Statistical Methods for Biodiversity Assessment

Ratchaneewan Kumphakarm

September 2016

A thesis submitted for the degree of

Doctor of Philosophy

School of Mathematics, Statistics and Actuarial Science

University of Kent

Abstract

This thesis focuses on statistical methods for estimating the number of species which is a natural index for measuring biodiversity. Both parametric and nonparametric approaches are investigated for this problem. Species abundance models including homogeneous and heterogeneous model are explored for species richness estimation. Two new improvements to the Chao estimator are developed using the Good-Turing coverage formula.

Although the homogeneous abundance model is the simplest model, the species are collected with different probability in practice. This leads to overdispersed data, zero inflation and a heavy tail. The Poisson-Tweedie distribution, a mixed-Poisson distribution including many special cases such as the negative-binomial distribution, Poisson, Poisson inverse Gaussian, Pólya-Aeppli and so on, is explored for estimating the number of species. The weighted linear regression estimator based on the ratio of successive frequencies is applied to data generated from the Poisson-Tweedie distribution. There may be a problem with sparse data which provides zero frequencies for species seen i times. This leads to the weighted linear regression not working. Then, a smoothing technique is considered for improving the performance of the weighted linear regression estimator. Both simulated data and some real data sets are used to study the performance of parametric and nonparametric estimators in this thesis.

Finally, the distribution of the number distinct species found in a sample is hard to compute. Many approximations including the Poisson, normal, COM-Poisson Binomial, Altham's multiplicative and additive-binomial and Pólya distribution are used for approximating the distribution of distinct species. Under various abundance models, Altham's multiplicative-binomial approximation performs well. Building on other recent work, the maximum likelihood and the maximum pseudo-likelihood estimators are applied with Altham's multiplicative-binomial approximation and compared with other estimators.

Acknowledgement

I would like to express my gratitude to my supervisor Professor Martin Ridout, for his patience, support and encouragement of my PhD study. He helped to guide me all the time for my research and writing my thesis. I sincerely appreciate Professor Martin Ridout as my supervisor. I could not have done this successfully without his guidance. My sincere thanks also goes to Dr.Alfred Kume who is my co-supervisor for his helpful advice and proof reading throughout my writing stages. It would not have been possible to complete the thesis without his guidance.

I am extremely grateful to my parents for encouragement all my life even during their illness. I would like to thank Claire Carter and Derek Baldwin for their helps and suggestions during four years of my PhD. I would also like to thank the School of Mathematics, Statistics and Actuarial Science at the University of Kent for providing the facilities for all the research students.

I am extremely grateful to the National Science and Technology Development Agency, Thai Government, for the scholarship to do my PhD in United Kingdom. I would like to recognise my workplace, Maejo University, Thailand, for encouraging me to undertake the PhD. Finally, I would like to thank my friends, Bill, Kyle, Sam and Tita, for sharing their ideas, knowledge and time for coffee. Koi and Kung, my dear friends, who reminded me to do important things for the PhD.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Biodiversity is a critical feature of an ecosystem. Currently, there are many studies focused on measuring biodiversity. One particular measure is species richness – "the number of species in a community, in a landscape or marinescape, or in a region" (Colwell, 2009). Species richness is one of the primary indicators which measures biodiversity and represents a feature of community ecology (Longino et al., 2002). In addition, estimating the number of species provides significant information for planning conservation and handling natural resources (Boulinier et al., 1998).

Bunge and Fitzpatrick (1993) present a survey of methods for estimating the number of species. There are different sampling models including hypergeometric, Bernoulli, multinomial, Poisson and multiple Bernoulli distribution. Data analytic methods using extrapolation of curves is another approach used to estimate the number of species. The number of observed species is plotted as a function of the number of individuals in the sample and extrapolated to give the number of species as the sample size tends to infinity.

As a result of anthropogenic and environmental changes such as physical, chemical and biological factors, local extinctions of some species occur and new species emerge (El-Shaarawi et al., 2011). Researchers have studied and developed many methods to estimate species richness. The key issue that makes species richness complicated to estimate is that there may be species that escape detection. In addition, each species is likely to have a different level of abundance in the population. Hence, there is a need for appropriate methods that can incorporate these issues.

Although of great interest to ecologists, conservationists and biologists, species richness estimation is fundamentally a statistical problem and has attracted considerable attention from statisticians. Both parametric and nonparametric estimators have been proposed for species richness estimation.

Nonparametric estimators are attractive for this problem because they do not require assumptions about the distribution of the abundance data. Chao (1984) proposed a nonparamatric estimator for estimating the number of species and it is called the Chao1 estimator in this thesis. The Chao1 estimator is used for estimating a lower bound of species richness. It performs well for a homogeneous population or for large sample size. The Chao1 estimator is improved by Chiu et al. (2014) using a modified Good-Turing frequency and called it the iChao1 estimator. The performance in terms of bias and mean square error are improved especially in a highly heterogeneous population. Other nonparametric estimators such as Good-Turing, the first-order, the second order jackknife are explored in this thesis.

Alternatively, the maximum likelihood estimation (MLE) is discussed for estimating the unknown parameter. The Poisson distribution can be used for homogeneous abundance data. Due to heterogeneous abundance in prac-

tice, Fisher et al. (1943) considered mixed-Poisson models such as the gamma mixed-Poisson known as the negative binomial distribution for estimating the number of species.

El-Shaarawi et al. (2011) investigated the Poisson-Tweedie (PT) distribution for abundance data, the mixed-Poisson distribution between the Poisson and Tweedie distribution. It includes many special cases such as the Poisson, negative binomial, Poisson-inverse Gaussian, Neyman Type A, Pólya-Aeppli and so on.

Additionally, the zero-truncated mixed-Poisson distribution is another way used to estimate the number of species. Cruyff and van der Heijden (2008) investigated the zero-truncated negative binomial distribution to estimate the population size. Bunge and Barger (2008) investigated the zero-truncated mixed-Poisson distribution including the log-normal mixed-Poisson, the inverse Gaussian mixed-Poisson, the Pareto mixed-Poisson distribution and so on. However, the MLE approach might lead to convergence problems in optimization.

Rocchetti et al. (2011) proposed the weighted linear regression (WLR) estimator based on the ratio of successive counts for heterogeneous model. For small sample size, there might be zero frequencies that cause the WLR approach to fail. Rocchetti et al. (2011) used truncated data in analysis for avoiding this problem. Smoothing data using the kernel estimation is another way to handle this issue. This choice is investigated in this thesis.

Hidaka (2014) introduced another parametric estimator of species richness using maximum pseudo-likelihood estimation. The distribution of observed species is considered under the occupancy distribution. Williamson (2012)

explored some approximations to the occupancy distribution based on the classical occupancy problem including the Poisson and normal distribution.

The question about "How many species are there?" is studied in this thesis. Many species richness estimators, both nonparametric and parametric approach, are explored. In this thesis, alternative species richness estimators under the closed population and various species abundance models are developed and applied to real data sets.

1.2 Real data examples

In this thesis, we select some examples from many fields including ecology, social science and environment. Species abundance data for animal and plant are used to estimate the number of species. Additionally, capture-recapture data is used in this thesis for estimating the population size. We select heroin users data who were treated at health treatment centres to estimate the number of total drug users. Other example about environment is used to compare our approach. In the following tables f_i denotes the number of species seen i times and K denotes the number of distinct species in the sample.

1.2.1 Malaysian butterfly data

Malaysian butterfly data (Fisher et al., 1943) is a large data set collected in Malaysia. It is used in many studies about species richness estimation. The frequencies of the butterflies are observed from 9031 individuals and representing 620 species as shown in Table 1.1.

Table 1.1: Frequency counts for Malaysian Butterfly Data (Fisher et al., 1943)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	118	74	44	24	29	22	20	19	20	15	12	14	6

i	14	15	16	17	18	19	20	21	22	23	24	24+	K
f_i	12	6	9	9	6	10	10	11	5	3	3	119	620

1.2.2 Pollutant data

In Table 1.2, the frequency of occurrence of organic compounds identified in water between 1970 and 1976 is shown. There are 5720 observations which are classified as 1258 organic compounds.

Table 1.2: Frequency counts for Pollutant Data (Janardan and Schaeffer, 1981)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	503	238	133	80	56	46	20	14	15	18	15	16	10

i	14	15	16	17	18	19	20	21	22	23	24	24+	K
f_i	10	9	4	12	6	7	4	4	1	4	0	33	1258

1.2.3 Christmas bird data

These data were collected at Fort Myers in Florida. The number of Christmas bird species has been investigated from this data set classified as 126 species from 20042 individuals (Chao and Bunge, 2002) (Table 1.3).

Table 1.3: Frequency counts for the Christmas bird data at Fort Myers, Florida, USA. (Chao and Bunge, 2002)

i	1	2	3	4	5	6	7	8	9	10	11	15	16
f_i	12	9	6	6	2	2	5	1	2	3	3	1	2

i	17	18	19	20	21	22	25	25+	K
f_i	1	2	1	1	2	1	2	62	126

1.2.4 Heroin users data

In Table 1.4, data that was collected in 2002 by the Office of the Narcotics Control Board in Thailand (Lanumteang and Böhning, 2011) is shown. There are 9302 unique drug users who were treated from a total of 39086 contacts at health treatment centres.

Table 1.4: Frequency counts for the heroin user data in Thailand (Lanumteang and Böhning, 2011)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	2176	1600	1278	976	748	570	455	368	281	254	188	138	99

i	14	15	16	17	18	19	20	21	K
f_i	67	44	34	17	3	3	2	1	9302

1.2.5 Beetle data

The beetle data set is separated into two sites, Osa second growth and Osa old growth, and collected in southwestern Costa Rica (Janzen, 1973). There are 976 individuals collected from 140 species in the Osa second growth site. For the Osa old growth, there are 237 individuals collected from 112 species as shown in Table 1.5.

Table 1.5: Frequency counts for the beetle data collected from two sites in southwestern Costa Rica (Janzen, 1973)

Osa second growth (M=976)

i	1	2	3	4	5	6	7	8	9	10	11	12	14
f_i	70	17	4	5	5	5	5	3	1	2	3	2	2

i	17	19	20	21	24	26	40	57	60	64	71	77	K
f_i	1	2	3	1	1	1	1	2	1	1	1	1	140

Osa old growth (M=237)

i	1	2	3	4	5	6	7	8	14	42	K
f_i	84	10	4	3	5	1	2	1	1	1	112

1.2.6 Tropical trees data

Norden et al. (2009) present the frequencies of tropical trees data from three forest sites in northeastern Costa Rica (Table 1.6). A total of 943 individuals were collected in Lindero EL Peje (LEP) old growth which included 152 distinct species. The tropical trees in the second site collected from LEP second growth which found 106 district species from a total of 1263 individuals. Another site, the data is collected from Lindero sur second growth site which has 76 distinct species found from 1020 individuals.

Table 1.6: Frequency counts for the tropical tree data observed from three forest sites in northeastern Costa Rica (Norden et al., 2009)

LEP old growth (M=943)

i	1	2	3	4	5	6	7	8	9	10	11	13	15
f_i	46	30	16	12	6	6	3	4	5	4	1	3	1

i	16	18	19	20	25	38	39	40	46	52	55	K
f_i	1	1	1	4	3	1	1	1	1	1	1	152

LEP older second growth (M=1263)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
f_i	33	15	13	4	5	3	3	1	2	1	4	2	2

i	14	15	16	17	20	22	39	45	57	72	88	132	133	178	K
f_i	1	2	1	1	1	1	1	1	1	1	2	1	1	1	104

Lindero Sur younger second growth growth (M=1020)

i	1	2	3	4	5	7	8	10	11	12	13	15	31
f_i	29	13	5	2	3	4	1	2	2	1	2	2	1

i	33	34	35	66	72	78	127	131	174	K
f_i	1	1	1	1	1	1	1	1	1	76

1.3 Thesis Structure

This thesis consists of eight chapters including an introduction as Chapter 1, six core chapters and conclusions as the final Chapter. The first Chapter

presents the background of the study, real data examples and thesis structure.

Chapter 2 reviews the literature on species richness estimation. We initially introduce the models of species sample frequency such as the multinomial model and the Poisson model. After that, the distribution of the number of observed species is discussed. Additionally, species abundance models such as the Zipf, Zipf-Mandelbrot, exponential-decay, broken-stick and log-normal models are reviewed. In this chapter, species richness estimation with nonparametric estimators is discussed. Two alternative estimators of species richness are developed and compared with Chao1, iChao1, the first-order and the second-ordered jackknife estimators. We also applied these nonparametric estimators to some real data examples.

Chapter 3 presents maximum likelihood estimation (MLE) for estimating species richness. The mixed-Poisson distribution and the zero-truncated mixed-Poisson distribution are considered for the MLE approach. Several problems about estimating the number of species using the MLE approach are presented including flat likelihood function, boundary problem and so on. For avoiding these problems in MLE, the weighted linear regression (WLR) analysis is investigated in the next Chapter.

Chapter 4 considers the mixed-Poisson distribution such as the Poisson-Tweedie (PT) distribution that exhibits overdispersion, zero inflation and heavy right tails to fit the model for species abundance data. We have focused on the WLR estimator to estimate the number of species based on the PT distribution. The PT distribution and its sub-families is introduced. The probability generating function is used to define the probability mass function of the PT distribution. In a separate section, we discuss the reparametrization of the PT distribution. Additionally, The `tweeDEseq` package in R is used to generate data and com-

pute the probability mass function in a simulation study. The WLR estimator based on the PT distribution is compared to the other estimators both in real and simulated data.

In Chapter 5, we improve the WLR estimator using kernel smoothing. Discrete kernel estimators and bandwidth selection are considered. The frequencies are smoothed using the kernel of Wang and Van Ryzin (1981) and Li and Racine (2010) before estimating the number of species by the WLR estimator. Abundance data are generated from the PT distribution. In addition, the `np` package in R is used for density estimation. In a simulation study, we investigate the performance of the WLR estimator with smoothing method. We then summarize the results of kernel smoothing and compare them with the nonsmoothing method and the Chao1 estimator.

Chapter 6 considers estimating the number of unseen species based on the occupancy distribution. The occupancy distribution and the classical occupancy problem are reviewed. Some approximations such as the Possion, the normal, the COM-Poisson Binomial, Altham's multiplicative and additive binomial and the Pólya distribution are explored. We apply the approximations to the example about birthday coincidences in Feller (1950). Then, we investigate the performance of approximations for both homogeneous and heterogeneous models in the simulation study and conclude the results.

In Chapter 7, the number of species is estimated using the pseudo-likelihood estimation based on the occupancy distribution. The distribution of observed species is considered for constructing the pseudo-likelihood function. The Hidaka (2014) study is extended. The pseudo-likelihood function and some approximations such as the Poisson-binomial and Altham's multiplicative binomial distribution are investigated. Additionally, the least squares estimation

is used to estimate the number of species. Then, we investigate the performance of the pseudo-likelihood and the least square estimator based on various approximations. Under the homogeneous abundance model, these approaches are compared with some nonparametric estimators in simulation study.

In this thesis, the computational work is carried out using R. Conclusion and suggestions for future work are included in the final chapter.

Chapter 2

Species richness estimation

2.1 Introduction

Species richness is a natural index and the simplest indicator for biodiversity assessment (Gotelli and Colwell, 2011; Chao and Jost, 2012). Although of great interest to ecologists, conservationists and biologists, species richness estimation is fundamentally a statistical problem and has attracted considerable attention from statisticians. Both parametric and nonparametric estimators have been proposed for species richness estimation (Chao and Bunge, 2002).

The Chao1 estimator is a very popular nonparametric estimator for species richness estimation, given a random sample from the population. It is approximately unbiased for a homogeneous abundance model. Additionally, the performance of the Chao1 estimator is good for a large sample size but depends on the under lying abundance model, as illustrated by results later in this Chapter. However, it is negatively biased for heterogeneous models or small sample size. A recent paper Chiu et al. (2014) describes a new improved estimator which is called the iChao1 estimator. It attempts to reduce the bias of the original Chao1 estimator by using additional data. In this Chapter, an alternative estimator which is intended to perform similarly to the iChao1

estimator but uses the same data as the original Chao1 estimator is developed and the results are shown later.

In this Chapter, the literature on species richness estimation is reviewed as follows. In Section 2.2, models of species sample frequency are discussed including the multinomial and the Poisson models. Species abundance models particularly the heterogeneous models which are used in practice, are discussed in Section 2.3. Nonparametric estimators of species richness are reviewed in Section 2.4. Two novel alternative species richness estimators designed to improve upon the Chao1 estimator are introduced in Section 2.5. The mean relative abundance of species seen k times is estimated using various approaches and their performance are investigated in Section 2.6. In Section 2.7, the performance of these new estimators is compared with Chao (1984), Chiu et al. (2014) and two jackknife estimators in a simulation study and applied to real data sets in Section 2.8. Finally, conclusions are summarized in Section 2.9.

2.2 Sampling Models

Let N denote the true species richness, the total number of species in the population, and p_i $(i = 1, \ldots, N)$ be the relative species abundance for species i or the probability of species i being observed, $\sum_{i=1}^{N} p_i = 1$. In ecological applications, this will depend on the difficulty of capturing this species as well as the relative abundance of the species, but we use relative abundance as a convenient shorthand term.

The sample size M denotes the number of individuals collected independently with replacement from the population of N species. Suppose that there are K distinct species in the sample. Let X_i denote the frequency with which species i is detected in the sample, so that, $M = \sum_{i=1}^{N} X_i$. When M is fixed,

the X_i's have a multinomial distribution which is often called *the multinomial model*. Alternatively, we may consider *the Poisson model* that arises when M is itself a random variable with a Poisson distribution. In this model, the X_i's are independent Poisson random variables with $X_i \sim Poi(\lambda p_i) \equiv Poi(\lambda_i)$ and then $M \sim Poi(\sum \lambda p_i) \equiv Poi(\lambda)$. In the multinomial model, the X_i's are not independent because they add up to the fixed total M. Note also that in the multinomial model, the marginal distribution of a particular X_i is $X_i \sim Bin(M, p_i)$. Another connection between the two models is that if M is large and p_i is small then the binomial distribution of X_i will be approximated well by a Poisson distribution with the same mean.

Let f_k be the frequency of species seen k times, $k = 1, 2, \ldots, M$. We have,

$$K = \sum_{k=1}^{k_{\max}} f_k = \sum_{i=1}^{N} I(X_i > 0),$$

where k_{max} is the maximum number of times that any species is seen and $I(X_i > 0) = 1$ if the event $X_i > 0$ occurs (species i occurs in the sample) and 0 otherwise. The total number of species can be written as

$$N = \mathbb{E}(K) + \mathbb{E}(f_0), \tag{2.1}$$

which is a common idea for species richness estimation, where $\mathbb{E}(K)$ is the expected number of observed species and $\mathbb{E}(f_0)$ is the expected number of unobserved species. $\mathbb{E}(K)$ can be estimated by the number of seen species, K,from the data.

2.3 Species abundance model

Species abundance is a simple method to describe biodiversity. Different ecology influences the abundance of a species (Huang and Zhan, 2014). The com-

monness and rarity of species have been described using species abundance models (McGill et al., 2007). The homogeneous model $p_i = \frac{1}{N}$ $(i = 1, \ldots, N)$ is the simplest model to fit the abundance data. However, the chances of collecting different species are typically far from equal in practice. Species abundance data normally exhibit overdispersion, zero inflation and a heavy right tail. These features indicate that a heterogeneous model is required rather than the homogeneous model.

Many models such as negative binomial, log-series, log-normal distributions, broken-stick, Zipf, Zipf-Mandelbrot models and so on have been developed to fit the species abundance data by ecologists (Huang and Zhan, 2014). The p_i can be defined as a function of different abundance model, $p_i = f(i)$. When the X_i's follow the Poisson distribution, the model of p_i's are discussed into two groups as follows:

2.3.1 Deterministic models

Deterministic models are used to describe the rank-ordered probabilities (ie. $p_1 \geq p_2 \geq \ldots \geq p_N$) and include the Zipf, Zipf-Mandelbrot, exponential decay and power decay (a special case of Zipf) models.

The Zipf model

The Zipf model describes the relative abundance rank of the N species. The Zipf model is a discrete probability distribution which is used to model the species abundance distribution and is based on Zipf's law. It is also known as the power-decay model (Chao et al., 2013). The relative abundance of the i^{th} ranked species based on the Zipf model is given by

$$p_i = \frac{c}{i^{\alpha}} \qquad (i = 1, \ldots, N),$$

where c is the normalising constant, $c = \sum_{i=1}^{N}(1/i^{\alpha})$, and $\alpha \geq 1$. When $\alpha = 0$, it gives the homogeneous model, $p_i = 1/N$ (Chao and Chiu, 2014).

The Zipf-Mandelbrot model

The Zipf-Mandelbrot model is another model of ranked abundance, which can be defined by

$$p_i = \frac{c}{(i+q)^{\alpha}}, \qquad (i = 1, \ldots, N),$$

where $q > -1$, $\alpha \geq 1$ and c is a normalising constant, $c = \sum_{i=1}^{N}(1/(i+q)^{\alpha})$ (Mouillot and Lepretre, 2000). When $q = 0$, it reduces to the Zipf model.

Exponential-decay model

The exponential decay model has

$$p_i = c\, e^{-\beta i} \qquad (i = 1, \ldots, N),$$

where β is the decay rate parameter, $\beta > 0$, i is ranked abundance and c is the normalising constant.

2.3.2 Random models

In random models, the p_i are drawn as a random sample from some probability distribution. The resulting p_i values are not ordered.

Broken-stick model

A natural distribution to choose is the Dirichlet distribution, since this automatically gives $\sum_{i=1}^{N} p_i = 1$. The general form of the Dirichlet distribution has parameters $\theta_1, \ldots, \theta_N$ and is generated as

$$p_i = \frac{Z_i}{\sum_{i=1}^{N} Z_i}$$

where Z_i are independent $Ga(\theta_i, 1)$ random variables. The broken stick model has all $\theta_i = 1$ so that the Z_i are independent exp(1) variables.
Therefore, the broken-stick model describes the pattern of species abundance which is given by

$$p_i = cZ_i \qquad (i = 1, \ldots, N),$$

where $(Z_1, Z_2, \ldots, Z_N)$ are a random sample from the exponential distribution with mean 1, and c is the normalising constant (Chao et al., 2013).

Log-normal model

The log-normal model is another distribution used widely for species abundance, and is given by

$$p_i = cV_i \qquad (i = 1, \ldots, N),$$

where $(V_1, V_2, \ldots, V_N)$ are a random sample from the log-normal distribution with parameters, μ and σ, and c is the normalising constant. In the study of Chao et al. (2013), species abundances are simulated using this model with parameters $\mu = 0$ and $\sigma = 1$.

Negative binomial model

Let $U_1, U_2, \ldots, U_N$ are a random sample from the negative binomial distribution with parameter s and r. Then, the species abundance model is given by

$$p_i = cU_i \qquad (i = 1, \ldots, N),$$

where the probability density function of the negative binomial is

$$f(U) = \frac{(U-1)!}{(s-1)!(U-s)!}(1-r)^{U-s}r^s.$$

2.3.3 Some numerical examples about species abundance models

Magurran and Henderson (2011) mention that species abundance data can be presented using a rank abundance plot which is also called a Whittaker plot. The pattern of species abundance is displayed similarly for different models as shown in Figure 2.1.

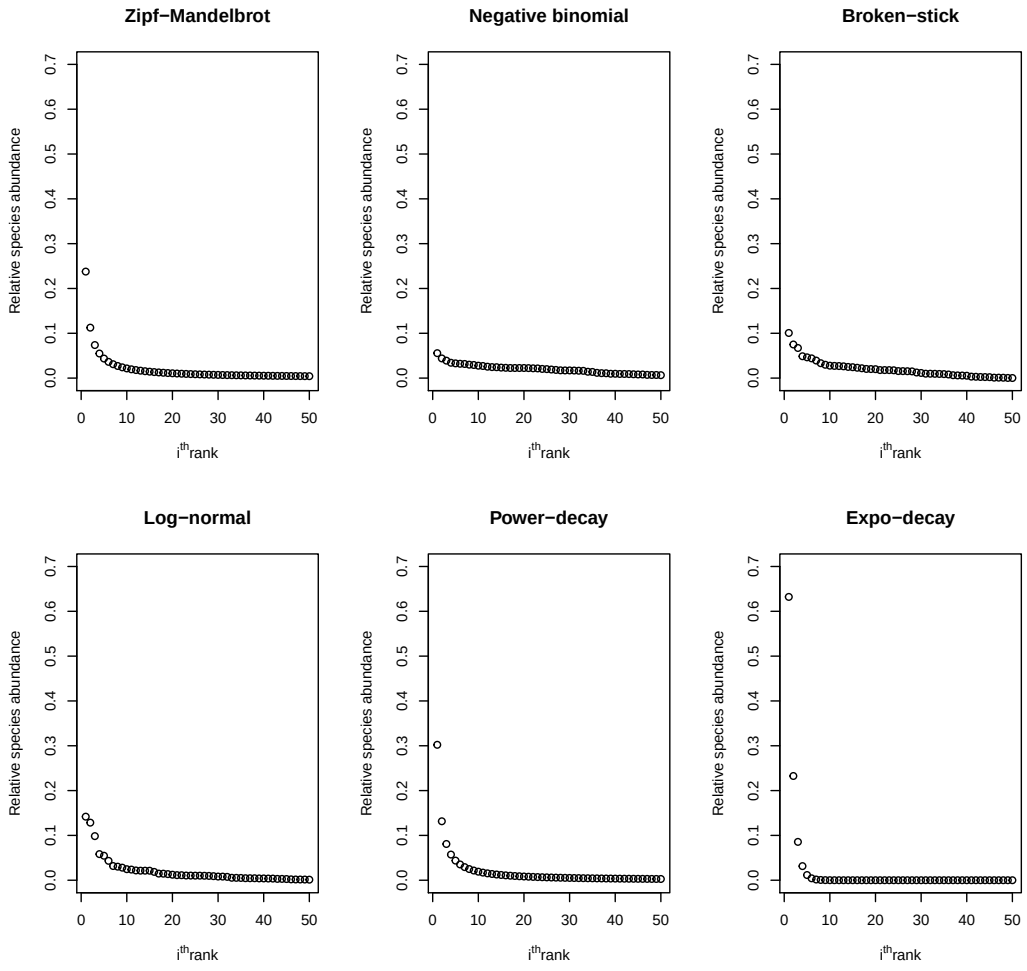


Figure 2.1: Probability p_i for distinct species $i = 1, 2, \ldots, N$, with $N = 50$ using different models, Zipf-Mandelbrot $p_i = 1/(i - 0.1)$, negative binomial(4,0.04), broken-stick (or Dirichlet(1)), log-normal(0,1), power-decay $p_i = 1/i^{1.2}$ and expo-decay $p_i = \exp(-i)$.

The most abundant species is presented at rank 1, the second most abundant species at rank 2 and so on. The exponential-decay model has a long right tail

with the highest first rank of abundance. The shape of rank abundance plot decreases rapidly compared to other models. For the log-normal, broken-stick and negative-binomial models, relative abundance decreases gradually and p_i is in the range 0 to 0.1.

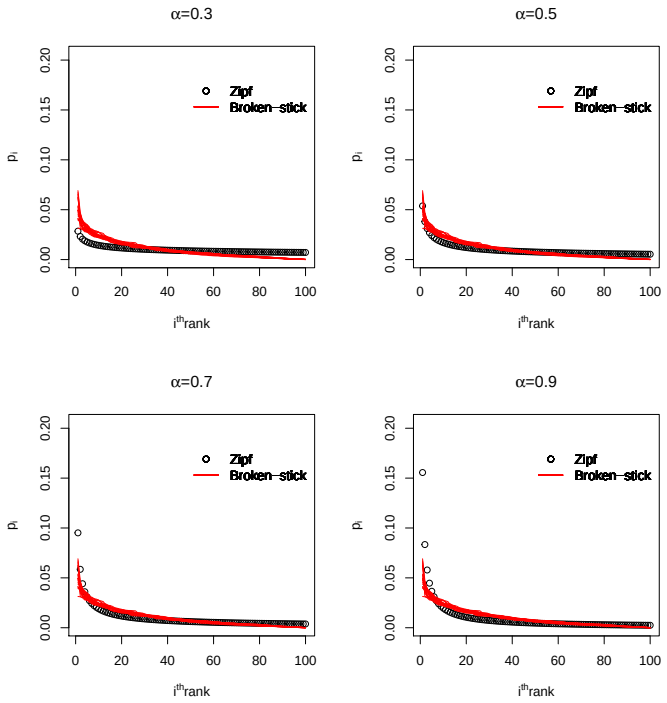


Figure 2.2: Plot of ranked p_i's values for the Zipf model with $N = 100$, α in the range [0.3,0.9] and broken-stick model with 20 simulations.

Relative abundance for Zipf model depends on the parameter α and $p_i = c/i^{\alpha}$ $(i = 1, \ldots, N)$. This model explains species abundance data with a similar shape to other models when choosing an appropriate value of α. For example, the Zipf model with $\alpha = 0.5$ provides the rank species abundance similar to log-normal(0,1) and broken-stick model (Figures 2.2 and 2.3). When $\alpha = 0.4$, the species abundance curve for the Zipf model displays the same results as negative binomial model NB(4,0.04) (Figure 2.4). When $\alpha = 2$, the Zipf model gives the species abundance which are similar the expo-decay model (Figure 2.5).

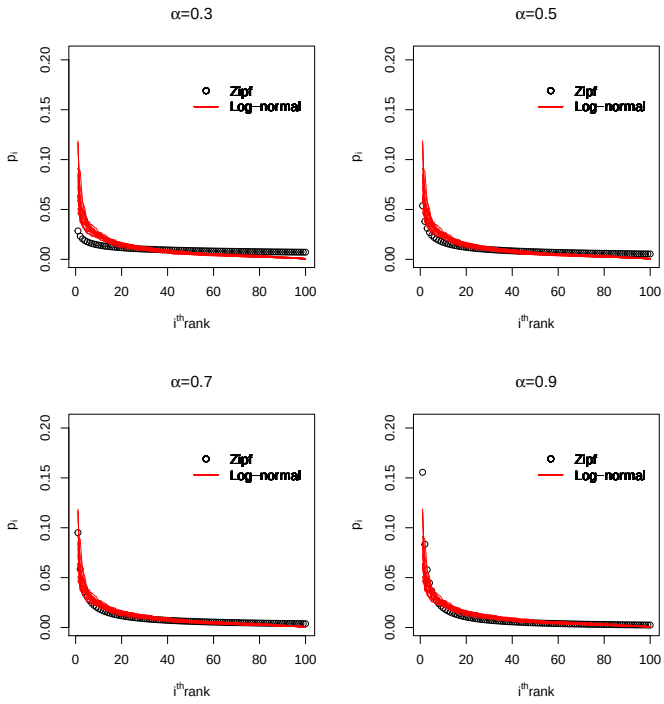


Figure 2.3: Plot of ranked p_i's values for the Zipf model with $N = 100$, α in the range [0.3,0.9] and log-normal(0,1) model 20 simulations.

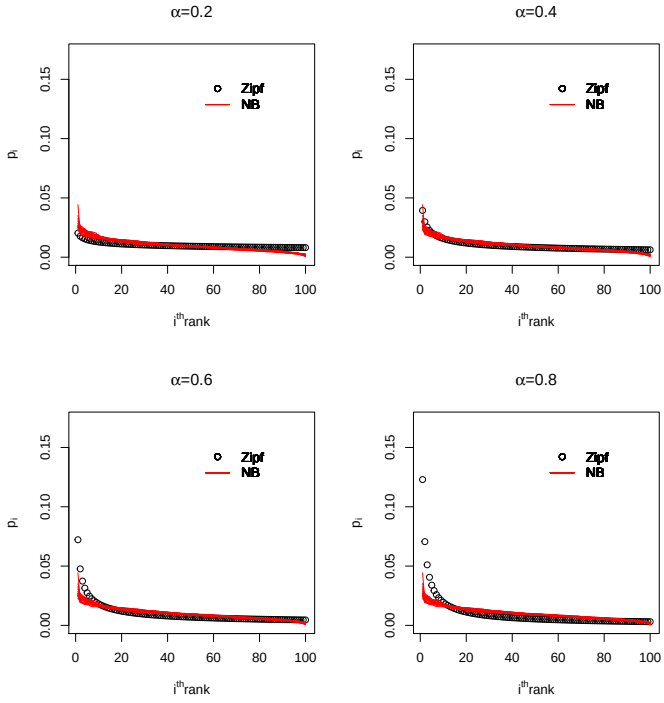


Figure 2.4: Plot of ranked p_i's values for the Zipf model with $N = 100$, α in the range [0.2,0.8] and NB(4,0.04) model with 20 simulations.

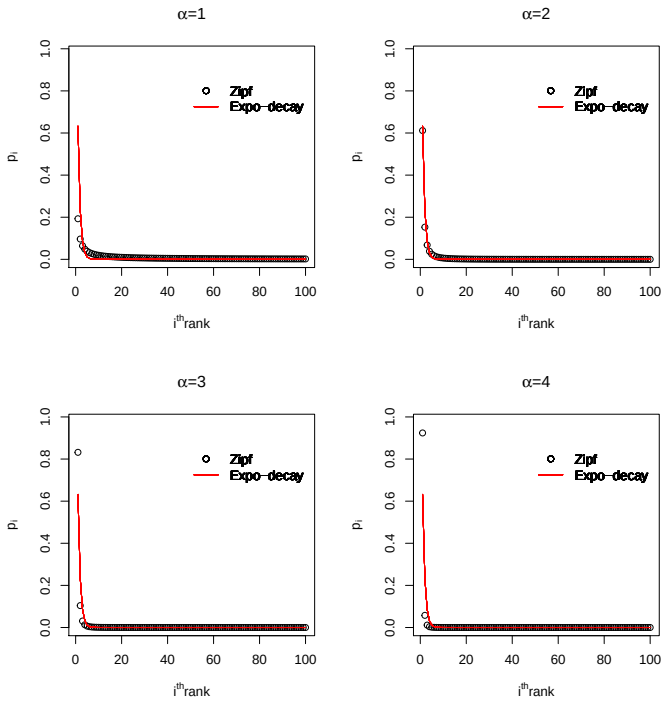


Figure 2.5: Plot of ranked p_i's values for the Zipf model and expo-decay model with $N = 100$, α in the range [1,4].

2.4 Nonparametric approach

Nonparametric estimators are useful methods as there are no requirements about assumptions for community structure (Chiarucci et al., 2003). Many estimators have been proposed for estimating the number of species and these are constructed based on the number of seen and unseen species. In particular, the number of unseen species is key for species richness estimation. The following nonparametric estimators are reviewed in this section.

2.4.1 Good-Turing estimator

Good-Turing estimation is a simple technique that estimates the number of unseen species using the frequency of singletons (species observed exactly once) in the sample, $f_1 = \sum_{i=1}^{N} I(X_i = 1)$. Because Good (1953) credits this idea to Turing, it is now known as the Good-Turing estimator.

The following explanation of the Good-Turing estimator is based on Chiu et al. (2014). Recall that M is the sample size or the number of individuals observed, $M = \sum_{k=1}^{k_{\max}} k f_k$, where f_k is the frequency of species seen k times. The mean relative abundance of the species seen k times in the sample, denoted as α_k, is

$$\alpha_k = \frac{\sum_{i=1}^{N} p_i I(X_i = k)}{f_k}, \qquad k = 0, 1, \ldots.$$

Good (1953) proposed that α_k could be estimated by

$$\widehat{\alpha}_k = \frac{(k+1)}{M} \frac{f_{k+1}}{f_k}. \tag{2.2}$$

By definition of α_k, the total relative abundance of all species seen k times is $\alpha_k f_k$ which can be estimated by

$$\widehat{\alpha}_k f_k = \frac{(k+1) f_{k+1}}{M}.$$

In particular, for $k = 0$,

$$\widehat{\alpha}_0 f_0 = \frac{f_1}{M}$$

is the estimated total relative abundance of all unseen species, $\sum_{i=1}^{N} p_i I(X_i = 0)$. Then, the expected number of unobserved species is

$$\mathrm{E}(f_0) = M \sum_{i=1}^{N} p_i I(X_i = 0) = M \alpha_0 f_0 = f_1.$$

Hence, the Good-Turing estimator of the number of species based on equation (2.1) is

$$\widehat{N}_G = K + f_1. \tag{2.3}$$

This form of the Good-Turing estimator is given for example by Hidaka (2014) as his estimator $\widehat{N}_{GT}$. This is also approximately the first-order jackknife estimator in Section 2.4.4, if the factor $\dfrac{(M-1)}{M}$ is omitted, see in Chiu et al.

(2014).

2.4.2 Chao1 Estimator

Chao (1984) proposed an estimator of a lower bound for species richness, although in practice it is often used as an estimator of species richness itself. Rare species have been considered in order to construct this estimator, which is based only on the number of species seen once and twice. Recall that X_i is the species frequency for species i in the sample and p_i is the probability that a randomly selected individual belongs to species i. The estimator can be derived under the multinomial and the Poisson sampling models as follows:

Multinomial Model

Under the multinomial sampling model, $X_i \sim Bin(M, p_i)$ which implies that

$$\begin{aligned} \mathbb{E}(f_k) &= \mathbb{E}\left[\sum_{i=1}^{N} I(X_i = 0)\right] \\ &= \sum_{i=1}^{N} \binom{M}{k} p_i^k (1-p_i)^{M-k}, \qquad k = 0, 1, 2, \ldots, M. \end{aligned} \tag{2.4}$$

The Cauchy-Schwarz inequality states that for any $a_i, b_i \in \mathbb{R}$,

$$\sum_{i=1}^{N} \left(a_i^2\right) \sum_{i=1}^{N} \left(b_i^2\right) \geq \left(\sum_{i=1}^{N} a_i b_i\right)^2. \tag{2.5}$$

Setting $a_i = (1-p_i)^{M/2}$, $b_i = p_i(1-p_i)^{M/2-1}$ and $a_i b_i = p_i(1-p_i)^{M-1}$, gives

$$\left[\sum_{i=1}^{N}(1-p_i)^M\right]\left[\sum_{i=1}^{N} p_i^2(1-p_i)^{M-2}\right] \geq \left[\sum_{i=1}^{N} p_i(1-p_i)^{M-1}\right]^2,$$

$$\begin{aligned} \mathbb{E}(f_0)\,\frac{1}{\binom{M}{2}}\mathbb{E}(f_2) &\geq \frac{1}{M^2}\left[\mathbb{E}(f_1)\right]^2, \\ \mathbb{E}(f_0) &\geq \frac{(M-1)}{M}\frac{\left[\mathbb{E}(f_1)\right]^2}{2\mathbb{E}(f_2)}. \end{aligned}$$

Using equation (2.1), a lower bound for the number of species becomes

$$N \geqslant \mathbb{E}(K) + \frac{(M-1)}{M}\frac{[\mathbb{E}(f_1)]^2}{2\mathbb{E}(f_2)},$$

which can be estimated using the observed data as

$$\widehat{N}_{Chao1} = K + \frac{(M-1)}{M}\frac{f_1^2}{2f_2}, \tag{2.6}$$

where f_1 and f_2 are the number of species seen once and twice.

The standard asymptotic approach known as the delta method is used for estimating the variance of $\widehat{N}_{Chao1}$ (Chiu et al., 2014).

$$\widehat{\text{var}}(\widehat{N}_{Chao1}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial \widehat{N}_{Chao1}}{\partial f_i}\,\frac{\partial \widehat{N}_{Chao1}}{\partial f_j}\,\widehat{\text{cov}}(f_i, f_j),$$

where

$$\widehat{\text{cov}}(f_i, f_j) = \begin{cases} f_i(1 - f_i/\widehat{N}_{Chao1}), & \text{if } i = j; \\ -f_i f_j/\widehat{N}_{Chao1}, & \text{if } i \neq j. \end{cases}$$

After some algebra, the variance estimator is derived as

$$\widehat{\text{var}}(\widehat{N}_{Chao1}) = f_2\left[\frac{1}{4}\left(\frac{M-1}{M}\right)^2\left(\frac{f_1}{f_2}\right)^4 + \left(\frac{M-1}{M}\right)^2\left(\frac{f_1}{f_2}\right)^3 + \frac{1}{2}\left(\frac{M-1}{M}\right)\left(\frac{f_1}{f_2}\right)^2\right] \tag{2.7}$$

Poisson Model

When M is large and p is small, the expected number of species seen k times can be approximated using the Poisson distribution with $\lambda_i = Np_i$ which gives

$$\mathbb{E}(f_k) \;=\; \sum_{i=1}^{N}\frac{\lambda_i^k\, e^{-\lambda_i}}{k!}, \qquad k = 0, 1, 2, \ldots, M. \tag{2.8}$$

Under the Cauchy-Schwarz inequality with $a_i = e^{-\lambda_i/2}$, $b_i = \lambda_i e^{-\lambda_i/2}$ and

$a_i b_i = \lambda_i e^{-\lambda_i}$, the lower bound for $\mathbb{E}(f_0)$ is given by

$$\begin{aligned} \left[\sum_{i=1}^{N} e^{-\lambda_i}\right]\left[\sum_{i=1}^{N} \lambda_i^2 e^{-\lambda_i}\right] &\geq \left[\sum_{i=1}^{N} \lambda_i e^{-\lambda_i}\right]^2, \\ \mathbb{E}(f_0)\ 2\mathbb{E}(f_2) &\geq [\mathbb{E}(f_1)]^2, \\ \mathbb{E}(f_0) &\geq \frac{[\mathbb{E}(f_1)]^2}{2\mathbb{E}(f_2)}, \end{aligned}$$

which again leads to the estimator

$$\widehat{N}_{Chao1} = K + \frac{f_1^2}{2f_2}. \tag{2.9}$$

When M is large, the variance estimator of $\widehat{N}_{Chao1}$ in equation (2.7) can be reduced as (Chao, 1987)

$$\mathrm{Var}(\widehat{N}_{Chao1}) = f_2 \left[\frac{1}{4}\left(\frac{f_1}{f_2}\right)^4 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{2}\left(\frac{f_1}{f_2}\right)^2\right]. \tag{2.10}$$

When the estimator breaks down at $f_2 = 0$, a modified bias-corrected estimator is proposed

$$\widehat{N}_{Chao1} = K + \frac{f_1(f_1 - 1)}{2(f_2 + 1)}. \tag{2.11}$$

The Chao1 estimator is extended in the study of Chiu et al. (2014) using the first four frequencies of distinct species and by Lanumteang and Böhning (2011) using the first three frequencies of distinct species. In the next section, the improved Chao1 estimator by Chiu et al. (2014) is investigated and compared to the original Chao1 estimator.

2.4.3 iChao1 estimator

An improved Chao1 estimator called iChao1 is developed by Chiu et al. (2014) based on a modified Good-Turing frequency formula. The new estimator obtains a new lower bound using the number of singletons, doubletons, tripletons and quadrupletons (i.e. f_1, f_2, f_3 and f_4). The improved estimator by Chiu et al. (2014) usually outperforms the traditional Chao1 estimator with reduced

bias, in particular when relative abundances are highly heterogeneous.

Chiu et al. (2014) proposed estimating the true mean relative abundance of species seen k times as

$$\widehat{\alpha}_k = \frac{(k+1)f_{k+1}}{(M-k)f_k + (k+1)f_{k+1}}, \qquad k = 1, 2, \ldots, \tag{2.12}$$

The new lower bound is derived by considering the magnitude of the first-order bias of $\widehat{N}_{Chao1}$ which can be derived as

$$\begin{aligned}
|\text{bias}(\widehat{N}_{Chao1})| &= \mathbb{E}(f_0) - \frac{(M-1)}{M}\frac{\mathbb{E}(f_1)^2}{2\mathbb{E}(f_2)} \\
&= \frac{\mathbb{E}(f_0)\left\{2\mathbb{E}(f_2)/[M(M-1)]\right\} - [\mathbb{E}(f_1)/M]^2}{2\mathbb{E}(f_2)/[M(M-1)]} \\
&\approx \left[\frac{1-\alpha_3}{\alpha_3}\frac{1-\alpha_1}{\alpha_1} - \left(\frac{1-\alpha_3}{\alpha_3}\right)^2\right] \\
&\quad \times \left[\sum_{i=1}^{N} p_i(1-p_i)^{n-1}\right] \times \left[\sum_{i=1}^{N} p_i^3(1-p_i)^{n-3}\right],
\end{aligned}$$

and applying the Cauchy Schwarz inequality yields

$$\left[\sum_{i=1}^{N} p_i(1-p_i)^{n-1}\right] \times \left[\sum_{i=1}^{N} p_i^3(1-p_i)^{n-3}\right] \geq \left[\sum_{i=1}^{N} p_i^2(1-p_i)^{n-2}\right]^2.$$

Hence, the approximate bias of the estimator becomes

$$\begin{aligned}
|\text{bias}(\widehat{N}_{Chao1})| &\approx \left[\frac{1-\alpha_3}{\alpha_3}\frac{1-\alpha_1}{\alpha_1} - \left(\frac{1-\alpha_3}{\alpha_3}\right)^2\right] \times \left[\sum_{i=1}^{N} p_i^2(1-p_i)^{n-2}\right]^2 \\
&\approx \frac{1-\alpha_3}{\alpha_3}\left[\frac{1-\alpha_1}{\alpha_1} - \frac{1-\alpha_3}{\alpha_3}\right]\frac{2\mathbb{E}(f_2)}{M(M-1)}.
\end{aligned} \tag{2.13}$$

Using the modified Good-Turing frequency in equation (2.12), we obtain the

improved Chao1 estimator as $\widehat{N}_{Chao1} + |\text{bias}(\widehat{N}_{Chao1})|$, that is,

$$\widehat{N}_{iChao1} = \widehat{N}_{Chao1} + \frac{(M-3)}{4M}\frac{f_3}{f_4} \times \max\left(f_1 - \frac{(M-3)}{2(M-1)}\frac{f_2 f_3}{f_4}, 0\right). \quad (2.14)$$

When $f_4 = 0$, it is replaced by $f_4 + 1$. For large sample size or equal species abundance, the iChao1 estimator is close to being an asymptotically unbiased estimator which leads to good approximation (Chiu et al., 2014). On the other hand, a negative bias may exist for unequal species abundance or small sample size (Chao and Chiu, 2014).

