UNIVERSITY OF SOUTHAMPTON

FACULTY OF HUMAN, SOCIAL AND MATHEMATICAL SCIENCES

School of Mathematical Sciences

Capture-Recapture Modelling for Zero-truncated Count Data Allowing for Heterogeneity

by

Orasa Anan

Thesis for the degree of Doctor of Philosophy

September 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF HUMAN, SOCIAL AND MATHEMATICAL SCIENCES
School of Mathematical Sciences

Doctor of Philosophy

CAPTURE-RECAPTURE MODELLING FOR ZERO-TRUNCATED COUNT DATA ALLOWING FOR HETEROGENEITY

by Orasa Anan

Capture-recapture modelling is a powerful tool for estimating an elusive target population size. This thesis proposes four new population size estimators allowing for population heterogeneity. The first estimator is developed under the zero-truncated of generalised Poisson distribution (ZTGP), called the MLEGP. The two parameters of the ZTGP are estimated by using a maximum likelihood with the Expectation-Maximisation algorithm (EM algorithm).

The second estimator is the population size estimator under the zero-truncated Conway-Maxwell-Poisson distribution (ZTCMP). The benefits of using the Conway-Maxwell-Poisson distribution (CMP) are that it includes the Bernoulli, Poisson and geometric distribution as special cases. It is also flexible for over- and under-dispersions relative to the original Poisson model. Moreover, the parameter estimates can be achieved by a simple linear regression approach. The uncertainty in estimating variances of the unknown population size under new estimator is studied with analytic and resampling approaches.

The geometric distribution is one of the nested models under the Conway-Maxwell-Poisson distribution, the Turing and the Zelterman estimators are extended for the geometric distribution and its related model, respectively. Variance estimation and confidence intervals are constructed by the normal approximation method.

An uncertainty of variance estimation of population size estimators for single marking capture-recapture data is studied in the final part of the research. Normal approximation and three resample approaches of variance estimation are compared for the Chapman and the Chao estimators. All of the approaches are assessed through simulations, and real data sets are provided as guidance for understanding the methodologies.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Orasa Anan, declare that the thesis entitled *Capture-Recapture Modelling for Zero-truncated Count Data Allowing for Heterogeneity* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: Anan et al. (2016)

Signed:.....

Date:.....

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me, and it would not have been possible to do without the support and guidance that I received from many people.

First and foremost, I would like to express my special appreciation and thanks to my supervisor, Professor Dankmar Böhning for providing me with the opportunity to study in the University of Southampton under his supervision. His guidance, understanding and patience lead to this great thesis project. Also, I would like to thank Dr. Antonello Maruotti, who is my second supervisor, for his spirited supervision and memorable supporting me. Without them, I could not have completed this thesis.

I would also like to thank my committee members, Professor Heinz Holling and Professor Peter G M Van der Heijden for your valuable comments and suggestions.

I am deeply thankful to my parents, Prakob and Jeaw Anan, Nakpalad family especially Thungpang and Numpan for love, support, and sacrifices. Without them, this thesis would never have been written.

I would also like to thank all of my friends, Dr. Waraporn Ratsameepakai, Dr. Wanpen Chantarangsi, Dr. Saowapa Chaiwong, L.T Prin Kanyoo, Dr. Jarunee - Dr. Wiriya Duangsuwan, Mathawee Srisawat, Jitayu Lawaruyttanakorn, Jaruwan Heangmanee, Lt.Cdn Sarawuth Srinakaew, Assist. Prof. Dr. Amonsak Sawusdee and Thai SOTON family in Southampton, who have supported me for writing up, and encouraged me to strive towards my goal.

Lastly, I gratefully acknowledge the Ministry of Science and Technology and the Royal Thai Government for providing Ph.D funding for studying in the UK.

Chapter 1

Introduction

1.1 Introduction

Information about the size of the target population is utilised in many areas. It can be used to evaluate the impact of threats, assess response to management actions designed to alleviate threats, and highlight areas where further research is required. The capture-recapture approach has been ordinarily used to estimate the elusive target population size of various species of wildlife in ecological science. In recent decades, this approach has also been applied to other areas such as epidemiology and surveillance (Ballivet et al., 2000; Böhning et al., 2005; Van Hest et al., 2008), clinical studies (Spoor et al., 1996), social science (Van der Heijden et al., 2003; Farcomeni et al., 2013) and computer system science (Liu et al., 2015), with the aim of estimating the sizes of particular populations.

The capture-recapture approach requires a series of data of repeated counts where each count reflects the number of times that a unit has been observed. In practice, some mechanisms such as a live trapping, a register or a screening test carry out a number of observations but leave some members of the population undetected. These hidden members, which result in missing or unobserved data (zero counts) in a population, lead to zero-truncated count data. Recently, many capture-recapture models for zero-truncated data have been developed to estimate hidden target population sizes. In this thesis, we restrict the study to situations in which the identifying mechanism is based on counting repeated identifications of the same unit within a given time span. Zero-truncated count data can be modelled by a zero-truncated Poisson distribution with a location parameter λ, but the Poisson distribution rarely occurs in reality. For example, level of education may lead to different behavioural responses in the target population studies. Therefore, individual capture-recapture probabilities are likely to be heterogeneous. The heterogeneous model for capture-recapture data can be classified into two groups, the first group is an evidence of observed heterogeneity that can described by covariates such as sex, age, location or level of education etc. The second group is the presence of unobserved

heterogeneity. The further studies of heterogeneity can be found in Böhning (2000); Link (2003); Huggins and Hwang (2011); Stoklosa et al. (2011).

The occurrence of heterogeneity leads to a violation of the Poisson distribution property that the variance is equal to the mean, the Poisson distribution therefore might not be suitable for capture-recapture data. Importantly, ignorance of heterogeneity results in a negative bias and an underestimate of population size N (Van der Heijden et al., 2003; McCrea and Morgan, 2014; Tounkara and Rivest, 2015). Thus, to mitigate the potential bias in population size due to heterogeneity many capture-recapture models have been proposed (see Chao, 1987; Chao and Bunge, 2002; Cruyff and Van der Heijden, 2008). The presence of heterogeneity in capture-recapture data often results in variance that is greater than mean, known as over-dispersion. A classical approach to account for heterogeneity is a Poisson mixture model that provides a flexible model capable of dealing with this issue. The negative binomial, which is the gamma-Poisson mixture, has been used instead of the original Poisson model. For example, Chao and Bunge (2002); Rocchetti et al. (2011); Lanumteang and Böhning (2011) developed population size estimators under the zero-truncated negative binomial model. Nevertheless, the length of the dispersion parameter in the negative binomial model is limited. Alternatively, an extension of the zero-truncated Poisson distribution with dispersion parameter to account for the heterogeneity is studied in this thesis. We wish to contribute extending this branch of literature by proposing more general count distributions that capture a wider range of dispersion setting than the negative binomial distribution.

The research presented in this thesis focuses on exploring more appropriate alternative distributions for zero-truncated capture-recapture data. This leads to new population size estimators based on their suitable distributions. Additionally, alternative choices of variance estimation of population size estimator for capture-recapture data are developed which depend on a resampling approach.

1.2 Objectives of the study

(a) Main objective

To explore some estimators of elusive target population size based on zero-truncated count modelling that accounts for individuals heterogeneity.

(b) Secondary objectives

1. To explore some of the difficulties and limitations of zero-truncated Poisson modelling.
2. Using the graphical device of the ratio plot to explore some of the limitations of negative binomial modelling.

3. To explore the suitability and appropriateness of zero-truncated generalised Poisson modelling by using the graphical device of the ratio plot.

4. To develop a new estimator of population size based on the zero-truncated generalised Poisson distribution.

5. To extend the graphical device of the ratio plot to the log-ratio plot for investigating the validity of the zero-truncated Conway-Maxwell-Poisson model in capture-recapture study.

6. To develop a new estimator of population size based upon the zero-truncated Conway-Maxwell-Poisson distribution.

7. To examine the uncertainty of variance estimation for the new population size estimator under the zero-truncated Conway-Maxwell-Poisson distribution.

8. To extend the Turing and the Zelterman based-Poisson estimators to the Turing and Zelterman based-geometric estimators.

9. To examine the uncertainty of variance estimation for the single marking capture-recapture count data.

1.3 Basic assumptions made throughout the thesis

1. The target population is a closed system.

2. Sources or occasions of counting are independent.

3. Sampling of individuals is independent.

1.4 Thesis outline

The thesis consists of eight chapters. The first Chapter is the introduction of this work and is followed by the literature review of the capture-recapture approach in Chapter 2. Additionally, some difficulties and limitation of the Poisson and the negative binomial are explored in this chapter.

In Chapter 3, a new estimator of population size using the maximum likelihood estimation of generalised Poisson distribution (MLEGP) is discussed. The performance of the new estimator is compared with other well-known estimators and considered under a data generating mechanism using both Poisson and generalised Poisson models. Some capture-recapture data examples under this model are the Shakespeare's data and the cottontail rabbit data.

In Chapter 4, the population size estimator is proposed based on the Conway-Maxwell-Poisson distribution (CMP). A graphical device namely the log-ratio plot is modified as a simple tool for investigating a zero-truncated Conway-Maxwell-Poisson distribution (ZTCMP). As a benefit of the log-linear scale, the ratio plot can be used to estimate two parameters of the ZTCMP, leading to a straightforward approach to estimate the zero counts and the target population size. Real data examples illustrate the practical use of the new estimator under the CMP distribution.

In Chapter 5, the model uncertainty of estimating variance based on the new estimator in Chapter 4 is examined by using normal approximation and resampling approaches. Three bootstrap methods: true bootstrap, imputed bootstrap and reduced bootstrap, are discussed in depth and compared with the normal approximation approach.

In Chapter 6, since the geometric distribution is nested in the Conway-Maxwell-Poisson distribution, other two estimators, the original Turing based-geometric and the Zelterman based-geometric are proposed as the alternative estimators for the geometric and contaminated geometric models, respectively. Additionally, analytic variance formulas are derived for the new estimators. The simulation scenarios under the geometric distribution are provided in this chapter to evaluate the performance of the new estimators. Three case studies are provided as practical guidance for the capture-recapture data following zero-truncated geometric distribution.

In Chapter 7, an extension of resampling approaches is applied for investigating variance estimation for single marking capture-recapture data. Simulation scenarios are set under the assumption of independence between two occasions and its violations. Four case studies are provided as illustration of use in practice.

The thesis ends with a chapter of conclusions and future work in Chapter 8.

1.5 Notation

Table 1.1: List of notation used throughout in the thesis

Symbol	Notation and Definition
i	An index for the individual i of the target population
j	An index of capture occasion
X_i	The random variable denoting the number of times that the i^{th} individual is identified during the study period; $X_i \in \{0, 1, 2, 3, ...\}$
m	The largest number of observed count in the population
N	A population size
$\widehat{N}$	A population size estimator
n	The number of observed individuals
S	The total number of identification during study period
f_x	The frequency of number of identifying individual exactly x times
f_0	The frequency of unobserved members of the target population member with count of observations $x = 0$

Chapter 2

Review of Capture-recapture Approach

This chapter provides a review of the background of capture-recapture. It describes a zero-truncated count approach to the modelling of capture-recapture data and their distributions, and two types of structure for capture-recapture data are introduced in the second section. In addition, some interesting estimators of population size are provided in section three. This is followed by some examples demonstrating the application of the capture-recapture approach in section four. Section five presents the graphical device of the ratio plot, as a tool for investigating the validity of the zero-truncated Poisson, the zero-truncated negative binomial and the zero-truncated geometric models. Additionally, some limitations of negative binomial modelling are provided in the last section.

2.1 The Zero-truncated count of capture-recapture data

The concept of the census approach is a method of measuring population size. However, it has limitations to measure elusive populations such as a wildlife population or human illegal inhabitants. A procedure can select some elements from an elusive target population but leave other elements undetected in the hidden part of the target population. For example, in the situation of using a registration mechanism to record drug users, some may remain undetected. Not every drug user uses a treatment centre on at least one occasion, hence some users remain undetected. Therefore, the recording system catches only the patients who come to treatment institutions. The existence of unobserved zero counts in a population is called *zero-truncated count data*. Statistical inference is applied to estimate an elusive target population size by using the zero-truncated count information.

2.1.1 The zero-truncated count distribution

A zero-truncated count distribution is defined by a conditional probability function, it is often used to model observed frequency data (McCrea and Morgan, 2014). Let $Pr(X = x)$ present the probability of observable variable X take on value x and $Pr(X|X > 0)$ is the probability of observing $X = x$ given that $X > 0$, and $Pr(X > 0)$ represents the probability of $X > 0$:

$$Pr(X|X > 0) = \frac{Pr(X = x)}{Pr(X > 0)} = \frac{Pr(X = x)}{1 - Pr(X = 0)}. \tag{2.1}$$

This can be written as

$$p_x^+ = \frac{p_x}{1 - p_0}, \tag{2.2}$$

where the zero-truncated probability function is denoted p_x^+, p_0 is the probability at $x = 0$, and p_x represents the discrete mass probability function $p_x = P(X = x)$.

For example, the zero-truncated Poisson distribution, $Po_+(x; \lambda)$, for count data means the distribution is truncated at $x = 0$. According to the above (2.2) we have that

$$Po_+(x; \lambda) = \frac{Po(x; \lambda)}{Po(x > 0; \lambda)} = \frac{Po(x; \lambda)}{1 - Po(0; \lambda)} = \frac{\exp(-\lambda)\lambda^x}{x!\,\{1 - \exp(-\lambda)\}}, \tag{2.3}$$

where $Po(0; \lambda) = \exp(-\lambda)$ and $1 - Po(0; \lambda) = 1 - \exp(-\lambda)$. An estimator $\widehat{\lambda}$ for λ can be obtained by fitting the zero-truncated Poisson distribution.

2.1.2 Capture probability and probability of a zero count

In a capture-recapture study, p_0 is defined as the probability of a zero count. Also, the probability of a capture is given as $1 - p_0$. This leads to size of population, N, being divided into

$$\underbrace{N}_{\text{population}} = \underbrace{N(1 - p_0)}_{\text{observed}} + \underbrace{Np_0}_{\text{unobserved}} \approx n + Np_0 \tag{2.4}$$

where $N(1 - p_0)$ is the observed part which can be estimated by n, that is $E(n) = N(1 - p_0)$. Consequently, the Horvitz-Thompson estimator of N is $\widehat{N} = \dfrac{n}{1 - p_0}$.

2.2 Structure of the capture-recapture data

The idea of capture-recapture approach is applied to estimate the size of an elusive target population by using some mechanisms such as live trapping and registration. Each individual can be repeatedly identified or recaptured over the study period. Generally, capture-recapture data can be divided into two types of structures, some arising from repeated count data, and others from different sources of data.

2.2.1 Capture-recapture data based upon numbers of repeated counting data

Suppose that m denotes the number of counting occasions over a period of study. Let Y_{ij} be the indicator of the individual i observed on occasion j, where $i = 1, 2, 3, ..., N$ and $j = 1, 2, 3, ..., m$. Hence,

$$Y_{ij} = \begin{cases} 1, & \text{if the individual } i \text{ is observed on the occasion } j \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

Note that Y_{ij} is only observed if $Y_{ij} > 0$ at least for one j and $j = 1, 2, 3, ..., m$.

According to the capture-recapture history, Table 2.1 shows the count for each individual and how often it was identified during the study period on m occasions. The marginal $X_i = \sum_{j=1}^{m} Y_{ij}$ is the number of times that the individual i is identified. Thus, the marginal frequency count X_i has possible values $0, 1, 2, 3, ..., m$. Hence the target population $(X_1, X_2, ..., X_N)$ is divided into two parts, the first part named the observed or the untruncated part $(X_1, X_2, X_3, ..., X_n)$, and the second part is the unobserved or truncated part $(X_{n+1}, X_{n+2}, ...X_N)$. In practice, the objective of estimating the size of a target population means an attempt to estimate unobserved data that we do not know, and the total target population is estimated from the estimated observed samples and unobserved information.

Table 2.1: Capture-recapture history

Individual (i)	Occasion (j)					$X_i = \sum_{j=1}^{m} Y_{ij}$
	1	2	3	...	m	
1	$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	...	$Y_{1,m}$	X_1
2	$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	...	$Y_{2,m}$	X_2
3	$Y_{3,1}$	$Y_{3,2}$	$Y_{3,3}$	...	$Y_{3,m}$	X_3
4	$Y_{4,1}$	$Y_{4,2}$	$Y_{4,3}$	...	$Y_{4,m}$	X_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Y_{n,1}$	$Y_{n,2}$	$Y_{n,3}$	...	$Y_{n,m}$	X_n
$n+1$	$Y_{n+1,1}$	$Y_{n+1,2}$	$Y_{n+1,3}$	...	$Y_{n+1,m}$	X_{n+1}
$n+2$	$Y_{n+2,1}$	$Y_{n+2,2}$	$Y_{n+2,3}$	...	$Y_{n+2,m}$	X_{n+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	$Y_{N-1,1}$	$Y_{N-1,2}$	$Y_{N-1,3}$	...	$Y_{N-1,m}$	X_{N-1}
N	$Y_{N,1}$	$Y_{N,2}$	$Y_{N,3}$	...	$Y_{N,m}$	X_N

To model the arising count distribution, if each individual can be repeatedly counted only at a specific time point in the study period the binomial distribution seems to be appropriate to model the marginal random variable X_i. The binomial distribution

$Bin(m, p)$ arises when on each of m occasions identification takes place independently with homogeneous capture probability p. However, when each individual is counted at any time throughout the study period, a fixed number of counting occasions cannot be specified. It might be that the largest count is m. The Poisson distribution $Po(\lambda)$ is potentially suitable to fit the marginal frequency X_i, and a parameter λ denotes the mean of each individual being counted in the period of interest. The capture-recapture data can be written in form frequency of frequencies Table 2.2 as:

Table 2.2: The frequency distribution

x	0	1	2	3	4	5	6	7	8	9	10	...	m
f_x	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	...	f_m

Suppose that f_x is the frequency of units identified at exactly x occasions or times where $x = 0, 1, 2, 3, ..., m$, and m is the largest count. Therefore, the total number of observed units will be

$$f_1 + f_2 + f_3 + ... + f_m = \sum_{x=1}^{m} f_x = n,$$

and the total number of identifications is $\sum_{x=0}^{m} x f_x$, but f_0 is a missing part. As a consequence, the size of population can be rewritten as

$$N = f_0 + f_1 + f_2 + f_3 + ... + f_m = f_0 + \sum_{x=1}^{m} f_m = f_0 + n.$$

An example of capture-recapture history is set out in Table 2.3.

Example 1: Individual capture history of 38 deer mice with six capture occasions Amstrup et al. (2010).

Table 2.3: Capture-recapture history table of 38 deer mice with six occasions.

Individual (i)	Occasion (j)						$X_i = \sum_{j=1}^{m} Y_{ij}$
	1	2	3	4	5	6	
1	1	1	1	1	1	1	6
2	1	0	0	1	1	1	4
3	1	1	0	0	1	1	4
4	1	1	0	1	1	1	5
5	1	1	1	1	1	1	6
6	1	1	0	1	1	1	5
7	1	1	1	1	1	0	5
8	1	1	1	0	0	1	4
9	1	1	1	1	1	1	6
10	1	1	0	1	1	1	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
33	0	0	0	0	1	0	1
34	0	0	0	0	1	0	1
35	0	0	0	0	1	0	1
36	0	0	0	0	0	1	1
37	0	0	0	0	0	1	1
38	0	0	0	0	0	1	1
39	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$N-1$	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0

According to Table 2.3, $X_1 = 6$ means the sample unit 1 was counted six times or on six occasions. $X_2 = 4$ means the sample unit 2, which has count series 100111, was counted four times in occasion 1, 4, 5 and 6. This table also shows the 0 series of unobserved data that means we do not observe the 39^{th}, 40^{th} ,..., N, units, and they need to be estimated. The associated frequency table is given in Table 2.4.

Table 2.4: The frequency distribution of deer mice

x	0	1	2	3	4	5	6	n
f_x	–	9	6	7	6	6	4	38

Note that $f_1+f_2+f_3+f_4+f_5+f_6 = n = 38$ is the observed sample size.

Example 2: This study used all data on heroin drug users from 61 health treatment centres in the Bangkok metropolitan region collected by the Office of the Narcotics Control Board (ONCB), Ministry of the Prime Minister, which occurred from 1st October to 31st December in 2001 (Böhning et al., 2004).

Table 2.5: The observed frequency distribution of visits to health treatment clinics in Bangkok by heroin users

x	1	2	3	4	5	6	7	8	9	10	
f_x	2,176	1,600	1,278	976	748	570	455	368	281	254	
x	11	12	13	14	15	16	17	18	19	20	21
f_x	188	138	99	67	44	34	17	3	3	2	1

2.2.2 Capture-recapture data with different sources

Capture-recapture data can be obtained from different sources, when the data are identified at specified sources and each source overlaps with the others. As a consequence, the occasion is replaced by different sources. For illustration, the sources of identified individuals from two and three lists can be summarised as in Table 2.6 and Table 2.8, respectively.

The capture history of two sources can be summarised in a 2x2 contingency table with frequencies given as Table 2.6

Table 2.6: The capture-recapture history with two sources

	Source 2		
Source 1	Yes	No	Total
Yes	f_{11}	f_{10}	n_1
No	f_{01}	f_{00}	
Total	n_2		N

where

- f_{00} denotes the unobserved frequency.
- f_{10} denotes the frequency of individuals identified only at the first occasion.
- f_{01} denotes the frequency of individuals identified only at the second occasion.
- f_{11} denotes the frequency of individuals identified at both occasions.

The population size can be calculated as: $N = f_{00} + f_{10} + f_{01} + f_{11}$, or $N = f_0 + f_1 + f_2$, when f_x denotes the number of individuals presented in x source, $x \in \{0, 1, 2\}$. Therefore, $f_0 = f_{00}$, $f_1 = f_{10} + f_{01}$ and $f_2 = f_{11}$.

Example 3: Two sources. Abegaz et al. (2014) studied the incidence of road traffic injuries and deaths in Ethiopia between June 2012 and May 2013. Two sources of data come from the traffic police and hospital injury surveillance. It can be represented with 2x2 contingency table as in Table 2.7.

Table 2.7: Ethiopia data

	Source 2		
Source 1	Yes	No	Total
Yes	50	103	153
No	69	f_{00}	
Total	84		N

It is simple to calculate the number of deaths in Ethiopia which is given as $N = 50 + 103 + 69 + f_{00}$. However, the parameters N and f_{00} are unknown and required to be estimated.

For the three sources example, the capture history can be written as Table 2.8 below. The indicator 1 denotes that an individual is identified in source a or b or c, and $a, b, c \in \{0, 1\}$. For example, the series 1 0 0 means that this individual is observed only in source a , and f_{abc} means the total number of individuals are identified from a, b and c, respectively.

Table 2.8: The capture-recapture history with three sources

Sources			Frequency count
a	b	c	f_{abc}
0	0	0	f_{000}
1	0	0	f_{100}
0	1	0	f_{010}
0	0	1	f_{001}
1	1	0	f_{110}
1	0	1	f_{101}
0	1	1	f_{011}
1	1	1	f_{111}

The population size is

$$N = \underbrace{f_{000}}_{unobserved} + \underbrace{f_{100} + f_{010} + f_{001} + f_{110} + f_{101} + f_{011} + f_{111}}_{observed} = f_{000} + n,$$

where f_{000} denotes an unobserved frequency, and n is the observed sample size that is calculated from $n = f_{100} + f_{010} + f_{001} + f_{110} + f_{101} + f_{011} + f_{111}$

Additionally, suppose that f_x is the frequency of individuals identified exactly x times, $x = 1, 2, 3..., m$, and m is the number of sources. In this type of capture-recapture data, the number of counts is fixed and known before the capture-recapture sampling, and the

largest count is the number of sources. However, the number of unobserved individuals, f_0, is unknown and may be required to be estimated. We can summarise the frequencies in Table 2.9 as:

Table 2.9: The frequency distribution for a three source design

x	0	1	2	3
f_x	f_0	f_1	f_2	f_3

f_x is connected to f_{abc} as follows: $f_0 = f_{000}$, $f_1 = f_{100}+f_{010}+f_{001}$, $f_2 = f_{110}+f_{101}+f_{011}$, and $f_3 = f_{111}$. Thus, the population size can be achieved from

$$N = \underbrace{f_0}_{unobserved} + \underbrace{f_1 + f_2 + f_3}_{\text{observed}} = f_0 + n \tag{2.6}$$

Example 4: Three sources. Héraud-Bousquet et al. (2012) estimated the number of 216 new HIV diagnoses in children under 13 years old from three sources: EPF be ANRS-French Perinatal Cohort, DOVIH denotes Mandatory HIV case reporting and LaboVIH : Laboratory surveillance of HIV testing activity, in mainland France from January 2003 to December 2006. The three sources of data are shown as follows in Figure 2.1:

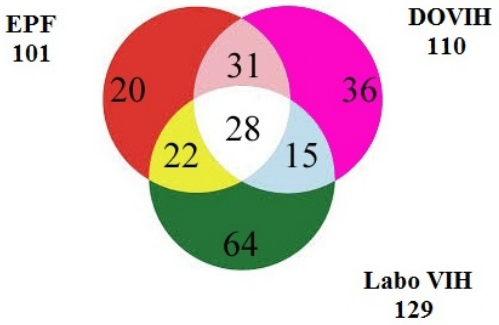

Figure 2.1: The number of new HIV cases in mainland France

We can rewrite the capture-recapture history with three sources as Table 2.10

Table 2.10: The capture-recapture history with three sources in a study of HIV cases in France

Source			Frequency count
DOVIH (a)	LaboVIH (b)	EPF (c)	f_{abc}
0	0	0	?
1	0	0	20
0	1	0	36
0	0	1	64
1	1	0	31
1	0	1	22
0	1	1	15
1	1	1	28

where $n = 216$. The associated frequency distribution is summarised as Table 2.11.

Table 2.11: The frequency table of new HIV cases in France

x	0	1	2	3
f_x	–	120	68	28

2.2.3 Notable differences between two types of capture-recapture data

Noteable differences exist between the two types of capture-recapture data (repeated counts and different sources)

- There are often more occasions than sources.
- The chronological order of occasions.

2.3 Overview of estimators

Estimating the size of an elusive population under the capture-recapture approach has been applied in various areas. A simple model is often used such as the homogeneous Poisson. However, it is often not appropriate in reality because of subpopulations created by covariates such as age, gender, size, location etc., or latent variables in which the capture probability takes on different values. This situation is referred to as heterogeneity. In a capture-recapture study, the zero counts disappear from a counting system as well as might do covariate information and unobserved heterogeneity.

Several estimators have been applied to estimate the size of target populations in capture-recapture count data. This section focuses on the majority of estimators based on homogeneous Poisson and heterogeneity models. Maximum likelihood estimator under the Poisson distribution, Turing's estimator and McKendrick's estimator are estimators for the homogeneous Poisson model. Heterogeneity models are estimated by using Zelterman estimator, Mantel-Haenszel estimator, Chao lower bound estimator and Chao and Bunge estimator.

2.3.1 Horvitz-Thompson's estimator

Horvitz-Thompson estimator (HT) was introduced by Horvitz and Thompson in 1952, in sampling theory. This well-known approach is frequently used in capture-recapture analysis (McCrea and Morgan, 2014). Let I_i be the counting indicator of each individual from population size N. Hence

$$I_i = \begin{cases} 1, & \text{if individual } i \text{ is observed (sampled)} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that the identification mechanism is able to identify each individual with probability $(1 - p_0)$, where p_0 denotes the probability of not identifying a unit. Suppose that every unit has the same probability, then the observed sample size can be written as $n = \sum_{i=1}^{N} I_i$. Note that $E(\sum_{i=1}^{N} I_i) = N(1 - p_0)$, so that the moment estimator $\frac{n}{1-p_0} = \widehat{N}$ arises. Therefore, it is simple to estimate the size of a population by the Horvitz-Thompson estimator as:

$$\widehat{N}_{HT} = \frac{n}{1 - p_0}. \tag{2.7}$$

Note that p_0 will often be unknown. Hence it is necessary to concentrate on the following ways to estimate p_0.

2.3.2 Maximum likelihood estimator under Poisson distribution

Suppose that the capture-recapture count X can be modelled as a Poisson distribution with density

$$p_x = \frac{\exp(-\lambda)\lambda^x}{x!}.$$

The maximum likelihood estimation of zero-truncated count data has been widely used as it generates small variance. From the previous section, the zero-truncated Poisson distribution is defined as

$$p_0^+ = \frac{\exp(-\lambda)\lambda^x}{x!\,\{1 - \exp(-\lambda)\}},$$

where, $p_0 = \exp(-\lambda)$. Hence, the size of target population can be achieved by

$$\widehat{N}_{MLEPoi} = \frac{n}{1 - \exp\left(-\widehat{\lambda}_{MLEPoi}\right)}, \tag{2.8}$$

where $\widehat{\lambda}_{MLEPoi}$ is a parameter estimated from the zero-truncated Poisson distribution. One of the well-known approaches to deal with the missing data is the Expectation-Maximisation algorithm or the EM algorithm (Dempster et al., 1977), that comprises two steps procedure, in the first E-step, the missing data is imputed and in the second M-step, the parameters are estimated. Therefore, the EM algorithm is applied to the zero-truncated Poisson distribution count data as follows:

The E-step requires $Q(\lambda) = E(\log L_c | f_1, f_2, ...f_m)$. The expected and associated complete data log-likelihood can be derived as follows.

The complete data log likelihood is

$$\begin{aligned} Q(\lambda) &= \log\left\{\prod_{x=0}^{m} p_x^{f_x}\right\} \\ &= f_0 \log p_0 + \sum_{x=1}^{m} f_x \log p_x, \end{aligned} \tag{2.9}$$

and replacing f_0 by $E(f_0 | f_1, f_2, f_3, ..., f_m; \lambda)$, we have

$$Q(\lambda) = E(f_0 | f_1, f_2, f_3, ..., f_m; \lambda) \log \frac{e^{-\lambda}\lambda^0}{0!} + \sum_{x=1}^{m} f_x \log \frac{e^{-\lambda}\lambda^x}{x!}, \tag{2.10}$$

where $E(\log L_c | f_1, f_2, ...f_m)$ denotes by $\widehat{f}_0$, and can be calculated as:

$$\begin{aligned} \widehat{f}_0 &= N p_0 \\ &= (n + \widehat{f}_0) \exp(-\lambda) \\ \widehat{f}_0 &= \frac{n \exp(-\lambda)}{1 - \exp(-\lambda)}. \end{aligned} \tag{2.11}$$

Substituting $\widehat{f}_0$ into (2.9), we achieve the expected complete log-likelihood which is of the form

$$Q(\lambda) = -\lambda \widehat{f}_0 - \lambda n + \log \lambda \sum_{x=1}^{m} x f x - \sum_{x=1}^{m} f_x \log(x!) \tag{2.12}$$

In the M-step, equating $\frac{\partial Q(\lambda)}{\partial \lambda}$ to zero, we obtain

$$\widehat{\lambda}_{MLEPoi} = \frac{\sum_{x=0}^{m} x f_x}{n + \widehat{f}_0}. \tag{2.13}$$

In summary, Böhning et al. (2005) provided the algorithm of maximum likelihood of zero-truncated Poisson distribution count data by using the EM algorithm technique as follows:

Step 0: Setting $k = 0$ and selecting some suitable initial values for $\widehat{\lambda}^{(k)}$, $\widehat{\lambda}^{(0)} = \frac{\sum_{x=1}^{m} x f_x}{n}$, has been suggested as an initial value.

Step 1: Compute

$$\widehat{f}^{(k+1)} = \frac{n \exp(-\widehat{\lambda}^{(k)})}{1 - \exp(-\widehat{\lambda}^{(k)})}$$

and

$$\widehat{N}^{(k+1)} = \frac{n}{1 - \exp(-\widehat{\lambda}^{(k)})}.$$

Step 2: Use the complete frequency table as Table 2.12

Table 2.12: The frequency distribution

x	0	1	2	3	...	m
f_x	$\widehat{f}_0^{(k+1)}$	f_1	f_2	f_3	...	f_m

to compute the estimator $\widehat{\lambda}^{(k+1)}$,

$$\widehat{\lambda}^{(k+1)} = \frac{0\widehat{f}_0 + 1f_1 + 2f_2 + ... + mf_m}{\widehat{N}^{(k+1)}},$$

and set $k = k + 1$ and repeat from step 1 until parameter estimates converge.

The variance of the population estimator under maximum likelihood method of estimation can be derived as:

$$\widehat{Var}(\widehat{N}_{MLEPoi}) = \frac{\widehat{N}_{MLEPoi}}{\left\{ \exp\left(\frac{\sum_{x=1}^{m} x f_x}{\widehat{N}_{MLEPoi}} \right) - \frac{\sum_{x=1}^{m} x f_x}{\widehat{N}_{MLEPoi}} - 1 \right\}}. \tag{2.14}$$

(see Chao and Lee, 1992)

2.3.3 Turing's estimator

Let f_x be the number of individuals identified exactly x times and m denote the largest observed count, so that the total number of identifications is given as

$$\sum_{x=1}^{m} x f_x = S.$$

The application of the Turing estimator can be used in a homogeneous Poisson distribution. Then in terms of a homogeneous Poisson distribution with parameter λ, we have

$$p_0 \quad = \quad \exp(-\lambda) = \frac{\lambda \exp(-\lambda)}{\lambda} = \frac{p_1}{E(X)} = \frac{E(f_1)/N}{E(S)/N} = \frac{E(f_1)}{E(S)}, \tag{2.15}$$

where $p_1 = \lambda \exp(-\lambda)$, and replacing these expected values by their observed quantities we have

$$\widehat{p}_0 = \frac{f_1}{S}. \tag{2.16}$$

If we plug $\widehat{p}_0$ into the Horvitz-Thompson estimator, we achieve the Turing estimator as:

$$\widehat{N}_{Turing} = \frac{n}{1 - f_1/S}. \tag{2.17}$$

The variance for Turing estimation is derived as

$$\widehat{Var}(\widehat{N}_{Turing}) = \frac{n\frac{f_1}{S}}{(1+\frac{f_1}{S})^2} + \frac{n^2}{(1+\frac{f_1}{S})^4}\left[\frac{f_1(1-\frac{f_1}{\widehat{N}})}{S^2} + \frac{f_1^2}{S^3}\right], \tag{2.18}$$

(see Lerdsuwansri, 2012).

The benefits of the Turing estimator are that it is easy to calculate, its value can be obtained in a straightforward way, and there is no need for an iterative procedure.

2.3.4 The McKendrick estimator

The McKendrick estimator, which was proposed by McKendrick (1925), is the estimator of population size based on the homogeneous Poisson distribution with parameter λ. He considered the example data of a cholera epidemic in India for which the population size was known. However, he found evidence of a lack of fit due to the enormous number of zero counts. McKendrick ignored all observed zeros and tried to fit a zero-truncated Poisson to an identified household affected by the epidemic. Let

$$S_1 = 0f_0 + 1f_1 + 2f_2 + ... + mf_m = \sum_{x=0}^{m} x f_x = \sum_{x=1}^{m} x f_x,$$

$$S_2 = 0^2 f_0 + 1^2 f_1 + 2^2 f_2 + ... + m^2 f_m = \sum_{x=0}^{m} x^2 f_x = \sum_{x=1}^{m} x^2 f_x.$$

An expected value and variance are computed under the Poisson distribution leading to

$$E(X) = Var(X) = \lambda,$$

so that the expected value can be written by the moment estimator as

$$E(X) = \lambda = \frac{E(S_1)}{N}, \tag{2.19}$$

and we have the variance $Var(X) = E(X^2) - [E(X)]^2$, which can be rewritten as

$$\begin{aligned} Var(X) + [E(X)]^2 &= E(X^2) \\ \lambda + \lambda^2 &= E(X^2) = \frac{E(S_2)}{N}. \end{aligned} \tag{2.20}$$

Hence, using simple arithmetic

$$\begin{aligned} \lambda &= \frac{E(S_1)}{N} \\ \lambda + \lambda^2 &= \frac{E(S_2)}{N} \\ \lambda(1+\lambda) &= \frac{E(S_2)}{N} \\ \frac{E(S_1)}{N}\left(1 + \frac{E(S_1)}{N}\right) &= \frac{E(S_2)}{N} \\ \left(1 + \frac{E(S_1)}{N}\right) &= \frac{E(S_2)}{N}\frac{N}{E(S_1)} \\ \frac{E(S_1)}{N} &= \frac{E(S_2)}{E(S_1)} - 1 \\ &= \frac{E(S_2) - E(S_1)}{E(S_1)} \\ N &= \frac{E(S_1^2)}{E(S_2) - E(S_1)}. \end{aligned} \tag{2.21}$$

Hence, using McKendrick's estimator arise by replacing moments by sample moments,

$$\widehat{N}_{McK} = \frac{S_1^2}{S_2 - S_1}. \tag{2.22}$$

By solving equations, McKendrick's estimator for λ

$$\lambda = \frac{E(S_1)}{E(S_1^2)/(E(S_2) - E(S_1))} = \frac{E(S_2) - E(S_1)}{E(S_1)} = \frac{E(S_2)}{E(S_1)} - 1, \tag{2.23}$$

and replacing moments by sample moments

$$\widehat{\lambda} = \frac{S_2}{S_1} - 1. \tag{2.24}$$

Moreover, Viwatwongkasem et al. (2008) pointed out the relation between McKendrick's estimator and Horvitz-Thompson estimator by using the Taylor's series theorem that

$$e^{-\widehat{\lambda}} = 1 - \frac{\widehat{\lambda}}{1!} + \frac{\widehat{\lambda}^2}{2!} - ...$$

Then we achieve

$$1 - \exp(-\widehat{\lambda}) \approx \widehat{\lambda},$$

The Horvitz-Thompson estimator based on $\widehat{\lambda}_{McK}$ is given as

$$\widehat{N}_{HT} = \frac{n}{1 - f(0, \exp(-\widehat{\lambda}_{McK}))} = \frac{n}{1 - \exp(-\widehat{\lambda}_{McK})} \approx \frac{n}{\widehat{\lambda}_{McK}} = \frac{nS_1}{S_2 - S_1}. \tag{2.25}$$

As we mention above, the heterogeneity is usually found in reality. Therefore, the maximum likelihood estimation under the Poisson distribution, Turing's estimator and the McKendrick's estimator might be not appropriate for capture-recapture data. The well-known problem of ignoring heterogeneity is underestimating the population size. Alternative estimators have been studied to deal with this problem with the expectation that they provide more realistic estimates using capture-recapture data.

2.3.5 The Zelterman estimator

Zelterman (1988) suggested an estimator under a truncated Poisson distribution. This is a well-known robust estimator under potential unobserved heterogeneity (Navaratna et al., 2008; Vilas and Böhning, 2008; Böhning, 2010). As the aim is to estimate population size, the parameter estimate of λ is required. Although a probability of the Poisson distribution between variables is invalid, it can be assumed to hold for small ranges of the count variable such as from x to $x+1$. Therefore, it can be used for the neighbouring frequencies f_x and f_{x+1} to estimate a parameter λ. Indeed, the zero-truncated Poisson probability can be calculated by

$$Po_+(x;\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!\,\{1 - \exp(\lambda)\}}. \tag{2.26}$$

It can then easily be seen that the ratio of the zero-truncated Poisson distribution is

$$\frac{Po_+(x+1;\lambda)}{Po_+(x;\lambda)} = \frac{\exp(-\lambda)\lambda^{x+1}/(x+1)!\,\{1 - \exp(-\lambda)\}}{\exp(-\lambda)\lambda^x/x!\,\{1 - \exp(-\lambda)\}} = \frac{\lambda}{(x+1)},$$

and the ratio of the untruncated Poisson distribution is

$$\frac{Po(x+1;\lambda)}{Po(x;\lambda)} = \frac{\exp(-\lambda)\lambda^{x+1}/(x+1)!}{\exp(-\lambda)\lambda^{x}/x!} = \frac{\lambda}{(x+1)}.$$

These are identical, therefore, the parameter λ can be written as

$$\lambda = \frac{(x+1)Po(x+1;\lambda)}{Po(x;\lambda)} = \frac{(x+1)Po_{+}(x+1;\lambda)}{Po_{+}(x;\lambda)}. \tag{2.27}$$

An estimator for the parameter λ can be achieved by replacing $Po_{+}(x;\lambda)$ with the empirical relative frequency $\frac{f_x}{n}$ leading to

$$\widehat{\lambda}_x = \frac{(x+1)f_{x+1}/n}{f_x/n} = \frac{(x+1)f_{x+1}}{f_x}, \tag{2.28}$$

since n cancels out. Zelterman suggested using $\widehat{\lambda}_1 = \frac{2f_2}{f_1}$. As Kuhnert and Böhning (2009) pointed out that there are two reasons for choosing $\widehat{\lambda}_1$ in the Zelterman parameter. Firstly, the majority of frequency count for capture-recapture studies are usually represented in terms of ones and twos (f_1 andf_2 are used), and count data greater than two have no effect on this estimator. Secondly, $\widehat{\lambda}_1$ is the closest neighbour of the target point of estimation f_0. The Zelterman estimator of population size is ultimately provided as

$$\widehat{N}_{Zel} = \frac{n}{1-\exp\left(-\frac{2f_2}{f_1}\right)}. \tag{2.29}$$

Zelterman's estimator has been widely used since it is easy to understand, and it is a robust estimator because it uses only the first and second order of frequencies. However, it might be not a good estimator for long tail count data (Lanumteang, 2010). The population size estimate may overestimate with a large variance (Vilas and Böhning, 2008). The variance of the Zelterman Estimator was given as

$$\widehat{Var}(\widehat{N}_{Zel}) = nG(\widehat{\lambda})\left[1 + G(\widehat{\lambda})\widehat{\lambda}^2\left(\frac{1}{f_1+f_2}\right)\right], \tag{2.30}$$

where $G(\widehat{\lambda}) = \dfrac{\exp(-\widehat{\lambda})}{\{1-\exp(-\widehat{\lambda})\}^2}$ and $\widehat{\lambda} = \dfrac{2f_2}{f_1}$ (see Böhning, 2008a).

2.3.6 The extension to Mantel-Haenszel estimator

Mantel-Haenszel's estimator is an extension of the Zelterman estimator under the truncated Poisson estimator, which is proposed by Wannasirikul (2005). The main idea of this extension is that weight is added into the truncated Poisson estimator of Zelterman. This means that the Zelterman estimator of parameter λ is $\widehat{\lambda} = \dfrac{(x+1)f_{x+1}}{f_x}$, and adding

weight to that estimator results in the Mantel-Haenzel's estimator. It is defined as

$$\widehat{\lambda}_{MH} = \frac{\sum_{x=1}^{m-1} W_x(x+1)\frac{f_x+1}{f_x}}{\sum_{x=1}^{m-1} W_x}. \tag{2.31}$$

Using $W_x = f_x$, that means using the frequency as weight for each class x, we have

$$\begin{aligned}
\widehat{\lambda}_{MH} &= \frac{\sum_{x=1}^{m-1} f_x(x+1)\frac{f_x+1}{f_x}}{\sum_{x=1}^{m-1} f_x} \\
&= \frac{\sum_{x=1}^{m-1}(x+1)f_{x+1}}{\sum_{x=1}^{m-1} f_x} \\
&= \frac{2f_2 + 3f_3 + 4f_4 + ... + mf_m}{f_1 + f_2 + f + 3 + ... + f_{m-1}} \\
&= \frac{S - f_1}{n - f_m},
\end{aligned} \tag{2.32}$$

where $S = \sum_{x=1}^{m} x f_x$. Note that the attraction of the Mantel-Haenszel estimators is that it takes sums before ratios and allows for more information on frequency counts to be included. The population size estimator according to the Horvitz-Thompson method is given as:

$$\widehat{N}_{MH} = \frac{n}{1 - \left\{\exp\left(-\frac{S - f_1}{n - f_m}\right)\right\}}. \tag{2.33}$$

2.3.7 Chao's estimator

Chao (1987, 1989) introduced the alternative estimator of population size under unobserved heterogeneity of the Poisson parameter. This means that the count probabilities can be considered as a mixed Poisson model with arbitrary mixing density $g(\lambda)$

$$p_x = \int_0^{\infty} \frac{\exp(-\lambda)\lambda^x}{x!} g(\lambda) d\lambda, x = 0, 1, 2, ... \tag{2.34}$$

The Cauchy-Schwarz inequality of any two random variables X and Y, states that

$$[E(XY)]^2 \leq E(X^2)E(Y^2). \tag{2.35}$$

Suppose that $g(\lambda)$ denotes arbitrary densities on parameter λ, and let X and Y be specific to our situation. In fact, we take $X = (e^{-\lambda}\lambda^2)^{\frac{1}{2}}$ and $Y = (e^{-\lambda})^{\frac{1}{2}}$. then as a consequence,

$$E(XY) = \int_0^\infty e^{-\lambda}\lambda g(\lambda)d\lambda,$$

$$E(X^2) = \int_0^\infty e^{-\lambda}\lambda^2 g(\lambda)d\lambda,$$

$$E(Y^2) = \int_0^\infty e^{-\lambda} g(\lambda)d\lambda.$$

Then any mixed Poisson probability leads to

$$\left(\int_0^\infty e^{-\lambda}\lambda g(\lambda)d\lambda\right)^2 \leq \left(\int_0^\infty e^{-\lambda} g(\lambda)d\lambda\right)\left(\int_0^\infty e^{-\lambda}\lambda^2 g(\lambda)d\lambda\right).$$

Since $p_0 = \int_0^\infty e^{-\lambda} g(\lambda)d\lambda$, $p_1 = \int_0^\infty e^{-\lambda}\lambda g(\lambda)d\lambda$, and $p_2 = \int_0^\infty \frac{e^{-\lambda}\lambda^2}{2!} g(\lambda)d\lambda$. It follows that

$$p_1^2 \leq p_0(2p_2). \tag{2.36}$$

Hence a lower bound for p_0 is achieved as

$$\frac{p_1^2}{2p_2} \leq p_0. \tag{2.37}$$

Multiplying probabilities with N leads to

$$\frac{(Np_1)^2}{2(Np_2)} \leq Np_0.$$

Replacing Np_1 and Np_2 by the observed frequencies f_1 and f_2 leads to the lower bound estimator $\frac{f_1^2}{2f_2}$, then

$$\widehat{f_0} = \frac{f_1^2}{2f_2}. \tag{2.38}$$

Moreover, this leads to a lower bound for estimating the population size. Chao's lower bound estimator of the population size N is

$$\widehat{N}_{Chao} = n + \frac{f_1^2}{2f_2}. \tag{2.39}$$

Note that only f_1 and f_2 are used in Chao's lower bound estimator. Chao's estimator represents lower bound estimates, if heterogeneity based on Poisson is present (see Böhning and Vilas, 2008; Böhning, 2010; Böhning et al., 2013b).

The variance of Chao's estimator is given as

$$\widehat{Var}(\widehat{N}_{Chao}) = \left(\frac{1}{4}\right)^2 \frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \left(\frac{1}{2}\right) \frac{f_1^2}{f_2} \tag{2.40}$$

(see Böhning, 2008a).

An extended version of the Chao estimator has been recently developed for covariate information by Böhning et al. (2013b) and based on the log-normal distribution in Chiu et al. (2014).

2.3.8 Chao and Bunge's estimator

Chao and Bunge (2002) introduced an estimator using a Poisson-Gamma distribution or a negative binomial distribution. Let θ be the repeated proportion of counts in the sample, and let X follow the negative binomial distribution. Referring to the Poisson mixture model as

$$p_x(\lambda) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} g(\lambda) d\lambda,$$

where $g(\lambda)$ is a gamma mixing distribution then the negative binomial distribution with shape parameter α and event parameter β is

$$p_{\alpha,\beta}(x) = \frac{\beta^\alpha \Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)(\beta+1)^{x+\alpha}},$$

where $x = 0, 1, 2, ...$ More precisely, let $\theta = 1 - (p_0 + p_1)$ be the repeated fraction count in the sample, hence the unobserved frequency f_0 can be calculated in the following way:

$$\begin{aligned}
\theta &= 1 - (p_0 + p_1) \\
p_0 + p_1 &= 1 - \theta \\
E(f_0) + E(f_1) &= N(1-\theta) \\
&= N - N\theta \\
&= N\theta(\frac{1}{\theta} - 1) \\
&= (\frac{1}{\theta} - 1) \sum_{x=2}^{m} E(f_x) \\
E(f_0) &= (\frac{1}{\theta} - 1) \sum_{x=2}^{m} E(f_x) - E(f_1) \\
\widehat{f_0} &= (\frac{1}{\theta} - 1) \sum_{x=2}^{m} f_x - f_1,
\end{aligned} \tag{2.41}$$

where $N\theta = \sum_{x=2}^{m} E(f_x)$. In addition, the expected proportion of individual counts in the sample can be calculated from the negative binomial distribution as

$$p_0 = \frac{\beta^\alpha \Gamma(\alpha)}{\Gamma(1)(\beta+1)^\alpha \Gamma(\alpha)} = \frac{\beta^\alpha \Gamma(\alpha)}{(\beta+1)^\alpha \Gamma(\alpha)} = \left(\frac{\beta}{\beta+1}\right)^\alpha, \tag{2.42}$$

and

$$\begin{aligned} p_1 &= \frac{\beta^\alpha \Gamma(1+\alpha)}{\Gamma(2)\Gamma(\alpha)(\beta+1)^{1+\alpha}} = \frac{\beta^\alpha \alpha \Gamma(\alpha)}{\Gamma(2)\Gamma(\alpha)(\beta+1)^{1+\alpha}} \\ &= \frac{\alpha\beta^\alpha}{(\beta+1)^(1+\alpha)} = \left(\frac{\beta}{\beta+1}\right)^\alpha \frac{\alpha}{\beta+1}, \end{aligned} \tag{2.43}$$

hence,

$$\begin{aligned} p_0 + p_1 &= \left(\frac{\beta}{\beta+1}\right)^\alpha + \left(\frac{\beta}{\beta+1}\right)^\alpha \frac{\alpha}{\beta+1} \\ &= \left(\frac{\beta}{\beta+1}\right)^\alpha \left(1 + \frac{\alpha}{\beta+1}\right) \\ &= \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{\alpha+\beta+1}{\beta+1}\right), \end{aligned} \tag{2.44}$$

given $\theta = 1 - (p_0 + p_1)$, then

$$\theta = 1 - \left[\left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{\alpha+\beta+1}{\beta+1}\right)\right].$$

Since $E(X) = \frac{\alpha}{\beta}$ and $Var(X) = \left(\frac{\beta+1}{\beta}\right)\left(\frac{\alpha}{\beta}\right) = \frac{\alpha(\beta+1)}{\beta^2}$, so that

$$\begin{aligned} E(X^2) &= Var(X) + E[(X)]^2 \\ &= \frac{\alpha(\beta+1)}{\beta^2} + \frac{\alpha^2}{\beta^2} \\ &= \frac{\alpha}{\beta}\left(\frac{\alpha+\beta+1}{\beta}\right) \end{aligned} \tag{2.45}$$

$$\begin{aligned} p_1 \frac{E(X^2)}{[E(X)]^2} &= p_1 \frac{\frac{\alpha}{\beta}\left(\frac{\alpha+\beta+1}{\beta}\right)}{\frac{\alpha^2}{\beta^2}} \\ &= p_1 \frac{\alpha+\beta+1}{\alpha} \\ &= \left[\left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{\alpha}{\beta+1}\right)\right]\left(\frac{\alpha+\beta+1}{\alpha}\right) \\ &= \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1+\alpha+\beta}{\beta+1}\right). \end{aligned} \tag{2.46}$$

Hence,

$$p_1 \frac{E(X^2)}{[E(X)]^2} = p_0 + p_1. \tag{2.47}$$

As a result, the recaptured fraction θ can be calculated as

$$\theta = 1 - p_1 \frac{E(X^2)}{[E(X)]^2}. \tag{2.48}$$

In practice, the probability of repeated counts can be estimated by using the empirical frequencies of exactly m identifications. Let $S_1 = \sum_{x=1}^{m} x f_x$ and $S_2 = \sum_{x=1}^{m} x^2 f_x$ then,

$$\begin{aligned} E(p_1) &= \frac{E(f_1)}{N} \\ \widehat{p_1} &= \frac{f_1}{N} \\ \widehat{E(X)} &= \frac{S_1}{N} \\ \widehat{E(X^2)} &= \frac{S_2}{N} \end{aligned} \tag{2.49}$$

$$\begin{aligned} \widehat{p_1} \frac{\widehat{E(X^2)}}{[\widehat{E(X)}]^2} &= \frac{f_1}{N} \frac{S_2/N}{(S_1/N)^2} \\ &= f_1 \frac{S_2}{S_1^2}. \end{aligned} \tag{2.50}$$

Hence,

$$\widehat{\theta} = 1 - f_1 \frac{S_2}{S_1^2} \tag{2.51}$$

The Chao-Bunge (C-B) estimator can be calculated as

$$\begin{aligned}
\widehat{N}_{C-B} &= \widehat{f}_0 + n \\
&= (\frac{1}{\widehat{\theta}} - 1)\sum_{x=2}^{m} f_x - f_1 + n \\
&= (\frac{1}{\widehat{\theta}} - 1)\sum_{x=2}^{m} f_x - f_1 + (f_1 + f_2 + f_3+, ..., +f_m) \\
&= (\frac{1}{\widehat{\theta}} - 1)\sum_{x=2}^{m} f_x + (f_2 + f_3+, ..., +f_m) \\
&= (\frac{1}{\widehat{\theta}} - 1)\sum_{x=2}^{m} f_x + \sum_{i=2}^{m} f_x \\
&= \sum_{x=2}^{m} \frac{f_x}{\widehat{\theta}} \\
\widehat{N}_{C-B} &= \frac{\sum_{x=2}^{m} f_x}{1 - \frac{f_1 S_2}{S_1^2}}. \qquad (2.52)
\end{aligned}$$

A limitation of this estimator is that it requires a large fraction of overlap between count occasions. If the overlap is small, the population size estimate may be negative.

2.4 Some applications of capture-recapture modelling

Capture-recapture approaches are applied in many areas with the aim of estimating the population size. Some examples using the population size estimators above are provided in this section.

2.4.1 Illegal immigrants study in the Netherlands

Van der Heijden et al. (2003) studied the number of illegal immigrants in the Netherlands using the truncated Poisson regression model based on the number of times each illegal immigrant was recorded by police. These records come from four cities: Rotterdam, The Hague, Amsterdam and Utrecht in 1995. The data are summarised in Table 2.13.

Table 2.13: The number of illegal immigrant population

x	1	2	3	4	5	6	n
f_x	4,074	257	44	14	2	1	4,392

We now apply the discussed estimators to estimate the size of the illegal immigrant population in the Netherlands. It is clearly seen that Zelterman's estimator gives the largest population size, while the smallest population size is given by Chao and Bunge's estimator as shown in Table 2.14.

Table 2.14: Estimated size of the illegal immigrant population

Estimator	Total of population	unobserved population
Poisson		
McKendrick	22,602	18,210
Turing	29,313	24,921
MLEPoi	27,049	22,657
Heterogeneity		
Zelterman	37,054	32,662
Mantel-Haenzel	29,116	24,724
Chao	36,683	32,291
Chao and Bunge	20,537	16,145

2.4.2 Methamphetamine users in Bangkok, Thailand

Methamphetamine users in Thailand were recorded by the Office of the Narcotics Control Board (ONCB) from 61 health treatment centres in Bangkok, Thailand (Böhning et al., 2004). Here interest is in estimating the number of hidden methamphetamine users.

Table 2.15: The number of methamphetamine users in Bangkok

x	1	2	3	4	5	6	7	8	9	10	11	12	n
f_x	3,114	163	23	20	9	3	3	3	4	3	0	1	3,346

It can be seen from Table 2.16 that McKendrick's estimator gives the lowest number of methamphetamine users in Bangkok, whereas Zelterman's estimator provided the highest number of drug users in Bangkok.

Table 2.16: Estimated size of methamphetamine users in Bangkok

Estimator	Total of population	unobserved population
Poisson		
McKendrick	7,279	3,933
Turing	19,110	15,764
MLEPoi	15,313	19,967
Heterogeneity		
Zelterman	33,664	30,318
Mantel-Haenzel	18,661	15,315
Chao	33,091	29,745
Chao and Bunge	18,219	14,873

2.4.3 Domestic violence data

Incidents of domestic violence in the Netherlands between 2005 and 2007 were considered using data taken from the Dutch police register system. The frequency of offenders is given in Table 2.17

Table 2.17: The number of domestic violence data

x	1	2	3	4	5	6	7	8	9	n
f_x	43,117	5,411	1,035	279	76	22	11	5	1	49,957

It is clear from Table 2.18 that McKendrick's estimator gives the lowest number of illegal immigrants in the Netherlands, whereas Chao and Bunge's estimator result in the highest number of domestic violence.

Table 2.18: Estimated incidents of family violence

Estimator	Total of population	Unobserved population
Poisson		
McKendrick	147,909	97,952
Turing	187,321	137,364
MLEPoi	175,360	125,403
Heterogeneity		
Zelterman	225,061	175,104
Mantel-Haenzel	185,435	135,478
Chao	221,744	171,787
Chao and Bunge	349,501	299,544

2.4.4 Three sources of data

Héraud-Bousquet et al. (2012) estimated the number of 216 new HIV diagnoses in children under 13 years old from three sources: EPF be ANRS-French Perinatal Cohort, DOVIH denotes Mandatory HIV case reporting and LaboVIH : Laboratory surveillance of HIV testing activity, in mainland France from January 2003 to December 2006. It can be seen three sources of data was presented previously in Example 4 in Section 2 of this chapter. It can be seen in Table 2.19 that McKendrick's estimator provides the largest number of new cases of HIV in France, whereas Chao and Bunge's estimator result in the lowest.

Table 2.19: Estimated size of family violence

Estimator	Total of population	unobserved population
Poisson		
McKendrick	381	165
Turing	334	118
MLEPoi	345	129
Heterogeneity		
Zelterman	319	103
Mantel-Haenzel	314	98
Chao	322	106
Chao and Bunge	290	74

From the applications, it can be seen a big different result in estimated population size when we use different methods. Therefore, it might be stated that capture-recapture models play the key factor for estimating a target population size in capture-recapture study.

As we mention in the previous sections, the basis of counting distributions for capture-recapture studies are the Poisson and the binomial models. However, the heterogeneity in detection of capture probabilities have been widely discussed as a violation of homogeneous modelling (Link, 2003). It is well known that a misspecification of the model in capture-recapture studies might lead to biased population size estimators (Huggins and Hwang, 2007; Thompson, 2013). As a consequence, a variety of mixture models were considered for estimating population sizes. The model selection becomes an important part for constructing population size estimators. In the next section, a graphical technique is discussed as a device to investigate count data modelling, and is particularly useful for zero-truncated count data in capture-recapture modelling.

2.5 The graphical device of the ratio plot for identifying a distribution

The foundation for identifying a distribution in statistics is graphical statistics because this approach is simple and quick. For example, Dubey (1966); Ord (1967) proposed a graphical technique for detecting discrete distributions (i.e. the binomial distribution, the Poisson distribution and the Pascal distribution). A graphical testing is valuable for gaining prior information concerning the frequency distribution in statistical analysis. In capture-recapture study, the graphical device, named the ratio plot was developed for investigating the homogeneous Poisson and heterogeneous models of capture-recapture data by Böhning et al. (2013a). If we know the distribution, its properties can be used for appropriately developing estimators.

2.5.1 The ratio plot of the zero-truncated distribution

A graphical device for identifying a distribution in capture-recapture study can be extended to the zero-truncated distribution. The ratio plot for untruncated probability is expressed as

$$r_x = (x+1)\frac{p_{x+1}}{p_x}. \tag{2.53}$$

In capture-recapture studies the observed sample frequencies $f_1, f_2, ..., f_m$ arise from the zero-truncated distribution since zero counts are truncated. The zero-truncated distribution in capture-recapture studies is defined as

$$p_+(x) = \frac{p_x}{1-p_0}.$$

Then, the ratio plot for the zero-truncated probability is given by

$$r_x = \frac{(x+1)p_{x+1}/(1-p_0)}{p_x/(1-p_0)}.$$

However, the ratio plot for the truncated and untruncated distributions is identical. That is

$$r_x = \underbrace{\frac{(x+1)p_{x+1}}{p_x}}_{Untruncated} = \underbrace{\frac{(x+1)p_{x+1}/(1-p_0)}{p_x/(1-p_0)}}_{Zero-truncated}. \tag{2.54}$$

Hence $r_x^* = (x+1)\frac{f_{x+1}}{f_x}$ can be used to check for both zero-truncated and completed count distributions.

2.5.2 The ratio plot of Poisson distribution

There are many methods of exploring the distribution of count data, one of which is a ratio plot. In this section the ratio plot for the Poisson distribution is discussed. As introduced previously, the Poisson distribution is provided by

$$p_x = \frac{\exp(-\lambda)\lambda^x}{x!}. \tag{2.55}$$

According to the Poisson distribution, if the ratio of neighbouring Poisson is computed, the ratio plot must be *constant*. Indeed, the ratio plot can be calculated as follows

$$\frac{p_{x+1}}{p_x} = \frac{\exp(-\lambda)\lambda^{x+1}/(x+1)!}{\exp(-\lambda)\lambda^x/x!} = \frac{1}{x+1}\lambda. \tag{2.56}$$

Hence the ratio plot of a Poisson, r_x, is the ratio of neighbouring Poisson probabilities multiplied by the value of the larger neighbour count and becomes a constant. It is represented as

$$r_x = (x+1)\frac{p_{x+1}}{p_x} = \lambda, \tag{2.57}$$

and shows that the ratio r_x is constant with varying count x. In practice, the ratio plot is estimated by

$$r_x^* = (x+1)\frac{\widehat{p}_{x+1}N}{\widehat{p}_x N} = (x+1)\frac{f_{x+1}/N}{f_x/N} = (x+1)\frac{f_{x+1}}{f_x}, \tag{2.58}$$

where f_x is the frequency of count x. See also Figure 2.2.

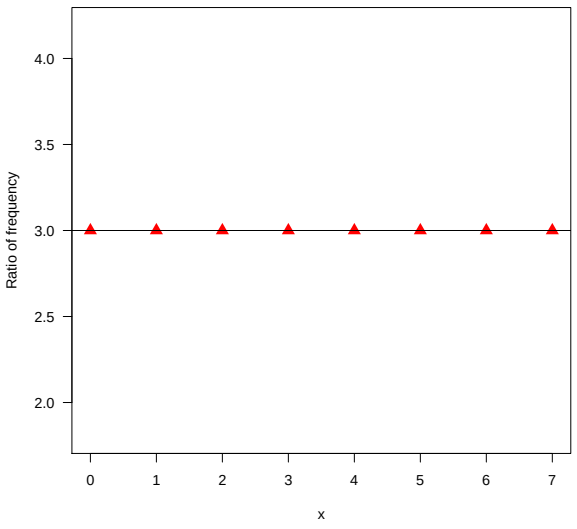

Figure 2.2: Ratio plot and regression line for 100,000 simulated data from a Poisson distribution with $\lambda = 3$

The graph $r_x^* = (x+1)\frac{f_{x+1}}{f_x}$ against x can be used to consider the Poisson distribution. If the ratio plot shows a pattern of *a horizontal line*, this can be taken an indication of a *Poisson distribution*. However, the occurrence of Poisson distribution is rare in practice, since population heterogeneity affects the inclusion of units in a capture-recapture study. This heterogeneity usually invalidates the assumption that $X_1, X_2, ..., X_N$ are identically distributed, and the estimation of population size will be biased if the heterogeneity is ignored (Van der Heijden et al., 2003; Hwang and Huggins, 2005; McCrea and Morgan, 2014). Typically, the well-known occurrence of heterogeneity in the Poisson parameter results in over-dispersion, where the variance is greater than the mean. This problem is a violation of the Poisson property. Thus, the modelling of heterogeneity seems to be more appropriate than the Poisson model.

2.5.3 The ratio plot of heterogeneous model

In reality, the target population may have many subpopulations leading to the capture probabilities depending on the individual heterogeneity. Chao (1987) suggested a mixed Poisson model with arbitrary mixing density $g(\lambda)$ to model the form of heterogeneity in capture-recapture as:

$$p_x = \int_0^\infty \frac{\exp(-\lambda)\lambda^x}{x!} g(\lambda) d\lambda, \tag{2.59}$$

where $x = 0, 1, 2, ...$ In the case of heterogeneity, Böhning et al. (2013a) stated that the ratio should exhibit a monotone increasing pattern under arbitrary mixed Poisson as follows:

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq ... \leq \frac{(x+1)p_{x+1}}{p_x} \leq ... \tag{2.60}$$

The simple monotone structure heterogeneity is a straight line with positive slope, which was defined as *structured heterogeneity*. The ratio plot exhibits structured heterogeneity if

$$r_x = \beta_0 + \beta_1 x, \tag{2.61}$$

where $\beta_1 > 0$. This evidence is indicative of *structured heterogeneity*, and if $\beta_1 = 0$, the model reduces to the original Poisson model (see Böhning et al., 2013a). In the capture-recapture literature, the ratio plot of the negative binomial shows the structure heterogeneity for capture-recapture data as we can see in Rocchetti et al. (2011); Lanumteang (2010); Lerdsuwansri (2012).

2.5.4 The ratio plot for the negative binomial

Considering more general distributions than the original Poisson distribution is an interesting approach to account for heterogeneous Poisson parameters. This approach considers the unobserved heterogeneity problem in which random variation in Poisson

parameters cannot be explained by the observed data. The negative binomial distribution can be used as an alternative to the homogeneous Poisson distribution. In practice, over-dispersion often occurs in a heterogeneous population. A mixed model approach to model over-dispersion starts with a standard Poisson model and adds a random effect to account for the unobserved heterogeneity, which is distributed according to a gamma distribution. Since the negative binomial distribution has one parameter more than the Poisson, the second parameter can be used to indicate the variance independently from the mean.

1) The Poisson-Gamma as the negative binomial distribution

The gamma distribution is selected as the arbitrary mixing density $g(\lambda)$ in a mixed Poisson model. The negative binomial distribution arises from the combination of a gamma distribution and a Poisson family distributions. Suppose that p_x is the Poisson-gamma mixture, let $f(x;\lambda) = Po(X;\lambda)$ and $g(\lambda) = Gam(\lambda;\theta,k) = \dfrac{\theta^{-k}\lambda^{k-1}e^{\frac{-\lambda}{\theta}}}{\Gamma(k)}$. This means that the variable λ has a gamma distribution with scale parameter θ and shape parameter k. According to the parametric mixture model above, the joint density of X and λ is

$$\begin{aligned}
p_x &= \int_0^\infty f(x;\lambda)g(\lambda)d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\frac{\theta^{-k}\lambda^{k-1}e^{\frac{-\lambda}{\theta}}}{\Gamma(k)}d\lambda \\
&= \frac{\theta^{-k}}{\Gamma(x+1)\Gamma(k)}\int_0^\infty e^{-\lambda(1+\frac{1}{\theta})}\lambda^{(x+k)-1}d\lambda \\
&= \frac{\theta^{-k}}{\Gamma(x+1)\Gamma(k)}\frac{\Gamma(x+k)}{(1+\frac{1}{\theta})^{x+k}}\int_0^\infty \frac{(1+\frac{1}{\theta})^{x+k}}{\Gamma(x+k)}e^{-\lambda(1+\frac{1}{\theta})}\lambda^{(x+k)-1}d\lambda \\
&= \frac{\theta^{-k}}{\Gamma(x+1)\Gamma(k)}\frac{\Gamma(x+k)}{(1+\frac{1}{\theta})^{x+k}} \\
&= \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}\left(\frac{\theta}{\theta+1}\right)^x\frac{\theta^{-k}}{(1+\frac{1}{\theta})^k} \\
&= \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}\left(\frac{\theta}{\theta+1}\right)^x\frac{1}{\theta^k\frac{(\theta+1)^k}{\theta^k}} \\
&= \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}\left(\frac{\theta}{\theta+1}\right)^x\left(\frac{1}{\theta+1}\right)^k, \qquad (2.62)
\end{aligned}$$

where $\int_0^\infty \frac{(1+\frac{1}{\theta})^{x+k}}{\Gamma(x+k)}e^{-\lambda(1+\frac{1}{\theta})}\lambda^{(x+k)-1}d\lambda = 1$. Hence, the mixed Poisson-Gamma with shape parameter k and scale parameter $\theta = \dfrac{1-p}{p}$ can be rewritten as

$$p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}p^k(1-p)^x, \qquad (2.63)$$

which is the probability density function of a negative binomial. In addition, this mixed

distribution is one of the structured heterogeneity models for capture-recapture modelling. The graphical technique is now applied to consider whether a negative binomial distribution is appropriate for the capture-recapture data. According to the negative binomial distribution we have

$$p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}(1-p)^x p^k, \tag{2.64}$$

for random count X with $x = 0, 1, 2, ...$, $p \in (0, 1)$ and $k > 0$. The expected value and variance of X are defined as:

$$E(X) = \frac{k(1-p)}{p} = \mu \tag{2.65}$$

and

$$Var(X) = \frac{k(1-p)}{p^2} = \mu + \frac{1}{k}\mu^2. \tag{2.66}$$

The negative binomial distribution has the variance greater than the mean, so that we can call this density an over-dispersion model. The graphical device of the ratio plot can be applied to consider the distribution of the random variable.

2) The graphical device of the ratio plot for identifying the negative binomial distribution

The ratio of neighbouring negative binomial probabilities with parameter k and p can be calculated as follows

$$(x+1)\frac{p_{(x+1)}}{p_x} = (x+1)\frac{\frac{\Gamma(x+1+k)}{\Gamma(x+2)\Gamma(k)}(1-p)^{x+1}p^k}{\frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}(1-p)^x p^k} = (x+k)(1-p),$$

leading to

$$r_x = (x+1)\frac{p_{x+1}}{p_x} = \underbrace{(1-p)k}_{intercept} + \underbrace{(1-p)}_{slope}x = \beta_0 + \beta_1 x. \tag{2.67}$$

If the ratio plot represents a straight line with positive slop, it can be assumed that the count data follow the negative binomial distribution. An estimation of the dispersion parameter k from (2.67) is

$$\begin{aligned} (1-p)k &= \beta_0 \\ (1-p) &= \beta_1 \\ \beta_1 k &= \beta_0 \\ k &= \frac{\beta_0}{\beta_1}, \end{aligned} \tag{2.68}$$

we then have

$$\widehat{k} = \frac{\widehat{\beta_0}}{\widehat{\beta_1}}, \tag{2.69}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the first and the second coefficients from a least square estimation from a linear regression model.

3) The graphical device of the ratio plot for the geometric distribution
The special case of the negative binomial distribution when $k = 1$ is the Poisson-exponential mixture or the geometric distribution. If the ratio plot is passing through the origin, then

$$\begin{aligned} r_x &= (1-p)(1) + (1-p)x \\ &= (1-p)(x+1) \\ &= \underbrace{(1-p)}_{slope} x^{'} = \beta_1 x^{'}, \end{aligned} \tag{2.70}$$

where $x^{'} = (x+1)$. Indeed, the equivalent ratio plot of the geometric distribution can be calculated as

$$\begin{aligned} r_x^{'} = x\frac{p_x}{p_{x-1}} &= x\frac{p(1-p)^x}{p(1-p)^{x-1}} \\ &= (1-p)x \\ &= [x + (k-1)](1-p) \\ &= \underbrace{(k-1)(1-p)}_{intercept} + \underbrace{(1-p)}_{slope} x \end{aligned} \tag{2.71}$$

where the geometric distribution is given as $p_x = p(1-p)^x$. Therefore, the ratio plot passes through the origin, which is indicative of exponential mixing or the geometric distribution.

2.6 Some of limitations of negative binomial modelling

Since structured heterogeneity might be more realistic than the Poisson model for capture-recapture count data, negative binomial modelling would seem to be more appropriate. However, it has some disadvantages to modelling capture-recapture data since its dispersion parameter constraints might not make it feasible. This can be seen in the ratio plot when the intercept estimate is lower than zero. This evidence shows that $\beta_0 < 0$, leading to the estimator value of parameter $k < 0$ will lie on the boundary. Hence, it might be said that if the ratio plot provides an the intercept lower than zero, this is evidence that the negative binomial will not be appropriate to model the count data.

Example 5: Domestic violence data

To illustrate the violation of dispersion parameter k, data relating to the number of domestic violence offenders in the Netherlands in 2007 were analysed. The data shown is from the Dutch police register system and frequencies of offenders is given in Table 2.20 (see Van der Heijden et al., 2014)

Table 2.20: Frequency counts f_x of offenders in the Netherlands with exactly x offences

x	1	2	3	4	5	6+
f_x	15,169	1,957	393	99	28	16

Using the graphical approach for identifying a distribution, the ratio plot is shown in Figure 2.3.

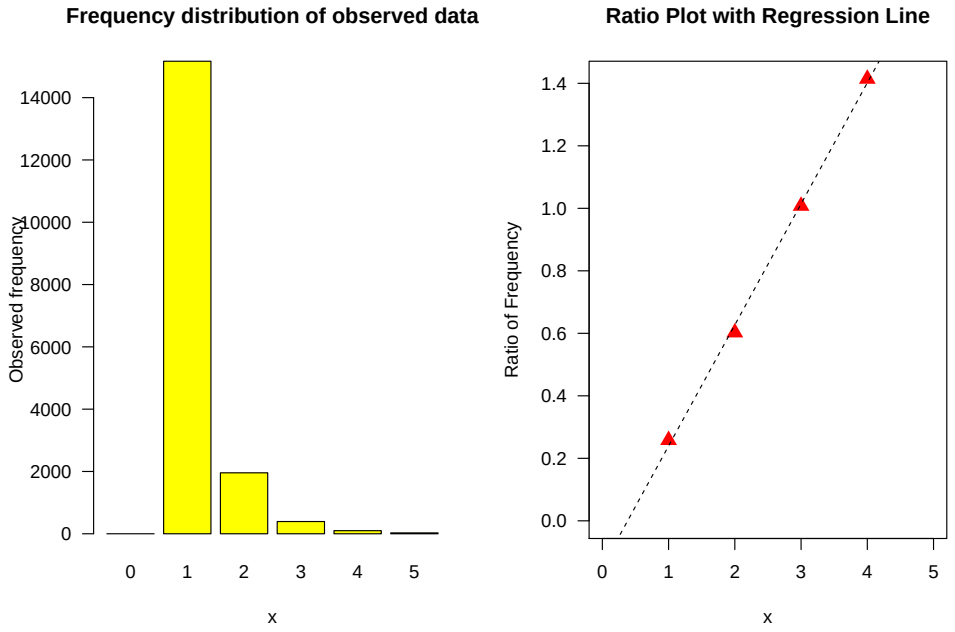

Figure 2.3: Frequency distribution and scatter plot with regression line of $r_x = (x+1)\frac{f_{x+1}}{f_x}$ versus x, x being the number of offences per offender for domestic violance offenders in the Netherlands

The ratio plot shows a straight line with a positive slope, this evidence seems to support the negative binomial distribution. Nevertheless, the ratio plot shows an intercept point lower than zero. This is a violation of the dispersion parameter (k) constraint since $\widehat{\beta}_0 < 0$ leading to $\widehat{k} < 0$. Therefore, the negative binomial is not appropriate as distribution for this set of count data. It is clear that although the negative binomial is suitable to model over-dispersion count data, it might not be appropriate in some cases of capture-recapture data because of the violation of parameter constraints. Therefore,

the evidence of a straight line with positive slope may indicate another appropriate distribution.

Chapter 3

Population Size Estimation under the Generalised Poisson Distribution

A new population size estimator, namely an estimate of the population size under maximum likelihood estimation of the zero-truncated generalised Poisson distribution (MLEGP), is proposed in this chapter. The MLEGP estimator is introduced for solving the violated dispersion parameter problem in the negative binomial model. The ratio plot is used as a device for investigating the validity of the generalised Poisson (GP) distribution as well as the zero-truncated generalised Poisson (ZTGP) distribution. Since the capture-recapture data do not provide zero counts, an Expectation-Maximisation (EM) algorithm is used to estimate the two parameters of ZTGP model. The behaviour of the new population size estimator is evaluated through simulation and real data examples

3.1 Introduction and problem formulation

The capture-recapture approach is a powerful tool for estimating the elusive population sizes in various fields. Since each individual of the target population can be repeat-counted during the study period, the Poisson distribution is used as a basic model for modelling capture-recapture count data. However, it is recognised that the count distribution for capture-recapture data often displays over-dispersion which results in a biased estimator of the target population size if not appropriately modelled (Baksh et al., 2011). The negative binomial distribution is usually used as an alternative to the Poisson distribution in the case of over-dispersion, because the negative binomial distribution adds one more parameter which can be used to adjust the variance independently of the mean. Nevertheless, the negative binomial distribution has some limitations in respect of

capture-recapture data such as the range of dispersion parameter estimation as pointed out in Chapter 2.

Alternative heterogeneity models will be researched to develop a realistic fit of capture-recapture data. Thus, the generalised Poisson distribution is a candidate since its ratio plot can also represent the heterogeneous structure (linear trend with positive slope) in cases of over-dispersion. Indeed, not only the negative binomial, but also the generalised Poisson distribution are used to cope with over-dispersion. Another advantage is that the ratio plot of the generalised Poisson distribution allows for a negative slope in case of under-dispersion which might occur in capture-recapture data . Thus, a population size estimator based upon the generalised Poisson distribution is developed.

3.2 Generalised Poisson distribution for capture-recapture data

3.2.1 Generalised Poisson distribution

The generalised Poisson (GP) distribution was introduced by Consul and Jain (1973a,b) for modelling count or frequency data. The GP model has been used to model frequency data in biological, ecological, social and marketing studies. Here, we apply it to the capture-recapture study setting. The benefit of the GP distribution is that it is more flexible than the negative binomial distribution since both over-dispersion and under-dispersion can be incorporated by the GP model. In the GP model, one more parameter α is added to the original Poisson distribution to provide a free variance parameter, and this is called the generalised Poisson distribution. More precisely, let X_i denote the number of times that individual i is observed from the elusive target population, of size N, during the study period. Hence, $i \in \{0, 1, 2, 3, ..., N\}$ and $X \in \{0, 1, 2, ...\}$. For the case of count X the probability density function of X following the GP distribution, with event parameter θ and dispersion parameter α, is defined as:

$$p_x(\theta, \alpha) = \theta(\theta + \alpha x)^{x-1} \left\{ \frac{\exp(-\theta - \alpha x)}{x!} \right\}, \tag{3.1}$$

where $x = 0, 1, 2, 3, ...$, $\theta > 0$, $\max(-1, \frac{-\theta}{4}) \leq \alpha < 1$. Then, the mean and variance of X are given by

$$E(X) = \frac{\theta}{1 - \alpha}, \tag{3.2}$$

$$Var(X) = \frac{\theta}{(1 - \alpha)^3}. \tag{3.3}$$

It is remarkable that when $\alpha < 0$, $Var(X) < E(X)$, resulting in the occurrence of under-dispersion. On the other hand, over-dispersion occurs when $\alpha > 0$, that is

$Var(X) > E(X)$. Lastly, the evidence of equi-dispersion occurs when $\alpha = 0$. The latter is also the original Poisson model (see Consul and Shoukri, 1985; LuValle, 1990).

3.2.2 Zero-truncated generalised Poisson distribution

Commonly, zero counts are unknown in natural capture-recapture data and need to be estimated, therefore, let $x = 0$ denote an individual that cannot be identified from the target population with probability p_0. Hence, under the assumption the zero-truncated generalised Poisson (ZTGP) distribution we have:

$$\begin{aligned} p_x^+ &= \frac{p_x}{1-p_0} \\ &= \frac{\theta(\theta+\alpha x)^{x-1}\left\{\frac{\exp(-\theta-\alpha x)}{x!}\right\}}{1-\exp(-\theta)} \\ &= \frac{\theta(\theta+\alpha x)^{x-1}}{x!}\left\{\frac{\exp(-\theta-\alpha x)}{1-\exp(-\theta)}\right\} \\ &= \frac{\theta(\theta+\alpha x)^{x-1}}{x!}\left\{\frac{\exp(-\theta)}{\exp(-\theta)}\right\}\left\{\frac{\exp(-\alpha x)}{\exp(\theta)-1}\right\} \\ &= \frac{\theta(\theta+\alpha x)^{x-1}}{x!}\left\{\frac{\exp(-\alpha x)}{\exp(\theta)-1}\right\}, \end{aligned} \tag{3.4}$$

where $p_0 = \exp(-\theta)$ and parameter estimates of θ and α can be obtained by fitting the ZTGP distribution.

