UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Capture-Recapture Estimation and Modelling for One-Inflated Count Data

by

Panicha Kaskasamkul

Thesis for the degree of Doctor of Philosophy

March 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

CAPTURE-RECAPTURE ESTIMATION AND MODELLING FOR ONE-INFLATED COUNT DATA

by Panicha Kaskasamkul

Capture-recapture methods are used to estimate the unknown size of a target population whose size cannot be reasonably enumerated. This thesis proposes the estimators and the models specifically designed to estimate the size of a population for one-inflated capture-recapture count data allowing for heterogeneity. These estimators can assist with overestimation problems occurring from one-inflation that can be seen in several areas of researches. The estimators are developed under three approaches.

The first approach is based on a modification by truncating singletons and applying the conventional Turing and maximum likelihood estimation approach to the one-truncated geometric data for estimating the parameter p_0. These $\hat{p}_0$ are applied to the Horvitz-Thompson approach for the modified Turing estimator (T_OT) and the modified maximum likelihood estimator (MLE_OT).

The second approach is the model-based approach. It focuses on developing a statistical model that describes the mechanism to generate the extra of count ones. The new estimator MLE_ZTOI is developed from a maximum likelihood approach by using the nested EM algorithm based upon the zero-truncated one-inflated geometric distribution

The last approach focuses on modifying a classical Chao's estimator to involve the frequency of counts of twos and threes instead of the frequency of counts of ones and twos. The modified Chao estimator (MC) is asymptotic unbiased estimator for a power series distribution with and without one-inflation and provides a lower bound estimator under a mixture of power series distributions with and without one-inflation. The three bias-correction versions of the modified Chao estimator have been developed to reduce the bias when the sample size is small. A variance approximation of MC and MC3 are also constructed by using a conditioning technique.

All of the proposed estimators are assessed through simulation studies. The real data sets are provided for understanding the methodologies.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Panicha Kaskasamkul , declare that the thesis entitled *Capture-Recapture Estimation and Modelling for One-Inflated Count Data* 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: Böhning et al. (2017)

Signed:.....

Date:.....

Acknowledgements

I would like to thank all the people who contributed in some way to the work described in this thesis.

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Dankmar Böhning, for accepting me into his research group and giving me opportunity to study in university of Southampton. During my tenure, he supported my study and research continuously. He contributed to a rewarding graduate school experience by giving me intellectual freedom in my work, supporting my attendance at various conferences and engaging me in new ideas. I really appreciate his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor for my PhD study.

Besides my supervisor, I would like to thank my thesis committee, Professor Herwig Friedl and Professor Peter G M Van der Heijden, for their encouragement, insightful comments, and suggestions.

Besides my supervisor, I would like to thank all colleagues in the School of Mathematical Sciences for their help and support. I would also like to extend a great deal of appreciation to my friends, Dr. Natchalee Srimaneekarn, Dr. Kim Lee and Dr. Olga Egorova for their constant support and friendship.

I am greatly thankful to my parents, Thana and Busarin Kaskasamkul, my family especially Wiroonwat Charanpattanapon for their constant love, understanding and emotional support. They have been the inspiration for all that I have achieved so far.

Finally, I gratefully acknowledge the Ministry of Science and Technology and the Royal Thai Government for the financial support for studying in the UK.

Chapter 1

Introduction

This chapter outlines the context of study and the objectives. The structure of thesis is also shown in the final section.

1.1 Introduction

Estimation of the size of an elusive target population is of great interest in several areas such as biology, ecology, epidemiology, public health and social science. For example, ecologists and biologists might investigate the number of species in a wildlife population as well as to estimate the animal abundance. In social science the interest is in determining the number of drug users and the number of violators of a law. In addition, there is a great interest in estimating the number of outbreaks of disease in public health.

Capture-recapture methods have been applied to estimate the size of populations which are difficult to approach. They have a long history and were traditionally applied in wildlife, biology and ecology to estimate the animal abundance and the size of wildlife populations. To estimate the population size N, capture-recapture surveys are conducted by using an identifying mechanism. Each individual is noted of presence or absence. For example, capture-recapture methods use the information available from animal captured on a number of surveys. Animals in a trap are marked, released and allowed to mix with the population. After a period of time, a second survey is taken and the number of animals captured is counted and marked again. Repeated surveys are carried out and the number of animals being marked from all surveys are obtained as the capture-recapture history. This provides the observed frequency of identifying individuals. Accordingly, the capture history is used to estimate the total population size or the number of cases which are never caught at any occasion. Typically, the survey is within a short period so that evolution of new cases or extinction of existing cases

is unlikely to occur during the study period. This is referred to as the case of a closed population. These concepts have been applied to human populations in social science and criminology to estimate, for example, the size of an illicit drug-using population or the number of violators of a law (see Van der Heijden et al. (2003b) and Hser (2001)), in public health science for estimating the disease prevalence (see Gallay et al. (2000) and Böhning et al. (2004)) and to estimate the number of unreported diseases, as well as infection rate of AIDS in epidemiology (Brookmeyer and Gail (1988)), also estimating the number of unknown errors in a software in system engineering (see Liu et al. (2015)). In these situations, the population size can be determined by using a number of different sources (lists) as a survey occasion or identifying mechanism such as hospital lists, treatment center registries or pharmacy records. This similarity to wildlife capture-recapture is in these cases that the role of the trap is taken by the register (cancer occurrence), the police (violations of a law) or the reviewer (software error). Typically the number of cases that does not appear in either list is unknown and need to be estimated (Brittain and Böhning (2009)).

From the capture-recapture history, a count x as the number of individuals identified exactly x times is obtained. A counting distribution arises when a frequency table is constructed from summarizing how often a particular individual was identified. This is usually referred to as capture-recapture data in the form of *frequencies of frequencies*. However, some individuals do not appear since they have never been identified so the zero count data are missing. According to Böhning and Kuhnert (2006), zero truncation arises regularly in many practical situations. There are underlying two different categories of zero truncation:

- Zero counts cannot occur because of the observational model such as counts of occupants of passing cars. A telephone interviewing survey asking for the number of telephones in the household will have only non-zero counts in a sample (Grogger and Carson (1991); Cameron and Trivedi (2013); Winkelmann (2008)).
- *Zero-truncated count models* are normally used in capture-recapture studies. From the capture-recapture history of the individual, we try to predict the frequency of units missed by the sample. For example, suppose that the police is keeping records on the number of times a person has been identified with deviant behavior in a particular community. It is clear that deviant persons who have never been identified will not be present in database. A truncated count model can be used to predict this quantity. Furthermore, it can be found in other applications such as the estimation of drug users in a community or the number of illegal workers in labour studies. This situation has been investigated with emphasis on the Poisson distribution (Böhning and Kuhnert (2006); Van der Heijden et al. (2003b); see also the review of Bunge and Fitzpatrick (1993)).

The frequency count data is $\{(x, f_x) \mid x \geq 1\}$ where f_x is the frequency of individuals captured exactly x times. Consequently, the frequency distribution is a zero-truncated count distribution. Based upon a zero-truncated model, it is assumed that all individuals in the population of interest have the same parameter determining probabilities to be captured once, twice and so on. This is defined as the case of *homogeneity* and often modelled by the Poisson or binomial distribution (see the review of Bunge and Fitzpatrick (1993)). The parameter is unknown and can be estimated by various methods. If an estimate of the parameter is derived, then the probability of zero counts is obtained leading to an estimate of the hidden as well as of the total size of population. However, the homogeneous model rarely holds in practice because of the fact that the population frequently composes of various subpopulations. Each subpopulation has the same distribution but different parameters. This case is the so-called *heterogeneity* case. Capture probabilities under a heterogeneous model are likely to differ for each individual. Approaches that take into account heterogeneous models are introduced by Chao (1987), Zelterman (1988) and Chao and Bunge (2002). The problem of heterogeneity should not be ignored as it can cause severe underestimation of the true population size (see Van der Heijden et al. (2003a) and Böhning and Schön (2005)).

There have been many statistical models developed for estimating the population size N. The classical modelling approach stems from the Lincoln-Petersen approach which uses the independent information of two identifying sources or lists in closed populations (Seber (2002)). In this model, each source provides a binary variable taking value 1 for presence and 0 for absence. Capture and recapture samples are then formed in a 2x2 contingency table. Finally, the Lincoln-Petersen's estimator can be constructed by multiplying the number of individuals found on each source and dividing the outcome by the number of individuals identified by both sources. Throughout the years, the numerous models and estimators were developed and proposed to improve inferences in capture-recapture studies which always rely on certain assumptions but are violated in real situations due to time effect, heterogeneity or behavioural response among others. Some examples include the maximum likelihood estimator, Good-Turing estimator, Zelterman's estimator (Zelterman (1988)) and Chao's lower bound estimator (Chao (1987)).

In some capture-recapture studies, we can notice from the observed data that there is some sort of *one-inflation* in the count distribution (see e.g. Farcomeni and Scacciatelli (2013)). Some portion of the population is mostly captured only once. This may be a consequence of the fact that the probability of recapturing the same individual is very low, especially in large cities/areas and generally within a short period of survey. Secondly, the first capture can lead to a behavioral response for some individuals to no longer be observed. For example, individuals are stressed from the first capture and learn to avoid recapture further on. Under a serious law enforcement, more serious legal penalties are expected after the second time an individual is reported as perpetrator.

Individuals may be abrogated their driver's licence, pay a fine and/or take part in treatment programs or entry visa may be invoked for foreigners. In contrast, an individual may get as consequence only a warning by the judge if they are identified the first time. It is not surprising if individuals may show *trap avoidance* after the first capture. Thirdly, the frequency of count one (singleton) may not be reliably observed in some applications such as in microbial diversity. One-inflation arises, especially, in data derived from modern high-throughput DNA sequencing. A new taxa may be assigned incorrectly from the error of sequences instead of being matched to the observed taxa. This leads to an artificially inflated frequency of count one as shown in terms of one-inflation (see Bunge et al. (2012)). As the result of one-inflation being present in the count data, some models suffer from a boundary problem when fitted and some estimators provide extreme overestimation of the population size (see Godwin (2017)), particularly for Chao's lower bound estimator which seemingly adjusted for heterogeneity.

Research in this thesis focuses on models specifically designed to estimate the size of a population for one-inflated capture-recapture count data allowing for heterogeneity. This provides new estimators based upon their suitable distributions. The modified Chao estimator and its variance are also presented as a new version of classical Chao's lower bound estimator based on non-parametric approach for one-inflation.

1.2 Basic assumptions for this thesis

1. The target population is in a closed system (closed population; no births, no deaths and no migration).
2. Individuals are sampled independently.
3. Repeated identification occurs independently (This assumption will be relaxed in certain cases).

1.3 Aims and objectives of the study

The aim of this research is to develop models and estimators to estimate population size which take into account potential one-inflation. There are a number of objectives that must be realized in order to achieve this aim.

1. To motivate the one-inflation problem in capture-recapture studies.
2. Using the ratio plot, to investigate the presence of one-inflation in the capture-recapture count data.

3. To investigate the performance of conventional estimators when prone to one-inflation.

4. To show that conventional approaches, seemingly good approaches adjusting for potential heterogeneity such as Chao's lower bound estimator, fail drastically when frequency counts of ones (singleton) is excess.

5. To develop a distributional model for counts with one-inflation.

6. To develop new estimators for estimating the size of population under the one-truncated geometric distribution.

7. To develop a full distributional approach under the geometric with one-inflation for estimating population size.

8. To develop a new modified Chao estimator and its variance approximation for the case of one-inflated count data.

9. To develop bias-correction versions for the modified Chao estimator.

1.4 Outline of thesis

The thesis consists of seven chapters. The first chapter begins by introducing the context, research objectives and the outline of the study. The remaining chapters are given according to respective context as follows:

In Chapter 2, a literature review of the capture-recapture approach is presented, particularly with emphasis on well-known estimators of population size that are used in capture-recapture methodology. Then, the examples of count data with one-inflation and the ratio plots are given in the following section.

Chapter 3 shows the weak performance of some classical estimators under one-inflation. The construction of new estimators based on the one-truncated geometric model is examined. This chapter also provides an investigation of relative bias, relative variance and relative mean square error of estimators under a variety of simulated conditions. Several empirical applications are also considered in order to illustrate the use of the new proposed estimators in real life situations.

In Chapter 4, the problem formulation and derivation of the proposed estimators based upon zero-truncated one-inflated geometric distribution are presented. In addition, the nested EM algorithm for estimating parameter of the model is also discussed. Finally, a simulation study is included to study the performance of the proposed estimators under a one-inflation problem and also illustrate these estimators in a variety of case studies.

Chapter 5 shows the construction of the new proposed estimator for one-inflation count data by extending the idea of Chao's lower bound estimator. The bias correction versions of the modified Chao estimator are given to improve the performance when the sample size is small. A variance approximation of the modified Chao estimators is also presented in Chapter 6. Performance evaluations of the new proposed estimators are undertaken under the models of geometric and mixture of geometric with and without one-inflation. This chapter ends with some applications.

Finally, Chapter 7 gives some concluding remarks and suggests potential directions for related future research.

1.4.1 Notation and definitions

There are many parameters and statistics involved in statistical methods for capture-recapture study using throughout in this thesis. Therefore, in order to easily understanding the statistical terms, some general notations and definitions are arranged as follows:

N the unknown population size of the target population
$\hat{N}$ the estimator of the size of the target population
m the total number of trapping occasions over the study period
n the total number of distinct observed individuals or the number of sample units
X_{ij} the indicator variable of the i^{th} unit being identified in the j^{th} occasion,

$$\text{where} \quad X_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ individual is identified on the } j^{th} \text{ occasion} \\ 0 & \text{otherwise} \end{cases}$$

X_i the number of times that the i^{th} individual was identified over the study period
p_x the capture probability of individuals that were identified exactly x times
f_x the frequency of identifying individuals exactly x times over the study period
f_0 the frequency of unobserved individuals
S the total number of identification during study period

Chapter 2

Review of capture-recapture

In this chapter, a review of the general background of capture-recapture is provided. It contains objective, basic idea, assumptions and characteristic of count data in capture-recapture, including model classification and summary of estimating the size of a target population. Some interesting estimators of population size are considered under homogeneous and heterogeneous Poisson models such as maximum likelihood, Chao's lower bound and Good-Turing estimator. This is followed by the examples of use and application of capture-recapture method of which some may have one-inflation form. The graphical device of the ratio plot as a tool to investigate the validity of the model is shown in the last section.

2.1 Objective and basic idea of capture-recapture

The capture-recapture method has a long history. It was developed to improve the limitation of census when we cannot take a complete census of the entire units in the target population. Some units are detected but some remain hidden or undetected such as sampling in a wildlife population, a human population with illicit habit or a human population with a disease which is hard to detect. The capture-recapture approach has been traditionally applied in wildlife, biology and ecology to estimate the animal abundance and the size of a wildlife population. The basic idea of capture-recapture is to sample or capture individuals, mark and identify, release and allow to mix with the population and then, on a second survey, recapture individuals, count and mark again. After that, the number of individuals are noted which have already been marked on the first sampling occasion. This capture-recapture method can continue to m surveys. The number of animals being marked from all surveys are obtained as the *capture-recapture history*. This provides the observed frequency of identified units. Correspondingly, the observed frequencies from the capture history are used to estimate the total population size N or the number of units which are never caught at any occasion, f_0. However,

at present, the capture-recapture approach is widely applied in a variety of other fields such as estimating the number of outbreaks of disease in public health and epidemiology, estimating biodiversity in bioinformatics, estimating the number of drug users, the size of homeless populations and the number of violators of a law in social science and criminology, as well as estimating the number of unknown errors in a software in system engineering (Böhning et al. (2013b)). Registration can be conducted to create a list of units in the particular population of interest. In clinical studies, for example, AIDS registries contain the number of contacts during diagnosis, treatment and after-care. Recording might fail if patients keep away from the process of remedy. As a result, the registries are incomplete and show only some part of the population. Questions arise about the total number of units in the population and the number of missing units. From this situation, under-reporting arises since the number of units reported is less than the actual number. One might analyze under-reporting using a binomial approach (see Cameron and Trivedi (2013)), regression approaches for the binomial model and the beta-binomial model (see Neubauer and Friedl (2006)) or a beta-Poisson regression model by Neubauer et al. (2011). However, the capture-recapture method is also an efficient tool to estimate the population size for these cases. Some mechanisms are used to identify a repeated unit such as register, surveillance system and life trapping. Each source is treated as a survey occasion. For example, animals were repeatedly captured using traps and the number of animals captured on trapping was reported (Otis et al. (1978)). In case of human populations, a registration system could be used as an identifying mechanism to identify units having a characteristic of interest such as police database recorded the number of illegal immigrants in the Netherlands (Van der Heijden et al. (2003a)) or drug treatment centre and arrest records collects the number of illicit drug users (Hser (2001)). Moreover, this recapturing works either in time or in cluster. In time, there is the period of observation in which each individual of the target population can be detected on several occasions. On the other hand, in cluster, recapturing base on multiple detection within a cluster such as a household, village, or herd. In a fixed observational period, n sample units are independently observed by a given registration. Finally, the capture-recapture history and the observed frequencies are provided for estimating the size of population and the number of unobserved units.

From the capture-recapture history, a distribution of counts arises when a frequency table is constructed from summarizing how often a particular unit was identified. This is usually referred to as capture-recapture data in the form of *frequencies of frequencies*. However, some units have never been identified so the zero count data are missing and this is called *zero-truncated count data*. The frequency count data is $\{(x, f_x) \mid x \geq 1\}$ where f_x is the frequency of units captured exactly x times. Consequently, the frequency distribution is a *zero-truncated count distribution* which is defined by a conditional probability function. It is frequently used to model frequency data (see McCrea and Morgan (2014)). Let $Pr(X = x) = p_x$ denote the capture probability for random variable X to take on value x or the probability of a unit being caught exactly

x times in any trap or appearing exactly x times on any registry. $Pr(X|X > 0)$ or p_x^+ is the conditional probability of $X = x$ given $X > 0$. It can be formulated as

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

or it can be written as

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

For example, suppose that the random variable X has a Poisson distribution. As f_0 is unknown and missing, the distribution is truncated at $x = 0$. Therefore, the zero-truncated Poisson probability function is

$$Pr(X = x|X > 0, \lambda) = \frac{Pr(X = x; \lambda)}{Pr(X > 0; \lambda)} = \frac{e^{-\lambda}\lambda^x}{x!(1 - e^{-\lambda})}. \tag{2.3}$$

In practice, the capture probability is not necessary the same for all units. There are possible sources of variation in the probability such as factors of age, social status or effects of weather. Suppose that p_0 is defined as the probability of a zero count (unobserved units) and the probability of unit identified is given by $1 - p_0$. It is assumed that every unit has the same probability of being identified. The population size N composes of the number of unobserved and observed units.

$$\begin{aligned} N &= \underbrace{N(1 - p_0)}_{\text{observed}} + \underbrace{Np_0}_{\text{unobserved}} \\ &= n + Np_0. \end{aligned} \tag{2.4}$$

The observed part $N(1-p_0)$ can be estimated by the number of observed units n where $E(n) = N(1-p_0)$. p_0 is unknown and required to be estimated, thereby the population size N in (2.4) can easily be solved and leads to the well-known Horvitz-Thompson estimator in (2.5)

$$\hat{N} = \frac{n}{1 - \hat{p}_0}. \tag{2.5}$$

2.2 Assumption

Almost all statistical theories generally require some assumptions for their models. Also, capture-recapture requires essential assumptions:

1. **Closed population:** It is assumed that there is no change to the population during the investigation. That is no birth, no death and no migration. This means that the population size remains constant throughout the study periods.

2. **Independence between subjects:** All individuals have the same probability of being captured in each trapping occasion. That is there is no dependence between different subjects.

3. **Independence between captures:** Lists or sources identify independently and repeated identification occurs independently. That is capture in the first sample does not affect capture in the second sample: samples are independent.

4. **Homogeneity of capture probability:** For a given source, every case has the same chance of being captured or it is called *equal catchability*.

These assumptions are important and should not be violated because it can affect estimation.

- If the *closed population* assumption is violated, individuals found in the first sample may not be possible to be found in the other sample. This reduces the probability of recapture and will lead to an overestimation of N.
- In case of *independence*, if there is *positive dependence*, N will be underestimated. Contrarily, if there is *negative dependence*, N will be overestimated.
- Individuals found in both samples must be *reliably identified and matched*. If true matches are missed, the number of recaptured units is falsely reduced leading to an overestimation of N. If false matches are created, the number of recaptured units is falsely increased leading to an underestimation of N.
- In terms of *equal catchability* assumption, if some individuals have a low probability of being found by either method, N will be underestimated.

2.3 Data structure of capture-recapture

Capture-recapture studies substantially need to sample observed units from a target population at least two times. The raw data are the capture records of all units identified during the study periods. The classical model of capture-recapture is a single marking study which has only two trapping occasions(two lists/samples). Furthermore, the multiple marking and multiple sources study are more complex by allowing to capture units more than twice. For all models, the variable of interest is the frequency count of identified units (f_x) and the latter is used to estimate the total number of a target population (N).

2.3.1 Single mark or two sources

The simplest model of a capture-recapture methodology is the Lincoln-Petersen model, also known as *dual systems estimation* used in the two sources situation. Based on this model, the Lincoln-Petersen estimator for the unknown population size (N) can be considered on the basis of the odds ratio (Brittain and Böhning (2009)) or the hypergeometric distribution (Seber (2002)). There are two trapping occasions and it can be summarised in 2x2 contingency table with frequencies given in Table 2.1.

Table 2.1: The two-occasion situation

		Occasion 1		
		1	0	
Occasion 2	1	f_{11}	f_{01}	n_2
	0	f_{10}	$f_{00} =?$	
		n_1		

where

- f_{11} denotes the frequency of individuals identified at both occasion.
- f_{10} denotes the frequency of individuals identified once at the 1^{st} occasion.
- f_{01} denotes the frequency of individuals identified once at the 2^{nd} occasion.
- f_{00} denotes the frequency of unobserved individuals.
- n_i denotes the number observed in source i; $i = 1, 2$.

Two main steps of sampling:

1. At the 1^{st} occasion, some units are caught as a first sample of size n_1 from the target population. Then, all of the sampled units are marked or indicated uniquely for future recapturing occasion. After that they are released back to mix with the population. So the marked proportion is $\frac{n_1}{N}$.
2. After some time, has elapsed a second sample of size n_2 is chosen and it is clear that this second sample composes of a number of marked units (f_{11}), and unmarked units (f_{01}), where $f_{11} + f_{01} = n_2$. Hence, the second marked proportion is $\frac{f_{11}}{n_2}$.

Let $m_2 = f_{11}$ be the number of observed individuals in both occasions. Under the assumption of *independence*, the proportion of marked individuals in the second sample is equal to the population proportion of marked individuals $\frac{m_2}{n_2} = \frac{n_1}{N}$, the Lincoln-Petersen estimator of N, can be obtained as:

$$\hat{N}_{\text{LP}} = \frac{n_1 n_2}{m_2} \tag{2.6}$$

and the variance of this estimator is provided as:

$$Var(\hat{N}_{\text{LP}}) = \frac{n_1 n_2 (n_1 - m_2)(n_2 - m_2)}{m_2^3} \tag{2.7}$$

Although the Lincoln-Petersen estimator is a simple approach, this method requires essential assumptions as follows:

1. The population is closed. It means the number of target population (N) is constant.
2. It does not matter that each individual is marked, all individuals have the same chance of being captured in each trapping occasion.
3. Marks are not lost and each individual is correctly identified on both occasions and successfully matched.
4. It does not matter that each individual is marked, there are no effects on individual's chances of being caught, so capture sample and recapture sample are independent.

However, the Lincoln-Petersen estimator has the drawback that if there is no overlap between sources or no marked individuals are trapped on the second occasion ($m_2 = 0$), the Lincoln-Petersen estimator for population size cannot be computed. A modified form of this estimator is the Chapman estimator, which is giving by:

$$\hat{N}_{\text{CPM}} = \frac{(n_1 + 1)(n_2 + 1)}{(m_2 + 1)} - 1. \tag{2.8}$$

This estimator is less affected by small value of m_2 and less biased than the Lincoln-Petersen estimator. A variance estimate of the estimator is given in (2.9):

$$Var(\hat{N}_{\text{CPM}}) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_2 - m_2)}{(m_2 + 1)^2 (m_2 + 2)}. \tag{2.9}$$

Example 1: Two sources. Sekar and Deming (1949) used the capture-recapture method to estimate the birth and death rates for residents of an area known as the Singur Health Centre, near Calcutta, India by using two lists; 1) the registration list (R) and 2) the interviews list (I) obtained from a complete house to house canvass. The data can be represented with 2x2 contingency table as in Table 2.2.

Table 2.2: Number of death in the Singur Health Centre

	I-List		Total
R-List	Present	Absent	
Present	439	427	866
Absent	421	f_{00}	
Total	860		N

The estimate of the total number of deaths is $\hat{N}_{\text{LP}} = \dfrac{(860)(866)}{439} = 1,596.$

2.3.2 Multiple mark

Multiple mark capture-recapture methodology is simply defined by the fact that the target population is sampled more than two times over the period of study, also known as *multiple systems estimation*. Suppose that m denotes the number of trapping occasions over a period of study and let N be the size of target population, so that each individual is indexed from $1, 2, 3, ..., n, n+1, ..., N$. It is assumed that all trapping occasions j have all individuals in the population available for trapping, $j = 1, 2, 3, ..., m$, due to the assumption of closed population. Hence, X_{ij} is the indicator variable of individual i observed on occasion j where

$$X_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ individual is identified on the } j^{th} \text{ occasion} \\ 0 & \text{otherwise.} \end{cases}$$

The capture-recapture history can be arranged in a matrix X and $X = [X_{ij}]_{N\times m}$ or it can easily be presented in the form of Table 2.3

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \\ x_{(n+1)1} & x_{(n+1)2} & x_{(n+1)3} & \cdots & x_{(n+1)m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{Nm} \end{bmatrix}_{N\times m} .$$

The matrix X composes of N rows corresponding to the individuals and m columns for all trapping occasions. It consists of only the values of zeros and ones indicating unidentified and identified individuals, respectively, during the study period. The first n rows relate to the capture-recapture history of each individual that is caught at least once over the given study period. The remaining rows $(n+1, n+2, ..., N)$, which contain only zeros, are unobserved and the number of these rows is unknown.

Table 2.3: Capture-recapture history

Individual i	Occasion j					$X_i = \sum_{j=1}^{m} X_{ij}$
	1	2	3	...	m	
1	x_{11}	x_{12}	x_{13}	...	x_{1m}	x_1
2	x_{21}	x_{22}	x_{23}	...	x_{2m}	x_2
3	x_{31}	x_{32}	x_{33}	...	x_{3m}	x_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	...	x_{nm}	x_n
$n+1$	$x_{(n+1)1}$	$x_{(n+1)2}$	$x_{(n+1)3}$	...	$x_{(n+1)m}$	x_{n+1}
$n+2$	$x_{(n+2)1}$	$x_{(n+2)2}$	$x_{(n+2)3}$	...	$x_{(n+2)m}$	x_{n+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	x_{N1}	x_{N2}	x_{N3}	...	x_{Nm}	x_N

Let X_i be the number of times that the i^{th} individual was caught or identified over the study period with m trapping occasions where $X_i = \sum_{j=1}^{m} X_{ij}$; $X_i = 0, 1, 2, ..., m$ as can be seen from Table 2.3. The number of unobserved individuals remains to be the unknown parameter of the study. Here, the random variable X is the main interest as it generates the marginal frequency. Consequently, we let $f_1, f_2, f_3, ..., f_m$ be the frequencies of distinct individuals being identified exactly x times for over period of study, $x = 1, 2, 3, ..., m$. Additionally, f_0 denotes the frequency of unobserved individuals. The population size N can be readily obtained as $N = f_0 + f_1 + f_2 + f_3 + ... + f_m = f_0 + n$ when $n = \sum_{j=1}^{m} f_j$ is the total number of observed individuals.

Example 2: Multiple marks. Table 2.4 shows the example of the capture-recapture history for each individual of 38 deer mice with six capture occasions (Amstrup et al. (2005)). It seems that assuming a closed population is reasonable since the duration of survey is short. It can be seen that the first individual was caught in all trapping occasions leading to $x_1 = 6$. The second individual was trapped in the first trapping, and was recaptured again in the forth, fifth and sixth occasion; $x_2 = 4$. Similarly, for the remaining distinct individuals. Note that (0,0,0,0,0,0), representing an individual not caught in any of the six trapping occasions, does not appear. Therefore, the number of identifications with $x = 0$ is unknown. The frequency of counts is summarized in Table 2.5. Since the number of trapping occasions is fixed and known prior to the capture-recapture sampling, the largest observed count m is known. The observed counts $1, 2, 3, ..., m$ provide $f_1, f_2, f_3, ..., f_m$. However, the frequency of unidentified individuals (f_0) is unknown and becomes important part for estimation.

Generally, there are two types of structured data in capture-recapture studies; 1) the discrete time data or capture-recapture data with different sources and 2) the continuous time data or repeated counting data. In the first type, recaptured identifications can occur only at *specific time points* within the study period, above is an example. *The Binomial distribution*, $B(m, p)$, is a reasonable option as a basic model for the capture

Table 2.4: Capture-recapture history of 38 deer mice with six cature occasions

Individual	Occasion						$X_i = \sum_{j=1}^{t} X_{ij}$
i	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
11	1	1	0	1	1	1	5
12	1	1	1	0	1	1	5
13	1	1	1	1	1	1	6
14	1	0	1	1	1	0	4
15	1	0	0	1	0	0	2
16	0	1	0	0	1	0	2
17	0	1	1	0	0	1	3
18	0	1	0	0	0	1	2
19	0	1	0	1	0	1	3
20	0	1	1	0	1	0	3
21	0	1	0	1	0	1	3
22	0	1	0	0	0	1	2
23	0	1	0	0	1	1	3
24	0	0	1	0	0	0	1
25	0	0	1	1	1	1	4
26	0	0	1	0	1	1	3
27	0	0	1	1	1	1	4
28	0	0	1	0	1	0	2
29	0	0	1	0	0	0	1
30	0	0	0	1	0	0	1
31	0	0	0	1	1	1	3
32	0	0	0	1	1	0	2
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

Table 2.5: The frequency count of deer mice

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

probability of the random variable X. In this case, m is the number of trapping occasions and p is the probability that each individual is identified on each trapping occasion.

Another structure of capture-recapture data occurs if recaptured identifications can occur *any time* during the study period and individuals are identified with probability $1 - p_0$ repeatedly by the same mechanism. In this case, the number of times that an individual is identified over a given period of the study takes the value $1, 2, 3, ...$ As a consequence, it is impossible to know the largest possible count of identifications such as how often patient coming to a treatment institution for the treatment of disease. Patients can usually go to the treatment institution to receive the treatment any time. It might be impossible to determine the largest contact count during the treatment period (see more examples in Hay and Smit (2003) and Norris and Pollock (1996)). For this particular type of data, *the Poisson distribution* is usually chosen to fit the capture probability (Böhning (2008)).

2.3.3 Multiple sources

The capture-recapture data can be obtained from the listing and recording systems in multiple sources (more than 2 sources) where the data are identified at different sources and matched with each others. These are now widely used in several areas. The lists of identified individuals from three sources can be merged and summarized as shown in Table 2.6

Table 2.6: The three-source situation

Source 1	Source 3			
	1		0	
	Source 2		Source 2	
	1	0	1	0
1	f_{111}	f_{101}	f_{110}	f_{100}
0	f_{011}	f_{001}	f_{010}	f_{000} =?

where 0 and 1 indicates an unidentified and identified individual, respectively. Similar to two sources and multiple marks,

$$f_0 = f_{000} \quad \text{is the unobserved frequency and unknown.}$$

$$f_1 = f_{100} + f_{010} + f_{001}$$

$$f_2 = f_{110} + f_{101} + f_{011}$$

$$f_3 = f_{111}$$

$$n = \sum_i \sum_j \sum_k f_{ijk} - f_{000} = f_1 + f_2 + f_3; \quad i, j, k \geq 0.$$

Therefore, the target population size can be calculated as:

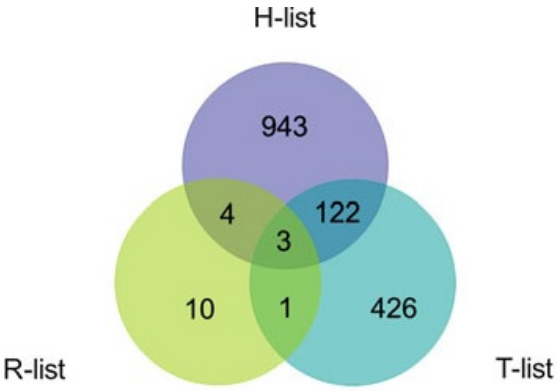

Figure 2.1: The numbers of hospitalizations captured and overlaps in the three lists (R, H and T)

$$N = f_0 + f_1 + f_2 + f_3 = f_0 + n.$$

Example 3: Three sources. Jouanjus et al. (2012) studied about addictive behaviours that are often assorted with hidden characteristics. This target population is difficult to detect, hence multiple sources were used to search these cases and crossed to identify eligible hospitalizations. A capture-recapture method was used to estimate the frequency of hospitalizations related to drug abuse. The data are shown in Figure 2.1, where

- Source 1 : Spontaneous reports of drug of abuserelated disorders (NotS), called R list
- Source 2 : Computerised hospital database Programme de Medicalisation des Systemes dInformation (PMSI), called H list
- Source 3 : Toxicological analyses (TA), called T list.

Table 2.7: Capture-recapture history with three sources of hospitalizations related to drug abuse study

Source			Frequency count
NotS (R)	PMSI (H)	TA (T)	f_{RHT}
0	0	0	?
1	0	0	10
0	1	0	943
0	0	1	426
1	1	0	4
1	0	1	1
0	1	1	122
1	1	1	3

The capture-recapture history can be written as Table 2.7 and the associated frequency counts of identified cases can be accounted as

$$f_0 = f_{000} \quad \text{is unknown.}$$

$$f_1 = f_{100} + f_{010} + f_{001} = 10 + 943 + 426 = 1,379$$

$$f_2 = f_{110} + f_{101} + f_{011} = 4 + 1 + 122 = 127$$

$$f_3 = f_{111} = 3, \quad \text{where} \quad n = f_1 + f_2 + f_3 = 1,509$$

The frequency of these counts is summarized in Table 2.8 and it is clear that the number of hospitalizations that were not seen f_0 is unobservable.

Table 2.8: The frequency table of hospitalizations

x	0	1	2	3	n
f_x	?	1,379	127	3	1,509

Example 4: Four sources. Bruno et al. (1994) used multiple sources to identify known cases of diabetes among the residents of the area of Casale Monferrato in northern Italy on October 1, 1988. There are four sources and the data are shown in Table 2.9, where

- Source 1 : Diabetic clinic and/or family physician
- Source 2 : Hospital discharges
- Source 3 : Prescriptions
- Source 4 : Reagent strips and insulin syringes.

Table 2.9: Data from prevalent cases of known diabetes mellitus for resident in Italy

Ascertainment			Source 1			
			yes		no	
			Source 2		Source 2	
			yes	no	yes	no
Source 3 yes	Source 4	yes	58	46	14	8
		no	157	650	20	182
Source 3 no	Source 4	yes	18	12	7	10
		no	104	709	74	?

From Table 2.9, the frequency counts of identified cases can be calculated as

$$f_0 = f_{0000} \quad \text{is unknown.}$$

$$f_1 = f_{1000} + f_{0100} + f_{0010} + f_{0001} = 709 + 74 + 182 + 10 = 975$$

$$f_2 = f_{1100} + f_{1010} + f_{1001} + f_{0110} + f_{0101} + f_{0011}$$

$$= 104 + 650 + 12 + 20 + 7 + 8 = 801$$

$$f_3 = f_{1110} + f_{1101} + f_{1011} + f_{0111} = 157 + 18 + 46 + 14 = 235$$

$$f_4 = f_{1111} = 58$$

$$n = f_1 + f_2 + f_3 + f_4 = 2,069.$$

The frequency distribution for these counts is summarized in Table 2.10 and the number of diabetes that were not seen f_0 need to be estimated.

Table 2.10: The frequency count of diabetes

x	0	1	2	3	4	n
f_x	?	975	801	235	58	2,069

Since the population size consists of observed and unobserved units, $N = n + f_0$, the estimate of f_0 leads to the estimate of population size N. Modelling and estimating p_0 is one of major concern for estimating the size N of a population according to $\hat{N} = n/(1 - \hat{p}_0)$. Let p_x be described by some model, e.g. $p_x = p_x(\theta)$. Then, an estimate $\hat{\theta}$ of the model parameter is derived. Hence, $p_0(\hat{\theta})$ is obtained and leads to the estimate of the population size $\hat{N} = n/(1 - p_0(\hat{\theta}))$ as well as the estimate of the hidden $\hat{f}_0 = \hat{N} - n$. Some methods to estimate p_0 and N are reviewed in Section 2.5.

2.4 The geometric model with truncation

The geometric distribution is a remarkably simple and flexible distribution. Although it has been often ignored for modelling count distributions, it is popular in survival analysis for life time data and also interesting through its memoryless property. Moreover, the geometric provides a more flexible model than the Poisson due to the fact that it arises as a mixture of the Poisson when the Poisson parameter is mixed with an exponential distribution that allows for some heterogeneity in the count data (see Niwitpong et al. (2013)).

The geometric distribution has a major interesting property that turns out to be useful for the truncated process.

- *Let* $(1 - p)^x p$ *be the geometric for* $x = 0, 1, ...$ *Then the zero-truncated geometric is again a geometric having the form*

$$\frac{(1 - p)^x p}{1 - p} = (1 - p)^{x-1} p \tag{2.10}$$

for $x = 1, 2, ...$

There is suspicion that counts of one are inflated. Hence, it might be appropriate to exclude ones from the estimation. The density is a geometric again.

- *Let* $(1-p)^{x-1}p$ *be the geometric for* $x = 1, 2, ...$ *Then the one-truncated geometric is again a geometric of the form*

$$\frac{(1-p)^{x-1}p}{1-p} = (1-p)^{x-2}p \tag{2.11}$$

  for $x = 2, 3, ...$

This truncation process can be continued with higher counts also leading to a geometric density. The first proposed model is based on the one-truncated geometric distribution that excludes the count of ones for the estimation and uses only the other counts for estimating p. Then use the estimate $\hat{p}$ of p to find the estimate of population size:

$$\hat{N} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\hat{p}}, \quad \text{since} \quad p_0 = (1-p)^0 p = p. \tag{2.12}$$

2.5 Overview of estimators

Classical capture-recapture methods usually focus on finding some appropriate models for the count probability distribution. Various estimators for estimating the population size have been proposed. Although there are two types of data sets as mentioned in Section 2.3, the Poisson model is reasonably chosen for the probability density function of the model because the binomial distribution can be widely approximated by a Poisson distribution if the number of trapping occasions m is large with the small success probability. In this section, therefore, the majority of estimators based on homogeneous Poisson, homogeneous geometric and heterogeneous models are examined. Maximum likelihood estimator and Good-Turing estimator are estimators for the homogeneous Poisson and geometric model whereas Chao's lower bound estimator and Zelterman estimator are proposed for heterogeneous models.

2.5.1 Horvitz-Thompson's estimator ($\hat{N}_{\mathbf{HT}}$)

Horvitz and Thompson (1952) introduced a basic technique for estimating means, totals and proportions of a finite population for any sampling design, with or without replacement. This approach is applied for capture-recapture studies to estimate the size N of target population. Let X_i be the identifying indicator variable of the i^{th} unit in the population, where $X_i = 1$ if i^{th} individual is identified, otherwise $X_i = 0$. Consequently, $\sum_{i=1}^{N} X_i = n$ is the number of observed units. Suppose that each unit is observed independently with identical probability $1 - p_0$ hence the probability of

observing exactly n units is the Binomial distribution. Moreover, we can note that $N(1-p_0)$ is the expected number of observed cases which can be estimated by n which $E(\sum_{i=1}^{N} X_i) = N(1-p_0) = n$. This leads to the simple equation to estimate the population size N. We can write

$$N = Np_0 + N(1-p_0) \approx Np_0 + n. \tag{2.13}$$

This equation can be solved for estimating N. Consequently, the Horvitz-Thompson Estimator is provided in the form:

$$\hat{N}_{\text{HT}} = \frac{n}{1-p_0}. \tag{2.14}$$

For more detail, see Bishop et al. (1975) and Van der Heijden et al. (2003a). However, p_0 is regularly unknown and need to be estimated for using in (2.14). There are many ways to estimate p_0 that will be discussed in following subsections.

2.5.2 Maximum likelihood estimator ($\hat{N}_{\text{MLE}}$)

The maximum likelihood method is the well-known traditional technique used to derive estimators (see Casella and Berger (2008)). Let $X_1, X_2, ..., X_n$ be a random sample with probability density function $f(x;\theta)$, the likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{2.15}$$

The maximum likelihood estimator (MLE) for unknown parameter θ can be obtained by maximizing the function $L(\theta)$, differentiation $L(\theta)$ with respect to θ and equated to zero. For capture-recapture study, the zero-truncated count data are considered because zeroes have not been observed in the identifying systems as mentioned in Section 2.1. Let X_i be the number of times that i^{th} individual was identified over the study period, where $i = 1, 2, 3, ..., n$. Count data X is often modelled by the *Zero-truncated Poisson Distribution* with probability function

$$Po^{+}(x;\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!(1-\exp(-\lambda))}; \qquad \lambda > 0, \qquad x = 1, 2, 3, ... \tag{2.16}$$

Additionally, in the sense of frequency data, let f_x denotes the frequencies of units observed x times over the study period, where $x = 1, 2, \ldots, m$ and $\sum_{x=1}^{m} f_x = n$. Then, the likelihood function for this zero-truncated count density is

$$L(\lambda) = \prod_{x=1}^{m} \left(\frac{Po(x,\lambda)}{1-\exp(-\lambda)} \right)^{f_x} \quad , x = 1, 2, 3, ..., m. \tag{2.17}$$

Therefore, the log-likelihood function of (2.17) is

$$l(\lambda) = -n\lambda + \log\lambda \sum_{x=1}^{m} x f_x - \sum_{x=1}^{m} f_x \log(x!) - n\log\left(1 - \exp\left(-\lambda\right)\right). \tag{2.18}$$

We equate the derivative of the log-likelihood to zero and get an expression for the maximum likelihood estimate $\hat{\lambda}$ of λ as

$$\frac{\partial l}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{x=1}^{m} x f_x - \frac{n\exp(-\lambda)}{1-\exp(-\lambda)} = 0$$

$$\frac{1}{n}\sum_{x=1}^{m} x f_x = \frac{\lambda}{1-\exp(-\lambda)}.$$

$$\bar{x} = \frac{\hat{\lambda}}{1-\exp(-\hat{\lambda})} \qquad \text{or} \qquad \hat{\lambda} = \bar{x}(1-\exp(-\hat{\lambda})). \tag{2.19}$$

Clearly, there is no close form solution for the maximum likelihood estimate $\hat{\lambda}$ in (2.19). We can find the approximate value of $\hat{\lambda}$ by a Taylor series approximation as follows

$$\hat{\lambda} = 2\left(\frac{\bar{x}-1}{\bar{x}}\right). \tag{2.20}$$

Another method to solve $\hat{\lambda}$ is an iterative method via EM algorithm. The likelihood function (2.17) can be maximized with algorithm between the E-Step and the M-Step:

(i) Expectation (E-Step)

The expected value of unobserved case f_0 given the observed variables and the current estimates of likelihood parameter are derived in this step.

$$\begin{aligned} \hat{f}_0 &= E(f_0|f_1, f_2, ..., f_m; \lambda) \\ &= p_0 N \\ &= \exp(-\lambda)(n + \hat{f}_0) \end{aligned} \tag{2.21}$$

Hence,

$$\hat{f}_0 = \frac{np_0}{1-p_0} = \frac{n\exp(-\lambda)}{1-\exp(-\lambda)} \tag{2.22}$$

given λ and n.

(ii) Maximization (M-Step)

In this step, the *unobserved, complete data* likelihood function is maximized by using observed cases (n) and unobserved cases (f_0) that is imputed from initial value in first

iteration and from using $\hat{f}_0$ from **E-Step** for next iteration. The estimate of λ in **M-Step** is

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{n + f_0}(0f_0 + 1f_1 + 2f_2 + ... + mf_m) = \frac{1}{\hat{N}} \sum_{x=0}^{m} x f_x \tag{2.23}$$

where $n = \sum_{x=1}^{m} f_x$ is the total number of all observed individuals and the conditional upon $f_0 = \hat{f}_0$. The EM-algorithm requires iterating between *E-Step* and *M-Step* until convergence in $\hat{\lambda}_{\text{MLE}}$ and $\hat{f}_0$. Moreover, the initial value is an important value to start the procedure; so it should be selected suitably and carefully. Frequently, the initial value is simply set by a sample mean. As a result of replacing $\hat{\lambda}_{\text{MLE}}$ in Horvitz-Thompson approach in (2.14), the population size estimator with regard to maximum likelihood is

$$\hat{N}_{\text{MLE\_P}} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MLE}})}. \tag{2.24}$$

The variance of (2.24) can be estimated by

$$\widehat{Var}(\hat{N}_{\text{MLE\_P}}) = \frac{\hat{N}_{\text{MLE\_P}}}{\left(\exp\left(\frac{\sum x f_x}{\hat{N}_{\text{MLE\_P}}}\right) - \frac{\sum x f_x}{\hat{N}_{\text{MLE\_P}}} - 1\right)}. \tag{2.25}$$

(see Böhning et al. (2005); Chao and Lee (1992); Meng (1997); Viwatwongkasem et al. (2008)).