When M is large, equation (2.14) can be simplified to

$$\widehat{N}_{iChao1} = \widehat{N}_{Chao1} + \frac{f_3}{4f_4} \times \max\left(f_1 - \frac{f_2 f_3}{2f_4}, 0\right). \quad (2.15)$$

The variance of iChao1 estimator can be approximated using the delta method by

$$\widehat{\text{var}}(\widehat{N}_{iChao1}) \approx \bigtriangledown g \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}^T \widehat{cov} \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} \bigtriangledown g \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}, \quad (2.16)$$

where

$$\bigtriangledown g \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} \frac{\partial \widehat{N}}{\partial f_0} & \frac{\partial \widehat{N}}{\partial f_1} & \frac{\partial \widehat{N}}{\partial f_2} & \frac{\partial \widehat{N}}{\partial f_3} & \frac{\partial \widehat{N}}{\partial f_4} \end{pmatrix}^T,$$

with $\widehat{N} = \widehat{N}_{iChao1}$. The partial derivatives $\dfrac{\partial \widehat{N}}{\partial f_i}$ for $j = 0, 1, 2, 3, 4$ are

$$\frac{\partial \widehat{N}}{\partial f_0} = -1,$$

$$\frac{\partial \widehat{N}}{\partial f_1} = \frac{1}{4}\left[\frac{4f_1 f_4(M-1) + f_2 f_3(M-3)}{M f_2 f_4}\right],$$

$$\frac{\partial \widehat{N}}{\partial f_2} = -\frac{1}{8}\left[\frac{4f_1^2 f_4^2(M-1)^2 + f_2^2 f_3^2(M-3)^2}{M(M-1)f_2^2 f_4^2}\right],$$

$$\frac{\partial \widehat{N}}{\partial f_3} = \frac{(M-3)}{4}\left[\frac{(f_1 f_4(M-1) - f_2 f_3(M-3))}{M(M-1)f_4^2}\right],$$

$$\frac{\partial \widehat{N}}{\partial f_4} = -\frac{(M-3)f_3}{4}\left[\frac{(f_1 f_4(M-1) - f_2 f_3(M-3))}{M(M-1)f_4^3}\right].$$

and the variance-covariance matrix of the multinomial vector $(f_0, f_1, f_2, f_3, f_4)^T$ can be estimated by

$$\widehat{cov}\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{bmatrix} f_0\left(1-\frac{f_0}{N}\right) & -\frac{f_0 f_1}{N} & -\frac{f_0 f_2}{N} & -\frac{f_0 f_3}{N} & -\frac{f_0 f_4}{N} \\ -\frac{f_0 f_1}{N} & f_1\left(1-\frac{f_1}{N}\right) & -\frac{f_1 f_2}{N} & -\frac{f_1 f_3}{N} & -\frac{f_1 f_4}{N} \\ -\frac{f_0 f_2}{N} & -\frac{f_1 f_2}{N} & f_2\left(1-\frac{f_2}{N}\right) & -\frac{f_2 f_3}{N} & -\frac{f_2 f_4}{N} \\ -\frac{f_0 f_3}{N} & -\frac{f_1 f_3}{N} & -\frac{f_2 f_3}{N} & f_3\left(1-\frac{f_3}{N}\right) & -\frac{f_3 f_4}{N} \\ -\frac{f_0 f_4}{N} & -\frac{f_1 f_4}{N} & -\frac{f_2 f_4}{N} & -\frac{f_3 f_4}{N} & f_4\left(1-\frac{f_4}{N}\right) \end{bmatrix}.$$

For practical calculation, f_0 and N can be replaced by $\hat{f}_0 = \frac{(M-1)f_1^2}{2Mf_2} + |\text{bias}(\widehat{N}_{Chao1})|$ and $\widehat{N}_{iChao1}$. For the homogeneous model, the expected value of $f_1 - f_2 f_3/2f_4$ tends to zero as the sample size increases. Then, the iChao1 estimator can be replaced by the Chao1 estimator (Chiu et al., 2014) .

2.4.4 Jackknife estimators

Jackknife estimators were proposed by Quenouille (1949) and expanded by Tukey (1958). Suppose we have a biased estimator, $\widehat{\theta}$, of a parameter θ. The basic idea of the jackknife method is to calculate a series of estimators $\widehat{\theta}_{-i}$, missing out the i^{th} sample observation and calculate the new estimators

$$\widehat{\theta}_J^{(1)} = M\widehat{\theta} - \left(\frac{M-1}{M}\right)\sum_{i=1}^{M}\widehat{\theta}_{-i}.$$

This estimator is known as the first-order jackknife method and has reduced

bias compared to $\widehat{\theta}$.

Jackknife estimators of species richness were introduced by Burnham and Overton (1978). The basic estimator is $\widehat{\theta} = K$, the observed number of species. Let $\widehat{\theta}_{-i}$ be the number of distinct species by leaving out species i, which is given by

$$\widehat{\theta}_{-i} = \begin{cases} K-1 & \text{if species } X_i \text{ seen only once,} \\ K & \text{otherwise.} \end{cases}.$$

The first-order jackknife estimator can therefore be derived as

$$\begin{aligned} \widehat{N}_J^{(1)} &= M\widehat{\theta} - \left(\frac{M-1}{M}\right)\sum_{i=1}^{M}\widehat{\theta}_{-i} \\ &= MK - \left(\frac{M-1}{M}\right)\{f_1(K-1) + (M-f_1)K\} \\ &= MK - \left(\frac{M-1}{M}\right)\{MK - f_1\} \\ &= MK - (M-1)K + \left(\frac{M-1}{M}\right)f_1 \\ &= K + \left(\frac{M-1}{M}\right)f_1. \end{aligned} \tag{2.17}$$

It is also possible to derive higher-order jackknife estimators by omitting more than one observation from the sample. The second-order jackknife estimator involves estimators $\widehat{\theta}_{-ij}$, calculated by excluding each pair of observations i, j from the sample. The general formula for the second-order jackknife estimator is

$$\widehat{\theta}_J^{(2)} = \frac{1}{2}\left\{M^2\widehat{\theta} - \frac{2(M-1)^2}{M}\sum_{i=1}^{M}\widehat{\theta}_{-i} + \frac{2(M-2)^2}{M(M-1)}\sum_{i<j}\widehat{\theta}_{-ij}\right\}.$$

To apply this to species sampling, let $\widehat{\theta}_{-ij}$ be the number of distinct species

by leaving out samples i and j

$$\widehat{\theta}_{-ij} = \begin{cases} K-2 & \text{if species } X_i \text{ and } X_j \text{ both seen once,} \\ K-1 & \text{if either species } X_i \text{ or species } X_j \text{ seen once,} \\ K & \text{otherwise.} \end{cases}$$

Burnham and Overton (1978) show that this leads to the estimator

$$\begin{aligned} \widehat{N}_J^{(2)} &= \frac{1}{2}\left\{ M^2\widehat{\theta} - \frac{2(M-1)^2}{M}\sum_{i=1}^{M}\widehat{\theta}_{-i} + \frac{2(M-2)^2}{M(M-1)}\sum_{i<j}\widehat{\theta}_{-ij} \right\} \\ &= K + \frac{(2M-3)f_1}{M} - \frac{(M-2)^2 f_2}{M(M-1)}. \end{aligned} \tag{2.18}$$

In practice, simplified forms of these estimators are often used, based on large values of M

$$\widehat{N}_J^{(1)} = K + f_1 \tag{2.19}$$

$$\widehat{N}_J^{(2)} = K + 2f_1 - f_2. \tag{2.20}$$

The result in equation (2.22) shows that the first-order Jackknife estimator is identical to the Good-Turing estimator in equation (2.3) when M is large.

Burnham and Overton (1978) proposed the general simplified formula for the k^{th}-order jackknife estimator which is given by

$$\widehat{N}_J^{(k)} = K + \sum_{j=1}^{k}(-1)^{j+1}\binom{k}{j}f_j. \tag{2.21}$$

Ji-Ping Wang developed the R package called SPECIES in 2011 which provides a function `jackknife` to calculate these estimators.

The first-order jackknife estimator is constructed using the number of rare

species which are seen only once. For the second order jackknife estimator, it is formed using both the number of species seen once and the number seen twice. The bias and variance of estimator are balanced by choosing the kth-order. A higher order is appropriate for improving bias. However, this might lead to a higher variance of estimator (Wang, 2011).

Under the distribution of K and the expectation of f_1 in equation (2.4), the expected value of the first-order jackknife estimator can be derived as

$$\begin{aligned} E(\widehat{N}_J^{(1)}) &= \mathrm{E}(K) + \frac{M-1}{M}\,\mathrm{E}(f_1) \\ &= \mathrm{E}(K) + \frac{M-1}{M}\sum_{i=1}^{N}\tbinom{M}{1}p_i(1-p_i)^{M-1} \\ &= \mathrm{E}(K) + (M-1)\sum_{i=1}^{N} p_i(1-p_i)^{M-1}. \end{aligned}$$

where $\mathrm{E}(K) = N - \sum_{i=1}^{N}(1-p_i)^M$ (Hidaka, 2014). Then, we have

$$\begin{aligned} \mathrm{Bias}(\widehat{N}_J^{(1)}) &= \mathrm{E}(\widehat{N}_J^{(1)}) - N \\ &= \sum_{i=1}^{N}(1-p_i)^M + (M-1)\sum_{i=1}^{N} p_i(1-p_i)^{M-1}. \end{aligned} \tag{2.22}$$

Considering the same approach, the bias of the second-order jackknife estima-

tor can be written as

$$\begin{aligned} E(\widehat{N}_J^{(2)}) &= E(K) + \frac{(2M-3)}{M}\,\mathrm{E}(f_1) - \frac{(M-2)^2}{M(M-1)}\,\mathrm{E}(f_2) \\ &= N - \sum_{i=1}^{N}(1-p_i)^M + \frac{(2M-3)}{M}\sum_{i=1}^{N}\tbinom{M}{1}p_i(1-p_i)^{M-1} - \\ &\quad \frac{(M-2)^2}{M(M-1)}\sum_{i=1}^{N}\tbinom{M}{2}p_i(1-p_i)^{M-2} \\ &= N - \sum_{i=1}^{N}(1-p_i)^M + (2M-3)\sum_{i=1}^{N}p_i(1-p_i)^{M-1} - \\ &\quad \frac{(M-2)^2}{2}\sum_{i=1}^{N}p_i^2(1-p_i)^{M-2}, \end{aligned}$$

and this gives

$$\begin{aligned} \mathrm{Bias}(\widehat{N}_J^{(2)}) &= E(\widehat{N}_J^{(2)}) - N \\ &= \sum_{i=1}^{N}(1-p_i)^M + (2M-3)\sum_{i=1}^{N}p_i(1-p_i)^{M-1} - \\ &\quad \frac{(M-2)^2}{2}\sum_{i=1}^{N}p_i^2(1-p_i)^{M-2}. \end{aligned} \tag{2.23}$$

2.4.5 Horvitz-Thompson estimator

The Horvitz-Thompson (HT) estimator is an unbiased estimator of the population size N proposed by Horvitz and Thompson (1952). It is applied in many fields including the problem of estimating the number of species in ecology and estimating vocabulary size in linguistics, for example Böhning (2008), Cruyff and van der Heijden (2008) and Hidaka (2014). Assume π_i is the probability that species i is included in the sample, termed the inclusion probability. The estimator of species richness is given by

$$\widehat{N}_H = \sum_{i=1}^{M}\frac{f_i}{\pi_i}. \tag{2.24}$$

An unbiased estimator of variance is

$$\widehat{\mathrm{Var}}(\widehat{N}_H) = \sum_{i=1}^{M} \left(\frac{1-\pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^{M} \sum_{i \neq j} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} \right) y_i y_j, \tag{2.25}$$

where $\pi_i = 1 - (1-p_i)^M$ is the inclusion probability of species i and $\pi_{ij} = \pi_i + \pi_j - \left[1 - (1-p_i-p_j)^M\right]$ is the inclusion probability for species i and j.

The species abundance model or the probability of species i collected is unknown in practice. The Horvitz-Thompson-Like estimator is an alternative estimator which can be used instead of the Horvitz-Thompson estimator. The unknown probability p_i is replaced by i/M. Then, the Horvitz-Thompson-Like estimator is given by (McCrea and Morgan, 2014)

$$\widehat{N}_{HL} = \sum_{i=1}^{M} \frac{f_i}{1 - \left(1 - \frac{i}{M}\right)^M}. \tag{2.26}$$

2.5 An alternative improvement to the Chao1 estimator

In this section, two new estimators of species richness are developed using $\widehat{\alpha}_k$ based on the Good-Turing coverage estimator. The sample coverage is the proportion of all individuals in the population belonging to the observed species in the sample. The concept of the sample coverage is presented in an example of Chao and Jost (2012) who discussed the sample coverage of a terrestrial arthropod community with 50 species. Assume that the relative abundance of species 1 is 0.3, species 2 is 0.1, species 3 through 5 is 0.05 each and species 6 through 50 are 0.01 each. In sample of 20 individuals, there are 12 species collected (e.g. species 1, 2, 3, 4, 5, 6, 9, 14, 23, 27, 41, 47) and then the sample coverage is 62% $(0.3 + 0.1 + 0.05 \times 3 + 0.01 \times 7)$. This means there is 62% of

all individuals belonging to the observed species in the sample.

Good (1953) proposed estimating the sample coverage using

$$\widehat{C} = \sum_{i=1}^{N} p_i I(X_i = 0) = 1 - \alpha_0 f_0 = 1 - \frac{f_1}{M}.$$

Then, an estimator for the true mean relative abundance of species seen k times based on the Good-Turing coverage approach is given by

$$\widehat{\alpha}_k = \left(1 - \frac{f_1}{M}\right) \frac{k}{M}. \tag{2.27}$$

The following two new estimators, $\widehat{N}_{new1}$ and $\widehat{N}_{new2}$, are constructed using the same idea of the iChao1 estimator. The bias of $\widehat{N}_{Chao1}$ is estimated using equation (2.13). The $\widehat{N}_{new1}$ estimator is constructed using $\widehat{\alpha}_1$ by Chiu et al. (2014) in equation (2.12) and $\widehat{\alpha}_3$ by equation (2.27). This provides

$$\widehat{N}_{new1} = \widehat{N}_C + \frac{1}{9}\left[\frac{(3Mf_1(M-1) - 2Mf_2(M-3) - 3f_1^2(M-1) - 6f_1f_2)\,(M^2 - 3(M-f_1))}{M(M-1)(M-f_2)^2}\right]. \tag{2.28}$$

The $\widehat{N}_{new2}$ estimator estimates both $\widehat{\alpha}_1$ and $\widehat{\alpha}_3$ by equation (2.27), giving

$$\widehat{N}_{new2} = \widehat{N}_C + \frac{4}{9}\left[\frac{(M^2 - 3M + 3f_1)Mf_2}{(M-1)(M-f_1)^2}\right]. \tag{2.29}$$

The delta method is used to estimate the variance of both alternative estimators followings the same approach as for the iChao1 estimator. The formulae are lengthy and are not given here, but are incorporated into R code.

2.6 Comparing previous model using simulation

The mean relative abundance of species seen k times (α_k) can be estimated using the estimator as follows

- Crude estimator : $\widehat{\alpha}_k = \dfrac{k}{M}$
- Good-Turing (GT) estimator : $\widehat{\alpha}_k = \dfrac{(k+1)}{M}\dfrac{f_{k+1}}{f_k}$
- Modified Good-Turing (GT_{chiu}) estimator : $\widehat{\alpha}_k = \dfrac{(k+1)f_{k+1}}{(M-k)f_k + (k+1)f_{k+1}}$
- Good-Turing coverage (GT_{cov}) estimator : $\widehat{\alpha}_k = \left(1 - \dfrac{f_1}{M}\right)\dfrac{k}{M}$.
- Chao and Jost (2012) coverage (CJ_{cov}) estimator : $\widehat{\alpha}_k = \widehat{C} \times \dfrac{k}{M}$, where

$$\widehat{C} = \begin{cases} 1 - \dfrac{f_1}{M}\left[\dfrac{(M-1)f_1}{(M-1)f_1 + 2f_2}\right], & \text{if } f_2 > 0, \\ 1 - \dfrac{f_1}{M}\left[\dfrac{(M-1)(f_1-1)}{(M-1)(f_1-1) + 2f_2}\right], & \text{if } f_2 = 0. \end{cases}$$

Here, these estimators above are applied to abundance data simulated from the negative binomial model with parameter (4,0.04). The performance of $\widehat{\alpha}_1$ and $\widehat{\alpha}_3$ is compared by plotting boxplots of their root mean square error (rmse) and of the bias (Figures 2.6 and 2.7). In the simulation study, the bias is calculated as the mean of $(\widehat{\alpha}_k - \alpha_k)$, where α_k is a sample quantity that varies from sample to sample, rather than a fixed parameter.

The results indicate that crude estimator outperforms other estimators for small N while GT and GT_{chiu} estimate $\widehat{\alpha}_1$ well for large N (Figure 2.6). Although GT_{cov} estimator is not very good for $\widehat{\alpha}_1$, it works very well with $\widehat{\alpha}_3$ for both small and large sample size, $M = 200$ and 400 respectively (Figure 2.7).

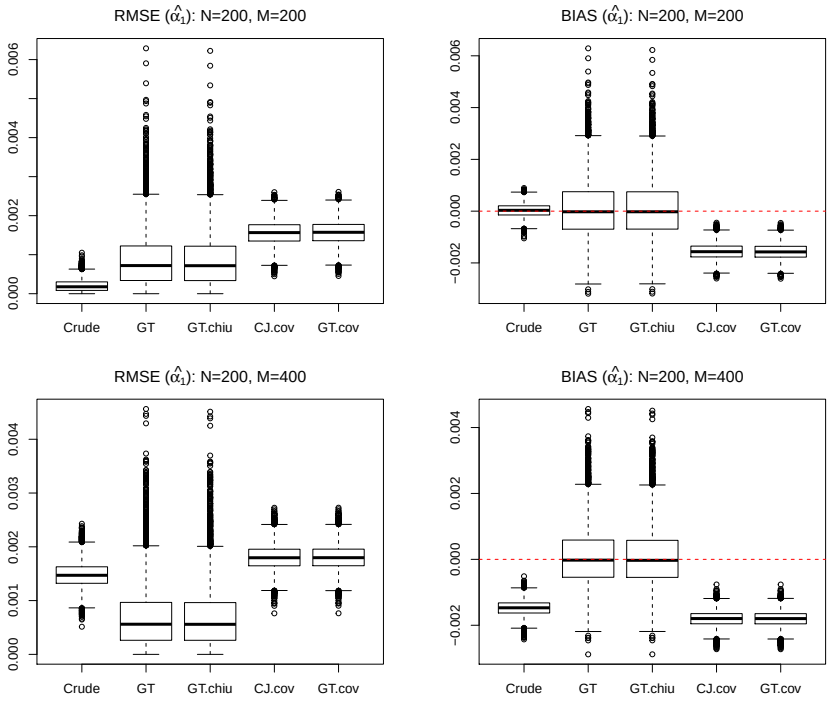


Figure 2.6: RMSE and Bias of $\widehat{\alpha}_1$ based on the Negative Binomial model with parameter $k = 4$ and $r = 0.04$, $N = 200$, M =200 and 400 with 10000 simulations.

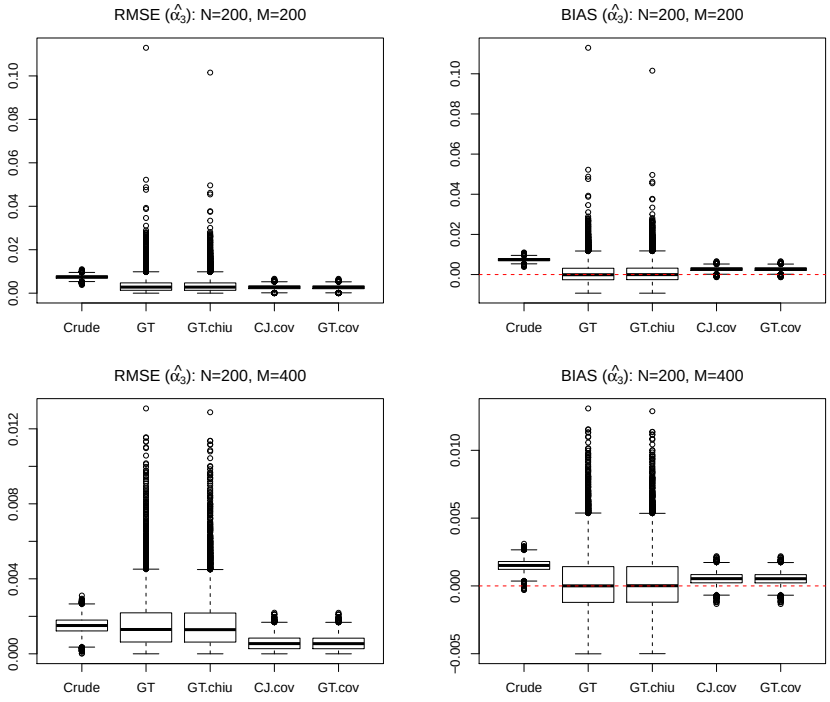


Figure 2.7: RMSE and Bias of $\widehat{\alpha}_3$ based on the negative binomial model with parameter $k = 4$ and $r = 0.04$, $N = 200$, M =200 and 400 with 10000 simulations.

Tables 2.1 and 2.2 present the bias and root means square error (RMSE) ($\times$10000) under various species abundance models for $\widehat{\alpha}_1$ and $\widehat{\alpha}_3$ respectively. When considering the performance in terms of RMSE for small sample size $M = 200$, crude estimator for $\widehat{\alpha}_1$ approximates well based on NB, expo-decay and broken-stick models.

For the GT_{cov}, it approximates well when using the log-normal and the Zipf-Mandelbrot model, while the GT and GT_{chiu} yield good approximations for the power-decay model. For example, the broken-stick model with $N = 200$, the crude estimator gives the best accuracy with smallest RMSE=3.64 while GT, GT_{chiu}, CJ_{cov} and GT_{cov} obtain large RMSE with 10.49, 10.44, 11.18 and 11.24 receptively (Table 2.1). It clearly outperforms other approaches. Additionally, it obtains smallest bias with 0.99, in contrast to CJ_{cov} and GT_{cov} which gives a negative bias with -11.15 and -11.21 respectively.
On the other hand, the performance of estimators of $\widehat{\alpha}_1$ for large sample size show that the GT_{chiu} estimator approximates the best for all models in terms of RMSE and bias. For example for the broken-stick model and $N = 800$, the GT_{chiu} estimator gives the smallest RMSE and smallest bias 4.43 and 0.63 respectively. Other estimators have higher RMSE and bias, see in Table 2.1.
When considering $\widehat{\alpha}_3$, it is clear that GT_{cov} estimator works well for both small and large sample size (Table 2.2). The results show small RMSE for GT_{cov} estimator. For example with $N = 200$, GT_{cov} estimator can approximate very well with the smallest RMSE for all models. The RMSE for the broken-stick model and using the GT_{cov} estimator is 13.32 while the Crude, GT, GT_{chiu} and CJ_{cov} estimators are 44.80, 46.20, 45.71 and 13.40 respectively (Table 2.2).

Table 2.1: Bias and RMSE of $\widehat{\alpha}_1$ ($\times$ 10000) with 10000 times

M	Model	Crude	GT	GT_{chiu}	CJ_{cov}	GT_{cov}
Bias						
200	Negative Binomial	0.27	0.68	0.67	-15.57	-15.64
	Expo-decay	-1.49	0.97	0.94	-14.61	-14.68
	Log-normal	7.68	0.99	1.00	-5.06	-5.11
	Zipf-Mandelbrot	19.48	0.97	1.02	8.08	8.05
	Power	22.68	0.87	0.92	13.72	13.70
	Broken-stick	0.99	1.05	1.03	-11.15	-11.21
400	Negative Binomial	-14.77	0.62	0.55	-18.03	-18.04
	Expo-decay	-10.44	0.84	0.79	-13.24	-13.25
	Log-normal	-4.35	0.69	0.67	-7.52	-7.53
	Zipf-Mandelbrot	3.56	0.43	0.43	0.11	0.10
	Power	7.48	0.39	0.40	4.57	4.56
	Broken-stick	-6.06	0.90	0.87	-8.85	-8.86
800	Negative Binomial	-16.20	1.21	1.15	-16.62	-16.62
	Expo-decay	-8.50	0.86	0.83	-8.97	-8.97
	Log-normal	-7.03	0.59	0.57	-7.65	-7.65
	Zipf-Mandelbrot	-3.48	0.36	0.36	-4.32	-4.32
	Power	0.55	0.22	0.22	-0.28	-0.28
	Broken-stick	-6.49	0.64	0.63	-7.03	-7.03
RMSE						
200	Negative Binomial	2.09	8.61	8.56	15.57	15.64
	Expo-decay	3.29	10.22	10.17	14.61	14.68
	Log-normal	7.77	9.17	9.14	5.43	5.47
	Zipf-Mandelbrot	19.48	8.22	8.21	8.17	8.13
	Power	22.68	8.90	8.89	13.74	13.71
	Broken-stick	3.64	10.49	10.44	11.18	11.24
400	Negative Binomial	14.77	6.81	6.78	18.03	18.04
	Expo-decay	10.44	7.18	7.15	13.24	13.25
	Log-normal	4.41	5.91	5.89	7.52	7.53
	Zipf-Mandelbrot	3.63	4.50	4.49	1.54	1.54
	Power	7.49	4.44	4.44	4.62	4.61
	Broken-stick	6.09	6.65	6.62	8.85	8.86
800	Negative Binomial	16.20	6.56	6.53	16.62	16.62
	Expo-decay	8.50	5.12	5.10	8.97	8.97
	Log-normal	7.03	4.08	4.07	7.65	7.65
	Zipf-Mandelbrot	3.48	3.03	3.03	4.32	4.32
	Power	0.99	2.50	2.50	0.88	0.88
	Broken-stick	6.49	4.44	4.43	7.03	7.03

Table 2.2: Bias and RMSE of $\widehat{\alpha}_3$ ($\times$ 10000) with 10000 times

M	Model	Crude	GT	GT_{chiu}	CJ_{cov}	GT_{cov}
Bias						
200	Negative Binomial	74.50	7.90	8.22	27.00	26.77
	Expo-decay	55.59	8.31	8.51	16.23	16.02
	Log-normal	52.38	12.51	12.45	14.18	14.02
	Zipf-Mandelbrot	48.87	27.03	25.92	14.66	14.56
	Power	41.36	36.82	34.67	14.49	14.41
	Broken-stick	44.78	10.75	10.77	8.36	8.18
400	Negative Binomial	15.07	2.47	2.51	5.30	5.26
	Expo-decay	8.71	3.61	3.58	0.32	0.29
	Log-normal	15.25	4.21	4.21	5.74	5.72
	Zipf-Mandelbrot	22.94	6.72	6.69	12.59	12.57
	Power	20.28	11.32	11.08	11.53	11.51
	Broken-stick	6.62	4.04	3.97	-1.76	-1.79
800	Negative Binomial	-4.97	1.35	1.31	-6.24	-6.25
	Expo-decay	-4.55	2.23	2.18	-5.96	-5.97
	Log-normal	1.75	1.75	1.73	-0.09	-0.10
	Zipf-Mandelbrot	9.22	1.45	1.46	6.72	6.71
	Power	9.52	2.55	2.54	7.04	7.03
	Broken-stick	-0.16	2.00	1.97	-1.77	-1.77
RMSE						
200	Negative Binomial	74.50	35.81	35.65	27.04	26.80
	Expo-decay	55.59	40.57	40.25	17.38	17.21
	Log-normal	52.59	51.97	51.41	20.10	20.01
	Zipf-Mandelbrot	51.56	80.16	78.61	27.68	27.64
	Power	48.77	100.15	97.62	33.25	33.23
	Broken-stick	44.80	46.20	45.71	13.40	13.32
400	Negative Binomial	15.07	15.85	15.76	5.82	5.79
	Expo-decay	9.21	19.80	19.67	4.98	4.97
	Log-normal	15.46	21.10	20.98	7.77	7.76
	Zipf-Mandelbrot	23.25	26.68	26.53	14.22	14.20
	Power	21.67	35.79	35.44	15.12	15.11
	Broken-stick	8.44	22.51	22.34	6.36	6.37
800	Negative Binomial	5.08	9.76	9.71	6.27	6.28
	Expo-decay	5.00	11.86	11.80	6.17	6.17
	Log-normal	3.33	10.39	10.35	2.97	2.97
	Zipf-Mandelbrot	9.26	9.70	9.68	6.89	6.88
	Power	9.75	12.43	12.39	7.57	7.56
	Broken-stick	3.39	11.61	11.56	3.65	3.66

2.7 Simulation Study

In a simulation study, the $\widehat{N}_{\text{new}_1}$ and $\widehat{N}_{\text{new}_3}$ estimators have been investigated. Their performances are compared with other estimators including the first-order jackknife, the second-order jackknife, Chao1 and iChao1 in terms of root mean square error (RMSE) and the coverage and width of 95% confidence intervals (C.I.) for N of the form

$$[K + (\widehat{N} - K)/R,\ K + (\widehat{N} - K) \times R],$$

where $R = \exp\left\{1.96\left[\log(1 + \frac{\widehat{\text{Var}}(\widehat{N})}{(\widehat{N} - K)^2}\right]\right\}$ (Chiu et al., 2014).

The following species abundance models were used to construct simulated data sets with $N = 200$, $M = 200, 400, 800$ and 1600. There were 10000 simulated data sets for each combination of N and M.

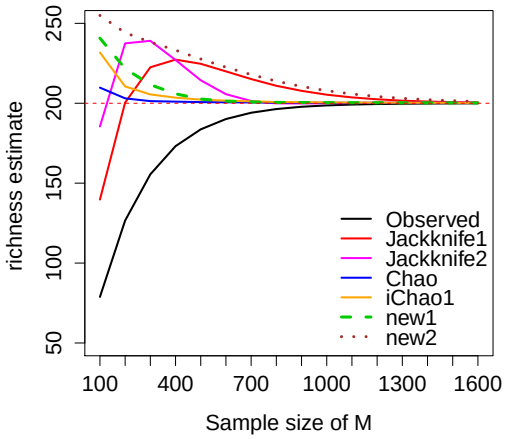
- model 1 : homogeneous model with $p_i = 1/N$
- model 2 : negative binomial model with parameter (0,0.04)
- model 3 : power-decay model with $p_i = c/i^{1.2}$
- model 4 : log-normal model with parameters $\mu = 0$ and $\sigma = 1$
- model 5 : Zipf-Mandelbrot model with $p_i = c/(i - 0.1), i = 1, 2, \ldots, N$
- model 6 : broken-stick model or Dirichlet$(1, 1, \ldots, 1)$ model

In Figure 2.8, the estimated number of species using various estimators are plotted against the sample size, M. For the homogeneous model, the Chao1 estimator estimates the number of species very well (Figure 2.8a). It outperforms other estimators with small bias for all M. When M tends to infinity, the estimated the number of species for all estimators is close to the true species richness $N = 200$. In Table 2.3, the Chao1 estimator yields the estimated the

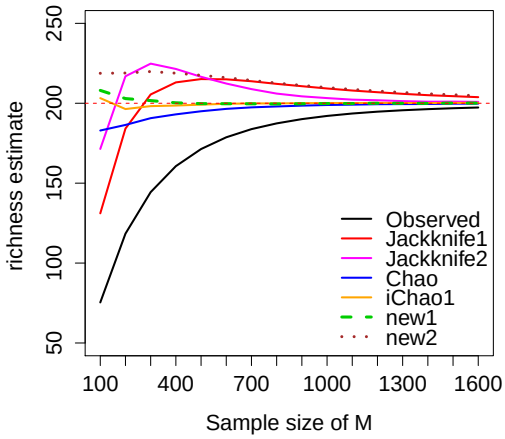
number of species close to $N = 200$ for all $M = 200, 400, 800, 1600$, with mean $\widehat{N}_{Chao1} = 202.74, 201.03, 200.37, 200.16$ respectively.

When addressing other estimators under the homogeneous model, the performance is not stable for all M. For example, the first-order Jackknife estimator can approximate well for $M = 200$, more bias for $M = 400, 800$ before becoming close to the true number of species at $M = 1600$, with mean $\widehat{N}_{J1} = 200.25$, 227.28, 210.93 and 200.46, respectively.

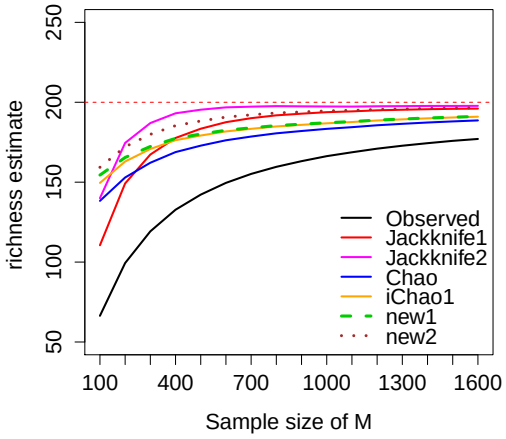
For the heterogeneous models, the new$_1$ estimator performs very similar to the iChao1 estimator. Figure 2.8b shows the performance of various estimators based on the negative binomial model. The results show the iChao1 and the new$_1$ estimators work well with the negative binomial model and are close to true number of species for all M. For the Chao1 estimator, it fits well when M is large. For example, when using the negative binomial model with $M = 200$, the new$_1$ estimator performs well and has a good coverage probability of the 95% confidence interval, with mean $\widehat{N} = 200.45$ and coverage 0.9464 respectively (Table 2.4). When $M = 800$, the Chao1 estimator performs the best in terms of RMSE and coverage probability of the 95% confidence interval, with values of 6.21 and 0.9443 respectively, while the new$_1$ estimator yields 7.22 and 0.9261 respectively.



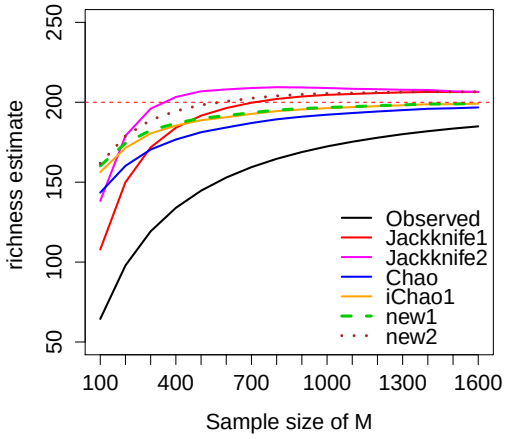
(a) Homogeneous model



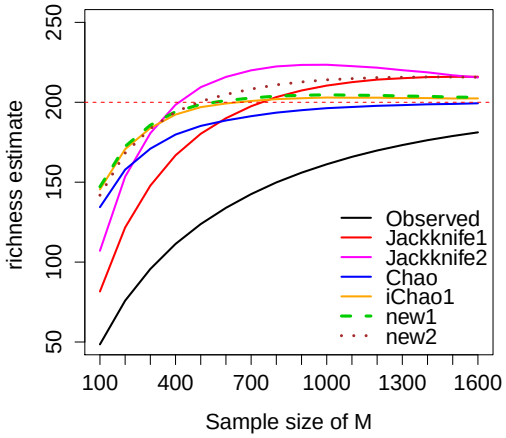
(b) Negative binomial model



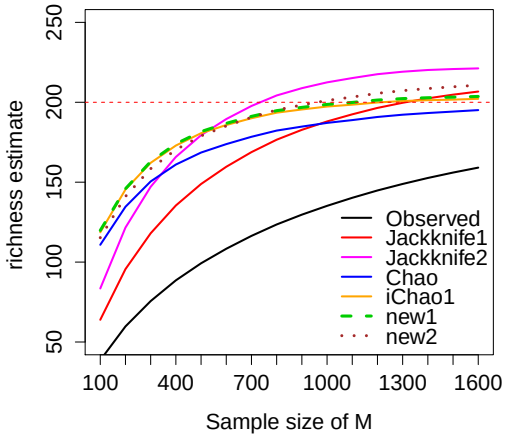
(c) Broken-stick model



(d) Log-normal model



(e) Zipf-Mandelbrot model



(f) Power-decay model

Figure 2.8: Comparison of biases for species richness estimators under homogeneous, negative binomial (NB), broken-stick, log-normal model, Zipf-Mandelbrot and power-decay models N=200, M=100-1600 and repeated 10000 times.

For the broken-stick model, all estimators underestimate the true number of species. The results indicate that the second-order jackknife performs well when compared with other estimators (Figure 2.8c). The new_2 estimator performs the second best and better than the new_1 and the iChao1 estimator. As shown in Table 2.8, the second-order jackknife estimator outperforms other estimators in terms of bias, RMSE and the coverage probability of the 95% confidence interval especially when $M = 200, 400$. For example, when $M = 400$, the lower bound is estimated by the Chao1 estimator as mean 169 species. The results show the best approximation is mean 193.32 by the second-order jackknife with RMSE=17.38.

For the log-normal abundance model, the first-order jackknife, the second-order jackknife and the new_2 estimators underestimate true number of species for small M before overestimating for large M. For other estimators, they underestimate for small M and tend to true number of species for large M (Figure 2.8d). Table 2.6, the iChao1 and the new_1 estimators have a similar results and approximate well for large M. For example with $M = 1600$, the results show the Chao1 estimator gives the lower bound of species richness as mean $\widehat{N}_{Chao1} = 196.82$, $\widehat{N}_{iChao1} = 199.17$ and $\widehat{N}_{\text{new}_1} = 199.27$. The first-order jackknife, the second-order jackknife and the new_2 overestimate with mean $\widehat{N}_{J1} = 206.47$, $\widehat{N}_{J2} = 206.57$ and $\widehat{N}_{\text{new}_2} = 206.59$.

For the Zipf-Mandelbrot and the power-decay, the results show the new_1, the new_2 and the iChao1 estimators perform similarly and better than other estimators when M is small (Figures 2.8e and 2.8f). In Tables 2.5 and 2.7, the new_1 estimator approximate better than other estimators for small sample size with $M = 200$ while the Chao1 estimator yields the estimated species richness very well for large M. For example with the Zipf-Mandelbrot model, the new_1 estimator gives the best results with mean $\widehat{N}_{\text{new}_1} = 172.46$ and the coverage

Table 2.3: Comparison of the mean of species richness estimators based on the homogeneous model $p_i = 1/N$ with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	126.54	Jackknife1	**200.25**	10.91	12.14	**10.91**	**96.94**
		Jackknife2	237.16	19.78	21.03	42.09	42.57
		Chao1	202.74	23.83	24.22	23.99	94.40
		iChao1	210.21	29.05	32.30	30.79	91.60
		new_1	221.31	28.99	29.75	35.98	86.30
		new_2	243.69	24.84	26.90	50.26	47.58
400	173.12	Jackknife1	227.28	8.78	10.41	28.65	88.20
		Jackknife2	227.15	16.91	18.03	31.98	44.81
		Chao1	**201.03**	10.05	10.20	**10.10**	**94.37**
		iChao1	203.49	12.16	12.87	12.65	92.26
		new_1	205.93	12.81	13.20	14.12	88.40
		new_2	233.15	9.93	11.99	34.60	35.20
800	196.37	Jackknife1	210.93	4.09	5.40	11.67	12.35
		Jackknife2	**199.97**	5.85	9.35	5.85	37.49
		Chao1	200.37	2.98	3.11	**3.00**	88.29
		iChao1	200.83	3.45	3.74	3.55	86.98
		new_1	200.52	3.27	3.54	3.32	**88.36**
		new_2	213.84	3.15	5.12	14.20	11.60
1600	199.93	Jackknife1	200.46	0.77	1.03	0.90	56.88
		Jackknife2	200.32	0.98	1.78	1.03	**75.53**
		Chao1	**200.16**	0.64	1.10	**0.66**	60.01
		iChao1	200.19	0.69	1.18	0.72	59.99
		new_1	200.20	0.75	1.25	0.77	56.89
		new_2	201.10	0.84	1.55	1.39	14.89

probability of the 95% interval is 0.8628, while the Chao1 estimator has mean $\widehat{N}_{Chao} = 199.35$ and the coverage probability of the 95% interval is 0.9455.

Table 2.4: Comparison of the mean of species richness estimators based on the negative binomial (4, 0.04) model with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	118.35	Jackknife1	183.98	10.72	11.46	**19.28**	81.36
		Jackknife2	216.61	18.91	19.84	25.17	81.45
		Chao1	186.06	22.57	22.90	26.52	90.51
		iChao1	195.85	27.74	30.45	28.05	93.10
		new_1	**202.50**	27.13	27.79	27.25	**95.15**
		new_2	218.47	22.98	24.41	29.48	87.60
400	160.66	Jackknife1	213.13	9.43	10.24	16.16	65.64
		Jackknife2	221.63	17.09	17.74	27.57	62.09
		Chao1	193.20	12.15	12.04	**13.92**	92.06
		iChao1	198.74	15.09	15.54	15.15	93.20
		new_1	**200.45**	15.05	15.17	15.05	**94.64**
		new_2	218.96	11.81	12.94	22.34	56.57
800	187.40	Jackknife1	212.12	5.93	7.03	13.50	35.73
		Jackknife2	206.05	11.39	12.18	12.90	75.94
		Chao1	197.98	5.87	5.91	**6.21**	**94.43**
		iChao1	**199.90**	7.02	7.30	7.02	93.19
		new_1	199.54	7.21	7.49	7.22	92.61
		new_2	212.60	5.60	6.92	13.79	29.67
1600	197.35	Jackknife1	203.83	2.92	3.60	4.82	52.12
		Jackknife2	200.89	4.56	6.24	4.65	46.67
		Chao1	199.82	2.76	3.08	**2.76**	**89.72**
		iChao1	200.26	3.13	3.55	3.14	87.88
		new_1	**200.14**	3.21	3.65	3.21	88.69
		new_2	204.71	2.65	3.80	5.40	39.72

Table 2.5: Comparison of the mean of species richness estimators based on the power decay model $p_i = c/i^{1.2}$ with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	59.73	Jackknife1	95.58	9.49	8.47	104.85	-
		Jackknife2	121.54	15.04	14.66	79.89	63.00
		Chao1	134.06	41.31	44.48	77.81	61.34
		iChao1	144.95	45.73	50.14	71.57	**73.75**
		new_1	**145.33**	43.68	46.76	**69.98**	70.77
		new_2	140.52	40.36	43.72	71.88	61.75
400	88.47	Jackknife1	135.26	10.65	9.67	65.61	4.00
		Jackknife2	165.68	16.95	16.76	**38.28**	62.28
		Chao1	161.04	31.40	32.13	50.04	75.28
		iChao1	173.06	35.73	38.21	44.75	**86.54**
		new_1	**174.01**	34.02	34.82	42.81	86.38
		new_2	170.32	30.35	31.27	42.45	78.53
800	123.43	Jackknife1	176.39	10.87	10.29	26.00	51.97
		Jackknife2	**203.95**	17.63	17.82	**18.07**	93.33
		Chao1	181.88	22.41	22.47	28.82	86.40
		iChao1	193.11	26.35	27.22	27.23	93.50
		new_1	194.30	25.06	25.33	25.70	**94.53**
		new_2	194.78	21.36	21.74	21.99	92.93
1600	159.18	Jackknife1	206.73	9.31	9.75	**11.49**	85.54
		Jackknife2	221.05	16.02	16.89	26.46	62.56
		Chao1	194.93	13.77	14.01	14.68	93.86
		iChao1	**201.92**	16.77	17.60	16.88	93.45
		new_1	203.43	16.18	16.68	16.54	**93.99**
		new_2	210.60	12.88	13.72	16.68	85.40

Table 2.6: Comparison of the mean of species richness estimators based on the log-normal(0,1) model with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	97.91	Jackknife1	150.04	10.44	10.21	51.04	12.00
		Jackknife2	**179.12**	17.55	17.69	**27.28**	**87.58**
		Chao1	160.26	24.58	24.52	46.73	67.94
		iChao1	171.54	29.20	30.30	40.77	83.90
		new_1	174.44	28.29	28.40	38.12	85.72
		new_2	178.90	24.05	24.42	32.00	83.53
400	133.92	Jackknife1	184.12	10.02	10.02	18.78	76.00
		Jackknife2	**203.51**	17.04	17.35	17.39	**93.50**
		Chao1	176.88	16.61	16.50	28.46	75.89
		iChao1	185.59	19.87	20.32	24.54	90.25
		new_1	187.08	19.48	19.60	23.37	91.41
		new_2	194.70	15.88	16.36	**16.74**	93.03
800	164.59	Jackknife1	**202.01**	8.37	8.65	**8.61**	93.22
		Jackknife2	209.27	14.68	14.98	17.36	83.68
		Chao1	189.19	11.08	10.90	15.48	85.72
		iChao1	194.27	13.21	13.36	14.39	93.22
		new_1	194.91	13.28	13.30	14.22	**95.04**
		new_2	203.90	10.44	10.97	11.14	93.29
1600	184.98	Jackknife1	206.47	5.99	6.56	8.82	72.20
		Jackknife2	206.57	11.02	11.36	12.84	78.50
		Chao1	196.82	7.07	7.08	**7.75**	92.84
		iChao1	199.17	8.32	8.55	8.36	**93.22**
		new_1	**199.27**	8.57	8.74	8.60	91.72
		new_2	206.59	6.55	7.36	9.29	77.18

Table 2.7: Comparison of the mean of species richness estimators based on the Zipf-Mandelbrot model $p_i = c/(i - 0.1), i = 1, 2, \ldots, N$ with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	75.81	Jackknife1	121.57	10.22	9.57	79.09	-
		Jackknife2	153.32	16.58	16.57	49.54	35.11
		Chao1	157.93	38.01	38.86	56.70	75.79
		iChao1	170.74	43.13	46.71	52.12	85.96
		new_1	**172.46**	41.16	42.03	49.52	**86.28**
		new_2	168.32	36.91	38.03	**48.64**	79.40
400	111.25	Jackknife1	166.61	10.92	10.52	35.13	23.26
		Jackknife2	**198.19**	17.87	18.23	**17.97**	**95.91**
		Chao1	179.52	26.01	26.13	33.10	86.48
		iChao1	191.98	30.45	31.91	31.49	93.47
		new_1	193.81	29.03	29.35	29.69	94.51
		new_2	193.69	24.94	25.40	25.72	92.86
800	149.80	Jackknife1	203.20	10.09	10.33	**10.58**	92.35
		Jackknife2	222.32	17.21	17.90	28.18	63.43
		Chao1	193.30	16.03	16.09	17.37	92.89
		iChao1	**201.68**	19.47	20.32	19.54	93.57
		new_1	203.63	18.71	19.06	19.06	**94.22**
		new_2	210.74	15.12	15.77	18.55	88.07
1600	181.25	Jackknife1	216.03	7.18	8.34	17.57	30.46
		Jackknife2	215.47	13.68	14.45	20.65	61.83
		Chao1	**199.35**	8.16	8.34	**8.19**	**94.55**
		iChao1	202.36	9.96	10.56	10.24	91.75
		new_1	203.06	10.13	10.53	10.58	90.32
		new_2	215.74	7.55	8.85	17.46	33.24

Table 2.8: Comparison of the mean of species richness estimators based on the broken-stick model (or Dirichlet$(1, 1, \ldots, 1)$) with $N = 200$ and 10000 simulations.

