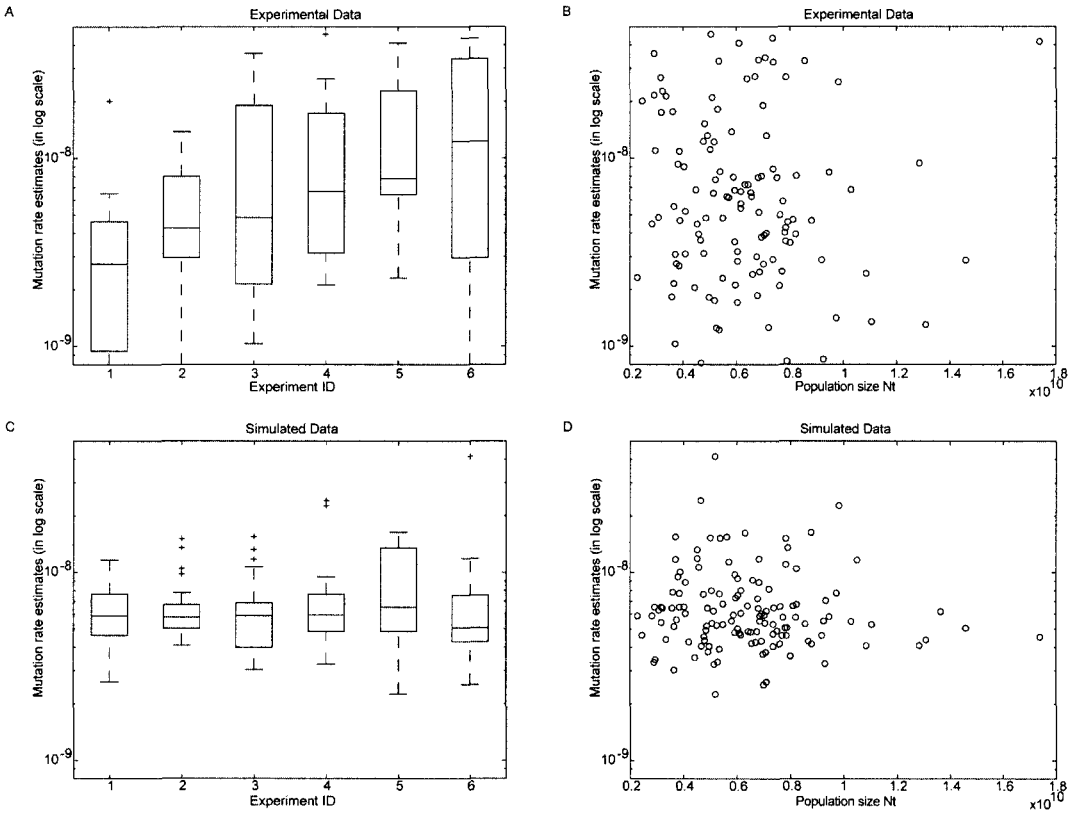


Figure 2.8: Box plot of mutation rate estimates and scatter plot of mutation rate estimates versus population sizes. (A) Bacterial data, box plot, (B) Bacterial data, scatter plot; (C) Simulated data using constant mutation rate assumption, box plot, (D) Simulated data, scatter plot.

growth conditions to become limiting over this growths period. The Luria-Delbrück model may be extended by allowing different growth rates of wild-type and mutant cells, as seen in [44]. For more general cases where cell growth rate (for both mutant and wild-type cells) and mutation rate may be affected by several factors, a deep exploration to the modeling methodology is needed.

However, returning to the colony size-dependence issue, it is noted that bacterial laboratory populations used in fluctuation experiments are generally larger and the process being monitored in this experiment has a lower mutation rate than in the chromosome instability assay in yeast. Usually, a *S. cerevisiae* colony (the parallel culture used in this assay) contains 10^6-10^7 cells and the mutation rates in this assay (chromosome loss and/or recombination) are on the order of 10^{-5}, whereas the bacterial cultures contain 10^9-10^{10} cells and the typical mutation rates are on the order of 10^{-8}. In Figure 2.7, Part A, the inverse relationship holds only for small ($< 8 \times 10^6$) values of N_t, but it tends to disappear for high values of N_t, which may be the reason it did not manifest itself in bacterial data (the right-hand side tail of the scatter plot in Figure 2.7, Part A seems consistent with Figure 2.8, Part B).

To minimize this variability, it might be possible to use assays with higher values of N_t. This will result in higher k-counts (on the order of 10^3-10^4), particularly for high mutation rates, such as the chromosome instability rates investigated in [38] and in the current paper. Unfortunately, because of the long right-hand tail of the LD distribution, this may lead to computational problems. In such situations, it may

prove practical to use the approximate expression of Lea and Coulson [23].

This analysis has implications for other experimental systems when N_t may be limited. We note that Kimmel and Axelrod reviewed several data sets used for fluctuation analysis in mammalian cells, mostly in a mutation to resistance context [20]. In these experiments, N_t varies within the 10^5-10^7 range, similar to the yeast experiments of Strome *et al.* [38], and therefore it will be important to determine whether there is a dependence of the mutation rate on population size in these experimental systems as well. Future studies will focus on relaxing the model assumptions to reduce the variability in mutation rate estimates.

2.2 Modeling by Two-Type Markov branching processes

An alternative way to model clonal growth and mutation in cell population is by two-type Markov branching processes (MBP). We assume that in a cell population, the life spans of cells are i.i.d. exponential random variables. Every cell (in general context of MBP, it is also called individual or particle) produces a number of offspring at its death. The number of offspring is a non-negative discrete random variable. There are two types of cell in the population: wild-type and mutant. Depending on its parental type, a new-born cell can mutate into the other type with a probability.

2.2.1 Preliminaries of Markov branching processes

Suppose the cell population can be modeled by a continuous-time MBP Z_t. The process Z_t has offspring distribution with pgf $f(s) = \sum_{k=0}^{\infty} p_k s^k$. Suppose the life length of each cell follows exponential distribution with parameter a. Denote the pgf of the process Z_t by $F(s,t) = E_1\left[s^{Z_t}\right]$. Here subscript 1 means that the process starts from a single cell. Then $F(s,t)$ satisfies the forward Kolmogorov equation

$$\frac{\partial F(s,t)}{\partial t} = \phi(s)\frac{\partial F(s,t)}{\partial s},$$

where $\phi(s) = a(f(s) - s)$, and the backward Kolmogorov equation

$$\frac{\partial F(s,t)}{\partial t} = \phi(F(s,t)), \tag{2.3}$$

with boundary condition $F(s,0) = s$; see [3] for details.

Denote the mean of the offspring distribution by m. Taking derivatives on both sides of Equation (2.3) with respect to s, as $s \uparrow 1$, we obtain such an ordinary differential equation (ODE) for $E_1[Z_t]$:

$$\frac{dE_1[Z_t]}{dt} = a(m-1)E_1[Z_t].$$

Solving this equation, we see that

$$E_1[Z_t] = e^{\lambda t}, \tag{2.4}$$

where $\lambda = a(m-1)$ is the Malthusian parameter. Equation 2.4 describes the exponential growth of continuous-time MBP. This equation will be of use further on in the next section.

2.2.2 Estimation of Mutation Probability

We model the clonal growth by a supercritical MBP, starting from one wild-type ancestral cell. As usual, we denote the offspring pgf by $f(s)$ and the parameter of the exponential life span by a. This process is completely described by the backward Kolmogorov equation (2.3).

We further assume that: (i) every wild-type cell can mutate with probability μ; (ii) mutation is non-reversible; and (iii) mutant cells have the same offspring distribution and life span as wild-type cells. Denote the offspring pgf initiated by a wild-type cell by $f^{(0)}(\boldsymbol{s})$, and the offspring pgf initiated by a mutant cell by $f^{(1)}(\boldsymbol{s})$. Similarly, denote the pgf of the process initiated by a wild-type cell by $F_0(\boldsymbol{s},t)$, the pgf of the process initiated by a mutant cell by $F_1(\boldsymbol{s},t)$, and write $\boldsymbol{F}(\boldsymbol{s},t) = (F_0(\boldsymbol{s},t), F_1(\boldsymbol{s},t))$. The backward Kolmogorov equations of this two-type MBP have the form:

$$\begin{cases} \frac{\partial F_0(\boldsymbol{s},t)}{\partial t} = a[f^{(0)}(\boldsymbol{F}(\boldsymbol{s},t)) - F_0(\boldsymbol{s},t)] \\ \frac{\partial F_1(\boldsymbol{s},t)}{\partial t} = a[f^{(1)}(\boldsymbol{F}(\boldsymbol{s},t)) - F_1(\boldsymbol{s},t)] \end{cases}, \tag{2.5}$$

with boundary conditions $F_0(\boldsymbol{s},0) = s_0$, $F_1(\boldsymbol{s},0) = s_1$.

Now suppose the offspring distribution is of simple binary-fission type (Yule process). Suppose the mutation is non-reversible and occurs with probability μ, so that $f^{(0)}(\boldsymbol{s}) = (1-\mu)^2 s_0^2 + 2\mu(1-\mu)s_0 s_1 + \mu^2 s_1^2$, $f^{(1)}(\boldsymbol{s}) = s_1^2$, and $F_1(\boldsymbol{s},t)$ is only a function of s_1 and t. Accordingly, (2.5) becomes:

$$\begin{cases} \frac{\partial F_0(\boldsymbol{s},t)}{\partial t} = a[(1-\mu)^2 F_0^2(\boldsymbol{s},t) + 2\mu(1-\mu)F_0(\boldsymbol{s},t)F_1(s_1,t) + \mu^2 F_1^2(s_1,t) - F_0(\boldsymbol{s},t)] \\ \frac{\partial F_1(s_1,t)}{\partial t} = a[F_1^2(s_1,t) - F_1(s_1,t)] \end{cases} \tag{2.6}$$

The analytic solution to (2.6) may be found in Kimmel and Axelrod [21]. It requires some preliminary knowledge of Riccatti-type equations. In our context, the purpose is to estimate parameter μ using observed data, i.e., the number of wild-type and mutant cells at time t in repeated experiments. For this purpose, the explicit pgf solution is difficult to apply in the maximum likelihood sense. Therefore, we turn to the method of moments, which needs only the moment information instead of the whole pgf, so that derivation of the explicit pgf is not necessary here.

Taking derivatives on both sides of the first equation in (2.6) with respect to s_0 and s_1, we obtain ODEs for the expected numbers of wild-type and mutant cells at time t in a population when initiated by a wild-type cell. The solution turns out to be:

$$\begin{cases} M_0(t) = e^{a(1-2\mu)t} \\ M_1(t) = e^{at} - e^{a(1-2\mu)t} \end{cases} . \tag{2.7}$$

This result is easy to interpret. For Markov branching processes, it can be seen by Equation (2.4) that the mean process behaves as an exponential function of time. If the initial population size is 1, then at time t, the expected population size $E[Z_t] = e^{\lambda t}$. For the binary-fission case, suppose the two-type MBP is initiated by a wild-type cell, it is easy to see that the expected total population size at time t is e^{at}, since $m = 2$. Under the above assumptions (i), (ii) and (iii), the non-mutant process has $m = 2(1-\mu)$, so the expected non-mutant population size at time t is $e^{a(1-2\mu)t}$. This explains $M_0(t)$ and $M_1(t)$ in (2.7).

Based on this result, the parameter μ can be estimated by the method of moments

if data from repeated experiments are available. Note, the meaning of repeated experiments here is equivalent to parallel cultures in fluctuation experiments, see Section 2.1.1. Suppose we perform n repeated experiments, and count wild-type and mutant cells at the same time for all the experiments. Denote the numbers of wild-type and mutant cells at time t in experiment i by $m_0^{(i)}(t)$ and $m_1^{(i)}(t)$, and define the average counts:

$$\begin{aligned} m_0(t) &= \frac{1}{n}\sum_{i=1}^{n} m_0^{(i)}(t), \\ m_1(t) &= \frac{1}{n}\sum_{i=1}^{n} m_1^{(i)}(t). \end{aligned}$$

By the method of moments, the mutation probability can be estimated as:

$$\hat{\mu} = \frac{1}{2}\left[1 - \frac{\ln m_0(t)}{\ln(m_0(t) + m_1(t))}\right]. \tag{2.8}$$

A slight extension is to allow cell death. Suppose the offspring distribution is birth-death with death probability q, i.e., $f^{(0)}(\boldsymbol{s}) = (1-q)(1-\mu)^2 s_0^2 + 2(1-q)\mu(1-\mu)s_0 s_1 + (1-q)\mu^2 s_1^2 + q$, $f^{(1)}(\boldsymbol{s}) = (1-q)s_1^2 + q$. Then (2.5) becomes:

$$\begin{cases} \frac{\partial F_0}{\partial t} = a[(1-q)(1-\mu)^2 F_0^2 + 2(1-q)\mu(1-\mu)F_0 F_1 + (1-q)\mu^2 F_1^2 + q - F_0] \\ \frac{\partial F_1}{\partial t} = a[(1-q)F_1^2 + q - F_1] \end{cases}.$$

Similarly, we obtain the expected numbers of wild-type and mutant cells at time t when initiated by a wild-type cell:

$$\begin{cases} M_0(t) = e^{a(1-2\mu-2q+2q\mu)t} \\ M_1(t) = e^{a(1-2q)t} - e^{a(1-2\mu-2q+2q\mu)t} \end{cases}.$$

Therefore,

$$\hat{\mu} = \frac{\ln(m_0(t) + m_1(t)) - \ln m_0(t)}{\ln(m_0(t) + m_1(t)) + at}.$$

Clearly, in this case, the composite parameter at must be known to estimate μ or q.

In a more general case, wild-type and mutant cells may have different life spans a_0, a_1 and different death rates q_0, q_1. We also allow backward mutations. Denote the backward mutation rate by ν, then (2.5) becomes:

$$\begin{cases} \frac{\partial F_0}{\partial t} = a_0[(1-q_0)(1-\mu)^2F_0^2 + 2(1-q_0)\mu(1-\mu)F_0F_1 + (1-q_0)\mu^2F_1^2 + q_0 - F_0] \\ \frac{\partial F_1}{\partial t} = a_1[(1-q_1)(1-\nu)^2F_1^2 + 2(1-q_1)\nu(1-\nu)F_0F_1 + (1-q_1)\nu^2F_0^2 + q_1 - F_1] \end{cases}.$$

Again, the expected numbers of wild-type and mutant cells at time t when initiated by a wild-type cell can be obtained by solving such an ODE array:

$$\begin{cases} \frac{dM_0(t)}{dt} = a_0\left[(1-2\mu-2q_0+2q_0\mu)M_0(t) + 2(1-q_0)\mu N_0\right] \\ \frac{dM_1(t)}{dt} = a_0\left[(1-2\mu-2q_0+2q_0\mu)M_1(t) + 2(1-q_0)\mu N_1\right] \\ \frac{dN_0(t)}{dt} = a_1\left[(1-2\nu-2q_1+2q_1\nu)N_0(t) + 2(1-q_1)\nu M_0\right] \\ \frac{dN_1(t)}{dt} = a_1\left[(1-2\nu-2q_1+2q_1\nu)N_1(t) + 2(1-q_1)\nu M_1\right] \end{cases},$$

where N_0 and N_1 represent the expected numbers of wild-type and mutant cells when initiated by a mutant cell.

2.2.3 Simulation Study and Discussion

We perform a simulation study to illustrate the model of Yule process. The simulation of MBP requires a self-recurrent technique to realize cell reproduction procedure. This programming is done in Matlab. Started from an ancestral cell

(generation 0), we record the birth and death time for every cell in later generations, as well as its allelic type and genealogy information, up to some time limit. In the simulation study, we set the life time of every cell as an exponential random variable with parameter 1, and set the mutation probability $\mu = 0.05$. Finally we count the number of wild-type and mutant cells and use the average among repeated simulations to obtain $m_0(t)$ and $m_1(t)$. The simulation is repeated for 100 times, and the counting procedure is done every 0.03 time unit. Due to computational limitation, we only simulate a population with 14 generations.

Figure 2.9 illustrates the trajectories of average population size for both wild-type and mutant cells over time, based on 100 simulations. We see that the empirical averages match the expected curves obtained through Equation (2.7), which shows that our simulation is indeed based on MBP. We then use Equation (2.8) to estimate the mutation probability μ at every time point and plot the result in Figure 2.10. We see from Figure 2.10 that the estimates do follow the true mutation probability, except for the initial stage when the population size is very small.

Similarly as in Kimmel and Axelrod [21], we list in Table 2.2 some characteristics related to the estimation of mutation rate for different models. Derivations concerning the two-type MBP model is easy to obtain from Section 2.2.2. The other derivations can be found in Kimmel and Axelrod [20].

We note that there are some limitations for the estimation of mutation probability using the two-type MBP model. First, in order to apply the method of moments,

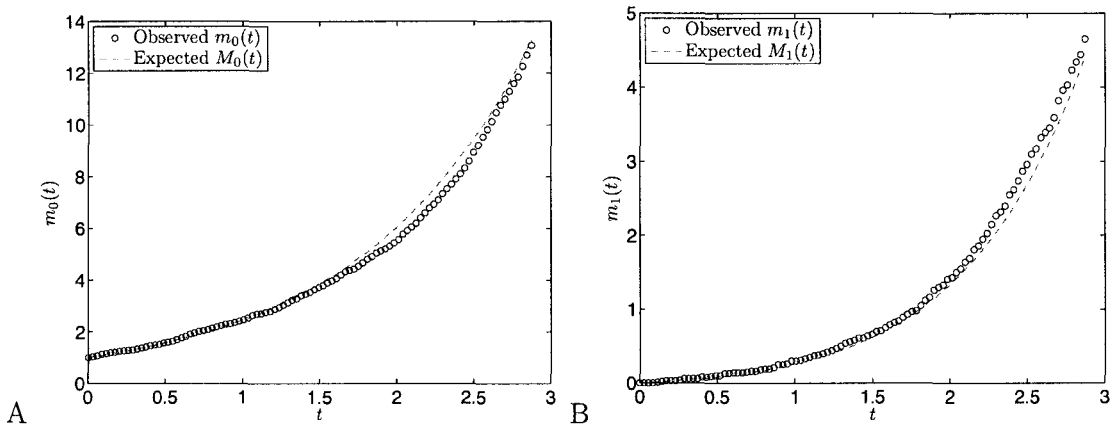


Figure 2.9: Average counts over time for wild-type and mutant cells. (A) number of wild-type cells $m_0(t)$; (B) number of mutant cells $m_1(t)$.

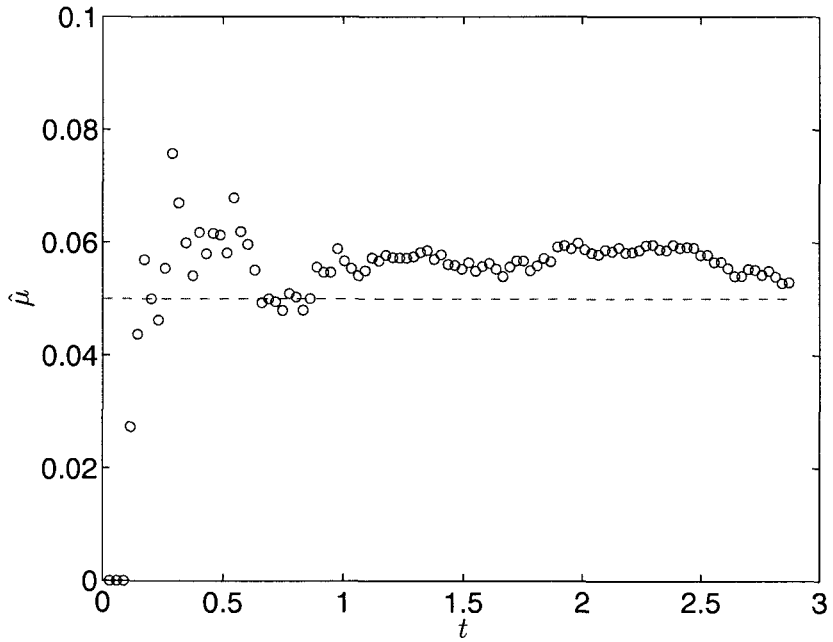


Figure 2.10: Estimates of mutation probability at every time point.

Table 2.2: Comparison of different models in describing cell population dynamics.

Model	Fluctuation Analysis	MBP (Yule)	MBP (Birth-Death)
Expected number of all viable cells	$e^{\beta_1 t}$	e^{at}	$e^{a(1-2q)t}$
Expected number of wild-type cells	$(1-\mu t)e^{\beta_1 t}$	$e^{a(1-2\mu)t}$	$e^{a(1-2\mu-2q+2q\mu)t}$
Expected number of mutant cells	$\mu t e^{\beta_1 t}$	$e^{at}-e^{a(1-2\mu)t}$	$e^{a(1-2q)t}-e^{a(1-2\mu-2q+2q\mu)t}$

data from repeated experiments must be obtained at the same time t. This condition seems to be always satisfied for parallel cultures in fluctuation experiments. However, when we repeat the same experiment for multiple times, as done in Section 2.1.4, these data collected from different days cannot be integrated for the purpose of mutation probability estimation. For this reason, it is not appropriate to apply this model to the yeast experimental data. Second, as a moment-based estimation method, it encounters the same robustness problem as the mean estimator in fluctuation analysis has. Therefore, in real experiments where noise always plays a role, the estimation may be biased and not reliable. Third, it is not easy to provide confidence intervals in the estimation of mutation probability based on the two-type MBP model. Alternatively, we may be able to use variance (second moment) information from the method of moments.

Chapter 3

Extinction of Markov Branching Processes

Extinction problems in branching processes have been studied for a long time since the theory of branching processes started in the middle of the nineteenth century. Besides the well-known facts about the extinction probability and its asymptotics, a well-developed topic is the time T to extinction, in a range of situations (see, e.g. [17, 31, 33]). However, the time to extinction alone is not enough to characterize the full picture of a branching process on its path towards extinction. There are still some other problems concerning extinction that need to be explored. Recently, Jagers *et al.* studied population size partway (or u-way) to extinction, namely Z_{uT}, under a suitable normalization, for subcritical general branching processes [17]. To depict the last stage of extinction in more detail, in their another paper, the same authors

obtained results on the "path on the verge of extinction", i.e., Z_{T-u}, for subcritical Markov branching processes [16].