Here, we consider maximum likelihood estimation under the geometric model. We assume that count data X is modelled by a geometric distribution with probability function

$$p_x = (1-p)^x p \qquad ; \qquad x = 0, 1, 2, ...$$

and the zero-truncated geometric likelihood is of the form

$$L(p) = \prod_{x=1}^{m} \left((1-p)^{x-1} p\right)^{f_x}.$$

The log-likelihood function is

$$\log L(p) = \log(1-p) \sum_{x=1}^{m} f_x(x-1) + \log p \sum_{x=1}^{m} f_x. \tag{2.26}$$

To find the maximum likelihood estimator (MLE) of unknown parameter p, differentiation of (2.26) with respect to p is equated to 0. This leads to

$$\frac{\partial l}{\partial p} = -\frac{\sum_{x=1}^{m} f_x(x-1)}{1-p} + \frac{\sum_{x=1}^{m} f_x}{p} = 0$$

$$\hat{p} = \frac{n}{S}$$

Hence, under the assumption of zero-truncated geometric model, the population size estimator with the maximum likelihood approach is

$$\hat{N}_{\text{MLE\_G}} = \frac{n}{1 - n/S} \tag{2.27}$$

where $S = \sum_{x=1}^{m} x f_x$. The variance estimation of the MLE_G in (2.27) can be estimated as

$$\widehat{Var}(\hat{N}_{\text{MLE\_G}}) = \frac{S^2 n^2}{(S-n)^3}. \tag{2.28}$$

(see Niwitpong et al. (2013)).

2.5.3 Turing estimator $(\hat{N}_{\mathbf{T}})$

Initially, Turing estimation is formulated to estimate the number of classes or species of animals which is defined as the sum of probabilities of observed classes. This estimator can also be applied to estimate the total number of populations. Let f_x be the frequency of individuals detected exactly x times, $x = 0, 1, 2, ..., m$ where m is the largest observed count. The total number of observed cases in the sample is $n = \sum_{x=1}^{m} f_x$ and the total number of captured cases can be defined as

$$S = f_1 + 2f_2 + 3f_3 + \ldots + mf_m = \sum_{x=1}^{m} x f_x.$$

Let p_x denote the probability that individual identified exactly x times. Assume that X has a homogeneous Poisson distribution with parameter λ so $p_0 = \exp(-\lambda)$ and $p_1 = \lambda \exp(-\lambda)$. We can write p_0 as

$$p_0 = e^{-\lambda} = \frac{e^{-\lambda}\lambda}{\lambda} = \frac{p_1}{E(X)}. \tag{2.29}$$

The estimate of p_0 can be calculated from observed frequency as follows

$$\hat{p}_0 = \frac{f_1/N}{S/N} = \frac{f_1}{S}. \tag{2.30}$$

Thus, if we plug $\hat{p}_0$ into the Horvitz-Thompson estimator, Turing estimator for estimating the population size is given by

$$\hat{N}_{\text{T\_P}} = \frac{n}{1 - f_1/S}. \tag{2.31}$$

The variance for Turing estimator can be estimated as

$$\widehat{Var}(\hat{N}_{\text{T\_P}}) = \frac{n f_1/S}{(1 - f_1/S)^2} + \frac{n^2}{(1 - f_1/S)^4}\left(\frac{f_1(1 - f_1/N)}{S^2} + \frac{f_1^2}{S^3}\right). \tag{2.32}$$

Here, Turing estimation under geometric homogeneity is considered. Let X have a marginal pmf following the geometric distribution with parameter p where $p_0 = p, p_1 = (1-p)p$ and $E(X) = \dfrac{1-p}{p}$, so that

$$\frac{p_1}{E(X)} = \frac{p(1-p)}{(1-p)/p} = p^2, \tag{2.33}$$

$$\sqrt{\frac{p_1}{E(X)}} = p = p_0.$$

The estimate of p_0 can be calculated from observed frequency as follows

$$\hat{p}_0^* = \sqrt{\frac{f_1}{S}} \tag{2.34}$$

Therefore, the extension of Turing estimator for estimating the population size under geometric model is given by

$$\hat{N}_{\text{T\_G}} = \frac{n}{1 - \sqrt{f_1/S}}. \tag{2.35}$$

The variance of $\hat{N}_{\text{T\_G}}$ can be derived as

$$\widehat{Var}(\hat{N}_{\text{T\_G}}) = \frac{n\sqrt{f_1/S}}{\left(1 - \sqrt{f_1/S}\right)^2} + n^2 \left(\frac{S + f_1}{4S^2 \left(1 - \sqrt{f_1/S}\right)^4} \right). \tag{2.36}$$

2.5.4 Chao's lower bound estimator ($\hat{N}_{\mathbf{C}}$)

The previous estimators are developed under the homogeneous Poisson model. However, it seems to be rarely met in practice and it is more suitable to incorporate heterogeneity. It is more reasonable to assume that the target population may be composed of a variety of subgroups. Chao (1987) proposed a lower bound for the population size N under the heterogeneous Poisson population. The capture probability is assumed to follow a Poisson mixture:

$$p_x = \int_0^\infty p\left(x|\lambda\right) q\left(\lambda\right) d\lambda,$$

$$p_x\left(x|\lambda\right) = \int_0^\infty \frac{e^{(-\lambda)}\lambda^x}{x!} q\left(\lambda\right) d\lambda, \tag{2.37}$$

where $q(\lambda)$ represents an arbitrary density of the model parameter λ in the population. Chao's estimator is derived in the sense of a nonparametric way by using the **Cauchy-Schwarz inequality** of any two random variables X and Y

$$\left[E(XY)\right]^2 \leq E\left(X^2\right) E\left(Y^2\right). \tag{2.38}$$

It is assumed that $X = u(\lambda)$ and $Y = v(\lambda)$ are function in λ where λ is assumed to be a continuous random quantity with density $q(\lambda)$ defined on the support $(0, \infty)$. We have

that

$$\left(\int_0^\infty u(\lambda)v(\lambda)q(\lambda)d\lambda\right)^2 \leq \left(\int_0^\infty u(\lambda)^2 q(\lambda)d\lambda\right)\left(\int_0^\infty v(\lambda)^2 q(\lambda)d\lambda\right) \tag{2.39}$$

Let $u(\lambda) = (e^{-\lambda}\lambda^{x-1})^{\frac{1}{2}}$ and $v(\lambda) = (e^{-\lambda}\lambda^{x+1})^{\frac{1}{2}}$. We have that $u(\lambda)v(\lambda) = e^{-\lambda}\lambda^x$. Then, the inequality in (2.39) can be written as

$$\left(\int_0^\infty e^{-\lambda}\lambda^x q(\lambda)d\lambda\right)^2 \leq \left(\int_0^\infty e^{-\lambda}\lambda^{x-1} q(\lambda)d\lambda\right)\left(\int_0^\infty e^{-\lambda}\lambda^{x+1} q(\lambda)d\lambda\right),$$

or

$$\left(\frac{x!}{x!}\int_0^\infty e^{-\lambda}\lambda^x q(\lambda)d\lambda\right)^2 \leq \left(\frac{(x-1)!}{(x-1)!}\int_0^\infty e^{-\lambda}\lambda^{x-1} q(\lambda)d\lambda\right)\left(\frac{(x+1)!}{(x+1)!}\int_0^\infty e^{-\lambda}\lambda^{x+1} q(\lambda)d\lambda\right),$$

or

$$(x!p_x)^2 \leq (x-1)!p_{x-1}(x+1)!p_{x+1},$$

and finally

$$\frac{xp_x}{p_{x-1}} \leq \frac{(x+1)p_{x+1}}{p_x}. \tag{2.40}$$

Replacing the probability p_x by their associated observed frequency specifically, for $x = 1$, leads to Chao's inequality $p_0 \geq \frac{p_1^2}{2p_2}$. The lower bound for the estimate of the number of unobserved units is provided as

$$\hat{f}_0 = \frac{f_1^2}{2f_2} \tag{2.41}$$

where the inequality $\hat{f}_0 \leq f_0$ will hold on in its expected value asymptotically. Finally, adding the estimator $\hat{f}_0$ to the number of observed cases n leads to Chao's lower bound estimator as

$$\hat{N}_{\text{C\_P}} = n + \frac{f_1^2}{2f_2}. \tag{2.42}$$

Chao also provided an approximate variance formula for estimator in (2.42) which is given as

$$\widehat{Var}(\hat{N}_{\text{C\_P}}) = \left(\frac{1}{4}\right)\frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \left(\frac{1}{2}\right)\frac{f_1^2}{f_2}. \tag{2.43}$$

It is interesting to note that Chao's lower bound estimator is simple to calculate, uses only f_1 and f_2. It represents lower bound estimates if heterogeniety based on Poisson is present and a mixing distribution is not required to be specified and to be estimated. Hence it is a truly non-parametric way (see Böhning (2010); Böhning et al. (2013b)). However, model selection is difficult due to the absence of the likelihood-based goodness of fit statistics.

Now Chao's lower bound estimator is considered under a geometric heterogeneity to estimate f_0 (see Niwitpong et al. (2013)). Let $g(x|p)$ be the geometric density with

parameter p and $k(p)$ is an arbitrary density of the model parameter p in the target population. The mixture geometric probability density is

$$q_x(p) = \int_0^1 g(x|p)k(p)dp = \int_0^1 \{(1-p)^x p\}\, k(p)dp.$$

The moment inequality under the **Cauchy-Schwarz inequality** is

$$[E(XY)]^2 \leq E\left(X^2\right) E\left(Y^2\right).$$

Let $X^2 = p$ and $Y^2 = p(1-p)^2$, so the inequality can be given as

$$\begin{aligned}
[E(XY)]^2 &\leq E\left(X^2\right) E\left(Y^2\right) \\
[E(p(1-p))]^2 &\leq E(p)E\left[p(1-p)^2\right] \\
E(p) &\geq \frac{[E(p(1-p))]^2}{E\left[p(1-p)^2\right]}.
\end{aligned}$$

Replacing expected values by frequencies leads to

$$f_0 = \frac{f_1^2}{f_2} \tag{2.44}$$

and to Chao's lower bound estimate

$$\hat{N}_{\text{C\_G}} = n + \frac{f_1^2}{f_2}. \tag{2.45}$$

A variance estimate of $\hat{N}_{\text{C\_G}}$ in (2.45) can be found as

$$\widehat{Var}(\hat{N}_{\text{C\_G}}) = \frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2}. \tag{2.46}$$

2.5.5 Zelterman's estimator ($\hat{N}_{\mathbf{Z}}$)

Zelterman (1988) proposed a series of robust estimators of the parameter λ under the zero-truncated Poisson probability, $Po^+(x;\lambda) = \frac{\exp(-\lambda)\lambda^x}{x!(1-\exp(-\lambda))}$. Zelterman's estimator is very popular and has a simple expression and is robust under potential unobserved heterogeneity. It is frequently used in socio-economical applications and illicit drug use research in the social sciences (see Navaratna et al. (2008); Böhning (2010); Farcomeni (2017)). By using Horvitz-Thompson approach to estimate population size; $\hat{N} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\exp(-\lambda)}$, the estimate of λ is required. Although the Poisson assumption is frequently invalid in reality, it can be assumed that homogeneity of Poisson probability can hold for small range of count variable from x to $x+1$. For example, singletons (f_1) and doubletons (f_2) follow a homogeneous Poisson distribution whereas other counts might be arbitrarily distributed. Therefore, the neighbouring frequencies f_x and f_{x+1} can be

used to estimated a parameter λ by considering the ratio

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

The parameter λ can be written as

$$\lambda = \frac{(x+1)Po^+(x+1;\lambda)}{Po^+(x;\lambda)}. \tag{2.47}$$

$Po^+(x;\lambda)$ and $Po^+(x+1;\lambda)$ are replaced by the empirical relative frequencies f_x/N and f_{x+1}/N, respectively to obtain an estimator for λ. Thus, we have that

$$\hat{\lambda} = \frac{(x+1)f_{x+1}/N}{f_x/N} = \frac{(x+1)f_{x+1}}{f_x}. \tag{2.48}$$

Zelterman claimed that individuals never seen should be more similar to rarely seen individuals than individuals captured many times hence he suggested to use $x = 1$ and $\hat{\lambda} = \frac{2f_2}{f_1}$. In addition, Kuhnert and Böhning (2009) supported the idea of using $x = 1$ by giving 2 reasons. First is that using $x = 1$ gives the closest vicinal frequencies (f_1 and f_2) to estimate the target parameter f_0. Second is that the majority of counts usually fall into count of ones and twos in many applications. Therefore, the counts larger than two do not affect the estimator. As the result, Zelterman's estimator for estimating the population size is

$$\hat{N}_{\text{Z\_P}} = \frac{n}{1-\exp(\frac{-2f_2}{f_1})}. \tag{2.49}$$

The variance estimate for the Zelterman's estimator in (2.49) is

$$\widehat{Var}(\hat{N}_{\text{Z\_P}}) = nG(\hat{\lambda})\left[1 + nG(\hat{\lambda})\hat{\lambda}^2\left(\frac{1}{f_1}+\frac{1}{f_2}\right)\right], \tag{2.50}$$

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

Here, Zelterman's estimator based on the geometric distribution is considered. Let $g(x|p) = g_x = (1-p)^x p$ be the geometric distribution with parameter p where $g_0 = p$. The zero-truncated geometric distribution is

$$g_x^+ = \frac{g_x}{1-g_0} = \frac{(1-p)^x p}{1-p} = (1-p)^{x-1}p \tag{2.51}$$

and the ratio of neighbouring zero-truncated geometric distribution can be calculated as

$$\frac{g_{x+1}^+}{g_x^+} = \frac{(1-p)^{x+1}p/(1-p)}{(1-p)^x p/(1-p)} = 1-p. \tag{2.52}$$

Replacing g_x^+ and g_{x+1}^+ in (2.52) by the respective relative frequencies

$$\frac{f_{x+1}/N}{f_x/N} = 1 - \hat{p}, \tag{2.53}$$

and it can be written as

$$1 - \hat{p} = \frac{f_{x+1}}{f_x} = 1 - \hat{g}_0. \tag{2.54}$$

Using a similar reasoning as for the Poisson distribution above, $x = 1$ is chosen to estimate g_0 hence $1 - \hat{g}_0 = \frac{f_2}{f_1}$. Using Horvitz-Thompson approach, the Zelterman estimator for estimating population size based on geometric distribution is given as

$$\hat{N}_{\text{Z\_G}} = \frac{n}{1 - \hat{g}_0} = \frac{n}{f_2/f_1} = \frac{nf_1}{f_2}. \tag{2.55}$$

A variance estimate for the Zelterman estimator under geometric distribution in (2.55) is

$$\widehat{Var}(\hat{N}_{\text{Z\_G}}) = \frac{nf_1(f_1 - f_2)}{f_2^2} + n^2 \left(\frac{f_1}{f_2^2} + \frac{f_1^2}{f_2^3} \right) \tag{2.56}$$

(see Anan (2016)).

2.6 Application concerning capture-recapture models

Capture-recapture methods are applied in many research areas to estimate the unknown population size. In this section, some examples are examined in order to illustrate an application of all estimators above, as well as to show that some data sets might have the problem of one-inflation.

2.6.1 Snowshoe hares in north-central Alberta

Keith and Meslow (1968) present data on the number of times individual snowshoe hares were captured and recaptured from live trapping at six different square mile study areas during 1962-1967 in north-central Alberta. Regular trapping periods included midwinter, spring and summer and the frequency counts are shown in Table 2.11.

Table 2.11: The frequency count of snowshoe hares

	f_1	f_2	f_3	f_4	f_5	f_6	n
Midwinter	72	19	2	1	1	0	95
Spring	109	45	19	5	3	0	181
Summer	184	55	14	4	4	0	261

The estimators based on the Poisson and geometric model in Section 2.5 are applied to estimate the abundance of snowshoe hares. The results are shown in Table 2.12. It can

be seen that all estimators under the geometric distribution give larger population sizes than estimators under the Poisson, while the smallest population size is given by Turing under the Poisson model for all seasons.

Table 2.12: Estimated sizes of a snowshoe hares population in north-central Alberta based on Poisson and geometric model

Estimator	Poisson			Geometric		
	Midwinter	Spring	Summer	Midwinter	Spring	Summer
MLE	249	343	578	396	479	875
Turing	224	289	516	394	467	880
Chao	231	313	569	368	445	877
Zelterman	232	322	580	360	438	873

2.6.2 Cottontail rabbits: data from known size experiment

Edwards and Eberhardt (1967) study the capture-recapture data of cottontail rabbits from an experiment with known size of population. They penned 135 wild cottontail rabbits in a four-acre rabbit proof enclosure and conducted live trapping for 18 sequential nights. The frequencies of capture-recapture were recorded as shown in Table 2.13 (see more detail in Chao (1987)).

Table 2.13: The frequency count of cottontail rabbits

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	...	f_{18}	n
43	16	8	6	0	2	1	0	...	0	76

Table 2.14: Estimated sizes of cottontail rabbits population based on the Poisson and geometric model

Estimator	Estimated Population Size	
	Poisson	Geometric
MLE	126	164
Turing	110	169
Chao	**134**	192
Zelterman	145	205

In this case of study, the target population size N is known; 135 cases. There were only 76 caught individuals within 18 nights of trapping occasions; $n = 76$. It is clear that 59 individuals were unobserved; $f_0 = 59$. The estimated values of the population size from estimators discussed in Section 2.5 are shown in Table 2.14. It is clearly seen that Poisson model is more suitable than the geometric model. We now consider only estimators based on the Poisson model and it is found that Chao's estimator yields a remarkable reasonable estimate which is almost equal to the true population size. Zelterman's estimator gives an overestimation whereas others show underestimation.

2.6.3 Illegal immigrants in the Netherlands

Van der Heijden et al. (2003a) presented the capture-recapture data of illegal immigrants in the Netherlands from police records in order to estimate population size by using the truncated Poisson regression model. These records contain information on the number of times each illegal immigrant was apprehended by the police according to Table 2.15.

Table 2.15: The frequency count of illegal immigrants

f_1	f_2	f_3	f_4	f_5	f_6	n
1,645	183	37	13	1	1	1,880

It can be seen from Table 2.16 that all estimators under the Poison model give smaller estimates than the geometric model. Additionally, Zelterman's estimator gives the largest population size for under both models whereas the smallest estimation is given by Turing and MLE under the Poisson and geometric model, respectively.

Table 2.16: Estimated sizes of illegal immigrants population based on the Poisson and geometric model

Estimator	Estimated Population Size	
	Poisson	Geometric
MLE	7,722	13,469
Turing	7,608	14,208
Chao	9,274	16,668
Zelterman	9,425	16,900

2.6.4 Methamphetamine use in Thailand

Rocchetti et al. (2011) show the data of drug abuse for 61 health centers in the Bangkok metropolitan region from the Office of the Narcotics Control Board (ONCB). Table 2.17 presents the number of methamphetamine users for each count of treatment episodes (see more detail in Böhning et al. (2004)). In this study case, the maximum observed frequency was 10 and the total number of methamphetamine users is estimated.

Table 2.17: The frequency of Methamphetamine use in Thailand

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	n
3,114	163	23	20	9	3	3	3	4	3	3,345

From Table 2.18, it can be seen that the MLE based on Poisson model gives the smallest number of methamphetamine users whereas Zelterman's estimator under geometric model provides the largest number.

Table 2.18: Estimated sizes of methamphetamine users in Bangkok based on the Poisson and geometric model

Estimator	Estimated Population Size	
	Poisson	Geometric
MLE	16,802	30,113
Turing	19,395	37,039
Chao	33,091	62,836
Zelterman	33,654	63,904

2.6.5 Microbial diversity in the Gotland Deep

Rocchetti et al. (2011) show the data on microbial diversity that stem from a recent count by Stock et al. (2009) as shown in Table 2.19. The maximum observed frequency is 53 and the number of observed individuals (n) is 83. In this case the number of different genes (DNA sequences) N in particular environments is estimated by a variety of estimators under the Poisson and geometric model. The results are shown in Table 2.20.

Table 2.19: The frequency of Microbial diversity in the Gotland Deep

f_1	f_2	f_3	f_4	f_6	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{16}	f_{17}	f_{18}	f_{20}	f_{29}	f_{42}	f_{53}
48	9	6	2	2	2	1	2	1	1	1	2	1	1	1	1	1	1

It is clear from Table 2.20 that Turing is estimator based on the Poisson gives the lowest estimate of microbial diversity whereas Zelterman's estimator under the geometric model provides the highest number of microbial diversity.

Table 2.20: Estimated microbial diversity in the Gotland Deep based on the Poisson and geometric model

Estimator	Estimated Population Size	
	Poisson	Geometric
MLE	105	105
Turing	95	128
Chao	211	399
Zelterman	266	443

It can be clearly noticed from above applications that there can be large differences between the results of estimating population size from different estimators based on different models. Therefore, the important key for estimating a target population size is the capture-recapture model that we use to fit the data. As it is mentioned in the previous section, the basis distributions for capture-recapture are Poisson and binomial models. However, a violation of homogeneous modelling have been widely discussed as it leads to bias in estimation. A variety of mixture models are offered for estimating

population size with heterogeneity. As a consequence, the model selection plays a crucial role for a process of estimation. The next section will provide a graphical device to investigate count data modelling.

2.7 The ratio plot

Statistical graphics are a fundamental and essential tools for statistical data analysis although they are often overlooked. Graphs are simple instruments for preliminary exploration of a dataset to perceive and understand the features and structure of data. It also provides insight into influential aspects of statistical inference such as invalid distributional assumptions and latent patterns. Graphs can assess quickly and efficiently these aspects. In capture-recapture study, the ratio plot has been developed as a graphical device for investigating models and choosing methods for estimating population size. The basic concept is derived from a homogeneous Poisson distribution and expanded to heterogeneous models by Böhning et al. (2013a).

Assume a count distribution $p_x = p_x(\lambda)$ for the generation of f_x. Here λ is reflecting some parametric model such as the Poisson

$$p_x = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $\lambda > 0$ is an unknown parameter. Then, using $E(f_0 \mid f_1, ..., f_m; \lambda) = Np_0$ and a Horvitz-Thompson estimator of N in (2.5), we can find a Horvitz-Thompson-type estimate of f_0 via

$$\hat{f}_0 = n\frac{p_0(\hat{\lambda})}{1 - p_0(\hat{\lambda})}.$$

In capture-recapture studies, the zero counts are truncated hence $\hat{\lambda}$ is an estimate based on the observed frequencies $f_1, ..., f_m$, potentially one of the estimates discussed in Section 2.5, m being the largest observed count, and $n = f_1 + ... + f_m$. Consider the ratios

$$\frac{p_{x+1}}{p_x} = \frac{p_{x+1}/(1 - p_0)}{p_x/(1 - p_0)} = \frac{\lambda}{x + 1} \tag{2.57}$$

It can be seen from (2.57) that the ratio for the zero-truncated and non-truncated distribution is identical and leading to

$$r_x = (x + 1)\frac{p_{x+1}}{p_x} = \lambda. \tag{2.58}$$

The ratio r_x is constant with varying count x. It is straightforward to estimate r_x by $\hat{r}_x$

$$\hat{r}_x = (x + 1)\frac{f_{x+1}}{f_x} \tag{2.59}$$

where f_x is the frequency of count x and $N = f_0 + f_1 + ... + f_m$. The graph x against $\hat{r}_x = (x+1)\frac{f_{x+1}}{f_x}$ is called the *ratio plot*. It can be used as a *diagnostic device* for the Poisson distribution. If the ratio plot shows a pattern of *a horizontal line*, it can be taken as indicative for the presence of a Poisson distribution. Conversely, departures from a horizontal line provide evidence for invalidation of Poisson homogeneity. For constructing the ratio plot, both of f_{x+1} and f_x should be positive. If any of the two is zero, the ratio is undefined and we will give some blanks in ratio plot.

The occurrence of homogeneous Poisson distribution is rare in practice. If the ratio plot is a monotone pattern, indeed λ is distributed with arbitrary density $q(\lambda)$. Then

$$p_x = \int_0^1 \frac{e^{-\lambda}\lambda^x}{x!} q(\lambda) d\lambda \tag{2.60}$$

has the monotonicity property (Böhning et al. (2013a))

$$1\frac{p_1}{p_0} \leq 2\frac{p_2}{p_1} \leq 3\frac{p_3}{p_2} \leq ...$$

Hence, the ratio plot must be monotone increasing. If we consider as mixing density $q(\lambda)$ the exponential then

$$p_x = \int_0^1 \frac{e^{-\lambda}\lambda^x}{x!} \frac{1}{\mu} e^{-\lambda/\mu} d\lambda = (1-p)^x p. \tag{2.61}$$

The geometric distribution arises with event parameter $p = 1/(1+\mu)$. For aspects of heterogeneity modelling see also Dorazio and Royle (2003). Based on zero-truncated and non-truncated geometric distribution, the ratio becomes

$$r'_x = \frac{p_{x+1}}{p_x} = \frac{p_{x+1}/(1-p_0)}{p_x/(1-p_0)} = 1 - p \tag{2.62}$$

which can be easily estimated by

$$\hat{r}'_x = \frac{f_{x+1}}{f_x}. \tag{2.63}$$

Hence plotting $\hat{r}'_x$ against x leads to the geometric ratio plot and would serve as a diagnostic device for the geometric distribution.

Chapter 3

Estimators Based Upon One-Truncated Geometric Distribution

This chapter illustrates one-inflation and the ratio plot as a diagnostic device. It also shows the inferior performance of classical estimators when data experience one-inflation. To cope with this situation two new estimators of population size based on the one-truncated geometric distribution are introduced. One is modified from Turing estimation (T_OT) and another one is developed from the maximum likelihood approach (MLE_OT). Simulation technique is applied to study the performance of proposed estimators. The simulation results show that the $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ can improve the efficiency of the original estimators under one-inflation especially the $\hat{N}_{\text{T\_OT}}$ has better performance than $\hat{N}_{\text{MLE\_OT}}$. Overall, the proposed estimators give the smallest relative bias, relative variance and relative mean square error for all conditions of study.

3.1 Introduction

Based on capture-recapture models, the identifying system generally provides a count $X_i > 0$ of how many times the i^{th} individual has been captured, for $i = 1, 2, ..., n$ and $X_i = 0$ denotes unobserved cases in the system for $i = n+1, n+2, ..., N$. Therefore, it can be written that the total number of a target population (N) consists of an observed part (zero-truncated) of size n and unobserved part of unknown size $f_0 = N - n$ as well as $N = n + f_0$. In order to investigate an estimate of N based on an available sample $X_1, X_2, ..., X_n$, it is usually required to assume a model for the capture probability of X, $p_i = \text{Prob}(X = i)$. Moreover, statistical models for biology and ecology assume that the population can be divided into a finite number of classes. Each member of the population is identified with one class. Here, we can say that the size of target population N is the

total number of existing classes. A drawn sample from such a population will typically have repeated observations of the various classes. That is, some may be observed only once, other twice and so on, while many classes may not appear in the sample at all. The *frequency count data* is $\{(x, f_x) \mid x \geq 1\}$ where f_x is the number of sample classes of size x. For example the data set $\{(1, 10), (2, 4), (3, 2), ..., (7, 1)\}$ has ten *singletons*, four *doubletons*, ..., and one class occurring seven times in the sample. The problem is how to estimate the total number of classes due to not all classes are observed. Some classes remain undetected and the purpose is to provide an estimate of the frequency f_0 of different classes that remain unobserved (see details in Bunge and Fitzpatrick (1993) or Böhning and Vilas (2008)). The Poisson distribution is used as a basic model for fitting capture-recapture data. However, it is recognised that many datasets in some capture-recapture application have a large number of count ones or the data are in the form of one-inflation. According to Farcomeni and Scacciatelli (2013) and Bunge et al. (2012), this may be the results of the fact that: 1) the recapture probability of the same individual is very low, 2) individuals may show trap avoidance after the first capture, and 3) individuals may be assigned to incorrect class due to the error of matching leading to an artificially inflated frequency of count one such as the data of microbial diversity.

As it was mentioned in Section 2.7, the ratio plot can be used as a diagnostic device for the Poisson or geometric model. If the pattern of the ratio plot is a horizontal line, it indicates the presence of the distribution of interest. The estimate for the Poisson is $\hat{r}_x = (x + 1)f_{x+1}/f_x$ whereas $\hat{r}'_x = f_{x+1}/f_x$ is the estimate for the geometric. To illustrate the ratio plot for one-inflation and the potential of large bias in the estimate of Chao, we consider synthetic data of a population with size $N = 15,000$ with $10,000$ counts generated from the Poisson with parameter 2 merged with $5,000$ extra-ones. The frequency distribution is $f_0 = 1,377, f_1 = 7,823, f_2 = 2,614, f_3 = 1,736, f_4 = 894, f_5 = 354, f_{6+} = 202$. In this case, the observed sample size is $n = 13,623$. We ignore the fact that f_0 is known and estimate it by the conventional Chao estimator $\hat{f}_0 = f_1^2/2f_2 = 11,706$ and finally the population size estimate is $\hat{N} = n + \hat{f}_0 = 25,329$. It can be seen clearly that Chao's estimator gives a serious overestimate of the true $f_0 = 1,377$ and $N = 15,000$, respectively. The associated ratio plot and frequency chart are presented in Figure 3.1. The ratio plot shows clear evidence of one-inflation since the first point is far away from the best horizontal line. Here, the explanation is that there are a lot more counts of one. Therefore, from this example, the ratio plot can be used as a rough diagnostic device of one-inflation. Additionally, we can use the ratio plot for the geometric to investigate the suitability of a geometric distribution with one-inflation in a similar way. Note that the frequency chart (right panel in Figure 3.1) would not allow an easy identification of one-inflation as the ratio plot does.

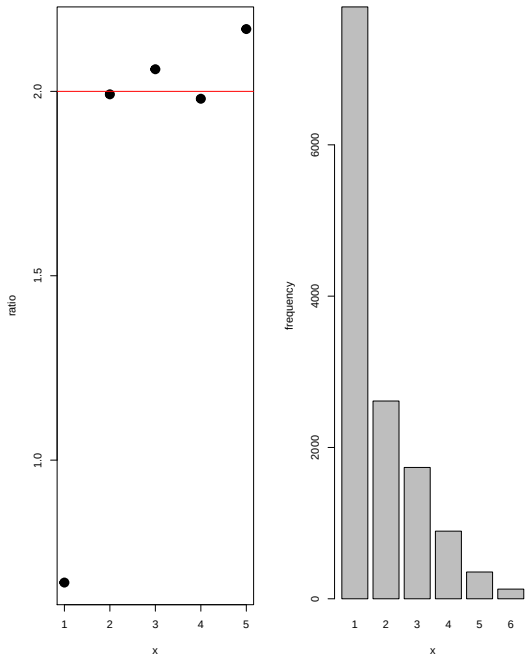

Figure 3.1: Ratio plot (left panel) and corresponding frequency chart (right panel) for $N = 15,000$ simulated Poisson counts with mean 2 and 50% one-inflation

3.2 Examples of applications with one-inflated count data and ratio plot

Example 1 In the context of animal disease surveillance, the data on scrapie-infected holdings in France are obtained from the French classical scrapie surveillance programme (Vergne et al. (2012)). Here, we are interested in estimating the total number of holdings with scrapie infection in France. Table 3.1 presents the frequency distribution of detection among holdings where at least one infected animal was detected. Here f_x represents the number of detected holdings with exactly x infected sheep. The total number of detected holdings is $n = 141$. There are 121 holdings with exactly one infected sheep, 13 holdings with exactly two infected sheep and so forth.

Table 3.1: Zero-truncated count data of French scrapie-infected holding in 2006

x	1	2	3	4
f_x	121	13	5	2

Figure 3.2 left panel shows the two ratio plots, the first one using $\hat{r}_x = (x+1)f_{x+1}/f_x$ for the diagnosis of a Poisson and the second one using $\hat{r}'_x = f_{x+1}/f_x$ for the diagnosis of a geometric. It is clear that the ratio plot for a Poisson shows a monotone increasing pattern, in particular, we can say that it does not show a horizontal line pattern. Hence

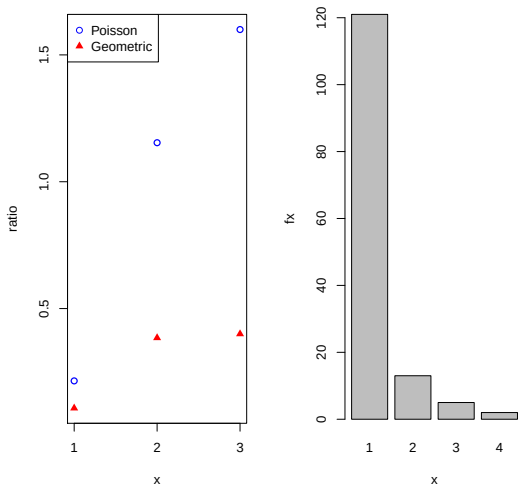

Figure 3.2: Ratio plot (left panel) and corresponding frequency chart (right panel) for the scrapie infected holding data

the Poisson model may not be suitable with this data whereas the ratio plot for a geometric is much closer to a horizontal line. However, it should be noticed that the first value of the geometric ratio plot $\hat{r}_1' = f_2/f_1$ is very low if compared with the others values in the graph. This could be explained by the fact that there are a lot more counts of holding with one infected sheep due to one-inflation as corresponding with the frequency chart in right panel. Therefore, it is indicated to use the geometric model under one-inflation estimating the total number of holding with scrapie infection.

Example 2 Van der Heijden et al. (2014) study the prevalence of domestic violence in the Netherlands for the year 2009 by using capture-recapture methods to estimate the total population size of offenders. The perpetrator study is reported with the data given in Table 3.2. The total number of observed offenders is $n = 17,662$. There are 15,169 offenders identified exactly once in a domestic violence incident, 1,957 exactly twice and so forth. From the data and the the frequency chart in Figure 3.3 right panel, it is noticed that the observed data may be contaminated with errors due to inflation in count one.

Table 3.2: The frequency count of a domestic violence incident in the Netherlands

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

In Figure 3.3 left panel, the two ratio plots are used to investigate the models appropriate for the data, one using for the diagnosis of a Poisson and another using for the diagnosis of a geometric. It appears to be clear that the Poisson model might not be appropriate for these data due to the ratio plot for the Poisson does not show a horizontal line

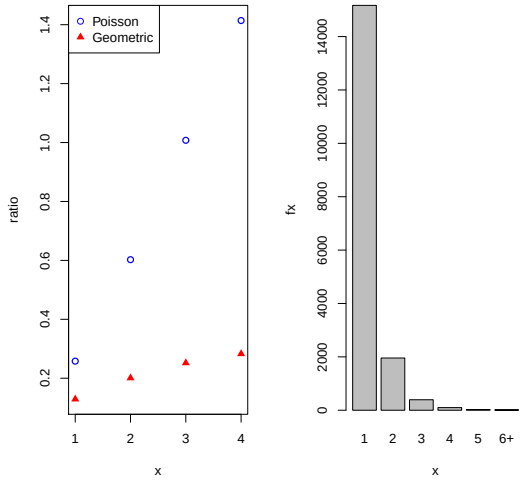

Figure 3.3: Ratio plot (left panel) and corresponding frequency chart (right panel) for the domestic violence data

pattern. Contrarily, the ratio plot for the geometric is much closer to a horizontal line, so it is interesting to use the geometric model fitting these data. Although, we cannot see the signal of one-inflation from the ratio plot, it is suspected that the counts of ones might be exceed due to the number of singletons is almost an amplitude higher than the number of doubletons.

Example 3 Phage diversity analyses represent a new level of population diversity beyond what is encountered in other areas of microbial ecology. We illustrate the situation for a contig spectrum from a swine fecal metagenome (Allen et al. (2011)). The contig spectrum was generated using Circonspect via the CAMERA pipeline (Sun et al. (2011)). Here, we are interested in estimating the taxonomic diversity of this metagenome. The complete frequency count data is in Table 3.3.

Table 3.3: The frequency count data of phage metagenome

x	1	2	3	4	5	6	7	8	9	10	11	12	13
f_x	4736	521	152	69	46	27	21	18	16	10	9	8	7
x	14	15	16	17	18	19	20	21	22	23	24	25	26
f_x	6	5	4	4	3	3	3	3	2	2	3	3	1
x	27	28	29	30	31	32	33	34	35	36	37	38	39
f_x	2	1	2	2	1	1	1	1	1	1	1	1	1
x	40	41	42	43	44	45	46	47	48	49	52	51	52
f_x	1	1	0	1	0	1	0	0	0	0	0	0	1

The total number of observed taxa is $n = 5,703$. Bunge et al. (2012) state: "It is clear even without graphing the data that the sample diversity is high: for instance, the number of singletons is almost an order of magnitude higher than the number of doubletons.

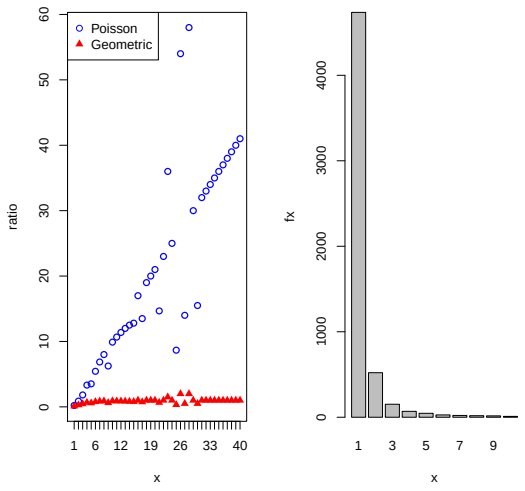

Figure 3.4: Ratio plot (left panel) and corresponding frequency chart (right panel) for the phage metagenome data

There is some basis to believe that the experimental and bioinformatic procedures that generated these data are prone to erroneous inflation of the low frequency counts" (p.5). Furthermore, it can be found in Figure 3.4 right panel that the number of singletons is about nine times the number of doubletons and there is a long and sparse tail to the right. This data shows an uncertainty in the low-frequency counts that is the salient characteristics of data in microbial ecology. The ratio plots in Figure 3.4 left panel are now considered and it is clear that the Poisson does not appear to be suitable with this data as its plots depart from a horizontal line pattern whereas the geometric is much closer to a horizontal line although some points are not that of a line pattern. Moreover, one should notice that the lower counts such as count one, two, three and four are very low when compared with other counts in the graph. Therefore, we can say that the ratio plot shows evidence of one-inflation in these data. Besides one-inflation, another notice should be considered that the error variance will increase with increasing counts.

The results of these examples show distinctly the effect of one-inflation in conventional estimators. Hence, we should be concerned about how to cope with this problem. In this chapter, we will focus on models specifically designed to estimate the size of a population for one-inflated capture-recapture count data allowing for heterogeneity. We also provide an inferential approach of estimators under one-inflation. These models and proposed estimators are based upon the geometric distribution.

3.3 One-truncated geometric model

Under the assumption that the frequency of count one is inflated, some estimators are developed under the one-truncated geometric model. The first proposed estimator is

provided in form of a Turing estimator and another one is developed by the maximum likelihood approach.

3.3.1 One-truncated Turing estimator ($\hat{N}_{\mathbf{T\_OT}}$)

Let f_x be the frequency of individuals identified exactly x times. Also, $n = \sum_{x=1}^{m} f_x$ is the total number of observed cases in the sample, and $S = f_1 + 2f_2 + 3f_3 + \ldots + mf_m = \sum_{x=1}^{m} xf_x$ is the total number of captured cases. The estimate of p_0 and population size N can be calculated from the observed frequencies as follows:

$$\hat{p}_0 = \frac{f_1/N}{S/N} = \frac{f_1}{S}$$

$$\hat{N}_T = \frac{n}{1 - f_1/S}.$$

This is the conventional Turing estimator developed under the Poisson model. Under the geometric distribution, let $p_x = (1-p)^x p \ ; \ x = 0, 1, 2, \ldots$. The Turing estimator of p can be derived as follows:

$$\frac{p_1}{E(X)} = \frac{(1-p)p}{(1-p)/p} = p^2,$$

or

$$\sqrt{\frac{p_1}{E(X)}} = p = p_0.$$

It follows that

$$\hat{p} = \sqrt{\frac{f_1}{S}}. \tag{3.1}$$

Consider the case of a one-truncated geometric distribution. Let us write

$$p_y = (1-p)^{y-1} p \, ; \quad y = 1, 2, 3, \ldots$$

in the form

$$p_x = (1-p)^x p \, ; \quad x = 0, 1, 2, \ldots$$

with $x = y - 1$. From formula in (3.1) follows that

$$\hat{p} = \sqrt{\frac{\hat{p}_1}{\hat{E}(X)}} = \sqrt{\frac{f_{x=1}}{S_x}}.$$

Transform the random variable x to y so that

$$\hat{p} = \sqrt{\frac{f_2}{0f_{x=0} + 1f_{x=1} + 2f_{x=2} + ... + (m-1)f_{x=m-1}}}$$
$$= \sqrt{\frac{f_2}{0f_1 + 1f_2 + 2f_3 + ... + (m-1)f_m}}.$$

Hence, the estimate of p can be calculated from the observed frequencies as:

$$\hat{p}_{\text{T\_OT}} = \sqrt{\frac{f_2}{f_2 + 2f_3 + 3f_4 + ... + (m-1)f_m}} \tag{3.2}$$

Thus, the one-truncated Turing estimator for estimating the population size is given by

$$\hat{N}_{\text{T\_OT}} = \frac{n}{1 - \hat{p}_{\text{T\_OT}}} \tag{3.3}$$

The formula in (3.3) is simply derived in terms of the Horvitz-Thompson estimator in (2.12) by replacing $\hat{p}$ by $\hat{p}_{\text{T\_OT}}$ in (3.2) assuming there is one-inflation in the capture probability. Expanding to k-truncated geometric distribution, the k-truncated Turing estimator (T_KT) for p is of the form:

$$\hat{p}_{\text{T\_KT}} = \sqrt{\frac{f_{k+1}}{\sum_{y=k+1}^{m}(y-k)f_y}} \tag{3.4}$$

3.3.2 One-truncated maximum likelihood estimator ($\hat{N}_{\text{MLE\_OT}}$)

Let X be the number of times that a unit was identified over the study period. Count X is modelled with a geometric distribution having probability function

$$p_x = (1-p)^x p \qquad ; \qquad x = 0, 1, 2, ...$$

Since the observed sample from a capture-recapture study contains only non-zero counts, the associated probability function becomes a zero-truncated geometric. Additionally, in the sense of frequency data, the observed data are given as f_x where $x = 1, 2, 3, ..., m$ where m is the largest observed count. The zero-truncated geometric likelihood is of the form

$$L(p) = \prod_{x=1}^{m} \left[(1-p)^{x-1}p\right]^{f_x}.$$

The log-likelihood function is

$$\log L(p) = \log(1-p)\sum_{x=1}^{m} f_x(x-1) + \log p \sum_{x=1}^{m} f_x. \tag{3.5}$$

To find the maximum likelihood estimator (MLE) of the unknown parameter p, the derivative of (3.5) with respect to p is equated to 0:

$$\frac{dl}{dp} = -\frac{\sum_{x=1}^{m} f_x(x-1)}{1-p} + \frac{\sum_{x=1}^{m} f_x}{p} = 0$$

This leads to

$$\hat{p}_{\text{MLE\_ZT}} = \frac{n}{S}.$$

Hence, under the assumption of a zero-truncated geometric model, the population size estimator based on the maximum likelihood estimation is

$$\hat{N}_{\text{MLE\_ZT}} = \frac{n}{1-n/S}. \tag{3.6}$$

Similarly, we assume that the count X is modelled as one-truncated geometric distribution with probability function

$$p_x = (1-p)^{x-2}p \qquad ; \qquad x = 2, 3, 4, ...$$

The log-likelihood function is

$$\log L(p) = \log(1-p)\sum_{x=2}^{m} f_x(x-2) + \log p \sum_{x=2}^{m} f_x. \tag{3.7}$$

To find the maximum likelihood estimator (MLE) of unknown parameter p, the derivative of (3.7) with respect to p is equated to 0:

$$\frac{dl}{dp} = -\frac{\sum_{x=2}^{m} f_x(x-2)}{1-p} + \frac{n-f_1}{p} = 0$$

so that

$$\hat{p}_{\text{MLE\_OT}} = \frac{n-f_1}{S-n}$$

arises. Hence, under the assumption of one-truncated geometric model, the population size estimator based on the maximum likelihood estimation is

$$\hat{N}_{\text{MLE\_OT}} = \frac{n}{1-(n-f_1)/(S-n)}. \tag{3.8}$$

In a similar way, the general form of maximum likelihood estimator for unknown parameter p under k-truncated geometric distribution is derived as

$$\hat{p}_{\text{MLE\_KT}} = \frac{n-\sum_{x=1}^{k} f_x}{S-kn+\sum_{x=1}^{k-1}(k-x)f_x}. \tag{3.9}$$

3.4 Goodness of fit

The goodness of fit (GOF) of a statistical model describes how well it fits a set of observations. The measures of GOF regularly summarize the discrepancy between observed values, which are the frequency of a class from a sample and the estimated or fitted frequencies, which is calculated under the claimed model. In order to measure departure of the observed data from the model, an asymptotic χ^2 goodness of fit statistic is used. Let G be

$$G = \sum_{x=1}^{m} \frac{(f_x - \hat{e}_x)^2}{\hat{e}_x}, \tag{3.10}$$

where f_x and e_x are the observed and fitted frequency in the x^{th} class, respectively.