M	K	Estimator	$\widehat{N}$	$Se(\widehat{N})$	$\widehat{Se(\widehat{N})}$	RMSE	95% CI coverage
200	99.34	Jackknife1	149.12	10.08	9.98	51.87	79.00
		Jackknife2	**174.46**	17.13	17.28	**30.75**	**81.44**
		Chao1	152.93	21.20	21.22	51.63	53.40
		iChao1	162.92	25.20	26.29	44.83	75.35
		new_1	165.44	24.75	24.94	42.50	76.33
		new_2	172.06	20.79	21.33	34.83	74.02
400	132.72	Jackknife1	177.74	9.33	9.49	24.14	51.49
		Jackknife2	**193.32**	16.05	16.43	**17.38**	**97.29**
		Chao1	169.05	14.84	14.74	34.32	58.19
		iChao1	176.57	17.69	17.95	29.36	80.60
		new_1	177.68	17.51	17.62	28.37	81.37
		new_2	185.58	14.13	14.68	20.19	83.10
800	159.66	Jackknife1	191.91	7.78	8.03	**11.22**	91.05
		Jackknife2	**197.74**	13.78	13.91	13.96	**96.97**
		Chao1	180.74	10.36	10.13	21.87	65.85
		iChao1	185.04	12.28	12.30	19.35	83.60
		new_1	185.59	12.41	12.33	19.02	86.06
		new_2	193.46	9.72	10.19	11.72	90.50
1600	177.09	Jackknife1	196.07	5.88	6.16	**7.07**	95.60
		Jackknife2	**197.73**	10.41	10.67	10.66	**95.91**
		Chao1	188.66	7.41	7.42	13.55	75.76
		iChao1	190.98	8.55	8.76	12.42	86.50
		new_1	191.16	8.82	8.99	12.49	89.74
		new_2	196.54	6.91	7.52	7.73	94.08

It is concluded that the Chao1 estimator works very well for homogeneous model. The new_1 and the iChao1 estimator perform similarly and are appropriate for the negative binomial, the Zipf-Mandelbrot and the power-decay abundance model. The second-order jackknife and the new_2 estimator approx-

imates well for the broken-stick model.

2.8 Real Data Examples

The species richness estimators that we consider in this Chapter are applied to some real data sets including Malaysian butterfly data (Fisher et al., 1943), the Pollutants data (Janardan and Schaeffer, 1981), Christmas bird count data (Chao and Bunge, 2002), Bangkok heroin users (Lanumteang and Böhning, 2011), beetle species abundance frequency counts data (Janzen, 1973) and species abundance frequency counts data for tree samples (Norden et al., 2009). The results are shown in Table 2.9.

The results indicate that the iChao1 and the new_1 estimators for N are the same results for Malayan butterfly, Christmas bird, heroin users, Tropical tree1, Tropical tree2 and Tropical tree3 data. When considering the Chao1 estimator as a lower bound, the first-order Jackknife estimator estimates lower than the Chao1 estimator in many data sets (e.g. pollutants, beetle site1, beetle site2, Tropical trees2 and tropical tree3).

2.9 Conclusion

It is clear that the performance of each estimator depends on species abundance model and sample size. Their performance improves when the sample size tends to infinity. In ecology, the heterogeneous population is considered in practice for species richness estimation. The iChao1 estimator is an improved the Chao1 estimator which can reduce bias and perform well especially under highly heterogeneous abundance. However, it requires more informations than the Chao1 estimator by adding the number of tripletons and quadrupletons in order to construct the iChao1 estimator.

In this thesis, we developed an alternative improvement to the Chao1 estimator. The performance of the new_1 is similar to the iChao1 estimator, but the new_1 estimator requires less information than the iChao1 estimator. The new_1 performs well with the negative binomial, the power-decay and the Zipf-Mandelbrot. The new_2 estimator approximates better than both the iChao and the new_1 estimators for the broken-stick model. When sample size is large, the new_2 estimator gives similar results to the first-order jackknife and the second-order jackknife estimator.

Although there is a larger variance of the higher order of jackknife, the second-order jackknife gives a better bias than the first-order jackknife estimator. In our simulation study, the second-order jackknife estimator was found to work well with the broken-stick.

Problem about estimating species richness has been investigated using non-parametric approaches. In this Chapter, we have investigated nonparametric estimation of species richness. In the next chapter, we consider the some problem using the maximum likelihood estimation in a parametric approach. The distribution of the number of individuals seen for species i have been investigated. Mixed Poisson models have been considered for the maximum likelihood estimation based on the heterogeneous model.

Table 2.9: Comparison of six estimators of total number for real data sets.

M	K	Estimator	$\widehat{N}$	$\widehat{Se(\widehat{N})}$	LC	UC
Malayan Butterfly						
9031	620	Jackknife1	738	15.36	711.51	772.12
		Jackknife2	782	26.60	737.65	843.02
		Chao1	714	22.66	679.06	769.85
		iChao1	737	28.81	692.78	808.30
		new_1	737	26.91	695.01	802.61
		new_2	745	21.90	711.58	798.42
Pollutants						
5720	1258	Jackknife1	1761	31.71	1702.50	1827.00
		Jackknife2	2026	54.92	1925.54	2141.26
		Chao1	1789	62.96	1679.66	1927.80
		iChao1	1916	74.13	1786.22	2078.25
		new_1	1910	72.50	1782.34	2067.89
		new_2	1917	60.79	1807.77	2046.87
Christmas Bird						
20042	126	Jackknife1	138	4.90	131.56	151.90
		Jackknife2	141	8.48	131.34	168.13
		Chao1	134	6.04	128.15	155.82
		iChao1	136	7.03	128.82	160.63
		new_1	136	7.38	128.75	162.41
		new_2	138	5.93	130.81	155.98
Heroin users						
39086	9302	Jackknife1	11478	65.97	11352.44	11611.13
		Jackknife2	12054	114.26	11838.97	12287.16
		Chao1	10782	82.90	10627.87	10953.24
		iChao1	11151	100.60	10964.14	11358.96
		new_1	11151	102.14	10961.41	11362.28
		new_2	11579	82.77	11422.49	11747.13
Beetle Site1						
976	140	Jackknife1	210	11.82	190.32	237.18
		Jackknife2	263	20.47	228.81	309.90
		Chao1	284	50.47	213.87	420.60
		iChao1	297	52.52	222.56	436.92
		new_1	305	53.57	228.44	446.62
		new_2	293	49.23	222.46	422.85
Beetle Site2						
237	112	Jackknife1	196	12.92	173.91	225.02
		Jackknife2	269	22.33	231.03	319.24
		Chao1	463	136.27	280.66	843.78
		iChao1	489	139.78	298.39	873.63
		new_1	501	141.58	306.91	888.61
		new_2	474	134.33	291.00	843.83

M	K	Estimator	$\widehat{N}$	$\widehat{Se(\widehat{N})}$	LC	UC
Tropical trees1						
943	152	Jackknife1	198	9.58	182.67	220.86
		Jackknife2	214	16.59	188.99	255.76
		Chao1	187	13.58	168.98	225.08
		iChao1	196	16.21	173.77	240.45
		new_1	196	16.35	173.72	241.03
		new_2	202	13.29	181.91	235.38
Tropical trees2						
1263	104	Jackknife1	137	8.12	124.50	157.05
		Jackknife2	155	14.06	133.97	190.64
		Chao1	140	16.84	119.26	190.21
		iChao1	147	24.57	119.37	225.97
		new_1	148	19.05	123.56	203.17
		new_2	147	16.17	125.32	191.91
Tropical trees3						
1020	76	Jackknife1	105	7.61	93.46	124.07
		Jackknife2	121	13.17	101.61	154.90
		Chao1	108	16.02	88.89	156.99
		iChao1	116	20.44	91.76	178.93
		new_1	115	18.10	92.55	168.78
		new_2	114	15.39	94.04	157.84

Chapter 3

Estimating the number of species using maximum likelihood estimation

3.1 Introduction

Maximum likelihood estimation (MLE) is a parametric approach which has a long history in statistics for estimating unknown parameters. For species richness estimation, the MLE has been used for estimating the number of species. If the abundance of each species is the same, the Poisson model is used to construct the likelihood function. In practice, there is usually heterogeneity in the abundance of different species that leads to an overdispersed model, in which the variance exceeds the mean (Bunge and Barger, 2008). Mixed Poisson models have been proposed for this issue, which is discussed in Section 3.2.

Fisher et al. (1943), Bunge and Barger (2008) and Cruyff and van der Heijden (2008) proposed alternative models for overdispersion and heterogeneous data. For species richness estimation, when the number of unseen species is unknown, the zero truncated model based on the mixed model (e.g. gamma-Poisson) is

used to construct the likelihood function. In Section 3.3, the MLE based on the zero truncated mixed Poisson model is investigated. Although the MLE approach is the preferred method of estimation in statistics in general, in species richness estimation problems of nonconvergence and the so-called boundary problem can arise in practice, particularly when the sample size is small. These problems with the MLE approach are reviewed in Section 3.4. Finally, conclusions are summarized in Section 3.5.

3.2 Mixed Poisson Models

Estimating the number of species has been discussed in many studies (e.g. Fisher et al. (1943), Chao (1984), Bunge and Barger (2008) and Colwell et al. (2012)). It is similar to the problem of the population size estimation in other fields such as estimating the size of vocabulary in linguistics (Hidaka, 2014), estimating the number of drug users in social science such as Hay and Smit (2003) and Lanumteang and Böhning (2011).

In ecological applications, the population is divided into N groups, just as individual plants and animals are classified into N species. The simplest assumption is that the number of individuals of each species which were found in the sample or trap or region of interest follows a Poisson distribution (Fisher et al., 1943). The Poisson is the appropriate distribution for discrete count data that result from a process of random and independent incidents that occur in a fixed time period and a limited area of space (Valero et al., 2010). For example, this might apply to the number of moths of a given species that enter a light trap during one night.

One of the features of the Poisson distribution is equality of the mean and variance. In practice, the variance of observed species counts usually exceeds

the mean as a consequence of variability or heterogeneity between species of the parameter of the Poisson distribution λ, the expected number of observed individuals seen for each species in the sampled region. When the variance exceeds the mean, the distribution is called 'overdispersed'. If the homogeneous Poisson model is fitted when the data are overdispersed, it leads to underestimation of the number of species (Cruyff and van der Heijden, 2008).

To accommodate heterogeneity, Fisher et al. (1943) proposed using a gamma mixed Poisson distribution which is also known as the negative binomial distribution. It has been applied by many researchers including Grogger and Carson (1991), Cruyff and van der Heijden (2008), Bunge and Barger (2008), Rocchetti et al. (2011) and Vergne et al. (2012).

However, there are many other mixed Poisson models that can be used. Bunge and Barger (2008) compared a variety of mixed Poisson models for estimating the number of species, specifically the standard (unmixed) Poisson, the gamma mixed Poisson, the lognormal mixed Poisson, the inverse Gaussian mixed Poisson, the Pareto mixed Poisson, the exponential mixed Poisson and mixtures of two or three exponential mixed Poisson distributions.

The variation in the behaviour of organisms in each species and/or heterogeneity in species abundance (i.e. some species are abundant while some species are rare), has an effect on the overdispersion. For a given sampling effort, a simple assumption is that the number of individuals seen for species i follows a Poisson distribution with a single parameter, λ_i, which is the discovery rate for species i (i.e. the average of the number of individuals seen for the species during the sampling period). Variability of the Poisson parameter between species reflects differences in the species abundance, but may also reflect the difficulty of sampling the species because some species may be more difficult

to detect.

We use the following notation:

N	the true number of species present
X_i	the number of individuals seen for species i $(i = 1, \ldots, N)$
λ_i	the Poisson parameter or discovery rate of species i $(i = 1, \ldots, N)$
M	the total number of individuals observed $(M = \sum_{i=1}^{N} X_i)$
f_k	the number of species seen k times $(k = 0, 1, \ldots, M)$
K	the number of distinct species seen $(K = \sum_{i=1}^{M} f_k)$

Although the discovery rates $\lambda_1, \lambda_2, ..., \lambda_N$, are expected to vary between species, there is insufficient data for each species, including an unknown frequency of undetected species, to make any progress with estimating a separate parameter λ_i for each species. Therefore, we define the parameters λ_i to be random variables which are based on some distribution with probability density function $f(\lambda)$. This means that the discovery rates $\lambda_1, \lambda_2, ..., \lambda_N$, are a random sample from this distribution and are treated as independent and identically distributed random variables.

Let X_i denote the number of individuals seen for species i, for $i = 1, 2, \ldots N$. Then, conditional on λ_i

$$X_i|\lambda_i \sim \text{Pois}(\lambda_i)$$

and the probability mass function (pmf) of X_i given by

$$\Pr(X_i = x|\lambda_i) = f(x|\lambda_i) = \frac{e^{-\lambda_i}\,\lambda_i^x}{x!}.$$

If λ_i is a random variable with the probability density function $f(\lambda;\theta)$, the marginal distribution of X_i is called a mixed Poisson distribution, where the

distribution of λ_i is termed the mixing distribution. Then, the mixed Poisson model has probability density function as

$$\Pr(X = x) = \int_0^\infty \frac{e^{-\lambda_i}\ \lambda_i^x}{x!}\ f(\lambda;\theta)d\lambda, \qquad x = 0, 1, \ldots$$

The probability for unseen species is

$$\Pr(X = 0) = \int_0^\infty e^{-\lambda}\ f(\lambda;\theta)d\lambda.$$

The mean of the mixed Poisson distribution is given by

$$\mathbb{E}_\lambda[\lambda] = \int_0^\infty \lambda f(\lambda;\theta)d\lambda = \mu.$$

Under the overdispersed model, $\Pr(X = 0) \geqslant e^{-\mu}$. Then,

$$\int_0^\infty f(\lambda;\theta)e^{-\lambda}d\lambda \geqslant e^{(-\int_0^\infty \lambda f(\lambda;\theta)d\lambda)}. \tag{3.1}$$

(Böhning and Schön, 2005).

In Figure 3.1, the empirical probability mass function is compared with the Poisson distribution. 400 individuals are selected with replacement from the population consisting of 200 species using the broken-stick abundance model. The results show the smaller value for the Poisson distribution when compared to the true probability with 0.1353 and 0.310 respectively. Therefore, using the Poisson distribution for abundance data is not appropriate. This leads to underestimation of the number of species.

The mixed Poisson distribution is considered to improve the performance of the Poisson distribution for overdispersed data. The Poisson-gamma distribution is a common mixed Poisson model known as the negative binomial

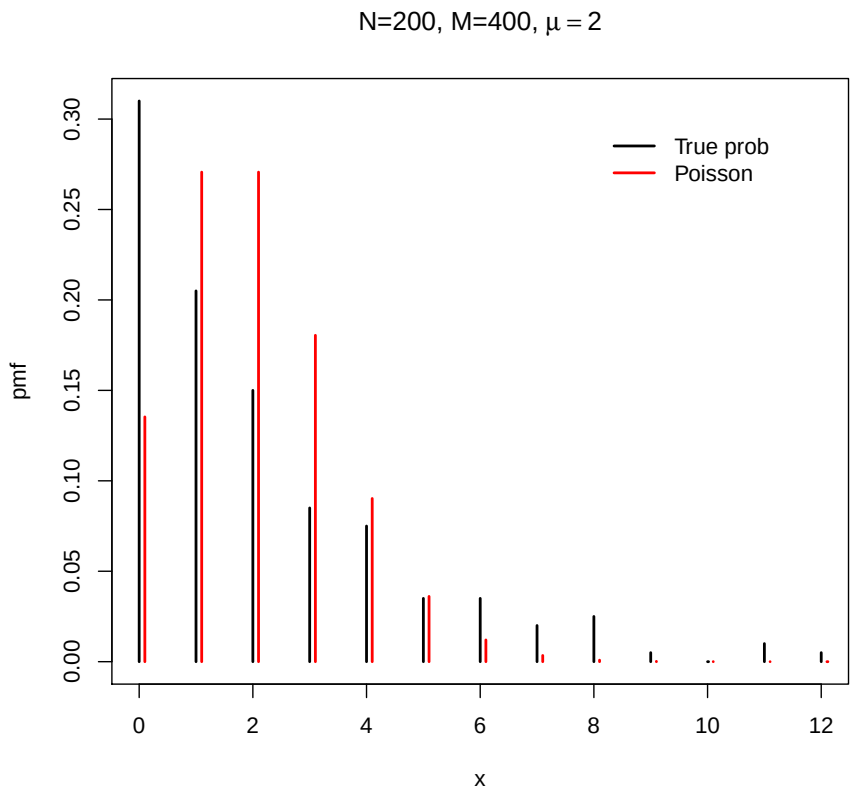


Figure 3.1: Plot of probability mass function under the overdispersed data with $N = 200$, $M = 400, \mu = 2$ and the estimated probability from the Poisson distribution with mean=2.

distribution. It has the variance grater than mean which is suitable for the overdispersed data (Cruyff and van der Heijden, 2008).

When λ_i is generated by the gamma distribution with shape parameter α and rate parameter β, the probability density function (pdf) of species discovery rate is given by

$$f(\lambda_i; \alpha, \beta) = \frac{\beta^{\alpha}\ \lambda_i^{\alpha-1}\ e^{-\lambda_i \beta}}{\Gamma(\alpha)}, \qquad \lambda_i > 0, \alpha > 0, \beta > 0,$$

and the probability mass function of the gamma mixed Poisson distribution is

$$
\begin{aligned}
Pr(X_i = x) &= \int_0^\infty f(x, \lambda_i)\, d\lambda_i \\
&= \int_0^\infty f(x|\lambda_i)\, f(\lambda; \alpha, \beta_i)\, d\lambda_i \\
&= \int_0^\infty \frac{e^{-\lambda_i}\, \lambda_i^x}{x!}\, \frac{\beta^\alpha\, \lambda_i^{\alpha-1}\, e^{-\lambda_i \beta}}{\Gamma(\alpha)}\, d\lambda_i \\
&= \frac{\beta^\alpha}{x!\, \Gamma(\alpha)} \int_0^\infty \lambda_i^{x+\alpha-1}\, e^{-\lambda_i(1+\beta)} d\lambda_i \\
&= \frac{\beta^\alpha}{x!\, \Gamma(\alpha)}\, \frac{\Gamma(x+\alpha)}{(1+\beta)^{x+\alpha}}
\end{aligned}
$$

This can be written as (Cruyff and van der Heijden, 2008)

$$
Pr(X_i = x) = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\, \Gamma(\alpha)}\, \left(\frac{\beta}{1+\beta}\right)^\alpha\, \left(\frac{1}{1+\beta}\right)^x . \tag{3.2}
$$

This is the negative binomial distribution with mean $\mu = \dfrac{\alpha}{\beta}$ and variance $\dfrac{\alpha(1+\beta)}{\beta^2} = \mu + \mu^2/\alpha$, where the parameter α refers to the heterogeneity of the parameter λ in the Poisson process. The smaller the value of α, the more heterogeneity in the population when the mean is kept fixed. There are many studies that have used the negative binomial model to estimate the number of species. For example, fitting the model to counts of Malaysian butterflies, it was found that this model can work well (Fisher et al., 1943). Although the gamma mixed Poisson model can perform well for heterogeneous data, it might give a biased estimator that overestimates N when the mean count, μ, is small (Cruyff and van der Heijden, 2008).

3.3 Maximum likelihood estimation based on zero-truncated Mixed-Poisson distribution

Assume that Θ represents the unknown parameters of the mixture distribution. Thus, for the negative binomial, this is a vector of two parameters, α and β. Let f_k denote the number of species seen k times for $k = 0, 1, 2, \ldots, M$. Therefore, the number of observed species is $K = \Sigma_{k=1}^{M} f_k$. We can write the multinomial likelihood function for θ and N as (Chao and Bunge, 2002)

$$L(N, \theta) = \frac{N!}{(N-K)! \prod_{k \geqslant 1} f_k!} P_0(\theta)^{N-K} \prod_{k \geqslant 1} (P_k(\theta))^{f_k},$$

which is the full likelihood for N and θ, where $P_0(\theta)$ refers to the probability of not observing a species and $P_k(\theta)$ is the probability of observing a species k times for $k \geqslant 1$. This likelihood function can be partitioned as $L(N, \theta) = L_1(N, \theta) \times L_2(\theta)$, where the first term can be written as a binomial likelihood

$$L_1(N, \theta) = \binom{N}{K} P_0\ (\theta)^{N-K} \left[1 - P_0(\theta)\right]^K$$

and the second term formulated as a multinomial likelihood given by

$$L_2(\theta) = \frac{K!}{\prod_{k \geqslant 1} f_k!} \prod_{k \geqslant 1} \left(\frac{P_k(\theta)}{1 - P_0(\theta)} \right)^{f_k},$$

which is the multinomial likelihood based on the zero-truncated distribution for the number of species seen k times, conditional on $k \geqslant 1$ (Chao and Bunge, 2002).

Since K and f_k are known

$$L_2(\theta) \propto \prod_{k \geqslant 1} \left(\frac{P_k(\theta)}{1 - P_0(\theta)} \right)^{f_k}$$

for computational work.

Sanathanan (1977) proposed two types of maximum likelihood estimators used to estimate N, known as the unconditional and the conditional MLEs. The unconditional MLE maximizes the full likelihood $L(N, \theta)$ as a function of θ and N, which obtains $\widehat{\theta}$ and $\widehat{N}$. The conditional MLE maximizes $L_2(\theta)$ first, for finding $\widehat{\theta}$, and then maximizes $L_1(N, \widehat{\theta})$ for finding the estimator of N. This yields the estimator of N which is given by $\widehat{N}_H = \dfrac{K}{1 - P_0(\widehat{\theta})}$ and known as the Horvitz-Thompson estimator. The conditional approach has been often used, as a consequence of easier calculation and the fact that it usually gives very similar values of $\widehat{N}$ (Sanathanan, 1977).

As the number of unseen species is unknown, the idea of the zero-truncated model is considered. The simplest such model is the zero-truncated Poisson (ZTP), which is appropriate when the λ_i are homogeneous ($\lambda_i = \lambda$) that leads to the equality of the mean and variance in the non-truncated distribution. For the heterogeneous case, where λ_i are not assumed equal, the zero-truncated mixed Poisson distributions, such as the zero-truncated negative binomial (ZTNB) are used for fitting the model, to incorporate the overdispersion (van der Heijden et al., 2003).

The probability function of a the zero truncated distribution is given by

$$Pr(X_i = x | X_i > 0,\ \theta) = \frac{P_x(\theta)}{1 - P_0(\theta)},$$

where θ denotes the parameter(s) of the untruncated distribution, and the likelihood function for the K observed species is

$$L_2(\theta) \propto \prod_{i=1}^{K} \frac{P_x(\theta)}{1 - P_0(\theta)} = \prod_{k=1}^{M} \left\{ \frac{P_k(\theta)}{1 - P_0(\theta)} \right\}^{f_k}$$

Then the log-likelihood for the zero-truncated distribution can be written as

$$\begin{aligned}
\log\ L_2(\theta) &= C + \sum_{k=1}^{M} f_k \log\left[\frac{P_k(\theta)}{1-P_0(\theta)}\right] \\
&= C + \sum_{k=1}^{M} f_k \left[\log\ P_k(\theta) - \log\left(1-P_0(\theta)\right)\right] \\
&= C + \sum_{k=1}^{M} f_k \log\ P_k(\theta) - \left(\sum_{k=1}^{M} f_k\right)\ \log\left(1-P_0(\theta)\right) \\
&= C + \sum_{k=1}^{M} f_k \log\ P_k(\theta) - K\ \log\left(1-P_0(\theta)\right), \qquad (3.3)
\end{aligned}$$

where C is a constant and $P_k(\theta)$ is the probability mass function of the untruncated distribution. For the zero truncated Poisson model, $P_0(\theta) = e^{-\lambda_i}$ and $P_k(\theta) = \dfrac{e^{-\lambda_i}\lambda_i^k}{k!}$. For the zero truncated negative binomial model, we have $P_0(\theta) = \left(\dfrac{\beta}{1+\beta}\right)^{\alpha}$ and $P_k(\theta) = \dfrac{\Gamma(k+\alpha)}{\Gamma(k+1)\,\Gamma(\alpha)}\ \left(\dfrac{\beta}{1+\beta}\right)^{\alpha}\ \left(\dfrac{1}{1+\beta}\right)^{k}$.

In the conditional likelihood approach to estimating N, the zero-truncated model is fitted based on the observed counts and then the point estimator of the total number of species can be written as

$$\widehat{N}_H = \frac{K}{1-P_0(\widehat{\theta})} \qquad (3.4)$$

where $\widehat{\theta}$ (i.e. $\widehat{\alpha}$ and $\widehat{\beta}$ for the zero-truncated negative binomial) refers to the MLE of the parameters of the zero-truncated model. If the number of observed species is large, the estimated total number of species $\widehat{N}_H$ and the estimated parameters of the abundance model, $\widehat{\theta}$, are approximately unbiased estimators (van der Heijden et al., 2003).

The variance of $\widehat{N}_H$ can be estimated by

$$\mathrm{Var}(\widehat{N}_H) = \mathrm{E}[\mathrm{Var}(\widehat{N}_H|I_i)] + \mathrm{Var}(\mathrm{E}[\widehat{N}_H|I_i]) \qquad (3.5)$$

which is formulated using the law of total variance (van der Heijden et al., 2003).

For example with the zero truncated negative binomial (ZTNB) model with parameters α and β, the first term can be estimated using the delta method by

$$\widehat{\text{Var}}(\widehat{N}_H|I_i) = a(\widehat{\theta})^T J(\widehat{\theta})^{-1} a(\widehat{\theta}) \mid_{\widehat{\theta}}$$

where $a' = \left(\frac{\partial \widehat{N}_H}{\alpha}, \frac{\partial \widehat{N}_H}{\beta}\right)$ and $J(\widehat{\theta}) = -\begin{pmatrix} \frac{\partial^2 \ell(\alpha,\beta)}{\partial\beta\partial\beta'} & \frac{\partial^2 \ell(\alpha,\beta)}{\partial\beta\partial\alpha} \\ \frac{\partial^2 \ell(\alpha,\beta)}{\partial\alpha\partial\beta'} & \frac{\partial^2 \ell(\alpha,\beta)}{\partial\alpha^2} \end{pmatrix}$ and the second term can be defined as

$$\text{Var}(E[\widehat{N}_H|I_i]) = \sum_{i=1}^{N} I_i \, \frac{1 - \Pr(Y_i > 0)}{[\Pr(Y_i > 0)]^2},$$

which may be estimated as

$$\widehat{\text{Var}}(E[\widehat{N}_H|I_i]) = K \times \frac{P_0(\widehat{\theta})}{[1 - P_0(\widehat{\theta})]^2}$$

(van der Heijden et al., 2003).

We assume that $\log(\widehat{N} - K)$ is normally distributed when M is large and $\widehat{N} > K$, following Chiu et al. (2014). Therefore, an approximate 95% confidence interval for N is given by

$$[K + (\widehat{N} - K)/R, K + (\widehat{N} - K)R],$$

where $R = \exp\left\{1.96[1 + \widehat{Var}(\widehat{N})/(\widehat{N} - K)^2]^{1/2}\right\}$. Since $\widehat{N}$ is not normally distributed, this formulae is used for finding confidence intervals for $\widehat{N}$ instead.

Table 3.1: Estimated N, estimated standard error of N, $\widehat{Se}(\widehat{N})$, 95% confidence interval of N and AIC criterion.

	$\widehat{N}$	$\widehat{Se}(\widehat{N})$	LC	UC	AIC
Butterfly (K=620, M=9031)					
Chao	714	22.66	670	758	-
Poisson	621	0.92	620	625	4362
NB	913	17.13	882	949	2792
Pollutant (K=1258, M=5720)					
Chao	1789	62.96	1680	1928	-
Poisson	1299	6.41	1288	1314	7445
NB	*	*	*	*	*
Christmas Bird (K=126, M=20042)					
Chao	134	6.04	128	156	-
Poisson	128	1.38	127	133	274
NB	154	5.28	145	166	239
Heroin (K=9302, M=39086)					
Chao	10782	82.90	10628	10953	-
Poisson	9453	12.30	9431	9479	50092
NB	11581	47.74	11489	11677	42870
Beetle site1 (K=140, M=976)					
Chao	284	50.47	214	421	-
Poisson	152	3.47	147	161	579
NB	*	*	*	*	*
Beetle site2 (K=112, M=237)					
Chao	463	136.27	282	845	-
Poisson	168	7.50	155	185	284
NB	*	*	*	*	*
Tropical tree1 (K=152, M=943)					
Chao	187	13.58	169	225	-
Poisson	160	2.90	156	168	596
NB	258	10.28	239	280	517
Tropical tree2 (K=104, M=1263)					
Chao	140	16.84	119	190	-
Poisson	106	1.31	104	110	598
NB	323	14.81	296	354	421
Tropical tree3 (K=76, M=1020)					
Chao	108	16.02	89	157	-
Poisson	104	5.33	96	117	133
NB	159	9.11	143	179	130

Note: * is optimization fail.

Table 3.1 displays the ZTP and ZTNB models applied to several real data sets. The 95 % confidence interval of N is computed. Akaike Information Criterion (AIC) is used to select the best model, which is calculated by 2(no.parameter)-ln(L), where L is the maximum value of the likelihood function. Considering the Chao1 estimator as a lower bound of N, the results indicated that ZTP model underestimates for all data set. The ZTNB model performs better than ZTP with smaller AIC. For example with heroin data, the Chao1 estimator estimates $\widehat{N}_C = 10782$ while the ZTP model underestimates with $\widehat{N}_P = 9453$ and $\text{AIC}_P = 50092$. The ZTNB model yields $\widehat{N}_{NB} = 11581$ and performs better with $\text{AIC}_{NB} = 42870$. However, the ZTNB model cannot be used to estimate N in some data sets including Pollutant, Beetle site1 and site2. There is a numerical problem in optimization. This yields very large estimated N for the ZTNB model.

3.4 Problems with maximum likelihood estimation

In statistics, it is generally accepted that the MLE is the preferred method of estimating the parameters in frequentist inference. Bayesian approaches for estimating the number of species have been also used (e.g. Barger and Bunge (2008)), but are not explored in this thesis. However, when applied to estimating the number of species there might be several problems with the MLE, especially for small sample size. Cruyff and van der Heijden (2008) note that the coverage probabilities of confidence intervals calculated for the NB model using the MLE approach decrease in a small sample. The potential problems with maximum likelihood for estimating the number of species are as follows:

1. It can be difficult to calculate the MLE. The convergence of MLE may be

slow because the likelihood function is very flat in some situations; consequently, the variance of the estimator is very large. Li and Sudjianto (2005) and Coull and Agresti (1999) note that the cause of the flat likelihood is a large heterogeneity which leads to an unstable estimate. When applying to the ZTNB model with small Poisson parameter and small sample size, the MLE may fail to converge (van der Heijden et al., 2003). This problem occurs in many capture-recapture studies. To illustrate, Rocchetti et al. (2011) mention that the MLE method cannot fit the negative binomial model in some cases and gives examples involving scrapie, methamphetamine use and microbial data. The variation in population results in the numerical algorithms failing to converge. Hence, the estimator cannot be computed. Recently, Böhning (2015) has investigated the difficulties with the negative binomial in more detail.

2. Extreme heterogeneity in species discovery rates can lead to a problem known as the boundary problem (Pledger and Phillpot, 2008). It can be found in the mixed exponential family of discrete distributions such as the mixed binomial and the mixed Poisson model. This results in very large estimates of the total number of species $\widehat{N}$. Wang and Lindsay (2008) handled the boundary problem in estimating species richness using a penalized conditional nonparametric maximum likelihood estimator (NPMLE) and Wang later developed the `SPECIES` package in R to implement this method. This package includes a function to estimate the number of species under the conditional likelihood of the mixed Poisson model.

3. As a result of the previous problems, there is a possible lack of robustness. Even if the model can be fitted, it is difficult to know whether the model that has been assumed is really appropriate.

4. Another issue is the problem of lack of identifiability; there may be sev-

eral different models where the distribution of non-zero counts is basically the same, but the probability of a zero count is different. So all the models would appear to fit the observed data equally well, but they would all give different estimates of N. However, this last point is really a general issue about the difficulty of estimating species richness; it is not specific to the maximum likelihood approach. See Link (2003) for related discussion in connection with capture-recapture data.

3.5 Conclusion

The Poisson distribution is considered for a homogeneous population. It is usually well-fitting for count data that exhibit the equivalence of the mean and variance. For species abundance data, it cannot work well as a result of the variation in population. The mixed Poisson distribution is considered instead of the original Poisson to estimate the total number of species. Models whose variance exceeds the mean are proposed including the Poisson-gamma distribution or the negative binomial distribution.

The number of unseen species is unknown in species richness estimation. Modelling the number of species based on the zero truncated mixed Poisson model is proposed. The zero truncated negative binomial distribution is used for overdispersed data without zero frequency. Maximum likelihood estimation for truncated data is an approach used to estimate the unknown parameter in the model. Although in principle, maximum likelihood estimation can be used based on a mixed Possion model, the results indicate that this approach sometimes leads to poor inference about the number of species, especially for small sample sizes. There might be several problems such as flatness of the likelihood function and the boundary problem that lead to a lack of robustness. The penalized NPMLE by Wang and Lindsay (2008) are proposed to

avoid the boundary problem.

Other mixed Poisson models for overdispersion have been investigated in many studies including log-normal mixed Poisson, inverse Gaussian mixed Poisson and so on. In Chapter 4, the Poisson-Tweedie distribution, where the mixing distribution is the flexible 3-parameters Tweedie distribution is considered for estimating the number of species. Therefore, the weighted linear regression analysis is investigated instead in order to avoid MLE problems.

Chapter 4

Estimating the number of species using Poisson-Tweedie model

4.1 Introduction

Species abundance data are often described using overdispersed model such as the negative binomial distribution. In this Chapter, we investigated species abundance data using the Poisson-Tweedie (PT) distribution. It is a mixed Poisson model where the mixing distribution is the Tweedie distribution. This model is useful for modelling species abundance data. It often exhibits overdispersion, zero inflation and a heavy right tail. Not only is the negative binomial distribution in the PT family, it includes many well-known discrete distributions such as the Poisson, Poisson inverse Gaussian, discrete stable, Pólya-Aeppli, Neyman Type A and so on (El-Shaarawi et al., 2011). In Section 4.2, the PT model including sub families, the probability mass function, mean, variance, dispersion, skewness, and reparametrization are reviewed.

For inference, in addition to maximum likelihood estimation (MLE), We have

focused on the weighted linear regression (WLR) approach to model the number of unseen species for estimating the total number of species. This approach has been proposed recently for avoiding the numerical problems which occur in some circumstances when using MLE under the negative binomial distribution (Rocchetti et al., 2011). Although, the MLE is well known in statistics for estimating the unknown parameter, there might be problems about flatness of the likelihood and the boundary problem especially for small sample size. Therefore, the WLR estimator is used instead and is more robust than the MLE. In Section 4.3, the WLR model, which is based on ratios of successive counts, is discussed. The performance of the WLR estimator based on the PT distribution is investigated in a simulation study. The WLR approach is compared with other estimators including Chao1, iChao1, new_1 and new_2 estimators and applied to real data sets. The results are shown in Section 4.4. The conclusions are presented in Section 4.5.

4.2 Poisson-Tweedie (PT) model for overdispersed data

Environmental changes including physical, chemical and biological result in the heterogeneity of abundance data (El-Shaarawi et al., 2011). The models for overdispersed data are discussed instead of Poisson model which has the variance equal to the mean. Dispersion index can be defined as the variance divided by the mean and denoted by ϕ. When $\phi > 1$, it represents ovedispersed model, the variance exceeds the mean. Then, the Poisson distribution with $\phi = 1$ cannot fit well for overdispersed data. El-Shaarawi et al. (2011) proposed the PT distribution with three parameters to model species abundance data. It is a flexible model which can describe overdispered data. The PT distribution is a mixed Poisson model using the mixing model from the Tweedie distribution

which is a family of exponential dispersion models.

4.2.1 Tweedie distribution

A distribution belongs to the family of exponential dispersion models(Jorgensen, 1987) if it has probability density function of the form

$$f(y;\theta,\phi) = a(y,\phi)\exp\left[\frac{1}{\phi}\{y\theta - K(\theta)\}\right], \tag{4.1}$$

where θ is a canonical parameter, ϕ is a dispersion parameter ($\phi > 0$) and $K()$ is a cumulant function. If Y follows the exponential dispersion model, the relationship between the mean and the variance of the exponential dispersion model is given by

$$\text{var}(Y) = \phi \text{var}(\mu),$$

where μ is the mean of the distribution, $\mu = K'(\theta)$, and $\text{var}(\mu) = K''(\theta)$ is called the variance function of the exponential dispersion model, where $K'(\theta)$ and $K''(\theta)$ denote the first and the second derivative of the cumulant function (Dunn and Smyth, 2005).

The Tweedie distribution is a member of the family of exponential dispersion models. The second derivative of the cumulant function can be calculated by $K''(\theta) = d\mu/d\theta = \mu^p$ and then $\text{var}(\mu) = \mu^p$. Jorgensen (1987) named it the Tweedie distribution. If Y follows the Tweedie distribution, the parameters (μ, ϕ, p) of the Tweedie distribution satisfy $\mu > 0$, $\phi > 0$ and the power parameter varies outside the interval (0,1). The probability density function of the Tweedie distribution is given by equation (4.1) based on

$$\theta = \begin{cases} \dfrac{\mu^{(1-p)}}{1-p}, & \text{for } p \neq 1 \\ \log\mu, & \text{for } p = 1 \end{cases}$$

and

$$K(\theta) = \begin{cases} \dfrac{\mu^{(2-p)}}{2-p}, & \text{for } p \neq 2 \\ \log \mu, & \text{for } p = 2. \end{cases}$$

The Tweedie distribution for values of $p < 0$ seems to be used less in practice and Dunn and Smyth (2005) investigated the model for values of $p > 1$ in their study. The value of p can be used to define the special case of the Tweedie distribution which includes normal distribution ($p = 0$), Poisson distribution ($p = 1$), compound Poisson-gamma distribution ($1 < p < 2$), gamma distribution ($p = 2$), positive stable distribution ($2 < p < 3$ and $p > 3$) and inverse Gaussian distribution ($p = 3$) (Jorgensen, 1987).

In general it is difficult to evaluate the function $a(y, \phi)$. Considering the value of $p > 1$, Dunn and Smyth (2005) have focused on the series expansions for $1 < p < 2$ and $p > 2$. For the case of $1 < p < 2$, the probability density function of the Tweedie distribution can be written by

$$f(y; \mu, \phi, p) = \begin{cases} \exp\left\{-\dfrac{\mu^{2-p}}{\phi(2-p)}\right\}, & \text{for } y = 0 \\ a(y; \phi) \exp\left[\dfrac{1}{\phi}\left\{y\dfrac{\mu^{1-p}}{1-p} - \dfrac{\mu^{2-p}}{2-p}\right\}\right], & \text{for } y > 0 \end{cases}$$

where $a(y; \phi) = \dfrac{1}{y}\sum_{j=1}^{\infty} \dfrac{y^{-j\alpha}(p-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-p)^j j!\Gamma(-j\alpha)}$ and $\alpha = \dfrac{2-p}{1-p}$.

For the case of $p > 2$, the probability can be defined similarly using

$$a(y; \phi) = \frac{1}{\pi y}\sum_{k=1}^{\infty} \frac{\Gamma(1+\alpha k)\phi^{k(\alpha-1)}(p-1)^{\alpha k}}{\Gamma(1+k)(p-2)^k y^{\alpha k}}(-1)^k \sin(-k\pi\alpha)$$

and $0 < \alpha < 1$.

4.2.2 Poisson-Tweedie distribution

The PT distribution is the mixture model between Poisson and Tweedie distributions which is flexible and suitable for count data that exhibit over-dispersion, zero-inflation and heavy right tail, in particular, species abundance data that are sampled from heterogeneous population. (El-Shaarawi et al., 2011).

The Poisson-Tweedie distribution has been investigated in many studies. It was discovered and called by different names: the generalised negative binomial distribution by Gerber (1992), the Poisson Gaussian by Hougaard et al. (1997), the Poisson-Tweedie by Kokonendji et al. (2004) and the Tweedie-Poisson family by Johnson et al. (2005). The PT family includes many standard discrete distributions as special cases, such as the Poisson, negative binomial, Poisson-inverse Gaussian (PIG), Neyman Type A, Pólya-Aeppli and Poisson Pascal.

Let X be a random variable generated by the PT distribution with parameters a, b, c. The probability mass function (pmf) of the PT distribution is impossible to define in an explicit form. However, it can be defined in terms of the probability generating function (pgf), which can be written as

$$G_X(t; a, b, c) = \exp\left\{\frac{b}{a}[(1-c)^a - (1-ct)^a]\right\}, \tag{4.2}$$

where $a \leqslant 1$, $b > 0$ and $0 \leqslant c < 1$ (El-Shaarawi et al., 2011), as explained in Section 4.2.5.

4.2.3 Sub-families of the PT distribution

The pgf of the PT distribution gives rise to several special cases which are defined by different ranges of the parameters a, b and c as follows (El-Shaarawi et al., 2011):

- When $c = 1$ and $0 < a \leqslant 1$, the pgf of the PT distribution reduces to $G_X(t; a, b, 1) = \exp\left\{\frac{-b}{a}(1-t)^a\right\}$ and is called the discrete stable distribution.

- When $c = 0$, it becomes the degenerate distribution with the pgf which is given by $G_X(t; a, b, 0) = 1$.

- When $a = 1$, and $c \neq 0$, it represents the Poisson distribution. The pgf can be written as $G_X(t; 1, b, 1) = \exp\{bc(t-1)\}$.

- When $a = 1/2$, $b = \frac{\lambda}{2m}\sqrt{1 + \frac{2m^2}{\lambda}}$ and $c = \frac{2m^2}{\lambda} \Big/ \left(1 + \frac{2m^2}{\lambda}\right)$, it gives the Poisson inverse Gaussian distribution. Then,

$$G_X(t; 1/2, b, c) = \exp\left\{\frac{\lambda}{m}\left[1 - \left(1 + \frac{2m^2}{\lambda}(1-t)\right)^{1/2}\right]\right\}, \qquad \lambda > 0, m > 0.$$

- When $0 < a < 1$ and $0 < c < 1$, it is the generalized Poisson inverse Gaussian distribution.

- When $a = 0$, and $0 < c < 1$, it becomes the negative binomial distribution with pgf $G_X(t; 0, b, c) = \left(\frac{1-c}{1-ct}\right)^b$. These parameters are related to the form of the negative binomial distribution given in the previous Chapter by equation (3.1) with the relationship $c = \frac{1}{1+\beta}$ and $b = \alpha$.

- When $a < 0$ and $0 < c < 1$, it is the pgf of the Poisson-Pascal distribution. For $a = -1$, it is the Pólya-Aeppli distribution.

- When $a \to -\infty$, $b \to \infty$ and $c \to 0$, it represents the Neyman Type A distribution.

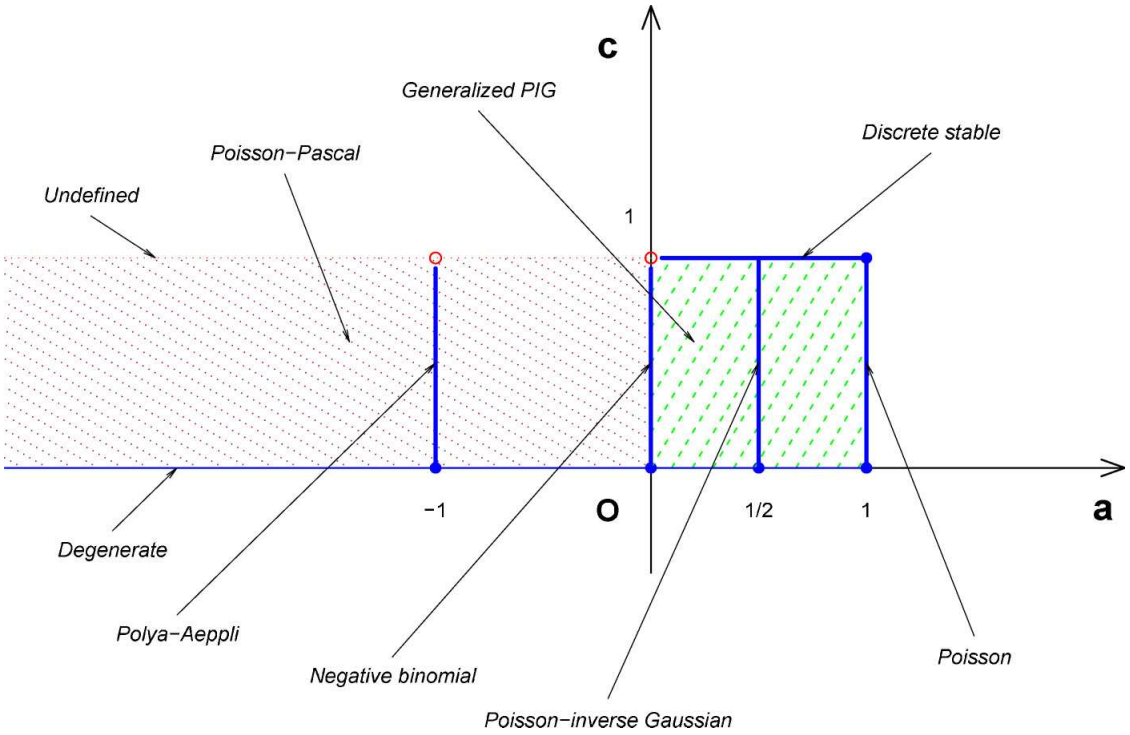


Figure 4.1: Partition of sub-families of the PT distribution based on parameters a and c (El-Shaarawi et al., 2011)

.

Figure 4.1, from El-Shaarawi et al. (2011), illustrates a partition of sub-families of the PT distribution based on the values of the parameters a and c above. This gives the special cases including distributions that can model overdispersed and heavy-tailed data. The Poisson inverse Gaussian, the generalized Poisson inverse Gaussian and the discrete stable distribution can model extremely heavy-tailed count data. Overdispersed data with shorter right tail can be fitted using the negative binomial and Poisson-Pascal distribution.