This chapter summarizes the three topics of extinction in subcritical MBP, i.e., time to extinction, path to extinction and path verging on extinction, and provides systematic proofs. Moreover, results concerning these topics are extended to the critical case with finite variance. Monte Carlo simulations are performed to illustrate the theoretical results for both subcritical and critical cases.

This study is motivated by the approximation of genetic drift using branching processes.

3.1 Motivation: Modeling Genetic Drift

Random genetic drift refers to the fluctuations in allele frequency, occurring particularly in small populations as a result of random sampling among gametes [10, 13]. Other systematic evolutionary forces, such as selection, mutation and migration, cause nonrandom changes in allele frequency. There are different approaches to model genetic drift. In particular, we are interested in Nagylaki's approximation of genetic drift for rare alleles.

3.1.1 Background Concerning Genetic Drift

Genetic drift was first introduced by Sewall Wright, one of the founders in the field of population genetics. It is known that genetic drift is one of several evolutionary

forces which lead to changes in allele frequencies over time, for example, certain alleles becoming fixed and others being lost. The role of genetic drift in the evolutionary scheme has raised a vigorous debate among population geneticists.

The Wright-Fisher model describes the process of genetic drift in a finite population [9, 40]. The model assumes that gametes are chosen randomly each generation from an effectively infinite gamete pool reflecting the parental allele frequencies, and the sampling is binomial (for two-allele case) without considering the effect of mutation or selection. Suppose that we have a population of N diploid individuals. For a single locus, there are two alleles A_1, A_2. In this finite population in which drift alone is acting, if initially there are i copies of the A_1 allele, then the probability that this population ends up with j copies of the A_1 allele after one generation is given by

$$P_{ij} = \begin{pmatrix} 2N \\ j \end{pmatrix} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

3.1.2 Nagylaki's Theory

Nagylaki [29] described and discussed the multinomial-sampling model for selection, mutation and random genetic drift at a single multiallelic locus in a panmictic (i.e., random-mating), monoecious (i.e., unisexual, hermaphroditic), diploid population with discrete, non-overlapping generations. He presented four different approximations for large populations. For frequent alleles (i.e., allelic frequencies are of order one as population size $N \to \infty$), the standard diffusion approximation holds if all the evolutionary forces are comparable [7]; the Gaussian approximation applies if the

deterministic forces, though still weak, dominate random drift [28]. For rare alleles (i.e., allelic frequencies tend to zero in probability as $N \to \infty$), if the allelic numbers are small (of order one as $N \to \infty$), the multiallelic Wright-Fisher model reaches a limit of a branching process with immigration; if they are moderate (tends to infinity with positive probability), the diffusion approximation of [8] holds.

We briefly introduce the basic model and the approximation for rare alleles using branching processes. The model follows [28], where a Markov chain was derived for the allelic frequencies from that of the genotypic frequencies. The life cycle starts with N monoecious adults. We focus on a single locus with r alleles and denote the frequency of the unordered genotype $A_iA_j, i \leq j$, just before reproduction by P_{ij}. The frequency of A_i at this stage is

$$p_i = P_{ii} + \frac{1}{2}\sum_{j:j>i} P_{ij} + \frac{1}{2}\sum_{j:j<i} P_{ji}.$$

After panmictic reproduction, the adults produce an infinite number of gametes, which then form zygotes according to the Hardy-Weinberg law with unordered genotypic frequencies $(2-\delta_{ij})p_ip_j$, where δ_{ij} represents the Kronecker delta.

Selection plays a role in the growth of zygotes. Denote the viability of A_iA_j by w_{ij}, then after selection the genotypic frequencies are

$$P^*_{ij} = \frac{(2-\delta_{ij})w_{ij}p_ip_j}{\sum_{k\leq l}(2-\delta_{kl})w_{kl}p_kp_l},$$

and the population size remains infinite.

Mutation comes in next. Denote the probability of A_i mutates to A_j by u_{ij}, and

suppose $u_{ii} = 0$, we obtain the genotypic frequencies after mutation

$$P_{ij}^{**} = \frac{1}{2}(2 - \delta_{ij}) \sum_{k \leq l} (R_{ki}R_{lj} + R_{kj}R_{li})P_{kl}^{*},$$

where

$$R_{ij} = \left(1 - \sum_{k} u_{ik}\right) \delta_{ij} + u_{ij}.$$

The life cycle is completed by random sampling, which reduces the population from infinity back to N adults with unordered genotypic frequencies P'_{ij}. By multinomial sampling assumption, we obtain the transition probabilities of the Markov chain of genotypic frequencies, in terms of pgf:

$$E\left[\prod_{i \leq j} s_{ij}^{NP'_{ij}} | P\right] = \left(\sum_{i \leq j} P_{ij}^{**} s_{ij}\right)^{N}. \tag{3.1}$$

Equation (3.1) reveals that the vector of gene frequencies $\mathbf{p}(n)$, where $n = 0, 1, 2, \cdots$ is time in generations, is Markovian. Suppose $A_1, \cdots, A_{r-1}$ are mutants and A_r is the wild-type allele. Let $y_i = 2Np_i, i = 1, \cdots, r-1$ be the observed mutant numbers. Nagylaki [29] has shown that: For rare alleles, if the allelic numbers are of order 1, as $N \to \infty$, and assuming independence among alleles, then

$$E\left[\prod_{i=1}^{r-1} s_i^{Y'_i} | \mathbf{Y} = \mathbf{y}\right] = \exp\left[\sum_{i=1}^{r-1} (\mu_{ri} + w_{ir}y_i)(s_i - 1)\right], \tag{3.2}$$

where $\mu_{ij} = 2Nu_{ij}$ represent the expected numbers of mutations in the population. Detailed proof is given in Appendix A in [29]. Therefore we see that (i) the allelic number $Y_i(n)$ forms a branching process, (ii) $Y_i(n)$ has a Poisson offspring distribution with mean w_{ir}, and (iii) $Y_i(n)$ is augmented by a Poisson influx of mutants at the rate μ_{ri}.

Therefore genetic drift for rare alleles can be approximated by a branching process with immigration if the allelic numbers are small. It should be noted that in Equation (3.2), w_{ir} represents the mean of the Poisson offspring distribution, which determines the type of branching process, critical or super/subcritical, whereas μ_{ri} describes the particle immigration in the branching process. Current analysis concerns the version of branching process without immigration (i.e., $\mu_{ri} = 0$). This corresponds to a single or multiple mutant alleles introduced at time $t = 0$. The version with immigration will be considered in the future.

To uncover the relation of the type of branching process to the type of selection in the Nagylaki model, we need to develop more general theoretical approach for different cases of monoecious/dioecious population with discrete overlapping and further continuous generations. Our ultimate goal is to use the theoretical result as a guidance to study the dynamics of extinction of disease-causing variants of genes in human populations.

Motivated by the approximation of genetic drift using branching processes, we concentrate on a study of the extinction problems in MBP.

3.2 Facts about Markov Branching Processes

In this section, we consider known facts about MBP. We briefly introduce the main results about the $\pi(s)$ function, the survival probability function $Q(t)$ and the conditional limit distribution of the process Z_t, as was done in [16]. Interestingly,

each result can be extended to the critical case with finite variance.

A MBP Z_t (unless specifically stated, we assume that it starts from a single particle) is completely determined by its offspring distribution and expected life span, given that the offspring are independent, identically distributed. Suppose the life length of each particle follows exponential distribution with rate parameter a, and the pgf of the offspring distribution is given by $f(s) = \sum_{k=0}^{\infty} p_k s^k$. Denote the pgf of the process Z_t by $F(s,t) = E_1\left[s^{Z_t}\right]$. $F(s,t)$ satisfies the forward Kolmogorov equation

$$\frac{\partial F(s,t)}{\partial t} = \phi(s)\frac{\partial F(s,t)}{\partial s},$$

where $\phi(s) = a(f(s) - s)$, and the backward Kolmogorov equation

$$\frac{\partial F(s,t)}{\partial t} = \phi(F(s,t)), \tag{3.3}$$

with boundary condition $F(s,0) = s$; see [3] for details. Let m and σ^2 denote the mean and variance of the offspring distribution and write $r = a(1-m)$. It is easy to check by Equation (3.3) that

$$E_1\left[Z_t\right] = e^{-rt}. \tag{3.4}$$

3.2.1 The $\pi(s)$ Function

For both the subcritical and critical cases, we define a specific function

$$\pi(s) = \int_0^s \frac{dv}{\phi(v)},$$

for $0 \leq s < 1$. It turns out that the $\pi(s)$ function plays an important role in the extinction of MBP.

Lemma 3.2.1. *In the subcritical case,*

$$\pi(s) \sim -r^{-1}\ln(1-s), \ \ as\ s \uparrow 1. \tag{3.5}$$

Furthermore,

$$\pi(s) = -r^{-1}[\ln(1-s) + \ln b] + o(1), \ \ as\ s \uparrow 1, \tag{3.6}$$

where $1 < b < \infty$ is a constant, if and only if the $x \log x$-condition holds:

$$\sum_{k=2}^{\infty}(k \ln k)p_k < \infty. \tag{3.7}$$

In the critical case, if $\sigma^2 < \infty$, then

$$\pi(s) \sim \frac{2}{a\sigma^2(1-s)} \sim -\frac{2}{a\sigma^2 \ln s}, \ \ as\ s \uparrow 1. \tag{3.8}$$

Proof. In the subcritical case,

$$\lim_{s\uparrow 1}\frac{\pi(s)}{\ln(1-s)} = \lim_{s\uparrow 1}\frac{\int_0^s \frac{dv}{a(f(v)-v)}}{\ln(1-s)} = -\frac{1}{a}\lim_{s\uparrow 1}\frac{1-s}{f(s)-s} = -\frac{1}{a(1-m)} = -r^{-1}.$$

It follows that $\lim_{s\uparrow 1}\frac{\pi(s)}{-r^{-1}\ln(1-s)} = 1$, hence $\pi(s) \sim -r^{-1}\ln(1-s)$, as $s \uparrow 1$.

Furthermore, $\sum_{k=2}^{\infty}(k \ln k)p_k < \infty$, which is equivalent to $\int_0^1 \left[\frac{1}{(1-m)(1-s)} - \frac{1}{f(s)-s}\right] ds < \infty$, see [3] page 26, and the integrand is a continuous function when $0 \leq s \leq 1$. Therefore, the $x \log x$-condition is equivalent to $\lim_{s\uparrow 1}[-r^{-1}\ln(1-s) - \pi(s)] = \text{constant}$.

Denote this limit by $r^{-1}\ln b, 1 < b < \infty$. Then, $\pi(s) = -r^{-1}[\ln(1-s) + \ln b] + o(1)$, as $s \uparrow 1$.

In the critical case, we need to apply the l'Hopital's rule twice.

$$\begin{aligned}\lim_{s\uparrow 1}(1-s)\pi(s) &= \lim_{s\uparrow 1}\frac{\int_0^s \frac{dv}{a(f(v)-v)}}{1/(1-s)} = \lim_{s\uparrow 1}\frac{1/[a(f(v)-v)]}{1/(1-s)^2} \\ &= \lim_{s\uparrow 1}\frac{-2(1-s)}{a(f'(s)-1)} = \lim_{s\uparrow 1}\frac{2}{af''(s)} = \frac{2}{a\sigma^2}.\end{aligned}$$

It follows that $\lim_{s\uparrow 1}\frac{a\sigma^2(1-s)\pi(s)}{2} = 1$ hence $\pi(s) \sim \frac{2}{a\sigma^2(1-s)}$, as $s \uparrow 1$. The second aymptotic equivalence is straightforward. □

Lemma 3.2.1 tells us of the different asymptotic behavior of $\pi(s)$ function as $s \uparrow 1$ in the subcritical and critical cases. To visualize this conclusion, we plot the $\pi(s)$ function and its asymptotic equivalent for both cases in Figure 3.1, assuming that the offspring distribution is binary fission or Poisson.

3.2.2 The Survival Probability Function

Proposition 3.2.2. *Suppose Z_t is a Markov branching process and define the survival probability $Q(t) = P_1(Z_t > 0)$. In the subcritical case, if (3.7) holds, we have*

$$Q(t) \sim b^{-1}e^{-rt}, \text{ as } t \to \infty. \tag{3.9}$$

In the critical case, if $\sigma^2 < \infty$, we have

$$Q(t) \sim \frac{2}{a\sigma^2 t}, \text{ as } t \to \infty. \tag{3.10}$$

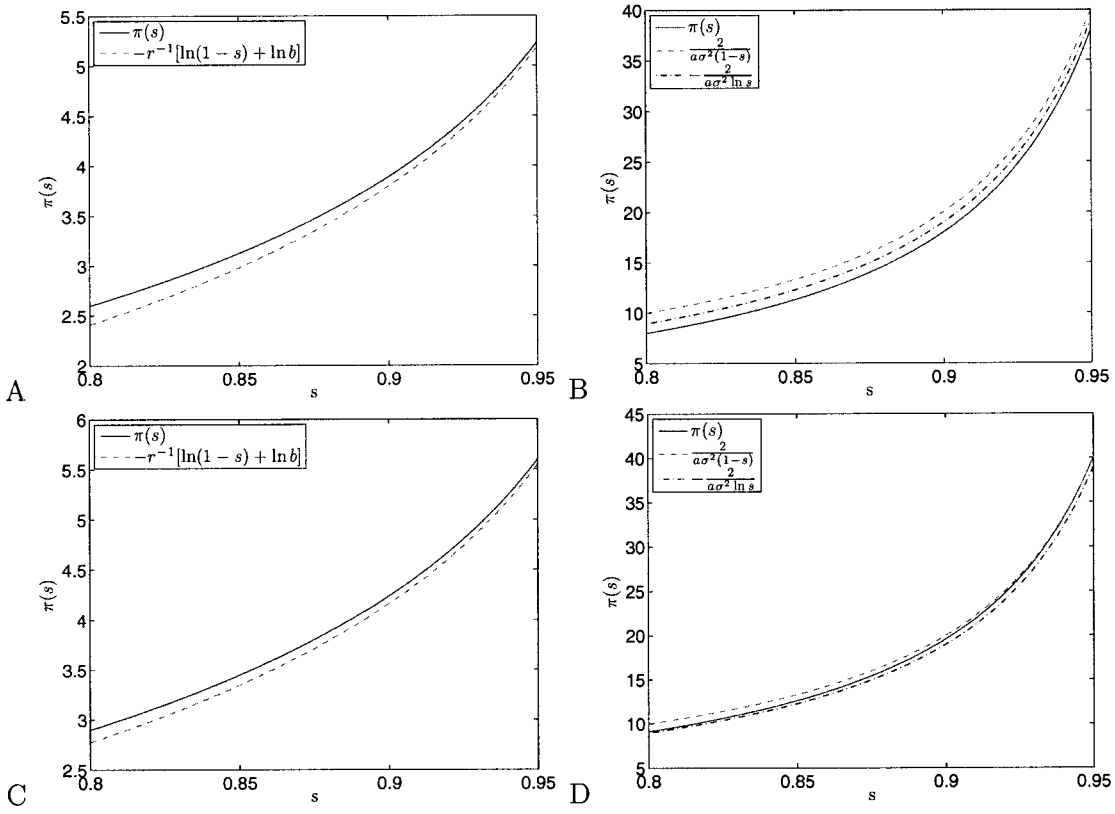


Figure 3.1: $\pi(s)$ function and its asymptotically equivalent functions. (A) binary fission, subcritical, $m = 0.5$; (B) binary fission, critical; (C) Poisson, subcritical, $m = 0.5$; (D) Poisson, critical.

Proof. By Equation(3.3), $\frac{\partial}{\partial t}\left[\int_0^{F(s,t)} \frac{dv}{\phi(v)}\right] = 1$. By solving this differential equation and using the boundary condition $F(s,0) = s$, we obtain

$$\pi(F(s,t)) = \pi(s) + t. \tag{3.11}$$

Therefore

$$\pi(F(0,t)) = t. \tag{3.12}$$

In the subcritical case, as $t \to \infty$, Equation (3.12) and (3.6) lead to $-r^{-1}[\ln Q(t) + \ln b] + o(1) = t$, which means $\lim_{t\to\infty} r^{-1}[\ln Q(t) + \ln b] + t = 0$, hence $Q(t) \sim b^{-1}e^{-rt}$ as $t \to \infty$.

In the critical case, if $\sigma^2 < \infty$, as $t \to \infty$, (3.8) and (3.12) lead to $\lim_{s\uparrow 1} \frac{a\sigma^2 Q(t)t}{2} = 1$, hence $Q(t) \sim \frac{2}{a\sigma^2 t}$ as $t \to \infty$. □

This result is well known and has been proved for Markov, age-dependent and general Crump-Mode-Jagers branching processes (see [36] Chapter II.2, Theorem 1 and Theorem 3, [3] Chapter IV.7, Theorem 1 and Chapter IV.6, Theorem 1, [15] Theorem 6.7.2 and Theorem 6.6.11).

We use simulations to illustrate the result. The programming is done in Matlab, using the same self-recurrent technique as in Section 2.2.3. We first simulate a MBP Z_t starting from a single particle, and record its status, extinct or not, at different time points. After repeating 20 times, the frequency of non-extinction status at different time points in all the simulations is calculated as an estimate of the survival probability function of this process, i.e., $\hat{Q}(t) = \frac{\text{\# of non-extinction simulations}}{\text{\# of simulations}}$. To obtain a

more accurate estimate, we assume the initial population size x larger than 1. Under this setting, $\hat{Q}(t) = 1 - \left(\frac{\text{\# of non-extinction simulations}}{\text{\# of simulations}}\right)^{1/x}$. Simulation results for both cases of subcritical and critical with binary fission and Poisson offspring distribution are shown in Figure 3.2. $\hat{Q}(t)$ is shown in solid lines, with its 2.5% and 97.5% percentiles represented by the dotted lines. The dashed lines represent the asymptotically equivalent functions. In this simulation study, we set $x = 20$ for the subcritical case, and $x = 10$ for the critical case.

Simulation of the critical case may encounter computational obstacles because of the late occurrence of extinction. This problem is solved by setting an upper bound for the generation count of the simulated process. In a simulated process, if extinction has not been reached before this generation bound, say 100, then we terminate this simulation and treat the corresponding time to extinction as "lost to follow-up". This will not affect the result of the simulation of survival probability, but it will play a role when simulating T, Z_{uT} and Z_{T-u}, as we will see in the simulations of the critical case in Section 3.4.

Another comment concerning the simulations in the subcritical case is the calculation of b. Haccou *et al.* [12] give an example of calculating b in MBP with binary fission offspring. Nevertheless we provide derivations here using the previously defined $\pi(s)$ function, for both binary fission and Poisson cases.

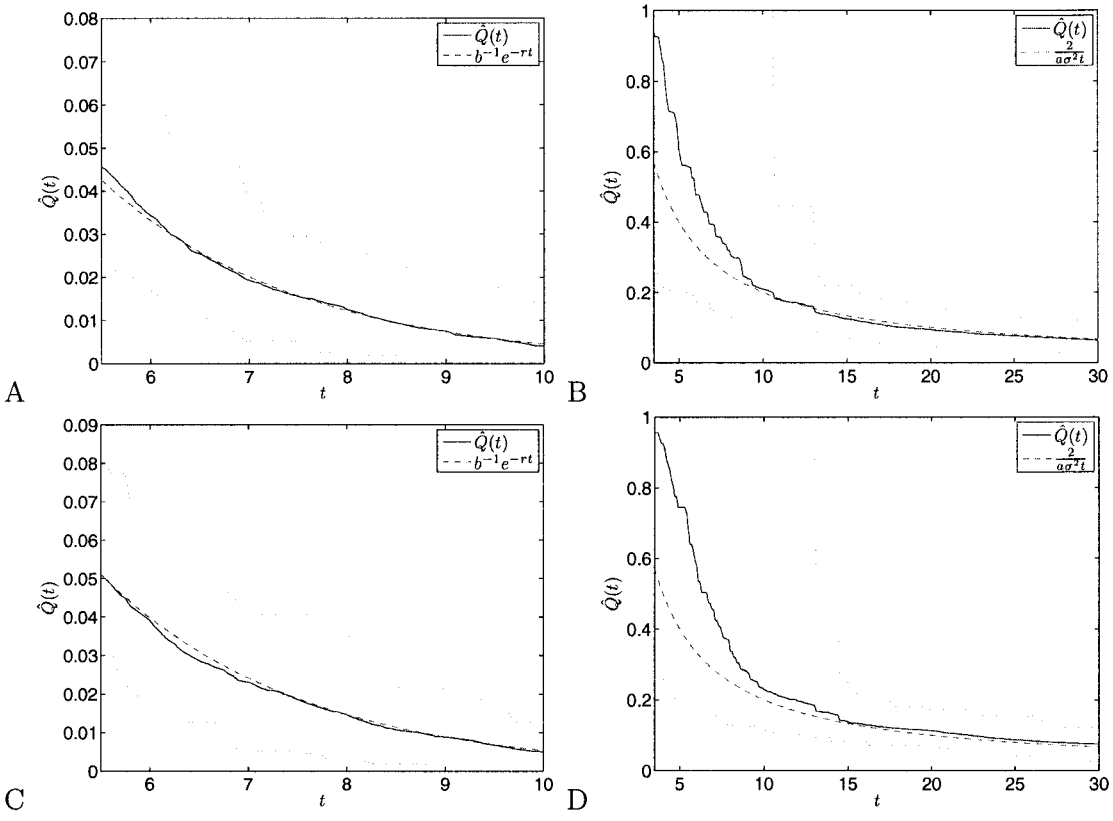


Figure 3.2: Survival probability $\hat{Q}(t)$ of simulated MBP, based on 20 simulations and its asymptotically equivalent function. $\hat{Q}(t)$ is shown in solid lines, with its 2.5% and 97.5% percentiles represented by the dotted lines. The dashed lines represent the asymptotically equivalent functions. (A) binary fission, subcritical, $m = 0.5, x = 20$; (B) binary fission, critical, $x = 10$; (C) Poisson, subcritical, $m = 0.5, x = 20$; (D) Poisson, critical, $x = 10$.