3.2.3 Graphical device of the ratio plot for investigating the validity of the generalised Poisson distribution

Graphical techniques have been used to visualise quantitative data, and applied to the problem of selecting a suitable model in statistics since they are quick and uncomplicated to understand. A graphical approach, namely *the ratio plot* was suggested as a method for choosing a model by Böhning et al. (2013a). It can be extended for the zero-truncated count model which is common in capture-recapture studies. Thus, the ratio plot is applied in order to test the validity of the GP and the ZTGP distributions. The ratio plot is defined as the ratio of neighbouring probabilities multiplied by the value of the larger neighbour count, therefore, we develop the ratio of probabilities for a GP

distribution as follow:

$$
\begin{aligned}
r_x = (x+1)\frac{p_{x+1}}{p_x} &= (x+1)\frac{\theta\{\theta+\alpha(x+1)\}^{x+1-1}\left(\frac{\exp\{-\theta-\alpha(x+1)\}}{(x+1)!}\right)}{\theta(\theta+\alpha x)^{x-1}\left(\frac{\exp(-\theta-\alpha x)}{x!}\right)} \\
&= (x+1)\frac{\{\theta+\alpha(x+1)\}^{x}}{(\theta+\alpha x)^{x-1}}\exp\left(\frac{-\theta-\alpha x-\alpha}{-\theta-\alpha x}\right)\frac{x!}{(x+1)!} \\
&= \frac{\{\theta+\alpha(x+1)\}^{x}}{(\theta+\alpha x)^{x-1}}\exp(-\alpha) \\
&= \exp(-\alpha)\{\theta+\alpha(x+1)\}\left\{\frac{\theta+\alpha(x+1)}{\theta+\alpha x}\right\}^{x-1} \\
&= \exp(-\alpha)\{\theta+\alpha(x+1)\}\left\{\frac{\theta+\alpha x+\alpha}{\theta+\alpha x}\right\}^{x-1} \\
&= \exp(-\alpha)\{\theta+\alpha x+\alpha\}\left\{1+\frac{\alpha}{\theta+\alpha x}\right\}^{x-1} \\
&= \underbrace{\underbrace{\exp(-\alpha)}_{\text{constant}}\underbrace{\left(1+\frac{\alpha}{\theta+\alpha x}\right)^{x-1}}_{\text{e=2.71828182846;x}\to\infty}}_{\text{constant: c}}\{(\theta+\alpha)+\alpha x\} \\
&= c\{(\theta+\alpha)+\alpha x\} \\
&= (c\theta+c\alpha)+c\alpha x \\
&= \underbrace{c^{'}}_{\text{intercept}}+\underbrace{c\alpha}_{\text{slope}}x, \qquad (3.5)
\end{aligned}
$$

where c is a positive constant, defined as $\exp(-\alpha)\left(1+\frac{\alpha}{\theta+\alpha x}\right)^{x-1}$ and $c^{'}=c\theta+c\alpha$. When x is large we have that $\left(1+\frac{\alpha}{\theta+\alpha x}\right)^{x-1}$ approaches $e\approx 2.71828$ (see appendix A). It can be seen from (3.5) that r_x takes the from

$$r_x=\beta_0+\beta_1 x, \qquad (3.6)$$

where $\beta_0=c^{'}=c\theta+c\alpha$ denotes an intercept point, and $\beta_1=c\alpha$ is a slope for linear regression between r_x and x. In practice, relative frequencies are used to estimate capture probabilities, $p_x's$. Hence, the ratio plot of the GP distribution can be obtained using

$$r_x^*=(x+1)\frac{\widehat{p}_{x+1}}{\widehat{p}_x}=(x+1)\frac{f_{x+1}/N}{f_x/N}=(x+1)\frac{f_{x+1}}{f_x}, \qquad (3.7)$$

where f_x is the frequency of count x and $N=f_0+f_1+f_2+...+f_m$.

In the remainder of this section the ratio plot of r_x^* against x is used as a diagnostic tool for investigating the validity of the original Poisson and the generalised Poisson distributions. Indeed, when the ratio plot illustrates a linear line with positive slope this is evidence of the GP model in the case of over-dispersion. In contrast, if the ratio plot

shows a linear line with negative slope, it represents the generalised Poisson model with under-dispersion. Finally, the horizontal line of the ratio plot indicates equi-dispersion of generalised Poisson model or the original Poisson model. The significant property of the ratio plot, as mentioned in Chapter 2, is that the ratio plots of untruncated and zero-truncated distributions is identical. Then, the fitted linear line following the GP model can be applied for predicting the unobserved frequency count, (f_0), in capture-recapture data.

Example 1: Shakespeare data

Take, for example, the estimation of how many words Shakespeare actually knew, but did not appear in his works. This data set was originally analysed as a capture-recapture history of species problem by Efron and Thisted (1976). The data are shown in Table 3.1. It can be seen that 14,376 word types are used only once, 4,343 word types appear twice and so forth. The total number of words appearing in Shakespeare literature was 884,647 with 31,534 different word types. Let X be the number of times each word is appearing in Shakespeare's work, and f_x is the number of words appearing exactly x times.

Table 3.1: Frequency distribution of Shakespeare's words data

x	1	2	3	4	5	6	7	8	9	10
f_x	14,376	4,343	2,292	1,463	1,043	837	638	519	430	364
x	11	12	13	14	15	16+	n			
f_x	305	259	242	223	187	4,013	31,534			

The ratio plot can be used to investigate the validity of a zero-truncated count model for Shakespeare's data. Basically, parametric models can achieve a good fit by using only a main part of data and treat others parts as outliers (Bunge et al., 2014). For example, Chao and Bunge (2002) set the truncation point m at 10 and collapse $m > 10$ to one value. For this case study, the analysis of the truncation point is set at 16 giving collapsed $f_{x>15} = 4,013$.

The ratio plot of $r_x^* = (x+1)\dfrac{f_{x+1}}{f_x}$ against x is next used to investigate the validity of the models of count data in the capture-recapture approach. It can be seen from Figure 3.1 that the ratio plot presents a straight line with positive slope, but the intercept point is less than zero. This is a violation of dispersion parameter k of the negative binomial model for the Shakespeare data. Additionally, the linear model r_x^* of the negative binomial model offers $\widehat{k} = \dfrac{\widehat{\beta}_0}{\widehat{\beta}_1} = \dfrac{-0.3447}{0.9785} = -0.3523$ (Efron and Thisted (1976) gave $\widehat{k} = -0.3954$).

Although the ratio plot r_x^* shows a very good fit of the negative binomial model for observed frequencies, it will be limited for predicting the unobserved frequency in capture-recapture studies due to the boundary problem for the dispersion parameter. Thus, when the ratio plot shows a straight line with positive slope, it does not guarantee that the data follow the negative binomial distribution.

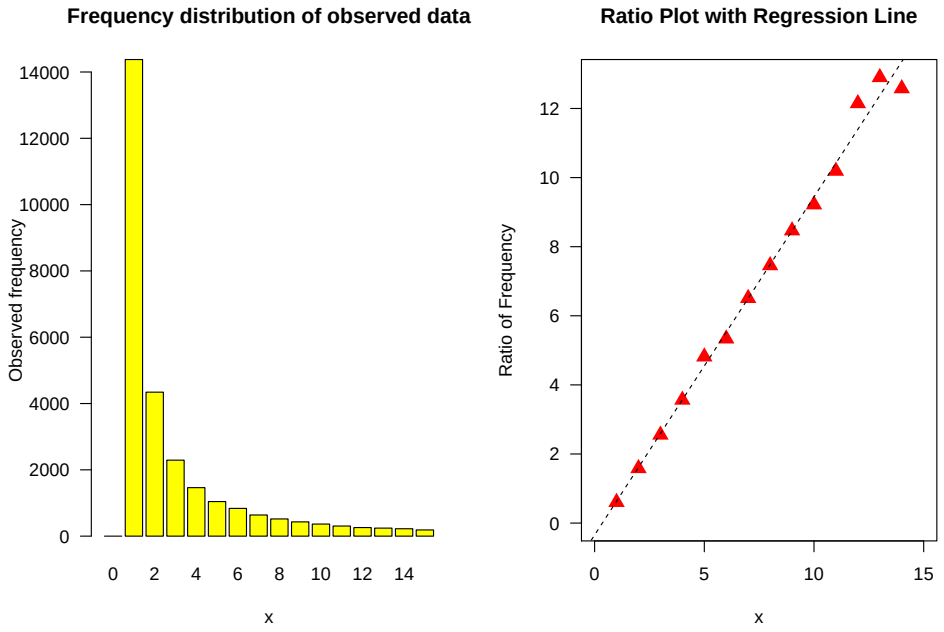

Figure 3.1: The observed frequency (left) and the ratio plot (r_x^*) (right) of Shakespeare data

3.3 Model based inference

Maximum likelihood estimation (MLE) is a well-known method for finding the value of unknown parameters (Scholz, 1985). Several statistical packages support this with excellent algorithms for estimating parameters from the MLE function. For example, the `GPseq package` in `R` is developed for computing parameters in the complex of maximum likelihood of the GP model. In this section the `GP package` is applied to estimate two parameters based on the maximum likelihood estimator of the zero-truncated generalised Poisson distribution in the M-step of the EM-algorithm.

3.3.1 The maximum likelihood of zero-truncated generalised Poisson modelling

Let f_x be the frequency counts with value x times and m is the largest observed count. It is simple to find the number of observed data n by calculating $n = f_1 + f_2 + ... + f_m$.

The likelihood function distribution of the ZTGP distribution based on n observations is given by

$$L(x;\theta,\alpha) = \prod_{x=1}^{m}\left[\left\{\frac{\theta(\theta+\alpha x)^{x-1}}{x!}\right\}\left\{\frac{\exp(-\alpha x)}{\exp(\theta)-1}\right\}\right]^{f_x}, \tag{3.8}$$

leading to the log-likelihood function of ZTGP distribution which is a conditional on n, defined as:

$$\begin{aligned}
\log L(x;\theta,\lambda) &= \log\left[\prod_{x=1}^{m}\left[\left\{\frac{\theta(\theta+\alpha x)^{x-1}}{x!}\right\}\left\{\frac{\exp(-\alpha x)}{\exp(\theta)-1}\right\}\right]^{f_x}\right] \\
&= \sum_{x=1}^{m}\log\left\{\frac{\theta(\theta+\alpha x)^{x-1}}{x!}\frac{\exp(-\alpha x)}{\exp(\theta)-1}\right\}^{f_x} \\
&= \sum_{x=1}^{m} f_x\{\log(\theta)+(x-1)\log(\theta+\alpha x)-\log(x!) \\
&\quad -\alpha x-\log(\exp(\theta)-1)\}.
\end{aligned} \tag{3.9}$$

Differentiating (3.9) with respect to θ and α, respectively, we have that

$$\frac{\partial}{\partial\theta}\log L(x;\theta,\alpha) = \sum_{x=1}^{m} f_x\left\{\frac{1}{\theta}+\frac{(x-1)}{\theta+\alpha x}-\frac{\exp(\theta)}{\exp(\theta)-1}\right\}, \tag{3.10}$$

and

$$\frac{\partial}{\partial\alpha}\log L(x;\theta,\alpha) = \sum_{x=1}^{m} f_x\left\{\frac{x(x-1)}{\theta+\alpha x}-x\right\}. \tag{3.11}$$

Setting the (3.10) and (3.11) equal to zero is not successful in achieving closed form expressions. Unfortunately, the estimators of θ and α under the zero-truncated generalised Poisson do not provide closed form solutions. Therefore, alternative approaches are need to deal with this problem. One of the effective approaches is the Expectation Maximisation algorithm or the EM algorithm, it is an iterative algorithm technique developed by Dempster et al. (1977). The EM algorithm consists of two components: E-step and M-step. The first step is for estimating the missing data. Then, the missing data are replaced by the conditional expected value, given the observed frequency counts and current parameter. This is followed by the M-step, in which the likelihood function is maximised by using both the observed and attributed information. The E-step and M-step are alternated until the likelihood parameters converge. It is important to bear in mind that the EM algorithm for the zero-truncated distribution requires a maximisation following the zero-truncated likelihood function (Böhning and Schön, 2005).

3.3.2 The EM algorithm of the zero-truncated generalised Poisson distribution

The objective of this study is to estimate the size of a target population N for capture-recapture data. The unobserved frequency, f_0, disappears in the counting procedure, yet it can be estimated under the observed frequencies. We start by calculating the expected value of unobserved frequency ($\widehat{f_0}$) in the E-Step.

- **Expectation: E-Step**

We start by estimating the unobserved frequency (f_0) by using an expected value of the unobserved frequency ($\widehat{f_0}$), given the observed data and the current estimates of likelihood parameters. That is

$$\begin{aligned} \widehat{f_0} &= E(f_0|\text{observed};\theta,\alpha) \\ &= E(f_0|f_1,f_2,f_3,...,f_m;\theta,\alpha) \\ &= Np_0. \end{aligned} \tag{3.12}$$

The size of population N can be estimated by $n+\widehat{f_0}$, where the sample size is denoted by $n=\sum_{x=1}^{m} f_x$. Therefore,

$$\begin{aligned} \widehat{f_0} &= (n+\widehat{f_0})p_0 \\ &= np_0+\widehat{f_0}p_0 \\ \widehat{f_0}-\widehat{f_0}p_0 &= np_0 \\ \widehat{f_0}(1-p_0) &= np_0 \\ \widehat{f_0} &= \frac{np_0}{1-p_0}. \end{aligned} \tag{3.13}$$

According to the generalised Poisson distribution, we have $p_0=\exp(-\theta)$, substituting in to (3.13) as:

$$\begin{aligned} \widehat{f_0} &= n\frac{\exp(-\theta)}{1-\exp(-\theta)} \\ &= \left\{\frac{\exp(-\theta)}{\exp(-\theta)}\right\}\left\{\frac{n}{\exp(\theta)-1}\right\} \\ &= \frac{n}{\exp(\theta)-1}. \end{aligned} \tag{3.14}$$

As a consequence, the expected value of unobserved data is estimated by

$$\widehat{f_0}=\frac{n}{\exp(\widehat{\theta})-1}. \tag{3.15}$$

- **Maximization: M-Step**.

In the M step of the EM algorithm, the log-likelihood function is required in which the f_0 is replaced by $\widehat{f}_0$ in E-Step as Table 3.2.

Table 3.2: Frequency distribution

x	0	1	2	3	...	m
f_x	$\widehat{f}_0$	f_1	f_2	f_3	...	f_m

The associated complete likelihood of the generalised Poisson function, $L_c(x;\theta,\alpha)$, consists of two parts: unobserved and observed data, given as

$$\begin{aligned} L_c(x;\theta,\alpha) &= p_0^{f_0}\prod_{x=1}^{m} p_x^{f_x} \\ &= \prod_{x=0}^{m}\left\{\theta(\theta+\alpha x)^{x-1}\left(\frac{\exp(-\theta-\alpha x)}{x!}\right)\right\}^{f_x}. \end{aligned} \tag{3.16}$$

This leads to the associated complete log-likelihood of the generalised Poisson distribution given by

$$\begin{aligned} \log L_c(x;\theta,\alpha) &= \log\left[\prod_{x=0}^{m}\left\{\theta(\theta+\alpha x)^{x-1}\left(\frac{\exp(-\theta-\alpha x)}{x!}\right)\right\}^{f_x}\right] \\ &= \sum_{x=0}^{m} f_x\left[\log\theta+(x-1)\log(\theta+\alpha x)-\theta-\alpha x-\log(x!)\right]. \end{aligned} \tag{3.17}$$

For estimating the MLE of parameters θ and α, the function (3.17) requires to be maximised. Hence, we take the partial derivative of complete log-likelihood function as follows:

$$\begin{aligned} \frac{\partial}{\partial\theta}\log L_c(x;\theta,\alpha) &= \frac{\partial}{\partial\theta}\left[\sum_{x=0}^{m} f_x\left\{\log\theta+(x-1)\log(\theta+\alpha x)-\theta-\alpha x-\log(x!)\right\}\right] \\ &= \sum_{x=0}^{m} f_x\left\{\frac{1}{\theta}+\frac{x-1}{\theta+\alpha x}-1\right\} \\ &= \frac{1}{\theta}\sum_{x=0}^{m} f_x+\sum_{x=0}^{m} f_x\left(\frac{x-1}{\theta+\alpha x}\right)-\sum_{x=0}^{m} f_x. \end{aligned} \tag{3.18}$$

Then, we equate (3.18) to zero

$$\frac{1}{\theta}\sum_{x=0}^{m} f_x+\sum_{x=0}^{m} f_x\left(\frac{x-1}{\theta+\alpha x}\right)-\sum_{x=0}^{m} f_x=0. \tag{3.19}$$

In addition, the parameter α can be estimated by taking derivative (3.17) with respect to α, that is

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log L_c(x;\theta,\alpha) &= \frac{\partial}{\partial\alpha}\left[\sum_{x=0}^{m} f_x\{\ \log(\theta)+(x-1)\log(\theta+\alpha x)-\theta-\alpha x-log(x!\,)\}\right]\\ &= \sum_{x=0}^{m} f_x\left\{\frac{x(x-1)}{\theta+\alpha x}-x\right\}\\ &= \sum_{x=0}^{m} f_x\left\{\frac{x(x-1)}{\theta+\alpha x}\right\}-\sum_{x=0}^{m} x f_x. \end{aligned} \tag{3.20}$$

Equating (3.20) to zero, we have that

$$\sum_{x=0}^{m} f_x\left\{\frac{x(x-1)}{\theta+\alpha x}\right\}-\sum_{x=0}^{m} x f_x=0. \tag{3.21}$$

Again, there are no closed form solutions of θ and α under the completely generalised Poisson distribution. However, statistical software programs can support parameter estimation such as the maximum likelihood estimation of the generalised Poisson distribution. One of these is the `GPseq package` in `R` which can be applied to the EM-algorithm in the M-Step to estimate two parameters; θ and α.

Finally, the size of target population can be achieved using the Horvitz-Thompson estimator as

$$\widehat{N}_{MLEGP}=\frac{n}{1-\exp(-\widehat{\theta})}, \tag{3.22}$$

where $\widehat{p}_0=\exp(-\widehat{\theta})$.

3.3.3 Algorithm

In summary, the EM algorithm procedure under the GP distribution for estimating the elusive population size in capture-recapture is given as follows:

Step 0 : Setting $l=0$ and then choosing a initial value for unobserved frequency $\widehat{f}_0^{(l)}$. Initial values taken as $\widehat{f}_0^{(0)}=\widehat{f}_{(0)Turing}=\widehat{N}\widehat{p}_0=\widehat{N}(\frac{f_1}{S})$ where $S=\sum_{x=1}^{m} x f_x$ may be useful when implementing the EM algorithm. Turing's estimator is constructed under the original Poisson model and easy to compute, hence it might save time in the computational procedure. The algorithm is repeated until the log-likelihood function of ZTGP distribution in (3.9) converges to a constant with an acceptable error, therefore, its initial value is set as

$$\log L(x;\theta,\alpha)^{(l)}=\log L(x;\theta,\alpha)^{(0)}=-\infty,$$

where a very large negative number is used in practice.

Step 1: Substituting $\widehat{f}_0^{(l)}$ in a completed frequency distribution table as Table 3.3 for computing new maximum likelihood estimators $\widehat{\theta}^{(l+1)}$ and $\widehat{\alpha}^{(l+1)}$.

Table 3.3: The frequency distribution of complete data

x	0	1	2	3	...	m
f_x	$\widehat{f}_0^{(l)}$	f_1	f_2	f_3	...	f_m

As suggested above, the maximum likelihood estimators are computed by using `GPseq package` in `R`, in particular `the generalized poisson likelihood () function`. This leads to new maximum likelihood estimators.

Step 2: Computing the new unobserved frequency and the size of the target population, that is

$$\widehat{f}_0^{(l+1)} = \frac{n}{\exp\{\widehat{\theta}^{(l+1)}\} - 1}$$

and

$$\widehat{N}_{MLEGP}^{(l+1)} = \frac{n}{1 - \{\exp(-\widehat{\theta}^{(l+1)})\}}.$$

Step 3 : Checking the condition of algorithm by plugging $\widehat{\theta}^{(l+1)}$ and $\widehat{\alpha}^{(l+1)}$ into the log-likelihood of ZTGP function:

$$\begin{aligned} \log L(x;\theta,\alpha)^{(l+1)} &= \sum_{x=1}^{m} f_x[\log(\widehat{\theta}^{(l+1)}) + (x-1)\log(\widehat{\theta}^{(l+1)} + \widehat{\alpha}^{(l+1)}x) - \log(x!) \\ &\quad -\alpha x - \log(\exp(\widehat{\theta}^{(l+1)}) - 1)], \end{aligned} \tag{3.23}$$

and comparing

$$dif = \left|\log L(x;\theta,\alpha)^{(l+1)} - \log L(x;\theta,\alpha)^{(l)}\right| < 0.0001,$$

setting $l = l + 1$. Then, if $dif > 0.0001$ return to step 1 so that new maximum likelihood estimators are updated. The algorithm is repeated until the log likelihood function of ZTGP converge to a constant with an acceptable error, that is $dif < 0.0001$.

It is noticed that the EM-algorithm for the ZTGP distribution sometimes does not converge, we suggest to experiment with different starting points in the E-step.

A numerical study to evaluate the performance of proposed estimator MLEGP is provided in the next section.

3.4 Simulation study

To investigate the behaviour of the proposed estimators, and compare with some well-known estimators, we perform various simulation scenarios by generating data covering the original Poisson and the generalised Poisson models which are assumed as true models.

3.4.1 Simulation scenarios

1) Poisson distribution
As the Poisson distribution is a special case of the GP distribution when the dispersion parameter $\alpha = 0$, data are generated from the Poisson distribution with ten different value parameters:

$$\lambda \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0\}.$$

2) Generalised Poisson distribution
Counts are generated from the generalised Poisson distribution using the function `rzigp ()` in `R` with parameters

$$\theta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0\}$$

and

$$\phi \in \{1.5, 2.0, 2.5, 3.0\},$$

where $\phi = \dfrac{Var[X]}{E[X]}$, so that $\phi = \left\{\dfrac{\theta}{(1-\alpha)^3}\right\} \div \left\{\dfrac{\theta}{(1-\alpha)}\right\} = \dfrac{1}{(1-\alpha)^2}$.

Additionally, the population size of the studies is fixed as $N = 200$ representing a small size study, $N = 1,000$ for a medium study size and $N = 10,000$ for a large study size. Each simulation scenario is repeated for $T = 1,000$ data sets. Any occurrences of zero counts are truncated, before going to the estimation procedure. Additionally, six estimators of population size are compared: the maximum likelihood estimator under the Poisson distribution (MLEPoi), the Turing estimator (Turing), the McKendrick estimator (McK), the Chao lower bound estimator (Chao), the Zelterman estimator (Zel) and the maximum likelihood estimator under the zero-truncated generalised Poisson (MLEGP).

3.4.2 Statistical investigation

Estimators of the population size N are evaluated in terms of accuracy and precision. Accuracy refers to the closeness of an expected value $E(\widehat{N})$ to the true value of parameter N. A difference between the expected and true value of a parameter indicates bias,

used in this thesis as a measure of accuracy, in the parameter estimation. A positive bias indicates, an overestimation, while a negative bias represents an underestimation. Precision refers to variability or variance, and is typically defined as the corresponding variance of an estimator, $Var(\widehat{N})$. Indeed, precision is an inverse of variance, and its value will be greater than zero. The lower relative variance, the more precise an estimator will be.

Bias and precision describe the performance of an estimator. The less biased and the more accurate an estimator is, the better its overall ability to produce an accurate point estimate. Another important measurement that incorporates concepts of both bias and precision is the root mean square error (RMSE). Both quantities are important and need to be as small as possible to achieve a good estimator.

As an increasing population size would be expected to lead to higher absolute bias and greater variability, this is addressed by calculating the relative bias (RBias), relative variance (RVar) and relative root mean square error (RRMSE) of population size estimators. These three measures are calculated within each simulation dataset

1) Relative bias of the population size estimator

Suppose that $\widehat{N}_{(t)}$ denotes the estimated value of the population size at replication t where $t = 1, 2, 3, ..., T$. The expected value of population size estimator ($RBias(\widehat{N})$) is achieved by

$$E(\widehat{N}) = \frac{1}{T}\sum_{t=1}^{T}\widehat{N}_{(t)}.$$

Hence, the relative bias of population size estimator is defined as

$$RBias(\widehat{N}) = \frac{1}{N}\left\{E(\widehat{N}) - N\right\} = \frac{1}{N}\left\{bias(\widehat{N})\right\}, \tag{3.24}$$

where $bias(\widehat{N}) = E(\widehat{N}) - N$.

2) Relative variance of population size estimator

The relative variable of the population size estimator ($RVar(\widehat{N})$) is defined as

$$RVar(\widehat{N}) = \frac{1}{N^2}\left[\frac{1}{T-1}\sum_{t=1}^{T}\left\{\widehat{N}_{(t)} - E(\widehat{N})\right\}^2\right] = \frac{1}{N^2}Var(\widehat{N}), \tag{3.25}$$

where $Var(\widehat{N}) = \dfrac{1}{T-1}\displaystyle\sum_{t=1}^{T}\left\{\widehat{N}_{(t)} - E(\widehat{N})\right\}^2$.

3) Relative root mean square error of population size estimator

The mean squre error ($MSE(\widehat{N})$) is defined as:

$$MSE(\widehat{N}) = Var(\widehat{N}) + \{bias(\widehat{N})\}^2,$$

taking root of $MSE(\widehat{N})$, gives the root mean square error as:

$$RMSE(\widehat{N}) = \sqrt{Var(\widehat{N}) + \{bias(\widehat{N})\}^2}.$$

The relative root mean square error $RRMSE(\widehat{N})$ is then:

$$RRMSE(\widehat{N}) = \frac{1}{N}\sqrt{Var\{\widehat{N}\} + \{bias(\widehat{N})\}^2}. \tag{3.26}$$

3.4.3 Simulation result

Simulation results are split into two parts; the original Poisson and generalised Poisson distributions. As described above three measures are used to evaluate the performance of estimators; relative bias, relative variance and relative root mean square error.

1) Poisson distribution

The simulation results compare the proposed estimator MLEGP to its competitors. We hypothesise that the MLEGP estimator will perform well when data are generated under the Poisson distribution because the Poisson distribution is a sub-model of the generalised Poisson distribution. The Turing, the MLEPoi and the McK estimators are also expected to have high performance under the Poisson model. According to the relative bias, overall, it is clear from Table A.1 in Appendix A and Figure 3.2 that all estimators display asymptotic unbiasedness with respect to population size. The relative bias from the MLEPoi, the Turing and the McK show that these estimators underestimate population size for the small population size as value of event parameter λ decrease. On the other hand, the MLEGP estimator provides a severe overestimation of population size, especially for the a small population size or/and and a small value of λ

Considering the relative variance of the six estimators, it can be seen from Table A.2 in Appendix A and Figure 3.3 that the MLEPoi estimator provides the smallest variance followed by the Turing and the McK estimators. The proposed MLEGP estimator shows high relative variance but it tends to be smaller than the Zelterman and the Chao estimators when λ and population size increase. Moreover, all estimators are affected by the Poisson parameters and population sizes. That is, increasing Poisson parameters and population sizes result in decreased variance.

A good estimator should have minimum variance and bias. Relative root mean square error has been used as a measurement for comparing the performance of population size estimators. As can be seen from Table A.3 in Appendix A and Figure 3.4, the MLEPoi is the best-performing estimator for the Poisson model with the smallest value of relative root mean square error for all conditions. The proposed estimator MLEGP might be a efficient estimator for estimating population size based-Poisson for λ greater than 1.

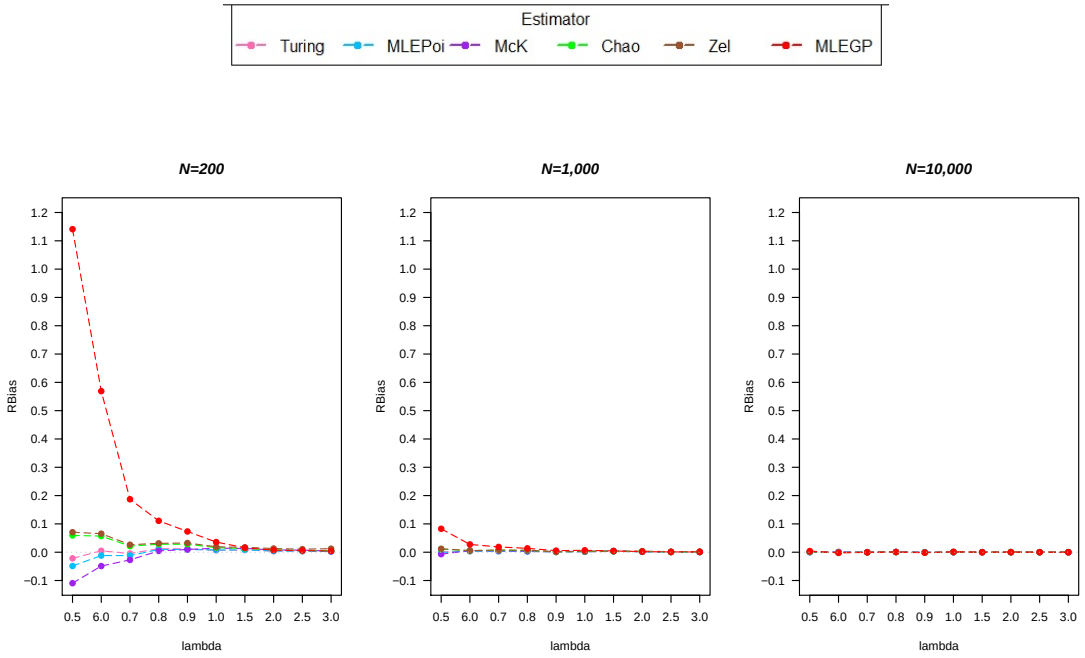

Figure 3.2: The relative bias of six estimators with different parameters following the Poisson distribution, when $N = 200$, $N = 1,000$ and $N = 10,000$.

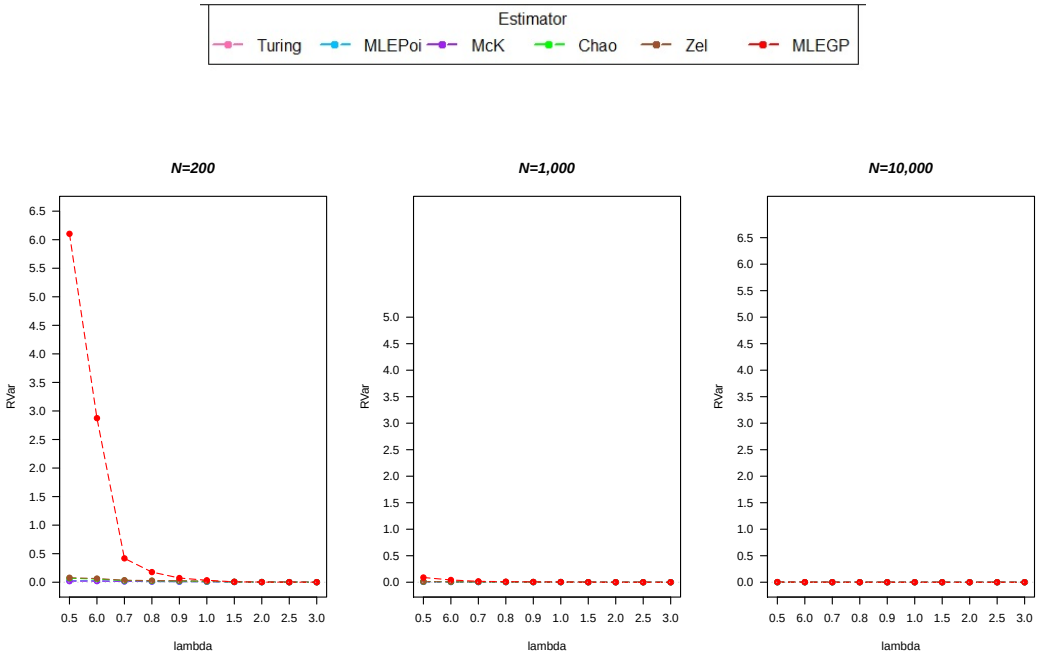

Figure 3.3: The relative variance six estimators with different parameters following the Poisson distribution, when $N = 200$, $N = 1,000$ and $N = 10,000$.

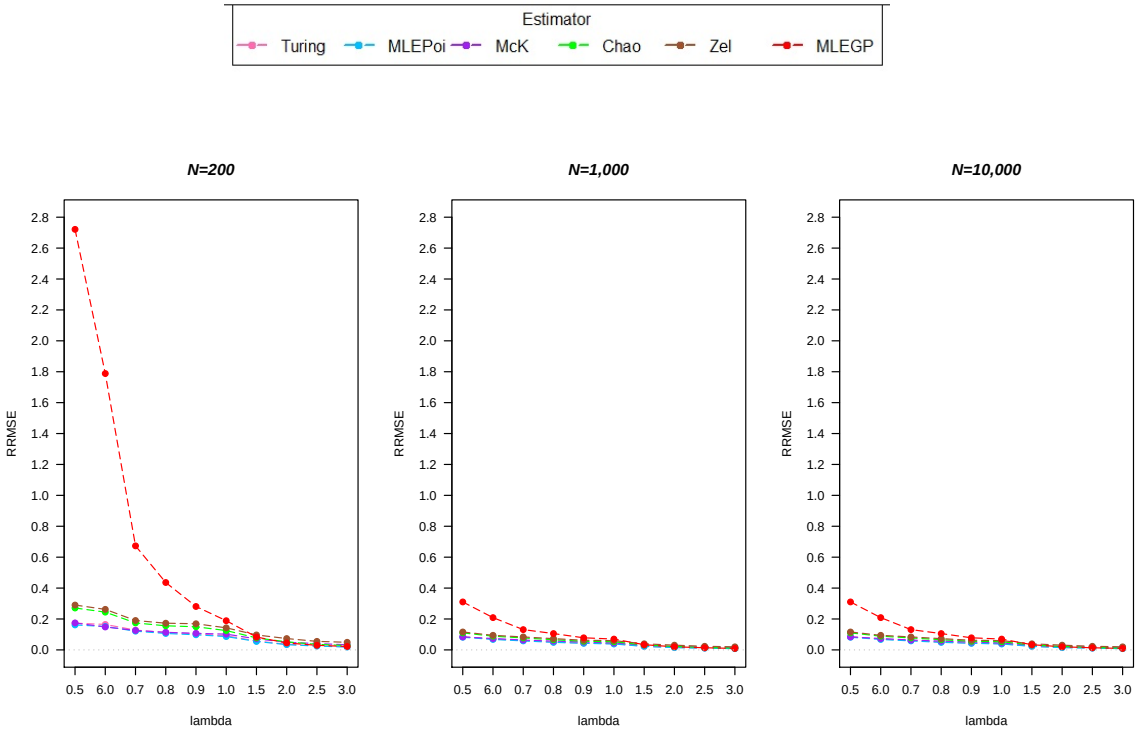

Figure 3.4: The relative root mean square six estimators with different parameters following the Poisson distribution.

2) The generalised Poisson distribution.

The generalised Poisson model is also taken as the true model. The maximum likelihood estimator under the zero-truncated generalised Poisson (MLEGP) estimator might be expected to be an appropriate estimator since it is derived from the generalised Poisson distribution.

According to Table A.4 in Appendix A and Figure 3.5, under the generalised Poisson distribution almost all estimators show an underestimation, except the MLEGP estimator which shows an overestimate for all conditions. The proposed MLEGP estimator provided the smallest bias among medium and large population sizes, followed by the Zelterman and the Chao estimators respectively. Only the proposed MLEGP estimator tends to be an asymptotically unbiased estimator under the generalised Poisson model with respect to N. Considering the effects of parameters and size of populations, it can be seen that increasing parameter ϕ leads to an increase in bias for the small population size but this situation tends to be less evident for larger population sizes or as parameter θ increases.

Regarding relative variance, which is presented in Table A.5 in Appendix and Figure 3.6, the MLEPoi estimator shows the highest level of precision with the smallest relative variance whereas the proposed estimator MLEGP has the worst precision for a small population size. However, it tends to be more precise when the population increases.

Lastly, it can be seen from Table A.6 in Appendix A and Figure 3.7, for a small population size the Zelterman estimator is the best performing with the smallest value of RRMSE, whereas the proposed MLEGP estimator has the largest RRMSE. For the medium size of population, the MLEGP estimator tends to have the best performance with smallest RRMSE when $theta \geq 1$ and $\phi \leq 2.5$. Furthermore, the MLEGP estimator is the best performing estimator with the smallest RRMSE when N is large. Therefore, the proposed estimator outperforms the competing estimators and is suitable for the large population size.

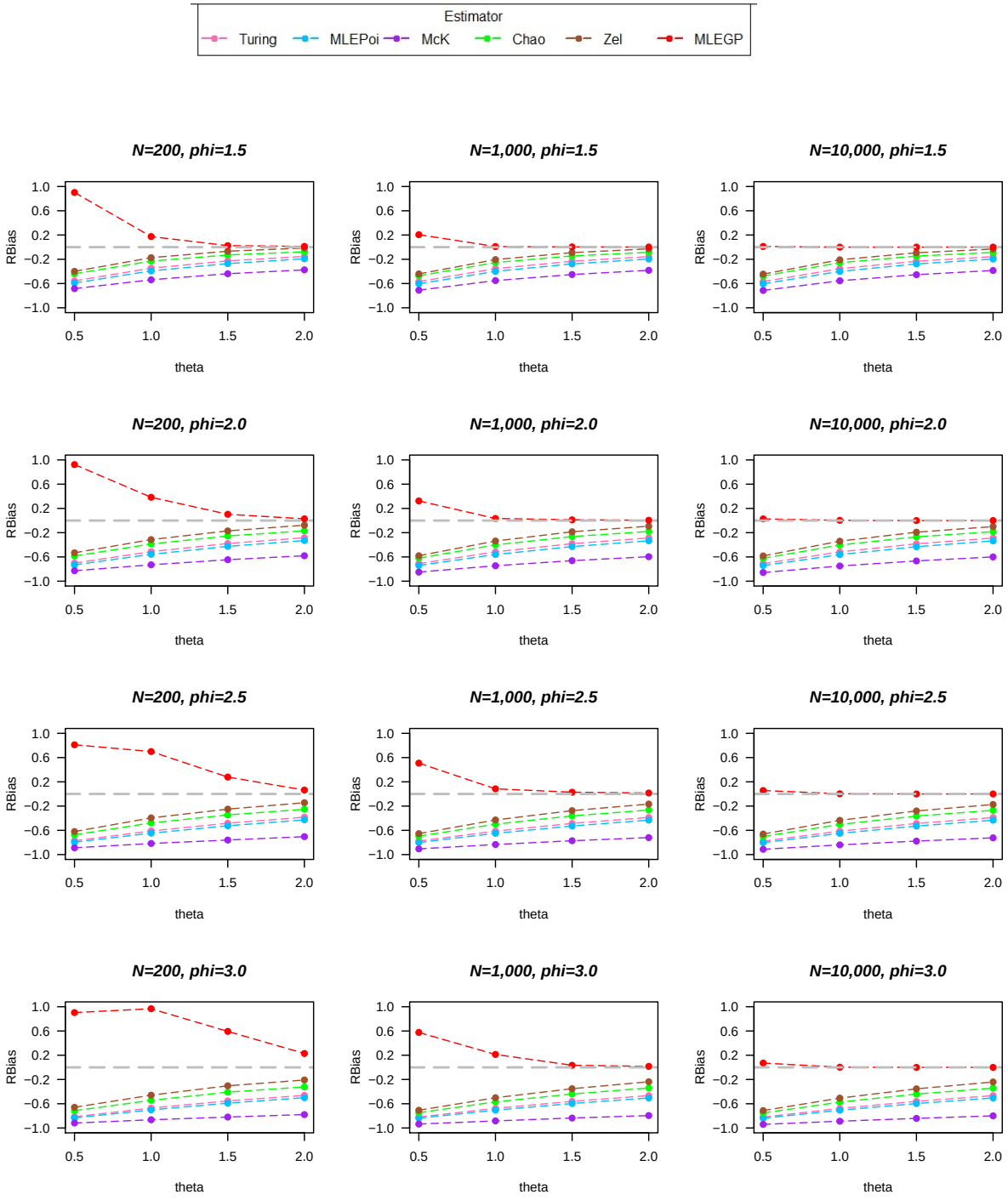

Figure 3.5: Relative bias of six estimators with different parameters in the generalised Poisson distribution, when $N = 200, N = 1,000$ and $N = 10,000$

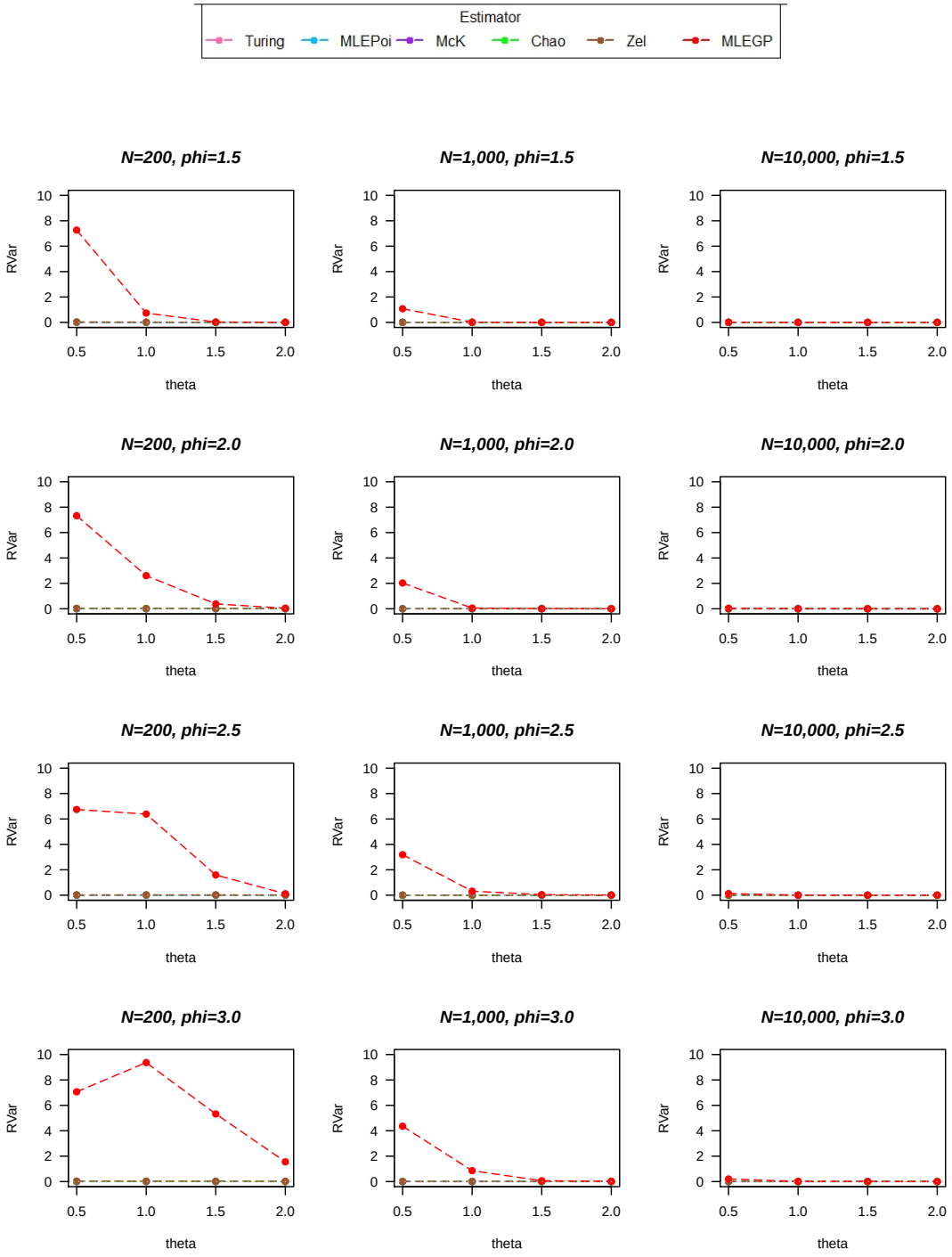

Figure 3.6: Relative variance of six estimators with different parameters in the generalised Poissonn distribution, when $N = 200, N = 1,000$ and $N = 10,000$

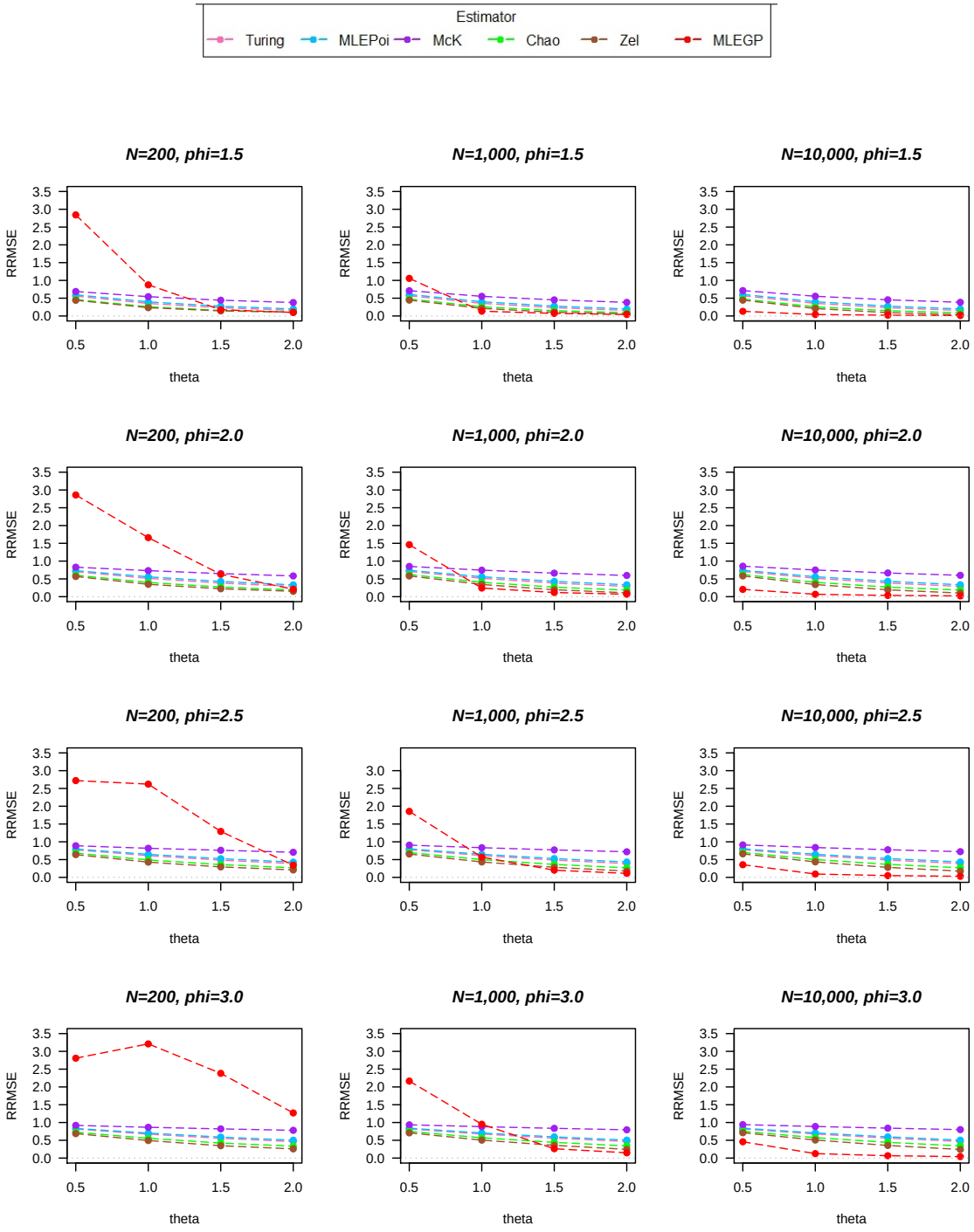

Figure 3.7: Relative root mean square error of six estimators with different parameters in the generalised Poisson distribution, when $N = 200, N = 1,000$ and $N = 10,000$

3.5 Real data examples

In this section application of the population size estimators to the real data samples, the Shakespeare data and Cottontail, are illustrated. The population size estimations from the new estimator is compared with the overviews of population size estimators presented in Chapter 2. Additionally, the expected value of frequencies are fitted based on the zero-truncated distribution, that is $\widehat{f}_x = n\widehat{p}_x^+$. Finally, we used the Chi-square goodness of fit $\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \widehat{f}_x)^2}{\widehat{f}_x}$ under the null hypothesis of Poisson and generalised Poisson distributions to compare the models fit.

3.5.1 Shakespeare data

The Shakespeare data from Example 1 in Section 2 of this chapter, are used to first illustrate the application of the estimators in the estimation of how many words Shakespeare actually knew, but did not use in his works. As mentioned above, the ratio plot represents a straight line with positive slope, yet the intercept point is lower than zero (see Figure 3.1). Thus, it might be more appropriate to model the observed data by the zero-truncated model of generalised Poisson distribution for over-dispersion. The estimation of maximum likelihood estimator under the zero-truncated generalised Poisson distribution is used to estimate the value of the two model parameters as: $\widehat{\theta} = 0.09$ and $\widehat{\alpha} = 0.63$. This corresponds to the total number of words that Shakespeare knew being estimated as $365,736$ words.

The aim of the study is to compare the performance of the MLEGP estimator with other estimators. The Poisson model is misspecified since the ratio plot is not a horizontal line. Hence, the Turing and the MLEPoi estimators will not be valid. On the other hand, the Chao, the Zelterman and the MLEGP estimators might be appropriate for estimating the number of words. According to Table 3.4, each estimator provides a large difference in the estimated total number of words. The MLEPoi and the Turing are lowest: $35,521$ and $38,694$, respectively. Whereas the proposed estimator provides the largest number (365,736). The Chao estimator, which is known as the lower bound estimator for heterogeneity model, shows a smaller number of words than the Zelterman and the MLEGP estimators.

Table 3.4: Point estimates of the number of words that Shakespeare knew.

Estimator	$\widehat{N}$	$\widehat{f}_0$
Homogeneous Poisson		
Turing	38,694	7,160
MLEPoi ($\widehat{\lambda} = 2.19$))	35,521	3,982
Heterogeneous Poisson		
Chao	55,328	23,794
Zelterman	69,537	38,003
MLEGP ($\widehat{\theta} = 0.09$, $\widehat{\alpha} = 0.63$)	365,736	291,723

To investigate the model fit to the data, the observed frequencies and the fitted frequencies of the zero-truncated Poisson and zero-truncated generalised models are provided in Table 3.5 and the comparison of fitted frequencies shown in Figure 3.8. Indeed, the fitted frequencies are achieved by comparing the expected frequencies conditioning on the observed data and the estimator of the zero-truncated distribution.

To illustrate how to calculate the fitted frequency count of one based on the ZTGP distribution; $\widehat{f}_1 = n \times \widehat{p}_1(\widehat{\theta}, \widehat{\alpha}) = 27,521 \times 0.5092 = 14,013.69$. As a consequence, the zero-truncated generalised Poisson model provides a better fit than the original Poisson due to adding a dispersion parameter which is more realistic for the long-tail data. Moreover, the numerical study supports that the MLEGP estimator tends to be a excellent estimator with small bias and variance for large sample size.

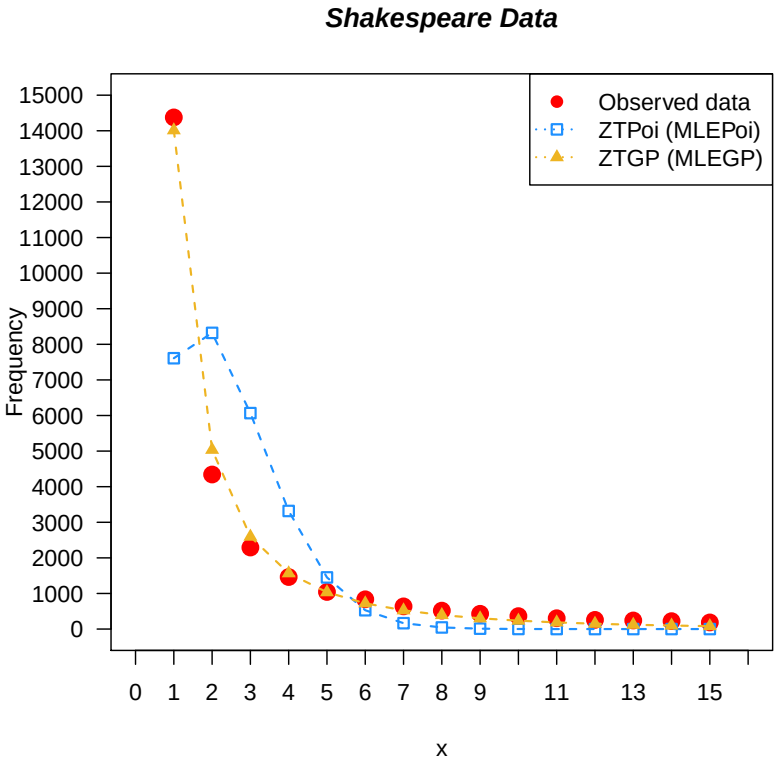

Figure 3.8: Observed frequencies with fitted frequencies under the zero-truncated Poisson (ZTPoi) and zero-truncated generalised Poisson (ZTGP) distributions for Shakespeare data, ignoring $f_{x\geq 16}$

Table 3.5: Observed and fitted frequency distribution for Shakespeare data.

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (MLEGP)
1	14,376	7,609	14,014
2	4,343	8,322	5,037
3	2,292	6,067	2,596
4	1,463	3,318	1,567
5	1,043	1,451	1,035
6	837	529	723
7	638	165	526
8	519	45	394
9	430	11	302
10	364	2	236
11	305	0	187
12	259	0	150
13	242	0	121
14	223	0	99
15	187	0	82
χ^2		9,943,648	917

3.5.2 Cottontail data

The original data are from Edwards and Eberhardt (1967) and were also analysed by Chao (1987). The live-trapping study knew the population size; 135 cottontail rabbits were penned in a limited area. The recorded frequency data given as in Table 3.6

Table 3.6: Frequency distribution of cottontail data

x	1	2	3	4	5+	n
f_x	43	16	8	6	3	76

The ratio plot $(r_x^*, x = 1, 2, 3, 4)$ as Figure 3.9 represents a linear regression line with a intercept point $\widehat{\beta}_0 = -0.5078$ and slope $\widehat{\beta}_1 = 1.1279$. Based on the zero-truncated negative binomial model the dispersion parameter can easily be estimated from $\widehat{k} = \dfrac{\widehat{\beta}_0}{\widehat{\beta}_1} = -0.45$, this is a boundary parameter violation in negative binomial model. Consequently, the zero-truncated generalised Poisson distribution is used as an alternative approach, assuming that the MLEGP will be an appropriate method for estimating the total number of cottontail rabbits. The MLEGP is used to estimate the value of the two parameters as, $\widehat{\theta} = 0.78$ and $\widehat{\alpha} = 0.14$. This results in the total number of cottontail rabbits being approximated as 139, which is slight overestimation.

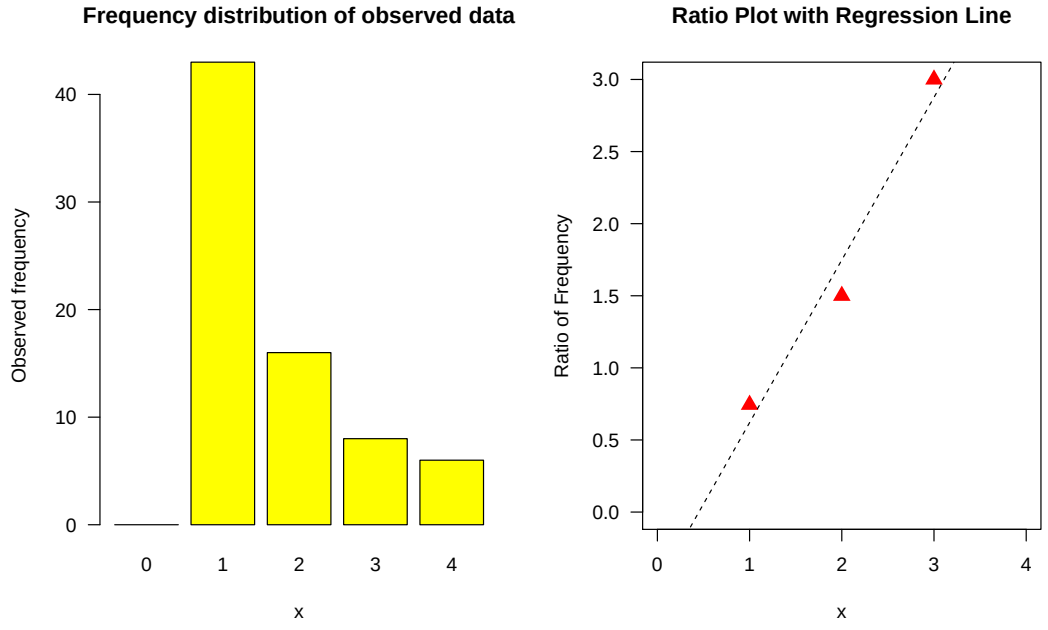

Figure 3.9: The observed frequency (left) and the ratio plot of $(x+1)\frac{f_{x+1}}{f_x}$ versus x (right) of cottontail data

As can be seen from Table 3.7 the Turing, the McK and the MLEPoi estimators, which assume a misspecified model, underestimate the population size of cottontail rabbit $(117, 100, 117)$. The Chao method shows the most accuracy with the smallest bias of -1, but still results in a slightly underestimated value. The proposed MLEGP estimator has more accuracy than the Zelterman estimator with respect to bias. However, both of them result in overestimation. Supporting evidence of the better fitting zero-truncated generalised Poisson distribution compared to the zero-truncated Poisson distribution can be seen as Table 3.8 and Figure 3.10.

Table 3.7: Estimated size of cottontail , $N = 135$

Estimator	$\widehat{N}$	$\widehat{f}_0$	*Bias*
Poisson			
Turing	117	41	-18
McK	100	24	-35
MLEPoi $(\widehat{\lambda} = 1.06)$	117	41	-18
Heterogeneity			
Chao	134	58	-1
Zelterman	145	69	10
MLEGP $(\widehat{\theta} = 0.78, \widehat{\alpha} = 0.14)$	139	60	4

Table 3.8: Observed and fitted frequency distribution for cottontail data.

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (MLEGP)
1	43	41	42
2	16	22	19
3	8	8	8
4	6	2	3
χ^2		9.73	3.50

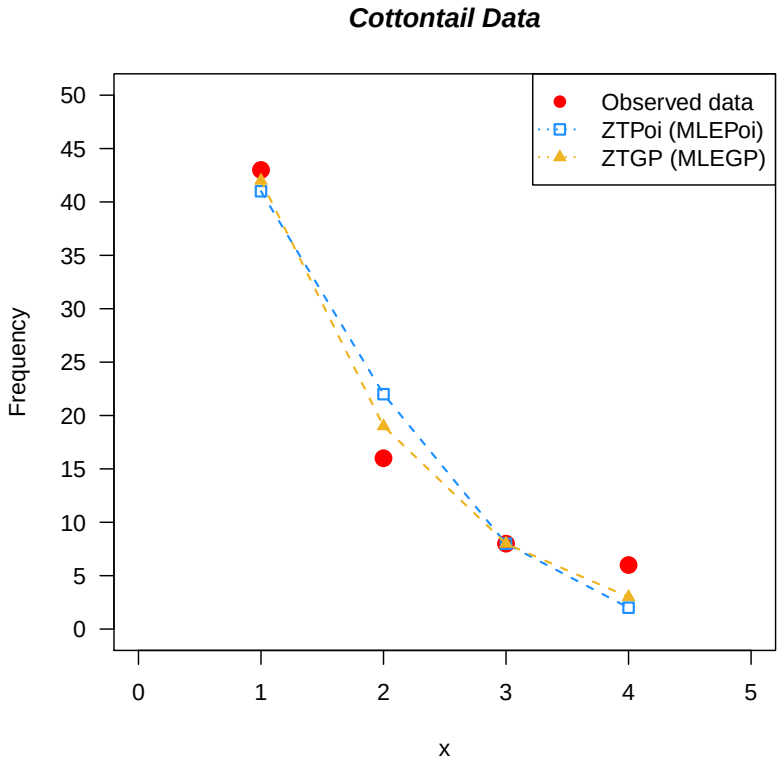

Figure 3.10: Observed and fitted frequencies under the zero-truncated Poisson (ZTPoi) and zero-truncated generalised Poisson (ZTGP) distributions for cottontail data, ignoring $f_{x\geq 5}$

3.6 Conclusion

Although the Poisson distribution is considered as the basic model for the capture-recapture data, it is often useless in real situations. The heterogeneity of populations which can be described by covariates or latent variables might mean that capture probabilities are heterogeneity. The mixture model is one of the well-known approaches that has been applied to deal with the heterogeneity under the assumption that some components of the capture probabilities arise from a mixed distribution. Alternatively, the generalised Poisson (GP) distribution is expected that it might be a powerful model for counting data where the event parameter is not homogeneous. The advantage of the GP model is that it is extending the Poisson model by adding a dispersion parameter so, it can cope with longer tail count data.

In this chapter new population sizes based on the GP model are derived using maximum likelihood estimation, resulting in the proposed MLEGP estimator. In addition, a simulation study is provided to consider the behaviour of the new proposed estimator under the Poisson and generalised Poisson models. The simulation results confirm

that the MLEGP estimator is an asymptotically valid estimator under both the Poisson and generalised Poisson distributions. However, it has substantial bias and very large variance for the small population size.

In summary, the MLEGP estimator works very well for estimating the target population size based on the Poisson distribution when the value of event parameter $\lambda \geq 1$ and performs well for the generalised Poisson distribution when population size 10,000 and above.

It is remarkable that there are two main drawbacks for using the population size estimator under the generalised Poisson distribution. Firstly, the non-convergence problem can occur when we use the EM algorithm to estimate parameters of the zero-truncated generalised Poisson distribution. Secondly, the MLEGP estimator works well under the true model when population size is large only. Therefore, it seems that the MLEGP estimator has limitation to be useful in capture-recapture count data analysis.

Chapter 4

Population size estimation based on the CMP distribution

This chapter furthers the work in the previous chapter using the capture-recapture approach for estimating the population size under a truncated count model that accounts for heterogeneity. Accordingly, a proposed estimator based on the zero-truncated Conway-Maxwell-Poisson distribution is provided in this chapter. Parameter estimates based on the CMP distribution can be obtained by exploiting the ratios of successive frequency counts in a weighted least squares regression framework. The results of the comparisons with the Turing, the maximum likelihood Poisson (MLEPoi), the Zelterman (Zel) and the Chao estimators reveal that our proposal can be beneficially used. Furthermore, the proposed estimator outperforms its competitors in heterogeneous settings. The empirical examples consider the homogeneous case and several heterogeneous cases, each with its own features, and provide interesting insights into the behaviour of the estimators.

4.1 Introduction

The Poisson distribution is the basic model which has been used for modelling count or frequency capture-recapture data. However, many real data sets do not follow assumption of equi-dispersion that underlines the Poisson distribution. The negative binomial distribution has become popular for capturing over dispersion in capture recapture data (see Schwarz and Arnason, 1996; Chao and Bunge, 2002; Cruyff and Van der Heijden, 2008; Rocchetti et al., 2011; Lanumteang and Böhning, 2011). Although the generalised Poisson model was discussed in the previous chapter, it is not suitable for all of real data sets. In this chapter another two-parameter model generalised from the Poisson distribution, called the Conway-Maxwell-Poisson (CMP) distribution is considered (Shmueli

et al., 2005). It is expected that the CMP distribution can account for heterogeneity. The benefits of the CMP distribution include the fact that it contains important sub-models (i.e. the Poisson, the Bernoulli and geometric distributions) and that it generalises the Poisson distribution allowing for over-dispersion as well as under-dispersion. The Conway-Maxwell-Poisson model has been used in various studies such as analysis of word length (Wimmer et al., 1994), internet search engine visits (Telang et al., 2004), and modelling electric power system reliability (Guikema and Coffelt, 2008). Here, it is applied to capture-recapture data.

4.2 The Conway-Maxwell-Poisson distribution for the capture-recapture data

Capture-recapture analysis depends on the repeated sampling of a target population of size N. The count distribution of a repeated identifications can be associated with the frequency distribution f_x where f_x is the frequency of individuals identified exactly x times and f_0 is an unobserved frequency. Therefore, $N = \sum_{x=0}^{m} f_x = n + f_0$. The frequency f_0 needs to be estimated based on a valid model. In practice, the assumption of a homogeneous Poisson parameter is rarely satisfied. Over-dispersion in capture-recapture studies may result in underestimation of population size (Böhning, 2008b; Kake et al., 2008; Baksh et al., 2011; Bronner et al., 2013), whereas under-dispersion may lead to overestimated size population (Anderson, 2008; Cecala et al., 2012). Thus, it is significant to select the more suitable distributions allowing for the individual heterogeneity. The Conway-Maxwell-Poisson distribution has been used less frequently, compared with other generalised models. However, it is useful for heterogeneous capture probabilities.

4.2.1 The Conway-Maxwell-Poisson distribution

The CMP distribution is an extension of the Poisson distribution, suggested by Conway and Maxwell (1962). It generalises the Poisson distribution by adding an extra parameter ν, which accounts for the cases of over and under-dispersion. For the population size N, consider a sample of counts $X_1, X_2, X_3, ..., X_N \in \{0, 1, 2, 3, ...\}$, arising from the CMP distribution. Hence, the CMP estimator with two parameters has probability mass function as follows:

$$p_x = \frac{\lambda^x}{(x!)^\nu} \frac{1}{z(\lambda, \nu)}, \lambda > 0, \nu \geq 0, \tag{4.1}$$

where $x = 0, 1, 2, 3, ...,$ and the function $z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ denotes a normalisation constant. The CMP distribution has been ignored for a long time due to the complexity

in the infinite series involved in the normalising constant. However, some approximations have been provided as special cases of the CMP distribution that are given as:

1. For $\nu = 1$, we have that $z(\lambda, \nu) = \exp(\lambda)$. The CMP distribution reduces to an ordinary Poisson distribution with parameter λ.

2. For $\nu \to \infty$, we obtain $z(\lambda, \nu) \to 1 + \lambda$. The CMP distribution reaches a Bernoulli distribution with a parameter of success; $P(X = 1) = \frac{\lambda}{1+\lambda}$.

3. For $\nu = 0$ and $0 < \lambda < 1$, gives from the geometric series $z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j = \frac{1}{1-\lambda}$. Consequently, this is a geometric distribution with probability of successive $1 - \lambda$. Its distribution can be rewritten as $p_x = \lambda^x(1 - \lambda)$.

Note that $\nu = 0$ and $\lambda \geq 1$, $z(\lambda, \nu)$ does not converge and the distribution is not defined (see Shmueli et al., 2005; Gupta et al., 2014).

The Conway-Maxwell-Poisson distribution does not have a closed-form expression for its moments in term of the parameters λ and ν. Shmueli et al. (2005) suggested an approximation of its mean and variance by using an asymptotic approximation for $z(\lambda, \nu)$ as:

$$E(X) \approx \lambda^{\frac{1}{\nu}} + \frac{1}{2\nu} - \frac{1}{2}, \tag{4.2}$$

and

$$Var(X) \approx \frac{1}{\nu}\lambda^{\frac{1}{\nu}}. \tag{4.3}$$

It should be noted that both E(X) and Var(X) are increasing functions of λ but inversely related to ν. By substituting $\mu = \lambda^{\frac{1}{\nu}}$ in (4.2) and (4.3), then mean and variance take on the form

$$E(X) \approx \mu + \frac{1}{2\nu} - \frac{1}{2}, \tag{4.4}$$

and

$$Var(X) \approx \frac{\mu}{\nu}. \tag{4.5}$$

To study behaviour of the CMP model, the location parameter is fixed at $\lambda = 0.8$. The graphs in Figure 4.1 show the shape of the CMP distribution where λ is fixed, as the value of ν changes from $0 - 1.5$, and as N varies between 100 and $10,000$. The frequency distributions become longer tailed when the value ν decreases.

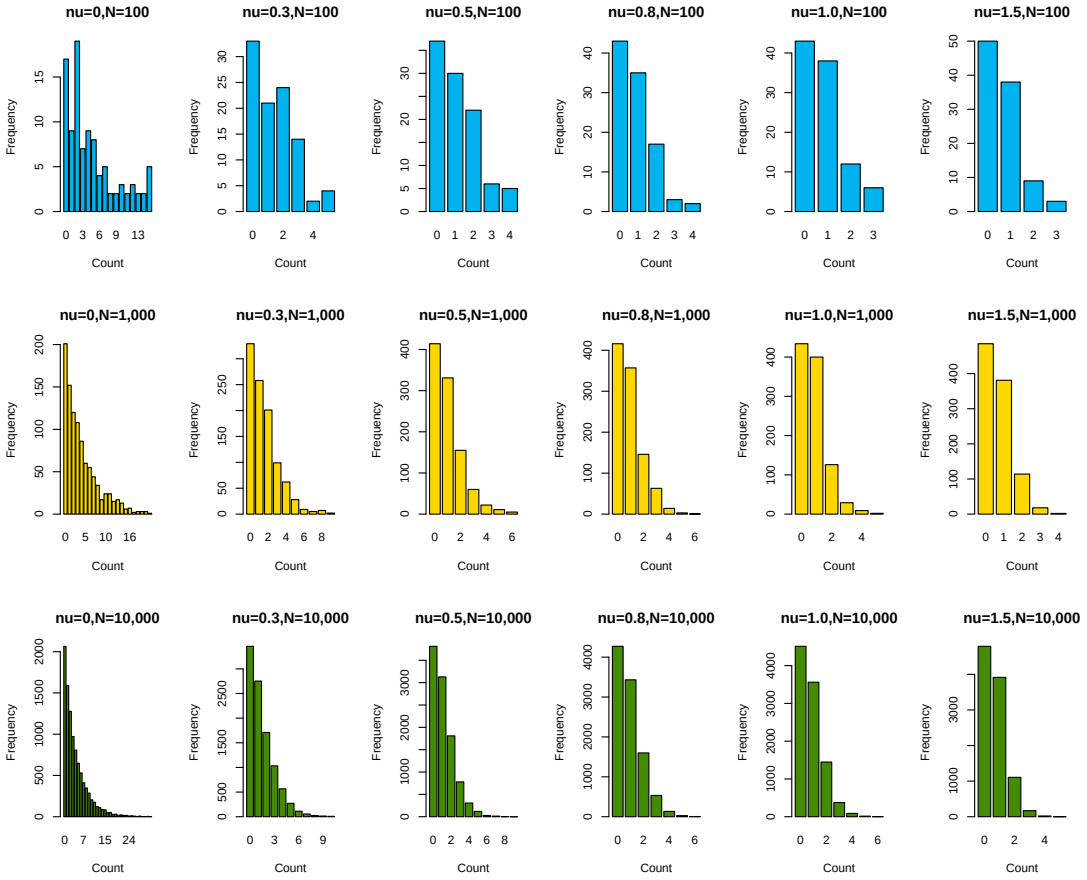

Figure 4.1: Simulated frequency distributions based on the Conway-Maxwell-Poisson with $CMP(0.8, \nu)$ and value of $\nu = 0$, 0.3, 0.5, 0.8, 1.0, 1.5 for $N = 100$, $1,000$ and $10,000$

4.2.2 The zero-truncated Conway-Maxwell-Poisson distribution

Only positive counts can be observed in capture-recapture data. Thus, unobserved units need to be estimated under parametric modelling, conditioning on the observed distribution or zero-truncated distribution. The zero-truncated Conway-Maxwell-Poisson (ZTCMP) distribution is defined by a probability function conditional on $X > 0$. Then the random variable X is the ZTCMP distribution if its p.m.f is given as

$$p_x^+ = \frac{p_x}{1 - p_0} = \frac{\frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}}{1 - \frac{1}{z(\lambda,\nu)}} = \frac{\frac{\lambda^x}{(x!)^\nu}}{z(\lambda,\nu) - 1}. \tag{4.6}$$

The likelihood function for the ZTCMP distribution is given by

$$L(x;\lambda,\nu) = \prod_{x=1}^{m} p_x = \prod_{x=1}^{m} \left\{ \frac{\frac{\lambda^x}{(x!)^\nu}}{z(\lambda,\nu) - 1} \right\}. \tag{4.7}$$

Its log-likelihood function is much simpler to work with:

$$\begin{aligned}\log L(x;\lambda,\nu) &= \sum_{x=1}^{m} \log\left\{\frac{\frac{\lambda^x}{(x!)^\nu}}{z(\lambda,\nu)-1}\right\} \\ &= \sum_{x=1}^{m}\left[\log\left(\frac{\lambda^x}{(x!\,)^\nu}\right) - \log\left\{z(\lambda,\nu)-1\right\}\right] \\ &= \sum_{x=1}^{m}\left[x\log\lambda - \nu\log x! - \log\left\{z(\lambda,\nu)-1\right\}\right]. \end{aligned} \tag{4.8}$$

Although the approximation of the normalisation $z(\lambda,\nu)$ has been discussed and approximated by mean of computational techniques, it represents a complication for the ZTCMP distribution. To avoid numerical issues, a simple graphical technique to estimate the two parameters of the ZTCMP distribution is considered.

4.2.3 Some properties of Conway-Maxwell-Poisson distribution for capture-recapture data

Dubey (1966); Ord (1967), and Holmes and Haggett (1977) proposed a graphical technique for detecting discrete distributions such as the binomial distribution, the Poisson distribution and the Pascal distribution. For statistical analysis, a graphical test is valuable for prior information concerning the frequency distribution due to the fact that it is a simple procedure and also uncomplicated to understand. The proposed method is based on a ratio of successive probabilities, which was originally suggested by Böhning et al. (2013a). The benefit of using the ratio plot is an identical property between the zero-truncated and the complete distributions. That is

$$r_x = (x+1)\frac{p_{x+1}}{p_x} = (x+1)\frac{p_{x+1}/(1-p_0)}{p_x/(1-p_0)}.$$

This property has been applied to investigate a suitable capture-recapture model by many researchers (see Rocchetti et al., 2011; Lanumteang and Böhning, 2011; Lerdsuwansri, 2012; Niwitpong et al., 2013; Rocchetti et al., 2014; Böhning, 2015).

The ratio plot for the CMP distribution is

$$\begin{aligned} r_x &= (x+1)\frac{p_{x+1}}{p_x} \\ &= (x+1)\frac{\frac{\lambda^{x+1}}{\{(x+1)!\}^\nu}\frac{1}{z(\lambda,\nu)}}{\frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}} \\ &= (x+1)\frac{\frac{\lambda^{x+1}}{\{(x+1)!\}^\nu}}{\frac{\lambda^x}{(x!)^\nu}} \\ &= (x+1)\frac{\lambda^x\lambda}{\lambda^x}\frac{(x!)^\nu}{\{(x+1)!\}^\nu} \\ &= (x+1)\frac{\lambda^x\lambda}{\lambda^x}\frac{(x!)^\nu}{(x+1)^\nu(x!)^\nu} \\ &= \frac{\lambda(x+1)}{(x+1)^\nu}. \end{aligned} \tag{4.9}$$

Importantly, the formula (4.9) does not depend on the complex normalising constant $z(\lambda,\nu)$ and allows for a non-linear relation between the ratio of successive probabilities and the count x. Using a logarithm transformation of both sides, it becomes a linear model as follows:

$$\begin{aligned} \log\{r_x\} &= \log\{(x+1)\frac{p_{x+1}}{p_x}\} \\ &= \log\left\{\frac{\lambda(x+1)}{(x+1)^\nu}\right\} \\ &= \log\lambda + \log(x+1) - \nu\log(x+1) \\ &= \log\lambda + (1-\nu)\log(x+1) \\ &= \beta_0 + \beta_1\log(x+1). \end{aligned} \tag{4.10}$$

Here there are using two new parameters β_0 and β_1, such that $\log\lambda = \beta_0$ or $\lambda = \exp(\beta_0)$, and $1-\nu = \beta_1$ or $\nu = 1-\beta_1$. However, some restrictions on the parameters based on the CMP distribution have to be concerned. The first restriction is $\lambda > 0$ in the CMP distribution which is always satisfied when $\beta_0 \in (-\infty, +\infty)$. Another constraint is the dispersion parameter $\nu \geq 0$ or $1-\nu \leq 1$, leading to limitation of the regression parameter $\beta_1 \leq 1$. Furthermore, it is necessary to mention that if $\beta_1 = 0$, then $\nu = 1$, and then the CMP distribution is reduced to the Poisson distribution. The linear regression for the Poisson distribution can be expressed this case as:

$$\begin{aligned} \log\{r_x\} &= \log\lambda \\ &= \beta_0, \end{aligned} \tag{4.11}$$

which is a constant. Additionally, the CMP distribution goes to the geometric distribution when $\beta_1 = 1$ and $0 < \lambda < 1$. The linear regression for the geometric distribution

can be expressed as

$$\begin{aligned} \log\{r_x\} &= \log\lambda + \log(x+1) \\ &= \beta_0 + \log(x+1), \end{aligned} \tag{4.12}$$

which is a linear regression with slope 1. In practice, the capture probabilities are approximated by relative frequencies, therefore, the ratio in (4.9) can be computed as

$$r_x^* = (x+1)\frac{\widehat{p}_{x+1}}{\widehat{p}_x} = (x+1)\frac{f_{x+1}/N}{f_x/N} = (x+1)\frac{f_{x+1}}{f_x}, \tag{4.13}$$

similarly, the ratio in (4.10), (4.11) and (4.12) can be computed as

$$\log\{r_x^*\} = \log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\}. \tag{4.14}$$

The graph of $\log(r_x^*)$ against $\log(x+1)$ can be used as a diagnostic tool for detecting the validity of the Conway-Maxwell-Poisson model and is called *the log-ratio plot*. It can be clearly assumed that if the log-ratio plot shows a straight line with a positive slope, it can be taken as indicative for the presence of the (zero-truncated) Conway-Maxwell-Poisson model in the case of an over-dispersion. On the other hand, in the event of the Conway-Maxwell-Poisson model for an under-dispersion, the log-ratio plot is indicated by a straight line with a negative slope. The final case is that the log-ratio plot represents a horizontal line, this is evidence of the Conway-Maxwell-Poisson distribution with equi-dispersion, or the original Poisson distribution.

4.3 Model based inference

The use of the log ratio in (4.10) goes beyond a simple graphical technique to check for under or over-dispersion for capture-recapture data. Indeed, it can be used as a tool for estimating model parameters. Thus, let us consider the basic equation (4.10), and fit the following model

$$\log(r_x^*) = \underbrace{\beta_0 + \beta_1\log(x+1)}_{Systematic} + \underbrace{\epsilon_x}_{Random}, \tag{4.15}$$

where β_0 and β_1 are the intercept and the slope parameters respectively, and ϵ_x is the error term. Commonly, a least-squares estimation (LS) method is used to provide estimates of β_0 and β_1. However, model (4.15) does not satisfy the classical linear regression assumptions. The reason is that the observed frequencies for capture-recapture data often have $f_1 >> f_2 > f_3 > \ldots$, and heteroscedasticity might occur in a heterogeneous population due to e.g. unobserved information (see Rocchetti et al., 2011). This issue is relevant and should be accounted for. Thus, these issues are addressed by using weighted least squares (WLS) techniques to estimate the regression parameters β_0 and β_1, and

accordingly λ and ν. These are obtained by minimising

$$\sum_{x=1}^{m-1} W_x \left[\log(r_x^*) - \beta_0 - \beta_1 \log(x+1)\right]^2,$$

where W_x denotes the x-th element of an appropriate weights matrix. In other words, we take

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}, \tag{4.16}$$

where

$$\mathbf{Y} = \begin{pmatrix} \log\frac{2f_2}{f_1} \\ \log\frac{3f_3}{f_2} \\ \vdots \\ \log\frac{mf_m}{f_{m-1}} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & \log(2) \\ 1 & \log(3) \\ \vdots & \vdots \\ 1 & \log(m) \end{pmatrix},$$

and m is the maximum count used in the estimator.