We can calculate the fitted frequency by

$$\hat{e}_x = nP(X = x)$$

for $x = 1, ..., m-1$; and

$$f_{m+} = \sum_{j=m}^{\infty} f_j$$

and

$$\hat{e}_{m+} = n \sum_{j=m}^{\infty} P(X = j)$$

for the last cell.

The asymptotic distribution of G is χ^2_ν where $\nu = m - p - 1$ is the degree of freedom for a model with p parameters. If any of the expected class frequencies are less than five, classes are binned. Starting from the lowest class frequency, classes are binned one by one until the expected frequency is greater than or equal to five.

3.5 Estimating an unknown population size

There is a large number of estimators which are derived under homogeneity and heterogeneity of the target population. Examples of estimators based on geometric homogeneity are $\hat{N}_{\text{MLE}} = n/(1 - n/S)$, $\hat{N}_{\text{T}} = n/(1 - \sqrt{f_1/S})$ where $S = 0f_0 + 1f_1 + ... + mf_m$. Furthermore, the popular estimator which allow population heterogeneity is Chao's lower bound given by $\hat{N}_{\text{C}} = n + f_1^2/f_2$ under geometric model. As presented in the previous section, the one-truncated Turing estimator $\hat{N}_{\text{T\_OT}} = \frac{n}{1-\hat{p}_{\text{T\_OT}}}$ where $\hat{p}_{\text{T\_OT}} = \sqrt{\frac{f_2}{f_2+2f_3+...+(m-1)f_m}}$ and the one-truncated maximum likelihood estimator $\hat{N}_{\text{MLE\_OT}} = \frac{n}{1-(n-f_1)/(S-n)}$ allow for one-inflation.

To compare the suggested estimators with existing estimators, we look at the synthetic data of a population size $N = 20$ with 20% of one-inflation which generated from geometric distribution with parameter 0.1 as shown in Table 3.4. Table 3.5 provides population size estimated by conventional and proposed estimators, respectively. As can be seen, the proposed estimators (T_OT and MLE_OT) can effectively improve the estimates of population size from conventional estimators (Turing and MLE) and provide values closest to the parameter of interest $N = 20$. Therefore, the proposed estimators are viable and become candidates for use under one-inflation situation.

Table 3.4: The frequency of zero-truncated count data with 20% one-inflation from Section 3.1

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_{11}	f_{13}	f_{14}	n
(1)	7	1	1	1	1	3	1	1	1	1	1	19

Table 3.5: Estimates for the data in Table 3.4

Estimator	Estimated population size
Chao	68
Turing	26.24
MLE	23.95
T_OT	**21.52**
MLE_OT	**22.74**

Further simulation work is conducted to investigate how well these estimators are performing in term of bias, variance and mean square error.

3.6 Simulation study

The main purpose of this section is to study the performance of proposed estimators and to compare their behaviours with the other well-known estimators: Chao's lower bound estimator (C), Turing estimator (T) and maximum likelihood estimator (MLE).

3.6.1 Scope of study

1. The data were generated by the Monte Carlo technique using the program R, each condition was repeated 1,000 times.
2. The target population data were generated from a geometric model.
3. The population size was $N = 20, 50, 100, 500$ and $1,000$.
4. There were 2 levels of one-inflation: 20% and 50% represent low and high level of one-inflation respectively.

5. The criteria of comparing the performance of each estimator was relative bias (RBias), relative variance (RVar) and relative mean square error (RMSE).

3.6.2 A simulation plan

This simulation study was undertaken to investigate the performance of two proposed estimators and to compare these with other conventional estimators by means of the Monte Carlo method. The total size of the target population for each level of one-inflation was assumed to be 20, 50, 100, 500 and 1,000. The heterogeneity populations were generated from geometric distribution (arising from the mixture of Poisson distribution with an exponential distribution) with parameter $p \in \{0.1, 0.2, 0.3, ..., 0.6\}$. The simulation procedure is as follows:

1. Generate random number for X_i from geometric distribution with 20% and 50% of extra-ones for each (p); where $i = 1, 2, ..., N$. For example, 20% extra-ones with a population of 500 is generated by $N = 400$ follows a geometric distribution and the one-inflation has size 100.

2. Count the frequencies of each value of X_i begin with f_0 the count of $X_i = 0$, until f_m the frequencies of maximum value of X_m, then f_0 is dropped or truncated before going to the next step of estimator computation.

3. Estimate the total number of population sizes by means of Chao, Turing, MLE and the suggested estimators: one-truncated Turing (T_OT) and one-truncated MLE (MLE_OT).

4. Repeat the procedure (1) to (3) for 1,000 times

5. Compute the relative bias, relative variance and relative mean square error of population size estimator of each method

6. Rank the best performance of each condition by means of smallest relative bias, relative variance and relative mean square error.

3.6.3 Statistical investigation

For each scenario, all estimates were computed. Expected values and variance were determined as

$$Mean(\hat{N}) = E(\hat{N}) = \frac{1}{1000}\sum_{k=1}^{1000}\hat{N}_k$$

$$Var(\hat{N}) = \frac{\sum_{k=1}^{1000}(\hat{N}_k - E(\hat{N}))^2}{1000}.$$

The performance of population size estimators is evaluated in terms of bias, variance and mean square error. Due to the fact that the expected values and variance increase with increasing N, we take the relative bias, relative variance and relative mean square error as follows:

Relative Bias:

$$\begin{aligned} RBias(\hat{N}) &= \frac{E(\hat{N}) - N}{N} \\ &= \frac{(\frac{1}{1000}\sum_{k=1}^{1000}\hat{N}_k) - N}{N} \end{aligned} \tag{3.11}$$

Relative Variance:

$$\begin{aligned} RVar(\hat{N}) &= \frac{E(\hat{N} - N)^2}{N^2} \\ &= \frac{\frac{1}{1000}\sum_{k=1}^{1000}(\hat{N}_k - N)^2}{N^2} \end{aligned} \tag{3.12}$$

Relative Mean Square Error:

$$\begin{aligned} RMSE(\hat{N}) &= \frac{E(\hat{N} - N)^2}{N^2} \\ &= \frac{\frac{1}{1000}\sum_{k=1}^{1000}(\hat{N}_k - N)^2}{N^2} \end{aligned} \tag{3.13}$$

3.6.4 Simulation results

The results of the simulation study are divided into two parts; 1) 20% one-inflation and 2) 50% one-inflation. Each part reports the investigation of relative bias (Rbias), relative variance (RVar) and relative mean square error (RMSE) for all conditions of the study. Due to the fact that the results of two parts are similar, both parts will be summarized together. To explore preliminary the behaviour of estimators, we consider the mean of estimates of population size from all estimators.

According to the results provided in Table 3.6, it is noticeable that the results of two parts are similar (20% and 50% one-inflation). Clearly, all of the conventional estimators (Chao, Turing and MLE) show an overestimation of population size for all conditions of the study particularly it is severe in Chao's lower bound estimator. Turing and MLE estimators are less affected by one-inflation than Chao's lower bound. The proposed estimators $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ yield satisfying outcomes which are close to the true value of population size N. We can summarize the performance of estimators by ordering them as $\hat{N}_{\text{C}} > \hat{N}_{\text{T}} > \hat{N}_{\text{MLE}} > \hat{N}_{\text{MLE\_OT}} > \hat{N}_{\text{T\_OT}}$. It is also clear that newly proposed estimators $\hat{N}_{\text{MLE\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ perform considerably better than the others. This can indicate that the proposed estimators have a good performance under one-inflation, for both low and high level. However, if comparing among new proposed estimators,

$\hat{N}_{\text{T\_OT}}$ performs better than $\hat{N}_{\text{MLE\_OT}}$. We can make a preliminary conclusion that the proposed estimators can cope with one-inflation situation satisfactorily. Next, further investigation will be provided.
Note "-" in Table 3.6 - 3.9 is defined as no results from simulation study.

1) Investigate of relative bias (RBias)

The relative bias (RBias) is commonly defined as the difference between the estimated value and true value of N, and then scaled by the true value. Consequently, a good estimate of population size will have an associated relative bias close to zero. Additionally, the positive value of RBias presents an overestimation whereas the negative value of RBias shows an underestimation. As can be seen from Table 3.7, almost all estimators provide an overestimation for all conditions. There is only $\hat{N}_{\text{T\_OT}}$ that gives an underestimate in case of small population sizes ($N = 20, 50, 100$) under 20% one-inflation. Furthermore, the proposed estimators $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ show the highest performance of accuracy, respectively, by giving the smallest RBias among the other estimators for all the geometric parameter p and population sizes N. It is also found that $\hat{N}_{\text{C}}$ has the worst performance of accuracy as the difference between the expected value of the estimator and the true value of N is largest for all conditions of study. This can confirm in a severe overestimation of Chao's lower bound estimator under one-inflation situation as mentioned in beginning. In addition, consider an effect of the geometric parameter and the population size. Certainly, increasing the geometric parameter leads to a slight increase in bias for all estimators except $\hat{N}_{\text{C}}$ which slightly decrease in the beginning before increase. Conversely, an increase in population size lead to a slight decrease in bias for all estimators except $\hat{N}_{\text{T\_OT}}$.

2) Investigate of relative variance (RVar)

Variance is the common measure of variation. Variance of each estimator is the squared difference in average between an individual value of estimator and the expected value of estimator. Hence, a small variance of estimator can indicate that most individual values of estimators are close to their mean. To compare the variation among estimators for different population sizes, the relative variance (RVar) is calculated as the ratio of the variance and the expected value of estimator squared, see equation (3.12). As can be seen from Table 3.8, the RVar of 20% and 50% one-inflation for all estimators are similar patterns in the study. It is clearly seen that the suggested estimators $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ perform the best with the smallest RVar where $\text{RVar}(\hat{N}_{\text{MLE\_OT}}) > \text{RVar}(\hat{N}_{\text{T\_OT}})$. Performance of classical estimators $\hat{N}_{\text{T}}$ and $\hat{N}_{\text{MLE}}$ are fairly close to suggested estimators whereas $\hat{N}_{\text{C}}$ performs the worst with the largest RVar and significantly different from other estimators. Additionally, similar to the results of RBias, increasing the geometric parameter leads to a slight increase in RVar whereas an increase in population size leads

Table 3.6: Monte Carlo means of the population size estimates ($Mean(\hat{N})$) based upon geometric distribution with 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT
20%	20	0.1	43	23	21	**20**	21
		0.2	44	25	23	**19**	22
		0.3	50	28	26	**19**	-
	50	0.1	127	57	53	**49**	**51**
		0.2	113	62	57	**49**	53
		0.3	113	67	61	**47**	56
		0.4	123	75	69	**45**	-
	100	0.1	237	114	105	**100**	103
		0.2	207	122	112	**99**	105
		0.3	207	133	122	**97**	110
		0.4	221	148	136	**95**	117
		0.5	245	170	158	**90**	129
		0.6	296	208	194	**83**	-
	500	0.1	1056	569	525	**507**	511
		0.2	952	608	557	**514**	525
		0.3	971	658	602	**523**	543
		0.4	1014	725	665	**530**	572
		0.5	1124	822	756	**546**	607
		0.6	1303	985	913	**530**	671
	1000	0.1	2082	1138	1051	**1018**	1022
		0.2	1906	1220	1117	**1040**	1051
		0.3	1922	1315	1203	**1063**	1087
		0.4	2028	1446	1325	**1098**	1136
		0.5	2225	1641	1509	**1113**	1208
		0.6	2549	1943	1797	**1153**	1314
50%	20	0.1	123	30	25	**20**	22
		0.2	124	40	32	**20**	-
	50	0.1	501	72	60	**51**	53
		0.2	403	89	72	**50**	57
		0.3	392	115	94	**51**	64
		0.4	409	152	127	**51**	-
	100	0.1	1110	143	118	**102**	106
		0.2	761	178	144	**104**	114
		0.3	702	220	179	**104**	124
		0.4	709	281	233	**106**	140
		0.5	894	389	329	**108**	165
		0.6	1160	578	506	**109**	-
	500	0.1	4421	711	588	**521**	528
		0.2	3169	871	705	**547**	563
		0.3	2992	1073	872	**581**	609
		0.4	3159	1370	1128	**605**	675
		0.5	3605	1822	1530	**648**	764
		0.6	4512	2585	2229	**699**	909
	1000	0.1	8493	1420	1174	**1047**	1056
		0.2	6214	1738	1408	**1106**	1126
		0.3	5924	2149	1743	**1173**	1217
		0.4	6214	2713	2231	**1254**	1340
		0.5	7074	3584	3005	**1377**	1507
		0.6	8926	5159	4437	**1515**	1779

Table 3.7: Relative bias of five population size estimators based upon geometric distribution with 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT
20%	20	0.1	1.1651	0.1548	0.0687	**-0.0132**	0.0367
		0.2	1.1924	0.2441	0.1490	**-0.0547**	0.0791
		0.3	1.4577	0.3887	0.2813	**-0.0997**	-
	50	0.1	1.5486	0.1423	0.0540	**-0.0122**	0.0234
		0.2	1.2611	0.2308	0.1312	**-0.0280**	0.0643
		0.3	1.2570	0.3359	0.2297	**-0.0639**	0.1160
		0.4	1.4593	0.4921	0.3748	**-0.0954**	-
	100	0.1	1.3569	0.1416	0.0543	**-0.0015**	0.0251
		0.2	1.0700	0.2221	0.1202	**-0.0120**	0.0539
		0.3	1.0685	0.3301	0.2179	**-0.0295**	0.0995
		0.4	1.2093	0.4762	0.3563	**-0.0506**	0.1684
		0.5	1.4593	0.7045	0.5771	**-0.0951**	0.2912
		0.6	1.9613	1.0826	0.9449	**-0.1671**	-
	500	0.1	1.1111	0.1383	0.0505	**0.0146**	0.0219
		0.2	0.9036	0.2160	0.1147	**0.0279**	0.0500
		0.3	0.9425	0.3167	0.2035	**0.0453**	0.0865
		0.4	1.0284	0.4494	0.3301	**0.0608**	0.1433
		0.5	1.2484	0.6446	0.5125	**0.0918**	0.2136
		0.6	1.6052	0.9705	0.8251	**0.0606**	0.3417
	1000	0.1	1.0817	0.1382	0.0507	**0.0177**	0.0223
		0.2	0.9060	0.2199	0.1167	**0.0397**	0.0506
		0.3	0.9225	0.3154	0.2033	**0.0637**	0.0873
		0.4	1.0278	0.4464	0.3246	**0.0981**	0.1359
		0.5	1.2251	0.6414	0.5088	**0.1131**	0.2079
		0.6	1.5488	0.9433	0.7970	**0.1531**	0.3136
50%	20	0.1	5.1590	0.5061	0.2380	**0.0157**	0.0776
		0.2	5.2127	0.9808	0.6178	**0.0142**	-
	50	0.1	9.0240	0.4460	0.1910	**0.0164**	0.0602
		0.2	7.0616	0.7890	0.4489	**0.0081**	0.1403
		0.3	6.8339	1.2990	0.8843	**0.0269**	0.2812
		0.4	7.1889	2.0316	1.5483	**0.0165**	-
	100	0.1	10.1014	0.4310	0.1813	**0.0220**	0.0582
		0.2	6.6093	0.7802	0.4417	**0.0411**	0.1384
		0.3	6.0221	1.1979	0.7921	**0.0441**	0.2414
		0.4	6.0930	1.8060	1.3315	**0.0623**	0.4024
		0.5	7.9437	2.8877	2.2871	**0.0801**	0.6529
		0.6	10.5987	4.7808	4.0617	**0.0892**	-
	500	0.1	7.8420	0.4218	0.1753	**0.0413**	0.0567
		0.2	5.3389	0.7414	0.4103	**0.0938**	0.1264
		0.3	4.9847	1.1470	0.7434	**0.1627**	0.2185
		0.4	5.3181	1.7420	1.2570	**0.2107**	0.3491
		0.5	6.2104	2.6430	2.0597	**0.2959**	0.5270
		0.6	8.0233	4.1704	3.4580	**0.3992**	0.8185
	1000	0.1	7.4933	0.4198	0.1738	**0.0470**	0.0558
		0.2	5.2137	0.7375	0.4078	**0.1065**	0.1259
		0.3	4.9243	1.1487	0.7433	**0.1729**	0.2175
		0.4	5.2138	1.7126	1.2313	**0.2543**	0.3397
		0.5	6.0734	2.5844	2.0053	**0.3774**	0.5071
		0.6	7.9257	4.1587	3.4367	**0.5150**	0.7790

Table 3.8: Relative variance of five population size estimators based upon geometric distribution with 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT
20%	20	0.1	1.0832	0.0117	0.0068	**0.0045**	0.0057
		0.2	1.5992	0.0515	0.0343	**0.0109**	0.0245
		0.3	2.4632	0.1236	0.0891	**0.0175**	-
	50	0.1	1.8850	0.0046	0.0028	**0.0024**	**0.0024**
		0.2	1.5986	0.0139	0.0087	**0.0050**	0.0066
		0.3	1.8110	0.0356	0.0262	**0.0094**	0.0214
		0.4	2.2841	0.0948	0.0757	**0.0157**	-
	100	0.1	1.0718	0.0023	0.0013	0.0012	**0.0011**
		0.2	0.6283	0.0071	0.0046	**0.0031**	0.0034
		0.3	0.4570	0.0166	0.0112	**0.0057**	0.0081
		0.4	0.5752	0.0392	0.0288	**0.0098**	0.0228
		0.5	1.2910	0.1087	0.0891	**0.0147**	0.0874
		0.6	2.4855	0.3422	0.3100	**0.0201**	-
	500	0.1	0.0904	0.0005	0.0003	0.0003	**0.0002**
		0.2	0.0499	0.0013	0.0008	0.0007	**0.0006**
		0.3	0.0489	0.0030	0.0021	0.0017	**0.0016**
		0.4	0.0608	0.0068	0.0052	**0.0036**	0.0038
		0.5	0.1112	0.0163	0.0128	**0.0074**	0.0100
		0.6	0.2212	0.0496	0.0423	**0.0125**	0.0354
	1000	0.1	0.0417	0.0002	**0.0001**	**0.0001**	**0.0001**
		0.2	0.0255	0.0006	0.0004	0.0004	**0.0003**
		0.3	0.0257	0.0016	0.0011	0.0010	**0.0008**
		0.4	0.0369	0.0037	0.0027	0.0019	**0.0018**
		0.5	0.0520	0.0077	0.0060	**0.0039**	0.0045
		0.6	0.0966	0.0212	0.0180	**0.0088**	0.0156
50%	20	0.1	8.5233	0.1108	0.0515	**0.0031**	0.0081
		0.2	10.6954	0.6657	0.4525	**0.0079**	-
	50	0.1	31.4384	0.0194	0.0074	**0.0015**	0.0021
		0.2	25.9054	0.0804	0.0395	**0.0040**	0.0083
		0.3	31.3470	0.4388	0.3094	**0.0081**	0.0577
		0.4	31.0232	1.2350	1.0519	**0.0119**	-
	100	0.1	57.6632	0.0086	0.0032	**0.0010**	**0.0010**
		0.2	23.8094	0.0373	0.0187	**0.0029**	0.0039
		0.3	22.0421	0.1134	0.0683	**0.0062**	0.0133
		0.4	15.5799	0.3139	0.2355	**0.0108**	0.0608
		0.5	37.1924	1.3921	1.0667	**0.0185**	0.3544
		0.6	70.9817	6.9650	6.2770	**0.0379**	-
	500	0.1	4.5555	0.0015	0.0005	0.0003	**0.0002**
		0.2	1.5720	0.0060	0.0028	0.0009	**0.0007**
		0.3	1.1085	0.0173	0.0103	0.0024	**0.0022**
		0.4	1.2783	0.0578	0.0389	**0.0048**	0.0070
		0.5	1.9032	0.1659	0.1225	**0.0119**	0.0210
		0.6	4.1770	0.6265	0.5235	**0.0333**	0.1103
	1000	0.1	1.9314	0.0007	0.0002	**0.0001**	**0.0001**
		0.2	0.6574	0.0028	0.0014	0.0005	**0.0003**
		0.3	0.5639	0.0094	0.0052	0.0013	**0.0010**
		0.4	0.6431	0.0279	0.0182	**0.0031**	0.0033
		0.5	0.9498	0.0767	0.0573	**0.0082**	0.0108
		0.6	2.0447	0.3160	0.2499	**0.0268**	0.0386

Table 3.9: Relative mean square error of five population size estimators based upon geometric distribution with 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT
20%	20	0.1	2.4394	0.0356	0.0115	**0.0046**	0.0071
		0.2	3.0195	0.1110	0.0565	**0.0139**	0.0307
		0.3	4.5856	0.2745	0.1682	**0.0274**	-
	50	0.1	4.2813	0.0249	0.0057	**0.0026**	0.0030
		0.2	3.1874	0.0671	0.0259	**0.0058**	0.0108
		0.3	3.3891	0.1484	0.0789	**0.0135**	0.0348
		0.4	4.4114	0.3368	0.2161	**0.0248**	-
	100	0.1	2.9118	0.0224	0.0042	**0.0012**	0.0017
		0.2	1.7724	0.0564	0.0190	**0.0032**	0.0063
		0.3	1.5982	0.1256	0.0587	**0.0066**	0.0080
		0.4	2.0372	0.2659	0.1557	**0.0124**	0.0511
		0.5	3.4193	0.6050	0.4221	**0.0238**	0.1721
		0.6	6.3297	1.5139	1.2024	**0.0480**	-
	500	0.1	1.3250	0.0196	0.0028	**0.0005**	0.0007
		0.2	0.8664	0.0479	0.0139	**0.0015**	0.0031
		0.3	0.9371	0.1033	0.0436	**0.0038**	0.0091
		0.4	1.1184	0.2087	0.1141	**0.0073**	0.0243
		0.5	1.6695	0.4317	0.2754	**0.0158**	0.0556
		0.6	2.7978	0.9913	0.7230	**0.0162**	0.1522
	1000	0.1	1.2117	0.0193	0.0027	**0.0005**	0.0006
		0.2	0.8463	0.0490	0.0140	**0.0020**	0.0028
		0.3	0.8767	0.1011	0.0425	**0.0050**	0.0085
		0.4	1.0932	0.2030	0.1080	**0.0115**	0.0203
		0.5	1.5527	0.4192	0.2649	**0.0167**	0.0478
		0.6	2.4953	0.9109	0.6533	**0.0322**	0.1139
50%	20	0.1	35.1301	0.3668	0.1081	**0.0034**	0.0141
		0.2	37.8571	1.6271	0.8338	**0.0080**	-
	50	0.1	112.8388	0.2183	0.0438	**0.0018**	0.0047
		0.2	75.7459	0.7028	0.2410	**0.0041**	0.0279
		0.3	78.0177	2.1259	1.0911	**0.0088**	0.1368
		0.4	82.6723	5.3611	3.4481	**0.0121**	-
	100	0.1	159.6436	0.1944	0.0361	**0.0015**	0.0044
		0.2	67.4689	0.6460	0.2138	**0.0046**	0.0230
		0.3	58.2859	1.5483	0.6956	**0.0082**	0.0716
		0.4	52.6894	3.5753	2.0081	**0.0147**	0.2227
		0.5	100.2575	9.7297	6.2962	**0.0250**	0.7803
		0.6	183.2422	29.8142	22.7683	**0.0458**	-
	500	0.1	66.0486	0.1794	0.0313	**0.0020**	0.0034
		0.2	30.0747	0.5557	0.1712	**0.0097**	0.0167
		0.3	25.9548	1.3329	0.5630	**0.0289**	0.0500
		0.4	29.5591	3.0922	1.6189	**0.0492**	0.1289
		0.5	40.4701	7.1514	4.3646	**0.0995**	0.2987
		0.6	68.5455	18.0183	12.4810	**0.1927**	0.7802
	1000	0.1	58.0789	0.1769	0.0305	**0.0023**	0.0032
		0.2	27.8394	0.5468	0.1677	**0.0118**	0.0162
		0.3	24.8117	1.3289	0.5577	**0.0312**	0.0483
		0.4	27.8260	2.9607	1.5343	**0.0678**	0.1187
		0.5	37.8369	6.7557	4.0783	**0.1506**	0.2679
		0.6	64.8600	17.6102	12.0607	**0.2920**	0.6458

to a slight decrease in RVar for all estimators except $\hat{N}_{\mathrm{C}}$.

3) Investigate of relative mean square error (RMSE)

Relative mean square error (RMSE) shows the ratio of the squared difference between each value of estimator and the true value of parameter over the true value of parameter, averaged over the sample space. The estimator, which gives a smallest value of RMSE, normally indicates that this estimator shows the highest efficient estimation, on average is closest to the true value of parameter of interest. The RMSE of each estimator under consideration are presented in Table 3.9. Similar to the results in the investigation of RVar, the suggested estimators $\hat{N}_{\mathrm{T\_OT}}$ and $\hat{N}_{\mathrm{MLE\_OT}}$ also perform the best by giving the smallest value of RMSE whereas $\hat{N}_{\mathrm{C}}$ performs the worst with the largest RMSE. As a summary of performance with regard to relative mean square error, ordering of the estimators is $\mathrm{RMSE}(\hat{N}_{\mathrm{C}}) > \mathrm{RMSE}(\hat{N}_{\mathrm{T}}) > \mathrm{RMSE}(\hat{N}_{\mathrm{MLE}}) > \mathrm{RMSE}(\hat{N}_{\mathrm{MLE\_OT}}) > \mathrm{RMSE}(\hat{N}_{\mathrm{T\_OT}})$ for all studied cases. Moreover, increasing the geometric parameter leads to an increase in RMSE whereas an increasing the population size leads to decreasing in RMSE for all estimators except $\hat{N}_{\mathrm{C}}$.

3.7 An application for estimating the population size

The aim of this section is to apply the proposed estimators (T_OT and MLE_OT) to real data where we suspect one-inflation is ongoing, and compare their performance with conventional estimators (Chao, Turing and MLE).

1) Estimating the total number of scrapie infected holding in France

According to the case study of estimating the total number of holdings with scrapie infection in France (see Table 3.1 in Section 3.2 for the data and Figure 3.2 left panel for the ratio plot of one-inflation), the results of estimation ($\hat{f}_0$ and $\hat{N}$) from a variety of methods are shown in Table 3.11. As we expect, due to an effect of one-inflation, there is a large difference between the conventional estimators and those accounting for one-inflation. Chao's lower bound estimator yields with $1,267$ a huge number of estimated scrapie-infected holdings. Next, the conventional Turing and MLE estimator give a more moderate estimate; 902 and 827, respectively. Finally, the smallest estimates are given by the proposed estimators $\hat{N}_{\mathrm{T\_OT}} = 427$ and $\hat{N}_{\mathrm{MLE\_OT}} = 454$. It can be seen clearly that the proposed estimators can reduce the overestimation associated with conventional estimators.

Although the simulation studies indicate that the suggested estimators can properly deal with one-inflation, we should also consider goodness-of-fit in model fitting when we apply these estimators in real situations. Figure 3.5 left panel shows the fitted values for this data set with all estimators. It is clear that the estimated values from T_OT

Table 3.10: The data of French scrapie-infected holdings from section 3.2

x	1	2	3	4
f_x	121	13	5	2

and MLE_OT, represented by purple line and orange line on the graph, fit the data very well and better than the conventional estimators. This agrees with the p-value from the goodness of fit statistics in Table 3.11.

Table 3.11: Results for scrapie-infected holdings in France

Estimator	$\hat{f}_0$	$\hat{N}$	Chi-square	p-value
Chao[1]	1,126	1,267	27.195	0.00000
Turing	761	902	8.487	0.01436
MLE	686	827	6.781	0.03369
T_OT	**286**	**427**	**0.283**	**0.59474**
MLE_OT	313	454	0.507	0.47644

2) Estimating the total number of domestic violence in the Netherlands

Coming back to the application of estimating the population size of domestic violence offenders in Section 3.2, we show the data again in Table 3.12. The results of estimation from the classical and proposed estimators are shown in Table 3.13. As we expect, the pattern of results is similar to the previous application, $\hat{N}_{\mathrm{C}} > \hat{N}_{\mathrm{T}} > \hat{N}_{\mathrm{MLE}} > \hat{N}_{\mathrm{T\_OT}} > \hat{N}_{\mathrm{MLE\_OT}}$, but here estimators are very few in their magnitude which corresponds to the unclear one-inflation signal of ratio plot in Figure 3.3. Nevertheless, it is clear that the proposed estimators can reduce overestimation associated with the conventional estimators.

Table 3.12: The data of domestic violence from Section 3.2

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

Table 3.13: Results for domestic violence study

Estimator	$\hat{f}_0$	$\hat{N}$	Chi-square	p-value
Chao[1]	117,577	135,223	317.537	0.00000
Turing	103,233	120,879	166.795	0.00000
MLE	98,788	116,434	144.797	0.00000
T_OT	65,573	83,219	7.227	0.02696
MLE_OT	**64,754**	**82,400**	**6.649**	**0.03599**

[1]For GOF-test, $\hat{p}_0 = \frac{\hat{f}_0}{\hat{N}}$ and $p = p_0$ for geometric model

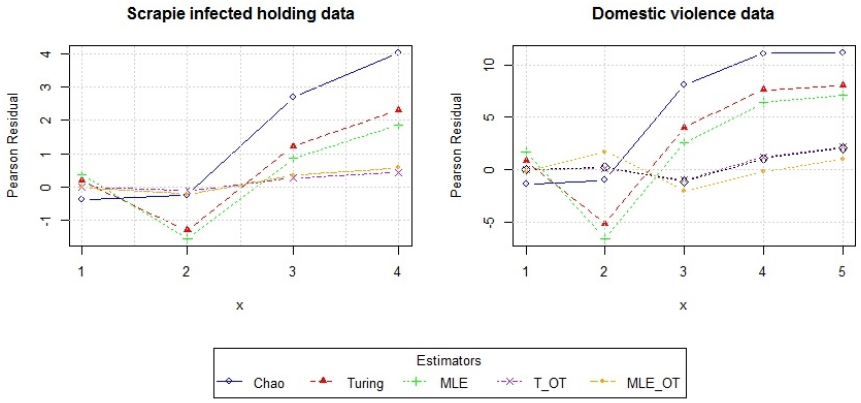

Figure 3.5: Square root of goodness of fit test charts for all estimators for scrapie infected holding data (left panel) and domestic violence data (right panel)

In terms of statistical model fitting, Figure 3.5 right panel shows the fitted values for this data set by using standardized residuals with all estimators. The conventional estimators are shown by blue, red and green lines, whereas the purple and orange lines are for the proposed estimators. It can be clearly seen from the graph and the p-value from the goodness of fit statistics in Table 3.13 that the estimated values from proposed estimators can fit the data well and definitely improve upon the fitting of the conventional estimators. Nonetheless, the p-values of the proposed estimators indicate that the one-truncated geometric model may not be able to fit this data set good enough if we consider the level of significance at 0.05 or 0.10.

3.8 Conclusion/Discussion

Chao's lower bound, Turing and maximum likelihood estimator are some of the most popular estimators used to estimate the elusive target population size in capture-recapture. Turing and maximum likelihood estimation are developed under the Poisson homogeneity assumption whereas Chao's lower bound is developed allowing heterogeneity. In this chapter, it is shown that these estimators can show a weak performance by producing an overestimation bias when these estimators experience one-inflation, particularly a severe overestimation for a Chao's lower bound. To cope with this problem two new estimators are proposed based upon modified forms of Turing and maximum likelihood estimation under the one-truncated geometric model.

To evaluate the performance of the proposed estimators, comparisons are done among existing conventional estimators. The simulation study considers the population generated from geometric distribution with 20% and 50% one-inflation. The size of population is 20, 50, 100, 500 and 1,000. The simulation results provides evidence that the proposed estimators show a good performance of accuracy and perform best with the smallest

mean square error for all conditions of study. If we compare only between the two newly proposed estimators, modified Turing is better than modified MLE all of case studies. As an application, we illustrate two case studies with one-inflation which were investigated previously by the ratio plot. The first case study looks at the total number of scrapie infected holdings in France, while the second is interested in the size of domestic violence in the Netherlands. Both examples show that the proposed estimators can cope with the problem of one-inflation by providing smaller estimates than conventional estimators. Nevertheless, only smaller estimates cannot confirm that the proposed estimators can be viable in real situations. Hence it requires considering statistical model fitting. Here the goodness of fit statistic is used for checking. It is found that the fitted values on the basis of the proposed estimators can suit the data with one-inflation well and better than conventional estimators in both of case studies.

To sum up, the proposed estimators under a one-truncated geometric distribution show a good performance in both, simulation and applications. However, in this chapter we applied the proposed estimators with only two case studies. It may not be appropriate and good enough for other applications. Therefore, it cannot guarantee that the proposed estimators will be practicable with one-inflation data in all real life situations. Moreover, the proposed estimators are not developed by a model approach, it is just based on modification by truncating counts of one and then applying the conventional Turing and MLE approach to the one-truncated data. It does not involve a model mechanism that describes how extra ones are generated. This point will be the project for Chapter 4 where we focus on developing a model that describes the mechanism for extra one generation. Furthermore, the proposed estimators in this chapter will be included in simulation experiment in Chapter 4 and we will see whether there are any benefits of the model-based approach. The EM algorithm that will be used in Chapter 4 is also more complex and more computational demanding so it would be beneficial if we could also deal with simpler approach provided in this chapter. However, final conclusions can only be reached after the results of Chapter 4 are available and we need to postpone final assessments to the next chapter.

Chapter 4

Zero-Truncated One-Inflated Geometric Distribution

This chapter focuses on developing a new model for capture-recapture estimation in order to deal with a one-inflation. This model describes the statistical mechanism for generating an extra of singletons. It is denoted as the zero-truncated one-inflated geometric model (ZTOI). A new estimator of the population size is also developed by the maximum likelihood approach based on the ZTOI geometric model ($\hat{N}_{\text{MLE\_ZTOI}}$). The nested EM algorithm is discussed for maximum likelihood estimation as no closed form solutions are available. As an evaluation, performance of the new proposed estimator ($\hat{N}_{\text{MLE\_ZTOI}}$) is investigated and compared to the previously proposed estimators from Chapter 3 ($\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$) in a simulation study. The applications are illustrated in the last section and the likelihood ratio test is used to check the presence of one-inflation in real data. Success of the proposed estimators is shown by simulation and applications.

4.1 Introduction

Knowledge of population size is of key importance in many fields of researches such as animal ecology, evolution, conservation biology, public health, epidemiology and criminology. For natural elusive populations, it is rarely possible to count all individuals. Therefore, capture-recapture estimation approach usually is used for estimating population size. The performance of capture-recapture models depends on their assumptions; these assumptions can be violated in many fields as it was mentioned previously. Critical assumptions are whether capture probability remains constant, changes with time or as behavioural response to previous experience, or varies among individuals. These might affect to the data being in one-inflation form. Furthermore, the reliability of statistical inference in capture-recapture studies depends on the quality of observed data.

This means that the correctness of the population size estimate depends on the initial identification of the sampled individuals. If sample individuals are classified into wrong classes, it can also cause the problem of one-inflation.

The variables of interest in a capture-recapture experiment are the frequency counts of identified individuals (f_x) and it was already defined in Chapter 2 that $f_1, f_2, f_3, ..., f_m$ represent the frequencies of different individuals identified exactly $1, 2, 3, ..., m$ times during the study period. Moreover, f_0 is the frequency of individuals that were not identified or observed in the study. Therefore, the unknown population size (N) can be calculated by $N = f_0+f_1+f_2+...+f_m$ or we can say that $N = f_0+n$ where $n = \sum_{x=1}^{m} f_x$. In fact, estimating f_0 leads to an estimate of population size N. Let $p_x = P(X = x)$ denote probability for identifying an individual exactly x times. Accordingly, p_0 is the probability of identifying an individual 0 times. As a result, the unknown population size can be defined as $N = Np_0 + E(n)$, where we treat n as a random variable. It can easily be solved for N and replacing $E(n)$ by its moment estimator n leads to the Horvitz-Thompson estimator

$$\hat{N} = \frac{n}{1 - p_0}. \tag{4.1}$$

Generally, p_0 is unknown and depends on model parameter so modelling for count probability p_x becomes one of main concerns. Maximum likelihood approach is a popular statistical method for estimating unknown parameters of a probability model. A parameter is a descriptor of the model. Likelihood is defined to be a quantity proportional to the probability of observing the data given the model. Thus, we can calculate the probability the observations which have actually been observed as a function of the model if we have a model (general, specific or modified model). Maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for capture-recapture estimation situations. Maximum likelihood methods have desirable mathematical and optimality properties. Specifically, they become minimum variance unbiased estimators as the sample size increases. These good properties are interesting and inducing to use for developing estimator. To cope with the problem of one-inflation, hence, the basic idea of this chapter is that a statistical model is built that describes the mechanism to generate the extra of count ones. A new estimator is developed from the maximum likelihood approach by using the nested EM algorithm based upon the zero-truncated one-inflated geometric distribution.

4.2 Zero-truncated one-inflated geometric model

A one-inflation model is a statistical model based on a probability distribution which allows for frequent one observations. A one-inflation model employs two components that correspond to two one-generating processes. The first process is governed by a binary distribution that generates structural ones. The second process is generated by

a probability density function of model $f_x(\theta)$ that generates counts, some of which may be one. The two components of a one-inflation model for θ are described as follows:

$$p_x = \begin{cases} \omega f_x(\theta) & , \quad \text{if} \quad x \neq 1 \\ (1-\omega) + \omega f_x(\theta) & , \quad \text{if} \quad x = 1 \end{cases}$$

where ω is an unknown weight parameter; $0 \leq \omega \leq 1$. Assume that $x_1, x_2, ..., x_n$ are observed and drawn from a geometric distribution with mean $(1-\theta)/\theta$, where $f_x(\theta) = (1-\theta)^x\theta\,;\, x = 0, 1, 2, ...$ Thus, a one-inflation geometric probability density function is

$$p_x = \begin{cases} \omega(1-\theta)^x\theta & , \quad \text{if} \quad x \neq 1 \\ (1-\omega) + \omega(1-\theta)^x\theta & , \quad \text{if} \quad x = 1. \end{cases} \tag{4.2}$$

The parameter $1-\omega$ represents the proportion of extra-ones present in the population which are not generated by the mechanism provided by $f_x(\theta)$ or a geometric distribution. However, due to the fact that over the study period of the capture-recapture experiment all observed units were identified at least once, we need to incorporate zero-truncation of the one-inflation geometric distribution and results in:

$$p_x^{1+} = \begin{cases} \omega(1-\theta)^x\theta/[1-\omega\theta] & , \quad \text{if } x \neq 1 \\ [(1-\omega) + \omega(1-\theta)^x\theta] \,/\, [1-\omega\theta] & , \quad \text{if } x = 1. \end{cases} \tag{4.3}$$

The observed, incomplete data log-likelihood for a zero-truncation-one-inflation geometric distribution is

$$\begin{aligned} l_A(\omega, \theta) &= \sum_{x=1}^{m} f_x \log p_x^{1+} \\ &= f_1 \log \left\{ \frac{(1-\omega) + \omega(1-\theta)\theta)}{1-\omega\theta} \right\} + \sum_{x=2}^{m} f_x \log \left\{ \frac{\omega(1-\theta)^x\theta}{1-\omega\theta} \right\} \\ &= f_1 \log \left\{ \frac{(1-\omega) + \omega(1-\theta)\theta)}{1-\omega\theta} \right\} + (n-f_1)\left\{\log\omega + \log\theta - \log(1-\omega\theta)\right\} \\ &\quad + (S-f_1)\log(1-\theta) \end{aligned}$$

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

4.3 Zero-truncated one-inflated maximum likelihood estimator via an EM algorithm

The EM algorithm is a popular method for maximum likelihood estimation. McLachlan and Krishnan (1997) stated that a general purpose of the EM algorithm is to cope with incomplete-data problem for maximum likelihood estimation. In addition, it composes of two steps, the Expectation (E-step) and the Maximization (M-step). In the E-step, we replace all missing data by their expected values that are calculated from the observed

data and the current estimates of likelihood parameters. In the M-step, we maximize the likelihood function by using both the observed and imputed data. The EM algorithm is an iterative method, so the procedure alternates between E-step and M-step until estimates of the likelihood parameters converge.

Here, we wish to fit the zero-truncated one-inflated geometric distribution to the frequency data in capture-recapture. The complete data log-likelihood is required. On defining the complete data as $f_x, x = 0, 1, 2, ..., m$, this situation can be viewed as a missing data problem since f_0 is unobserved. If f_0 is given, the maximum likelihood estimators are available. The EM algorithm can be used by imputing a value for f_0 and then maximize the non-zero-truncated distribution. Iterating through these two steps gives us a maximum likelihood estimate for θ and ω. The likelihood for the one-inflated distribution can be maximized by means of the EM algorithm. Embedding another EM into the M-step of the outer EM algorithm gives us a nested EM.

4.3.1 EM algorithm for zero-truncated part (Outer part)

The first step is to specify an initial values by letting $\hat{\omega}_{(0)} = 1/2$ and finding the initial value for $\hat{\theta}_{(0)}$ from $E(X)$; $X \sim Geo(\theta)$

$$E(X) = \frac{1-\theta}{\theta} = \frac{1}{\theta} - 1$$

$$\frac{1}{\hat{\theta}_{(0)}} = \frac{\sum_{x=0}^{m} x f_x}{n} + 1 = \frac{\sum_{x=0}^{m} x f_x + n}{n}$$

$$\hat{\theta}_{(0)} = \frac{n}{\sum_{x=0}^{m} x f_x + n} = \frac{1}{1+\bar{x}}$$

Thus, the estimated probability $X = 0$ given the observed data is

$$\hat{p}_{0(0)} = \hat{\omega}_{(0)}\hat{\theta}_{(0)} = \frac{1}{2(1+\bar{x})}.$$

E-step: In order to estimate f_0, the EM algorithm is used as an instrument to solve this problem. By the E-step, the unobserved frequency f_0 is replaced by its expected value given observed frequencies, $(n = f_1 + f_2 + ... + f_m)$, and current estimates of likelihood estimators. Let $\hat{f}_0$ denotes the estimate of the expected value of f_0 which can

be achieved as follows:

$$\begin{aligned}\hat{f}_0 &= E(f_0|\text{observed data}\,;\,\theta)\\ &= E(f_0|f_1, f_2, ..., f_m\,;\,\theta)\\ &= Np_0\\ &= (n+\hat{f}_0)p_0\\ &= np_0 + \hat{f}_0 p_0\end{aligned}$$

The expected frequency of zero counts is

$$\hat{f}_0 = \frac{np_0}{1-p_0},$$

where $n = \sum_{x=1}^{m} f_x$ is the number of observed units and $\hat{N} = n + \hat{f}_0$.

M-step: The associated complete data log-likelihood is

$$l(\omega, \theta) = \sum_{x=0}^{m} f_x \log p_x$$

where p_x is a one-inflated geometric probability density function, see (4.2). We need to find $\hat{\omega}$ and $\hat{\theta}$ that maximize $l(\omega, \theta)$ to complete the M-step. Unfortunately, M-step cannot be solved in closed form. Therefore, we use another EM algorithm to solve the M-step.

4.3.2 EM algorithm for one-inflated part (Inner part)

This can be accomplished by introducing a binary indicator variable z_i defined as

$$z_i = \begin{cases} 1 & \text{, if the sample value one is from the extra-ones population} \\ 0 & \text{, otherwise.} \end{cases}$$

This leads to the unobserved, complete likelihood function given as:

$$L(x; \omega, \theta) = \prod_{x_i=1} (1-\omega)^{z_i} [\omega(1-\theta)^{x_i}\theta]^{1-z_i} \prod_{x_i \neq 1} [\omega(1-\theta)^{x_i}\theta]. \tag{4.4}$$

The log-likelihood is

$$\begin{aligned} l(x; \omega, \theta) = &\sum_{x_i=1} [z_i log(1-\omega) + (1-z_i) log\omega + (1-z_i)x_i log(1-\theta) + (1-z_i) log\theta] \\ &+ \sum_{x_i \neq 1} [log\omega + x_i log(1-\theta) + log\theta] \end{aligned}$$

which can be simplified to

$$\begin{aligned} l(x;\omega,\theta) = &\sum_{x_i=1} z_i[log(1-\omega) - log\omega] + Nlog\omega + \sum_{i=1}^{N} x_i log(1-\theta) + Nlog\theta \\ &- \sum_{x_i=1} z_i[x_i log(1-\theta) + log\theta]. \end{aligned} \tag{4.5}$$

Nested E-step: The unobserved indicator z_i is treated as missing data. In the E-step, z_i is replaced by its expected value e_i conditional upon the observed data and current values of ω and θ. Moreover, e_i can be determined as the posterior probability that observation i belongs to extra-ones and can be calculated by the following version of Bayes's theorem:

$$\begin{aligned} e_i &= E(z_i \mid x_i;\omega,\theta) = P(z_i = 1 \mid x_i = 1;\omega,\theta) \\ &= \frac{P(x_i = 1 \mid z_i = 1;\omega,\theta)P(z_i = 1 \mid \omega,\theta)}{[P(x_i = 1 \mid z_i = 1)P(z_i = 1) + P(x_i = 1 \mid z_i = 0)P(z_i = 0)]} \\ &= \frac{1-\omega}{[(1-\omega) + \omega f_1(\theta)]}, \end{aligned}$$

where $f_1(\theta)$ is the geometric probability for a one, so

$$e_i = P(z_i = 1 \mid x_i = 1;\omega,\theta) = \frac{1-\omega}{[(1-\omega) + \omega(1-\theta)\theta]}. \tag{4.6}$$

Now z_i is replaced by its expected values e_i.