By (3.9) and (3.12),

$$\begin{aligned} b^{-1} &= \lim_{t\to\infty} e^{rt}Q(t) \\ &= \lim_{t\to\infty} e^{rt}[1-F(0,t)] \\ &= \lim_{t\to\infty} e^{rt}[1-\pi^{-1}(t)]. \end{aligned}$$

Therefore, b can be obtained in terms of function $\pi(s)$, if its inverse can be found.

For the binary fission case, $f(s) = p_0 + (1-p_0)s^2$, $m = 2(1-p_0)$, $r = a(1-m) = a(2p_0-1)$, so

$$\begin{aligned} \pi(s) &= \int_0^s \frac{dv}{a[f(v)-v]} \\ &= \frac{1}{a}\int_0^s \frac{dv}{p_0 - v + (1-p_0)v^2} \\ &= \frac{1}{a}\int_0^s \frac{dv}{[p_0-(1-p_0)v][1-v]} \\ &= \frac{1}{a(1-2p_0)}\int_0^s \left[\frac{1-p_0}{p_0-(1-p_0)v} - \frac{1}{1-v}\right] dv \\ &= -r^{-1}\ln\frac{p_0(1-s)}{p_0-(1-p_0)s}. \end{aligned}$$

Therefore the inverse of $\pi(s)$ can be obtained,

$$\pi^{-1}(t) = \frac{p_0(e^{-rt}-1)}{(1-p_0)e^{-rt}-p_0}.$$

Hence $b = \frac{p_0}{2p_0-1}$.

For the Poisson case, $f(s) = e^{\lambda(s-1)}$, $m = \lambda$, $r = a(1-m) = a(1-\lambda)$, so

$$\begin{aligned} \pi(s) &= \int_0^s \frac{dv}{a[f(v)-v]} \\ &= \frac{1}{a}\int_0^s \frac{dv}{e^{\lambda(v-1)}-v}. \end{aligned}$$

The inverse of $\pi(s)$ cannot be easily solved. However, when $\lambda \approx 0$, the denominator of the integrand can be approximated by $(1-v)\left[1-\lambda+\frac{\lambda^2}{2}(1-v)\right]$, therefore

$$\pi^{-1}(t) \approx \frac{\left(1-\lambda+\frac{\lambda^2}{2}\right)(e^{-rt}-1)}{\frac{\lambda^2}{2}e^{-rt}-\left(1-\lambda+\frac{\lambda^2}{2}\right)}.$$

Hence $b \approx \frac{1-\lambda+\frac{\lambda^2}{2}}{1-\lambda}$, for $\lambda \approx 0$.

3.2.3 Conditional Limit Laws

Proposition 3.2.3. *Suppose Z_t is a Markov branching process.*

In the subcritical case, if (3.7) holds, then

$$P_1(Z_t = k | Z_t > 0) \to b_k, k \geq 1, \text{ as } t \to \infty, \tag{3.13}$$

where the constant sequence b_k satisfies $\sum_{k=1}^{\infty} b_k = 1$.

Consequently,

$$E_1[Z_t | Z_t > 0] \to b \text{ as } t \to \infty, \tag{3.14}$$

where $b = \sum_{k=1}^{\infty} k b_k$.

In the critical case, if $\sigma^2 < \infty$, then

$$P_1\left(\frac{Z_t}{t} > z \middle| Z_t > 0\right) \to e^{-\frac{2}{a\sigma^2}z} \text{ as } t \to \infty. \tag{3.15}$$

Consequently,

$$E_1\left[\frac{Z_t}{t} \middle| Z_t > 0\right] \to \frac{a\sigma^2}{2} \text{ as } t \to \infty. \tag{3.16}$$

Proof. By (3.11) and (3.12), $\pi(F(0,t+\pi(s))) = t+\pi(s) = \pi(F(s,t))$. Since $\pi(s)$ is a monotone function, we conclude that

$$F(s,t) = F(0,t+\pi(s)) = 1 - Q(t+\pi(s)). \tag{3.17}$$

Consider the pgf of the process Z_t, $E_1[s^{Z_t}]$,

$$\begin{aligned} & E_1[s^{Z_t}] = E_1[s^{Z_t}|Z_t > 0]P(Z_t > 0) + P(Z_t = 0) \\ \Rightarrow \quad & F(s,t) = E_1[s^{Z_t}|Z_t > 0)Q(t) + 1 - Q(t) \\ \Rightarrow \quad & E_1[s^{Z_t}|Z_t > 0] = 1 - \frac{1-F(s,t)}{Q(t)} = 1 - \frac{Q(t+\pi(s))}{Q(t)}. \end{aligned} \tag{3.18}$$

In the subcritical case, if (3.7) holds, then by (3.9), $Q(t) \sim b^{-1}e^{-rt}$ as $t \to \infty$. If we apply this relation to (3.18), it follows that

$$\begin{aligned} \lim_{t\to\infty} E_1[s^{Z_t}|Z_t > 0] &= 1 - \lim_{t\to\infty} \frac{Q(t+\pi(s))}{Q(t)} \\ &= 1 - \lim_{t\to\infty} \left\{ \frac{Q(t+\pi(s))}{b^{-1}e^{-r[t+\pi(s)]}} \cdot \frac{b^{-1}e^{-rt}}{Q(t)} \cdot \frac{e^{-r[t+\pi(s)]}}{e^{-rt}} \right\} \\ &= 1 - \lim_{t\to\infty} \frac{Q(t+\pi(s))}{b^{-1}e^{-r[t+\pi(s)]}} \cdot \lim_{t\to\infty} \frac{b^{-1}e^{-rt}}{Q(t)} \cdot \lim_{t\to\infty} \frac{e^{-r[t+\pi(s)]}}{e^{-rt}} \\ &= 1 - e^{-r\pi(s)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{t\to\infty} P_1(Z_t = k|Z_t > 0) &= \lim_{t\to\infty} \frac{d^k}{k!ds^k} E_1[s^{Z_t}|Z_t > 0]_{s=0} \\ &= \frac{d^k}{k!ds^k} \left\{ \lim_{t\to\infty} E_1[s^{Z_t}|Z_t > 0] \right\}_{s=0} \\ &= \frac{d^k}{k!ds^k} \left\{ 1 - e^{-r\pi(s)} \right\}_{s=0} \\ &=: b_k, k \geq 1. \end{aligned}$$

Relation (3.14) can be easily achieved by (3.4), (3.9) and $E_1[Z_t|Z_t>0]=\frac{E_1[Z_t]}{P[Z_t>0]}$.

In the critical case, if $\sigma^2<\infty$, by (3.10), $Q(t)\sim\frac{2}{a\sigma^2 t}$ as $t\to\infty$. Consider function $\frac{Q(t+\pi(s))}{Q(t)}$, it is easy to see that $\frac{Q(t+\pi(s))}{Q(t)}\sim\frac{t}{t+\pi(s)}$ as $t\to\infty$. By (3.18), that is equivalent to $1-E_1[s^{Z_t}|Z_t>0]\sim\frac{t}{t+\pi(s)}$ as $t\to\infty$. To construct the moment generating function (mgf) of $\frac{Z_t}{t}$, let $s=e^{-\frac{2\nu}{a\sigma^2 t}}$ where $0<\nu<\infty$ so that $0<s<1$, we have $1-E_1\left[e^{-\frac{2}{a\sigma^2}\frac{Z_t}{t}\nu}|Z_t>0\right]\sim\frac{t}{t+\pi\left(e^{-\frac{2\nu}{a\sigma^2 t}}\right)}$ as $t\to\infty$. Therefore

$$\begin{aligned}\lim_{t\to\infty}E_1\left[e^{-\frac{2}{a\sigma^2}\frac{Z_t}{t}\nu}|Z_t>0\right] &= 1-\lim_{t\to\infty}\frac{t}{t+\pi\left(e^{-\frac{2\nu}{a\sigma^2 t}}\right)}\\ &= 1-\frac{1}{1+\lim_{t\to\infty}\frac{\pi\left(e^{-\frac{2\nu}{a\sigma^2 t}}\right)}{t}}.\end{aligned}$$

As $t\to\infty, e^{-\frac{2\nu}{a\sigma^2 t}}\uparrow 1$, and by (3.8), $\pi(s)\sim-\frac{2}{a\sigma^2\ln s}$ as $s\uparrow 1$, therefore

$$\begin{aligned}&\lim_{t\to\infty}-\frac{\pi(e^{-\frac{2\nu}{a\sigma^2 t}})a\sigma^2[\ln e^{-\frac{2\nu}{a\sigma^2 t}}]}{2}=1\\ \Rightarrow\ &\lim_{t\to\infty}\left\{\frac{\pi(e^{-\frac{2\nu}{a\sigma^2 t}})}{t}\cdot\frac{a\sigma^2[\frac{2\nu}{a\sigma^2 t}]t}{2}\right\}=1\\ \Rightarrow\ &\lim_{t\to\infty}\frac{\pi(e^{-\frac{2\nu}{a\sigma^2 t}})}{t}=\frac{1}{\nu}.\end{aligned}$$

Accordingly, $E_1\left[e^{-\frac{2}{a\sigma^2}\frac{Z_t}{t}\nu}|Z_t>0\right]\to 1-\frac{1}{1+1/\nu}=\frac{1}{1+\nu}$ as $t\to\infty$. By uniqueness and continuity of mgf, $\frac{2}{a\sigma^2}\frac{Z_t}{t}$ converges in distribution to a standard exponential random variable, hence $P_1\left(\frac{Z_t}{t}>z|Z_t>0\right)\to e^{-\frac{2}{a\sigma^2}z}$ as $t\to\infty$. Relation(3.16) is a direct consequence of (3.15), but it can also be easily obtained using (3.10), $E_1[Z_t]=1$ and $E_1[Z_t|Z_t>0]=\frac{E_1[Z_t]}{P[Z_t>0]}$. □

Expressions (3.13) and (3.15) are known as the Yaglom's Theorem and exponential limit law, respectively. Although these limit theorems have been proved for Markov,

age-dependent and general Crump-Mode-Jagers branching processes (see [36] Chapter II.4 Theorem 1 and Chapter II.5 Theorem 1, [3] Chapter IV.10 Theorem 1 and Chapter IV.9 Theorem (Goldstein), [15] Theorem 6.7.3 and Theorem 6.6.11) branching processes (see [3] page 159 and 163), it seems instructive to see the proof using Lemma 3.2.1 and Proposition 3.2.2.

3.3 Extinction of Subcritical Markov Branching Processes

Based mainly on Jagers' work [16, 17], this section summarizes the three topics on extinction: *time to extinction*, *path to extinction* and *path verging on extinction*, for subcritical MBP.

3.3.1 Time to Extinction

The time to extinction problem has been studied in [17], [31] and [33]. These papers considered the different cases of Galton-Watson, Markov and general branching processes. In this section, we summarize the result for both subcritical and critical cases in MBP and extend it to age-dependent or general branching processes.

We consider a continuous time branching process Z_t which is subcritical or critical and therefore eventually dies out with probability one. Let x denote the initial population size of Z_t, and T denote the time to extinction of this process. Intuitively,

T should perform as an increasing function of x. Classical extreme value theory determines the trend and randomness of this relation:

Proposition 3.3.1. *Suppose Z_t is a subcritical Markov branching process, with initial population size x and time to extinction T. If (3.7) holds, then*

$$rT - \ln x - \ln c \xrightarrow{d} \eta \text{ as } x \to \infty, \tag{3.19}$$

where $0 < c < 1$, η is distributed as standard Gumbel and $\xrightarrow{d}$ denotes convergence in distribution.

The basic idea of proving using extreme value theory is to consider random variable T as the maximum of independent, identically distributed random variables $T_1, T_2, ..., T_x$, which are the times to extinction given that the process starts from single particle $i, i = 1, ..., x$. By classical extreme value theory, distribution of random variable T is directly determined by the tail behavior of the distribution function (d.f.) of $T_i, i = 1, ..., x$, denoted as $G(t)$.

Proof. If (3.7) holds, by Proposition 3.2.2, $Q(t) \sim b^{-1}e^{-rt}$, as $t \to \infty$. Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1 - G(t + r^{-1}y)}{1 - G(t)} &= \lim_{t\to\infty} \frac{Q(t + r^{-1}y)}{Q(t)} \\
&= \lim_{t\to\infty} \frac{Q(t + r^{-1}y)}{b^{-1}e^{-r(t+r^{-1}y)}} \cdot \frac{b^{-1}e^{-rt}}{Q(t)} \cdot e^{-y} \\
&= e^{-y}, \forall y \in \mathbb{R}.
\end{aligned}$$

Theorem 1.6.2 of [24] tells us that $G(t)$ belongs to the Type-I (Gumbel) domain of attraction. By Corollary 1.6.3 of [24], the normalizing constants a_x, b_x can be

determined by $a_x = r$ and $b_x = G^{-1}(1 - 1/x) = -r^{-1}(\ln b - \ln x)$, which gives (3.19). □

We illustrate the distribution of the time to extinction by simulations. For a prespecified initial population size x, firstly we generate a MBP and obtain its time to extinction T. After repeating this procedure 500 times, we plot the empirical cumulative distribution function (cdf) for the simulated T's and compare it to the theoretical cdf based on Proposition 3.3.1. Figure 3.3 shows the simulation results.

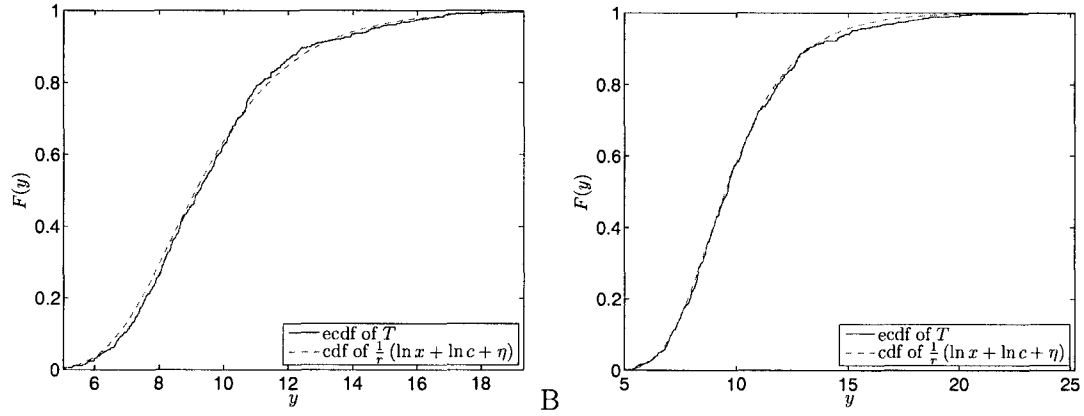


Figure 3.3: Empirical cdf of the time to extinction T in subcritical ($m = 0.5$) MBP with initial population size $x = 100$, based on 500 simulations. (A) binary fission; (B) Poisson.

Figure 3.4 shows the dynamic relation between x and T in our simulations. Instead of checking the distribution of T as in Figure 3.3, we make a loose verification on its first moment. We see that, as x changes, the trajectory of the simulated mean time to extinction follows the theoretical result $E[T] = r^{-1}(\ln x + \ln c + \gamma)$, where $\gamma \approx 0.577$ represents the Euler constant, which is the expected value of standard Gumbel distribution.

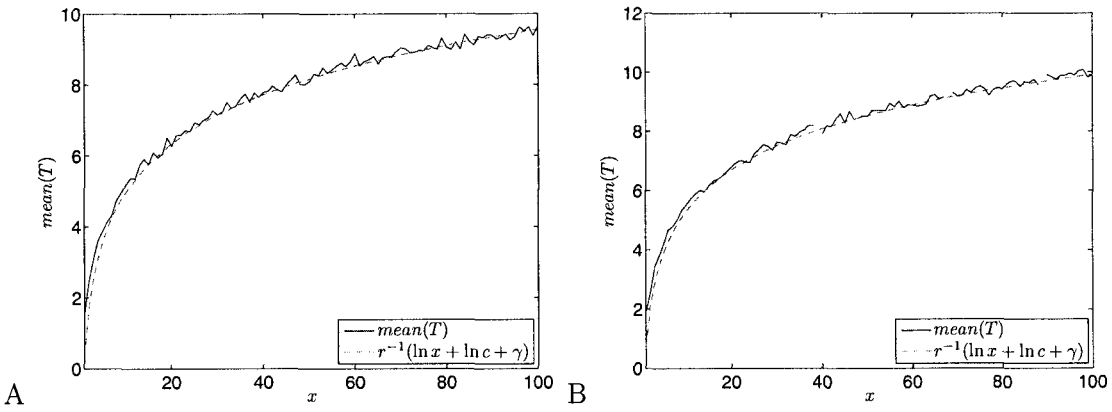


Figure 3.4: Scatter plot of x versus sample mean of T in subcritical ($m = 0.5$) MBP, based on 500 simulations. (A) binary fission; (B) Poisson.

Given the known facts on survival probability under some appropriate conditions, the above proof also applies to the age-dependent branching process and furthermore, the general Crump-Mode-Jagers branching process. Recently, Jagers *et al.* [17] proposed a different proof for the subcritical case of the general branching process, from the viewpoint of $E[T]$. To demonstrate this rule in a direct way, here we provide another proof for the age-dependent process using fundamental probability principles. Connection and difference between the subcritical and critical cases can be easily seen from the proof. This proof can also be extended to the general branching process, given appropriate conditions for achieving relations (3.9) and (3.10) (see [15] Theorem 6.7.2 and Theorem 6.6.11).

We start from a well known fact.

Lemma 3.3.2. *Suppose $f(x)$ is a function defined on positive integer numbers. If $f(x) = o(1)$, then $\lim_{x\to\infty}\left[1 + \frac{f(x)}{x}\right]^x = 1$.*

Proof.

$$\begin{aligned}\lim_{x\to\infty}\left[1+\frac{f(x)}{x}\right]^x &= \lim_{x\to\infty}\left\{\left[1+\frac{f(x)}{x}\right]^{\frac{x}{f(x)}}\right\}^{f(x)} \\ &= \exp\left(\lim_{x\to\infty} f(x)\cdot\ln\left\{\left[1+\frac{f(x)}{x}\right]^{\frac{x}{f(x)}}\right\}\right) \\ &= 1.\end{aligned}$$

□

Theorem 3.3.3. *For an age-dependent branching process, Proposition 3.3.1 also holds.*

Proof. Let $T^* = rT - \ln x - \ln c$. We consider the cdf of T^*, $P_x(T^* \le t) = P_x[T \le r^{-1}(t + \ln x + \ln c)]$. Let $\tau = r^{-1}(t + \ln x + \ln c)$. Then $P_x(T^* \le t) = P_x(T \le \tau) = [P_1(T \le \tau)]^x = [1 - P_1(Z_\tau > 0)]^x$. In age-dependent branching processes, given suitable conditions, $P_1(Z_\tau > 0) \sim ce^{-r\tau}$ as $\tau \to \infty$ (see [3] page 163). So, $P_1(Z_\tau > 0) - ce^{-r\tau} = o(ce^{-r\tau})$. Apply $\tau = r^{-1}(t + \ln x + \ln c)$, and notice that for fixed t, $x \to \infty$ implies $\tau \to \infty$. We obtain $P_1(Z_\tau > 0) - \frac{e^{-t}}{x} = o\left(\frac{e^{-t}}{x}\right)$. Note, here we consider $P_1(Z_\tau > 0)$ as a function of x since τ is a function of x, given t. So, $[1 - P_1(Z_\tau > 0)] - \left[1 - \frac{e^{-t}}{x}\right] = o\left(\frac{e^{-t}}{x}\right)$. For simplicity, write $1 - P_1(Z_\tau > 0)$ as $f(x)$.

Then

$$
\begin{aligned}
& f(x) - \left[1 - \frac{e^{-t}}{x}\right] = o\left(\frac{e^{-t}}{x}\right) \\
\Rightarrow \quad & \frac{f(x) - \left[1 - \frac{e^{-t}}{x}\right]}{\left[1 - \frac{e^{-t}}{x}\right]} = o\left(\frac{e^{-t}/x}{\left[1 - \frac{e^{-t}}{x}\right]}\right) \\
\Rightarrow \quad & \frac{f(x) - \left[1 - \frac{e^{-t}}{x}\right]}{\left[1 - \frac{e^{-t}}{x}\right]} = o\left(\frac{e^{-t}}{x}\right) \text{ since } \frac{e^{-t}/x}{\left[1 - \frac{e^{-t}}{x}\right]} = O\left(\frac{e^{-t}}{x}\right) \\
\Rightarrow \quad & \frac{f(x)}{1 - \frac{e^{-t}}{x}} - 1 = o\left(\frac{e^{-t}}{x}\right) = o(1/x) \text{ for fixed } t.
\end{aligned}
$$

By Lemma 3.3.2, we have $\lim_{x\to\infty}\left[\frac{f(x)}{1-\frac{e^{-t}}{x}}\right]^x = 1$. So, $\lim_{x\to\infty}[1 - P_1(Z_\tau > 0)]^x = e^{-e^{-t}}$. This implies $\lim_{x\to\infty} P_x(T^* \le t) = e^{-e^{-t}}$, hence $rT - \ln x - \ln c \xrightarrow{d} \eta$ as $x \to \infty$, where η is distributed as standard Gumbel. □

3.3.2 Path to Extinction

Process $Z_{uT}, 0 \le u < 1$ is called the "u-way" path to extinction for branching process Z_t. As an example, Figure 3.5 shows population size, Z_t and Z_{uT}, of a subcritical MBP with binary fission offspring distribution.

For subcritical MBP, Jagers *et al.* [16] derived the distribution of Z_{uT} and the limit distribution of $x^{u-1}Z_{uT}$ when the initial population size approaches infinity. The latter result was also proposed in their earlier paper [17], but for general branching processes, under additional assumptions. This section summarizes their work.