4.2.4 Mean, Variance, Dispersion and Skewness

We can calculate the derivatives of the pgf when $t = 1$ to obtain the mean, variance, dispersion and skewness. We have

$$\begin{aligned} G'_X(1; a, b, c) &= \mathrm{E}[X] \\ G''_X(1; a, b, c) &= \mathrm{E}[X(X-1)] \\ G'''_X(1; a, b, c) &= \mathrm{E}[X(X-1)(X-2)] \end{aligned}$$

These derivatives are used to derive the mean, variance, dispersion (ratio of variance to mean) and skewness for the PT model, which are given by (El-Shaarawi et al., 2011).

$$\mu = \frac{bc}{(1-c)^{1-a}},$$

$$\sigma^2 = \frac{bc(1-ac)}{(1-c)^{2-a}},$$

$$D = \frac{\sigma^2}{\mu} = \frac{1-ac}{1-c},$$

$$\psi = \frac{E[(X-E(X))^3]}{\sigma^3} = \frac{a^2c^2-3ac+c+1}{\sqrt{bc(1-c)^a(1-ac)^3}}.$$

4.2.5 The probability mass function

El-Shaarawi et al. (2011) derived the probability mass function of the PT distribution using a recursive algorithm. The first derivative of the pgf of the PT distribution

$$G_X^{'}(t;a,b,c) = bc(1-ct)^{a-1}G_X(t;a,b,c),$$

is considered under Taylor expansion. This gives the formula of the probability mass function which can be generated recursively as

$$p_0 = \begin{cases} \exp\left\{\frac{b}{a}[(1-c)^a-1]\right\} & \text{for } a \neq 0 \\ (1-c)^b & \text{for } a = 0 \end{cases},$$

$$p_1 = bcp_0,$$

$$p_{k+1} = \frac{1}{k+1}\left(bcp_k + \sum_{i=1}^{k} ir_{k+1-i}p_i\right) \text{ for } k = 1,2,\ldots,$$

where $r_1 = (1-a)c$ and $r_{j+1} = \left(\frac{j-1+a}{j+1}\right)cr_j$ (for $j = 1,2,\ldots$), are the

recursive relationships among the $r_j s$.

The likelihood function is constructed using these probabilities in order to estimate the unknown parameters of the PT distribution a, b, c . For example, when a is fixed as zero, the count data follows the negative binomial distribution. The MLE method is used to estimate the parameter b and c. Other sub-families are identified by different constraints, but there are similar procedures for estimating the parameters. Therefore, the PT distribution is used as a tool to investigate the robustness issue.

4.2.6 The Reparametrization (μ, D, a)

Additionally, we can reparametrize the parameters (a, b, c) to new parameters (μ, D, a), where μ is the mean, D is the dispersion index and a is the shape parameter that can determine the distribution of count data. These calculations are implemented in the R package `tweeDEseq`. The parameter (a, b, c) is reparametrized to (μ, D, a) using the relationship

$$b = \frac{\mu(1-a)^{1-a}}{(D-1)(D-a)^{-a}}, \qquad c = \frac{D-1}{D-a}$$

In `tweeDEseq` package, count data can be generated based on the PT distribution using the function `rPT()`. The probability mass function can be calculated using the function `dPT()`. The parameters of the PT distribution, μ, D, a, can estimated by optimizing the likelihood function.

When the parameters μ and D are fixed in the range, $\mu > 0$ and $D > 1$, the parameter a is called the family index and identifies different PT families. When $a = -1$, it becomes Pólya-Aeppli distribution. When $a < 0$, it becomes Poisson-Pascal distribution. When $a = 0$, it becomes the negative binomial distribution. When $a = 0.5$, it becomes the Poisson inverse Gaussian distri-

bution. When $0 < a < 1$, it becomes the generalized Poisson inverse Gaussian distribution. When $a \to -\infty$, it becomes the Neyman Type A (El-Shaarawi et al., 2011).

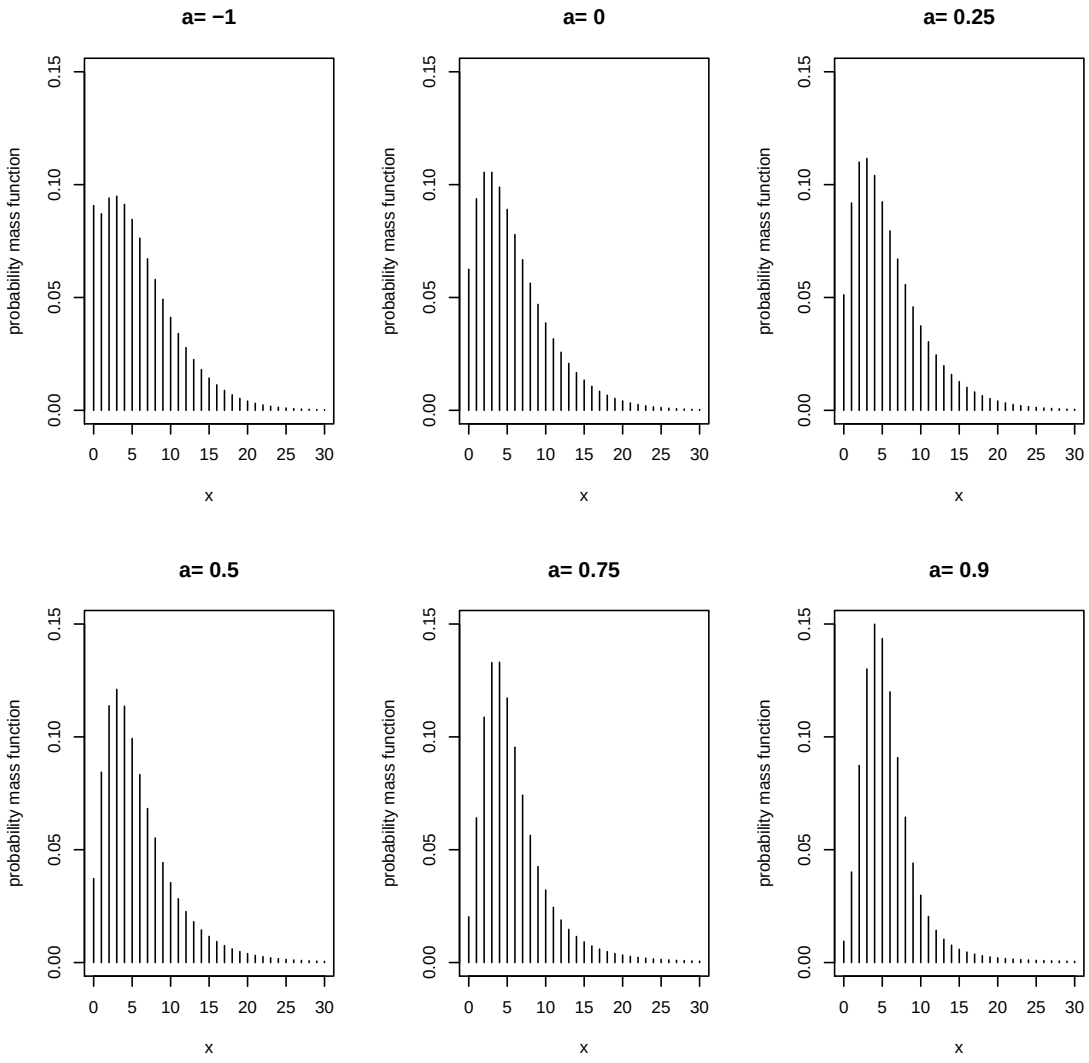


Figure 4.2: Comparison of the probability mass function for the PT distribution when $\mu = 6$, $D = 4$ and $a = -1, 0, 0.25, 0.5, 0.75, 0.9$.

Figure 4.2 shows the probability mass function of the PT distribution based on varying parameter a with $\mu = 6$ and $D = 4$ fixed, which changes the shape of the PT distribution. The probability for unseen species decreases when the value of a decreases. For $a = -1$, the probability for unseen species is the highest. When a increases to 0.9, the probability for unseen species has the

lowest value when compared to other models.

When $a = -1$, $p_0 > p_1$ and $p_1 < p_2$, then we found the PT distribution can be bimodal. (El-Shaarawi et al., 2011) mention that it is probably unimodal and bimodal for the PT distribution when $0 \leq a < 1$ and $a < 0$ respectively. Additionally, the probability mass function of the PT distribution has longer tail when a is close to 1. For example, when looking at the negative binomial distribution ($a = 0$) and the Poisson Inverse Gaussian distribution ($a = 0.5$), the pmf for the negative binomial decreases more slowly for the Poisson Inverse Gaussian distribution. Then, the Poisson Inverse Gaussian distribution is more suitable for a right long tail than the negative binomial distribution.

4.3 Models based on ratios of successive counts

The ratios of probabilities of successive counts are considered for estimating the number of species in the study of Rocchetti et al. (2011). The ratio of successive probabilities is considered for the Katz family of distributions, of which the Poisson, the binomial and the negative binomial distribution are special cases. Let p_x denote the probability distribution of X over the nonnegative integers. In the Katz family, this ratio is given by

$$\frac{p_{x+1}}{p_x} = \frac{\gamma + \delta x}{x+1}, \qquad x = 0, 1, 2, \ldots,$$

and it follows that

$$r_x = (x+1)\frac{p_{x+1}}{p_x} = \gamma + \delta x, \tag{4.3}$$

where $\gamma > 0$ and $\delta < 1$. Under the condition $\delta < 0$, the distribution of X becomes the binomial distribution. When $\delta = 0$, the distribution of X represents the Poisson distribution. When $0 < \delta < 1$, the negative binomial distribution arises. The ratio r_x is a monotone increasing pattern for the mixed-Poisson

distribution (Rocchetti et al., 2011). Then, this ratio is a linear function of x.

The idea of ratio plot is considered for estimating the number of species. When p_x is unknown, it can be estimated as $\widehat{p}_x = f_x/N$. Considering the ratio of probability of successive counts, N cancels out and the ratio in equation (4.3) can be estimated without N as follows:

$$\widehat{r}_x = (x+1)\frac{\widehat{p}_{x+1}}{\widehat{p}_x} = (x+1)\frac{f_{x+1}}{f_x}, \tag{4.4}$$

where f_x is the number of species seen x times. When $x = 0$, the number of unseen species can be estimated by $\widehat{f}_0 = \dfrac{f_1}{\widehat{\gamma}}$.

Although equation (4.4) gives a linear regression of $\widehat{r}_x$ on x, the logarithmic transformation of the response r_x is considered to avoid the possibility of negative predicted values from the model (Rocchetti et al., 2011). So equation (4.4) is replaced by

$$\log\ \widehat{r}_x = \gamma + \delta x. \tag{4.5}$$

This results in $\widehat{r}_x = \exp\left\{\widehat{\gamma} + \widehat{\delta}x\right\}$ and we have $\dfrac{(x+1)f_{x+1}}{f_x} = \exp\left\{\widehat{\gamma} + \widehat{\delta}x\right\}$. Thus, under the log-scale of the ratio, the number of unseen species is given by

$$\widehat{f}_0 = \frac{f_1}{\exp\{\widehat{\gamma}\}}. \tag{4.6}$$

Therefore, we have

$$\widehat{N} = K + \widehat{f}_0. \tag{4.7}$$

where K be the number of seen species in the sample. The delta method is used to approximate the variance of $\widehat{f}_0$ based on the conditional variance (Böhning 2008) which gives

$$\mathrm{Var}(\widehat{f}_0) \approx \exp\{-\widehat{\gamma}\}^2 f_1 \left[\mathrm{Var}(\widehat{\gamma}) f_1 + 1\right], \tag{4.8}$$

Let K denote a binomial random variable with parameters N and $(1-p_0)$. The variance of K is $\text{Var}(K) = N(1-p_0)p_0$ and the estimated variance of K based on the delta method is given by

$$\text{Var}(K) \approx \frac{K\widehat{f}_0}{\widehat{N}}. \tag{4.9}$$

Therefore, the variance of $\widehat{N}$ can be estimated, by

$$\text{Var}(\widehat{N}) \approx \text{Var}(K) + \text{Var}(\widehat{f}_0) = \frac{K\widehat{f}_0}{\widehat{N}} + \exp\left\{-\widehat{\gamma}\right\}^2 f_1[\text{Var}(\widehat{\gamma})f_1 + 1] \tag{4.10}$$

In Figures 4.3 and 4.4, count data are generated under the PT distribution with parameters $\mu = 1$, $D = 2$ and $a = -1, 0, 0.25, 0.5, 0.75, 0.9$. The results show the relationship of the ratio of successive counts after using log-transformation and show that r_x is non linear. The Poisson-Tweedie distribution is not a member of the Katz family in general, as is seen in Figures 4.3 and 4.4, where the points do not lie on straight lines. Therefore some bias is expected. There is no simple expression for r_x for the PT distribution.

For example with $a = 0$, it presents the negative binomial distribution which is given by

$$p_x = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\,\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{x}.$$

The ratio of successive probability can be written as

$$\frac{p_{x+1}}{p_x} = \frac{\dfrac{\Gamma(x+\alpha+1)}{\Gamma(x+2)\,\Gamma(\alpha)} \left(\dfrac{\beta}{1+\beta}\right)^{\alpha} \left(\dfrac{1}{1+\beta}\right)^{x+1}}{\dfrac{\Gamma(x+\alpha)}{\Gamma(x+1)\,\Gamma(\alpha)} \left(\dfrac{\beta}{1+\beta}\right)^{\alpha} \left(\dfrac{1}{1+\beta}\right)^{x}}$$

$$= \left(\frac{x+\alpha}{x+1}\right)\left(\frac{1}{1+\beta}\right). \tag{4.11}$$

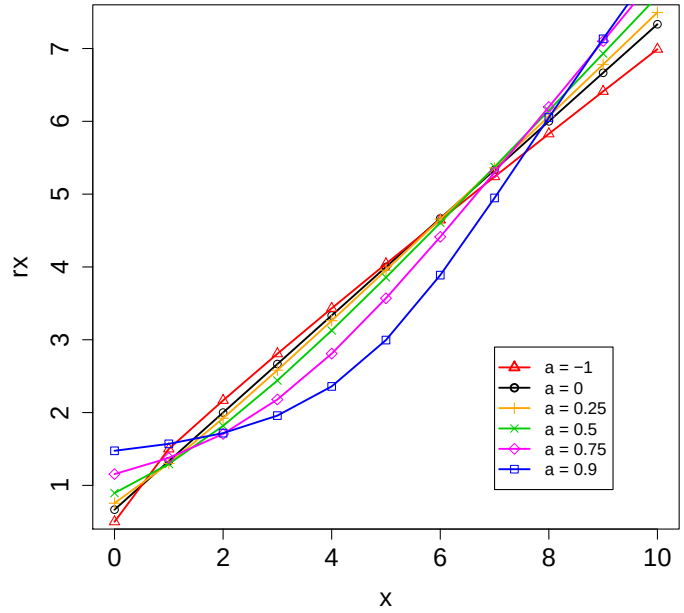


Figure 4.3: The ratio of successive frequencies based on the true probability of PT distribution with the parameters $\mu = 1$, $D = 2$ and $a = -1$, 0, 0.25, 0.5, 0.75, 0.9.

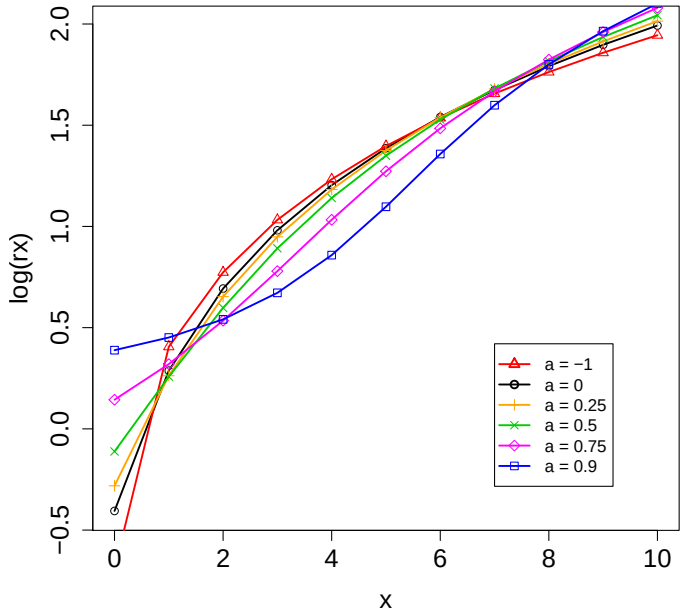


Figure 4.4: The logarithmic transformation of the ratio of successive frequencies based on the true probability of PT distribution with the parameters $\mu = 1$, $D = 2$ and $a = -1$, 0, 0.25, 0.5, 0.75, 0.9.

From the equation (4.5), the log-transformation gives

$$\begin{aligned}\log\ \widehat{r}_x &= \log\left\{\frac{(x+1)p_{x+1}}{p_x}\right\} \\ &= \log(x+1) + \log\left(\frac{p_{x+1}}{p_x}\right) \\ &= \log(x+\alpha) + \log\left(\frac{1}{1+\beta}\right). \end{aligned} \tag{4.12}$$

The first order Taylor expansion of $\log(x+\alpha)$ around α is

$$\log(x+\alpha) \approx \log(\alpha) + \frac{x}{\alpha}.$$

Therefore,

$$\log\ \widehat{r}_x \approx \log\left(\frac{\alpha}{1+\beta}\right) + \frac{x}{\alpha} \tag{4.13}$$

which is the linear regression model with $\gamma = \log\left(\frac{\alpha}{1+\beta}\right)$ and $\delta = \frac{1}{\alpha}$ (Rocchetti et al., 2011). When considering the situation for unseen species, this approximation in equation (4.13) has the model similar to equation (4.12).

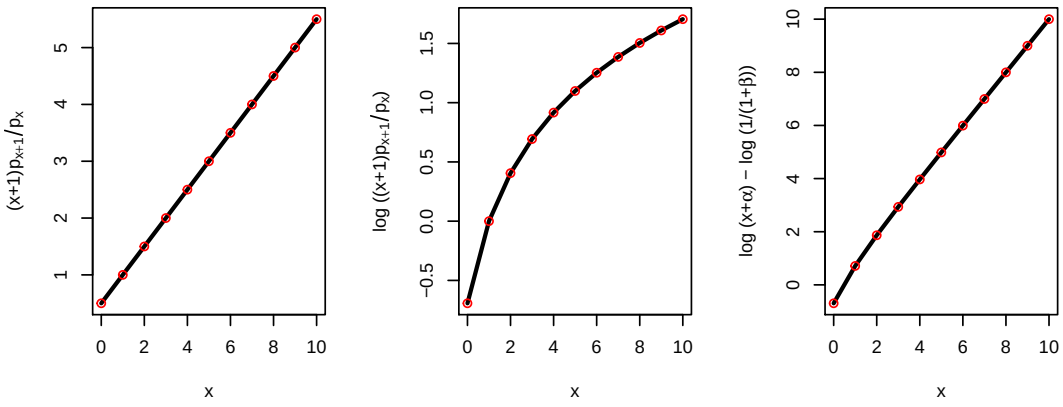


Figure 4.5: The ratios r_x, $\log\left((x+1)\frac{p_{x+1}}{p_x}\right)$ and $\log\left(\frac{\alpha}{1+\beta}\right) + \frac{x}{\alpha}$ under the PT distribution; $\mu = 1$, $D = 2$, $a = 0$

Figure 4.5 presents the relation between log-transformation of the ratio of

successive and x. Count data is generated from the PT distribution with $a = 0$ which displays the negative binomial distribution. When using the first order Taylor expansion for the the negative binomial distribution, it is clear that the linear approximation works well for the negative binomial model as shown in Figure 4.5 on the right hand side. Hence, the WLR approach should work well for this model.

4.4 Weighted Linear Regression Analysis

Multiple linear regression is a technique used to explain the relationship between the continuous response variable and two or more independent variables. The least square method is used to estimate the regression parameter. However, it is not appropriate for fitting the model in equation (4.5) as a result of condition of assumptions especially the problem about dependence and heteroscedasticity (Rocchetti et al., 2011).

Since the elements of the response are correlated and have unequal variance, the weighted linear regression is more appropriate than ordinary unweighted regression. For the analysis, the two parameters γ and δ are estimated using the weighted least squares method that is given by

$$\begin{pmatrix} \widehat{\gamma} \\ \widehat{\delta} \end{pmatrix} = (X^T W X)^{-1} X^T W Y. \tag{4.14}$$

where

$$Y = \begin{pmatrix} \log(r_1) \\ \log(r_2) \\ \vdots \\ \log(r_{m-1}) \end{pmatrix}, X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & m-1 \end{pmatrix}.$$

where $\log(r_i) = \log\left(\frac{(i+1)f_{i+1}}{f_i}\right)$ and m refers to the number of frequencies

used for fitting model, e.g. for truncated data, m is the maximum frequency used to estimate population size. The count data which arise from the PT distribution might have a heavy right tail and this sometimes leads to problems with sparse data. We might find zero frequencies that lead to the WLR not working. The truncation of data is proposed for this issue. Rocchetti et al. (2011) chose the truncation point to be the smallest m such that $f_m > 0$ and $f_{m+1} = 0$ for the WLR approach.

Rocchetti et al. (2011) considered setting the weight matrix W to be an approximation to the inverse of the covariance matrix of Y, $W \approx (\text{cov}(Y))^{-1}$, obtained using the delta method and ignoring the terms of the off-diagonal of cov(Y). The precision is slightly lost when dropping the off-diagonal term. the weight matrix W can be approximated by (Rocchetti et al., 2011)

$$W = \begin{bmatrix} \frac{1}{f_1} + \frac{1}{f_2} & 0 & 0 & 0 \\ 0 & \frac{1}{f_2} + \frac{1}{f_3} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{f_m} + \frac{1}{f_{m+1}} \end{bmatrix}^{-1}$$

4.5 Simulation study and real data examples

4.5.1 Simulation study

We have investigated the performance of various estimators applied to simulated data from the PT distribution. Simulations were carried out as follows.

- Data are simulated using the PT distribution with parameters $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$, $N = 100, 200, 1000$ and repeated 10000 times.

- The WLR estimator is used to estimate species richness and compared with the Chao1, the iChao1, the new$_1$ and the new$_2$ estimators. For

the WLR estimator, truncated data is used in the analysis. Data are truncated at the first m frequencies where m is the smallest value such that $f_m > 0$ and $f_{m+1} = 0$. For the Chao1, the iChao1, the new_1 and the new_2 estimators, when the number of doubletons and/or bias of the estimator is zero, the Chao1 estimator is used instead.

- The performance of the different estimators is measured in terms of the root mean square error (RMSE) and the bias.

Considering the results in Tables 4.1, 4.2 and 4.3, the performance of the WLR estimator based on the PT distribution with fixed $\mu = 1$ depends on the choice of the parameters D and a. For small N, the performance with small a has good approximation. When $\mu > 1$, $D > 1$ and $a = -1$, the PT distribution becomes the Pólya-Aeppli distribution. The results show the WLR estimator works well with the Pólya-Aeppli distribution for small N. Table 4.1 shows the result for $N = 100$. The model with $a = 0$ gives the best estimates when $D = 1.1$ in terms of RMSE and the model with $a = -1$ gives the smallest RMSE when $D = 1.25, 1.5$ and 2. Table 4.2 presents the performance for $N = 200$. The results show the model with $a = -1$ leads to the best estimates when $D = 1.1, 1.25$ and 1.5.

Considering the results in Tables 4.1, 4.2, 4.3 and 4.4, the performance of the WLR estimator based on the PT distribution depends on the value of the parameters D and a. The value of parameter μ affects the performance of the WLR estimator significantly. The results show the WLR estimator based on the PT distribution with $\mu = 2$ performs better than $\mu = 1$ significantly. Additionally, the performance of the WLR estimator is improved when using small a and small D.

In table 4.1, when $a = -1$, the PT distribution becomes the Pólya-Aeppli

Table 4.1: Performance of $\widehat{N}_{WLR}$ based on the PT distribution with $N = 100$, $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Bias	RMSE	$se(\widehat{N})$	$\widehat{se}(\widehat{N})$
$N = 100$	$D = 1.1$	-1	**10.88**	**36.01**	34.33	42.40
		0	11.68	37.36	35.48	46.51
		0.25	11.43	37.36	35.57	46.60
		0.50	12.29	37.59	35.53	46.82
		0.75	13.23	39.10	36.80	50.79
		0.90	13.64	38.71	36.23	45.87
	$D = 1.5$	-1	**7.83**	**34.78**	33.89	45.61
		0	9.64	37.49	36.23	48.46
		0.25	10.06	37.42	36.04	46.55
		0.50	11.84	39.10	37.26	47.63
		0.75	13.62	41.85	39.57	60.72
		0.90	16.29	43.52	40.35	51.04
	$D = 1.5$	-1	**-0.89**	**31.26**	31.24	40.53
		0	4.03	35.96	35.74	44.94
		0.25	5.64	35.32	34.87	45.51
		0.50	9.51	38.39	37.19	48.64
		0.75	15.69	44.18	41.30	57.18
		0.90	17.66	44.77	41.14	50.48
	$D = 2$	-1	-18.22	**29.94**	23.76	30.24
		0	-7.99	31.27	30.23	39.51
		0.25	**-2.94**	32.79	32.66	43.56
		0.50	4.71	38.79	38.51	52.40
		0.75	15.95	49.42	46.77	57.62
		0.90	20.76	51.49	47.12	61.15

distribution. The results show the WLR estimator works well with the Pólya-Aeppli distribution with $N = 100$ and $\mu = 1$. In particular, the performance of the estimator improves when D increases. When looking at $N = 100$ and $\mu = 2$, the performance of the WLR estimator with the same D provides similar results for various a as shown in Table 4.2. When comparing between $\mu = 1$ and $\mu = 2$, RMSE of the WLR estimator for $\mu = 2$ reduces around three times.

Table 4.3, when $N = 1000$, the WLR estimator gives the positive bias when $D = 1.1$ and it gives the negative bias when $D > 1.1$ in most situations. When the parameters change, the best model is different. When $D = 1.1$, the model

Table 4.2: Performance of $\widehat{N}_{WLR}$ based on the PT distribution with $N = 100$, $\mu = 2$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Bias	RMSE	$se(\widehat{N})$	$\widehat{se}(\widehat{N})$
$N = 100$	$D = 1.1$	-1	**2.08**	9.22	8.98	9.43
		0	2.19	9.19	8.93	9.43
		0.25	2.22	**9.03**	8.76	9.45
		0.50	2.16	9.27	9.02	9.44
		0.75	2.28	9.17	8.89	9.39
		0.90	2.47	9.24	8.90	9.46
	$D = 1.5$	-1	**1.50**	9.71	9.60	10.17
		0	1.77	9.68	9.51	10.02
		0.25	1.89	**9.56**	9.37	9.93
		0.50	2.14	9.91	9.67	10.07
		0.75	2.75	9.95	9.57	10.16
		0.90	3.01	9.99	9.52	10.09
	$D = 1.5$	-1	**-0.56**	**10.10**	10.09	10.66
		0	0.58	10.24	10.22	10.85
		0.25	1.24	10.53	10.45	11.05
		0.50	2.09	10.72	10.52	11.25
		0.75	3.13	10.99	10.53	11.13
		0.90	3.63	10.83	10.20	10.86
	$D = 2$	-1	-6.75	12.44	10.45	10.77
		0	-2.87	**11.75**	11.40	11.89
		0.25	**-1.18**	11.85	11.79	12.27
		0.50	1.25	12.12	12.05	12.99
		0.75	3.92	12.59	11.97	12.89
		0.90	4.25	11.98	11.21	11.72

with small a gives a better performance. When $D = 2$, the WLR estimator can estimate well when a increases. Considering the standard error of $\widehat{N}_{WLR}$ and the estimated standard error of $\widehat{N}$ in equation (4.10), the results for large N give a good approximation.

Table 4.4 shows the performance of the WLR estimator for $N = 1000$ and $\mu = 2$. The bias and RMSE are improved when comparing to $\mu = 1$ around four times. The results show the best estimator for each situation is similar to $\mu = 1$.

Table 4.3: Performance of $\widehat{N}_{WLR}$ based on the PT distribution with $N = 1000$, $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	bias	RMSE	$se(\widehat{N})$	$\widehat{se}(\widehat{N})$
$N = 1000$	$D = 1.1$	-1	**8.94**	**75.82**	75.29	76.50
		0	12.44	76.82	75.80	77.96
		0.25	13.76	77.99	76.76	77.51
		0.50	16.06	78.44	76.77	78.01
		0.75	24.07	81.32	77.68	79.27
		0.90	36.37	87.46	79.54	81.67
	$D = 1.25$	-1	-25.49	74.34	69.83	72.03
		0	-11.59	**73.51**	72.59	73.56
		0.25	**-3.80**	74.17	74.07	75.30
		0.50	10.24	76.30	75.61	75.98
		0.75	35.75	86.69	78.97	78.91
		0.90	63.11	102.31	80.53	83.26
	$D = 1.5$	-1	-100.16	117.56	61.55	63.33
		0	-63.19	91.10	65.62	68.11
		0.25	-44.00	80.75	67.70	70.21
		0.50	**-12.39**	**72.15**	71.08	73.62
		0.75	42.42	88.59	77.77	78.59
		0.90	86.17	120.63	84.41	85.23
	$D = 2$	-1	-254.02	258.42	47.44	47.12
		0	-175.15	183.61	55.09	58.17
		0.25	-133.97	146.78	59.98	63.05
		0.50	-71.08	97.11	66.16	70.01
		0.75	**32.39**	**84.32**	77.85	78.45
		0.90	102.93	133.83	85.53	87.92

Tables 4.5 - 4.8 present the performance of the WLR estimator compared with other estimators for $N = 100$ and 1000. The results show the approximation of the WLR estimator is a poor fit based on the PT distribution with $\mu = 1$ especially for small N.

When $N = 100$ and $\mu = 1$, the Chao1 estimator outperforms other estimators with the smallest RMSE when $D = 1.1$ and $D = 1.25$ for all values of a. The new_2 estimator works well with the PT distribution with $D = 1.5$, $a < 0.75$ and $D = 2$, $a < 0.9$ (Table 4.5). When $N = 100$ and $\mu = 2$, the Chao estimator is the best performance in most situations. However, it is found that the

Table 4.4: Performance of $\widehat{N}_{WLR}$ based on the PT distribution with $N = 1000$, $\mu = 2$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	bias	RMSE	$se(\widehat{N})$	$\widehat{se}(\widehat{N})$
$N = 1000$	$D = 1.1$	-1	**1.56**	**24.04**	23.99	23.17
		0	2.36	24.17	24.05	23.21
		0.25	2.49	24.11	23.98	23.27
		0.50	3.29	24.39	24.17	23.32
		0.75	4.97	24.57	24.07	23.35
		0.90	6.57	24.93	24.05	23.50
	$D = 1.25$	-1	-4.93	25.94	25.47	24.52
		0	-1.99	25.72	25.64	24.65
		0.25	**-0.23**	**25.46**	25.46	24.79
		0.50	2.36	25.68	25.57	24.75
		0.75	7.95	27.05	25.86	24.94
		0.90	13.93	29.06	25.50	25.00
	$D = 1.5$	-1	-26.49	37.75	26.89	25.79
		0	-15.95	31.58	27.25	26.19
		0.25	-9.73	29.46	27.81	26.50
		0.50	**-1.63**	**27.47**	27.42	26.66
		0.75	12.08	30.47	27.98	26.91
		0.90	21.33	34.60	27.24	26.84
	$D = 2$	-1	-91.03	95.05	27.35	26.02
		0	-56.20	62.95	28.37	28.07
		0.25	-39.48	49.55	29.94	28.75
		0.50	-16.73	34.51	30.19	29.25
		0.75	**13.91**	**33.25**	30.20	29.16
		0.90	28.77	41.05	29.28	28.66

performance of the WLR estimator becomes the second best and outperforms the iChao1, the new$_1$ and the new$_2$ in terms of RMSE especially when $D < 2$ (Table 4.6).

Table 4.5: RMSE and bias of five estimators based on the PT distribution with $N = 100$ $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Chao1	iChao1	new1	new2	WLR
$N = 100$							
$D = 1.1$	RMSE	-1	17.61	21.21	22.54	25.29	36.01
		0	17.92	21.63	23.07	25.88	37.36
		0.25	17.90	21.57	22.95	25.67	37.36
		0.50	17.84	21.69	23.16	26.13	37.59
		0.75	18.43	22.54	24.04	26.94	39.10
		0.90	18.37	22.54	24.28	27.60	38.71
	bias	-1	-0.97	3.58	7.94	17.61	10.88
		0	-0.47	4.18	8.54	18.15	11.68
		0.25	-0.72	3.91	8.23	17.85	11.43
		0.50	-0.13	4.49	8.94	18.51	12.29
		0.75	0.44	5.20	9.61	19.16	13.23
		0.90	1.10	5.91	10.41	20.12	13.64
$D = 1.25$	RMSE	-1	18.40	20.53	20.78	20.37	34.78
		0	18.65	21.16	21.58	21.46	37.49
		0.25	18.27	20.90	21.34	21.44	37.42
		0.50	18.71	21.70	22.36	22.65	39.10
		0.75	18.90	22.40	23.24	24.07	41.85
		0.90	19.23	23.51	24.79	26.54	43.52
	bias	-1	-5.99	-1.29	2.24	10.13	7.83
		0	-4.99	-0.12	3.44	11.22	9.64
		0.25	-4.57	0.36	3.91	11.73	10.06
		0.50	-3.53	1.50	5.15	12.86	11.84
		0.75	-2.06	3.06	6.90	14.69	13.62
		0.90	0.32	5.57	9.68	17.94	16.29
$D = 1.5$	RMSE	-1	21.97	21.87	20.87	16.77	31.26
		0	20.84	21.39	20.69	17.48	35.96
		0.25	20.47	21.27	20.74	17.86	35.32
		0.50	20.26	22.04	21.74	19.51	38.39
		0.75	20.18	23.40	23.67	22.56	44.18
		0.90	19.44	23.80	24.66	25.50	44.77
	bias	-1	-14.25	-9.78	-7.18	-1.09	-0.89
		0	-11.58	-6.72	-4.01	1.74	4.03
		0.25	-10.49	-5.58	-2.73	3.07	5.64
		0.50	-8.00	-2.76	0.19	5.78	9.51
		0.75	-3.79	1.88	5.10	10.91	15.69
		0.90	-0.36	5.36	9.11	16.13	17.66
$D = 2$	RMSE	-1	31.67	29.73	28.44	23.61	29.94
		0	27.63	26.10	24.87	20.56	31.27
		0.25	25.79	24.82	23.62	19.44	32.79
		0.50	24.14	24.25	23.32	19.61	38.79
		0.75	21.75	24.04	23.87	21.31	49.42
		0.90	21.26	25.88	26.45	26.02	51.49
	bias	-1	-28.08	-24.49	-22.86	-18.61	-18.22
		0	-21.71	-17.28	-15.42	-11.77	-7.99
		0.25	-18.60	-13.80	-11.82	-8.33	-2.94
		0.50	-14.01	-8.70	-6.49	-3.20	4.71
		0.75	-6.67	-0.74	2.04	5.80	15.95
		0.90	-0.25	5.99	9.43	14.95	20.76

Table 4.6: RMSE and bias of five estimators based on the PT distribution with $N = 100$ $\mu = 2$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Chao1	iChao1	new1	new2	WLR
$N = 100$							
$D = 1.1$	RMSE	-1	7.85	9.68	10.32	17.41	9.22
		0	7.80	9.67	10.30	17.45	9.19
		0.25	7.79	9.66	10.31	17.47	9.03
		0.50	7.91	9.77	10.39	17.49	9.27
		0.75	7.76	9.58	10.24	17.62	9.17
		0.90	7.80	9.63	10.35	17.93	9.24
	bias	-1	0.35	2.27	3.31	15.58	2.08
		0	0.43	2.33	3.38	15.68	2.19
		0.25	0.49	2.41	3.47	15.72	2.22
		0.50	0.38	2.28	3.32	15.68	2.16
		0.75	0.51	2.37	3.44	15.86	2.28
		0.90	0.68	2.51	3.61	16.20	2.47
$D = 1.25$	RMSE	-1	8.55	10.35	10.75	15.63	9.71
		0	8.37	10.14	10.62	15.84	9.68
		0.25	8.16	9.94	10.41	15.84	9.56
		0.50	8.42	10.28	10.80	16.21	9.91
		0.75	8.27	10.27	10.82	16.83	9.95
		0.90	8.25	10.28	10.92	17.57	9.99
	bias	-1	-0.91	1.41	2.38	13.26	1.50
		0	-0.73	1.53	2.57	13.57	1.77
		0.25	-0.64	1.63	2.67	13.69	1.89
		0.50	-0.41	1.84	2.91	13.98	2.14
		0.75	0.12	2.38	3.47	14.75	2.75
		0.90	0.54	2.65	3.80	15.62	3.01
$D = 1.5$	RMSE	-1	9.63	10.77	10.90	12.58	10.10
		0	9.41	10.83	11.03	13.43	10.24
		0.25	9.39	10.99	11.25	14.03	10.53
		0.50	9.28	11.06	11.43	14.71	10.72
		0.75	9.04	11.05	11.50	15.95	10.99
		0.90	8.68	10.79	11.46	17.41	10.83
	bias	-1	-3.57	-0.86	0.04	9.10	-0.56
		0	-2.70	0.06	1.02	10.18	0.58
		0.25	-2.17	0.63	1.63	10.85	1.24
		0.50	-1.52	1.25	2.30	11.75	2.09
		0.75	-0.58	2.04	3.21	13.31	3.13
		0.90	0.44	2.73	4.04	15.22	3.63
$D = 2$	RMSE	-1	13.84	13.37	13.09	9.50	12.44
		0	12.29	12.64	12.51	10.78	11.75
		0.25	11.66	12.43	12.38	11.52	11.85
		0.50	11.11	12.51	12.64	13.05	12.12
		0.75	10.22	12.36	12.74	15.38	12.59
		0.90	9.16	11.34	12.03	17.21	11.98
	bias	-1	-9.82	-6.88	-6.11	0.51	-6.75
		0	-6.74	-3.49	-2.59	4.13	-2.87
		0.25	-5.43	-2.09	-1.14	5.74	-1.18
		0.50	-3.50	-0.11	0.98	8.20	1.25
		0.75	-1.07	2.11	3.42	11.84	3.92
		0.90	0.23	2.73	4.24	14.73	4.25

Table 4.7: RMSE and bias of five estimators based on the PT distribution with $N = 1000$, $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Chao1	iChao1	new1	new2	WLR
$N = 1000$							
$D = 1.1$	RMSE	-1	60.38	72.09	80.69	159.10	75.82
		0	59.61	72.82	82.68	161.21	76.82
		0.25	59.40	73.52	83.77	162.36	77.99
		0.50	58.78	73.66	84.88	163.99	78.44
		0.75	56.71	75.96	89.34	169.43	81.32
		0.90	54.43	79.17	97.43	179.69	87.46
	bias	-1	-34.15	7.60	53.25	150.36	8.94
		0	-32.17	10.44	55.68	152.54	12.44
		0.25	-31.22	11.76	56.79	153.67	13.76
		0.50	-29.87	12.81	58.41	155.31	16.06
		0.75	-24.97	19.25	64.31	160.95	24.07
		0.90	-16.65	27.63	74.19	171.41	36.37
$D = 1.25$	RMSE	-1	100.74	75.61	58.25	86.81	74.34
		0	93.78	72.73	59.59	95.75	73.51
		0.25	90.07	71.76	60.57	100.52	74.17
		0.50	83.24	71.23	64.13	109.95	76.30
		0.75	71.67	74.42	73.74	128.80	86.69
		0.90	58.42	80.95	89.86	156.61	102.31
	bias	-1	-88.76	-34.70	-8.61	71.64	-25.49
		0	-79.79	-22.88	2.13	81.38	-11.59
		0.25	-75.14	-16.99	7.77	86.70	-3.80
		0.50	-65.95	-5.39	18.71	96.94	10.24
		0.75	-49.25	12.85	38.55	117.25	35.75
		0.90	-27.07	32.08	64.46	147.33	63.11
$D = 1.5$	RMSE	-1	175.59	126.20	113.90	59.78	117.56
		0	152.37	101.77	90.39	49.11	91.10
		0.25	140.55	90.78	80.14	48.79	80.75
		0.50	121.94	76.90	67.20	56.82	72.15
		0.75	91.00	70.23	64.80	89.82	88.59
		0.90	64.84	84.77	88.68	139.91	120.63
	bias	-1	-169.73	-111.41	-100.38	-38.65	-100.16
		0	-144.93	-80.96	-70.78	-11.96	-63.19
		0.25	-131.95	-65.10	-55.45	2.46	-44.00
		0.50	-111.21	-40.63	-30.90	26.12	-12.39
		0.75	-73.85	0.89	12.98	72.14	42.42
		0.90	-34.01	35.15	58.88	128.17	86.17
$D = 2$	RMSE	-1	308.46	263.17	258.85	215.14	258.42
		0	251.28	196.09	192.75	154.98	183.61
		0.25	222.61	163.84	160.64	124.97	146.78
		0.50	180.45	118.99	115.33	82.57	97.11
		0.75	116.32	73.22	67.93	59.68	84.32
		0.90	69.97	85.04	85.85	122.05	133.83
	Bias	-1	-305.92	-258.55	-254.59	-211.47	-254.02
		0	-247.35	-188.44	-185.55	-148.59	-175.15
		0.25	-217.42	-153.20	-150.52	-115.63	-133.97
		0.50	-173.19	-101.62	-98.84	-65.63	-71.08
		0.75	-101.87	-20.77	-16.00	21.23	32.39
		0.90	-41.22	35.52	53.30	107.90	102.93

Table 4.8: RMSE and bias of five estimators based on the PT distribution with $N = 1000$, $\mu = 2$, $D = 1.1, 1.25, 1.5, 2$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$ and 10000 simulations.

		a	Chao1	iChao1	new1	new2	WLR
$N = 1000$							
$D = 1.1$	RMSE	-1	23.90	31.30	35.99	148.10	24.04
		0	23.89	31.58	36.68	149.03	24.17
		0.25	23.85	31.53	36.73	149.11	24.11
		0.50	23.72	31.64	37.16	149.99	24.39
		0.75	23.32	31.73	37.87	151.79	24.57
		0.90	23.24	31.61	38.48	154.08	24.93
	bias	-1	-6.78	5.87	20.71	146.32	1.56
		0	-6.06	6.57	21.52	147.25	2.36
		0.25	-5.96	6.61	21.64	147.34	2.49
		0.50	-5.31	7.27	22.34	148.21	3.29
		0.75	-4.02	8.39	23.71	150.06	4.97
		0.90	-2.93	8.47	24.50	152.37	6.57
$D = 1.25$	RMSE	-1	31.41	33.14	33.02	124.89	25.94
		0	30.12	33.40	34.14	127.88	25.72
		0.25	29.24	33.69	34.79	129.73	25.46
		0.50	28.32	33.67	35.64	132.51	25.68
		0.75	26.43	34.61	38.11	139.04	27.05
		0.90	24.52	34.47	40.41	148.53	29.06
	bias	-1	-19.82	1.28	11.75	122.54	-4.93
		0	-17.52	3.46	14.29	125.55	-1.99
		0.25	-16.10	4.74	15.90	127.47	-0.23
		0.50	-14.32	5.98	17.63	130.28	2.36
		0.75	-9.98	8.95	22.01	136.92	7.95
		0.90	-4.78	10.68	26.30	146.61	13.93
$D = 1.5$	RMSE	-1	53.89	38.48	34.59	84.27	37.75
		0	46.80	35.80	32.87	94.30	31.58
		0.25	42.96	35.37	33.38	100.52	29.46
		0.50	38.14	35.28	34.49	108.72	27.47
		0.75	31.39	37.52	39.19	125.12	30.47
		0.90	26.34	37.18	42.99	143.63	34.60
	bias	-1	-47.10	-17.46	-11.81	80.23	-26.49
		0	-38.70	-8.45	-2.30	90.69	-15.95
		0.25	-33.64	-3.27	3.31	97.01	-9.73
		0.50	-27.48	2.46	9.89	105.60	-1.63
		0.75	-16.70	9.93	20.51	122.39	12.08
		0.90	-6.57	12.56	28.31	141.43	21.33
$D = 2$	RMSE	-1	115.00	84.38	82.25	28.39	95.05
		0	86.81	56.17	53.86	38.73	62.95
		0.25	74.44	47.27	44.63	52.59	49.55
		0.50	58.38	39.44	36.94	73.37	34.51
		0.75	38.93	40.06	40.25	107.48	33.25
		0.90	28.53	40.40	45.94	138.64	41.05
	Bias	-1	-111.56	-76.95	-74.96	-7.70	-91.03
		0	-81.94	-43.41	-41.10	26.89	-56.20
		0.25	-68.14	-28.57	-25.69	43.81	-39.48
		0.50	-50.18	-10.23	-6.12	67.38	-16.73
		0.75	-25.87	9.77	18.08	103.71	13.91
		0.90	-8.72	14.55	30.73	136.07	28.77

For large N, it is found the best estimator for different situations. Table 4.7 presents the performance of various estimator for $N = 1000$ and $\mu = 1$. The Chao1 estimator outperforms other estimators when $D = 1.1$. The new_1 estimator performs well when $D = 1.25$ while the new_2 estimator works well when $D > 1.25$. When looking at the WLR estimator, it performs as the second best in some situations. For example, when $D = 1.5$ and $a = -1$, the results show the new_2 estimator is the best compared to the other estimators with RMSE=59.78. The WLR estimator has $\text{RMSE}_{WLR} = 117.56$ which is better than the Chao1 the iChao1 and the new_1 estimators with RMSE as 175.59, 126.20 and 113.90 respectively (Table 4.7).

Table 4.8 shows the results for $N = 1000$ and $\mu = 2$. It seems the WLR estimator approximates better when compared to $\mu = 1$. When $D = 1.1$, the results indicate that the Chao1 estimator is a good approximation while the WLR estimator performs well as the second best. However, the WLR estimator performs the best in many situations for $D > 1.1$. For the new_2 estimator, it approximates not good especially for $D < 2$. For example with $D = 1.25$ and $a = 0$, the WLR estimator is the best performance with RMSE=25.94 while the new_2 estimator has a bigger RMSE as five times with 127.88.

4.5.2 Real data example

Rocchetti et al. (2011) use the chi-squared goodness of fit statistic to assess the overall fit of the regression model. The estimated frequencies are obtained recursively by

$$\widehat{f}_{x+1} = \frac{\widehat{f}_x \exp(\widehat{y}_x)}{(x+1)},$$

where $x = 1, 2, \ldots, m-1$ and m is the truncation point. The chi-square

statistic for goodness of fit is defined by

$$\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \widehat{f}_x)^2}{\widehat{f}_x}$$

with degrees of freedom $m-2$ as a result of estimating the parameters α and γ in the regression model (Rocchetti et al., 2011).

Figures 4.6 and 4.7 show the log-scale of the ratio on x for real data sets with the weighted linear regression line. The trend in some data sets are clearly not linear such as pollutants, heroin and beetle site1 data. The inclusion of quadratic terms is alternative way to model these data, but we do not consider this approach in this thesis.

In Table 4.9, the WLR estimator is used to estimate the number of population size by applying to the real data sets. The results of goodness of fit indicate that the WLR estimator can be used to fit data for Christmas bird data, tropical trees1 data and tropical tree3 data with $\widehat{N}_{WLR} = 135(K = 126, p = 0.1026)$, $\widehat{N}_{WLR} = 195(K = 152, p = 0.5495)$ and $\widehat{N}_{WLR} = 137(K = 76, p = 0.4333)$ respectively.