Nested M-step: Let $\sum_1 = \sum_{x_i=1} e_i$. To find MLEs of ω and θ, the log-likelihood with z_i replaced by e_i in (4.5) is maximized by taking a derivative with respect to ω and setting it equal to 0,

$$\begin{aligned} \frac{\partial l}{\partial \omega} = -\frac{\sum_1}{1-\omega} - \frac{\sum_1}{\omega} + \frac{\hat{N}}{\omega} &= 0 \\ \frac{\hat{N}}{\omega} = \frac{\sum_1}{1-\omega} + \frac{\sum_1}{\omega} \\ 1 - \omega = \frac{\sum_1}{\hat{N}} \end{aligned}$$

Hence,

$$\hat{\omega} = 1 - \frac{\sum_1}{\hat{N}} \tag{4.7}$$

Then taking a derivative with respect to θ and setting it equal to 0, we yield

$$\frac{\partial l}{\partial \theta} = -\frac{\sum_{i=1}^{\hat{N}} x_i}{1-\theta} + \frac{\hat{N}}{\theta} + \frac{\sum_1}{1-\theta} - \frac{\sum_1}{\theta} = 0,$$

or

$$\frac{\hat{N}}{\theta} - \frac{\sum_1}{\theta} = \frac{\sum_{i=1}^{\hat{N}} x_i}{1-\theta} - \frac{\sum_1}{1-\theta},$$

or

$$\frac{1-\theta}{\theta} = \frac{\sum_{i=1}^{\hat{N}} x_i - \sum_1}{\hat{N} - \sum_1},$$

finally

$$\frac{1}{\theta} - 1 = \frac{\sum_{i=1}^{\hat{N}} x_i - \sum_1}{\hat{N} - \sum_1}.$$

Hence,

$$\hat{\theta} = \frac{\hat{N} - \sum_1}{\hat{N} + \sum_{i=1}^{\hat{N}} x_i - 2\sum_1} \tag{4.8}$$

In summary, we have

$$\hat{\omega} = 1 - \frac{f_1}{\hat{N}}(1-\omega)/[(1-\omega) + \omega(1-\theta)\theta] \tag{4.9}$$

and

$$\hat{\theta} = \frac{\hat{N} - f_1(1-\omega)/[(1-\omega) + \omega(1-\theta)\theta}{\hat{N} + \sum_{i=1}^{\hat{N}} x_i - 2f_1(1-\omega)/[(1-\omega) + \omega(1-\theta)\theta]}. \tag{4.10}$$

The equation (4.9) and (4.10) have to be interpreted in the way that ω represents the current value and $\hat{\omega}$ is the solution from the M-step for the new iteration. Note also that f_1 is the frequency of ones. Also, $\hat{N}$ refers to the current value of $\hat{f}_0$ leading to $\hat{N} = \hat{f}_0 + n$.

Convergence Criterion determines when iterations are stopped. For the outer EM, iterations are ceased when

$$\mid \hat{f}_{0(k)} - \hat{f}_{0(k-1)} \mid < \varepsilon$$

For the inner EM, iterations are ceased when all parameter estimates meet the criteria

$$\mid \hat{\omega}_{(l)} - \hat{\omega}_{(l-1)} \mid < \varepsilon$$

and

$$\mid \hat{\theta}_{(l)} - \hat{\theta}_{(l-1)} \mid < \varepsilon$$

Consequently, the population size estimator based upon zero-truncated one-inflated geometric model through the Horvitz-Thompson approach is

$$\hat{N}_{ZTOI} = \frac{n}{1-\hat{p}_0} \qquad \text{where} \quad \hat{p}_0 = \hat{\omega}\hat{\theta}$$

In summary, the algorithm used to compute the estimate of population size is given as follows.

Step 0 : choose initial values for $\hat{\omega}_{(0)}$ and $\hat{\theta}_{(0)}$, and set $k = 0, l = 0$.

Table 4.1: The complete frequency table

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

Here, we set $\hat{\omega}_{(0)} = 1/2$ and using the complete frequency data to calculate initial value of θ, where $\hat{\theta}_{(0)} = 1/(1+\bar{x})$; $\bar{x} = \sum_{x=0}^{m} x f_x/n$, hence $\hat{p}_{0(0)} = 1/2(1+\bar{x})$.

Step 1 : E-step

Set $k = k+1$, compute $\hat{f}_{0(k)}$ by using $\hat{\omega}_{(k-1)}$ and $\hat{\theta}_{(k-1)}$

$$\hat{f}_{0(k)} = \frac{n\hat{p}_{0(k-1)}}{1-\hat{p}_{0(k-1)}}$$

$$\hat{N}_{(k)} = n + \hat{f}_{0(k)}$$

Step 2 : M-step

Using the complete frequency table $\hat{f}_{0(k)}, f_1, f_2, ..., f_m$ computes the new maximum likelihood estimator $\hat{\omega}_{(k)}$ and $\hat{\theta}_{(k)}$. The unobserved, complete likelihood function is

$$L(x;\omega,\theta) = \prod_{X_i=1} (1-\omega)^{z_i}[\omega(1-\theta)^{x_i}\theta]^{1-z_i} \prod_{X_i\neq 1} [\omega(1-\theta)^{x_i}\theta]$$

and the log-likelihood function is

$$\begin{aligned} l(x;\omega,\theta) = &\sum_{x_i=1} z_i[log(1-\omega) - log\omega] + Nlog\omega + \sum_{i=1}^{N} x_i log(1-\theta) + Nlog\theta \\ &- \sum_{x_i=1} z_i[x_i log(1-\theta) + log\theta] \end{aligned}$$

Finding updated ω and θ by maximizing this log-likelihood function

$$\hat{\omega}_{(k)} = 1 - \frac{\sum_{1(k)}}{\hat{N}_{(k)}} \quad ; \quad \sum_{1(k)} = \sum_{x_i=1} z_{i(k)}$$

$$\hat{\theta}_{(k)} = \frac{\hat{N}_{(k)} - \sum_{1(k)}}{\hat{N}_{(k)} + \sum_{i=1}^{N} x_i - 2\sum_{1(k)}} \quad ; \quad \hat{N}_{(k)} = n + \hat{f}_{0(k)}$$

Note that now we cannot calculate $\hat{\omega}_{(k)}$ and $\hat{\theta}_{(k)}$ because we do not know $\sum_{1(k)}$, so we have to do another EM step for z_i.

Step 2.1 : Nested E-step

Set $l = l + 1$ and $\hat{z}_{i(k)}$ are computed by using their expected values.

$$E(z_i \mid x_i; \omega, \theta) = e_{i(l)} = \frac{1 - \hat{\omega}_{(l-1)}}{[(1 - \hat{\omega}_{(l-1)}) + \hat{\omega}_{(l-1)}(1 - \hat{\theta}_{(l-1)})\hat{\theta}_{(l-1)}]}$$

Step 2.2 : Nested M-step

Updated $\hat{\omega}_{(k)}$ and $\hat{\theta}_{(k)}$ are obtained by

$$\hat{\omega}_{(l)} = 1 - \frac{f_1 e_{i(l)}}{\hat{N}_{(k)}}$$

$$\hat{\theta}_{(l)} = \frac{\hat{N}_{(k)} - f_1 e_{i(l)}}{\hat{N}_{(k)} + \sum_{i=1}^{N} x_i - 2 f_1 e_{i(l)}}$$

lead to $\hat{p}_{0(l)} = \hat{\omega}_{(l)} \hat{\theta}_{(l)}$.

Checking
$$\begin{aligned} &1) \mid \hat{\omega}_{(l)} - \hat{\omega}_{(l-1)} \mid \; < \; \varepsilon \\ &\quad \mid \hat{\theta}_{(l)} - \hat{\theta}_{(l-1)} \mid \; < \; \varepsilon \\ &2) \mid \hat{f}_{0(k)} - \hat{f}_{0(k-1)} \mid \; < \; \varepsilon \end{aligned}$$

Then, going back to step 1. These steps alternate continuously until $\hat{\omega}, \hat{\theta}$ and $\hat{f}_0$ converge to a MLE with an acceptable error. Here ε is set equal to 10^{-6}.

4.4 Likelihood-ratio test (LRT)

The likelihood-ratio test is a test statistic used to compare the goodness of fit of two models, one of which; the null model, is a special case of the other; the alternative model. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. The likelihood-ratio test can be presented as a difference in the log-likelihood as follows:

$$\begin{aligned} LRT &= -2 \ln \left(\frac{L_0}{L_A} \right) \\ &= -2 \ln(L_0) + 2 \ln(L_A) \\ &= -2 l_0 + 2 l_A \end{aligned}$$

where l_0 and l_A denote the log-likelihood function under null and alternative hypothesis respectively. Generally, the probability distribution of the test statistic is approximately a chi-square distribution with degree of freedom equal to $df_A - df_0$, where df_0 and df_A

represent the number of free parameters under the null model and the alternative model respectively. Now, the hypothesis that we consider here is

$$H_0 : \text{data are from ZT geometric distribution} \quad (\omega = 0)$$
$$H_A : \text{data are from ZTOI geometric distribution} \quad (\omega > 0)$$

Hence, LRT is determined as:

$$LRT = -2l_0(0, \tilde{\theta}) + 2l_A(\hat{\omega}, \hat{\theta}) \tag{4.11}$$

where $\tilde{\theta}$ is the MLE under a zero-truncated (ZT) geometric model, whereas $\hat{\omega}$ and $\hat{\theta}$ are the MLEs under a zero-truncated one-inflated (ZTOI) geometric model. It can be seen from the null hypothesis that the true parameter value $\omega = 0$ is on the boundary of parameter space ($0 \leq \omega \leq 1$). Therefore, the asymptotic distribution of the test statistic LRT in (4.11) is the mixture of the one point distribution with all its mass equal to zero (χ^2_0) and the chi-square distribution with one degree of freedom (χ^2_1) with equal weights, $LRT \backsim \frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0$, while the upper percentiles of the null distribution of LRT are approximately equal to the $(1 - 2\alpha)100$ percentiles of χ^2_1 (see more details in Böhning et al. (1994) and Self and Liang (1987)). The log-likelihood of null and alternative models are shown in next subsection.

4.4.1 A zero-truncated geometric model

The probability density function of a zero-truncated geometric distribution is

$$p_x = (1-p)^{x-1}p\,; \quad x = 1, 2, 3, ...$$

and the likelihood function is

$$L = \prod_{x_i=1}^{m} (p_x)^{f_x}\,.$$

Hence, the observed, incomplete data log-likelihood based on null model is determined as:

$$\begin{aligned} l_0(\theta) &= \sum_{x=1}^{m} f_x \log[(1-p)^{x-1}p] \\ &= \sum_{x=1}^{m} f_x(x-1)\log(1-\theta) + \sum_{x=1}^{m} f_x \log\theta \end{aligned} \tag{4.12}$$

4.4.2 A zero-truncated one-inflated geometric model

The probability density function of a zero-truncated one-inflated geometric distribution is

$$p_x^{1+} = \begin{cases} \omega(1-\theta)^x\theta/[1-\omega\theta] & \text{if } x \neq 1 \\ [(1-\omega)+\omega(1-\theta)^x\theta]\,/\,[1-\omega\theta] & \text{if } x = 1 \end{cases}$$

and the likelihood function is

$$L(\omega,\theta) = \prod_{x_i=1}^{m} \left(p_x^{1+}\right)^{f_x}.$$

The observed, incomplete data log-likelihood based on alternative model is shown as:

$$\begin{aligned} l_A(\omega,\theta) &= \sum_{x=1}^{m} f_x \log p_x^{1+} \\ &= f_1 \log\left\{\frac{(1-\omega)+\omega(1-\theta)\theta)}{1-\omega\theta}\right\} + \sum_{x=2}^{m} f_x \log\left\{\frac{\omega(1-\theta)^x\theta}{1-\omega\theta}\right\} \\ &= f_1 \log\left\{\frac{(1-\omega)+\omega(1-\theta)\theta)}{1-\omega\theta}\right\} + (n-f_1)\left\{\log\omega + \log\theta - \log(1-\omega\theta)\right\} \\ &\quad + (S-f_1)\log(1-\theta) \end{aligned} \tag{4.13}$$

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

4.5 The performance of the newly proposed estimator

The following examples are given to demonstrate finding the newly proposed maximum likelihood estimator based on a zero-truncated one-inflated geometric model (MLE_ZTOI) under 20% and 50% one-inflation. Then, this newly proposed estimator is compared with the conventional estimators and the formerly proposed estimators from Chapter 3.

Example 4.1 The data in Table 4.2 are generated from the zero-truncated geometric distribution with parameter $p = 0.2$ and $N = 100$ under 20% of one-inflation.

Table 4.2: The data with 20% one-inflation

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{11}	f_{12+}	n
(13)	32	14	7	7	9	5	3	3	1	3	3	87

We consider only zero-truncated counts and $n = 87$. The computation of MLE_ZTOI starts with setting the initial values of ω and θ. The complete frequency table $f_0, f_1, ..., f_m$ is used to calculate the maximum likelihood estimators of $\hat{\omega}, \hat{\theta}$ and $\hat{f}_0$ as following:

Step 0 : Set $\hat{\omega}_{(0)} = 1/2$, $\hat{\theta}_{(0)} = 1/(1+\bar{x}) = 0.2112$,where $\bar{x} = 3.7356$. Hence, $\hat{p}_{0(0)} = \hat{\omega}_{(0)}\hat{\theta}_{(0)} = 0.5(0.2112) = 0.1056$

First outer iteration $(k = 1)$

Step 1 : E-step

Computing $\hat{f}_{0(1)}$ and $\hat{N}_{(1)}$ by using $\hat{\omega}_{(0)}$ and $\hat{\theta}_{(0)}$, hence

$$\hat{f}_{0(1)} = \frac{n\hat{p}_{0(0)}}{1-\hat{p}_{0(0)}} = \frac{87(0.1056)}{1-0.1056} = 10.27$$

$$\hat{N}_{(1)} = n + \hat{f}_{0(1)} = 87 + 10.27 = 97.27$$

Step 2 : M-step

Using the new complete frequency table $\hat{f}_{0(1)}, f_1, f_2, ..., f_m$ computes the new maximum likelihood estimator $\hat{\omega}_{(1)} = 1 - \frac{\sum_{1(1)}}{\hat{N}_{(1)}}$ and $\hat{\theta}_{(1)} = \frac{\hat{N}_{(1)} - \sum_{1(1)}}{\hat{N}_{(1)} + \sum_{i=1}^{N} x_i - 2\sum_{1(1)}}$. Now we do not know the value of $\sum_{1(1)}$; $\sum_{1(1)} = \sum_{x_i=1} z_{i(1)}$ so we need to do another EM-step.

First inner iteration $(l = 1)$

Step 2.1 : Nested E-step

Computing $z_{i(1)}$ by using their expected values

$$e_{1(1)} = \frac{1-\hat{\omega}_{(0)}}{[(1-\hat{\omega}_{(0)}) + \hat{\omega}_{(0)}(1-\hat{\theta}_{(0)})\hat{\theta}_{(0)}]} = \frac{1-0.5}{[(1-0.5) + 0.5(1-0.2112)(0.2112)]} = 0.8572$$

Step 2.2 : Nested M-step

New update $\hat{\omega}_{(1)}$ and $\hat{\theta}_{(1)}$ are obtained by

$$\hat{\omega}_{1(1)} = 1 - \frac{f_1 e_{1(1)}}{\hat{N}_{(1)}} = 1 - \frac{32(0.8572)}{97.27} = 0.718$$

$$\hat{\theta}_{1(1)} = \frac{\hat{N}_{(1)} - f_1 e_{1(1)}}{\hat{N}_{(1)} + \sum_{i=1}^{N} x_i - 2 f_1 e_{i1(1)}} = \frac{97.27 - 32(0.8572)}{97.27 + 325 - 2(32)(0.8572)} = 0.1901$$

Checking $\quad | \hat{\omega}_{1(1)} - \hat{\omega}_{1(0)} | = | 0.718 - 0.5 | = 0.218 \quad > \quad 10^{-6}$

$$| \hat{\theta}_{1(1)} - \hat{\theta}_{1(0)} | = | 0.1901 - 0.2112 | = 0.0211 \quad > \quad 10^{-6}$$

Due to both estimates cannot meet the criteria so we need to go back to step 2.1 for the second inner iteration.

Second inner iteration $(l = 2)$

Step 2.1 : Nested E-step

Computing $z_{i(2)}$ by using their expected values

$$\begin{aligned} e_{1(2)} &= \frac{1-\hat{\omega}_{1(1)}}{[(1-\hat{\omega}_{1(1)})+\hat{\omega}_{1(1)}(1-\hat{\theta}_{1(1)})\hat{\theta}_{1(1)}]} \\ &= \frac{1-0.718}{[(1-0.718)+0.718(1-0.1901)(0.1901)]} \\ &= 0.7184 \end{aligned}$$

Step 2.2 : Nested M-step

New maximum likelihood estimates $\hat{\omega}_{(2)}$ and $\hat{\theta}_{(2)}$ are obtained by

$$\hat{\omega}_{1(2)} = 1 - \frac{f_1 e_{1(2)}}{\hat{N}_{(1)}} = 1 - \frac{32(0.7184)}{97.27} = 0.7637$$

$$\hat{\theta}_{1(2)} = \frac{\hat{N}_{(1)} - f_1 e_{1(2)}}{\hat{N}_{(1)} + \sum_{i=1}^{N} x_i - 2 f_1 e_{1(2)}} = \frac{97.27 - 32(0.7184)}{97.27 + 325 - 2(32)(0.7184)} = 0.1974$$

Checking $\quad | \hat{\omega}_{1(2)} - \hat{\omega}_{1(1)} | = | 0.7637 - 0.718 | = 0.0458 \quad > \quad 10^{-6}$

$$| \hat{\theta}_{1(2)} - \hat{\theta}_{1(1)} | = | 0.1974 - 0.1901 | = 0.0073 \quad > \quad 10^{-6}$$

Go to step 2.1 for the third inner iteration ($l = 3$). Continue these inner steps until $\hat{\omega}$ and $\hat{\theta}$ converge to a MLE. This example takes 20 inner iterations for the first outer iteration. It provides $\hat{\omega}_{(1)} = 0.80058$ and $\hat{\theta}_{(1)} = 0.20307$ so $\hat{p}_{0(1)} = \hat{\omega}_{(1)}\hat{\theta}_{(1)} = 0.16257$. Then we move on the second outer iteration.

Second outer iteration ($k = 2$)

Step 1 : E-step

Computing $\hat{f}_{0(2)}$ and $\hat{N}_{(2)}$ by using $\hat{\omega}_{(1)}$ and $\hat{\theta}_{(1)}$, hence

$$\hat{f}_{0(2)} = \frac{n\hat{p}_{0(1)}}{1-\hat{p}_{0(1)}} = \frac{87(0.16257)}{1-0.16257} = 16.8897$$

$$\hat{N}_{(2)} = n + \hat{f}_{0(2)} = 87 + 16.8897 = 103.8897$$

Checking $\quad | \hat{f}_{0(2)} - \hat{f}_{0(1)} | = | 16.8897 - 10.27001 | = 6.61969 \quad > \quad 10^{-6}$ go to step 2.

Step 2 : M-step

Using the new complete frequency table $\hat{f}_{0(2)}, f_1, f_2, ..., f_m$ computes the new maximum likelihood estimator $\hat{\omega}_{(2)} = 1 - \frac{\sum_{1(2)}}{\hat{N}_{(2)}}$ and $\hat{\theta}_{(2)} = \frac{\hat{N}_{(2)} - \sum_{1(2)}}{\hat{N}_{(2)} + \sum_{i=1}^{N} x_i - 2\sum_{1(2)}}$. Using another EM-step calculates $\sum_{1(2)}$; $\quad \sum_{1(2)} = \sum_{x_i=1} z_{i(2)}$.

First inner iteration $l = 1$ **for** $k = 2$

Step 2.1 : Nested E-step

Let $\hat{\omega}_{2(0)} = \hat{\omega}_{(1)}$ and $\hat{\theta}_{2(0)} = \hat{\theta}_{(1)}$. Computing $z_{i(2)}$ by using their expected values

$$\begin{aligned} e_{2(1)} &= \frac{1-\hat{\omega}_{2(0)}}{[(1-\hat{\omega}_{2(0)}) + \hat{\omega}_{2(0)}(1-\hat{\theta}_{2(0)})\hat{\theta}_{2(0)}]} \\ &= \frac{1-0.80058}{[(1-0.80058) + 0.80058(1-0.20307)(0.20307)]} \\ &= 0.60618 \end{aligned}$$

Step 2.2 : Nested M-step

New maximum likelihood estimates $\hat{\omega}_{2(1)}$ and $\hat{\theta}_{2(1)}$ are obtained by

$$\hat{\omega}_{2(1)} = 1 - \frac{f_1 e_{2(1)}}{\hat{N}_{(2)}} = 1 - \frac{32(0.60618)}{103.8897} = 0.81329$$

$$\hat{\theta}_{2(1)} = \frac{\hat{N}_2 - f_1 e_{2(1)}}{\hat{N}_{(2)} + \sum_{i=1}^{N} x_i - 2 f_1 e_{2(1)}} = \frac{103.8897 - 32(0.60618)}{103.8897 + 325 - 2(32)(0.60618)} = 0.21659$$

Checking $\quad | \hat{\omega}_{2(1)} - \hat{\omega}_{2(0)} | = | 0.81329 - 0.80058 | = 0.01271 \quad > \quad 10^{-6}$

$$| \hat{\theta}_{2(1)} - \hat{\theta}_{2(0)} | = | 0.21659 - 0.20307 | = 0.01352 \quad > \quad 10^{-6}$$

Go to step 2.1 for $l = 2$, These inner steps are repeated until $\hat{\omega}$ and $\hat{\theta}$ converge to a constant. Finally, we get $\hat{\omega}_{(2)} = 0.83525$ and $\hat{\theta}_{(2)} = 0.21987$ so $\hat{p}_{0(2)} = \hat{\omega}_{(2)}\hat{\theta}_{(2)} = 0.18365$. Then going on step 1 for iteration 3, we calculate $\hat{f}_{0(3)}$ by using $\hat{\omega}_{(2)}, \hat{\theta}_{(2)}$ and checking its convergence. Both outer and inner steps are repeated until $\hat{\omega}, \hat{\theta}$ and $\hat{f}_0$ converge to a constant or the difference between present and previous values are less than 10^{-6}. The all iterations of EM-algorithm for maximum likelihood estimation of this example are shown in Table 4.3.

Finally, the estimates of unobserved frequency ($\hat{f}_0$) and population size ($\hat{N}$) from maximum likelihood estimation based on a zero-truncated one-inflated geometric model (MLE_ZTOI) are equal to 21.50 and 108.50 respectively. To compare this newly suggested estimator with existing estimators, Table 4.4 provides the population size estimated by the conventional estimators and the proposed estimators based on OT and ZTOI model, respectively. It can be seen clearly that the newly suggested estimator ($\hat{N}_{MLE\_ZTOI}$) yields value closest to the parameter of interest $N = 100$. Although the Turing_OT and MLE_OT produce the overestimation more than MLE_ZTOI, their estimates are not much different. Therefore, we can say that the newly proposed estimator can show the best performance and all proposed estimators can effectively cope with one-inflation problem.

Table 4.3: The maximum likelihood estimation for Example 4.1

k	$\hat{f}_{0(k)}$	$\hat{N}_{(k)}$	l	$e_{k(l)}$	$\hat{\omega}_{k(l)}$	$\hat{\theta}_{k(l)}$	$\hat{p}_{0(k)}$
0	-	-	-	-	0.5	0.21116	0.10558
1	10.27001	97.27001	1	0.857211	0.717994	0.1900861	
			2	0.718407	0.763658	0.1974025	
			3	0.661408	0.782409	0.2003079	
			⋮	⋮	⋮	⋮	
			19	0.6061827	0.8005773	0.2030701	
			20	0.6061808	0.8005780	0.2030702	0.16257
2	16.88972	103.88972	21	0.6061797	0.8132852	0.2165937	
			22	0.5750135	0.8228849	0.2180355	
			⋮	⋮	⋮	⋮	
			40	0.5348744	0.8352486	0.2198709	0.1836468
3	19.57152	106.57152	41	0.5348731	0.8393948	0.2251363	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	21.49810	108.49810	175	0.4834359	0.8574173	0.2310924	0.1981427

Table 4.4: Estimates for the data in Example 4.1 with true $N = 100$

Estimator	Estimated population size
Chao	160.14
Turing	126.78
MLE	118.8
Turing_OT	114.85
MLE_OT	113.15
MLE_ZTOI	**108.50**

Example 4.2 We also apply the newly proposed estimator with 50% one-inflation situation. Table 4.5 shows the data that are generated from the zero-truncated geometric distribution with parameter $p = 0.2$ and $N = 100$ under 50% one-inflation.

Table 4.5: The data with 50% one-inflation

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_{10}	f_{12}	f_{15}	f_{17}	n
(11)	67	5	3	2	4	1	1	1	2	1	1	1	89

The newly proposed estimator for this example can be calculated in same way with the Example 4.1. It can be seen that the procedure of EM algorithm is repeated 24 rounds for outer part and 104 rounds for inner part until reaching the constant of estimators and it produces $\hat{f}_0 = 6.88$ and $\hat{N}_{MLE\_ZTOI} = 95.88$. The details and outcomes are shown in Table 4.6.

As can be seen in Table 4.7, there is only one estimator, MLE_ZTOI, which provides an underestimation ($\hat{N}_{MLE\_ZTOI} = 95.88$) and closest to the parameter $N = 100$

Table 4.6: The maximum likelihood estimation for Example 4.2

k	$\hat{f}_{0(k)}$	$\hat{N}_{(k)}$	l	$e_{k(l)}$	$\hat{\omega}_{k(l)}$	$\hat{\theta}_{k(l)}$	$\hat{p}_{0(k)}$
0	-	-	-	-	0.5	0.3090278	0.1545139
1	16.26489	105.2649	1	0.8240425	0.4755056	0.2582192	
			2	0.8520416	0.4576844	0.2534471	
			⋮	⋮	⋮	⋮	
			13	0.8679196	0.4475783	0.2506563	0.1121883
2	11.246484	100.24648	14	0.8679199	0.4199235	0.2301009	
			15	0.8863331	0.4076169	0.2264110	
			⋮	⋮	⋮	⋮	
			34	0.9080619	0.3780488	0.2111513	0.09017878
3	8.821416	97.82142	35	0.9080621	0.3709712	0.2061625	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	6.875001	95.87500	104	0.9179104	0.3585398	0.2000000	0.07170796

whereas Turing_OT and MLE_OT produce the overestimation, $\hat{N}_{Turing\_OT} = 113.12$ and $\hat{N}_{MLE\_OT} = 111.25$, but they do not give the severe overestimation as much as the conventional estimators. It can be shown that the all proposed estimators can improve the overestimation associated with conventional estimators.

Table 4.7: Estimates for the data in Example 4.2

Estimator	Estimated population size
Chao	986.8
Turing	212.03
MLE	161.01
Turing_OT	113.12
MLE_OT	111.25
MLE_ZTOI	**95.88**

It is clear that the newly suggested estimator (MLE_ZTOI) can perform effectively under one-inflation and better than the previously proposed estimators from Chapter 3 (T_OT and MLE_OT). However, these are only examples from two data sets, it is necessary to do a further investigation in terms of bias, variance and mean square error by simulation studies and it is shown in next section.

4.6 Simulation study

The simulation study was undertaken to investigate the performance of three proposed estimators: the Turing estimator ($\hat{N}_{\text{T\_OT}}$), the maximum likelihood estimator ($\hat{N}_{\text{MLE\_OT}}$) based on the one-truncated geometric model, and the maximum likelihood

estimator based on zero-truncated one-inflated geometric model ($\hat{N}_{\text{MLE\_ZTOI}}$). In addition, three conventional estimators namely Chao's lower bound, the conventional Turing and maximum likelihood estimator are included to create a comprehensive comparison of all estimators affected by the one-inflation problem. The heterogeneous populations were generated from a geometric distribution (arising from the mixture of a Poisson distribution with an exponential distribution) with parameter $\theta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ and population sizes $N = 20, 50, 100, 500, 1000$ for two levels of one-inflation (20% and 50%). Each case is repeated 1,000 times. To evaluate the performance of estimation, the following criteria are used:

1) Relative bias ($RBias(\hat{N}) = \frac{E(\hat{N})-N}{N}$)

2) Relative variance ($RVar(\hat{N}) = \frac{E(\hat{N}-E(\hat{N}))^2}{N^2}$)

2) Relative mean square error ($RMSE(\hat{N}) = \frac{E(\hat{N}-N)^2}{N^2}$)

The results of simulation study are presented in Table 4.8 - 4.11 and Figure 4.1 - 4.2. Due to the fact that the results of two one-inflation levels are similar, both parts are summarized together. To explore preliminary the behaviour of estimators, we consider the mean of estimates of population size. According to the results provided in Table 4.8, all of the conventional estimators (Chao, Turing and MLE) show clearly an overestimation of population size for all conditions of the study, particularly, it is severe in Chao's lower bound estimator. Conventional Turing and MLE estimators are less affected by one-inflation than Chao's lower bound. All proposed estimators yield satisfying outcomes which are close to the true value of population size N with a slight tendency of overestimating except $\hat{N}_{\text{T\_OT}}$ which gives slight underestimates for the small population sizes ($N = 20, 50, 100$) in the case of 20% one-inflation. In addition, $\hat{N}_{\text{MLE\_ZTOI}}$ yields the best estimation results for almost all studied conditions. Correspondingly, $\hat{N}_{\text{MLE\_ZTOI}}$ produces the smallest RBias in all studied cases as Table 4.9 and Figure 4.1 - 4.2 left panel show. We can rank the performance of proposed estimators in terms of accuracy as $\hat{N}_{\text{MLE\_ZTOI}}, \hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$. This could indicate that the $\hat{N}_{\text{MLE\_ZTOI}}$ can cope with the one-inflation situation better than $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ in both, low and high level, one-inflation scenarios. According to RVar (see Table 4.10), the $\hat{N}_{\text{T\_OT}}$ tends to provide the minimum RVar in the case of small population size ($N = 20, 50, 100$) whereas $\hat{N}_{\text{MLE\_ZTOI}}$ yields the minimum RVar for the large size of population ($N = 500, 1000$). However, all proposed estimators give relatively small RVar in all conditions if compared with the conventional estimators as shown in Figure 4.1 - 4.2 middle panel. Similar to the results of RVar, the $\hat{N}_{\text{T\_OT}}$ seems to provide the smallest RMSE for the small population size whereas $\hat{N}_{\text{MLE\_ZTOI}}$ gives the smallest RMSE for the large size of population as Table 4.11 and Figure 4.1 - 4.2 right panel show. However, overall the efficiency of $\hat{N}_{\text{T\_OT}}$ seems to be reduced if the level of one-inflation is increasing which is opposite to $\hat{N}_{\text{MLE\_ZTOI}}$. Furthermore, it can be noticed that with increase of the population size, there is a decline in the RBias, RVar and RMSE for all proposed estimators. On the

other hand, with increasing geometric parameter θ there is an increase in the RBias, RVar and RMSE for all proposed estimators.

Note "-" in Table 4.8 - 4.11 is defined as no results from simulation study.

Table 4.8: Monte Carlo means of the population size estimates ($Mean(\hat{N})$) under 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT	MLE_ZTOI
20%	20	0.1	42.89	22.96	21.26	19.61	20.63	**20.18**
		0.2	46.31	25.27	23.24	18.90	21.74	**20.66**
		0.3	47.28	27.78	25.73	17.89	-	**21.63**
	50	0.1	127.70	57.04	52.73	49.23	51.25	**50.14**
		0.2	112.46	61.63	56.61	48.98	53.22	**50.75**
		0.3	113.43	67.04	61.52	47.45	55.54	**51.33**
		0.4	122.31	75.12	69.20	45.54	59.93	**53.31**
	100	0.1	231.04	113.93	105.26	99.30	102.37	**100.15**
		0.2	205.59	122.04	111.86	98.73	105.22	**100.29**
		0.3	210.15	132.64	121.53	97.50	109.82	**101.52**
		0.4	216.72	146.61	135.11	95.24	117.00	**104.62**
		0.5	242.10	168.66	156.12	**91.22**	128.72	109.93
		0.6	300.93	206.86	192.18	**83.27**	-	118.53
	500	0.1	1053.44	569.26	525.72	506.84	511.55	**500.49**
		0.2	963.08	609.47	558.00	515.16	525.04	**500.07**
		0.3	971.41	659.11	602.42	518.37	543.61	**500.71**
		0.4	1020.69	724.57	663.84	532.26	569.39	**503.55**
		0.5	1119.12	818.73	753.22	539.65	605.51	**508.49**
		0.6	1303.91	981.44	907.65	545.36	666.00	**521.65**
	1000	0.1	2092.53	1138.30	1050.46	1017.62	1021.92	**999.64**
		0.2	1907.28	1218.46	1115.75	1038.31	1050.11	**1000.22**
		0.3	1933.49	1316.74	1203.44	1057.25	1086.36	**1000.56**
		0.4	2027.43	1445.50	1323.87	1094.42	1135.09	**1002.55**
		0.5	2224.43	1642.99	1510.24	1132.57	1208.15	**1008.72**
		0.6	2565.39	1953.83	1807.09	1180.50	1319.75	**1024.08**
50%	20	0.1	120.34	29.79	24.56	**20.17**	21.50	20.21
		0.2	123.95	38.38	31.39	**20.07**	-	21.13
	50	0.1	494.75	72.15	59.53	50.68	53.09	**50.19**
		0.2	403.30	89.65	72.78	51.48	57.27	**50.65**
		0.3	399.49	113.63	92.37	**50.56**	62.89	51.62
		0.4	420.22	152.91	127.72	**50.70**	69.75	57.09
	100	0.1	1038.28	143.13	118.28	102.20	105.94	**100.21**
		0.2	742.85	176.13	142.85	104.76	113.43	**100.62**
		0.3	684.94	219.17	178.15	106.47	123.40	**101.50**
		0.4	742.94	282.58	233.44	109.48	138.51	**103.77**
		0.5	898.43	387.43	327.56	**108.30**	-	113.77
	500	0.1	4371.18	710.60	587.27	521.62	527.90	**499.98**
		0.2	3182.76	872.46	706.55	547.70	563.69	**500.56**
		0.3	2999.10	1074.84	872.24	578.09	609.01	**501.27**
		0.4	3141.97	1361.50	1120.47	616.55	672.10	**504.07**
		0.5	3610.54	1812.17	1521.01	648.45	761.10	**507.99**
		0.6	4522.17	2613.77	2257.55	681.05	919.27	**529.34**
	1000	0.1	8471.62	1419.90	1173.77	1047.30	1055.66	**999.92**
		0.2	6278.96	1744.01	1412.57	1107.76	1127.99	**1001.85**
		0.3	5872.90	2138.95	1736.09	1167.56	1215.55	**1001.32**
		0.4	6253.75	2719.71	2235.84	1246.69	1339.53	**1004.02**
		0.5	7120.06	3618.72	3034.65	1383.39	1514.05	**1008.41**
		0.6	8869.88	5175.92	4463.78	1438.39	1803.81	**1031.95**

Table 4.9: Relative bias of six population size estimators under 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT	MLE_ZTOI
20%	20	0.1	1.1447	0.1481	0.0629	-0.0197	0.0315	**0.0090**
		0.2	1.3154	0.2636	0.1619	-0.0550	0.0870	**0.0329**
		0.3	1.3641	0.3889	0.2867	-0.1053	-	**0.0814**
	50	0.1	1.5541	0.1407	0.0546	-0.0154	0.0250	**0.0027**
		0.2	1.2492	0.2326	0.1322	-0.0204	0.0643	**0.0150**
		0.3	1.2687	0.3409	0.2304	-0.0510	0.1108	**0.0267**
		0.4	1.4463	0.5024	0.3841	-0.0893	0.1985	**0.0662**
	100	0.1	1.3104	0.1393	0.0526	-0.0070	0.0237	**0.0015**
		0.2	1.0559	0.2204	0.1186	-0.0127	0.0522	**0.0029**
		0.3	1.1015	0.3264	0.2153	-0.0250	0.0982	**0.0152**
		0.4	1.1672	0.4661	0.3511	-0.0476	0.1700	**0.0462**
		0.5	1.4210	0.6866	0.5612	**-0.0878**	0.2872	0.0993
		0.6	2.0093	1.0686	0.9218	**-0.1673**	-	0.1853
	500	0.1	1.1069	0.1385	0.0514	0.0137	0.0231	**0.0010**
		0.2	0.9262	0.2189	0.1160	0.0303	0.0501	**0.0001**
		0.3	0.9428	0.3182	0.2048	0.0367	0.0872	**0.0014**
		0.4	1.0414	0.4491	0.3277	0.0645	0.1388	**0.0071**
		0.5	1.2382	0.6375	0.5064	0.0793	0.2110	**0.0170**
		0.6	1.6078	0.9629	0.8153	0.0907	0.3320	**0.0433**
	1000	0.1	1.0925	0.1383	0.0505	0.0176	0.0219	**-0.0004**
		0.2	0.9073	0.2185	0.1157	0.0383	0.0501	**0.0002**
		0.3	0.9335	0.3167	0.2034	0.0572	0.0864	**0.0006**
		0.4	1.0274	0.4455	0.3239	0.0944	0.1351	**0.0025**
		0.5	1.2244	0.6430	0.5102	0.1326	0.2081	**0.0087**
		0.6	1.5654	0.9538	0.8071	0.1805	0.3198	**0.0241**
50%	20	0.1	5.0169	0.4895	0.2281	**0.0083**	0.0748	0.0104
		0.2	5.1975	0.9188	0.5693	**0.0033**	-	0.0565
	50	0.1	8.8950	0.4430	0.1905	0.0137	0.0619	**0.0039**
		0.2	7.0660	0.7930	0.4555	0.0296	0.1454	**0.0130**
		0.3	6.9899	1.2726	0.8474	**0.0112**	0.2578	0.0325
		0.4	7.4045	2.0582	1.5544	**0.0140**	0.4043	0.1419
	100	0.1	9.3828	0.4313	0.1828	0.0220	0.0594	**0.0021**
		0.2	6.4285	0.7613	0.4285	0.0476	0.1343	**0.0062**
		0.3	5.8494	1.1917	0.7815	0.0647	0.2340	**0.0150**
		0.4	6.4294	1.8258	1.3344	0.0948	0.3851	**0.0377**
		0.5	7.9843	2.8743	2.2756	**0.0830**	-	0.1377
	500	0.1	7.7424	0.4212	0.1745	0.0432	0.0558	**0.0000**
		0.2	5.3655	0.7449	0.4131	0.0954	0.1274	**0.0011**
		0.3	4.9982	1.1497	0.7445	0.1562	0.2180	**0.0025**
		0.4	5.2839	1.7230	1.2409	0.2331	0.3442	**0.0081**
		0.5	6.2211	2.6243	2.0420	0.2969	0.5222	**0.0160**
		0.6	8.0443	4.2275	3.5151	0.3621	0.8385	**0.0587**
	1000	0.1	7.4716	0.4199	0.1738	0.0473	0.0557	**-0.0001**
		0.2	5.2790	0.7440	0.4126	0.1078	0.1280	**0.0019**
		0.3	4.8729	1.1390	0.7361	0.1676	0.2156	**0.0013**
		0.4	5.2537	1.7197	1.2358	0.2467	0.3395	**0.0040**
		0.5	6.1201	2.6187	2.0346	0.3834	0.5141	**0.0084**
		0.6	7.8699	4.1759	3.4638	0.4384	0.8038	**0.0320**

Table 4.10: Relative variance of six population size estimators under 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT	MLE_ZTOI
20%	20	0.1	1.2344	0.0126	0.0070	**0.0043**	0.0056	0.0052
		0.2	1.6830	0.0476	0.0334	**0.0108**	0.0263	0.0235
		0.3	2.1594	0.1152	0.0920	**0.0160**	-	0.0739
	50	0.1	1.9136	0.0050	0.0030	**0.0024**	0.0025	0.0024
		0.2	1.6089	0.0144	0.0092	**0.0057**	0.0069	0.0067
		0.3	1.5709	0.0401	0.0280	**0.0101**	0.0219	0.0203
		0.4	2.1363	0.0968	0.0709	**0.0154**	0.0570	0.0521
	100	0.1	0.8588	0.0023	0.0013	0.0012	0.0011	**0.0010**
		0.2	0.4538	0.0068	0.0045	**0.0032**	0.0036	0.0034
		0.3	0.9350	0.0171	0.0119	**0.0060**	0.0086	0.0086
		0.4	0.5790	0.0391	0.0292	**0.0097**	0.0225	0.0248
		0.5	0.8630	0.1092	0.0885	**0.0163**	0.0966	0.0865
		0.6	2.7078	0.2989	0.2624	**0.0197**	-	0.2931
	500	0.1	0.0967	0.0004	0.0003	0.0003	0.0002	**0.0002**
		0.2	0.0513	0.0013	0.0008	0.0007	0.0006	**0.0006**
		0.3	0.0552	0.0031	0.0021	0.0016	0.0015	**0.0015**
		0.4	0.0760	0.0078	0.0055	**0.0035**	0.0037	0.0040
		0.5	0.1078	0.0153	0.0124	**0.0078**	0.0103	0.0116
		0.6	0.2153	0.0490	0.0424	**0.0141**	0.0364	0.0420
	1000	0.1	0.0409	0.0002	0.0001	0.0001	0.0001	**0.0001**
		0.2	0.0252	0.0007	0.0004	0.0004	0.0003	**0.0003**
		0.3	0.0261	0.0014	0.0010	0.0009	0.0007	**0.0008**
		0.4	0.0352	0.0037	0.0027	0.0018	0.0018	**0.0018**
		0.5	0.0589	0.0092	0.0072	**0.0048**	0.0048	0.0050
		0.6	0.1025	0.0215	0.0183	**0.0102**	0.0161	0.0179
50%	20	0.1	7.4757	0.0797	0.0352	**0.0033**	0.0074	0.0040
		0.2	10.1349	0.4444	0.2996	**0.0073**	-	0.0835
	50	0.1	30.5648	0.0179	0.0067	0.0015	0.0018	**0.0013**
		0.2	27.9088	0.0749	0.0399	**0.0045**	0.0092	0.0047
		0.3	31.9849	0.2735	0.1637	**0.0077**	0.0366	0.0202
		0.4	35.7446	1.5514	1.0792	**0.0140**	0.2580	0.2362
	100	0.1	45.7641	0.0077	0.0028	0.0009	0.0009	**0.0006**
		0.2	19.6035	0.0332	0.0168	0.0029	0.0038	**0.0020**
		0.3	13.2992	0.1108	0.0627	**0.0056**	0.0119	0.0058
		0.4	20.1043	0.3522	0.2379	**0.0115**	0.0443	0.0182
		0.5	39.7080	1.3324	1.0135	**0.0183**	-	0.2355
	500	0.1	4.5381	0.0015	0.0005	0.0003	0.0002	**0.0001**
		0.2	1.4321	0.0058	0.0028	0.0010	0.0007	**0.0004**
		0.3	1.1298	0.0192	0.0111	0.0023	0.0022	**0.0010**
		0.4	1.3901	0.0519	0.0336	0.0051	0.0066	**0.0027**
		0.5	2.0034	0.1663	0.1244	0.0106	0.0235	**0.0078**
		0.6	4.3476	0.6947	0.5811	**0.0231**	0.1228	0.0424
	1000	0.1	1.8840	0.0007	0.0003	0.0001	0.0001	**0.0001**
		0.2	0.7029	0.0032	0.0015	0.0005	0.0003	**0.0002**
		0.3	0.4896	0.0091	0.0053	0.0011	0.0010	**0.0005**
		0.4	0.6253	0.0259	0.0175	0.0032	0.0034	**0.0013**
		0.5	0.9114	0.0773	0.0572	0.0080	0.0102	**0.0033**
		0.6	1.8886	0.3106	0.2568	0.0219	0.0486	**0.0133**

Table 4.11: Relative mean square error of six population size estimators under 20% and 50% one-inflation