For subcritical MBP Z_t, let the cdf of the time T to extinction be

$$G(t) = P_1(T \le t) = P_1(Z_t = 0) = F(0, t).$$

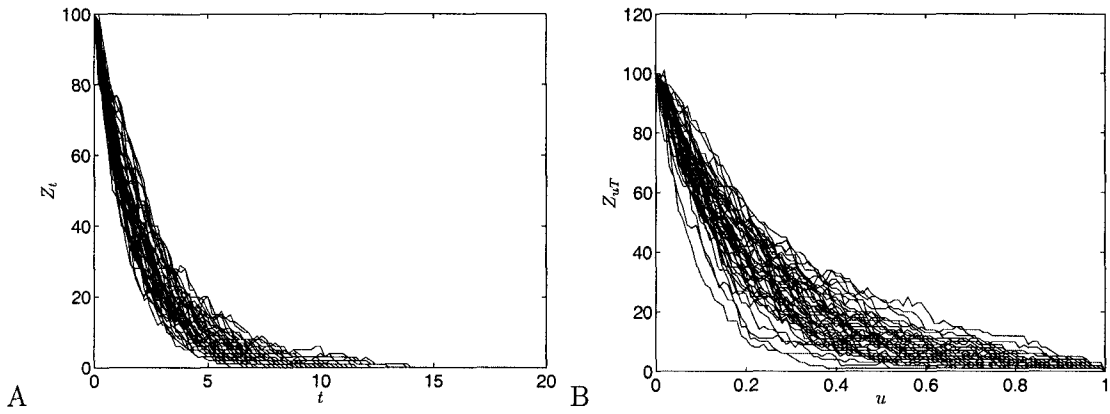


Figure 3.5: Path of subcritical ($m = 0.5$) binary fission MBP with initial population size $x = 100$, based on 50 simulations. (A) Z_t; (B) Z_{uT}.

Let $g(t) = G'(t)$. The backward Kolmogorov equation (3.3) yields $g(t) = \phi(F(0,t))$.

For the pgf $F(s,t)$, define $F'(s,t) = \frac{\partial}{\partial s}F(s,t)$.

Theorem 3.3.4. *Let* $x \in \mathcal{Z}^+, 0 \leq u < 1$.

$$E_x[s^{Z_{uT}}] = sx \int_0^\infty F^{x-1}(sG((1-u)t), ut)F'(sG((1-u)t), ut)g((1-u)t)dt. \quad (3.20)$$

Proof. By the law of total probability,

$$\begin{aligned}
E_x[s^{Z_{uT}}] &= \int_0^\infty E_x[s^{Z_{uT}}, T \in dt] \\
&= \int_0^\infty E_x[s^{Z_{ut}}, T \in dt] \\
&= \int_0^\infty \sum_y s^y P_x(Z_{ut} = y, T \in dt) \\
&= \int_0^\infty \sum_y s^y P_x(Z_{ut} = y) P_x(T \in dt | Z_{ut} = y) \\
&= \int_0^\infty \sum_y s^y P_x(Z_{ut} = y) P_y(T' + ut \in dt), \text{ define } T' = (1-u)T \\
&= \int_0^\infty \sum_y s^y P_x(Z_{ut} = y) \frac{d}{d(1-u)t} G^y((1-u)t)dt \\
&= \int_0^\infty s \frac{\partial}{\partial s_t} F^x(s_t, ut) g((1-u)t)dt, \text{ define } s_t = sG((1-u)t) \\
&= sx \int_0^\infty F^{x-1}(sG((1-u)t), ut) F'(sG((1-u)t), ut) g((1-u)t)dt.
\end{aligned}$$

□

The following theorem is then obtained by using Lemma 3.2.1, Proposition 3.2.2, Equation (3.17) and Theorem 3.3.4.

Theorem 3.3.5. *Suppose Z_t is a subcritical Markov branching process, with initial population size x and time to extinction T. If (3.7) holds, then*

$$\left\{x^{u-1} Z_{uT}, 0 \le u < 1\right\} \xrightarrow{fdd} \left\{b^u e^{-u\eta}, 0 \le u < 1\right\} \text{ as } x \to \infty,$$

where η is distributed as standard Gumbel, and $\xrightarrow{fdd}$ denotes convergence in finite dimensional distribution.

Detailed proof of this theorem can be found in [16] and [17], hence is skipped here.

To illustrate one-dimensional convergence in Theorem 3.3.5, for prespecified u, we plot the empirical cdf of the simulated process $x^{u-1}Z_{uT}$ in Figure 3.6, and compare it to the theoretical cdf according to Theorem 3.3.5. To show finite-dimensional convergence, Figure 3.7 plots the simulated trajectories of $x^{u-1}Z_{uT}$ as a comparison to the simulated trajectories of $b^u e^{-u\eta}$. In order to observe the dynamic relation between $x^{u-1}Z_{uT}$ and x, we further plot the mean and variance process in Figure 3.8 as x changes.

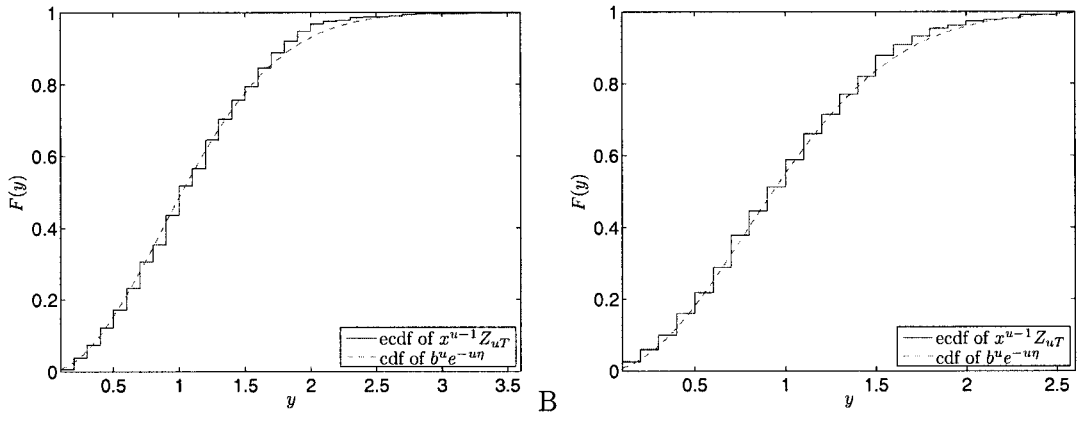


Figure 3.6: Empirical cdf of the path to extinction in subcritical ($m = 0.5$) MBP with initial population size $x = 100$, based on 500 simulations. (A) binary fission, $u = 0.5$; (B) Poisson, $u = 0.5$.

3.3.3 Path Verging on Extinction

Jagers *et al.* [16] derived the limit distribution for process Z_{T-u} in subcritical MBP when the initial population size approaches infinity, which follows a *time reversed* branching process. Furthermore, they proposed a theorem about the asymptotic

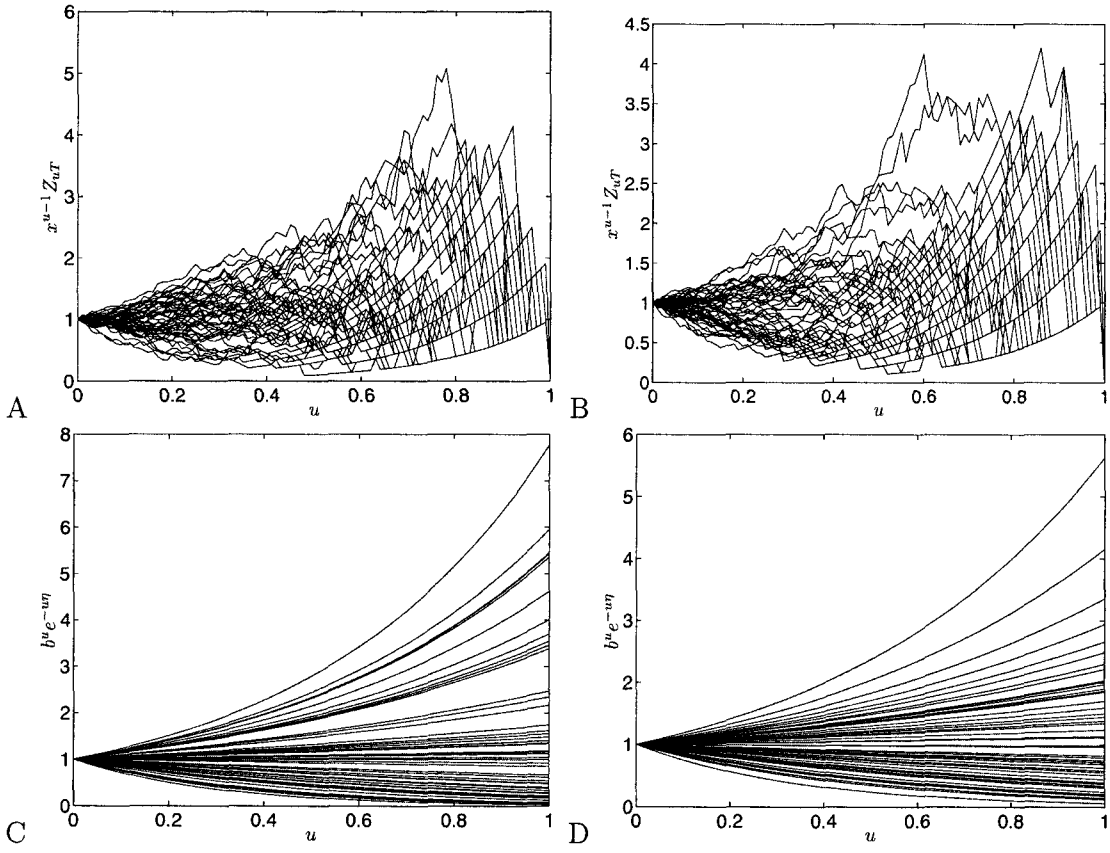


Figure 3.7: Comparison between $x^{u-1}Z_{uT}$ and $b^u e^{-u\eta}$ in subcritical ($m = 0.5$) MBP with initial population size $x = 100$, based on 50 simulations. (A) $x^{u-1}Z_{uT}$, binary fission; (B) $x^{u-1}Z_{uT}$, Poisson; (C) $b^u e^{-u\eta}$, binary fission; (D) $b^u e^{-u\eta}$, Poisson.

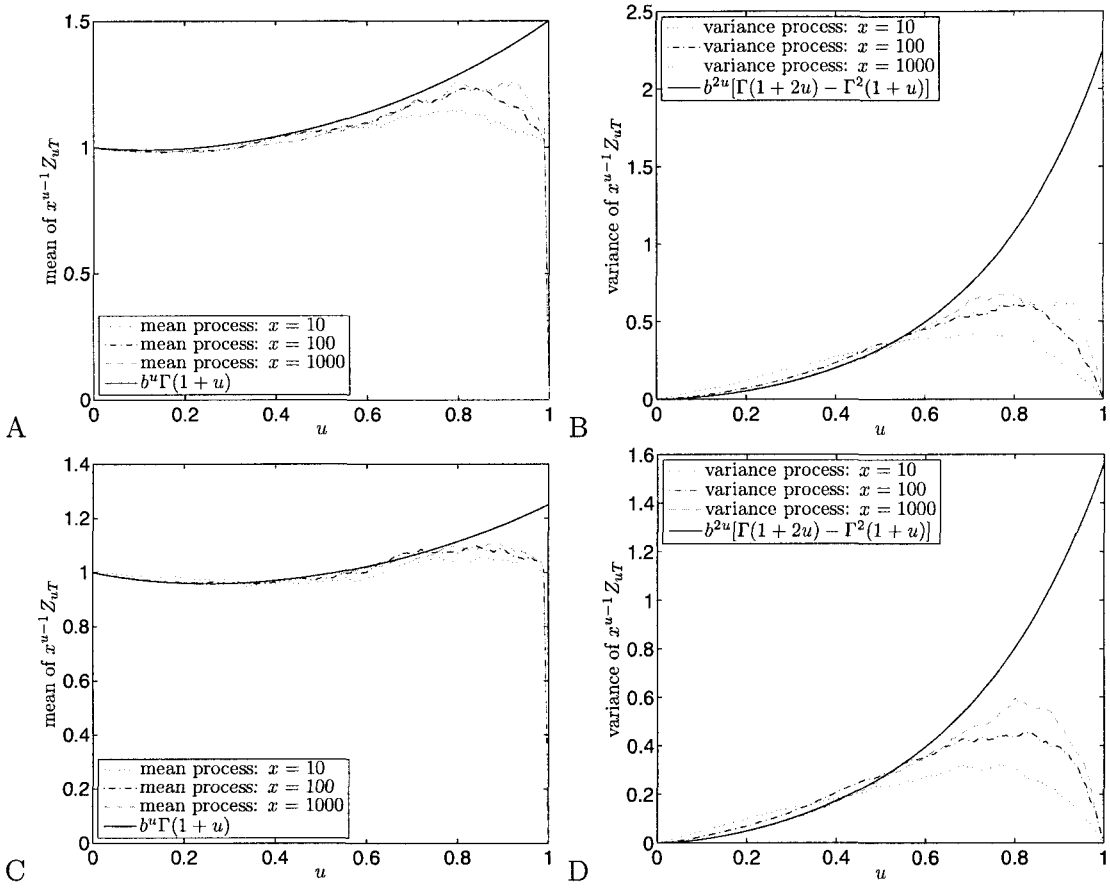


Figure 3.8: Mean and variance of $x^{u-1}Z_{uT}$ for different x in subcritical ($m = 0.5$) MBP, based on 500 simulations. (A) mean process, binary fission; (B) variance process, binary fission; (C) mean process, Poisson; (D) variance process, Poisson.

behavior of this limit under appropriate normalization (multiply by e^{-ru}) as $u \to \infty$.

Theorem 3.3.6. *Suppose Z_t is a subcritical Markov branching process, with initial population size x and time to extinction T. If (3.7) holds, then Z_{T-u} converges in distribution, $u \geq 0$, as $x \to \infty$. Furthermore, suppose Y_u denote the limit of Z_{T-u} as $x \to \infty$, then $e^{-ru}Y_u \xrightarrow{d} be^{-\eta}$ as $u \to \infty$, where η is distributed as standard Gumbel. In other words, denote the distribution function of the process $e^{-ru}Z_{T-u}$ by $F_{x,u}$, then* $\lim_{u\to\infty}[\lim_{x\to\infty} F_{x,u}(z)] = 1 - e^{-z/b}$.

From an intuitive perspective, the second conclusion, $e^{-ru}Y_u \xrightarrow{d} be^{-\eta}$ as $u \to \infty$, is another representation of Proposition 3.3.1. The reason is that process Z_{T-u} makes sense only when $0 \leq u < T$, therefore as $u \to T$, intuitively, $Y_u \to x$. Noting that condition $u \to \infty$ is a composition of $u \to T$ and $x \to \infty$, these two results are equivalent through a logarithm transformation.

Detailed proof can be found in [16]. It is worth noting that a direct proof is also possible, without introducing Y_u as a *time reversed* branching process. Reader with interest may refer to the proof of Theorem 3.4.6.

Simulation results for $e^{-ru}Z_{T-u}$ in the subcritical case are shown in Figure 3.9.

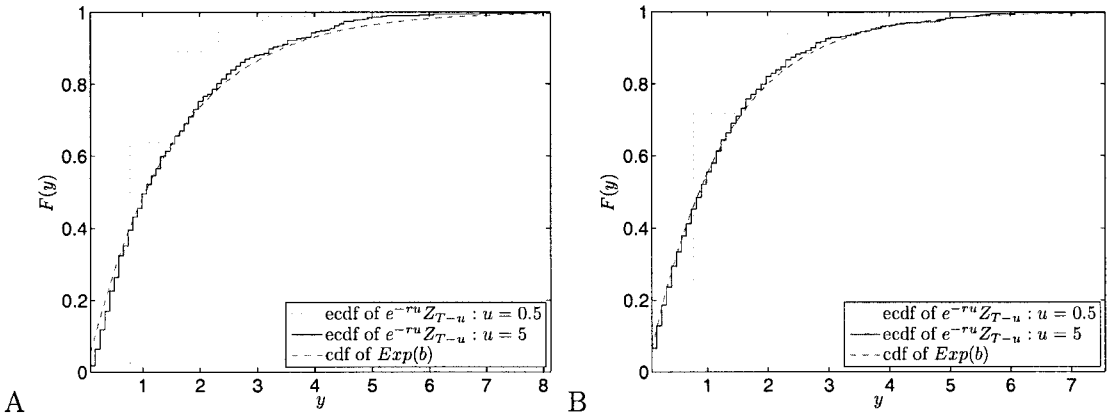


Figure 3.9: Empirical cdf of $e^{-ru}Z_{T-u}$ in subcritical ($m = 0.5$) MBP with initial population size $x = 100$, based on 500 simulations. (A) binary fission, $u = 0.5$ and 5; (B) Poisson, $u = 0.5$ and 5.

3.4 Extinction of Critical Markov Branching Processes

All the results in the previous section can be extended to the critical case with finite variance. In this section, we use similar techniques to find out the extinction pattern of critical MBP, which turns out to be analogous to that of subcritical MBP. Simulations are also performed to illustrate the theoretical results. However, it should be noted that, the simulation of critical MBP usually encounters computational problems since it is "harder" for the process to reach its extinction.

3.4.1 Time to Extinction

Proposition 3.4.1. *Suppose Z_t is a critical Markov branching process, with initial population size x and time to extinction T. If $\sigma^2 < \infty$, then*

$$\frac{a\sigma^2}{2x}T \xrightarrow{d} \xi \text{ as } x \to \infty, \tag{3.21}$$

where ξ is distributed as standard Frechet with shape parameter 1.

Proof. If $\sigma^2 < \infty$, by Proposition 3.2.2, $Q(t) \sim \frac{2}{a\sigma^2 t}$, as $t \to \infty$. Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1-G(ty)}{1-G(t)} &= \lim_{t\to\infty} \frac{Q(ty)}{Q(t)} \\
&= \lim_{t\to\infty} \frac{Q(ty)}{\frac{2}{a\sigma^2 ty}} \cdot \frac{\frac{2}{a\sigma^2 t}}{Q(t)} \cdot y^{-1} \\
&= y^{-1}, \forall y > 0.
\end{aligned}$$

By Theorem 1.6.2 of [24], $G(t)$ belongs to the Type-II (Frechet) domain of attraction. By Corollary 1.6.3 of [24], the normalizing constants a_x, b_x can be determined by $a_x = \frac{1}{G^{-1}(1-1/x)} = \frac{a\sigma^2}{2x}$ and $b_x = 0$, which gives (3.21). □

Figure 3.10 illustrates the simulation results. Due to computational constraints, we only set $x = 10$ in the critical case. We also plot the relation between x and T in the simulations. In the critical case, the first moment of T does not exist, so we use its median instead. Again we obtain a satisfying match on $median(T) = \frac{2x}{a\sigma^2 \ln 2}$, where $\ln 2$ comes from the median of standard Frechet distribution. Note that the simulated trajectory has a bias due to restricting x to small values, as we mentioned before. An alternative way to check the relation between x and T in the critical case

is to show the simulated trajectory of the mean of $1/T$, which theoretically follows standard exponential distribution. Although this method may not be as robust as the previous one, it works when the censoring times are large enough.

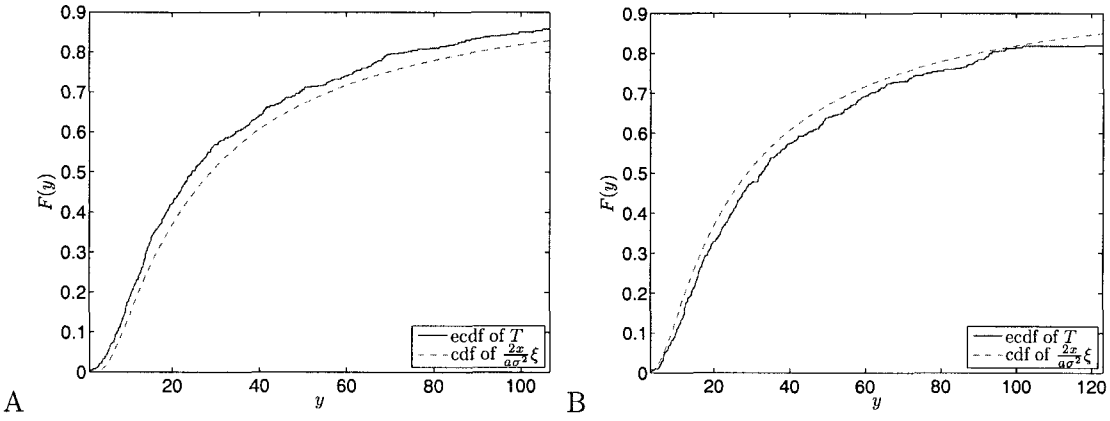


Figure 3.10: Empirical cdf of the time to extinction T in critical MBP with initial population size $x = 10$, based on 500 simulations. (A) binary fission; (B) Poisson. There are 89 simulations lost to follow up.

Similarly to the subcritical case, we have such a theorem for the age-dependent process, as an extension to Proposition 3.4.1.