The application of weighted least square requires the specification of $\mathbf{W} \approx cov(\mathbf{Y})^{-1}$ to reduce the mean square error. Following Meurant (1992) and Rocchetti et al. (2011), covariances between adjacent log-ratios did not play a large role in reducing mean square error, and thus it was suggested dropping the off-diagonal terms in $cov(\mathbf{Y})$ in approximating $\mathbf{W}$, with little loss of efficiency. Accordingly, let matrix $\mathbf{W}$ be a diagonal matrix containing weights:

$$\mathbf{W} = \begin{bmatrix} \frac{1}{f_1}+\frac{1}{f_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_2}+\frac{1}{f_3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{f_{m-1}}+\frac{1}{f_m} \end{bmatrix}^{-1}. \tag{4.17}$$

To see that (4.17) is the right choice, let $w_x = [Var\{\log(r_x^*)\}]^{-1}$, we have

$$\begin{aligned} Var\left\{\log(r_x^*)\right\} &= Var\left[\log\left\{(x+1)\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right\}\right] \\ &= Var\left\{\log(x+1) + \log(\widehat{p}_{x+1}) - \log(\widehat{p}_x)\right\} \\ &= Var\left\{\log(\widehat{p}_{x+1})\right\} + Var\left\{\log(\widehat{p}_x)\right\} - 2Cov\left\{\log(\widehat{p}_{x+1}), \log(\widehat{p}_x)\right\}. \end{aligned}$$

Using the delta method for variance estimation, that is

$$\begin{aligned} Var\left\{\log(r_x^*)\right\} &\approx \frac{1}{\widehat{p}_{x+1}^2} Var(\widehat{p}_{x+1}) + \frac{1}{\widehat{p}_x^2} Var(\widehat{p}_x) - \frac{2Cov(\widehat{p}_{x+1}, \widehat{p}_x)}{(\widehat{p}_{x+1})\widehat{p}_x} \\ &= \frac{1}{\widehat{p}_{x+1}^2}\left\{\frac{\widehat{p}_{x+1}(1-\widehat{p}_{x+1})}{n}\right\} + \frac{1}{\widehat{p}_x^2}\left\{\frac{\widehat{p}_x(1-\widehat{p}_x)}{n}\right\} + \frac{\frac{2(\widehat{p}_{x+1})\widehat{p}_x}{n}}{(\widehat{p}_{x+1})\widehat{p}_x} \\ &= \frac{1-\widehat{p}_{x+1}}{n\widehat{p}_{x+1}} + \frac{1-\widehat{p}_x}{n\widehat{p}_x} + \frac{2}{n} \\ &= \frac{1}{n\widehat{p}_{x+1}} - \frac{\widehat{p}_{x+1}}{n\widehat{p}_{x+1}} + \frac{1}{n\widehat{p}_x} - \frac{\widehat{p}_x}{n\widehat{p}_x} + \frac{2}{n} \\ &= \left(\frac{1}{n\widehat{p}_{x+1}} + \frac{1}{n\widehat{p}_x}\right) + \left(\frac{2}{n} - \frac{\widehat{p}_{x+1}}{n\widehat{p}_{x+1}} - \frac{\widehat{p}_x}{n\widehat{p}_x}\right) \\ &= \left(\frac{1}{n\widehat{p}_{x+1}} + \frac{1}{n\widehat{p}_x}\right) + \left(\frac{2}{n} - \frac{1}{n} - \frac{1}{n}\right), \end{aligned} \tag{4.18}$$

where n is the number of observations from the target population. Therefore, the variance of log-ratio is given by

$$Var\left\{\log(r_x^*)\right\} \approx \frac{1}{n\widehat{p}_{x+1}} + \frac{1}{n\widehat{p}_x}.$$

In practice, $\widehat{p}_{x+1}$ and $\widehat{p}_x$ can be estimated by relative observed frequency $\frac{f_{x+1}}{n}$ and $\frac{f_x}{n}$, respectively. Hence

$$\widehat{Var}\left\{\log(r_x^*)\right\} = \frac{1}{n\frac{f_x}{n}} + \frac{1}{n\frac{f_{x+1}}{n}} = \frac{1}{f_x} + \frac{1}{f_{x+1}}.$$

Thus, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained from (4.16), in which $\mathbf{W}$ is given by (4.17). Accordingly, the unknown f_0 can be then estimated by considering that

$$\begin{aligned} \log\left(\frac{f_1}{f_0}\right) &= \widehat{\beta}_0 \\ \frac{f_1}{f_0} &= \exp\left(\widehat{\beta}_0\right) \\ \widehat{f}_0 &= f_1 \exp\left(-\widehat{\beta}_0\right), \end{aligned} \tag{4.19}$$

where $\widehat{f}_0$ is the estimator of the unobserved frequency. Hence, a target population size estimator using a linear regression estimation based on the Conway-Maxwell-Poisson distribution (LCMP) can be readily estimated as

$$\widehat{N}_{LCMP} = n + \widehat{f}_0 = n + f_1 \exp(-\widehat{\beta}_0). \tag{4.20}$$

We also obtain an estimated probability of the count to be zero (unobserved) as:

$$\widehat{p}_0 = \frac{\widehat{f}_0}{\widehat{N}_{LCMP}}. \tag{4.21}$$

We point out here that the Conway-Maxwell-Poisson distribution includes as a special case the geometric ($\nu = 0$) so that an associated weighted least squares estimator is available for the geometric. In detail, the linear regression fitted based on the geometric distribution is given as

$$\log(r_x^*) = \beta_0 + \log(x+1) + \varepsilon_x,$$

using weighted lest squares (WLS) technique to estimate linear regression coefficients, required minimisation of the sum of square residuals (SSR),

$$\begin{aligned} SSR = \sum_{x=1}^{m-1} W_x \varepsilon_x^2 &= \sum_{x=1}^{m-1} W_x \left\{\log(r_x) - \beta_0 - \log(x+1)\right\}^2 \\ &= \sum_{x=1}^{m-1} W_x \left\{\log\left((x+1)\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right) - \beta_0 - \log(x+1)\right\}^2 \\ &= \sum_{x=1}^{m-1} W_x \left\{\log(x+1) + \log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right) - \beta_0 - \log(x+1)\right\}^2 \\ &= \sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right) - \beta_0\right\}^2 \\ \frac{\partial(SSR)}{\partial\beta_0} &= -2\sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right) - \beta_0\right\}. \end{aligned} \tag{4.22}$$

Setting (4.22) equal to zero,

$$\begin{aligned} -2\sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right) - \beta_0\right\} &= 0 \\ \beta_0 \sum_{x=1}^{m-1} W_x &= \sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right)\right\} \\ \beta_0 &= \frac{\displaystyle\sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{\widehat{p}_{x+1}}{\widehat{p}_x}\right)\right\}}{\displaystyle\sum_{x=1}^{m-1} W_x}. \end{aligned}$$

In practice, the capture probabilities can be approximated by the relative frequencies, leading to the following simplified form:

$$\widehat{\beta}_0 = \frac{\displaystyle\sum_{x=1}^{m-1} W_x \left\{\log\left(\frac{f_{x+1}}{f_x}\right)\right\}}{\displaystyle\sum_{x=1}^{m-1} W_x},$$

where W_x is the $x-$th diagonal element of (4.17). Besides, the weighted lest square regression can be modified for the original Poisson distribution case by

$$\widehat{\beta}_0 = \frac{\sum_{x=1}^{m-1} W_x \left\{ \log\left((x+1)\frac{f_{x+1}}{f_x} \right) \right\}}{\sum_{x=1}^{m-1} W_x}.$$

Next, it is shown that the proposed LCMP estimator is asymptotically unbiased under the Conway-Maxwell-Poisson distribution as well as for the spacial cases, the Poisson and the geometric distributions, as follows:

Theorem 4.1. *The LCMP estimator is an asymptotically unbiased estimator under the Conway-Maxwell-Poisson distribution:* $p_x = \frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}$ *where* $\nu \geq 0$ *and* $\lambda > 0$.

We anticipate that $\widehat{N}_{LCMP}$ is asymptotically unbiased in the sense

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{LCMP})}{N} \to 1,$$

if the sample arises from the Conway-Maxwell-Poisson distribution.

Proof. $E(f_1/N)$ converges with $N \to \infty$ to p_1. Suppose that β_0 would be known, then

$$E\left(\frac{\widehat{f}_{0,LCMP}}{N}\right) = E\{(f_1 \exp(-\widehat{\beta}_0))/N\}$$

and, as $\exp(-\widehat{\beta}_0)$ converges to $\exp(-\beta_0) = 1/\lambda$, we have that

$$E\left(\frac{\widehat{f}_{0,LCMP}}{N}\right) \to \exp(-\beta_0)p_1 = \frac{p_1}{\lambda} = p_0 = \frac{1}{z(\lambda,\nu)}.$$

As a consequence, $E(\widehat{N}_{LCMP}/N) = E\{(n+\widehat{f}_0)/N\} = E(n/N) + E(\widehat{f}_0/N)$ converges to $\left(1 - \frac{1}{z(\lambda,\nu)}\right) + \frac{1}{z(\lambda,\nu)} = 1$. □

Theorem 4.2. *The LCMP estimator is an asymptotically unbiased estimator under the geometric distribution;* $p_x = \lambda^x(1-\lambda)$ *which is a specific case of the CMP distribution for* $\nu = 0$ *and* $0 < \lambda < 1$.

We have that

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{LCMP})}{N} \to 1$$

Theorem 4.3. *The LCMP estimator is an asymptotically unbiased estimator under the Poisson distribution,* $p_x = \frac{\exp(-\lambda)\lambda^x}{x!}$ *which is a specific case of the CMP distribution for* $\nu = 1$.

We have that

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{LCMP})}{N} \to 1$$

A small simulation study below compares the levels of precision and accuracy between weighted least squares and unweighed least squares for the simple linear regression estimation by computing the bias, variance and mean squared error of $\widehat{N}_{LCMP}$ as follows.

Example 1 Frequency data were drawn from a Conway-Maxwell-Poisson with parameters $\lambda = 1.5$ and $\nu = 0.4$. The number of replications is $1,000$, and Table 4.1 shows results for $N = 100, 500, 1,000, 5,000$ and $10,000$.

Table 4.1: The performance of weighted least square and unweighted least squares estimators for CMP(1.5,0.4)

N	Bias($\widehat{N}$)		Var($\widehat{N}$)		MSE($\widehat{N}$)	
	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted
100	2.61	3.46	43.99	41.92	50.82	53.89
500	2.61	6.70	197.90	222.14	204.69	267.07
1,000	3.54	9.13	374.63	494.71	387.15	577.99
5,000	4.69	0.90	1,821.59	4,640.23	1,843.58	4,641.03
10,000	3.39	-7.20	3,700.52	11,327.70	3,711.99	11,379.48

The simulation result from Table 4.1 provides evidence that the weighted least squares estimator is more accurate and precise in the estimation of population size than the unweighted lest squares estimator, therefore, the weighted least squares method is preferred to estimate parameters of the zero-truncated Conway-Maxwell-Poisson distribution using a linear regression approach.

4.4 Simulation study

This section provides a comprehensive assessment of population size estimators performance. The LCMP estimator proposed in this chapter is compares with other well-established estimators highlighted in the literature review. The simulation study is designed to cover scenarios with different underlying null models, with varying population sizes: $N = 100, 250$ for small sizes, $N = 500, 1,000$ for medium sizes, and $N = 5,000, 10,000$ for large sizes, and degree of dispersion is measured by the inverse of dispersion parameter ν (see Anan et al., 2016). Random variable X_i where $i = 1, 2, 3, ...N$ is generated following the Poisson, the geometric, the CMP and the negative distributions as follows:

4.4.1 Simulation scenarios

In detail, the following data generation settings are used in the simulation study:

i) The Poisson distribution: counts are generated from the Poisson distribution with parameters

$$\lambda \in \{0.5, 0.8, 1.0, 1.2, 1.5, 2.0\}.$$

ii) The geometric distribution: counts are generated from the geometric distribution with parameters

$$\lambda \in \{0.2, 0.5, 0.8\}.$$

where $\lambda = 1 - p$, and p is a probability of success.

iii) The Conway-Maxwell-Poisson distribution: counts are generated from Conway-Maxwell-Poisson distribution with parameters

$$\lambda \in \{0.8, 1.0, 1.2\},$$

and

$$\nu \in \{0.1, 0.5, 0.8, 1.25\}$$

iv) The negative binomial distribution: the negative binomial contains as a special case, the Poisson distribution as $k \to \infty$ and the geometric for $k = 1$. Then, counts are generated from a negative binomial distribution

$$p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}(1-\lambda)^k\lambda^x,$$

with parameters

$$\lambda \in \{0.2, 0.4, 0.6, 0.8\},$$

dispersion parameters

$$k \in \{2, 3, 4\},$$

expected value and variance are given respectively by

$$E(X) = \frac{k\lambda}{1-\lambda} = \mu,$$

and

$$Var(X) = \frac{k\lambda}{(1-\lambda)^2} = \mu + \frac{1}{k}\mu^2.$$

Population size is estimated conditioning on the observed data with the zero counts ($x_i = 0$) truncated before the estimation procedure. As the settings i)-ii)-iii) cover situations where the data generation is from a special case of the CMP distribution, we include setting iv) to investigate what happens when data follows a distribution outside

the CMP family, e.g. a negative binomial distribution. We draw T=5,000 samples from each *null* model. Any occurrences of zero counts are truncated, and five estimators of population size are compared: Turing's estimator (Turing), the maximum likelihood estimation under the zero-truncated Poisson model (MLEPoi), Chao's lower bound estimator (Chao), Zelterman's estimator (Zel) and weighted linear regression estimator under the zero-truncated Conway-Maxwell-Poisson model (LCMP). Three measurements: the relative bias (RBias), the relative variance (RVar) and the relative root mean square error (RRMSE) of population size estimations, are used to evaluate the performance of the population size estimators.

4.4.2 Simulation results when data are generated from the Poisson distribution

The homogeneity model assumes that each individual of the target population has the same capture probabilities and follows the Poisson distribution. The proposed LCMP estimator is hypothesised as the alternative choice to estimate the hidden population size based on the homogeneous Poisson distribution since the Poisson is a sub-model of the CMP distribution.

The simulation results show that all five estimators are asymptotically unbiased for the Poisson distribution with respect to the population size as we can see that the relative biases (RBias) converge to zero (see Figure 4.2). Indeed, the LCMP, the MLEPoi and the Turing estimators for the small populations ($N \leq 250$) have a smaller bias and variance than Chao and Zelterman estimators. The MLEPoi estimator gives the smallest variance of population size as was found in Lanumteang (2010), however, it tends to be identical to the LCMP and the Turing estimators when λ and/or the population size increases.

Given the results on bias and variance of population size, it is no surprise that the LCMP, the Turing and the MLEPoi estimators perform the best when considering the RRMSE. On the other hand, the estimator of Zelterman and Chao, which allow for population heterogeneity, are persistently biased and provide the highest RRMSE for all conditions (see Figure 4.3 and Figure 4.4).

Note that population size is not the only factor influencing the performance of the estimators. Increasing parameter λ leads to a decrease in both bias and variance for all estimators. This may be due to the fact that the observed sample n is approaching the population size, N.

1) Relative bias of population size estimators when data are generated from the Poisson distribution

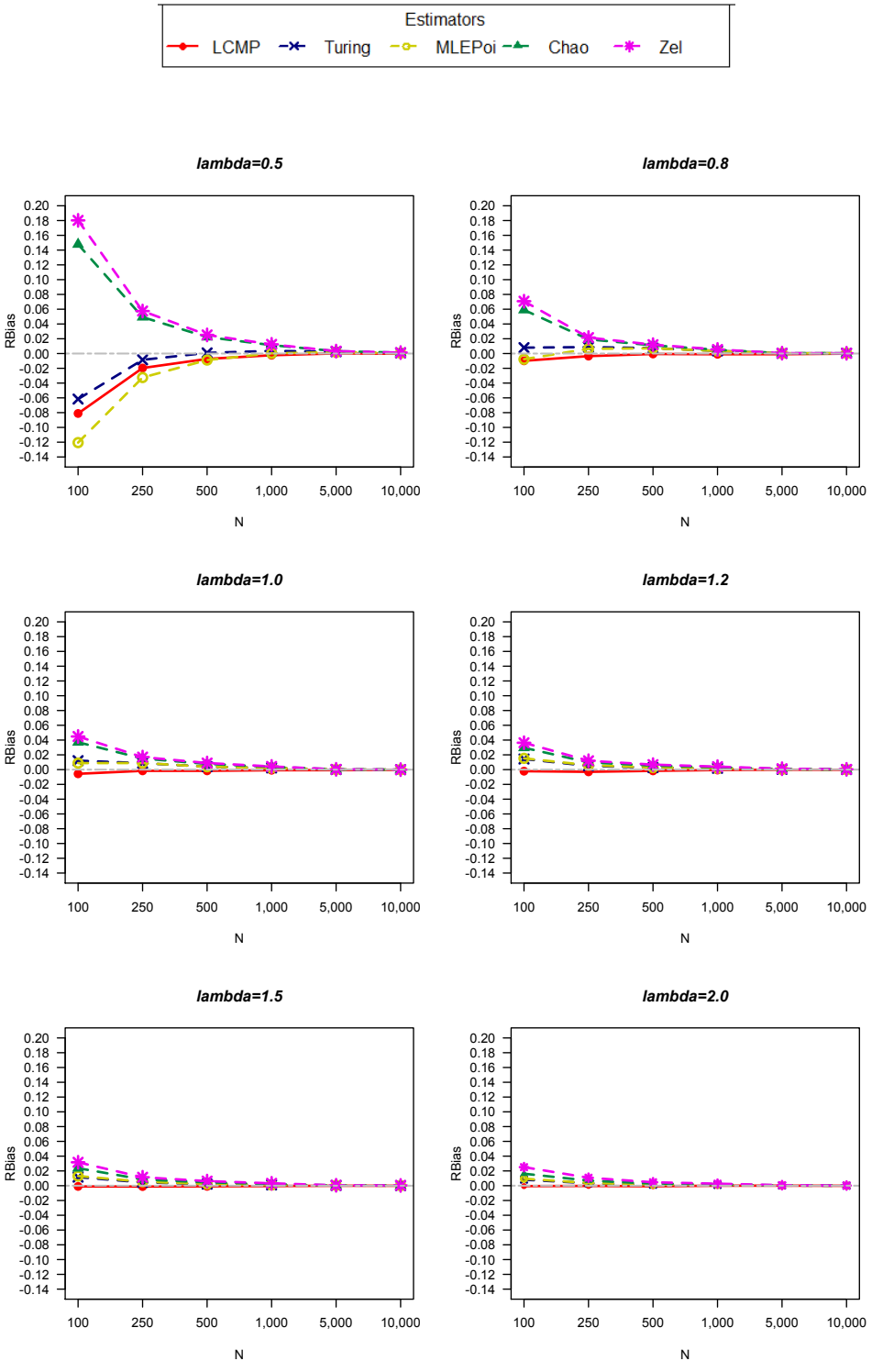

Figure 4.2: Relative bias of five estimators for counts drawn from $Poi(\lambda)$

2) Relative variance of population size estimators when data are generated from the Poisson distribution

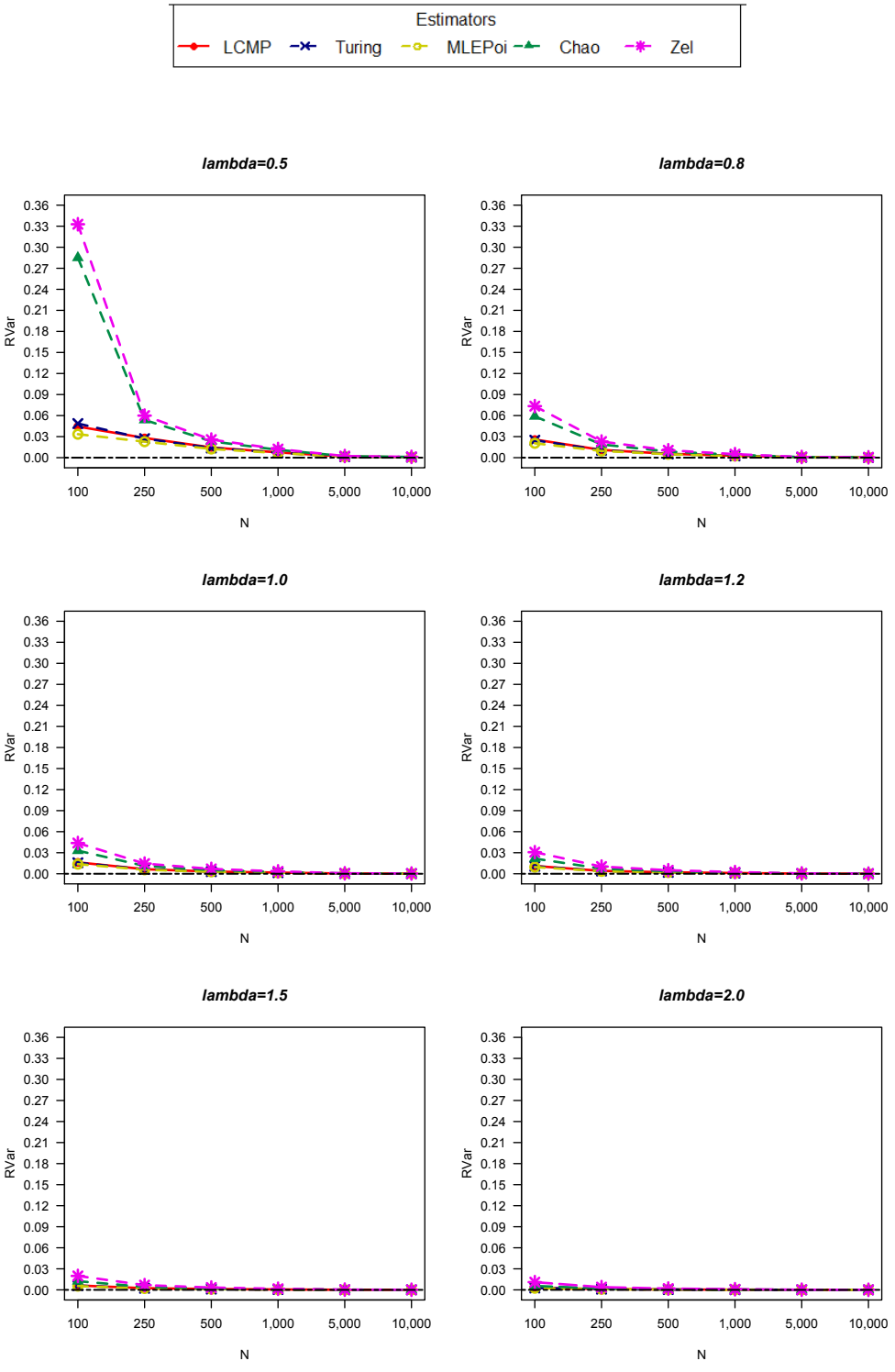

Figure 4.3: Relative variance of five estimators for counts drawn from $Poi(\lambda)$

3) Relative root mean square error of population size estimators when data are generated from the Poisson distribution

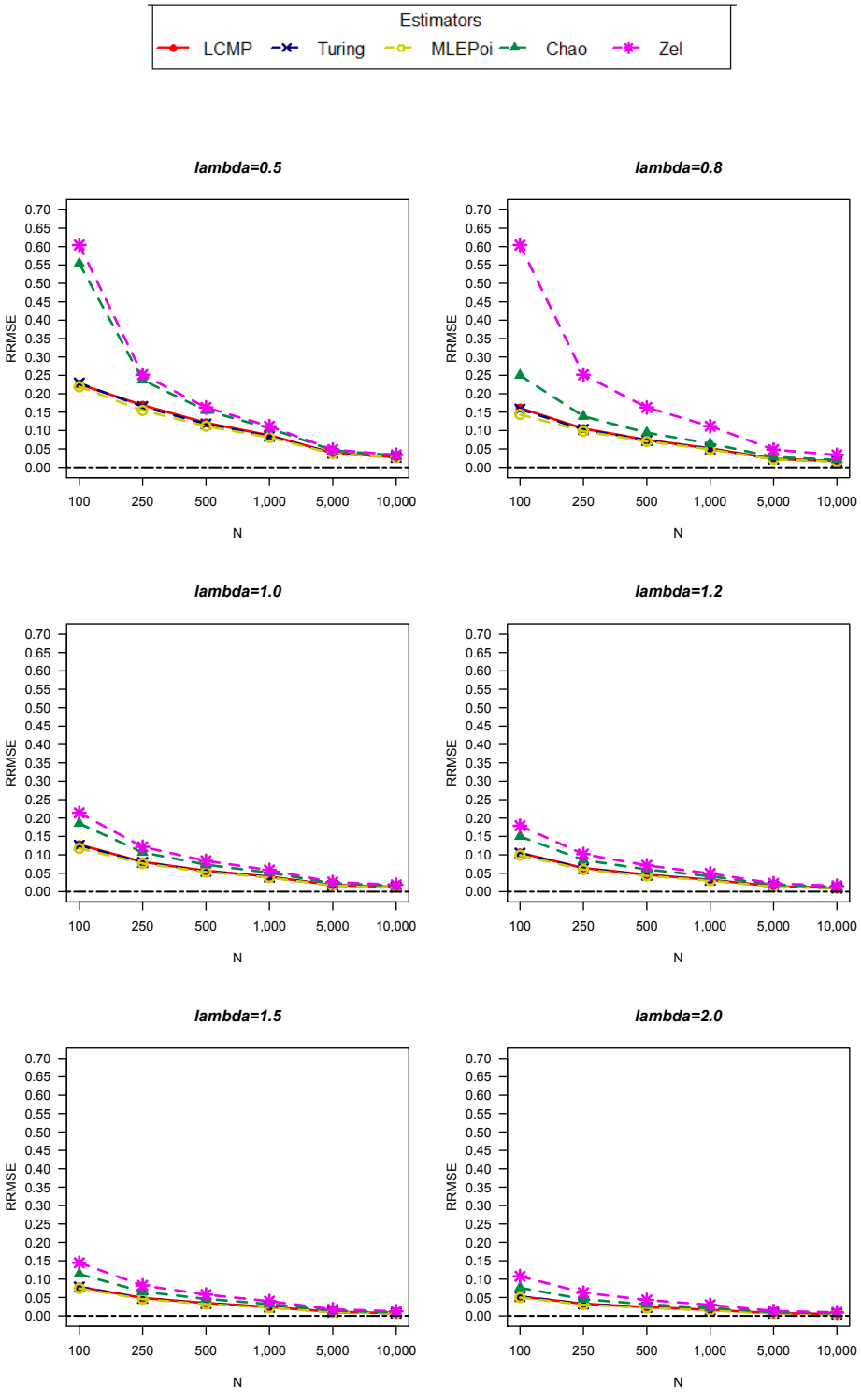

Figure 4.4: Relative root mean square error of five estimators for counts drawn from $Poi(\lambda)$

4.4.3 Simulation results when data are generated from the geometric distribution

In reality, the Poisson distribution is often not appropriate in the capture-recapture study. The geometric distribution is an excellent alternative to the Poisson as the variance is greater than the mean. As the CMP distribution can be considered as a generalisation of the geometric when $\nu = 0$ and $0 < \lambda < 1$ therefore, it can be assumed that the LCMP estimator is one of the best choices for estimating population size in the geometric model.

The simulation results in Figure 4.5 provide evidence that only the LCMP represents an asymptotic unbiased estimator based on the geometric distribution. Overall, other estimators give an underestimation of population size except for the Zelterman estimator when the location parameter λ value is high. The Zelterman tends to agree with the LCMP estimator for increasing value of λ. This means that the observed counts increase.

According to Figure 4.6, the relative variance from the LCMP estimator is higher than the MLEPoi and the Turing estimators. However, it tends to get close to zero when λ or/and population size increase. From the RBias and Rvar above, it is no surprise that the LCMP estimator clearly shows the best performance for the geometric-based model as its RRMSE is smallest for all situations (see Figure 4.7).

1). Relative bias of population size estimators when data are generated from the geometric distribution

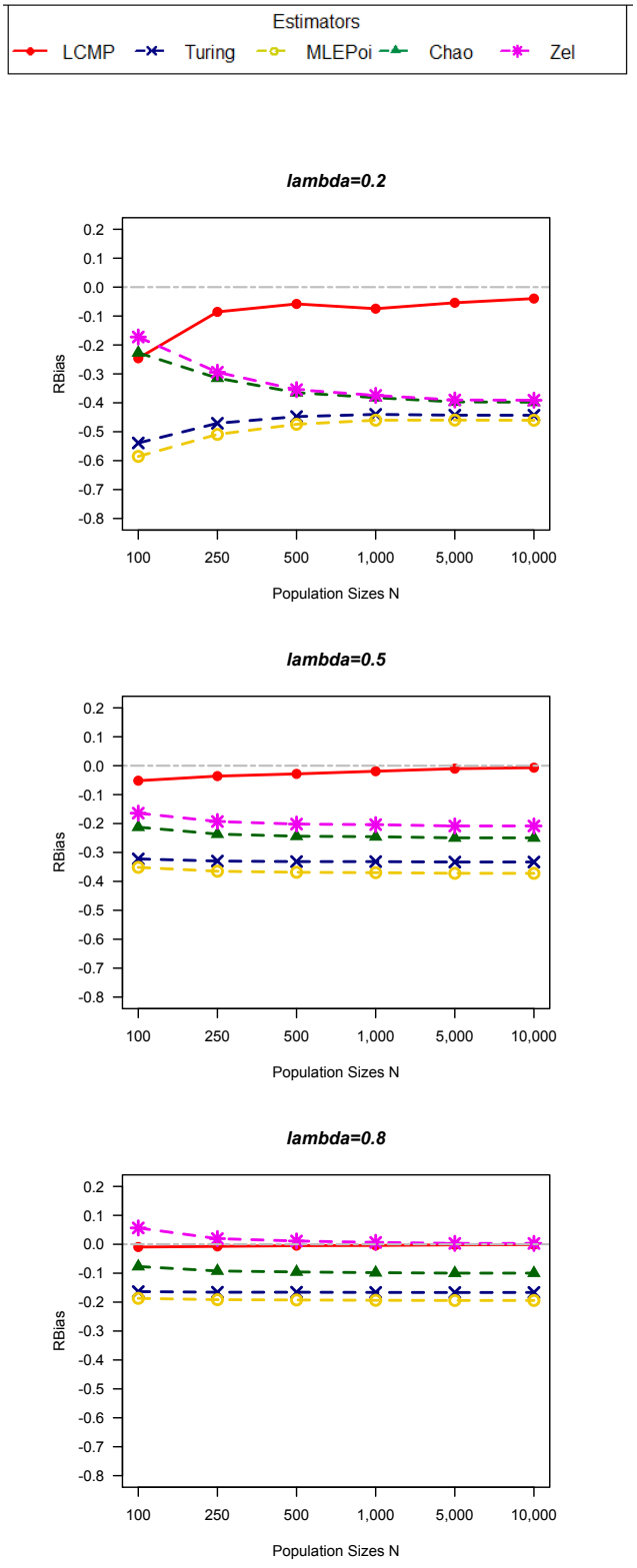

Figure 4.5: Relative bias of five estimators for counts drawn from $Geo(\lambda)$

2). Relative variance of population size estimators when data are generated from the geometric distribution

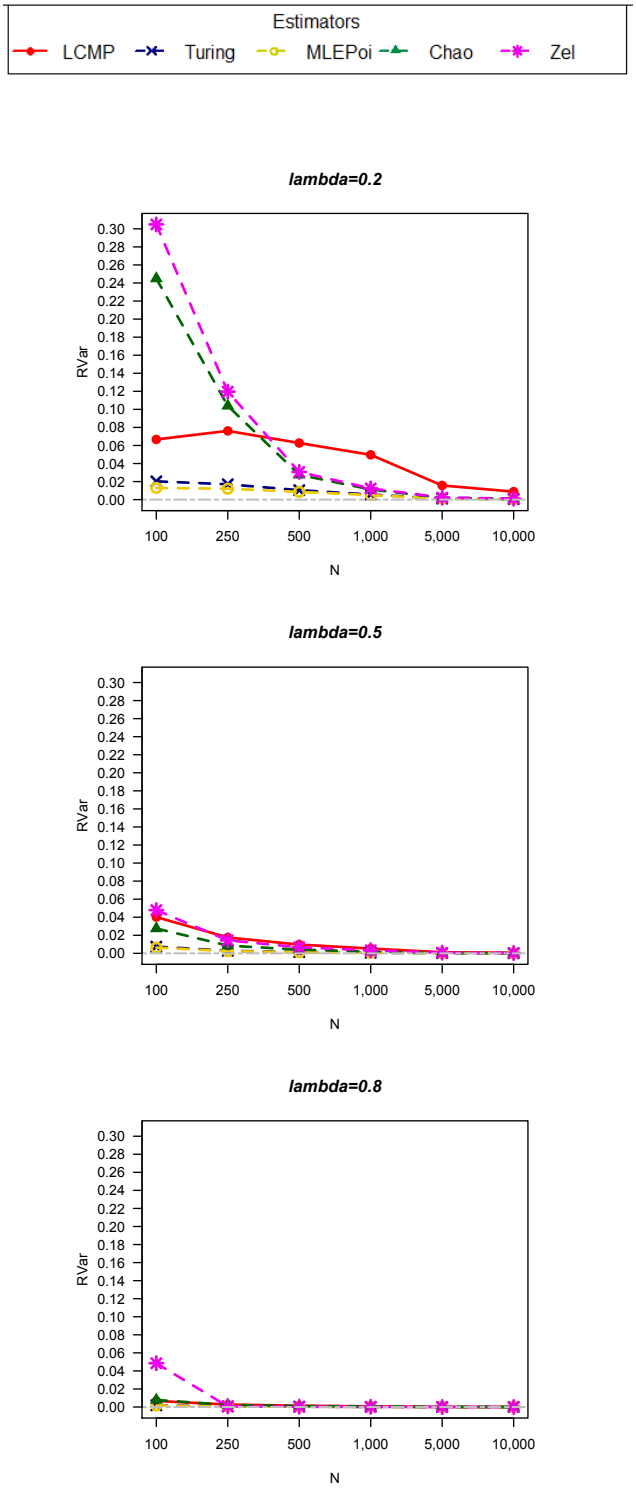

Figure 4.6: Relative variance of five estimators for counts drawn from $Geo(\lambda)$

3). Relative root mean square error of population size estimators when data are generated from the geometric distribution

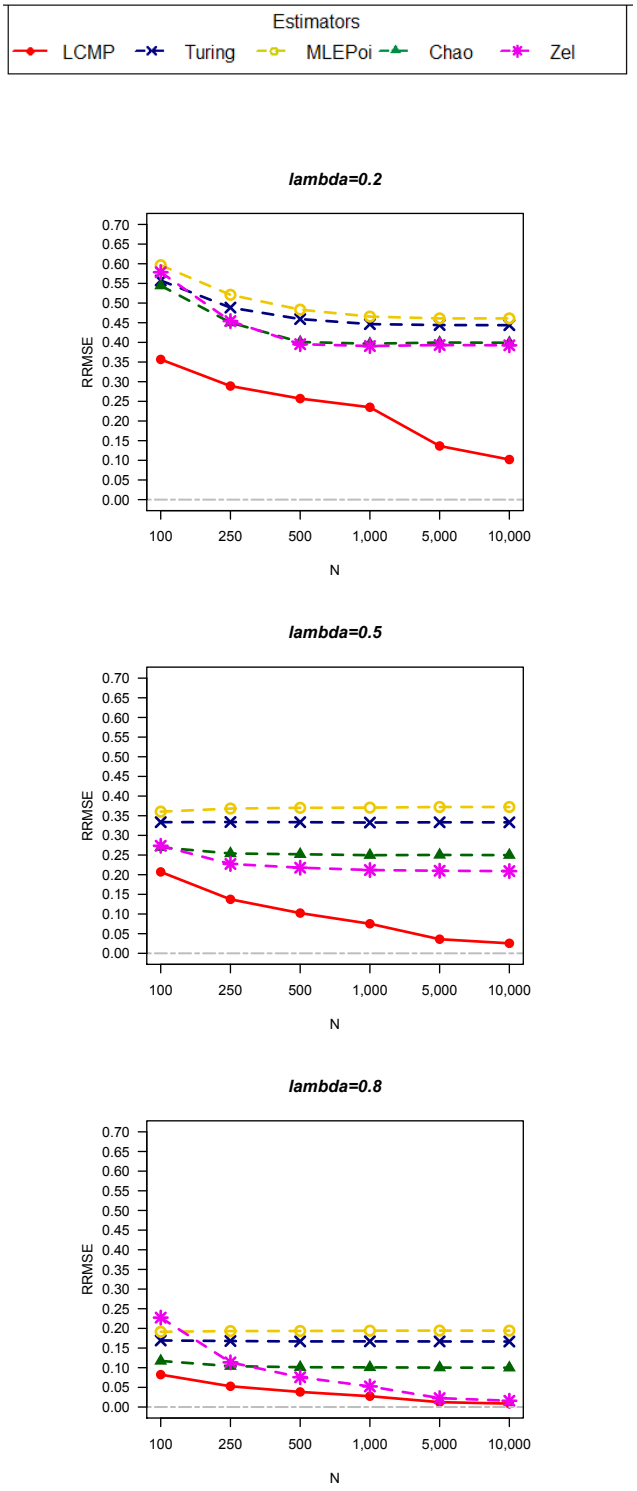

Figure 4.7: Relative root mean square error of five estimators for counts drawn from $Geo(\lambda)$

4.4.4 Simulation results based on the data generated from the Conway-Maxwell-Poisson distribution

As over-dispersion is common in capture recapture data, to study estimator's performance in the presence of both over- and under-dispersion, data are also generated from the CMP distribution. It is expected that LCMP, Chao and Zelterman estimators outperform the Turing and the MLEPoi estimators under heterogeneous simulation schemes. Simulation results are provided in Figures 4.8, 4.9, and 4.10.

The simulation results point out that while samples are generated by the CMP distribution, the LCMP estimator is not doing well generally. According to the investigation of the accuracy of parameters, it can be seen that the LCMP estimator provides an asymptotic unbiased estimator with respect to the population size. It has the least bias when a dispersion parameter ν is close to zero. However, the LCMP estimator has reduced accuracy for small populations and as ν increases, indicating a lack of fit for short-tail frequency distributions. Also, other estimators based on heterogeneity as well as homogeneity tend to underestimate the population size in the over-dispersion case, and overestimate for the under-dispersion scenario.

The simulation results also point out that not only biases but also the variances of proposed estimator increase when the value of dispersion parameter increases. Indeed, the LCMP estimator for under-dispersion starts to do well for $1,000 \leq N \leq 5,000$ and performs the best when $N > 5,000$.

Overall, the use of LCMP estimator is suitable in the presence of strong over-dispersion or strong heterogeneity as we can see that the LCMP estimator shows the most accurate and precise behaviour leading to the best performance with smallest RRMSE for $\nu = 0.1$. However, it is also useful for moderate to weak over-dispersion and under-dispersion if the population size is large enough.

1) Relative bias of population size estimators when data are generated from the CMP distribution

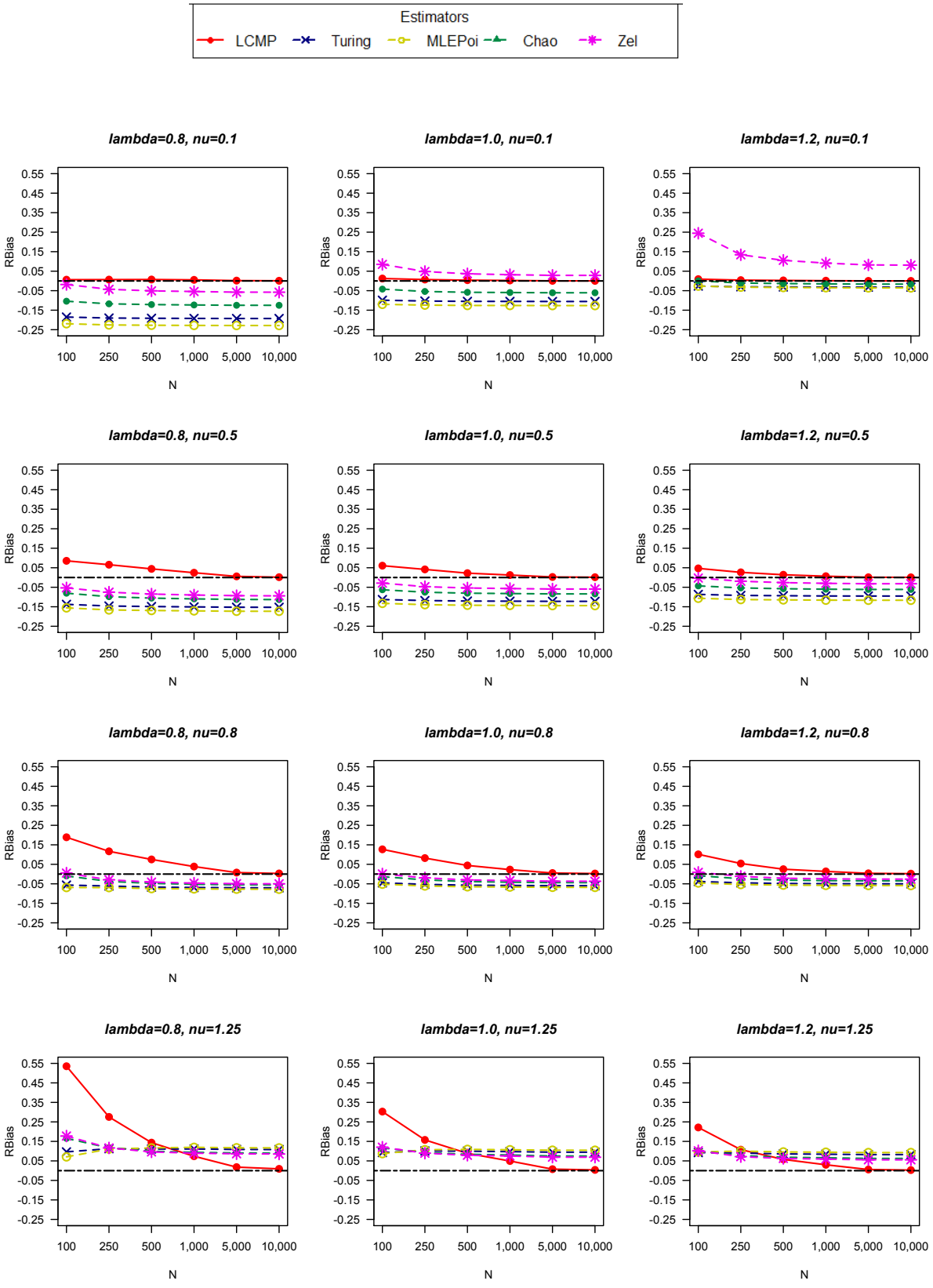

Figure 4.8: Relative bias of five estimators for counts drawn from $CMP(\lambda, \nu)$

2) Relative variance of population size estimators when data are generated from the CMP distribution

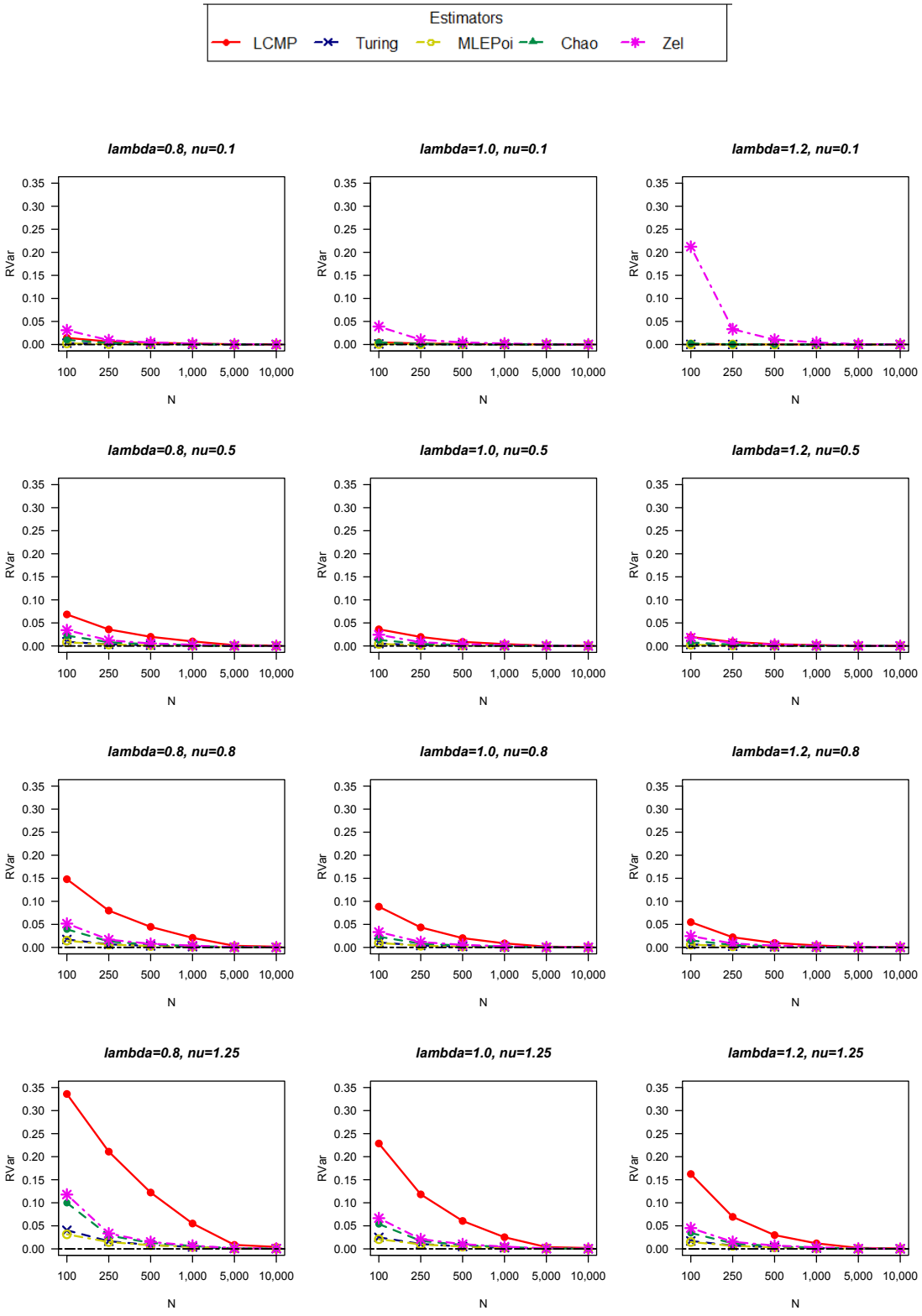

Figure 4.9: Relative variance of five estimators for counts drawn from $CMP(\lambda, \nu)$

3) Relative root mean square error of population size estimators when data are generated from the CMP distribution

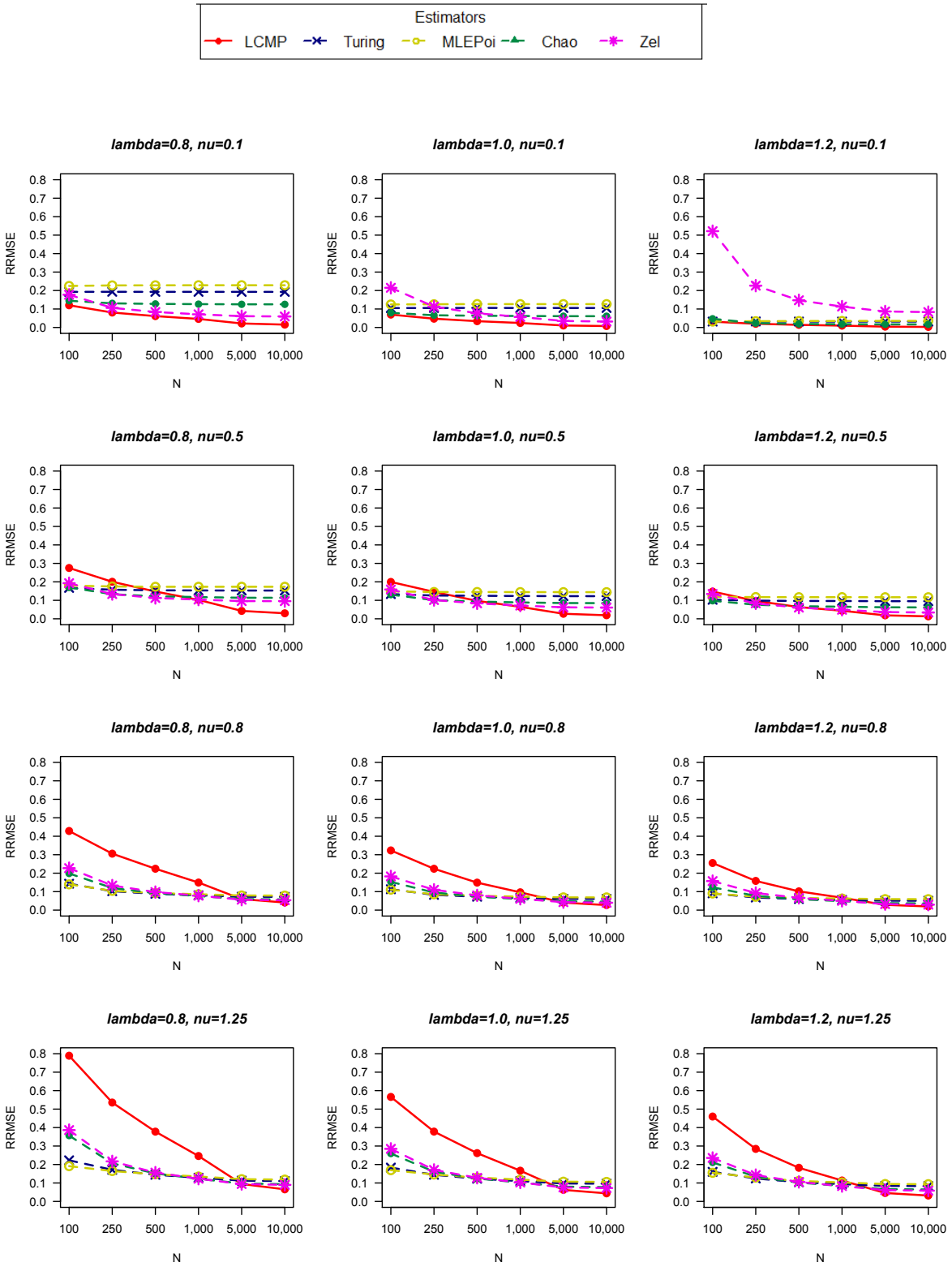

Figure 4.10: Relative root mean square error of five estimators for counts drawn from $CMP(\lambda, \nu)$

4.4.5 Simulation results when data are generated from the negative binomial distribution

In practice, the negative binomial is often chosen for fitting count data in the case of over-dispersion relative to the Poisson distribution. Therefore, the negative binomial is used for capture-recapture data analysis by many authors. The limitation is as pointed out before that there might be no feasible value of the dispersion parameter k. It is expected the LCMP estimator can be used for estimating population size because the negative binomial can be viewed as a generalised from the geometric distribution. The simulation results in Figures 4.11, 4.12, and 4.13 show that the new estimator tends to have less bias and performs much better than competitors when the population size or λ increases. Also, the LCMP estimator tends to be an unbiased estimator with respect to the population size. In contrast, it can be said that the MLEPoi, the Turing and the Chao estimators provide a clear underestimation of population size under the negative binomial distribution. This is similar to the results found in Lanumteang and Böhning (2011).

Interestingly, the parameter λ plays the important role of the accuracy, precision and performance of the LCMP estimator. The smaller value of $\lambda \leq 0.4$, or the larger the number of unobserved counts, affect to the less accurate and precise of the LCMP estimator. However, if the population size is large ($N \geq 5,000$), it is reasonable to estimate population size by the LCMP even under the negative binomial distribution. Another case is when λ is greater than 0.4. Here, we can see that the LCMP represents one of the best choices for estimating population size based on the negative binomial model with small RRMSE.

1) Relative bias of population size estimators when data are generated from the negative binomial distribution

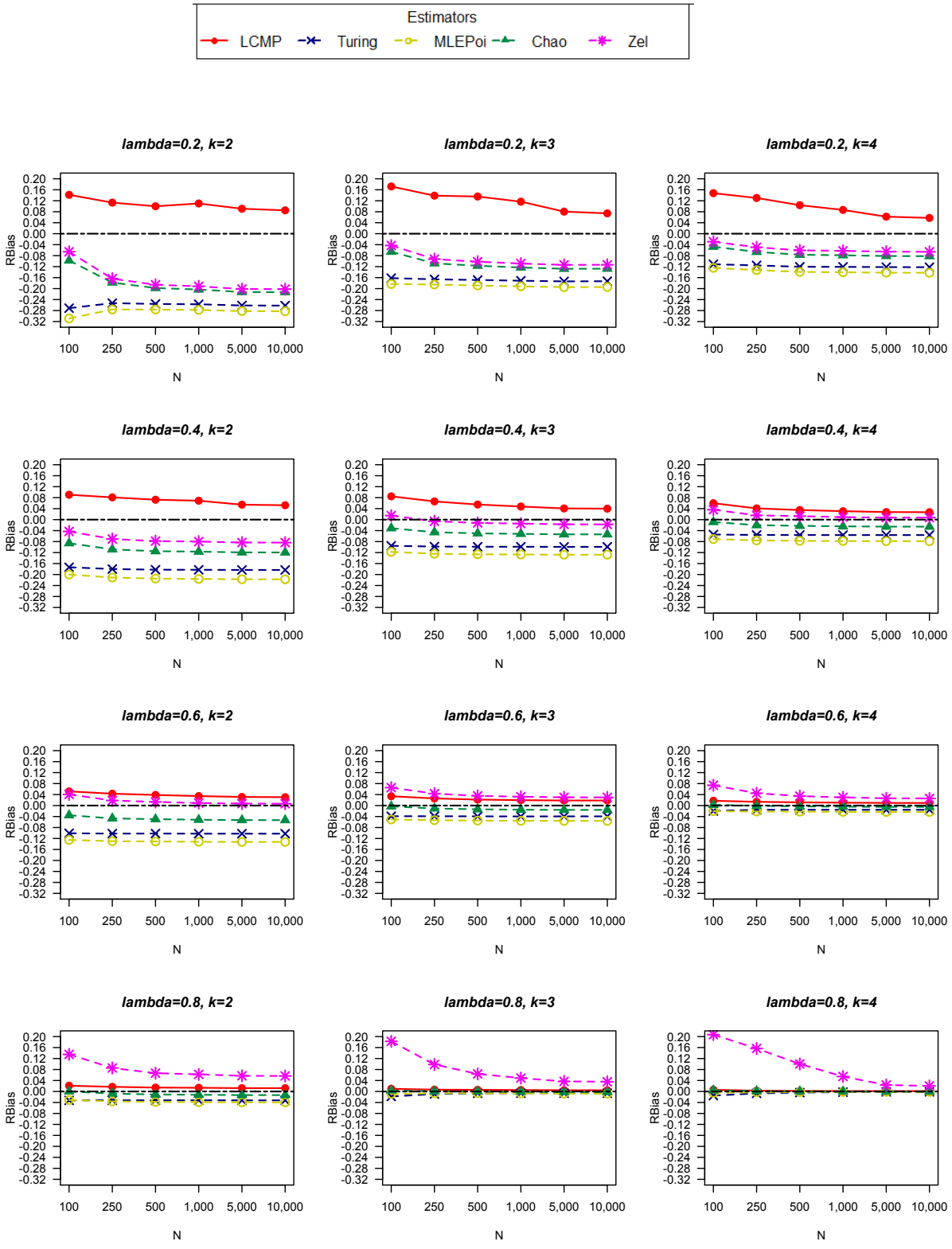

Figure 4.11: Relative bias of five estimators for counts drawn from $NB(\lambda, k)$

2) Relative variance of population size estimators when data are generated from the negative binomial distribution

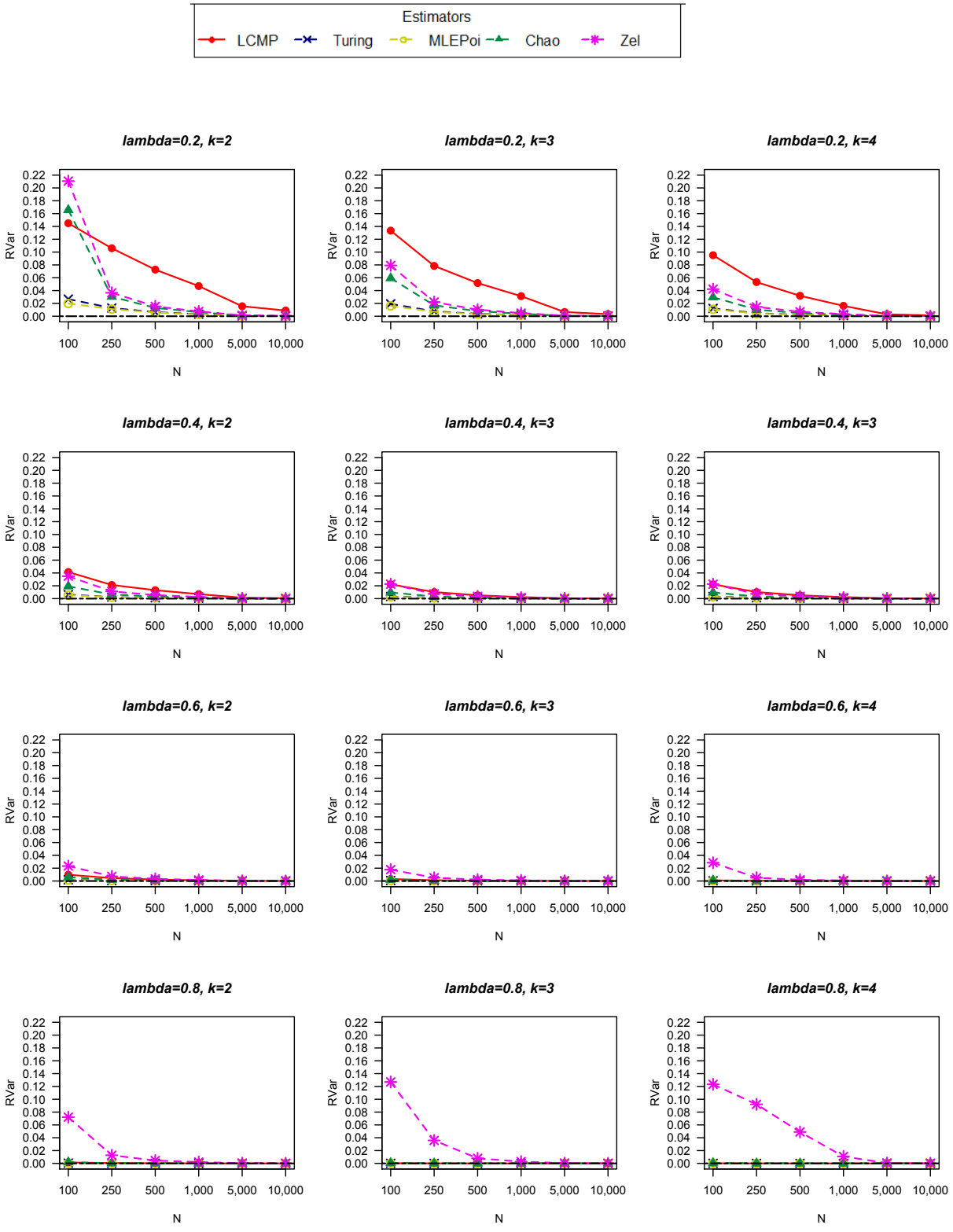

Figure 4.12: Relative variance of five estimators for counts drawn from $NB(\lambda, k)$

3) Relative root mean square error of population size estimators when data are generated from the negative binomial distribution

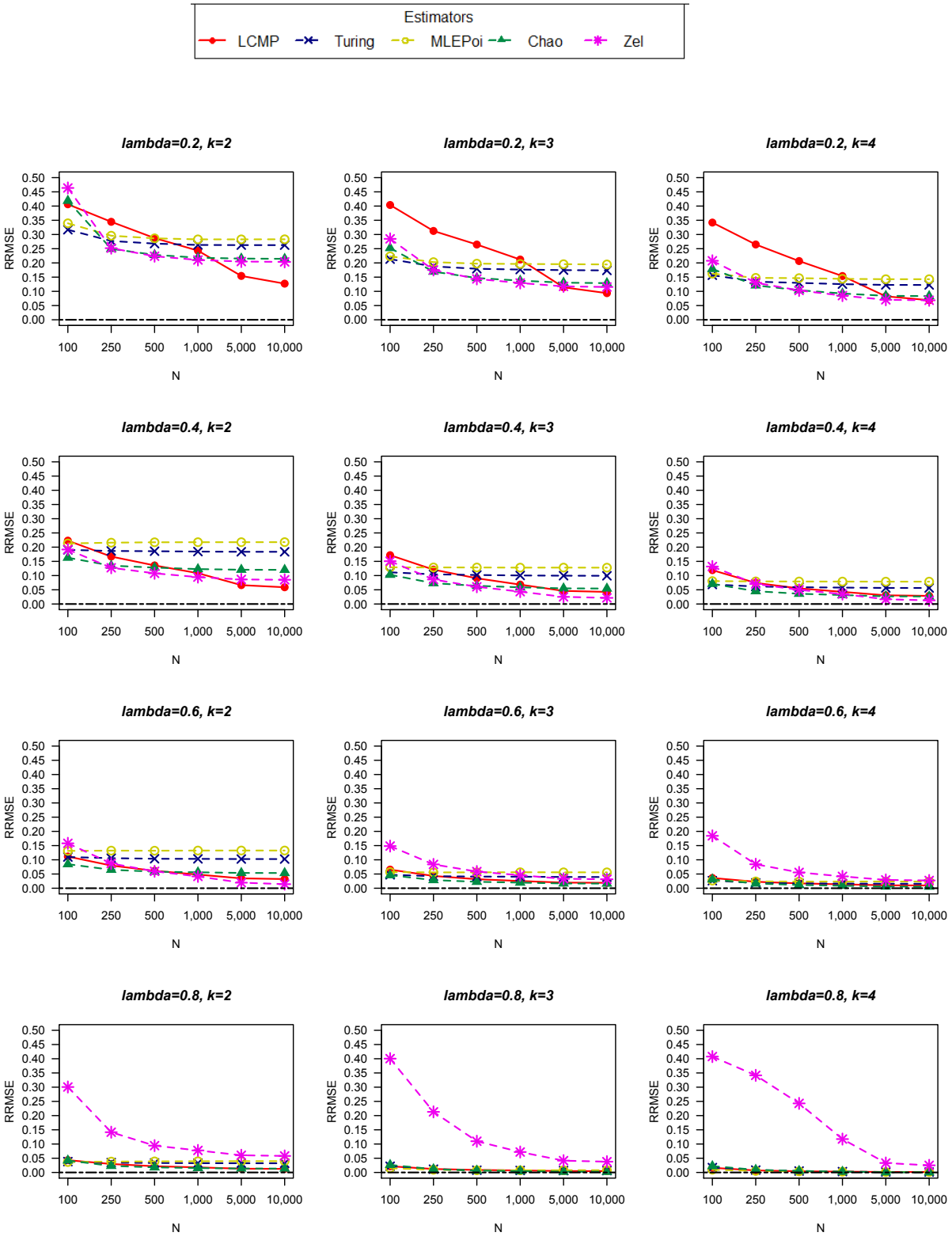

Figure 4.13: Relative root mean square error of five estimators for counts drawn from $NB(\lambda, k)$

4.5 Real data examples

In this section the proposed estimator LCMP is compared with the competing estimators by using real data examples. Five datasets are considered: the cholera data; the golf tees data; the artificial data used in the Link study; drug users in Bangkok data and snowshoe hare data. Additionally, the fitted frequencies of the zero-truncated distributions (Poisson and Conway-Maxwell-Poisson), conditioning on the observed data are compared in terms of model fit. The Chi-square goodness of fit under the null hypothesis of zero-truncated distribution is computed by $\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \widehat{f}_x)^2}{\widehat{f}_x}$, where $\widehat{f}_x$ denotes the fitted frequencies count of x, calculated by $\widehat{f}_x = n\widehat{p}_x^{+}$.

4.5.1 Cholera data

The example stems from Mao and Lindsay (2003) and has been discussed previously in Blumenthal et al. (1978); Viwatwongkasem et al. (2008), and others. A cholera epidemic affected in a village with 223 totally households in India. Originally, the data set was analysed by McKendrick (1925). The original research acknowledged that the data have been provided for the homogeneity. Data are provided in Table **??**.

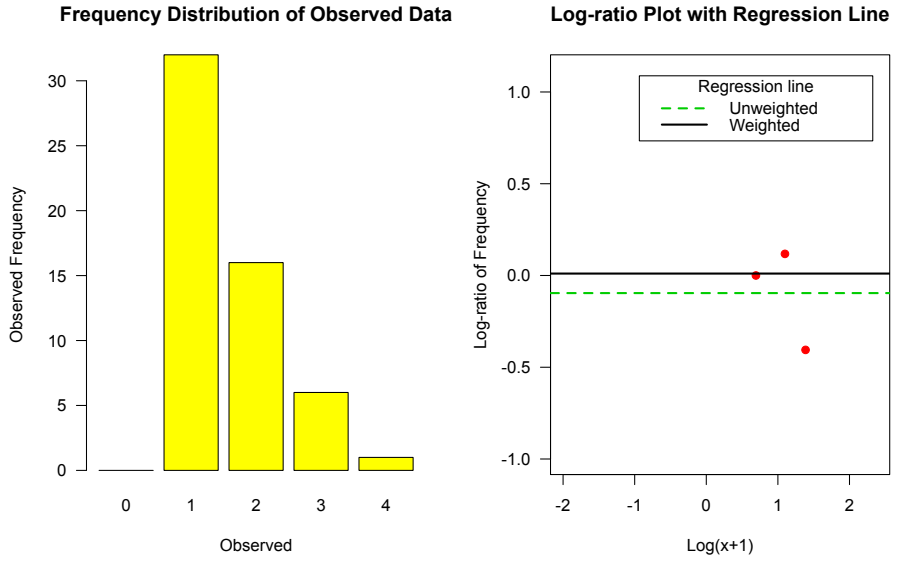

Figure 4.14: Frequency distribution (left) and the log-ratio plot with linear line (right) for cholera data

Table 4.2: Population size estimation for the cholera data

Estimator	$\widehat{N}$	$\widehat{f}_0$
Poisson		
Turing	88	33
MLEPoi ($\widehat{\lambda} = 0.98$)	89	34
LCMP	87	32
($\widehat{\lambda} = 1.01$ and $\widehat{\nu} = 1$)		
Heterogeneity		
Chao	87	32
Zelterman	88	33

It is known that the proposed estimator with $\nu = 1$ can be used as the estimator for homogeneous Poisson case. Therefore, it can be seen in Table 4.3 that the fitted frequencies of the zero-truncated Poisson distribution using the MLEPoi estimator and the zero-truncated Conway-Maxwell-Poisson distribution using the LCMP estimator are identical. Interestingly, only the count of three from both estimators provide a tiny underestimation. The graphical representation of estimated and observed frequencies is shown in Figure 4.15 to support the validity of both models but preferably the Poisson model.

Table 4.3: Observed data and fitted frequencies following the zero-truncated Poisson (ZTPoi) and the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distributions for cholera data

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (LCMP)
1	32	32	32
2	16	16	16
3	6	5	5
4	1	1	1
	χ^2	0.2	0.2

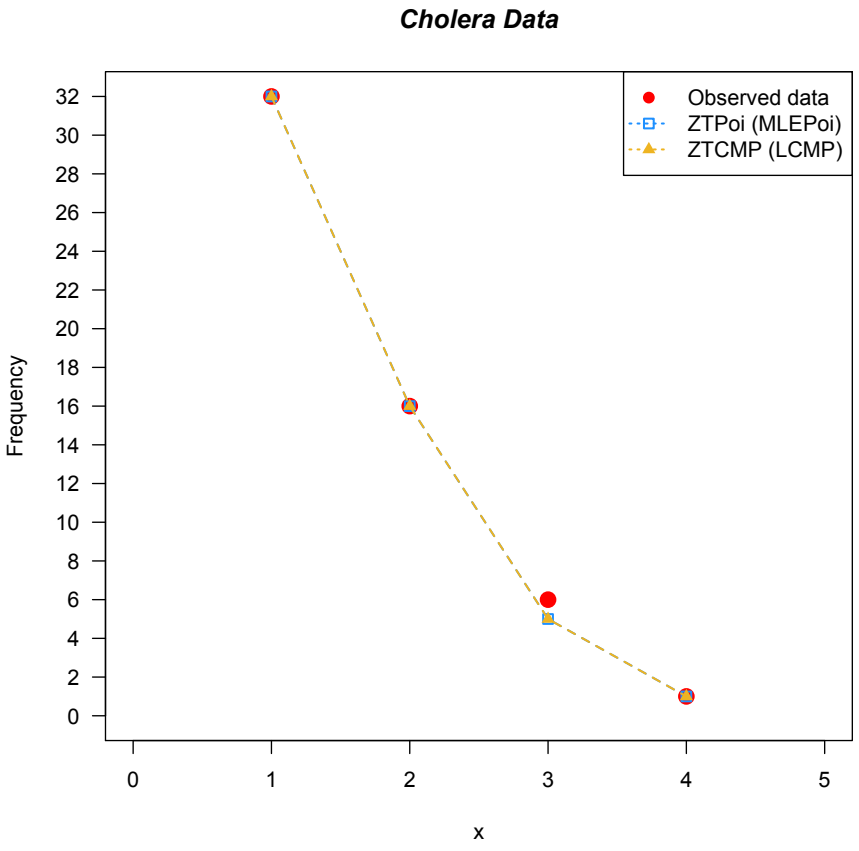

Figure 4.15: Observed frequencies with fitted frequencies under the zero-truncated Poisson and zero-truncated Conway-Maxwell-Poisson for cholera data

4.5.2 Golf tees data

In a field experiment, $N = 250$ groups of golf tees were placed in a survey region, either exposed above the surrounding grass or hidden by it. They were surveyed by the 1999 statistics honour class at the University of St Andrews (Scotland), (see Borchers et al., 2002). A total of $n = 162$ groups of tees were observed, but a (potentially unknown) number were missed and needs to be estimated. Table 4.4 shows the corresponding frequency distribution. Figure 4.16 provides a plot of the log-ratios of successive frequencies and the count distribution. It is clear that the log-ratio plot displays a linear relationship between log-ratios and log-counts, with a positive slope. It is reasonable to assume that a heterogeneous model would be suitable to estimate population size, such as the LCMP estimator proposed in this Chapter.

Table 4.4: The frequency distribution of golf-tees

x	0	1	2	3	4	5	6	7	8	N
f_x	88	46	28	21	13	23	14	6	11	250

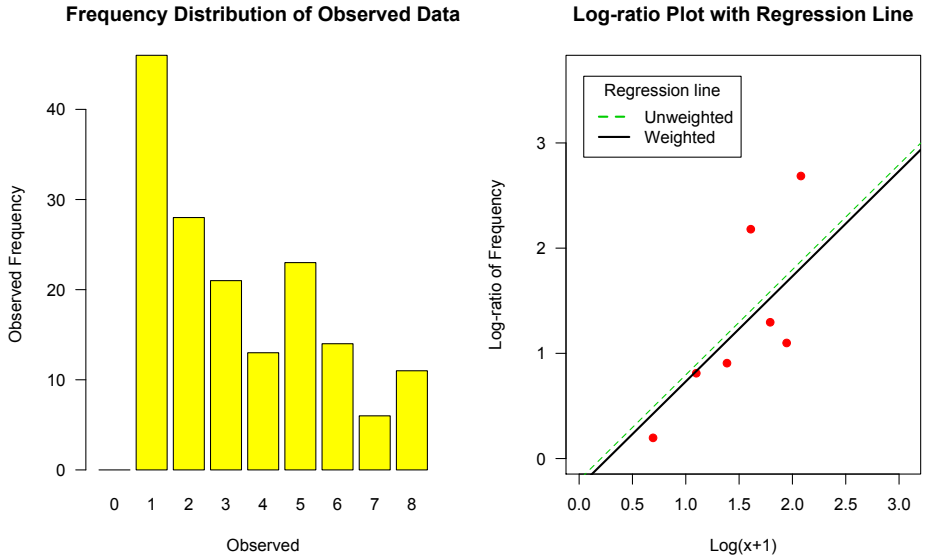

Figure 4.16: Frequency distribution (left) and the log-ratio plot with linear line (right) for golf tees data

Looking at other estimators performance, Zelterman's estimator shows the highest degree of accuracy in terms of having the smallest bias, followed by the LCMP estimator. The Turing and the MLEPoi estimators provide the least accuracy since they show a very large bias. This can be expected as the log-ratio plot suggests any estimator based on a homogeneous Poisson-based model should be avoided. It should be remarked that the imposed (and necessary) constraint on β_1 may limit the capacity of the LCMP estimator to recover the true population size if the underlying count distribution is far from being geometrically-distributed. However, the LCMP-based estimator allows for heterogeneity. It provides better estimates than homogeneous population-based Poisson estimators.

Table 4.5: Population size estimation of the golf tees data

Estimator	$\widehat{N}$	Bias	$\widehat{f}_0$
Poisson			
Turing	177	73	15
MLEPoi ($\widehat{\lambda} = 3.23$)	169	81	7
Heterogeneity			
LCMP	223	27	61
($\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$)			
Chao	200	50	38
Zel	231	19	67

A way to examine the goodness-of-fit of the model is to compare the observed and the estimated frequency as provided in Table 4.6. Overall, the estimated frequencies based on the zero-truncated Conway-Maxwell-Poisson distribution are slightly different from the observed data. However, it can be stated that it is closer to the empirical data than the traditional zero-truncated Poisson model, as it can be seen that the Chi-square goodness of fit from the zero-truncated Conway-Maxwell-Poisson is much less than the zero-truncated Poisson distribution. Moreover, the estimated regression parameter estimates are $\widehat{\beta}_0 = -0.268$ and $\widehat{\beta}_1 = 1$, that is $\widehat{\beta}_1$ is on the boundary of the parameter space. Accordingly, the parameters of the zero-truncated Conway-Maxwell-Poisson model are $\widehat{\lambda} = 0.765$ and $\widehat{\nu} = 0$, and so it can be said that the geometric distribution is obtained (see Table 4.5). In Figure 4.17, estimated frequencies of the zero-truncated Conway-Maxwell-Poisson distribution under the LCMP estimator with the zero-truncated Poisson distribution are compared with MLEPoi estimator. It is clear from the graph that the truncated Poisson distribution is not suitable for these data.

Table 4.6: Observed data and fitted frequencies following the zero-truncated Poisson (ZTPoi) and the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distributions for for golf tees data

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (LCMP)
1	46	22	38
2	28	35	29
3	21	37	22
4	13	30	17
5	23	20	13
6	14	11	10
7	6	5	8
8	11	2	6
	χ^2	86.10	16.66

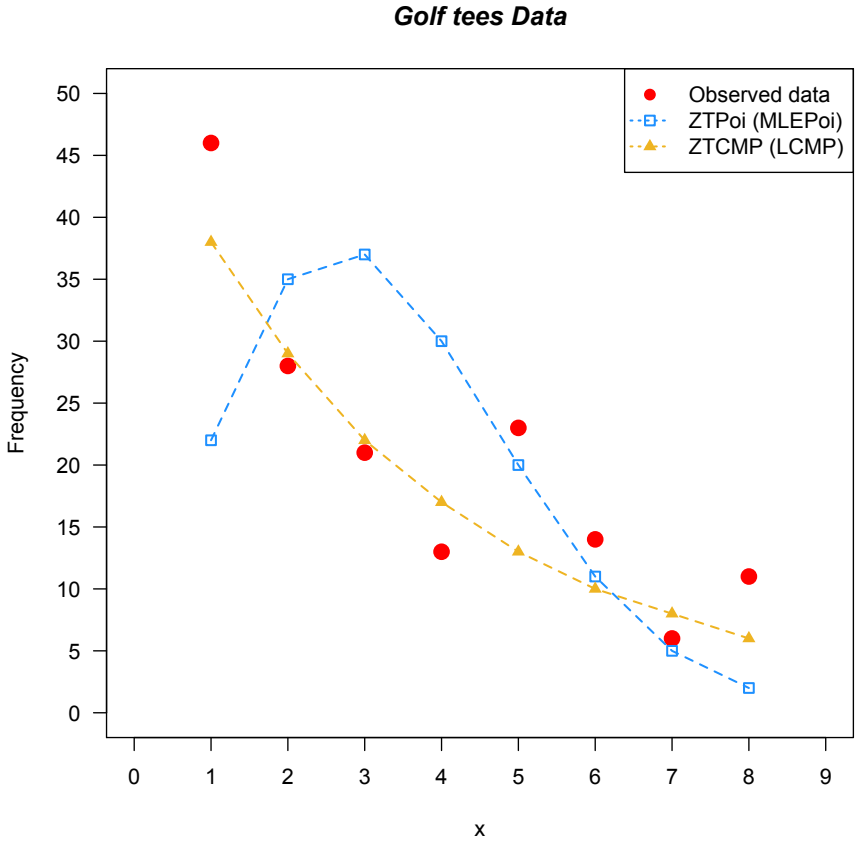

Figure 4.17: Observed frequencies with fitted frequencies under the zero-truncated Poisson and zero-truncated Conway-Maxwell-Poisson for golf tees data

4.5.3 Heroin users in Bangkok data

This study used all data on drug use from 61 health treatment centres in Bangkok metropolitan region collected by the Office of the Narcotics Control Board (ONCB), Ministry of the Prime Minister, which occurred from 1 October to 31 December in 2001. This data set has been analysed by Böhning et al. (2004) using the Poisson mixture model. The data are presented in Table 4.7, and the log-ratio plot is used as a tool for investigating the appropriateness of the model. As can be seen in Figure 4.18, the log-ratio plot suggests the a heterogeneous model would be more appropriate and the CMP distribution seems to fit the data well.

Table 4.7: Observed frequency distribution of the count of contacts heroin users for the 2001 drug user data of Bangkok.

x	1	2	3	4	5	6	7	8	9	10	
f_x	2,176	1,600	1,278	976	748	570	455	368	281	254	
x	11	12	13	14	15	16	17	18	19	20	21
f_x	188	138	99	67	44	34	17	3	3	2	1

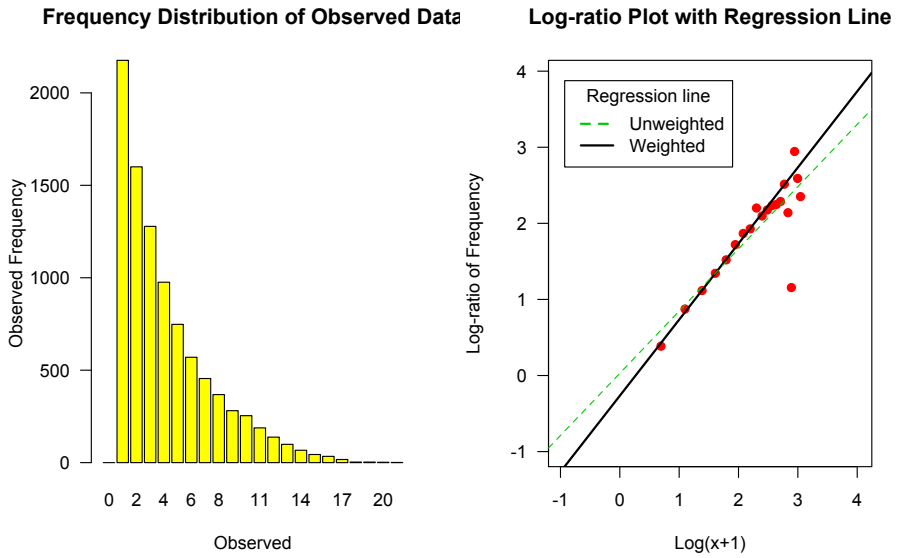

Figure 4.18: Frequency distribution (left) and the log-ratio plot with linear line (right) of heroin users in Bangkok data

Estimates of the population size under different assumptions are calculated using the various estimators. Results are displayed in Table 4.8. As in the golf-tees data, the geometric distribution is obtained as a special case of the LCMP for $\widehat{\nu} = 0$. The Zelterman estimator does not differ too much from the LCMP estimator, while the other estimators provide smaller population sizes estimates. These results do not surprise, given the results of the simulations.

Table 4.8: Population size estimation of the heroin users in Bangkok data

Estimator	$\widehat{N}$	$\widehat{f}_0$
Poisson		
Turing	9,850	548
MLEPoi ($\widehat{\lambda} = 4.13$)	9,454	152
Heterogeneity		
Chao	10,782	1,480
Zel	12,077	2,775
LCMP	12,141	2,839
($\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$)		

Observed and predicted frequencies are represented in Table 4.9. The homogeneous Poisson gives the poorest goodness-of-fit compared with the zero-truncated Conway-Maxwell-Poisson distribution. Additionally, Figure 4.19 supports that zero-truncated count distribution with long tail can be fitted more closely by the zero-truncated CMP distribution than by the zero-truncated Poisson model.

Table 4.9: Observed data and fitted frequencies following the zero-truncated Poisson (ZTPoi) and the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distributions for Heroin Users in Bangkok, 2001

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (LCMP)
1	2,176	626	2,179
2	1,600	1,293	1,670
3	1,278	1,783	1,281
4	976	1,843	982
5	748	1,524	753
6	570	1,050	577
7	455	620	442
8	368	321	339
9	281	147	260
10	254	61	199
11	188	23	153
12	138	8	117
13	99	3	90
14	67	1	69
15	44	0	53
16	34	0	40
17	17	0	31
18	3	0	24
19	3	0	18
20	2	0	14
21	1	0	11
	χ^2	26,844.98	94.56

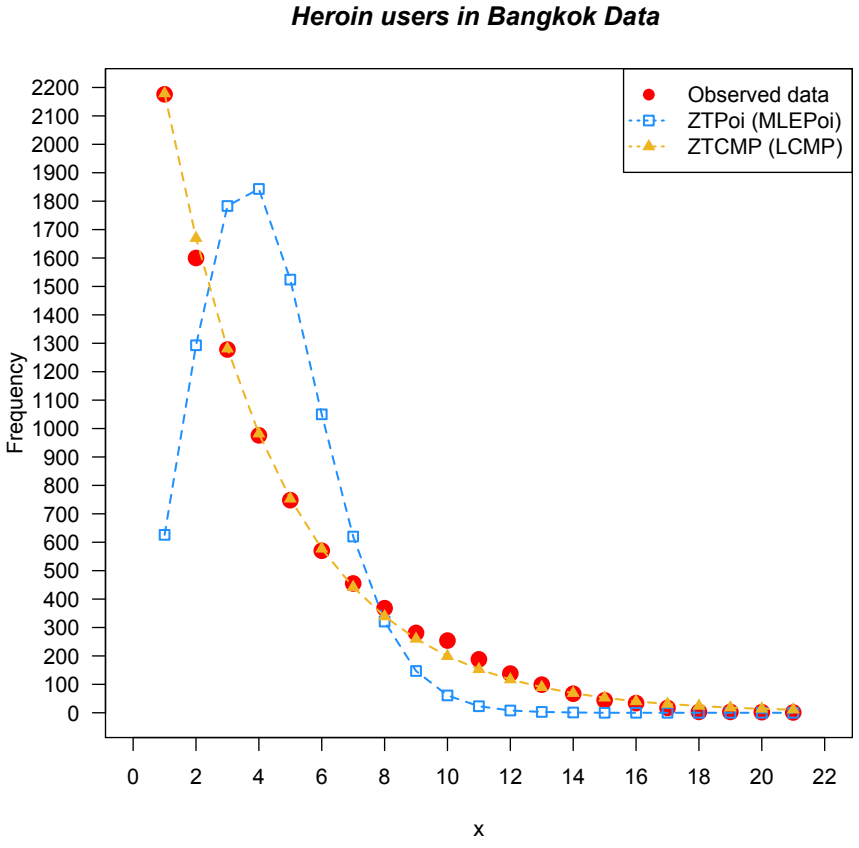

Figure 4.19: Observed frequencies with fitted frequencies under the zero-truncated Poisson and zero-truncated Conway-Maxwell-Poisson for heroin users in Bangkok Thailand in 2001

4.5.4 Link-3 data

Another interesting example is an artificial data set originally considered in Link (2003), see Table 4.10. These data are of particular interest as they show substantial heterogeneity (see Figure 4.20) and provide a dataset with a large number of recaptures. Thus, it is expected the MLEPoi and Turing estimators underestimate the population size. Indeed, the long tail of the count variable may lead to biassed estimates even for the Zelterman estimator. The log ratio plot as Figure 4.20 suggests a heterogeneous model such as the CMP distribution.