Extra-ones	N	p	Chao	Turing	MLE	T_OT	MLE_OT	MLE_ZTOI
20%	20	0.1	2.5436	0.0345	0.0109	**0.0047**	0.0066	0.0053
		0.2	3.4116	0.1170	0.0595	**0.0138**	0.0338	0.0246
		0.3	4.0180	0.2663	0.1741	**0.0271**	-	0.0805
	50	0.1	4.3268	0.0248	0.0059	0.0026	0.0031	**0.0024**
		0.2	3.1679	0.0684	0.0267	**0.0061**	0.0111	0.0069
		0.3	3.1788	0.1563	0.0811	**0.0127**	0.0341	0.0210
		0.4	4.2258	0.3490	0.2184	**0.0234**	0.0963	0.0564
	100	0.1	2.5751	0.0217	0.0041	0.0013	0.0017	**0.0010**
		0.2	1.5682	0.0554	0.0186	**0.0033**	0.0063	0.0035
		0.3	2.1474	0.1236	0.0582	**0.0067**	0.0182	0.0089
		0.4	1.9408	0.2563	0.1525	**0.0120**	0.0514	0.0269
		0.5	2.8814	0.5805	0.4034	**0.0240**	0.1790	0.0963
		0.6	6.7422	1.4406	1.1118	**0.0477**	-	0.3272
	500	0.1	1.3217	0.0196	0.0029	0.0005	0.0007	**0.0002**
		0.2	0.9090	0.0492	0.0143	0.0017	0.0031	**0.0006**
		0.3	0.9440	0.1043	0.0440	0.0029	0.0091	**0.0015**
		0.4	1.1604	0.2095	0.1129	0.0077	0.0230	**0.0040**
		0.5	1.6409	0.4216	0.2689	0.0140	0.0548	**0.0119**
		0.6	2.8002	0.9761	0.7071	**0.0223**	0.1466	0.0439
	1000	0.1	1.2345	0.0193	0.0027	0.0005	0.0006	**0.0001**
		0.2	0.8483	0.0484	0.0138	0.0019	0.0028	**0.0003**
		0.3	0.8975	0.1017	0.0424	0.0041	0.0082	**0.0008**
		0.4	1.0908	0.2022	0.1076	0.0107	0.0200	**0.0018**
		0.5	1.5580	0.4227	0.2675	0.0224	0.0481	**0.0050**
		0.6	2.5529	0.9312	0.6697	0.0428	0.1184	**0.0184**
50%	20	0.1	32.6378	0.3192	0.0872	**0.0034**	0.0130	0.0041
		0.2	37.1391	1.2882	0.6234	**0.0073**	-	0.0866
	50	0.1	109.6550	0.2141	0.0430	0.0017	0.0057	**0.0013**
		0.2	77.8097	0.7036	0.2473	0.0053	0.0303	**0.0049**
		0.3	80.8113	1.8928	0.8817	**0.0079**	0.1030	0.0212
		0.4	90.5354	5.7861	3.4944	**0.0142**	0.2645	0.2561
	100	0.1	133.7561	0.1937	0.0362	0.0014	0.0044	**0.0006**
		0.2	60.9101	0.6128	0.2004	0.0051	0.0219	**0.0021**
		0.3	47.5016	1.5308	0.6734	0.0098	0.0667	**0.0060**
		0.4	61.4216	3.6855	2.0182	0.0205	0.1925	**0.0196**
		0.5	103.4174	9.5929	6.1908	**0.0251**	-	0.2542
	500	0.1	64.4777	0.1789	0.0310	0.0021	0.0033	**0.0001**
		0.2	30.2195	0.5607	0.1734	0.0101	0.0169	**0.0004**
		0.3	26.1107	1.3410	0.5654	0.0267	0.0497	**0.0010**
		0.4	29.3088	3.0206	1.5735	0.0594	0.1250	**0.0027**
		0.5	40.7032	7.0533	4.2942	0.0987	0.2961	**0.0081**
		0.6	69.0547	18.5662	12.9364	0.1542	0.8258	**0.0458**
	1000	0.1	57.7072	0.1770	0.0305	0.0024	0.0032	**0.0001**
		0.2	28.5697	0.5567	0.1717	0.0121	0.0167	**0.0002**
		0.3	24.2343	1.3063	0.5471	0.0292	0.0475	**0.0005**
		0.4	28.2265	2.9833	1.5448	0.0641	0.1187	**0.0013**
		0.5	38.3656	6.9350	4.1970	0.1550	0.2744	**0.0033**
		0.6	63.8217	17.7486	12.2543	0.2140	0.6947	**0.0144**

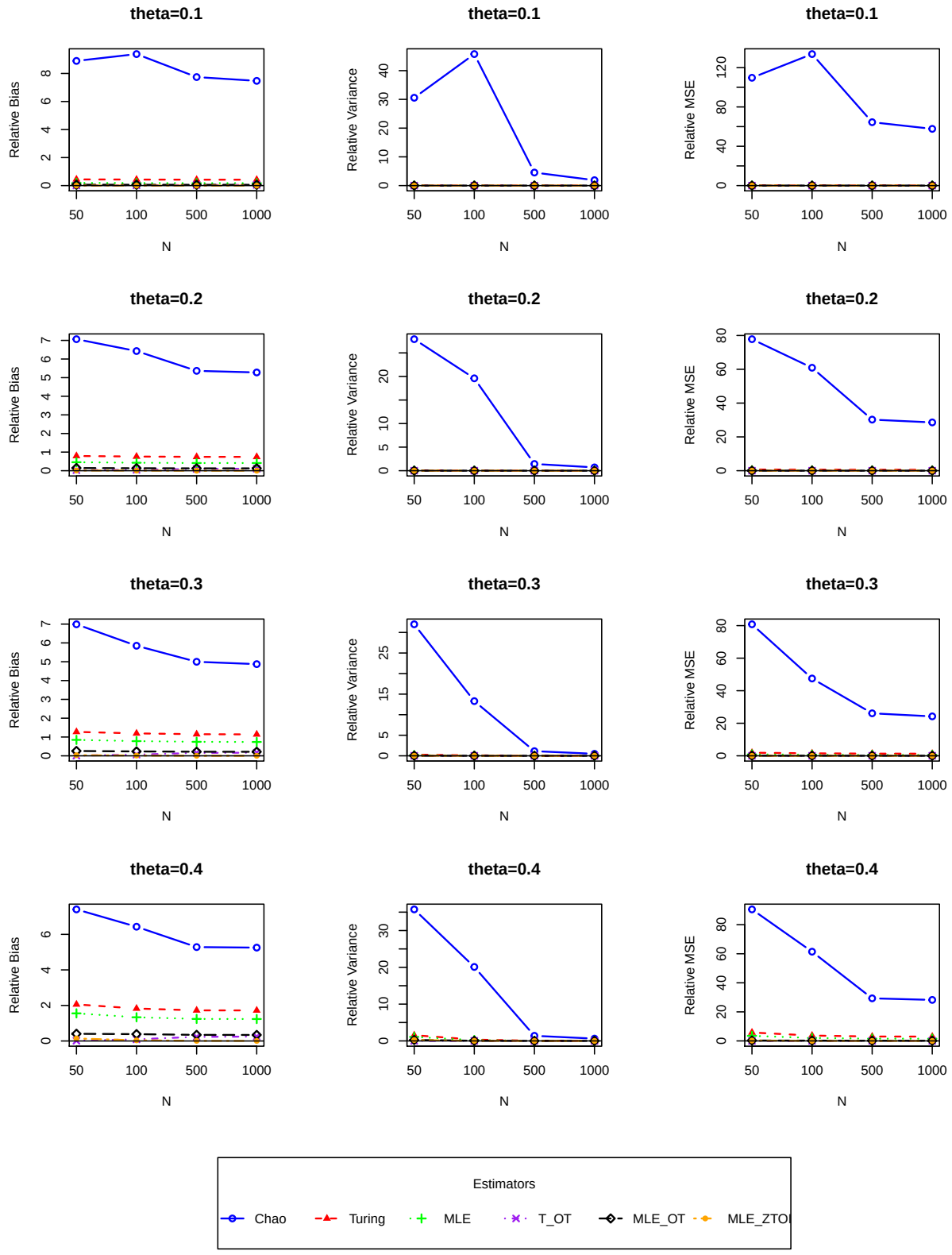

Figure 4.1: *RBias, RVar and RMSE of six estimators for counts drawn from geometric(θ) with 20% one-inflation*

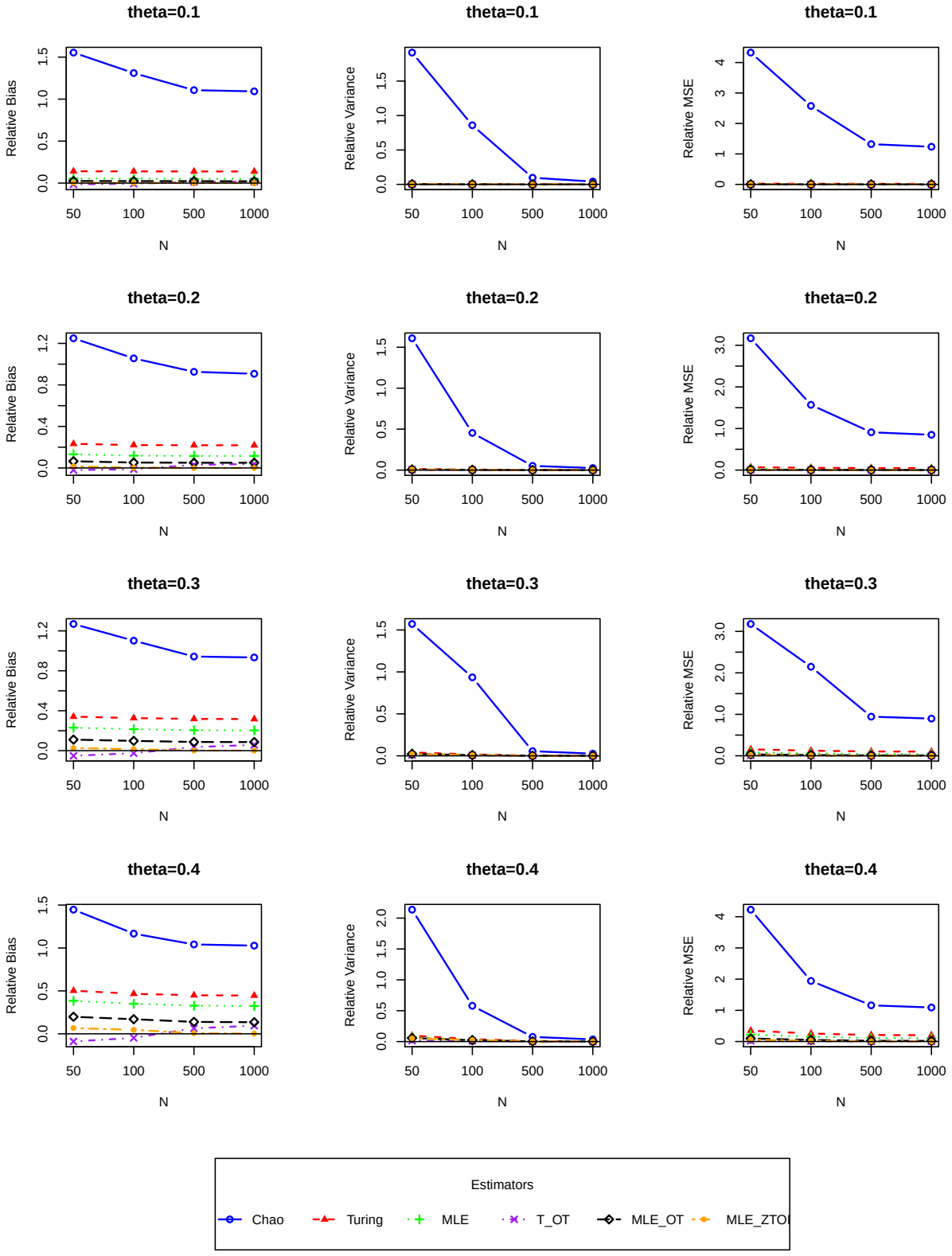

Figure 4.2: *RBias, RVar and RMSE of six estimators for counts drawn from geometric(θ) with 50% one-inflation*

4.7 Real-data examples

In this section, the two data sets of Section 3.7 and a new case study are used to examined the newly proposed estimator in actual data of one-inflation and comparing the estimate with other methods.

Example 4.3 Referring to the first case study in Chapter 3, we wish to estimate the total number of scrapie-infected holdings in France. The details of observed counts are presented again in Table 4.12.

Table 4.12: The data of French scrapie-infected holdings

x	1	2	3	4
f_x	121	13	5	2

The ratio plot in Figure 3.2 left panel shows that this data may experience one-inflation form. Now, we can check this suspicion again by using the likelihood ratio test as follows:

H_0 : data are from zero-truncated geometric distribution

H_A : data are from zero truncation one inflation geometric distribution

Set $\alpha = 0.05$ and use the test statistic

$$\begin{aligned} LRT &= -2l_0(0,\tilde{\theta}) + 2l_A(\hat{\omega},\hat{\theta}) \\ &= -2(77.6590) + 2(75.5607) \\ &= 4.1966. \end{aligned}$$

with a critical value of $\chi^2_{.90,1} = 2.706$. We come to the decision to reject H_0 since $LRT > 2.706$ and $\frac{p-value}{2}(0.02025) < \alpha(0.05)$. We conclude that this data set are from zero-truncated one-inflated geometric distribution at 0.05 significance level.

From the evidence provided by ratio plot and likelihood ratio test, the presence of one-inflation can be conjectured. Therefore, all proposed estimators should be appropriate for this data set, particularly the newly developed estimator. The results of estimating the total number of scrapie-infected holdings and the goodness of fit statistics from all estimators are shown in Table 4.13. As we expect, the newly suggested estimator $\hat{N}_{\text{MLE\_ZTOI}}$ can definitely reduce the overestimation assorted with conventional estimators by producing a distinctly smaller estimate. It clearly reveals that the MLE_ZTOI provides the smallest estimate while the estimate of T_OT is in between the estimate of MLE_ZTOI and MLE_OT but slightly more close to the proposed MLE_OT estimator. Moreover, the goodness of fit statistics and the graph in Figure 4.3 show that the estimated values from MLE_ZTOI, shown by the orange line on the graph, can fit the data very well and as good as the T_OT estimator.

Table 4.13: Results for scrapie-infected holdings in France

Estimator	$\hat{f}_0$	$\hat{N}$	Chi-square	p-value
Chao[1]	1126	1267	27.195	0.00000
Turing	761	902	8.487	0.01436
MLE	686	827	6.781	0.03369
T_OT	**286**	**427**	**0.283**	**0.59474**
MLE_OT	313	454	0.507	0.47644
MLE_ZTOI	**120**	**261**	**0.316**	**0.57402**

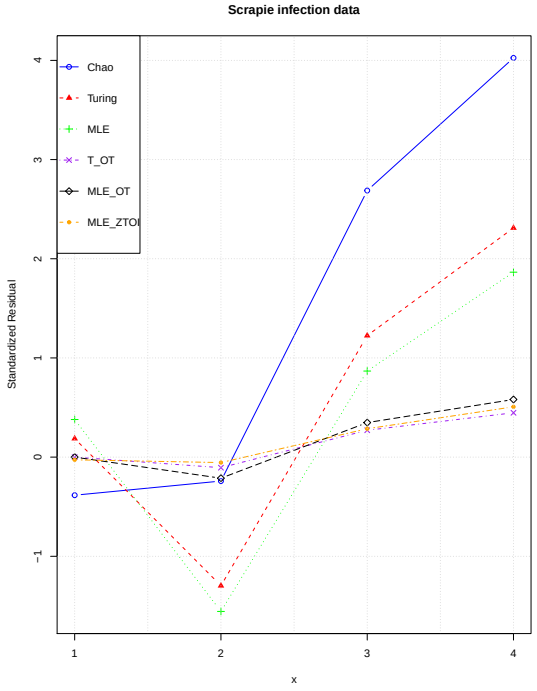

Figure 4.3: *Residual plot with all estimators for scrapie infection data*

Example 4.4 According to the case study of domestic violence in the Netherlands which was previously discussed in Chapter 3, Section 3.7, the frequency of domestic violence offenders can be seen in Table 3.12. The ratio plot is shown in Figure 3.3 left panel and it shows unclear one-inflation so here we investigate again by the likelihood ratio test and the result of testing shows the presence of one inflation; $LRT = 98.9135$ and $p - value < 0.001$. It can be assumed that the newly proposed estimator is viable and suitable with this data set. The results of estimation from the classical and newly proposed estimators are shown in Table 4.14. The pattern of results for all proposed estimators is different from the example 1, $\hat{N}_{\text{T\_OT}} > \hat{N}_{\text{MLE\_OT}} > \hat{N}_{\text{MLE\_ZTOI}}$. It is clearly seen that the estimate of MLE_ZTOI is smallest and obviously different from the estimate of T_OT and MLE_OT.

[1]For GOF-test, $\hat{p}_0 = \frac{\hat{f}_0}{\hat{N}}$ and $p = p_0$ for geometric model

Table 4.14: Results for domestic violence study

Estimator	$\hat{f}_0$	$\hat{N}$	Chi-square	p-value
Chao[1]	117,577	135,223	317.537	0.00000
Turing	103,233	120,879	166.795	0.00000
MLE	98,788	116,434	144.797	0.00000
T_OT	65,573	83,219	7.227	0.02696
MLE_OT	64,754	82,400	6.649	0.03599
MLE_ZTOI	**35,085**	**52,731**	**8.097**	**0.01745**

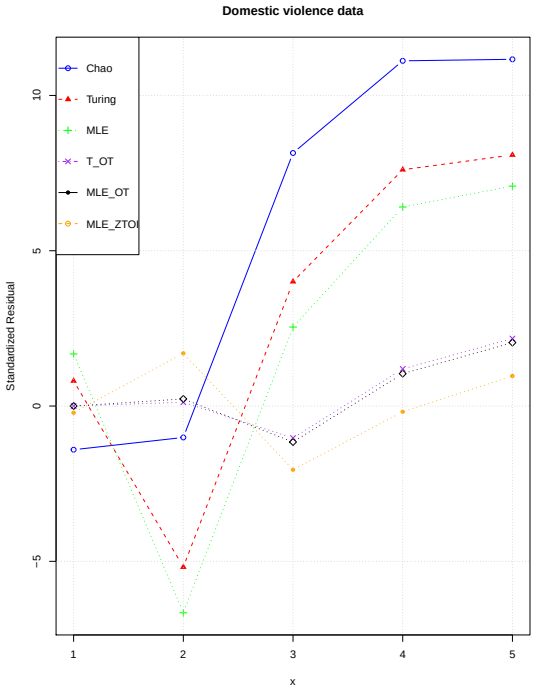

Figure 4.4: *Residual plot with all estimators for domestic violence data*

In terms of statistical model fitting, Figure 4.4 shows the standardized residual plot for this data set with all estimators. The fitted values of MLE_ZTOI are shown by the orange line, whereas the purple and black lines are for T_OT and MLE_OT respectively. It can be clearly seen from the graph and the p-value from the goodness of fit statistics in Table 4.4 that the estimated values from newly proposed estimator can fit the data well but it cannot clearly improve upon the fitting of the former proposed estimators.

Example 4.5 From the example of capture-recapture data in Section 2.6, the data of illegal immigrants in the Netherlands from police records are shown again in Table 4.15 in order to estimate the population size by all proposed estimators comparing with the classical estimators.

It can be noticed that the number of singletons is considerably higher than the number of doubletons. This indicates that the data may experience one-inflation. Then, we

Table 4.15: The data of illegal immigrants in the Netherlands

f_1	f_2	f_3	f_4	f_5	f_6	n
1645	183	37	13	1	1	1880

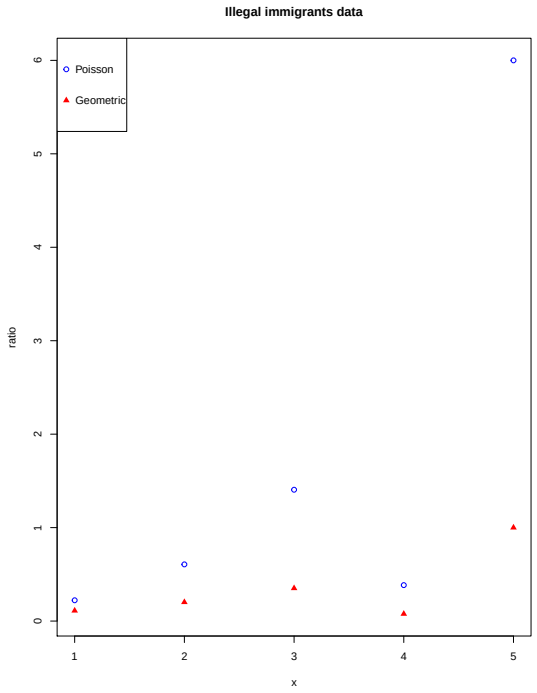

Figure 4.5: *Ratio plot for illegal immigrants data*

look at the ratio plot as shown in Figure 4.5. We found that a geometric distribution might be more suitable with this data than a Poisson distribution but we cannot see the evidence of one-inflation. However, the likelihood ratio test indicates that this data set undergoes one-inflation by $LRT = 20.8471$ and $p - value = 4.97 \times 10^{-6}$. Hence, all proposed estimators are applied to this data and the results of estimation from all estimators are shown in Table 4.16.

Table 4.16: Results for illegal immigrants study

Estimator	$\hat{f}_0$	$\hat{N}$	Chi-square	p-value
Chao[1]	14,787	16,667	92.009	0.00000
Turing	12,327	14,270	43.811	0.00000
MLE	11,588	13,468	36.326	0.00000
T_OT	6,461	8,341	3.085	0.37870
MLE_OT	6,311	8,191	2.917	0.40460
MLE_ZTOI	**2,983**	**4,863**	**2.859**	**0.40779**

As similar with previous examples, we consider the results in two parts; estimation and model fitting. In terms of estimation, the estimates from conventional estimators are

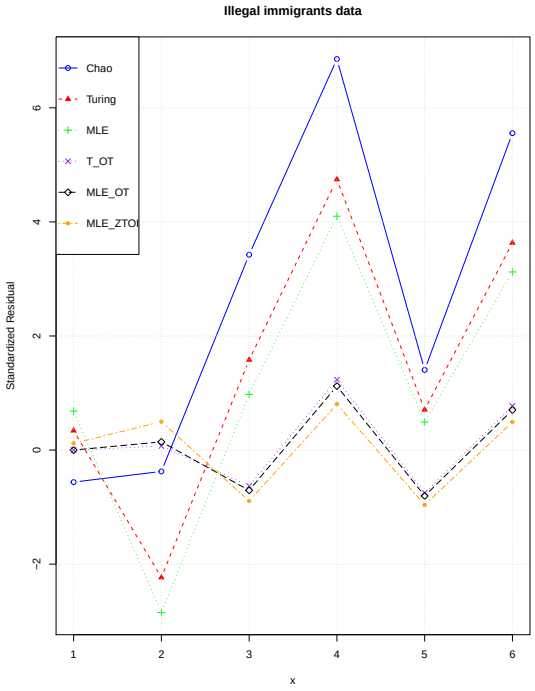

Figure 4.6: *Residual plot with all estimators for illegal immigrants data*

about double in size in comparison to the proposed estimators due to the effect of one-inflation as we expect. Interestingly, the estimates of T_OT and MLE_OT are similar; $\hat{N}_{\text{T\_OT}} = 8,341$ and $\hat{N}_{\text{MLE\_OT}} = 8,191$, whereas they are about double of MLE_ZTOI $\hat{N}_{\text{MLE\_ZTOI}} = 4,863$. The estimation of MLE_ZTOI seems to be the best in term of model fitting with $\chi^2 = 2.859$ and $p - value = 0.40779$. This corresponds with the graph in Figure 4.6. Nevertheless, it can be seen that the fitted values for count twos, threes, fours and so on seem to be identical under the two models as shown in Table 4.17. The question arises why the estimate for count zeros are so different. The reason is that the estimates of p_0 from two models are different although the estimates of the two models parameters are identical, $\hat{p_0}(\text{OT}) = \hat{\theta}$ whereas $\hat{p_0}(\text{ZTOI}) = \hat{\omega}\hat{\theta}$.

Table 4.17: Fitted frequencies under the MLE_OT and MLE_ZTOI estimators for illegal immigrants data

x	f_x	MLE_OT	MLE_ZTOI
1	1,645		1,647.02
2	183	181.07	176.39
3	37	41.55	42.84
4	13	9.54	10.41
5	1	2.19	2.53
6	1	0.50	0.61

4.8 Discussion and conclusion

In capture-recapture studies, the reliability of the population size estimate depends on the correctness of initial identification of sample individuals. Some applications undergo the situation of identifying individuals to wrong classes or some may face a trap-avoidance situation and these can cause the problem of one-inflation. To estimate the size N of an elusive population under one-inflation, two concepts are suggested to deal with. The first is based on a modification by truncating singletons and applying the conventional Turing and MLE approach to the one-truncated geometric data ($\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$). These are examined in Chapter 3. On the other hand, another concept, the model-based approach, focuses on developing a statistical model that describes the mechanism for extra one generation as shown in Chapter 4. In this chapter, a new estimator ($\hat{N}_{\text{MLE\_ZTOI}}$) is developed from the maximum likelihood approach by using the nested EM algorithm based upon the zero-truncated one-inflated geometric distribution. Chapter 3 shows that $\hat{N}_{\text{T\_OT}}$ and $\hat{N}_{\text{MLE\_OT}}$ can solve the problem of one-inflation and $\hat{N}_{\text{T\_OT}}$ performs better than $\hat{N}_{\text{MLE\_OT}}$. To evaluate the performance of the newly proposed estimators $\hat{N}_{\text{MLE\_ZTOI}}$, simulation studies are done again in order to compare the performance of all newly suggested and also existing conventional estimators. The simulation results provides evidence that $\hat{N}_{\text{MLE\_ZTOI}}$ shows a good performance in accuracy and perform best for all conditions under study. In addition, $\hat{N}_{\text{MLE\_ZTOI}}$ provides the smallest variance and mean square error for the big size of population ($N = 500, 1000$) whereas $\hat{N}_{\text{T\_OT}}$ provides the smallest for the small population size ($N = 20, 50, 100$). Overall it can be concluded that $\hat{N}_{\text{MLE\_ZTOI}}$ is better than $\hat{N}_{\text{T\_OT}}$ especially in case of high level of one-inflation and the large size of population.

Furthermore, we applied the newly proposed estimator with the two data sets from Chapter 3 and a new data set of illegal immigrants in the Netherlands. Also we compare the estimate with other estimators. All examples show that the newly proposed estimators can cope with the problem of one-inflation by providing smaller estimates than conventional estimators and also smaller than the previous suggested estimators except example 1. In term of statistical model fitting, it is found that the fitted values of the newly developed estimator can fit the data with one-inflation well and better than the conventional estimators in all of the cases studies particularly in the last example.

To sum up, it can be seen clearly that both concepts can solve the problem of one-inflation and each concept has a different strength. The first concept is simpler whereas the second concept uses a model-based approach to explain the extra-ones. Although the latter approach is more complex and more computational demanding, it produces the best estimates, especially for the large population size and high level of one-inflation. However, in case of a small size of population, although the first approach seems to be better than the second approach, the differences between the two are almost negligible.

Hence, both approaches seem reasonable to use with a slight benefit to the first one as it is the simpler concept.

Chapter 5

A Modified Chao Estimator for Zero-Truncated One-Inflated Count Distribution

As it is shown in Chapter 2, Chao's lower bound estimator is widely used to estimate population size in capture-recapture. One reason is that its formula is easy to calculate. It involves only the frequency of counts one and two. Moreover, it is not only asymptotically unbiased if the count distribution is a member of the power series family such as Poisson, binomial, exponential and geometric distribution but also provides a lower bound estimator if the count distribution is a mixture of the power series family. Nevertheless, Chao's lower bound estimator can severely overestimate if the count data have an excess number of ones, called one-inflation. To avoid the overestimation caused by one-inflation, Chao's estimator is modified to involve the frequency of counts of twos and threes instead of the frequency of counts of ones and twos. The modified Chao estimator shows a good performance in simulation studies under the geometric model, geometric model with one-inflation, mixture of geometric model, mixture of geometric with one-inflation, and it is reasonable to use in applications. Furthermore, it retains its good properties; asymptotically it is an unbiased estimator for a power series distribution with and without one-inflation and provides a lower bound estimator under a mixture of power series distributions with and without one-inflation. However, all Chao estimators are biased estimators when the sample size is small, so the bias-reduction versions of all Chao estimators have been developed that can reduce the bias considerably.

5.1 Introduction

The estimation of the size N of a closed elusive population using capture-recapture techniques is an important topic in many research areas. The problem consists in extrapolating a value for the number of units that have been missed, using the information gathered on the captured units at m occasions where m might be known or not. The data set for analysis consists of the empirical frequencies $f_1, f_2, ..., f_m$ where f_x is the frequency of distinct units identified exactly x times during the study period and m is the largest count observed in the sample. The predicted value of f_0 depends on the model for the capture of units based on a zero-truncated count distribution. The typical model is the Poisson or the binomial. However, heterogeneity in the capture probabilities is a common occurrence. It appears to be general agreement that a simple model $p(x|\lambda)$ is not flexible enough to capture the variation in the recapture probability for the distinct units of real-life population (see e.g. Pledger (2005) for the discrete mixture model and Dorazio and Royle (2003) for the continuous mixture model). It can be seen that in fact the mixture $p_x = \int_0^\infty p(x|\lambda)f(\lambda)d\lambda$ is a natural model for modelling a heterogeneous population. Nevertheless, there has also been discussion on the problem of identifiability of the mixture model (see Link (2003)). Several models for the individual capture probabilities have been investigated by several researchers. Huggins (2001) and Link (2003), for example, showed that the population size is not estimable in the presence of heterogeneity. Even though m is large, two models fitting the data evenly well can give entirely different estimates for N. In addition, boundary problems may occur in the maximum likelihood approach for finite mixture models (Wang and Lindsay (2005)). These emphasize the importance of Chao's lower bound estimator; a lower bound estimate on the population size might be the best result achievable for a heterogeneous population. Moreover, the lower bound approach has neither an identifiability problem, nor is there need to specify a mixing distribution; it is completely nonparametric. For these advantages the conventional estimators $f_1^2/(2f_2)$ for the unobserved frequency f_0 of zero-counts and $n + f_1^2/(2f_2)$ of the population size N (Chao (1987)) are popular and frequently used. It is asymptotically unbiased if count X follows a Poisson distribution and represents a lower bound if X follows a mixture of Poisson distributions. The purpose of this chapter is to present a modification of the Chao's estimator in the case of one-inflation as it can severely overestimate as shown in previous chapters. This effect is in contrast to the expectation of users of the estimator as it is expected to provide a lower bound that is relatively close to the true population size.

5.2 Power Series and Mixture of Power Series Distribution

The family of power series distribution is important due to the fact that it provides a very elegant and perceptive formulation of several classical discrete distributions that are used

in statistical research including the capture-recapture area. Most of the special, discrete distributions are included such as Poisson, binomial, geometric, negative-binomial with known shape parameter and others. The discrete random variable X is said to have a power series distribution if

$$p_x(\theta) = b_x\theta^x/g(\theta), \tag{5.1}$$

where $b_x > 0$ is a known, non-negative coefficient, θ is a positive parameter and x ranges over the set of non-negative integers. Here $g(\theta) = \sum_{x=0}^{\infty} b_x\theta^x$ is the normalizing constant. The specific member of the power series is defined by the coefficient b_x, for example, the Poisson is defined by $b_x = 1/x!$ whereas $b_x = \binom{m}{x}$ for $x = 0, 1, ..., m$ defines the binomial with positive integer m ($b_x = 0$ for $x > m$). The geometric is defined as $b_x = 1$. In case of the negative binomial, $p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}\theta^x(1-\theta)^k$ is also a member of the power series family with $b_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}$ for a known value of shape parameter $k > 0$ and $\theta \in (0, 1)$ is the event parameter. For $k = 1$ the negative binomial becomes the geometric distribution, conversely, for $k \to \infty$ the negative binomial approaches the Poisson distribution. However, a model which is a member of the power series distribution cannot adequately describe the target population of interest with heterogeneity especially when heterogeneity is unobserved. Hence, the mixture of power series distribution is considered in the sense that the unobserved heterogeneity is described by a latent variable T. The joint probability density of the count random variable X and the latent variable T is given by $f(x, t) = f(x|t)f(t)$ and $f(x|t) = f(x, t)/f(t)$ where $f(t) = \int_x f(x, t)dx$. The marginal density $\int_t f(x|t)f(t)dt$ of the count x is considered instead of the joint density because the information of latent variable is unknown. If we define the conditional density $f(x|t)$ by a power series density $p_x(\theta)$ and identify the latent variable t by the parameter θ, the mixture model for the power series family is obtained by

$$m_x = \int_\theta p_x(\theta)f(\theta)d\theta \tag{5.2}$$

where $p_x(\theta)$ is the mixture kernel and $f(\theta)$ is the mixing distribution. The mixture of power series distribution seems to be more flexible but it also involves a lack of identifiability or the boundary problem in maximum likelihood estimation. Hence, the lower bound estimation might be the better choice to avoid these problems. The original concept of lower bound estimation is to apply nonparametric statistical inference based upon the Cauchy-Schwarz inequality in the context of zero-truncated count mixture modelling by keeping the mixture distribution unspecified. To give some ideas of the lower bound approach by consider the Poisson mixture kernel $\exp(-\lambda)\lambda^x/x!$, the Cauchy-Schwarz inequality can be written as

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

and it is equivalent to $p_1^2 \leq p_0(2p_2)$. Replacing the theoretical probabilities p_i by their sample estimates f_i/N for $i = 1, 2$, leads to Chao's estimator for f_0 and N as $f_1^2/(2f_2)$

and $n + f_1^2/(2f_2)$, respectively. This idea is taken up again and developed further for the one-inflated count distributions in Section 5.3.

5.2.1 The Monotonicity of the Mixed Power Series Probability Ratio

The power series in (5.1) has an interesting property. If we consider the ratios of neighbouring probabilities for the power series multiplied by a known factor over the range of x, they provide a constant value which is equal to the unknown parameter θ as follows:

$$\begin{aligned}\frac{p_x}{p_{x+1}} &= \frac{b_x\theta^x/g(\theta)}{b_{x+1}\theta^{x+1}/g(\theta)}\\ &= \frac{b_x}{b_{x+1}}\frac{1}{\theta}\end{aligned}$$

and

$$\begin{aligned}\frac{p_{x-1}}{p_x} &= \frac{b_{x-1}\theta^{x-1}/g(\theta)}{b_x\theta^x/g(\theta)}\\ &= \frac{b_{x-1}}{b_x}\frac{1}{\theta}\end{aligned}$$

so

$$\theta = \frac{b_x}{b_{x+1}}\frac{p_{x+1}}{p_x} = \frac{b_{x-1}}{b_x}\frac{p_x}{p_{x-1}} = r_x. \tag{5.3}$$

In capture-recapture studies, zero-counts are truncated. Let $p_x^+(\theta) = p_x(\theta)/[1 - p_0(\theta)]$ and $m_x^+(\theta) = m_x(\theta)/[1 - m_0(\theta)]$ are the zero-truncated density for the power series and the mixture of power series distributions, respectively. The ratio r_x is also identical to the zero-truncated power series distribution since

$$r_x = \frac{b_x}{b_{x+1}}\frac{p_{x+1}^+}{p_x^+} = \frac{b_x}{b_{x+1}}\frac{p_{x+1}/[1-p_0(\theta)]}{p_x/[1-p_0(\theta)]} = \frac{b_x}{b_{x+1}}\frac{p_{x+1}}{p_x}.$$

The ratio r_x is estimated by $\hat{r}_x = \dfrac{b_x}{b_{x+1}}\dfrac{f_{x+1}}{f_x}$ where f_x is the frequency of observed value of count x. Böhning et al. (2013a) developed the graph of x against $\hat{r}_x$ as a diagnostic device for departure of a distribution and it is called the ratio plot (see also Chapter 2, Section 2.7). The horizontal line is consistent with homogeneous power series distributional observations whereas departures from a horizontal line provide evidence for the occurrence of unobserved heterogeneity leading to a mixture of power series distribution. Similarly, we can consider the ratio plot for mixtures

$$r_x = \frac{b_x}{b_{x+1}}\frac{m_{x+1}}{m_x}, \tag{5.4}$$

where the coefficients b_x is associated with the mixture kernel $p_x(\theta)$. It can be seen that the estimate of r_x will not change but the observed pattern in the ratio plot will change

and it will be interpreted in a different way due to the property of monotonicity in the mixture of power series as follows (see Chao (1987) and more general detail in Böhning (2008) and Böhning (2015)).

Theorem 1. Let $m_x = \int_\theta p_x(\theta) f(\theta) d\theta$ where $p_x(\theta)$ is a member of the power series family and $f(\theta)$ is an arbitrary density. Then, for $r_x = \frac{b_x}{b_{x+1}} \frac{m_{x+1}}{m_x}$ we have the following monotonicity:

$$r_x \leq r_{x+1}$$

for all $x = 0, 1, ...$

The result from Theorem 1 shows that the ratio plot will no longer show a horizontal line pattern in the case of a mixture of power series distributions but it will display a monotone non-decreasing pattern with an increase in x. Therefore, the ratio plot can be taken as an indicator for presence of heterogeneity if a monotone increasing pattern occurs in the ratio plot.

5.3 Modified Chao Estimation

From the ratio in (5.4) and a consequence of the result in Theorem 1, we obtain

$$\frac{b_{x-1}}{b_x} \frac{m_x}{m_{x-1}} \leq \frac{b_x}{b_{x+1}} \frac{m_{x+1}}{m_x}.$$

For $x = 1$ we have that

$$\frac{b_0}{b_1} \frac{m_1}{m_0} \leq \frac{b_1}{b_2} \frac{m_2}{m_1}$$

or

$$m_0 \geq \frac{b_0 b_2}{b_1^2} \frac{m_1^2}{m_2}. \tag{5.5}$$

Replacing the theoretical quantities m_x by their sample estimates f_x/N leads to the lower bound of Chao estimate for f_0 and N as (Chao (1987) and Chao (1989))

$$\hat{f}_0 = \frac{b_0 b_2}{b_1^2} \frac{f_1^2}{f_2} \quad \text{and} \quad \hat{N}_{\mathrm{C}} = n + \frac{b_0 b_2}{b_1^2} \frac{f_1^2}{f_2}, \tag{5.6}$$

respectively. It can be seen that the estimate in (5.6) provides a lower bound of the number of missing units in population and a lower bound of the population size if compares with the interpretation in (5.5). This estimate is the most popular and frequently used, particularly, when the assumption is based on a Poisson distribution, where $b_x = 1/x!$ then $\hat{f}_0 = f_1^2/2f_2$ and $\hat{N}_{\mathrm{C}} = n + f_1^2/2f_2$. However, if we apply the monotonicity property in Theorem 1 with all possible ratios of x, we get

$$\frac{b_0}{b_1} \frac{m_1}{m_0} \leq \frac{b_1}{b_2} \frac{m_2}{m_1} \leq \frac{b_2}{b_3} \frac{m_3}{m_2} \leq \frac{b_3}{b_4} \frac{m_4}{m_3} \leq \frac{b_4}{b_5} \frac{m_5}{m_4} ...$$

As a consequence many other lower bounds for m_0 are possible:

$$
\begin{aligned}
m_0 &\geq \frac{b_0 b_2}{b_1^2}\frac{m_1^2}{m_2} \quad \text{(For Chao estimation)} \\
m_0 &\geq \frac{b_0 b_3}{b_1 b_2}\frac{m_1 m_2}{m_3} \\
m_0 &\geq \frac{b_0 b_4}{b_1 b_3}\frac{m_1 m_3}{m_4} \\
m_0 &\geq \frac{b_0 b_5}{b_1 b_4}\frac{m_1 m_4}{m_5} \\
&\vdots
\end{aligned}
$$

It can be seen clearly that the lower bound of Chao is the largest, hence it is the best lower bound estimate of f_0 and N. Indeed, for example

$$
\begin{aligned}
&\text{from} && \frac{b_1}{b_2}\frac{m_2}{m_1} \leqslant \frac{b_2}{b_3}\frac{m_3}{m_2} \\
&\text{follows} && \frac{b_2}{b_1}\frac{m_1}{m_2} \geq \frac{b_3}{b_2}\frac{m_2}{m_3} \\
&\text{and finally} && \frac{b_0 b_2}{b_1^2}\frac{m_1^2}{m_2} \geq \frac{b_0 b_3}{b_1 b_2}\frac{m_1 m_2}{m_3}.
\end{aligned}
$$

Nevertheless, Chao's estimator can experience a severe problem of overestimation when there is one-inflation occurrence due to its form involves f_1^2 as shown in Chapter 3. Let $m_x^{'}$ be the one-inflated model described as follows:

$$
m_x^{'} = \begin{cases} \omega m_x & \text{for} \quad x \neq 1 \\ (1-\omega) + \omega m_x & \text{for} \quad x = 1 \end{cases} \tag{5.7}
$$

where m_x is the mixture of a power series member and $1-\omega$ represents the proportion of one-inflation. Note that $m_x^{'}$ in (5.7) can be written as $m_x^{'} = (1-\omega)\delta_1(x) + \omega m_x$ for $x = 0, 1, 2, ...$ and $\delta_1(x) = 1$ for $x = 1$ and zero otherwise. For a one-inflation model, counts of one will be no more compatible with the non-parametric mixture model as it is outside the class of non-parametric mixtures. Hence, Chao's estimator is no longer a lower bound estimator as Theorem 1 no longer holds. This point is different from a zero-inflated model as every zero-inflated power series distribution can be written as the mixture $(1-\omega)\delta_0(x) + \omega m_x = (1-\omega)b_x 0^x/g(\theta) + \omega m_x$ which is within the class of non-parametric mixtures of power series distribution.

To cope with this problem and avoid using f_1 for estimation, the monotonicity property is considered in the following another way:

$$
\frac{b_1}{b_2}\frac{m_2}{m_1} \leq \frac{b_2}{b_3}\frac{m_3}{m_2} \tag{5.8}
$$

So

$$m_1 \geq \frac{b_1 b_3}{b_2^2} \frac{m_2^2}{m_3}. \tag{5.9}$$

This bound has never been used nor enlarged on since it seems aimless as the counts of ones are observed and its bound is not required. However, if we replace m_1 in (5.5) with the bound of counts one given in (5.9), we yield

$$m_0 \geq \frac{b_0 b_2}{b_1^2} \left(\frac{b_1 b_3}{b_2^2} \frac{m_2^2}{m_3} \right)^2 \frac{1}{m_2}. \tag{5.10}$$

The bound can be simplified to

$$m_0 \geq \frac{b_0 b_3^2}{b_2^3} \frac{m_2^3}{m_3^2}, \tag{5.11}$$

then plugging the relative frequencies leads to

$$\hat{f}_0^* = \frac{b_0 b_3^2}{b_2^3} \frac{f_2^3}{f_3^2} \tag{5.12}$$

and

$$\hat{N}_{\mathrm{MC}} = n + \frac{b_0 b_3^2}{b_2^3} \frac{f_2^3}{f_3^2}. \tag{5.13}$$

From the Theorem 1, $\hat{f}_0^*$ can be expected to be smaller than $\hat{f}_0$ as

$$\hat{f}_0 \geq \frac{b_0 b_2}{b_1^2} \frac{f_1^2}{f_2} \geq \frac{b_0 b_3^2}{b_2^3} \frac{f_2^3}{f_3^2}. \tag{5.14}$$

Here inequalities are meant w.r.t. expected values. The generalised modified Chao estimator in (5.11) and (5.12) can be transformed to the specific forms for the mixture of particular power series member by substituting associated coefficients. For example, if m_x is

a Poisson mixture then	$\hat{f}_0^* = \frac{2}{9} \frac{f_2^3}{f_3^2}$,
a geometric mixture then	$\hat{f}_0^* = \frac{f_2^3}{f_3^2}$,
a binomial mixture then	$\hat{f}_0^* = \frac{(m-2)^2}{m(m-1)} \frac{2}{9} \frac{f_2^3}{f_3^2}$.

Note that $\hat{f}_0^*$ for a binomial mixture and a Poisson mixture is identical if m becomes large. Additionally, both $\hat{f}_0$ and $\hat{f}_0^*$ are asymptotically unbiased under a power series distribution, in the sense of no mixing involved.

According to (5.14), the bound $\hat{f}_0^*$ could be of interest as it will typically provide an even lower bound than the conventional Chao lower bound estimator $\hat{f}_0$. The advantage of the new lower bound estimator is shown here.

Theorem 2. Assume a one-inflation model $m_x^{'}$ as given in (5.7), where $m_x = \int_\theta p_x(\theta) f(\theta) d\theta$ where $p_x(\theta)$ is a member of the power series family and $f(\theta)$ is an arbitrary density. Then

$$m_0^{'} \geq \frac{b_0 b_3^2}{b_2^3} \frac{m_2^{'3}}{m_3^{'2}}. \tag{5.15}$$

Proof. For the non-inflated component we have that

$$m_0 \geq \frac{b_0 b_3^2}{b_2^3} \frac{m_2^3}{m_3^2}$$

and multiplying both sides with ω gives

$$\omega m_0 \geq \frac{b_0 b_3^2}{b_2^3} \frac{(\omega m_2)^3}{(\omega m_3)^2}$$

which is the result as $m_x^{'} = \omega m_x$ for $x \neq 1$. This ends the proof.

As a consequence of this theorem, we can expect that $\hat{f}_0^*$ is a lower bound estimator in the mean under heterogeneity of the power series distribution *and under one-inflation.*

Consider the case of power series distribution with one-inflation, in other words

$$m_x^{'} = (1-\omega)\delta_1(x) + \omega p_x.$$

Then, the conventional Chao estimator has asymptotic bias

$$\frac{b_0 b_2}{b_1^2} \frac{[(1-\omega) + \omega p_1]^2}{\omega p_2} N - b_0/g(\theta) N$$

Proof.

We have from (5.11) that

$$\begin{aligned}
E(\hat{N}_{\mathrm{C}}) &= E\left(n + \frac{b_0 b_2}{b_1^2} \frac{f_1^2}{f_2}\right) \\
&= E(n) + \frac{b_0 b_2}{b_1^2} E(f_1^2/f_2) \\
&= N(1-p_0) + \frac{b_0 b_2}{b_1^2} E(f_1^2/f_2) \\
&= N[1 - b_0/g(\theta)] + \frac{b_0 b_2}{b_1^2} \frac{[(1-\omega) + \omega p_1]^2}{\omega p_2} N \\
&= N + \left\{ \frac{b_0 b_2}{b_1^2} \frac{[(1-\omega) + \omega p_1]^2}{\omega p_2} N - b_0/g(\theta) N \right\},
\end{aligned}$$

which ends the proof.