Theorem 3.4.2. *For an age-dependent branching process, Proposition 3.4.1 also holds. Here a should be substituted by $1/\mu$, where μ is the mean of offspring life length distribution.*

Proof. Let $T^* = \frac{\sigma^2}{2\mu x}T$. The cdf of T^* is $P_x(T^* \leq t) = P_x\left(T \leq \frac{2\mu x}{\sigma^2}t\right)$. Let $\tau = \frac{2\mu x}{\sigma^2}t$. Then $P_x(T^* \leq t) = [1 - P_1(Z_\tau > 0)]^x$. By the known fact that $P_1(Z_\tau > 0) \sim \frac{2\mu}{\tau\sigma^2}$ as $\tau \to \infty$ (see [3] page 159), we obtain $P_1(Z_\tau > 0) - \frac{2\mu}{\tau\sigma^2} = o(\frac{2\mu}{\tau\sigma^2})$. Using the transformation $\tau = \frac{2\mu x}{\sigma^2}t$, for fixed t, $x \to \infty$ implies $\tau \to \infty$. Therefore, $P_1(Z_\tau >$

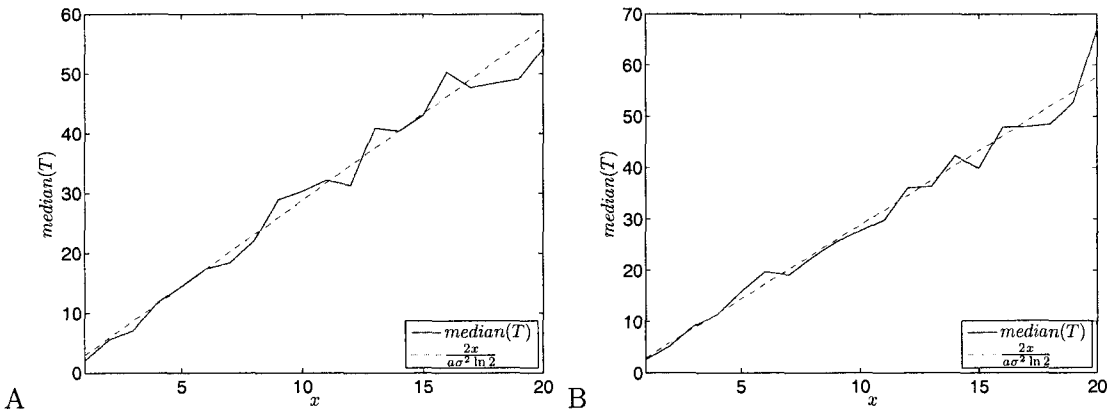


Figure 3.11: Scatter plot of x versus sample median of T in critical MBP, based on 500 simulations. (A) binary fission; (B) Poisson.

$0) - \frac{1}{tx} = o(\frac{1}{tx})$. Again, for simplicity, write $1 - P_1(Z_\tau > 0)$ as $f(x)$. Then

$$
\begin{aligned}
& f(x) - \left[1 - \frac{1}{tx}\right] = o\left(\frac{1}{tx}\right) \\
\Rightarrow \quad & \frac{f(x) - \left[1 - \frac{1}{tx}\right]}{\left[1 - \frac{1}{tx}\right]} = o\left(\frac{\frac{1}{tx}}{\left[1 - \frac{1}{tx}\right]}\right) \\
\Rightarrow \quad & \frac{f(x) - \left[1 - \frac{1}{tx}\right]}{\left[1 - \frac{1}{tx}\right]} = o\left(\frac{1}{tx}\right) \text{ since } \frac{\frac{1}{tx}}{\left[1 - \frac{1}{tx}\right]} = O\left(\frac{1}{tx}\right) \\
\Rightarrow \quad & \frac{f(x)}{1 - \frac{1}{tx}} - 1 = o\left(\frac{1}{tx}\right) = o(1/x) \text{ for fixed } t.
\end{aligned}
$$

By Lemma 3.3.2, we have $\lim_{x\to\infty} \left[\frac{f(x)}{1-\frac{1}{tx}}\right]^x = 1$. So, $\lim_{x\to\infty} [1 - P_1(Z_\tau > 0)]^x = e^{-1/t}$. This implies $\lim_{x\to\infty} P_x(T^* \leq t) = e^{-1/t}$, hence $\frac{\sigma^2}{2\mu x} T \xrightarrow{d} \xi$ as $x \to \infty$, where ξ is distributed as Frechet with shape parameter 1. □

3.4.2 Path to Extinction

Comparing to the subcritical case, it seems more difficult to predict the path in the critical case because the extinction occurs in a slowgoing way. Figure 3.12 shows

an example of population size, Z_t and Z_{uT}, of a critical MBP with binary fission offspring distribution. It is worth noting that in the simulation of the critical MBP, extinctions may occur very late. Therefore the computing resources for storing the whole simulated process may be exhausted before the extinction has been reached, especially when we set a large initial population size. In such a situation, we only can treat the time to extinction in the simulation as "lost to follow-up" if the extinction occurs later than a certain generation bound. The trajectories lost to follow-up are not shown for the processes Z_{uT} in Figure 3.12.

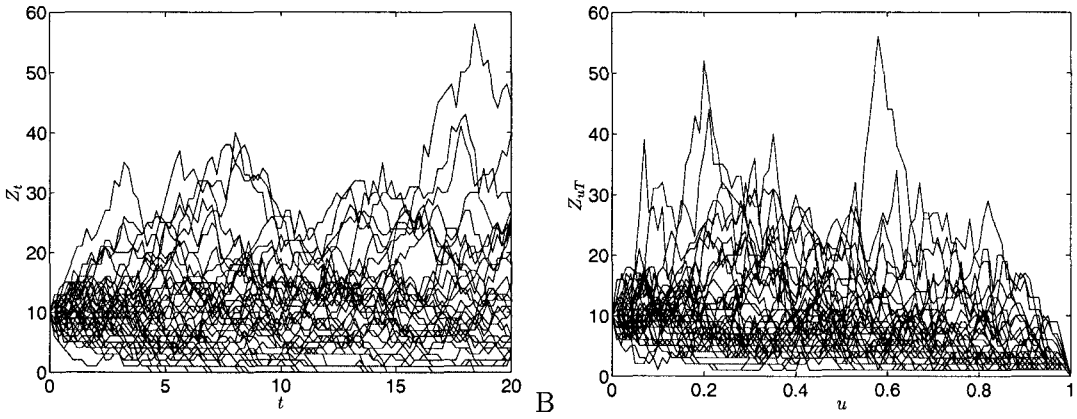


Figure 3.12: Path of critical binary fission MBP with initial population size $x = 10$, based on 50 simulations. (A) Z_t; (B) Z_{uT}. There are 3 simulations lost to follow up.

Let us now consider the process Z_{uT}/T.

Theorem 3.4.3. *The mgf of Z_{uT}/T is given by*

$$E_x[e^{\nu Z_{uT}/T}] = x \int_0^\infty e^{\nu/t} F^{x-1}(e^{\nu/t} G((1-u)t), ut) F'(e^{\nu/t} G((1-u)t), ut) g((1-u)t) dt, \tag{3.22}$$

where $\nu \leq 0, x \in \mathcal{Z}^+, 0 \leq u < 1$.

Proof. By the law of total probability

$$\begin{aligned}
E_x[s^{Z_{uT}/T}] &= \int_0^\infty E_x[s^{Z_{uT}/T}, T \in dt] \\
&= \int_0^\infty E_x[s^{Z_{ut}/t}, T \in dt] \\
&= \int_0^\infty \sum_y s^{y/t} P_x(Z_{ut} = y, T \in dt) \\
&= \int_0^\infty \sum_y s^{y/t} P_x(Z_{ut} = y) P_x(T \in dt | Z_{ut} = y) \\
&= \int_0^\infty \sum_y s^{y/t} P_x(Z_{ut} = y) P_y(T' + ut \in dt), \text{ define } T' = (1-u)T \\
&= \int_0^\infty \sum_y s^{y/t} P_x(Z_{ut} = y) \frac{d}{d(1-u)t} G^y((1-u)t) dt \\
&= \int_0^\infty s^{1/t} \frac{\partial}{\partial s_t} F^x(s_t, ut) g((1-u)t) dt, \text{ define } s_t = s^{1/t} G((1-u)t) \\
&= x \int_0^\infty s^{1/t} F^{x-1}(s^{1/t} G((1-u)t), ut) F'(s^{1/t} G((1-u)t), ut) g((1-u)t) dt.
\end{aligned}$$

Replacing s by e^ν then gives the mgf. □

Consequently, we obtain such a theorem:

Theorem 3.4.4. *For critical MBP* Z_t *with* $\sigma^2 < \infty$, *if* $Z_0 = x$ *and* Z_t *hits zero at time* T, *then* Z_{uT}/T *converges in distribution as* $x \to \infty$, *for each* $0 \leq u < 1$, *to a random variable with mgf* $\frac{1}{1-u} \cdot \frac{1}{1-\frac{a\sigma^2}{2}(1-u)\nu} - \frac{u}{1-u} \cdot \frac{1}{1-\frac{a\sigma^2}{2}u(1-u)\nu}$. *As* $u \downarrow 0$, *the limit distribution approaches exponential with parameter* $\frac{2}{a\sigma^2}$.

This result is parallel to the subcritical-case version (Theorem 3.3.4), in which Z_{uT} is normalized by x^{u-1} and the process $x^{u-1}Z_{uT}$ has trajectories with a weak limit

$b^u e^{-u\eta}$, where η is a standard Gumbel random variable (rv). The second conclusion, the limit distribution approaches exponential with parameter $\frac{2}{a\sigma^2}$ as $u \downarrow 0$, can be considered an alternative representation of Proposition 3.4.1. Intuitively, $Z_{uT} \to x$ almost surely as $u \downarrow 0$. Therefore we expect Z_{uT}/T to perform similarly as x/T, which converges in distribution as $x \to \infty$, to $\frac{a\sigma^2}{2\xi}$, that is, an exponential rv with parameter $\frac{2}{a\sigma^2}$. Of course this intuition is true only when the two limits in x and u are exchangeable.

Proof. First, by (3.8) and (3.10), as $t \to \infty$

$$\begin{aligned}
\pi(e^{\nu/t}G((1-u)t)) &\sim \frac{2}{a\sigma^2[1-e^{\nu/t}G((1-u)t)]} \\
&\sim \frac{2}{a\sigma^2\left[-\frac{\nu}{t}+\frac{2}{a\sigma^2(1-u)t}\right]} \\
&= \frac{t}{-\frac{a\sigma^2}{2}\nu+\frac{1}{1-u}}.
\end{aligned}$$

Second, as $t \to \infty$

$$\begin{aligned}
Q(ut+\pi(e^{\nu/t}G((1-u)t))) &\sim \frac{2}{a\sigma^2[ut+\pi(e^{\nu/t}G((1-u)t))]} \\
&\sim \frac{1}{t\left[\frac{a\sigma^2}{2}u+\frac{1}{-\nu+\frac{2}{a\sigma^2(1-u)}}\right]}.
\end{aligned}$$

By (3.17), for $t \to \infty$, we obtain the first term in (3.22)

$$\begin{aligned}
F^{x-1}(e^{\nu/t}G((1-u)t), ut) &= [1-Q(ut+\pi(e^{\nu/t}G((1-u)t)))]^{x-1} \\
&\sim \left\{1-\frac{1}{\frac{a\sigma^2}{2}t\left[u+\frac{1}{-\frac{a\sigma^2}{2}\nu+\frac{1}{1-u}}\right]}\right\}^{x-1}.
\end{aligned}$$

By l'Hopital's rule, (3.8) and (3.10) also indicate that as $t \to \infty$

$$\begin{aligned}
\pi'(e^{\nu/t}G((1-u)t)) &\sim \frac{2}{a\sigma^2[1-e^{\nu/t}G((1-u)t)]^2} \\
&\sim \frac{2}{a\sigma^2\left[-\frac{\nu}{t}+\frac{2}{a\sigma^2(1-u)t}\right]^2} \\
&= \frac{t^2}{\frac{a\sigma^2}{2}\left[-\nu+\frac{2}{a\sigma^2(1-u)}\right]^2}.
\end{aligned}$$

$$\begin{aligned}
-Q'(ut+\pi(e^{\nu/t}G((1-u)t))) &\sim \frac{2}{a\sigma^2[ut+\pi(e^{\nu/t}G((1-u)t))]^2} \\
&\sim \frac{1}{\frac{a\sigma^2}{2}t^2\left[u+\frac{1}{-\frac{a\sigma^2}{2}\nu+\frac{1}{(1-u)}}\right]^2}.
\end{aligned}$$

Therefore, the second term

$$\begin{aligned}
F'(e^{\nu/t}G((1-u)t),ut) &= -Q'(ut+\pi(e^{\nu/t}G((1-u)t)))\pi'(e^{\nu/t}G((1-u)t)) \\
&\sim \frac{1}{\frac{a\sigma^2}{2}t^2\left[u+\frac{1}{-\frac{a\sigma^2}{2}\nu+\frac{1}{(1-u)}}\right]^2}\cdot\frac{t^2}{\frac{a\sigma^2}{2}\left[-\nu+\frac{2}{a\sigma^2(1-u)}\right]^2} \\
&= \frac{1}{(\frac{1}{1-u}-\frac{a\sigma^2}{2}u\nu)^2}.
\end{aligned}$$

The third term $g((1-u)t) = -Q'((1-u)t) \sim \frac{2}{a\sigma^2(1-u)^2t^2}$.

With $t = \frac{2z}{a\sigma^2}x$, we obtain

$$\begin{aligned}
&xe^{\nu/t}F^{x-1}(e^{\nu/t}G((1-u)t),ut)F'(e^{\nu/t}G((1-u)t),ut)g((1-u)t) \\
\sim\ &\exp\left[-\frac{1}{z\left(u+\frac{1}{-\frac{a\sigma^2}{2}\nu+\frac{1}{1-u}}\right)}\right]\cdot\frac{1}{(\frac{1}{1-u}-\frac{a\sigma^2}{2}u\nu)^2}\cdot\frac{a\sigma^2}{2(1-u)^2z^2x} \text{ as } x\to\infty.
\end{aligned}$$

If dominated convergence applies, then identity (3.22) can be simplified as

$$\begin{aligned} E_x[e^{\nu Z_{uT}/T}] &\rightarrow \frac{1}{(1-\frac{a\sigma^2}{2}u(1-u)\nu)^2}\int_0^\infty \exp\left[-\frac{1}{z\left(u+\frac{1}{-\frac{a\sigma^2}{2}\nu+\frac{1}{1-u}}\right)}\right]\cdot\frac{1}{z^2}dz \\ &= \frac{1}{1-u}\cdot\frac{1}{1-\frac{a\sigma^2}{2}(1-u)\nu}-\frac{u}{1-u}\cdot\frac{1}{1-\frac{a\sigma^2}{2}u(1-u)\nu} \end{aligned} \tag{3.23}$$

We see that the right hand side of (3.23) is a linear combination of two exponential mgf's, one with parameter $\frac{2}{a\sigma^2(1-u)}$, the other with parameter $\frac{2}{a\sigma^2u(1-u)}$. By continuity of the mgf, this proves the convergence of Z_{uT}/T. The proof of the limiting exponential distribution as $u \downarrow 0$ is straightforward.

The proof of dominated convergence follows the same way as in [16]. That is,

$$\begin{aligned} 0 &\leq xF^{x-1}(e^{\nu/t}G((1-u)t),ut)F'(e^{\nu/t}G((1-u)t),ut)g((1-u)t) \\ &\leq c_1x\exp(-(x-1)(1-F(e^{\nu/t}G((1-u)t),ut)))F'(1,ut)\frac{1}{t^2} \\ &\leq c_2x\exp(-c_3x(1-F(e^{\nu/t}G((1-u)t),ut)))E_1[ut]\frac{1}{t^2} \\ &\leq c_4\exp(-c_3xQ(ut+\pi(e^{\nu/t}G((1-u)t))))\frac{1}{z^2}\frac{1}{x} \\ &\leq c_4\exp(-c_5\frac{x}{t})\frac{1}{z^2}\frac{1}{x} \\ &= c_4e^{-\frac{c_6}{z}}\frac{1}{z^2}\frac{1}{x}, \end{aligned}$$

where the c_i are suitable positive constants. □

We use simulation to illustrate this result. Figure 3.13 plots the empirical cdf of the simulated Z_{uT}/T process at fixed u's to check, as x increases, whether the distribution of Z_{uT}/T approaches the limiting distribution by Theorem 3.4.4. Note that the trajectories lost to follow up will bias the empirical cdf, so large initial

population size x is not practical. In the simulation, we set x to be three different values 2, 5 and 10, and set the generation bound to be 500 to reduce the effect of losing tracking. We see that for a given u, as x increases, the empirical cdf is getting close to its theoretical limit. In particular, as shown in Figure 3.13 panel (A), when $u = 0$, the theoretical limit becomes an exponential distribution with parameter $\frac{2}{a\sigma^2}$.

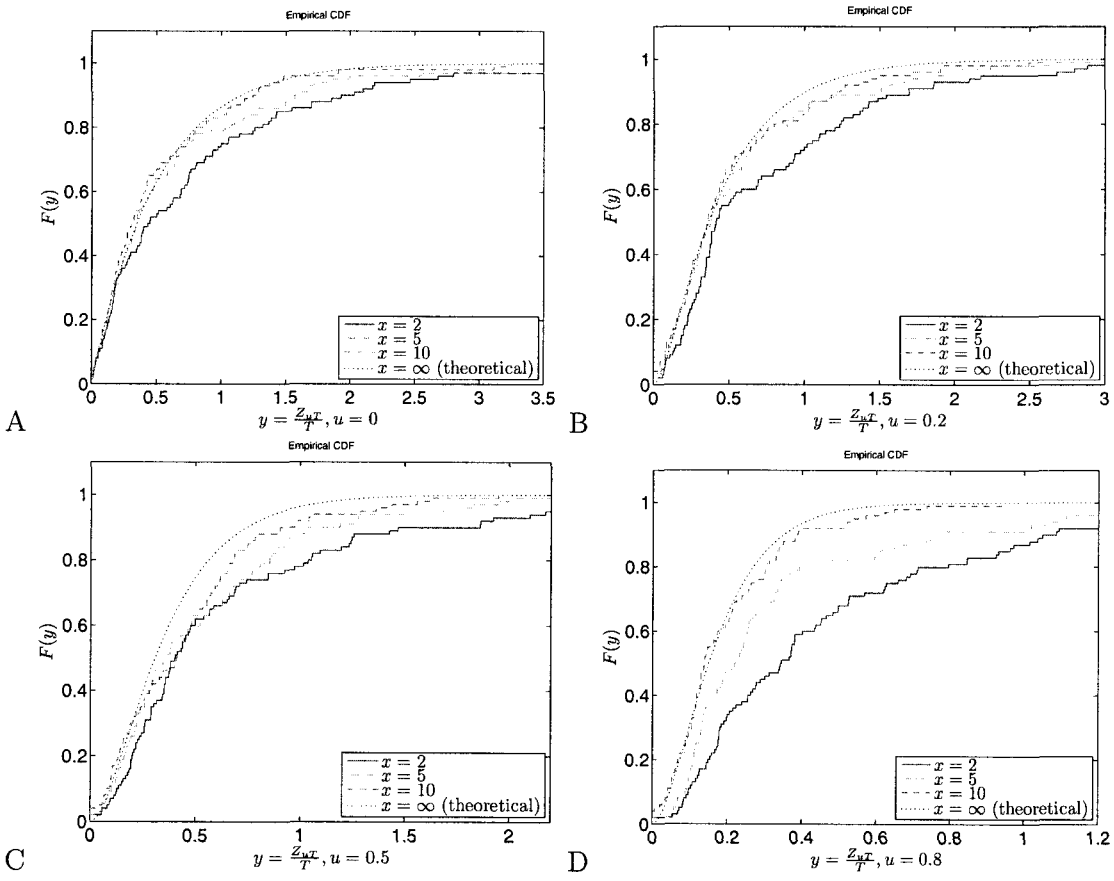


Figure 3.13: Empirical cdf of Z_{uT}/T in critical binary fission MBP, based on 100 simulations. (A) $u = 0$; (B) $u = 0.2$; (C) $u = 0.5$; (D) $u = 0.8$. There are $2, 2, 4$ out of the 100 simulations lost to follow-up when the initial population size $x = 2, 5, 10$, correspondingly.

3.4.3 Path Verging on Extinction

Let us now consider the limiting behavior of the process Z_{T-u}, for $u > 0$.

Theorem 3.4.5. *The pgf of Z_{T-u} is given by*

$$E_x[s^{Z_{T-u}}] - P_x(T < u) = sg(u)x \int_u^\infty F^{x-1}(sG(u), t-u)F'(sG(u), t-u)dt, \quad (3.24)$$

where $P_x(T < u) \downarrow 0$, as $x \to \infty$.

This result holds for both subcritical and critical MBP. The proof follows the same pattern as in the proof of Theorem 3.3.4.

Theorem 3.4.6. *For critical MBP Z_t with finite variance, Z_{T-u} converges in distribution as $x \to \infty$, for $u > 0$, to a rv Y_u with $E[s^{Y_u}] = sg(u)\pi'(sG(u))$. As $u \to \infty$, Y_u/u converges in distribution to a gamma rv with shape parameter 2 and scale parameter $a\sigma^2/2$.*

We note that the first conclusion in Theorem 2, i.e., the pgf of Y_u has already been proved by Jagers *et al.* (see Proposition 3, Equation (33) in [16]). However, it is instructive to see that a direct proof is also possible, for both subcritical and critical cases, without introducing Y_u as a *time reversed* branching process. This proof also leads directly to the derivation of the limit of Y_u/u.

Proof. By (3.17) and (3.10), as $t \to \infty$

$$\begin{aligned} & F^{x-1}(sG(u), t-u)F'(sG(u), t-u) \\ = \; & -[1 - Q(t-u+\pi(sG(u)))]^{x-1} \cdot Q'(t-u+\pi(sG(u)))\pi'(sG(u)) \\ \sim \; & \left[1 - \frac{2}{a\sigma^2(t-u+\pi(sG(u)))}\right]^{x-1} \cdot \frac{2\pi'(sG(u))}{a\sigma^2[t-u+\pi(sG(u))]^2}. \end{aligned}$$

Since $\forall u > 0, 0 \leq sG(u) < 1$, with $t = \frac{2zx}{a\sigma^2}$, we obtain

$$\begin{aligned} & F^{x-1}(sG(u), t-u)F'(sG(u), t-u) \\ \sim & \left[1 - \frac{2}{2zx + a\sigma^2(-u + \pi(sG(u)))}\right]^{x-1} \cdot \frac{2\pi'(sG(u))}{a\sigma^2\left[\frac{2zx}{a\sigma^2} - u + \pi(sG(u))\right]^2}, \end{aligned}$$

as $x \to \infty$.