Table 4.10: Frequency distribution of Link (2003) data

x	1	2	3	4	5	6	7	8	9	10
f_x	679	531	379	272	198	143	99	67	46	32
x	11	12	13	14	15	16	n			
f_x	22	14	9	5	3	1	2,500			

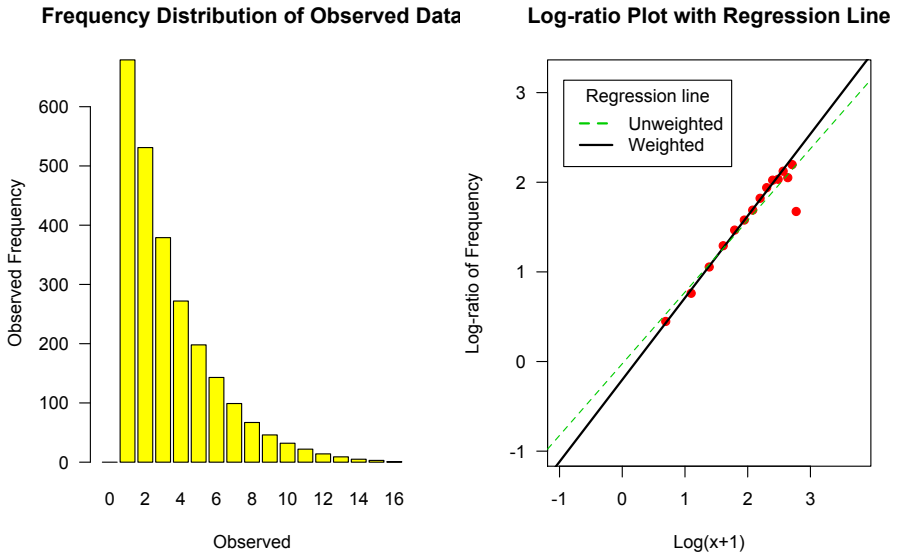

Figure 4.20: Frequency distribution (left) and the log-ratio plot with linear line (right) for Link-3 data

Population size estimates are displayed in Table 4.11. Homogeneous population-based estimators show very low estimates of $\widehat{N}$ with few unobserved data (similar behaviour is found in the simulation study). Chao's estimator provides a lower bound for $\widehat{N}$ in the presence of heterogeneity, and the population size of Zelterman's estimator does not differ too much, but is smaller than that from the LCMP estimator. This result does not surprise because the two estimated parameters; $\widehat{\lambda} = 0.82$ and $\widehat{\nu} = 0.09$, which implies a strong over dispersion case for the CMP distribution. The result is similar to the simulation study. The LCMP-based estimator seems to fit the data, and provide an estimate of N in line with the values obtained in Link (2003) under other parametric distributions accounting for heterogeneity. That was given $N = 3,494$.

Table 4.11: Population size estimation of the Link-3 data

Estimator	$\widehat{N}$	$\widehat{f}_0$
Poisson		
Turing	2,719	219
MLEPoi ($\widehat{\lambda} = 3.24$)	2,602	102
Homogeneity		
Chao	2,935	435
Zel	3,162	662
LCMP ($\widehat{\lambda} = 0.82$ and $\widehat{\nu} = 0.09$)	3,333	833

For considering the model fitted on the Link data as Table 4.12 and Figure 4.21, it can be seen that the fitted frequencies of zero-truncated CMP distribution provide an acceptable goodness-of fit ($\chi^2 = 2.14, df = 13$), each of predicted frequency is fairly closed to the observed frequency.

Table 4.12: Observed data and fitted frequencies following the zero-truncated Poisson (ZTPoi) and the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distributions for Link 3 data

x	Observed data	$\widehat{f}_x$ (MLEPoi)	$\widehat{f}_x$ (LCMP)
1	679	330	679
2	531	534	522
3	379	578	388
4	272	468	281
5	198	303	200
6	143	164	140
7	99	76	97
8	67	31	66
9	46	11	45
10	32	4	30
11	22	1	20
12	14	0	13
13	9	0	9
14	5	0	6
15	3	0	4
16	1	0	2
	χ^2	2,031.12	2.14

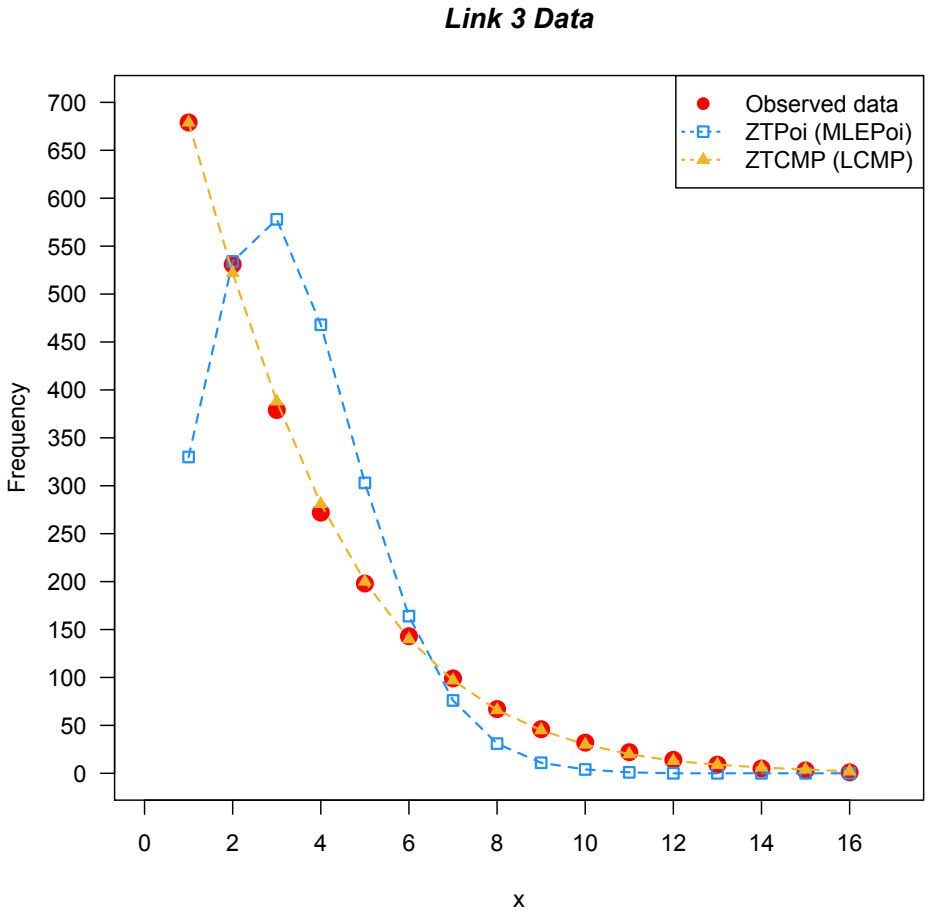

Figure 4.21: Observed frequencies with fitted frequencies under the zero-truncated Poisson and zero-truncated Conway-Maxwell-Poisson for Link 3 data

4.5.5 Snowshoe hare data

The snowshoe hare data have been analysed in Morgan and Ridout (2008) with the new mixture based on seven occasions, however the original study Cormack (1989) suggested to remove the 2 hares caught 6 times as outliers and then use the observed counts 1 to 5 as a zero-truncated distribution.

Table 4.13: Frequency distribution of snowshoe hare data

x	1	2	3	4	5	6
f_x	25	22	13	5	1	2

To investigate possible models for snowshoe hare data by using the log ratio ($x = 1, 2, 3, 4, 5$) as Figure 4.22. It can be seen that the log ratio plot suggests a straight line with negative slope indicating an under-dispersion with respect to the homogeneous

Poisson distribution. Consequently, it is not surprising that the estimated parameters are $\widehat{\nu} = 1.25$ and $\widehat{\lambda} = 2.16$ as they confirm the presence of under-dispersion. From the simulation results we expect that all estimators provide the overestimation of population size, however, this data set provide a slightly different of population size (78 – 83). Indeed, the population size based on the LCMP estimator gives the same result as Morgan and Ridout (2008)'s study, that is $\widehat{N} = 78$.

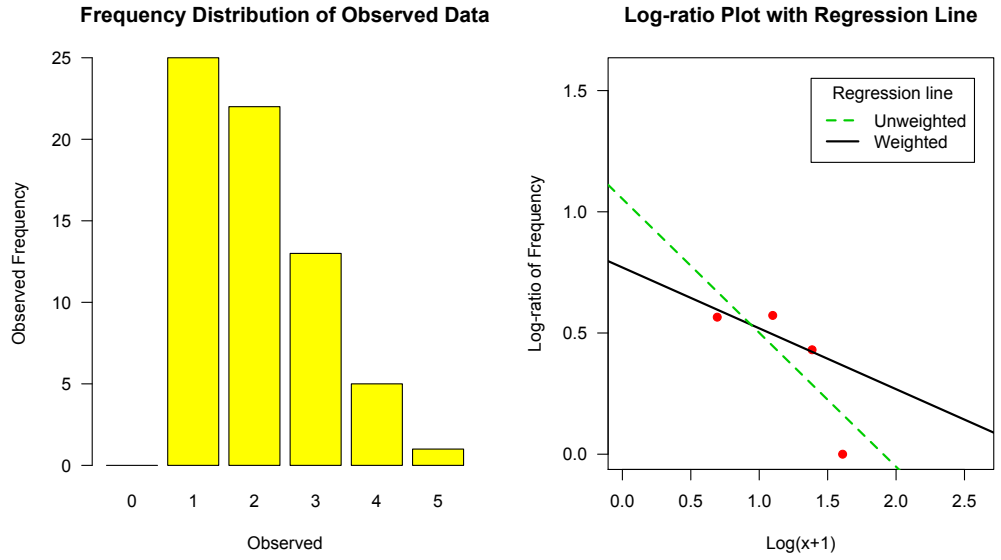

Figure 4.22: Frequency distribution (left) and the log-ratio plot with linear line (right) for the snow shoe hare data, ignoring f_6

Table 4.14: Population size estimation of the snowshoe hare data

Estimator	$\widehat{N}$	$\widehat{f}_0$
Poisson		
Turing	81	15
MLEPoi ($\widehat{\lambda} = 1.62$)	83	17
Homogeneity		
Chao	81	15
Zel	80	14
LCMP ($\widehat{\lambda} = 2.16$ and $\widehat{\nu} = 1.25$)	78	12

Apart from the parameters estimation of the parametric models (i.e. the Poisson and the CMP distributions), we need to consider the goodness of fit. It can be seen that the fitted frequencies of the zero-truncated CMP distribution shows a slightly better fit than the zero-truncated Poisson distribution, and there is a negligible difference of fitted frequencies only for the count of two and three as Table 4.15 and Figure 4.23 show.

Table 4.15: Observed data and fitted frequencies following the zero-truncated Poisson (ZTPoi) and the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distributions for snowshoe hares data

x	Observed data	$\widehat{f}_x$ MLEPoi	$\widehat{f}_x$ (LCMP)
1	25	26	25
2	22	21	23
3	13	12	12
4	5	5	5
5	1	2	1
	χ^2	0.67	0.13

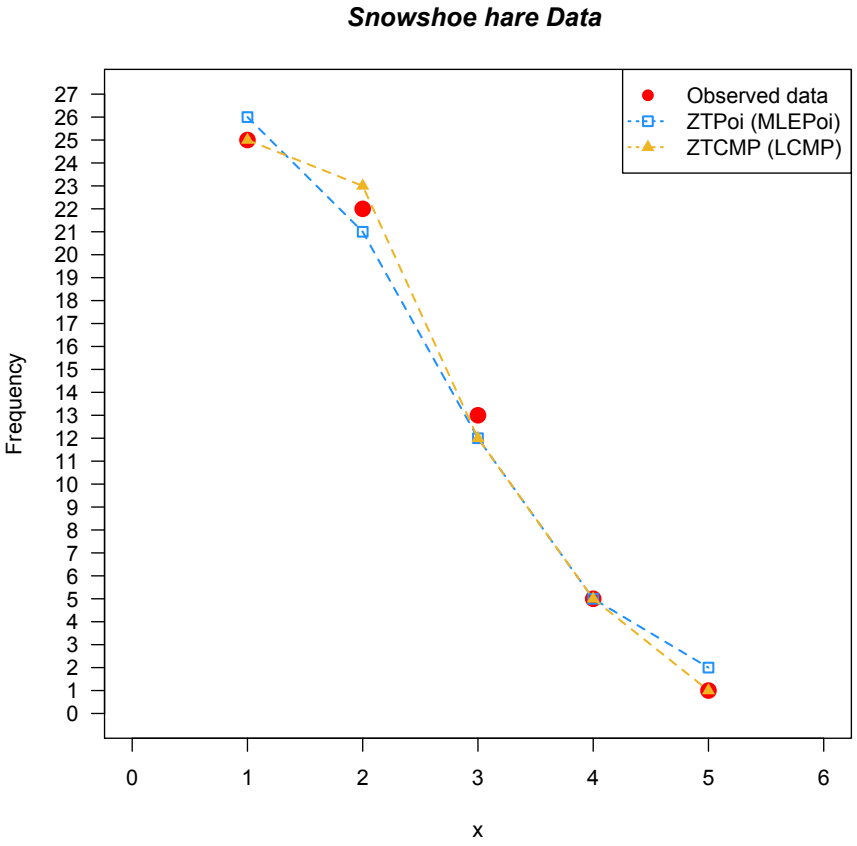

Figure 4.23: Observed frequencies with fitted frequencies under the zero-truncated Poisson and zero-truncated Conway-Maxwell-Poisson for snowshoe hares data, ignoring f_6

4.6 Conclusion

The challenge of population size estimation is to select an estimator which will perform well. One of the crucial issues in capture-recapture data is that capture probabilities do not always follow the assumption of homogeneity. As we know, ignoring heterogeneity in population can lead to biased estimators of population size. The negative binomial as well as the geometric distributions have been suggested as flexible models to deal with over-dispersion in capture-recapture. However, the failure of a dispersion parameter estimation using the negative binomial distribution have been demonstrated in various capture-recapture data sets. Consequently, the generalised Poisson distribution was proposed in the previous chapter. For the benefit of improving the population size parameter estimation for capture-recapture data, a new method of estimating the population size allowing for heterogeneity based on the Conway-Maxwell-Poisson distribution is proposed. The main advantage of using the Conway-Maxwell-Poisson distribution is that it includes the geometric, the Poisson and the Bernoulli distributions as sub-model. Additionally, it allows for under and over-dispersion relative to the Poisson model.

A new parametric population size estimator for capture-recapture data under the zero-truncated Conway-Maxwell-Poisson distribution, namely the LCMP estimator is proposed. The two parameters of the LCMP estimator can be obtained by exploring the ratio of successive frequency counts by the weighted least squares regression. Moreover, the graphical diagnostic tool called the log ratio plot is utilised as a tool for detecting the validity of the Conway-Maxwell-Poisson distribution. The analytic and simulation results confirm that the LCMP estimator is asymptotically unbiased under the Conway-Maxwell-Poisson distribution as well as in the spacial cases: the geometric and the Poisson distributions.

In addition, the simulation results suggest that the proposed estimator gives high quality estimates under the homogeneous zero-truncated Poisson distribution, although it produces a slight underestimation for small population sizes similarly to the traditional estimators of the MLEPoi and the Turing estimators. For the geometric distribution, the LCMP estimator works very well compared with others, and its behaviour gets close to the robust Zelterman estimator when the parameter λ approaches one.

We also consider the performance of the LCMP estimator for the original true model that is the zero-truncated Conway-Maxwell-Poisson distribution. The results suggest that the LCMP estimator is the best choice for a long tail frequency distribution with the dispersion parameter close to zero. However, the LCMP estimator limits for under-dispersion, it start to do well for $N \leq 1,000$ when $\lambda > 1$. It performs well for $N > 5,000$.

Interestingly, the proposed estimator LCMP can be use for estimating population size under the Negative Binomial in some cases. It shows good performance for $\lambda > 0.4$ where $\lambda = 1 - p$ is the event parameter of the negative binomial distribution. Nevertheless, in

the case that the frequency of unobserved counts is large ($\lambda \leq 0.4$), the LCMP estimator remains a good alternative given the population size is no less than $5,000$.

Chapter 5

Variance and Confidence Interval for the LCMP Estimator

Capture-recapture approaches have been applied for estimating target population sizes by many authors. A new estimator, namely the LCMP estimator is proposed in Chapter 4. In this chapter, we consider variance estimation, for the LCMP estimator, constructed by an normal approximation approach and a resampling-based approach. The analytic confidence interval and its corrected confidence intervals are presented in the first section of this chapter. This is followed by the percentile confidence interval based on three resampling methods: true bootstrap, imputed bootstrap and reduced bootstrap. The second section provides a simulation study to investigate the behaviour of variance estimation approaches for the new LCMP estimator. The data are generated from the Conway-Maxwell-Poisson (CMP) distribution as well as the Poisson and the geometric distributions which are particular cases nested within the CMP distribution. Additionally, the negative binomial distribution is included in the simulation study for measuring the ability of the new estimator under model misspecification. Finally, real data examples are presented in the final section.

5.1 Variance estimation approaches

As the LCMP estimator is constructed by the ratio regression approach, it is possible to derive variance estimation from the variances of the regression coefficients. Four methods for constructing variance estimates are summarised below:

Method 1 (M1) : Variance estimation based on the normal approximation
Let $\widehat{N}$ be the population size estimator, its variance can be derived by using the conditional technique. Referring to Böhning (2008a), the variance of $\widehat{N}_{LCMP} = n + f_1 e^{-\widehat{\beta}_0}$ arises from two sources: the random variable n and the estimator $\widehat{f}_0$. Therefore a simple

formula for variance of population size estimator is given as:

$$Var(\widehat{N}) = Var_n\{E(\widehat{N}|n)\} + E_n\{Var(\widehat{N}|n)\}, \tag{5.1}$$

where E_n and Var_n refer to the first and the second moments of the marginal distribution of n. As the approximation $E(\widehat{N}|n)$ can be estimated by $n + \widehat{f}_0$ with the delta method (see Bishop et al., 2007; Böhning, 2008a), it follows that

$$Var_n\{E(\widehat{N}|n)\} \approx Var_n\{n + \widehat{f}_0\} = Var_n(n) = N(1-p_0)p_0. \tag{5.2}$$

Since $E(n) = N(1-p_0)$ and $p_0 = E(f_0/N)$, leading to $\widehat{p}_0 = \dfrac{\widehat{f}_0}{n+\widehat{f}_0}$, can be estimated by

$$\widehat{Var}_n\{E(\widehat{N}|n\} = \frac{n\widehat{f}_0}{n+\widehat{f}_0} = \frac{nf_1e^{-\widehat{\beta}_0}}{n+f_1e^{-\widehat{\beta}_0}}. \tag{5.3}$$

$E_n\{Var(\widehat{N}|n)\}$ can also be approximated using the delta method. Therefore, $E_n\{Var(\widehat{N}|n)\} \approx Var(\widehat{N}|n) = Var(\widehat{f}_0) = Var\{f_1e^{-\widehat{\beta}_0}\}$, and so

$$E_n\{Var(\widehat{N}|n)\} = Var\{f_1e^{-\widehat{\beta}_0}\}. \tag{5.4}$$

Applying the conditional technique gives

$$Var(f_1e^{-\widehat{\beta}_0}) = Var_{f_1}\{E(f_1e^{-\widehat{\beta}_0})|f_1\} + E_{f_1}\{Var(f_1e^{-\widehat{\beta}_0})|f_1\}. \tag{5.5}$$

Then,

$$\begin{aligned} Var_{f_1}\{E(f_1e^{-\widehat{\beta}_0})|f_1\} &\approx Var(f_1e^{-\widehat{\beta}_0}) = (e^{-\widehat{\beta}_0})^2Var(f_1) \\ &= (e^{-\widehat{\beta}_0})^2Np_1(1-p_1) \approx (e^{-\widehat{\beta}_0})^2f_1\left(1-\frac{f_1}{N}\right), \end{aligned} \tag{5.6}$$

and, $E_{f_1}\{Var(f_1e^{-\widehat{\beta}_0})|f_1\}$ can be estimated by $Var\{(f_1e^{-\widehat{\beta}_0})|f_1\}$, so that

$$E_{f_1}\{Var(f_1e^{-\widehat{\beta}_0})|f_1\} \approx Var\{(f_1e^{-\widehat{\beta}_0})|f_1\} = f_1^2Var(e^{-\widehat{\beta}_0}). \tag{5.7}$$

Using the delta method, it is shown that $Var(e^{-\widehat{\beta}_0}) \approx (e^{-\widehat{\beta}_0})^2Var(\widehat{\beta}_0)$. Hence

$$E_{f_1}\{Var(f_1e^{-\widehat{\beta}_0})|f_1\} \approx f_1^2(e^{-\widehat{\beta}_0})^2Var(\widehat{\beta}_0),$$

where $Var(\widehat{\beta}_0)$ is readily available from linear regression. The approximate expression for the variance of the new estimator $\widehat{N}_{LCMP}$ is therefore given as

$$\widehat{Var}(\widehat{N}_{LCMP}) = \frac{nf_1e^{-\widehat{\beta}_0}}{n+f_1e^{-\widehat{\beta}_0}} + (e^{-\widehat{\beta}_0})^2f_1\left(1-\frac{f_1}{N}\right) + f_1^2(e^{-\widehat{\beta}_0})^2Var(\widehat{\beta}_0). \tag{5.8}$$

As $1 - \frac{f_1}{N} \leq 1$, a conservative asymptotic variance estimator of $\widehat{N}_{LCMP}$ is obtained as follows:

$$\widehat{Var}(\widehat{N}_{LCMP}) = \frac{nf_1e^{-\widehat{\beta}_0}}{n + f_1e^{-\widehat{\beta}_0}} + (e^{-\widehat{\beta}_0})^2 f_1[1 + f_1 Var(\widehat{\beta}_0)]. \tag{5.9}$$

An approximate 95% confidence interval for the true population size N based on the proposed LCMP estimator is given as

$$\widehat{N}_{LCMP} \pm z_{0.975}\widehat{S.E.}(\widehat{N}_{LCMP}), \tag{5.10}$$

where $\widehat{S.E.}(\widehat{N}_{LCMP}) = \sqrt{\widehat{Var}(\widehat{N}_{LCMP})}$ is the estimated standard error. This method is commonly used for estimating a confidence interval based on the assumption of an asymptotic normal distribution. We call this the symmetric confidence interval method (SYM). However, the unsuitability of the symmetric confidence interval in capture-recapture studies has been highlighted by several researchers such as Chao (1987) and Burnham and Overton (1978). Additionally, Zwane and Van der Heijden (2003) confirmed that a skewed distribution was commonly demonstrated in the literature describing capture recapture studies.

To adapt to the fact that $\widehat{N}$ has a skewed distribution, a second approach, namely the Burnham confidence interval (BH), was introduced (see Burnham and Overton, 1978; Chao, 1987; Tounkara and Rivest, 2015). The Burnham confidence interval is produced by taking the log of $(\widehat{N} - n)$ to improve coverage probability. The 95% Burnham confidence interval for the true population size N can be obtained by

$$\left(n + \frac{(\widehat{N}_{LCMP} - n)}{c}, n + (\widehat{N}_{LCMP} - n)c\right), \tag{5.11}$$

where $c = \exp\left\{z_{0.975}\left[\log\left(1 + \frac{\widehat{S.E.}(\widehat{N}_{LCMP})}{(\widehat{N}_{LCMP} - n)^2}\right)^{1/2}\right]\right\}$, and $z_{0.975}$ is the 97.5% quantile of a standard normal distribution.

Another approach for dealing with the failure of the assumption of a symmetric distribution is a log-transformation of population size estimator $\widehat{N}$. This method was suggested by Köse et al. (2014) to correct coverage probability in the case where the sample size is close to the estimated population size ($\widehat{N} - n = \widehat{f}_0 \to 0$). The 95% confidence interval of N is given by

$$\log \widehat{N}_{LCMP} + \frac{1}{2}\log[1 + \widehat{Var}(\widehat{N}_{LCMP})/\widehat{N}^2_{LCMP}] \pm z_{0.975}\sqrt{\log[1 + \widehat{Var}(\widehat{N}_{LCMP})/\widehat{N}^2_{LCMP}]}. \tag{5.12}$$

Then, taking the anti-log gives the final form of the log transformed confidence interval (LOG) for N.

The estimated variance of the LCMP estimator by an approximation method will clearly be valid asymptotically. However, large sample sizes are required for such an assumption to hold. Target population sizes in capture-recapture studies can be small and this might lead to underestimation or overestimation of the variance. Additionally, model uncertainty in variance estimation is often ignored since the analytical approach depends on data from only one sample. Although the construction of the confidence interval might be improved by the transformation procedures, it can only be used for the specific range of parameter value.

These limitations of the analytic variance approach might lead to invalid confidence intervals for the true population size and provide an unsatisfactory coverage probability (Regal and Hook, 1991). The resampling based approaches can provide an alternative method for calculating variance and standard error estimates in the context of capture-recapture studies. Parametric and non-parametric bootstrap methods have been proposed to estimate the variance of capture-recapture estimators by several authors including Buckland (1984); Buckland and Garthwaite (1991); Norris III and Pollock (1996); Zwane and Van der Heijden (2003), but have never been fully investigated. Here, the performance of different resampling-based and approximation-based approaches are compared in the context of variance and confidence interval estimation for the LCMP estimator.

In the following section, three bootstrap methods: the true bootstrap (M2), the imputed bootstrap (M3) and the reduced bootstrap (M4) are examined as alternative methods for approximating the variance of $\widehat{N}_{LCMP}$ and its corresponding 95% confidence interval. The bootstrap methods are proposed based on the basic assumption that the capture-recapture history can be defined by the multinomial likelihood. That is

$$\binom{N}{f_0 \quad f_1 \quad f_2 \quad \cdots \quad f_m} p_0^{f_0} p_1^{f_1} ... p_m^{f_m}.$$

In other words, the sampling distribution of capture-recapture data is assumed to follow a multinomial distribution with $m+2$ parameters $(N, p_0, p_1, ...p_m)$, where N is the target population size and $p_x's$ denote capture probabilities $(X \in \{0, 1, 2, .., m\})$. In this setting, the algorithm of the true bootstrap, the imputed bootstrap and the reduced bootstrap for variance estimation techniques proceed as follows.

Method 2 (M2): True bootstrap method

For the case where the target population size is known, the true parametric bootstrap can be used for estimating the variance of population size estimator. Each individual is drawn from the the multinomial distribution with full probability model $(N, p_0, p_1, ...p_m)$, and the parameters are estimated, including our proposed estimator $\widehat{N}_{LCMP}$, as follows:

Step 1: Capture-recapture probabilities ($\widehat{\mathbf{p}}$) are estimated using the relative frequencies as follows:

$$\widehat{\mathbf{p}}_{TB} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2, ..., \widehat{p}_m) = \left(\frac{f_0}{N}, \frac{f_1}{N}, \frac{f_2}{N}, ..., \frac{f_m}{N}\right).$$

Step 2: Resampling associated frequencies ($\mathbf{f}^*$) under the multinomial distribution with parameters $(N, \widehat{\mathbf{p}})$, that is

$$\mathbf{f}^{*(b)} = \left(f_0^{*(b)}, f_1^{*(b)}, f_2^{*(b)}, f_3^{*(b)}, ..., f_m^{*(b)}\right) \sim Multinomial(N, \widehat{\mathbf{p}})$$

can be accomplished by using the function `rmultinom ()` in `R`

Step 3: The frequency at zero ($f_0^{*(b)}$), is truncated and the population size based on the LCMP estimator is found as:

$$\widehat{N}_{LCMP}^{*(b)} = n^{*(b)} + f_1^{*(b)} \exp\left(-\widehat{\beta}_0^{*(b)}\right),$$

where the $\widehat{\beta}_0^{*(b)}$ is estimated using the weighted least squares regression method, and $n^{*(b)} = f_1^{*(b)} + f_2^{*(b)} + f_3^{*(b)} + ... + f_m^{*(b)}$. If $\widehat{N}_{LCMP}^{*(b)}$ is not an integer, it has been suggested by Buckland and Garthwaite (1991) that it should be rounded to the nearest integer value.

Step 4: Step 2 and the step 3 are repeated B times where $b \in \{1, 2, 3, ..., B\}$. Here, $B = 1,000$ is used to obtain a set of population sizes estimated using LCMP estimator as:

$$\widehat{N}_{LCMP}^{*(1)}, \widehat{N}_{LCMP}^{*(2)}, \widehat{N}_{LCMP}^{*(3)}, ..., \widehat{N}_{LCMP}^{*(1,000)}.$$

Step 5: From this set of resampled estimates of size B the following statistics are computed:

1). Mean of population size

$$E(\widehat{N}_{M2}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{N}_{LCMP}^{(b)}\right).$$

2). Median of population size

$$Median(\widehat{N}_{M2}) = Median\{\widehat{N}_{LCMP}^{*(1)}, \widehat{N}_{LCMP}^{*(2)}, \widehat{N}_{LCMP}^{*(3)}, ..., \widehat{N}_{LCMP}^{*(1,000)}\}.$$

3).The bootstrap variance of population size is then calculated as

$$\widehat{Var}(\widehat{N}_{M2}) = \frac{1}{B-1}\left[\sum_{b=1}^{B}\left(\widehat{N}^{*(b)} - E(\widehat{N}_{M2})\right)^2\right],$$

and *the standard error of population* size is estimated by $\widehat{S.E.}(\widehat{N}_{M2}) = \sqrt{\widehat{Var}(\widehat{N}_{M2})}$.

Step 6: A robust confidence interval can then be obtained by using the bootstrap percentile interval method (Engel, 2010). The approximate 95% confidence interval of the population size N is obtained as follow: order $N_{(b)}$ from the smallest to largest, and denote the ordered list by $N_{(b)}$. The approximate 95% confidence limits are then given by $N_{(B+1)0.025}$ and$N_{(B+1)0.975}$, both rounded to the nearest integer value (Buckland and Garthwaite, 1991).

Method 3 (M3): Imputed bootstrap method

The imputed bootstrap is a modified form of *the parametric bootstrap approach* in Zwane and Van der Heijden (2003) and *Method 3* in Norris III and Pollock (1996) which requires an estimator to estimate full probabilities. This resampling method depends on the estimated population size ($\widehat{N}_{LCMP}$), indicating the limits of this approach if the proposed estimator is unsatisfactory. The reason why the predicted unobserved frequency is included in the multinomial model arises from using the using the conditional technique (as in M1) in the variance estimation. Since the variance of population size estimator arises from the sampling of n units from a population of size N and from estimating f_0 by using n observed units. These need to be incorporated into the imputed bootstrap. This means that it is necessary to add $\widehat{f}_0$ to the observed n before drawing a sample, and we use the observed values instead of the fitted values for $f_1, f_2, f_3, ..., f_m$. The imputed bootstrap algorithm is given as:

Step 1: Population size N is estimated by

$$\widehat{N}_{LCMP} = n + \widehat{f}_0 = n + f_1 \exp(-\widehat{\beta}_0),$$

from the zero-truncated count distribution, and $n = \sum_{x=1}^{m} f_x$ is the total number of observed individuals.

Step 2: The capture probabilities $\mathbf{p}$ are estimated by the relative frequencies, given as

$$\widehat{\mathrm{p}} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2, ..., \widehat{p}_m) = \left\{ \frac{\widehat{f}_0}{\widehat{N}}, \frac{f_1}{\widehat{N}}, \frac{f_2}{\widehat{N}}, ..., \frac{f_m}{\widehat{N}} \right\}$$

Step 3: Since the capture-recapture data can be summarised using the multinomial likelihood, the associated frequencies are generated under the multinomial distribution with parameters $\widehat{N}$, $\widehat{\mathbf{p}}$. That is

$$\mathbf{f}^{*(b)} = \left(f_0^{*(b)}, f_1^{*(b)}, f_2^{*(b)}, f_3^{*(b)}, ..., f_m^{*(b)} \right) \sim Multinomial(\widehat{N}, \widehat{\mathbf{p}}).$$

Step 4: The frequency of zero count is then truncated. The bootstrap population size is then estimated as:

$$\widehat{N}^{*(b)}_{LCMP} = n^{*(b)} + f_1^{*(b)} \exp\left(-\widehat{\beta}_0^{*(b)}\right)$$

where $n^{*(b)} = \sum_{x=1}^{m} f_x^{*(b)}$ and $\widehat{\beta}_0^{*(b)}$ is obtained from the weighted least squares regression method.

Step 5-7: There are the same as the step 4-6 in Method 2.

Method 4 (M4): Reduced bootstrap method

The concept of the reduced bootstrap is based on the resample approach conditioning on observed counts n. The major difference from $Method3$ is that n is treated as fixed. This means that the variance estimation can be approximated from only the second source in equation (5.1), i.e the variance estimation of f_0, conditional upon n using the model selection. This approach resembles *the nonparametric bootstrap* which was suggested by Zwane and Van der Heijden (2003) and *Method 1* in Norris III and Pollock (1996). It uses a resampling technique conditioning on the observed units. However, this approach is defined to be a reduced bootstrap in our study. Again each individual is generated from a multinomial distribution, therefore, the algorithm is summarised below.

Step 1: The zero-truncated capture probabilities $\mathbf{p}^{+} = (p_1, p_2, p_3, ..., p_m,)$ are estimated by the relative frequencies. That is

$$\widehat{\mathbf{p}}^{+} = \left(\frac{f_1}{n}, \frac{f_2}{n}, \frac{f_3}{n}, ..., \frac{f_m}{n}\right),$$

where $n = \sum_{x=1}^{m} f_x$.

Step 2: Resampling associated frequencies $\mathbf{f}^{*+(b)}$, are then generated from a multinomial distribution:

$$\mathbf{f}^{*+(b)} = \left(f_1^{*(b)}, f_2^{*(b)}, f_3^{*(b)}, ..., f_m^{*(b)}\right) \sim Multinomial(n, \widehat{\mathbf{p}}^{+})$$

Step 3 - 6: are the same as step 3 - 6 in Method 2.

5.2 Simulation study

A simulation study is also performed to explore the performance of various techniques for estimating the variance and confidence intervals for the proposed population sizes estimator (LCMP). The simulation study is designed to cover scenarios with different underlying *null* models with varying population sizes ($N = 100; 250$ for small sizes, $N = 500; 1000$ for medium sizes, and $N = 5000; 10000$ for large sizes). It is presented in Chapter 4, 1,000 data sets are drawn for each null model.

The objective is to evaluate the variance and standard error of the estimation method, each simulation scenario is repeated $1,000$ times (T = 1,000). Therefore, the true variance estimator is given by

$$Var(\widehat{N})_{True} = \frac{1}{T-1}\sum_{t=1}^{T}\left(\widehat{N}_{(t)} - E(\widehat{N})\right)^2, \tag{5.13}$$

and the true standard error is simply achieved as $S.E.(\widehat{N})_{True} = \sqrt{\left\{Var(\widehat{N})_{True}\right\}}$.

To study the behaviour and performance of variance estimation approaches, the estimated standard error are derived. $E\left\{SE(\widehat{N})_{M1}\right\}$ is defined as the expected value of the approximated standard error from the normal approximation approach (M1) given in (5.9). Its expected values can be calculated by

$$E\left\{\widehat{S.E.}(\widehat{N})_{M1}\right\} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{Var}(\widehat{N}_{(LCMP,t)})\right]}.$$

Additionally,

$$E\left\{\widehat{S.E.}(\widehat{N})_{M2}\right\} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{Var}(\widehat{N}_{(M2,t)})\right]},$$

$$E\left\{\widehat{S.E.}(\widehat{N})_{M3}\right\} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{Var}(\widehat{N}_{(M3,t)})\right]},$$

and

$$E\left\{\widehat{S.E.}(\widehat{N})_{M4}\right\} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{Var}(\widehat{N}_{(M4,t)})\right]},$$

are defined as expected values of the approximate standard error of population size from the true bootstrap, imputed bootstrap and reduced bootstrap, respectively. For the convenience, *the ratio of standard error of estimation*, which is defined as the estimated standard error from each approach divided by the true standard error. $\frac{E[\widehat{S.E.}(\widehat{N})]}{S.E.(\widehat{N})_{True}}$ is provided for comparing the behaviours and the performances of each variance estimation approach. The reference value for this ratio is equal to one.

The performance of a confidence interval (CI) method is assessed by how often the confidence interval covers the true value of the population sizes. The coverage probability is defined as the probability that the confidence interval contains the true value. The reference of comparison is the nominal coverage probability, set at 0.95 or 95%. Within the simulation study two-sided confidence intervals for N, as defined above, will be calculated. There are the symmetric confidence interval (SYM), the Burnham confidence interval (BH), and the log-transform of population size estimator $\widehat{N}$ (LOG), respectively. Additionally, the coverage probability produced from the robust percentile confidence intervals of the three bootstrap approaches are presented. Since the simulation has $T = 1,000$ replication runs, the (actual) coverage probability can be calculated as

$$Cov = \frac{\sum_{t=1}^{T} A_{(t)}}{T} \times 100, \tag{5.14}$$

where $A_{(t)}$ equal to 1 if the true value N is within the target confidence interval, and 0 otherwise.

5.2.1 Simulation results based on the Poisson distribution

1) Relative bias and relative variance when data are generated from a Poisson distribution

In the previous chapter was shown that the LCMP estimator is asymptotically unbiased under the Poisson distribution. Therefore in the first simulation scenario data are generated from a Poisson distribution. The results of the simulation study highlight the need to be cautious when using the LCMP estimator to estimate population size for small target populations and those with a large number of zero counts (λ is close to zero). In these situations the estimator underestimated population size and produced large variances (see Figure 5.1).

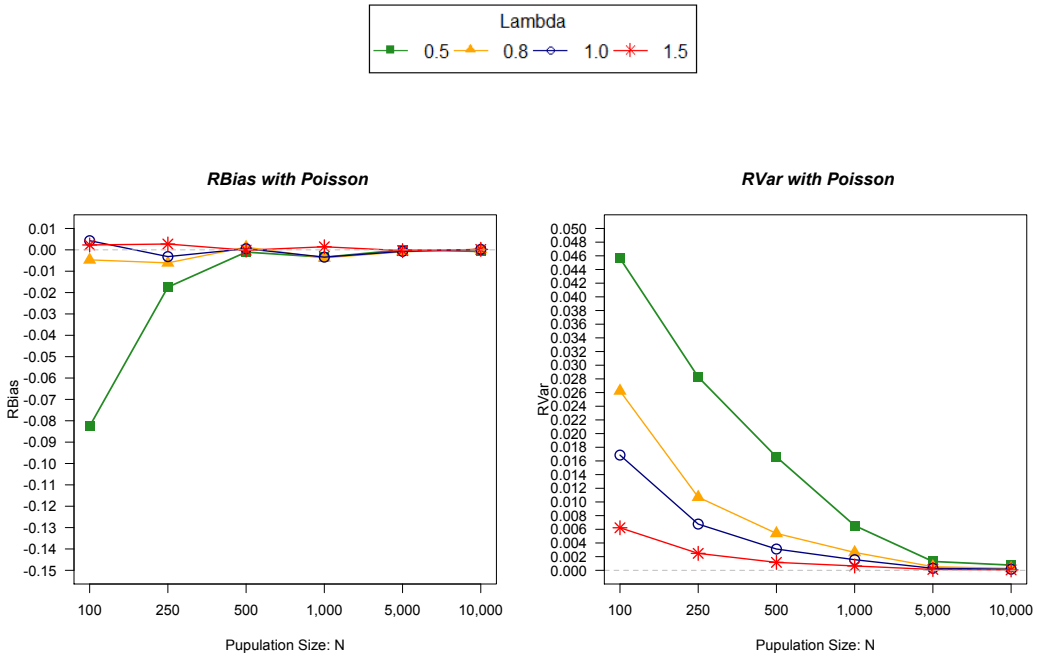

Figure 5.1: Relative bias (left) and the relative variance (right) of population size estimation based on the Poisson distribution

2) Variance and standard error of population size estimator when data are generated from a Poisson distribution

The aim of the simulation study is to investigate behaviour and performance for the variance estimation approaches of the LCMP estimator. Table 5.1 provides standard errors from 1,000 replication runs, and the ratio of standard errors are represented in Figure 5.2 when data are generated under a Poisson distribution. Overall, it can be seen from the simulation results that the true bootstrap and the imputed bootstrap perform the best with the ratio of standard errors close to one for all conditions. Interestingly, the variance estimation based on the normal approximation method (M1) gives an underestimation of the standard error, except in the case of small population size ($N \leq 250$) and $\lambda = 0.5$ providing an overestimation. Moreover, it is no surprise that the reduced bootstrap (M4) gives an underestimation of the standard error due to the condition that only the observed data are used. This finding corroborates the ideas of nonparametric bootstrap by Zwane and Van der Heijden (2003) and Norris III and Pollock (1996), who proposed the bootstrap conditioning on the observed data.

Table 5.1: Comparison of the standard errors of four methods with the true standard error of the LCMP estimator when data are generated from the Poisson distribution ($Poi(\lambda)$)

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
		M1	M2	M3	M4
		$Poi(0.5)$			
100	21.362	25.418	19.377	19.162	14.958
250	42.037	42.379	41.665	41.520	36.602
500	64.393	56.808	62.187	62.012	55.327
1,000	80.896	74.889	87.269	87.249	77.962
5,000	180.105	177.044	194.334	194.517	173.486
10,000	276.338	247.105	274.150	274.067	244.883
		$Poi(0.8)$			
100	16.198	14.812	16.246	16.192	13.434
250	25.863	23.032	26.280	26.258	22.096
500	36.735	33.422	37.373	37.327	31.385
1,000	51.169	46.910	52.177	52.218	43.766
5,000	115.483	106.183	116.263	116.315	97.380
10,000	169.231	148.938	164.418	164.565	137.351
		$Poi(1.0)$			
100	12.983	11.438	13.099	13.093	10.603
250	20.560	18.321	20.490	20.427	16.523
500	27.897	25.708	28.820	28.822	23.205
1,000	39.410	36.685	40.327	40.254	32.347
5,000	86.139	81.552	89.974	90.080	72.214
10,000	125.947	115.135	127.576	127.425	102.143
		$Poi(1.5)$			
100	7.890	7.095	7.910	7.900	5.808
250	12.389	11.244	12.389	12.414	9.050
500	17.060	15.844	17.394	17.367	12.586
1,000	25.274	22.614	24.540	24.557	17.782
5,000	56.074	50.079	54.600	54.734	39.506
10,000	77.405	71.258	77.478	77.422	55.982

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

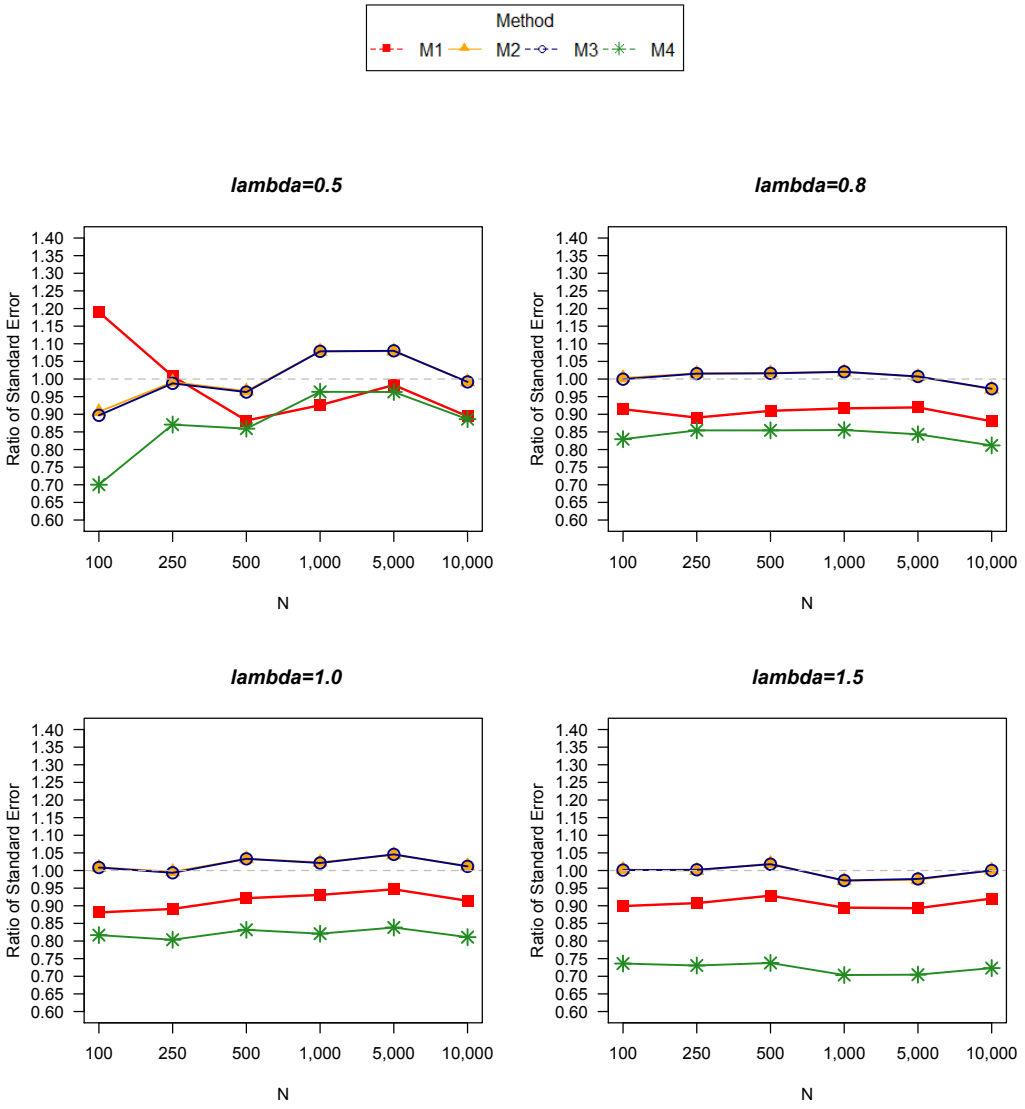

Figure 5.2: Ratio of standard errors from four methods to the true standard error when data are generated under a Poisson distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

3) Coverage probability when data are generated from a Poisson distribution

The coverage probabilities of the 95% confidence intervals based on the Poisson distribution are provided in Table 5.2 and Figure 5.3. A comparison of the means ($\widehat{N}_{Mean}$) and the medians ($\widehat{N}_{Med}$) in Table 5.2 shows that the distribution of the population size estimator is slightly positively skewed for the small population size and tends to be more symmetric when the parameter λ increases. Overall, we can see that the coverage probabilities for the true bootstrap are the closest to the nominal level for all simulation schemes based on the Poisson model. Importantly, the coverage probability of the imputed bootstrap becomes close to the true bootstrap when the population size increases. The reason for this might be due to the asymptotic property of the LCMP estimator

under the Poisson distribution and the efficiency of standard error estimation of the imputed bootstrap. The reduced bootstrap shows suboptimal behaviour for almost cases, resulting in lower coverage probabilities. In particular, the reduced bootstrap behaves poorly when $\lambda \geq 1$.This might be due to the substantial underestimation of standard error.

For the intervals constructed under variance estimation based on the normal approximation approach, it can be seen that the symmetric confidence interval approach, M1(SYM), results in coverage probabilities which are lower than those from the corrected confidence interval method (i.e. M1(BH) and M1(LOG)) for small population size and/or a high rate of unobserved data (λ is small). However, all of the confidence intervals constructed based on M1 method provide lower coverage probabilities in comparison to the nominal level for all conditions. It is found that the M1(BH) is more robust for $\lambda = 0.5$ and very small population size ($N = 100$), where coverage probability was closest to the nominal level. The robustness of this method might be explained by the fact that the M1(BH) is transformed to deal with the positive skew in the estimated population size.

Table 5.2: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from Poisson distribution ($Poi(\lambda)$)

			Approximate Normal (M1)			Bootstrap Method		
N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	SYM	BH	LOG	M2	M3	M4
			$Poi(0.5)$					
100	91.7	88.0	81.2	90.5	86.9	87.4	85.4	78.1
250	245.6	241.0	86.2	90.0	88.1	93.5	92.7	89.5
500	499.5	495.5	85.9	87.5	87.3	93.3	93.1	90.1
1000	996.6	994.0	87.6	88.0	87.7	95.9	95.4	93.0
5000	4,999.9	4,990.0	90.4	90.6	90.4	96.7	96.6	93.6
10000	9,992.5	9,985.5	88.1	87.4	87.7	94.9	95.3	92.2
			$Poi(0.8)$					
100	99.5	98.0	86.9	89.8	88.5	93.9	93.6	88.0
250	248.5	246.5	87.0	88.8	88.3	94.1	93.6	89.4
500	500.6	499.0	90.0	90.4	89.8	94.8	94.6	91.1
1000	996.2	995.0	90.0	91.2	91.0	95.4	94.7	90.3
5000	4,995.4	4,990.0	90.2	90.6	90.7	94.9	95.0	89.9
10000	10,003.1	10,004.0	90.4	90.0	90.3	94.2	94.6	88.7
			$Poi(1.0)$					
100	100.4	100.0	87.6	88.8	87.9	93.7	92.8	86.1
250	249.2	247.0	89.7	90.7	89.8	94.3	93.2	86.3
500	500.3	499.0	91.1	91.3	91.2	95.3	95.1	88.5
1000	996.6	995.0	90.3	91.1	90.2	95.2	94.6	89.2
5000	4,996.5	4,994.5	92.2	92.6	92.5	95.6	95.5	90.2
10000	10,002.0	9,995.0	90.9	91.0	90.9	95.7	95.9	89.0
			$Poi(1.5)$					
100	100.2	100.0	90.4	90.6	91.0	93.7	85.4	82.1
250	250.7	250.0	91.4	92.5	92.0	94.3	92.7	83.8
500	499.9	500.0	91.7	92.4	91.9	95.3	93.1	84.3
1000	1,001.4	1,001.0	91.8	91.4	92.0	95.2	95.4	82.8
5000	4,998.2	4,998.0	91.0	91.8	91.3	95.6	96.6	84.1
10000	10,002.0	10,002.0	92.7	92.7	92.9	95.7	95.3	84.8

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

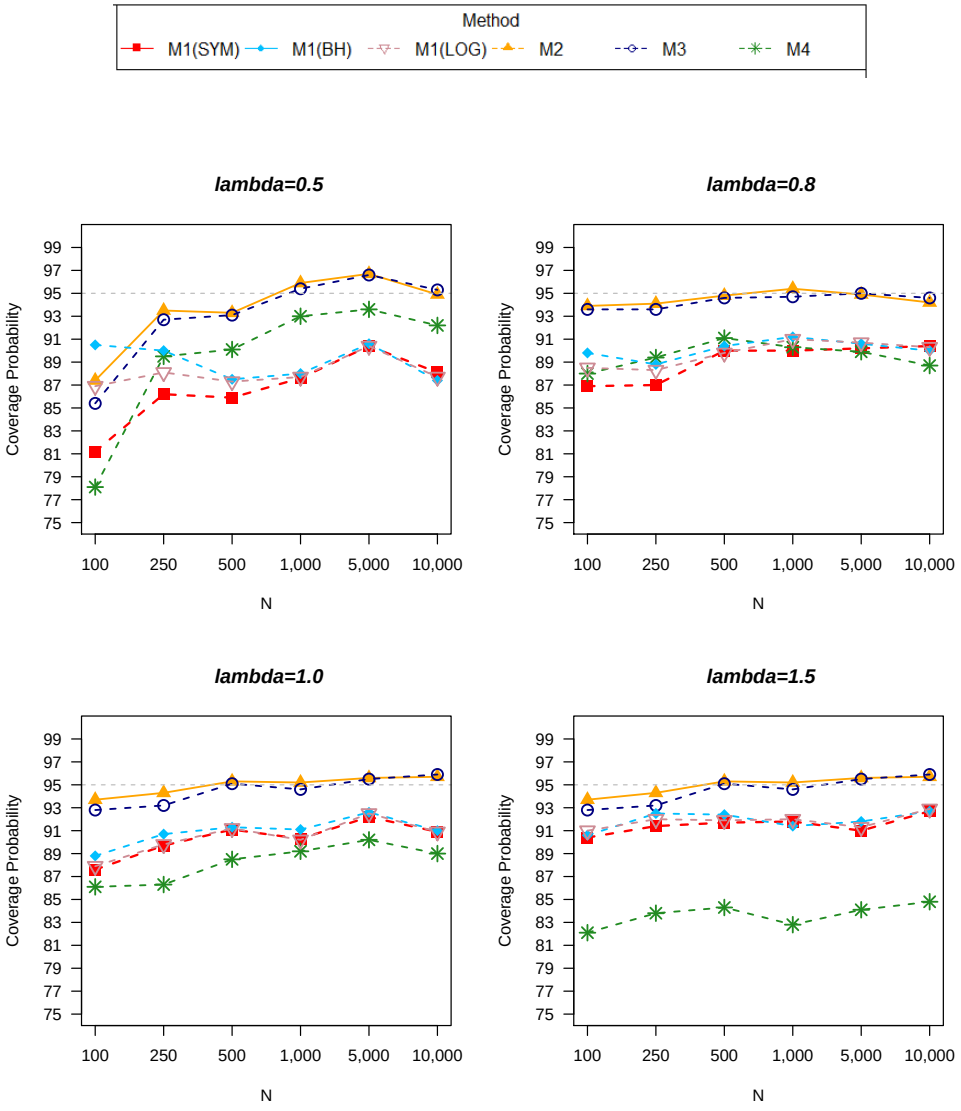

Figure 5.3: Coverage probabilities of 95% confidence interval when data are generated from the Poisson distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

5.2.2 Simulation results based on the geometric distribution

1) Relative bias and relative variance of the LCMP estimator when data are generated from the geometric distribution

Since the geometric distribution is one of the special cases of the CMP distribution when dispersion parameter $\nu = 0$ and $0 < \lambda < 1$, data are generated from the geometric distribution. Again, as shown in Chapter 4, the LCMP estimator is asymptotically unbiased with respect to the population size under the geometric distribution. Additionally, it is found that the level of accuracy increases when λ increases. In particular, the precision of the LCMP estimator is very high even for small population size and/or high level of unobserved probability (see Figure 5.4).

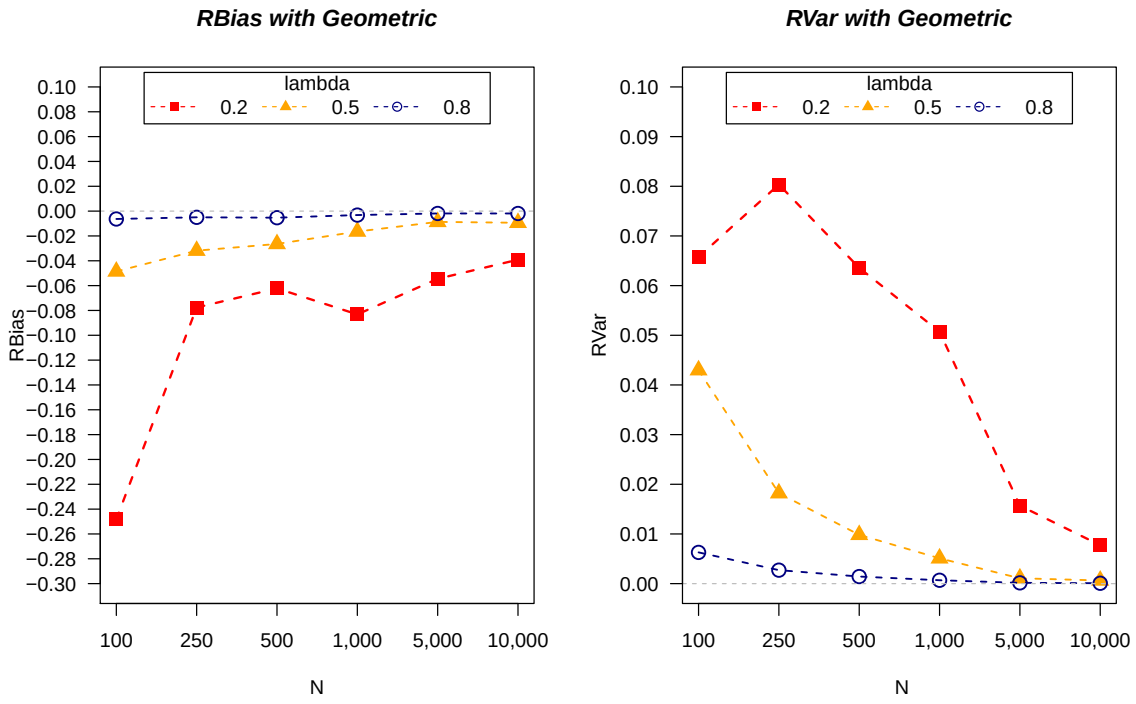

Figure 5.4: Relative bias (left) and relative variance (right) of population size estimation based on geometric distribution, $Geo(\lambda)$

2) Variance estimation and standard error of LCMP estimator when data are generated from the geometric distribution

To explore the behaviour of variance estimates of the LCMP estimator, standard errors derived using four approaches are compared with the true value. The estimated standard errors are summarised in Table 5.3 and in Figure 5.5. The variance estimation based on the normal approximation approach shows a significant overestimation of standard error especially for small population size and/or a high number of unobserved data (λ is small). The true bootstrap and the imputed bootstrap both give results close to the true standard errors. The reduced bootstrap often results in an underestimation of the standard error, but the simulation results suggest that it might be useful for small λ and $N \geq 500$.

Table 5.3: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the geometric distribution: $Geo(\lambda)$

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
		M1	M2	M3	M4
		$Geo(0.2)$			
100	25.646	159.104	23.147	22.614	15.649
250	70.866	364.607	63.972	63.888	54.728
500	126.067	485.015	119.372	119.282	110.798
1,000	225.214	624.888	208.529	208.270	199.691
5,000	626.445	1320.431	610.528	610.078	593.016
10,000	879.202	1,851.765	905.151	905.371	881.499
		$Geo(0.5)$			
100	20.730	45.662	20.397	20.286	18.028
250	33.731	55.851	33.987	33.868	30.281
500	49.557	77.354	49.378	49.360	44.262
1,000	71.383	109.866	71.567	71.522	64.056
5,000	164.102	247.637	168.934	168.791	153.129
10,000	251.886	348.331	245.386	245.442	223.983
		$Geo(0.8)$			
100	7.930	18.297	8.351	8.284	6.730
250	13.026	15.492	13.237	13.174	10.654
500	18.877	22.315	18.880	18.819	15.226
1,000	26.121	31.968	26.934	26.858	21.806
5,000	62.786	72.507	61.697	61.641	50.609
10,000	90.015	103.133	88.158	88.225	72.411

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

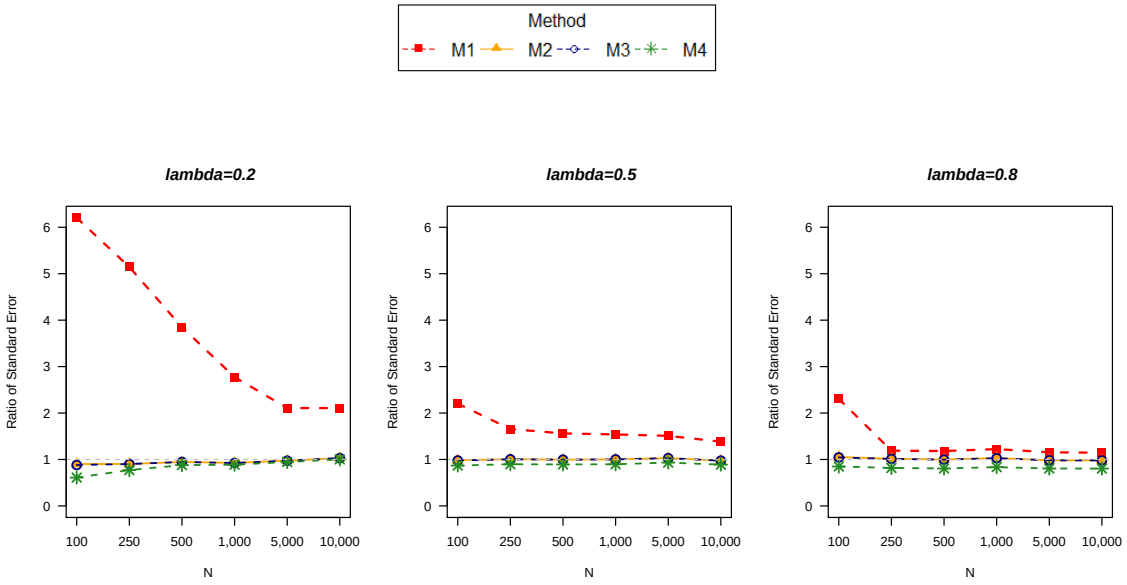

Figure 5.5: Ratio of standard errors when data are generated from the geometric distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

3) Coverage probability of the LCMP estimation when data are generated from the geometric distribution

The constructed 95% confidence intervals for the LCMP estimator for the geometric model are compared via their coverage probabilities, which are summarised in 5.4 and Figure 5.6. Although the LCMP estimator is an asymptotic unbiased estimator with respect to the population size, its estimation shows slight underestimation of population sizes in some simulation scenarios. The results show that the distribution of the estimated parameter is likely to be skewed for $\lambda = 0.2$. This might be due to the inefficiency of the LCMP estimator for large numbers of unobserved counts. However, the estimated population size tends to be distributed more symmetric when the number in the unobserved population is small (as λ increases) as can be seen by comparing the $\widehat{N}_{Mean}$ and $\widehat{N}_{Med}$ of parameter estimation.

For large values of unobserved population size with $\lambda = 0.2$, the symmetrical confidence interval, M1(SYM), as well as its corrected confidence intervals: M1(BH) and M1(LOG), show the best performance for $N < 1,000$, with the coverage probabilities being close to the nominal level. This result might stem from the very large overestimation of standard error leading to very wide length in confidence intervals.

Furthermore, for λ no less than 0.5, and although the M1(BH) provides the highest coverage probability, it is not a good choice for constructing the confidence interval. It is recommend that the true bootstrap method is used in the case of known true population and the imputed bootstrap is used for unknown population size. The reason for this is that the simulation results suggest that the true and imputed bootstrap give a

very good performance on average, and we can see their coverage probabilities are close to the nominal level.

The reduced bootstrap performs worst overall based on the geometric distribution, and results in low coverage probability compared with the nominal level.

Table 5.4: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from the geometric distribution: $Geo(\lambda)$

			Approximate Normal (M1)			Bootstrap Method		
N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	SYM	BH	LOG	M2	M3	M4
				$Geo(0.2)$				
100	75.2	73.0	93.3	97.6	97.0	72.8	69.0	42.4
250	230.6	223.0	96.1	97.7	97.6	91.0	89.3	85.1
500	468.9	470.0	92.4	95.4	95.1	92.3	92.1	89.7
1,000	916.8	938.5	86.9	89.8	89.6	93.2	93.3	91.6
5,000	4,727.4	4,859.0	91.1	94.0	94.0	93.6	92.8	90.8
10,000	9,608.8	9,797.0	94.8	96.4	96.4	95.5	95.0	93.7
				$Geo(0.5)$				
100	95.1	94.0	89.2	94.0	91.5	92.9	92.0	88.9
250	242.0	242.0	92.8	97.0	95.0	94.8	93.4	90.5
500	486.8	490.0	93.4	96.4	94.7	94.4	94.3	90.9
1,000	983.6	989.5	94.6	97.3	96.1	94.5	94.6	90.6
5,000	4,956.6	4,972.5	96.4	96.7	96.7	95.1	95.2	92.4
10,000	9,906.0	9,942.0	95.3	96.2	95.9	93.1	93.3	89.3
				$Geo(0.8)$				
100	99.4	99.0	92.0	98.1	93.6	94.8	93.5	87.7
250	248.8	249.0	93.8	96.8	94.4	94.5	93.7	87.5
500	497.3	498.0	94.3	96.3	95.2	94.0	92.8	86.4
1,000	996.8	997.0	95.6	96.7	96.0	94.4	94.0	88.9
5,000	4,990.6	4,994.0	96.4	97.1	96.8	94.6	94.4	87.8
10,000	9,981.0	9,983.5	96.4	96.6	96.5	94.4	93.9	86.9

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

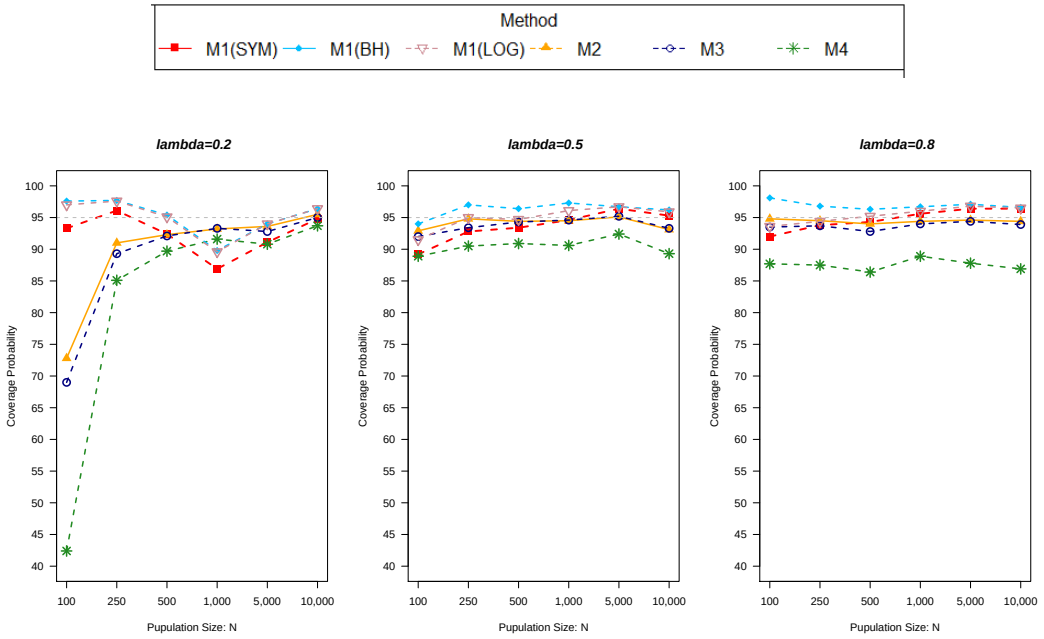

Figure 5.6: Coverage probabilities of 95% confidence interval when data are generated from the geometric distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

5.2.3 Simulation results based on the Conway-Maxwell-Poisson distribution

1) Relative bias and relative variance of the LCMP estimator when data are generated from the CMP distribution

The LCMP estimator is an asymptotic unbiased estimator with respect to the population size as pointed out by means of analytical and simulation studies as provided in Chapter 4. Indeed, the LCMP estimator works very well for strong over-dispersion or/and large population size. As can be seen in Figures 5.7 and 5.8, its relative bias and relative variance are close to zero. We remark that there is a decrease in the level of dispersion with increasing value of ν, resulting in lower accuracy and precision for small and medium population sizes.

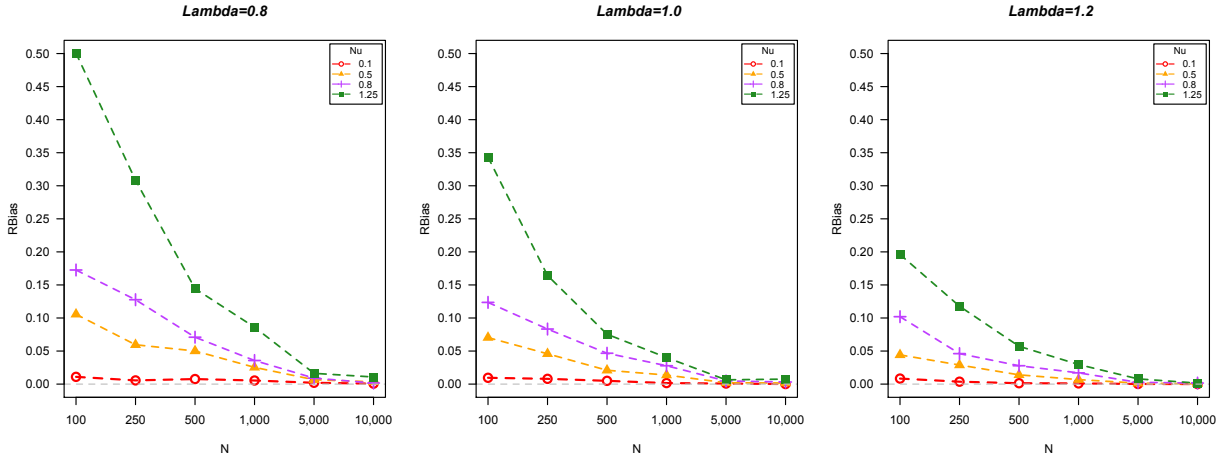

Figure 5.7: Relative bias of the LCMP estimator when data are generated from the CMP distribution

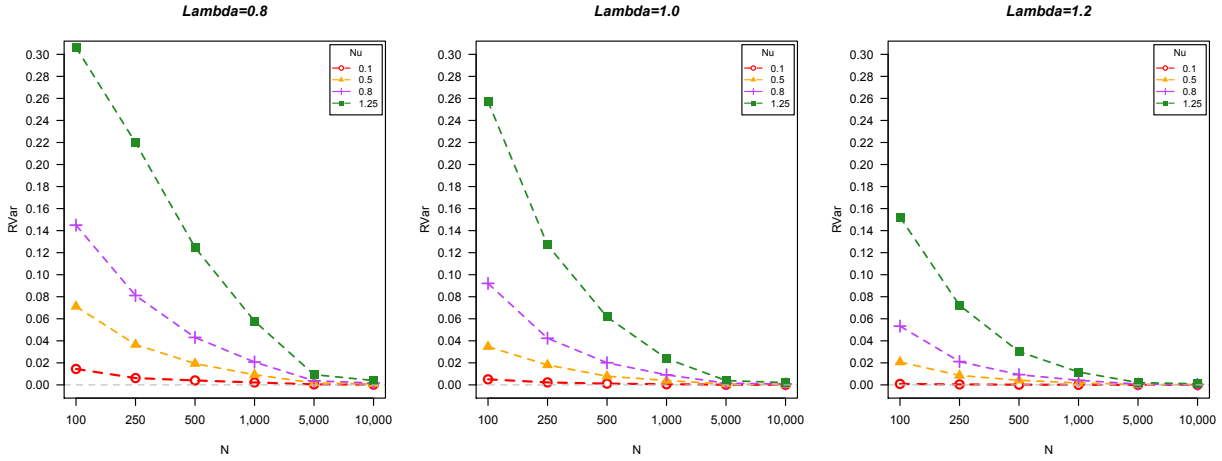

Figure 5.8: Relative variance of the LCMP estimator when data are generated from the CMP distribution

2) Variance estimation and standard error of the LCMP estimator when data are generated from the CMP distribution

The simulation results are represented in Tables 5.5, 5.6, 5.7, and the ratio of standard errors in Figure 5.9. The findings suggest that the true bootstrap might be the best choice for estimating the standard error of the LCMP estimator based on the CMP distribution. However, in an application, the true bootstrap can only be used if the population size is known. In real data analysis, the true population size is often unknown, and so the imputed bootstrap would be the more realistic and valid method for estimating standard error.

The reduced bootstrap tends to result in an underestimation for the higher level of dispersion (ν tends to zero), and it performs more poorly than variance estimation based on the normal approximation method. However, the reduced bootstrap might be useful for situations with weak dispersion (i.e. $\nu \to 1$) for the medium and large population sizes. Nevertheless, the reduced bootstrap is not recommended for estimating the variance of the LCMP estimator.

It is important to highlight that the approximate normal variance approach provides a poor standard error, particularly for the small population size. However, it is still useful for estimating standard error for large population size. Moreover, it is not complicated and does not require an intensive computational approach.

Table 5.5: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the CMP distribution: $CMP(\lambda, \nu)$

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
		M1	M2	M3	M4
$CMP(0.8, 0.1)$					
100	12.023	21.715	11.741	11.754	10.020
250	19.678	24.049	21.508	19.817	19.556
500	31.760	34.434	29.870	29.894	26.732
1,000	47.005	49.188	43.922	43.988	39.437
5,000	109.311	112.037	108.204	108.284	99.232
10,000	153.994	156.747	153.395	153.115	140.677
$CMP(0.8, 0.5)$					
100	26.680	38.825	24.497	24.779	22.828
250	47.845	53.472	43.327	43.513	41.063
500	69.560	73.611	68.177	68.353	65.476
1,000	96.083	98.496	101.024	100.986	97.541
5,000	207.516	208.044	215.034	215.211	207.459
10,000	292.139	292.074	298.754	298.572	287.687
$CMP(0.8, 0.8)$					
100	38.080	53.399	34.778	35.185	33.187
250	71.200	79.817	66.163	66.524	64.279
500	103.777	104.119	101.537	101.642	99.227
1,000	144.014	135.617	148.517	148.494	145.763
5,000	304.545	281.586	305.187	305.256	298.997
10,000	414.295	398.580	417.793	417.233	408.376
$CMP(0.8, 1.25)$					
100	55.348	119.345	49.122	50.057	46.506
250	117.211	131.931	106.868	107.384	104.935
500	176.528	149.266	167.407	167.580	165.497
1,000	240.457	216.875	249.640	250.158	247.585
5,000	482.323	391.332	480.428	480.879	476.291
10,000	630.449	579.953	654.980	654.504	648.635

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.6: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the CMP distribution: $CMP(\lambda, \nu)$

		$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
N	$S.E.(\widehat{N})_{True}$	M1	M2	M3	M4
		$CMP(1.0, 0.1)$			
100	7.102	11.385	6.724	6.741	5.460
250	11.875	12.190	11.126	11.205	9.250
500	16.831	17.068	16.521	16.588	14.003
1,000	24.531	24.134	24.087	24.127	20.673
5,000	56.692	55.169	54.559	54.561	47.074
10,000	75.288	77.867	76.584	76.566	66.087
		$CMP(1.0, 0.5)$			
100	18.629	25.009	17.861	18.081	16.510
250	33.635	34.029	32.079	32.583	30.296
500	44.377	44.483	46.640	46.711	44.280
1,000	63.076	61.778	65.715	65.788	62.402
5,000	140.760	136.390	137.565	137.535	129.966
10,000	186.795	192.018	192.613	192.742	181.595
		$CMP(1.0, 0.8)$			
100	30.351	36.043	26.959	27.273	25.624
250	51.457	52.013	49.357	49.548	47.681
500	70.809	66.736	73.137	73.372	71.323
1,000	95.335	90.835	100.812	100.884	98.018
5,000	198.338	186.427	200.759	200.510	193.984
10,000	279.543	267.505	279.865	279.671	270.243
		$CMP(1.0, 1.25)$			
100	50.715	68.601	44.260	44.891	42.724
250	89.082	84.672	82.818	83.135	81.315
500	124.120	105.120	121.300	121.411	119.724
1,000	154.622	132.239	167.733	168.005	165.673
5,000	313.537	285.387	319.611	319.416	314.477
10,000	452.885	402.916	443.416	443.595	435.482

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.7: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the CMP distribution: $CMP(\lambda, \nu)$

		$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
N	S.E$(\widehat{N})_{True}$	M1	M2	M3	M4
		$CMP(1.2, 0.1)$			
100	2.960	4.914	2.995	3.115	2.313
250	5.119	5.440	4.889	4.934	3.772
500	6.897	6.939	6.980	6.980	5.445
1,000	9.796	9.847	9.988	10.005	7.873
5,000	22.344	21.939	21.808	21.837	17.042
10,000	31.205	31.099	30.814	30.766	24.038
		$CMP(1.2, 0.5)$			
100	14.398	16.239	13.212	13.345	12.065
250	23.177	22.732	22.815	22.919	21.172
500	32.268	29.669	32.406	32.468	30.103
1,000	41.782	41.498	43.854	44.005	40.660
5,000	91.651	92.734	93.232	93.361	85.787
10,000	129.361	130.847	130.802	130.689	120.152
		$CMP(1.2, 0.8)$			
100	23.093	27.101	21.512	21.727	20.361
250	36.364	34.409	36.652	36.801	35.237
500	48.192	47.664	52.383	52.507	50.560
1,000	64.592	62.866	69.536	69.560	66.695
5,000	136.824	133.268	141.713	141.771	134.982
1,0000	197.144	187.446	198.334	198.075	188.635
		$CMP(1.2, 1.25)$			
100	39.017	42.861	36.412	36.827	35.426
250	67.196	61.757	65.403	65.610	64.224
500	86.901	75.436	91.542	91.558	89.988
1,000	108.155	96.815	119.082	119.207	116.777
5,000	229.720	205.916	231.600	231.594	226.356
10,000	305.133	291.161	316.419	316.831	308.882

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

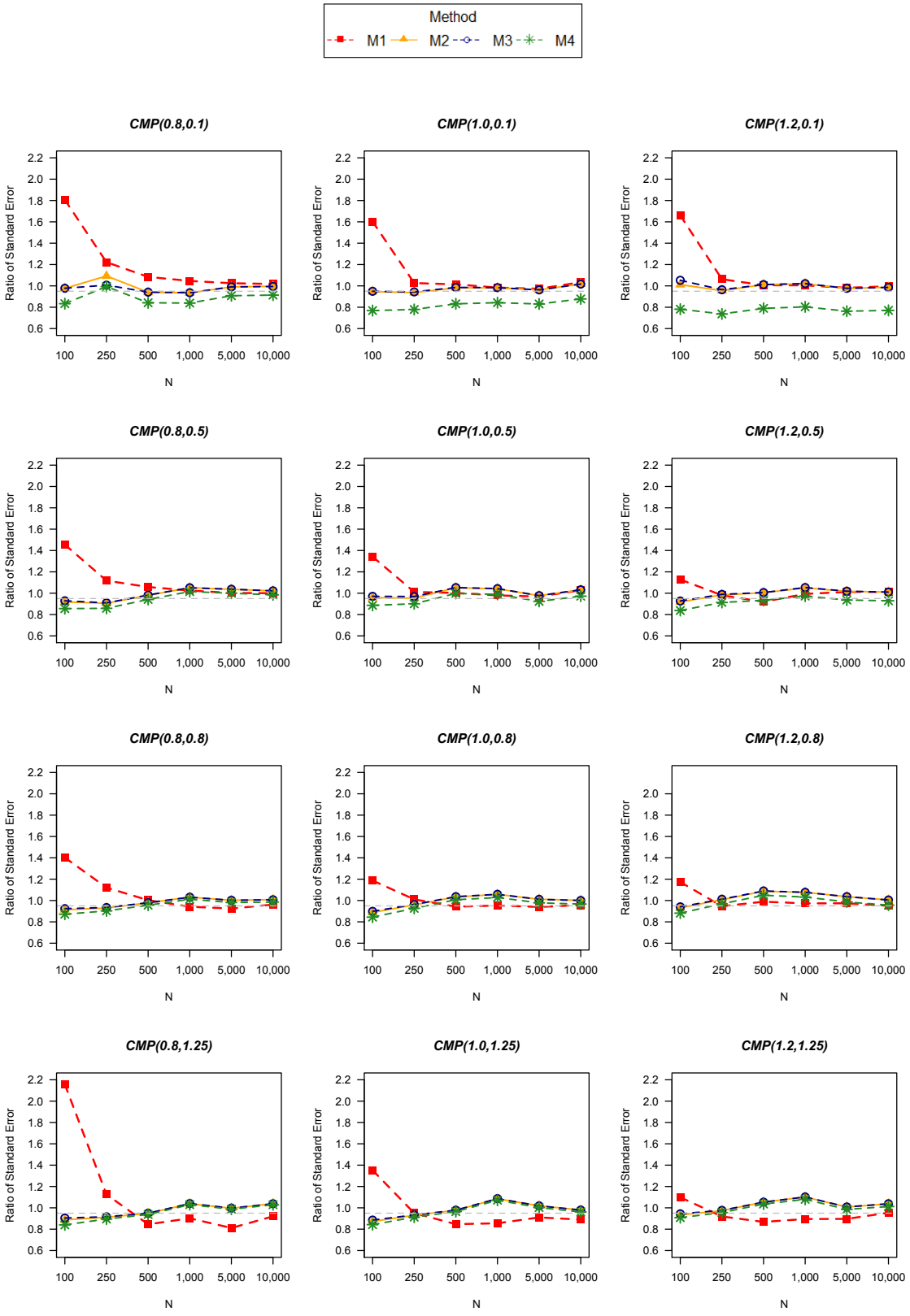

Figure 5.9: Ratio of standard errors when data are generated from the CMP distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

3) Coverage Probability of the LCMP estimator when data are generated from the CMP distribution

The performance of confidence interval estimation using coverage probability and are summarised in Tables 5.8, 5.9, 5.10 and Figure 5.10. It can be seen that all methods provide confidence intervals with some biases, the true bootstrap gives the best coverage probability on average which is close to the nominal confidence interval level set at 95%. However, in reality, the imputed bootstrap should be used since the true population size N in the capture-recapture is not usually available and needs to be estimated.

Indeed, in the case of a high degree of over-dispersion, ($\nu = 0.1$), the normal approximation method is useful for medium and large population sizes, particularly when λ is not less than 1.0. In these situations the LCMP estimator is also previously shown to perform very well with largely unbiased estimates of population size and low standard errors. Therefore, if the dispersion parameter close to zero and population size medium or large the recommended method for confidence interval estimation is the symmetric normal approximation approach, M1(SYM), because it is relatively easy and not computationally intensive.

Interestingly, when the dispersion parameter is increased, all of the normal approximations tend to be inappropriate. The reason, as we stated before, is that the population size estimations might be effected by the increasing dispersion parameter value. This evidence leads to lower coverage of the confidence interval. Although the normal approximations out perform over the bootstrap approaches for $CMP(0.8, 1.25)$ for $N = 100$, the substantial method to construct intervals for the LCMP estimator should based on the resampling approach. Overestimation of standard error can lead to wider length of the confidence interval.