Considering the Chao1 estimator as the lower bound, the WLR estimator underestimates the true number of species and the goodness of fit results indicate that there are some real data sets for which the model fits poorly including Malaysian butterfly data, pollutants data, heroin users data and tropical tree2 data. Amongst other estimators, the results of the iChao1 and the new_1 estimator are similar for most data sets. For example, for Malaysian butterfly data ($K = 620$), the WLR can estimate the species richness with $\widehat{N}_{WLR} = 692$ while other estimators give $\widehat{N}_{Chao1} = 714$, $\widehat{N}_{iChao1} = 737$, $\widehat{N}_{new1} = 737$ and $\widehat{N}_{new2} = 748$.

In particular, the WLR estimator overestimates when compared with the Chao1 estimator. Particularly, for the beetle site2 data, the WLR estimator gives $\widehat{N}_{WLR} = 617$ greater than the Chao1 estimator $\widehat{N}_{Chao1} = 463$. The estimated standard error of the WLR estimator is very large for this data set with 280.32, which is twice as large as that for other estimators.

Table 4.9: Comparison of six estimators of total number for real data sets and p-value from χ^2 goodness of fit test for the WLR estimator.

Data	$\widehat{N}_{Chao1}$	$\widehat{N}_{iChao1}$	$\widehat{N}_{new1}$	$\widehat{N}_{new2}$	$\widehat{N}_{WLR}$	p-value$_{wlr}$
Malaysian Butterfly	714	737	737	748	692	0.0000
Pollutants	1789	1916	1910	1917	1738	0.0000
Christmas Bird	134	136	136	138	135	0.1026
Heroin users	10782	11151	11151	11579	10648	0.0000
Beetle Site1	284	297	305	293	295	0.0030
Beetle Site2	463	489	501	474	617	0.0027
Tropical tree1	187	196	196	202	195	0.5495
Tropical tree2	140	147	148	147	134	0.0043
Tropical tree3	108	116	115	114	137	0.4333

Data	$\widehat{se}_{Chao1}$	$\widehat{se}_{iChao1}$	$\widehat{se}_{new1}$	$\widehat{se}_{new2}$	$\widehat{se}_{WLR}$
Malaysian Butterlfy	22.66	28.81	26.91	21.90	15.17
Pollutants	62.96	74.13	72.50	60.79	62.12
Christmas Bird	6.04	7.03	7.38	5.93	4.95
Heroin users	82.90	100.60	102.14	82.77	119.15
Beetle Site1	50.47	52.52	53.57	49.23	53.23
Beetle Site2	136.27	139.78	141.58	134.33	280.32
Tropical tree1	13.58	16.21	16.35	13.29	10.83
Tropical tree2	16.84	24.57	19.05	16.17	9.77
Tropical tree3	16.02	20.44	18.10	15.39	24.41

Figure 4.6: Scatter plot with the weighted linear regression line of $\log(\mathrm{r}_x)$ on x for Malaysian butterfly, pollutants, Christmas bird, heroin users and beetle data sets.

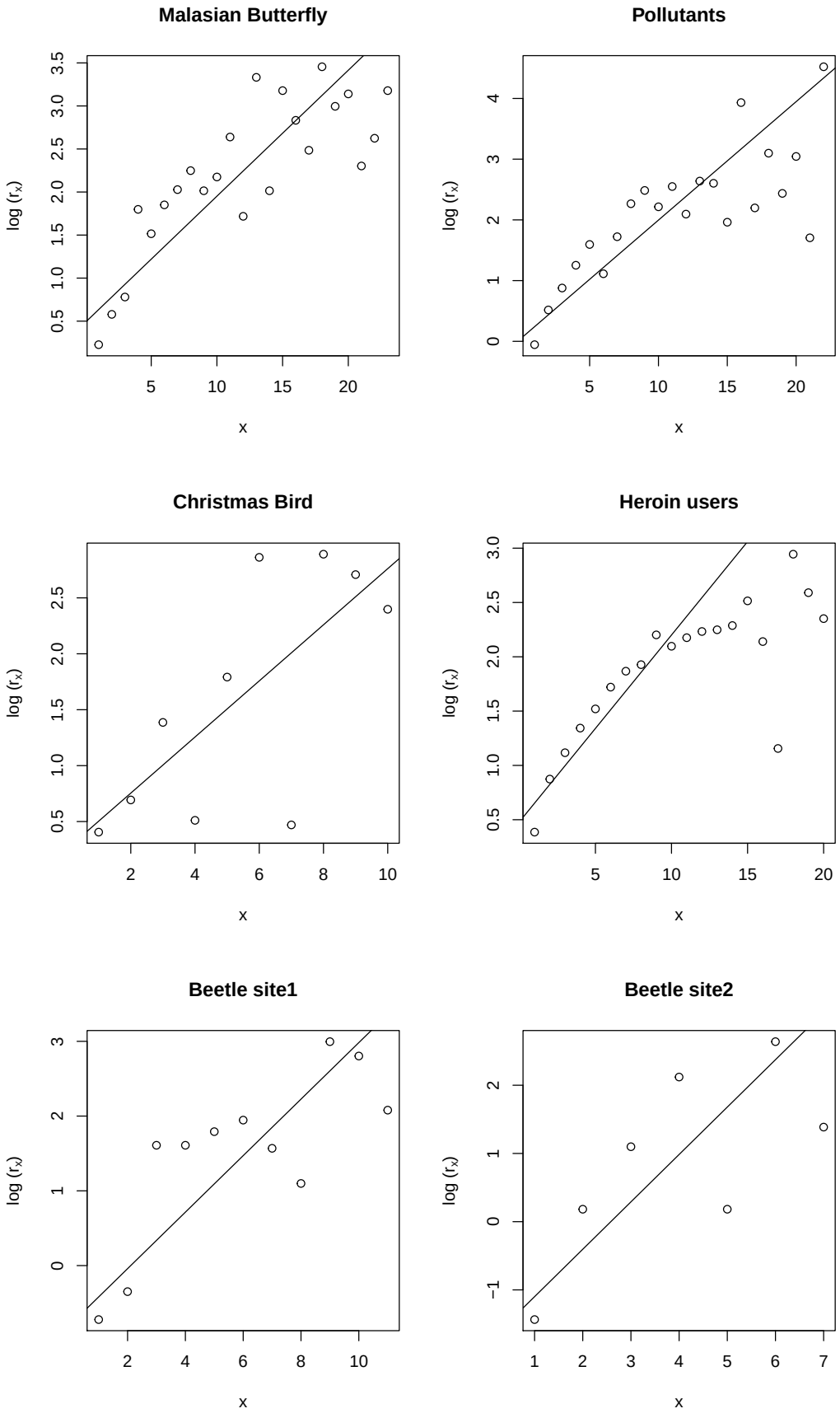
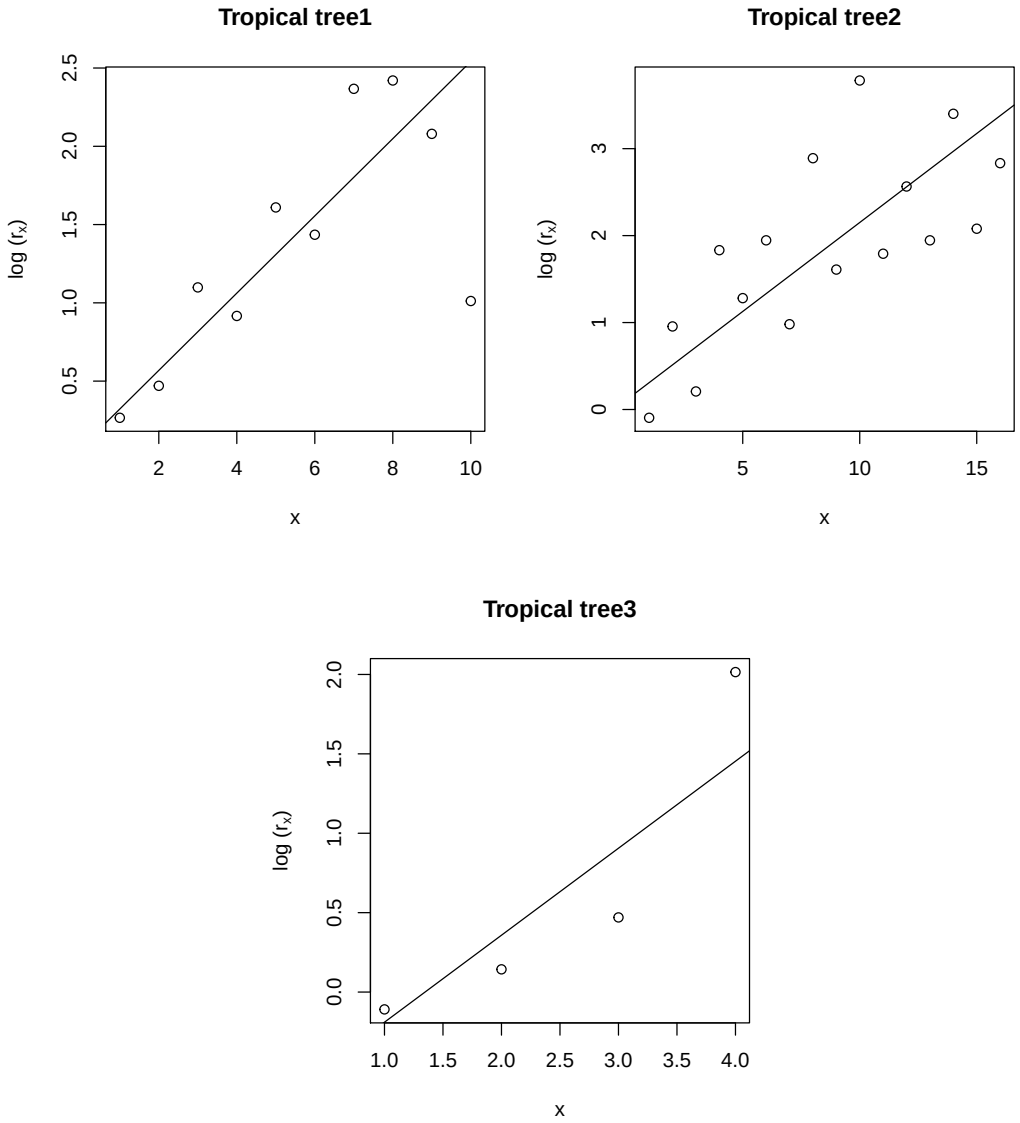


Figure 4.7: Scatter plot with the weighted linear regression line of $\log(\mathrm{r}_x)$ on x for tropical tree data sets.



4.6 Conclusion

Estimating the number of species can provide a measure of biodiversity in an ecosystem. The variation in environment affects the species abundance. That leads us to consider the PT distribution, a model which has the property of a flexible model for count data that exhibits overdispersion, zero-inflation and long right-tail. The difficulty for the number of species estimation is how to

estimate the frequency of unseen species. The WLR analysis can be used for this issue and can address problems with MLE.

Because the probabilities of the PT distribution are only available recursively, the relationship between the log ratio and the independent variable is complex. In this study, we only considered the linear model. Using a nonlinear regression model is an alternative approach which could improve estimation of the number of species Böhning (2015).

In some situations, it is difficult to fit the model using the WLR approach, especially for small sample size. Sometimes there are many small frequencies that lead to the WLR not working well. In particular, the analysis does not work when the frequency of f_k is zero for $k > 0$. An alternative way to handle this issue is smoothing the observed frequencies for improving the model. In Chapter 5, We have investigated the use of nonparametric kernel estimation to smooth the frequencies.

Chapter 5

Data Smoothing

5.1 Introduction

The idea of smoothing frequencies in relation to estimating the number of species dates back to Good (1953), who proposed an estimator based on smoothed frequencies for the probability that the next species detected will be a previously unseen species. Simonoff (1995) mentions that for sparse data, data smoothing can lead to an estimated probability function that has better performance in analysis than simply using sample proportions. Many approaches for smoothing data have been proposed for discrete data including the empirical estimator, shrinkage estimators, Bayes methods, penalized likelihood and kernel estimator.

In this chapter, smoothing the observed frequencies is considered for improving estimation of the number of species. We have focused on the use of nonparametric kernel estimation to smooth the frequencies. The kernel smoothing method is considered in order to estimate the probability function. Although it is used mostly for continuous data, the discrete kernel estimator has been explored in many studies such as Aitchison and Aitken (1976), Aitken (1983), Wang and Van Ryzin (1981), Racine and Li (2004), Li and Racine (2010) and

Kokonendji and Kiessé (2011).

Discrete kernel estimators are reviewed in Section 5.2. Several weight functions for discrete kernel estimation including the uniform, geometric and Li and Racine (2010) kernel are considered along with the estimators from the R package **np**. The performance of the discrete kernel estimators measured using mean integrated squared error (MISE) is discussed in Section 5.3. The performance of the kernel estimators depend on the bandwidth parameter. Bandwidth selection is presented in Section 5.4. In Section 5.5, the `np` package in R for density estimation has been explored . This package is proposed for nonparametric and semiparametic kernel estimation. In our simulation study under the Poisson-Tweeidie distribution, the performance of the weighted linear regression (WLR) estimator with smoothing is compared with nonsmoothing. We compare the WLR with the Chao1 estimator which is used as a lower bound for estimating the number of species. Their performance is measured using mean squared error, bias and a risk function, which is defined later in this chapter. All the results are shown in Section 5.6. The performance of kernel estimators is summarised in Section 5.7

5.2 Discrete kernel estimator

The probability mass function of a discrete random variable which is unknown can be estimated by

$$\tilde{p}_x = \frac{1}{Mh} \sum_{i=1}^{M} K\left(\frac{x - X_i}{h}\right), \qquad i = 1, 2, \ldots, M, \tag{5.1}$$

where $X_1, X_2, \ldots, X_M$ denote a set of independent and identically distributed discrete random variables, $K\left(\frac{x - X_i}{h}\right)$ is a kernel function and h is a smoothing parameter, also called the bandwidth, $h > 0$ (Kokonendji and Kiessé, 2011).

The kernel estimator in equation (5.1) can be written in terms of the associated discrete kernel as

$$\tilde{p}_x = \frac{1}{M}\sum_{i=1}^{M} w(h, x, X_i), \tag{5.2}$$

where $w(h, x, X_i)$ is the weight function or associated discrete kernel function for count data, $w(h, x, X_i) = \frac{1}{h}K\left(\frac{x - X_i}{h}\right) > 0$ and $\sum_{i=1}^{M} w(h, x, X_i) = 1$ (Kokonendji and Kiessé, 2011).

5.2.1 Weight functions

Some weight functions for discrete kernel estimation are presented below.

- **Empirical or naive estimator** is the simplest weight function for discrete kernel smoothing, for any $h \geqslant 0$ (Kokonendji and Kiessé, 2011), it is given by

$$w(h, x, X_i) = \begin{cases} 1 \text{ for } x = X_i, \\ 0 \text{ for } x \neq X_i. \end{cases} \tag{5.3}$$

- **Aitchison and Aitken (1976)** introduced the following weight function

$$w(h, x, X_i) = \begin{cases} 1 - h \text{ for } x = X_i, \\ \dfrac{h}{c-1} \text{ for } x \neq X_i, \end{cases} \tag{5.4}$$

  where c is the number of outcomes of x, for $x \in \{0, 1, \ldots, c-1\}$.

- **Wang and Van Ryzin (1981)** proposed several discrete weight functions including *the uniform weight function* which is defined by

$$w_u(h,x,X_i) = \begin{cases} 1-h & \text{for } x = X_i, \\ \frac{h}{2k} & \text{for } |x - X_i| = 1,2,\ldots,k, \\ 0 & \text{for } |x - X_i| > k, \end{cases} \tag{5.5}$$

where $0 \leq h \leq 1$, k is a fixed integer ($k \geq 1$) and *the geometric weight function* is expressed as

$$w_g(h,x,X_i) = \begin{cases} 1-h & \text{for } x = X_i, \\ \frac{1}{2}(1-h)h^{|x-X_i|} & \text{for } x \neq X_i, \end{cases} \tag{5.6}$$

where $0 \leq h \leq 1$.

- **Li and Racine (2010)** presented another weight function which is given by

$$w(h,x,X_i) = \begin{cases} 1 & \text{for } x \neq X_i, \\ h^{|x-X_i|} & \text{for } x \neq X_i, \end{cases} \tag{5.7}$$

where again, $0 \leq h \leq 1$.

5.2.2 Other discrete kernels

Kokonendji and Zocchi (2010) recommend the *discrete triangular kernel* based on the symmetric discrete triangular distribution with mode x, arm a can be defined by

$$w_{t_1}(h,x,X_i) = \frac{(a+1)^h}{D(a,h)}\left[1 - \frac{|X_i - x|^h}{(a+1)^h}\right], \qquad X_i = x, x \pm 1, \ldots, x \pm a, \tag{5.8}$$

where $D(a,h) = (2a+1)(a+1)^h - 2\sum_{k=1}^{a} k^h$ and $h > 0$. It becomes the empirical estimators, equation (5.3), when $h \to 0$ and the discrete uniform distribution when $h \to \infty$.

Kokonendji and Kiessé (2011) presented other discrete kernels using Poisson, binomial and negative binomial distributions. Categorical data with small sample size be used to estimate can estimate the probability mass function using these kernels. Under the Poisson distribution with parameter $\lambda = x + h$, the Poisson weight function can be written as

$$w_p(h, x, X_i) = \frac{(x+h)^{X_i} e^{-(x+h)}}{X_i!}, \qquad h > 0. \tag{5.9}$$

The binomial kernel is a discrete kernel based on binomial distribution with parameters $\left(x+1, \frac{x+h}{x+1}\right)$, which is given by (Kokonendji and Kiessé, 2011)

$$w_b(h, x, X_i) = \frac{(x+1)!}{X_i!(x+1-X_i)!} \left(\frac{x+h}{x+1}\right)^{X_i} \left(\frac{1-h}{x+1}\right)^{x+1-X_i}, \qquad 0 < h \leqslant 1. \tag{5.10}$$

For the negative binomial kernel, it follows the negative binomial distribution with parameters $\left(x+1, \frac{x+1}{2x+1+h}\right)$ (Kokonendji and Kiessé, 2011). The weight function is expressed by

$$w_{nb}(h, x, X_i) = \frac{(x+X_i)!}{x!X_i!} \left(\frac{x+h}{2x+1+h}\right)^{X_i} \left(\frac{x+1}{2x+1+h}\right)^{x+1}, \qquad h > 0. \tag{5.11}$$

Figure 5.1 shows the sample frequencies and the smoothed frequencies using the Li and Racine (2010) kernel function, for data were simulated from a PT distribution with $N = 100$, $\mu = 2$, $D = 1.25$, $a = 0$. When applying the kernel estimation to smooth the frequencies, the kernel estimator can improve the zero frequencies or small frequencies. Most smoothed frequencies are closer to the expected value than the sample frequencies. For example, when $X = 6$, there is sample frequency of zero. After using kernel estimation for smoothing data, it can lead to improved estimated frequencies around one.

For sparse data, the number of species seen k times might be zero. Smoothing

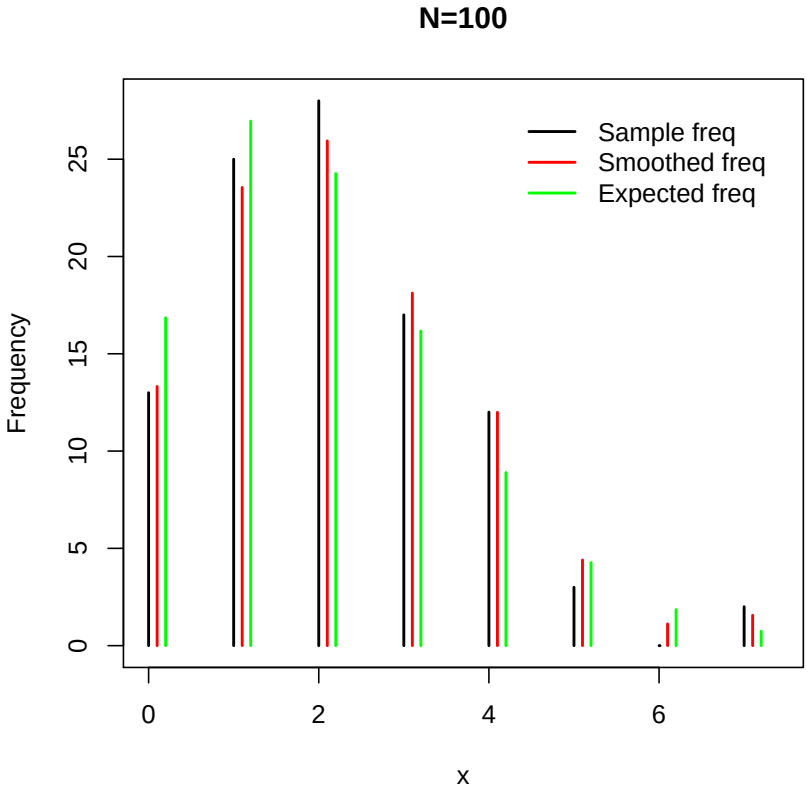


Figure 5.1: Plot of the unsmoothed and smoothed frequencies comparing to the expected frequencies based on data simulated from the PT distribution with $N = 100$, $\mu = 2$, $D = 1.25$, $a = 0$. The smoothed frequencies were estimated using the kernel estimator by Li and Racine (2010) .

data is an alternative way to handle this problem. When smoothing data, we can sometimes increase the number of frequencies used in analysis, which may improve the performance of the WLR method.

5.3 The performance measurement of the estimator

The difference between the true probability function and kernel estimator is used to measure the performance of the smoothing method. Mean integrated squared error (MISE) is a widely used criterion. This criterion assesses the

error of kernel estimation in terms of expected total mean squared error (MSE) which can be written as

$$\begin{aligned} \text{MISE}(\tilde{p}_x) &= \text{E}\left[\sum_x [\tilde{p}_x - p_x]^2\right] \\ &= \sum_x \text{MSE}(\tilde{p}_x) \\ &= \sum_x \text{var}[\tilde{p}_x] + \sum_x \text{bias}^2[\tilde{p}_x]. \end{aligned} \tag{5.12}$$

5.4 Bandwidth Selection

The choice of the smoothing parameter is a crucial factor that affects the performance of a kernel estimator. A smoothing parameter that is too small ($h \to 0$) can give an undersmoothed estimator while a smoothing parameter that is too large ($h \to \infty$) can lead to oversmoothing. Plug-in methods are a common approach for bandwidth selection. The optimal bandwidth can be estimated by minimizing the mean integrated squared error (MISE) in equation (5.12). Thus, we have that

$$h_{MISE} = \arg\min_{h>0} \text{MISE}(\tilde{p}_x) \tag{5.13}$$

is the optimal bandwidth.

Wang and Van Ryzin (1981) used this choice to derive a formula for the optimal bandwidth for the uniform kernel, which is given by

$$h_u = \alpha_1 \left(1 + \frac{1}{2n} + (K-1)\alpha_{10}\right)^{-1}, \qquad n = 1, 2, \ldots \tag{5.14}$$

where $\alpha_1 = 1 - \sum_x p_x^2 + \frac{1}{2}A_1/n$, $\alpha_{10} = \sum_x p_x^2 - A_1/n + \frac{1}{4}A_0/n^2$,
$A_0 = \sum_x \left(\sum_{|x-Xi|=1}^{n} p_{X_i}\right)^2$, $A_1 = \sum_x \sum_{|x-X_j|=1}^{n} p_x p_{X_i}$ and K is the number of seen species.

For the geometric kernel, the corresponding formula is given by

$$h_g = \beta_1 \left(\frac{3}{2} + B_1 - B_2 + (K-1)\beta_{10} \right)^{-1} \tag{5.15}$$

where $\beta_1 = 1 - \sum_x p_x^2 + \frac{1}{2}B_1$, $\beta_{10} = \sum_x p_x^2 - \beta_1 + \frac{1}{4}B_0$, $B_0 = \sum_x (p_{x-1} + p_{x+1})^2$, $B_1 = \sum_x p_x(p_{x-1} + p_{x+1})$, $B_2 = \sum_x p_x(p_{x-2} + p_{x+2})$ and K is the number of seen species.

However, these expressions depend on the true probability function, which is unknown. Therefore, the empirical probabilities or relative frequencies are used to estimate the optimal bandwidth.

Another bandwidth selection method which is widely used for kernel estimation is least squares cross validation (LSCV). The accuracy between $\tilde{p}_x$ and p_x is measured using the integrated squared error (ISE), which is given by

$$\begin{aligned} \text{ISE}(\tilde{p}_x) &= \sum_x [\tilde{p}_x - p_x]^2 \\ &= \sum_x (\tilde{p}_x)^2 - 2\sum_x \tilde{p}_x p_x + \sum_x p_x^2 \end{aligned} \tag{5.16}$$

where $\sum_x \tilde{p}_x p_x$ is the expected value of $\tilde{p}_x$. The last term is independent of the bandwidth parameter so only the first two terms are considered for estimating the optimal bandwidth. Rudemo (1982), Bowman (1984) and Stone (1984) developed the LSCV approach by minimizing the first two terms of ISE. This criterion can be used to find the optimal bandwidth for a discrete kernel (Kokonendji and Kiessé, 2011). Let $\tilde{p}_{-i}$ denote the estimator of p_i when cell i is omitted. The second term, $\sum_x \tilde{p}_x p_x$, can be replaced by $\frac{1}{M}\sum_{i=1}^{M} \tilde{p}_{-i}(X_i)$, where $\tilde{p}_{-i}(X_i) = \frac{1}{M-1}\sum_{i=1}^{M}\sum_{i \neq j} w(h, X_i, X_j)$. Then, we have

$$h_{lscv} = \arg\min_{h>0} \text{LSCV}(h) \tag{5.17}$$

where

$$\begin{aligned} \text{LSCV}(h) &= \sum_{x=1}^{M} (\tilde{p}_x)^2 - \frac{2}{M} \sum_{i=1}^{M} \tilde{p}_{-i}(X_i) \\ &= \sum_{x=1}^{M} \left\{ \frac{1}{M} \sum_{i=1}^{M} w(h, x, X_i) \right\}^2 - \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{i \neq j} w(h, X_i, X_j). \end{aligned}$$

5.5 The np package for density estimation

Hayfield and Racine (2008) created the `np` package in R to estimate the density which is unknown using kernel estimation, including nonparametric and semiparametric estimators. The package can estimate both univariate and multivariate distributions. We have investigated this package for estimating the probability mass function for categorical data. Nonparametric density estimation with optimal bandwidth selection is available in this package.

The `np` package in R can calculate nonparametric kernel density estimates with the function `npudens()`. The discrete kernel functions such as the geometric and Li and Racine (2010) kernel are available in the `np` package. Automatic bandwidth selection procedures, such as LSCV are also available. For large sample sizes, the procedure requires quite a long computation time.

5.6 Simulation study

In this Section, we explored the performance of the kernel estimator for smoothing data. Data were simulated from the PT distribution and analysed using applied with the WLR estimator. The result for the WLR estimator with non-smoothing and smoothing are compared to the Chao1 estimator. This study was conducted as follows:

- The count data were generated under the PT distribution with the parameters $\mu = 1, 2,\ D = 2, 1.5, 1.25, 1.1$ and $a = -1, 0, 0.25, 0.5, 0.75, 0.9$

and $N = 100, 1000$.

- 1000 simulations were run for each set of parameters; $2 \times 4 \times 6 \times 2$ combinations.

- The frequencies were smoothed using the uniform kernel, the geometric kernel and Li and Racine (2010) kernel functions. The bandwidth selection of the uniform kernel function is chosen by equation (5.14). For the geometric and Li and Racine (2010) kernel function, the probability mass function is estimated using the `np` package with the LSCV method for bandwidth selection.

- After smoothing the data, the frequency is greater than zero. Then, all data can be used potentially in the WLR method. However, if the smoothed frequencies are less than 0.5, the count data is cut at the first smoothed m frequencies at $f_m > 0$ and $f_{m+1} < 0.5$ for use in the analysis.

- The Chao1 estimator is used to estimate species richness as a lower bound and compared with the WLR estimator with nonsmoothing (WLR), the WLR with smoothing by the uniform kernel (WLR_u), the geometric kernel (WLR_g) and Li and Racine (2010) kernel (WLR_l). For the WLR approach, when m = 2,3, the Chao1 estimator is used to estimate species richness.

- The performance of each estimator was summarised by the root mean square error (RMSE), the bias, and the estimated and true standard error for each approach.

Figures 5.2 - 5.3 show RMSE of the WLR estimator using the kernel of Li and Racine (2010) (WLR_l) for $\mu = 1$ and $\mu = 2$. The results indicate that the performance of the WLR_l estimator for $\mu = 2$ is better than $\mu = 1$. The performance of the WLR_l improves significantly when $\mu = 2$. In Figure 5.2,

when $N = 100$ and $\mu = 1$, the WLR_l estimator performs the best under the PT distribution with $a = 0$. When $N = 100$ and $\mu = 2$, the WLR_l estimator works well with $a = 0.9$ as shown in Figure 5.3.

For large N, the WLR_l estimator performs well with the different value of a. When $N = 1000$ and $\mu = 1$, the PT distribution with $a = 0.5$ provides a small RMSE when $1.1 \leq D \leq 1.5$ and the model $a = 0.75$ gives the best approximation when $D = 2$ (Figure 5.4). When $\mu = 2$, the PT distribution with $a = 0.75$ is appropriate for the WLR_l estimator (Figure 5.5).

The Chao1 estimator is considered as a lower bound for species richness estimation. In our simulation study, it is used to compare the performance with the WLR estimator for both with and with out smoothing. Tables 5.1 - 5.4 show the performance of various estimators based on the PT distribution. When $a = 0$ and $a = 0.5$, they present the negative binomial distribution and the Poisson inverse Gaussian distribution respectively. The performance of all estimators for $\mu = 2$ is much more accurate than for $\mu = 1$. There is some decrease in RMSE and bias for $\mu = 2$, a reduction of around three or four times when compared to $\mu = 1$.

In Tables 5.1 and 5.3, when $\mu = 1$, the results indicate that the Chao1 estimator performs the best when compared with other estimators for both $a = 0$ and $a = 0.5$. However, when $\mu = 2$, the WLR estimator with smoothing outperforms the Chao1 estimator for both $a = 0$ and $a = 0.5$ especially the kernel of Li and Racine (2010) as shown in Tables 5.2 and 5.4.

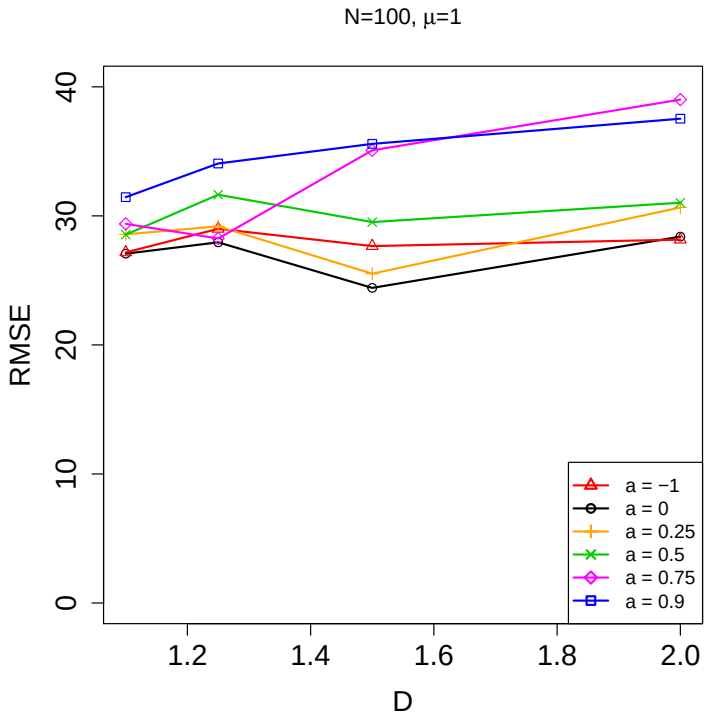


Figure 5.2: RMSE for the WLR estimator using the kernel of Li and Racine (2010) based on data from the PT distribution; $N = 100$, $\mu = 1$, $D = 2, 1.5, 1.25, 1.1$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$.

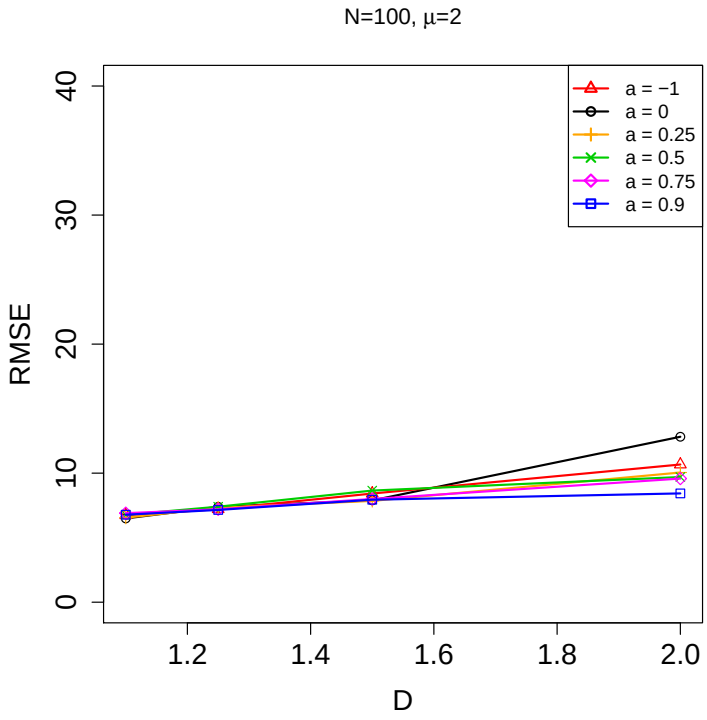


Figure 5.3: RMSE for the WLR estimator using the kernel of Li and Racine (2010) based on data from the PT distribution; $N = 100$, $\mu = 2$, $D = 2, 1.5, 1.25, 1.1$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$.

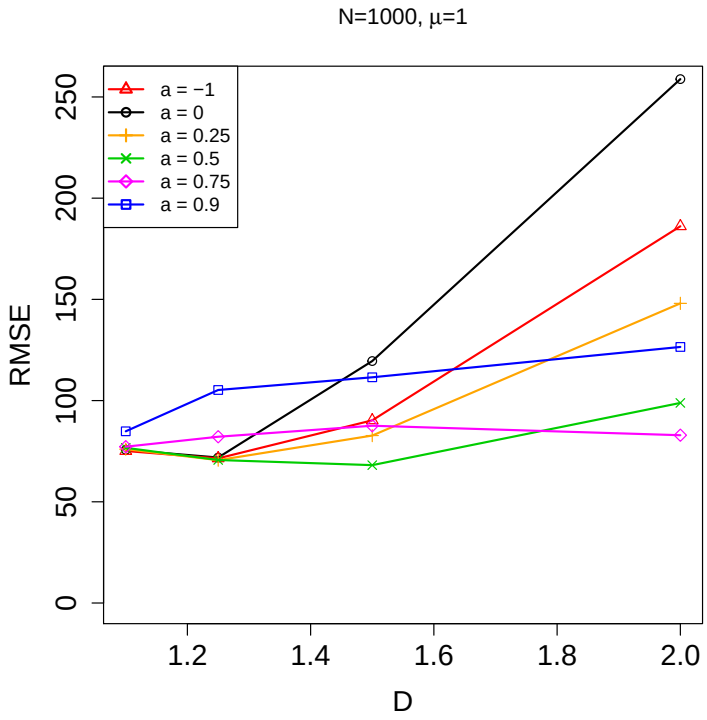


Figure 5.4: RMSE for the WLR estimator using the kernel of Li and Racine (2010) based on data from the PT distribution; $N = 1000$, $\mu = 1$, $D = 2, 1.5, 1.25, 1.1$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$.

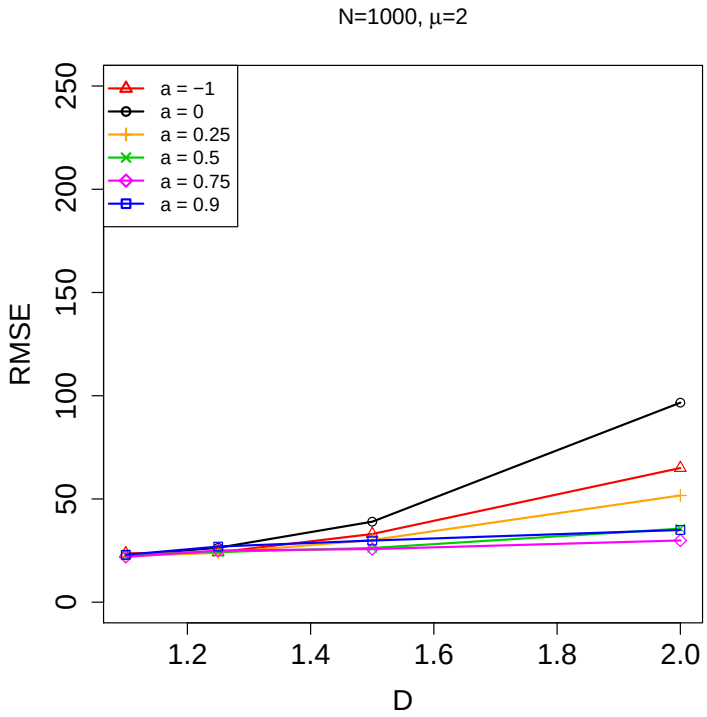


Figure 5.5: RMSE for the WLR estimator using the kernel of Li and Racine (2010) based on data from the PT distribution; $N = 1000$, $\mu = 2$, $D = 2, 1.5, 1.25, 1.1$, $a = -1, 0, 0.25, 0.5, 0.75, 0.9$.

Table 5.1: RMSE, bias, true standard error and estimated standard error for $\widehat{N}$ based on the WLR estimator with nonsmoothing, the WLR with smoothing and the Chao1 estimator ; $N = 100, 1000$, $\mu = 1$, $D = 1.1.1.25.1.5.2$, $a = 0$ using 1000 simulations.

	D	WLR	WLR_u	WLR_g	WLR_l	Chao1
$N = 100$						
RMSE	1.1	35.47	32.14	27.53	27.17	17.56
	1.25	36.66	33.09	29.26	29.00	18.04
	1.5	33.26	30.48	27.78	27.66	20.92
	2	30.98	29.47	28.17	28.16	27.05
bias	1.1	11.39	8.33	6.17	5.93	-0.50
	1.25	9.84	6.44	5.07	4.94	-4.61
	1.5	2.69	-0.81	-0.80	-0.78	-12.04
	2	-6.90	-10.71	-8.67	-8.59	-20.86
$\text{se}(\widehat{N})$	1.1	33.59	31.04	26.83	26.52	17.55
	1.25	35.31	32.46	28.82	28.57	17.44
	1.5	33.15	30.47	27.76	27.65	17.11
	2	30.21	27.46	26.81	26.82	17.22
$\widehat{se}(\widehat{N})$	1.1	45.41	39.06	32.97	32.46	18.59
	1.25	48.66	41.66	36.32	35.90	18.19
	1.5	40.64	34.24	31.66	31.45	17.48
	2	45.08	37.59	36.73	36.67	17.59
$N = 1000$						
RMSE	1.1	77.23	76.28	75.11	75.10	59.85
	1.25	72.49	72.25	71.47	71.46	92.05
	1.5	89.95	91.64	90.24	90.24	150.93
	2	186.10	188.85	186.13	186.12	252.26
bias	1.1	11.04	8.82	7.84	7.83	-32.44
	1.25	-7.98	-10.63	-10.47	-10.46	-77.92
	1.5	-61.57	-64.52	-62.90	-62.89	-143.22
	2	-177.38	-180.38	-177.52	-177.51	-248.20
$\text{se}(\widehat{N})$	1.1	76.44	75.77	74.70	74.69	50.30
	1.25	72.05	71.46	70.70	70.69	49.01
	1.5	65.57	65.08	64.71	64.71	47.64
	2	56.30	55.91	55.96	55.96	45.08
$\widehat{se}(\widehat{N})$	1.1	79.51	78.13	77.34	77.33	50.21
	1.25	74.84	73.48	73.30	73.30	48.76
	1.5	68.31	66.95	67.39	67.39	46.59
	2	57.91	56.56	57.58	57.58	42.68

Table 5.2: RMSE, bias, true standard error and estimated standard error for $\widehat{N}$ based on the WLR estimator with nonsmoothing, the WLR with smoothing and the Chao1 estimator ; $N = 100, 1000$, $\mu = 2$, $D = 1.1.1.25.1.5.2$, $a = 0$ using 1000 simulations.

	D	WLR	WLR_u	WLR_g	WLR_l	Chao1
N=100						
RMSE	1.1	9.08	8.15	6.96	6.74	7.66
	1.25	9.35	8.41	7.41	7.21	8.47
	1.5	10.83	9.81	8.60	8.42	9.90
	2	12.00	11.50	10.82	10.67	12.08
bias	1.1	2.07	1.86	0.37	0.08	0.58
	1.25	0.91	0.51	-1.06	-1.32	-0.88
	1.5	0.40	-0.40	-2.14	-2.35	-2.25
	2	-4.06	-5.18	-6.36	-6.36	-6.81
$\text{se}(\widehat{N})$	1.1	8.85	7.94	6.95	6.74	7.64
	1.25	9.31	8.40	7.33	7.09	8.43
	1.5	10.82	9.80	8.33	8.09	9.64
	2	11.29	10.27	8.76	8.56	9.98
$\widehat{se}(\widehat{N})$	1.1	9.49	7.96	6.90	6.58	8.06
	1.25	9.90	8.34	7.06	6.77	8.52
	1.5	11.29	9.37	7.92	7.63	9.41
	2	11.75	9.70	8.83	8.64	10.16
N=1000						
RMSE	1.1	25.09	24.65	23.77	23.68	24.78
	1.25	24.95	24.55	24.21	24.15	29.99
	1.5	32.51	32.60	33.07	33.03	46.75
	2	63.81	65.00	65.02	64.98	86.80
bias	1.1	1.79	2.10	-1.11	-1.36	-6.22
	1.25	-2.19	-2.42	-5.66	-5.87	-18.01
	1.5	-15.46	-16.30	-18.85	-18.92	-37.90
	2	-57.12	-58.59	-58.96	-58.92	-81.96
$\text{se}(\widehat{N})$	1.1	25.03	24.56	23.75	23.64	23.99
	1.25	24.86	24.43	23.54	23.42	23.98
	1.5	28.59	28.23	27.17	27.08	27.37
	2	28.44	28.17	27.41	27.39	28.57
$\widehat{se}(\widehat{N})$	1.1	23.37	22.65	21.77	21.62	22.93
	1.25	24.39	23.67	22.80	22.69	24.13
	1.5	26.35	25.59	24.88	24.82	25.80
	2	28.41	27.63	27.50	27.49	27.95

For example with $N = 100$, $\mu = 1$, $D = 2$ and $a = 0$ (Table 5.1), the WLR without smoothing yields RMSE and bias as 35.47 and 11.39. When smoothed is used with Li and Racine (2010) kernel, it can improve RMSE as 27.17and 5.93. For the Chao1 estimator, it gives the best performance with RMSE 17.56 and bias -0.50. For another example with $N = 100$, $\mu = 2$, $D = 2$ and $a = 0$ (Table 5.2), the best estimator is the WLR with the Li and Racine (2010) kernel. The WLR_l gives RMSE and bias as 6.74 and 0.08, while the Chao1 estimator results are 7.66 and 0.58 respectively.

When N is large, all WLR estimators approximate well with the PT distribution and outperform the Chao1 estimator especially when $D > 1.1$. Smoothing technique does not improve the WLR estimators by much. The performance of the WLR with smoothing is close to the results of nonsmoothing. In some situations such as $N = 1000$ and $D = 2$, the smoothing technique does not improve the performance of the WLR estimator.

For example, when $N = 1000$, $\mu = 1$, $D = 1.5$ and $a = 0.5$ (Table 5.3), the WLR with nonsmoothing gives RMSE and bias as 68.68 and -12.95. When using theLi and Racine (2010) kernel, the WLR estimator has improved results with RMSE and bias as 68.11 and -14.88. For the Chao1 estimator, it yields large RMSE and bias are 120.92 and -111.14.

It is clear that the performance of the WLR_l estimator depends on the parameters of the PT distribution. The Li and Racine (2010) kernel outperforms the uniform and geometric kernel. However, the performance of the WLR estimator with smoothing is improved only a little. Nonsmoothing approach is sufficient for the WLR estimator based on the PT distribution.

Table 5.3: RMSE, bias, true standard error and estimated standard error for $\widehat{N}$ based on the WLR estimator with nonsmoothing, the WLR with smoothing and the Chao1 estimator ; $N = 100, 1000$, $\mu = 1$, $D = 1.1, 1.25, 1.5, 2$, $a = 0.5$ using 1000 simulations.

	D	WLR	WLR_u	WLR_g	WLR_l	Chao1
N=100						
RMSE	1.1	37.06	33.57	28.93	28.55	18.12
	1.25	40.47	36.68	31.96	31.64	18.69
	1.5	36.33	32.90	29.67	29.52	19.72
	2	36.10	32.73	31.05	31.02	23.36
bias	1.1	12.95	9.84	7.54	7.30	0.37
	1.25	12.01	8.52	6.70	6.54	-3.49
	1.5	7.32	3.56	3.17	3.19	-8.60
	2	2.53	-1.80	-0.47	-0.39	-15.11
$\text{se}(\widehat{N})$	1.1	34.73	32.09	27.93	27.60	18.12
	1.25	38.65	35.68	31.25	30.96	18.36
	1.5	35.59	32.70	29.50	29.34	17.75
	2	36.01	32.68	31.05	31.02	17.81
$\widehat{se}(\widehat{N})$	1.1	48.12	41.54	35.26	34.72	18.83
	1.25	49.83	42.67	36.36	35.88	18.82
	1.5	44.95	37.99	35.32	35.13	18.71
	2	45.52	37.85	37.61	37.56	18.86
N=1000						
RMSE	1.1	79.20	78.03	76.69	76.67	57.75
	1.25	72.33	71.49	70.66	70.65	81.24
	1.5	68.68	68.87	68.12	68.11	120.92
	2	98.63	100.85	98.86	98.85	180.37
bias	1.1	20.02	17.74	16.64	16.63	-27.64
	1.25	9.02	6.33	6.23	6.24	-65.41
	1.5	-12.95	-16.07	-14.89	-14.88	-111.14
	2	-71.71	-75.08	-72.63	-72.62	-172.59
$\text{se}(\widehat{N})$	1.1	76.63	75.98	74.86	74.85	50.70
	1.25	71.76	71.21	70.38	70.38	48.17
	1.5	67.45	66.97	66.47	66.47	47.63
	2	67.72	67.33	67.06	67.06	52.39
$\widehat{se}(\widehat{N})$	1.1	81.11	79.73	78.89	78.88	50.69
	1.25	74.86	73.54	73.27	73.27	49.75
	1.5	71.69	70.38	70.61	70.61	49.52
	2	70.07	68.69	69.41	69.41	49.79

Table 5.4: RMSE, bias, true standard error and estimated standard error for $\widehat{N}$ based on the WLR estimator with nonsmoothing, the WLR with smoothing and the Chao1 estimator ; $N = 100, 1000$, $\mu = 2$, $D = 1.1, 1.25, 1.5, 2$, $a = 0.5$ using 1000 simulations.