In contrast, the suggested modified Chao estimator is asymptotically unbiased, even if the power series distribution is contaminated by one-inflation as can be seen in the following:

$$\begin{aligned}
E(\hat{N}_{\mathrm{MC}}) &= E\left(n + \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}\right) \\
&= E(n) + \frac{b_0 b_3^2}{b_2^3} E(f_2^3/f_3^2) \\
&= N(1-m_0) + \frac{b_0 b_3^2}{b_2^3} E(f_2^3/f_3^2) \\
&= N(1-\omega p_0) + \frac{b_0 b_3^2}{b_2^3}\frac{(\omega p_2)^3}{(\omega p_3)^2} N \\
&= N[1-\omega b_0/g(\theta)] + \frac{b_0 b_3^2}{b_2^3}\left[\frac{\omega^3 (b_2\theta^2/g(\theta))^3}{\omega^2 (b_3\theta^3/g(\theta))^2}\right] N \\
&= N[1-\omega b_0/g(\theta)] + [\omega b_0/g(\theta)]N \\
&= N
\end{aligned}$$

5.4 Bias Correction

The limitation of Chao's estimator is that it can have severe bias when the sample size is small. To reduce the bias of the modified Chao estimator, we need, firstly, to understand the occurrence of bias and the idea of bias-reduction for classical Chao estimator, then apply to the newly modified Chao estimator so we go back to consider the original Chao estimator in (5.5) again.

5.4.1 Classical Chao estimator with bias correction

As the arguments used to reduce bias are not easily available in the published literature, it is presented here. The idea is to attempt to estimate $Nm_1^2/m_2 = [E(f_1)]^2/E(f_2)$ by using f_1^2/f_2 but $E(f_1^2/f_2)$ is not necessarily close to $[E(f_1)]^2/E(f_2)$ except that f_1/N and f_2/N are close to m_1 and m_2, respectively; say if N is large enough. If N is small, hence, we cannot use f_1^2 to estimate $[E(f_1)]^2$ directly as $[E(f_1)]^2$ will not equate to $E(f_1^2)$. Indeed, we use an equidispersion property of Poisson assumption: "Mean = Variance",

$$Var(f_1) = E(f_1^2) - [E(f_1)]^2 = E(f_1). \tag{5.16}$$

It follows that $[E(f_1)]^2 = E(f_1^2) - E(f_1)$ which can be estimated as $f_1^2 - f_1$ leading to the numerator of the bias-corrected Chao estimator. Turning to the denominator, we also notice that the interest is in $1/\lambda = 1/E(f_2)$, but we can use $1/f_2$ estimate $E(1/f_2)$ only if the latter exists or $f_2 \neq 0$. Alternatively, $1/(f_2+1)$ will estimate $E[1/(f_2+1)]$

which can be evaluated using the Poisson assumption for f_2 as

$$\begin{aligned} E\left(\frac{1}{f_2+1}\right) &= \sum_{f_2=0}^{\infty} \frac{1}{f_2+1}\frac{\exp(-\lambda)\lambda^{f_2}}{f_2!} \\ &= \sum_{f_2=0}^{\infty} \frac{\exp(-\lambda)\lambda^{f_2}}{(f_2+1)!}\frac{\lambda}{\lambda} \\ &= \frac{1}{\lambda}\sum_{f_2=0}^{\infty} \frac{\exp(-\lambda)\lambda^{f_2+1}}{(f_2+1)!} \\ &= \frac{\exp(-\lambda)}{\lambda}(\exp(\lambda)-1) \\ &= 1/\lambda - \exp(-\lambda)/\lambda \\ &\approx \frac{1}{E(f_2)}, \end{aligned} \tag{5.17}$$

with the approximation error less than 0.001 for $\lambda > 5$. This leads to the bias-corrected classical Chao estimator

$$\hat{N}_{\mathrm{CC}} = n + \frac{b_0 b_2}{b_1^2}\frac{f_1(f_1-1)}{f_2+1}. \tag{5.18}$$

5.4.2 Modified Chao estimator with bias correction 1

In a similar way, the modified Chao estimator in (5.11) is considered again by separating m_2^3/m_3^2 into 2 parts as $(m_2^2/m_3)(m_2/m_3)$. We apply the Poisson assumptions in (5.16) and (5.17) for the numerator and denominator in the first part, respectively and use the property in (5.17) again for the denominator in the second part as follows

$$\begin{aligned} \frac{Nm_2^3}{m_3^2} &= \left(\frac{Nm_2^2}{m_3}\right)\left(\frac{m_2}{m_3}\right) \\ &= \left(\frac{[E(f_2)]^2}{E(f_3)}\right)\left(\frac{E(f_2)}{E(f_3)}\right) \\ &\approx \left(\frac{E(f_2^2)-E(f_2)}{E(f_3+1)}\right)\left(\frac{E(f_2)}{E(f_3+1)}\right) \end{aligned}$$

which can be estimated by $\left(\frac{f_2^2-f_2}{f_3+1}\right)\left(\frac{f_2}{f_3+1}\right)$ or $\frac{f_2^2(f_2-1)}{(f_3+1)^2}$. This provides the first version of bias correction for the modified Chao estimator

$$\hat{N}_{\mathrm{MC1}} = n + \frac{b_0 b_3^2}{b_2^3}\frac{f_2^2(f_2-1)}{(f_3+1)^2}. \tag{5.19}$$

5.4.3 Modified Chao estimator with bias correction 2

Here we give some details of developing another version for the bias-reduction of the modified Chao estimator. We use the property of third moment of the Poisson distribution to estimate $m_2^3 = E(f_2)^3$ for the numerator and keep the denominator identical as in the bias correction 1 in (5.19). We note that

$$\begin{aligned} E[f_2 - E(f_2)]^3 &= E[f_2^3 - 3f_2^2 E(f_2) + 3f_2 E(f_2)^2 - E(f_2)^3] \\ &= E(f_2^3) - 3E(f_2^2)E(f_2) + 3E(f_2)E(f_2)^2 - E(f_2)^3 \\ &= E(f_2^3) - 3E(f_2^2)E(f_2) + 2E(f_2)^3. \end{aligned}$$

Using a Poisson assumption for f_2, $E[f_2 - E(f_2)]^3 = E(f_2)$, we yield

$$E(f_2) = E(f_2^3) - 3E(f_2^2)E(f_2) + 2E(f_2)^3$$

Using the Poisson assumption of variance once more, we have that $E(f_2^2) = E(f_2) + E(f_2)^2$ so that

$$\begin{aligned} E(f_2) &= E[f_2^3] - 3[E(f_2) + E(f_2)^2]E(f_2) + 2E(f_2)^3 \\ 2E(f_2)^3 &= E(f_2) - E(f_2^3) + 3E(f_2)^2 + 3E(f_2)^3 \\ E(f_2)^3 &= E(f_2^3) - E(f_2) - 3E(f_2)^2 \end{aligned}$$

using the Poisson assumption for $E(f_2)^2 = E(f_2)^2 - E(f_2)$ again

$$\begin{aligned} E(f_2)^3 &= E(f_2^3) - E(f_2) - 3E(f_2^2) + 3E(f_2) \\ &= E(f_2^3) + 2E(f_2) - 3E(f_2^2) \end{aligned}$$

which can be estimated by $f_2^3 - 3f_2^2 + 2f_2$. Finally, we obtain the second version of bias correction for modified Chao estimator

$$\hat{N}_{\text{MC2}} = n + \frac{b_0 b_3^2}{b_2^3} \frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3 + 1)^2}. \tag{5.20}$$

5.4.4 Modified Chao estimator with bias correction 3

The last version of bias reduction, the numerator is maintained identical as in the bias correction version 2 in (5.20). For the denominator we note that $E[1/(f_3 + 1)(f_3 + 2)]$

can be evaluated using the Poisson assumption as

$$\begin{aligned}
E\left(\frac{1}{(f_3+1)(f_3+2)}\right) &= \sum_{f_3=0}^{\infty} \frac{1}{(f_3+1)(f_3+2)} \frac{\exp(-\lambda)\lambda^{f_3}}{f_3!} \\
&= \sum_{f_3=0}^{\infty} \frac{\exp(-\lambda)\lambda^{f_3}}{(f_3+2)!} \frac{\lambda^2}{\lambda^2} \\
&= \frac{1}{\lambda^2} \sum_{f_3=0}^{\infty} \frac{\exp(-\lambda)\lambda^{f_3+2}}{(f_3+2)!} \\
&= \frac{\exp(-\lambda)}{\lambda^2} \sum_{f_3=0}^{\infty} \frac{\lambda^{f_3+2}}{(f_3+2)!} \\
&= \frac{\exp(-\lambda)}{\lambda^2}(\exp(\lambda) - 1 - \lambda) \\
&= \frac{1}{\lambda^2} - \frac{\exp(-\lambda)}{\lambda^2} - \frac{\exp(-\lambda)}{\lambda} \\
&\approx \frac{1}{\lambda^2} = \frac{1}{E(f_3)^2},
\end{aligned}$$

which is giving an excellent approximation if $\lambda \geq 5$. Hence we derived the modified Chao estimator with bias correction version 3

$$\hat{N}_{\mathrm{MC3}} = n + \frac{b_0 b_3^2}{b_2^3} \frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3+1)(f_3+2)}. \tag{5.21}$$

5.5 Simulation study

A simulation experiment is undertaken to study the performance of all proposed estimators. To demonstrate how well the modified Chao estimator and all bias reduction versions work, we focus on the geometric distribution with and without one-inflation as it can incorporate the form of heterogeneity; the mixture of the Poisson and exponential distribution. Morever, due to all versions of bias reduction have been developed under the Poisson assumptions for the frequency f_x, it needs to be investigated if it works well outside the Poisson sampling for X; say whether the Poisson distribution is valid for the frequency. Note that the simulation is for the geometric distribution only, all the terms $\frac{b_0 b_2}{b_1^2}$ and $\frac{b_0 b_3^2}{b_2^3}$ are exactly one and thus disappear in the definition of the estimators. Six estimators of population size are compared:

1. Classical Chao estimator (C)
$$\hat{N}_{\mathrm{C}} = n + \frac{f_1^2}{f_2}$$
2. Classical Chao estimator with bias correction (CC)
$$\hat{N}_{\mathrm{CC}} = n + \frac{f_1(f_1-1)}{f_2+1}$$

3. Modified Chao estimator (MC)

$$\hat{N}_{\mathrm{MC}} = n + \frac{f_2^3}{f_3^2}$$

4. Modified Chao estimator with bias reduction 1 (MC1)

$$\hat{N}_{\mathrm{MC1}} = n + \frac{f_2^2(f_2 - 1)}{(f_3 + 1)^2}$$

5. Modified Chao estimator with bias reduction 2 (MC2)

$$\hat{N}_{\mathrm{MC2}} = n + \frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3 + 1)^2}$$

6. Modified Chao estimator with bias reduction 3 (MC3)

$$\hat{N}_{\mathrm{MC3}} = n + \frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3 + 1)(f_3 + 2)}$$

5.5.1 Simulation Scenarios

The scope of study covers six different scenes with different parameters.

1. The homogeneous geometric distribution with four parameter $\theta = 0.1, 0.2, 0.3, 0.4$.
2. The homogeneous geometric distribution as scene 1 with 20% one-inflation; or $1 - \omega = 0.2$. It means that the probability of taking only count one is 0.2 and the probability of the count is taken from homogeneous geometric (ω) is 0.8.
3. The homogeneous geometric distribution as scene 1 with 50% one-inflation; or $1 - \omega = 0.5$.
4. The equally weighted mixture of two geometric distributions. The six two-component mixture populations were considered: $(\theta_1, \theta_2) = (0.1, 0.2), (0.1, 0.3), (0.1, 0.4), (0.2, 0.3), (0.2, 0.4), (0.3, 0.4)$, where θ_1 and θ_2 is parameter of the geometric from the first and second component, respectively.
5. The equally weighted mixture of two geometric distributions as scene 4 with 20% one-inflation.
6. The equally weighted mixture of two geometric distributions as scene 4 with 50% one-inflation.

All scenarios are studied at three population sizes $N = 50, 100$ and $1,000$. Each scenario is repeated 5,000 times to eliminate any random error due to the simulation. Performance

is investigated by comparing relative bias (RBias), relative variance (RVar) and Relative mean square error (RMSE) to allow for comparisons across different sized populations (see more details in Chapter 3, Section 3.6.3).

5.5.2 Simulation Results

All results of simulation study are shown in Table 5.1 - 5.6 and also in Figure 5.1 - 5.12. The salient findings of the simulation study are summarised here.

- **Models without one-inflation**
  The results for this case are from scene 1 and 4, the geometric and mixture of geometric models. The results show that all six estimators are asymptotically unbiased with respect to the population size as we can see that the relative biases (RBias) converge to zero (see Figure 5.1 and 5.4 - 5.6). Indeed, Table 5.1 and 5.4 present that the MC estimator for the small populations ($N = 50, 100$) has the largest bias, variance and mean square error whereas the CC estimator gives the smallest. However, the MC has a good performance, similar to other estimators when population size is large ($N = 1,000$). Moreover, all estimators tend to be identical when the population size increases if comparing between general version and bias reduction version; (C and CC) and (MC and MC1, MC2, MC3), as we expect.

  Note that the population size is not the only factor influencing the performance of the estimators but also the parameter θ. Increasing parameter θ leads to a increase in bias, variance and mean square error for all estimators. This may be due to the fact that the observed counts show more excess of count zero as the mean converges to zero.

- **Models with one-inflation**
  The results of scene 2, 3, 5 and 6, the geometric and mixture of geometric with 20% and 50% one-inflation, are summarised here. Figure 5.2-5.3 and 5.7-5.12 provide evidences that the C and CC estimators give a severe overestimation of population size whereas the MC, MC1, MC2 and MC3 are asymptotic unbiased estimators based on one-inflated count distribution. It is clear that the larger one-inflation, the higher the overestimation bias of C and CC estimators. Now, we focus on the modified Chao estimator and its 3 bias reduction versions (MC, MC1, MC2 and MC3). Table 5.2-5.3 and 5.5-5.6 show that the MC estimator has the largest bias, variance and mean square error when the population size is small ($N = 50, 100$) whereas the MC3 estimator provides the best performance. However, all modified Chao estimators are identical when the population size is large. Similarly to models without one-inflation, increasing parameter θ leads to an increase in bias, variance and mean square error for all estimators.

Table 5.1: RBias, RVar and RMSE of six population size estimators under geometric model

N	θ	C	CC	MC	MC1	MC2	MC3
			Relative Bias				
50	0.1	0.0614	**-0.0026**	0.3434	0.1022	0.0361	-0.0090
	0.2	0.0718	**-0.0050**	0.5491	0.1792	0.0912	0.0106
	0.3	0.0871	**-0.0044**	0.7767	0.1703	0.0687	-0.0159
	0.4	0.1184	**-0.0037**	1.0672	0.2824	0.1352	-0.0209
100	0.1	0.0314	**0.0004**	0.1606	0.0462	0.0170	-0.0020
	0.2	0.0341	**0.0010**	0.1868	0.0647	0.0290	0.0029
	0.3	0.0394	**-0.0001**	0.2680	0.0882	0.0419	0.0031
	0.4	0.0537	**0.0025**	0.4267	0.1261	0.0625	0.0002
1000	0.1	0.0019	-0.0005	0.0084	0.0041	0.0016	**0.0002**
	0.2	0.0024	-0.0004	0.0109	0.0051	0.0020	**0.0000**
	0.3	0.0036	**0.0001**	0.0141	0.0058	0.0017	-0.0013
	0.4	0.0040	**-0.0005**	0.0245	0.0117	0.0060	0.0013
			Relative Variance				
50	0.1	0.0768	**0.0198**	1.5439	0.6513	0.3943	0.1122
	0.2	0.0859	**0.0341**	5.8897	2.1906	1.4911	0.4423
	0.3	0.1263	**0.0594**	10.7251	1.0830	0.7328	0.3244
	0.4	0.2657	**0.1117**	14.8403	4.6637	3.2516	0.9714
100	0.1	0.0175	**0.0087**	0.6149	0.0834	0.0585	0.0328
	0.2	0.0252	**0.0183**	1.3790	0.2016	0.1624	0.1040
	0.3	0.0381	**0.0292**	5.7567	0.5283	0.4398	0.2484
	0.4	0.0634	**0.0479**	5.8556	0.7630	0.6227	0.3725
1000	0.1	0.0007	**0.0007**	0.0023	0.0021	0.0020	0.0019
	0.2	0.0015	**0.0015**	0.0052	0.0050	0.0048	0.0047
	0.3	0.0027	**0.0027**	0.0110	0.0104	0.0102	0.0099
	0.4	0.0045	**0.0044**	0.0214	0.0199	0.0195	0.0190
			Relative Mean Square Error				
50	0.1	0.0774	**0.0198**	1.6315	0.6616	0.3955	0.1123
	0.2	0.0900	**0.0342**	6.1776	2.2223	1.4991	0.4423
	0.3	0.1332	**0.0594**	11.3164	1.1118	0.7374	0.3246
	0.4	0.2770	**0.1117**	15.9291	4.7425	3.2693	0.9717
100	0.1	0.0184	**0.0087**	0.6401	0.0855	0.0588	0.0328
	0.2	0.0264	**0.0183**	1.4136	0.2058	0.1632	0.1040
	0.3	0.0397	**0.0292**	5.8274	0.5360	0.4414	0.2484
	0.4	0.0663	**0.0479**	6.0365	0.7788	0.6265	0.3724
1000	0.1	0.0007	**0.0007**	0.0023	0.0021	0.0020	0.0019
	0.2	0.0015	**0.0015**	0.0054	0.0050	0.0048	0.0047
	0.3	0.0027	**0.0027**	0.0112	0.0104	0.0102	0.0099
	0.4	0.0045	**0.0044**	0.0220	0.0201	0.0195	0.0190

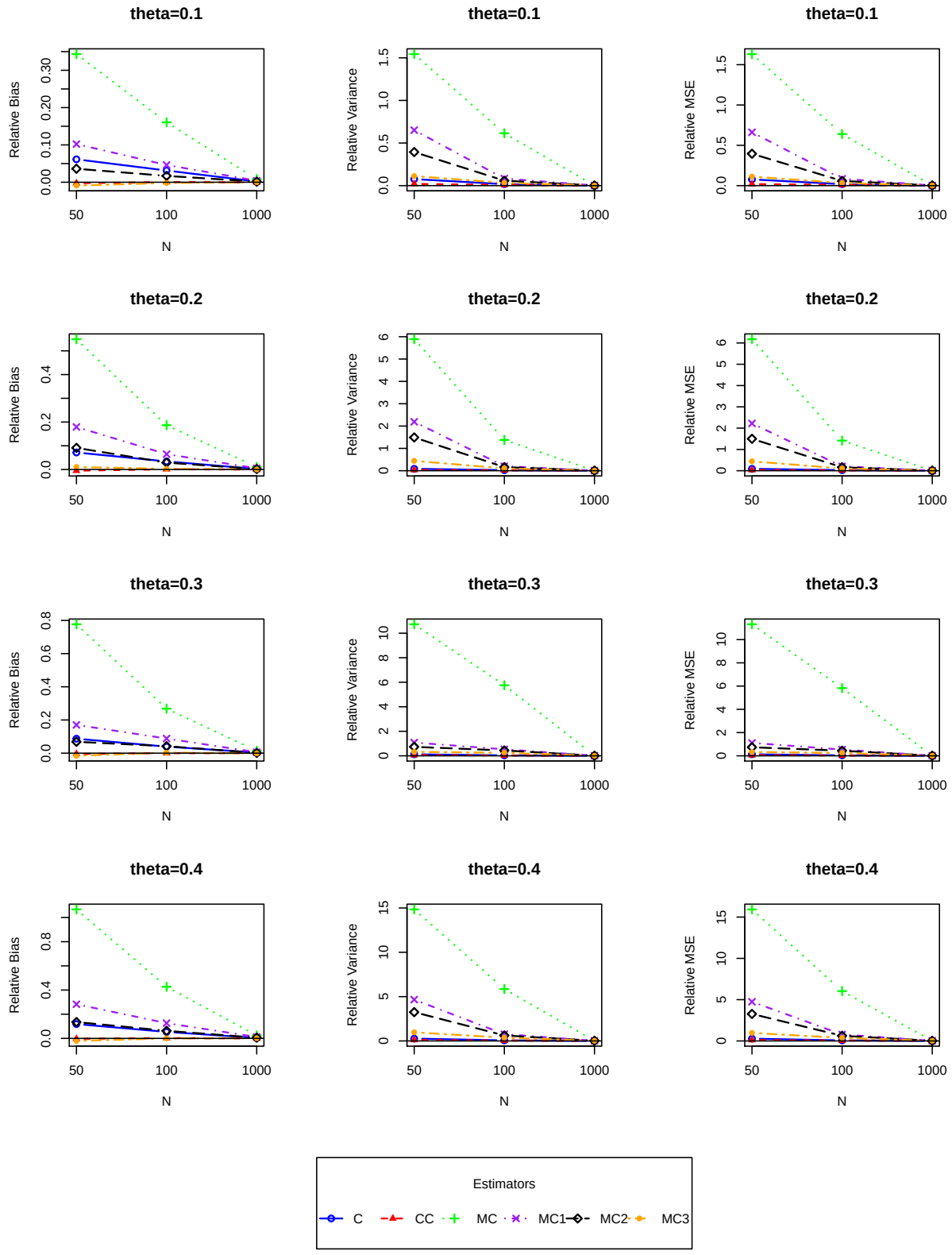

Figure 5.1: *RBias, RVar and RMSE of six estimators for counts drawn from geometric(θ)*

Table 5.2: RBias, RVar and RMSE of six population size estimators under geometric model with 20% one-inflation

N	θ	C	CC	MC	MC1	MC2	MC3
				Relative Bias			
50	0.1	1.5052	1.0036	0.2895	0.0764	0.0159	**-0.0187**
	0.2	1.2799	0.8846	0.5367	0.1622	0.0736	**-0.0007**
	0.3	1.2831	0.8982	0.7343	0.1761	0.0698	**-0.0199**
	0.4	1.4311	0.9962	0.9301	0.2220	0.0817	**-0.0493**
100	0.1	1.4118	1.0856	0.1814	0.0461	0.0161	**-0.0039**
	0.2	1.0662	0.8958	0.2251	0.0755	0.0374	**0.0060**
	0.3	1.0683	0.9117	0.2946	0.0848	0.0377	**-0.0015**
	0.4	1.2131	1.0268	0.5181	0.1345	0.0680	**0.0004**
1000	0.1	1.0841	1.0617	0.0085	0.0041	0.0016	**0.0002**
	0.2	0.9063	0.8925	0.0109	0.0050	0.0019	**-0.0001**
	0.3	0.9260	0.9129	0.0148	0.0064	0.0023	**-0.0007**
	0.4	1.0279	1.0127	0.0233	0.0103	0.0046	**-0.0002**
				Relative Variance			
50	0.1	2.1832	0.8768	0.9730	0.2994	0.1567	**0.0462**
	0.2	1.6987	0.6534	4.5151	1.2425	0.7641	**0.2310**
	0.3	1.8997	0.6369	6.7181	1.3116	0.8233	**0.2761**
	0.4	2.4456	0.9023	8.5397	2.6935	1.7622	**0.5139**
100	0.1	1.4664	0.5883	0.7497	0.0954	0.0621	**0.0287**
	0.2	0.4946	0.2967	1.5423	0.4108	0.3089	**0.1259**
	0.3	0.4664	0.3005	2.7356	0.3038	0.2372	**0.1396**
	0.4	0.7068	0.4150	8.6535	0.9034	0.7039	**0.3461**
1000	0.1	0.0419	0.0396	0.0019	0.0017	0.0016	**0.0016**
	0.2	0.0247	0.0238	0.0044	0.0041	0.0040	**0.0039**
	0.3	0.0269	0.0261	0.0092	0.0085	0.0083	**0.0081**
	0.4	0.0343	0.0333	0.0181	0.0166	0.0161	**0.0156**
				Relative Mean Square Error			
50	0.1	3.8734	1.8838	1.0067	0.3052	0.1570	**0.0465**
	0.2	3.2241	1.4358	4.7708	1.2686	0.7693	**0.2309**
	0.3	3.4636	1.4436	7.2204	1.3423	0.8280	**0.2764**
	0.4	4.3150	1.8945	9.3116	2.7423	1.7686	**0.5162**
100	0.1	3.4296	1.7666	0.7806	0.0975	0.0623	**0.0287**
	0.2	1.6296	1.0991	1.5924	0.4164	0.3103	**0.1259**
	0.3	1.6076	1.1317	2.8219	0.3110	0.2386	**0.1396**
	0.4	2.1750	1.4693	8.9193	0.9213	0.7084	**0.3461**
1000	0.1	1.2172	1.1668	0.0020	0.0017	0.0016	**0.0016**
	0.2	0.8459	0.8204	0.0046	0.0042	0.0040	**0.0039**
	0.3	0.8844	0.8594	0.0094	0.0086	0.0083	**0.0081**
	0.4	1.0908	1.0587	0.0186	0.0167	0.0162	**0.0156**

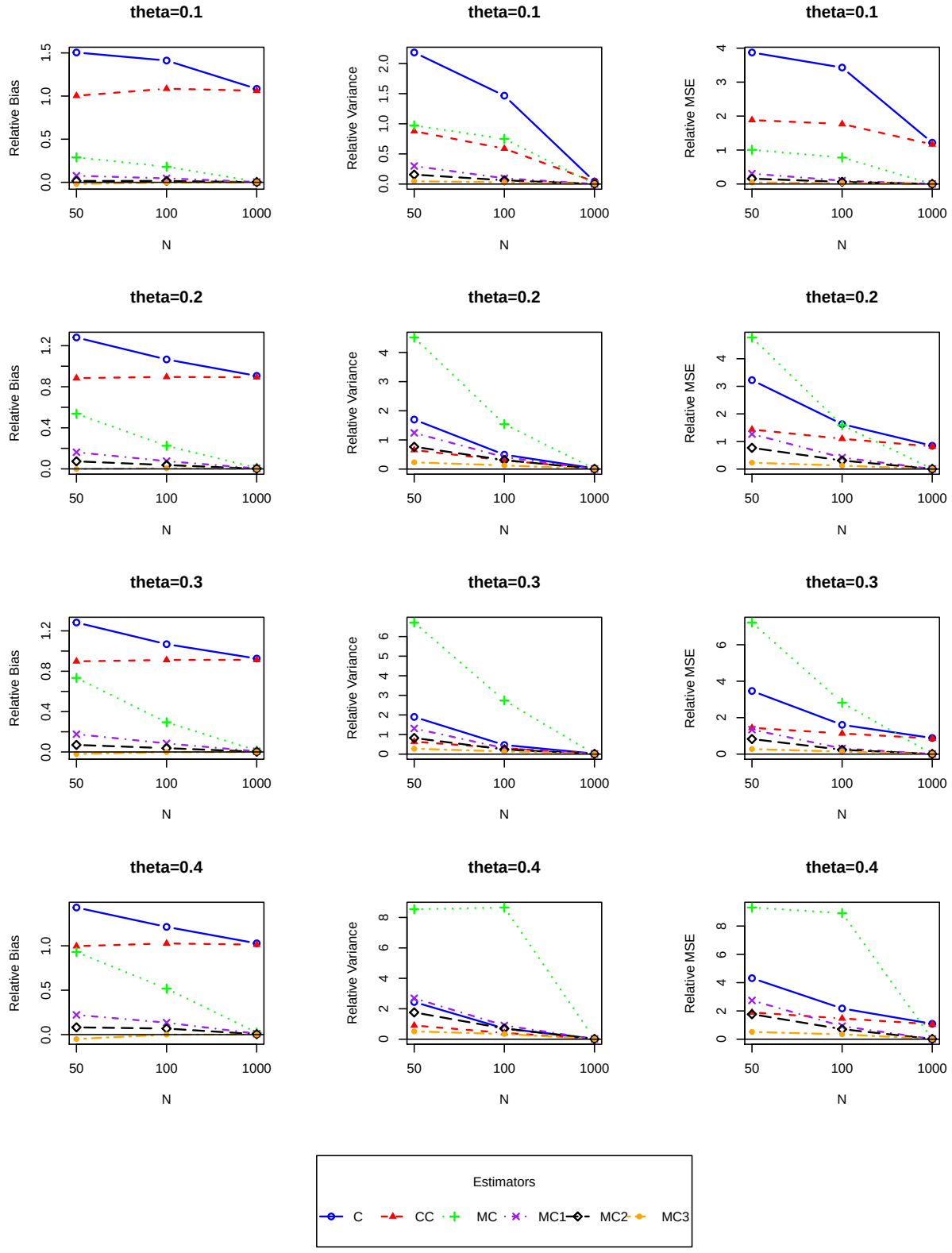

Figure 5.2: *RBias, RVar and RMSE of six estimators for counts drawn from geometric(θ) with 20% one-inflation*

Table 5.3: RBias, RVar and RMSE of six population size estimators under geometric model with 50% one-inflation

N	θ	C	CC	MC	MC1	MC2	MC3
				Relative Bias			
50	0.1	8.5678	6.2765	0.1827	0.0440	**-0.0030**	-0.0226
	0.2	7.1358	4.9910	0.3654	0.1014	0.0255	**-0.0225**
	0.3	6.5368	4.6201	0.4749	0.1310	**0.0328**	-0.0367
	0.4	7.0650	5.0781	0.5267	0.1264	**0.0079**	-0.0749
100	0.1	9.7085	7.1170	0.1690	0.0476	0.0156	**-0.0058**
	0.2	6.3522	5.0746	0.2354	0.0639	0.0245	**-0.0068**
	0.3	5.9645	4.8689	0.3586	0.0976	0.0443	**-0.0042**
	0.4	6.3117	5.1454	0.5515	0.1477	0.0729	**-0.0074**
1000	0.1	7.4840	7.2799	0.0090	0.0043	0.0017	**0.0003**
	0.2	5.2695	5.1751	0.0107	0.0046	0.0014	**-0.0006**
	0.3	4.9253	4.8468	0.0154	0.0066	0.0025	**-0.0006**
	0.4	5.2223	5.1373	0.0248	0.0111	0.0054	**0.0005**
				Relative Variance			
50	0.1	39.5137	19.8072	0.5022	0.1035	0.0420	**0.0122**
	0.2	32.7721	13.6394	1.5070	0.5301	0.2940	**0.0823**
	0.3	27.5307	10.5799	2.3636	0.9027	0.5247	**0.1468**
	0.4	35.5242	15.6737	2.4809	0.8467	0.4555	**0.1287**
100	0.1	52.7731	21.8199	0.4341	0.1368	0.0798	**0.0235**
	0.2	18.9726	7.9329	1.1083	0.2922	0.1950	**0.0632**
	0.3	15.2945	6.4314	1.9647	0.5362	0.3692	**0.1228**
	0.4	18.1564	7.7890	3.6750	1.1290	0.7738	**0.2331**
1000	0.1	1.9320	1.7462	0.0015	0.0012	0.0011	**0.0011**
	0.2	0.6511	0.6126	0.0030	0.0026	0.0025	**0.0024**
	0.3	0.5417	0.5138	0.0067	0.0059	0.0057	**0.0054**
	0.4	0.6168	0.5844	0.0122	0.0106	0.0102	**0.0096**
				Relative Mean Square Error			
50	0.1	75.9645	59.1974	0.5190	0.1054	0.0420	**0.0128**
	0.2	71.9171	38.5464	1.5542	0.5402	0.2946	**0.0828**
	0.3	61.6887	31.9233	2.4784	0.9197	0.5257	**0.1481**
	0.4	71.2714	41.4573	2.5835	0.8626	0.4555	**0.1342**
100	0.1	139.0425	72.4677	0.4492	0.1390	0.0800	**0.0235**
	0.2	58.6119	33.6831	1.1591	0.2962	0.1956	**0.0633**
	0.3	50.2873	30.1363	2.0862	0.5456	0.3711	**0.1228**
	0.4	56.9091	34.2630	3.9586	1.1506	0.7790	**0.2331**
1000	0.1	57.9412	54.7430	0.0015	0.0012	0.0011	**0.0011**
	0.2	28.4186	27.3938	0.0031	0.0027	0.0025	**0.0024**
	0.3	24.8000	24.0049	0.0069	0.0060	0.0057	**0.0054**
	0.4	27.8892	26.9757	0.0129	0.0107	0.0102	**0.0096**

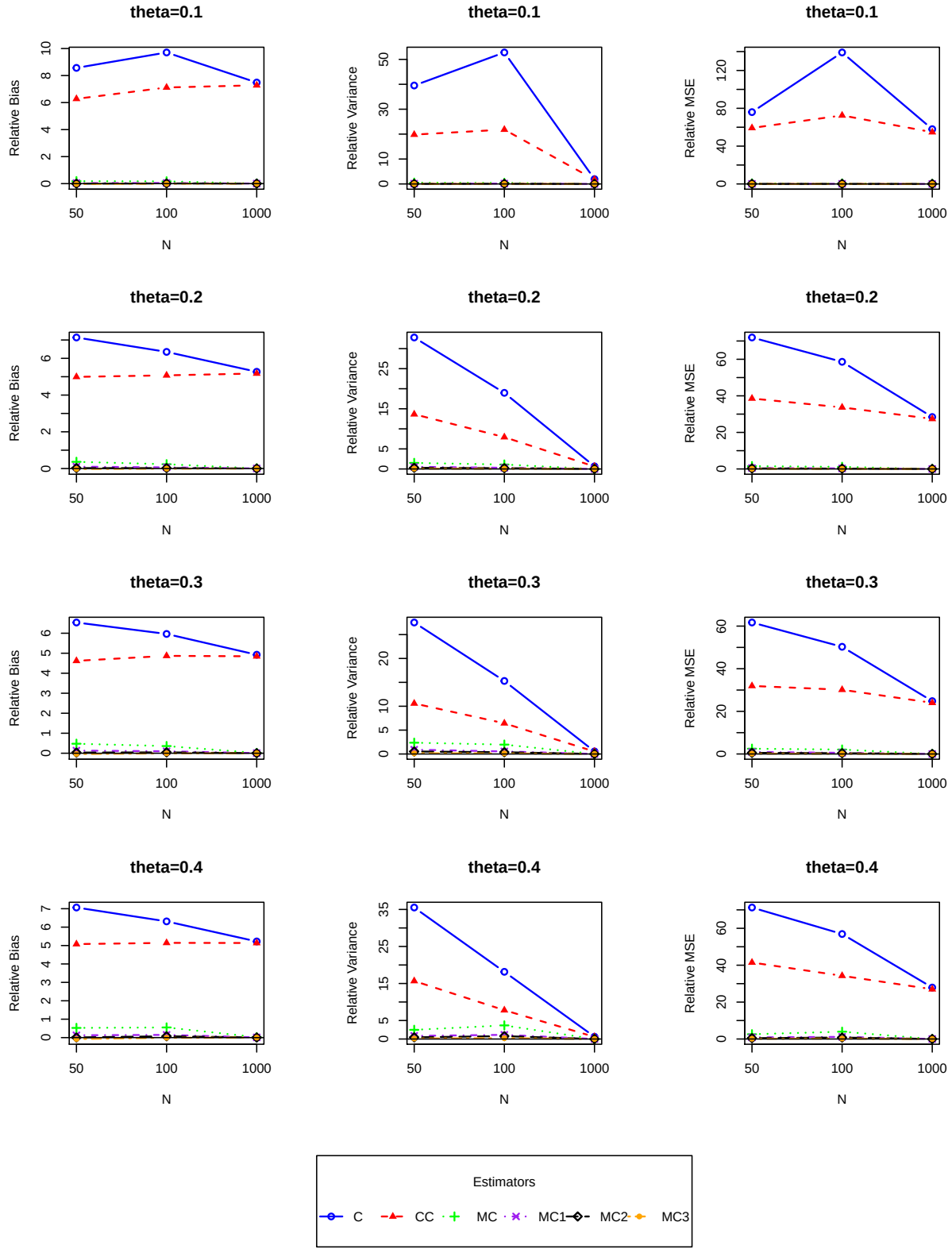

Figure 5.3: *RBias, RVar and RMSE of six estimators for counts drawn from geometric(θ) with 50% one-inflation*

5.6 Applications

In the following, population sizes are estimated through the classical and modified Chao estimator including bias correction versions so far considered in four well-known datasets: the data of H5N1 epidemic in Thailand, the data of scrapie-infected holding in France, the domestic violence data and the illegal immigrants data. Data are provided in Table 5.7. The Poisson and geometric ratio plots are provided in Figure 5.13 and for all applications we can see that the geometric model is more appropriate than Poisson model as it shows a horizontal line and the first point of ratio also shows the evidence of one-inflation. Therefore we apply all classical and modified Chao estimators based upon the geometric model. Population size estimates and unobserved units estimates are reported in Table 5.8.

5.6.1 H5N1 data

The data are from Vergne et al. (2014) that provides the number of highly pathogenic avian influenza (HPAI) H5N1 outbreak that were reported at subdistrict level in Thailand during the second epidemic wave (July 2004 - May 2005). The large epidemic occurred through out the country especially in the Central Plain for about two years, causing huge mortality in chickens and ducks. More than 65 million birds were culled and over US$ 130 million was spent compensating farmers' losses during 2004-2005. First two columns in Table 5.7 shows the spatial distribution of the number of H5N1 outbreaks reported in each subdistrict. There are 6,587 subdistricts with no outbreak. However, it seems to be suspected that these subdistricts might include a subdistrict where at least one outbreak occurred but none were reported. Our interest is to estimate the number of unobserved subdistricts with outbreak (f_0). The results are shown in Table 5.8 row 3-4. As we expect there are the large effects of one-inflation on classical Chao estimates ($\hat{f}_0^{\mathrm{C}} = 1,044$ and $\hat{N}_{\mathrm{C}} = 1,813$) while modified Chao estimator and its bias reduction versions can avoid overestimation by giving smaller estimates. However, the modified Chao estimator with bias reduction 3 (MC3) gives the smallest estimates ($\hat{f}_0^{\mathrm{MC3}} = 522$ and $\hat{N}_{\mathrm{MC3}} = 1,291$) similar to simulation results. Hence it is reasonable to use estimates from MC3 as it is the best performing.

5.6.2 Scrapie Infection data

The data on scrapie-infected holdings in France are obtained from the French classical scrapie surveillance programme (Vergne et al. (2012)). We are interested to evaluate the number of infected units that remain undetected by the surveillance system and also to estimate the total number of scrapie-infected holdings. Vergne et al. (2012) detected that there is a large amount of heterogeneity in the count data, corresponding to the top

left ratio plot in Figure 5.13, making the use of the simple zero-truncated Poisson model inappropriate. Hence, the zero-truncated geometric model with one-inflation might be more suitable for this example. Table 5.8 row 6-7 gives the estimates for f_0 and N of six estimators. Similarly, we can separate the estimates to 2 groups as they are definitely different: severe overestimates in C and CC, and, those not affected by one-inflation in MC, MC1, MC2 and MC3. All modified Chao estimators can cope with the problem of one-inflation particularly MC3 as it gives the smallest estimates.

5.6.3 Domestic Violence data

Van der Heijden et al. (2014) study the prevalence of domestic violence in the Netherlands for the year 2009 by using capture-recapture methods to estimate the total population size of offenders. The study is reported with the data given in Table 5.7 column 5-6. The total number of observed culprits is $n = 17,662$. There are 15,169 culprits identified exactly once in a domestic violence incident, 1,957 exactly twice and so forth. From the data and ratio plot in Figure 5.13 bottom left, it is noticed that the observed data may experience forms of one-inflation. It seems likely that a portion of perpetrators caught the first time change their behaviour and will no longer recapture as perpetrators again. The results are not surprising. It is similar to two first examples. The largest estimates comes from C ($\hat{f}_0^{\mathrm{C}} = 117,577$ and $\hat{N}_{\mathrm{C}} = 135,293$) whereas the smallest estimates are from MC3 ($\hat{f}_0^{\mathrm{MC3}} = 48,085$ and $\hat{N}_{\mathrm{MC3}} = 65,747$).

5.6.4 Illegal Immigrants data

We revisit the capture-recapture data of illegal immigrants in the Netherlands from police records (Van der Heijden et al. (2003b)) and use this data set to compare all proposed modified Chao estimators of population size with the classical Chao estimators. The data records contain information on the number of times each illegal immigrant was apprehended by the police (see Table 5.7 column 7-8). It can be noticed that the number of singletons is considerably higher than the number of doubletons. This indicates that the data may experience one-inflation as it is also supported by the ratio plot in Figure 5.13 right bottom (see also in Chapter 3 Section 3.7). Hence, all proposed modified estimators are applied to this data and the results of estimation are compared with results from classical estimator as shown in Table 5.8 row 12-13. Similarly to the previous examples, the pattern of results for all estimators is $\hat{N}_{\mathrm{C}} > \hat{N}_{\mathrm{CC}} > \hat{N}_{\mathrm{MC}} > \hat{N}_{\mathrm{MC1}} > \hat{N}_{\mathrm{MC2}} > \hat{N}_{\mathrm{MC3}}$. Here, the estimate using MC3 is smallest and obviously different from the estimate of C and CC as we expect.

5.7 Discussion and conclusion

The main issue of population size estimation is selecting an estimator (or several) from various approaches which should perform well and flexible even if the assumptions fail to hold for the data at hand. One of the crucial assumptions in capture-recapture is homogeneity. We know that heterogeneity in the capture probabilities is often occurring and ignoring heterogeneity can lead to biased estimations. Many approaches have been developed and offered to cope with this problem. The most popular estimator for heterogeneity is Chao's lower bound estimator as its formula is easy to calculate and involves only the frequency of count ones and twos. Moreover, Chao's estimator is asymptotically unbiased for a count distribution being a member of the power series family and also provides a lower bound if the count distribution is a mixture of the power series family. However, Chao's estimator seems to face the big problem of overestimation when the count data experience one-inflation.

A modified Chao estimator is developed to avoid overestimation stemming from one-inflation by using the frequency of count twos and threes instead of the frequency of count ones and twos. The main advantage of using the modified Chao estimation is that it retains the good properties of the classical Chao estimator; asymptotically it is an unbiased estimator for a power series distribution with and without one-inflation and provides a lower bound estimator under a mixture of power series distributions with and without one-inflation, as we can see from theoretical, analytic and simulation results. However, both classical and modified Chao estimators have a limitation. They are biased estimators when the sample size is small. Hence three versions of bias correction for modified Chao estimation have been developed. It is assumed that the frequency of counts follows a Poisson distribution which is a conventional assumption in frequency table analysis. The properties of the Poisson distribution are used to reduce the bias; equidispersion (mean = variance) and the third moment of the Poisson distribution. To investigate the performance of the modified Chao estimator and demonstrate how well all bias reduction versions work, the geometric and the mixture of geometric distribution with and without one-inflation are considered in a simulation study. The simulation results show that the larger the one-inflation, the higher the overestimation bias of the classical Chao estimator. On the other hand, the modified Chao estimator can avoid the effect of one-inflation as it shows a good performance for all situations except when the sample size is small. Furthermore, all bias reduction versions of the modified Chao estimator have a good performance for all cases of study, especially good is the last version of bias reduction (MC3).

In summary, the modified Chao estimator can reduce the bias from one-inflation and all bias-reduction versions can reduce the bias for small sample size settings considerably. Hence it is reasonable to use all proposed modified Chao estimators for one-inflation count data in practice.