Table 5.8: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from Conway-Maxwell-Poisson distribution: $CMP(\lambda, \nu)$

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	Approximate Normal (M1)			Bootstrap Method		
			SYM	BH	LOG	M2	M3	M4
$CMP(0.8, 0.1)$								
100	101.1	101.0	91.7	98.1	93.8	95.5	95.1	91.0
250	251.4	252.0	93.1	96.4	93.9	95.3	95.3	92.8
500	503.8	503.0	93.7	94.6	93.8	94.7	94.6	91.3
1,000	1,005.4	1,004.5	94.0	94.6	93.9	94.3	94.8	92.6
5,000	5,010.0	5,004.0	95.7	95.2	95.5	94.8	94.8	92.4
10,000	10,006.8	10,001.0	94.6	94.8	94.8	94.6	94.4	92.3
$CMP(0.8, 0.5)$								
100	110.6	108.0	87.0	88.0	87.5	95.5	95.8	94.2
250	264.9	257.0	89.4	89.5	89.0	93.8	94.0	93.0
500	525.1	518.0	93.0	91.4	92.1	93.6	93.4	92.7
1,000	1,025.4	1,018.0	92.3	91.1	92.2	96.0	95.5	94.9
5,000	5,033.0	5,022.5	92.5	91.3	91.7	94.9	94.5	93.9
10,000	10,034.5	10,030.0	92.1	92.6	92.3	95.9	95.7	94.8
$CMP(0.8, 0.8)$								
100	117.3	112.0	82.2	82.8	81.0	93.6	94.1	92.5
250	281.9	269.0	85.7	82.5	82.7	94.4	94.6	94.0
500	535.5	515.0	84.9	83.2	83.6	94.4	95.1	94.4
1,000	1,035.7	1,014.0	87.4	86.0	86.2	94.3	94.2	93.2
5,000	5,045.1	5,025.5	87.9	87.6	87.6	93.2	93.7	93.1
10,000	10,015.7	10,002.5	88.0	88.2	88.1	94.9	95.0	94.6
$CMP(0.8, 1.25)$								
100	150.1	146.0	90.8	85.9	86.2	83.6	84.4	82.3
250	327.1	304.0	81.6	72.4	74.3	91.3	91.4	91.3
500	572.2	532.5	74.9	72.6	72.0	94.3	94.1	93.7
1,000	1,086.0	1,041.0	79.8	77.5	78.4	94.9	94.6	94.6
5,000	5,081.5	5,030.5	79.5	79.4	78.9	94.2	94.2	93.6
10,000	10,107.4	10,075.0	83.5	83.8	83.5	95.1	94.8	94.3

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.9: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from Conway-Maxwell-Poisson distribution: $CMP(\lambda, \nu)$

			Approximate Normal (M1)			Bootstrap Method		
N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	SYM	BH	LOG	M2	M3	M4
			$CMP(1.0, 0.1)$					
100	101.0	101.0	91.7	96.4	92.8	93.1	92.9	86.5
250	252.0	252.0	94.3	93.5	94.6	93.8	93.8	88.0
500	502.5	502.0	94.9	93.8	95.2	95.3	95.1	91.0
1,000	1,001.6	1,001.0	94.3	94.7	94.7	94.7	94.6	91.1
5,000	5,004.3	5,001.5	94.3	94.3	94.2	93.6	93.5	89.9
10,000	10,000.8	9,996.5	96.0	96.0	96.0	95.2	95.3	92.0
			$CMP(1.0, 0.5)$					
100	107.0	104.5	91.3	92.0	92.0	94.8	95.5	93.9
250	261.5	258.0	91.0	90.3	90.3	94.3	94.2	92.3
500	510.5	504.0	92.1	91.4	92.1	95.1	95.1	93.4
1,000	1,013.7	1,008.5	92.5	92.3	91.9	96.0	95.5	94.7
5,000	5,010.0	5,005.0	91.2	91.6	91.5	93.5	93.9	92.6
10,000	10,016.3	10,013.0	93.9	94.1	94.1	95.5	95.4	94.0
			$CMP(1.0, 0.8)$					
100	112.4	107.0	82.8	82.0	81.6	94.5	95.0	92.4
250	270.8	262.0	84.9	83.8	84.2	94.1	94.0	93.5
500	523.3	514.0	88.2	87.6	87.5	94.5	94.7	93.5
1,000	1,027.8	1,016.0	88.8	87.6	87.8	94.4	93.9	93.8
5,000	5,023.8	5,014.5	88.9	89.2	89.4	95.4	95.4	94.7
10,000	10,031.8	10,020.5	89.8	89.6	89.6	95.1	94.7	94.1
			$CMP(1.0, 1.25)$					
100	134.3	122.0	84.3	77.2	79.3	88.6	87.9	86.9
250	291.1	267.0	77.6	75.5	74.9	94.3	94.2	94.1
500	537.7	512.5	78.9	78.8	78.2	94.1	94.2	93.8
1,000	1,040.2	1,017.5	80.5	79.2	79.5	94.1	94.0	93.6
5,000	5,032.4	4,991.5	84.4	83.7	83.9	94.9	94.8	93.9
10,000	10,075.4	10,043.0	83.8	83.6	83.5	94.0	93.9	93.5

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.10: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from Conway-Maxwell-Poisson distribution: $CMP(\lambda, \nu)$

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	Approximate Normal (M1)			Bootstrap Method		
			SYM	BH	LOG	M2	M3	M4
$CMP(1.2, 0.1)$								
100	100.8	101.0	93.8	93.0	94.0	94.6	94.1	77.2
250	250.9	251.0	92.1	92.7	92.7	92.2	91.5	81.9
500	500.7	500.0	93.9	93.6	94.3	94.5	94.3	84.4
1,000	1,001.1	1,001.0	94.6	95.3	95.1	94.8	94.3	86.1
5,000	5,000.9	5,001.0	94.3	94.4	94.9	94.1	93.3	85.0
10,000	10,000.6	10,000.0	94.5	94.8	94.8	94.4	94.1	85.6
$CMP(1.2, 0.8)$								
100	104.4	103.0	90.4	92.3	90.7	94.3	95.0	92.0
250	257.2	254.0	92.2	91.3	91.7	94.4	94.2	91.5
500	507.0	504.0	91.1	90.6	90.7	93.5	94.0	91.8
1,000	1,006.8	1,004.0	92.3	93.0	93.0	95.1	95.3	92.8
5,000	5,007.0	5,002.0	93.8	94.3	94.0	94.9	94.5	92.6
10,000	10,004.8	10,002.5	94.3	94.3	94.3	95.6	95.6	93.3
$CMP(1.2, 0.8)$								
100	110.2	106.0	88.2	86.0	87.0	94.1	94.7	92.0
250	261.5	255.0	88.7	86.6	87.2	94.3	95.3	93.1
500	513.9	507.5	91.5	90.1	91.1	94.8	94.9	93.5
1,000	1,017.0	1,012.0	91.1	89.5	89.9	94.3	94.2	93.2
5,000	5,013.2	5,007.0	91.7	91.0	91.7	95.4	95.7	94.2
10,000	10,015.5	10,005.5	90.0	90.4	90.4	94.7	95.1	93.4
$CMP(1.2, 1.25)$								
100	119.6	110.0	81.2	78.1	78.0	93.8	94.2	92.7
250	279.4	264.0	81.4	78.9	79.2	94.0	94.5	93.4
500	528.6	514.0	82.8	81.8	82.1	92.8	93.4	92.8
1,000	1,029.6	1,016.0	85.0	83.9	84.4	95.0	95.1	94.4
5,000	5,039.5	5,025.5	85.1	85.6	84.9	93.5	93.7	93.3
10,000	10,015.1	10,005.0	88.1	88.1	88.1	95.4	95.3	95.6

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

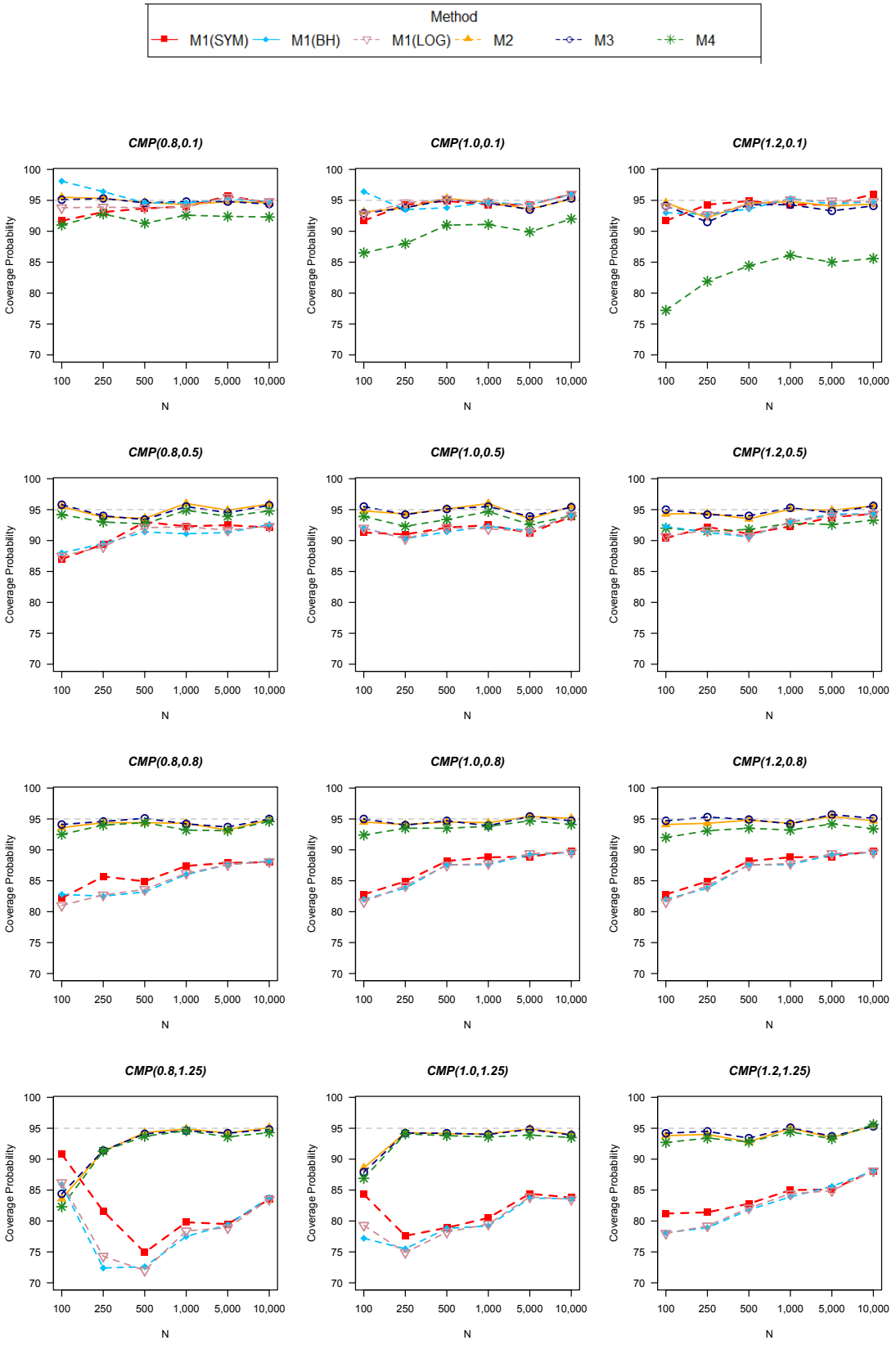

Figure 5.10: Coverage probabilities of 95% confidence interval when data are generated from the Conway-Maxwell-Poisson distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

5.2.4 Simulation results based on the negative binomial distribution

1) Relative bias and relative variance of the LCMP estimator when data are generated from the negative binomial distribution

The simulations suggest that the LCMP can be used for the negative binomial distribution, although it tends to largely overestimate for small population sizes and small λ as shown in Figure 5.11. Additionally, the relative variance decreases dramatically to zero when population size becomes large or/and λ approaches one (see Figure 5.12).

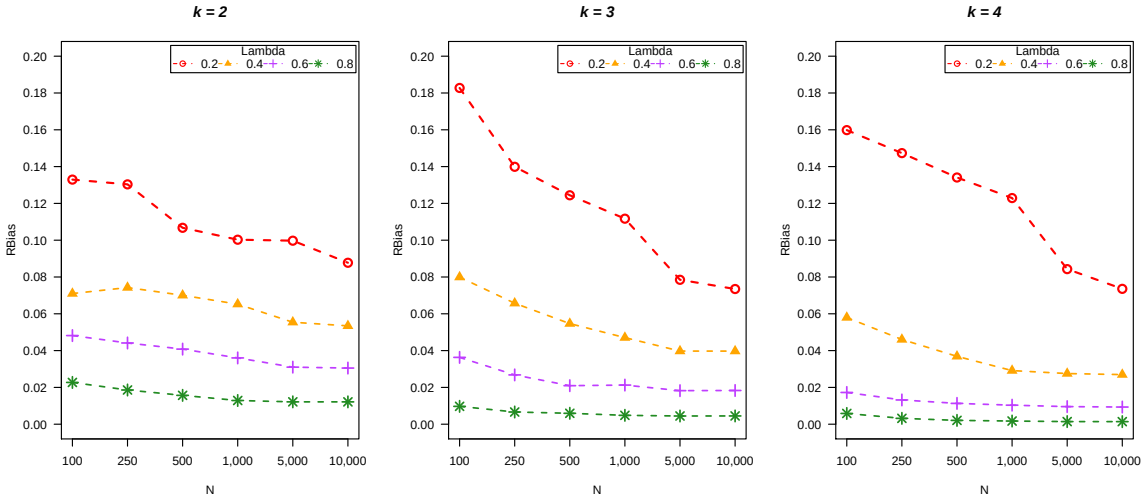

Figure 5.11: Relative bias of the LCMP estimator when data are generated based on the negative binomial distribution

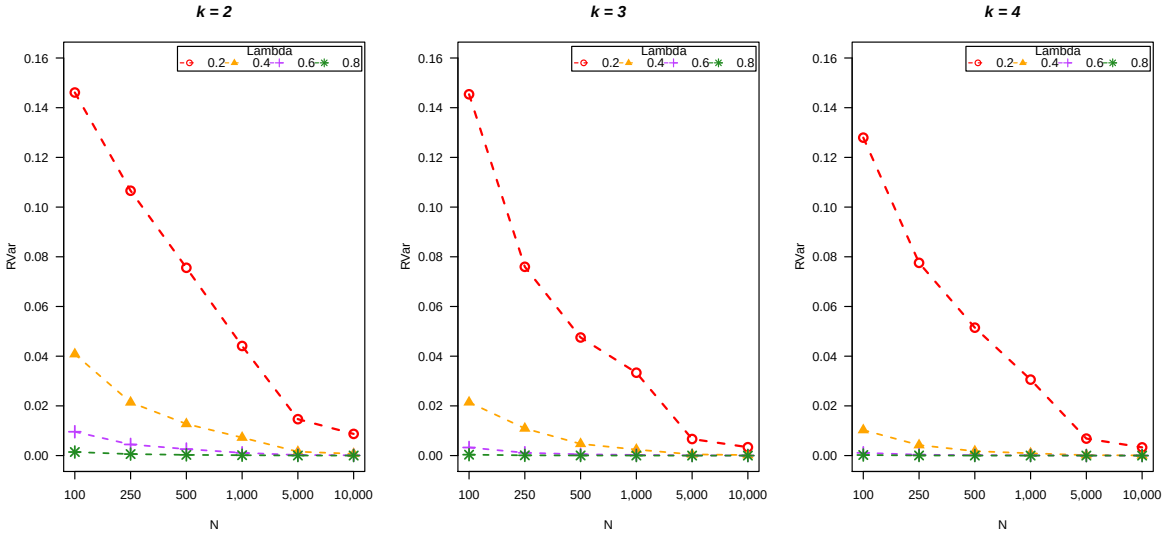

Figure 5.12: Relative variance of LCMP estimator when data are generated based on the negative binomial distribution

2) Variance estimation and standard errors of the LCMP estimator when data are generated from the negative binomial distribution

The performance of the different methods of standard error estimation were assessed in the simulation study, and the results are tabulated in Tables 5.11, 5.12 and 5.13, with the ratio of standard errors are given in Figure 5.13. It can be seen that the variance estimation based on the normal approximation method tends to overestimate the standard error for the small population and low values of λ on average. The true bootstrap method remains the best way for estimating standard error of the LCMP estimator if the true population size is given. As previously pointed out, this makes it useless in real life situations. Therefore, the imputed bootstrap is more realistic and should be used as the method of choice for small population sizes. Nevertheless, the variance estimation based on the normal approximation method can be used if the population size increase and parameters increase.

Table 5.11: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the $NB(k, \lambda)$ distribution

		$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
N	S.E$(\widehat{N})_{True}$	M1	M2	M3	M4
		$NB(2, 0.2)$			
100	38.225	130.943	34.274	34.685	29.915
250	81.619	182.437	75.288	75.715	71.108
500	137.464	238.091	125.015	125.261	120.328
1,000	210.028	292.100	199.949	200.274	193.730
5,000	605.269	665.696	576.017	575.635	566.374
10,000	934.069	950.235	895.921	895.692	882.719
		$NB(2, 0.4)$			
100	20.208	33.925	19.508	19.685	17.721
250	36.586	45.022	34.109	34.437	31.549
500	56.433	61.376	52.760	53.119	49.526
1,000	85.127	87.981	80.837	81.205	76.781
5,000	197.975	186.919	194.718	195.248	186.370
10,000	269.745	270.972	272.060	273.302	260.587
		$NB(2, 0.6)$			
100	9.793	14.194	9.340	9.572	8.140
250	16.726	16.852	15.968	16.322	14.249
500	25.553	24.074	23.763	24.200	21.475
1,000	33.345	33.338	34.283	34.828	31.180
5,000	75.567	74.214	74.564	75.657	67.457
10,000	100.202	107.380	104.870	106.330	94.699
		$NB(2, 0.8)$			
100	3.786	5.530	3.698	3.969	3.061
250	6.119	6.180	6.051	6.374	5.067
500	8.497	8.572	8.555	8.968	7.176
1,000	12.084	11.956	11.903	12.383	9.896
5,000	24.892	26.908	26.061	27.162	21.552
10,000	38.062	38.336	36.693	38.271	30.346

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.12: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the $NB(k, \lambda)$ distribution

		$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
N	S.E$(\widehat{N})_{True}$	M1	M2	M3	M4
		$NB(3, 0.2)$			
100	38.132	75.975	33.969	34.408	31.538
250	68.911	102.606	64.062	64.403	61.090
500	109.003	133.832	103.993	104.401	100.636
1,000	182.646	187.859	163.315	163.538	159.173
5,000	408.257	402.482	420.152	420.142	411.538
10,000	581.158	572.471	585.028	586.374	574.587
		$NB(3, 0.4)$			
100	14.646	19.316	13.852	14.147	12.658
250	26.129	25.403	24.041	24.422	22.292
500	34.345	34.867	35.834	36.212	33.608
1,000	49.179	47.488	50.218	50.700	47.082
5,000	109.327	107.190	107.203	107.981	99.860
10,000	148.543	156.291	150.606	151.998	140.330
		$NB(3, 0.6)$			
100	5.699	6.260	5.530	5.809	4.806
250	8.411	8.625	8.898	9.217	7.790
500	11.298	11.719	12.090	12.471	10.484
1,000	16.363	16.784	16.860	17.450	14.590
5,000	36.424	37.821	36.358	37.577	31.022
10,000	50.772	54.697	51.175	52.930	43.731
		$NB(3, 0.8)$			
100	1.902	2.545	2.081	2.328	1.663
250	2.477	2.723	2.641	2.946	2.116
500	3.294	3.574	3.358	3.768	2.678
1,000	4.632	4.920	4.570	5.052	3.557
5,000	10.066	11.069	10.038	11.091	7.755
10,000	14.192	15.821	14.184	15.677	10.977

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.13: Comparison of the standard errors of the four methods with the true standard error of the LCMP estimator when data are generated from the $NB(k, \lambda)$ distribution

		$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
N	S.E$(\widehat{N})_{True}$	M1	M2	M3	M4
		$NB(4, 0.2)$			
100	35.771	70.478	33.805	34.173	31.483
250	69.630	103.827	65.269	65.690	62.472
500	113.449	136.658	102.652	103.162	99.187
1,000	174.871	188.070	163.119	164.001	159.139
5,000	413.570	401.568	421.925	422.587	414.454
10,000	575.136	573.154	586.487	586.139	575.690
		$NB(4, 0.4)$			
100	10.139	11.287	9.770	10.015	8.830
250	16.240	15.580	16.323	16.630	15.018
500	21.055	21.173	22.543	22.916	20.644
1,000	30.455	29.053	30.110	30.589	27.346
5,000	64.133	66.520	65.302	66.295	58.905
10,000	89.125	96.791	91.791	93.082	82.660
		$NB(4, 0.6)$			
100	3.322	3.367	3.283	3.509	2.759
250	4.720	4.702	4.764	5.054	3.952
500	5.944	6.495	6.408	6.804	5.227
1,000	8.833	9.167	8.874	9.413	7.165
5,000	20.506	20.823	19.388	20.544	15.530
10,000	26.617	30.147	27.254	28.995	21.842
		$NB(4, 0.8)$			
100	1.488	2.225	1.802	2.006	1.421
250	1.699	1.947	1.898	2.138	1.455
500	1.959	2.092	1.979	2.270	1.468
1,000	2.269	2.487	2.217	2.618	1.661
5,000	4.309	4.965	4.254	5.026	3.187
10,000	5.761	7.036	5.977	7.043	4.465

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

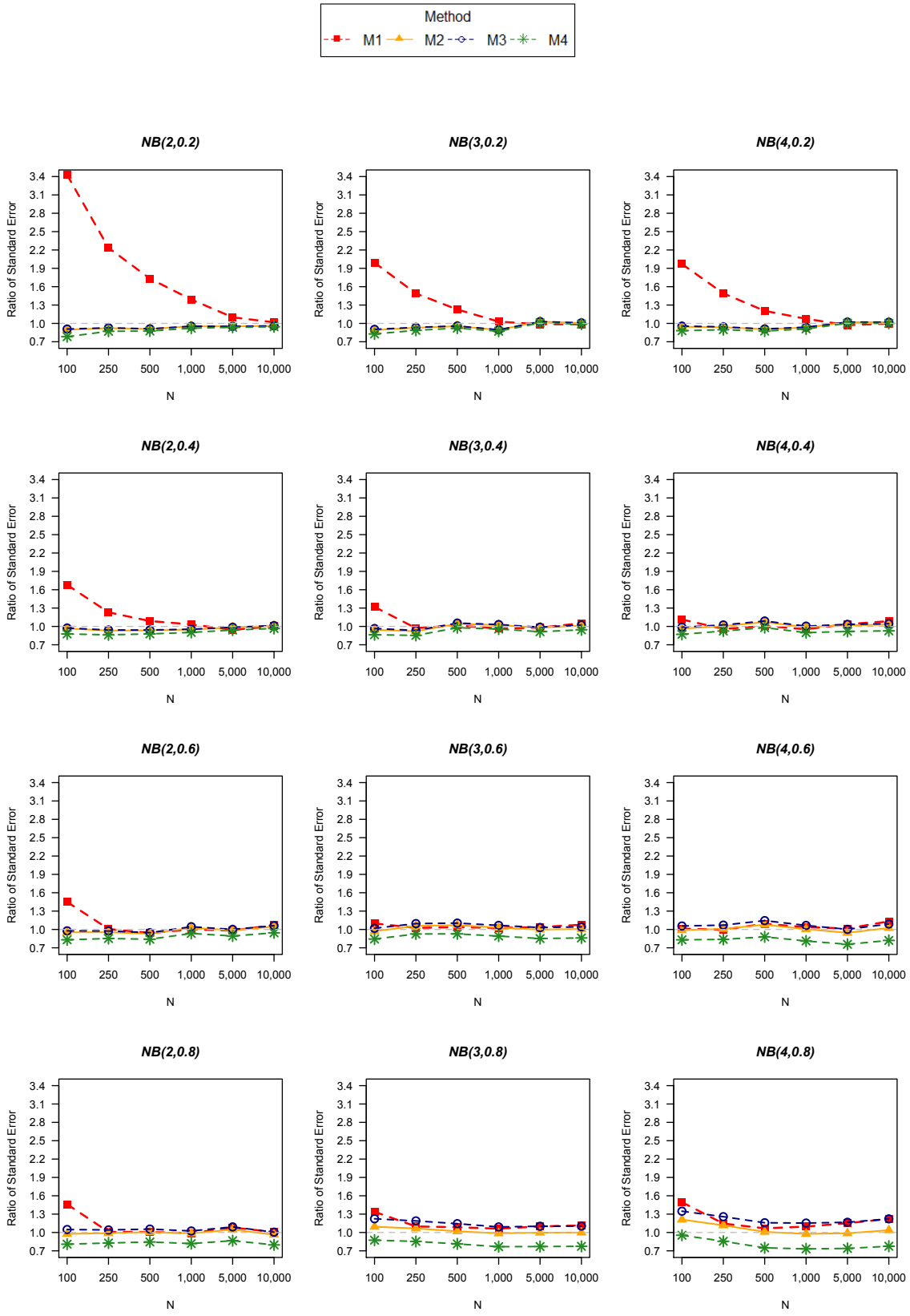

Figure 5.13: Ratio of standard errors when data are generated from the negative binomial distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

3) Coverage probability of the LCMP estimator when data are generated from the negative binomial distribution

The simulation results show that the confidence intervals obtained by the imputed bootstrap (M3), M1(SYM) and M1(LOG) perform better than their competitors with less bias in confidence intervals for the population size $N \leq 1,000$ (see Figure 5.14). This is supported by their coverage probabilities being close to the nominal level of 95%. Additionally, it was found that the variance estimation based on the normal approximation performs well for $\lambda \geq 0.4$ and/or $k \geq 2$. It might be that the estimated population size become more symmetric as indicated by the fact that means are approximately equivalent to medians (see Tables 5.14, 5.15, and 5.16). Although the log-transform of $\widehat{N}$, M1(LOG), can improve the traditional symmetric normal approximation (SYM) in some cases, it resulted in minimal differences within this simulation study.

However, when the population size tends to be large, all of the approaches provide poor confidence intervals as we can see by the fact that the coverage probabilities decrease and reach to zero. The possible reason for this could be that when the observed data approaches the population size, the efficiency of the standard errors improves and leads to narrow confidence intervals. This results in the confidence interval not including the true population size.

The results indicate that the imputed bootstrap (M3) is the best method on average. The reduced bootstrap is not recommended for constructing confidence intervals under model misspecification. It exhibits the worst performance for almost all situations, particular when the event parameter λ increases.

Table 5.14: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from $NB(k, \lambda)$ distribution

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	Approximate Normal (M1)			Bootstrap Method		
			SYM	BH	LOG	M2	M3	M4
$NB(2, 0.2)$								
100	113.3	110.0	94.1	96.0	95.2	95.9	95.8	92.2
250	282.6	287.0	90.8	92.2	92.1	93.7	94.6	92.2
500	553.4	567.0	88.7	90.2	89.7	94.4	94.2	93.1
1,000	1,100.3	1,118.5	88.2	86.8	86.3	93.5	93.2	92.1
5,000	5,498.7	5,480.0	86.4	82.7	83.9	88.8	88.5	86.9
10,000	10,877.5	10,825.5	82.2	77.5	78.2	82.8	83.2	82.1
$NB(2, 0.4)$								
100	107.1	107.0	91.8	95.2	93.3	95.9	95.7	93.4
250	268.6	266.0	93.2	91.8	92.4	93.5	94.2	91.3
500	535.0	532.0	92.1	88.2	90.0	91.7	91.3	89.0
1,000	1,065.2	1,058.0	91.2	84.4	87.6	87.5	87.6	85.9
5,000	5,277.1	5,267.5	68.6	63.5	66.1	66.5	66.7	64.8
10,000	10,534.5	10,527.0	49.0	44.9	47.4	43.8	43.8	39.6
$NB(2, 0.6)$								
100	104.8	104.0	95.4	94.1	96.3	92.2	94.3	89.1
250	261.0	260.0	93.8	85.8	92.6	90.9	92.9	86.9
500	520.4	518.0	89.2	80.6	87.6	85.2	86.3	80.5
1,000	1,036.0	1,034.0	85.2	77.3	83.4	82.6	84.1	76.3
5,000	5,155.1	5,151.0	45.5	40.2	44.8	40.9	42.0	32.6
10,000	10,305.1	10,305.0	16.7	14.6	16.5	13.1	13.9	8.8
$NB(2, 0.8)$								
100	102.3	102.0	96.1	94.1	96.5	89.6	95.1	75.4
250	254.6	254.0	94.1	85.8	94.4	86.8	92.5	74.8
500	507.8	507.0	89.8	80.6	91.0	82.5	88.6	70.7
1,000	1,012.9	1,012.0	83.1	77.3	84.3	78.6	82.5	67.2
5,000	5,060.7	5,060.0	35.3	40.2	36.2	31.0	33.4	17.3
10,000	10,121.3	10,121.0	9.2	14.6	9.7	7.6	8.3	2.6

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.15: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from $NB(k, \lambda)$ distribution

			Approximate Normal (M1)			Bootstrap Method		
N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	SYM	BH	LOG	M2	M3	M4
			$NB(3, 0.2)$					
NB(3,0.2)								
100	118.3	115.5	88.7	91.0	89.3	94.1	94.2	91.4
250	285.0	286.0	87.8	86.4	86.9	94.4	94.4	93.1
500	562.2	557.0	88.9	85.7	86.3	92.9	92.7	91.6
1,000	1,111.7	1,096.5	87.1	82.2	83.7	89.9	90.4	89.2
5,000	5,392.4	5,380.0	82.5	77.3	78.4	83.2	83.4	81.9
10,000	10,734.7	10,696.0	71.8	67.5	69.5	70.3	70.8	69.5
			$NB(3, 0.4)$					
100	108.0	106.0	95.3	92.2	95.1	93.1	95.1	91.2
250	266.4	263.5	91.4	84.1	89.9	89.5	91.8	86.8
500	527.3	523.0	91.6	88.2	90.0	88.8	90.5	84.9
1,000	1,047.1	1,045.0	85.3	76.0	83.2	82.8	84.5	78.3
5,000	5,198.7	5,200.0	54.9	48.0	52.9	48.3	49.1	42.4
10,000	10,397.2	10,395.0	26.6	22.7	25.8	19.3	19.6	14.6
			$NB(3, 0.6)$					
100	103.6	103.0	96.4	82.1	96.1	88.5	93.7	78.6
250	256.7	256.0	94.7	80.3	94.8	85.7	92.0	77.1
500	510.5	510.0	91.8	79.0	92.0	84.6	88.8	74.6
1,000	1,021.3	1,020.0	79.2	67.0	79.6	72.2	76.0	60.1
5,000	5,091.2	5,090.0	30.6	24.9	31.2	24.8	26.0	14.5
10,000	10,183.2	10,181.5	6.4	3.9	6.4	3.2	3.3	1.4
			$NB(3, 0.8)$					
100	101.0	101.0	97.5	82.1	97.8	97.9	98.5	44.5
250	251.7	251.0	97.1	80.3	98.0	90.3	97.1	57.5
500	503.0	503.0	92.1	79.0	96.4	80.9	94.1	57.2
1,000	1,004.8	1,005.0	86.1	67.0	90.9	75.8	88.0	55.1
5,000	5,022.3	5,022.0	42.8	24.9	48.0	35.6	44.3	15.9
10,000	10,044.9	10,045.0	11.7	3.9	13.6	9.3	11.4	2.6

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

Table 5.16: Comparison of the six methods to construct confidence intervals for the LCMP estimator when data are generated from $NB(k, \lambda)$ distribution

			Approximate Normal (M1)			Bootstrap Method		
N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	SYM	BH	LOG	M2	M3	M4
			$NB(3, 0.2)$					
100	116.0	116.0	88.3	90.8	89.0	95.3	95.8	93.1
250	286.8	287.0	87.3	87.1	86.5	94.6	94.5	93.5
500	567.0	565.0	86.1	83.3	83.5	90.9	90.9	89.7
1,000	1,122.9	1,107.5	86.3	81.0	81.9	91.5	91.4	90.9
5,000	5,421.4	5,394.5	79.1	73.3	75.2	80.6	79.9	79.4
10,000	10,735.7	10,706.5	74.9	70.1	71.5	71.9	71.7	71.1
			$NB(4, 0.4)$					
100	105.8	105.0	94.7	88.5	94.0	91.7	94.0	86.9
250	261.5	260.0	92.2	82.1	90.0	88.0	90.5	82.7
500	518.4	517.0	89.9	80.8	88.8	87.1	88.2	79.8
1,000	1,029.1	1,026.0	85.0	78.5	84.2	81.8	83.0	75.1
5,000	5,137.6	5,135.0	44.8	39.0	44.2	38.5	39.3	30.3
10,000	10,269.5	10,272.0	18.3	15.8	18.2	11.7	12.4	8.4
			$NB(4, 0.6)$					
100	101.7	101.0	96.3	80.1	96.4	90.7	96.1	66.5
250	253.3	253.0	93.2	77.2	95.0	83.1	92.6	66.0
500	505.7	505.0	92.5	77.3	94.3	83.5	91.3	64.5
1,000	1,010.3	1,010.0	82.4	70.6	85.2	75.4	82.0	57.2
5,000	5,047.8	5,048.0	31.6	25.8	33.6	26.1	28.9	14.5
10,000	10,093.2	10,093.0	8.7	6.9	9.1	5.5	6.3	1.2
			$NB(4, 0.8)$					
100	100.6	100.0	98.4	96.9	98.8	99.0	99.1	34.8
250	250.8	251.0	97.8	90.7	97.8	97.4	98.2	38.2
500	501.0	501.0	96.9	85.8	97.4	93.5	96.8	41.0
1,000	1,001.7	1,002.0	97.4	78.2	98.5	86.8	97.6	45.8
5,000	5,007.1	5,007.0	65.4	46.9	77.0	54.2	72.8	28.6
10,000	10,014.0	10,014.0	41.2	27.6	47.3	32.3	44.3	11.8

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

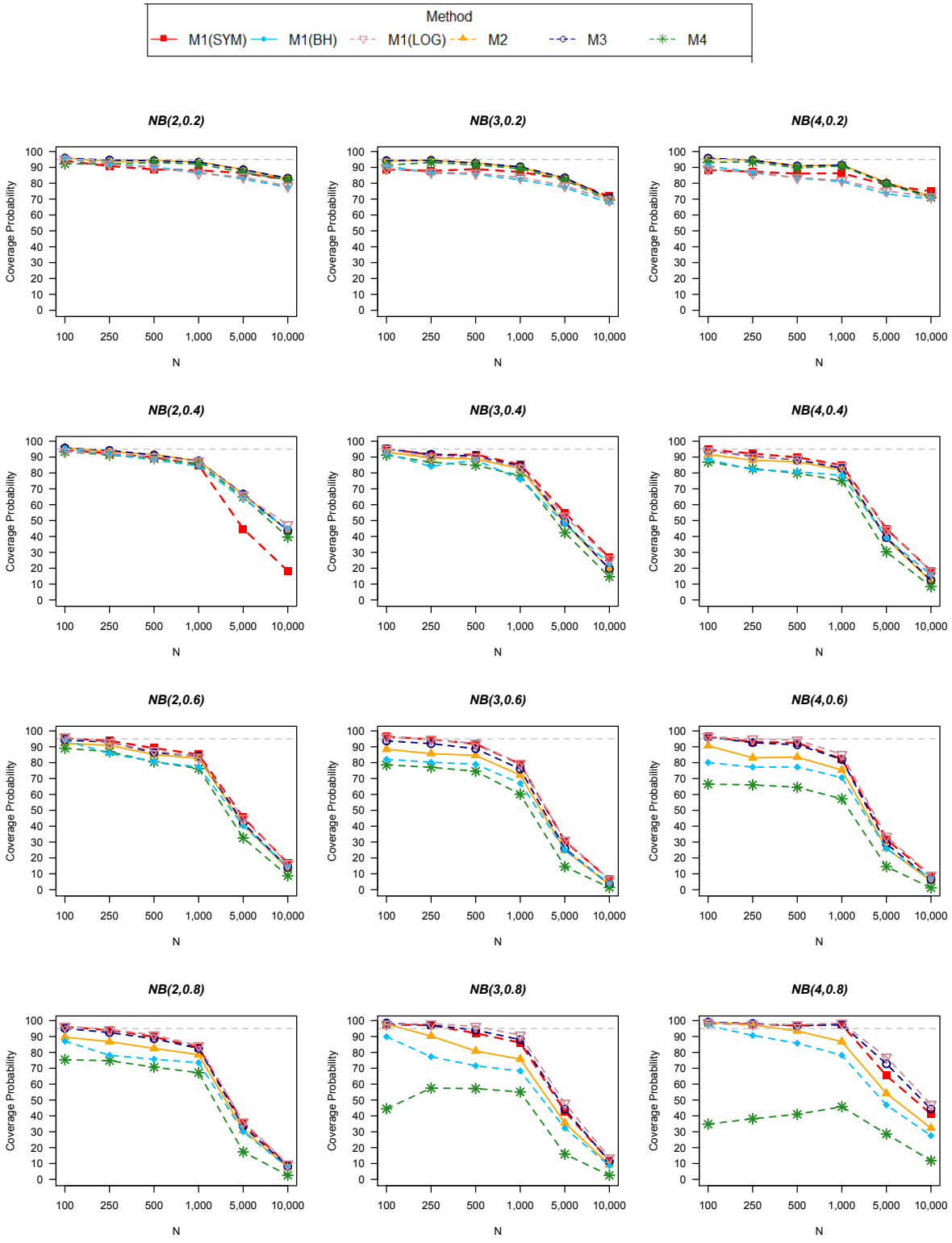

Figure 5.14: Coverage probabilities of 95% confidence interval when data is generated from the negative binomial distribution; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap and M4: Reduced bootstrap

5.3 Real data examples

For the application of the variance estimation approach for the LCMP estimator, the real data sets presented in Chapter 4 are reconsidered. These datasets were: the cholera data, the golf tees data, the heroin drug users in Bangkok data and the artificial Link-3 data. Additionally, The snowshoes hare data in which under-dispersion is also considered. The last example is the taxicab data to compare the variances approaches for the medium population size. We provide the results regarding the standard error and confidence intervals at the 95% level constructed from five methods. Furthermore, we use the length of confidence intervals to compare the capacity of confidence interval from variance estimation methods.

5.3.1 Cholera epidemiology in India data

The first application is to demonstrate how the LCMP estimator works based on the zero-truncated Poisson distribution. A cholera epidemic affected a village in India. The count of cholera cases per household is taken as an example of a zero-truncated Poisson distribution with $\widehat{\lambda} = 0.98$. A summary of data analysis is provided in Table 5.17. Since the true population size is unknown, we can not use the true bootstrap to construct the standard error and the confidence interval for this case.

As can be seen, the variance estimation based on the normal approximation gives the smallest of standard error estimation (7.59). This leads to the three confidence intervals with shorter lengths when compared to the other two computational methods. The imputed bootstrap provides a negligible difference of standard error, confidence interval and length of confidence interval compared to the reduced bootstrap method. The imputed bootstrap seem to be the most appropriate for estimating standard error and confidence interval. However, it is not significantly different to the estimates produced by the normal approximation approach.

In the previous Chapter 4, it was shown that the total number of cholera cases was $89(67 - 112)$ using the MLEPoi estimator and $91(67 - 115)$ by McKedrick's estimator (Viwatwongkasem et al., 2008). The confidence intervals associated with these estimates do not differ too much from the imputed and reduced bootstrap approaches.

Table 5.17: Comparison of four methods of variance estimation for cholera epidermic in India data

Method	$(\widehat{N})_{Mean}$	$(\widehat{N})_{Med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	87	-	7.59	$(73-102)^{(a)}$	29
				$(73-103)^{(b)}$	30
				$(74-104)^{(c)}$	30
M 2	-	-	-	-	-
M 3	87	86	11.90	$(67-114)$	47
M 4	87	86	11.89	$(67-113)$	46

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.3.2 Golf tees data

The golf tees data were also analysed in the previous chapter and were found to follow a of zero-truncated geometric distribution with $\widehat{\lambda} = 0.77$ or correspondingly a spacial case of zero-truncated Conway-Maxwell-Poisson distribution with $\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$. The true population size is provided as 250. Estimates of the standard errors and confidence intervals are provided in Table 5.18, the standard error of the normal approximation is the largest (33.09), causing all of three confidence intervals constructed based on this approach to be very wide, but all including the true N. The true bootstrap and the imputed bootstrap results is identical, while the reduced bootstrap shows the worst performance since it's confidence interval does not cover the true N.

To compare the analysis with the previous study, Niwitpong et al. (2013) proposed the Chao estimator based on the geometric distribution and gave the number of golf tees to be 230 with the 95% of CI (207 – 253). Also, a regression estimator based on the beta-binomial distribution from Rocchetti et al. (2014) estimated a lower number of golf tees, that is 216 with the confidence intervals (193 – 238) from an asymptotic approximation and (188 – 247) from the nonparametric approach. It can be seen that both of confidence intervals do not cover the true N.

Under the assumption of the geometric distribution, the reliability of inference might be supported by looking at the simulation study when $\widehat{\lambda} = 0.8$ and $N = 250$. In this situation, the true bootstrap and the imputed bootstrap perform the best with the highest accuracy of standard errors and good coverage of confidence intervals. On the other hand, the confidence intervals from the variance estimation based on the normal approximation also shows the high value of coverage probability, as a result of the overestimation of the standard error approximation.

The reduced bootstrap can not be satisfying, given the underestimation of standard error and low coverage probability with respect to the nominal level of 95%. For the approximated number of golf tees by the LCMP estimator, it would therefore be suggested that

the standard error and confidence interval should be calculated using the true bootstrap or the imputed bootstrap method. In short, it might be said that the LCMP estimator provides an estimated number of golf tees with 95% of CI is 223 (195–252).

Table 5.18: Comparison of four methods of variance estimation for golf tees data

Method	$(\widehat{N})_{Mean}$	$(\widehat{N})_{Med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	223	-	33.09	$(159-288)^{(a)}$	129
				$(168-298)^{(b)}$	130
				$(169-301)^{(c)}$	132
M 2	223	223	15.11	$(195-252)$	57
M 3	223	222	14.41	$(195-252)$	57
M 4	222	222	11.16	$(202-246)$	44

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.3.3 Heroin users in Bangkok data

The heroin drug users in Bangkok data are taken as a real situation representing the case of the zero truncate Conway-Maxwell-Poisson distribution, which reduces to the zero-truncated geometric distribution when the dispersion parameter equal to zero and $0 < \lambda < 1$ in Chapter 4. It was proved by using the log-ratio plot in the previous chapter that this data set follows the zero-truncated Conway-Maxwell-Poisson distribution with $\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$. It was also shown that the estimated number of drug users is 12,141. From the simulation result, we expect that the imputed bootstrap will be the best choice for estimating standard error and confidence intervals.

It is not surprising that the symmetric normal approximation method produces the highest standard error, leading to the widest confidence interval (see Table 5.19). As the estimated population size is quite large and the value of λ is high, it can be assumed that the estimated population size is normally distributed. As a consequence, all of the confidence intervals constructed under the normal estimation are not different from each other. On the other hand, the two computational approaches show standard errors significantly lower than analytic approaches resulting in narrower confidence interval.

To give a comparison the population size estimation from the LCMP estimator is similar to the result obtained by the Zelterman estimator in Anan et al. (2016) which estimated the population size to be $12,077$ with an analytic confidence interval (11,715 - 12,439). Lanumteang and Böhning (2011) proposed a new estimator by extending the Chao estimator under the negative binomial distribution. Using this method the number of drug users were reported as $11,714$ with a standard error of bootstrap, $\widehat{S.E.}(\widehat{N}_{Zel}) = 265.07$ and a bootstrap percentile (11,256 - 12,279). It can be seen that this method gave a lower population size, higher standard error and wider confidence interval than

the LCMP estimation. The estimate obtained through the mixture of truncated count by Viwatwongkasem et al. (2008) is 11,130 with an normal approximation confidence interval (11,014 - 11,246). This case provided a smaller population size and a lower confidence interval than the LCMP estimator.

Looking back to the simulation study with data generated following the geometric distribution with $\lambda = 0.8$ and $N = 5,000$. It can be seen that the reduced bootstrap has a underestimation of standard error and provides the coverage probability lower than the nominal level setting of 95%. Then, the imputed bootstrap seems to be the best choice to estimate standard error and construct the confidence interval for this data set. Therefore, it might be suitable to report that by using the LCMP estimator, the number of heroin drug users in Bangkok, Thailand in 2002 is estimated 12,141, with a 95% of the imputed bootstrap confidence interval between 11,918 and 12,320.

Table 5.19: Comparison of four methods of variance estimation for heroin users in Bangkok data

Method	$(\widehat{N})_{Mean}$	$(\widehat{N})_{Med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	12,141	-	210.2371	$(11,729 - 12,554)^{(a)}$	825
		-		$(11,736 - 12,560)^{(b)}$	824
		-		$(11,738 - 12,562)^{(c)}$	824
M 2	-	-	-	-	-
M 3	12,132	12,136	101.31	(11,918 – 12,320)	402
M 4	12,131	12,138	81.93	(11,944 – 12,277)	333

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.3.4 Link-3 data

In the previous chapter it was shown using the Link 3 data that the estimated population size was $3,333$ based on the zero-truncated Conway-Maxwell-Poisson distribution with $\widehat{\lambda} = 0.82$ and $\widehat{\nu} = 0.09$. The estimated population size is relatively large so it is no surprise that all analytic approximations provide the same confidence interval ranges whereas the two computational approaches give higher standard errors and widths (see Table 5.20).

Table 5.20: Comparison of four methods of variance estimation for Link data

Method	$(\widehat{N})_{mean}$	$(\widehat{N})_{med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	3,333	-	45.21	$(3,245-3,422)^{(a)}$	177
				$(3,246-3,423)^{(b)}$	177
				$(3,246-3,423)^{(c)}$	177
M 2	-	-	-		-
M 3	3,341	3,338	80.69	(3,189 – 3,506)	317
M 4	3,340	3,337	72.98	(3,204 – 3,492)	288

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.3.5 Snowshoe hare data

The snowshoe hare data have been analysed in Cormack (1989) and Agresti (1994). From a graphical inspection by means of the ratio plot in Chapter 4, if we remove the two hares caught six times and treat them as outliers following Cormack (1989), the under-dispersion is estimated ($\widehat{\lambda} = 2.16$; $\widehat{\nu} = 1.25$), resulting in $\widehat{N} = 78$. The population size estimation from the LCMP is close to the new mixture model estimator of 79 proposed by Morgan and Ridout (2008). However, the LCMP estimator remains flexible and can be used for the under-dispersion case.

Leaving two hares out of the two hares left out of the analysis is considered non-representative of the unobserved part of the population. Bootstrap intervals, which do not require the exclusion of the two hares, are larger than those obtained by normal approximation in line with the simulation result (see Figure 5.10). The Burnham and the log-transformed-based intervals are similar to the symmetric approach, confirming that, there is under-dispersion, a symmetric confidence interval may lead to unreliable inference. Nevertheless, the percentiles bootstrap approaches provided the widest length of confidence intervals. In the perspective of standard error and confidence interval, the imputed bootstrap seems to provide a potential and realistic method in the light of small population size and under-dispersion.

Table 5.21: Comparison of four methods of variance estimation for the snowshoes hare (reduce) data

Method	$(\widehat{N})_{mean}$	$(\widehat{N})_{med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	78	-	4.58	$(70-87)^{(a)}$	17
				$(70-88)^{(b)}$	18
				$(70-88)^{(c)}$	18
M 2	-	-	-	-	-
M 3	84	80	14.09	(66 – 121)	55
M 4	84	79	13.50	(69 – 121)	52

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.3.6 Taxicabs in Edinburgh data

Carothers (1973) reported that 420 taxicabs are registered in Edinburgh, Scotland during his mark-recapture study. This closed population was sampled for ten consecutive days with observation points and times varied across the study period. Sighting a cab was considered a capture. No taxis were observed on more than six occasions. The data are shown in Table 5.22. As can be seen the fitted frequencies ($\widehat{f}_x$) of the zero-truncated Conway-Maxwell-Poisson distribution are not much more different from the observed frequencies. This is supported by Chi-square goodness of fit test with $\chi^2 = 7.81$, $df = 3$, leading to a $p - value = 0.05011$. The number of taxicabs can be assumed following the CMP distribution, therefore, the LCMP estimator given population of taxicabs to be 428 with $\widehat{\lambda} = 1.32$ and $\widehat{\nu} = 0$.

Table 5.22: Frequency distribution of Taxicabs data in Edinburgh

x	0	1	2	3	4	5	6	χ^2
f_x	137	142	81	49	7	3	1	
$\widehat{f}_x$ (ZTCMP)	-	139	84	39	15	5	1	7.81

To compare the performance of approximated standard errors and confidence interval methods, we look at the results in Table 5.23. The confidence intervals under the variance estimation based on the normal approximation methods display larger length than the bootstrap methods, yet all the CIs cover the true population size. The true bootstrap and imputed bootstraps are identical, providing similar standard errors and confidence intervals. Although the reduced bootstrap provides the smallest length, this is not a good choice for estimating standard error and confidence interval. All simulation results indicated that the reduced bootstrap has an underestimated standard error for $\lambda > 1$ and $\nu \to 0$.

Table 5.23: Comparison of four methods of variance estimation for the taxicabs data in Edinburgh data

Method	$(\widehat{N})_{mean}$	$(\widehat{N})_{med}$	$\widehat{S.E.}(\widehat{N})$	$95\%CI$	Length
M 1	428	-	91.28	$(250 - 607)^{(a)}$	357
		-		$(284 - 648)^{(b)}$	364
		-		$(284 - 662)^{(c)}$	372
M 2	449	438	65.75	(348 – 600)	252
M 3	449	438	65.85	(348 – 600)	252
M 4	449	438	64.12	(353 – 597)	244

(a) : Symmetric confidence interval
(b) : Burnham confidence interval
(c) : Logarithm transformation confidence interval

5.4 Conclusion

The desirable properties for a good population size estimator should be small biased and high precision. The LCMP estimator is mentioned as an asymptotically unbiased estimator with respect to the population size under the Conway-Maxwell-Poisson distribution which includes the Poisson and geometric models as nested models. This chapter describes and compares methods for variance estimations for the LCMP estimator, including a normal approximation approach and re-sampling techniques. The three bootstrap resample approaches: true bootstrap, imputed bootstrap and reduced bootstrap, are examined as the robust methods in comparison to the variance estimation based on the normal approximation, particular for the small population size.

The true bootstrap approach exhibits the best performance for estimating the variance of the LCMP estimator on average, it often provides the smallest bias compared with the true variance. However, this approach is often not useful in practice since the the population size N usually needs to be estimated for capture-recapture data. The imputed bootstrap seems to be the best choice to estimate variance of the LCMP estimator, it works similarly to the true bootstrap, but it requires a valid model. The variance estimation based on the normal approximation approach is useful for medium or large population sizes or/and the event rate parameter increase due to the normal approximation property. It is recommended that the reduced bootstrap is not appropriate for estimating the variance of the LCMP estimator in any situation.

The six methods are investigating to construct confidence intervals for the LCMP estimator. This is supported by the simulation results that the true bootstrap performs the best for constructing confidence interval of the LCMP estimator, but it is limited in reality. Alternatively, the imputed bootstrap works similarly to the true bootstrap under the true model. Therefore it might be the most suitable method for constructing the confidence interval in capture-recapture study since target population sizes usually are not provided . The imputed bootstrap gives a coverage probability close to the nominal level at 95%. Moreover, it is very clear that the reduced bootstrap often underestimates the true variance and has a very poor confidence interval which is consistent with Buckland and Garthwaite (1991) and Zwane and Van der Heijden (2003).

Chapter 6

Population Size Estimation based on the Geometric Distribution

As the geometric distribution is a special case of the Conway-Maxwell-Poisson distribution, a framework based on the geometric distribution is specified in this chapter. The traditional Turing estimator and Zelterman estimator are developed based on the geometric distribution. Additionally, the normal approximation variance formulas as well as confidence intervals are derived for these new estimators. Furthermore, the performance of the proposed estimators are compared with alternative estimators in a simulation study. The results show that the Turing estimator based on the geometric distribution can improve the efficiency of estimation. The extension of Zelterman estimator provides some bias, however, it is useful for estimating a large population size for capture-recapture data under the geometric model. Three empirical data sets: the wood mice data, the golf tees data and the heroin users in Bangkok data are used as case studies for the geometric distribution to illustrate the performance of these estimators.

6.1 Introduction

The capture-recapture approach aims to estimate elusive target population sizes. As we mentioned before that the basic model for capture-recapture data is the Poisson distribution but as it rarely occurs in reality, the heterogeneity problem becomes a very important issue for statistical modelling as well as for capture-recapture study. If the heterogeneous population is unobserved, the target population size will be underestimate (Böhning and Schön, 2005; Huggins and Hwang, 2011). To account for the heterogeneity for estimating population size, the Poisson mixture has been applied to heterogeneous capture-recapture data. For example, the exponential Poisson mixture leads to the geometric distribution (see Lanumteang and Böhning, 2011; Niwitpong et al., 2013; Vidal-Diez, 2015). Let the count of identifications be $X_i \in \{0, 1, 2, 3, ...\}$ for

$i = 1, 2, 3, ..., N$. The geometric distribution arises as a mixture of Poisson distributions with an exponential density. The mixture model is given by

$$g_x = \int_0^\infty p(x;\lambda)h(\lambda;\theta)d\lambda, \tag{6.1}$$

where the mixture kernel is the Poisson distribution $p(x;\lambda) = \dfrac{\exp(-\lambda)\lambda^x}{x!}$, and the mixing density comes from the exponential density $h(\lambda;\theta) = \dfrac{1}{\theta}\exp(-\dfrac{\lambda}{\theta})$. Then, the associated marginal density is obtained as:

$$g_x(p) = (1-p)^x p, \tag{6.2}$$

where $p = \dfrac{1}{1+\theta} \in (0,1)$ is the event parameter, and $x = 0, 1, 2, ...$ Mean and variance are $E(X) = \dfrac{1-p}{p}$ and $Var(X) = \dfrac{1-p}{p^2}$, respectively. One of the interesting properties of geometric distribution is the ratio of neighbouring probabilities which is defined as

$$r_x = \frac{g_{x+1}}{g_x} = \frac{(1-p)^{x+1}p}{(1-p)^x p} = 1-p. \tag{6.3}$$

As can be seen that (6.3) is a constant depending on p. It is then straightforward to estimate the probabilities by the their relative frequencies and use $r_x^* = \dfrac{f_{x+1}}{f_x}$ as a graphical tool for investigating the geometric distribution (see Niwitpong et al., 2013; Böhning et al., 2013a; Böhning, 2015).

Example 1: Wood mice data

To investigate a appropriate parametric model for zero-truncated count data, a wood mice data example which was analysed by Morgan and Ridout (2008) is utilised. The data are summarised in Table 6.1.

Table 6.1: The observed frequency distribution of the wood mice data

x	1	2	3	4	5	6	7	8	9	10
f_x	71	59	41	39	20	26	19	12	9	5
x	11	12	13	14	15	16	17	18		
f_x	8	4	9	2	1	3	3	3		

The two ratio plots which represent the Poisson ratios; $r_x^* = (x+1)\dfrac{f_{x+1}}{f_x}$, and the geometric ratios $r_x^* = \dfrac{f_{x+1}}{f_x}$ are used as a diagnostic tool and as shown in Figure 6.1. Clearly, the ratio plot of the successive neighbouring geometric is close to a horizontal line pattern with a negligible value of residual heterogeneity. Indeed, it also can provide further evidence for the geometric model by fitting a simple linear regression and testing the relationship between the count x and the ratio of successive frequencies r_x^*. Therefore, a test for slope equal to zero is used, and in this case gives a $p - value = 0.26$, $df = 15$.

From the ratio plot we conclude that it is reasonable to assume that the data are generated under the geometric distribution.

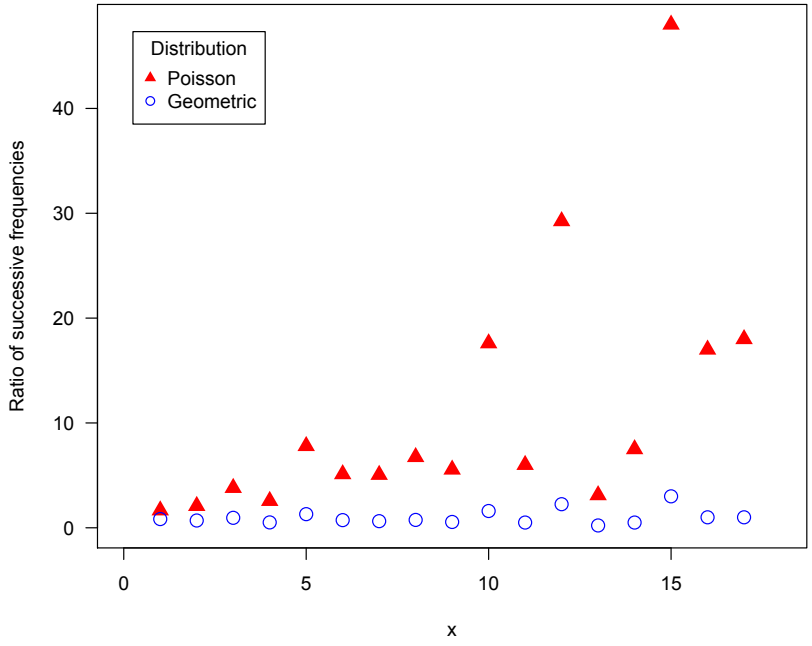

Figure 6.1: Poisson ratio plot (red) and geometric ratio plot (blue) for the wood mice data

6.2 Population sizes estimators based on the geometric distribution

In the Chapter 4 and Chapter 5, a performance of the new proposed estimator LCMP for the specific case of the geometric model is compared with the two generalised heterogeneity estimators of Chao and Zelterman. Indeed, the Chao estimator based on the mixture geometric (CG) approach as well as the maximum likelihood estimation for the geometric distribution (MLEGeo) were provided in Niwitpong et al. (2013). For this section we are interested in the Zelterman estimator which is known as a robust estimator using only count of ones and twos, but the traditional Zelterman estimator was constructed based on the zero-truncated Poisson distribution. For the fairness of comparison, its original formulation is extended to the zero-truncated geometric distribution. Additionally, the Turing estimator for the zero-truncated geometric model is modified in this section.

6.2.1 The extension of Zelterman's estimation based on the zero-truncated geometric distribution

The traditional Zelterman estimator (Zel) was derived under the zero-truncated Poisson distribution, allowing the Poisson to be contaminated. It is a robust estimator and performs well for several zero-truncated count distributions (Navaratna et al., 2008). The Zelteman estimator is given as $\widehat{N} = \dfrac{n}{1-\widehat{p}_0(\lambda)}$, where $\widehat{p}_0(\lambda)$ is the estimated probability of an individual being unobserved distributed under the Poisson distribution with parameter λ. As a consequence, this idea is extended to a new population size estimator called the extended Zelterman's estimator based on the geometric distribution (ZG). The moment method is used to construct Zelterman estimator based on the geometric distribution as follows.

Start by considering the zero-truncated geometric distribution (g_x^+) with parameter p:

$$g_x^+ = \frac{g_x}{1-g_0} = \frac{(1-p)^x p}{1-p} = (1-p)^{x-1}p, \tag{6.4}$$

where $g_0 = p$. Additionally, the ratio of the zero-truncated geometric distribution is simple to calculate as

$$\frac{g_{x+1}}{g_x} = \frac{(1-p)^{x+1}p/(1-p)}{(1-p)^x p/(1-p)} = 1-p. \tag{6.5}$$

As we know that the probabilities can be estimated by relative frequencies:

$$\frac{E(f_{x+1}/N)}{E(f_x/N)} = 1-p. \tag{6.6}$$

Hence, it can be rewritten as

$$\frac{f_{x+1}}{f_x} = 1-\widehat{p} = 1-\widehat{g}_0. \tag{6.7}$$

Parameter g_0 can be estimated as $1-\dfrac{f_{x+1}}{f_x}$. For similar reasons as in the construction of the original Zelterman estimator as pointed out in Chapter 2, we choose the first and second frequencies are chosen or x is set as $x=1$ for estimating g_0. Therefore, $\widehat{g}_0 = 1-\dfrac{f_2}{f_1}$. The extension of Zelterman estimator based on the geometric estimator (ZG) is given as

$$\widehat{N}_{ZG} = \frac{n}{1-\widehat{g}_0} = \frac{n}{1-(1-\frac{f_2}{f_1})} = \frac{n}{\frac{f_2}{f_1}} = \frac{nf_1}{f_2}, \tag{6.8}$$

where $n = \sum_{x=1}^{m} f_x$ is the total number of observed counts.

Theorem 6.1. *The ZG estimator is the asymptotically unbiased under the geometric distribution with parameter p:* $g_x = (1-p)^x p$ *for* $x = 0, 1, 2, ...$ *That is*

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{ZG})}{N} \to 1.$$

Proof. For the ZG estimator, we set $x = 1$, then we have that $E\left(\frac{f_2/N}{f_1/N}\right)$ converges as $N \to \infty$ to $\frac{(1-p)^2 p}{(1-p)p} = (1-p)$, and $E(\frac{n}{N})$ converges to $1 - g_0 = 1 - p$. Consequently,

$$\begin{aligned} E\left(\frac{\widehat{N}_{ZG}}{N}\right) &= E\left\{\frac{\frac{n}{1-(1-\frac{f_2}{f_1})}}{N}\right\} = E\left\{\left(\frac{n}{N}\right)\left(\frac{1}{1-(1-\frac{f_2}{f_1})}\right)\right\} \\ &\overset{N\to\infty}{\to} (1-p)\frac{1}{1-[1-(1-p)]} = (1-p)\frac{1}{(1-p)} \\ &= 1. \end{aligned} \tag{6.9}$$

□

6.2.2 Variance of the ZG estimator

A variance estimator for the ZG estimator is now constructed. The use of the conditional technique together with the delta method is utilised for deriving the variance formula for the population size estimator. Variance of the ZG estimator can be produced under the conditional technique as follows:

$$Var(\widehat{N}_{ZG}) = Var_n\{E(\widehat{N}|n)\} + E_n\{Var(\widehat{N}|n)\}, \tag{6.10}$$

where Var_n and E_n are estimated from the distribution of n. As the approximation $E(\widehat{N}|n)$ can be justified by the delta-method method, the expected value of the transformed random variable can be estimated by transformation of the the expected value (see Bishop et al., 2007; Böhning, 2008a), then $E(\widehat{N}|n) \approx \frac{n}{1-g_0}$, and so it follows that

$$Var_n\{E(\widehat{N}|n)\} \approx Var_n\left\{\frac{n}{1-g_0}\right\} = \frac{1}{(1-g_0)^2} Var(n) = \frac{N g_0(1-g_0)}{(1-g_0)^2}. \tag{6.11}$$

Since $E(n) = N(1-g_0)$, the variance given in (6.11) can be estimated as

$$\widehat{Var}_n\{E(\widehat{N}|n)\} = \frac{n g_0}{(1-g_0)^2}. \tag{6.12}$$

Then $\widehat{g}_0 = 1 - \frac{f_2}{f_1}$, so that

$$\widehat{Var}_n\{E(\widehat{N}|n)\} = \frac{n(1-\frac{f_2}{f_1})}{(1-(1-\frac{f_2}{f_1}))^2} = \frac{\frac{n}{f_1}(f_1-f_2)}{\frac{f_2^2}{f_1^2}} = \frac{n(f_1^2-f_1f_2)}{f_2^2}. \tag{6.13}$$

For the second term in (6.10), starting from a basic property of the variance, $Var(\widehat{N}|n)$ is determined as

$$Var(\widehat{N}|n) = Var(\frac{nf_1}{f_2}|n) \approx n^2 Var(\frac{f_1}{f_2}). \tag{6.14}$$

Then, applying the bivariate delta method to obtain $Var(\frac{f_1}{f_2})$, the approximation is obtained as

$$Var(\frac{f_1}{f_2}) \approx \bigtriangledown h(f_1,f_2)^T \mathbf{cov}(f_1,f_2) \bigtriangledown h(f_1,f_2), \tag{6.15}$$

where $h(f_1,f_2) = \frac{f_1}{f_2}$ and $\bigtriangledown h(f_1,f_2)^T = (\frac{\partial}{\partial f_1}\frac{f_1}{f_2}, \frac{\partial}{\partial f_2}\frac{f_1}{f_2}) = (\frac{1}{f_2}, -\frac{f_1}{f_2^2})$. In addition, the covariance matrix of the multinomial distribution, conditional on n for frequencies one and two, provided the covariance matrix as

$$\widehat{\mathbf{cov}}(f_1,f_2) = \begin{pmatrix} var(f_1) & cov(f_1,f_2) \\ cov(f_1,f_2) & var(f_2) \end{pmatrix} = \begin{pmatrix} f_1(1-\frac{f_1}{n}) & -\frac{f_1f_2}{n} \\ -\frac{f_1f_2}{n} & f_2(1-\frac{f_2}{n}) \end{pmatrix}. \tag{6.16}$$

Hence,

$$\begin{aligned} \bigtriangledown h(f_1,f_2)^T \widehat{\mathbf{cov}}(f_1,f_2) &= \begin{pmatrix} \frac{1}{f_2} & -\frac{f_1}{f_2^2} \end{pmatrix} \begin{pmatrix} f_1(1-\frac{f_1}{n}) & -\frac{f_1f_2}{n} \\ -\frac{f_1f_2}{n} & f_2(1-\frac{f_2}{n}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{f_1}{f_2}(1-\frac{f_1}{n}) + \frac{f_1^2f_2}{nf_2^2} & -\frac{f_1f_2}{nf_2} - \frac{f_1f_2}{f_2^2}(1-\frac{f_2}{n}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{f_1}{f_2}(1-\frac{f_1}{n}) + \frac{f_1^2}{nf_2} & -\frac{f_1}{n} - \frac{f_1}{f_2}(1-\frac{f_2}{n}) \end{pmatrix}. \end{aligned} \tag{6.17}$$

As a consequence,

$$\begin{aligned}
\bigtriangledown h(f_1,f_2)^T \widehat{\mathbf{cov}}(f_1,f_2) \bigtriangledown h(f_1,f_2) &= \left(\frac{f_1}{f_2}(1-\frac{f_1}{n})+\frac{f_1^2}{nf_2} \quad -\frac{f_1}{n}-\frac{f_1}{f_2}(1-\frac{f_2}{n})\right)\begin{pmatrix}\frac{1}{f_2}\\ -\frac{f_1}{f_2^2}\end{pmatrix}\\
&= \frac{f_1}{f_2^2}(1-\frac{f_1}{n})+\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{f_2^3}(1-\frac{f_2}{n})\\
&= \frac{f_1}{f_2^2}-\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{f_2^3}-\frac{f_1^2f_2}{nf_2^3}\\
&= \frac{f_1}{f_2^2}+\frac{f_1^2}{nf_2^2}+\frac{f_1^2}{f_2^3}-\frac{f_1^2}{nf_2^2}\\
&= \frac{f_1}{f_2^2}+\frac{f_1^2}{f_2^3}. \qquad (6.18)
\end{aligned}$$

Finally, the variance estimation of the ZG estimator is given as

$$\widehat{Var}(\widehat{N}_{ZG}) = \frac{nf_1(f_1-f_2)}{f_2^2}+n^2\left(\frac{f_1}{f_2^2}+\frac{f_1^2}{f_2^3}\right). \qquad (6.19)$$

6.2.3 An extension of Turing's estimator based on the geometric distribution

The generalised Turing estimator was introduced by Böhning et al. (2013a) based upon the negative binomial distribution. It is adapted here to provide a Turing estimator based on the geometric distribution (TG). Let $X_1, X_2, X_3, ..., X_N$ have a marginal density following the geometric distribution with parameter p, which can be rewritten as $X \sim Geo(p)$. Then, $g_0 = p$, $g_1 = (1-p)p$ and $E(X) = \dfrac{1-p}{p}$, giving

$$g_0 = p = \sqrt{p^2} = \sqrt{\frac{(1-p)p^2}{(1-p)}} = \sqrt{\frac{p(1-p)}{(1-p)/p}} = \sqrt{\frac{g_1}{E(X)}}. \qquad (6.20)$$

In practice, the capture probabilities can be estimated by the relative frequencies so that the unobserved probability $\widehat{g}_0$ is achieved by

$$\widehat{g}_0 = \sqrt{\frac{f_1/N}{S/N}} = \sqrt{\frac{f_1}{S}}, \qquad (6.21)$$

when $S = \sum_{x=0}^{m} xf_x = \sum_{x=1}^{m} xf_x$. Hence, the extension of Turing's estimator for the geometric distribution (TG) is given as

$$\widehat{N}_{TG} = \frac{n}{1-\sqrt{\frac{f_1}{S}}}. \qquad (6.22)$$

The advantages of the TG estimator are that it uses all the information available from the observed data (S), and so it seems to be more natural than the ZG estimator. Another positive point is that the TG estimator is a straightforward approach getting the estimated population size, and it is simple to calculate.

Theorem 6.2. *The TG estimator is the asymptotically unbiased estimator under the geometric distribution;* $g_x = (1-p)^x p$ *for* $x = 0, 1, 2, ...$

Then, we achieve that

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{TG})}{N} \to 1.$$

Proof. We have that $E(X) = E(S/N) = (1-p)/p$, $E(f_1) = Ng_1 = Np(1-p)$ so that $\sqrt{\frac{E(f_1/N)}{E(S/N)}} = \sqrt{\frac{p(1-p)}{(1-p)/p}} = p = g_0$ and $E(n/N) = (1-g_0) = (1-p)$. Therefore, for sufficiently large N,

$$\begin{aligned} E\left(\frac{\widehat{N}_{TG}}{N}\right) &= E\left(\frac{\frac{n}{1-\sqrt{f_1/S}}}{N}\right) = E\left(\frac{n}{N}\frac{1}{1-\sqrt{f_1/S}}\right) \\ &\overset{N\to\infty}{\to} (1-p)\frac{1}{1-p} \\ &= 1. \end{aligned} \tag{6.23}$$

□

Theorem 6.3. *The TG estimator is larger than the original Turing estimator:*

$$\widehat{N}_{TG} > \widehat{N}_{Turing}.$$

Proof. The estimated probability of zero counts for the Turing estimator is $\widehat{p}_{0(Turing)} = \frac{f_1}{S}$ whereas the probability of zero count of the TG estimator can be estimated as $\widehat{p}_{0(TG)} = \sqrt{\frac{f_1}{S}}$. As we know that in capture-recapture studies $f_1 < S$ where $S = \sum_{x=1}^{m} x f_x$, therefore $\frac{f_1}{S} < \sqrt{\frac{f_1}{S}}$. Then, we have that

$$\widehat{N}_{TG} = \frac{1}{1-\sqrt{\frac{f_1}{S}}} > \frac{1}{1-\frac{f_1}{S}} = \widehat{N}_{Turing}.$$

□

6.2.4 Variance of the TG estimator

Let us recall the composition of variance. The first source of variance estimation comes from choosing the sample size n from the population size N so that the probability follows the binomial distribution with parameter $1-g_0$, and the second arises from estimating $\widehat{g}_0$:

$$Var(\widehat{N}) = Var_n\left\{E(\widehat{N}|n)\right\} + E_n\left\{Var(\widehat{N}|n)\right\}. \tag{6.24}$$

Starting with the first term in (6.24), the approximation $E(\widehat{N}|n)$ can be justified by the delta-method, so that

$$E(\widehat{N}|n) \approx \frac{n}{1-g_0},$$

and we have that

$$\begin{aligned} Var_n\left\{E(\widehat{N}|n)\right\} &\approx Var_n\left\{\frac{n}{1-g_0}\right\} \\ &= \frac{1}{(1-g_0)^2}Var(n) \\ &= \frac{N(1-g_0)g_0}{(1-g_0)^2}. \end{aligned} \tag{6.25}$$

Since $E(n) = N(1-g_0)$ and $\widehat{g}_{0(TG)} = \sqrt{\frac{f_1}{S}}$, the variance in (6.25) can be estimated as:

$$\widehat{Var}_n\left\{E(\widehat{N}|n)\right\} = \frac{n\sqrt{\frac{f_1}{S}}}{\left(1-\sqrt{\frac{f_1}{S}}\right)^2}. \tag{6.26}$$

Additionally, we assume that $E_n\left\{Var(\widehat{N}|n)\right\}$ can be estimate by $Var(\widehat{N}|n)$. We have that

$$Var(\widehat{N}|n) = Var\left(\frac{n}{1-\sqrt{\frac{f_1}{S}}}|n\right) = n^2 Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right). \tag{6.27}$$

We know that $Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right)$ can be computed by the delta-method. Let $y = \frac{f_1}{S}$ and we take $h(y) = \frac{1}{1-\sqrt{y}}$. Therefore,

$$h'(y) = -(1-y^{1/2})^{-2}\left(-\frac{1}{2}y^{-1/2}\right) = \frac{1}{2\sqrt{y}(1-\sqrt{y})^2}.$$

Then,

$$Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right) \approx \left(\frac{1}{2\sqrt{\frac{f_1}{S}}(1-\sqrt{\frac{f_1}{S}})^2}\right)^2 Var\left(\frac{f_1}{S}\right)$$
$$= \left(\frac{1}{4\frac{f_1}{S}(1-\sqrt{\frac{f_1}{S}})^4}\right) Var\left(\frac{f_1}{S}\right). \quad (6.28)$$

In the next step we use the variance conditional technique to estimate $Var\left(\frac{f_1}{S}\right)$, that is

$$Var\left(\frac{f_1}{S}\right) = Var_{f_1}\left\{E(\frac{f_1}{S})|f_1\right\} + E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\}. \quad (6.29)$$

Assuming that $E\left(\frac{f_1}{S}|f_1\right) = f_1E(\frac{1}{S}) \approx \frac{f_1}{S}$, then

$$\begin{aligned} Var_{f_1}\left\{E\left(\frac{f_1}{S}|f_1\right)\right\} &\approx Var_{f_1}\left(\frac{f_1}{S}\right) \\ &= \frac{1}{S^2}Var(f_1) \\ &= \frac{1}{S^2}Np_1(1-p_1) \\ &= \frac{1}{S^2}\left(N\frac{f_1}{N}(1-\frac{f_1}{N})\right) \\ &= \frac{f_1}{S^2}\left(1-\frac{f_1}{N}\right). \end{aligned} \quad (6.30)$$

Again, assuming that $E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\}$ can be estimated by $Var\left(\frac{f_1}{S}|f_1\right)$ so that

$$\begin{aligned} E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\} &\approx Var\left(\frac{f_1}{S}|f_1\right) \\ &= f_1^2Var\left(\frac{1}{S}\right). \end{aligned} \quad (6.31)$$

Using the delta method,

$$\begin{aligned} Var\left(\frac{1}{S}\right) &\approx \frac{1}{S^4}Var(S) \\ &= \frac{1}{S^4}Var(N\bar{X}) \\ &= \frac{1}{S^4}N^2Var(\bar{X}) \\ &= \frac{1}{S^4}N^2\frac{Var(X)}{N}. \end{aligned} \quad (6.32)$$

Since $X \sim Geo(p)$, it follows that $E(X) = \dfrac{1-p}{p}$ and $Var(X) = \dfrac{1-p}{p^2}$.

$$\begin{aligned} Var\left(\frac{1}{S}\right) &\approx \frac{1}{S^4} N^2 \frac{(\frac{1-p}{p^2})}{N} \\ &= \frac{1}{S^4} N^2 \frac{(\frac{E(X)}{p})}{N} \\ &= \frac{1}{S^4} N^2 \frac{(\frac{E(S/N)}{p})}{N} \\ &= \frac{1}{pS^3}. \end{aligned} \tag{6.33}$$

Note that

$$\begin{aligned} E(X) &= \frac{1-p}{p} \\ E(\frac{S}{N}) &= \frac{1-p}{p} \\ \frac{S}{N} &\approx \frac{1-p}{p} \\ Sp &\approx N - Np \\ p(S+N) &\approx N \\ p &\approx \frac{N}{S+N} \\ \frac{1}{p} &\approx \frac{S+N}{N} \end{aligned} \tag{6.34}$$

Hence,

$$\widehat{Var}\left(\frac{f_1}{S}|f_1\right) = \frac{f_1^2}{S^3}\left(\frac{S+N}{N}\right) \tag{6.35}$$

Substituting (6.30) and (6.34) into (6.29), this leads to

$$\begin{aligned} \widehat{Var}\left(\frac{f_1}{S}\right) &= \frac{1}{S^2}\left\{f_1\left(1-\frac{f_1}{N}\right)\right\} + \frac{f_1^2}{S^3}\left(\frac{S+N}{N}\right) \\ &= \frac{f_1}{S^2}\left\{\frac{N+f_1}{N} + \frac{f_1}{S}\left(\frac{S+N}{N}\right)\right\} \\ &= \frac{f_1}{S^2}\left\{\frac{NS - Sf_1 + f_1S + f_1N}{NS}\right\} \\ &= \frac{f_1}{S^2}\left\{\frac{N(S+f_1)}{NS}\right\} \\ &= \frac{f_1S + f_1^2}{S^3}. \end{aligned} \tag{6.36}$$

We have that

$$\begin{aligned}
\widehat{Var}\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right) &= \left\{\frac{1}{\frac{4f_1}{S}\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}\left\{\frac{f_1S+f_1^2}{S^3}\right\} \\
&= \widehat{Var}\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right) = \left\{\frac{S}{4f_1\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}\left\{\frac{f_1S+f_1^2}{S^3}\right\} \\
&= \frac{Sf_1+f_1^2}{4f_1S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4} \\
&= \frac{S+f_1}{4S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4}. \qquad (6.37)
\end{aligned}$$

Finally, the variance of TG estimator is given as

$$\widehat{Var}(\widehat{N}_{TG}) = \frac{n\sqrt{\frac{f_1}{S}}}{\left(1-\sqrt{\frac{f_1}{S}}\right)^2} + n^2\left\{\frac{S+f_1}{4S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}. \qquad (6.38)$$

6.3 Alternative estimators based on the geometric distribution

Three alternative estimators based on the geometric distribution are given in this section. There are two estimators proposed by Niwitpong et al. (2013), the maximum likelihood estimator base on the geometric distribution and Chao estimator for the geometric mixture. Another estimator is the LCMP estimator which is the novel estimator introduced in Chapter 4 and Chapter 5.

6.3.1 Maximum likelihood estimator based on the geometric distribution

The maximum likelihood estimator based on the zero-truncated geometric distribution (MLEGeo) was proposed by Niwitpong et al. (2013) as

$$\widehat{N}_{MLEGeo} = \frac{n}{1-\widehat{p}_0} = \frac{n}{1-\frac{n}{S}} = \frac{nS}{S-n}, \qquad (6.39)$$

where $\widehat{p}_0$ is estimated by using the maximum likelihood approach based on the zero-truncated geometric distribution and $S = \sum_{x=1}^{m} xfx$. The variance estimation of the MLEGeo is given as

$$\widehat{Var}(\widehat{N}_{MLEGeo}) = \frac{s^2 n^2}{(S-n)^3}. \tag{6.40}$$

An advantage of MLEGeo estimator is that it often provides a small variance for the true model. In contrast, it will underestimate the variance where there is evidence of a contaminated geometric distribution.

6.3.2 Linear regression estimation based on the Conway-Maxwell-Poisson distribution

A population size estimator based on the zero-truncated Conway-Maxwell-Poisson distribution, called LCMP estimator, is discussed in Chapter 4 and Chapter 5. This estimator can be used as population size estimator for a geometric distribution with parameter of success $p = 1 - \lambda$. The LCMP estimator is given

$$\widehat{N}_{LCMP} = n + \widehat{f}_0 = n + f_1 \exp(-\widehat{\beta}_0), \tag{6.41}$$

where $\widehat{\beta}_0$ is the intercept, achieved by fitting weighted lest square regression between $\log(r_x^*) = \log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\}$ against $\log(x+1)$, i.e

$$\log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\} = \widehat{\beta}_0 + \widehat{\beta}_1 \log(x+1).$$

Also, a normal approximation variance of LCMP estimator is given as:

$$\widehat{Var}(\widehat{N}_{LCMP}) = \frac{n f_1 e^{-\widehat{\beta}_0}}{n + f_1 e^{-\widehat{\beta}_0}} + (e^{-\widehat{\beta}_0})^2 f_1[1 + f_1 Var(\widehat{\beta}_0)]. \tag{6.42}$$

6.3.3 Chao estimator for the geometric mixture

The Chao estimator for the geometric mixture model was proposed by Niwitpong et al. (2013). The basic idea came from adding the mixing distribution $h^*(p)$ into the geometric model to create a more flexible model. Then, the geometric mixture is given as

$$k_x = \int_0^1 g_x(p) h^*(p) dp = \int_0^1 \{(1-p)^x p\} h^*(p) dp. \tag{6.43}$$

The moment inequality under the Cauchy-Schwarz inequality is

$$[E(WZ)]^2 \leq E(W^2)E(Z^2).$$

Let $W^2 = p$ and $Z^2 = p(1-p)^2$, so that $[E(WZ)] = \int_0^1 p(1-p)h^*(p)dp$, $E(W^2) = \int_0^1 ph^*(p)dp$ and $E(Z^2) = \int_0^1 p(1-p)^2h^*(p)dp$. As a consequence, the geometric mixture model is given as

$$\left(\int_0^1 p(1-p)h^*(p)dp\right)^2 \leq \left(\int_0^1 ph^*(p)dp\right)\left(\int_0^1 p(1-p)^2h^*(p)dp\right). \tag{6.44}$$

From (6.43) above, it follows that $k_0 = \int_0^1 ph^*(p)dp$, $k_1 = \int_0^1 p(1-p)h^*(p)dp$, and $k_2 = \int_0^1 p(1-p)^2h^*(p)dp$, and substituting in (6.44), it becomes $k_1^2 \leq k_0k_2$. After this, replacing the probabilities by their relative frequencies, gives the lower bound estimator $\widehat{f}_0 \geq \frac{f_1^2}{f_2}$. Finally, the Chao estimator based on the geometric mixture $\widehat{N}_{CG}$ is defined as

$$\widehat{N}_{CG} = n + \frac{f_1^2}{f_2}. \tag{6.45}$$

In addition, the variance estimation of $\widehat{N}_{CG}$ is given as

$$\widehat{Var}(\widehat{N}_{CG}) = \frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2}, \tag{6.46}$$

(see Niwitpong et al., 2013). The CG estimator is appropriate to estimate population size with geometric heterogeneity (contaminate based-geometric) models, however, it gives a larger variance compared with the MLEGeo estimator.

6.4 Simulation study

A simulation study is undertaken to investigate the performance of the proposed estimators and their competitors. The count data sets are generated following the geometric distribution with a variety of parameters. That is $X \sim Geo(p)$ where $p = 0.1, 0.15, 0.2, 0.25, 0.3, 0.5$. The population size N is set at $N = 100, 250$ for small sizes, $N = 500, 1,000$ for medium sizes, and $N = 5,000, 10,000$ for large sizes. Each data set is then rearranged in form of frequencies f_0, f_1, f_2, f_3, ..., f_m, corresponding to the counts 0, 1, 2, 3, ..., m. The frequency of zero count f_0 was omitted before estimating population sizes $\widehat{N}$.

6.4.1 Simulation result to investigate the performance of estimators

To study the performance of the estimator, each scenario is repeated T = 5,000 times. Thereafter, for each scenario the expected value and variance of is calculated as

$E(\widehat{N}) = \dfrac{1}{5,000}\sum_{t=1}^{5,000}(\widehat{N}_t)$ and $Var(\widehat{N}) = \dfrac{1}{4,999}\sum_{t=1}^{5,000}\left(\widehat{N}_t - E(\widehat{N})\right)^2$, respectively. Three measurements are used to evaluate the performance of population size estimators: the relative bias (RBias), the relative variance (RVar) and the relative root mean square error (RRMSE), defined in the same manner as Chapter 3 and Chapter 4.