Given dominated convergence, Lemma 3 then follows

$$\begin{aligned} E_x[s^{Z_{T-u}}] &\to sg(u)\pi'(sG(u)) \int_0^\infty e^{-\frac{1}{z}} \frac{1}{z^2} dz \\ &= sg(u)\pi'(sG(u)). \end{aligned}$$

This yields the desired convergence in distribution to Y_u. The proof of dominated convergence is quite similar to that of Theorem 2 and is thereby skipped here. Replacing s by $e^{\frac{\nu}{u}}$, as $u \to \infty$, by (3.8) and (3.10), we obtain

$$\begin{aligned} E_x[e^{\nu \frac{Y_u}{u}}] &= e^{\frac{\nu}{u}} g(u)\pi'(e^{\frac{\nu}{u}} G(u)) \\ &\sim e^{\frac{\nu}{u}} \cdot \frac{2}{a\sigma^2 u} \cdot \frac{2}{a\sigma^2\left[1 - e^{\frac{\nu}{u}} G(u)\right]^2} \\ &\sim e^{\frac{\nu}{u}} \cdot \left[\frac{2}{a\sigma^2 u\left[-\frac{\nu}{u} + \frac{2}{a\sigma^2 u}\right]}\right]^2 \\ &\to \frac{1}{\left[1 - \frac{a\sigma^2}{2}\nu\right]^2}. \end{aligned}$$

Starting line two, the argument is specific for the critical process. This completes the proof. □

Figure 3.14 plots the empirical cdf of the simulated Z_{T-u}/u process at different u's to check, as u increases, whether the distribution of Z_{T-u}/u approaches the limiting

Gamma distribution by Theorem 3.4.6. Due to computational limitation, we set $x = 10$. We see that the empirical cdf does follow its theoretical limit as u increases.

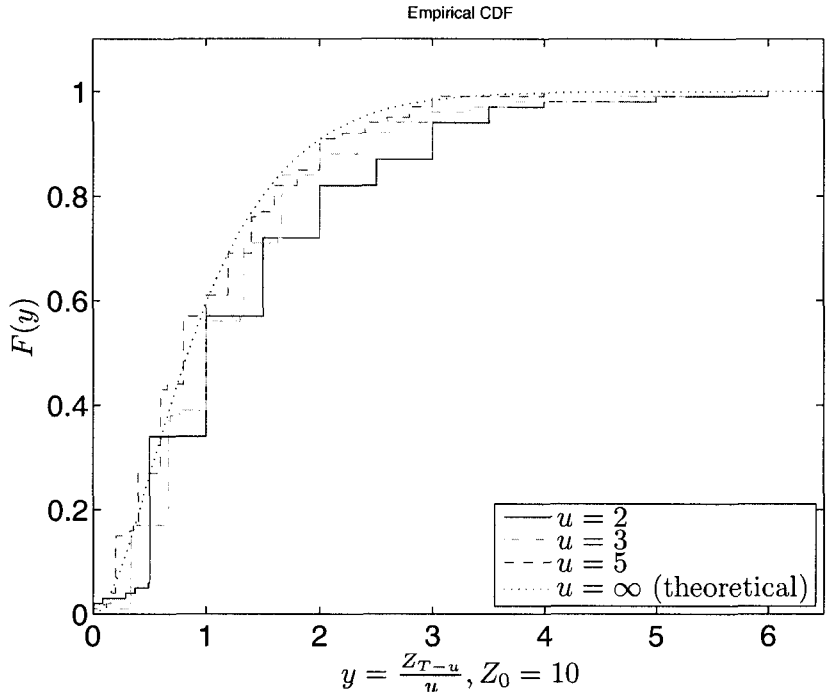


Figure 3.14: Empirical cdf of Z_{T-u}/u in critical binary fission MBP, based on 100 simulations. There are 6 out of the 100 simulations lost to follow-up when the initial population size $x = 10$.

3.5 Discussion

In Section 3.1.2, we briefly introduce the motivation of this work: Nagylaki's theory of genetic drift approximation using branching processes. In particular, we are interested in the case that (i) no mutation exists, i.e., $\mu_{ri} = 0$ in Equation (3.2); and (ii) selection coefficients $w_{ir} = 1$ for $i = 1, \cdots, r-1$. Under these settings, the

allelic number $Y_i(n)$ is a critical branching process with Poisson offspring distribution. It is then of special interest to see how the critical process of rare alleles dies out in the end.

The rest of this chapter is devoted to the study of the extinction of subcritical/critical Markov branching processes started from a large number of individuals. The theoretical result of the time and path to extinction in MBP may be applied to experimental genetic data to study the dynamics of extinction of disease-causing variants of genes in human populations. As a conclusion to our current-stage work, we summarize the results about the extinction of subcritical/critical MBP in Table 3.1. Concerning this direction, there are some extensions to MBP with immigration, multi-type MBP or continuous-state MBP which we would like to explore in the future.

Table 3.1: Summary of the extinction of Markov branching processes.

Characteristic	Subcritical: $m < 1, x \log x$-condition	Critical: $m = 1, \sigma^2 < \infty$	Theorem
$\pi(s)$	$\pi(s) \sim -r^{-1} \ln(1-s)$ as $s \uparrow 1$	$\pi(s) \sim \frac{2}{a\sigma^2(1-s)}$ as $s \uparrow 1$	Lemma 3.2.1
$Q(t)$	$Q(t) \sim b^{-1} e^{-rt}$ as $t \to \infty$	$Q(t) \sim \frac{2}{a\sigma^2 t}$ as $t \to \infty$	Proposition 3.2.2
Conditional Limit Laws	$P_1(Z_t = k \mid Z_t > 0) \to b_k$ as $t \to \infty$	$P_1\left(\frac{Z_t}{t} > z \mid Z_t > 0\right) \to e^{-\frac{2}{a\sigma^2}z}$ as $t \to \infty$	Proposition 3.2.3
T	$rT - \ln x - \ln c \xrightarrow{d} \eta$ as $x \to \infty$	$\frac{a\sigma^2}{2x} T \xrightarrow{d} \xi$ as $x \to \infty$	Proposition 3.3.1/3.4.1
Z_{uT}	$x^{u-1} Z_{uT} \xrightarrow{fdd} b^u e^{-u\eta}$ as $x \to \infty$	$\Phi_{\frac{Z_{uT}}{T}}(\nu) \to \frac{1}{1-u} \cdot \frac{1}{1-\frac{a\sigma^2}{2}(1-u)\nu} - \frac{u}{1-u} \cdot \frac{1}{1-\frac{a\sigma^2}{2}u(1-u)\nu}$	Theorem 3.3.5/3.4.4
Z_{T-u}	$e^{-ru} Z_{T-u} \xrightarrow{d} Exp(b)$ as $x \to \infty, u \to \infty$	$\frac{Z_{T-u}}{u} \xrightarrow{d} Gamma(2, \frac{a\sigma^2}{2})$ as $x \to \infty, u \to \infty$	Theorem 3.3.6/3.4.6

Chapter 4

The Infinite-Allele Markov Branching Process

In this chapter, we consider a process which grows according to branching rule, however individuals can mutate into novel allelic forms. More precisely, we assume that the mutation is independent of the previous history of the process, and the offspring distribution is independent of allelic type, i.e., the selection is neutral for all alleles. Under these settings, the process can be described as an "infinite-many-alleles" model. That is, whenever a mutation happens, it yields a new allele, which differs from all the existing ones. Because there is no fixed labeling of allelic types, analysis of such models is usually complicated. However, Pakes obtained some results for the Galton-Watson branching process [11] and the Markov branching process [32]. We are mainly interested in the result concerning the frequency spectrum because

the allele frequency information is usually available in many genetic processes.

4.1 Pakes' Theory

We briefly introduce Pakes' theory on infinite-allele MBP in this section. Suppose the MBP has exponential life span with parameter a, and the mean of its offspring distribution is m, regardless of allelic types. We assume further that the process starts from i individuals carrying the same allele, and a new-born individual is able to mutate into a novel allelic type with probability μ. Let $\alpha_t(j)$ be the number of alleles which are represented by j individuals at time t. Our objective is to find

$$\phi_{i,t}(j) := E[\alpha_t(j)],$$

from which the frequency spectrum is easy to achieve by normalization.

Let $T_1, T_2, \cdots$ be the successive split times, N_t be the number of split times till time t and U_n be the number of offspring produced at split time T_n. Consider that at time t, the alleles which are represented by j individuals are from two sources: the initial allele or the mutant alleles. Correspondingly, we define two indicator functions: $I_{0,j}(t) = 1$ if there are j individuals carrying the initial allele alive at time t; and $I_{n,k,j}(t) = 1$, for $n, k \geq 1$ if the kth individual born at time T_n ($T_n < t$) mutates to a novel allelic type and further produces j individuals carrying this allele t units later. Then

$$\alpha_t(j) = I_{0,j}(t) + \sum_{n=1}^{N_t} \sum_{k=1}^{U_n} I_{n,k,j}(t - T_n).$$

It is easy to see that $E[I_{0,j}(t)] = q_{ij}(t)$, and since for each n, $I_{n,k,j}(t)$ is independent of U_n and T_n, $E[I_{n,k,j}|U_n, T_n] = \mu q_{1j}(t)$. Therefore

$$\phi_{i,t}(j) = E[\alpha_t(j)] = q_{ij}(t) + m\mu E_i \left[\sum_{n=1}^{N_t} q_{1j}(t - T_n)\right].$$

In Lemma 3.1.1 of [32], Pakes has shown that

$$E_i \left[\sum_{n=1}^{N_t} \alpha(t - T_n)\right] = iae^{\lambda t} \int_0^t e^{-\lambda u \alpha(u)} du,$$

where $\lambda = a(m-1)$ and $\alpha(t)$ is a bounded continuous function. So

$$\phi_{i,t}(j) = q_{ij}(t) + iam\mu e^{\lambda t} \int_0^t e^{-\lambda x} q_{1j}(x) dx. \tag{4.1}$$

Since $0 \leq q_{ij}(t) \leq 1$, we see that it is possible to determine, asymptotically, the expected frequency spectrum for an infinite-allele MBP. Let

$$G_j(\lambda) = \int_0^\infty e^{-\lambda t} q_{1j}(t) dt, j \geq 0, \tag{4.2}$$

the limiting frequency spectrum can be obtained as

$$\psi(j) = \frac{G_j(\lambda)}{\sum_{j\geq 1} G_j(\lambda)} = \frac{\lambda G_j(\lambda)}{1 - \lambda G_0(\lambda)}, j \geq 1. \tag{4.3}$$

It should be noted that in such case the supercritical condition w.r.t parental allele is necessary, i.e., $M \geq 1$, where $M = m(1-\mu)$ is the mean number of offspring carrying the parental allele, so that $\lambda = a(m-1) > a(M-1) \geq 0$.

In general, the transition probability q_{1j} can rarely be determined explicitly because the Kolmogorov equations are usually hard to solve. Hence the limiting frequency spectrum $\psi(j)$ is difficult to obtain. However, there are two important cases where the explicit form of q_{1j} or its pgf can be obtained, namely the Yule process and the birth-death process.

4.2 Frequency Spectrum of The Infinite-Allele Birth-Death Process

4.2.1 Derivation of The Limiting Frequency Spectrum

For the birth-death process, (one form of) the offspring pgf w.r.t the parental allele can be written as

$$f(s) = (\alpha + \beta\mu + \beta(1-\mu)s)^2,$$

where α, β, μ stand for the death, birth and mutation probabilities for every individual, and $\alpha + \beta = 1$. Write $A = \alpha + \beta\mu, B = \beta(1-\mu)$, with $A + B = 1$. Then

$$f(s) = (A + Bs)^2.$$

By backward Kolmogorov equation, the process pgf $F(s,t)$ satisfies

$$\begin{aligned}
\frac{\partial F}{\partial t} &= a[(A+BF)^2 - F] \\
&= a[B^2F^2 - (1-2AB)F + A^2] \\
&= a[B^2F^2 - (A^2+B^2)F + A^2] \\
&= a(A^2+B^2)\left[\frac{1}{A^2+B^2}(A^2 + B^2F^2) - F\right].
\end{aligned}$$

Therefore, this process is equivalent to a birth-death process with $\tilde{a} = a(A^2+B^2)$ and offspring pgf $\tilde{f}(s) = \frac{1}{A^2+B^2}(A^2 + B^2s^2)$. Using the known results for the pgf of the birth-death process, we see that

$$F(s,t) = \frac{A^2(1-s) - (A^2 - B^2s)e^{-ct}}{B^2(1-s) - (A^2 - B^2s)e^{-ct}},$$

where $c = B^2a - A^2a = (B - A)a > 0$.

To obtain an explicit form for $G_j(\lambda)$, there may be two approaches. The first approach starts from finding the pgf of $G_j(\lambda)$, which then leads to its coefficients. The second approach attempts to find $q_{1j}(t)$ directly, and apply (4.2).

(1) Denote the pgf of $G_j(\lambda)$ by $\gamma(s)$; then

$$\begin{aligned}\gamma(s) &:= \sum_{j\geq 0} G_j(\lambda)s^j \\ &= \int_0^\infty \frac{A^2(1-s) - (A^2 - B^2 s)e^{-ct}}{B^2(1-s) - (A^2 - B^2 s)e^{-ct}} e^{-\lambda t} dt.\end{aligned}$$

Let $v = e^{-ct}$,

$$\begin{aligned}\gamma(s) &= \frac{1}{c}\int_0^1 \frac{A^2(1-s) - (A^2 - B^2 s)v}{B^2(1-s) - (A^2 - B^2 s)v} v^{\frac{\lambda}{c}-1} dv \\ &= \frac{1}{c}\int_0^1 \frac{A^2(1-v)}{B^2 - A^2 v} \cdot \frac{1 - \frac{A^2 - B^2 v}{A^2(1-v)}s}{1 - \frac{B^2(1-v)}{B^2 - A^2 v}s} v^{\frac{\lambda}{c}-1} dv \\ &= \frac{1}{c}\int_0^1 \frac{A^2(1-v)}{B^2 - A^2 v}\left[1 - \frac{A^2 - B^2 v}{A^2(1-v)}s\right] \sum_{j\geq 0}\left[\frac{B^2(1-v)}{B^2 - A^2 v}\right]^j s^j v^{\frac{\lambda}{c}-1} dv.\end{aligned}$$

From the term that does not contain s, we can easily read $G_0(\lambda)$ as

$$\begin{aligned}G_0(\lambda) &= \frac{1}{c}\int_0^1 \frac{A^2(1-v)}{B^2 - A^2 v} v^{\frac{\lambda}{c}-1} dv \\ &= \frac{A^2}{cB^2}\int_0^1 \frac{v^{\frac{\lambda}{c}-1}(1-v)}{1 - \frac{A^2}{B^2}v} dv.\end{aligned}$$

Since

$$\int_0^1 \frac{v^{b-1}(1-v)^{c-b-1}}{(1-vz)^a} dv = \frac{\Gamma(b)\Gamma(c-b)}{\Gamma(c)} F(a, b; c; z),$$

where $F(\cdot, \cdot; \cdot; \cdot)$ is the (Gauss's) hypergeometric function. It follows that

$$G_0(\lambda) = \frac{A^2}{cB^2}\frac{\Gamma(\frac{\lambda}{c})\Gamma(2)}{\Gamma(2 + \frac{\lambda}{c})} F\left(1, \frac{\lambda}{c}; 2 + \frac{\lambda}{c}; \frac{A^2}{B^2}\right).$$

Similarly, $G_j(\lambda)$ can be read from the coefficients:

$$\begin{aligned} G_j(\lambda) &= \frac{1}{c}\int_0^1 \frac{A^2(1-v)}{B^2-A^2v}\left\{\left[\frac{B^2(1-v)}{B^2-A^2v}\right]^j - \frac{A^2-B^2v}{A^2(1-v)}\cdot\left[\frac{B^2(1-v)}{B^2-A^2v}\right]^{j-1}\right\} v^{\frac{\lambda}{c}-1}dv \\ &= \frac{(B-A)^2}{cB^4}\int_0^1 \frac{v^{\frac{\lambda}{c}}(1-v)^{j-1}}{\left(1-\frac{A^2}{B^2}v\right)^{j+1}}dv \\ &= \frac{(B-A)^2}{cB^4}\frac{\Gamma(1+\frac{\lambda}{c})\Gamma(j)}{\Gamma(j+1+\frac{\lambda}{c})}F\left(j+1,1+\frac{\lambda}{c};j+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right), j\geq 1. \end{aligned}$$

(2) From the explicit form of $F(s,t)$, we can directly read $q_{1j}(t)$.

$$\begin{aligned} F(s,t) &= \frac{A^2(1-e^{-ct})}{B^2-A^2e^{-ct}+B^2(e^{-ct}-1)s} + \frac{(B^2e^{-ct}-A^2)s}{B^2-A^2e^{-ct}+B^2(e^{-ct}-1)s} \\ &= \frac{A^2(1-e^{-ct})}{B^2-A^2e^{-ct}}\cdot\frac{1}{1-\frac{B^2(1-e^{-ct})}{B^2-A^2e^{-ct}}s} + \frac{B^2e^{-ct}-A^2}{B^2-A^2e^{-ct}}\cdot\frac{s}{1-\frac{B^2(1-e^{-ct})}{B^2-A^2e^{-ct}}s}. \end{aligned}$$

Let $w_1 = \frac{A^2(1-e^{-ct})}{(B-A)e^{-ct}}$, $w_2 = \frac{B^2e^{-ct}-A^2}{(B-A)e^{-ct}}$ and $\tilde{p} = \frac{(B-A)e^{-ct}}{B^2-A^2e^{-ct}}$,

$$F(s,t) = w_1\frac{\tilde{p}}{1-s(1-\tilde{p})} + w_2\frac{s\tilde{p}}{1-s(1-\tilde{p})},$$

and $w_1+w_2=1$.

It is easy to see that this is a linear combination of two geometric pgf's with the same parameter $\tilde{p}$, one is supported on the set $\{0,1,\cdots\}$, the other is supported on the set $\{1,2,\cdots\}$. Therefore, $q_{1j}(t)$ can be read as:

$$\begin{aligned} q_{10}(t) &= w_1\tilde{p} = \frac{A^2(1-e^{-ct})}{B^2-A^2e^{-ct}}, \\ q_{1j}(t) &= w_1\tilde{p}(1-\tilde{p})^j + w_2\tilde{p}(1-\tilde{p})^{j-1} \\ &= \frac{(B-A)^2[B^2(1-e^{-ct})]^{j-1}e^{-ct}}{(B^2-A^2e^{-ct})^{j+1}}, j\geq 1. \end{aligned}$$

Hence,

$$\begin{aligned}
G_0(\lambda) &= \int_0^\infty e^{-\lambda t} q_{10}(t)dt \\
&= \int_0^\infty \frac{A^2(1-e^{-ct})}{(B^2-A^2e^{-ct})} e^{-\lambda t}dt \\
&= \frac{A^2}{cB^2}\int_0^1 \frac{v^{\frac{\lambda}{c}-1}(1-v)}{1-\frac{A^2}{B^2}v}dv \\
&= \frac{A^2}{cB^2}\frac{\Gamma(\frac{\lambda}{c})\Gamma(2)}{\Gamma(2+\frac{\lambda}{c})} F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), \\
G_j(\lambda) &= \int_0^\infty e^{-\lambda t} q_{1j}(t)dt \\
&= \int_0^\infty \frac{(B-A)^2[B^2(1-e^{-ct})]^{j-1}}{(B^2-A^2e^{-ct})^{j+1}} e^{-(\lambda+c)t}dt \\
&= \frac{(B-A)^2}{cB^4}\int_0^1 \frac{v^{\frac{\lambda}{c}}(1-v)^{j-1}}{\left(1-\frac{A^2}{B^2}v\right)^{j+1}}dv \\
&= \frac{(B-A)^2}{cB^4}\frac{\Gamma(1+\frac{\lambda}{c})\Gamma(j)}{\Gamma(j+1+\frac{\lambda}{c})} F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), j \geq 1. \quad (4.4)
\end{aligned}$$

We see that the two approaches give the same expression for $G_j(\lambda)$. Hence the limiting frequency spectrum is

$$\psi(j) = \frac{\lambda G_j(\lambda)}{1-\lambda G_0(\lambda)} = \frac{\frac{\lambda}{c}\frac{(B-A)^2}{B^4}\frac{\Gamma(1+\frac{\lambda}{c})\Gamma(j)}{\Gamma(j+1+\frac{\lambda}{c})} F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)}{1-\frac{\lambda}{c}\frac{A^2}{B^2}\frac{\Gamma(\frac{\lambda}{c})\Gamma(2)}{\Gamma(2+\frac{\lambda}{c})} F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)}. \quad (4.5)$$

In the birth-death process, $M = 2B = 2(1-\alpha)(1-\mu)$. By the supercritical condition $M \geq 1$, it is clear that the following constraint is required for the parameters:

$$(1-\alpha)(1-\mu) \geq \frac{1}{2},$$

for $0 \leq \alpha < \frac{1}{2}, 0 < \mu < \frac{1}{2}$. Figure 4.1 shows the domain of parameters α and μ.