Table 5.4: RBias, RVar and RMSE of six population size estimators under mixture of geometric model

N	θ_1	θ_2	C	CC	MC	MC1	MC2	MC3
				Relative Bias				
50	0.1	0.2	0.0718	**-0.0021**	0.4769	0.1220	0.0478	-0.0084
		0.3	0.0843	**-0.0036**	0.5613	0.1382	0.0494	-0.0210
		0.4	0.0891	**-0.0102**	0.6638	0.1381	0.0345	-0.0502
	0.2	0.3	0.0855	**-0.0003**	0.6617	0.1396	0.0509	-0.0191
		0.4	0.0918	**-0.0057**	0.7696	0.1571	0.0526	-0.0369
	0.3	0.4	0.1142	**0.0037**	0.9343	0.2100	0.0870	-0.0277
100	0.1	0.2	0.0291	**-0.0022**	0.1919	0.0712	0.0360	0.0072
		0.3	0.0362	**-0.0005**	0.1918	0.0531	0.0152	-0.0126
		0.4	0.0337	**-0.0092**	0.2593	0.0515	0.0063	-0.0311
	0.2	0.3	0.0330	**-0.0026**	0.2169	0.0779	0.0366	0.0047
		0.4	0.0369	**-0.0048**	0.2722	0.0741	0.0263	-0.0144
	0.3	0.4	0.0439	**-0.0011**	0.3194	0.0983	0.0441	-0.0041
1000	0.1	0.2	0.0032	**0.0005**	0.0087	0.0035	0.0007	-0.0011
		0.3	**0.0002**	-0.0029	0.0032	-0.0032	-0.0066	-0.0088
		0.4	**-0.0055**	-0.0091	-0.0099	-0.0181	-0.0221	-0.0250
	0.2	0.3	0.0022	**-0.0010**	0.0094	0.0024	-0.0012	-0.0036
		0.4	**-0.0008**	-0.0045	-0.0026	-0.0111	-0.0153	-0.0183
	0.3	0.4	0.0039	**-0.0001**	0.0104	0.0003	-0.0044	-0.0081
				Relative Variance				
50	0.1	0.2	0.0868	**0.0304**	3.6927	0.9183	0.5884	0.1873
		0.3	0.1149	**0.0398**	4.9532	0.9841	0.6157	0.2072
		0.4	0.1653	**0.0603**	6.1848	1.2303	0.7747	0.2557
	0.2	0.3	0.1335	**0.0593**	8.8799	0.8303	0.5545	0.2470
		0.4	0.1419	**0.0679**	9.4369	1.4284	0.9527	0.3469
	0.3	0.4	0.2125	**0.0930**	12.8778	2.0382	1.3688	0.4912
100	0.1	0.2	0.0183	**0.0118**	1.2129	0.4110	0.3064	0.1175
		0.3	0.0275	**0.0184**	0.7058	0.1524	0.1186	0.0782
		0.4	0.0371	**0.0245**	4.0213	0.3602	0.2839	0.1551
	0.2	0.3	0.0289	**0.0220**	0.8094	0.2212	0.1795	0.1254
		0.4	0.0396	**0.0300**	3.4121	0.4147	0.3353	0.2020
	0.3	0.4	0.0498	**0.0381**	2.3844	0.4539	0.3702	0.2403
1000	0.1	0.2	0.0011	**0.0011**	0.0039	0.0036	0.0035	0.0034
		0.3	0.0016	**0.0016**	0.0056	0.0052	0.0051	0.0049
		0.4	0.0023	**0.0022**	0.0084	0.0078	0.0076	0.0074
	0.2	0.3	0.0022	**0.0021**	0.0077	0.0073	0.0071	0.0069
		0.4	0.0028	**0.0028**	0.0104	0.0097	0.0095	0.0093
	0.3	0.4	0.0034	**0.0034**	0.0145	0.0136	0.0133	0.0130
				Relative Mean Square Error				
50	0.1	0.2	0.0884	**0.0304**	3.8436	0.9332	0.5907	0.1874
		0.3	0.1180	**0.0398**	5.1770	1.0032	0.6181	0.2076
		0.4	0.1660	**0.0604**	6.4542	1.2494	0.7759	0.2583
	0.2	0.3	0.1392	**0.0593**	9.2624	0.8498	0.5571	0.2474
		0.4	0.1482	**0.0680**	9.9537	1.4531	0.9555	0.3483
	0.3	0.4	0.2219	**0.0930**	13.6221	2.0822	1.3764	0.4920
100	0.1	0.2	0.0191	**0.0118**	1.2487	0.4160	0.3077	0.1176
		0.3	0.0288	**0.0184**	0.7425	0.1552	0.1188	0.0784
		0.4	0.0383	**0.0246**	4.0877	0.3629	0.2839	0.1561
	0.2	0.3	0.0300	**0.0220**	0.8564	0.2272	0.1809	0.1254
		0.4	0.0410	**0.0300**	3.4862	0.4202	0.3360	0.2022
	0.3	0.4	0.0517	**0.0381**	2.4865	0.4635	0.3721	0.2403
1000	0.1	0.2	0.0011	**0.0011**	0.0040	0.0036	0.0035	0.0034
		0.3	0.0016	**0.0016**	0.0056	0.0052	0.0051	0.0050
		0.4	0.0023	**0.0023**	0.0085	0.0081	0.0081	0.0080
	0.2	0.3	0.0022	**0.0021**	0.0078	0.0073	0.0071	0.0070
		0.4	0.0028	**0.0028**	0.0104	0.0098	0.0097	0.0096
	0.3	0.4	0.0034	**0.0034**	0.0146	0.0136	0.0133	0.0130

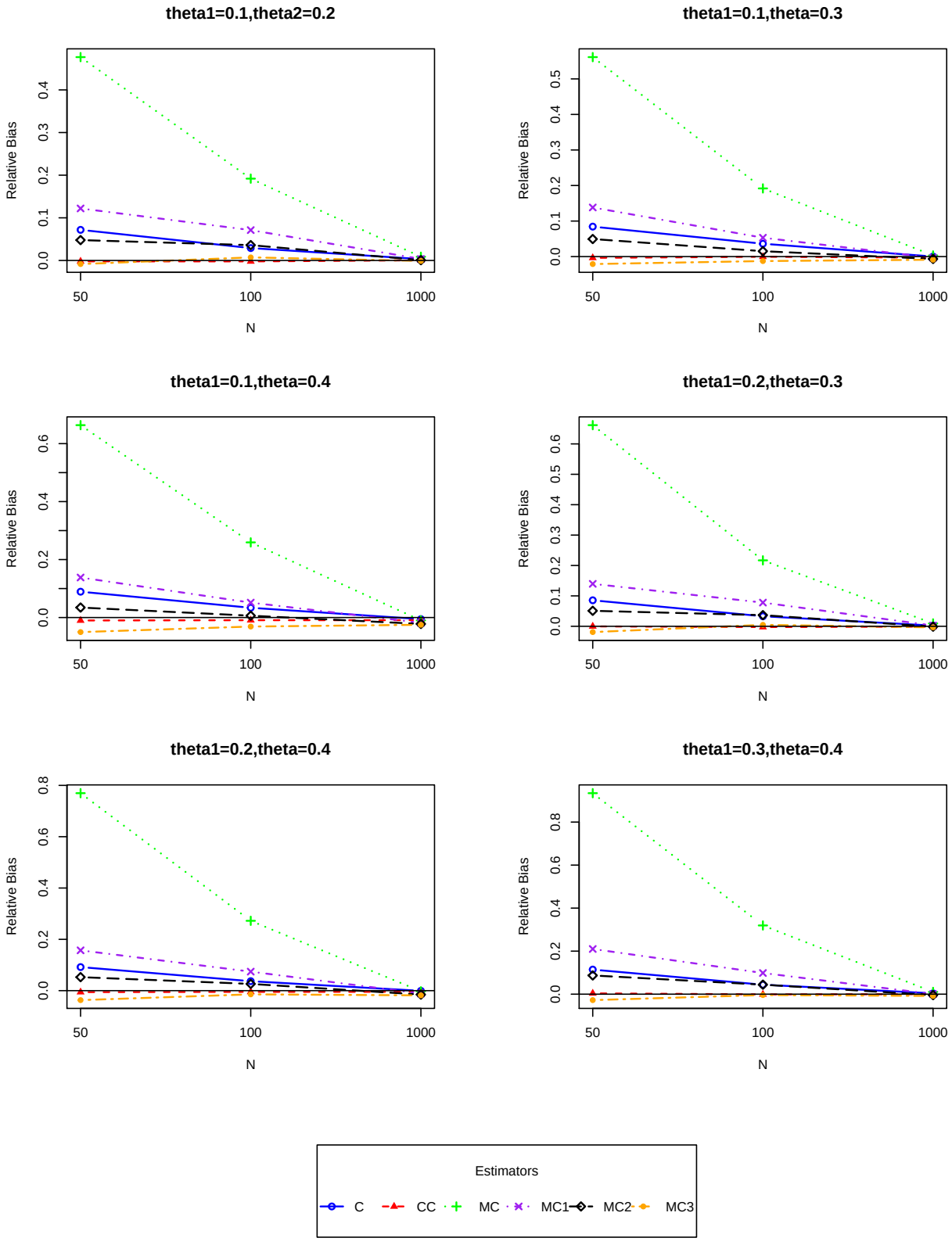

Figure 5.4: *RBias of six estimators for counts drawn from mixture of geometric*(θ_1, θ_2)

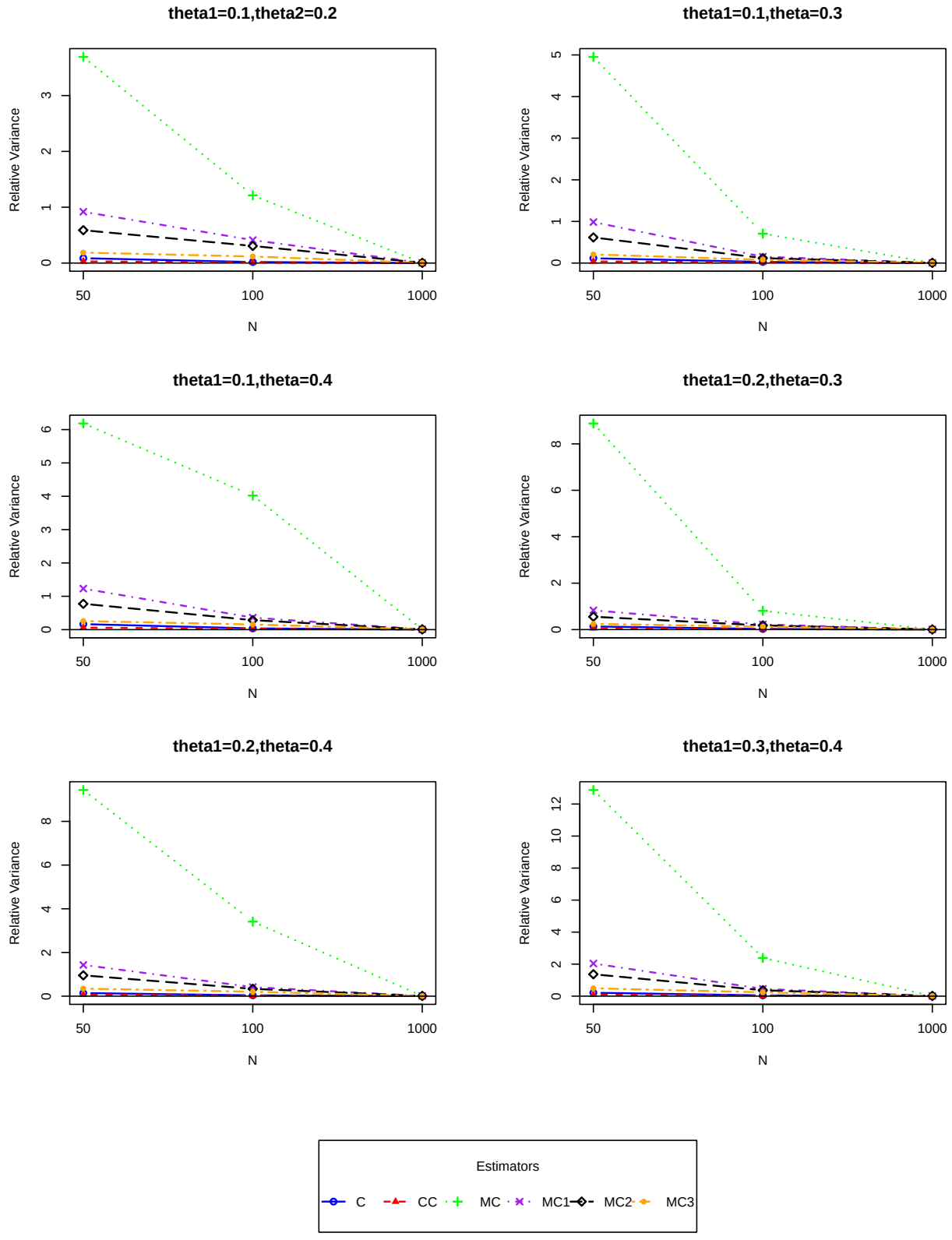

Figure 5.5: *RVar of six estimators for counts drawn from mixture of geometric*(θ_1, θ_2)

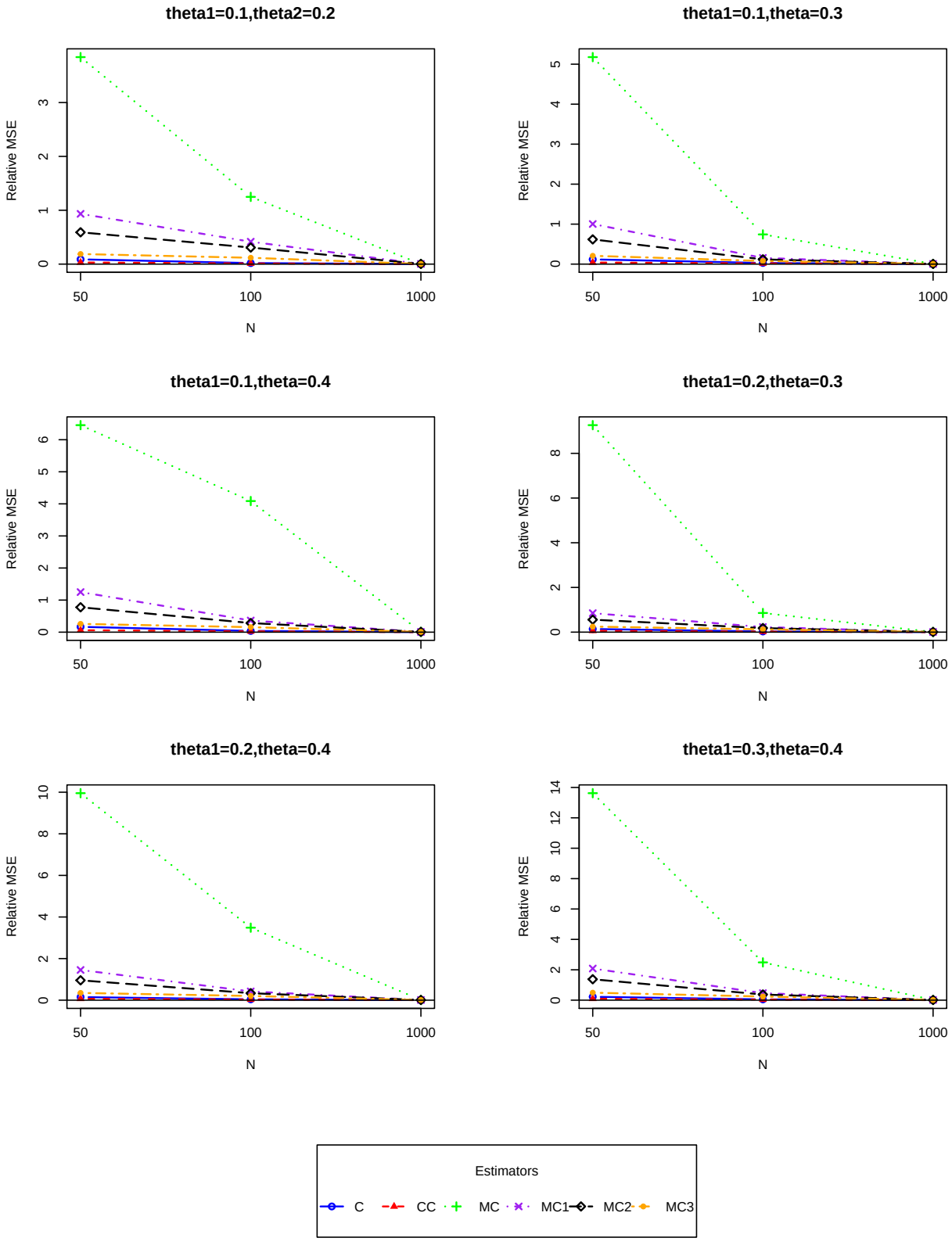

Figure 5.6: *RMSE of six estimators for counts drawn from mixture of geometric*(θ_1, θ_2)

Table 5.5: RBias, RVar and RMSE of six population size estimators under mixture of geometric model with 20% one-inflation

N	θ_1	θ_2	C	CC	MC	MC1	MC2	MC3
					Relative Bias			
50	0.1	0.2	1.4079	0.9503	0.3880	0.1080	0.0336	**-0.0185**
		0.3	1.3844	0.9358	0.5367	0.1191	0.0334	**-0.0280**
		0.4	1.4842	1.0178	0.5972	0.1190	0.0204	**-0.0525**
	0.2	0.3	1.3196	0.9084	0.6219	0.1525	0.0587	**-0.0176**
		0.4	1.4005	0.9740	0.6500	0.1668	0.0584	**-0.0366**
	0.3	0.4	1.3488	0.9506	0.8244	0.2313	0.0994	**-0.0234**
100	0.1	0.2	1.1773	0.9609	0.1936	0.0665	0.0311	**0.0018**
		0.3	1.1719	0.9666	0.2352	0.0563	0.0170	**-0.0123**
		0.4	1.2715	1.0432	0.3036	0.0577	**0.0101**	-0.0300
	0.2	0.3	1.0513	0.8931	0.2563	0.0758	0.0332	**-0.0008**
		0.4	1.1347	0.9559	0.2994	0.0723	0.0234	**-0.0184**
	0.3	0.4	1.1171	0.9521	0.3632	0.0926	0.0378	**-0.0107**
1000	0.1	0.2	0.9755	0.9587	0.0082	0.0030	**0.0002**	-0.0016
		0.3	0.9799	0.9637	0.0070	0.0004	**-0.0031**	-0.0053
		0.4	1.0412	1.0235	-0.0043	-0.0126	**-0.0167**	-0.0196
	0.2	0.3	0.9140	0.9006	0.0102	0.0031	**-0.0005**	-0.0030
		0.4	0.9665	0.9520	-0.0007	-0.0092	**-0.0134**	-0.0164
	0.3	0.4	0.9752	0.9611	0.0142	0.0039	**-0.0009**	-0.0047
					Relative Variance			
50	0.1	0.2	1.9539	0.7780	2.0791	0.6323	0.3634	**0.1075**
		0.3	1.9676	0.7673	3.2093	0.7264	0.4328	**0.1361**
		0.4	2.4558	1.0385	4.3556	0.9688	0.5914	**0.1823**
	0.2	0.3	1.9729	0.6636	5.5099	1.1691	0.7334	**0.2371**
		0.4	2.2912	0.9185	5.7805	1.8916	1.2273	**0.3622**
	0.3	0.4	2.0398	0.8328	8.0650	2.1248	1.3279	**0.4082**
100	0.1	0.2	0.7317	0.4203	0.9091	0.6121	0.4627	**0.1395**
		0.3	0.6591	0.3504	1.3838	0.1668	0.1218	**0.0683**
		0.4	0.8596	0.4588	2.2214	0.3197	0.2276	**0.1005**
	0.2	0.3	0.4072	0.2631	1.6437	0.2435	0.1864	**0.1093**
		0.4	0.6054	0.3473	2.2074	0.3221	0.2421	**0.1338**
	0.3	0.4	0.4868	0.3232	2.6291	0.3879	0.3032	**0.1822**
1000	0.1	0.2	0.0302	0.0289	0.0030	0.0028	0.0027	**0.0026**
		0.3	0.0315	0.0303	0.0050	0.0046	0.0045	**0.0043**
		0.4	0.0359	0.0345	0.0068	0.0062	0.0059	**0.0057**
	0.2	0.3	0.0256	0.0248	0.0062	0.0058	0.0056	**0.0054**
		0.4	0.0285	0.0276	0.0083	0.0076	0.0074	**0.0072**
	0.3	0.4	0.0301	0.0292	0.0124	0.0114	0.0111	**0.0108**
					Relative Mean Square Error			
50	0.1	0.2	3.6660	1.6809	2.1881	0.6438	0.3645	**0.1079**
		0.3	3.6997	1.6430	3.4514	0.7405	0.4338	**0.1368**
		0.4	4.3308	2.0742	4.6287	0.9828	0.5917	**0.1850**
	0.2	0.3	3.6231	1.4887	5.8660	1.1921	0.7367	**0.2374**
		0.4	4.0863	1.8669	6.1495	1.9190	1.2304	**0.3635**
	0.3	0.4	3.7000	1.7363	8.6683	2.1778	1.3375	**0.4087**
100	0.1	0.2	2.1084	1.3436	0.9456	0.6164	0.4636	**0.1394**
		0.3	2.0303	1.2846	1.4386	0.1700	0.1221	**0.0684**
		0.4	2.4682	1.5469	2.3120	0.3230	0.2276	**0.1014**
	0.2	0.3	1.5114	1.0606	1.7089	0.2492	0.1875	**0.1092**
		0.4	1.8910	1.2610	2.2963	0.3273	0.2426	**0.1342**
	0.3	0.4	1.7346	1.2297	2.7604	0.3964	0.3046	**0.1823**
1000	0.1	0.2	0.9819	0.9480	0.0031	0.0028	0.0027	**0.0026**
		0.3	0.9917	0.9591	0.0051	0.0046	0.0045	**0.0044**
		0.4	1.1198	1.0821	0.0068	0.0063	0.0062	**0.0061**
	0.2	0.3	0.8610	0.8359	0.0063	0.0058	0.0056	**0.0054**
		0.4	0.9627	0.9340	0.0083	0.0077	0.0076	**0.0074**
	0.3	0.4	0.9811	0.9528	0.0126	0.0115	0.0111	**0.0108**

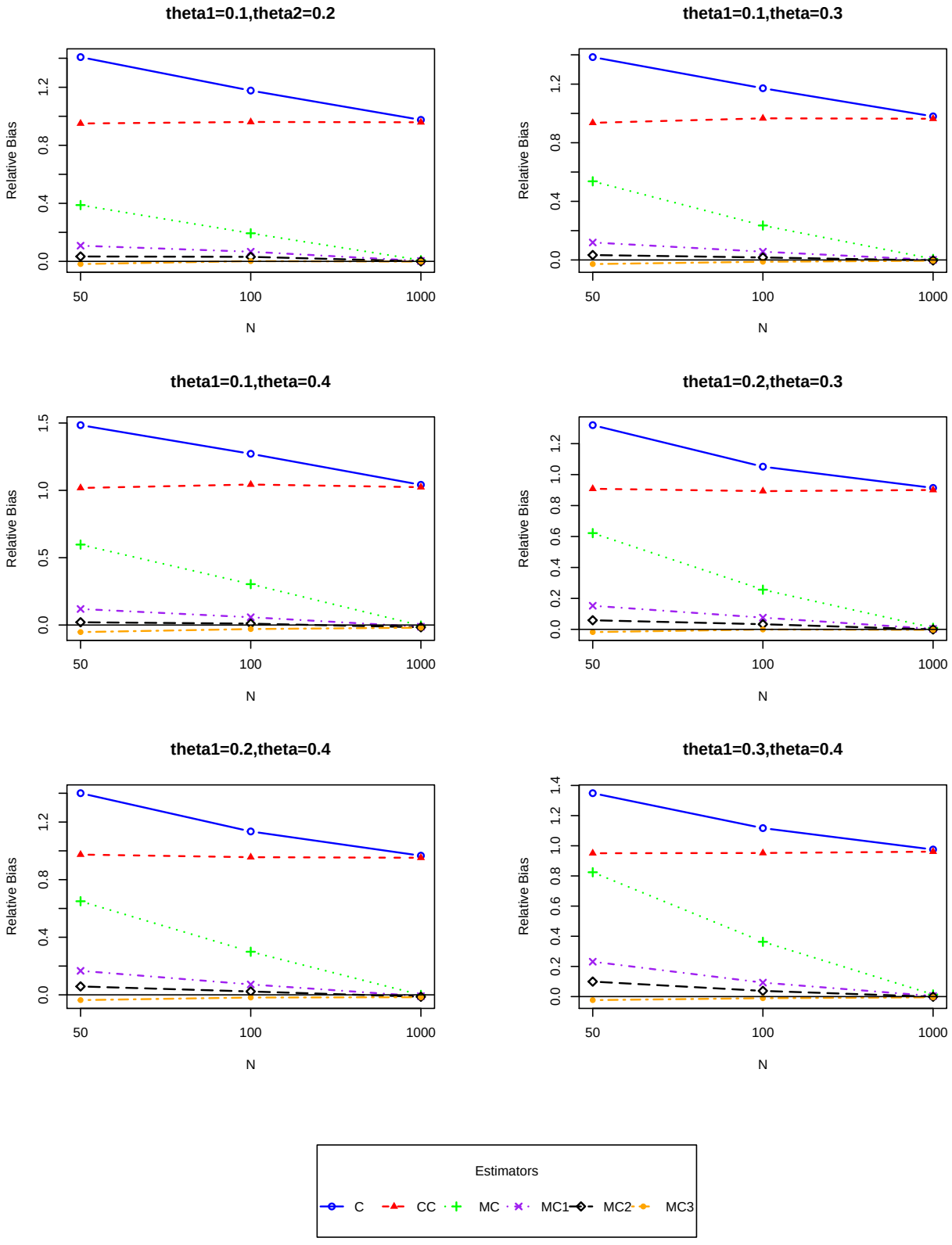

Figure 5.7: *RBias of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 20% one-inflation*

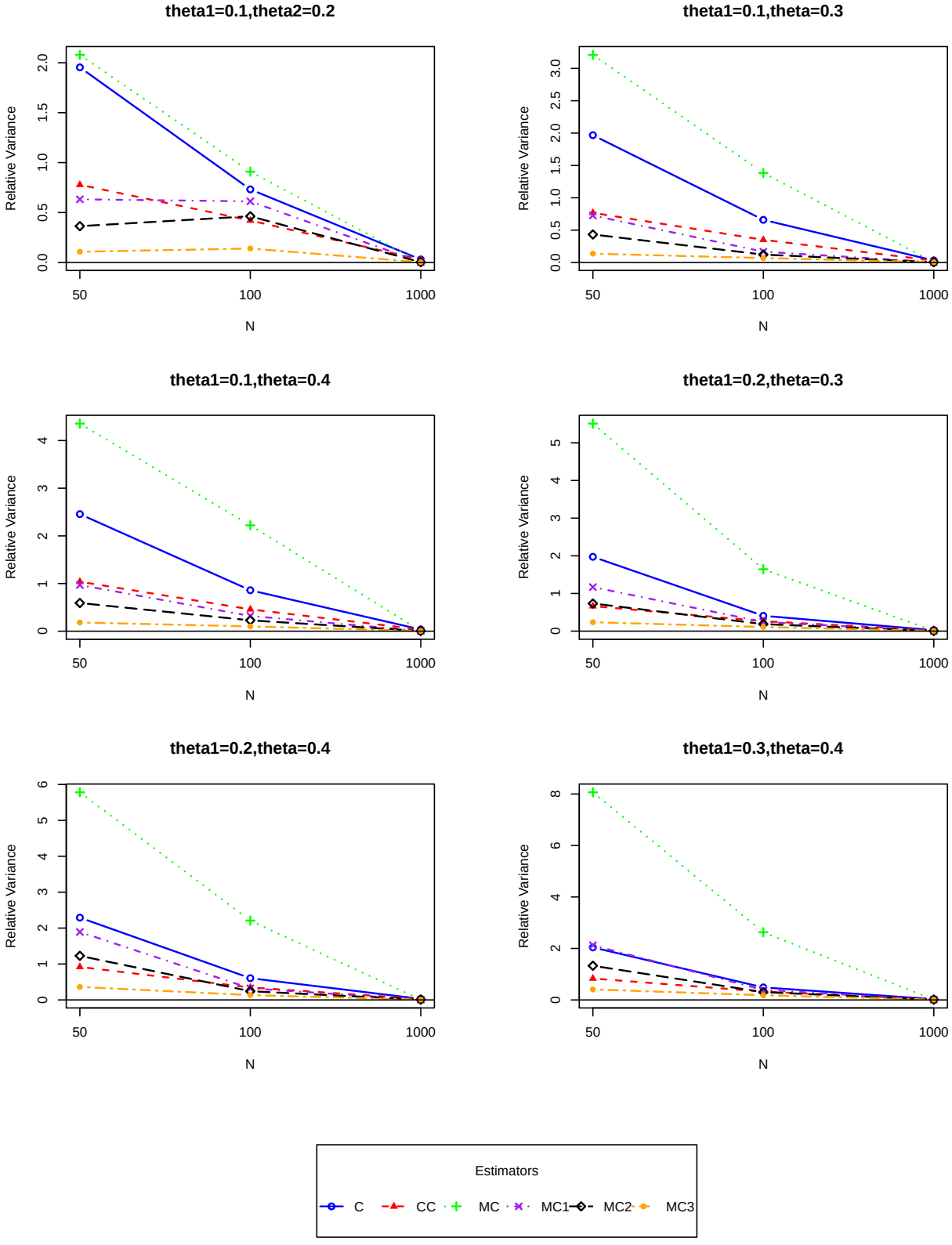

Figure 5.8: *RVar of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 20% one-inflation*

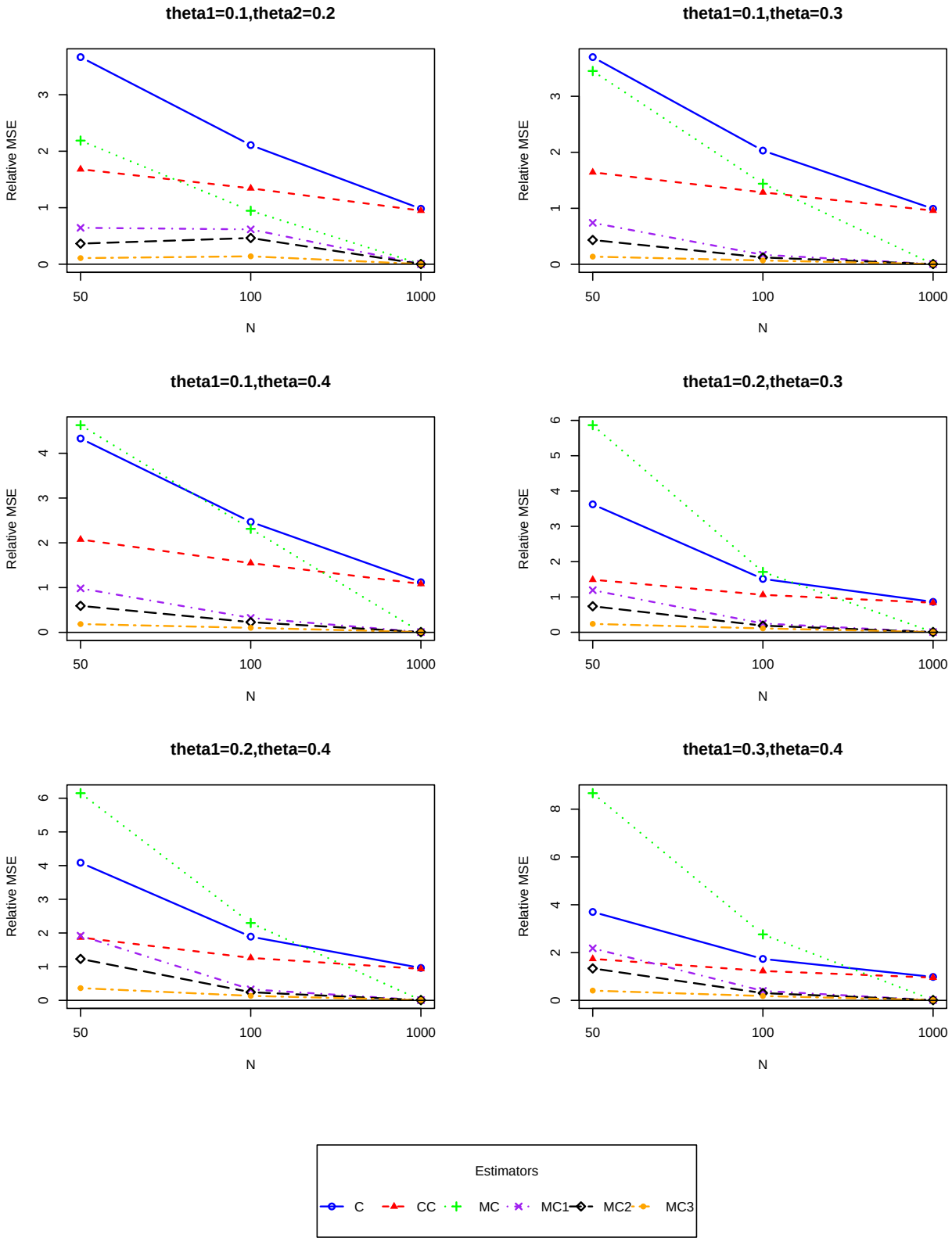

Figure 5.9: *RMSE of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 20% one-inflation*

Table 5.6: RBias, RVar and RMSE of six population size estimators under mixture of geometric model with 50% one-inflation

N	θ_1	θ_2	C	CC	MC	MC1	MC2	MC3
					Relative Bias			
50	0.1	0.2	7.6972	5.4523	0.2538	0.0593	**-0.0005**	-0.0298
		0.3	7.6606	5.4325	0.3011	0.0626	**-0.0065**	-0.0434
		0.4	7.7222	5.5449	0.3124	0.0679	**-0.0127**	-0.0589
	0.2	0.3	6.8114	4.7925	0.3733	0.0709	**-0.0047**	-0.0487
		0.4	7.1722	4.9878	0.3802	0.1132	**0.0161**	-0.0501
	0.3	0.4	6.9371	4.9361	0.4460	0.1037	**0.0018**	-0.0672
100	0.1	0.2	7.7507	5.9046	0.2244	0.0577	0.0209	**-0.0062**
		0.3	7.4630	5.7802	0.2700	0.0741	0.0299	**-0.0093**
		0.4	7.5967	5.8943	0.3230	0.0636	**0.0134**	-0.0261
	0.2	0.3	6.2668	5.0381	0.2662	0.0743	0.0287	**-0.0106**
		0.4	6.4426	5.1533	0.3363	0.0699	**0.0194**	-0.0238
	0.3	0.4	6.1307	5.0014	0.4137	0.0906	0.0327	**-0.0192**
1000	0.1	0.2	6.1167	5.9848	0.0083	0.0029	**0.0000**	-0.0018
		0.3	5.8251	5.7092	0.0089	0.0020	**-0.0015**	-0.0038
		0.4	6.0236	5.9022	0.0036	-0.0052	**-0.0093**	-0.0122
	0.2	0.3	5.0516	4.9668	0.0128	0.0054	0.0018	**-0.0008**
		0.4	5.2135	5.1248	0.0078	-0.0013	**-0.0055**	-0.0086
	0.3	0.4	5.0504	4.9693	0.0177	0.0068	0.0019	**-0.0019**
					Relative Variance			
50	0.1	0.2	34.6928	15.6734	0.8041	0.1831	0.0838	**0.0250**
		0.3	34.7046	15.5893	1.0302	0.2677	0.1303	**0.0387**
		0.4	35.6402	17.2157	1.3053	0.4075	0.2103	**0.0600**
	0.2	0.3	29.2484	11.9099	1.4363	0.3317	0.1692	**0.0525**
		0.4	34.7685	13.4350	1.5915	0.6744	0.3540	**0.0982**
	0.3	0.4	31.9114	13.6552	2.2137	0.7041	0.3785	**0.1082**
100	0.1	0.2	34.1997	12.8562	0.8330	0.1833	0.1134	**0.0371**
		0.3	33.0858	13.1726	1.3453	0.6165	0.4322	**0.1214**
		0.4	31.0391	12.1909	1.4695	0.2484	0.1543	**0.0541**
	0.2	0.3	19.0359	7.7209	1.3869	0.3974	0.2640	**0.0847**
		0.4	20.3165	7.2571	1.8407	0.3779	0.2532	**0.0876**
	0.3	0.4	18.8909	8.0653	2.6697	0.4650	0.3168	**0.1172**
1000	0.1	0.2	1.0273	0.9516	0.0021	0.0018	0.0017	**0.0016**
		0.3	0.8676	0.8100	0.0034	0.0030	0.0028	**0.0027**
		0.4	0.9534	0.8885	0.0050	0.0042	0.0040	**0.0038**
	0.2	0.3	0.5466	0.5166	0.0045	0.0040	0.0038	**0.0036**
		0.4	0.6145	0.5805	0.0061	0.0053	0.0051	**0.0048**
	0.3	0.4	0.5608	0.5317	0.0089	0.0078	0.0075	**0.0071**
					Relative Mean Square Error			
50	0.1	0.2	74.2285	45.3977	0.8065	0.1865	0.0838	**0.0259**
		0.3	75.1669	45.0979	1.0412	0.2715	0.1303	**0.0406**
		0.4	73.9131	47.9578	1.3117	0.4121	0.2105	**0.0634**
	0.2	0.3	66.5927	34.8755	1.4986	0.3366	0.1692	**0.0549**
		0.4	73.3605	38.3101	1.6388	0.6871	0.3542	**0.1007**
	0.3	0.4	68.3319	38.0179	2.2927	0.7147	0.3785	**0.1127**
100	0.1	0.2	92.0884	47.7181	0.8747	0.1866	0.1138	**0.0371**
		0.3	86.7972	46.5807	1.4074	0.6219	0.4330	**0.1215**
		0.4	86.1852	46.9309	1.5573	0.2524	0.1545	**0.0548**
	0.2	0.3	57.5793	33.1020	1.4523	0.4028	0.2647	**0.0848**
		0.4	61.2083	33.8122	1.9477	0.3827	0.2535	**0.0881**
	0.3	0.4	55.8452	33.0781	2.8320	0.4731	0.3178	**0.1176**
1000	0.1	0.2	38.4411	36.7695	0.0022	0.0018	0.0017	**0.0016**
		0.3	34.7996	33.4052	0.0035	0.0030	0.0028	**0.0027**
		0.4	37.2371	35.7243	0.0050	0.0043	0.0041	**0.0039**
	0.2	0.3	26.0655	25.1859	0.0046	0.0040	0.0038	**0.0036**
		0.4	27.7950	26.8440	0.0061	0.0053	0.0051	**0.0049**
	0.3	0.4	26.0671	25.2252	0.0092	0.0078	0.0075	**0.0071**

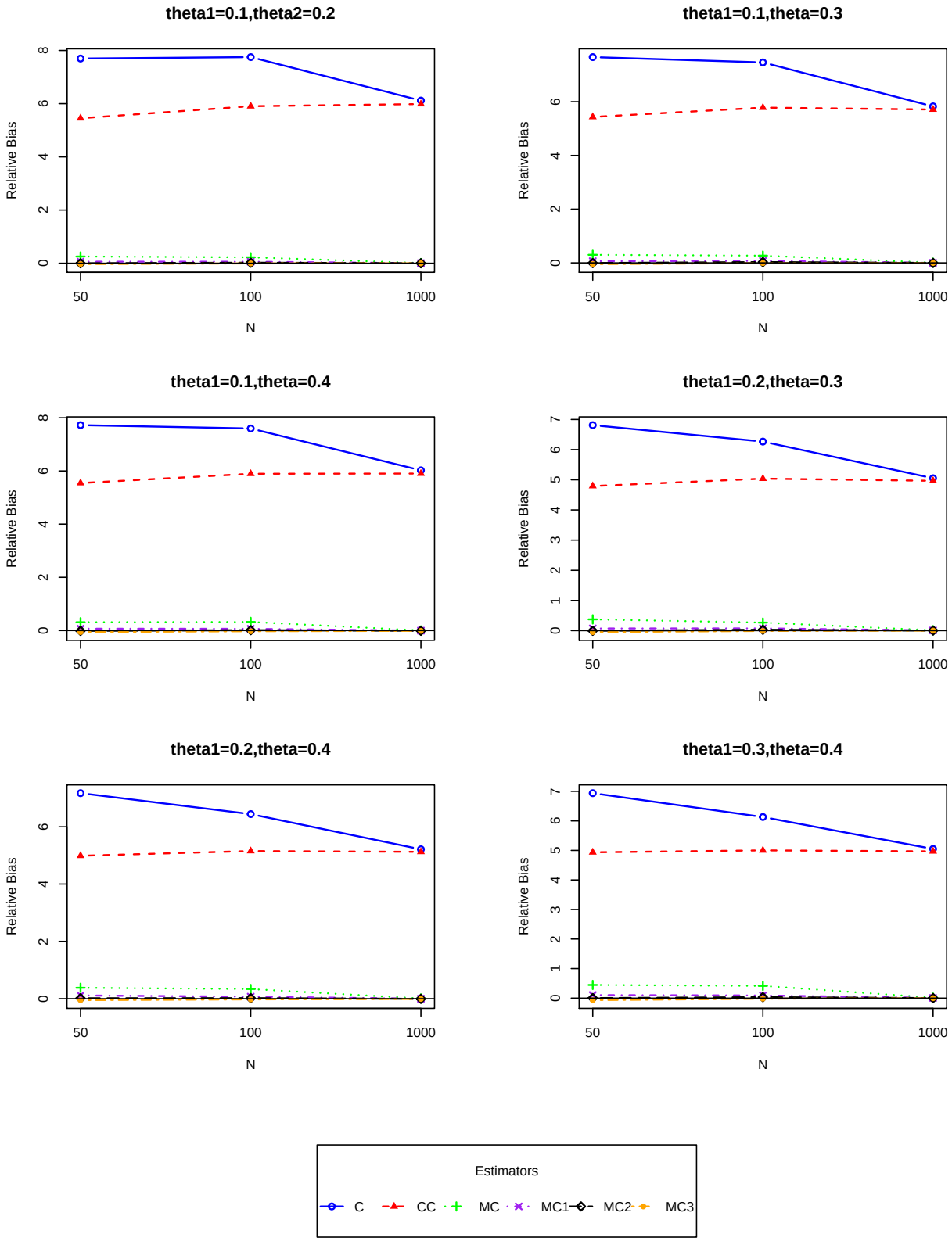

Figure 5.10: *RBias of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 50% one-inflation*

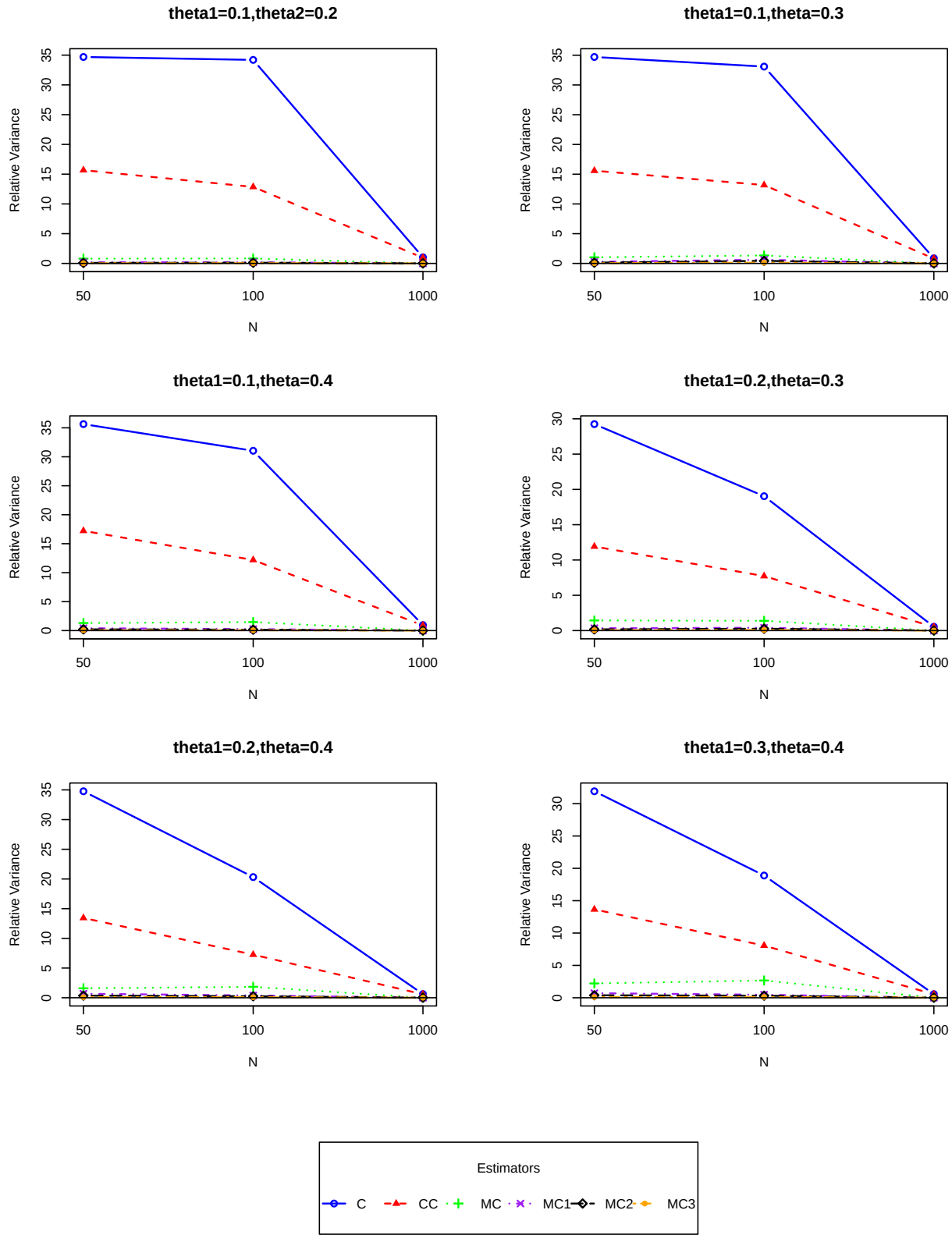

Figure 5.11: *RVar of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 50% one-inflation*

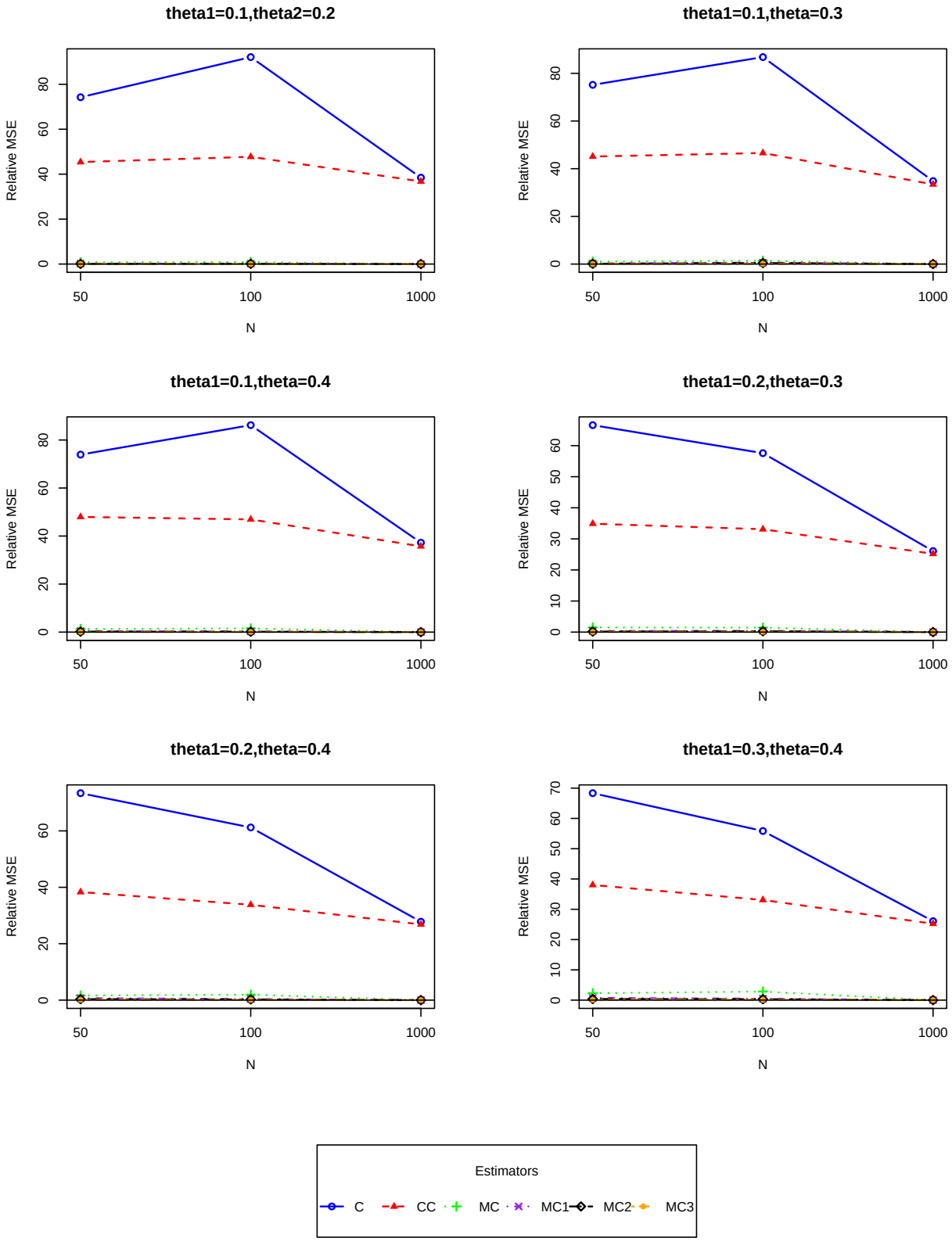

Figure 5.12: *RMSE of six estimators for counts drawn from mixture of geometric(θ_1, θ_2) with 50% one-inflation*

Table 5.7: Observed frequency distribution of the count of four applications

H5N1		Scrapie Infection		Domestic Violence		Illegal Immigrants	
x	f_x	x	f_x	x	f_x	x	f_x
0	6,587	1	121	1	15,169	1	1,645
1	410	2	13	2	1,957	2	183
2	161	3	5	3	393	3	37
3	87	4	2	4	99	4	13
4	46			5	28	5	1
5	26			6+	16	6	1
6	21						
7	8						
8	4						
9	6						

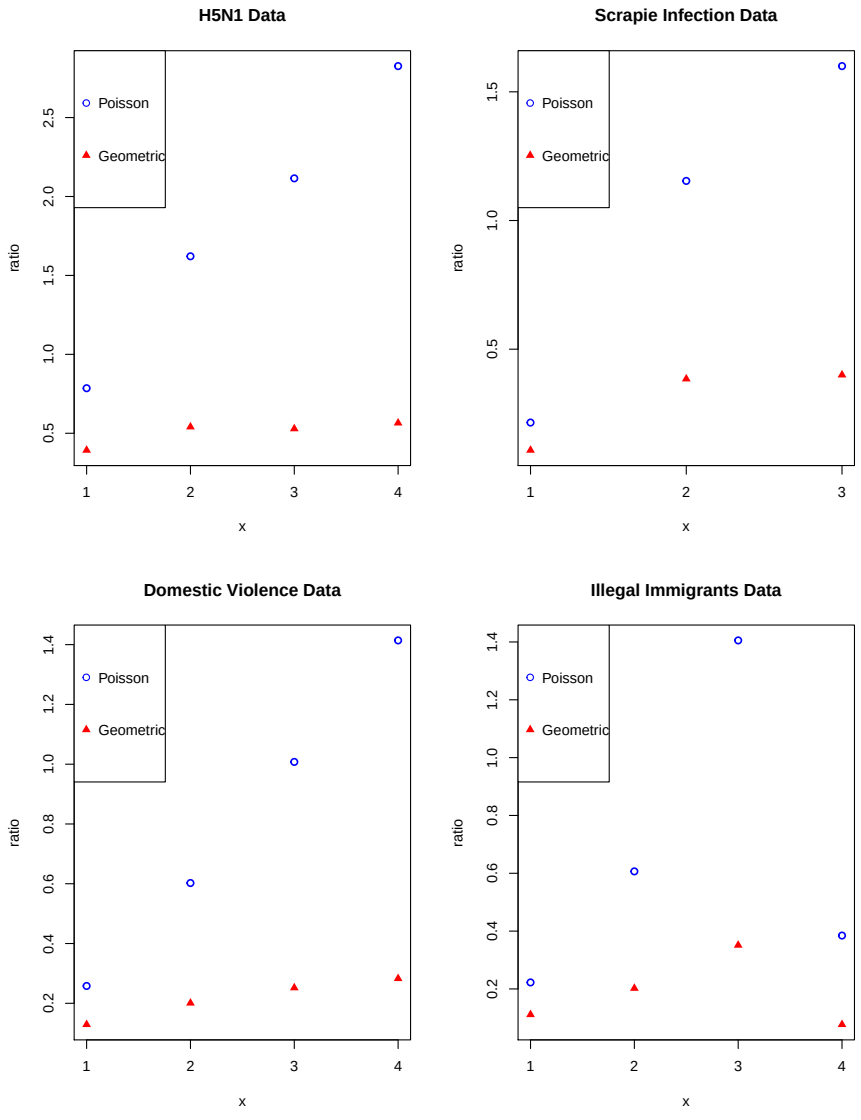

Figure 5.13: *Poisson and geometric ratio plot for real data examples*

Table 5.8: Population size estimation for four applications

Estimator	C	CC	MC	MC1	MC2	MC3
			H5N1			
$\hat{f}_0$	1,044	1,035	551	535	528	522
$\hat{N}$	1,813	1,804	1,320	1,304	1,297	1,291
			Scrapie Infection			
$\hat{f}_0$	1,126	1,037	88	56	48	41
$\hat{N}$	1,267	1,178	229	197	189	182
			Domestic Violence			
$\hat{f}_0$	117,577	117,509	48,527	48,257	48,207	48,085
$\hat{N}$	135,293	135,171	66,189	65,919	65,869	65,747
			Illegal Immigrants			
$\hat{f}_0$	14,787	14,698	4,477	4,221	4,175	4,068
$\hat{N}$	16,667	16,578	6,357	6,101	6,055	5,948

Chapter 6

Variance Estimation for Modified Chao Estimators

This chapter relates to the proposed estimators in Chapter 5. It provides two variance approximations for the modified Chao estimator and the modified Chao estimator with bias correction version 3.