1) Comparing the behaviour and performance of five estimators constructed under the geometric distribution

According to the result displayed in Figures 6.2, 6.3 and 6.4, all estimators represent an asymptotic unbiased property with respect to population size. As we can see almost all estimators except the LCMP provide an overestimation of population size for a small population and tend to be less biased when the population size increases. The TG estimator shows the most accurate property with the lowest bias on average.

In contrast, the ZG estimator is the worst performing with a dramatic overestimation and large variance, particularly for small population sizes. This leads to the highest value of the RRMSE as shown in the Figure 6.4. The RBias of the LCMP and the TG estimators tends to be identical when p is small (i.e $p = 0.1, 0.15$). The MLEGoe estimator provides the smallest variance in all circumstances as we expect. Another interesting point is that the new estimator, TG, not only has a small bias but also produces a variance close to the MLEGeo estimator.

In brief, for selecting a good estimator, a compromise between bias and precision might be appropriate. Therefore, we use the RRMSE as a measurement for comparing the performance of five estimators. The simulation results show that TG and MLEGeo estimators are likely to be the best choices to estimate population size, as they provide the smallest of RRMSE for all situations. The LCMP estimator is a more sensitive estimator in comparison to the TG and MLEGeo estimators. However, it remains a good choice for small values of p and performs better than the ZG and CG estimators under the geometric distribution.

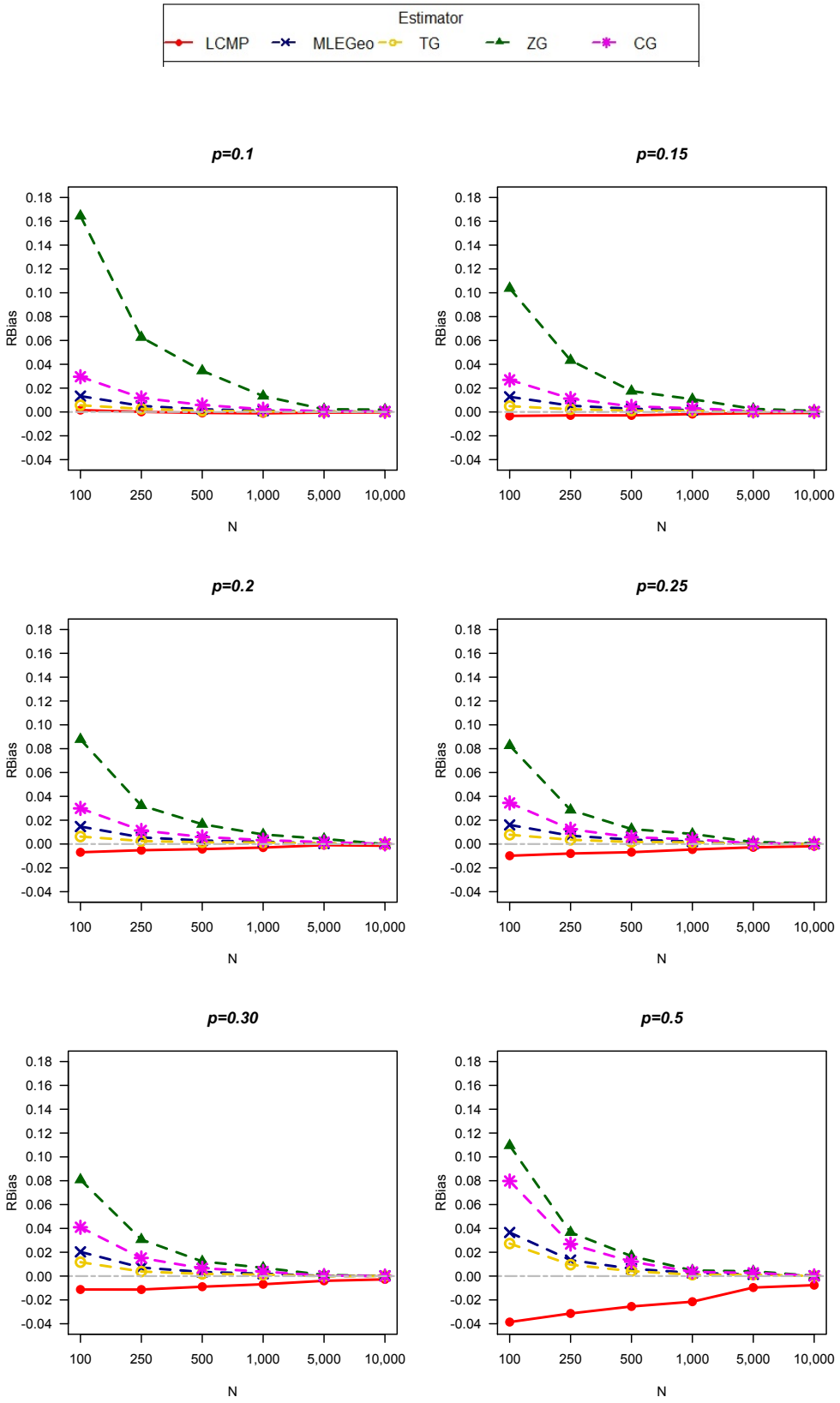

Figure 6.2: Relative bias of five estimators with different parameters following the geometric distribution

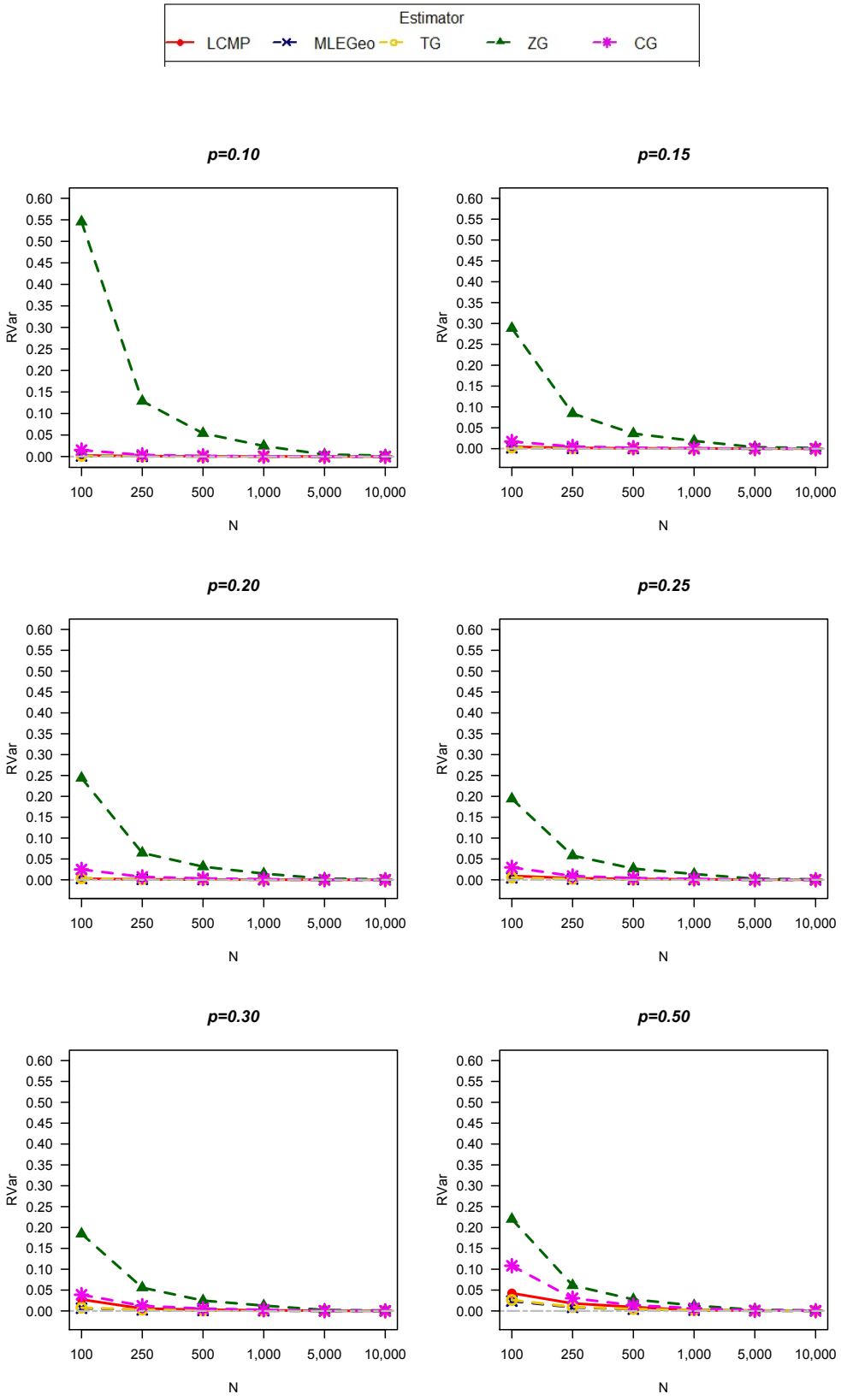

Figure 6.3: Relative variance of five estimators with different parameters following the geometric distribution

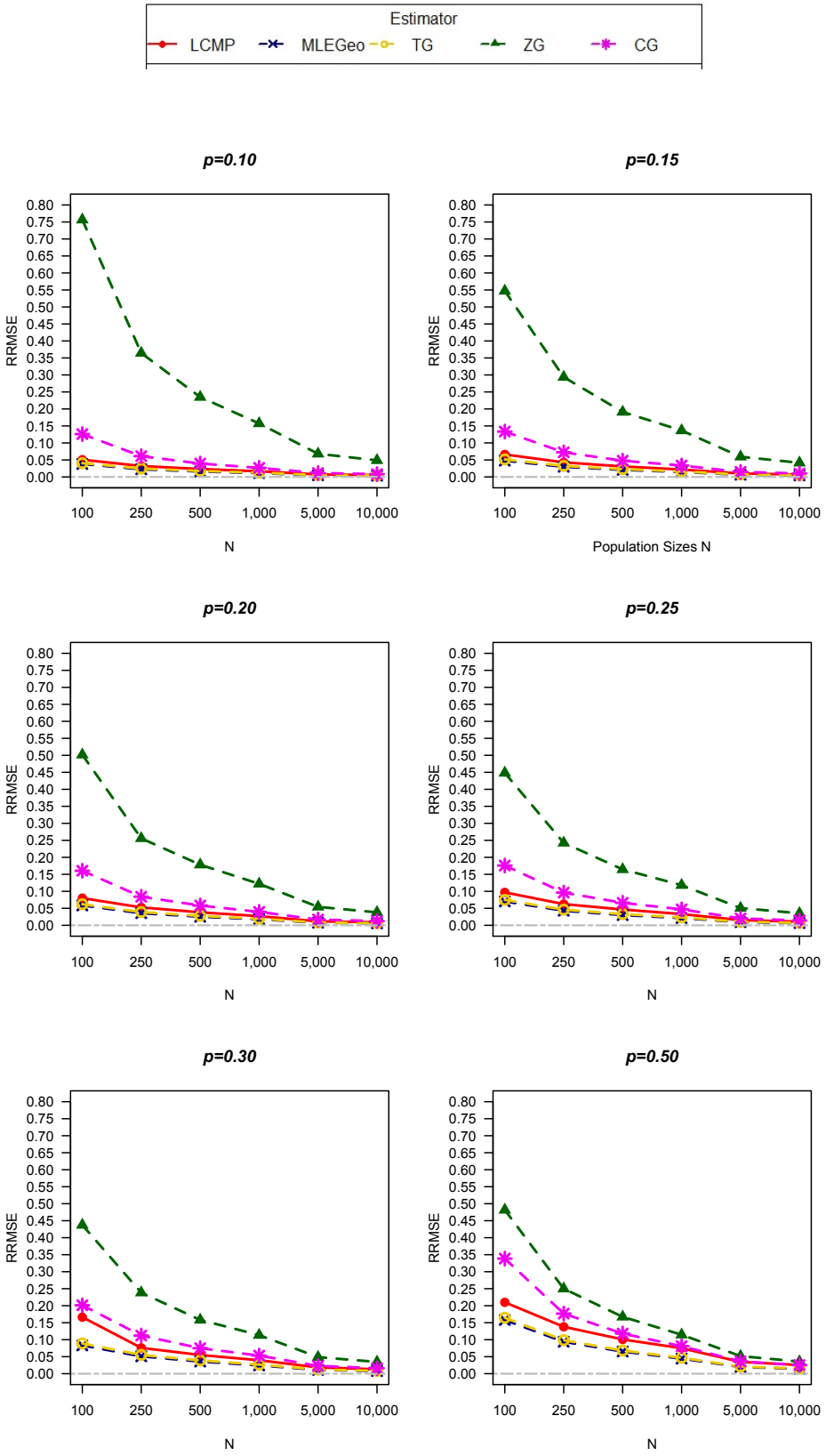

Figure 6.4: Relative root mean square error of five estimators with different parameters following the geometric distribution

2) Comparing the behaviour and performance of nine population size estimators

We add four more well-known population size estimators for capture-recapture data to the comparison. They are the MLEPoi estimator, the original Turing estimator allowing for the Poisson model, and the original Zelterman (Zel) estimator as well as the original Chao lower bound estimator developed for the contaminated Poisson disribution. The simulation results are provided in Figures 6.5, 6.6 and 6.7. It can be seen that the Turing, the MLEPoi and the Chao estimators provide an underestimation of population sizes as we expect. Moreover, they tend to dramatically underestimate of population size when the parameter p increase.

The new estimator, TG, performs better than the original Turing estimator with a smaller bias of population size estimate, and TG always provides the larger population size estimation than the original Turing estimator. The original Zelterman estimator is an asymptotic biased estimator in the case of the geometric distribution, providing overestimates of population size for small p and underestimation for large p. In contrast, the new estimator, ZG, shows a property of being asymptotic unbiased estimator. We found that both ZG and Zelterman estimators give larger variance than their competitors. However, increasing of population size leads to a decrease in the magnitude of variance for both estimators.

In conclusion, it might be recognised that the MLEPoi and the original Turing estimators are not suitable for estimating population size in the case of the geometric distribution as we can see the RRMSE being far away from zero for all situations. The TG and the MLEGeo estimators show the best performance to estimate population size with the smallest RRMSE whereas the LCMP estimator is appropriate only when p is close to zero. Indeed, the TG estimator using only f_1 and S, and being a nonparametric approach, it is easy to use in practice. The ZG estimator is also nonparametric in nature using only the counts of once and twos for estimating the population size. However, it requires a large population size N when it is used under the geometric distribution.

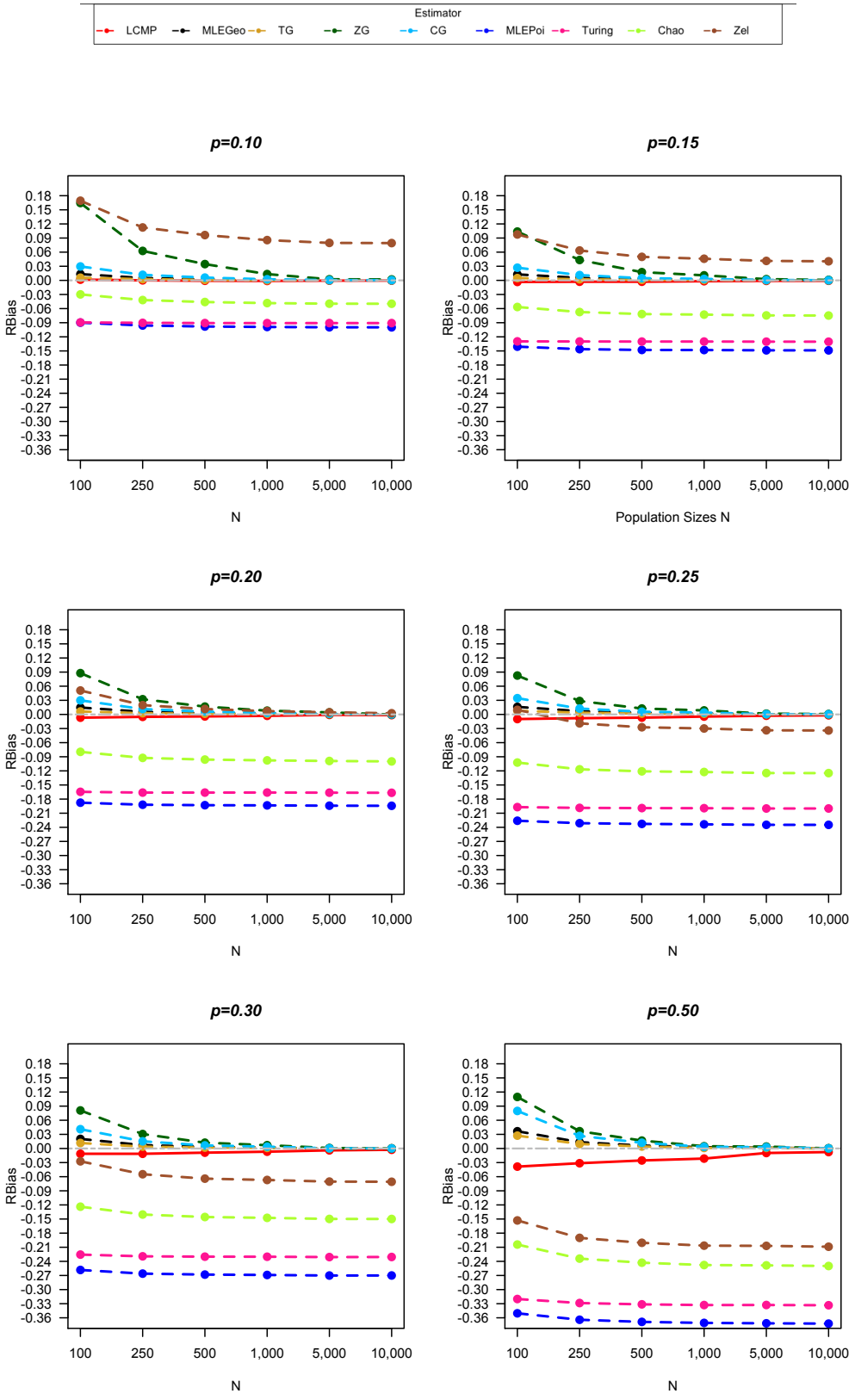

Figure 6.5: Relative bias of nine estimators with different parameters following the geometric distribution

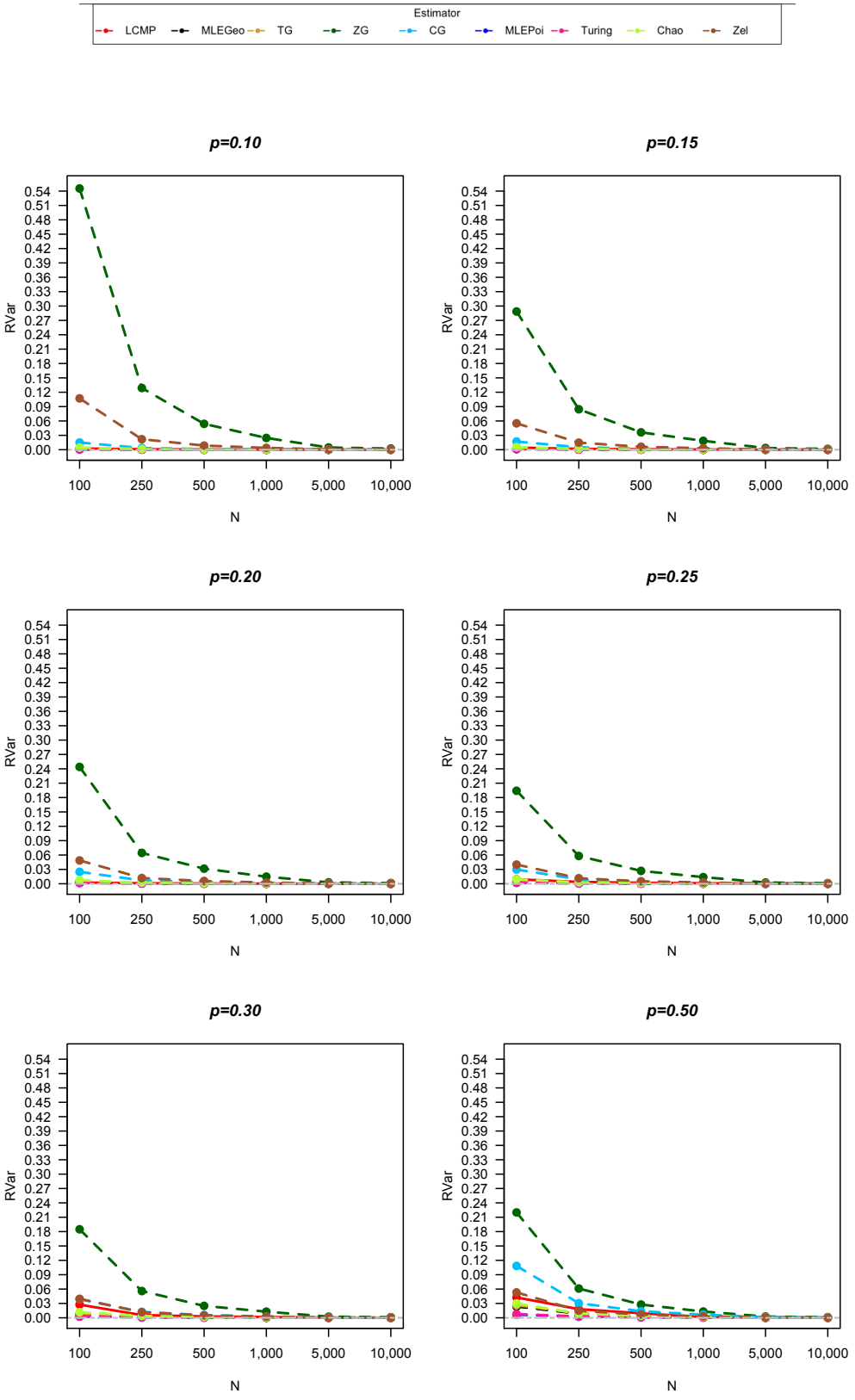

Figure 6.6: Relative variance of nine estimators with different parameters following the geometric distribution

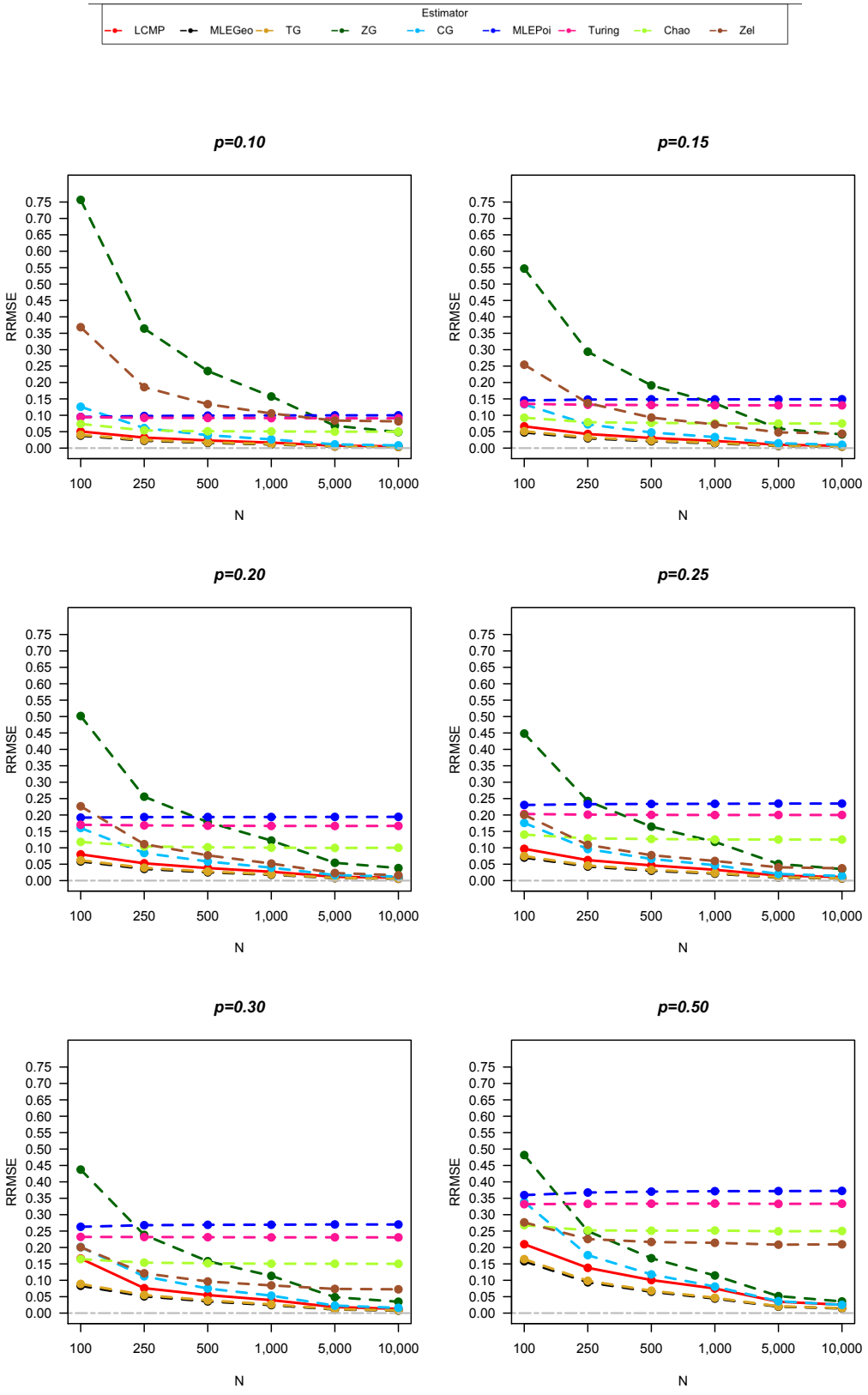

Figure 6.7: Relative root mean square error of nine estimators with different parameters following the geometric distribution

6.4.2 Simulation results for investigating the performance of variance approximation estimators

The aim of this part is to investigate the performance of the normal approximation variance estimators of the TG and the ZG estimators via the simulation study. As each simulation scenario was repeated T= 5,000 times, then the true variance can be calculated as

$$Var(\widehat{N}_{True}) = \frac{1}{T-1}\sum_{t=1}^{5,000}\left(\widehat{N}_{(t)} - E(\widehat{N})\right)^2,$$

hence, the true standard error is defined as $S.E.(\widehat{N}_{True}) = \sqrt{Var(\widehat{N}_{True})}$. Additionally, normal approximation variance is averaged as

$$E[\widehat{Var}(\widehat{N})] = \frac{1}{T}\sum_{t=1}^{5,000}\widehat{Var}(\widehat{N}_{(t)}),$$

where $\widehat{Var}(\widehat{N}_{(t)})$ is the normal approximation variance of estimator form replication t. Also, the expected of the approximate standard error is defined as

$$E[\widehat{S.E.}(\widehat{N})] = \sqrt{E[\widehat{Var}(\widehat{N})]}.$$

Moreover, the ratio of the standard error (RSE) is defined as

$$RSE = \frac{E[\widehat{S.E.}(\widehat{N})]}{S.E.(\widehat{N}_{True})}.$$

The standard error estimates of the two new estimators are provided in Table 6.3. The ZG estimator gives a larger standard error than the TG estimator for all situations. Additionally, the validity of the normal approximation of variance estimations is considered by using the ratio of standard error which are shown in Figure 6.8. The target value of this ratio is equal to one. Simulation results suggest that the normal approximation standard error of the ZG estimator is acceptable for population size 500 and above. It is remarkable that the normal approximation standard error from the ZG estimator provided a serious underestimation for small population sizes under the geometric distribution.

Investigating the performance of the normal approximation to the standard error of TG estimator, it is found that it tends to slightly underestimate on average, but it is likely to be a good approximation estimator for the true standard error. This is because the ratio of standard error from the TG estimator is fairly close to one.

Table 6.2: Comparing the standard errors with the true standard error of the TG, and the ZG estimators when data is generated from the geometric distribution; $Geo(p)$

N	TG estimator			ZG estimator		
	$E(\widehat{N})$	$S.E.(\widehat{N})$	$E[\widehat{S.E.}(\widehat{N})]$	$E(\widehat{N})$	$S.E.(\widehat{N})$	$\mathrm{E}[\widehat{S.E.}(\widehat{N})]$
			$Geo(0.10)$			
100	100.5	4.01	3.95	116.2	72.46	61.53
250	250.5	6.28	6.13	265.6	88.08	83.63
500	500.3	8.65	8.59	512.8	117.07	112.74
1,000	1,000.6	12.31	12.12	1,013.0	159.67	156.37
5,000	5,000.6	27.35	27.01	5,006.8	346.09	344.12
10,000	10,000.6	39.42	38.19	10,014.0	481.40	486.49
			$Geo(0.15)$			
100	100.6	5.05	5.02	112.1	60.67	49.72
250	250.7	8.04	7.82	260.7	71.75	69.75
500	500.7	11.17	10.99	509.7	96.81	95.43
1,000	1,001.0	16.03	15.51	1,011.0	133.69	133.28
5,000	5,000.6	35.87	34.59	5,011.6	298.73	294.56
10,000	10,001.0	49.18	48.89	10,002.1	420.09	415.50
			$Geo(0.20)$			
100	100.8	6.14	6.10	110.5	47.74	43.86
250	250.6	9.89	9.47	257.1	65.38	62.44
500	500.6	13.71	13.33	508.6	87.46	86.71
1,000	1,000.3	19.61	18.79	1,007.3	122.40	120.97
5,000	4,998.8	42.71	41.91	5,007.7	255.60	268.31
10,000	10,000.4	61.91	59.30	9,998.1	379.46	378.32
			$Geo(0.25)$			
100	100.8	7.45	7.19	108.3	43.85	40.42
250	250.8	11.59	11.20	258.6	60.42	59.34
500	500.6	16.24	15.74	508.5	82.02	81.93
1,000	1,001.3	23.15	22.24	1,009.2	115.04	114.42
5,000	5,001.1	50.17	49.57	5,006.6	251.57	253.23
10,000	10,001.8	73.98	70.08	10,009.5	362.64	357.87
			$Geo(0.30)$			
100	100.8	8.80	8.39	107.4	41.94	38.68
250	250.8	13.49	13.06	256.5	58.65	56.97
500	501.2	19.04	18.38	508.0	79.40	79.21
1,000	1,000.8	27.05	25.89	1,006.0	110.93	110.51
5,000	5,002.4	60.79	57.81	5,014.9	245.20	245.78
10,000	10,002.2	84.61	81.69	10,007.1	346.96	346.65
			$Geo(0.50)$			
100	102.5	16.02	16.08	109.8	46.68	41.68
250	252.3	24.31	24.37	257.9	62.19	59.82
500	501.8	33.73	33.91	507.8	83.91	82.59
1,000	1,003.2	48.19	47.79	1,007.4	116.64	115.23
5,000	5,004.2	109.55	106.35	5,017.7	259.35	256.20
10,000	9,998.8	150.20	150.06	10,003.6	361.75	360.86

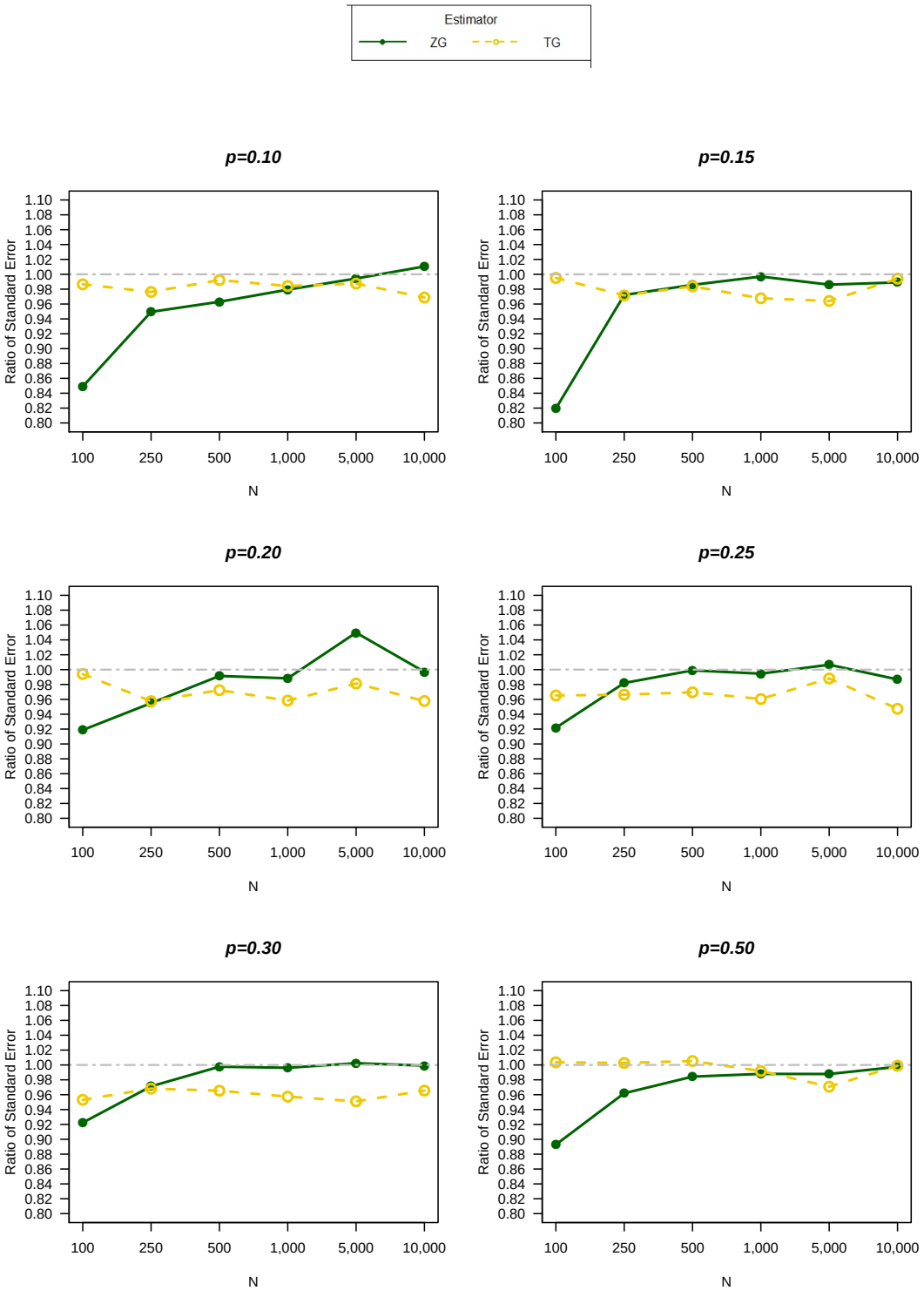

Figure 6.8: Ratio of standard errors for the TG and the ZG estimators

6.4.3 Simulation results for investigating the performance of confidence interval

In the simulation study confidence intervals for population size, N, are also considered. Confidence intervals for five estimators constructed based on the geometric distribution are compared. A 95% approximate confidence interval of population size N is constructed under the symmetric normal approximation as:

$$95\%CI = (\widehat{N}_L, \widehat{N}_U) = (\widehat{N} \pm Z_{.975} S.E.(\widehat{N})), \tag{6.47}$$

where $S.E.(\widehat{N})$ is appoximated by $\widehat{S.E.}(\widehat{N})$. The performance of the confidence intervals is quantified using the coverage probability (Cov) and the average length (AL) is defined as follows:

$$Cov = \frac{\sum_{t=1}^{T} A_{(t)}}{T} \times 100, \tag{6.48}$$

where $A_{((t))}$ equal to 1 if the true value N is in the target confidence interval, and 0 otherwise. The average length (AL) defined as

$$AL = \frac{\sum_{t=1}^{T} (\widehat{N}_U - \widehat{N}_L)}{T}. \tag{6.49}$$

The confidence intervals are produced based upon the assumption of asymptotic normality. It can be seen from the Figure 6.9 and Table 6.3 that the coverage probabilities from almost all estimators are lower than the nominal level of 95% when the population sizes are small and converge to the nominal level with increasing N.

Overall, the simulation results suggest that the MLEGeo estimator provides the best performance in perspective of confidence intervals. As can be seen, the coverage probability of the MLEGeo estimator is close to the nominal level on average and provides the shortest length. Another choice for the small population is the proposed TG estimator, it can be seen that it is satisfactory for small population sizes. The CG and the ZG estimators require population size $1,000$ or above to achieve confidence intervals which perform well at 95%. However, the ZG estimator has the widest average length compare with others.

Another interesting estimator as we discussed in Chapter 4 and 5 is the LCMP estimator. It can be seen that the coverage probability is seriously anti-conservative with much lower than the nominal levels for small population sizes especially for lower capture probabilities (p = 0.5). Additionally, it tends to be higher than the nominal levels when the population sizes increase, this is a consequence of the large standard errors which lead to wide confidence intervals. Accordingly, it is more appropriate to construct the

confidence interval of the LCMP estimator with the imputed bootstrap approach as previously suggested.

Table 6.3: Comparison of the performance of confidence intervals of six estimators when data is generated from the geometric distribution

	Coverage probability (%)					Average length				
N	LCMP	MLEGeo	TG	ZG	CG	LCMP	MLEGeo	TG	ZG	CG
	Geo(0.1)									
100	91.6	94.4	92.9	91.4	88.5	21.2	14.8	15.5	241.3	40.9
250	94.6	94.7	94.2	93.6	92.9	34.8	22.4	24.0	327.8	55.7
500	94.8	94.5	93.6	94.8	93.8	49.8	31.2	33.7	446.1	75.3
1,000	95.3	94.5	94.1	94.8	94.3	71.0	43.9	47.5	613.5	103.4
5,000	96.8	94.8	94.9	95.3	95.3	161.9	97.6	105.9	1,349.9	227.4
10,000	96.3	94.4	94.0	94.5	95.0	230.1	137.8	149.7	1,907.9	321.2
	Geo(0.15)									
100	91.8	94.2	93.0	91.2	89.5	28.9	18.9	19.7	189.9	47.3
250	93.7	94.2	93.5	94.0	92.8	47.2	28.9	30.6	274.0	68.2
500	94.1	94.7	93.8	94.5	93.4	67.6	40.4	43.1	373.8	92.8
1,000	95.5	94.2	93.9	94.8	94.4	96.9	56.8	60.8	522.4	130.0
5,000	96.2	94.3	93.8	94.7	94.4	221.1	126.5	135.6	1,154.9	287.0
10,000	96.9	95.1	94.8	94.7	95.1	314.2	178.7	191.7	1,630.1	404.9
	Geo(0.20)									
100	92.0	94.5	93.6	91.6	90.8	37.5	23.0	23.8	169.5	55.1
250	94.2	94.9	93.4	94.5	93.4	61.0	35.4	37.2	245.5	80.2
500	94.7	94.6	93.5	94.2	93.8	87.4	49.5	52.3	339.6	110.6
1,000	95.6	95.0	93.9	94.7	94.4	124.9	69.7	73.7	474.1	154.4
5,000	97.6	96.0	96.2	94.0	96.0	283.8	155.1	164.5	1,053.7	342.9
10,000	96.4	94.6	93.5	95.0	94.5	404.0	219.2	232.5	1,483.1	482.1
	Geo(0.25)									
100	91.6	94.8	93.3	92.8	91.4	47.2	27.4	28.2	158.2	63.2
250	94.0	95.0	94.0	94.4	94.0	76.4	42.2	43.9	231.0	92.3
500	94.1	94.9	93.7	94.4	94.0	108.8	59.0	61.7	319.3	127.2
1,000	95.6	94.7	93.8	94.0	94.5	156.6	83.1	87.2	448.3	178.8
5,000	96.3	95.2	94.0	94.9	94.9	355.0	184.9	194.2	993.3	395.6
10,000	96.9	95.5	94.7	94.9	95.2	504.6	261.4	274.7	1,402.6	558.9
	Geo(0.3)									
100	91.3	95.0	93.2	91.9	91.2	61.4	32.3	33.1	152.7	71.8
250	93.4	94.6	93.5	93.7	93.3	94.4	49.4	51.2	224.4	105.3
500	94.2	94.5	92.6	94.7	94.1	134.8	69.2	72.0	309.1	144.9
1,000	95.5	95.2	93.6	94.4	94.5	192.8	97.5	101.5	433.6	203.1
5,000	96.3	95.0	94.0	95.2	95.2	437.7	217.0	226.4	961.6	450.0
10,000	96.7	95.4	94.5	95.2	95.3	620.6	306.9	320.2	1,357.7	635.4
	Geo(0.50)									
100	89.5	95.5	94.6	92.8	92.3	148.8	59.8	63.4	165.9	117.9
250	92.5	94.8	94.5	94.3	94.2	219.5	90.1	95.4	235.9	167.1
500	94.2	94.5	94.6	94.9	94.7	308.9	125.7	133.1	324.1	229.4
1,000	94.0	95.3	95.0	94.9	94.7	433.4	176.5	186.9	450.7	318.8
5,000	97.2	97.4	96.6	95.0	95.4	969.0	393.1	416.8	1,004.7	710.5
10,000	96.7	94.7	94.9	95.3	95.2	1,380.2	554.8	588.3	1,414.1	1,000.0

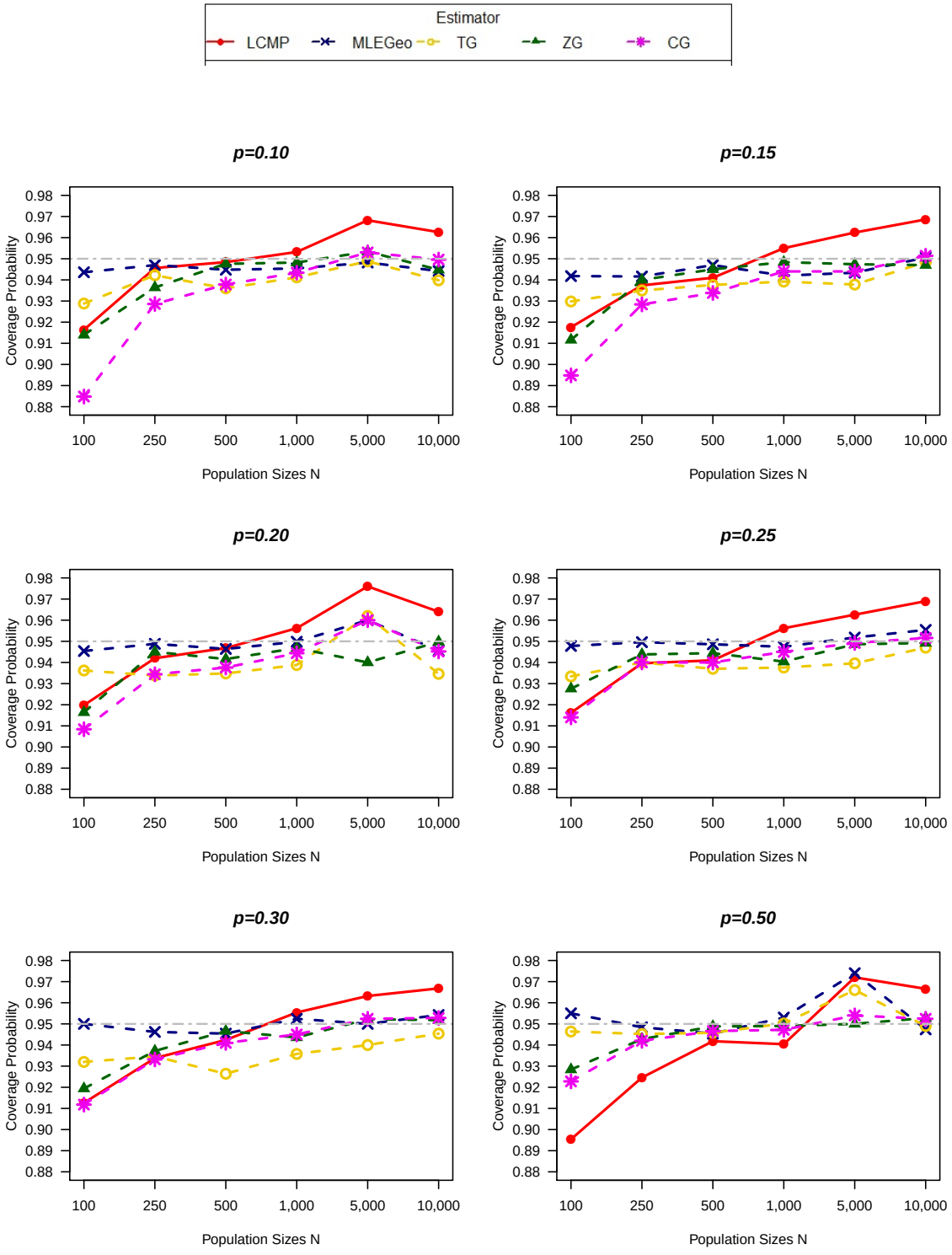

Figure 6.9: Coverage probabilities of 95% confidence intervals when data is generated from the geometric distribution

6.5 Real data examples

In this section, the estimators described above are applied to real data examples. Real data examples which follow a geometric distribution are selected, including: the golf tees data which has a small population size, and the heroin drug users in Bangkok data with a larger population. Additionally, the wood mice data are also considered. An analysis compares estimations from nine different estimators in terms of: the population size, standard error of population size, confidence interval and length of the confidence interval. The fitted frequencies and the Chi-square goodness of fit are used to evaluate goodness of fit under zero-truncated Poisson, zero-truncated geometric and zero-truncated Conway-Maxwell-Poisson distributions. Since the frequency of zero counts does not appear in the capture-recapture history, we predict the observed frequencies based on the observed counts n. More precisely, the chi-square goodness-of fit of the zero-truncated distribution is computed as $\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \widehat{f}_x)^2}{\widehat{f}_x}$, where $\widehat{f}_x = n\widehat{p}_x^+$, and $\widehat{p}_x^+$ is the capture probability estimated based upon the zero-truncated distribution.

6.5.1 Golf tees data

The first example is the golf tees data and is useful for comparing the performance of estimators as the true value population size, $N = 250$ is provided. It is assumed that the observed count distribution can be distributed by the zero-truncated geometric distribution. As can be seen in Figure 6.10, the geometric ratio plot ($\frac{f_{x+1}}{f_x}$ vs x) on left hand side provides a horizontal line with negligible residual. Additionally, using in the log ratio of Conway-Maxwell-Poisson distribution as shown on the right hand side with $\log[(x+1)\frac{f_{x+1}}{f_x}$ vs $\log(x+1)]$, there is a linear line with positive slope. It can, therefore, be assumed that the golf tees data follows the geometric or/and the Conway-Maxwell-Poisson distributions. Hence, it is expected that the estimators the MLEGeo, TG, ZG, LCMP and CG might be suitable estimators to estimate the number of golf tees.

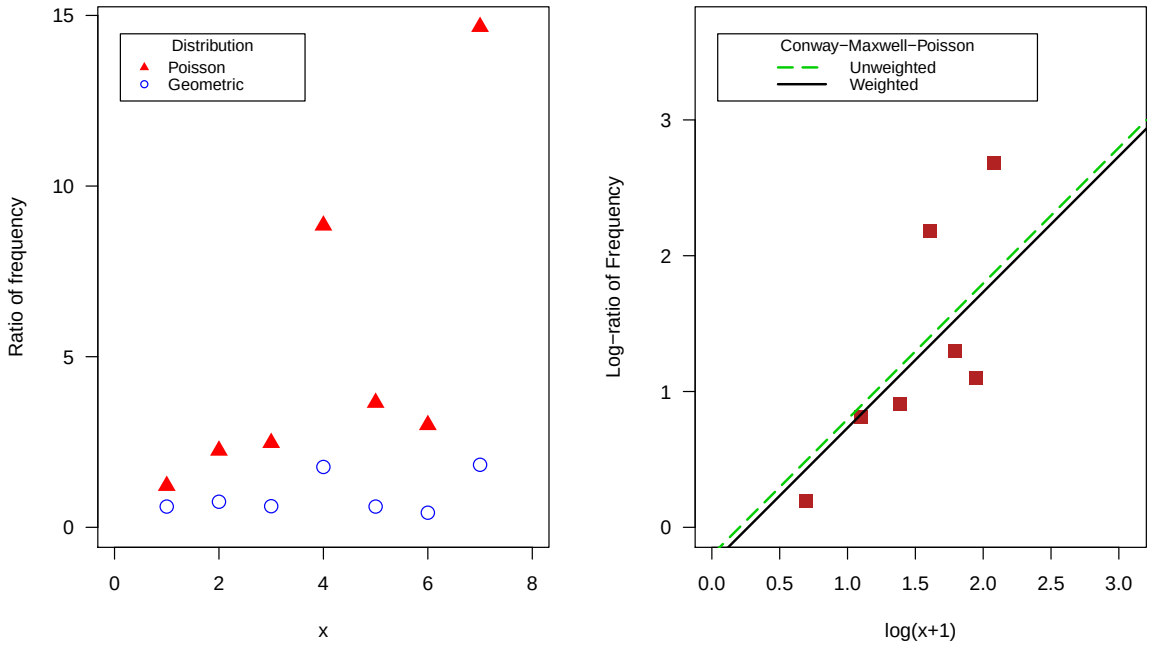

Figure 6.10: The left panel displays the Poisson ratio plot (red) and geometric ratio plot (blue), and the right panel shows the log-ratio plot for the CMP with linear line for the golf tees data

To evaluate the performance of population size estimations, nine estimators are compared and results are summarised in Table 6.4. The population size estimators for heterogeneity (i.e., Zel, TG, ZG, MLEGeo, LCMP and CG) estimate the population size to be fairly close to the true number $N = 250$. In detail, under the geometric distribution assumption it is found that the TG estimator and the MLEGeo estimator result in a negligible difference in the estimated population size and associated standard error. However, the geometric distribution might be not be the best fit as as the ratio plot exhibits some errors. In this case, the other two estimators which are constructed for the contaminated geometric (ZG and CG) are much more suitable for estimating the population size than the TG, LCMP and MLEGeo estimators.

It is no surprise that the Turing and MLEPoi estimators underestimate the population size, and we found that their confidence intervals do not cover the true value. Although the original Zelterman estimator and the modified ZG estimator provide the population size close to the true value, the standard errors are very large leading to very wide length of confidence intervals. The LCMP estimator might then be an appropriately alternative choice for estimating the number of golf tees although it leads to a light bias. However, when using the LCMP estimate the imputed bootstrap method is recommended as the method of choice for calculating the associated standard error and constructing confidence intervals. It might be suggested that the TG, the MLEGeo and the LCMP

estimators are the best choices for estimating the number of golf tees data as they provide acceptable values of population size estimation and the shorter length of confidence intervals compared with compared to the other estimators.

Table 6.4: The approximate number of golf tees data with standard errors, confidence intervals and lengths of confidence interval from nine estimators.

Estimator	$\widehat{N}$	Asymptotic		Length
		$\widehat{SE}(\widehat{N})$	$95\%CI$	
Poisson				
Turing	177	4.58	(168 – 186)	18
MLEPoi ($\widehat{\lambda} = 3.23$)	169	2.83	(163 – 175)	12
Contaminated Poisson				
Chao	200	13.09	(174 – 226)	52
Zel	231	29.90	(171 – 289)	118
Geometric				
MLEGeo($\widehat{p} = 0.3$)	230	11.75	(207 – 253)	46
TG	228	12.03	(204 – 252)	48
Conway-Maxwell-Poisson				
LCMP	223	33.09	(159 – 288)	129
($\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$)		15.11	$(195 - 252)^{(a)}$	57
Contaminated geometric				
ZG	266	65.12	(138 – 394)	256
CG	238	27.86	(183 – 293)	110

(a): the 95% percentile confidence interval by using the imputed bootstrap method

Model fit is assessed and the results presented in Table 6.5 where the observed frequencies are followed by the zero-truncated Poisson, the zero-truncated Conway-Maxwell-Poisson and the zero-truncated geometric distribution, respectively. As can be seen, the MLEPoi estimator under the zero-truncated Poisson distribution provides a serious underestimation for the count of one. On the other hand, The LCMP estimator which is constructed based on the zero-truncated Conway-Maxwell-Poisson distribution and the MLEGeo estimators of the zero-truncated geometric distribution show the better fit when compared to the MLEPoi (see Figure 6.11). Additionally, the Chi-square goodness of fit values conditioning on the observed data are also presented in the right-hand side column. The zero-truncated Conway-Maxwell-Poisson distribution provides the best fit with the smallest value of Chi-square for the golf tees data following by the zero-truncated geometric distribution. Interestingly, the the model fitted frequencies from the LCMP and the MLEGeo are slightly different. This reason might be affected from the estimated location parameters are not equivalence ($\widehat{p}_{(MLEGeo)} \neq 1 - \widehat{\lambda}_{(LCMP)}$) since we use the different methods for estimating a location parameter.

Table 6.5: Observed and fitted values for the Golf tees data

x	1	2	3	4	5	6	7	8	χ^2
Observed data	46	28	21	13	23	14	6	11	
$\widehat{f}_x$(ZTPoi)	22	35	37	30	20	11	5	2	86.10
$\widehat{f}_x$(ZTCMP)	38	29	22	17	13	10	8	6	16.66
$\widehat{f}_x$(ZTGeo)	48	34	24	17	12	8	6	4	29.29

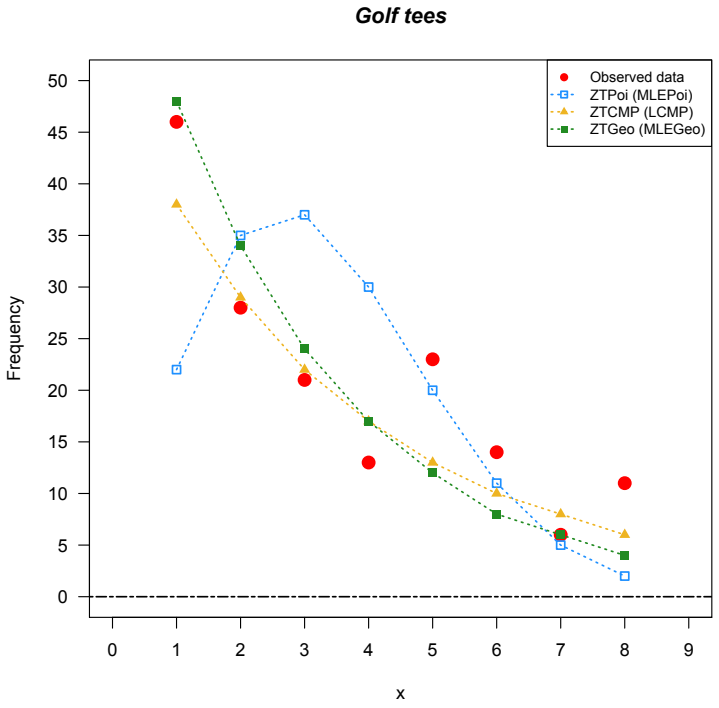

Figure 6.11: Observed frequencies with fitted frequencies based on the zero-truncated Poisson (ZTPoi), the zero-truncated Conway-Maxwell-Poisson (ZTCMP) and the zero-truncated geometric (ZTGeo) of golf tees data

6.5.2 Wood mice data

As shown earlier, the ratio plot of wood mice data provides evidence that the data follow a geometric distribution. Also, the log-ratio plot of the Conway-Maxwell-Poisson distribution is given in Figure 6.12 shows a linear line with positive slope and $\widehat{\beta}_1 = 1$. It can be considered as a special case of the CMP distribution specified by the geometric distribution with event parameter $p = 1 - \lambda$.

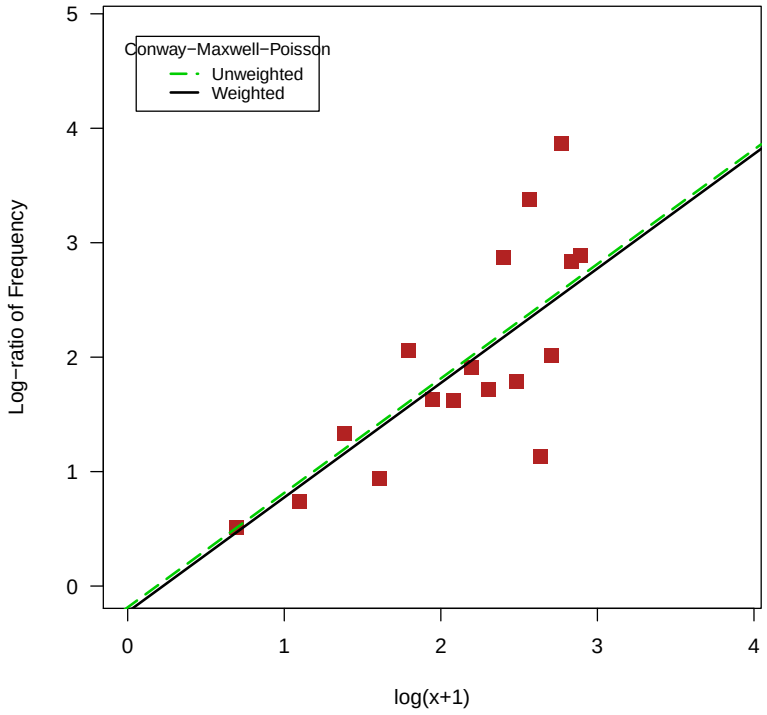

Figure 6.12: The log-ratio plot for the CMP with linear line for wood mice data

In order to compare the size of the wood mice population using nine estimators are compared in Table 6.6. The Turing and the MLEPoi estimators provide the smallest number of wood mice and the lowest standard errors resulting in a narrow length of 95% confidence interval. The number of wood mice from the TG, the MLEGeo and the LCMP estimators are slightly different, they are 426, 427 and 423, respectively. It can be seen that the standard error and confidence interval of MLEGeo and the TG estimators are of negligible difference. The LCMP estimator produces a higher value of standard error and wider confidence interval not only from the normal approximation method but also from the imputed bootstrap approach.

It is no surprise that the Chao estimator which assumes a Poisson mixture provides a smaller number of wood mice than the CG estimator which assumes a geometric mixture distribution. This is similar to the reason for the difference between the TG and Turing estimators; the former of which can detect heterogeneity. It is remarkable that all of the results appear to be supported by the simulation study results for the case of small p and medium population size.

Table 6.6: The approximate number of wood mice data with standard errors, confidence intervals and lengths of confidence interval from nine estimators.

Estimator	$\widehat{N}$	Asymptotic		Length
		$\widehat{SE}(\widehat{N})$	$95\%CI$	
Poisson				
Turing	350	4.52	(341 – 359)	18
MLEPoi ($\widehat{\lambda} = 4.51$)	338	1.96	(334 – 342)	8
Contaminated based-Poisson				
Chao	377	11.55	(354 – 400)	46
Zel	412	29.93	(353 – 471)	118
Geometric				
MLEGeo ($\widehat{p} = 0.22$)	427	12.33	(403 – 451)	48
TG	426	12.92	(401 – 451)	50
Conway-Maxwell-Poisson				
LCMP	423	24.64	(375 – 472)	97
($\widehat{\lambda} = 0.80$ and $\widehat{\nu} = 0$)		18.41	$(385 - 456)^{(a)}$	72
Contaminated based-geometric				
CG	419	24.91	(370 – 468)	98
ZG	402	71.38	(262 – 542)	280

(a): the 95% percentile confidence interval by using the imputed bootstrap method

To study the goodness-of-fit of the wood mice data, we compare the fitted frequencies under three zero-truncated count models as represented in Table 6.7 and shown in Figure 6.13. It might be concluded that the observed counts of wood mice data are close to the ones fitted by the geometric and the CMP distribution, leading to Chi-square values of the LCMP and the MLEGeo estimators that do not differ too much in terms of model fit. On the other hand, the zero-truncated Poisson is a misspecification for the wood mice data, given the large Chi-square value. Therefore, we do not suggest to estimate the number of wood mice by the MLEPoi and the Turing estimators.

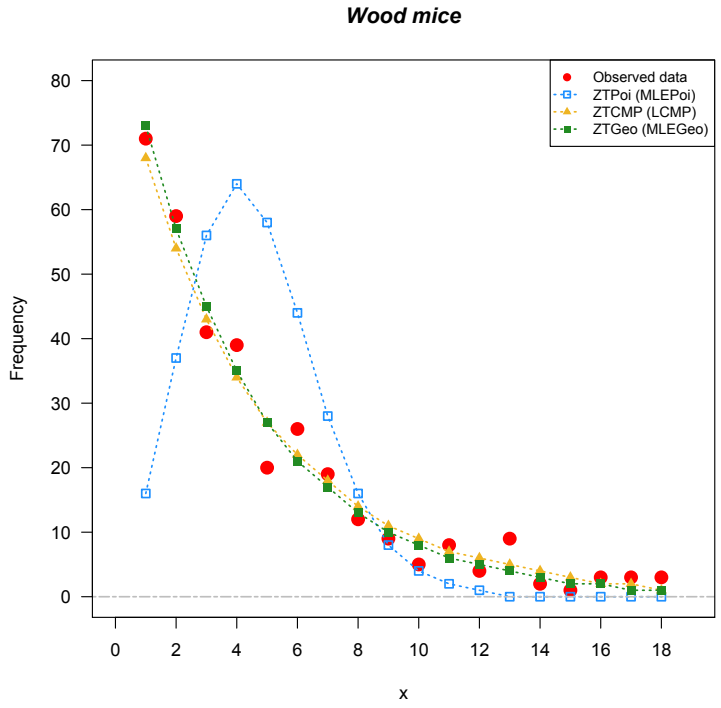

Figure 6.13: Observed frequencies with fitted frequencies based on the zero-truncated Poisson (ZTPoi), the zero-truncated Conway-Maxwell-Poisson (ZTCMP) and the zero-truncated geometric (ZTGeo) of wood mice data

Table 6.7: Observed and fitted frequency distribution for wood mice data

x	Observed data	$\widehat{f}_x$ (ZTPoi)	$\widehat{f}_x$ (ZTCMP)	$\widehat{f}_x$ (ZTGeo)
1	71	16	68	73
2	59	37	54	57
3	41	56	43	45
4	39	64	34	35
5	20	58	27	27
6	26	44	22	27
7	19	28	18	17
8	12	16	14	13
9	9	8	11	13
10	5	4	9	8
11	8	2	7	6
12	4	1	6	5
13	9	0	5	4
14	2	0	4	3
15	1	0	3	2
16	3	0	2	2
17	3	0	2	1
18	3	0	1	1
χ^2		279.46	17.791	21.93

6.5.3 Heroin users in Bangkok, Thailand

The final example of data that follow a geometric distribution is the heroin drug users in Bangkok data. As mentioned in Chapter 4, the geometric distribution is a special case of the Conway-Maxwell-Poisson distribution when the event parameter $p = 1 - \lambda$, $0 < \lambda < 1$ and dispersion parameter $\nu = 0$. The ratio plot of the CMP distribution is shown in the right-hand side in Figure 6.14. Additionally, the ratio plot for the geometric distribution is shown in the left hand side, indicates that the heroin drug users data can be modelled by the geometric distribution. Therefore, the population size estimators under the geometric are not too different. Moreover, the TG and the MLEGeo estimators give reasonable standard errors and 95% confidence interval lengths. Although the population size from the LCMP estimator is close to the MLEGeo and TG estimators, it shows the larger values for both, the standard errors and confidence intervals. Comparing the model fit in Table 6.9 and Figure 6.15, it can be seen that the zero-truncated geometric as well as the zero-truncated Conway-Maxwell-Poisson distribution provide the better fit with smaller values of the Chi-square than the zero-truncated Poisson. However, Chi-square values are slightly different since we use the different techniques for estimating the location parameters leading to $\widehat{p}_{(MLEGeo)} \neq 1 - \widehat{\lambda}_{(LCMP)}$.

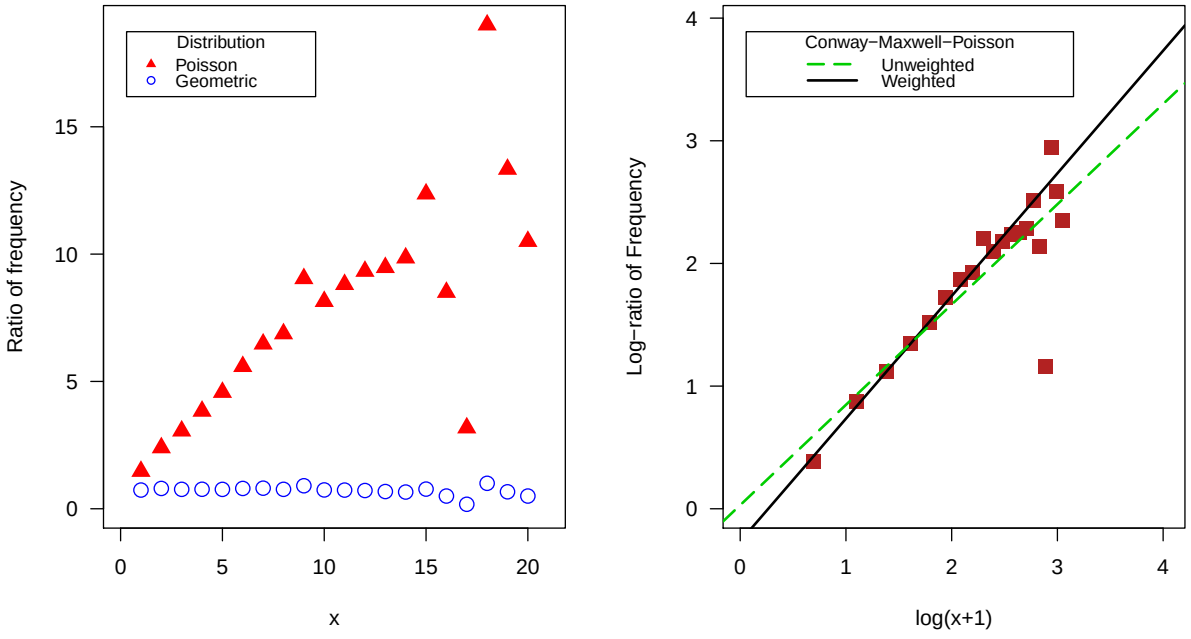

Figure 6.14: The left panel are Poisson ratio plot (red) and geometric ratio plot (blue), and the right panel is the log-ratio plot for the CMP with linear line for heroin users in Bangkok data

Table 6.8: The approximate number of heroin drug use in Bangkok data with standard errors, confidence intervals and lengths of confidence interval from nine estimators.

Estimator	$\widehat{N}$	Asymptotic		Length
		$\widehat{SE}(\widehat{N})$	$95\%CI$	
Poisson				
Turing	9,850	26.65	(9,798-9,850)	52
MLEPoi ($\widehat{\lambda} = 4.13$)	9,454	12.84	(9,429-9,479)	50
Contaminated based-Poisson				
Chao	10,782	71.86	(10,641 – 10,923)	282
Zel	12,077	184.54	(11,715 – 12,439)	724
Geometric				
MLEGeo ($\widehat{p} = 0.67$)	12,207	70.73	(12,068 – 12,346)	278
TG	12,175	73.99	(12,030 – 12,320)	290
Conway-Maxwell-Poisson				
LCMP	12,141	210.24	(11,729 – 12,554)	825
($\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$)		101.31	$(11,918 - 12,320)^{(a)}$	402
Contaminated based-geometric				
CG	12,261	156.63	(11,954 – 12,568)	614
ZG	12,651	416.62	(11,834 – 13,468)	1,634

(a): the 95% percentile confidence interval by using the imputed bootstrap method

Table 6.9: Observed and fitted frequency distribution for the heroin users in Bangkok data

x	Observed	$\widehat{f}_x$ (ZTPoi)	$\widehat{f}_x$ (ZTCMP)	$\widehat{f}_x$ (ZTGeo)
1	2,176	626	2,214	2,179
2	1,600	1,293	1,687	1,670
3	1,278	1,783	1,285	1,281
4	976	1,843	980	982
5	748	1,524	746	753
6	570	1,050	569	577
7	455	620	433	442
8	368	321	330	339
9	281	147	252	260
10	254	61	192	199
11	188	23	146	153
12	138	8	111	117
13	99	3	85	90
14	67	1	65	69
15	44	0	49	53
16	34	0	38	40
17	17	0	29	31
18	3	0	22	24
19	3	0	17	18
20	2	0	13	14
21	1	0	10	11
χ^2		26,844.98	106.13	94.78

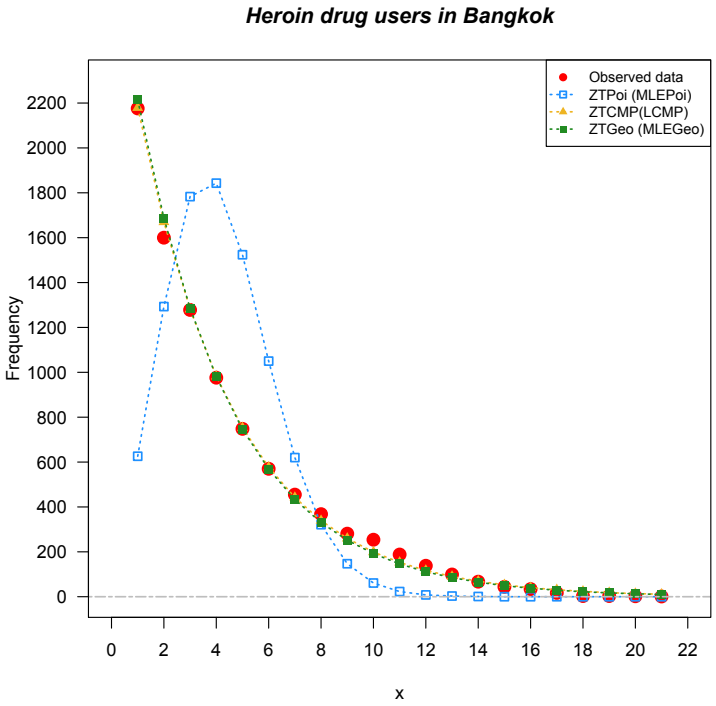

Figure 6.15: Observed frequencies with fitted frequencies based on the zero-truncated Poisson (ZTPoi), the zero-truncated Conway-Maxwell-Poisson (ZTCMP) and the zero-truncated geometric (ZTGeo) of heroin drug users in Bangkok data

6.6 Conclusion

Evidence of homogeneous probabilities under the Poisson distribution might not be available for natural capture-recapture data, which may instead be represented by a mixture of Poisson distributions. The Poisson mixture with the exponential density as mixing density is a well-known parametric mixture model, leading to the geometric distribution. It has been applied in the capture-recapture context by several authors (see Lanumteang, 2010; Niwitpong et al., 2013; Vidal-Diez, 2015), resulting in several population size estimators constructed using the geometric model. However, they have some limitations in their applications to real data. For example, the MLEGeo estimator is sensitive to the occurrence of contaminations based on the geometric distribution (Niwitpong et al., 2013). As a consequence, there is a need for appropriately modified estimators. Moreover, to allow for fairness of comparison to the performance of the LCMP estimator, they should be compared under the same basic assumption such as the geometric model. The Turing and the Zelterman estimators are extended to be the Turing estimator based on the geometric distribution (TG) and the Zelterman estimators based on the geometric distribution (ZG), respectively. The advantages of the original Turing estimator are the small bias and low variance, leading to excellent performance

for capture-recapture data under the Poisson distribution. Also, the original Zelterman estimator is one of the robust estimators in case of misspecification and occurrence of heterogeneity (Lanumteang, 2010). The two new modified estimators based on the Turing and Zelterman as well as their normal approximate variance estimators are proposed in Section 6.2.

The simulation study is carried out to investigate the performance of the proposed estimators and their behaviour is compared not only with the estimators based on the geometric distribution but also with other well-known estimators. The simulation results show that TG and ZG estimators are the asymptotic unbiased estimators with respect to the population size. The MLEPoi and the Turing estimators are not suitable for the heterogeneity. Estimates from the Chao estimator confirm the lower bound property as they are the lowest among the population size estimators for heterogeneity (i.e MLEGeo, Zel, ZG, CG, TG and LCMP). The simulation results also demonstrate that the TG estimator can detect the heterogeneity following the geometric distribution by increasing the estimate of population size compared with the original Turing. The ZG estimator can be used for the large population size when data follow the geometric distribution.

With respect to variance estimation of the new estimators, the simulation suggests that the variance estimation of the TG estimator provides a negligible underestimation of variance, comparing with the true variance. In contrast, the variance estimator of the ZG estimator shows a severe underestimation of variance when the population size is small. This may be due to a violation of the normal approximation variance occurs here as the approximation approach requires a large population size. As a consequence, the resampling approach for estimating the variance of the ZG estimator such as the bootstrap method should be considered in future study.

A 95% confidence interval for the proposed population size estimator of N is constructed based on their variance and population size estimations following the normal approximation approach. Overall, the coverage probability of the TG shows the better performance than the ZG estimator. However, it is less satisfactory than the results produced by the MLEGeo estimator.

In short, the TG and the MLEGeo estimators perform the best for population size estimation for capture-recapture data under the geometric distribution. They provide the most accurate and precise leading to the highest performance all situations. Additionally, it is acceptable to estimate variances and constructs confidence intervals with normal approximation methods for these estimators. The LCMP estimator is an alternative choice and work very well for the small p. The ZG estimator is constructed for the contaminated geometric distribution, then some contaminated model or a continuous heterogeneity distribution such as the negative binomial should be consider in future studies.

Chapter 7

Variance Estimation for Single Marking Capture-Recapture Data

7.1 Introduction

Parametric approaches for estimating a target population size consist of two main steps. The first step is the process of selecting the parametric model which best fits the data, e.g., the hypergeometric, the Poisson or the binomial distributions and so forth. The second step is that all parameters of interest are estimated under the selected model. The most important property of a good estimator is that it is close to the true value of the parameter being estimated, or, in other words, it is unbiased. Additionally, its variance should be low, meaning it has good precision which in turn results in narrow confidence intervals.

Various methodologies have been proposed to estimate the variances and standard errors of parameters of interest. The assumption of an approximately normal distribution is commonly used for constructing a variance estimator. This approach, however, is based on an approximation that is not always a good choice. Several competing methods are available that perform better, especially for small sample sizes.

A well-known bootstrap approach have been applied to quantify the precision of capture-recapture data by many authors (see Norris III and Pollock, 1996; Zwane and Van der Heijden, 2003; Buckland, 1984; Buckland and Garthwaite, 1991). However, evidence of the adequacy of the estimated standard errors and confidence intervals for single marking problems is limited. In this chapter, three bootstrap approaches based on the multinomial distribution for estimating the variance of population size (standard error) and construct confidence intervals of single marking population size estimators are proposed. Comparison of the behaviour and performance of variance estimation

methods and confidence intervals for the Chapman estimator and the Chao estimator based on the binomial mixture distribution are provided in this chapter.

7.2 Population size estimators for single marking capture recapture data

The origin of the capture-recapture approach is the single marking method that was developed to estimate wildlife populations. It has also been utilised in epidemiology, social science and surveillance (Brittain and Böhning, 2009; Vuillermoz et al., 2014; Xu et al., 2014; Jacquinet et al., 2015; Vergne et al., 2012). The single marking capture-recapture data is defined as a capture history by means of only two observed samples or data arising from two sources (occasions). In this study, we focus on the two sources design which is frequently applied in epidemiology and social science studies. The capture history of a single marking study can be viewed as a 2x2 contingency table with frequencies as illustrated in Table 7.1.

Table 7.1: Contingency table representing capture history from two sources

	Source 2		
Source 1	Yes	No	Total
Yes	f_{11}	f_{10}	n_1
No	f_{01}	f_{00}	
Total	n_2		N

where

- N denotes the target population size.
- f_{00} denotes the frequency for an individual not being observed.
- f_{10} denotes the frequency of individuals identified by the first source only.
- f_{01} denotes the frequency of individuals identified by the second source only.
- f_{11} denotes the frequency of individuals identified by both two sources.
- n_1 denotes the total number of individuals identified by source one.
- n_2 denotes the total number of individuals identified by source two.

A general likelihood function of population size (N) for the single marking model can be written as a multinomial likelihood with five parameters $(N, p_{00}, p_{10}, p_{01}, p_{11})$,

$$\binom{N}{f_{00} \quad f_{10} \quad f_{01} \quad f_{11}} p_{00}^{f_{00}} p_{10}^{f_{10}} p_{01}^{f_{01}} p_{00}^{f_{00}}, \tag{7.1}$$

where

- p_{00} denotes the probability of an individual not being observed.
- p_{10} denotes the capture probability of individuals identified by the first source only.
- p_{01} denotes the capture probability of individuals identified by the second source only.
- p_{11} denotes the capture probability of individuals identified by both sources.

The basic assumptions for single marking capture-recapture data are that the target population is closed (i.e. the population cannot change over the course of the study), individual homogeneity, and independence between two sources. In practice, the violation of assumptions usually occurs due to dependence between sources which results in overestimation or underestimation of population size (Brenner, 1995). For example, in a study of road traffic death rates, sources such as hospital injury surveillance and traffic police records are not always independent since sometimes the police might obtains some medical evidence from hospitals.

The Chapman estimator has been commonly used for estimating two sources capture-recapture data under the basic assumption of independence. Interestingly, Brittain and Böhning (2009) suggested the Chao estimator based on the binomial mixture distribution for two sources as a alternative method when the independent assumption is in doubt.

7.2.1 Chapman's estimator

The Lincoln-Petersen (LP) estimator was constructed based on the basic assumption of two independent sources. The odds ratio (OR) is used as a measurement of the independence between two sources. Then from the contingency table as Table 7.1 we have that $OR = 1 = \frac{f_{11} f_{00}}{f_{10} f_{01}}$. The number of unobserved individuals can be estimated as $\widehat{f}_{00,LP} = \frac{f_{10} f_{01}}{f_{11}}$. The Chapman (CM) estimator was developed from the LP estimator to address the situation that the number of observed units in both sources, (f_{11}) is small or zero. An extra individual is added into the frequencies of individuals identified by both sources. The odds ratio for the CM estimator was modified to $OR = 1 = \frac{(f_{11}+1) f_{00}}{f_{10} f_{01}}$, which leads to

$$\widehat{f}_{00,CM} = \frac{f_{10} f_{01}}{f_{11}+1}.$$

Also, the Chapman estimator for estimating population size is given by

$$\widehat{N}_{CM} = \frac{(n_1+1)(n_2+1)}{m_2+1} - 1, \tag{7.2}$$

where $m_2 = f_{11}$ denotes the number of individuals observed by both sources, $n_1 = f_{10} + f_{11}, n_2 = f_{01} + f_{11}$ are the number of observed by source one and two, respectively. Amstrup et al. (2010) pointed out that the CM estimator is unbiased for $n_1 + n_2 \geq N$. It can be said that the CM estimator is more flexible and less biased than the original LP estimator. An approximately unbiased variance estimator of the CM estimator was derived by Seber (1970, 2002) as follows:

$$Var(\widehat{N}_{CM}) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)}. \tag{7.3}$$

7.2.2 Chao's estimator based on the binomial mixture distribution

Chao's estimator was developed by Brittain and Böhning (2009) for estimating population size using two source data based on the mixture binomial distribution with size parameter two, $(m = 2)$. That is

$$p_x = \int_0^1 \binom{2}{x} p^x (1-p)^{2-x} h^*(p) dp, \tag{7.4}$$

where $x = 0, 1, 2$ and $h^*(p)$ is a mixing density. The lower bound of zero counts for the Chao estimator based on the binomial mixture distribution with size parameter two (CB) is provided as:

$$\widehat{f}_{0,CB} = \frac{f_1^2}{4f_2}.$$

The population size estimator of the CB estimator is given as:

$$\widehat{N}_{CB} = n + \frac{f_1^2}{4f_2}, \tag{7.5}$$

when $f_1 = f_{01} + f_{10}$, $f_2 = f_{11}$ and $n = f_{10} + f_{01} + f_{11}$. Additionally, Zelterman upper bound was proposed based on the zeo-truncated binomial distribution with two sources (ZB). It was given as:

$$\widehat{N}_{ZB} = \frac{n}{1 - \left(\frac{f_1}{f_1+2f_2}\right)^2}. \tag{7.6}$$

Since in two sources capture-recapture data, the number of capture samples is always equal to two, it can be shown that the CB and the ZB estimators are identical.

$$\begin{aligned}
\widehat{N}_{ZB} &= \frac{n}{1-\left(\frac{f_1}{f_1+2f_2}\right)^2} = \frac{f_1+f_2}{1-\left(\frac{f_1}{f_1+2f_2}\right)^2} \\
&= \frac{(f_1+f_2)(f_1+2f_2)^2}{4f_1f_2+4f_2^2} = \frac{(f_1+f_2)(f_1+2f_2)^2}{4f_2(f_1+f_2)} \\
&= \frac{(f1+2f_2)^2}{4f_2} = \frac{f_1^2+4f_1f_2+4f_2^2}{4f_2} \\
&= f_1+f_2+\frac{f_1^2}{4f_2} = n+\frac{f_1^2}{4f_2} = \widehat{N}_{CB}. \qquad (7.7)
\end{aligned}$$

In the single marking case, it can also be proven that the CB estimator is identical to the maximum likelihood estimator under the binomial distribution and the Mckendrick's moment estimator based on the binomial distribution with two sources. Therefore, it is suggested that only the CB estimator is considered as an alternative to the original CM estimator. An approximate variance estimator of the CB estimator, which is equal to the variance estimate of ZB estimator, is given as

$$\widehat{Var}(\widehat{N}_{ZB}) = \widehat{Var}(\widehat{N}_{CB}) = \frac{f_1^2}{4f_2}\left(\frac{f_1}{2f_2}+1\right)^2, \qquad (7.8)$$

(see Brittain and Böhning, 2009).

7.3 Variance estimation methods

Appropriate variance estimation techniques are essential for high quality population size estimates. If the associated estimation of the variance is poor, then a coverage probability of the confidence interval may falsely indicate poor estimation. This section describes and compares four methodologies that are used to estimate the variance of population size estimators for single marking capture recapture data. The first method is a variance estimation based on the normal approximation method or formula method. The competitors are three bootstrap methods: true bootstrap, imputed bootstrap and reduced bootstrap methods, which were presented in Chapter 5 but they are modified to the single marking capture-recapture data in this Chapter.