	D	WLR	WLR_u	WLR_g	WLR_l	Chao1
N=100						
RMSE	1.1	9.08	8.16	7.00	6.77	7.85
	1.25	10.04	8.95	7.65	7.40	8.68
	1.5	11.26	10.16	8.86	8.65	9.44
	2	11.92	10.94	9.89	9.70	11.05
bias	1.1	2.02	1.80	0.33	0.05	0.69
	1.25	2.15	1.68	-0.02	-0.29	-0.01
	1.5	1.10	0.36	-1.47	-1.69	-1.65
	2	-0.15	-1.37	-3.15	-3.28	-3.32
$\text{se}(\widehat{N})$	1.1	8.85	7.96	6.99	6.77	7.82
	1.25	9.81	8.79	7.65	7.39	8.68
	1.5	11.21	10.15	8.73	8.48	9.29
	2	11.92	10.86	9.37	9.12	10.54
$\widehat{se}(\widehat{N})$	1.1	9.27	7.82	6.75	6.44	8.07
	1.25	10.38	8.67	7.37	7.04	8.68
	1.5	11.00	9.25	8.00	7.69	9.28
	2	12.80	10.69	9.28	9.01	10.71
N=1000						
RMSE	1.1	24.46	24.06	23.02	22.92	24.12
	1.25	25.53	25.10	24.40	24.32	28.70
	1.5	26.95	26.70	26.32	26.25	38.10
	2	35.17	35.64	35.67	35.63	57.41
bias	1.1	3.41	3.73	0.50	0.25	-5.28
	1.25	0.43	0.23	-3.08	-3.29	-15.50
	1.5	-2.89	-3.73	-6.72	-6.84	-28.35
	2	-16.47	-17.97	-19.72	-19.72	-48.48
$\text{se}(\widehat{N})$	1.1	24.22	23.77	23.02	22.92	23.54
	1.25	25.52	25.09	24.21	24.09	24.15
	1.5	26.79	26.44	25.44	25.34	25.45
	2	31.07	30.78	29.72	29.67	30.75
$\widehat{se}(\widehat{N})$	1.1	23.21	22.50	21.65	21.50	22.92
	1.25	24.41	23.68	22.79	22.67	24.12
	1.5	26.68	25.93	25.13	25.05	26.04
	2	29.78	29.02	28.52	28.49	29.22

Figure 5.6 shows the performance of estimators based on the PT distribution with $a = 0$. When $N = 100$, Chao estimator performs better than the WLR estimator for both nonsmoothing and smoothing. When $N = 1000$, all WLR estimators perform well. Particularly when $D > 1.25$, all WLR estimators outperform Chao estimator significantly.

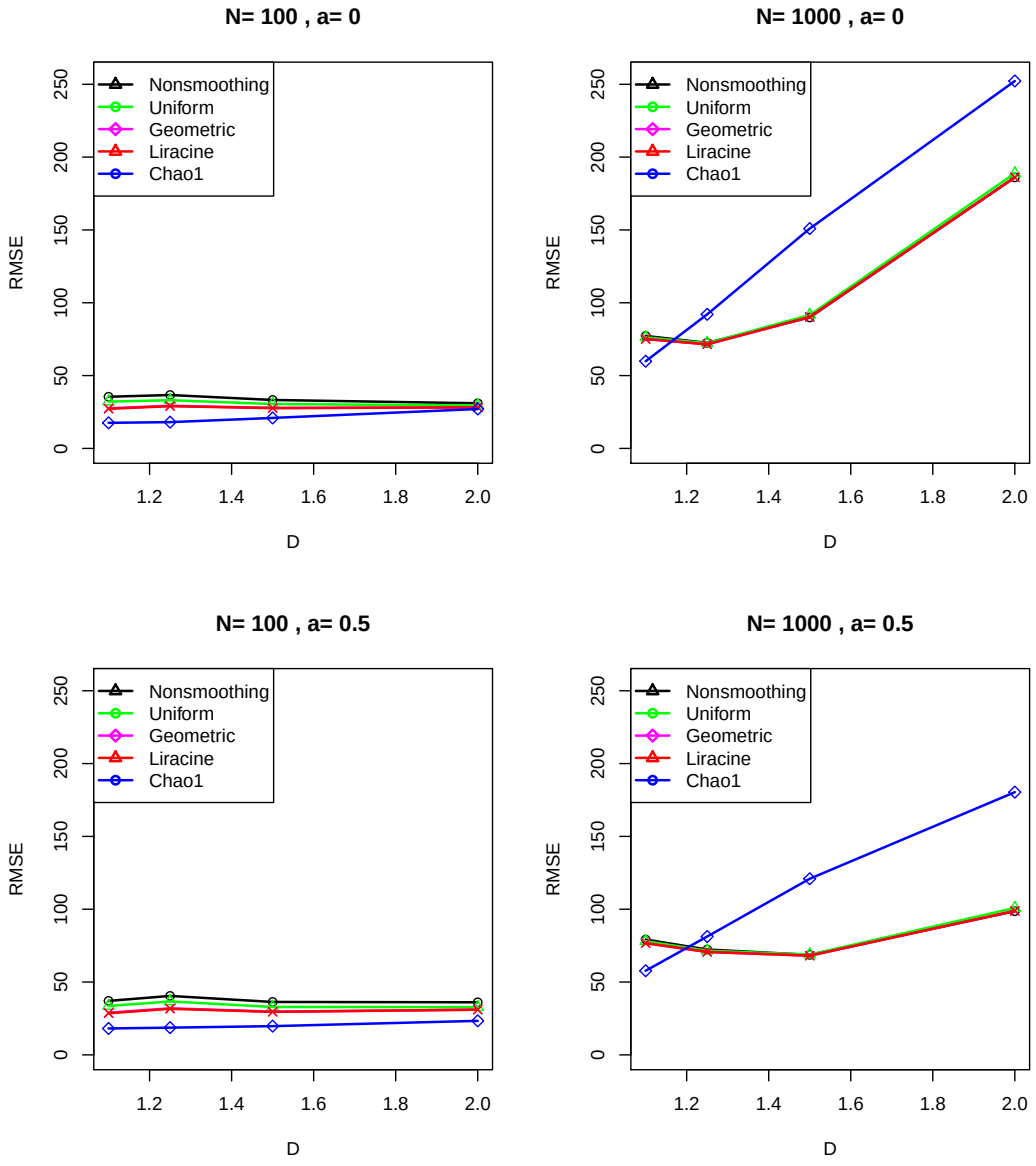


Figure 5.6: Comparison between the WLR with nonsmoothing and the WLR estimator with smoothing data and the Chao1 estimator, N=100,1000, $\mu = 1, D = 1.1, 1.25, 1.5, 2, a = 0, 0.5$.

In summary, the smoothing approach can improve the performance of the WLR estimator a little. Due to the need for using a long time in bandwidth

selection, the nonsmoothing approach is preferred when N is large using the PT model. For example with Table 5.4, we used 28 hours for computation.

5.7 Conclusion

The WLR estimator based on the PT distribution for estimating the species richness is considered with truncated data. Data is cut off at the frequency seen m times with the condition $f_m > 0$ and $f_{m+1} = 0$. Kernel smoothing approach is an alternative way to improve sparse data which has zero frequency count. This technique can handle the problem about zero frequency in the weighted linear regression analysis. Therefore, smoothing can be used in any situation. However, the kernel smoothing estimation can improve the performance of WLR estimator only a little. It is not always necessary to use smoothing.

The WLR estimator can work well with the PT distribution for large N. It outperforms Chao estimator significantly especially when D increases. When applying the smoothing technique, it takes a very long time to compute as a result of the optimal bandwidth parameter selection. The performance of the smoothing approach is not far from the nonsmoothing approach. Then, the WLR approach with nonsmoothing can be used for large N. When N is small, the Chao estimator is more appropriate for species richness estimation.

There are other kernel functions for discrete data such as triangular, binomial, Poisson and negative binomial kernel and so on, but they are not explored in our study. The boundary problem in optimization when choosing the bandwidth selection may need to be solved. Kokonendji and Zocchi (2010) proposed the kernel estimation for this problem. In future work, we could have explored other kernels with the WLR estimator.

In the next chapter, we investigate the distribution of the number of observed species. The methods of Hidaka (2014) and Williamson (2012) are explored. New approximations for the number of observed species are proposed and compared to the Poisson and normal approximations.

Chapter 6

New approximations for the number of observed species

6.1 Introduction

In this Chapter, we consider some alternative approximations for the distribution of the number of observed species which can be explained through the urn models. In probability theory, the *occupancy problem* arises from considering the distribution of the number of occupied urns when throwing M balls into N urns randomly. Each ball is thrown in urn i with the probability p_i and K denotes the number of occupied urns. When $p_1 = p_2 = \ldots = p_N = 1/N$, it is a special case of the distribution of occupied urns and called the *classical occupancy problem* (Johnson and Kotz, 1977).

The literature on occupancy problems considers not only the number of occupied urns, K, but also the number of urns occupied by exactly r balls, which we denote by m_r. In this notation, $K = N - m_0$. These random variables are also relevant in species sampling problems, because estimators of the number of species are based on these variables. For example, the Chao1 estimator is based on m_1 and m_2.

The Poisson and the normal approximations are the common approaches for approximating the discrete distribution. Williamson (2012) used both approximations to the exact distribution of K under the classical occupancy problem. Their performance depends on the ratio M/N. There are several situations in which the Poisson and the normal distributions arise as limiting distributions, see in Williamson (2012), Johnson and Kotz (1977) and Kolchin et al. (1978). In our study, we focus on new approximations for the distribution of the number of observed species based on two-parameter generalisations of the binomial distribution including Altham's multiplicative and additive-binomial, Pólya and COM-Poisson-Binomial distribution.

The exact probability distribution of K, along with the moment generating function from the literature on occupancy distributions, is reviewed in Section 6.2. In Section 6.3, the probability function for the classical occupancy problem and some its properties is discussed. The mean and variance of K can be derived from the low order moments of the number of occupied urns (David and Barton, 1962) and also derived from the moment generating function of the number of empty urns as well (Kolchin et al., 1978).

Various approximations to the distribution of K are presented in Section 6.4. An example about birthday coincidences is used to illustrate the various approximations in Section 6.5. Section 6.6 presents a simulation study to investigate the accuracy of the approximations for both homogeneous and heterogeneous models of p_i for occupancy distribution. The performance of the various approximations for the occupancy problem is compared in Section 6.6. Finally, in Section 6.7, the performance of approximations is summarised to indicate whether the true distribution can be approximated well by underdispersed binomial distributions.

6.2 Distribution of number of observed species

Suppose that an infinite population consists of N distinct species. Each individual is collected randomly with the probability p_i, where $\sum_{i=1}^{N} p_i = 1$. The random variable K is the number of different species encountered amongst the first M individuals. Therefore K is a positive integer in the range $1, 2, \ldots, n$, where n denotes the maximum number of species that could have been seen, defined by $n = \min(N, M)$. The exact probability function of K is given by (Hidaka, 2014)

$$\mathrm{P}(K|M,N) = \sum_{k=1}^{K}(-1)^{K-k}\binom{N-k}{N-K}\sum_{s\subseteq\bar{N}:|s|=k}\mathrm{P}_s(\boldsymbol{\theta})^M \tag{6.1}$$

where $s \subseteq \bar{N} : |s| = k$ denotes all subsets s of size k drawn from the set of $\bar{N}$, for $\bar{N} = \{1, \ldots, N\}$. $\mathrm{P}_s(\boldsymbol{\theta})$ be is the probability that an individual chosen at random from the population belongs to a species in subset s, $\mathrm{P}_s(\boldsymbol{\theta}) = \sum_{i\in s} p_i$, depending on unknown parameters $\boldsymbol{\theta}$. However, this distribution is computationally intractable when N and/or M are large.

Although the exact probabilities are difficult to compute, there are reasonably simple expressions for the exact mean and variance of K using

$$\mathrm{E}(K) = N - \sum_{i}^{N}(1-p_i)^M, \tag{6.2}$$

$$\mathrm{E}(K^2) = \mathrm{E}(K) + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\left\{1 + (1-p_{\{i,j\}})^M - (1-p_i)^M - (1-p_j)^M\right\}$$

and

$$\mathrm{Var}(K) = \mathrm{E}(K^2) - \mathrm{E}(K)^2, \tag{6.3}$$

where p_i and p_j are the relative frequencies or species abundances for the i^{th} species and the j^{th} species and $p_{\{i,j\}} = p_i + p_j$, $i, j = 1, 2, \ldots, N$. This suggests

using a generalized binomial-type approximation with parameters chosen to agree with the mean and variance of the true distribution. The generalized binomial distribution needs to be underdispersed (relative to the binomial distribution with the same mean), because it can be shown that

$$\mathrm{Var}(K) = \mathrm{E}(K^2) - \mathrm{E}(K)^2 < M\frac{\mathrm{E}(K)}{M}\left[1 - \frac{\mathrm{E}(K)}{M}\right].$$

Figures 6.1 and 6.2 are species accumulation curves with the x-axis showing the number of individuals in a sample and the y-axis showing the number of distinct species. As M tends to infinity, the number of distinct species increases and becomes increasingly close to the true species richness. The slope of the species accumulation curve depends on the species abundance model (Gotelli and Colwell, 2011). The species accumulation increases more rapidly for equal species abundance than for unequal species abundance as shown in Figures 6.1 and 6.2. The red line is the expected number of distinct species $\mathrm{E}(K)$ and the black dots are the number of distinct species K in samples of increasing size M from a simulation of the model. The results show that K and $\mathrm{E}(K)$ are similar. Then, it is clear that the expected number of distinct species can be estimated using the number of distinct species from data.

The model of p_i used in the occupancy distribution can be selected from both homogeneous and heterogeneous models. The classical occupancy problem is a special case of the occupancy distribution when p_i following the homogeneous model or $p_i = 1/N$. It is discussed in the next Sections. However, in ecology, the heterogeneity model is normally used as a result of unequal species abundance.

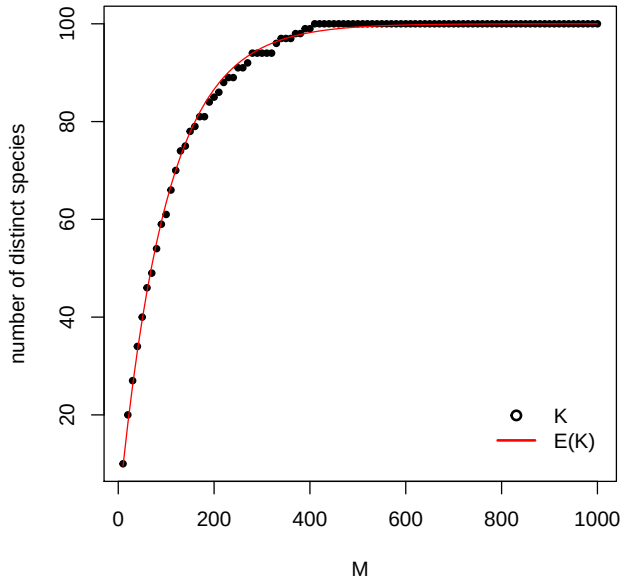


Figure 6.1: Example of species accumulation curve for $N = 100$ when all species are equally likely to be observed, M is the number of individuals collected or sample size.

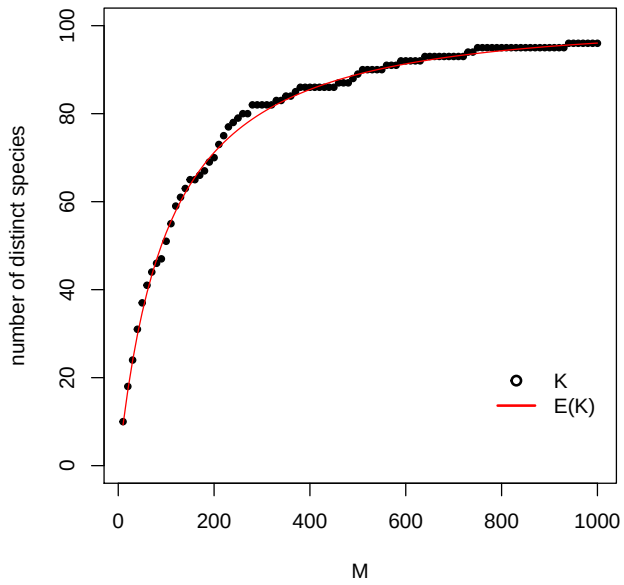


Figure 6.2: Example of species accumulation curve for $N = 100$ with unequal abundance following the broken-stick model, M is the number of individuals collected or sample size.

6.3 The classical occupancy problem

The classical occupancy model arises, when species abundance or the probability for seen species are equal for each species $p_i = \frac{1}{N}$. In relation to species sampling, the number of N urns represents the number of species while the number of balls, M, represents the number of individuals collected. The distribution of the number of urns containing at least one ball, which corresponds to the number of observed species is explored in many studies including David and Barton (1962). When $p_i = \frac{1}{N}$, equation (6.1) reduces to (Williamson, 2012)

$$P(K = x) = \binom{N}{x} x! \, \frac{S(M, x)}{N^M}, \qquad x = 1, 2, \ldots, n \tag{6.4}$$

where $n = \min(N, M)$, N and M are positive integer and $S(M, x)$ denotes a Stirling number of the second kind defined by

$$S(M, x) = \frac{1}{x!} \sum_{i=0}^{x} (-1)^i \binom{x}{i} (x - i)^M.$$

An alternative recursive relationship is as follows:

$$S(M, x) = x \; S(M - 1, x) + S(M - 1, x - 1).$$

For large M and N, accurate computation of the Stirling numbers is difficult using double precision arithmetic, such as in R. Programs such as Maple and Mathematica, that provide high precision arithmetic, are able to compute these numbers directly (Williamson, 2012).

The moment properties of the occupancy distribution can be obtained by introducing indicator variables for occupied and empty urns (David and Barton, 1962). Williamson (2012) and Samuel-Cahn (1974) present the mean and vari-

ance of K considering the probability of occupied urn i. Let $Z_i = 0$ if the i^{th} urn is empty whereas $Z_i = 1$ denotes that the i^{th} urn is occupied. Then $K = \sum_{i=1}^{N} Z_i$. When throwing M balls into N urns with the same probability, we have the probability that urn i is empty as $\mathrm{P}(Z_i = 0) = (1 - 1/N)^M$ and therefore the probability that urn i is occupied is $\mathrm{P}(Z_i = 1) = 1 - (1 - 1/N)^M$. The mean of K is given by $N\mathrm{P}(Z_i = 1)$. So

$$\mathrm{E}(K) = N - N\left(1 - \frac{1}{N}\right)^M \tag{6.5}$$

and the variance of K can be written as

$$\mathrm{Var}(K) = N\left(1 - \frac{1}{N}\right)^M + N(N-1)\left(1 - \frac{2}{N}\right)^M - N^2\left(1 - \frac{1}{N}\right)^{2M}. \tag{6.6}$$

Kolchin et al. (1978) discussed another way to find the mean and variance of K by considering the moment generating function of the number of the urns containing r balls, m_r. Let $U_{ri} = 1$ if there are r balls in urn i and $U_{ri} = 0$ otherwise. Then $m_r = \sum_{i=1}^{N} U_{ri}$. When $p_i = 1/N$, the probability that urn i occupied by r balls is given by the binomial probability

$$P(U_{ri} = 1) = \binom{N}{r} \frac{1}{N^r}\left(1 - \frac{1}{N}\right)^{M-r}. \tag{6.7}$$

Then the first moment of m_r is given by

$$\mathrm{E}(m_r) = N\mathrm{P}(U_{ri} = 1) = N\binom{N}{r}\frac{1}{N^r}\left(1 - \frac{1}{N}\right)^{M-r}. \tag{6.8}$$

If $r = 0$, we have

$$\mathrm{E}(m_0) = N\left(1 - \frac{1}{N}\right)^M,$$

where $m_0 = N - K$. The second moment of m_r is given by

$$\mathrm{E}(m_r^2) = E(m_r) + N(N-1)\frac{M^{[2r]}}{(r!)^2 N^{2r}}\left(1 - \frac{2}{N}\right)^{M-2r} \tag{6.9}$$

where factorial powers are defined by $M^{[k]} = M(M-1)\dots(M+k+1)$ and $M^{[0]} = 1$. From equation (6.8) and (6.9), if $r = 0$, the variance of m_0 is derived as

$$\begin{aligned}\mathrm{Var}(m_0) &= \mathrm{E}(m_0^2) - [\mathrm{E}(m_0)]^2 \\ &= N\left(1-\frac{1}{N}\right)^M + N(N-1)\left(1-\frac{2}{N}\right)^M - N^2\left(1-\frac{1}{N}\right)^{2m}\end{aligned}$$

which gives the same result as $\mathrm{Var}(K)$ in equation (6.6). The classical occupancy distribution is the simplest model which is relevant to species sampling, although only in the rather unrealistic case when species abundances are all equal. In practice, the heterogeneous model is generally applied more than homogeneous model in ecology as a result of unequal species abundance.

6.4 Approximation to the distribution of K

Several familiar distributions including the Poisson and normal distributions, have been discussed for approximating the distribution of K. We have investigated some alternative discrete distributions, specifically, the COM-Poisson-Binomial(CMPB) distribution, Altham's additive and multiplicative binomial distributions and the Pólya or extended beta-binomial distribution.

Let $p(x) = \Pr(K = x)$ be the true probability function of K and $p^*(x) = \Pr^*(K = x)$ be an approximating probability function. The performance of each approximation is measured by calculating a measure of the discrepancy between $p(x)$ and $p^*(x)$. There are three discrepancy criteria that are commonly used, which

are given by

$$d_1 = \sum_{i=1}^{n} \{p(x) - p^*(x)\}^2,$$

$$d_2 = \frac{1}{2} \sum_{i=1}^{n} |p(x) - p^*(x)|,$$

$$d_3 = \max_{x=1,\ldots,n} |p(x) - p^*(x)|.$$

The first two criterions are chosen following Williamson (2012), who used them to measure the performance of approximations. The measure d_2 is called the *total variation distance*. The third criterion is known as the *Kolmogorov distance*.

6.4.1 Poisson Approximation

Poisson approximations are suggested by several limit distributions for urn models as follows:

1. Let K be the number of urns occupied by the M balls and $m_0 = N - K$ the number of empty urns. Then

$$\Pr(K = x) = \Pr(m_0 = N - x).$$

If $M, N \to \infty$ and $Ne^{-M/N} \to \lambda$, then $m_0 = N - x \to \text{Poisson}(\lambda)$ (Johnson and Kotz, 1977). This suggests the following approximation to the distribution of K

$$\Pr^*_{pois}(K = x) = \frac{\exp(-\lambda)\lambda^{N-x}}{(N-x)!}, \qquad x = 1, 2, \ldots, n \tag{6.10}$$

where $n = \min(N, M)$ and $\lambda = Ne^{-M/N}$. Williamson (2012) investigated the accuracy of Poisson approximation to the occupancy distribution in the classical occupancy problem. The Poisson approximation to the

distribution of m_0 can work well for large M and N, in particular, $M > N$ or $\frac{M}{N} > 1$ (Williamson, 2012).

2. Williamson (2012) explored another result on limit distribution following Barbour and Holst (1989) when $M, N \to \infty$ and $\mathrm{E}(m_0) \to \lambda$. The distribution of m_0 can be approximated by the Poisson distribution with $\lambda = \mathrm{E}(m_0)$, where $\mathrm{E}(m_0) = N\left(1 - \frac{1}{N}\right)^M$.

3. Sevast'Yanov and Chistyakov (1964) show another results when $M, N \to \infty$. When $\frac{M}{N} - \ln M \to \ln \lambda$, the Poisson distribution can be used to approximate the distribution of m_0 similar to above results. Then the distribution of K in equation (6.10) can be used with parameter $\lambda = \frac{1}{N} e^{M/N}$.

4. Kolchin et al. (1978) discussed the result on the limit distribution when $M, N \to \infty$, $\frac{M}{N} \to 0$ and $\mathrm{Var}(K) \sim \frac{M^2}{2N} \to \lambda$. We have

$$m_0 - (N - M) = M - K \to \mathrm{Poisson}(\lambda)$$

Based on this result, the distribution of K can be approximated as follows:

$$\Pr^*_{pois}(K = x) = \frac{\exp(-\lambda)\lambda^{M-x}}{(M-x)!} \qquad , x = 1, 2, \ldots, n \tag{6.11}$$

where $\lambda = \mathrm{Var}(K)$ by equation (6.6).

These limit theorems suggest that the Poisson approximation can perform well for the classical occupancy problem under several different conditions. However, under other conditions it may not approximate the exact probability in equation (6.1) well including when the p_i's are unequal.

6.4.2 Normal Approximation

Under the condition $M \to \infty$ and $Ne^{-M/N}\left\{1-e^{-M/N}\left(1+\frac{M}{N}\right)\right\} \to \infty$, Samuel-Cahn (1974) proved that the distribution of K becomes the normal distribution. The mean and variance are given by equation (6.5) and equation (6.6) respectively. Williamson (2012) explored the performance of the normal approximation to the distribution of K. Good results are obtained when $M/N \leq 2$. However, the normal distribution is appropriate for continuous random variables. When used to approximate a discrete distribution, using a continuity correction improves the accuracy of the approximation. The approximation is

$$\Pr(K=x) \approx \Pr(x-\frac{1}{2} < W < x+\frac{1}{2}),$$

where W is the approximating normal variable. Therefore, the distribution of K can be approximated by

$$\begin{aligned} \Pr^*_{Norm}(K=x) &\approx \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{(w-\mu}{\sigma}\right)^2\right\} dw \\ &= \Phi\left(\frac{x+1/2-\mu}{\sigma}\right) - \Phi\left(\frac{x-1/2-\mu}{\sigma}\right), x \in \mathbb{R}. \end{aligned} \tag{6.12}$$

where $\mu = \mathrm{E}(W)$ and $\sigma^2 = \mathrm{Var}(W)$ are given by equation (6.5) and (6.6).

6.4.3 COM-Poisson-Binomial Approximation

Conway and Maxwell (1962) introduced a generalization of the Poisson distribution for use in queuing system problems. This distribution was rediscovered by Shmueli et al. (2005), who termed it the COM-Poisson (Conway-Maxwell Poisson) distribution. It is a flexible distribution which can be used to model both overdispersion and underdispersion. The COM-Poisson-binomial (CMPB) distribution is an analogous extension of the binomial distribution

which is discussed briefly by Shmueli et al. (2005) and more extensively by Borges et al. (2014). The probability function is given by

$$\Pr^*_{CMPB}(K=x) = \frac{\binom{n}{x}^{\nu} p^x(1-p)^{n-x}}{\sum_{x=0}^{n}\binom{n}{x}^{\nu} p^x(1-p)^{n-x}}, \qquad x=0,1,\ldots,n \tag{6.13}$$

where $n=\min(N,M)$, $p\in(0,1)$ and $\nu\in\mathbb{R}$. The distribution is overdispersed relative to the binomial when $\nu<1$ and underdispersed when $\nu>1$. For $\nu=1$, it becomes the binomial distribution.

Borges et al. (2014) considered the alternative parametrization using $\theta=\dfrac{p}{1-p}$ and divide terms of $(1-p)^n(n!)^\nu$ from equation (6.13). In terms of the parameters n,θ and ν, the distribution of K in equation (6.13) can be rewritten as

$$\Pr^*_{CMPB}(K=x) = \frac{1}{Z(\theta,\nu)}\frac{\theta^x}{x![(n-x)!]^\nu}, \tag{6.14}$$

where $Z(\theta,\nu)=\sum_{j=0}^{n}\dfrac{\theta^j}{j[(n-j)!]^\nu}$.

For the moments, there is no explicit form and they must be calculated numerically from the formula

$$\mathrm{E}(K^r) = \frac{1}{Z(\theta,\nu)}\sum_{x=0}^{n} x^r\frac{\theta^x}{x![(n-x)!]^\nu}.$$

There are potential computational issues with the distribution when $\nu>1$. These can be avoided by writing the probability function in terms of logarithms. Specifically, let $\xi_x=\nu\log\binom{n}{x}+x\log(p)+(n-x)\log(1-p)$. Then

$$\Pr^*_{CMPB}(K=x) = \exp\left\{\frac{\xi_x}{\sum_{j=0}^{n}\xi_j}\right\} \tag{6.15}$$

The R function `lchoose` is used to evaluate $\log \binom{n}{x}$.

6.4.4 Altham's multiplicative binomial Approximation

Altham (1978) developed two generalisations of the binomial distribution which are able to model both overdispersion and underdispersion. The first of these is termed the Altham's multiplicative-binomial distribution. The probability function is given by

$$\Pr{}^*_{MB}(K = x) = \frac{\binom{n}{x} p^x (1-p)^{n-x}\ \theta^{x(n-x)}}{\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}\ \theta^{x(n-x)}}, \qquad x = 0, 1, \ldots, n \tag{6.16}$$

where $p \in (0, 1)$ and $\theta > 0$. When $\theta = 1$, it reduces to the binomial distribution with parameters (n, p). This model allows for underdispersion when $\theta > 1$, and for overdispersion when $\theta < 1$. However, computation issues can arise, similarly to the CMPB distribution, for large θ. To avoid these, the probability function is again expressed in terms of logarithms as follows:

$$\Pr{}^*_{MB}(K = x) = \exp\left\{ \frac{\xi_x}{\sum_{j=0}^{n} \xi_j} \right\} \tag{6.17}$$

where $\xi_x = \log \binom{n}{x} + x \log(p) + (n-x) \log(1-p) + x(n-x) \log(\theta)$.

Let $F_n(n, p, \theta) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}\ \theta^{x(n-x)}$. The first and second moments of K are given by (Altham, 1978)

$$\mathrm{E}(K) = np(p + (1-p))^{n-1} F_n\left(\frac{p}{p + (1-p)\theta}, \theta, n-1 \right) / F_n(n, p, \theta),$$

$$\mathrm{E}[K(K-1)] = n(n-1)p^2(p+(1-p)\theta^2)^{n-2} F_n\left(\frac{p}{p + (1-p)\theta^2}, \theta, n-2 \right) / F_n(n, p, \theta).$$

6.4.5 Altham's additive binomial Approximation

The second generalized binomial distribution introduced by Altham (1978) is called the additive binomial distribution. The probability function of K can be written as

$$\Pr^*_{AB}(K=x) = \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{\alpha}{2}\left(\frac{x(x-1)}{p} + \frac{(n-x)(n-x-1)}{1-p} - n(n-1)\right) + 1\right] \tag{6.18}$$

where $x = 0, 1, \ldots, n$, $p \in (0,1)$, $n = \min(M, N)$ and to ensure a valid probability distribution, α must satisfy the conditions

$$-\min\left(\frac{p}{1-p}, \frac{1-p}{p}\right) \leq \alpha \leq 1, \qquad n = 2, \tag{6.19}$$

and

$$\frac{-2}{n(n-1)} \min\left(\frac{p}{1-p}, \frac{1-p}{p}\right) \leq \alpha \leq 2\left(n + \frac{(1-2p)^2}{4p(1-p)}\right)^{-1}, \qquad n > 2. \tag{6.20}$$

The mean and variance of K can be derived as

$$\mathrm{E}^*(K) = np \qquad \text{and} \qquad \mathrm{Var}^*(K) = np(1-p)[1 + (n-1)\alpha],$$

respectively (Altham, 1978).

6.4.6 Pólya distribution

Pólya distribution was proposed by Eggenberger and Pólya (1923) as an urn process. This refers to a sampling model with replacement from an urn containing initially a black balls and b white balls. When a ball is drawn from the urn, it is replaced along with c balls of the same color. This is repeated n times and the random variable K denotes the number of times a black ball is drawn. The probability distribution of K is given by (Johnson and Kotz,

1977)

$$\Pr^*_{Pol}(K = x) = \binom{n}{x} \frac{a(a+c)\dots(a+(x-1)c)\; b(b+c)\dots(b+(n-x-1)c)}{(a+b)(a+b+c)(a+b+2c)\dots(a+b+(n-1)c)}, \tag{6.21}$$

where $n = \min(N, M)$. Although a, b, c are integers in the urn model, they can be taken as real and equation (6.21) is still valid (with some restriction on a, b, c). If $c = 0$, this model represents the binomial distribution. If $c = -1$, it becomes a hypergeometric distribution.

Skipper et al. (2012) presented another form of this model which can be rewritten as

$$\Pr^*_{Pol}(K = x) = \binom{n}{x} \frac{p(p+\theta)\dots(p+(x-1)\theta)\; q(q+\theta)\dots(q+(n-x-1)\theta)}{(1+\theta)(1+2\theta)\dots(1+(n-1)\theta)} \tag{6.22}$$

where $q = 1 - p$, $p \in (0, 1)$ and $\theta \in \mathbb{R}$ with the constraint

$$\theta > -\min(p, q)/(n-1). \tag{6.23}$$

This constraint is needed to ensure that the probabilities given by equation (6.22) are non-negative. When $\theta = 0$, it reduces to the binomial distribution with parameter (n, p). Another special case is the hypergeometric distribution, if $n < a + b$, $p = a/(a+b)$ and $\theta = -1/(a+b)$. The mean and variance of the Pólya distribution are given by

$$\mathrm{E}^*(K) = np \qquad \text{and} \qquad \mathrm{Var}^*(K) = np(1-p)\left(1 + (n-1)\frac{\theta}{1+\theta}\right),$$

respectively (Skipper et al., 2012).

6.4.7 Choosing parameters for the approximating distribution

In order to use the distributions in Sections 6.4.3-6.4.6 to approximate the occupancy distribution, the key thing is choosing parameters of the distribution. The appropriate parameters are chosen so that the mean and variance of the approximation match the exact mean and variance.

For the COM-Poisson-Binomial and Altham's multiplicative-binomial distribution, the parameters can be chosen using optimization to find

$$\min\left\{(\mathrm{E}(K)-\mathrm{E}^*(K))^2+(\mathrm{Var}(K)-\mathrm{Var}^*(K))^2\right\},$$

where $\mathrm{E}(K)$ and $\mathrm{Var}(K)$ is given by equation (6.5) and (6.6), $\mathrm{E}^*(K)$ and $\mathrm{Var}^*(K)$ are the mean and variance of the approximation. This provides $\widehat{p}^*_C$, $\widehat{\nu^*_C}$ for CMPB model and $\widehat{p}^*_{MB}$, $\widehat{\theta}^*_{MB}$ for Altham's multiplicative-binomial model.

For Altham's additive-binomial and the Pólya distribution, their parameters can be selected easily from the explicit formulae for the mean and variance of K. The parameters of Altham's additive binomial distribution are given by $\widehat{p}_{AB}=\dfrac{\mathrm{E}(K)}{n}$ and $\widehat{\alpha}_{AB}=\dfrac{1}{N}\left(\dfrac{\mathrm{Var}(K)}{np(1-p)}-1\right)$, where $\mathrm{E}(K)$ and $\mathrm{Var}(K)$ are given by equations (6.5) and (6.6). However, $\widehat{\alpha}$ might not follow the equation (6.20). The parameters of the Pólya distribution given by $\widehat{p}^*_{Pol}=E(K)/n$ and $\widehat{\theta}^*_{Pol}=\widehat{\rho}/(1-\widehat{\rho})$, where

$$\widehat{\rho}=\frac{\mathrm{Var}(K)-\mathrm{E}(K)(1-p)}{\mathrm{E}(K)(1-p)(n-1)}$$

where $\mathrm{E}(K)$ and $\mathrm{Var}(K)$ is given by equation (6.6) and (6.7). However, the resulting value of $\widehat{\theta}^*_{Pol}$ is not guaranteed to satisfy the constraint in equation (6.23).

6.5 Example-birthday coincidences

Williamson (2012) compared the performance of the Poisson and normal approximations in the classical occupancy problem using the example about birthday coincidences in Feller (1950). In this problem, K is the number of days that are a birthday amongst a random sample of M people, N is the number of days in the year ($N = 365$). All days are assumed to be equally likely as birthdays. However, the methods of this Chapter could also be adopted to allow for seasonal variations in birth rate (e.g. Nunnikhoven (1992)).

In this example, we added the CMPB, Altham and Pólya distributions for comparing with the Poisson and normal approximations. In the classical occupancy problem, the exact probability from equation (6.4) is calculated instead of the full expression in equation (6.1). Maple is used to compute the Stirling number of the second kind. The performance of various approximations of K for $\frac{M}{N} \to 0$, small $\frac{M}{N}$ and large $\frac{M}{N}$ are shown and compared with the exact probability in Table 6.1.

As an example of large $\frac{M}{N}$, when $M = 2000, N = 365$, the Stirling number of the second kind in term of log scale is calculated as $\log(S(2000, 365)) = 10005.93113$. The exact probability from equation (6.4) can be computed as

$$\begin{aligned}
\Pr(K < 365) &= 1 - \Pr(K = 365) \\
&= 1 - \exp\left(\log(1) + \log(365!) + \log(S(2000, 365)) - 2000\log(365)\right) \\
&= 0.7839.
\end{aligned}$$

When the distribution of $m_0 = N - K$ is approximated by the Poisson distribution with parameter $\lambda = 365e^{-2000/365} = 1.5226$, equation (6.10) gives

$$\begin{aligned}\Pr^*_{Poi1}(X < 365)) &= 1 - \Pr(X = 365) \\ &= 1 - \Pr(N - X = 0) \\ &= 1 - \exp(1.5226) * (1.5226^0)/0! \\ &= 0.7819.\end{aligned}$$

When $N - K$ is approximated by the Poisson distribution with parameter $\lambda = 365(1 - 1/365)^{2000} = 1.5112$, equation (6.10) gives

$$\begin{aligned}\Pr^*_{Poi2}(K < 365) &= 1 - \Pr(K = 365) \\ &= 1 - \Pr(N - K = 0) \\ &= 1 - \exp(1.5112) * (1.5112^0)/0! \\ &= 0.7794.\end{aligned}$$

For Pois3, $M - K$ is approximated by the Poisson distribution with parameter $\lambda = \mathrm{Var}(K)$, where $\mathrm{Var}(K) = 1.470854$ and $M/N = 5.48$. As a result of large M/N, $\Pr^*_{Poi3}(K = 365) \to 0$ and the results by equation (6.11) tend to 1. Therefore, we don't consider Pois3 for this situation.

For the normal approximation, the continuity correction is used for this approximation in equation (6.12). The mean and variance are $\mu = E(K) = 363.4888$ and $\sigma^2 = \mathrm{Var}(K) = 1.470854$. Therefore

$$\Pr^*_{Norm}(K < 365) \approx \Phi\left(\frac{364.5 - 363.4888}{\sqrt{1.47084}}\right) = 0.8464.$$

Another approximation investigated is the CMPB distribution. The parameters p and ν found using optimization are $\widehat{p^*}_C = 0.9966$ and $\widehat{\nu}^*_C = 1.0385$ with $n = \min(365, 2000) = 365$, the approximate probability by equation (6.9) can

be computed as

$$\Pr^*_{CMPB}(K < 365) = 1 - P(K = 365) = 0.7847.$$

For Altham's multiplicative binomial approximation, the parameters p and θ are again calculated using optimization which give $\widehat{p}^*_{MB} = 0.9997$ and $\widehat{\theta}^*_{MB} = 1.0078$. The probability can be approximated using equation (6.16) as

$$\Pr^*_{MB}(K < 365) = 1 - P(K = 365) = 0.7839.$$

For Altham's additive binomial approximation, the parameters $\widehat{p}_{AB} = 0.9959$ and $\widehat{\alpha}_{AB} = -0.000062$. Then

$$\Pr^*_{AB}(K < 365) = 1 - P(K = 365) = 0.7838.$$

When K is distributed by the Pólya distribution, we have parameters $\widehat{p} = 0.9959$ and $\widehat{\theta} = -0.000063$. Then, the probability of birthday coincidences is calculated by

$$\Pr^*_{Pol}(K < 365) = 1 - P(K = 365) = 0.7839.$$

However, it is found that the constraints in equation (6.20) and (6.23) for Altham's additive binomial and the Pólya distribution are not satisfied, so that these approximations do not give a valid probability distribution over the full range of K. They will give negative probabilities for some values of K. Although it would be possible to constrain the parameters of the Pólya distribution, the constraint is awkward to work with. We have not investigated this further because other distributions, such as the multiplicative binomial distribution, work well without the need for constraints.

Additionally, we have investigated in small group of people, 10 and 40 people. The probability that at least two people have the same birthday can be calculated using various approaches as above. Table 6.1 shows the performance of all approximations for $M = 10$, $M = 40$ and $M = 2000$.

Table 6.1: Probability of birthday coincidences $\mathrm{P}(K < M)$ for the occupancy problem when $N = 365$

Probability	$M = 10$	$M = 40$	$M = 2000$
Exact probability	0.1169	0.8912	0.7839
Pois1: Pois$(Ne^{-M/N})$	0.9788	0.9780	0.7819
Pois2: Pois$(N(1-1/N)^M)$	0.9788	0.9780	0.7794
Pois3: Pois(Var(K))	0.1118	0.8331	-
Normal	0.1717	0.9065	0.8464
COM-Poisson Binomial	0.1170	0.8939	0.7847
Altham Multiplicative Binomial	0.1170	0.8915	**0.7839**
Altham Additive Binomial	**0.1169**	0.8911	0.7838
Pólya	**0.1169**	**0.8912**	**0.7839**
E(K)	9.8776	37.9353	363.4888
Var(K)	0.118534	1.79164	1.470854

Under the classical occupancy problem, the new approximations outperform the Poisson and the normal approximations. Particularly, the Pólya distribution provides the best approximation which is similar to the exact probability for all situations. For Altham distribution, both multiplicative and additive binomial approach give good approximation which is a slightly different from the exact one. For the COM-Poisson Binomial (CMPB) distribution, the approximated probability overestimates slightly. When $M = 2000$, the exact probability is 0.7839 while the CMPB distribution provides 0.7847.

For Pois1 and normal approximations, the results agree with Williamson (2012). The Pois1 approximation is appropriate for large $\frac{M}{N}$ while the Pois3 approxi-

mation performs well when $\frac{M}{N} \to 0$. For the Pois2 approximation, it performs similarly the Pois1 approximation. For the normal approximation, it is close to the exact probability when M/N is not large. For example in Table 6.1, the Pois1 approximation and the exact distribution give similar probabilities of 0.7819 and 0.7839 respectively. For Pois3 approximation, it overestimates when $M = 10$ with the probability 0.9788, while the exact probability is 0.1169.

6.6 Simulation Study

In this section, we explore the performance of the various approximations to the distribution of K. The useful approximations in the previous section are compared in the following simulation study. In the next Chapter, the best performing approximation is used for estimating the number of species. Due to varying species abundances in ecology, we have investigated both homogeneous and heterogeneous populations, involving both equal and unequal p_i. However, the exact distribution cannot be computed for unequal p_i. Instead, the empirical probability function is considered and compared with the approximations.

1. The exact distribution is difficult to compute, especially for large N or M. Here, the empirical probability distribution is used for approximating the exact distribution, by resampling M individuals from N species with replacement based on various relative frequency or species abundance models (p_i) and repeated in 500,000 simulations. This large number of simulations is used to ensure that the empirical distribution closely approximates the exact distribution, so that the accuracy of the various approximations can be assessed.

2. The following species abundance models for species $i = 1, 2, \ldots, N$ are considered

- model 1 : homogeneous model with $p_i = 1/N$
- model 2 : Zipf model with $p_i = c/i^{0.5}$ (Zipf1)
- model 3 : Zipf model with $p_i = c/i^2$ (Zipf2)
- model 4 : log-normal model with parameters $\mu = 0$ and $\sigma = 1$
- model 5 : broken-stick model or Dirichlet$(1, 1, \ldots, 1)$ model
- model 6 : exponential-decay model with $p_i = \exp(-i)$

3. Let $p(x)$=Pr$(K = x)$ and $p^*(x)$=Pr$^*(K = x)$. The accuracy of the approximation is measured using total variation distance $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, where $p^*(x)$ can be defined by the distribution as follows:

   - Pois1: m_0 is distributed by the Poisson $(\lambda = Ne^{-M/N})$
     P$(K = x)$ = P$(m_0 = N - x)$
   - Pois2: m_0 is distributed by the Poisson $(\lambda = N(1 - \frac{1}{n})^M)$
     P$(K = x)$ = P$(m_0 = N - x)$
   - Pois3: $m_0 - (N - M)$ is distributed by the Poisson $(\lambda = \text{Var}(K))$
     P$(K = x)$ = P$(m_0 - (N - M) = M - x)$
   - Norm: K is distributed by the normal$(\mu = \text{E}(K)$, $\sigma^2 = \text{Var}(K)))$
   - CMPB: K is distributed by the CMPB(p,ν)
   - MB: K is distributed by the Altham's multiplicative binomial(p,θ)
   - AB: K is distributed by the Altham's additive binomial(p,α)
   - Pólya: K is distributed by the Pólya(p,θ)

4. The simulations are divided into three groups. Firstly, the small populations are defined using $M, N \leq 100$. Secondly, the large populations are defined with $100 < M, N < 4500$. Lastly, the data are generated under small $M/N < 0.5$ and large $6 \leq M/N \leq 30$ (Williamson, 2012).

Figure 6.3 represents the performance of approximations for the homogeneous model, $p_i = \frac{1}{N}$. The results show that the total variation distance d_2 as a function of M/N for the Pólya and Altham's multiplicative distribution (MB) are quite similar. Both can approximate the empirical probability distribution very well. For the CMPB, it is a little bit worse than the Pólya and MB approximation. For the AB approximation, it is suitable for large M/N (i.e. $M/N > 3$). Additionally, the Pois1 and Pois2 can work well for large M/N while the Pois3 is very close to the empirical probability when M/N tends to 0. The normal approximation performs well when M/N is small (i.e. $0.5 < M/N < 1.5$). When M/N is greater than 1.5, its performance decreases significantly before improving again for large M/N.

On the other hand, the Pólya and AB approximations for occupancy distribution sometimes give negative probabilities because their parameters do not follow the conditions in equation (6.20) and (6.23). Therefore, these approximations do not necessarily give valid probability distributions, although the negative probabilities are usually close to zero. It may be possible to adapt the method of choosing the parameters of these distributions to ensure that the constraints are satisfied. However, this has not been investigated because the multiplicative binomial distribution perform as well as the other approximations and does not have complicated constraints on its parameters.