Figure 4.2 shows an example of the limiting frequency spectrum for $a = 1, \alpha = 0.25, \mu = 0.01$, based on Equation (4.5). In order to see the frequency spectrum varying with different parameter settings, we plot in Figure 4.3 panel (A) the $\psi(1)$ surface

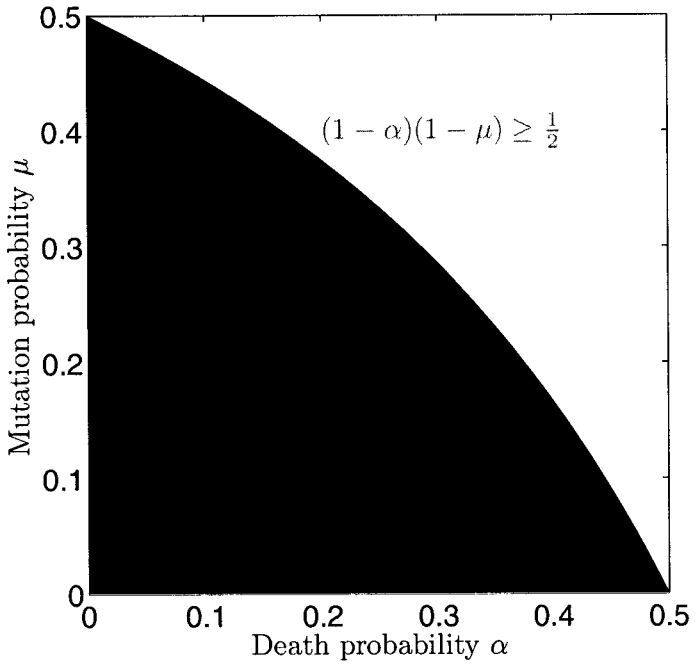


Figure 4.1: Domain of parameters α and μ in the infinite-allele birth-death process.

for α and μ in interval $[0, 0.5]$, and in Figure 4.3 panel (B) the corresponding contour plot. We see that increasing α or μ causes an increase of $\psi(1)$. In particular, the increase of μ has a more significant effect, because in the splitting of each individual, increasing μ will produce more alleles. This then tends to increase the number of alleles in the single copy ($j = 1$) class.

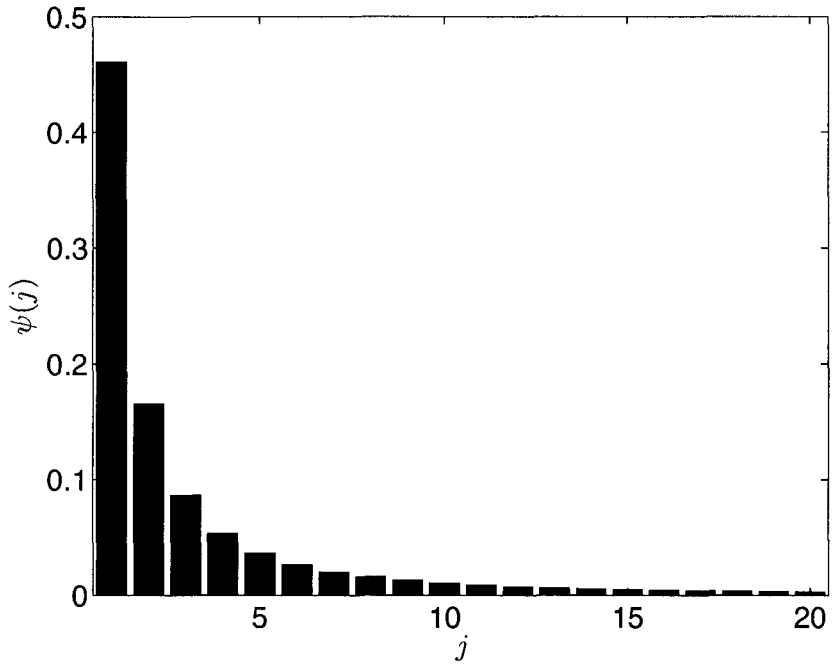


Figure 4.2: Limiting frequency spectrum of an infinite-allele birth-death process, $a = 1, \alpha = 0.25, \mu = 0.01$.

Remark 4.2.1. *Yule process is a special case of the birth-death process with* $\alpha = 0, \beta = 1, A = \mu$ *and* $B = 1 - \mu$. *For another birth-death process, where the offspring distribution is specified as* $f(s) = \alpha + \beta\mu + \beta(1-\mu)s^2$, *i.e., the linear term is removed, we replace* c *with* $B - A$. *There may be some other processes with explicit pgf's for*

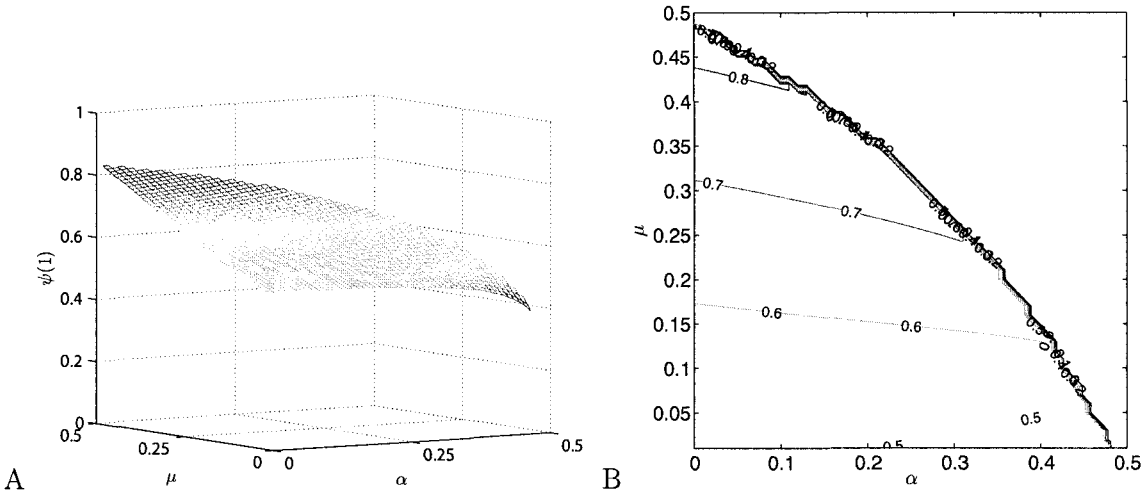


Figure 4.3: Surface of $\psi(1)$ at different α and μ. (A) 3D plot; (B) contour plot.

which we can derive frequency spectra. For example, in the Yule case, if the split of each individual produces k offspring ($k > 2$), i.e., $f(s) = s^k$, then the process pgf is given by $F(s,t) = \frac{e^{-at}s}{\left[1-\left(1-e^{-a(k-1)t}\right)s^{k-1}\right]^{k-1}}$. *Another example is when* $f(s) = 1-\sqrt{1-s}$, *which gives the process pgf* $F(s,t) = 1 - \left[1 - e^{-\frac{t}{2}} + e^{-\frac{t}{2}}(\sqrt{1-s})\right]^2$.

4.2.2 Tail Property and Numerical Solution

In his paper, Pakes [32] also obtained property for the tail of the frequency spectrum. Let $\phi(j) = am\mu G_j(\lambda)$, it has been shown that under the conditions: $M > 1$ and for some $\epsilon > 0, \sum p_j j^{\nu+\epsilon} < \infty$,

$$\lim_{j\to\infty} j^{1+\nu}\phi(j) = \frac{m\mu}{m-1}\mu_\nu,$$

where $\nu = \frac{m-1}{M-1}$ and $\mu_\nu = E(W^\nu)$, W stands for the limit of the normalized process, i.e., $e^{-a(M-1)t}X_t \overset{a.s.}{\to} W$. Figure 4.4 illustrates this property for the birth-death process, for $a = 1, \alpha = 0.25, \mu = 0.01$.

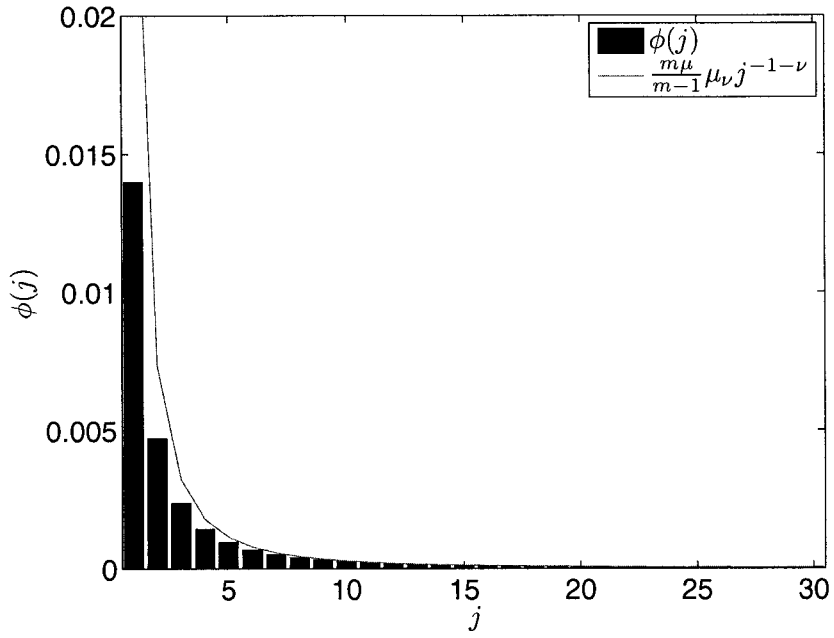


Figure 4.4: Tail behavior of the limiting frequency spectrum.

The limiting frequency spectrum is theoretically useful, however in practice, the frequency spectrum can only be observed in finite time. Therefore in many cases, we need to find $\int_0^t e^{-\lambda x} q_{1j}(x)dx$ instead of $\int_0^\infty e^{-\lambda x} q_{1j}(x)dx$. Usually this can only be accomplished numerically. Let us use the birth-death process as an example. Applying the same technique, we see that

$$\int_0^t e^{-\lambda x} q_{1j}(x)dx = \frac{(B-A)^2}{cB^4} \int_{e^{-ct}}^1 \frac{v^{\frac{\lambda}{c}}(1-v)^{j-1}}{\left(1-\frac{A^2}{B^2}v\right)^{j+1}} dv. \tag{4.6}$$

So we need to numerically compute this "incomplete" hypergeometric function. Figure 4.5 illustrates the difference between this numerical solution $\tilde{\psi}(j)$ at different t's and its asymptotic form $\psi(j)$ based on Equation (4.5) for $a = 1, \alpha = 0.25, \mu = 0.01$. Panel (A) shows the frequencies for the first 10 classes $1 \leq j \leq 10$. The spectra tails ($21 \leq j \leq 30$) are shown in Panel (B) on a different scale. We see that when $t = 17$ (corresponding approximately to 17 generations since $a = 1$), the numerically obtained spectrum is almost identical to the limiting frequency spectrum, but when $t = 10$, there exist some differences between the limiting frequency spectrum and the numerically obtained spectrum. This provides us some intuitions to answer the question: in order to safely use the limiting frequency spectrum, how long should the process history be?

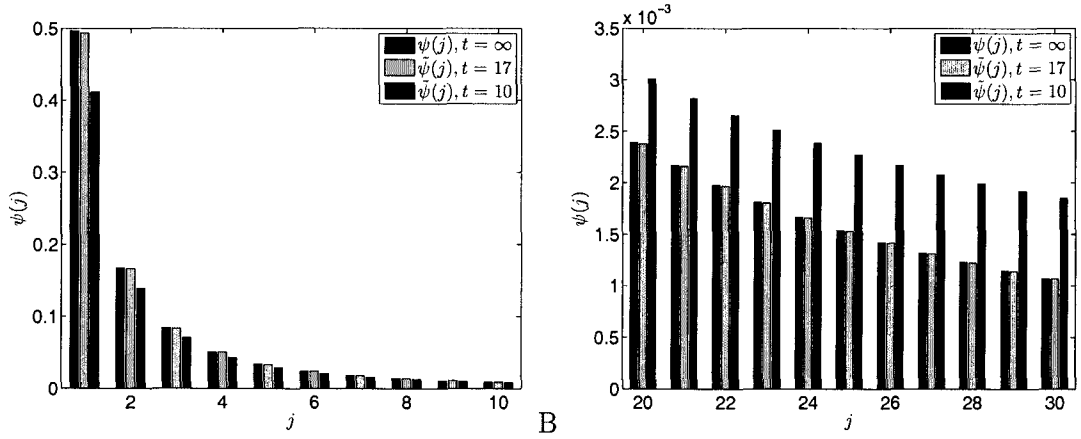


Figure 4.5: Frequency spectrum at time t and limiting frequency spectrum.

4.2.3 Simulation Study of The Frequency Spectrum

To illustrate the results of the frequency spectrum empirically, we perform a simulation study. The programming is done in Matlab, using the same self-recurrent technique as in Section 2.2.3 and Section 3.2.2. The only difference is that in this simulation, the number of alleles is not fixed. Each new allele will be labeled at its appearance, and this information needs to be stored for frequency spectrum calculation in the end. We first generate a Markov birth-death branching process starting from 1 individual, and set the life span parameter $a = 1$, the death probability $\alpha = 0.1$. Every new-born individual has probability $\mu = 0.4$ to mutate into a novel type. We simulate the process for 20 times, record the number of alleles represented by j individuals, $j = 1, 2, \cdots$, and finally obtain the simulated frequency spectrum.

Figure 4.6 shows a side-by-side bar plot of the simulated and numerically obtained $(t = 4)$ frequency spectrum. Based on Equation (4.1), the expected number of alleles in each class, and the observed number of alleles in each class in our simulation, we then perform a χ^2 goodness-of-fit test. This test gives a χ^2 statistic of 3.0903, which leads to a p-value of 0.6861. This shows that the theoretic result of the frequency spectrum of infinite-allele birth-death process is consistent with the simulated data. Note that since the expected spectrum has infinite many classes, we need to truncate the number of classes according to the observed spectrum.

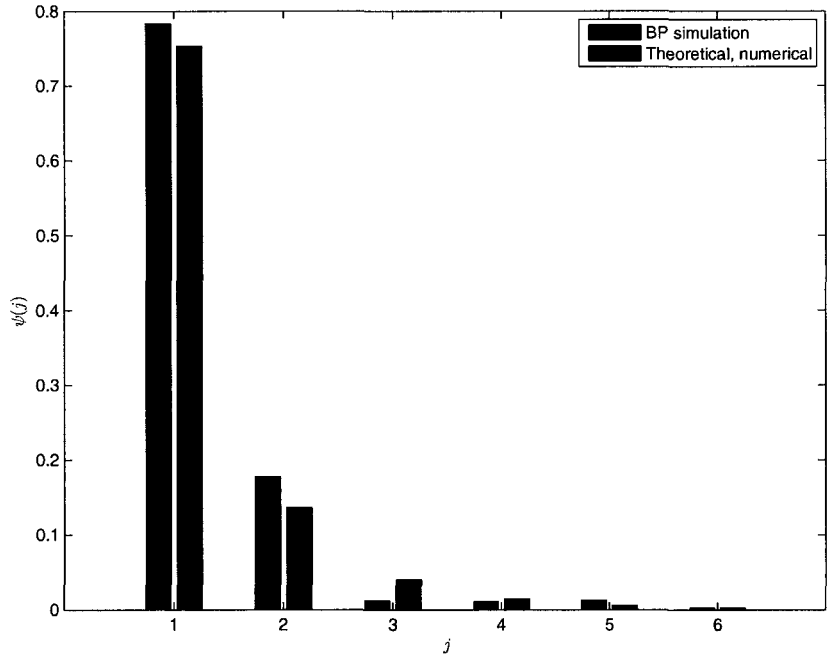


Figure 4.6: Comparison of the simulated and numerically obtained ($t = 4$) frequency spectrum.

4.2.4 Comparison to The Discrete Time Linear-Fractional Process

In previous sections, we are concerned only about the continuous time MBP model. The reason is that results for the continuous time MBP are analogous to those for the discrete time Galton-Watson branching process, and in many cases the MBP results are more complete and the continuous time MBP is more suitable to model genetic processes. However, it will be more instructive to compare the continuous time infinite-allele MBP to the discrete time infinite-allele MBP. We introduce briefly the discrete linear-fractional process. For more details about the Galton-Watson branching process model, readers are referred to the thesis work by Mathaes [27].

Using similar technique as in Section 4.1, Pakes has shown that in the Galton-Watson branching process, the number of alleles which are represented by j individuals at generation n is:

$$\phi_i^{(n)}(j) = q_{ij}^{(n)} + i\mu \sum_{r=0}^{n-1} m^{n-r} q_{1j}^{(r)}.$$

Suppose that

$$G_j = \sum_{r=0}^{\infty} m^{n-r} q_{1j}^{(r)}, j \geq 0.$$

The limiting frequency spectrum

$$\psi(j) = \frac{G_j}{\sum_{j\geq 1} G_j} = \frac{G_j}{\sum_{r=0}^{\infty} m^{-r} - G_0} = \frac{G_j}{\frac{1}{1-m^{-1}} - G_0}, j \geq 1.$$

The discrete linear-fractional process is the only non-trivial example in Galton-Watson branching process for which the n-th iterative generating function can be

computed explicitly. For the single-type linear-fractional process, the offspring pgf $f(s) = 1 - \frac{b}{1-p} + \frac{bs}{1-ps}$. Changing s to $\mu + (1-\mu)s$ gives (one form of) the offspring pgf w.r.t the parental allele:

$$f(s) = 1 - \frac{b}{1-p} + \frac{b[\mu + (1-\mu)s]}{1 - p[\mu + (1-\mu)s]} = 1 - \frac{\tilde{b}}{1-\tilde{p}} + \frac{\tilde{b}s}{1-\tilde{p}s},$$

where $\tilde{b} = \frac{b(1-\mu)}{(1-p\mu)^2}$ and $\tilde{p} = \frac{p(1-\mu)}{1-p\mu}$.

It is known that the n-th iterative generating function for the non-mutant linear fractional process is (for details, please see [3])

$$f_n(s) = 1 - m^n \left(\frac{1-s_0}{m^n - s_0}\right) + \frac{m^n \left(\frac{1-s_0}{m^n-s_0}\right)^2 s}{1 - \left(\frac{m^n-1}{m^n-s_0}\right) s},$$

where $s_0 = \frac{1}{p}\left(1 - \frac{b}{1-p}\right)$ is one of the two roots of equation $f(s) = s$ (the other root is 1 and for supercritical case, $s_0 < 1$). Therefore,

$$\begin{aligned} q_{10}^{(n)} &= 1 - \tilde{m}^n \left(\frac{1-\tilde{s}_0}{\tilde{m}^n - \tilde{s}_0}\right), \\ q_{1j}^{(n)} &= \tilde{m}^n \left(\frac{1-\tilde{s}_0}{\tilde{m}^n - \tilde{s}_0}\right)^2 \left(\frac{\tilde{m}^n - 1}{\tilde{m}^n - \tilde{s}_0}\right)^{j-1}, \end{aligned}$$

where $\tilde{m} = f'(1) = \frac{\tilde{b}}{(1-\tilde{p})^2}$, $\tilde{s}_0 = \frac{1}{\tilde{p}}\left(1 - \frac{\tilde{b}}{1-\tilde{p}}\right)$.

Finally, the limiting frequency spectrum has the form:

$$\begin{aligned} \psi(j) &= \frac{G_j}{\frac{1}{1-\tilde{m}^{-1}} - G_0} \\ &= \frac{\sum_{r=0}^{\infty} m^{n-r} q_{1j}^{(r)}}{\frac{1}{1-\tilde{m}^{-1}} - \sum_{r=0}^{\infty} m^{n-r} q_{10}^{(r)}} \\ &= (1-\tilde{s}_0) \frac{\sum_{r=0}^{\infty} \frac{(\tilde{m}^r - 1)^{j-1}}{(\tilde{m}^r - \tilde{s}_0)^{j+1}}}{\sum_{r=0}^{\infty} \frac{1}{\tilde{m}^r - \tilde{s}_0}}. \end{aligned} \tag{4.7}$$

Remark 4.2.2. *We note that in the discrete linear-fractional case,* $M = \frac{\tilde{b}}{(1-\tilde{p})^2}$. *By the supercritical condition* $M \geq 1$, *it is clear that the following constraint is required for the parameters:*

$$b(1-\mu) \geq (1-p)^2.$$

To illustrate the similarity between the limiting frequency spectrum (4.5) of the continuous time birth-death case and the limiting frequency spectrum (4.7) of the discrete time linear-fractional case, we plot a side by side bar plot in Figure 4.7, where for the linear-fractional case $b = 0.66, p = 0.3, \mu = 0.01$, and for the birth-death case $\alpha = 0.05, \mu = 0.016, a = 1$. We see that under these parameter settings, the two spectra differ mainly in the first three classes.

4.3 Estimation of Mutation/Death Probability

Given the limiting frequency spectrum, the parameters of the infinite-allele Markov branching process can be estimated using the method of moments. The idea is to equate the empirical mean frequency spectrum to the theoretical expected limiting frequency spectrum, and solve for the process parameters. For the birth-death model, we see that no explicit solutions exist, but numerical search is always applicable. As an example, we assume that $a = 1$, set the parameter true values $\alpha = 0.38, \mu = 10^{-4}$, and obtain the limiting frequency spectrum based on Equation (4.5). Using this spectrum as our data, we then search in the two dimensional parameter space for the best

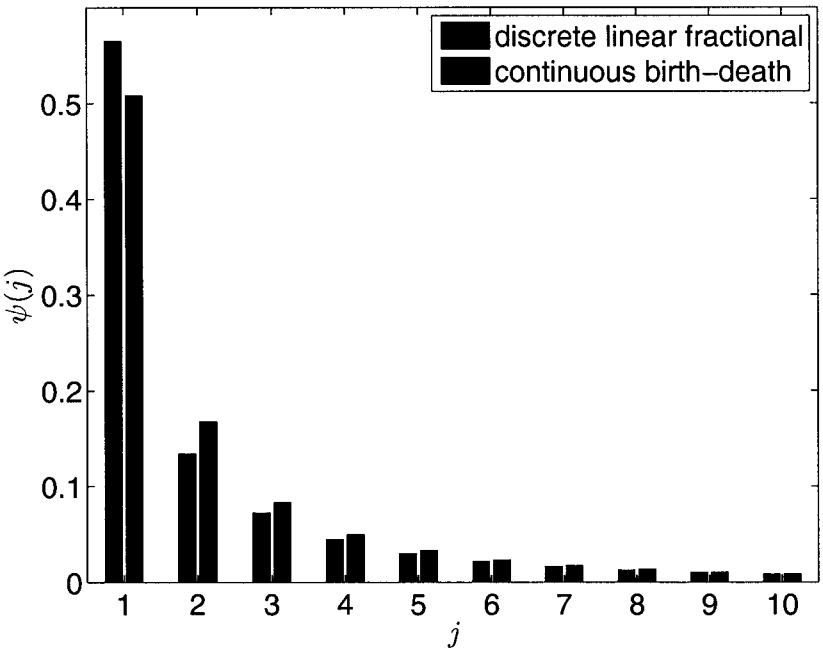


Figure 4.7: Comparison of the limiting frequency spectrum from continuous time birth-death process and the limiting frequency spectrum from discrete time linear-fractional process.

estimates of α and μ which minimizes the distance (square of l_2 norm) between the data spectrum and the candidate spectrum. That is,

$$\{\hat{\alpha}, \hat{\mu}\} = \underset{\alpha,\mu}{\operatorname{argmin}} ||\psi(j) - \hat{\psi}(j; \alpha, \mu)||^2.$$

Figure 4.8 shows the distance surface, where the searching is on a fine grid of $[0, 0.49] \times [10^{-5}, 2 \times 10^{-4}]$. It is clear that the algorithm find the minimum of the distance surface, which is exactly the true value of the parameters.