Method 1 (M1): Variance estimation based on the normal approximation method or formula method

This method utilises the approximate normal estimation variance formulas of $\widehat{N}_{CM}$ and $\widehat{N}_{CB}$, given in (7.3) and (7.8) respectively. This method which is commonly used to approximate confidence interval relies on the asymptotic normal distribution. Approximate normal $100(1-\alpha)\%$ confidence intervals based on population size estimators are

constructed as

$$\left(\widehat{N}_{(L)}, \widehat{N}_{(U)}\right) = \left(\widehat{N} - z_{1-\alpha/2}\widehat{S.E.}(\widehat{N}), \widehat{N} + z_{1-\alpha/2}\widehat{S.E.}(\widehat{N})\right), \tag{7.9}$$

where $\widehat{S.E.}(\widehat{N}) = \sqrt{\widehat{Var}(\widehat{N})}$ is the standard error of population size estimator, $\widehat{N}_{(L)}$ denotes the lower limit and $\widehat{N}_{(U)}$ denotes the upper limit. If we choose nominal 95% confidence interval of the populations size, then

$$\left(\widehat{N}_{(L)}, \widehat{N}_{(U)}\right) = \left(\widehat{N} - 1.96 \times \widehat{S.E.}(\widehat{N}), \widehat{N} + 1.96 \times \widehat{S.E.}(\widehat{N})\right). \tag{7.10}$$

However, the appropriateness of the approximate normal distribution with associated variance and confidence interval depends on the sample data and sample size. Sometimes the confidence interval based on the approximate normal distribution does not cover the true population size. The use of resampling methods along with the robust confidence interval percentile method have been studies by (Buckland and Garthwaite, 1991; Buckland, 1984) to deal with the variance estimation and the confidence interval construction for population size estimators. Also, resample technique has been deeply studied for the multiple marking in Chapter 5. However, the three bootstrap methods based on the multinomial distribution in the context of single marking capture-recapture for variance and confidence interval estimation are deeply studied below.

Method 2 (M2): True bootstrap method

If the population size is known, the true bootstrap could be applied for estimating the variance of the interested population size estimator. The sampling distribution of capture history is assumed to be multinomial with parameters $(N, \widehat{\mathbf{p}})$. Therefore, the algorithm of the true bootstrap is proposed below:

Step 1: Capture-recapture probabilities $(\widehat{\mathbf{p}})$ are estimated by the relative frequencies as follows:

$$\widehat{\mathbf{p}} = \begin{pmatrix} \widehat{p}_{11} & \widehat{p}_{10} \\ \widehat{p}_{01} & \widehat{p}_{00} \end{pmatrix} = \begin{pmatrix} \frac{f_{11}}{N} & \frac{f_{10}}{N} \\ \\ \frac{f_{01}}{N} & \frac{f_{00}}{N} \end{pmatrix}$$

Step 2: Resample associated frequencies $(\mathbf{f}^*)$ under the multinomial distribution with parameters $(N, \widehat{\mathbf{p}})$, that is

$$\mathbf{f}^{*(\mathbf{b})} = \begin{pmatrix} f_{11}^{*(b)} & f_{10}^{*(b)} \\ f_{01}^{*(b)} & f_{00}^{*(b)} \end{pmatrix} \sim Multinomial(N, \widehat{\mathbf{p}})$$

Step 3: Calculate

(a) the CM estimator

$$\widehat{N}_{CM}^{*(b)} = \frac{(n_1^{*(b)} + 1)(n_2^{*(b)} + 1)}{m_2^{*(b)} + 1} - 1,$$

where $n_1^{*(b)} = f_{10}^{*(b)} + f_{11}^{*(b)}$, $n_2^{*(b)} = f_{01}^{*(b)} + f_{11}^{*(b)}$ and $m_2^{*(b)} = f_{11}^{*(b)}$.

(b) the CB estimator

$$\widehat{N}_{CB}^{*(b)} = n^{*(b)} + \frac{(f_1^{*(b)})^2}{4f_2^{*(b)}},$$

where $f_1^{*(b)} = f_{10}^{*(b)} + f_{01}^{*(b)}$, $f_2^{*(b)} = f_{11}^{*(b)}$ and $n^{*(b)} = f_{10}^{*(b)} + f_{01}^{*(b)} + f_{11}^{*(b)}$.

Step 4: Repeat step 2 and step 3, B times where $b \in \{1, 2, 3, ..., B\}$. We use $\mathrm{B} = 1,000$ and achieve the estimated population sizes of the CM estimator:

$$\widehat{N}_{CM}^{*(1)}, \widehat{N}_{CM}^{*(2)}, \widehat{N}_{CM}^{*(3)}, ..., \widehat{N}_{CM}^{*(1,000)},$$

and for the CB estimator as:

$$\widehat{N}_{CB}^{*(1)}, \widehat{N}_{CB}^{*(2)}, \widehat{N}_{CB}^{*(3)}, ..., \widehat{N}_{CB}^{*(1,000)}.$$

Step 5: Calculate the statistical measures of interest, they are

(a) The CM estimator

i) Mean of population size from CM estimator

$$E(\widehat{N}_{CM}) = \frac{1}{1,000} \sum_{b=1}^{1,000} \left\{ N_{CM}^{*(b)} \right\}. \tag{7.11}$$

ii) Median of population size from the CM estimator

$$(\widehat{N}_{CM}) = Median(\widehat{N}_{CM}^{*(1)}, \widehat{N}_{CM}^{*(2)}, \widehat{N}_{CM}^{*(3)}, ..., \widehat{N}_{CM}^{*(1,000)}). \tag{7.12}$$

iii) The bootstrap variance of population size from CM estimator

$$\widehat{Var}(\widehat{N}_{CM}) = \frac{1}{999} \sum_{b=1}^{1,000} \left[\widehat{N}_{CM}^{*(b)} - E(\widehat{N}_{CM}) \right]^2, \tag{7.13}$$

then, *the standard error of population size* is estimated by $\widehat{S.E.}(\widehat{N})_{CM} = \sqrt{\widehat{Var}(\widehat{N}_{CM})}$.

b) The CB estimator

i) Mean of population size from CB estimator

$$E(\widehat{N}_{CB}) = \frac{1}{1,000} \sum_{b=1}^{1,000} \left\{ N_{CB}^{*(b)} \right\} \tag{7.14}$$

ii) Median of population size from the CB estimator

$$(\widehat{N}_{CB}) \quad = \quad Median[\widehat{N}_{CB}^{*(1)}, \widehat{N}_{CB}^{*(2)}, \widehat{N}_{CB}^{*(3)}, ..., \widehat{N}_{CB}^{*(1,000)}] \tag{7.15}$$

iii) The bootstrap variance of population size from CM estimator

$$\widehat{Var}(\widehat{N}_{CB}) \quad = \quad \frac{1}{999}\sum_{b=1}^{1,000}\left[\widehat{N}_{CB}^{*(b)} - E(\widehat{N}_{CB})\right]^2, \tag{7.16}$$

therefore, *the standard error of population size* of the CB estimator can be estimated as $\widehat{S.E.}(\widehat{N})_{CB} = \sqrt{\widehat{Var}(\widehat{N}_{CB})}$.

Step 6 The bootstrap nonparametric percentile confidence intervals are calculated, the lower and upper bound of the 95% of a confidence interval can be achieved from percentiles 2.5^{th} and 97.5^{th}, $(P_{0.025}, P_{0.975})$, respectively.

Method 3 (M3): Imputed bootstap

In practice, the target population size is usually unknown and requires to be inferred. The variance of the population size estimator arises from two sources (Böhning, 2008a). That is the random variation drawing n individuals from the target population size N, and the random variation due to estimating f_{00} by using n observed units. Then, the resampling distribution is generated including the unobserved but estimated f_{00}. This approach requires an excellent estimator of $\hat{f}_{00}$ and $\widehat{N}$. Again, a multinomial with parameters $(\widehat{N}, \widehat{\mathbf{p}})$ is used. An algorithm of imputed bootstrap is given as follows:

Step 1: Construct $\widehat{N}_{CM}$ and $\widehat{N}_{CB}$, then we have $\widehat{f}_{00(CM)}$ and $\widehat{f}_{00(CB)}$

Step 2: Capture-recapture probabilities $(\widehat{\mathbf{p}})$ are estimated by the relative frequencies as follows:

(a) The CM estimator:

$$\widehat{\mathbf{p}}_{CM} = \begin{pmatrix} \widehat{p}_{11} & \widehat{p}_{10} \\ \widehat{p}_{01} & \widehat{p}_{00} \end{pmatrix} = \begin{pmatrix} \frac{f_{11}}{\widehat{N}_{CM}} & \frac{f_{10}}{\widehat{N}_{CM}} \\ \\ \frac{f_{01}}{\widehat{N}_{CM}} & \frac{\hat{f}_{00}}{\widehat{N}_{CM}} \end{pmatrix}.$$

(b) The CB estimator:

$$\widehat{\mathbf{p}}_{CB} = \begin{pmatrix} \widehat{p}_{11} & \widehat{p}_{10} \\ \widehat{p}_{01} & \widehat{p}_{00} \end{pmatrix} = \begin{pmatrix} \frac{f_{11}}{\widehat{N}_{CB}} & \frac{f_{10}}{\widehat{N}_{CB}} \\ \\ \frac{f_{01}}{\widehat{N}_{CB}} & \frac{\hat{f}_{00}}{\widehat{N}_{CB}} \end{pmatrix}.$$

Step 3: Resample associated frequencies $(\mathbf{f}^*)$ under the multinomial distribution, that is

(a) The CM estimator:

$$\mathbf{f}^{*(\mathbf{b})} = \begin{pmatrix} f_{11}^{*(b)} & f_{10}^{*(b)} \\ f_{01}^{*(b)} & \widehat{f}_{00(CM)}^{*(b)} \end{pmatrix} \sim Multinomial(\widehat{N}_{CM}, \widehat{\mathbf{p}}_{\mathbf{CM}}).$$

(b) The CB estimator:

$$\mathbf{f}^{*(\mathbf{b})} = \begin{pmatrix} f_{11}^{*(b)} & f_{10}^{*(b)} \\ f_{01}^{*(b)} & \widehat{f}_{00(CB)}^{*(b)} \end{pmatrix} \sim Multinomial(\widehat{N}_{CB}, \widehat{\mathbf{p}}_{\mathbf{CB}}).$$

Step 4: Calculating
(a) The CM estimator:

$$\widehat{N}_{CM}^{*(b)} = \frac{(n_1^{*(b)} + 1)(n_2^{*(b)} + 1)}{m_2^{*(b)} + 1} - 1,$$

where $n_1^{*(b)} = f_{10}^{*(b)} + f_{11}^{*(b)}$, $n_2^{*(b)} = f_{01}^{*(b)} + f_{11}^{*(b)}$ and $m_2^{*(b)} = f_{11}^{*(b)}$.
(b) The CB estimator:

$$\widehat{N}_{CB}^{*(b)} = n^{*(b)} + \frac{(f_1^{*(b)})^2}{4f_2^{*(b)}},$$

where $f_1^{*(b)} = f_{10}^{*(b)} + f_{01}^{*(b)}$, $f_2^{*(b)} = f_{11}^{*(b)}$ and $n^{*(b)} = f_{10}^{*(b)} + f_{01}^{*(b)} + f_{11}^{*(b)}$.

Step 5-7 are executed the same way as the true bootstrap step 4 - 6

Method 4 (M4): Reduced bootstrap

The idea of the reduced bootstrap is that the variance is estimated conditional on the observed sample size. Hence, n is treated as fixed thus the variance comes from only one source of error that is the random variation due to estimating f_{00} by using n observed units. The sampling distribution of capture history is the multinomial, conditional on the observed sample size with four parameters $(n, p_{11}, p_{10}, p_{01})$. An algorithm for the reduced bootstrap method to estimate variance of population size under the CM and the CB estimators is given as follows:

Step 1: Capture-recapture probabilities $(\widehat{\mathbf{p}})$ are estimated by the relative frequencies as follows:

$$\widehat{\mathbf{p}} = \begin{pmatrix} \widehat{p}_{11} & \widehat{p}_{10} \\ \widehat{p}_{01} & - \end{pmatrix} = \begin{pmatrix} \frac{f_{11}}{n} & \frac{f_{10}}{n} \\ \frac{f_{01}}{n} & - \end{pmatrix}$$

Step 2: Resample associated frequencies $(\mathbf{f}^*)$ under the multinomial distribution, that is

$$\mathbf{f}^{*(\mathbf{t})} = \begin{pmatrix} f_{11}^{*(b)} & f_{10}^{*(b)} \\ f_{01}^{*(b)} & - \end{pmatrix} \sim Multinomial(n, \widehat{\mathbf{p}})$$

where $n = f_{11} + f_{01} + f_{10}$.

Step 3-6: are computed the same way as the step $3-6$ of the true bootstrap.

To illustrate the various approaches of variance estimation and confidence interval construction, a simulation technique is carried out to examine the properties and behaviour of the estimation methods under different conditions.

7.4 Simulation scheme

A simulation study was conducted to illustrate the performance of the variance and confidence interval estimation methods. The standard error and coverage probability of the confidence interval is used for evaluating and comparing the normal approximation of variance or formula method and three bootstrap methods in this section. Since the capture history can be represented as a multinomial likelihood, each simulation experiment is generated as $\mathbf{f} \sim Multinomial(N, \mathbf{p})$, where $\mathbf{f}$ and $\mathbf{p}$ denote a vector of frequencies, $(f_{00}, f_{01}, f_{10}, f_{11})$ and a vector of capture probabilities, $(p_{00}, p_{01}, p_{10}, p_{11})$ respectively. The population sizes were chosen as $N = 100, 250, 500, 1,000, 5,000$ and $10,000$. The frequencies are obtained using the function `rmultinom()` in `R`.

One of the assumptions of the single marking method is independence between sources, however, this assumption is frequently violated in reality. Therefore, the focus in the simulation study is on positive dependence, which occurs more often in capture-recapture data than negative dependence (Brenner, 1995) particularly in infectious disease capture-recapture data (Van Hest et al., 2008). Moreover, positive dependence results in underestimation of the population size. Simulation scenarios were set under assumptions of both independence and positive dependence between sources.

1. Independence between sources

Suppose that the single marking capture-recapture data are homogeneous. It can applied the odds ratio can be used as a tool for checking the independence between two sources, that is (OR), $\text{OR} = \dfrac{p_{00}p_{11}}{p_{10}p_{01}} = 1$. The simulation scenarios for independent cases are set up as Table 7.2, following Böhning and Van der Heijden (2015). The marginal probabilities for individuals identified in source one and source two are defined as $p_1 = p_{10} + p_{11}$ and $p_2 = p_{01} + p_{11}$, respectively. Six simulation scenarios are defined, and in the first and second the marginal capture probabilities are the same for the two sources. The third and the fourth are set such that the marginal probability of the second source is larger than the first one. For the last two cases, the marginal probability of the second one is smaller than the first one. The difference between the marginal probabilities, $D_p = p_1 - p_2$ is investigated, and it is expected that the D_p might influence some of the properties of the variance estimations.

Table 7.2: Design of the simulation study with capture probability $p_{00}, p_{10}, p_{01}, p_{11}$ when two sources are independent

Scenario	p_1	p_2	p_{00}	p_{10}	p_{01}	p_{11}	$D_p = p_1 - p_2$
1A	0.5	0.5	0.25	0.25	0.25	0.25	0.00
2A	0.3	0.3	0.49	0.21	0.21	0.09	0.00
3A	0.5	0.6	0.20	0.20	0.30	0.30	−0.10
4A	0.3	0.35	0.455	0.195	0.245	0.105	−0.05
5A	0.5	0.3	0.35	0.35	0.15	0.15	0.20
6A	0.3	0.1	0.63	0.27	0.07	0.03	0.20

2. Dependence between sources

The assumption of independence between two sources commonly fails, for example, in surveillance studies sources such as hospital registration and laboratory registration could be associated. This violated assumptions lead to biased estimates of population size, too small variance and too narrow confidence interval (Amstrup et al., 2010). Since the situation of completely independent sources is rare in reality, in the simulation study the effect of positive dependence is explored. The degree of dependence is qualified by the odds ratio, and it is expected that the odds ratio may be a key criterion to choose the suitable variance estimation and the confidence interval approaches. The design of dependent sources with capture probabilities with positive odds ratio was provided in Brittain and Böhning (2009), given as Table 7.3. This same design will be used within the simulations.

Table 7.3: Design of the simulation study with capture probability $p_{00}, p_{10}, p_{01}, p_{11}$ when two sources are dependent

Scenario	p_1	p_2	p_{00}	p_{10}	p_{01}	p_{11}	OR
1B	0.50	0.50	0.25	0.25	0.25	0.25	1.00
2B	0.45	0.50	0.30	0.20	0.25	0.25	1.50
3B	0.40	0.50	0.30	0.15	0.30	0.25	1.67
4B	0.45	0.45	0.35	0.20	0.20	0.25	2.18
5B	0.45	0.50	0.35	0.15	0.20	0.30	3.50
6B	0.45	0.55	0.35	0.10	0.20	0.35	6.12

7.5 Simulation results for investigating the performance of variance estimation approaches

In this section, variance and standard error of population size estimators are evaluated. Both independence and dependence data scenarios are simulated using 5,000 data sets

(T = 5,000). The true variance can be calculated as

$$Var(\widehat{N})_{True} = \frac{1}{T-1}\sum_{t=1}^{T}\left(\widehat{N}_{(t)} - E(\widehat{N})\right)^2, \tag{7.17}$$

and the standard error is then given as $S.E.(\widehat{N})_{True} = \sqrt{\left\{Var(\widehat{N})_{True}\right\}}$. Furthermore, to study behaviour and performance of variance estimation approaches, the expected values of estimated standard errors are calculated from the simulation study. $E\left\{S.E.(\widehat{N})_{M1}\right\}$ is defined as the expected value of the normal approximation method or the variance formulas, according to (7.3) and (7.8). Its expected values can be computed by the root of an expected variance, that is

$$E\left(\widehat{S.E.}(\widehat{N})_{M1}\right) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{Var}(\widehat{N}_{(M1,t)})\right]}.$$

Additionally, $E\left\{\widehat{S.E.}(\widehat{N})_{M2}\right\}$, $E\left\{\widehat{S.E.}(\widehat{N})_{M3}\right\}$ and $E\left\{\widehat{S.E.}(\widehat{N})_{M4}\right\}$ were defined as expected values of the approximate standard error of population size from the true bootstrap, imputed bootstrap and reduced bootstrap, respectively. *The ratio of standard error estimation* which is defined as the standard error estimation from each approach divided by the true standard error, $\frac{E[\widehat{S.E.}(\widehat{N})]}{S.E.(\widehat{N})_{True}}$, provided a means for comparing the behaviours and the performances of variance estimations. The reference value for this ratio is one.

7.5.1 Simulation results of variance estimation when two sources are independent

1) Variance estimations of the CM estimator when two sources are independent

The first set of results to be evaluated from the simulation study are the variance estimation methods for the CM estimator. Table 7.4 represents standard errors from 5,000 replication runs. For more convenient presentation the ratio of standard error are also presented in Figure 7.1. Overall, the simulation results demonstrate that the standard error from the formula method is significantly underestimated when population sizes are small.

For more details, scenario 1A and 2A are designed with equal marginal probabilities of two sources ($D_p = 0$). It is found that the approximate variance using the normal approximation formula of the CM estimator is slightly better than the true and imputed bootstraps among the small population size. However, the reduced bootstrap seems to be the best choice when the marginal probabilities reduce. Next, we consider the scenario 3A and 4A, these scenarios are designed such that the marginal probability

for the first source is smaller than the second one ($D_p < 0$). The true bootstrap and the imputed bootstrap perform the best for large marginal probabilities of both sources whereas the reduced bootstrap remains the best method for small population size and small marginal probabilities of two sources. The last two cases (scenario 5A and 6A) describe the situations that the probability of the first source is greater than the second one ($D_p > 0$). The simulation results suggest that the reduced bootstrap is suitable for estimating the variance of the CM estimator when the population size is very small ($N = 100$) and the marginal probabilities are high. On the other hand, if the marginal probabilities from both sources are small as the case of 6A, the standard error from the variance formula of the CM estimator performs better than all of the bootstrap approaches.

Table 7.4: Comparison of the standard errors of the four estimators with the true standard error for the CM estimator when the two sources are independent

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$ M1	M2	M3	M4
1A: $p_1 = 0.5$, $p_2 = 0.5$					
100	10.166	9.839	10.831	10.765	9.043
250	16.186	15.690	16.268	16.219	13.390
500	22.459	22.258	22.654	22.628	18.574
1,000	31.099	31.562	31.837	31.807	26.048
5,000	71.039	70.720	70.812	70.806	57.840
10,000	97.299	100.008	100.046	100.031	81.707
2A: $p_1 = 0.3$, $p_2 = 0.3$					
100	25.060	22.349	28.960	31.513	26.861
250	38.853	36.324	41.110	41.045	37.840
500	52.193	51.881	54.920	54.906	50.208
1,000	74.359	73.574	75.662	75.641	68.896
5,000	164.471	164.717	165.586	165.596	150.388
10,000	232.741	233.419	234.047	233.945	212.429
3A: $p_1 = 0.5$, $p_2 = 0.6$					
100	8.453	8.062	8.696	8.645	7.029
250	13.277	12.802	13.184	13.141	10.498
500	18.607	18.211	18.473	18.450	14.668
1,000	25.687	25.746	25.933	25.899	20.542
5,000	58.080	57.708	57.792	57.765	45.701
10,000	81.859	81.681	81.737	81.725	64.631
4A: $p_1 = 0.3$, $p_2 = 0.35$					
100	23.146	20.355	26.078	27.507	24.101
250	34.321	32.511	36.016	35.946	32.754
500	47.333	46.512	48.789	48.749	44.143
1,000	67.157	65.688	67.248	67.225	60.631
5,000	148.058	147.156	147.744	147.768	132.880
10,000	203.747	207.898	208.297	208.253	187.254
5A: $p_1 = 0.5$, $p_2 = 0.3$					
100	16.307	14.950	17.917	17.969	16.117
250	24.421	23.896	25.492	25.450	22.598
500	34.504	33.929	34.998	34.967	30.861
1,000	48.598	48.177	48.937	48.880	42.985
5,000	109.239	107.997	108.200	108.321	95.064
10,000	152.515	152.822	152.938	153.027	134.246
6A: $p_1 = 0.3$, $p_2 = 0.1$					
100	33.892	33.250	31.862	49.303	26.918
250	79.352	69.868	89.164	105.282	85.589
500	107.080	101.085	124.698	125.881	120.737
1,000	152.350	145.052	159.647	159.786	153.944
5,000	329.721	323.621	329.387	329.384	316.038
10,000	458.479	457.725	461.746	461.698	442.698

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

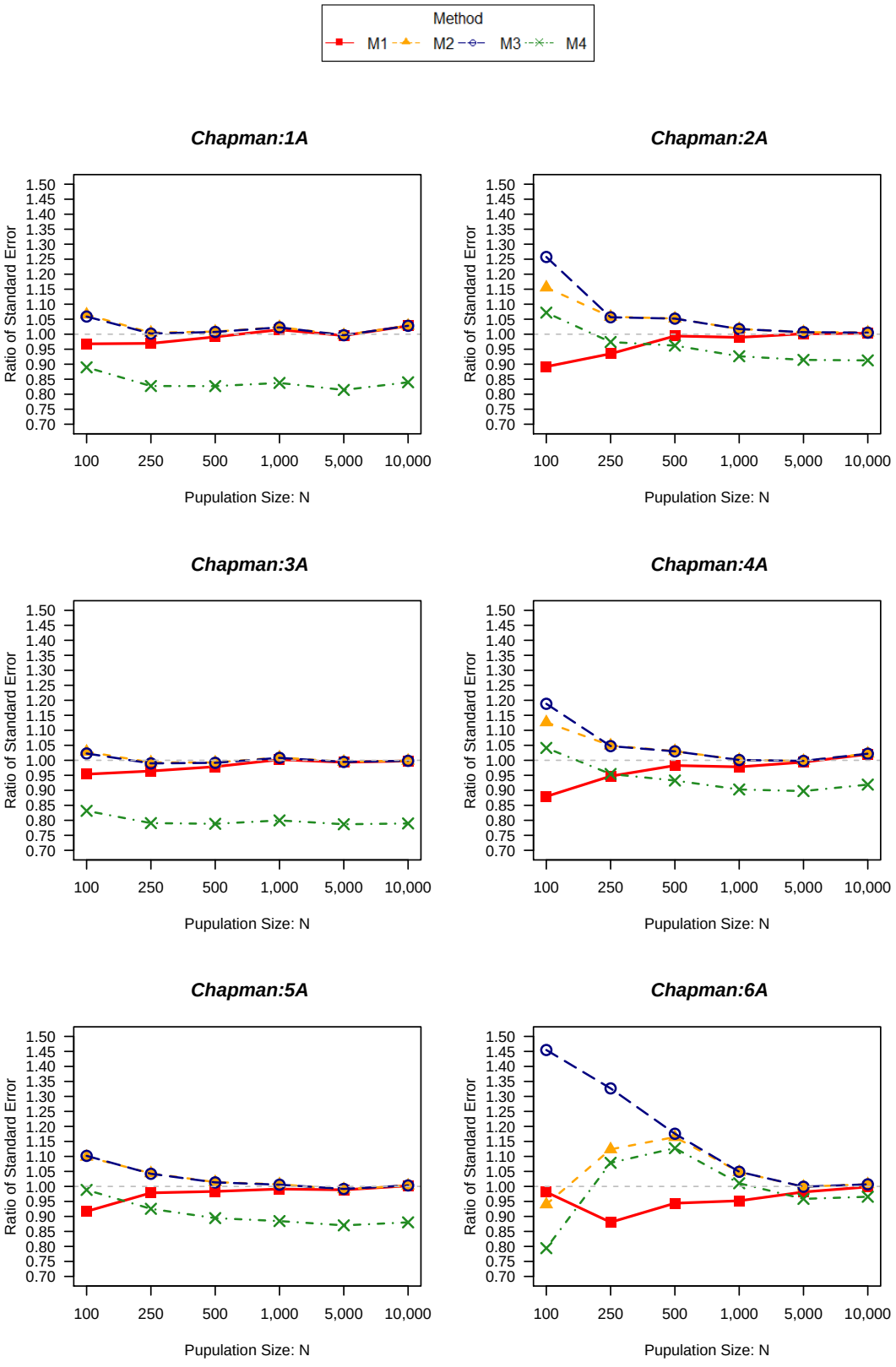

Figure 7.1: Ratio of standard errors of estimations, using the CM estimator where M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

2) Variance estimation of the CB estimator when two sources are independent

In the second part of the simulation study, properties of the variance approximation methods under independence of the two sources for the CB estimator are explored. The results are shown in Table 7.5 and in Figure 7.2. The simulation results suggest that the approximate variance formula method is the most suitable approach to estimate the variance of population size of the CB estimator for almost all cases. However, for the case 6A, which has very small marginal probabilities for both sources as well as a $D_p > 0$, the true and imputed bootstraps perform better than the variance formula of the CB estimator.

It can be seen that decreasing the marginal probability for each source is a key factor in the size of the bias of variance estimations of population size. This is particularly noticeable for the bootstrap resampling approaches for small sample size.

Table 7.5: Comparison of standard errors of four estimators and true standard error of the CB estimator when two sources are independently

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$			
		M1	M2	M3	M4
1A: $p_1 = 0.5$, $p_2 = 0.5$					
100	10.719	10.623	11.694	11.724	10.031
250	16.506	16.162	16.683	16.709	13.877
500	22.662	22.587	22.926	22.944	18.896
1,000	31.233	31.792	32.021	32.039	26.267
5,000	71.109	70.823	70.892	70.940	57.935
10,000	97.329	100.080	100.102	100.168	81.773
2A: $p_1 = 0.3$, $p_2 = 0.3$					
100	33.821	29.017	45.790	45.882	44.494
250	42.023	39.704	45.716	45.765	42.765
500	54.093	54.149	57.327	57.405	52.807
1,000	75.617	75.140	77.181	77.157	70.544
5,000	165.056	165.399	166.209	166.150	151.064
10,000	233.175	233.903	234.487	234.338	212.908
3A: $p_1 = 0.5$, $p_2 = 0.6$					
100	9.056	8.802	9.411	9.495	7.851
250	13.784	13.449	13.731	13.824	11.120
500	19.122	18.900	19.035	19.159	15.294
1,000	26.280	26.548	26.565	26.737	21.228
5,000	59.291	59.215	58.931	59.288	46.905
10,000	83.688	83.759	83.290	83.772	66.284
4A: $p_1 = 0.3$, $p_2 = 0.35$					
100	29.980	25.556	38.663	38.735	37.308
250	36.905	35.371	39.501	39.523	36.474
500	49.243	48.699	51.014	51.118	46.516
1,000	68.659	67.514	68.983	69.041	62.459
5,000	149.752	149.079	149.502	149.729	134.645
10,000	205.908	210.249	210.477	210.456	189.390
5A: $p_1 = 0.5$, $p_2 = 0.3$					
100	21.311	19.412	24.831	25.072	23.225
250	28.366	28.342	29.957	30.276	27.170
500	39.369	39.194	39.929	40.432	35.835
1,000	54.807	54.998	55.182	55.855	49.185
5,000	122.118	122.010	120.737	122.324	107.417
10,000	169.422	172.432	170.457	172.573	151.521
6A: $p_1 = 0.3$, $p_2 = 0.1$					
100	96.438	108.243	90.574	92.779	82.575
250	180.160	144.558	229.020	229.425	226.088
500	177.304	167.551	229.122	229.888	225.119
1,000	231.268	221.809	246.468	247.597	239.924
5,000	473.067	468.171	474.334	476.671	457.922
10,000	657.812	657.629	660.091	663.537	636.458

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

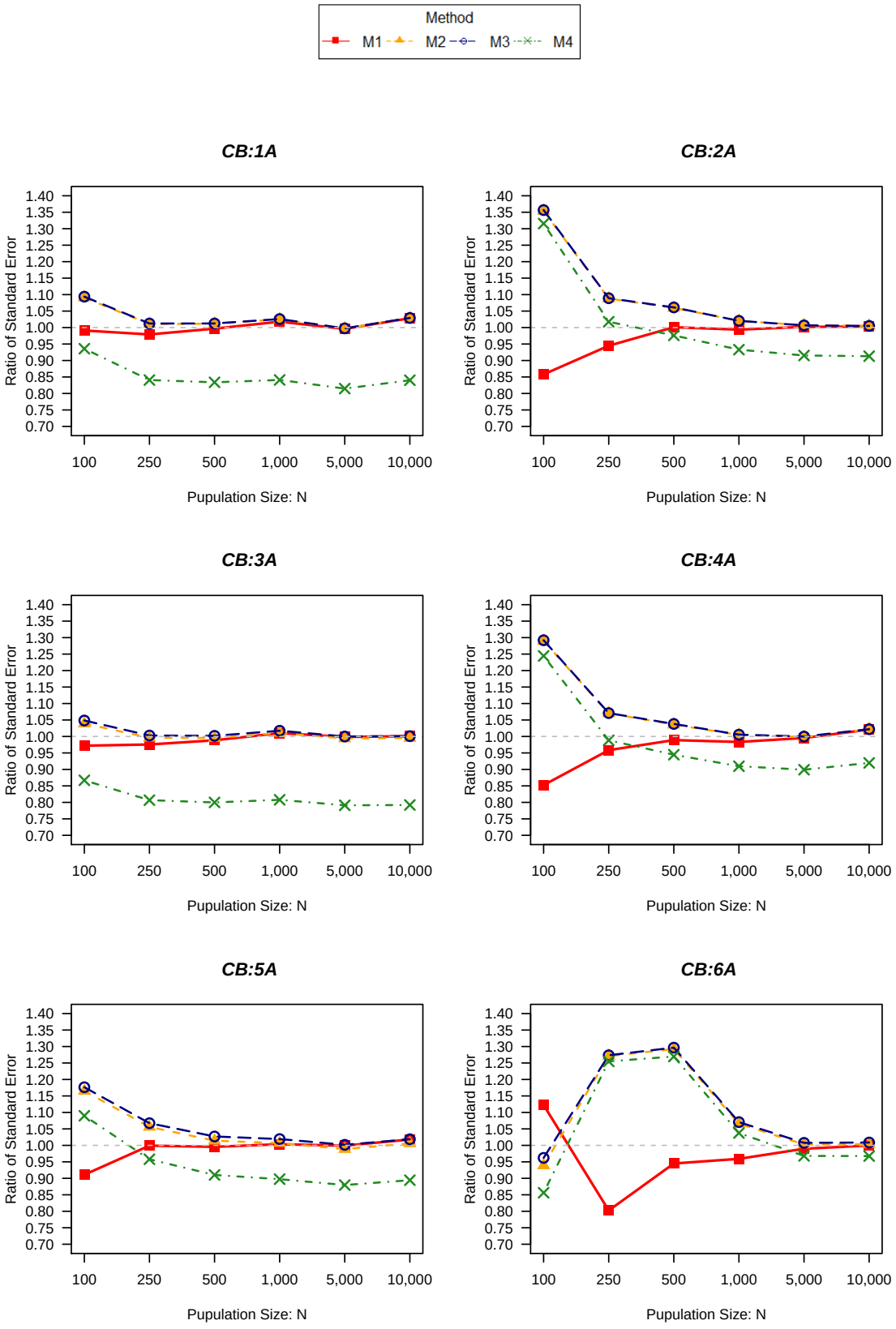

Figure 7.2: Ratio of standard errors of estimators over true standard error for the CB estimator where M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

7.5.2 Simulation results of variance estimations when two sources are dependent

This dependency can be measured using the odds ratio. The CB estimator is widely used to estimate population size with two sources assumed independent. Recently, Brittain and Böhning (2009) have proved that the CB estimator is the more robust than the CM estimator when the assumption of independence is in doubt. In the case of positive dependence they found that both the CM and CB estimators provided an underestimation of population size. However, the CB estimator gave the better results in terms of less bias and larger variance leading to a good coverage rate of confidence interval when it was compared with the CM estimator. In this part of the simulation study, we attempt to compare the performance of different variance estimation methods when two sources are dependent.

1) Variance estimation of the CM estimator when two sources are dependent
Results from the CM estimator with two dependent sources are summarised in Table 7.6 and in Figure 7.3. Overall, the estimated variance which uses the variance formula of the CM estimator is underestimated and strongly underestimated when the degree of dependence increases. The true bootstrap shows most robustness and most accuracy for variance estimation with the CM estimator under various dependencies between sources. It can be seen that the ratio of standard errors from the true bootstrap method are close to one. The imputed bootstrap method might be the best choice for estimating variance of the CM estimator in reality when $N \leq 1,000$, although this method tends to dramatically underestimate when the rate of dependence increases. If N is large, the approximate normal estimation is suggested instead other methods. The reduced bootstrap is not recommended for any two-source dependency situation as it seriously underestimates with respect to the true variance. It might lead to a too narrow confidence interval and low level of coverage of the associated confidence interval.

Table 7.6: Comparison of standard errors of four estimators and true standard error of the CM when two sources are dependent

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{SE}(\widehat{N})\right\}$ M1	M2	M3	M4
1B: $p_1 = 0.5$, $p_2 = 0.5$, Odds Ratio = 1.00					
100	10.285	9.825	10.800	10.767	9.030
250	15.959	15.776	16.347	16.320	13.476
500	22.372	22.257	22.670	22.633	18.574
1,000	31.971	31.575	31.844	31.819	26.046
5,000	70.505	70.688	70.777	70.787	57.825
10,000	99.235	100.051	100.133	100.154	81.772
2B: $p_1 = 0.45$, $p_2 = 0.5$, Odds Ratio = 1.50					
100	9.164	8.298	9.624	9.074	7.480
250	14.360	13.363	14.631	13.814	11.209
500	20.221	18.891	20.364	19.210	15.479
1,000	28.504	26.848	28.705	27.058	21.770
5,000	64.585	59.974	63.715	60.075	48.195
10,000	91.105	84.874	90.083	84.930	68.132
3B: $p_1 = 0.40$, $p_2 = 0.50$, Odds Ratio = 1.67					
100	8.754	7.826	9.198	8.553	7.046
250	13.326	12.463	13.904	12.886	10.435
500	19.289	17.779	19.502	18.082	14.570
1,000	27.383	25.144	27.383	25.351	20.378
5,000	59.877	56.195	60.788	56.240	45.148
10,000	84.520	79.527	85.984	79.532	63.812
4B: $p_1 = 0.45$, $p_2 = 0.45$, Odds Ratio = 2.18					
100	8.433	7.108	8.752	7.770	6.285
250	12.919	11.357	13.300	11.739	9.324
500	18.596	16.062	18.558	16.316	12.879
1,000	25.788	22.718	26.070	22.906	18.014
5,000	57.150	50.864	57.999	50.967	40.004
10,000	82.436	72.006	81.994	72.060	56.534
5B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio = 3.50					
100	6.837	4.964	6.899	5.337	4.034
250	10.461	7.827	10.587	8.046	5.973
500	14.937	11.095	14.860	11.247	8.293
1,000	20.899	15.766	20.975	15.873	11.679
5,000	46.218	35.356	46.789	35.398	26.004
10,000	66.607	49.959	66.165	49.974	36.699
6B: $p_1 = 0.45$, $p_2 = 0.55$, Odds Ratio = 6.12					
100	5.747	3.287	5.757	3.531	2.447
250	9.050	5.328	9.041	5.480	3.757
500	12.490	7.533	12.719	7.640	5.208
1,000	17.673	10.720	17.974	10.793	7.353
5,000	40.266	24.044	40.190	24.084	16.375
10,000	57.878	33.969	56.817	33.991	23.075

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

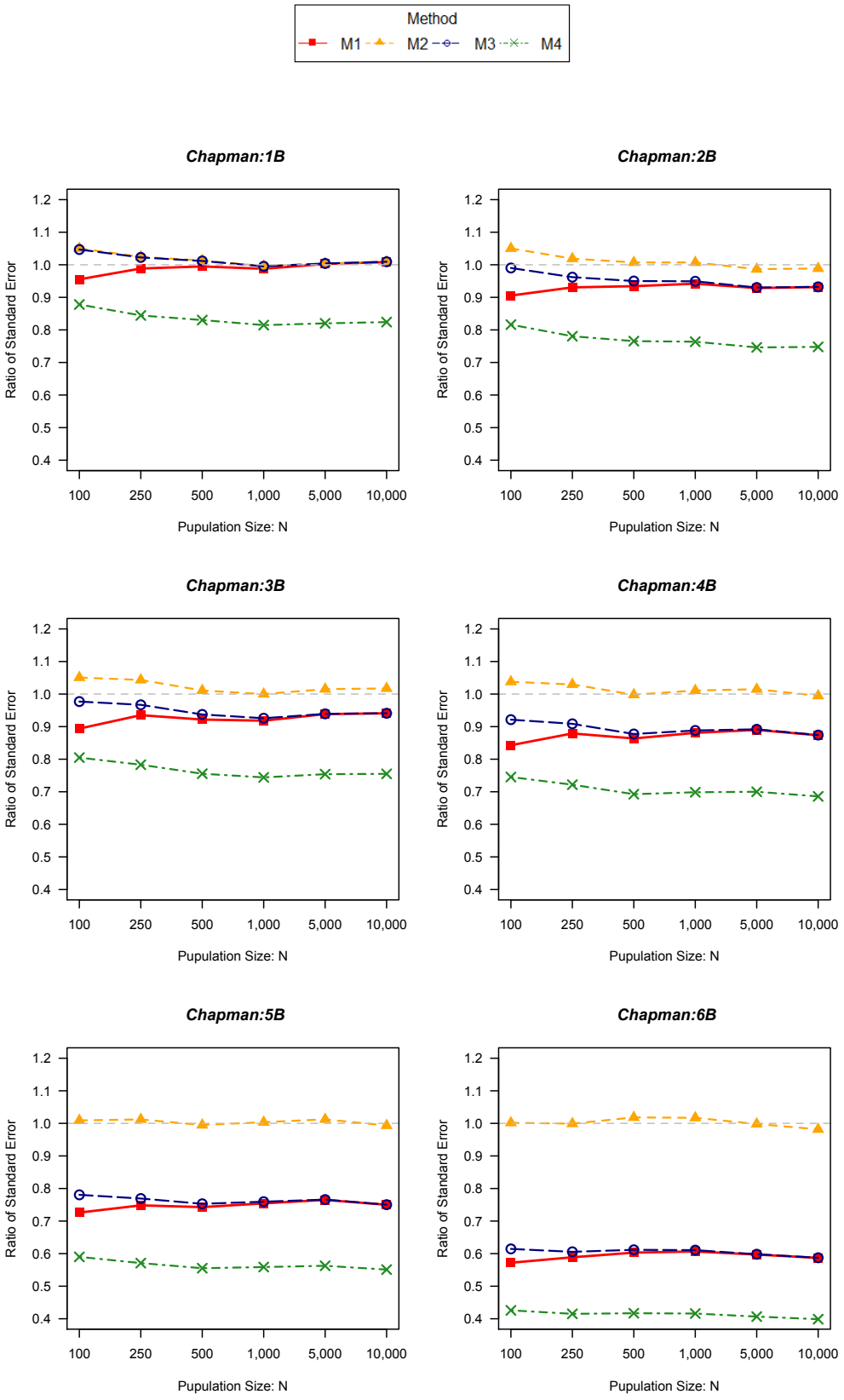

Figure 7.3: Ratio of standard errors of estimators over true standard error using CM estimator where M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

2) Variance estimation of the CB estimator when two sources are dependent
For the CB estimator, according to Table 7.7 and Figure 7.4, the approximate variance formula is affected by the violation of the assumption of independence between two sources. Overall, the true bootstrap shows the best accuracy of standard error for the dependent cases (2B - 6B). It can be seen that the ratio of standard errors from the true bootstrap are close to one.

However, the population size is usually unknown in reality and requires to be estimated. It is found that the imputed bootstrap is the most appropriate method for estimating variance of population size for the CB estimator, particular when the population sizes are no more than $1,000$. If the population sizes are greater than $1,000$, the variance formula of the CB estimator can be used to estimate the variance and standard error of the CB estimator. This is the same results as the CM estimator in the pervious part that the reduced method represents the dramatical underestimation of standard errors for all situations. Therefore, it is not recommended for approximating variance for the CB estimator when two sources are dependent.

Interestingly, the simulation results also show the variance estimations of the CB estimator are greater than the CM estimator for all situations under the condition of dependence between two sources. This result agrees with the original study presented by Brittain and Böhning (2009).

Table 7.7: Comparison of standard errors of four estimators with true standard error of the CB estimator when the two sources are dependent

N	$S.E.(\widehat{N})_{True}$	$E\left\{\widehat{S.E.}(\widehat{N})\right\}$ M1	M2	M3	M4
1B: $p_1 = 0.5$, $p_2 = 0.5$, Odds Ratio = 1.00					
100	10.847	10.605	11.655	11.702	10.021
250	16.283	16.251	16.767	16.791	13.970
500	22.603	22.586	22.942	22.943	18.897
1,000	32.100	31.806	32.030	32.062	26.268
5,000	70.557	70.791	70.858	70.869	57.920
10,000	99.269	100.123	100.189	100.185	81.839
2B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio = 1.50					
100	9.662	9.026	10.380	9.937	8.387
250	14.718	13.872	15.067	14.343	11.734
500	20.590	19.319	20.722	19.645	15.906
1,000	28.861	27.251	29.044	27.481	22.162
5,000	65.034	60.521	64.145	60.599	48.661
10,000	91.569	85.583	90.630	85.622	68.712
3B: $p_1 = 0.40$, $p_2 = 0.50$, Odds Ratio = 1.67					
100	9.709	9.079	10.418	10.004	8.438
250	14.181	13.791	14.991	14.258	11.650
500	20.589	19.390	20.783	19.720	15.966
1,000	28.988	27.201	29.003	27.432	22.108
5,000	63.204	60.449	64.066	60.568	48.597
10,000	88.978	85.497	90.563	85.577	68.612
4B: $p_1 = 0.45$, $p_2 = 0.45$, Odds Ratio = 2.18					
100	8.841	7.674	9.320	8.456	7.014
250	13.166	11.699	13.588	12.102	9.708
500	18.760	16.298	18.752	16.571	13.143
1,000	25.901	22.886	26.202	23.066	18.192
5,000	57.186	50.938	58.049	51.015	40.072
10,000	82.469	72.059	82.030	72.095	56.583
5B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio = 3.50					
100	7.046	5.361	7.214	5.786	4.485
250	10.614	8.125	10.778	8.342	6.247
500	15.085	11.368	15.023	11.510	8.518
1,000	21.044	16.050	21.146	16.141	11.903
5,000	46.470	35.819	47.076	35.858	26.351
10,000	66.943	50.578	66.549	50.601	37.158
2B: $p_1 = 0.45$, $p_2 = 0.55$,Odds Ratio = 6.12					
100	5.959	3.718	5.999	3.983	2.833
250	9.329	5.816	9.289	5.964	4.129
500	12.772	8.126	13.016	8.230	5.639
1,000	18.042	11.494	18.359	11.564	7.899
5,000	41.110	25.665	40.991	25.681	17.486
10,000	58.925	36.233	57.934	36.250	24.614

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

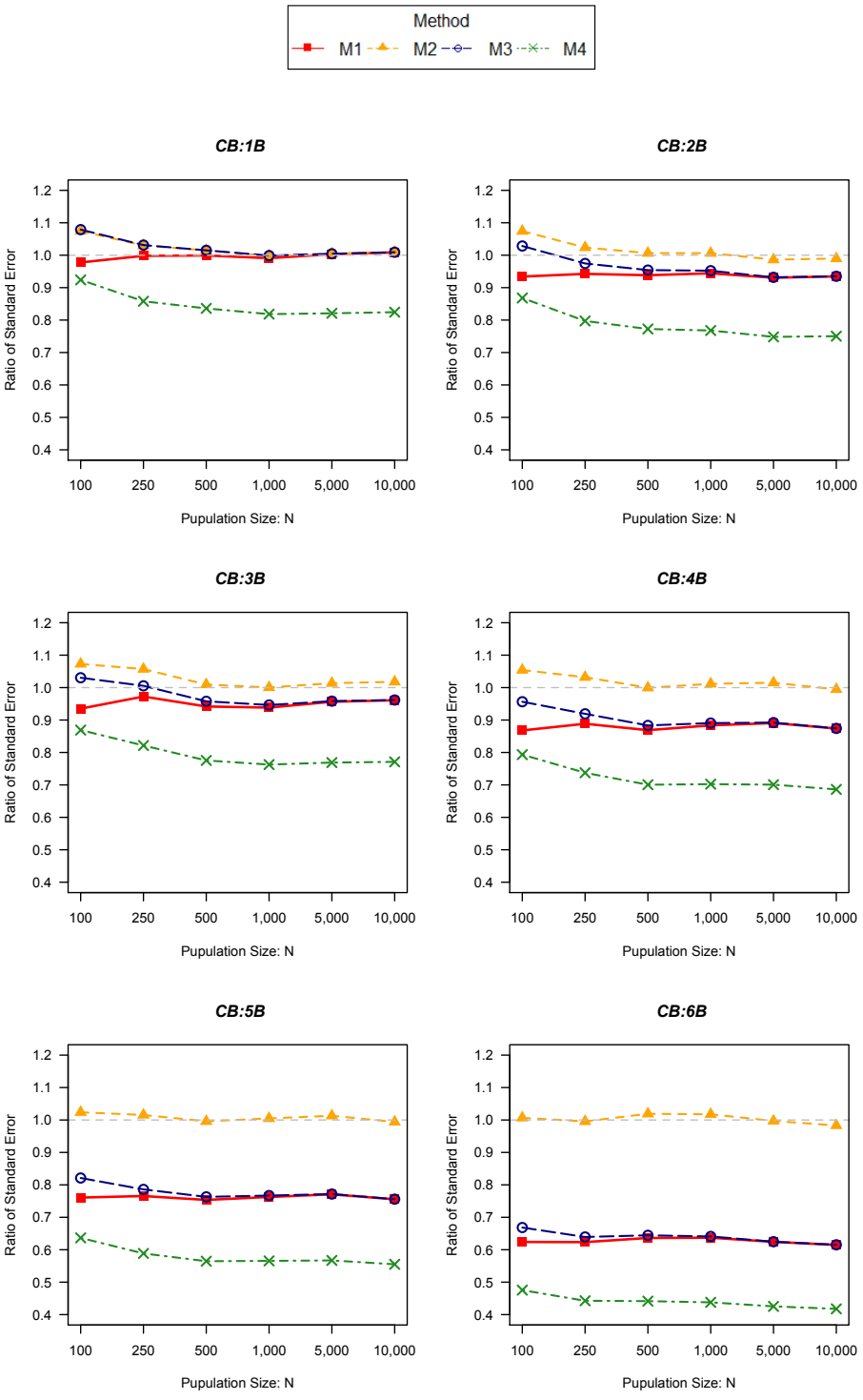

Figure 7.4: Ratio of standard errors of four estimators with the CB estimator where M1: Formula,M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

7.6 Comparison of confidence intervals for estimating population size

The approximate normal confidence interval is constructed on the basis of a large population size. Chao (1987) pointed out that the sampling distribution under the assumption of asymptotic normality is usually skewed, and the coverage probability is potentially unsatisfactory. It might be lower than the nominal level for small marginal probability or small population size (Köse et al., 2014). Consequently, the bootstrap approaches to estimate variance of population size and to construct confidence intervals are compared with the traditional normal approximation method in this section.

The performance of a confidence interval method can be evaluated using coverage probability and comparing with the nominal coverage probability. The nominal coverage probability is often set at 0.95, hence the simulation study is also investigating 95% two-sides confidence intervals for N. It is desirable for a well performing confidence interval that the coverage probability is close to the nominal level of 95% and that the average length of the confidence interval is narrow. Additionally, the short length will provide a more precise result. The (actual) coverage probability (Cov) and the average length (AL) are defined as:

Coverage probability for confidence interval:

$$Cov = \frac{\sum_{t=1}^{T} A_{(t)}}{T} \times 100\%, \tag{7.18}$$

where $A_{(t)} = 1$ if the confidence interval covers the true value N.

Average length (AL):

$$AL = \frac{\sum_{t=1}^{T} \left(\widehat{N}_{(k')} - \widehat{N}_{(k)} \right)}{T}, \tag{7.19}$$

when $\widehat{N}_{(k)}$ is the 2.5^{th} percentile of order $\widehat{N}_t$ and $\widehat{N}_{(k)'}$ is the 97.5^{th} percentile of the (estimated) distribution of $\widehat{N}$. We set T = 5,000 replications.

7.6.1 Simulation results for investigating confidence interval when two sources are independent

1) Investigating confidence intervals of the CM estimator when the two sources are independent

Table 7.8 and Figure 7.5 represent the coverage probabilities. Confidence intervals estimated using the true bootstrap are close to those from the imputed bootstrap method

for most scenarios. Additionally, both methods perform the best when constructing confidence intervals of the CM estimator, due to the fact that their coverage probabilities are close to the nominal level (95%), when compared with the variance formula and reduced bootstrap methods. The simulation results also confirm that the variance formula method provides lower coverage probabilities than the nominal level for small population size and tends to have even lower coverage as the marginal probabilities decrease.

A comparison of the means and medians in Table 7.8 shows that the distributions of the population size estimator in 2A, 4A and 6A are slightly positively skewed, especially for the small population. This results indicate that the variance formula method might be not appropriate for constructing confidence intervals for medium and small population sizes. It can be seen that the variance formula method require the population size to be 5,000 or above.

The reduce bootstrap provided all of the coverage probabilities lower than the nominal level. Therefore, it is not a good candidate for constructing the confidence interval of the CM estimator.

Table 7.8: Comparison of four methods of confidence interval construction for the CM estimator when two sources are independent

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	M1		M2		M3		M4	
			Cov	AL	*Cov*	AL	*Cov*	AL	*Cov*	AL
1A: $p_1 = 0.5$, $p_2 = 0.5$										
100	99.9	99.0	91.08	38.57	93.84	42.18	93.80	41.99	87.60	34.70
250	250.0	249.0	93.26	61.51	94.44	63.55	94.12	63.40	87.54	51.99
500	500.0	499.0	94.14	87.25	94.80	88.51	94.50	88.36	88.88	72.35
1000	999.6	999.0	94.72	123.72	95.00	124.29	94.94	124.16	89.24	101.57
5000	5,001.8	5,000.0	94.76	277.21	94.74	276.36	94.86	276.39	88.82	225.69
10000	10,000.7	10,001.0	95.44	392.02	95.44	390.47	95.30	390.35	89.76	318.77
2A: $p_1 = 0.3$, $p_2 = 0.3$										
100	99.6	95.0	88.02	87.63	94.16	110.42	94.20	116.08	90.68	100.87
250	249.3	244.0	91.26	142.40	94.10	158.83	94.08	158.49	91.18	145.08
500	500.2	495.0	94.04	203.37	95.46	213.87	95.50	213.73	92.62	194.83
1000	999.9	994.0	94.86	288.41	95.56	295.28	95.48	295.02	93.26	268.22
5000	4,998.0	4,993.0	94.92	645.67	95.04	646.07	94.94	645.88	92.48	586.40
10000	10,002.5	9,994.0	94.94	914.98	94.92	913.19	94.70	912.65	92.28	828.52
3A: $p_1 = 0.5$, $p_2 = 0.6$										
100	100.1	99.0	91.56	31.61	93.68	33.97	93.34	33.83	85.72	27.06
250	249.9	249.0	92.76	50.19	93.68	51.53	93.68	51.39	86.20	40.79
500	500.0	499.0	93.68	71.37	94.00	72.17	93.68	72.09	87.28	57.15
1000	999.6	999.0	94.50	100.91	94.90	101.26	94.90	101.12	87.68	80.10
5000	4,999.9	4,999.0	94.90	226.21	94.94	225.66	94.86	225.50	87.16	178.37
10000	10,000.4	10,000.0	94.88	320.18	94.90	319.08	94.84	319.04	87.74	252.19
4A: $p_1 = 0.3$, $p_2 = 0.35$										
100	100.2	96.0	88.38	79.82	93.62	99.34	93.48	102.06	90.08	90.35
250	249.7	246.0	91.80	127.45	94.40	139.66	94.42	139.44	91.30	126.00
500	501.1	499.0	93.64	182.34	95.18	190.13	95.02	190.08	91.88	171.46
1000	1,000.5	995.0	94.10	257.50	94.56	262.42	94.42	262.45	91.46	236.16
5000	5,001.3	4,998.5	94.48	576.84	94.40	576.74	94.50	576.47	91.38	518.43
10000	9,995.7	9,989.0	95.50	814.95	95.28	812.83	95.18	812.74	92.70	730.64
5A: $p_1 = 0.5$, $p_2 = 0.3$										
100	100.2	98.0	90.68	58.62	94.44	68.71	94.04	68.69	90.36	60.86
250	250.0	248.0	93.22	93.67	94.56	99.41	94.44	99.21	91.02	87.59
500	499.9	498.0	93.72	133.00	94.98	136.67	94.60	136.45	91.04	120.07
1000	999.6	997.0	94.12	188.86	94.46	190.99	94.46	190.79	90.80	167.54
5000	5,001.0	5,000.0	94.28	423.35	94.54	422.48	94.40	422.84	91.18	370.92
10000	10,003.8	10,001.0	94.80	599.05	94.66	597.01	94.74	597.07	91.10	523.66
6A: $p_1 = 0.3$, $p_2 = 0.1$										
100	90.2	83.0	76.70	58.62	90.48	122.72	92.60	187.71	86.88	101.08
250	249.7	233.0	86.84	93.67	94.24	340.51	94.02	385.58	92.76	326.88
500	500.9	482.0	90.96	133.00	94.38	473.02	94.76	474.06	93.24	455.71
1000	1,004.5	987.0	93.18	188.86	94.88	617.99	95.00	618.67	93.78	594.66
5000	4,999.7	4,980.5	94.42	423.35	94.62	1,284.78	94.48	1,284.56	93.52	1,232.21
10000	9,997.5	9,987.0	94.72	599.05	94.68	1,802.90	94.74	1,801.87	93.40	1,726.91

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

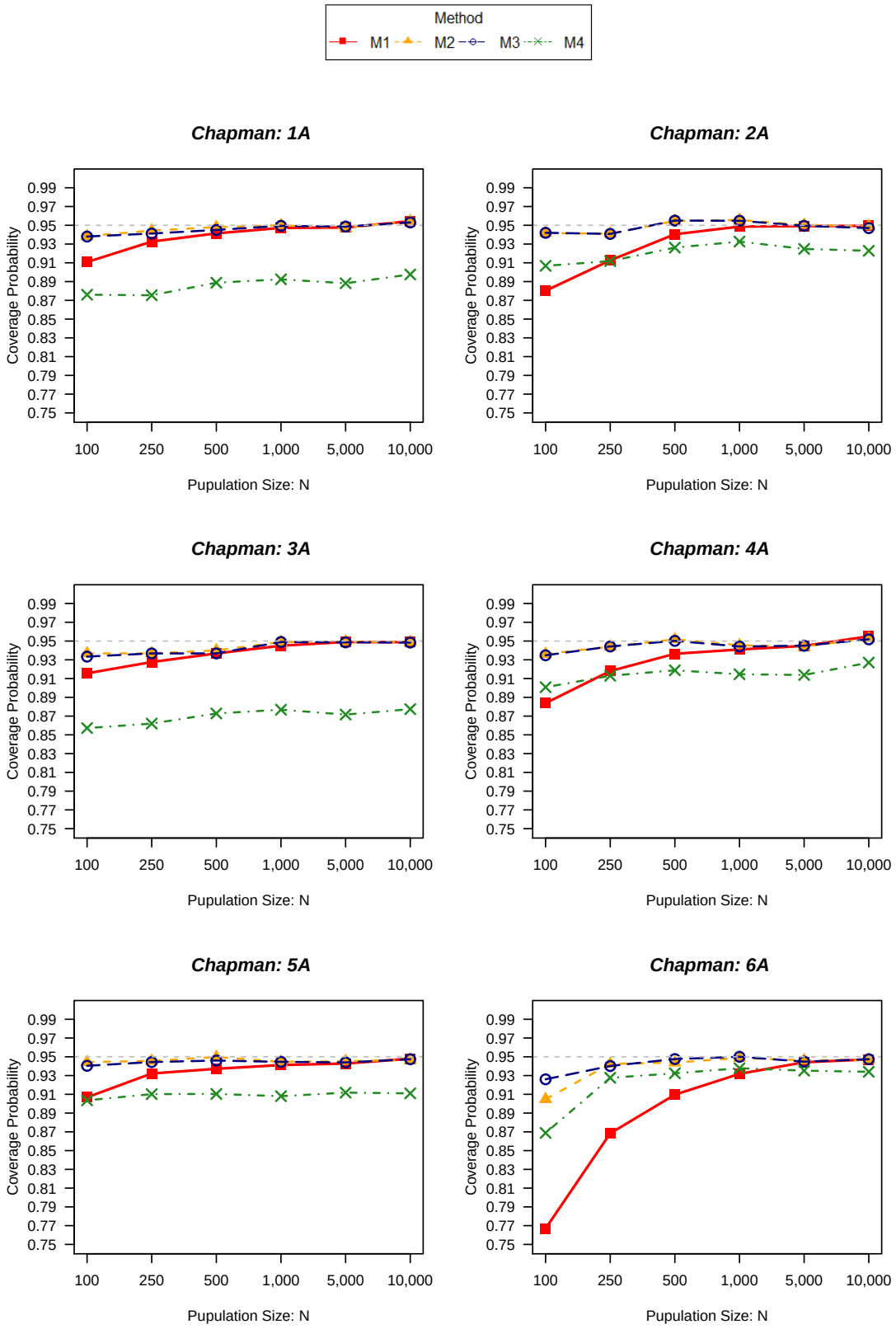

Figure 7.5: Coverage probability of confidence interval methods using the CM estimator when two sources are independent; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

2) Investigating confidence interval of the CB estimator when the two sources are independent

The model structure plays an important role in inference statistical modelling. It is known that the CB estimator for capture-recapture data with finite member of trapping occasions requires a binomial mixture distribution assumption for the recapture count. If this basic assumption is ignored, the population size can be biased. The simulation results show that the estimated population size from the CB estimator slightly overestimate for almost all cases, particular when $D_p \neq 0$.

To account for the coverage probability of confidence interval, conditioning on the independence between two sources as in Table 7.9, the true bootstrap provides the most acceptable coverage probabilities for all most circumstances, especially for the small and medium population sizes ($N \leqslant 1,000$). However, the true bootstrap is useless in reality since the number of population size is not know in advance. In light of this the most appropriate method for constructing confidence intervals for the CB estimator when the marginal probabilities are equal ($D_p = 0$) is the imputed bootstrap method. Look at the scenario 3A, 4A, 5A and 6A. The simulation has actual coverage probabilities for the bootstrap methods less than the nominal level if the population sizes increase. This means that the CB estimator is not an asymptotic unbiased estimator under these circumstances. The variance estimations are efficient when the observed data approach the population sizes. This leads to confidence intervals that do not cover the true value N and coverage probabilities that reduce to zero. The reason for this might be that the model is strongly misspecified.

The simulation results also show that the CB estimator should not be used under an assumption of independence between two sources when the marginal probabilities are not equal as 2A, 3A, 5A, and 6A. Since if $p_1 \neq p_2$, $\widehat{f}_{0,CB} \neq \dfrac{f_1^2}{4f_2}$

Table 7.9: Comparison of four methods of confidence interval construction for the CB estimator when the two sources are independent

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	M1		M2		M3		M4	
			Cov	AL	*Cov*	AL	*Cov*	AL	*Cov*	AL
1A: $p_1 = 0.5$, $p_2 = 0.5$										
100	102.1	101.0	87.60	34.70	94.76	45.42	94.94	45.62	87.02	38.29
250	252.2	251.0	87.54	51.99	94.54	65.17	94.66	65.28	87.06	53.85
500	502.2	501.0	88.88	72.35	94.88	89.58	94.86	89.60	88.42	73.56
1000	1,001.8	1,001.0	89.24	101.57	95.08	125.01	95.30	125.04	88.98	102.38
5000	5,004.0	5,002.0	88.82	225.69	94.78	276.68	94.84	276.87	88.62	226.08
10000	10,002.9	10,003.0	89.76	318.77	95.46	390.69	95.18	390.82	89.68	319.05
2A: $p_1 = 0.3$, $p_2 = 0.3$										
100	108.7	102.0	90.68	100.87	93.20	165.52	94.64	165.82	90.52	158.01
250	257.1	251.0	91.18	145.08	94.28	175.53	94.30	175.81	90.78	162.69
500	507.6	502.0	92.62	194.83	95.40	223.16	95.42	223.44	92.54	59.53
1000	1,007.2	1,001.0	93.26	268.22	95.56	301.21	95.58	301.20	92.36	274.51
5000	5,005.1	5,000.0	92.48	586.40	95.14	648.52	95.00	648.47	92.52	588.96
10000	10,009.6	10,002.0	92.28	828.52	94.98	914.92	94.82	914.30	92.26	830.33
3A: $p_1 = 0.5$, $p_2 = 0.6$										
100	102.7	102.0	85.72	27.06	95.54	36.71	94.42	37.08	84.00	30.07
250	253.6	253.0	86.20	40.79	94.74	53.63	93.86	54.02	84.78	43.12
500	505.9	505.0	87.28	57.15	94.80	74.37	93.96	74.82	85.24	59.53
1,000	1,009.6	1,009.0	87.68	80.10	94.52	103.74	93.80	104.45	84.74	82.78
5,000	5,043.1	5,043.0	87.16	178.37	89.70	230.05	89.04	231.40	77.40	183.06
10,000	10,085.2	10,083.0	87.74	252.19	83.22	325.05	82.58	326.95	69.02	258.65
4A: $p_1 = 0.3$, $p_2 = 0.35$										
100	108.2	103.0	90.08	90.35	93.66	139.27	94.42	139.30	89.38	131.78
250	257.6	253.0	91.30	126.00	94.82	152.71	94.34	152.79	91.02	139.45
500	510.3	508.0	91.88	171.46	95.56	198.77	95.12	199.29	91.56	180.43
1000	1,012.6	1,007.0	91.46	236.16	95.28	269.18	94.02	269.50	90.80	243.07
5000	5,037.0	5,035.0	91.38	518.43	94.64	583.66	94.00	584.36	90.64	525.13
10000	10,061.3	10,054.5	92.70	730.64	95.02	821.46	94.62	821.26	91.30	738.93
5A: $p_1 = 0.5$, $p_2 = 0.3$										
100	111.7	108.0	90.36	60.86	97.96	91.92	92.00	92.93	83.42	84.09
250	270.9	268.0	91.02	87.59	96.06	116.63	89.28	117.88	80.56	104.93
500	537.3	535.0	91.04	120.07	91.18	155.75	84.18	157.79	74.22	139.25
1000	1,070.7	1,068.0	90.80	167.54	79.82	215.31	73.96	217.96	61.54	191.48
5000	5,338.7	5,337.0	91.18	370.92	18.24	471.34	16.04	477.15	10.02	419.12
10000	10,675.3	10,671.0	91.10	523.66	1.50	665.46	1.34	673.61	0.52	591.10
6A: $p_1 = 0.3$, $p_2 = 0.1$										
100	167.1	137.0	86.88	101.08	97.52	337.78	90.28	349.21	82.40	263.96
250	384.8	343.0	92.76	326.88	99.54	844.39	77.20	845.68	70.16	831.41
500	709.4	674.0	93.24	455.71	99.28	831.72	64.80	833.84	58.04	808.83
1000	1,377.0	1,348.5	93.78	594.66	73.40	950.81	42.08	955.91	35.94	922.02
5000	6,701.6	6,671.0	93.52	1,232.21	0.74	1,849.85	0.30	1,857.94	0.20	1,785.37
10000	13,361.2	13,339.0	93.40	1,726.91	0.00	2,576.40	0.00	2,588.87	0.00	2,481.56

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

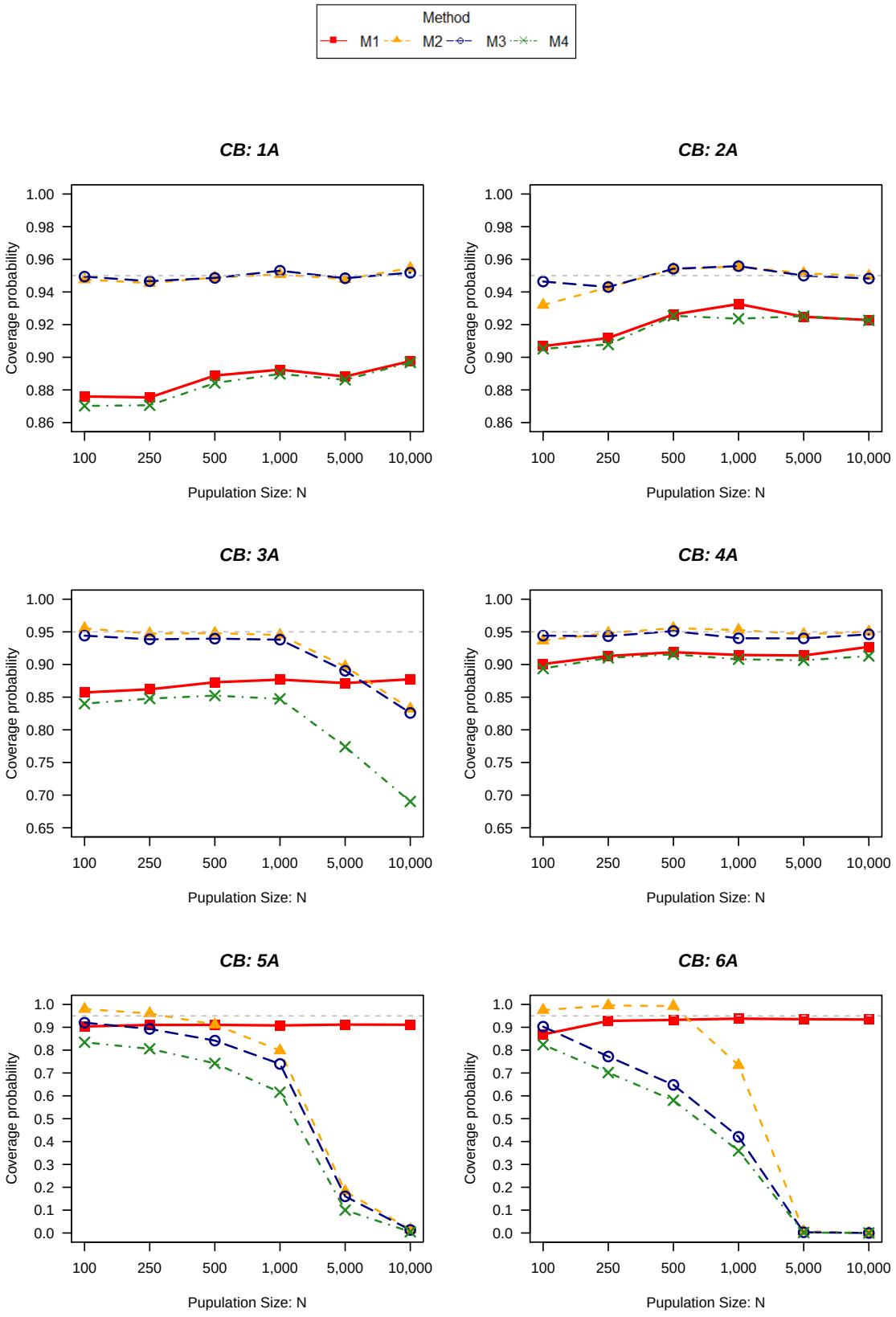

Figure 7.6: Coverage probability of confidence interval methods using the CB estimator; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

7.6.2 Simulation results for investigating confidence intervals when two sources are dependent

In this section, the performance of four confidence interval methods, constructed with the CM estimator and the CB estimator are compared. The dependencies are quantified in terms of the odds ratio.

1) Investigating confidence intervals of the CM estimator when two sources are dependent

The model assumption of independence remains the key factor, increasing dependence between the two sources is not only results in underestimating the number of population for the CM estimator, but also the coverage probability is lower than the nominal level (see Table 7.10 and Figure 7.7). Overall, the true bootstrap methods perform the best for every scenario since its larger standard error leads to wider average length. If the population size is unknown, the simulation results suggest that the imputed bootstrap is more suitable than the normal approximation of variance formula based on confidence interval and the reduced bootstrap for constructing the confidence interval of population size.

Another salient feature in the case of dependence between the two sources is that when the population size is increased the coverage probabilities tend to decrease. The actual reason for this is that the CM estimator is asymptotically biased when two sources are dependent, and the variance of estimation decreases. Therefore, the coverage of probability goes to zero.

Table 7.10: Comparison of the performance of four confidence interval methods for the CM estimaor when two sources are dependent

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	M1		M2		M3		M4	
			Cov	AL	*Cov*	AL	*Cov*	AL	*Cov*	AL
1B: $p_1 = 0.50$, $p_2 = 0.50$, Odds Ratio=1.00										
100	100.4	99.0	92.84	38.51	94.28	42.12	94.18	42.01	87.42	34.67
250	250.8	250.0	93.68	61.83	94.44	63.88	94.36	63.75	87.64	52.33
500	500.1	499.0	94.32	87.24	94.60	88.54	94.48	88.41	88.64	72.36
1,000	1,000.5	999.0	94.58	123.77	94.50	124.37	94.52	124.27	88.52	101.56
5,000	5,000.4	4,999.0	95.14	277.09	95.22	276.29	95.24	276.25	89.00	225.60
10,000	10,001.8	10,001.0	95.20	392.19	95.26	390.83	95.32	391.12	89.56	319.11
2B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio=1.50										
100	90.2	89.0	65.98	32.51	79.46	37.56	74.80	35.45	68.24	28.69
250	225.6	225.0	51.38	52.36	63.02	57.17	59.16	54.02	50.12	43.53
500	450.4	450.0	29.24	74.05	36.92	79.56	34.22	75.02	25.60	60.29
1,000	900.5	900.0	7.82	105.24	10.56	112.07	9.82	105.60	5.78	84.90
5,000	4,500.3	4,500.0	0.00	235.10	0.00	248.67	0.00	234.42	0.00	188.01
10,000	9,002.1	9,001.0	0.00	332.70	0.00	351.64	0.00	331.32	0.00	265.92
3B: $p_1 = 0.40$, $p_2 = 0.50$, Odds Ratio=1.67										
100	88.5	88.0	58.84	30.69	73.10	35.91	68.38	33.39	62.34	27.08
250	220.4	220.0	35.56	48.86	47.22	54.34	43.24	50.40	34.82	40.53
500	441.2	440.0	14.94	69.69	20.02	76.17	18.46	70.65	12.60	56.76
1,000	880.7	880.0	1.88	98.58	2.58	106.94	2.40	98.98	1.34	79.44
5,000	4,399.9	4,399.0	0.00	220.29	0.00	237.17	0.00	219.44	0.00	176.14
10,000	8,799.9	8,800.0	0.00	311.75	0.00	335.48	0.00	310.50	0.00	249.14
4B: $p_1 = 0.45$, $p_2 = 0.45$, Odds Ratio=2.18										
100	81.7	81.0	31.60	27.84	48.24	34.20	43.06	30.36	36.00	24.12
250	203.4	203.0	7.64	44.50	12.64	51.98	11.02	45.91	7.84	36.17
500	405.9	405.0	0.58	62.96	1.02	72.49	0.84	63.75	0.40	50.12
1,000	810.6	810.0	0.00	89.05	0.02	101.77	0.02	89.41	0.00	70.20
5,000	4,049.5	4,049.0	0.00	199.39	0.00	226.37	0.00	198.87	0.00	156.16
10,000	8,100.6	8,099.0	0.00	282.27	0.00	319.98	0.00	281.30	0.00	220.65
5B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio=3.50										
100	75.6	75.0	6.00	19.44	12.50	26.98	10.02	20.93	7.18	15.49
250	187.9	188.0	0.12	30.68	0.20	41.37	0.14	31.48	0.10	23.14
500	375.1	375.0	0.00	43.50	0.00	58.05	0.00	43.96	0.00	32.26
1,000	750.5	750.0	0.00	61.81	0.00	81.90	0.00	61.99	0.00	45.51
5,000	3,750.0	3,750.0	0.00	138.60	0.00	182.56	0.00	138.18	0.00	101.44
10,000	7,498.6	7,498.0	0.00	195.84	0.00	258.24	0.00	195.09	0.00	143.20
6B: $p_1 = 0.45$, $p_2 = 0.55$, Odds Ratio=6.12										
100	71.1	71.0	0.14	12.89	0.58	22.46	0.44	13.84	0.22	9.34
250	177.4	177.0	0.00	20.89	0.00	35.29	0.00	21.44	0.00	14.55
500	354.1	354.0	0.00	29.54	0.00	49.64	0.00	29.88	0.00	20.23
1,000	707.7	708.0	0.00	42.02	0.00	70.15	0.00	42.16	0.00	28.64
5,000	3,537.2	3,536.0	0.00	94.27	0.00	156.81	0.00	94.01	0.00	63.87
10,000	7,072.5	7,072.0	0.00	133.15	0.00	221.71	0.00	132.62	0.00	90.04

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

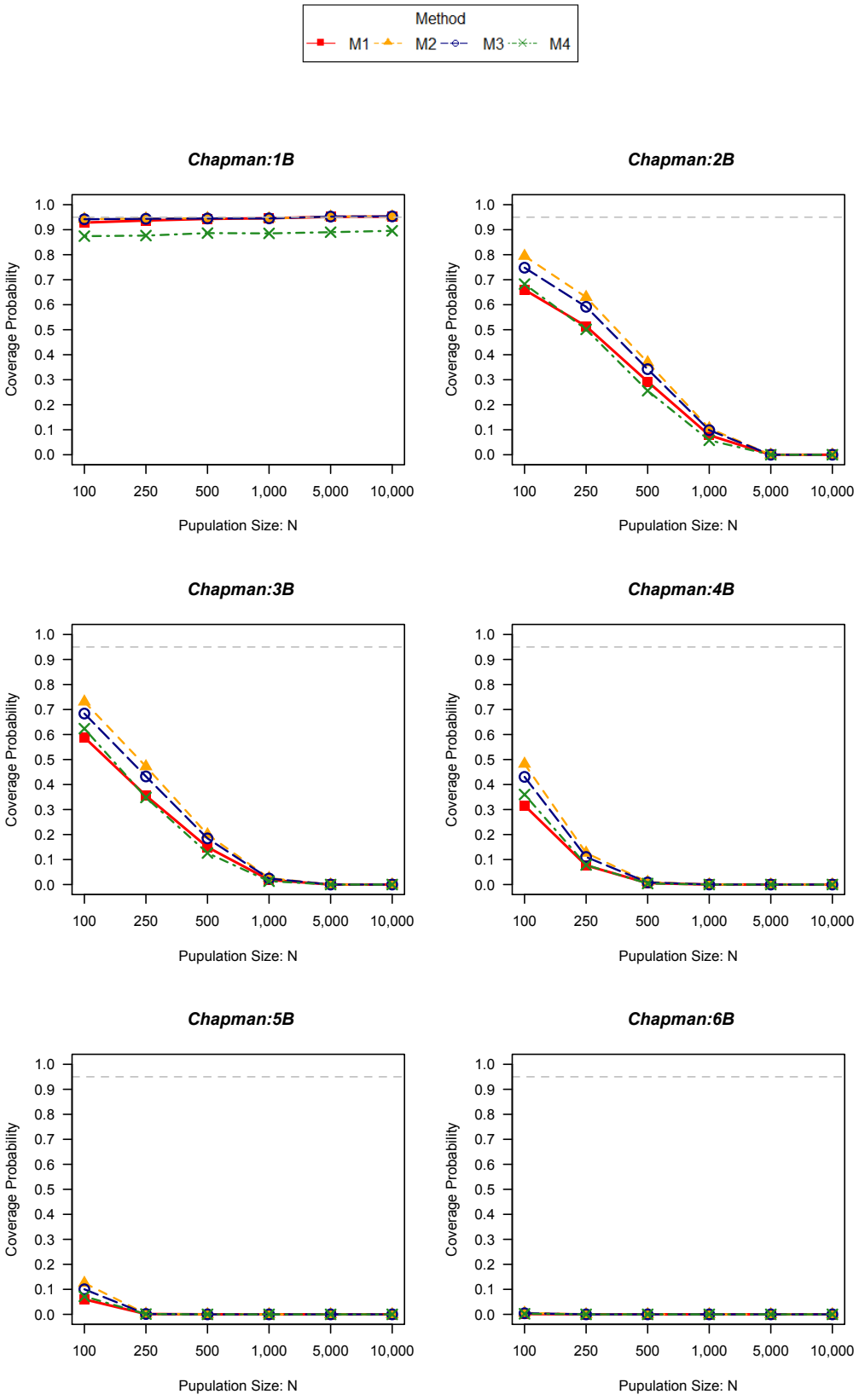

Figure 7.7: Coverage probability for the four methods using the CB estimator when two sources are independent; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

2) Investigating confidence intervals of the CB estimator when the two sources are dependent

The CB estimator is expected to be a more suitable for constructing a confidence interval than the CM estimator due to the fact that it might be the more accurate population size estimator with larger standard error in the case of violated assumption of independence. To select the best way for constructing the confidence interval of population size using the CB estimator, Table 7.11 and Figure 7.8 show that the confidence intervals achieved by the imputed bootstrap are more accurate than the other methods. A possible reason is that the CB estimator can adjust for dependency of sources, leading to less bias in population size estimation.

Interestingly, the CB estimator is more accurate than the CM estimator with respect to the population size, and the average lengths of the CB estimator are wider than the CM estimators for dependent cases. As a results in the CB estimator provides the better performance of coverage probabilities than the CM estimator.