Figure 6.4 compares the approximations with the empirical probability distribution for equal p_i. It is clear that the Pólya and Altham's multiplicative binomial approximations perform well for every situation considered. Although the CMPB and normal approximations are not as good as the Pólya and Altham's, they can be used for occupancy problem. The Pois1 and Pois2 approximations are suitable for large M and N ($M/N > 4$). For the Pois3, the probability is appropriate for $M/N \to 0$ (e.g. $N = 2000, M = 40$). When M/N is large,

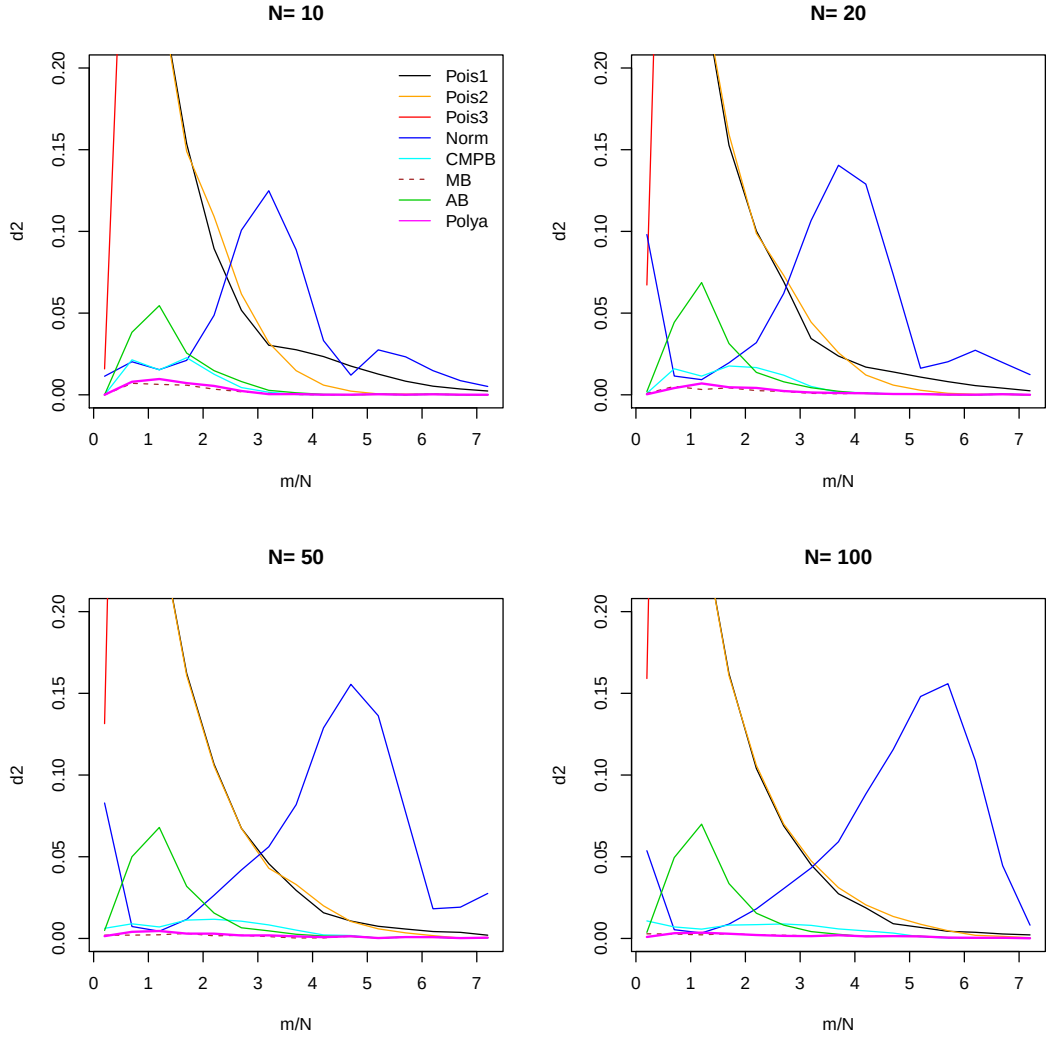


Figure 6.3: Total variation distance $d_2 = \frac{1}{2}\sum |P(K = x) - P^*(K = x)|$ for $N = 10, 20, 50, 100$ based on $p_i = 1/N$

the Pois3 cannot approximate close to the empirical probability distribution. When $N = 400$ and $M = 2000$, the Pois1, Pois2, Pólya, CMPB and Altham's are all very similar to the empirical probability.

Extending Williamson (2012), we have selected some situations to compare the performance of approximations measured using d_2 and d_3. Table 6.2 presents the distance measures d_2 and d_3 ($\times 10^5$) for $p_i = 1/N$ for various values of $M, N \leq 100$. The results indicate that the MB and Pólya distribution outperform others. The Pólya distribution outperforms the MB distribution for

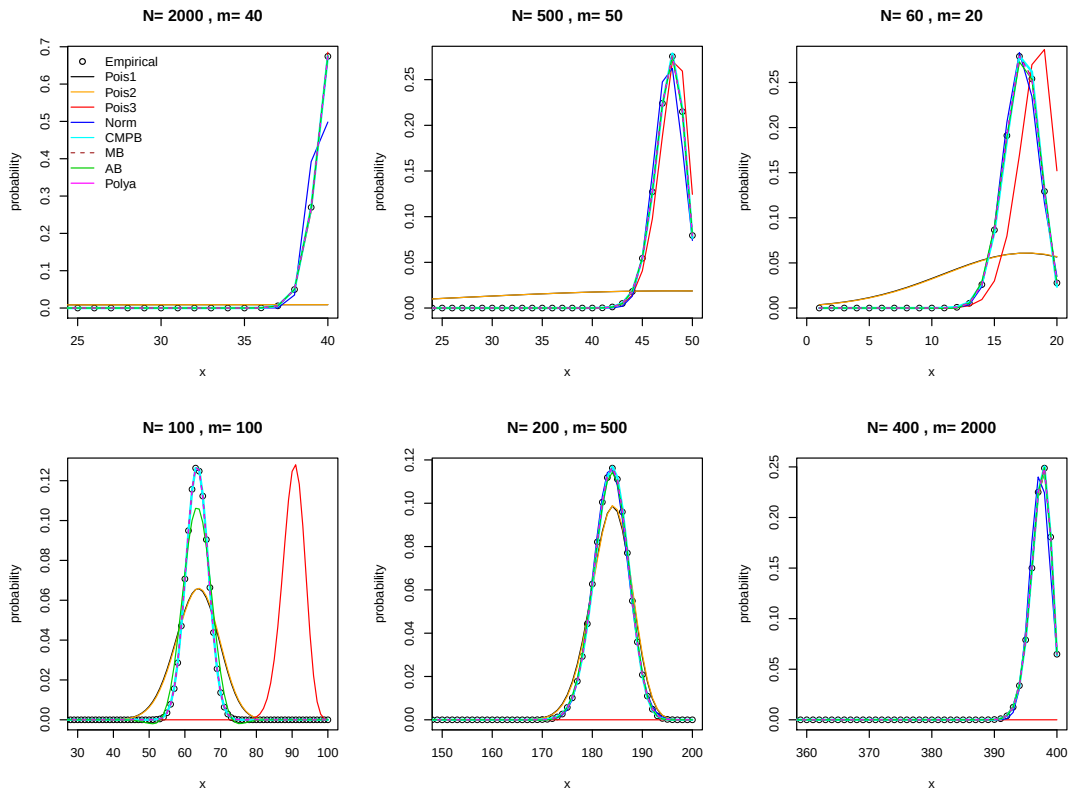


Figure 6.4: Distribution of K based on $p_i = \dfrac{1}{N}$ with various M and N

values of both d_2 and d_3 when $M/N < 1$ while the MB distribution outperforms the Pólya distribution when $M/N > 2.7$. When $1 \leq M/N \leq 2.7$, there is no clear preference between the MB and Pólya distribution.

When considering the Poisson distribution using $\lambda = Ne^{-M/N}$ (Pois1), $\lambda = N(1 - \dfrac{1}{n})^M$ (Pois2) and $\lambda = \text{Var}(K)$ (Pois3), their performance depends on the value of M/N. The performance of the Pois1 and Pois2 improves when M/N is large whereas Poi3 works well when M/N is very small. When M/N is large, Pois3 give $\Pr^*(K = x) \to 0$ for all values of K in the range 1 to n. Therefore, it is not appropriate for approximation when $M/N \to \infty$. For the normal distribution, it shows good approximation when M/N is between 0.8 and 1. For the CMPB, it can perform well when $M/N \geq 4$ and similar to the MB and the Pólya distributions. For the AB distribution, it cannot perform as well as the MB model until $M/N \geq 4$.

Table 6.2: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1-1/N)^M$), Poisson(Var(X)), Normal, CMPB, Altham's (MB and AB) and Pólya based on small N and $M \leq 100$ with $p_i = \frac{1}{N}$.

				d_2							
	N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
6	80	12	0.15	5548	5540	980	1004	61	24	21	8
18	80	20	0.25	5161	5139	2144	401	121	19	61	11
21	100	50	0.50	4868	4854	7182	112	95	26	268	20
24	60	48	0.80	3782	3783	9550	40	86	33	640	25
31	100	100	1.00	3095	3089	9999	35	40	27	982	43
33	50	100	2.00	1233	1248	5000	191	131	33	207	20
35	40	100	2.50	802	802	5000	380	115	14	100	31
38	37	100	2.70	617	687	5000	437	126	25	66	16
48	35	100	2.86	597	565	5000	525	112	14	46	19
51	20	60	3.00	437	575	5000	838	66	7	62	29
53	33	100	3.03	522	535	5000	704	88	10	47	19
54	30	100	3.33	350	446	5000	815	67	9	33	12
55	15	60	4.00	195	144	5000	1121	8	8	16	10
58	15	65	4.33	190	76	5000	765	4	3	3	2
60	12	60	5.00	141	16	5000	160	1	1	1	1
64	14	98	7.00	28	1	5000	98	1	1	1	1

				d_3							
	N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
6	80	12	0.15	3723	3722	962	1165	52	19	16	5
18	80	20	0.25	2583	2582	1254	335	56	15	28	10
21	100	50	0.50	1168	1167	1623	41	31	11	81	12
24	60	48	0.80	976	977	1750	17	32	15	198	11
31	100	100	1.00	625	626	1280	16	15	13	217	12
33	50	100	2.00	438	447	1928	75	53	14	81	9
35	40	100	2.50	390	374	2603	231	61	6	48	16
38	37	100	2.70	329	332	2904	325	65	17	35	14
48	35	100	2.86	351	354	3057	422	68	11	31	17
51	20	60	3.00	437	449	4116	978	56	6	56	28
53	33	100	3.03	262	338	3500	397	77	7	22	17
54	30	100	3.33	313	304	3983	952	60	7	24	9
55	15	60	4.00	195	144	7792	1266	6	7	15	10
58	15	65	4.33	190	76	8403	913	3	2	3	2
60	12	60	5.00	141	16	9364	209	1	1	1	1
64	14	98	7.00	28	1	9902	98	1	1	1	1

Table 6.3 shows the performance for heterogeneous models with unequal p_i for small M, N. It shows that the MB distribution dominates when the p_i is the Zipf1 model especially $1 \leq M/N \leq 4$. For the Pólya distribution, it works well with p_i from the log-normal and broken-stick model when $M/N \leq 2.7$. For the CMPB distribution, it can approximate domination for the Zipf2 model

Table 6.3: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1-1/N)^M$), Poisson(Var(K)), Normal, CMPB, Altham's (MB and AB) and Pólya based on small N and $M \leq 100$ with various unequal p_i.

				d_2					d_3				
Model	N	M	$\frac{M}{N}$	Norm	CMPB	MB	AB	Pólya	Norm	CMPB	MB	AB	Pólya
Zipf1 ($p_i = c/i^{0.5}$)													
6	80	12	0.15	684	45	7	19	14	794	39	6	17	9
21	100	50	0.50	90	69	24	207	13	31	19	9	57	5
31	100	100	1.00	27	23	21	780	29	13	7	8	157	11
33	50	100	2.00	141	92	21	234	22	54	32	11	67	12
38	37	100	2.70	319	104	12	110	25	160	46	5	50	14
54	30	100	3.33	475	96	10	45	23	404	60	6	36	12
55	15	60	4.00	1250	25	5	29	10	1481	22	4	25	7
60	12	60	5.00	901	4	5	11	7	1051	4	5	9	6
Zipf2 ($p_i = c/i^2$)													
6	80	12	0.15	375	203	266	538	349	213	122	166	334	193
21	100	50	0.50	294	102	202	856	338	116	44	80	318	125
31	100	100	1.00	252	75	178	983	319	95	29	65	281	110
33	50	100	2.00	231	95	168	799	259	93	37	62	264	96
38	37	100	2.70	211	104	162	729	224	73	30	51	229	71
54	30	100	3.33	199	119	158	651	199	94	56	71	231	86
55	15	60	4.00	159	182	162	485	152	92	90	83	233	76
60	12	60	5.00	88	210	152	348	108	45	91	80	178	57
Log-Normal													
6	80	12	0.15	566	84	31	32	22	357	76	29	23	21
21	100	50	0.50	40	81	44	292	27	12	24	16	81	13
31	100	100	1.00	33	26	28	822	29	9	8	8	164	8
33	50	100	2.00	45	83	34	521	24	16	22	8	145	6
38	37	100	2.70	56	105	45	448	16	25	37	16	144	9
54	30	100	3.33	166	133	16	451	90	86	72	10	210	40
55	15	60	4.00	367	188	51	140	30	312	160	37	116	19
60	12	60	5.00	352	202	76	118	13	282	172	61	101	11
Broken-stick													
6	80	12	0.15	622	93	23	30	4	432	79	18	21	3
21	100	50	0.50	48	76	42	315	25	19	23	12	81	11
31	100	100	1.00	23	26	26	947	23	8	10	10	187	9
33	50	100	2.00	37	72	38	780	19	13	30	20	221	7
38	37	100	2.70	71	99	36	416	27	36	30	13	145	16
54	30	100	3.33	98	112	24	593	72	59	49	13	229	40
55	15	60	4.00	297	29	169	701	318	198	16	76	322	173
60	12	60	5.00	292	251	63	311	105	267	215	53	164	82
Expo-decay													
6	80	12	0.15	361	113	210	1768	766	329	85	142	1469	621
21	100	50	0.50	375	104	268	3222	4757	374	90	217	1501	909
31	100	100	1.00	329	93	236	3323	*	323	65	170	2190	*
33	50	100	2.00	325	94	230	3201	2428	320	65	166	2117	1071
38	37	100	2.70	319	98	225	3104	1385	315	61	161	2059	961
54	30	100	3.33	318	106	225	3003	1171	310	70	163	2005	873
55	15	60	4.00	345	170	261	2557	666	289	143	219	1374	463
60	12	60	5.00	352	221	278	2263	494	302	188	237	1200	362

Note : * is huge value.

when $M/N < 4$ and dominates others for all M/N when choosing p_i from the expo-decay model. There is a problem about choosing the parameters of the Pólya distribution for the expo-decay models. The parameters are very small and violate the conditions for a valid probability distribution, which leads to

Table 6.4: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1 - 1/N)^M$), Poisson(Var(K)), Normal, CMPB, Altham's (MB and AB) and Pólya based on large N and M (fixed M and N) with $p_i = \frac{1}{N}$.

			d_2							
N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
160	500	3.12	490	483	5000	341	60	13	42	14
154	500	3.25	446	441	5000	382	63	19	45	19
150	500	3.33	404	411	5000	411	64	22	34	19
140	500	3.57	311	325	5000	503	55	18	18	20
350	1000	2.86	615	621	5000	189	46	25	76	28
300	1000	3.33	412	411	5000	294	53	19	38	18
290	1000	3.45	375	379	5000	335	46	14	35	17
280	1000	3.57	333	333	5000	375	38	16	27	18
250	1000	4.00	230	235	5000	492	46	20	23	20
500	1500	3.00	538	538	5000	187	36	30	58	32
500	1700	3.40	377	375	5000	231	47	22	29	20
500	1800	3.60	315	315	5000	296	32	24	24	25
500	1875	3.75	284	288	5000	303	43	23	27	22
1000	3800	3.80	269	272	5000	223	33	23	22	22
1000	3900	3.90	263	264	5000	239	42	31	37	31
1000	4000	4.00	236	239	5000	245	40	32	36	32
1000	4300	4.30	179	182	5000	311	30	19	20	19

			d_3							
N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
200	500	2.50	179	183	1161	41	18	9	30	8
160	500	3.12	158	159	1648	123	20	5	16	6
154	500	3.25	155	155	1761	140	24	7	15	8
150	500	3.33	154	148	1888	142	30	8	18	9
140	500	3.57	134	136	2089	226	22	6	8	7
350	1000	2.86	119	119	1007	38	11	8	18	8
300	1000	3.33	106	108	1321	74	16	7	11	7
290	1000	3.45	110	108	1424	96	12	7	14	7
280	1000	3.57	100	101	1511	106	17	5	10	6
250	1000	4.00	88	87	1969	202	22	9	8	8
500	1500	3.00	96	97	892	39	12	8	12	9
500	1700	3.40	80	81	1047	51	12	9	7	9
500	1800	3.60	74	72	1156	63	9	8	6	8
500	1875	3.75	69	74	1232	67	19	15	17	14
1000	3800	3.80	45	47	890	41	8	5	7	5
1000	3900	3.90	52	50	938	43	11	9	11	9
1000	4000	4.00	52	54	978	54	16	13	12	12
1000	4300	4.30	43	41	1129	65	7	6	4	6

a problem about huge positive and negative values of probability.

Table 6.4 summarizes the performance of approximations for large M and N using $p_i = 1/N$. The MB and Pólya distribution give similar results for fixed

M and N. Their performance is better than the CMPB by a factor of two or three. They often perform better than the Poisson and normal approximations. For example, when $N = 250$ and $M = 1000$, d_2 of the MB and Pólya equal 20 which is around 10 times smaller than the Pois1 and the Pois2 approximations (230 and 235, respectively).

Table 6.5 explores the situation for unequal p_i and fixed M, N. When considering large M and N, the results are different from small M and N in Table 6.3. The MB and Pólya distribution dominate others when using the Zipf and log-normal model; however, the MB distribution is a little bit worse than the Pólya distribution. When $M, N \to \infty$, the CMPB distribution using p_i from the Zipf model gives the same performance as the MB and Pólya distribution (e.g. fixed $N = 1000$ and $M/N \geq 4$).

Additionally, the MB distribution outperform others when using p_i from the broken-stick model. The CMPB distribution is the best approximation when considering the Zipf2 and expo-decay model for p_i. For example with the Zipf2, when $M = 250$ and $N = 1000$, $d_2 = 57$ for the CMPB distribution while the normal, the MB, AB and Pólya distribution give 153, 109, 865 and 175 respectively. When using the Pólya distribution with the expo-decay model, it is found the problem about choosing parameter as well. For the MB and AB distributions, they do not work well with the expo-decay model. Particularly, the parameter of the AB distribution does not follow the constraint for choosing parameter.

Table 6.5: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1 - 1/N)^M$), Poisson(Var(K)), Normal, CMPB, Altham's (MB and AB) and Pólya based on large N and M (fixed M and N) with various unequal p_i.

				d_2					d_3				
Model	N	M	$\frac{M}{N}$	Norm	CMPB	MB	AB	Pólya	Norm	CMPB	MB	AB	Pólya
Zipf1 ($p_i = c/i^{0.5}$)													
	200	500	2.50	26	42	34	259	30	7	11	9	37	8
	160	500	3.12	46	48	33	196	32	11	14	11	30	10
	150	500	3.33	47	45	27	175	24	12	12	8	27	7
	140	500	3.57	62	42	27	163	25	16	13	9	28	7
	350	1000	2.86	32	32	30	235	29	4	6	5	25	5
	300	1000	3.33	33	58	47	185	43	6	10	8	26	8
	280	1000	3.57	43	41	35	179	33	9	9	8	24	7
	250	1000	4.00	50	37	28	157	27	10	9	7	23	6
	500	1500	3.00	38	40	38	231	38	9	8	8	25	8
	500	1700	3.40	36	48	44	184	42	5	7	6	16	5
	500	1800	3.60	40	41	36	188	35	6	9	8	21	8
	500	1875	3.75	46	40	39	165	39	9	6	6	18	7
	1000	3800	3.80	37	43	40	176	39	4	6	5	14	5
	1000	3900	3.90	46	48	47	173	46	5	7	6	14	6
	1000	4000	4.00	50	46	46	175	46	6	6	6	15	6
	1000	4300	4.30	44	40	40	154	40	5	4	4	12	4
Zipf2 ($p_i = c/i^2$)													
	200	500	2.50	179	64	128	888	211	37	14	27	187	45
	160	500	3.12	166	57	115	861	191	45	19	33	173	49
	150	500	3.33	160	57	111	837	184	41	16	28	177	44
	140	500	3.57	182	75	132	822	202	41	18	28	179	44
	350	1000	2.86	146	45	100	900	177	29	11	19	150	33
	300	1000	3.33	147	45	102	881	175	29	13	22	154	33
	280	1000	3.57	137	42	95	875	163	23	9	15	148	27
	250	1000	4.00	153	57	109	865	175	33	14	24	150	37
	500	1500	3.00	140	52	100	916	169	32	15	24	133	36
	500	1700	3.40	132	45	91	914	160	23	12	17	134	28
	500	1800	3.60	121	37	82	907	147	23	8	16	134	26
	500	1875	3.75	130	46	92	905	155	21	9	15	126	24
	1000	3800	3.80	110	34	5000	932	136	18	10	698	109	20
	1000	3900	3.90	116	44	5000	941	141	19	11	693	110	22
	1000	4000	4.00	108	36	5000	929	132	15	6	689	109	18
	1000	4300	4.30	112	40	5000	925	136	17	8	682	110	21
Log-Normal													
	200	500	2.50	34	53	30	402	26	8	15	11	59	9
	160	500	3.12	46	62	36	351	28	18	12	10	66	13
	150	500	3.33	65	70	30	335	29	16	19	13	71	10
	140	500	3.57	75	77	29	357	27	19	20	8	69	8
	350	1000	2.86	49	44	35	350	36	10	9	7	42	7
	300	1000	3.33	38	54	34	350	32	10	10	7	47	6
	280	1000	3.57	66	42	31	291	35	12	10	5	45	7
	250	1000	4.00	81	41	21	271	31	15	8	4	40	6
	500	1500	3.00	41	36	29	342	30	7	6	4	36	5
	500	1700	3.40	52	46	38	302	38	10	9	7	35	7
	500	1800	3.60	51	55	41	338	39	9	13	11	39	10
	500	1875	3.75	41	42	27	319	25	7	7	4	35	4
	1000	3800	3.80	51	41	39	297	41	8	6	5	24	6
	1000	3900	3.90	45	51	40	261	39	7	7	6	24	5
	1000	4000	4.00	44	41	37	291	37	6	6	5	23	5
	1000	4300	4.30	46	38	35	328	38	8	6	5	29	6

				d_2					d_3				
Model	N	M	$\frac{M}{N}$	Norm	CMPB	MB	AB	Pólya	Norm	CMPB	MB	AB	Pólya
Broken-stick													
	200	500	2.50	36	51	32	622	33	8	13	9	96	7
	160	500	3.12	45	48	30	629	44	14	18	12	115	13
	150	500	3.33	29	62	26	705	31	8	13	7	132	9
	140	500	3.57	72	46	32	713	80	26	17	12	149	26
	350	1000	2.86	42	38	28	613	39	9	9	6	76	9
	300	1000	3.33	37	36	26	686	39	8	8	5	95	8
	280	1000	3.57	36	41	23	691	38	10	12	8	101	9
	250	1000	4.00	34	58	35	695	35	11	12	7	102	11
	500	1500	3.00	43	42	38	618	42	9	10	7	62	8
	500	1700	3.40	46	39	38	725	49	10	6	7	77	10
	500	1800	3.60	52	41	38	586	48	12	8	10	63	11
	500	1875	3.75	48	33	28	536	41	9	6	6	62	8
	1000	3800	3.80	49	47	45	572	47	7	7	5	47	7
	1000	3900	3.90	46	39	38	627	45	6	7	7	50	6
	1000	4000	4.00	43	45	40	616	42	5	7	6	49	5
	1000	4300	4.30	39	50	40	536	37	5	5	4	43	5
	500	1500	3.00	389	222	5000	5659	*	389	177	4441	2581	*
	500	1700	3.40	390	252	5000	5609	*	338	207	4787	2921	*
	500	1800	3.60	390	249	5000	5559	*	297	204	4939	3071	*
	500	1875	3.75	391	244	5000	5514	*	265	197	5043	3175	*
	1000	3800	3.80	487	115	5000	5560	*	487	94	5499	3665	*
	1000	3900	3.90	503	134	5000	5607	*	503	115	5477	3645	*
	1000	4000	4.00	511	138	5000	5659	*	511	121	5445	3616	*
	1000	4300	4.30	498	144	5000	5769	*	498	129	5307	3485	*
Expo-decay													
	200	500	2.50	365	147	5000	3926	*	239	123	4569	2361	*
	160	500	3.12	355	153	279	3901	*	245	130	229	2335	*
	150	500	3.33	349	138	263	3890	*	231	116	215	2340	*
	140	500	3.57	358	147	272	3892	*	241	127	225	2335	*
	350	1000	2.86	346	134	5000	4399	*	344	112	4129	2053	*
	300	1000	3.33	352	137	5000	4397	*	351	116	4137	2057	*
	280	1000	3.57	349	137	5000	4393	*	349	116	4134	2053	*
	250	1000	4.00	348	130	5000	4392	*	346	109	4131	2047	*
	500	1500	3.00	353	160	5000	4261	*	187	133	4671	2617	*
	500	1700	3.40	354	140	5000	4286	*	247	111	4758	2713	*
	500	1800	3.60	361	138	5000	4280	*	282	107	4765	2726	*
	500	1875	3.75	344	131	5000	4264	*	293	96	4745	2711	*
	1000	3800	3.80	357	167	5000	4512	*	222	142	4597	2657	*
	1000	3900	3.90	360	172	5000	4529	*	211	144	4630	2691	*
	1000	4000	4.00	366	166	5000	4539	*	188	133	4673	2734	*
	1000	4300	4.30	361	172	5000	4565	*	198	141	4717	2781	*

Note : * is huge value.

Table 6.6 shows the results of approximations for very small and very large value of M/N, it is indicated that the MB and Pólya distribution still dominate others and give similar performance. When addressing very large values of M/N for equal p_i, in Table 6.6, the distance measuring give the same results for all approximations (e.g. $M/N = 20$, $d_2 = d_3 = 0$) except the Pois3 distribution which can work when $M/N \to 0$. On the other hand, for very small

Table 6.6: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1-1/N)^M$), Poisson(Var(K)), Normal, CMPB, Altham's (MB and AB) and Pólya based on very small and very large $\frac{M}{N}$ with $p_i = \frac{1}{N}$.

			d_2							
N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
50	1500	30.00	0	0			0	0	0	246663
50	1000	20.00	0	0	5000	0	0	0	0	0
50	500	10.00	3	0	5000	20	0	0	0	0
50	400	8.00	12	1	5000	154	1	1	1	1
50	300	6.00	59	11	5000	412	6	5	5	5
400	50	0.12	6140	6135	1250	505	73	20	18	15
1000	50	0.05	6600	6599	363	897	24	10	15	12
2000	50	0.02	6483	6483	159	1476	14	8	7	8
5000	50	0.01	6121	6121	40	1349	3	2	2	2
1000	100	0.10	6418	6415	1216	467	43	13	15	12
2000	100	0.05	6826	6825	454	637	33	10	8	9
5000	100	0.02	6825	6825	146	1048	9	4	5	4
10000	100	0.01	6542	6542	60	1563	7	5	5	5

			d_3							
N	M	$\frac{M}{N}$	Pois1	Pois2	Pois3	Norm	CMPB	MB	AB	Pólya
50	1500	30.00	0	0			0	0	0	117762
50	1000	20.00	0	0	10000	0	0	0	0	0
50	500	10.00	3	0	9980	20	0	0	0	0
50	400	8.00	12	1	9846	154	1	1	1	1
50	300	6.00	59	11	8894	526	5	5	5	5
400	50	0.12	2217	2217	626	256	39	9	7	6
1000	50	0.05	3581	3581	359	881	19	7	8	7
2000	50	0.02	5310	5310	153	1701	11	8	7	8
5000	50	0.01	7763	7763	40	1480	3	2	2	2
1000	100	0.10	1780	1780	448	186	19	5	6	4
2000	100	0.05	2569	2569	235	419	15	5	5	5
5000	100	0.02	3674	3674	144	1235	6	3	3	3
10000	100	0.01	6050	6050	55	1825	5	5	5	5

M/N, it is clear that the Pois1 and Pois2 are not appropriate to approximate the exact probability. They give the same results (e.g. when $N = 10000$ and $M = 100$, $d_2 = 6542$) while the MB and Pólya distribution are very accurate (e.g. when $N = 10000$ and $M = 100$, $d_2 = 5$).

For unequal p_i in Table 6.7, the MB and Pólya distributions outperform others and their performance is quite similar. For example with the Zipf model, when $N = 2000$ and $M = 50$, the MB and Pólya distribution give the best value with

Table 6.7: Distance measures ($\times 10^5$), $d_2 = \frac{1}{2}\sum |p(x) - p^*(x)|$ and $d_3 = \max |p(x) - p^*(x)|$, for Poisson($Ne^{-M/N}$)), Poisson($N(1 - 1/N)^M$), Poisson(Var(K)), Normal, CMPB, Altham's (MB and AB) and Pólya based on very small and very large $\frac{M}{N}$ with various unequal p_i.

				d_2					d_3				
Model	N	M	$\frac{M}{N}$	Norm	CMPB	MB	AB	Pólya	Norm	CMPB	MB	AB	Pólya
Zipf1 ($p_i = c/i^{0.5}$)													
	50	1500	30.00	79	1	1	1	1	79	1	1	1	1
	50	1000	20.00	238	3	3	3	3	335	3	3	3	3
	50	500	10.00	861	43	8	11	7	593	40	6	8	4
	50	400	8.00	540	64	13	14	10	333	31	9	6	8
	50	300	6.00	371	80	19	51	12	160	32	8	26	8
	400	50	0.12	41	72	57	72	54	18	21	18	20	17
	1000	50	0.05	106	70	62	59	61	30	20	18	18	18
	2000	50	0.02	141	71	68	68	68	36	23	23	23	23
	5000	50	0.01	203	61	67	67	67	51	19	21	21	21
	1000	100	0.10	36	56	48	59	46	8	13	13	19	12
	2000	100	0.05	48	52	45	47	44	14	11	10	11	10
	5000	100	0.02	89	36	32	33	32	19	8	8	8	8
	10000	100	0.01	105	43	42	42	42	20	14	14	14	14
Zipf2 ($p_i = c/i^2$)													
	50	1500	30.00	52	102	59	221	40	20	29	21	60	16
	50	1000	20.00	36	110	81	336	62	14	31	23	95	18
	50	500	10.00	106	102	104	503	105	34	31	33	128	32
	50	400	8.00	137	109	120	554	131	42	34	38	145	40
	50	300	6.00	148	99	125	609	147	50	31	39	163	46
	400	50	0.12	287	96	196	799	318	119	39	74	294	122
	1000	50	0.05	301	102	203	780	329	125	40	80	286	128
	2000	50	0.02	295	97	198	784	323	123	39	79	287	126
	5000	50	0.01	303	104	206	790	330	124	39	80	291	126
	1000	100	0.10	273	87	187	868	315	87	32	62	260	93
	2000	100	0.05	272	91	186	856	312	94	33	63	247	102
	5000	100	0.02	257	71	171	836	297	81	24	53	251	91
	10000	100	0.01	273	90	187	843	313	90	28	58	255	98
Log-Normal													
	50	1500	30.00	1310	9	5	5	4	1444	8	5	5	3
	50	1000	20.00	812	34	49	193	117	946	31	43	170	103
	50	500	10.00	753	93	10	49	15	628	82	8	21	12
	50	400	8.00	395	124	27	74	15	228	61	19	38	8
	50	300	6.00	251	135	36	146	10	109	54	14	57	3
	400	50	0.12	350	47	22	22	21	128	19	10	8	9
	1000	50	0.05	555	39	22	23	21	298	29	18	18	18
	2000	50	0.02	909	12	12	12	12	561	6	7	7	7
	5000	50	0.01	1504	15	16	16	16	1704	14	16	16	16
	1000	100	0.10	291	32	19	20	18	73	13	8	8	7
	2000	100	0.05	431	17	17	17	17	129	7	7	7	7
	5000	100	0.02	662	21	22	22	22	397	15	15	15	15
	10000	100	0.01	1012	9	11	11	11	863	9	8	8	8
Broken-stick													
	50	1500	30.00	381	190	241	989	602	264	112	158	423	360
	50	1000	20.00	415	131	108	374	286	318	118	59	311	137
	50	500	10.00	183	182	24	684	189	98	116	13	344	85
	50	400	8.00	309	151	14	311	121	200	104	8	176	59
	50	300	6.00	170	43	79	963	255	86	22	42	353	109
	400	50	0.12	369	53	19	35	19	135	25	6	13	8
	1000	50	0.05	606	47	18	17	16	438	21	10	10	9
	2000	50	0.02	932	10	15	16	16	951	9	12	14	13
	5000	50	0.01	1568	6	4	5	4	1806	6	4	4	4
	1000	100	0.10	314	49	26	28	26	95	17	10	12	10
	2000	100	0.05	488	30	14	11	14	197	12	5	4	5
	5000	100	0.02	768	20	7	6	7	510	12	5	4	5
	10000	100	0.01	1061	10	11	11	11	1245	7	9	9	9

				d_2					d_3				
Model	N	M	$\frac{M}{N}$	Norm	CMPB	MB	AB	Pólya	Norm	CMPB	MB	AB	Pólya
Expo-decay													
	50	1500	30.00	354	173	258	3846	2192	178	140	203	2458	966
	50	1000	20.00	365	163	262	4153	3283	363	138	216	1935	880
	50	500	10.00	361	160	269	3709	3249	238	136	222	2232	885
	50	400	8.00	329	136	250	3902	4093	327	115	214	1786	854
	50	300	6.00	355	123	241	3694	2980	351	92	188	2180	977
	400	50	0.12	365	88	251	3228	4748	364	74	205	1490	899
	1000	50	0.05	364	83	247	3230	4748	362	70	204	1489	897
	2000	50	0.02	363	91	249	3220	4746	360	78	205	1487	896
	5000	50	0.01	356	86	243	3220	4739	355	73	200	1482	891
	1000	100	0.10	327	88	231	3326	*	323	64	168	2190	*
	5000	100	0.02	325	89	225	3326	*	316	67	161	2183	*
	10000	100	0.01	329	96	235	3318	*	318	71	171	2185	*

Note : * is huge value.

$d_2 = 68$ which is about half the value from the normal distribution, $d_2 = 141$. Additionally, the MB and Pólya distribution seem to perform very well with the log-normal and broken-stick model and very small M/N. For the CMPB distribution, it still works very well with Zipf2 the same as Table 6.3 and Table 6.5, in particular, when $M/N < 0.12$ and $6 \leq M/N \leq 10$. It dominates for both very small and very large M/N with the expo-decay model. It gives the best results when compared with the MB distribution by a factor of about two and the normal distribution around four (e.g. when $N = 1000$ and $M = 50$). The AB and Pólya distributions are not appropriate to approximate when using p_i from the expo-decay model for this situation either.

6.7 Conclusion

Considering the classical problem, the performance of the Poisson and the normal approximations in this study agree with the results of Williamson (2012). Poisson($Ne^{-M/N}$) and Poisson($N(1-1/N)^M$) work well for large M/N while Poisson(Var(K)) is appropriate for $M/N \to 0$. For the normal distribution, it can approximate well for small M/N not greater than 2. The value of M/N is the key factor which affects to the performance of approximations.

For the new approximations, they can approximate well and are suitable for different models of p_i. Most of them outperform the Poisson and the normal approximations. The Pólya and Altham's multiplicative distribution give a good approximation when selecting p_i from the Zipf1, log-normal and broken-stick model. They can approximate similarly, in particular, using large M, N and p_i from the Zipf1 and broken-stick model.

For the COM-Poisson-Binomial distribution, it can work well with the Zipf2 and expo-decay model for both small and large M/N. For the Altham's additive distribution, it cannot approximate as well as the Altham's multiplicative and the Pólya distribution because the parameters do not follow the condition in equation (6.20).

There is a potential computational problem for the COM-Poisson-binomial and the Altham's multiplicative distribution, but it can be resolved using logarithmic transformation. The Pólya distribution is quite easy for selecting the parameters because there is the formulae for estimating parameters. However, the parameter θ might be smaller than the lower limit given by equation (6.23) in some situations. Particularly, for the expo-decay model, Altham's multiplicative and additive and Pólya distribution are not appropriate for approximation to the occupancy distribution.

Chapter 7

Estimating the number of unseen species using approximations to the distribution of seen species

7.1 Introduction

In this Chapter, species richness is estimated using an approximation to the distribution of K, based on the work of the previous chapter Hidaka (2014), developed a method of estimating species richness based on an approximation to the distribution of K, and used this method to estimate the number of distinct words in the novel "Alice's Adventures in Wonderland". The distribution of occupied urns is considered as the exact distribution of the number of distinct words. As a result of intractable computation for the exact distribution, the asymptotic distribution, which is shown to be the Poisson binomial (PB) distribution, is proposed for approximation. For inference, a maximum pseudo-likelihood estimation (MPLE) method is developed to estimate the unknown population size. Data are separated into many subsets which are used

to construct the pseudo-likelihood function.

Hidaka's parametric method, including the pseudo-likelihood function and evaluation of the distribution of seen species, is presented in Section 7.1. Altham's multiplicative binomial (MB) distribution is considered as an alternative to approximate the distribution of K in Section 7.2. Section 7.3 presents a least squares method for estimating the number of species.Subsets of the data can be constructed using many schemes, some of which are described in Section 7.4. Measuring the accuracy of the maximum likelihood approach is presented in Section 7.5. In a simulation study, the performance of maximum likelihood (MLE), maximum pseudo-likelihood (MPLE), least squares (LS) and some nonparametric estimators are compared in Section 7.6. Finally, the results are summarised in Section 7.7.

7.2 Hidaka's parametric method

Let N denote the total number of species in the population and K denote the number of distinct species in a sample of M individuals. Hidaka (2014) proposed the MPLE approach to estimating N in his study. The pseudo-likelihood was developed originally by Besag (1975). In Hidaka's method, the pseudo likelihood function is constructed using m data sets $D_1, \ldots, D_m$ which are generated from the original data. Each data set D_r contains M_r individuals from K_r distinct species. The product of the probability function of the random variables K_i used to construct the pseudo-likelihood function

$$L(N|\boldsymbol{\theta}) = \prod_{r=1}^{m} \mathrm{P}(K_r|M_r, N, \boldsymbol{\theta}). \tag{7.1}$$

where $\boldsymbol{\theta}$ is a vector of parameters describing the distribution of relative abundances, m is the number of data sets and $m \leq M$. This is a pseudo-likelihood

rather than a true likelihood because the data sets $D_1, \ldots, D_m$ are typically not independent even though they are treated as if they were independent.

In this Chapter, we focus on the case where all species have the same relative abundance (the classical occupancy model) because in this case the results can be compared with maximum likelihood estimation.

Then, the pseudo log-likelihood function is given by

$$\ell(N, \boldsymbol{\theta}) = \sum_{r=1}^{m} \log \mathrm{P}(K_r | M_r, N, \boldsymbol{\theta}), \tag{7.2}$$

Finally, the unknown parameters N and θ can be estimated by maximizing equation (7.2).

An additional complication in evaluating the pseudo-likelihood is that the exact probabilities $\mathrm{P}(K_r | M_r, N, \boldsymbol{\theta})$ are generally intractable and are replaced instead by an approximation. Hidaka (2014) uses a Poisson binomial approximation for the exact distribution, but there are other possibilities, as discussed later.

7.2.1 Evaluation of $\mathrm{P}(K_r | M_r, N, \boldsymbol{\theta})$

Poisson-binomial approximation

This is the approximation used by Hidaka (2014). Assume that there are m independent Bernoulli trials with probability of successes $p_1, p_2, \ldots, p_m$ and $0 \leq p_r \leq 1$ $(r = 1, \ldots, m)$. For independent but not identical Bernoulli trials, where p_r for each trial is non equivalent, the Poisson-binomial distribution is the distribution of the number of successes in m trials (Wang, 1993). When the sample size M tends to infinity, the exact probability distribution of the

number of observed species, K, tends to the Poisson-binomial distribution which is given by (Hidaka, 2014)

$$Q(K|M,N) = \sum_{s\subseteq\bar{N}:|s|=K} \prod_{r\in s} q_{M,r} \prod_{j\in\bar{N}\setminus s} (1-q_{M,j}) \tag{7.3}$$

where $s_r = \{1,\ldots,N\} \setminus \{r\}$, the set of all species apart from species r and $q_{M,r} = 1-(1-p_r)^M$. The approximate pseudo-likelihood function based on the Poisson-binomial distribution is

$$L_{PB}(N|\boldsymbol{\theta}) = \prod_{r=1}^{m} Q(K_r|M_r,N,\boldsymbol{\theta}), \tag{7.4}$$

and the corresponding approximate log pseudo-likelihood function is

$$\ell_{PB}(N,\boldsymbol{\theta}) = \sum_{r=1}^{m} \log\left\{Q(K_r|M_r,N,\boldsymbol{\theta})\right\}. \tag{7.5}$$

The parameters θ and N are estimated by maximizing the pseudo log-likelihood function.

Direct computation of the Poisson-binomial probabilities form equation (7.3) is not simple. However, Hong (2014) has developed a package in R called `poibin` which provides the probability function for the Poisson-binomial distribution.

Altham's multiplicative binomial approximation

As discussed in the previous Chapter, Altham's multiplicative binomial distribution (Altham, 1978) is a two-parameter generalization of the binomial distribution with probability mass function

$$\Pr{}^{*}_{MB}(K=x|M,N) = \frac{\binom{n}{x} p^x(1-p)^{n-x}\ \phi^{x(n-x)}}{\sum_{x=0}^{n}\binom{n}{x} p^x(1-p)^{n-x}\ \phi^{x(n-x)}}, \qquad x=0,1,\ldots,n \tag{7.6}$$

where $n = \min(M, N), p \in (0, 1)$ and $\phi > 0$.

To use this distribution as an approximation to the distribution of K, we choose the parameters of the Altham's multiplicative-binomial distribution, p and ϕ, so that the mean and variance of the Altham's multiplicative-binomial distribution equal the exact mean and variance of K. Then, the approximate pseudo-likelihood function is given by

$$L_{MB}(N|\boldsymbol{\theta}) = \prod_{r=1}^{m} \Pr_{MB}(K_r|M_r, N, \boldsymbol{\theta}). \tag{7.7}$$

and the pseudo log-likelihood is

$$\ell_{MB}(N|\boldsymbol{\theta}) = \sum_{r=1}^{m} \log \left\{\Pr_{MB}(K_r|M_r, N, \boldsymbol{\theta})\right\}. \tag{7.8}$$

where K_r and M_r are the number of distinct species and the number of individuals in data set D_r.

7.2.2 Construction of the data sets $D_1, \ldots, D_m$

There are many possible schemes for constructing the data sets D_r $(r = 1, 2, \ldots, m)$ including overlapping and non-overlapping. For non-overlapping, each data set is separate and independent as follows:

- Non-overlapping 1 (Non1): The data are separated into m data sets with equal number of individuals following a sequence of sampling, for example with $m = 10$ for $M = 100$ and $m = 20$ for $M = 1000$.
- Non-overlapping 2 (Non2):The data are separated into 4 data sets with sizes in the proportion 1:2:3:4. For example with sample size $M = 100$, the data sets consist of the sample $S_{1:10}, S_{11:30}, S_{31:60}$ and $S_{61:100}$.
- Non-overlapping 3 (Non3): The data are separated into 5 data sets with

sizes in the proportion 1:1:2:3:3. For example with sample $M = 100$, the data sets consist the sample $S_{1:10}, S_{11:20}, S_{21:40}, S_{41:70}$ and $S_{71:100}$.

- Overlapping : Hidaka (2014) created the data sets D_r that overlap by selecting the first $[M/m] \times r$ individuals in sample, where $m \leq M$ and m is the number of data sets. For example, $m = 10$ and $M = 100$, there are 10 data sets, $D_1, \ldots, D_{10}$ constructed using Hidaka (2014) scheme. The data sets are constructed by adding 10 new samples in sequence. Then, the data sets consists the sample $S_{1:10}, S_{1:20}, \ldots,$ and $S_{1:100}$.

7.3 Least squares estimator (LS)

Least squares (LS) estimation is used to estimate unknown parameters by minimizing a sum of squares between observation and expectation. The LS is a common method for fitting the models to data. It is usually a simpler method computationally than the MLE method. While the MLE method requires the probability function for the likelihood function, the LS method requires only the mean for estimating the unknown parameter θ by minimizing the residual sum of squares (Morgan, 2008, p.130)

$$RSS = \sum_{r=1}^{m} \left(K_r - \mathrm{E}(K_r|M, N, \theta)\right)^2,$$

where K_r is the number of distinct species with the expected value $\mathrm{E}(K_r|M, N, \theta)$ and M_r is the number of individuals in data set D_r ($r = 1, \ldots, m$). To simplify notation, we replace $\mathrm{E}(K_r|M, N, \theta)$ by $\mathrm{E}(K_r)$, so that the LS criterion is given by

$$\min \left[\sum_{r=1}^{m} \left(K_r - \mathrm{E}(K_r)\right)^2\right] \tag{7.9}$$

where $\mathrm{E}(K_r) = \sum_{r=1}^{m} \left(1 - (1 - p_r)^{M_r}\right)$, p_r is the relative abundance describing the probability of species r being collected (e.g. p_r follows the Zipf distribution

with parameters α and N, $p_r \propto r^{-\alpha}$).

When the data sets D_r are an increasing sequence of subsets, an alternative approach is to consider the number of *new* distinct species observed for each data set D_r which is denoted as $K_r^{'}$ with the expectation $\mathrm{E}(K_r^{'})$, then the minimum residual sum of squares is given by

$$\min \left[\sum_{r=1}^{m} \left(K_r^{'} - \mathrm{E}(K_r^{'}) \right)^2 \right] \tag{7.10}$$

where $K_r^{'} = K_r - K_{r-1}$ and

$\mathrm{E}(K_r^{'}) = \sum_{r=1}^{m} \left(1 - (1-p_r)^{M_r}\right) - \sum_{r=1}^{m} \left(1 - (1-p_r)^{M_r - M_{r-1}}\right)$.

7.4 Measuring the accuracy of the MLE

Although the MLE method is more a complicated approach to estimate the unknown parameter, it is a preferred method that gives an efficient estimator. The performance of the estimator of MLE depends on the Fisher information which measures the amount information of observed data used to estimate the unknown parameter θ.