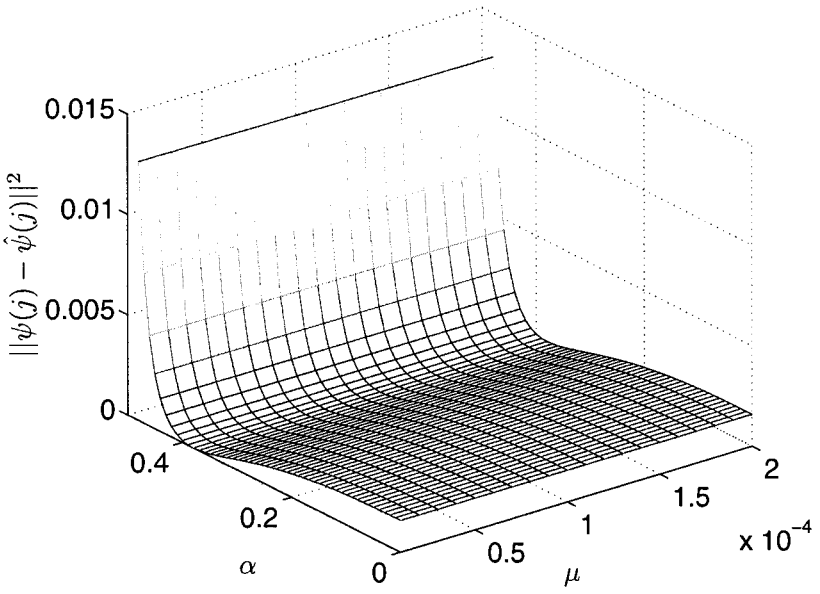


Figure 4.8: Distance surface between the true frequency spectrum and the fitted limiting frequency spectrum.

4.4 Discussion and Application

In practice, if the allele frequency spectrum of a genetic process is available, we can use the method introduced in the previous section to estimate the mutation and death probability in the evolution of the process. Based on the estimated parameters, we may check the goodness-of-fit of the infinite-allele birth-death process to the real frequency spectrum by a χ^2 test.

The χ^2 statistic is a sum of differences between observed and expected outcome frequencies, each squared and divided by the expectation:

$$\chi^2 = \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j},$$

where O_j represents the observed frequency for the j-th class, and E_j represents the expected frequency for the j-th class based on the infinite-allele birth-death model. Under the null hypothesis, i.e., the data come from the model, this test statistic follows a χ^2 distribution with degree of freedom $n-1$, where n is the total number of classes in the observed data.

The χ^2 goodness-of-fit test can be used to test if the observed spectrum came from the infinite-allele MBP model under neutral selection, however we should note that there may be some restrictions of applying this test, such as sufficient sample size. We also note that in our problem, the expected limiting frequency spectrum based on the estimated parameters has infinite many classes, therefore is not comparable to the observed spectrum which is of finite-length. A practical way to apply the χ^2 goodness-of-fit test is to use the counts of alleles in finite time numerically obtained

by Equation (4.6), and truncate the number of classes according to the observed frequency spectrum. As an example, we try to model the evolution of Alu sequence using the infinite-allele MBP model.

An Alu sequence is a short stretch (about 300 base pairs long) of DNA originally characterized by the action of the Alu restriction endonuclease. They are the most abundant (about eleven percent) mobile elements in the human genome, and are still growing in copy numbers [34]. It is known that Alu insertions have been implicated in several inherited human diseases, including various forms of cancer. Therefore there is a need for modeling the amplification, mutation and selection of Alu sequences.

Alu sequences amplify by retrotransposition, also known as "copy and paste" mechanism. Although the exact mechanisms for their retrotransposition are still not fully understood, researchers generally agree that Alu elements are non-autonomous and have to borrow the tools for retrotransposition from the L1 elements. The L1 endonuclease causes a nick at the $TTAAAA$ consensus site, after which Alu anneals directly to the site of integration [19]. A second nick on the other strand completes the insertion. These two staggered nicks introduce an identifiable characteristic of Alu elements. The newly inserted Alu element is surrounded by an identical set of direct repeats, which are also called target site duplications (TSDs). These direct repeats can range anywhere from 10 to 15 base pairs and are considered the prevalent feature of retrotranspositional insertion [6]. This process of integration, also known as target-primed reverse transcription (TPRT) [30], is responsible for the successful

amplification of Alu elements.

Based on certain diagnostic mutations, Alu elements are divided into different subfamilies. The three major families for Alu sequences are J, S and Y. The letters were chosen in alphabetical order to convey the different ages of each family. Alu sequences in the J family are the oldest, while Alu sequences in the Y family are the youngest family. Within each family, further distinctions are made based upon additional diagnostic mutations. For example, subfamily Ya5, indicates that Alu sequences from this subfamily differ from the consensus sequence of family Y by five diagnostic mutations.

We obtained a set of Alu sequence data from Dr. Jerzy Jurka, with nine different Alu subfamilies extracted from the March 2006 assembly of the USCS Human Genome database. These Alu subfamilies include: AluYa1, AluYa4, AluYa5, AluYa8, AluYb8, AluYc1, AluYd2, AluYe2, and AluYe5. For each subfamily, a consensus or reference Alu sequence was used to screen the entire human genome for matching sequences. A match occurred when stretches of nucleotides that include the main diagnostics mutations agreed with the Alu subfamily consensus sequences. After preprocessing, such as deleting poly-A tails and middle-A stretch, the average length of Alu sequences is about 260 base pairs. The number of Alu classes with the same copies is then counted, see Table (4.1).

Table 4.1: Spectrum of the Alu Data.

Alu	1	2	3	4	5	6	7	8	9	10	Percentage
Ya1	3761 (99.1%)	25 (0.7%)	2 (0.1%)	4 (0.1%)	1	1					100%
Ya4	426 (97%)	6 (1.4%)	2 (0.5%)	2 (0.5%)	1 (0.2%)						99.5%
Ya5	1713 (91.2%)	77 (4.1%)	16 (0.9%)	16 (0.9%)	11 (0.6%)	10 (0.5%)	5 (0.3%)	4 (0.2%)	2 (0.1%)	5 (0.3%)	99.9%
Ya8	28 (87.5%)	3 (9.4%)									96.9%
Yb8	1480 (91.4%)	72 (4.5%)	11 (0.7%)	15 (0.9%)	10 (0.6%)	11 (0.7%)	4 (0.3%)	4 (0.3%)	2 (0.1%)		99.4%
Yc1	3162 (98.1%)	42 (1.3%)	9 (0.3%)	4 (0.1%)	1	1	1	0	1		99.9%
Yd2	401 (99.7%)	1 (0.3%)									100%
Ye2	1130 (99.6%)	3 (0.3%)	1 (0.1%)								100%
Ye5	853 (97.6%)	10 (1.1%)	7 (0.8%)	2 (0.2%)	1 (0.1%)						99.9%
All	12952 (96.7%)	240 (1.8%)	48 (0.4%)	43 (0.3%)	25 (0.2%)	23 (0.2%)	10 (0.1%)	8 (0.1%)	6	5	99.8%

These counts can be used as the empirical frequency spectrum for analysis. We see that in Table (4.1), the frequencies of the first class in all Alu subfamilies are very high ($> 90\%$). However, we notice that in an infinite-allele birth-death process, the highest frequency of the first class is below 86%, as seen in Figure 4.3. Therefore, it seems not very appropriate to use the infinite-allele MBP model with neutral selectionto fit this Alu sequence dataset. One possible reason of the departure from the theoretical frequency spectrum may due to the non-neutral selection of the Alu evolution process. However, we also notice that the theory of infinite-allele MBP has a limitation, which may cause the lack of fit.

From Figure 4.3, we see that the first class reaches its highest frequency when death probability $\alpha \approx 0$ and mutation probability $\mu \approx \frac{1}{2}$, which corresponds to the case that the non-mutant process is near-critical. It is natural to make a conjecture that the first class frequency keeps increasing when the mean of the non-mutant process decreases. However, in current literature concerning the theory of infinite-allele MBP, the supercritical condition is required for the non-mutant process, i.e., $M \geq 1$. Therefore, an extension of the current supercritical case to subcritical case for the non-mutant process will be interesting and important. This question is left open for our future study.

Appendix A

Derivation of the modified median estimator

Suppose a fluctuation experiment has n parallel cultures, each starting from a single wild-type cell, and by the plating time t, the ith culture has size $N_{ti}, i = 1, ..., n$. Random variable $K_i, i = 1, ..., n$ represents the number of mutants in the ith culture at time t. Given N_{ti} and the underlying mutation rate μ, K_i has the Luria-Delbrück distribution defined in our Equation (1) in the paper.

Let k_i be the observed number of mutant cells in the ith culture. Our purpose is to estimate the mutation rate based on k_i and $N_{ti}, i = 1, ..., n$. Let $\hat{\mu}_i$ denote the median estimator of μ based purely on the ith culture, that is, $\hat{\mu}_i$ is a function of k_i and N_{ti} only. Mathematically, we have

$$\hat{\mu}_i : P(K_i \leq k_i | N_{ti}, \hat{\mu}_i) = \frac{1}{2}. \tag{A.1}$$

The modified median estimator is defined as the median of $\hat{\mu}_i, i = 1, ..., n$:

$$\hat{\mu} = median(\hat{\mu}_i). \tag{A.2}$$

To find the confidence intervals for the estimates, we need to derive the cumulative distribution function (cdf) of $\hat{\mu}$ under hypothesis $\mu = \mu_0$,

$$F_{\mu_0}(x) = P(\hat{\mu} \leq x | N_{t1}, \ldots, N_{tn};\ \mu = \mu_0),$$

where the dependence on $N_{t1}, \ldots, N_{tn}$ will be dropped for notational convenience. To find the cdf of $\hat{\mu}$, first we compute the cdf of $\hat{\mu}_i$ under hypothesis $\mu = \mu_0$,

$$F^{(i)}_{\mu_0}(x) = P(\hat{\mu}_i \leq x | N_{ti}, \mu = \mu_0).$$

From (A.1) it follows that $\hat{\mu}_i$ is a function of random variable K_i. Since K_i is integer-valued, $\hat{\mu}_i$ takes values from a discrete denumerable set with elements x_k such that

$$\{\hat{\mu}_i = x_k\} \iff \{K_i = k\}.$$

Because of strict monotonicity of $P(K_i \leq k_i | N_{ti}, \mu)$ with respect to discrete k_i and continuous μ, condition (A.1) implies that x_k is strongly increasing in k. Therefore $P(\hat{\mu}_i \leq x_k | N_{ti}, \mu = \mu_0) = P(K_i \leq k | N_{ti}, \mu = \mu_0)$. Therefore, $F^{(i)}_{\mu_0}(x)$ can be defined using the LD distribution for all $x = x_k$ and then extended to all $x \in \mathbb{R}_+$, as a piecewise constant, right continuous function.

Now that we know the distribution of $\hat{\mu}_i, i = 1, ..., n$, the distribution of their median, which is the median of independent, though not identical distributed random

variables, can be obtained. Denote the set $[n] = \{1, 2, ..., n\}$. Also, denote by $\mathcal{J}$ a subset of $[n]$, and by $|\mathcal{J}|$ the number of elements in $\mathcal{J}$. We have

$$\begin{aligned} F_{\mu_0}(x) &= P(\hat{\mu} \leq x | \mu = \mu_0) \\ &= \sum_{\mathcal{J} \subset [n], |\mathcal{J}| \geq \lceil \frac{n}{2} \rceil} \prod_{i \in \mathcal{J}} P(\hat{\mu}_i \leq x | \mu = \mu_0) \prod_{i \in [n] \setminus \mathcal{J}} P(\hat{\mu}_i > x | \mu = \mu_0) \\ &= \sum_{\mathcal{J} \subset [n], |\mathcal{J}| \geq \lceil \frac{n}{2} \rceil} \prod_{i \in \mathcal{J}} F_{\mu_0}^{(i)}(x) \prod_{i \in [n] \setminus \mathcal{J}} (1 - F_{\mu_0}^{(i)}(x)) \end{aligned} \tag{A.3}$$

The principle on which this expression is based, is that the median of a sample is less than or equal to x, if and only if at least half of the elements of the sample are less than or equal to x. Accordingly, the summation of probabilities in formula (A.3) is extended over all subsamples $\mathcal{J} \subset [n]$ satisfying $|\mathcal{J}| \geq \lceil \frac{n}{2} \rceil$, where $\lceil a \rceil$ is the ceiling of a. Replacement of the ceiling function by the floor function does not lead to substantial changes in the numerical computations. Expression (A.3) is a straightforward extension of the result on the distribution of the middle order statistic (the median) in the independent identically distributed case (Theorem 5.4.4 in ref. [18]).

Treating $F_{\mu_0}(x)$ as the distribution under the null hypothesis $\mu = \mu_0$, we use the method of hypothesis test inversion (Theorem 9.2.2 in [5]), to define a confidence interval $[a, b]$, at significance level α, by conditions

$$F_a(\hat{\mu}) = \alpha/2 \text{ and } 1 - F_b(\hat{\mu}) = \alpha/2. \tag{A.4}$$

Using expression (A.3), we solve for a, b such that conditions (A.4) are satisfied. It is always possible since for a given $x \in \mathbb{R}_+$, $F_{\mu_0}(x)$ is a continuous function of μ_0, with

values covering the $(0, 1)$ range. Interval $[a, b]$ is the $(1 - \alpha)$-confidence interval for the mutation rate μ.

Appendix B

Hypergeometric functions

The generalized hypergeometric function, also known as the Barnes extended hypergeometric function, is defined as

$$ {}_pF_q(a_1, \cdots, a_p; b_1, \cdots, b_q; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k \cdots (a_p)_k}{(b_1)_k (b_2)_k \cdots (b_q)_k} \frac{z^k}{k!} $$

where $(a)_k$ is the rising factorial

$$ (a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} = a(a+1)\cdots(a+k-1). $$

As a special case of the generalized hypergeometric function, the Gauss hypergeometric function $F(a, b; c; z)$ corresponds to $p = 2, q = 1$. The following is cited from [1]. Define

$$ F(a, b; c; z) = {}_2F_1(a, b; c; z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^k}{k!}. $$

This series has the circle of convergence $|z| = 1$. The convergence property is:

- Divergence when $\mathcal{R}(c-a-b) \leq -1$.

- Absolute convergence when $\mathcal{R}(c-a-b) > 0$.

- Conditional convergence when $-1 < \mathcal{R}(c-a-b) \leq 0$; the point $z = 1$ is excluded.

$F(a,b;c;z)$ can be written in the following integral representation:

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} dt.$$

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover Publications Inc. New York, 1972.

[2] G. Asteris and S. Sarkar. Bayesian procedures for the estimation of mutation rates from fluctuation experiments. *Genetics*, 142:313–326, 1996.

[3] K. B. Athreya and P. E. Ney. *Branching Processes.* Springer, Berlin, 1972.

[4] J. J. Boesen, M. J. Niericker, N. Dieteren, and J. W. Simons. How variable is a spontaneous mutation rate in cultured mammalian cells? *Mutation Research*, 307:121–129, 1994.

[5] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury, 2001.

[6] R. Cordaux, D. Srikanta, J. Lee, M. Stoneking, and M. A. Batzer. In search of polymorphic alu insertions with restricted geographic distributions. *Genomics*, 90:154–158, 2007.

[7] S. N. Ethier and T. Nagylaki. Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability*, 12:14–49, 1980.

[8] S. N. Ethier and T. Nagylaki. Diffusion approximations of Markov chains with two time scales and applications to population genetics, II. *Advances in Applied Probability*, 20:525–545, 1988.

[9] R. A. Fisher. *The genetical theory of natural selection.* Clarendon Press, Oxford, 1930.

[10] D. J. Futuyma. *Evolutionary Biology.* Sinauer Associates, 1998.

[11] R. C. Griffiths and A. G. Pakes. An infinite-alleles version of the simple branching process. *Advances in Applied Probability*, 20:489–524, 1988.

[12] P. Haccou, P. Jagers, and V. A. Vatutin. *Branching Processes: Variation, Growth, and Extinction of Populations.* Cambridge University Press, 2005.

[13] D. Hartl. *Principles of Population Genetics.* Sinauer Associates, INC, 1980.

[14] P. J. Hastings, H. J. Bull, J. R. Klump, and S. M. Rosenberg. Adaptive amplification: An inducible chromosomal instability mechanism. *Cell*, 103:723–731, 2000.

[15] P. Jagers. *Branching Processes with Biological Applications.* Wiley, New York, 1975.

[16] P. Jagers, F. C. Klebaner, and S. Sagitov. Markovian paths to extinction. *Advances in Applied Probability*, 39:569–587, 2007.

[17] P. Jagers, F. C. Klebaner, and S. Sagitov. On the path to extinction. *PNAS*, 104:6107–6111, 2007.

[18] M. E. Jones, J. Wheldrake, and A. Rogers. Luria-Delbrück fluctuation analysis: Estimating the Poisson parameter in a compound Poisson distribution. *Computers in Biology and Medicine*, 23:525, 1993.

[19] J. Jurka. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *PNAS*, 94:1872–1877, 1997.

[20] M. Kimmel and D. E. Axelrod. Fluctuation test for two-stage mutations: application to gene amplification. *Mutation Research*, 306:45–60, 1994.

[21] M. Kimmel and D. E. Axelrod. *Branching processes in biology.* Springer-Verlag, New York, 2002.

[22] H. L. Klein. Spontaneous chromosome loss in Saccharomyces cerevisiae is suppressed by DNA damage checkpoint functions. *Genetics*, 159:1501–1509, 2001.

[23] D. E. Lea and C. A. Coulson. The distribution of the number of mutants in bacterial populations. *Journal of Genetics*, 49:264, 1949.

[24] M. R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, 1983.

[25] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28:491, 1943.

[26] W. T. Ma, G. Vh. Sandri, and S. Sarkar. Analysis of the Luria-Delbrück distribution using discrete convolution. *Journal of Applied Probability*, 29:255–267, 1992.

[27] M. Mathaes. Statistical tests of neutrality based on snp and alu repeat data. *Ph.D. Thesis, Rice University*, 2009.

[28] T. Nagylaki. *The Gaussian approximation for random genetic drift, in "Evolutionary Process and Theory"*. Academic Press, New York, 1986.

[29] T. Nagylaki. Models and approximations for random genetic drift. *Theoretical Population Biology*, 37:192–212, 1990.

[30] E. M. Ostertag and H. H. Kazazian. Biology of mammalian l1 retrotransposons. *Annual Reviews in Genetics*, 35:501–538, 2001.

[31] A. G. Pakes. Asymptotic results for the extinction time of Markov branching processes allowing emigration. I. Random walk decrements. *Advances in Applied Probability*, 21:243–269, 1989.

[32] A. G. Pakes. An infinite alleles version of the markov branching process. *Journal of Australian Mathematical Society (Series A)*, 46:146–170, 1989.

[33] A. G. Pakes. On the asymptotic behaviour of the extinction time of the simple branching process. *Advances in Applied Probability*, 21:470–472, 1989.

[34] A. M. Roy-Engel, M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen, A. H. Salem, M. A. Batzer, and P. L. Deininger. Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*, 159:279–290, 2001.

[35] S. Sarkar, W. T. Ma, and G. vH. Sandri. On fluctuation analysis: a new simple and efficient method for computing the expected number of mutants. *Genetica*, 85:173, 1992.

[36] B. A. Sevastyanov. *Verzweigungsprozesse.* R. Oldenbourg Verlag Munchen Wien, 1975.

[37] F. M. Stewart. Fluctuation tests: How reliable are the estimates of mutation rates? *Genetics*, 137:1139, 1994.

[38] E. D. Strome, X. Wu, M. Kimmel, and S. E. Plon. Heterozygous screen in saccharomyces cerevisiae identifies dosage-sensitive genes that affect chromosome stability. *Genetics*, 178:1193–1207, 2008.

[39] S. K. Waghmare and C. V. Bruschi. Differential chromosome control of ploidy in the yeast Saccharomyces cerevisie. *Yeast*, 22:625–639, 2005.

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

[41] X. Wu, E. D. Strome, Q. Meng, P. J. Hastings, S. E. Plon, and M. Kimmel. A robust estimator of mutation rates. *Mutation Research*, 661:101–109, 2009.

[42] Y. Zang, M. Garre, K. Gjuracic, and C. V. Bruschi. Chromosome V loss due to centromere knockout or MAD2-deletion is immediately followed by restitution of homozygous diploidy in Saccharomyces cerevisiae. *Yeast*, 19:553–564, 2002.

[43] Q. Zheng. Progress of a half century in the study of the Luria-Delbrück distribution. *Mathematical Biosciences*, 162:1–32, 1999.

[44] Q. Zheng. Statistical and algorithmic methods for fluctuation analysis with SALVADOR as an implementation. *Mathematical Biosciences*, 176:237–252, 2002.

[45] Q. Zheng. New algorithms for Luria-Delbrück fluctuation analysis. *Mathematical Biosciences*, 196:198–214, 2005.