Table 7.11: Comparison of the four methods of confidence interval construction for the CB estimator for single marking when the two sources are dependent

N	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	M1		M2		M3		M4	
			Cov	AL	*Cov*	AL	*Cov*	AL	*Cov*	AL
1B: $p_1 = 0.50$, $p_2 = 0.50$, Odds Ratio=1.00										
100	102.2	101.0	87.42	34.67	95.02	45.35	94.84	45.50	86.86	38.29
250	252.5	251.0	87.64	52.33	94.70	65.50	94.84	65.63	87.56	54.21
500	501.7	501.0	88.64	72.36	94.50	89.60	94.90	89.57	88.52	73.60
1,000	1,002.2	1,001.0	88.52	101.56	94.78	125.11	94.60	125.21	88.46	102.40
5,000	5,002.1	5,001.0	89.00	225.60	95.32	276.61	95.22	276.71	88.94	225.94
10,000	10,003.5	10,003.0	89.56	319.11	95.34	391.05	95.06	390.80	89.60	319.37
2B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio=1.50										
100	92.0	91.0	68.24	28.69	73.80	40.45	82.54	38.72	76.24	31.98
250	227.7	227.0	50.12	43.53	57.72	58.88	66.32	56.08	56.94	45.47
500	453.1	452.0	25.60	60.29	33.76	80.96	40.44	76.72	30.64	62.09
1,000	904.3	904.0	5.78	84.90	10.42	113.36	13.04	107.26	7.56	86.36
5,000	4,514.2	4,513.0	0.00	188.01	0.00	250.33	0.00	236.55	0.00	189.87
10,000	9,028.3	9,027.0	0.00	265.92	0.00	353.76	0.00	334.32	0.00	268.17
3B: $p_1 = 0.40$, $p_2 = 0.50$, Odds Ratio=1.67										
100	92.2	91.0	62.34	27.08	74.82	40.57	83.04	39.00	76.94	32.18
250	227.4	227.0	34.82	40.53	56.96	58.60	65.16	55.70	56.00	45.16
500	453.9	453.0	12.60	56.76	35.72	81.17	41.88	77.02	31.90	62.09
1,000	904.4	904.0	1.34	79.44	9.94	113.23	12.38	107.13	7.72	86.12
5,000	4,513.5	4,513.0	0.00	176.14	0.00	250.01	0.00	236.47	0.00	189.61
10,000	9,026.4	9,027.0	0.00	249.14	0.00	353.33	0.00	333.85	0.00	267.84
4B: $p_1 = 0.45$, $p_2 = 0.45$, Odds Ratio=2.18										
100	82.9	83.0	36.00	24.12	39.10	36.35	52.50	32.96	44.72	26.75
250	204.5	204.0	7.84	36.17	9.48	53.10	14.00	47.33	9.80	37.59
500	407.0	406.0	0.40	50.12	0.70	73.24	1.16	64.76	0.66	51.08
1,000	811.7	811.0	0.00	70.20	0.00	102.26	0.02	90.02	0.00	70.89
5,000	4,050.5	4,050.5	0.00	156.16	0.00	226.57	0.00	199.15	0.00	156.47
10,000	8,101.7	8,100.0	0.00	220.65	0.00	320.11	0.00	281.38	0.00	220.81
5B: $p_1 = 0.45$, $p_2 = 0.50$, Odds Ratio=3.50										
100	76.6	76.0	7.18	15.49	9.42	28.20	15.68	22.65	11.58	17.17
250	189.1	189.0	0.10	23.14	0.14	42.11	0.20	32.67	0.14	24.19
500	376.8	376.0	0.00	32.26	0.00	58.69	0.00	44.95	0.00	33.11
1,000	753.2	753.0	0.00	45.51	0.00	82.57	0.00	63.04	0.00	46.37
5,000	3,761.1	3,761.0	0.00	101.44	0.00	183.67	0.00	139.95	0.00	102.79
10,000	7,520.0	7,520.0	0.00	143.20	0.00	259.69	0.00	197.49	0.00	145.0
6B: $p_1 = 0.45$, $p_2 = 0.55$, Odds Ratio=6.12										
100	72.3	72.0	0.22	9.34	0.60	23.44	1.44	15.63	0.80	10.82
250	179.6	179.0	0.00	14.55	0.00	36.27	0.00	23.34	0.00	15.97
500	358.1	358.0	0.00	20.23	0.00	50.80	0.00	32.15	0.00	21.90
1,000	715.2	715.0	0.00	28.64	0.00	71.65	0.00	45.18	0.00	30.78
5,000	3,573.4	3,572.5	0.00	63.87	0.00	159.94	0.00	100.29	0.00	68.24
10,000	7,144.3	7,145.0	0.00	90.04	0.00	226.01	0.00	141.46	0.00	96.05

M1: Formula, M2: True bootstrap
M3: Imputed bootstrap, M4: Reduced bootstrap

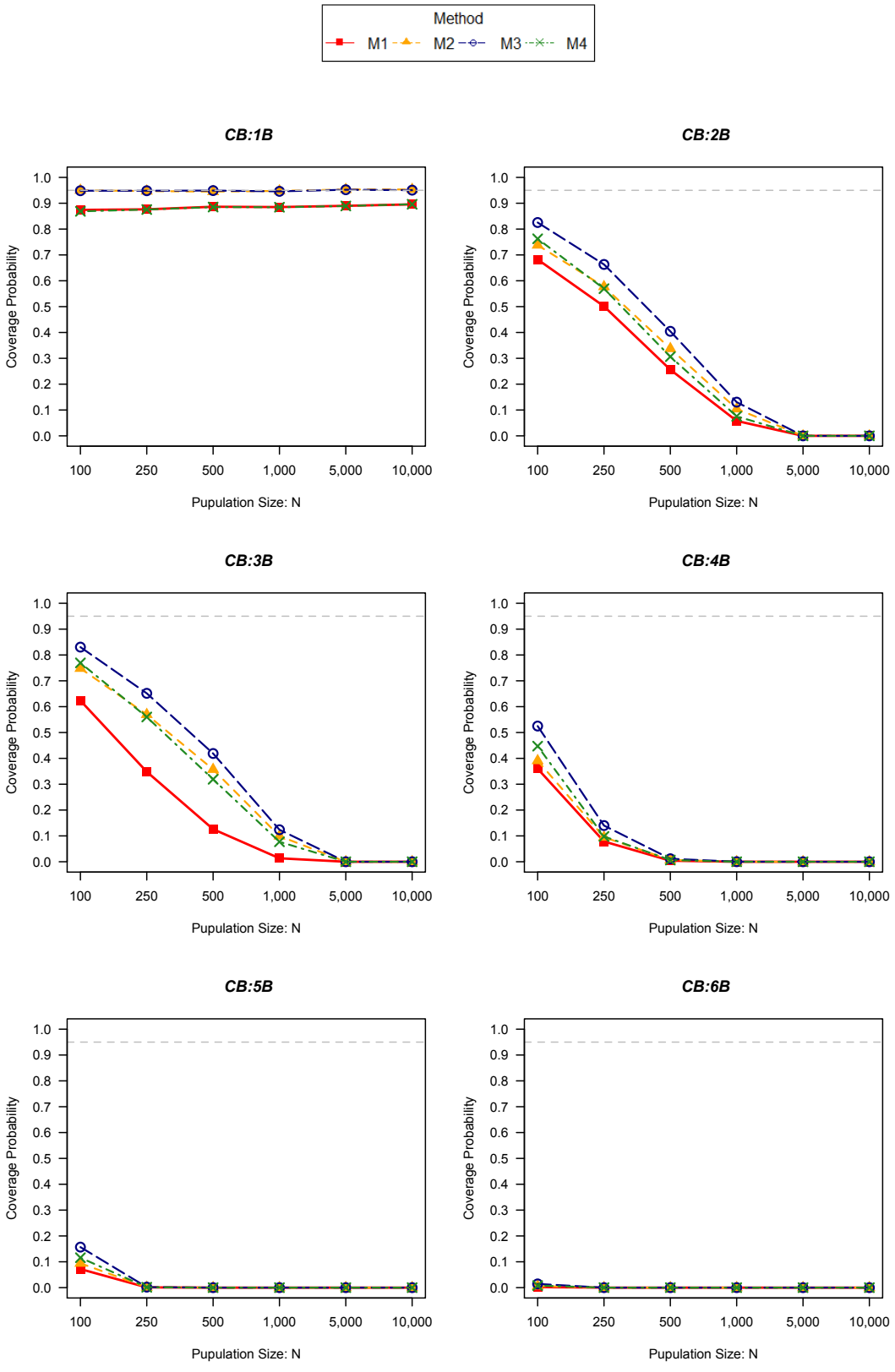

Figure 7.8: Coverage probability using CB estimator for four methods using the CB estimator; M1: Formula, M2: True bootstrap, M3: Imputed bootstrap, M4: Reduced bootstrap

7.7 Real data examples

In this section, the variance approximation approaches in four real data sets are compared. The cancer notification data in Saaland, Germany is taken as an example with known N. This is followed by the traffic death rates data in Ethiopia from June 2012 to May 2013. The third example is the homeless death rate in France data between 2008 and 2012. The last example is the number of Legionella cases in Wallon, Belgium in 2012. The population size estimators, standard errors, the 95% of confidence interval and the length of confidence intervals are calculated for the CM and the CB estimators as follows:

7.7.1 Patients with breast cancer in Saarland, Germany data

The first empirical example is the breast cancer registration in Saarland, Germany in 1970, 1975, 1980, and 1989. The study data have been previously analysed by Brenner (1995). The count distribution is given in Table 7.12

Table 7.12: The number of patients with breast cancer in Germany, according to the notifications by clinicians and death certificates in 1970, 1975, 1980, and 1989

	Clinician's report		
Death certificate	Yes	No	Total
Yes	492	124	616
No	722	307	1,029
Total	1,214	431	1,645

This case study, 1,645 patients with breast cancer were reported to the population based-centre registry of Saalrand, Germany by pathologists. The other two sources of notifications were provided by clinicians and death certificates. Interestingly, 492 patients who died were recorded in both clinician and death certificate notifications, and 722 patients and reported by clinicians but did not appear in death certificate records. 307 patients were not reported by the clinicians or death certificate notifications, but appeared in the population based-centre register. The odds ratio is utilised for a simple testing the assumption of homogeneity between two sources. We found that $OR = \dfrac{(492 \times 307)}{(722 \times 124)} = 1.69$, with a 95% confidence interval (1.33 – 2.14). The odds ratio is greater than one which indicates a positive dependence between clinician reports and death certificates. Consequently, we consider the CB estimator might be more appropriate than the CM estimator.

Typically, an unknown number of unobserved patients would be missed, however, in this case study we know that the number of unobserved patients with breast cancers

is exactly 307. We truncated the f_{00} and estimated the target population with using the CB and the CM estimators. According to Table 7.13 although the population size estimators under the CB estimator overestimate with large variances in all cases, it provides more accurate of population size than the CM estimator. Moreover, all of the confidence intervals based on the CB estimator cover the true population size. This confirms that this estimator can adjust in the case of dependence.

Table 7.13: The number of patients with breast cancer in Germany ($N = 1,645$)

Method	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	$\widehat{S.E.}(\widehat{N})$	95% of CI	Length
The CM estimator					
M1	1,520	-	23.64	(1,474–1,566)	92
M2	1,520	1,520	25.94	(1,471–1,573)	102
M3	1,520	1,520	23.71	(1,476–1,569)	93
M4	1,521	1,520	18.81	(1,485–1,561)	76
The CB estimator					
M1	1,702	-	35.47	(1,632–1,772)	140
M2	1,703	1,703	34.34	(1,638–1,772)	134
M3	1,703	1,702	35.39	(1,636–1,774)	138
M4	1,703	1,703	28.12	(1,650–1,763)	113

M1: Normal approximation method, M2: True bootstrap method,
M3: Imputed bootstrap method, M4: Reduced bootstrap method

Another interesting point is that $OR = 1.69$; this indicates that the two sources of data are dependent, leading to underestimation of the total number of patients with breast cancer in Saarland, Germany when using the CM estimator. A further implication of knowing the odds ratio is that we can adjust the total of patients with breast cancer as:

$$\begin{aligned} OR &= 1.69 \\ \frac{f_{11}f_{00}}{f_{10}f_{01}} &= 1.69 \\ \widehat{f}_{00} &= \frac{1.69 f_{10} f_{01}}{f_{11}}. \end{aligned} \tag{7.20}$$

Then the estimate for f_{00} of the CM estimator is $\widehat{f}_{00} = \frac{1.69 f_{10} f_{01}}{f_{11}+1}$. The number of patients with breast cancer can be adjusted by the odds ratio as follows:

$$\begin{aligned} \widehat{N}_{CM(adj)} &= f_{10} + f_{01} + f_{11} + \widehat{f}_{00} \\ &= f_{10} + f_{01} + f_{11} + \frac{1.69 f_{10} f_{01}}{f_{11}+1} \\ &= 124 + 722 + 492 + \frac{1.69 \times 124 \times 722}{492+1} \\ &= 1,644.90 \\ &\approx 1,645. \end{aligned} \tag{7.21}$$

Finally, we have that the number of patients with breast cancer from adjustment of the CM estimator by means of the odds ratio is equal to the real number of patients.

7.7.2 The road traffic death rates in Ethiopia data

The road traffic deaths data was established by the traffic police and hospital injury surveillance between June 2012 and May 2013, see also Table 7.14 . The previous study used the CM estimator to estimate the number of death, it was reported that the approximate number of death people by road traffic in Ethiopia was 256 with a 95% of confidence interval (219 - 293) (see Abegaz et al., 2014).

Table 7.14: The number of death people from road traffic in Ethiopia from 2012 to 2013

	Hospital		
Police	Yes	No	Total
Yes	50	103	153
No	44	$\widehat{f}_{00}$	
Total	84		

For the CM estimator to be valid it requires the assumption of independence between the two sources.. However, there was evidence of positive dependence as pointed out in the original study. In fact, sometimes the police required medical evidence from the hospital, referred to as a medical-legal issue. The CB estimator provides a potential more accurate population size estimate than the original one under dependence of the two sources. Then, the simulation results suggest that the imputed bootstrap is the best way to estimate variance and to construct confidence intervals (see Table 7.15).

Table 7.15: The number of road traffic deaths in Ethiopia

Method	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	$\widehat{S.E.}(\widehat{N})$	95% of CI	Length
The CM estimator					
M1	256	-	18.41	(220–292)	72
M2	-	-	-	-	-
M3	258	256	19.16	(225–299)	74
M4	257	255	16.26	(230–293)	63
The CB estimator					
M1	281	-	22.96	(236–326)	90
M2	-	-	-	-	
M3	283	281	24.49	(242–337)	95
M4	283	281	20.38	(269–327)	58

M1: Normal approximation method, M2: True bootstrap method,
M3: Imputed bootstrap method, M4: Reduced bootstrap method

7.7.3 The homeless deaths in France data

The number of homeless deaths in France between 2008 to 2012 was analysed by Vuillermoz et al. (2014). The data was collected from two sources: the French National Institute for Health and Medical Reach databases (CMDR), and all deaths certificated from the C'epiDc database (C'epiDc), summarised in Table 7.16. The total of homeless deaths estimated under the Lincoln-Petersen estimator was $6,730$ with 95% of confidence interval (4,381-9,079). The original research acknowledged that the data might have a positive dependence between the two sources. This led to the underestimation of the homeless death rates.

Table 7.16: The number of homeless deaths in France according to the two sources.

	C'epiDc database		
CMDR database	Yes	No	Total
Yes	27	727	754
No	214	$\widehat{f}_{00}$	
Total	241		

In Table 7.17, provides estimates of the population size based upon the CM and the CB estimators as well as standard error and confidence intervals based on the approximating normal approach, the imputed bootstrap and reduced bootstrap approach. According to the simulation result, it is argued that CB estimator is more appropriate to estimate the population size for the homeless deaths in France data than the CM estimator due to dependence between the sources. Then, it might be appropriate to construct the confidence interval by the imputed bootstrap.

Table 7.17: The number of homeless death people in France between 2008 - 2010

Method	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	$\widehat{S.E.}(\widehat{N})$	95% of CI	Length
The CM estimator					
M1	6,524	-	1,118.15	(4,332–8,716)	4,384
M2	-	-	-	-	-
M3	6,750	6,577	1,252.52	(4,822–9,779)	4,957
M4	6,693	6,544	1,203.25	(4,828–9,642)	4,814
The CB estimator					
M1	9,167	-	1,668.43	(5,897–12,437)	6,540
M2	-	-	-	-	-
M3	9,533	9,198	1,933.41	(6,720–14,210)	7,490
M4	9,454	9,167	1,813.64	(8,300–13,522)	5,222

M1: Normal approximation method, M2: True bootstrap method,
M3: Imputed bootstrap method, M4: Reduced bootstrap method

7.7.4 Legionnaires' disease in Belgium data

The total number of Legionella cases in Wallonia, Belgium in 2012 was estimated by Jacquinet et al. (2015). Data were collected from two sources, the notification of Walloon public health medical inspectors and the hospital in Wallonia. (see Table 7.18) This study used the CM estimator, the most commonly used estimator to deal with single marking capture recapture data. The total number of Legionella cases was estimated as 45 (95% CI: 41–48).

Table 7.18: The Legionnaires' disease cases according to the two sources

	Notification		
Hospital	Yes	No	Total
Yes	26	3	29
No	14	$\widehat{f}_{00}$	
Total	40		

Table 7.19: The number of Legionnaires' disease cases in Wallonia, Belgium, 2012

Method	$\widehat{N}_{Mean}$	$\widehat{N}_{Med}$	$\widehat{S.E.}(\widehat{N})$	95% of CI	Length
The CM estimator					
M1	45	-	1.60	(42–48)	6
M2	-	-	-	-	-
M3	45	45	1.71	(42–48)	6
M4	45	44	1.12	(42–51)	9
The CB estimator					
M1	46	-	2.2	(42–50)	8
M2	-	-	-	-	
M3	46	46	2.35	(42–51)	9
M4	46	46	1.67	(45–50)	5

M1: Normal approximation method, M2: True bootstrap method,
M3: Imputed bootstrap method, M4: Reduced bootstrap method

The recommended method here is M3, that is the imputed bootstrap, to construct standard errors and confidence intervals of population size under the CB or the CM estimators. In this case, there is not much of a difference between CM and CB estimators. In the original work a caution was raised that the two sources were associated due to the fact that microbiologists in the hospital catalysed the data for the second source. From the given data, this question of dependency cannot ultimately be answered.

7.8 Conclusion

Estimating the hidden population size for capture-recapture study by parametric approaches involoves two main procedures. Firstly, the best parametric model for count data set must be selected. Secondly, all parameters under the selected model are estimated. The problem of misspecification might occur for the capture-recapture model. For the single marking capture-recapture method, data are often used to estimate population size by Chapman's (CM) estimator. However, the CM estimator requires as basic assumption: that the count data are distributed by the hyper geometric distribution, individual homogeneity and independence between sources. Furthermore, the Chao estimator based on the binomial mixture distribution for two sources (CB) was proposed by Brittain and Böhning (2009) in 2009 as a better choice in case of a dependency problem in single marking data.

Variance estimations based on the normal approximation or formula method of the CM and the CB estimators were derived by Seber (1970) and Brittain and Böhning (2009), respectively. However, population size estimations of the CM and the CB estimators might not follow an asymptotically normal distribution for small population sizes. As a consequence, the resampling approach have been considered as alternatives (see Norris III and Pollock, 1996; Zwane and Van der Heijden, 2003). In this Chapter, the resampling approaches based on the multinomial distribution are proposed as alternative methods.

The aim was to investigate the variance approximation of the CM estimator under independence of two sources. We found that the variance estimation based on the normal approximation (formula) method provides the underestimation of variance estimation for small population size, but it is the best way to approximate variance for high marginal probabilities.

To approximated the variance of the CB estimator, the variance estimation based on the normal approximation (formula) method is the best way for estimating the variance of the CB data, except if the marginal probabilities are very small. The true bootstrap method or the imputed bootstrap are more appropriate for a small population size. However, it might be better to choose the CM estimator to estimate the population size under independent sources because the CM estimator is more accurate than the CB estimator.

If the assumption of independence between two sources fails, the original study claimed that the CB estimator is more suitable for estimating population size than the CM estimator. The simulation study leads to the recommendations that the true bootstrap should be used to approximate variance of the CM and the CB estimators. The imputed bootstrap should be used for unknown population size if population size is not large. In the case of population size is large enough, approximate normal formulas can be used

to estimate the variance of the CM and the CB estimators. Overall, the CB estimator leads to higher estimates of the variance than the CM estimator.

For constructing confidence intervals, it is recognised that we should use the CM estimator for estimating the target population size if two sources are independent. Additionally, the imputed bootstrap method should be used to calculate confidence intervals. The CB estimator is not suggested as being an appropriate method to use for independent sources. It might be better to use the CB estimator to estimate the target population size for the dependent case and constructed the confidence interval by the imputed bootstrap.

Chapter 8

Conclusions and Future Work

This chapter provides the conclusions and discussions of the thesis. Additionally, some future work for extension and developing the research is described.

8.1 Conclusions and discussions

A capture-recapture approach is a powerful tool for estimating the target population size with zero-truncated count data. An individual of the target population carries information on the number of times identified during a study period. A basic distribution for the observed counts is the binomial distribution for a fixed number of counting occasions or the Poisson when counting occasions can not be fixed in advance. This thesis, focus was on the case where the number of counting occasions is not known in advance. Therefore, the Poisson model is assumed as the basic model in this study. In real life capture-recapture data, the target population might be heterogeneous. For example, the target population might have different level of gender, regions or education, and they are usually recorded in the measuring process, resulting in over- or under-dispersion based on the basic model. The mixed Poisson model has been studied by many researchers (Chao, 1987; Böhning et al., 2005; Viwatwongkasem et al., 2008; Lanumteang, 2010). The negative binomial distribution is a popular choice for the distribution of a count data in the case of a case of over-dispersion in a Poisson model; however, a boundary problem for the dispersion parameter might occurs in some cases for a capture-recapture study. The aim of the thesis is to develop the population size estimator under zero-truncated count data accounting for heterogeneity. The two-parameter distributions, which are generalised from the Poisson were investigated in this thesis, these are the generalised Poisson (GP) distribution and the Conway-Maxwell-Poisson (CMP) distribution. The advantage of these two models is that they not only include the Poisson as the sub-model but also generalise the original Poisson allowing for over-dispersion and under-dispersion. Additionally, the GP and the CMP distributions can capture the potential skewness of the

underlying distribution without adding mixture components as in the Poisson mixture model.

The first model considered is a capture-recapture modelling with the generalised Poisson distribution (see Chapter 3). For estimating the parameters of the zero-truncated generalised Poisson (ZTGP) distribution, the EM-algorithm was modified for the GP distribution. The Horvitz-Thompson estimator was used and applied to a new estimator which is the MLEGP estimator. The simulation results suggest that the MLEGP estimator is an asymptotically valid estimator concerning the population size under the Poisson and generalised Poisson model. It works well for estimating the target population size based on the Poisson distribution when the value of event parameter $\lambda \geq 1$ and performs well for the generalised Poisson distribution when population size is large. However, it seems to be that the MLEGP is too limited to be useful in capture-recapture count data analysis.

The second model for the capture-recapture data is the Conway-Maxwell-Poisson model (see Chapter 4). The advantages of the CMP distribution are that it includes the Poisson and geometric distributions as sub-models. The second population size estimator, called the LCMP, was proposed under the zero-truncated Conway-Maxwell-Poisson distribution. The ratio plot is modified to the log ratio plot as a tool for investigating the suitability of the Poisson, the geometric or the CMP distributions. Moreover, the LCMP estimator can be obtained by exploiting the log-ratio of successive frequency counts using the weighted least squares method. Interestingly, simulation studies confirm that the LCMP estimator is an asymptotically valid estimator not only for the true models (i.e., Poisson, geometric and CMP) but also for the negative binomial distribution with respect to the population size. The LCMP estimator is recommended for estimating population size in populations following the Poisson, the geometric and the CMP distributions, particularly for long-tail frequencies data when the dispersion parameter is close to zero.

Considering variance estimation of the LCMP estimator, variance estimation based on the normal approximation method was derived and provided in Chapter 5. As the adequacy of the approximated normal distribution with associated variance and confidence interval depends on the sample size and sample data, this could lead to a confidence interval not covering the true population size. Then, it is no surprise that although the proposed estimator showed superior performance in terms of accuracy, it evidently also gave the largest variation. Nevertheless, the variation of the new estimator considerably decreases for medium or large population sizes (1,000 and more), as often occurs in real-world applications. Another drawback of normal variance approximation is that the model selection for capture-recapture data might be not a symmetric distribution such as the Poisson distribution or the Conway-Maxwell-Poisson distribution. The issue of asymmetric distributions in capture-recapture data has been discussed by many researchers such as Chao (1987); Tounkara and Rivest (2015) used the Burnham method

or Köse et al. (2014) proposed the log-normal distribution to construct the confidence interval for capture-recapture data. As a consequence, resampling methods, the true bootstrap, imputed bootstrap and the reduced bootstrap were proposed as alternative ways for estimating the variance and standard error as well as for constructing the confidence interval for the LCMP estimator. The simulation results suggest that the true bootstrap performs the best for estimating variance and constructing the confidence interval for the LCMP estimator on average, but this approach requires the population size to be known in advance. Therefore, the imputed bootstrap might be the best choice in real-life situations since it performs very similarly to the true bootstrap method. The normal approximation variance is still useful, it might be a simple way for estimating the variance and standard error of the LCMP estimator when the population is large. The reduced bootstrap is not recommended to be used for the LCMP estimator since it usually provides an underestimate of the variance and the coverage probability is often lower than the desired confidence level.

The geometric distribution $Geo(p)$ is nested in the Conway-Maxwell-Poisson distribution $CMP(\lambda, \nu)$ with event parameter $p = 1 - \lambda$ and $0 < \lambda < 1$. Then the traditional Turing and Zelterman estimators were developed under the geometric distribution in Chapter 6. The simulation study confirms that the Turing estimator under the geometric distribution (TG) performs well with less bias in population size and small variance. It is not surprising that the Zelterman estimator under the zero-truncated geometric distribution (ZG) is an asymptotically unbiased estimator, but it provides a very large bias and variance for small population size. This reason is that the ZG estimator is expected to work for the contaminated geometric distribution, and the ZG might be a robust estimator for misspecification of the original geometric model. To study the behaviour of the ZG estimator under the contaminated geometric distribution and other heterogeneity cases should be a topic for the future research. For real datasets, it is recommended that the variance and confidence intervals for the TG estimator are calculated using the variance formula. The ZG estimator is not recommended for use in real data which follows a geometric distribution if the population size is not large.

According to the uncertainty estimation in heterogeneous capture-recapture data in Chapter 5, we investigated the case that the number of occasions is fixed as two in Chapter 7. This is called the single marking capture-recapture approach which is often found in social and epidemiological studies. The traditional variance estimation based on the normal approximation method was compared with three resampling approaches. It was concluded that the Chapman estimator should be used to estimate population size for independent sources and construct variance estimates with the imputed bootstrap approach. On the contrary, the Chao estimator under a binomial distribution with two sources (CB) is the better choice to estimate the population size for dependent sources. The confidence interval of the population size of CB estimator should also be constructed by the imputed bootstrap method.

Another issue is the process of model selection and inference. Typically, the EM-algorithm is one of the important methods for estimating parameters for zero-truncated count distribution such as the EM-algorithm for the zero-truncated generalised Poisson in Chapter 3. However, Böhning et al. (2013a) suggested the ratio-based approach which can be applied to capture-recapture count data especially the Katz distributions. Many cases of the ratio of successive probabilities show a linear line pattern such as the negative binomial, the geometric and the extension of polynomial regression (see Lanumteang, 2010; Rocchetti et al., 2011; Böhning et al., 2016), and include the generalised Poisson and the Conway-Maxwell-Poisson distributions. The benefit of the ratio regression is not only for selecting an appropriate model but also for estimating the model parameters for capture-recapture data. The presentation of the long right-tail data such as the Shakespeare word data in Chapter 3 or the heroin drug users in Chapter 4, 5 and 6 are a challenge to the goodness-of-fit. As a consequence, the procedure of cutting off the truncated frequency count on the right-hand side might be used. The accuracy of population size depends on the model being correct. Therefore, the statistical testing to select the above-truncation point should be included before going to the population size estimation procedure.

The thesis includes several empirical data sets and has given practical advice step by step. The ratio plot is used as a basic tool for selecting the validation models for the empirical datasets. There are ratio-plots for the Poisson, the generalised Poisson, the Conway-Maxwell-Poisson and the geometric distribution. All procedures and algorithms involved in the calculations have been done by `R programming`.

8.2 Recommendation

The LCMP estimator is a powerful estimator for estimating both homogeneous and heterogeneous population size under the Conway-Maxwell-Poisson distribution. Additionally, the alternative choice is the TG estimator which is a nonparametric approach and suitable for heterogeneous populations following the geometric distribution. The benefit of the TG estimator is that it is very easy to calculate and uses only the observed data. The ZG estimator might be useful for the contaminated geometric distribution and misspecification of the geometric model.

We provided several insights into the behaviour of bootstrap methods for variance estimation. It is very clear that the reduced bootstrap does not work, in the sense that it provides underestimation with respect to the true variance. This is independent of whether the model holds or not. This result indicates that current practice using reduced bootstrap method in capture-recapture should be discontinued. This result is similar to Zwane and Van der Heijden (2003). However, Buckland and Garthwaite (1991) suggested that the nonparametric bootstrap (reduced bootstrap) might have a benefit for

relaxing the assumption of multinomial distribution such as the open population study. The true bootstrap works very well but it cannot be used in practice. The imputed bootstrap seems to work like the true bootstrap but only if the model is valid.

8.3 Future Work

Although the results presented in this thesis have demonstrated the effectiveness of capture-recapture modelling for zero-truncated count data for individual heterogeneuty, there are some important aspects that could be developed or extended in the future.

8.3.1 Developing the population size estimator for Conway-Maxwell-Poisson distribution with the maximum likelihood estimation

The proposed estimator $\widehat{N}_{LCMP} = n + f_1 \exp(-\widehat{\beta}_0)$ where $\widehat{\beta}_0$ is the intercept point from fitting the log-ratio plot of CMP distribution with WLS approach (see Chapter 4) should be investigated further in future work. There are some limitations for the ratio regression approach with weighted least squares approach which are that the frequencies of the observed count are non-negative ($f_x > 0$ for $x > 0$) and the number of observed classes is no less than four. Therefore, other methods for estimating parameters of zero-truncated Conway-Maxwell-Poisson distribution might be considered in future study. For example, using the maximum likelihood estimation approach. Indeed, the maximum likelihood approach (MLE) for estimating the two parameters of the zero-truncated Conway-Maxwell-Poisson (ZTCMP) distribution is limited in the current situation. Since the likelihood function of the ZTCMP distribution involves a normalising term, the computation and distribution theory are complex. Recently, the `compoisson package` in `R` has been developed for computing the log-likelihood for the CMP distribution. Thus, the EM-algorithm needs to be modified for estimating the two parameters of ZTCMP distribution under the MLE estimation in the future work. Moreover, the comparison of the population size estimator with WLS and MLE estimation for the ZTCMP distribution is a further topic for future research. In this study, Conway-Maxwell-Poisson parameter estimation (λ, ν) derived from the WLS approach could be used as starting values for a numerical method, extending to the maximum likelihood estimator in a future study.

The population size estimator for capture-recapture data under the Conway-Maxwell-Poisson distribution is given as

$$\widehat{N}_{MLEcmp} = \frac{n}{1 - \widehat{p}_{0(MLEcmp)}}, \tag{8.1}$$

where $\widehat{p}_{0(MLEcmp)} = \dfrac{1}{z(\widehat{\lambda},\widehat{\nu})}$, and two parameters need to be estimated under maximum likelihood estimation of zero-truncated Conway-Maxwell-Poisson distribution. However, they do not have a closed form solution. The EM-algorithm is applied to estimate ν and λ as follows:

Step 0 : Set $l = 0$ and choose the initial value for observed frequency $\widehat{f}_0^{(l)}$. $\widehat{f}_0^{(0)} = 0$ is suggested for the general case, or use the WLS approach to get the $\widehat{p}_0^{(0)} = \dfrac{1}{z(\widehat{\lambda}^{(0)},\widehat{\nu}^{(0)})}$, substituting into $\widehat{f}_0^{(0)} = \dfrac{n\widehat{p}_0^{(0)}}{1-\widehat{p}_0^{(0)}}$.

Step 1 : Substituting $\widehat{f}_0^{(l)}$ in a completed frequency distribution table as Table 8.1 for computing a new maximum likelihood estimators $\widehat{\lambda}^{(l+1)}$ and $\widehat{\nu}^{(l+1)}$.

Table 8.1: The frequency distribution of complete data

x	0	1	2	3	...	m
f_x	$\widehat{f}_0^{(l)}$	f_1	f_2	f_3	...	f_m

As suggested above, the maximum likelihood estimators are computed by using the function `com.fit()`, `compoisson package` in `R`. This leads to new maximum likelihood estimators.

Step 2 : Computing a new unobserved frequency and size of target population, that is

$$\widehat{f}_0^{(l+1)} = \frac{n\widehat{p}_0^{(l+1)}}{1-\widehat{p}_0^{(l+1)}},$$

where $\widehat{p}_0^{(l+1)} = \dfrac{1}{z(\widehat{\lambda}^{(l+1)},\widehat{\nu}^{(l+1)})}$, and

$$\widehat{N}_{MLEcmp}^{(l+1)} = \frac{n}{1-\widehat{p}_0^{(l+1)}}.$$

Step 3 : Checking the condition of the algorithm by plugging $\widehat{\theta}^{(l+1)}$ and $\widehat{\alpha}^{(l+1)}$ into the log-likelihood of ZTGP function:

$$\begin{aligned}\log L(x;\lambda,\nu)^{(l+1)} &= \sum_{x=1}^{m}[x\log\lambda^{(l+1)} - \nu^{(l+1)}\log x! \\ &\quad - \log\{z(\lambda^{(l+1)},\nu^{(l+1)}) - 1\}]. \qquad (8.2)\end{aligned}$$

and compare

$$dif = \left|\log L(x;\lambda,\nu)^{(l+1)} - \log L(x;\lambda,\nu)^{(l)}\right| < 0.0001,$$

then set $l = l + 1$. Then, if $dif > 0.0001$ return to step 1 so that new maximum likelihood estimators are updated. The algorithm is repeated until the log likelihood function of ZTCMP converge to a constant with an acceptable error, that is $dif < 0.0001$.

The CMP log-ratio plot is a simple tool for investigating validity of the zero-truncated Conway-Maxwell-Poisson distribution. However, the comparison of goodness of fit of parameters estimation approaches will be of interest in future study.

8.3.2 Developing the population size estimator for capture-recapture data with generalised Conway-Maxwell-Poisson which includes the negative binomial distribution allowing for heterogeneity

In practice, the appropriate model plays a key role for the accuracy of the parameter estimation. The proposed LCMP estimator in Chapter 4 performs well for the true models (i.e. Poisson, geometric and Conway-Maxwell-Poissn distribution), and useful for the negative binomial distribution. Recently, Imoto (2014) proposed the more flexible alternative distribution for modelling over- and under-dispersion. There is a generalised Conway-Maxwell-Poisson distribution which includes the negative binomial (GCMP) distribution and is an extension of the CMP distribution. A benefit of the GCMP distribution is that it includes the negative binomial as a special case, leading to a longer-tailed model than the CMP distribution. The new parameter r is added for controlling the length of tail. The GCMP distribution with three parameters has the probability mass function as follows:

$$p_x = \frac{\Gamma(r+x)^{\nu}\lambda^x}{x!\, C(\nu, r, \lambda)}, x = 0, 1, 2, ..., \tag{8.3}$$

where the normalised constant $C(\nu, r, \lambda) = \sum_{j=0}^{\infty} \frac{\Gamma(r+j)^{\nu}\lambda^j}{j!}$ for $\nu < 1, r > 0$ and $\lambda > 0$ or $\nu = 1, r > 0$ and $0 < \lambda < 1$. This distribution reduces to the CMP distribution with parameters $(1 - \nu)$ and λ when $r \to 1$. Moreover, the GCMP distribution is over-dispersed for $0 < \nu < 1$ and under-dispersed for $\nu < 0$. The GCMP ratio plot are formed as

$$\begin{aligned}
\frac{p_{x+1}}{p_x} &= \frac{\frac{\Gamma(r+x+1)^{\nu}\lambda^{x+1}}{(x+1)!\,C(r,\nu,\lambda)}}{\frac{\Gamma(r+x)^{\nu}\lambda^{x}}{x!\,C(r,\nu,\lambda)}} = \frac{\frac{\Gamma(r+x+1)^{\nu}\lambda^{x+1}}{(x+1)!}}{\frac{\Gamma(r+x)^{\nu}\lambda^{x}}{x!}} = \frac{\frac{\Gamma(r+x+1)^{\nu}\lambda^{x}\lambda}{(x+1)x!}}{\frac{\Gamma(r+x)^{\nu}\lambda^{x}}{x!}} \\
&= \left\{\frac{\Gamma(r+x+1)}{\Gamma(r+x)}\right\}^{\nu}\frac{\lambda}{(x+1)} = \left\{\frac{(r+x+1+1)!}{(r+x-1)!}\right\}^{\nu}\frac{\lambda}{(x+1)} \\
&= \left\{\frac{(r+x)!}{(r+x-1)!}\right\}^{\nu}\frac{\lambda}{(x+1)} = \left\{\frac{(r+x)(r+x-1)!}{(r+x-1)!}\right\}^{\nu}\frac{\lambda}{(x+1)} \\
&= (r+x)^{\nu}\frac{\lambda}{(x+1)} \\
(x+1)\frac{p_{x+1}}{p_x} &= (r+x)^{\nu}\lambda \qquad (8.4)
\end{aligned}$$

Let $r_x = (x+1)\frac{p_{x+1}}{p_x}$ be called the ratio plot of the GCMP distribution, as we can see that the equation (8.4) allows for non linear regression. Therefore, taking the log-transform results in a linear equation:

$$\log(r_x) = \log\left\{(x+1)\frac{p_{x+1}}{p_x}\right\} = \nu\log(r+x) + \log\lambda \qquad (8.5)$$

Taking the first-order Taylor expansion of $\log(r+x)$ around 0. Let $g(x) = \log(r+x)$, then $g'(x) = \frac{1}{r+x}$. We achieve $\log(r+x) \approx \log r + \frac{1}{r}x$. The log-ratio of GCMP distribution is given as:

$$\begin{aligned}
\log(r_x) \approx \nu(\log r + \frac{1}{r}x) + \log\lambda &= (\nu\log r + \log\lambda) + \frac{\nu}{r}x \qquad (8.6) \\
&= \beta_0 + \beta_1 x. \qquad (8.7)
\end{aligned}$$

It can be seen that the equation is a linear line when the new parameter $\beta_0 = \nu\log r + \log\lambda$ and $\beta_1 = \frac{\nu}{r}$. Therefore, it might be said that a graph of $\log(r_x)$ against x can be used as a tool for detecting the GCMP model and its nested models. Note that the GCMP distribution reduces to the CMP distribution with parameter $1-\nu$ when $r \to 1$.

We expect that the GCMP model might be useful for estimating population size heterogeneity in the case of short and long - tail skewed models. However, the three parameter estimation based on the zero-truncated generalised Conway-Maxwell-Poisson distribution is a challenge for future study.

8.3.3 Developing the population size estimator for capture-recapture data with the Poisson log-normal model allowing for heterogeneity

The parametric Poisson mixture such as the Poisson-log-normal model might be useful for capture-recapture modelling. As the negative binomial is the Poisson-gamma distribution, we replace the gamma density by the log-normal distribution leading to a Poisson log-normal (PLN) mixture distribution. The advantage of the PLN distribution is that it is a more natural interpretation of the parameter mean, relying on the central limit theorem (Trinh et al., 2014). Then let λ be a local parameter which can be different relying on sub-populations and is log-normally distributed.

$$f(\lambda;\theta,\sigma) = \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log\lambda-\theta)^2}{2\sigma^2}\right\}, \tag{8.8}$$

where θ is the mean and σ^2 and σ are the variance and standard deviation of normal distribution X where $X = \log(\lambda)$. Then p_x is the Poisson-log-normal mixture distribution, if

$$\begin{aligned} p_x &= \int_0^\infty \frac{\exp(-\lambda)\lambda^x}{x!} \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log\lambda-\theta)^2}{2\sigma^2}\right\} d\lambda \\ &= \frac{1}{x!\,\sigma\sqrt{2\pi}} \int_0^\infty \exp(-\lambda)\lambda^{x-1} \exp\left\{-\frac{(\log\lambda-\theta)^2}{2\sigma^2}\right\} d\lambda, \end{aligned} \tag{8.9}$$

where $x = 0, 1, 2, \ldots$. The integral in equation (8.9) cannot be expressed in a simple form. Also, the apparent expression for the generating function is not know. However, the l^{th} factorial moment of the PLN distribution is equal to the l^{th} ordinary moment, that is

$$\mu_{(l)} = \exp(\theta l + \frac{1}{2} l^2\sigma^2), \tag{8.10}$$

(see Bulmer, 1974). Then, the mean of PLN distribution is the

$$E(X) = \exp(\theta + \frac{\sigma^2}{2}), \tag{8.11}$$

and the variance is

$$Var(X) = \exp(\mu + \frac{\sigma^2}{2}) \left[1 + \left\{\exp(\mu + \frac{\sigma^2}{2})(\exp(\sigma^2) - 1)\right\}\right]. \tag{8.12}$$

The proposed estimator based on the zero-truncated PLN distribution will be achieved from

$$\widehat{N}_{PLN} = \frac{n}{1 - \widehat{p}_{o(PLN)}}, \tag{8.13}$$

where $\widehat{p}_{0(PLN)} = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_0^{\infty} \frac{e^{-\lambda}}{\lambda} \exp\left\{-\frac{(\log\lambda - \theta)^2}{2\sigma^2}\right\} d\lambda$ and $n = \displaystyle\sum_{x=1}^{m} f_x$. Therefore, the estimated parameters θ and σ under the PLN require to be estimated. However, their closed form results of the integral is not known. This leads to the EM-algorithm being applied to solve this problem. The basic knowledge of the EM-algorithm of the PLN distribution can be taken from Karlis (2005). Moreover, the `poilogMLE () function` from `poilog package` in `R` might be useful for estimating parameters θ and σ in the E-step.

It is expected that expect that the population size estimation based on the PLN distribution will be an alternative choice for estimating heterogeneous population size in capture recapture study.

All in all, parametric approaches for modelling individual heterogeneity have been studied by many researchers. The crucial issue is that each observed data set can be fitted by several models. The new population size estimators under the Poisson and its extension of the Poisson models for population heterogeneity in this thesis has shown excellent properties in a variety of situations. Also, the resampling technique was provided as a good choice for estimating variance estimation of population size for capture-recapture studies. It is expected that this knowledge will provide a set of alternative choices making real world capture-recapture data more realistic.

Appendix A

Generalised Poisson Distribution

A.1 The ratio plot for investigating the validity of generalised Poisson distribution

Graphical techniques have been used to visualise quantitative data, and applied for selecting a suitable model in statistics. Since it is quick and uncomplicated to understand, a graphical approach, namely *the ratio plot* was suggested as a method for choosing a model by Böhning et al. (2013a). It can be extended to zero-truncated modelling which is common in capture-recapture studies. As a consequence, the ratio plot was applied in order to investigate the validity of the generalised Poisson distribution and the zero-truncated generalised Poisson distribution. The ratio plot can be defined as the ratio of neighbouring probabilities multiplied by the value of the largest neighbour count. For the GP distribution, it is given as:

$$\begin{aligned} r_x &= (x+1)\frac{p_{x+1}}{p_x} \\ &= (x+1)\frac{\theta\{\theta+\alpha(x+1)\}^{x+1-1}\left(\frac{\exp(-\theta-\alpha(x+1))}{(x+1)!}\right)}{\theta(\theta+\alpha x)^{x-1}\left(\frac{\exp(-\theta-\alpha x)}{x!}\right)} \\ &= \exp(-\alpha)\left(1+\frac{\alpha}{\theta+\alpha x}\right)^{x-1}\{(\theta+\alpha)+\alpha x\}. \end{aligned} \tag{A.1}$$

Let $y = \left(1+\frac{\alpha}{\theta+\alpha x}\right)^{x-1}$ and taking natural logarithm of both sides, we achieve that

$$\begin{aligned} \ln(y) &= \ln\left(1+\frac{\alpha}{\theta+\alpha x}\right)^{(x+1)} \\ &= (x-1)\ln\left(1+\frac{\alpha}{\theta+\alpha x}\right). \end{aligned} \tag{A.2}$$

Next, multiply (A.2) by $\frac{\alpha}{\theta+\alpha x}/\frac{\alpha}{\theta+\alpha x}$ on the right hand side, that is

$$\ln(y) = (x-1)\left[\frac{\ln\left(1+\frac{\alpha}{\theta+\alpha x}\right)}{\frac{\alpha}{\theta+\alpha x}}\right]\left[\frac{\alpha}{\theta+\alpha x}\right]. \tag{A.3}$$

Let $g(x) = \ln\left(1+\frac{\alpha}{1+\alpha x}\right)$ and $z(x) = \frac{\alpha}{1+\alpha}$. Therefore,

$$g'(x) = \frac{1}{1+\frac{\alpha}{1+\alpha}}\left[-\frac{\alpha^2}{(1+\alpha x)^2}\right],$$

and

$$z'(x) = -\frac{\alpha^2}{(1+\alpha x)^2}.$$

Then,

$$\begin{aligned}\frac{g'(x)}{z'(x)} &= \frac{\frac{1}{1+\frac{\alpha}{1+\alpha x}}\left[-\frac{\alpha^2}{(1+\alpha x)^2}\right]}{-\frac{\alpha^2}{(1+\alpha x)^2}} \\ &= \frac{1}{1+\frac{\alpha}{1+\alpha x}}.\end{aligned} \tag{A.4}$$

Using L'Hospital's rule, $\lim_{x\to 0}\frac{g(x)}{z(x)} = \lim_{x\to 0}\frac{g'(x)}{z'(x)}$, we achieve as:

$$\lim_{x\to 0}\frac{1}{1+\frac{\alpha}{1+\alpha x}} = 1. \tag{A.5}$$

Hence, substituting (A.5) into (A.3)

$$\begin{aligned}\ln y &= (x+1)\frac{\alpha}{1+\alpha x} \\ &= \frac{\alpha x-\alpha}{\alpha x+1} \\ &= \frac{\alpha x+1-1-\alpha}{\alpha x+1} \\ &= \frac{\alpha x+1}{\alpha x+1}-\frac{\alpha+1}{\alpha x+1} \\ &= 1-\frac{\alpha+1}{\alpha x+1}.\end{aligned} \tag{A.6}$$

Taking a limit of (A.6) with x approaches to infinity as:

$$\begin{aligned} \lim_{x\to\infty} \ln y &= \lim_{x\to} \left(1 - \frac{\alpha+1}{\alpha x+1}\right) \\ &= 1 \\ y &= e. \end{aligned} \tag{A.7}$$

It can be seen that y will converse to the exponential function (e) when $x \to \infty$. Substuting (A.7) into (A.1), that is

$$\begin{aligned} r_x = (x+1)\frac{p_{x+1}}{p_x} &= \underbrace{\underbrace{\exp(-\alpha)}_{\text{constant}} \underbrace{\left(1 + \frac{\alpha}{\theta+\alpha x}\right)^{x-1}}_{\text{e=2.71828182846;x}\to\infty}}_{\text{constant}} \{(\theta+\alpha)+\alpha x\} \\ &= c\{(\theta+\alpha)+\alpha x\} \\ &= (c\theta + c\alpha) + c\alpha x \\ &= \underbrace{c^{'}}_{\text{intercept}} + \underbrace{c\alpha}_{\text{slope}} x, \end{aligned} \tag{A.8}$$

where c is a positive constant, calculated by $\exp(-\alpha)\left(1 + \frac{\alpha}{\theta+\alpha x}\right)^{x-1}$, and $c^{'} = c\theta + c\alpha$. When x is large we found that $\left(1 + \frac{\alpha}{\theta+\alpha x}\right)^{x-1}$ closed to $e \approx 2.71828$.

A.2 Simulation results

i) A homogeneous Poisson As the Poisson distribution is a special case of GP when the dispersion parameter $\alpha = 0$, data are generated data from the Poisson distribution with ten different value parameters:

$$\lambda \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0\}$$

ii) A heterogeneous based Poisson: counts are generated from the generalised Poisson distribution with parameters

$$\theta \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0\}$$

and

$$\phi \in \{1.5, 2.0, 2.5, 3.0\},$$

where $\phi = \frac{Var[X]}{E[X]}$, so that $\phi = \left\{\frac{\theta}{(1-\alpha)^3}\right\} \div \left\{\frac{\theta}{(1-\alpha)}\right\} = \frac{1}{(1-\alpha)^2}$.

1) Population size estimation when data are generated from the Poisson distribution

Table A.1: The relative bias $\{RBias(\widehat{N})\}$ of six estimators with different parameters in the Poisson distribution, when $N = 200$, $N = 1,000$ and $N = 10,000$.

Model	Turing	MLEPoi	MCK	Chao	Zel	MLEGP
			$N = 200$			
Poi(0.5)	-0.02176	-0.04872	-0.10951	0.05924	0.07056	1.14113
Poi(0.6)	0.00499	-0.01191	-0.04894	0.05694	0.06547	0.56892
Poi(0.7)	-0.00453	-0.01186	-0.02715	0.02147	0.02653	0.18701
Poi(0.8)	0.01231	0.00937	0.00451	0.02809	0.03133	0.11084
Poi(0.9)	0.01174	0.00974	0.00962	0.02807	0.03245	0.07326
Poi(1.0)	0.00686	0.00776	0.01424	0.01617	0.01912	0.03581
Poi(1.5)	0.00775	0.00819	0.01320	0.01339	0.01714	0.01572
Poi(2.0)	0.00394	0.00427	0.00792	0.00834	0.01301	0.00732
Poi(2.5)	0.00353	0.00403	0.00744	0.00601	0.01071	0.00558
Poi(3.0)	0.00240	0.00257	0.00376	0.00542	0.01250	0.00426
			$N = 1,000$			
Poi(0.5)	0.00133	-0.00144	-0.00696	0.01081	0.01210	0.08281
Poi(0.6)	0.00351	0.00384	0.00576	0.00557	0.00580	0.02745
Poi(0.7)	0.00376	0.00367	0.00490	0.00697	0.00768	0.01864
Poi(0.8)	0.00295	0.00275	0.00384	0.00618	0.00710	0.01361
Poi(0.9)	0.00056	0.00103	0.00291	0.00123	0.00131	0.00553
Poi(1.0)	0.00135	0.00143	0.00252	0.00304	0.00353	0.00673
Poi(1.5)	0.00295	0.00330	0.00448	0.00329	0.00356	0.00433
Poi(2.0)	0.00137	0.00136	0.00214	0.00253	0.00373	0.00206
Poi(2.5)	0.00034	0.00042	0.00075	0.00091	0.00200	0.00097
Poi(3.0)	0.00042	0.00046	0.00098	0.00101	0.00255	0.00073
			$N = 10,000$			
Poi(0.5)	0.00052	0.00018	0.00042	0.00115	0.00124	0.00407
Poi(0.6)	-0.00038	-0.00011	0.00113	-0.00104	-0.00120	-0.00245
Poi(0.7)	-0.00008	-0.00002	0.00058	-0.00003	-0.00001	-0.00066
Poi(0.8)	0.00065	0.00075	0.00110	0.00046	0.00038	0.00072
Poi(0.9)	-0.00067	-0.00044	0.00049	-0.00096	-0.00105	-0.00199
Poi(1.0)	0.00024	0.00011	-0.00016	0.00055	0.00063	0.00139
Poi(1.5)	0.00002	0.00014	0.00032	-0.00013	-0.00024	0.00016
Poi(2.0)	0.00012	0.00007	-0.00004	0.00025	0.00037	0.00031
Poi(2.5)	0.00002	0.00006	0.00010	0.00003	0.00006	0.00011
Poi(3.0)	-0.00012	-0.00012	-0.00001	-0.00006	0.00011	-0.00012

Table A.2: The relative variance $\{RVar(\widehat{N})\}$ six estimators with different parameters in the Poisson distribution, when $N = 200$, $N = 1,000$ and $N = 10,000$.

Model	Turing	MLEPoi	MCK	Chao	Zel	MLEGP
			$N = 200$			
Poi(0.5)	0.02946	0.02374	0.01843	0.07002	0.07925	6.10363
Poi(0.6)	0.02709	0.02281	0.01960	0.05647	0.06455	2.87448
Poi(0.7)	0.01634	0.01455	0.01510	0.03003	0.03545	0.41819
Poi(0.8)	0.01253	0.01122	0.01291	0.02352	0.02884	0.17791
Poi(0.9)	0.01071	0.00949	0.01140	0.02162	0.02755	0.07355
Poi(1.0)	0.00821	0.00740	0.01019	0.01539	0.02003	0.03438
Poi(1.5)	0.00322	0.00295	0.00495	0.00587	0.00908	0.00672
Poi(2.0)	0.00131	0.00114	0.00237	0.00263	0.00522	0.00195
Poi(2.5)	0.00067	0.00060	0.00162	0.00124	0.00293	0.00085
Poi(3.0)	0.00039	0.00034	0.00109	0.00075	0.00226	0.00047
			$N = 1,000$			
Poi(0.5)	0.00733	0.00669	0.00699	0.01202	0.01328	0.08946
Poi(0.6)	0.00491	0.00458	0.00541	0.00784	0.00887	0.04288
Poi(0.7)	0.00365	0.00339	0.00405	0.00595	0.00695	0.01689
Poi(0.8)	0.00252	0.00238	0.00322	0.00429	0.00524	0.01098
Poi(0.9)	0.00192	0.00182	0.00256	0.00313	0.00389	0.00609
Poi(1.0)	0.00160	0.00146	0.00201	0.00278	0.00360	0.00485
Poi(1.5)	0.00055	0.00050	0.00089	0.00099	0.00154	0.00111
Poi(2.0)	0.00025	0.00021	0.00047	0.00048	0.00092	0.00039
Poi(2.5)	0.00014	0.00012	0.00035	0.00024	0.00054	0.00019
Poi(3.0)	0.00008	0.00007	0.00023	0.00014	0.00039	0.00009
			$N = 10,000$			
Poi(0.5)	0.00072	0.00069	0.00080	0.00104	0.00114	0.00409
Poi(0.6)	0.00049	0.00045	0.00055	0.00077	0.00087	0.00232
Poi(0.7)	0.00033	0.00032	0.00041	0.00052	0.00061	0.00132
Poi(0.8)	0.00026	0.00025	0.00033	0.00041	0.00049	0.00089
Poi(0.9)	0.00019	0.00018	0.00026	0.00032	0.00040	0.00062
Poi(1.0)	0.00016	0.00014	0.00021	0.00027	0.00035	0.00047
Poi(1.5)	0.00006	0.00005	0.00009	0.00010	0.00015	0.00011
Poi(2.0)	0.00003	0.00002	0.00005	0.00005	0.00009	0.00004
Poi(2.5)	0.00001	0.00001	0.00003	0.00003	0.00006	0.00002
Poi(3.0)	0.00008	0.00007	0.00023	0.00014	0.00039	0.00009

Table A.3: The relative root mean square $\{RRMSE(\widehat{N})\}$ six estimators with different parameters in the Poisson distribution, when $N = 200$, $N = 1,000$ and $N = 10,000$.

Model	Turing	MLEPoi	MCK	Chao	Zel	MLEGP
			$N = 200$			
Poi(0.5)	0.1730	0.1616	0.1744	0.2712	0.2902	2.7214
Poi(0.6)	0.1647	0.1515	0.1483	0.2444	0.2624	1.7883
Poi(0.7)	0.1279	0.1212	0.1259	0.1746	0.1902	0.6732
Poi(0.8)	0.1126	0.1063	0.1137	0.1559	0.1727	0.4361
Poi(0.9)	0.1042	0.0979	0.1072	0.1497	0.1691	0.2809
Poi(1.0)	0.0909	0.0864	0.1019	0.1251	0.1428	0.1889
Poi(1.5)	0.0573	0.0549	0.0716	0.0778	0.0968	0.0835
Poi(2.0)	0.0364	0.0340	0.0493	0.0520	0.0734	0.0448
Poi(2.5)	0.0262	0.0249	0.0409	0.0357	0.0552	0.0298
Poi(3.0)	0.0200	0.0186	0.0332	0.0280	0.0492	0.0221
			$N = 1,000$			
Poi(0.5)	0.0856	0.0818	0.0839	0.1102	0.1159	0.3104
Poi(0.6)	0.0702	0.0678	0.0738	0.0887	0.0944	0.2089
Poi(0.7)	0.0605	0.0583	0.0638	0.0775	0.0837	0.1313
Poi(0.8)	0.0503	0.0489	0.0569	0.0658	0.0727	0.1057
Poi(0.9)	0.0438	0.0427	0.0507	0.0559	0.0624	0.0782
Poi(1.0)	0.0401	0.0383	0.0449	0.0528	0.0601	0.0700
Poi(1.5)	0.0237	0.0226	0.0301	0.0317	0.0394	0.0336
Poi(2.0)	0.0159	0.0147	0.0218	0.0222	0.0306	0.0200
Poi(2.5)	0.0118	0.0111	0.0186	0.0157	0.0233	0.0138
Poi(3.0)	0.0088	0.0081	0.0154	0.0119	0.0198	0.0095
			$N = 10,000$			
Poi(0.5)	0.0268	0.0262	0.0283	0.0323	0.0338	0.0641
Poi(0.6)	0.0220	0.0213	0.0235	0.0277	0.0295	0.0483
Poi(0.7)	0.0183	0.0178	0.0201	0.0228	0.0246	0.0363
Poi(0.8)	0.0162	0.0159	0.0181	0.0202	0.0221	0.0299
Poi(0.9)	0.0139	0.0134	0.0161	0.0180	0.0201	0.0251
Poi(1.0)	0.0125	0.0120	0.0146	0.0163	0.0186	0.0218
Poi(1.5)	0.0075	0.0072	0.0096	0.0099	0.0123	0.0106
Poi(2.0)	0.0051	0.0048	0.0072	0.0070	0.0097	0.0063
Poi(2.5)	0.0037	0.0034	0.0056	0.0051	0.0075	0.0043
Poi(3.0)	0.0027	0.0025	0.0047	0.0036	0.0058	0.0029

2) Population size estimation when data are generated from the generalised Poisson distribution

Table A.4: The relative bias ($RBias(\widehat{N})$) of six estimators with different parameters in the generalised Poisson distribution, when $N = 200, N = 1,000$ and $N = 10,000$

θ	ϕ	Turing	MLEPoi	McK	Chao	Zel	MLEGP
				$N = 200$			
0.50	1.5	-0.5566	-0.5919	-0.6842	-0.4395	-0.3995	0.9018
	2.0	-0.7010	-0.7309	-0.8277	-0.5852	-0.5321	0.9232
	2.5	-0.7727	-0.7959	-0.8881	-0.6732	-0.6202	0.8112
	3.0	-0.8143	-0.8327	-0.9182	-0.7155	-0.6592	0.9036
1.00	1.5	-0.3470	-0.3921	-0.5383	-0.2298	-0.1738	0.1746
	2.0	-0.5134	-0.5565	-0.7303	-0.3859	-0.3144	0.3839
	2.5	-0.6103	-0.6466	-0.8168	-0.4778	-0.3951	0.6994
	3.0	-0.6716	-0.7014	-0.8645	-0.5439	-0.4586	0.9671
1.50	1.5	-0.2268	-0.2707	-0.4388	-0.1293	-0.0669	0.0248
	2.0	-0.3789	-0.4257	-0.6460	-0.2545	-0.1706	0.1034
	2.5	-0.4825	-0.5247	-0.7602	-0.3461	-0.2503	0.2784
	3.0	-0.5533	-0.5899	-0.8193	-0.4091	-0.3041	0.5934
2.00	1.5	-0.1532	-0.1923	-0.3744	-0.0765	-0.0147	0.0125
	2.0	-0.2836	-0.3297	-0.5804	-0.1690	-0.0761	0.0286
	2.5	-0.3842	-0.4275	-0.7041	-0.2523	-0.1445	0.0645
	3.0	-0.4623	-0.5007	-0.7788	-0.3221	-0.2081	0.2291
				$N = 1,000$			
0.50	1.5	-0.5685	-0.6045	-0.7100	-0.4742	-0.4415	0.2054
	2.0	-0.7100	-0.7403	-0.8504	-0.6207	-0.5804	0.3241
	2.5	-0.7771	-0.8012	-0.9061	-0.6961	-0.6543	0.5088
	3.0	-0.8193	-0.8381	-0.9348	-0.7477	-0.7073	0.5761
1.00	1.5	-0.3545	-0.3990	-0.5515	-0.2538	-0.2051	0.0102
	2.0	-0.5164	-0.5600	-0.7451	-0.4015	-0.3379	0.0345
	2.5	-0.6126	-0.6493	-0.8340	-0.4985	-0.4289	0.0849
	3.0	-0.6758	-0.7057	-0.8833	-0.5691	-0.5010	0.2130
1.50	1.5	-0.2324	-0.2762	-0.4517	-0.1451	-0.0901	0.0042
	2.0	-0.3817	-0.4294	-0.6614	-0.2641	-0.1848	0.0123
	2.5	-0.4858	-0.5283	-0.7719	-0.3617	-0.2751	0.0287
	3.0	-0.5607	-0.5963	-0.8358	-0.4393	-0.3519	0.0344
2.00	1.5	-0.1564	-0.1959	-0.3821	-0.0853	-0.0289	0.0024
	2.0	-0.2868	-0.3334	-0.5961	-0.1796	-0.0933	0.0048
	2.5	-0.3881	-0.4318	-0.7198	-0.2651	-0.1666	0.0156
	3.0	-0.4663	-0.5044	-0.7948	-0.3397	-0.2372	0.0160
				$N = 10,000$			
0.50	1.5	-0.5685	-0.6050	-0.7138	-0.4765	-0.4445	0.0116
	2.0	-0.7104	-0.7414	-0.8570	-0.6225	-0.5831	0.0274
	2.5	-0.7788	-0.8031	-0.9125	-0.7011	-0.6611	0.0568
	3.0	-0.8201	-0.8390	-0.9403	-0.7513	-0.7126	0.0707
1.00	1.5	-0.3549	-0.3998	-0.5546	-0.2563	-0.2088	0.0000
	2.0	-0.5175	-0.5615	-0.7494	-0.4034	-0.3401	0.0037
	2.5	-0.6132	-0.6501	-0.8395	-0.5023	-0.4352	0.0032
	3.0	-0.6770	-0.7069	-0.8881	-0.5734	-0.5071	0.0025
1.50	1.5	-0.2330	-0.2770	-0.4544	-0.1469	-0.0929	0.0008
	2.0	-0.3840	-0.4316	-0.6661	-0.2692	-0.1920	-0.0001
	2.5	-0.4871	-0.5294	-0.7774	-0.3654	-0.2800	-0.0001
	3.0	-0.5611	-0.5968	-0.8414	-0.4409	-0.3542	-0.0001
2.00	1.5	-0.1572	-0.1969	-0.3847	-0.0871	-0.0317	0.0010
	2.0	-0.2886	-0.3352	-0.6002	-0.1836	-0.0997	0.0007
	2.5	-0.3890	-0.4326	-0.7236	-0.2695	-0.1735	0.0002
	3.0	-0.4664	-0.5045	-0.7995	-0.3418	-0.2410	0.0001

Table A.5: The relative variance ($RVar(\widehat{N})$) of six estimators with different parameters in the generalised Poisson distribution, when $N = 200, N = 1,000$ and $N = 10,000$

θ	ϕ	Turing	MLEPoi	McK	Chao	Zel	MLEGP
				$N = 200$			
0.50 0.50	1.5	0.004411	0.003629	0.005495	0.023060	0.035563	7.265677
	2.0	0.002123	0.001610	0.002244	0.020211	0.039861	7.329670
	2.5	0.001381	0.001036	0.001339	0.012205	0.025269	6.747437
	3.0	0.001080	0.000801	0.000997	0.017479	0.038211	7.068468
1.00	1.5	0.003541	0.002894	0.004944	0.013927	0.024101	0.735551
	2.0	0.002326	0.001834	0.003326	0.010944	0.022299	2.608979
	2.5	0.001758	0.001373	0.002112	0.011807	0.027788	6.384222
	3.0	0.001503	0.001162	0.001550	0.012012	0.033286	9.369726
1.50	1.5	0.002476	0.002057	0.004271	0.007941	0.015883	0.030520
	2.0	0.002023	0.001607	0.003290	0.008971	0.020429	0.385989
	2.5	0.001841	0.001438	0.002480	0.009683	0.024013	1.592902
	3.0	0.001552	0.001206	0.001850	0.010867	0.028190	5.322908
2.00	1.5	0.001662	0.001397	0.003094	0.004847	0.011579	0.009847
	2.0	0.001803	0.001431	0.003219	0.006764	0.016985	0.043402
	2.5	0.001766	0.001416	0.002686	0.008233	0.022877	0.113472
	3.0	0.001576	0.001296	0.002239	0.008063	0.023399	1.558719
				$N = 1,000$			
0.50	1.5	0.000782	0.000646	0.001152	0.002814	0.004074	1.076208
	2.0	0.000398	0.000306	0.000553	0.001954	0.003433	2.032277
	2.5	0.000290	0.000210	0.000301	0.001657	0.003172	3.186194
	3.0	0.000190	0.000142	0.000189	0.001256	0.002696	4.358476
1.00	1.5	0.000654	0.000536	0.001005	0.002040	0.003436	0.017991
	2.0	0.000466	0.000353	0.000680	0.001855	0.003495	0.057601
	2.5	0.000377	0.000286	0.000461	0.001735	0.003691	0.317794
	3.0	0.000286	0.000219	0.000348	0.001447	0.003350	0.859483
1.50	1.5	0.000451	0.000361	0.000852	0.001447	0.002860	0.005191
	2.0	0.000446	0.000347	0.000748	0.001699	0.003723	0.013594
	2.5	0.000362	0.000281	0.000598	0.001651	0.003969	0.040050
	3.0	0.000329	0.000259	0.000460	0.001694	0.004395	0.068025
2.00	1.5	0.000317	0.000256	0.000684	0.000985	0.002351	0.001719
	2.0	0.000377	0.000302	0.000706	0.001295	0.003259	0.004762
	2.5	0.000350	0.000273	0.000650	0.001549	0.003998	0.012435
	3.0	0.000346	0.000275	0.000506	0.001519	0.004109	0.021662
				$N = 10,000$			
0.50	1.5	0.000080	0.000065	0.000122	0.000281	0.000408	0.016908
	2.0	0.000039	0.000031	0.000059	0.000162	0.000281	0.041395
	2.5	0.000025	0.000018	0.000036	0.000137	0.000263	0.123479
	3.0	0.000021	0.000016	0.000025	0.000102	0.000210	0.204849
1.00	1.5	0.000068	0.000057	0.000112	0.000206	0.000340	0.001746
	2.0	0.000044	0.000034	0.000074	0.000186	0.000360	0.004465
	2.5	0.000038	0.000029	0.000052	0.000171	0.000357	0.008843
	3.0	0.000029	0.000022	0.000041	0.000149	0.000334	0.015790
1.50	1.5	0.000049	0.000040	0.000089	0.000143	0.000283	0.000473
	2.0	0.000042	0.000033	0.000073	0.000153	0.000331	0.001223
	2.5	0.000038	0.000030	0.000068	0.000154	0.000368	0.002523
	3.0	0.000033	0.000026	0.000049	0.000154	0.000384	0.004371
2.00	1.5	0.000035	0.000028	0.000069	0.000096	0.000213	0.000184
	2.0	0.000035	0.000028	0.000074	0.000128	0.000329	0.000471
	2.5	0.000032	0.000026	0.000068	0.000126	0.000341	0.000917
	3.0	0.000035	0.000028	0.000058	0.000146	0.000386	0.001713

Table A.6: The relative variance ($RRMSE(\widehat{N})$) of six estimators with different parameters in the generalised Poisson distribution, when $N = 200, N = 1,000$ and $N = 10,000$

θ	ϕ	Turing	MLEPoi	McK	Chao	Zel	MLEGP
				$N = 200$			
0.50 0.50	1.5	0.5606	0.5949	0.6882	0.4650	0.4418	2.8423
	2.0	0.7025	0.7320	0.8290	0.6022	0.5683	2.8604
	2.5	0.7736	0.7966	0.8888	0.6822	0.6403	2.7213
	3.0	0.8149	0.8331	0.9187	0.7276	0.6875	2.8080
1.00	1.5	0.3521	0.3958	0.5429	0.2583	0.2330	0.8752
	2.0	0.5157	0.5582	0.7325	0.3999	0.3480	1.6602
	2.5	0.6118	0.6476	0.8181	0.4900	0.4288	2.6217
	3.0	0.6727	0.7022	0.8654	0.5549	0.4936	3.2101
1.50	1.5	0.2322	0.2744	0.4436	0.1570	0.1427	0.1765
	2.0	0.3815	0.4276	0.6486	0.2715	0.2226	0.6298
	2.5	0.4844	0.5261	0.7618	0.3598	0.2944	1.2924
	3.0	0.5547	0.5909	0.8204	0.4222	0.3474	2.3822
2.00	1.5	0.1585	0.1959	0.3785	0.1034	0.1086	0.1000
	2.0	0.2867	0.3319	0.5831	0.1880	0.1509	0.2103
	2.5	0.3865	0.4292	0.7060	0.2681	0.2092	0.3430
	3.0	0.4640	0.5019	0.7803	0.3344	0.2583	1.2693
				$N = 1,000$			
0.50	1.5	0.5692	0.6050	0.7108	0.4772	0.4461	1.0575
	2.0	0.7103	0.7405	0.8508	0.6223	0.5833	1.4620
	2.5	0.7773	0.8013	0.9063	0.6973	0.6567	1.8561
	3.0	0.8194	0.8382	0.9349	0.7486	0.7092	2.1657
1.00	1.5	0.3554	0.3997	0.5524	0.2578	0.2133	0.1345
	2.0	0.5168	0.5604	0.7455	0.4038	0.3430	0.2425
	2.5	0.6130	0.6495	0.8342	0.5002	0.4332	0.5701
	3.0	0.6760	0.7059	0.8835	0.5704	0.5043	0.9512
1.50	1.5	0.2334	0.2768	0.4526	0.1500	0.1048	0.0722
	2.0	0.3823	0.4298	0.6620	0.2673	0.1946	0.1172
	2.5	0.4862	0.5285	0.7723	0.3640	0.2822	0.2022
	3.0	0.5610	0.5965	0.8361	0.4412	0.3581	0.2631
2.00	1.5	0.1574	0.1966	0.3830	0.0909	0.0564	0.0415
	2.0	0.2875	0.3338	0.5967	0.1832	0.1094	0.0692
	2.5	0.3885	0.4321	0.7202	0.2681	0.1782	0.1126
	3.0	0.4667	0.5047	0.7952	0.3419	0.2457	0.1480
				$N = 10,000$			
0.50	1.5	0.5686	0.6050	0.7139	0.4768	0.4450	0.1305
	2.0	0.7104	0.7414	0.8571	0.6227	0.5833	0.2053
	2.5	0.7788	0.8031	0.9125	0.7011	0.6613	0.3560
	3.0	0.8201	0.8390	0.9403	0.7514	0.7127	0.4581
1.00	1.5	0.3550	0.3998	0.5547	0.2567	0.2096	0.0418
	2.0	0.5175	0.5615	0.7495	0.4037	0.3407	0.0669
	2.5	0.6133	0.6501	0.8396	0.5025	0.4357	0.0941
	3.0	0.6771	0.7069	0.8881	0.5735	0.5075	0.1257
1.50	1.5	0.2331	0.2771	0.4545	0.1474	0.0945	0.0218
	2.0	0.3840	0.4317	0.6662	0.2694	0.1928	0.0350
	2.5	0.4871	0.5295	0.7774	0.3656	0.2807	0.0502
	3.0	0.5611	0.5968	0.8415	0.4411	0.3548	0.0661
2.00	1.5	0.1573	0.1969	0.3848	0.0877	0.0349	0.0136
	2.0	0.2887	0.3352	0.6003	0.1839	0.1013	0.0217
	2.5	0.3891	0.4326	0.7237	0.2697	0.1744	0.0303
	3.0	0.4664	0.5046	0.7995	0.3420	0.2418	0.0414

References

Abegaz, T., Berhane, Y., Worku, A., Assrat, A., and Assefa, A. (2014). Road traffic deaths and injuries are under-reported in Ethiopia: A capture-recapture method. *PloS one*, 9(7):e103001.

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, 50(2):494–500.

Amstrup, S. C., McDonald, T. L., and Manly, B. F. (2010). *Handbook of Capture-recapture Analysis*. Princeton University Press.

Anan, O., Böhning, D., and Maruotti, A. (2016). Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Statistical Methods & Applications*, 25(76):1–31.

Anderson, D. R. (2008). *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer.

Baksh, M. F., Böhning, D., and Lerdsuwansri, R. (2011). An extension of an over-dispersion test for count data. *Computational Statistics & Data Analysis*, 55(1):466–474.

Ballivet, S., Salmi, L. R., and Dubourdieu, D. (2000). Capture-recapture method to determine the best design of a surveillance system. Application to a thyroid cancer registry. *European Journal of Epidemiology*, 16(2):147–153.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media.

Blumenthal, J. A., Williams, R. B., Kong, Y., Schanberg, S. M., and Thompson, L. W. (1978). Type A behavior pattern and coronary atherosclerosis. *Circulation*, 58(4):634–639.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications*. Taylor & Francis.

Böhning, D. (2008a). A simple variance formula for population size estimators by conditioning. *Statistical Methodology*, 5(5):410–423.

Böhning, D. (2008b). Editorial-Recent developments in capture–recapture methods and their applications. *Biometric Journal*, 50(6):954–956.

Böhning, D. (2010). Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scandinavian Journal of Statistics*, 37(2):221–236.

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., Baksh, M. F., Lerdsuwansri, R., and Gallagher, J. (2013a). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics*, 22(1):135–155.

Böhning, D., Dietz, E., Kuhnert, R., and Schön, D. (2005). Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14(1):29–43.

Böhning, D., Rocchetti, I., Alfó, M., and Holling, H. (2016). A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics*, -(-):Available from: DOI: 10.1111/biom.12485.

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):721–737.

Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., and Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture–recapture approach using repeated entries in one list. *European Journal of Epidemiology*, 19(12):1075–1083.

Böhning, D. and Van der Heijden, P. G. (2015). Correspondence: some general points regarding Ledberg and Wennberg, BMC Medical Research Methodology 2014 April 27; 14: 58. *BMC Medical Research Methodology*, 15(51):1–5.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. (2013b). A generalization of Chao's estimator for covariate information. *Biometrics*, 69(4):1033–1042.

Böhning, D. and Vilas, V. J. D. R. (2008). Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological, and Environmental statistics*, 13(1):1–22.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations*. Springer Science & Business Media.

Brenner, H. (1995). Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*, 6(1):42–48.

Brittain, S. and Böhning, D. (2009). Estimators in capture–recapture studies with two sources. *AStA Advances in Statistical Analysis*, 93(1):23–47.

Bronner, A., Hénaux, V., Vergne, T., Vinard, J.-L., Morignat, E., Hendrikx, P., Calavas, D., and Gay, E. (2013). Assessing the mandatory bovine abortion notification system in France using unilist capture-recapture approach. *Plos One*, 8(5):e63246.

Buckland, S. T. (1984). Monte carlo confidence intervals. *Biometrics*, 40(3):811–817.

Buckland, S. T. and Garthwaite, P. H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, 47(1):255–268.

Bulmer, M. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30:101–110.

Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application*, 1:427–445.

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.

Carothers, A. (1973). Capture-recapture methods applied to a population with known parameters. *Journal of Animal Ecology*, 42(1):125–146.

Cecala, K. K., Price, S. J., and Dorcas, M. E. (2012). Modeling the effects of life-history traits on estimation of population parameters for a cryptic stream species. *Freshwater Science*, 32(1):116–125.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45(2):427–438.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, 58(3):531–539.

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217.

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.

Consul, P. and Jain, G. (1973a). On some interesting properties of the generalized Poisson distribution. *Biometrische Zeitschrift*, 15(7):495–500.

Consul, P. and Jain, G. C. (1973b). A generalization of the Poisson distribution. *Technometrics*, 15(4):791–799.

Consul, P. and Shoukri, M. (1985). The generalized Poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics-Simulation and Computation*, 14(3):667–681.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12(2):132–136.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45(2):395–413.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1):1–38.

Dubey, S. D. (1966). The teacher's corner: graphical tests for discrete Distributions. *The American Statistician*, 20(3):23–24.

Edwards, W. R. and Eberhardt, L. (1967). Estimating cottontail abundance from live-trapping data. *The Journal of Wildlife Management*, 31(1):87–96.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447.

Engel, J. (2010). On teaching bootstrap confidence intervals. In *Proceedings of the Eighth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute.*

Farcomeni, A., Scacciatelli, D., et al. (2013). Heterogeneity and behavioral response in continuous time capture–recapture, with application to street cannabis use in Italy. *The Annals of Applied Statistics*, 7(4):2293–2314.

Guikema, S. D. and Coffelt, J. P. (2008). Modeling count data in risk analysis and reliability engineering. In *Handbook of Performability Engineering*. Springer London.

Gupta, R. C., Sim, S., and Ong, S. (2014). Analysis of discrete data by conway–maxwell poisson distribution. *AStA Advances in Statistical Analysis*, 98(4):327–343.

Héraud-Bousquet, V., Lot, F., Esvan, M., Cazein, F., Laurent, C., Warszawski, J., and Gallay, A. (2012). A three-source capture-recapture estimate of the number of new hiv diagnoses in children in france from 2003–2006 with multiple imputation of a variable of heterogeneous catchability. *BMC Infectious Diseases*, 12(251):1–9.

Holmes, J. and Haggett, P. (1977). Graph theory interpretation of flow matrices: a note on maximization procedures for identifying significant links. *Geographical Analysis*, 9(4):388–399.

Huggins, R. and Hwang, W.-H. (2007). Non-parametric estimation of population size from capture–recapture data when the capture probability depends on a covariate. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4):429–443.

Huggins, R. and Hwang, W.-H. (2011). A Review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review*, 79(3):385–400.

Hwang, W.-H. and Huggins, R. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92(1):229–233.

Imoto, T. (2014). A generalized Conway–Maxwell–Poisson distribution which includes the negative binomial distribution. *Applied Mathematics and Computation*, 247:824–834.

Jacquinet, S., Denis, O., Soares, F. V., and Schirvel, C. (2015). Legionnaires' disease: overview of the situation concerning notification in Wallonia (Belgium) in 2012, a retrospective descriptive study based on a capture-recapture method. *Archives of Public Health*, 73(2):1–6.

Kake, T. R., Arnold, R., and Ellis, P. (2008). Estimating the prevalence of schizophrenia among New Zealand M a ori: a capture–recapture approach. *Australian and New Zealand Journal of Psychiatry*, 42(11):941–949.

Karlis, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. *Astin Bulletin*, 35(01):3–24.

Köse, T., Orman, M., Ikiz, F., Baksh, M., Gallagher, J., and Böhning, D. (2014). Extending the Lincoln–Petersen estimator for multiple identifications in one source. *Statistics in Medicine*, 33(24):4237–4249.

Kuhnert, R. and Böhning, D. (2009). CAMCR: Computer-Assisted Mixture model analysis for Capture–Recapture count data. *AStA Advances in Statistical Analysis*, 93(1):61–71.

Lanumteang, K. (2010). *Estimating of Size of a Target Population Using Capture-recapture Methods based upon Multiple Sources and Continuous Time Experiments*. PhD thesis, University of Reading.

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 & Data Analysis*, 55(7):2302–2311.

Lerdsuwansri, R. (2012). *Generalisation of the Lincoln-Petersen approach to Non-binary Source Variables*. PhD thesis, University of Reading.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130.

Liu, G., Rong, G., Zhang, H., and Shan, Q. (2015). The adoption of capture-recapture in software engineering: a systematic literature review. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, page 15. ACM Digital Library.

LuValle, M. (1990). Generalized Poisson Distributions: Properties and Applications. *Technometrics*, 32(3):346–347.

Mao, C. X. and Lindsay, B. G. (2003). Tests and diagnostics for heterogeneity in the species problem. *Computational Statistics & Data Analysis*, 41(3):389–398.

McCrea, R. S. and Morgan, B. J. (2014). *Analysis of capture-recapture data*. CRC Press.

McKendrick, A. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Meurant, G. (1992). A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM Journal on Matrix Analysis and Applications*, 13(3):707–728.

Morgan, B. J. and Ridout, M. S. (2008). A new mixture model for capture heterogeneity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(4):433–446.

Navaratna, W., del Rio Vilas, V. J., and Böhning, D. (2008). Extending Zelterman's approach for robust estimation of population size to zero-truncated clustered data. *Biometrical Journal*, 50(4):584–596.

Niwitpong, S.-a., Böhning, D., Van der Heijden, P. G., and Holling, H. (2013). Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika*, 76(4):495–519.

Norris III, J. L. and Pollock, K. H. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, 3(3):235–244.

Ord, J. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society. Series A (General)*, 130(2):232–238.

Regal, R. R. and Hook, E. B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, 10(5):717–721.

Rocchetti, I., Alfó, M., and Böhning, D. (2014). A regression estimator for mixed binomial capture–recapture data. *Journal of Statistical Planning and Inference*, 145:165–178.

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.

Scholz, F. (1985). Maximum likelihood estimation. *Encyclopedia of Statistical Sciences*.

Schwarz, C. J. and Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, 52(3):860–873.

Seber, G. (1970). The effects of trap response on tag recapture estimates. *Biometrics*, 26(1):13–22.

Seber, G. A. F. (2002). *The Estimation of Animal Abundance and Related Parameters*. The Blackburn Press.

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. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142.

Spoor, P., Airey, M., Bennett, C., Greensill, J., and Williams, R. (1996). Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ*, 313(7053):342–343.

Stoklosa, J., Hwang, W.-H., Wu, S.-H., and Huggins, R. (2011). Heterogeneous capture–recapture models with covariates: A partial likelihood approach for closed populations. *Biometrics*, 67(4):1659–1665.

Telang, R., Boatwright, P., and Mukhopadhyay, T. (2004). A mixture model for Internet search-engine visits. *Journal of Marketing Research*, 41(2):206–214.

Thompson, W. (2013). *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters*. Island Press.

Tounkara, F. and Rivest, L.-P. (2015). Mixture regression models for closed population capture–recapture data. *Biometrics*, 71(3):721–730.

Trinh, G., Rungie, C., Wright, M., Driesener, C., and Dawes, J. (2014). Predicting future purchases with the Poisson log-normal model. *Marketing Letters*, 25(2):219–234.

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.

Van der Heijden, P. G., Cruyff, M., and Böhning, D. (2014). Capture recapture to estimate criminal populations. In *Encyclopedia of Criminology and Criminal Justice*, pages 267–276. Springer.

Van Hest, N., Grant, A., Smit, F., Story, A., and Richardus, J. H. (2008). Estimating infectious diseases incidence: validity of capture–recapture analysis and truncated models for incomplete count data. *Epidemiology and Infection*, 136(01):14–22.

Vergne, T., Grosbois, V., Durand, B., Goutard, F., Bellet, C., Holl, D., Roger, F., and Dufour, B. (2012). A capture–recapture analysis in a challenging environment: assessing the epidemiological situation of foot-and-mouth disease in Cambodia. *Preventive Veterinary Medicine*, 105(3):235–243.

Vidal-Diez, A. (2015). *Development of Capture-recapture Estimators in Closed Populations Including Individual Covariate Information.* PhD thesis, University of Southampton.

Vilas, V. D. R. and Böhning, D. (2008). Application of one-list capture–recapture models to scrapie surveillance data in Great Britain. *Preventive Veterinary Medicine*, 85(3):253–266.

Viwatwongkasem, C., Kuhnert, R., and Satitvipawee, P. (2008). A comparison of population size estimators under the truncated count model with and without allowance for contaminations. *Biometrical Journal*, 50(6):1006.

Vuillermoz, C., Aouba, A., Grout, L., Vandentorren, S., Tassin, F., Vazifeh, L., Ghosn, W., Jougla, E., and Rey, G. (2014). Estimating the number of homeless deaths in France, 2008–2010. *BMC Public Health*, 14(1):1.

Wannasirikul, N. (2005). *A Comparison of Truncated Poisson Estimators of Population Size under Model Contaminations.* PhD thesis, Mahidol University.

Wimmer, G., Köhler, R., Grotjahn, R., and Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1(1):98–106.

Xu, Y., Fyfe, M., Walker, L., and Cowen, L. L. (2014). Estimating the number of injection drug users in greater Victoria, Canada using capture-recapture methods. *Harm Reduction Journal*, 11(9):1–7.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18(2):225–237.

Zwane, E. and Van der Heijden, P. G. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & Probability Letters*, 65(2):121–125.