7.4.1 Likelihood function of species sampling

Consider an infinite population consisting of a finite number of species, N, and where a randomly chosen individual is equally likely to belong to any of the N species. The likelihood function for estimating the number of different species from a random sample of M individuals is given by

$$\left[\frac{M!}{\prod_{i=1}^{K} c_i! \prod_{j=1}^{M} f_j!} \right] \left[\frac{N!}{(N-K)!} \left(\frac{1}{N} \right)^M \right], \tag{7.11}$$

where K is the number of distinct species in the sample, c_i is the number of individuals in species i and f_j is the number of species appearing j times (Lewontin and Prout, 1956). The likelihood function is therefore

$$\begin{aligned} L(N) &\propto \frac{N!}{(N-K)!}\left(\frac{1}{N}\right)^M \\ &= \prod_{i=0}^{K-1}(N-i) \times \left(\frac{1}{N}\right)^M. \end{aligned}$$

and, ignoring a constant term that does not depend on N, the log-likelihood function is

$$\ell(N) = -M\log(N) + \sum_{i=0}^{K-1}\log(N-i) \tag{7.12}$$

Differentiating with respect to N gives

$$\frac{\partial \ell}{\partial N} = \frac{-M}{N} + \sum_{i=0}^{K-1}\frac{1}{N-i}$$

and therefore the maximum likelihood estimator of N satisfies

$$\frac{M}{\widehat{N}} = \sum_{i=0}^{K-1}\frac{1}{\widehat{N}-i},$$

(Lewontin and Prout, 1956).

Letting $j = \widehat{N} - i$, this expression can equivalently be written as

$$\frac{M}{\widehat{N}} = \sum_{j=\widehat{N}-K+1}^{\widehat{N}}\frac{1}{j}. \tag{7.13}$$

7.4.2 Fisher information

The Fisher information is given by

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial N^2}\right] = \mathrm{E}\left[-\frac{M}{N^2} + \sum_{j=N-K+1}^{N}\frac{1}{j^2}\right]. \tag{7.14}$$

which can be simplified using the approximation $\sum_{j=N-K+1}^{N} \frac{1}{j^2} \cong \frac{K}{N(N-K+1)}$. Assume that U is the number of unseen species, $U = N - K$. If $M, N \to \infty$ and $Ne^{-M/N} \to \lambda$, the distribution of the number of unseen species converges to the Poisson (Johnson and Kotz, 1977), so that

$$P(U = u) \simeq \frac{e^{-\lambda}\lambda^u}{u!} \tag{7.15}$$

Based on this approximation

$$E(U) = \lambda = Ne^{-M/N} \tag{7.16}$$

and

$$\mathrm{E}(K) = N - E(u) = N(1 - e^{-M/N}). \tag{7.17}$$

Assuming that at least one species is collected, the probability distribution of U is given by

$$f(u) = \frac{P(u)}{1 - P(u = N)}$$

Then, the Fisher information becomes

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial N^2}\right] \cong -\frac{M}{N^2} + \mathrm{E}\left[\frac{N-u}{N(u+1)}\right]$$

$$= -\frac{M}{N^2} + \sum_{u=0}^{N-1}\left[\frac{N-u}{N(u+1)}\right] \times f(u)$$

$$= -\frac{M}{N^2} + \sum_{u=0}^{N-1}\left[\frac{N-u}{N(u+1)}\right] \times \frac{e^{-\lambda}\lambda^u}{u!} \times \left[\frac{1}{1 - P(u = N)}\right]$$

Letting $x = u + 1$, then

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial N^2}\right] = -\frac{M}{N^2} + \sum_{x=1}^{N}\left[\frac{N-x+1}{Nx}\right] \times \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \times \left[\frac{1}{1-P(x-1=N)}\right]$$

$$= -\frac{M}{N^2} + \frac{1}{\lambda N}\sum_{x=1}^{N}(N-x+1) \times \frac{e^{-\lambda}\lambda^{x}}{(x)!} \times \left[\frac{1}{1-P(x=N+1)}\right]$$

$$= -\frac{M}{N^2} + \frac{1}{\lambda N}E(N-x+1)$$

From equation (7.17), when letting $K = N - x + 1$, then

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial N^2}\right] = -\frac{M}{N^2} + \frac{1}{\lambda N}(N - Ne^{-M/N})$$

$$= -\frac{M}{N^2} + \frac{1-e^{-M/N}}{Ne^{-M/N}}$$

$$= \frac{1}{N}\left[e^{M/N} - \left(1 + \frac{M}{N}\right)\right]$$

which is the information for all observed data depending on the unknown parameter N.

When considering the observed data from data set D_r, the pseudo-likelihood function is used to approximate the exact one. The product of the probability functions for all K_r from data set D_r is used to construct the pseudo-likelihood function. There are many possible schemes both non-overlapping and overlapping for generating the data sets D_r. When comparing the information for all observed data and the data sets D_r, the split data contain less information.

To illustrate this, suppose that the observed data is divided into two non-overlapping sets with the same number of individuals, $M_r = M/2$. Let $y = e^{M/2N}$. Then, since $y > 1$

$$\begin{aligned}
(y-1)^2 &> 0 \\
y^2 - 2y + 1 &> 0 \\
2(y-1) &< y^2 - 1 \\
2(e^{M/2N} - 1) &< e^{M/N} - 1 \\
\frac{2}{N}\left[e^{M/2N} - \left(1 + \frac{M}{2N}\right)\right] &< \frac{1}{N}\left[e^{M/N} - \left(1 + \frac{M}{N}\right)\right] \\
\sum_{i=1}^{2} I_{M_r} &< I_M
\end{aligned}$$

Therefore, for the homogeneous abundance model, the likelihood based on the full data set gives more information for estimating the number of species, N, than splitting the data into two non-overlapping subsets.

It is concluded that, the performance of the MLE estimator based on the full likelihood function is better than the pseudo-likelihood function. The exact probability function of K in equation (6.1) is intractable to compute in general, but the likelihood function for the homogeneous model is not difficult to construct as shown in Section 7.5. The pseudo-likelihood function is an alternative way. Although the homogeneous case is not used in practice, we investigate it for comparing the performance of the MLE and MPLE methods. For heterogeneous models, the likelihood cannot be computed and only the MPLE method is available. However, heterogeneous models are not explored in the thesis.

7.5 Simulation study

In this section, the performance of different approaches to estimate N is explored. The simulated data for small and large N are generated using $N = 100$ and $N = 1000$. The accuracy of the estimators is measured using the root mean square error (RMSE) and bias. The conditions in the simulation study are as follows:

1. M individuals are collected with replacement randomly with $p_r = 1/N$ from the population for $N = 100$ with $M = 100, 200$, $N = 250$ with $M = 250, 500$ and $N = 500$ with $M = 500, 1000$. All situations are repeated for 100 simulations.

2. From the sample size M, the data sets $D_1, \ldots, D_n$ are generated using nonoverlapping and overlapping schemes as described in Section 7.4. This gives n pairs of (K_r, M_r), the number of distinct species and the number of individuals for data set D_r, which are used for the MLE and the MLPE approach.

3. From data set $D_1, \ldots, D_n$, we consider another pattern for (K_r, M_r). The number of new distinct species for each data set are counted, $K_{new1}, \ldots, K_{new_n}$, and are resampled themselves 1000 times. This gives 1000 new values of K_{new_i} and used for the LS approach.

4. Nonparametric estimators are used to estimate N including Chao1, iChao1, Good-Turing (GT), Horvitz-Thompson (HT) estimators, as shown in Chapter 2. Parametric estimators including the MLE and MPLE are used based on the Poisson-Binomial (MLE_{pb} and MPLE_{pb}) and Altham's multiplicative (MLE_{al}) distribution. The least squares estimator (LS) is another parametric approach used for estimating N.

5. The performance of all estimators are compared using the RMSE= $\sqrt{E(\widehat{N} - N)^2}$ and the bias =$\text{E}(\widehat{N}) - N$.

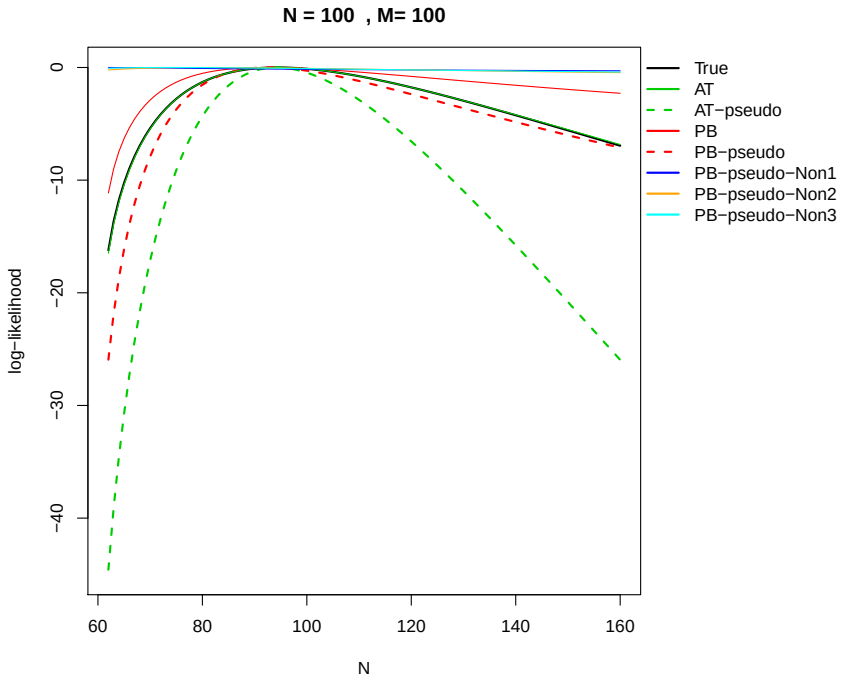
We have generated the data to represent the profile likelihood of overlapping and non-overlapping compared with the exact one. Data sets D_r ($r = 1, \ldots, 100$) are generated using $N = 100, M = 100$ to represent a small sample and $N = 1000, M = 1000$ to represent large sample, based on the homogeneous model.

Figure 7.1 shows the log-likelihood of various probability distributions for $N = 100$ and $M = 100$. The results indicate that the full log-likelihood using the Altham's multiplicative-binomial approximation (AT) provides the results similar to the true log-likelihood. For PB, the full log-likelihood is a little worse when compared with the true likelihood. When using the pseudo log-likelihood, the AT-pseudo and PB-pseudo have less accuracy than the AT with a narrow confidence interval compared to the true log-likelihood. When comparing log-likelihood using the Poisson-binomial approximation with overlapping (PB-Hidaka) and non-overlapping (PB-Non1, PB-Non2 and PB-Non3) scheme, the nonoverlapping schemes give a flat likelihood function.

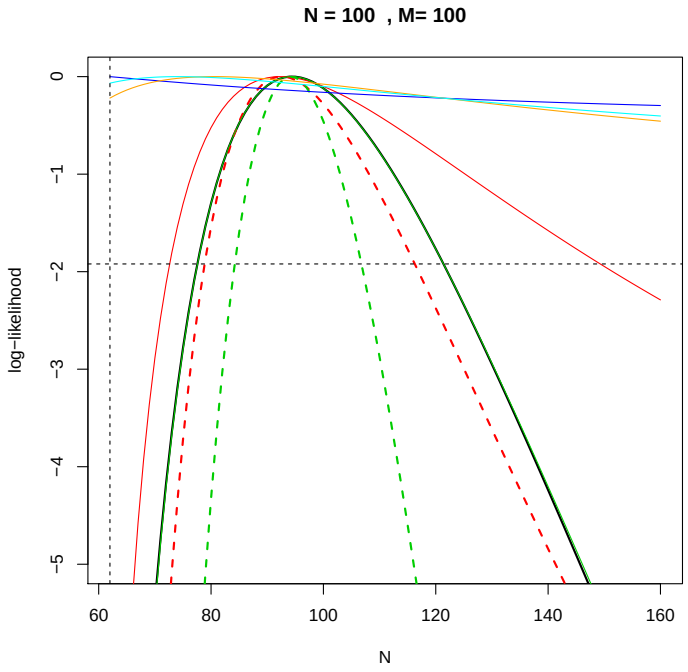
When considering large N and M, the log-likelihood curves show the similar results to small N and M (Figure 7.2). The Altham's multiplicative-binomial distribution has the log-likelihood very close to the true likelihood and outperforms other approaches. For the Poisson-binomial approximation, the log-likelihood curve is not smooth for some values of N for both the full and pseudo log-likelihoods.

We found the local optimization problem in some simulations. The SANN method is considered for handling the local optimization, but there are a few simulations that do not converge.

Table 7.1 shows the number of times that convergence was achieved in opti-

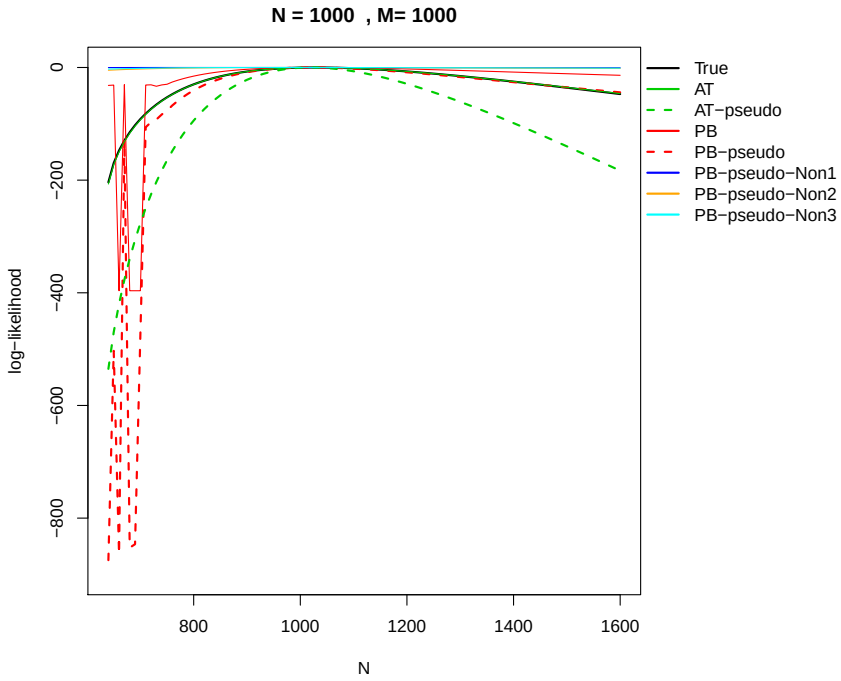


(a) Plot of log-likelihood

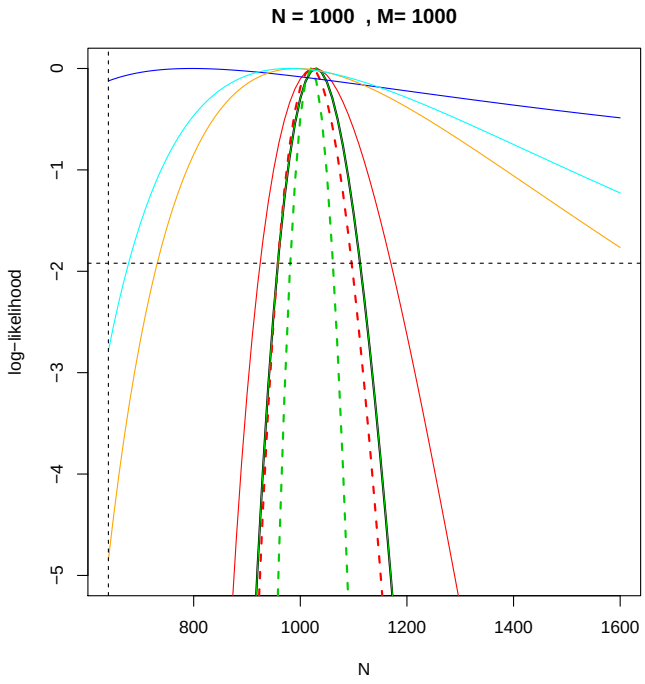


(b) Close ups of the Figure 7.1(a)

Figure 7.1: Plot of log-likelihood for $N = 100, M = 100$ using the Exact, Altham's, PB, PB with overlapping (PB-Hidaka) and PB with nonoverlapping data (PB-Non1,PB-Non2 and PB-Non3) distribution based on abundance data following the homogeneous model.



(a) Plot of log-likelihood



(b) Close ups of the Figure 7.2(a)

Figure 7.2: Plot of log-likelihood for $N = 1000, M = 1000$ using the Exact, Altham's, PB, PB with overlapping (PB-Hidaka) and PB with nonoverlapping data (PB-Non1,PB-Non2 and PB-Non3) distribution based on abundance data following the homogeneous model.

Table 7.1: Number of times that convergence was achieved of optimization using various estimators based on the abundance data following the homogeneous model with repeated 100 times.

N	M	a	MLE_{pb}	$MPLE_{pb}$	MLE_{al}	LS
100	100	0.0	100	100	100	99
100	150	0.0	100	100	100	100
100	200	0.0	100	100	100	100
250	250	0.0	100	100	100	100
250	500	0.0	100	100	100	100
500	500	0.0	100	100	100	99
500	1000	0.0	100	99	100	100

mization for 100 simulations. The MLE approach with the MPE_{pb} and MLE_{al} estimator achieved convergence for all situations. There is one case of non-convergence for the $MPLE_{pb}$ ($N = 500, M = 1000$) and two cases for the LS approach ($N = 100, M = 100$ and $N = 500, M = 500$).

The performance of various estimators are compared in Table 7.2. The results indicate that the GT estimator performs well for $N = M$ with smallest RMSE. The performance of the MLE_{pb} and MLE_{al} are similar and perform well for $M/N = 2$ (i.e. $N = 100, M = 200$, $N = 250, M = 500$ and $N = 500, M = 1000$). Both estimators are slightly better than the $MLPE_{pb}$. However, The $MLPE_{pb}$ does not work well for large N (i.e. $N = 500$). The Chao1 estimator cannot outperform the MLE_{pb} and MLE_{al} estimator. The iChao1 and LS estimator approximate poorly when compared with other estimators especially when $N = M$. For example with $N = 500$ and $M = 1000$, the MLE_{pb} and MLE_{al} estimators perform the best and yield similar RMSE as 9.10, While Chao1 estimator has RMSE as 17.80.

When looking at the bias, the HT, MLE_{pb} and MLE_{al} estimators give a negative bias in some situations as shown in Table 7.2. The bias of the LS estimator is large for all situation (Figure 7.3-7.8).

Table 7.2: BIAS and RMSE of $\widehat{N}$ using the Chao1, iChao1, Good-Turing(GT), Horviz-Tompson(HT), MLE with the PB and Altham distribution (MLE_{pb} and MLE_{al}, MPLE with the PB (MLPE_{pb}) and LS estimator with 100 simulations for $N = 100, 250$ and 500.

	N	M	Chao1	iChao1	GT	HT	MLE_{pb}	MLPE_{pb}	MLE_{al}	LS
BIAS($\widehat{N}$)										
	100	100	2.62	6.49	-0.61	-12.91	-3.64	-4.33	-1.46	9.38
	100	200	1.06	2.43	13.81	7.70	-0.94	-0.48	-0.36	2.91
	250	250	6.73	18.07	2.06	-29.19	0.16	0.61	2.69	31.69
	250	500	0.45	2.90	33.81	18.92	-1.02	0.38	-0.60	4.54
	500	500	7.59	24.04	1.70	-59.98	-0.75	5.31	1.53	37.89
	500	1000	3.66	9.14	70.12	39.66	0.41	6.27	1.18	9.58
RMSE($\widehat{N}$)										
	100	100	18.00	22.00	7.80	14.10	12.50	12.90	12.40	31.90
	100	200	6.90	8.40	15.00	8.80	4.50	4.70	4.20	10.00
	250	250	27.00	37.70	12.30	30.50	18.40	20.50	19.20	91.10
	250	500	9.20	11.30	35.00	20.30	7.80	8.90	8.00	12.10
	500	500	38.00	54.00	18.80	61.50	28.30	39.40	28.40	89.50
	500	1000	17.80	24.40	71.40	40.70	9.10	23.30	9.10	19.30

7.6 Conclusion

The exact distribution of the number of seen species is complicated. Altham's multiplicative-binomial distribution is an alternative approximation to the distribution of number of seen species. Particularly, when $N < M$, MLE estimator with Altham's multiplicative-binomial distribution can approximate well. The homogeneous abundance model is used in our study.

The simplest case of a homogeneous population is not of much practical interest in ecology, but a manageable expression for the exact distribution of the number of distinct species is available in this case. The heterogeneous abundance models could be applied in the next study including the Zipf, broken-stick,

Dirichlet and so on.

When using the MLE estimator with Altham's multiplicative-binomial distribution, it required a long time in computation as a result of complicated formula. Additionally, there is the local optimization problem for the MLPE and LS estimator in some situations. Although, the SANN method can handle this problem, this method used a long time for optimization and might not converge at all.

When splitting data, it provides less information than the full data. The pseudo-likelihood approach provides misleadingly narrow confidence interval compared to the true likelihood. This is likely to be because overlapping samples give to much weight to the initial observations. Further work is needed to investigate methods of correcting for this effect. For the Poisson-binomial approximation, it shows a nonsmooth result for the log-likelihood curve. For homogeneous abundance data, it is clear that the MLE approach performs better than the MPLE.

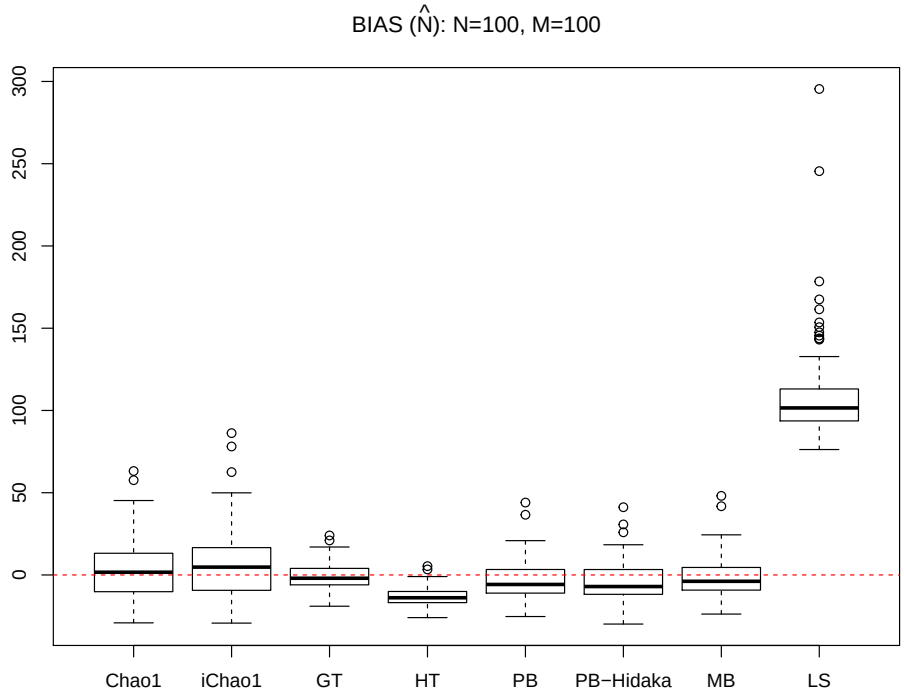


Figure 7.3: Bias of $\widehat{N}$ using various estimators, $N = 100, M = 100$ with homogeneous model.

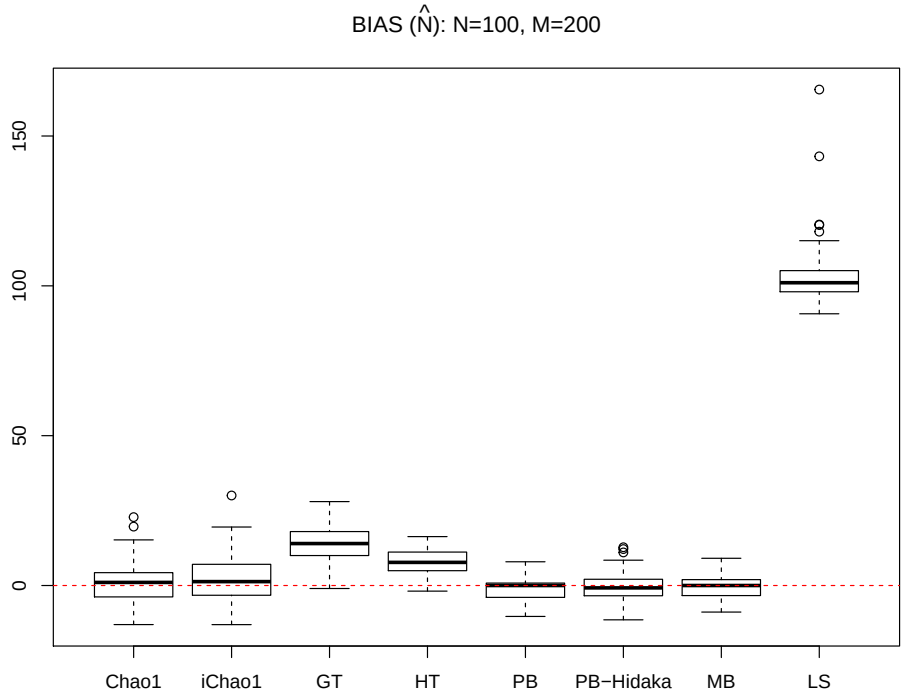


Figure 7.4: Bias of $\widehat{N}$ using various estimators, $N = 100, M = 200$ with homogeneous model.

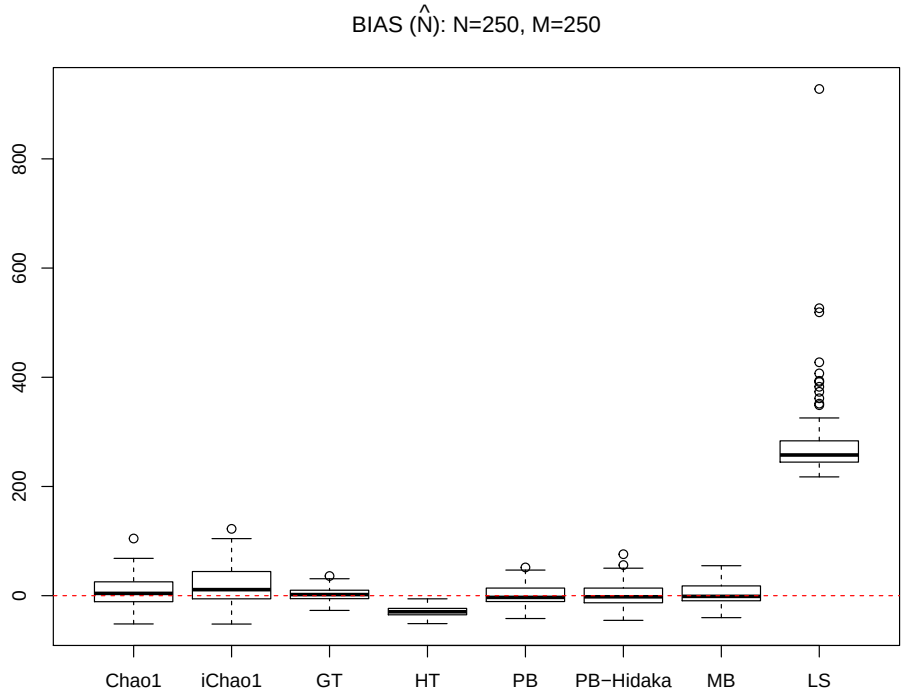


Figure 7.5: Bias of $\widehat{N}$ using various estimators, $N = 250, M = 250$ with homogeneous model.

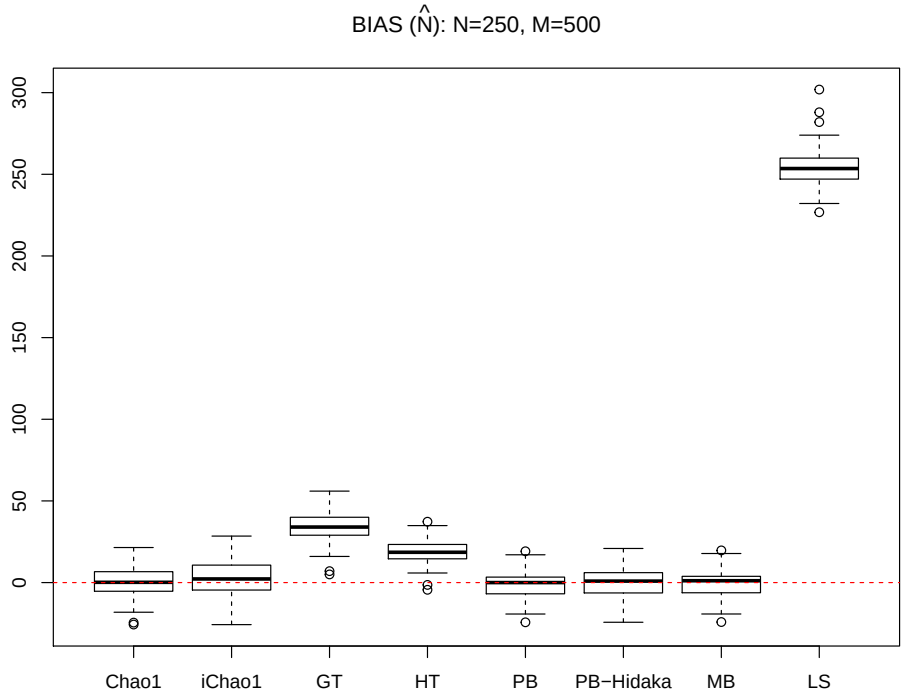


Figure 7.6: Bias of $\widehat{N}$ using various estimators, $N = 250, M = 500$ with homogeneous model.

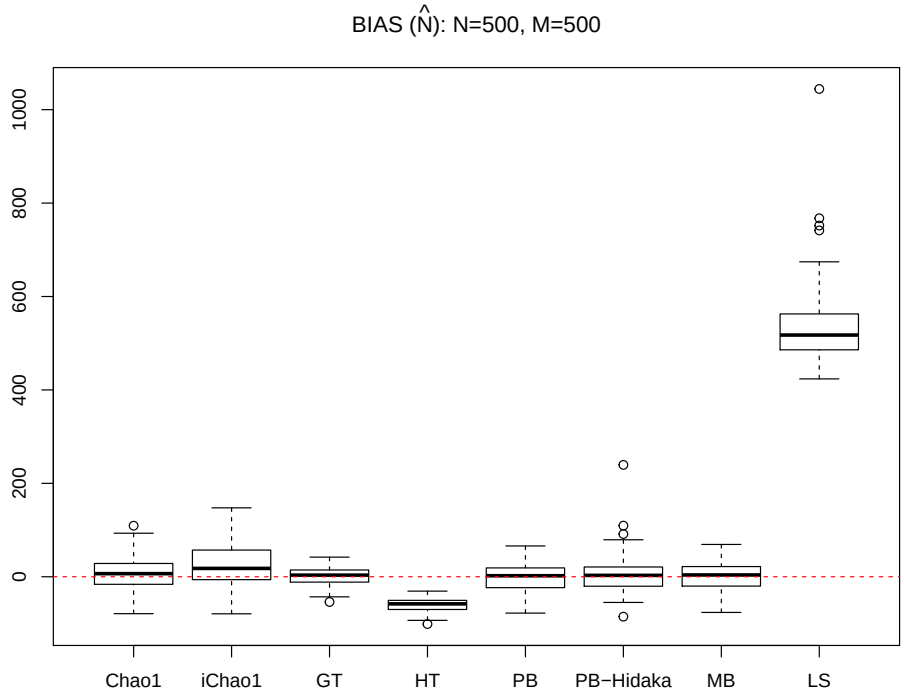


Figure 7.7: Bias of $\widehat{N}$ using various estimators, $N = 500, M = 500$ with homogeneous model.

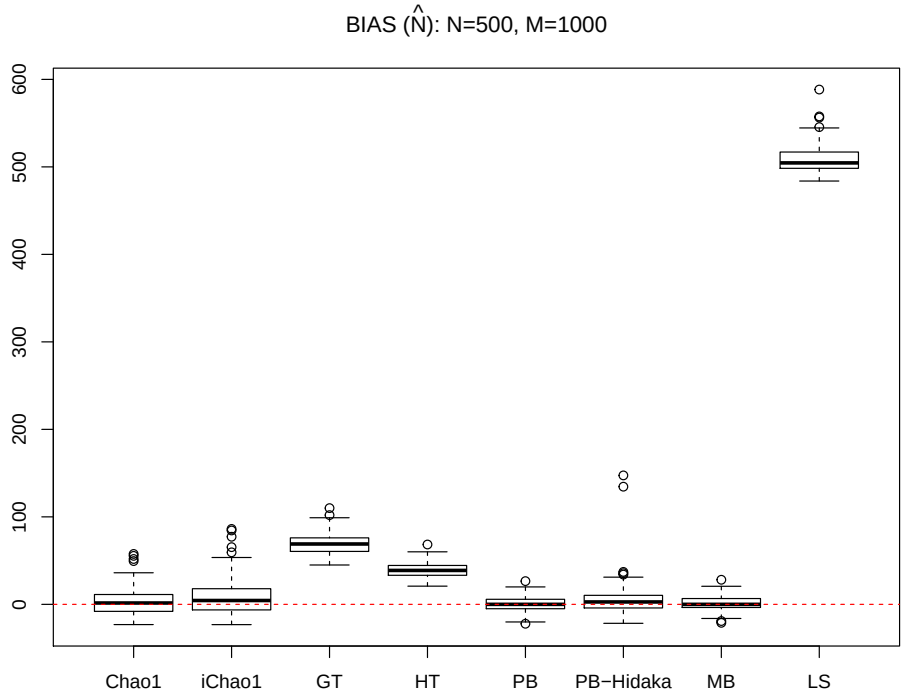


Figure 7.8: Bias of $\widehat{N}$ using various estimators, $N = 500, M = 1000$ with homogeneous model.

Chapter 8

Conclusion and Future work

8.1 Conclusion

In this thesis, we have examined the statistical methods for estimating the number of species in a closed population. Nonparametric and parametric estimators are investigated based on various species abundance models. Due to anthropogenic and environmental changes, these lead to unequal species detection probability. Therefore, the heterogeneity models are more appropriate in practice.

In species sampling, the numbers of species seen i times ($i = 1, 2, \ldots, k$) in the sample are used to estimated the number of unseen species is using various approaches. The Chao1 estimator is a nonparametric estimator used widely for estimating the total number of species as the lower bound and performs well for the homogeneous population.

Chiu et al. (2014) improves on the Chao1 estimator. It approximates well and outperforms the Chao1 estimator in term of bias and the mean square error especially for a highly heterogeneous population. The iChao estimator is constructed using $\widehat{N}_{Chao1} + |\text{bias}(\widehat{N}_{Chao1})|$. A modified Good-Turing frequency

formula is used in the second term. In this case, the number of species seen exactly once, twice, three and four times are used to estimate the number of species.

In Chapter 2, two new alternative improvements to the Chao1 estimators are developed using the same idea as the iChao1 estimator. New estimators, new_1 and new_2, are constructed using the Good-Turing coverage formula to approximate $|\text{bias}(\widehat{N}_{Chao1})|$. These estimators require only the number of species seen exactly once and twice which is very similarly to the Chao1 estimator. We found that the performance of the new_1 estimator is similar to the iChao1 estimator under heterogeneous models. The new_1 estimator works well with the negative binomial, the power-decay, the Zipf-Mandelbrot and log-series model. New_2 performs better than new_1 estimator for the broken-stick and exponential models.

We also considered using a parametric approach such as the MLE estimator based on the mixed-Poisson model to fit the model for estimating species richness. For the PT model, the MLE has problems due to the flat likelihood or the boundary problem in optimization. In Chapter 4, we addressed this problem using the WLR estimator. We showed that the WLR estimator works well with the PT distribution for large N . The WLR estimator does not work well for small sample size because frequencies are estimated poorly.

When applying nonparametric estimators to the PT distribution, the performance of the estimators depends on the dispersion parameter. From a simulation study, the new_2 estimator is a good approximation for the PT distribution with dispersion $D \geqslant 1.5$. For the lower dispersion, the new_2 estimator performs similarly or a little worse when compared with the new_1 estimator. Additionally, nonparametric estimators perform better than the WLR estimator under

the PT distribution.

The performance of the WLR estimator is improved by using the smoothing method in Chapter 5. Therefore, the problem about the zero and small frequencies are handled. The simulation study focuses on the uniform, the geometric and the Li and Racine (2010) kernel functions. It is clear that the performance of the WLR estimator is best improved under the Li and Racine (2010) kernel function. However, the computation requires very long time for bandwidth selection. The results show only small improvement in performance of the WLR estimator when applying smoothing. Therefore, it might be not necessary to apply smoothing method for improving the WLR estimator.

In Chapter 6, we investigated species sampling using the urn models. The occupancy distribution can explain the distribution of the number of seen species. Some approximations to the occupancy distribution are explored. Under the classical occupancy problem, each individual is drawn randomly from a population with equal probability. The simulation study shows that the Altham's multiplicative binomial and the Pólya distribution performs very well and provide similar results. The performance was particularly good for data generated from the Zipf-Mandelbrot distribution. The COM-Poisson-binomial distribution performs well when abundance data are generated from the Poisson, expo-decay and power-decay models (when $M/N \leq 10$). However, numerical issues occur for the Pólya distribution in some situations especially for the Poisson and expo-decay model.

Finally, we focussed on the pseudo-likelihood estimator under the classical occupancy problem. Hidaka (2014) proposed the pseudo-likelihood estimator based on the Poisson-binomial distribution for estimating the number of species. Multiple data sets are generated form the original data (i.e. non-

overlapping and overlapping data sets) in order to construct the pseudo-likelihood function. The MLE approach works well based on the Poisson-binomial and Altham's multiplicative distribution. The MPLE approach give less information. Therefore, the MLE estimates the number of species more accurately than the MPLE. When looking at the likelihood function the overlapping scheme for subset data provides better results than nonoverlapping.

When applying the MPLE method for estimating the number of species to data generated from the homogeneous model, we might only find local optima. Although the SANN method can handle this issue very well, it does not always work. The MLE estimator based on the Poisson-binomial and Altham's multiplicative binomial distribution are used for estimating the number of species very well when $N < M$. The MPLE works well when N is small. The Good-Turing estimator performs the best when $N = M$. The LS and iChao1 estimator approximate poorly in our study.

8.2 Future work

1. To improve the WLR estimator using the kernel estimation, we would like to examine other kernel discrete functions such as the Poisson, the binomial and the negative binomial kernel. For boundary problem in kernel estimation, the modified discrete triangular kernel in study of Kokonendji and Zocchi (2010) is probably useful guidance.

2. In Chapter 4, the ratio plot shows a nonlinear relation in the PT distribution (e.g. $a = 0.75, 0.9$). The WLR approach is not appropriate for use. Böhning et al. (2016) proposed using the fractional polynomials for the nonlinear regression model which could be applied with the PT distribution.

3. To explore other approximations to the distribution of number of seen species such as COM-Poisson-binomial distribution and so on for the MPLE and MLE estimators.

4. Only the homogeneous abundance model is investigated in Chapter 7. We would like to focus on heterogeneous abundance models such as the Zipf, log-normal, broken-stick model for the MPLE and MLE and LS estimators.

Bibliography

Aitchison, J. and Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Aitken, C. (1983). Kernel methods for the estimation of discrete distributions. *Journal of Statistical Computation and Simulation*, 16(3-4):189–200.

Altham, P. M. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, 27(2):162–167.

Barbour, A. and Holst, L. (1989). Some applications of the Stein-Chen method for proving Poisson convergence. *Advances in Applied Probability*, 21(1):74–90.

Barger, K. and Bunge, J. (2008). Bayesian estimation of the number of species using noninformative priors. *Biometrical Journal*, 50(6):1064–1076.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology*, 5(5):410–423.

Böhning, D. (2015). Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *Metron*, 73(2):201–216.

Böhning, D., Rocchetti, I., Alfó, M., and Holling, H. (2016). A flexible ratio re-

gression approach for zero-truncated capture–recapture counts. *Biometrics*, 72(3):697–706.

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Applied Statistics*, 54(4):721–737.

Borges, P., Rodrigues, J., Balakrishnan, N., and Bazán, J. (2014). A COM–Poisson type generalization of the binomial distribution and its properties and applications. *Statistics and Probability Letters*, 87:158–166.

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. (1998). Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, 79(3):1018–1028.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Bunge, J. and Barger, K. (2008). Parametric models for estimating the number of classes. *Biometrical Journal*, 50(6):971–982.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373.

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3):625–633.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, 58(3):531–539.

Chao, A. and Chiu, C. (2014). Estimation of species richness and shared species richness. In *Handbook of Methods and Applications of Statistics in the Atmospheric and Earth Sciences.* Wiley, New York, 76-111.

Chao, A. and Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547.

Chao, A., Wang, Y., and Jost, L. (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, 4(11):1091–1100.

Chiarucci, A., Enright, N., Perry, G., Miller, B., and Lamont, B. (2003). Performance of nonparametric species richness estimators in a high diversity plant community. *Diversity and Distributions*, 9(4):283–295.

Chiu, C.-H., Wang, Y.-T., Walther, B. A., and Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, 70(3):671–682.

Colwell, R. K. (2009). Biodiversity: Concepts, patterns, and measurement. In *the Princeton Guide to Ecology.* Wiley, New York, 257-263.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., and Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1):294–301.

Cruyff, M. J. and van der Heijden, P. G. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, 50(6):1035–1050.

David, F. N. and Barton, D. E. (1962). *Combinatorial Chance*. Griffin, London.

Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280.

Eggenberger, F. and Pólya, G. (1923). Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289.

El-Shaarawi, A. H., Zhu, R., and Joe, H. (2011). Modelling species abundance using the Poisson–Tweedie family. *Environmetrics*, 22(2):152–164.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications: Volume One*. Wiley, New York.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42–58.

Gerber, H. U. (1992). From the generalized gamma to the generalized negative binomial distribution. *Insurance: Mathematics and Economics*, 10(4):303–309.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Gotelli, N. J. and Colwell, R. K. (2011). Estimating species richness. *In Biological diversity: frontiers in measurement and assessment*, 12:39–54.

Grogger, J. T. and Carson, R. T. (1991). Models for truncated counts. *Journal of Applied Econometrics*, 6(3):225–238.

Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory*, 11(4):235–243.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32.

Hidaka, S. (2014). General type-token distribution. *Biometrika*, 101(4):999–1002.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hougaard, P., Lee, M.-L. T., and Whitmore, G. (1997). Analysis of overdispersed count data by mixtures of poisson variables and poisson processes. *Biometrics*, 53(4):1225–1238.

Huang, B. and Zhan, R. (2014). Species-abundance models for brachiopods across the OrdovicianœSilurian boundary of South China. *Estonian Journal of Earth Sciences*, 63(4):240–243.

Janardan, K. G. and Schaeffer, D. J. (1981). Methods for estimating the number of identifiable organic pollutants in the aquatic environment. *Water Resources Research*, 17(1):243–249.

Janzen, D. H. (1973). Sweep samples of tropical foliage insects: effects of seasons, vegetation types, elevation, time of day, and insularity. *Ecology*, 54(3):687–708.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions*, volume 444. Wiley, Chichester.

Johnson, N. L. and Kotz, S. (1977). *Urn models and their application: an approach to modern discrete probability theory*. Wiley, New York.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B*, 49(2):127–162.

Kokonendji, C. C., Dossou-Gbété, S., and Demétrio, C. G. (2004). Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demetrio classes. *Statistics and Operations Research Transactions*, 28(2):201–214.

Kokonendji, C. C. and Kiessé, T. S. (2011). Discrete associated kernels method and extensions. *Statistical Methodology*, 8(6):497–516.

Kokonendji, C. C. and Zocchi, S. S. (2010). Extensions of discrete triangular distributions and boundary bias in kernel estimation for discrete functions. *Statistics and Probability Letters*, 80(21):1655–1662.

Kolchin, V. F., Sevastyanov, B. A., and Chistyakov, V. P. (1978). *Random allocations*. Winston.

Lanumteang, K. and Böhning, D. (2011). An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics and Data Analysis*, 55(7):2302–2311.

Lewontin, R. C. and Prout, T. (1956). Estimation of the number of different classes in a population. *Biometrics*, 12(2):211–223.

Li, Q. and Racine, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory*, 26(06):1607–1637.

Li, R. and Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2):111–120.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130.

Longino, J. T., Coddington, J., and Colwell, R. K. (2002). The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology*, 83(3):689–702.

Magurran, A. and Henderson, P. (2011). Commonness and rarity. In *Biological diversity: frontiers in measurement and assessment.* Oxford University Press, Oxford, UK, 97-104.

McCrea, R. S. and Morgan, B. J. (2014). *Analysis of Capture-Recapture Data.* CRC Press, Bc ca Raton, USA.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, 10(10):995–1015.

Morgan, B. J. (2008). *Applied Stochastic Modelling.* CRC Press, London, 130.

Mouillot, D. and Lepretre, A. (2000). Introduction of relative abundance distribution (rad) indices, estimated from the rank-frequency diagrams (rfd), to assess changes in community diversity. *Environmental Monitoring and Assessment*, 63(2):279–295.

Norden, N., Chazdon, R. L., Chao, A., Jiang, Y.-H., and Vílchez-Alvarado, B. (2009). Resilience of tropical rain forests: tree community reassembly in secondary forests. *Ecology Letters*, 12(5):385–394.

Nunnikhoven, T. S. (1992). A birthday problem solution for nonuniform birth frequencies. *The American Statistician*, 46(4):270–274.

Pledger, S. and Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, 50(6):1022–1034.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(13):355–375.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Rocchetti, I., Bunge, J., and Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *The Annals of Applied Statistics*, 5(2B):1512–1533.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Samuel-Cahn, E. (1974). Asymptotic distributions for occupancy and waiting time problems with positive probability of falling through the cells. *The Annals of Probability*, 2(3):515–521.

Sanathanan, L. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association*, 72(359):669–672.

Sevast'Yanov, B. and Chistyakov, V. (1964). Asymptotic normality in the classical ball problem. *Theory of Probability and Its Applications*, 9(2):198–211.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conway–maxwell–poisson distribution. *Applied Statistics*, 54(1):127–142.

Simonoff, J. S. (1995). Smoothing categorical data. *Journal of Statistical Planning and Inference*, 47(1):41–69.

Skipper, M. et al. (2012). A Pólya approximation to the Poisson-binomial law. *Journal of Applied Probability*, 49(3):745–757.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *In Annals of Mathematical Statistics*, 29(2):614.

Valero, J., Pérez-Casany, M., and Ginebra, J. (2010). On zero-truncating and mixing Poisson distributions. *Advances in Applied Probability*, 42(4):1013–1027.

van der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., and van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, 3(4):305–322.

Vergne, T., Calavas, D., Cazeau, G., Durand, B., Dufour, B., and Grosbois, V. (2012). A Bayesian zero-truncated approach for analysing capture–recapture count data from classical scrapie surveillance in France. *Preventive Veterinary Medicine*, 105(1):127–135.

Wang, J.-P. and Lindsay, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology*, 5(1):30–45.

Wang, M.-C. and Van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, 68(1):301–309.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312.

Williamson, P. P. (2012). Usefulness of asymptotic distributions in the classical occupancy problem. *Communications in Statistics-Simulation and Computation*, 41(8):1501–1517.