6.1 Introduction

As it was shown in previous chapters, the crucial parameters in capture-recapture studies are N, f_0 and p_0 where N denotes the unknown population size, f_0 is the frequency of unobserved units and p_0 is the probability of not identifying a unit of the target population. It follows that $E(f_0) = Np_0$ and $\hat{f}_0 = \frac{np_0}{1-p_0}$, n is the number of observed units. The estimator of N is $\hat{N} = n/(1-p_0)$ or

$$\hat{N} = n + \hat{f}_0 \tag{6.1}$$

and depends on the approach used to estimate p_0. If we consider the variance of (6.1), it can be easily seen that there are two components of variation. The first variation is due to n and another variation is due to $\hat{f}_0$. The *total variance* of $\hat{N}$ can be calculated

as

$$\begin{aligned} var(\hat{N}) &= var(n + \hat{f}_0) \\ &= var(n + \frac{n\hat{p}_0}{1-\hat{p}_0}) \\ &= var(\frac{n}{1-\hat{p}_0}) \\ &= \left(\frac{1}{1-\hat{p}_0}\right)^2 var(n) \\ &= \left(\frac{1}{1-\hat{p}_0}\right)^2 Np_0(1-p_0) \end{aligned} \tag{6.2}$$

where $n \sim B(N, 1-p_0)$ and has variance $Np_0(1-p_0)$. However, it can be argued that there is nothing uncertain about n since it has been observed already. The interest should be regularly in the uncertainty attached to $\hat{f}_0$ as this quantity is unobserved and is predicted. The *partial variance* of prediction $\hat{f}_0$ can be easily computed as

$$var(\hat{f}_0) = \left(\frac{p_0}{1-p_0}\right)^2 var(n) = \left(\frac{p_0}{1-p_0}\right)^2 Np_0(1-p_0) \tag{6.3}$$

Note that the partial variance $var(\hat{f}_0)$ is smaller than the total variance $var(\hat{N})$ and they are related by $var(\hat{f}_0) = p_0^2 var(\hat{N})$. We find that the prediction variance is more appropriate than the total variance for capture-recapture experiments. Hence in order to assess the uncertainty of the proposed estimators in Chapter 5, the variance estimation for the modified Chao estimator is constructed by using the advantages of likelihood framework and considering only on a partial prediction variance $var(\hat{f}_0)$. This chapter focuses on $\hat{f}_0$ from the modified Chao estimator, $\hat{f}_{0(\mathrm{MC})} = \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}$, and the modified Chao estimator with bias correction version 3, $\hat{f}_{0(\mathrm{MC3})} = \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3+1)(f_3+2)}$, as it is the best version of bias reduction.

6.2 Likelihood framework

In this section a likelihood framework is developed in order to derive the variance of the modified Chao estimator. As the modified Chao estimator uses only frequencies with counts of twos and threes, we consider truncating all counts except counts of twos and threes. This truncated sample leads to a binomial log-likelihood

$$l = f_2 \log(p) + f_3 \log(1-p), \tag{6.4}$$

where $p = P(X = 2)$ and $1 - p = P(X = 3)$ which is uniquely maximized as

$$\frac{\partial l}{\partial p} = \frac{f_2}{p} - \frac{f_3}{1-p} = 0$$

$$\hat{p} = \frac{f_2}{f_2 + f_3}. \tag{6.5}$$

Let $E(f_x \mid f_2, f_3; p_x) = e_x$ for $x = 0, 1, 2, ..., m$ and we have

$$e_x = Np_x = (e_0 + e_1 + f_2 + f_3 + e_4^+)p_x \tag{6.6}$$

and

$$\begin{aligned} E(f_x \mid x \neq 2, 3) &= NP(X \neq 2, 3) \\ e_0 + e_1 + e_4^+ &= (e_0 + e_1 + f_2 + f_3 + e_4^+)(1 - p_2 - p_3) \\ &= [(e_0 + e_1 + e_4^+) + (f_2 + f_3)](1 - p_2 - p_3) \\ &= \frac{(1 - p_2 - p_3)}{p_2 + p_3}(f_2 + f_3). \end{aligned} \tag{6.7}$$

We know that $e_0 = (e_0 + e_1 + e_4^+ + f_2 + f_3)p_0$ from (6.6) and $e_0 + e_1 + e_4^+ = \frac{(1 - p_2 - p_3)}{p_2 + p_3}(f_2 + f_3)$ from (6.7) and finally e_0 can be obtained as

$$\begin{aligned} e_0 &= (e_0 + e_1 + e_4^+ + f_2 + f_3)p_0 \\ &= \left[\frac{(1 - p_2 - p_3)}{p_2 + p_3}(f_2 + f_3) + f_2 + f_3\right] p_0 \\ &= \frac{p_0}{(p_2 + p_3)}(f_2 + f_3). \end{aligned} \tag{6.8}$$

To develop this further we need to use power series $p_x = b_x\theta^x / g(\theta)$ only for $x = 0, 2, 3$ so

$$e_0 = \frac{b_0}{(b_2\theta^2 + b_3\theta^3)}(f_2 + f_3) \tag{6.9}$$

and replacing θ by its maximum likelihood estimator $(\hat{\theta})$ in (6.9):

$$\hat{e}_0 = \frac{b_0}{(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)}(f_2 + f_3). \tag{6.10}$$

Refer to the binomially-truncated log-likelihood in (6.4), $f_2 \log(p) + f_3 \log(1 - p)$, and maximum likelihood estimator of p in (6.5), $\hat{p} = \frac{f_2}{f_2 + f_3}$. Under the power series, $p_x = b_x\theta^x / g(\theta)$, we have

$$\begin{aligned} p &= \frac{Np_2}{Np_2 + Np_3} \\ &= \frac{p_2}{p_2 + p_3} \\ &= \frac{b_2\theta^2}{b_2\theta^2 + b_3\theta^3} \\ &= \frac{b_2}{b_2 + b_3\theta}. \end{aligned}$$

The invariance principle is used to find the estimator of θ as following:

$$\frac{f_2}{f_2+f_3} = \frac{b_2}{b_2+b_3\hat{\theta}}$$
$$\frac{b_2+b_3\hat{\theta}}{b_2} = \frac{f_2+f_3}{f_2}$$
$$\hat{\theta} = \frac{b_2}{b_3}\frac{f_3}{f_2}. \tag{6.11}$$

Replacing $\hat{\theta}$ from (6.11) in (6.10) and it follows that

$$\begin{aligned}\hat{e}_0 &= \frac{b_0}{b_2\frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^2} + b_3\frac{b_2^3}{b_3^2}\frac{f_3^3}{f_2^3}}(f_2+f_3)\\ &= \frac{b_0}{\frac{b_2^3}{b_3^2}\left(\frac{f_3^2}{f_2^2}+\frac{f_3^3}{f_2^3}\right)}(f_2+f_3)\\ &= \frac{b_0 b_3^2}{b_2^3}\frac{(f_2+f_3)f_2^3}{f_2f_3^2+f_3^3}\\ &= \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}.\end{aligned}$$

We can see that $\hat{e}_0 = \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}$ corresponds to the modified Chao estimator ($\hat{f}_{0(\mathrm{MC})}$) in Chapter 5.

6.3 Variance of the modified Chao estimator

The modified Chao estimator for f_0 can be written as

$$\hat{f}_{0(\mathrm{MC})} = \hat{e}_0 = \frac{b_0}{(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)}(f_2+f_3) = T(\hat{\theta})(f_2+f_3) \tag{6.12}$$

where $\hat{\theta} = \frac{b_2}{b_3}\frac{f_3}{f_2}$ and $T(\hat{\theta}) = \frac{b_0}{(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)}$.

Here the interest is in developing the variance of estimator $\hat{f}_{0(\mathrm{MC})}$ in (6.12) by mean of conditioning technique which has a general form for two random variables X and Y as follows:

$$var(X) = E[var(X \mid Y)] + var[E(X \mid Y)] \tag{6.13}$$

(see more details in Böhning (2008) and Van der Heijden et al. (2003a)).

We apply the concept in (6.13) by using $X = T(\hat{\theta})(f_2 + f_3)$ and $Y = f_2 + f_3$ so that we achieve

$$\begin{aligned} var(\hat{e}_0) &= var[T(\hat{\theta})(f_2 + f_3)] \\ &= E\left[var\left\{T(\hat{\theta})(f_2 + f_3) \mid (f_2 + f_3)\right\}\right] + var\left[E\left\{T(\hat{\theta})(f_2 + f_3) \mid (f_2 + f_3)\right\}\right] \end{aligned} \tag{6.14}$$

To solve the first term in (6.14), we assume that $E\left[var\left\{T(\hat{\theta})(f_2 + f_3) \mid (f_2 + f_3)\right\}\right] = var[T(\hat{\theta})(f_2 + f_3)]$. We have that

$$var[T(\hat{\theta})(f_2 + f_3)] = (f_2 + f_3)^2 var[T(\hat{\theta})]$$

The delta-method is applied here to deal with $var[T(\hat{\theta})]$. This leads to $var[T(\hat{\theta})] = T'(\hat{\theta})^2 var(\hat{\theta})$. Hence the first term in (6.14) can be written as

$$E\left[var\left\{T(\hat{\theta})(f_2 + f_3) \mid (f_2 + f_3)\right\}\right] = T'(\hat{\theta})^2 var(\hat{\theta})(f_2 + f_3)^2 \tag{6.15}$$

To consider the second term in (6.14), since $E[T(\hat{\theta})(f_2 + f_3) \mid (f_2 + f_3)] \approx T(\theta)(f_2 + f_3)$ so $var[T(\theta)(f_2 + f_3)]$ can be estimated as $T(\hat{\theta})^2(f_2 + f_3)$.

Finally, the partial prediction variance of modified Chao estimator can be derived from

$$var[T(\hat{\theta})(f_2 + f_3)] \approx \underbrace{T'(\hat{\theta})^2}_{(1)} \quad \underbrace{var(\hat{\theta})}_{(2)\text{use F-information}} \quad (f_2 + f_3)^2 + T(\hat{\theta})^2(f_2 + f_3). \tag{6.16}$$

Let us consider first term (1) in (6.16):

$$\begin{aligned} T(\hat{\theta}) &= \frac{b_0}{(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)} \\ T'(\hat{\theta}) &= -b_0(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)^{-2}(2b_2\hat{\theta} + 3b_3\hat{\theta}^2) \\ T'(\hat{\theta})^2 &= b_0^2(b_2\hat{\theta}^2 + b_3\hat{\theta}^3)^{-4}(2b_2\hat{\theta} + 3b_3\hat{\theta}^2)^2 \\ &= \frac{b_0^2\hat{\theta}^2}{\hat{\theta}^8}\frac{(2b_2 + 3b_3\hat{\theta})^2}{(b_2 + b_3\hat{\theta})^4} \end{aligned}$$

where $\hat{\theta} = \frac{b_2}{b_3}\frac{f_3}{f_2}$, hence

$$\begin{aligned} T'(\hat{\theta})^2 &= \frac{b_0^2 b_3^6}{b_2^6}\frac{f_2^6}{f_3^6}\frac{(2b_2 + 3b_2 f_3/f_2)^2}{(b_2 + b_2 f_3/f_2)^4} \\ &= \frac{b_0^2 b_3^6}{b_2^6}\frac{f_2^6}{f_3^6}\frac{(2f_2 + 3f_3)^2/f_2^2}{(f_2 + f_3)^4/f_2^4} \\ &= \frac{b_0^2 b_3^6}{b_2^8}\frac{f_2^8}{f_3^6}\frac{(2f_2 + 3f_3)^2}{(f_2 + f_3)^4}. \end{aligned}$$

Second term (2) in (6.16) can be derived by using the delta-method (see Bishop et al. (1975))as

$$var(\hat{\theta}) = var\left(\frac{b_2}{b_3}\frac{f_3}{f_2}\right)$$
$$\approx \bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix}^T cov\begin{pmatrix} f_2 \\ f_3 \end{pmatrix} \bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix}$$

where

$$\bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix}^T = \begin{pmatrix} -\frac{b_2}{b_3}\frac{f_3}{f_2^2} & \frac{b_2}{b_3}\frac{1}{f_2} \end{pmatrix}$$

$$cov\begin{pmatrix} f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} f_2\left(1-\frac{f_2}{n}\right) & -\frac{f_2 f_3}{n} \\ -\frac{f_2 f_3}{n} & f_3\left(1-\frac{f_3}{n}\right) \end{pmatrix}$$

and

$$\bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix}^T Cov\begin{pmatrix} f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} -\frac{b_2}{b_3}\frac{f_3}{f_2}\left(1-\frac{f_2}{n}\right) - \frac{b_2}{b_3}\frac{f_3}{n} & \frac{b_2}{b_3}\frac{f_3^2}{f_2 n} + \frac{b_2}{b_3}\frac{f_3}{f_2}\left(1-\frac{f_3}{n}\right) \end{pmatrix}$$

hence

$$\begin{aligned} var(\hat{\theta}) &\approx \bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix}^T cov\begin{pmatrix} f_2 \\ f_3 \end{pmatrix} \bigtriangledown g\begin{pmatrix} f_2 \\ f_3 \end{pmatrix} \\ &= \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^3}\left(1-\frac{f_2}{n}\right) + \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^2 n} + \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^2 n} + \frac{b_2^2}{b_3^2}\frac{f_3}{f_2^2}\left(1-\frac{f_3}{n}\right) \\ &= \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^3} - \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^2 n} + \frac{2b_2^2}{b_3^2}\frac{f_3^2}{f_2^2 n} + \frac{b_2^2}{b_3^2}\frac{f_3}{f_2^2} - \frac{b_2^2}{b_3^2}\frac{f_3^2}{f_2^2 n} \\ &= \frac{b_2^2}{b_3^2}\left(\frac{f_3^2}{f_2^3} + \frac{f_3}{f_2^2}\right) \\ &= \frac{b_2^2}{b_3^2}\frac{(f_2+f_3)f_3}{f_2^3} \end{aligned}$$

and finally

$$\begin{aligned} var[T(\hat{\theta})(f_2+f_3)] &\approx T'(\hat{\theta})^2 Var(\hat{\theta})(f_2+f_3)^2 + T(\hat{\theta})^2(f_2+f_3) \\ &= \frac{b_0^2 b_3^6}{b_2^8}\frac{f_2^8}{f_3^6}\frac{(2f_2+3f_3)^2}{(f_2+f_3)^4}\frac{b_2^2}{b_3^2}\frac{(f_2+f_3)^3 f_3}{f_2^3} + \frac{b_0^2 b_3^4}{b_2^6}\frac{f_2^6}{f_3^4(f_2+f_3)^2}(f_2+f_3) \\ &= \frac{b_0^2 b_3^4}{b_2^6}\frac{f_2^5}{f_3^5}\frac{(2f_2+3f_3)^2}{(f_2+f_3)} + \frac{b_0^2 b_3^4}{b_2^6}\frac{f_2^6}{f_3^4}\frac{1}{(f_2+f_3)} \\ &= \frac{b_0^2 b_3^4}{b_2^6}\frac{f_2^6}{f_3^4}\frac{1}{(f_2+f_3)}\left\{1 + \frac{(2f_2+3f_3)^2}{f_2 f_3}\right\} \end{aligned} \tag{6.17}$$

It can be seen from (6.17) that $\frac{b_0^2 b_3^4}{b_2^6}\frac{f_2^6}{f_3^4} = \hat{f}_{0(\mathrm{MC})}^2$ as we know from Chapter 5 that $\hat{f}_{0(\mathrm{MC})} = \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}$, hence, the prediction variance of modified Chao estimator (V1) is

$$\begin{aligned} \hat{var}_1(\hat{f}_0) &= \left\{\frac{b_0 b_3^2}{b_2^3}\frac{f_2^3}{f_3^2}\right\}^2 \frac{1}{(f_2+f_3)}\left\{1+\frac{(2f_2+3f_3)^2}{f_2 f_3}\right\} \\ &= \frac{\hat{f}_{0(\mathrm{MC})}^2}{f_2+f_3}\left\{1+\frac{(2f_2+3f_3)^2}{f_2 f_3}\right\}. \end{aligned} \tag{6.18}$$

The prediction variance for the modified Chao estimator can be obtained in another version by replacing $\hat{f}_{0(\mathrm{MC})}$ in (6.18) by the modified Chao estimator with bias correction version 3, $\hat{f}_{0(\mathrm{MC3})}$, where $\hat{f}_{0(\mathrm{MC3})} = \frac{b_0 b_3^2}{b_2^3}\frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3+1)(f_3+2)}$ and replacing $\frac{1}{f_2 f_3}$ by $\frac{1}{(f_2+1)(f_3+1)}$. Therefore, the second version of prediction variance for the modified Chao estimator (V2) is

$$\begin{aligned} \hat{var}_2(\hat{f}_0) &= \frac{\hat{f}_{0(\mathrm{MC3})}^2}{f_2+f_3}\left\{1+\frac{(2f_2+3f_3)^2}{(f_2+1)(f_3+1)}\right\} \\ &= \left\{\frac{b_0 b_3^2}{b_2^3}\frac{f_2^3-3f_2^2+2f_2}{(f_3+1)(f_3+2)}\right\}^2 \frac{1}{(f_2+f_3)}\left\{1+\frac{(2f_2+3f_3)^2}{(f_2+1)(f_3+1)}\right\}. \end{aligned} \tag{6.19}$$

Note that a 95% prediction interval based upon prediction variance can be calculated as $\hat{f}_0 \pm 1.96\sqrt{var(\hat{f}_0)}$ for f_0 and as $n + \hat{f}_0 \pm 1.96\sqrt{var(\hat{f}_0)}$ for N.

6.4 Simulation study

To explore the performance of two versions of variance estimation (V1 and V2) for modified Chao estimators (MC and MC3), the simulation study is designed to cover different models, geometric distribution and mixture of geometric distribution with and without one-inflation, with population size $N = 50, 100, 1,000$. Each simulation scenario is repeated 1,000 times ($B = 1,000$). Therefore, the Monte Carlo variance for two proposed estimator is given by

$$var(\hat{f}_{0(\mathrm{MC})})_{\mathrm{True}} = \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{f}_{0(\mathrm{MC})}^{(b)} - E(\hat{f}_{0(\mathrm{MC})})\right\}^2 \tag{6.20}$$

$$var(\hat{f}_{0(\mathrm{MC3})})_{\mathrm{True}} = \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{f}_{0(\mathrm{MC3})}^{(b)} - E(\hat{f}_{0(\mathrm{MC3})})\right\}^2 \tag{6.21}$$

where

$$\hat{f}_{0(\mathrm{MC})} = \frac{f_2^3}{f_3^2},$$

$$\hat{f}_{0(\mathrm{MC3})} = \frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3+1)(f_3+2)}$$

and

$$E(\hat{f}_{0(\mathrm{MC})}) = \frac{\sum_{b=1}^{B} \hat{f}_{0(\mathrm{MC})}^{(b)}}{B} \quad , \quad E(\hat{f}_{0(\mathrm{MC3})}) = \frac{\sum_{b=1}^{B} \hat{f}_{0(\mathrm{MC3})}^{(b)}}{B}.$$

The true standard error is simply computed as $s.e.(\hat{f}_0)_{\mathrm{True}} = \sqrt{var(\hat{f}_0)_{\mathrm{True}}}$.

To evaluate the behaviour and performance of two variance estimations in (6.18) and (6.19), the expected value of the approximated standard error $E\left\{s.e.(\hat{f}_0)\right\}$ can be calculated by

$$E\left\{\hat{s.e.}_1(\hat{f}_0)\right\} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \hat{var}_1(\hat{f}_0)} \tag{6.22}$$

$$E\left\{\hat{s.e.}_2(\hat{f}_0)\right\} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \hat{var}_2(\hat{f}_0)} \tag{6.23}$$

where

$$\hat{var}_1(\hat{f}_0) = \frac{(f_2^3/f_3^2)^2}{f_2+f_3}\left\{1 + \frac{(2f_2+3f_3)^2}{f_2 f_3}\right\}$$

and

$$\hat{var}_2(\hat{f}_0) = \left\{\frac{f_2^3 - 3f_2^2 + 2f_2}{(f_3+1)(f_3+2)}\right\}^2 \frac{1}{(f_2+f_3)}\left\{1 + \frac{(2f_2+3f_3)^2}{(f_2+1)(f_3+1)}\right\}.$$

Note that $\frac{b_0 b_3^2}{b_2^3}$ is equal to 1 for geometric distribution (see more detail in section 5.3).

The ratio of standard error of estimation $\frac{E[\hat{s.e.}(\hat{f}_0)]}{s.e.(\hat{f}_0)_{\mathrm{True}}}$ is provided for comparing the performance of two versions of variance estimation (V1 and V2) which are considered for two estimators (MC and MC3). R1 and R2 are the ratios of the approximated standard error from V1 and V2 to the true standard error from modified Chao estimator (MC), respectively, whereas R3 and R4 are the ratios of the approximated standard error from V1 and V2 to the true standard error from modified Chao estimator with bias correction version 3 (MC3), respectively. The comparison can simply be seen as following:

- Comparing $\hat{var}_1(\hat{f}_0)$ and $\hat{var}_2(\hat{f}_0)$ for MC

$$R1 = \frac{E[\hat{s.e.}_1(\hat{f}_0)]}{s.e.(\hat{f}_{0(\mathrm{MC})})_{\mathrm{True}}}$$

$$R2 = \frac{E[\hat{s.e.}_2(\hat{f}_0)]}{s.e.(\hat{f}_{0(\mathrm{MC})})_{\mathrm{True}}}$$

- Comparing $\hat{var}_1(\hat{f}_0)$ and $\hat{var}_2(\hat{f}_0)$ for MC3

$$R3 = \frac{E[\hat{se}_{.1}(\hat{f}_0)]}{s.e.(\hat{f}_{0(\text{MC3})})_{\text{True}}}$$

$$R4 = \frac{E[\hat{se}_{.2}(\hat{f}_0)]}{s.e.(\hat{f}_{0(\text{MC3})})_{\text{True}}}$$

The reference value for these ratios is equal to one. The more ratio is close to 1, the more estimation is close to the real value.

6.5 Simulation results

Table 6.1 and 6.2 provide the standard errors and the ratio of standard errors R1-R4 from 1,000 repeated simulation samples. Figure 6.1-6.4 illustrate the graphs of the ratio of standard errors where blue (R1) and red (R2) lines are used for comparing V1 and V2 for MC whereas green (R3) and purple (R4) lines are used for comparing V1 and V2 for MC3. The data are generated under the geometric and the mixture of geometric models with and without one-inflation. Overall, it can be seen from the simulation results that V2 performs the best with the ratio of standard errors close to one for all conditions especially the ratio of standard error from V2 to MC3 (R4). For more details, V2 gives an underestimation of the standard error for MC but gives slightly overestimation for MC3. On the other hand, V1 gives an overestimation of the standard error for both MC and MC3 particularly, severe overestimation for MC3 when population sizes are small (N=50, 100). There is no surprise that V1 is larger than V2 all cases of study due to V2 is derived from the bias correction version 3 for estimating the variance. However, the results also show that V1 and V2 are identical when the size of population are large and they are close to the true variance (MC and MC3 are also identical for large population size) as we can see that all ratios converge to one (see Figure 6.1-6.4). As a result, it is reasonable to state that the variance approximation V2 can be utilized to represent the true variance of both MC and MC3 for the geometric and the mixture of geometric models with and without one-inflation while the variance approximation V1 can be applied to stand for the true variance of MC and MC3 only the cases of large population sizes.

6.6 Conclusion

To determine the efficiency of an estimator in capture-recapture study, accuracy and precision are considered. Accuracy, is provided as the bias of estimator, refers to the closeness of an estimate to the true value. The estimator which is close to the parameter

would be more accurate than other estimators which provide a larger different value. The term precision, is defined as a variance of estimator, refers to the degree of variation for a series of estimates. The estimator which provides a small variation shows higher performance of estimation in terms of precision. In other words, the most effective estimator is the one among all possible estimators which has minimal bias and variance.

The modified Chao estimator (MC) is an asymptotically unbiased estimator and the modified Chao estimator with bias correction version 3 (MC3) is the best version of reducing bias when a sample size is small for a power series distribution with and without one-inflation as presented in Chapter 5. This chapter examines an approximation variance of MC and MC3. Variance estimators are simply derived by the conditioning method. It is clearly seen that the variation of both modified Chao estimators arise from two sources; the random variation of sampling n individuals from population and the variation from the predicted estimate $\hat{f}_0$. Here we focus on the partial variance of prediction $var(\hat{f}_0)$ as it has nothing uncertain about observed n. Variance of the proposed estimators, V1 or $\hat{var}_1(\hat{f}_0)$ and V2 or $\hat{var}_2(\hat{f}_0)$, are given in (6.18) and (6.19) respectively. The simulation study shows that V2 has the best performance for estimating the variance of MC and MC3 estimators on average as it provides the closest values to true variances. Although V1 give a severe overestimates for small sample size, it has a good performance when the sample size is large as it can be seen that the estimates of V1 and V2 are slightly different and they are close to true variances. Therefore, it can be sensibly stated that the variance estimation V2 in (6.19) represent well the true variance of MC and MC3 whereas the variance estimation V1 in (6.18) can stand for the true variance of MC and MC3 only in the case of large sample size.

Table 6.1: Comparison of the standard errors of two formulas with the true standard error of the modified Chao (MC) and the modified Chao with bias correction version 3 (MC3) under the geometric model and the geometric model with 20% and 50% one-inflation

N	θ	$s.e.(\hat{f}_0)_{\text{True}}$		$E[\hat{s.e.}(\hat{f}_0)]$		MC		MC3	
		MC	MC3	V1	V2	R1	R2	R3	R4
The geometric model									
50	0.1	64.827	8.738	150.376	13.517	2.320	**0.209**	17.210	**1.547**
	0.2	74.522	14.341	161.638	21.590	2.169	**0.290**	11.271	**1.505**
	0.3	199.850	30.545	434.860	46.571	2.176	**0.233**	14.237	**1.525**
	0.4	162.530	26.444	368.764	41.787	2.269	**0.257**	13.945	**1.580**
100	0.1	57.413	17.852	107.226	23.197	1.868	**0.404**	6.006	**1.299**
	0.2	47.626	22.329	69.334	29.388	**1.456**	0.617	3.105	**1.316**
	0.3	78.058	37.850	105.909	46.676	**1.357**	0.598	2.798	**1.233**
	0.4	241.840	60.654	465.971	83.738	1.927	**0.346**	7.682	**1.381**
1000	0.1	44.605	41.234	47.182	43.044	1.058	**0.965**	1.144	**1.044**
	0.2	72.657	68.776	73.182	68.779	**1.007**	0.947	1.064	**1.000**
	0.3	105.920	100.438	106.509	100.448	**1.006**	0.948	1.060	**1.000**
	0.4	139.395	131.084	147.045	137.632	1.055	**0.987**	1.122	**1.050**
The geometric model with 20% one-inflation									
50	0.1	47.452	5.963	114.530	9.305	2.414	**0.196**	19.207	**1.561**
	0.2	103.091	13.595	238.158	21.506	2.310	**0.209**	17.517	**1.582**
	0.3	136.309	19.853	308.888	31.031	2.266	**0.228**	15.559	**1.563**
	0.4	149.598	19.998	353.345	32.676	2.362	**0.218**	17.669	**1.634**
100	0.1	82.623	17.314	169.614	24.527	2.053	**0.297**	9.796	**1.417**
	0.2	57.640	20.052	100.371	27.673	1.741	**0.480**	5.005	**1.380**
	0.3	116.448	39.281	184.743	52.074	1.586	**0.447**	4.703	**1.326**
	0.4	132.047	38.704	241.790	54.949	1.831	**0.416**	6.247	**1.420**
1000	0.1	40.671	36.869	43.877	39.114	1.079	**0.962**	1.190	**1.061**
	0.2	70.172	65.453	68.415	63.280	**0.975**	0.902	1.045	**0.967**
	0.3	95.235	89.011	96.906	90.009	**1.018**	0.945	1.089	**1.011**
	0.4	133.457	123.463	137.098	126.111	**1.027**	0.945	1.110	**1.021**
The geometric model with 50% one-inflation									
50	0.1	33.435	3.365	84.004	5.373	2.512	**0.161**	24.963	**1.597**
	0.2	58.237	6.849	141.397	10.914	2.428	**0.187**	20.644	**1.593**
	0.3	76.461	8.489	190.984	14.160	2.498	**0.185**	22.498	**1.668**
	0.4	195.415	29.327	436.201	45.719	2.232	**0.234**	14.874	**1.559**
100	0.1	51.169	6.983	122.652	11.175	2.397	**0.218**	17.564	**1.600**
	0.2	116.431	18.932	256.318	28.522	2.201	**0.245**	13.539	**1.507**
	0.3	151.591	24.544	333.705	37.568	2.201	**0.248**	13.596	**1.531**
	0.4	212.765	33.200	472.461	51.474	2.221	**0.242**	14.231	**1.550**
1000	0.1	36.633	30.994	40.141	33.182	1.096	**0.906**	1.295	**1.071**
	0.2	55.353	49.523	57.691	50.867	**1.042**	0.919	1.165	**1.027**
	0.3	79.672	71.429	83.545	74.092	**1.049**	0.930	1.170	**1.037**
	0.4	115.963	101.869	118.971	103.684	**1.026**	0.894	1.168	**1.018**

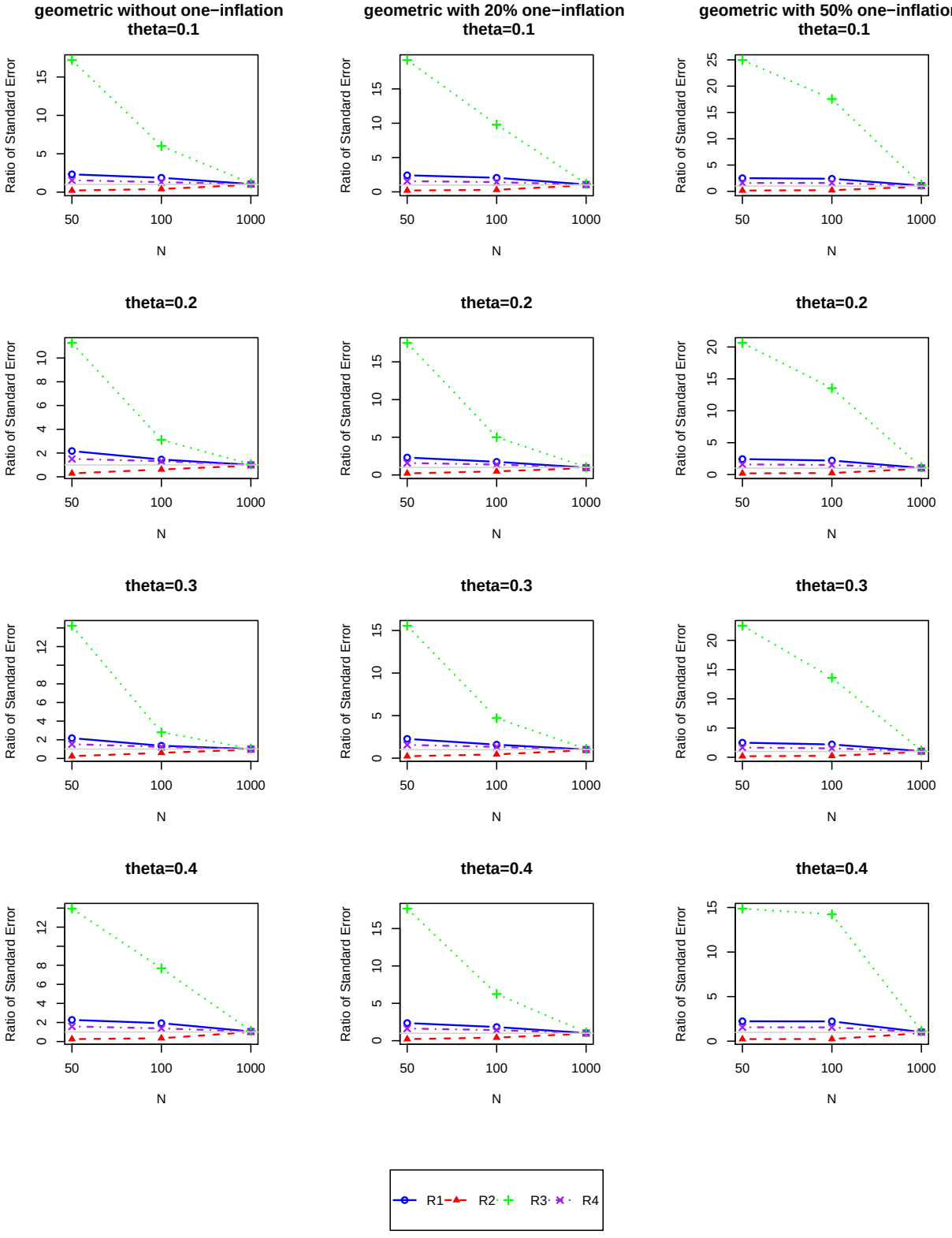

Figure 6.1: *Ratio of standard errors from two formulas (V1 and V2) to the true standard errors of MC and MC3 when data are generated from the geometric(θ) with and without one-inflation*

Table 6.2: Comparison of the standard errors of two formulas with the true standard error of the modified Chao (MC) and the modified Chao with bias correction version 3 (MC3) under the mixture of geometric model and the mixture of geometric model with 20% and 50% one-inflation

N	θ_1	θ_2	$s.e.(\hat{f}_0)_{\text{True}}$		$E[\hat{s.e.}(\hat{f}_0)]$		MC		MC3	
			MC	MC3	V1	V2	R1	R2	R3	R4
The mixture of geometric model										
50	0.1	0.2	100.032	13.049	229.203	20.751	2.291	**0.207**	17.565	**1.590**
		0.3	116.737	16.870	262.047	26.255	2.245	**0.225**	15.533	**1.556**
		0.4	90.156	13.730	210.868	21.722	2.339	**0.241**	15.359	**1.582**
	0.2	0.3	130.149	20.195	283.869	30.822	2.181	**0.237**	14.056	**1.526**
		0.4	232.759	33.972	512.965	52.022	2.204	**0.224**	15.099	**1.531**
	0.3	0.4	208.168	30.522	459.253	47.376	2.206	**0.228**	15.047	**1.552**
100	0.1	0.2	58.141	21.444	92.811	28.743	1.596	**0.494**	4.328	**1.340**
		0.3	108.749	35.927	173.895	47.283	1.599	**0.435**	4.840	**1.316**
		0.4	175.713	38.650	356.337	53.336	2.028	**0.304**	9.219	**1.380**
	0.2	0.3	168.506	38.522	343.785	51.062	2.040	**0.303**	8.924	**1.326**
		0.4	116.138	36.520	222.276	48.154	1.914	**0.415**	6.086	**1.319**
	0.3	0.4	77.939	38.535	112.229	49.549	1.440	**0.636**	2.912	**1.286**
1000	0.1	0.2	61.270	57.453	61.413	57.031	**1.002**	0.931	1.069	**0.993**
		0.3	69.772	65.467	74.751	69.605	1.071	**0.998**	1.142	**1.063**
		0.4	90.335	84.220	93.053	86.172	**1.030**	0.954	1.105	**1.023**
	0.2	0.3	90.528	85.757	89.096	83.891	**0.984**	0.927	1.039	**0.978**
		0.4	105.727	99.770	104.243	97.798	**0.986**	0.925	1.045	**0.980**
	0.3	0.4	119.845	113.350	123.975	116.546	1.034	**0.972**	1.094	**1.028**
The mixture of geometric model with 20% one inflation										
50	0.1	0.2	73.133	9.069	173.524	14.528	2.373	**0.199**	19.134	**1.602**
		0.3	77.976	11.162	182.095	17.415	2.335	**0.223**	16.314	**1.560**
		0.4	80.540	9.912	195.570	16.335	2.428	**0.203**	19.730	**1.648**
	0.2	0.3	163.569	22.641	366.425	34.858	2.240	**0.213**	16.184	**1.540**
		0.4	128.509	16.830	302.899	27.319	2.357	**0.213**	17.997	**1.623**
	0.3	0.4	127.299	17.990	297.321	29.150	2.336	**0.229**	16.527	**1.620**
100	0.1	0.2	170.650	27.485	365.677	40.456	2.143	**0.237**	13.305	**1.472**
		0.3	138.664	28.634	285.837	41.100	2.061	**0.296**	9.983	**1.435**
		0.4	137.594	27.055	293.109	40.748	2.130	**0.296**	10.834	**1.506**
	0.2	0.3	74.339	27.280	121.110	37.644	1.629	**0.506**	4.440	**1.380**
		0.4	184.055	39.756	380.738	56.244	2.069	**0.306**	9.577	**1.415**
	0.3	0.4	256.494	49.405	533.563	70.400	2.080	**0.274**	10.800	**1.425**
1000	0.1	0.2	54.436	50.171	57.697	52.580	1.060	**0.966**	1.150	**1.048**
		0.3	68.324	63.119	71.112	65.023	**1.041**	0.952	1.127	**1.030**
		0.4	83.883	77.021	85.667	77.837	**1.021**	0.928	1.112	**1.011**
	0.2	0.3	80.704	75.385	80.815	74.907	**1.001**	0.928	1.072	**0.994**
		0.4	89.588	83.421	94.545	87.290	1.055	**0.974**	1.133	**1.046**
	0.3	0.4	104.159	97.014	111.465	103.125	1.070	**0.990**	1.149	**1.063**
The mixture of geometric model with 50% one inflation										
50	0.1	0.2	38.490	3.970	97.554	6.489	2.535	**0.169**	24.576	**1.635**
		0.3	55.612	6.383	136.451	10.131	2.454	**0.182**	21.377	**1.587**
		0.4	48.758	5.326	124.124	8.614	2.546	**0.177**	23.303	**1.617**
	0.2	0.3	65.614	7.874	159.964	12.557	2.438	**0.191**	20.315	**1.595**
		0.4	57.531	6.630	143.518	10.678	2.495	**0.186**	21.646	**1.610**
	0.3	0.4	79.367	9.051	197.076	14.855	2.483	**0.187**	21.775	**1.641**
100	0.1	0.2	120.411	16.428	270.033	25.281	2.243	**0.210**	16.437	**1.539**
		0.3	99.384	16.855	222.252	25.314	2.236	**0.255**	13.186	**1.502**
		0.4	92.761	13.456	215.555	21.370	2.324	**0.230**	16.020	**1.588**
	0.2	0.3	155.439	23.073	343.368	35.457	2.209	**0.228**	14.882	**1.537**
		0.4	112.252	20.952	247.096	31.566	2.201	**0.281**	11.794	**1.507**
	0.3	0.4	176.719	27.999	388.222	42.705	2.197	**0.242**	13.865	**1.525**
1000	0.1	0.2	48.444	42.466	50.710	43.606	**1.047**	0.900	1.194	**1.027**
		0.3	61.014	53.237	61.523	53.006	**1.008**	0.869	1.156	**0.996**
		0.4	69.385	60.063	75.067	64.052	1.082	**0.923**	1.250	**1.066**
	0.2	0.3	64.319	57.619	69.113	61.166	1.075	**0.951**	1.199	**1.062**
		0.4	71.192	63.463	79.060	69.492	1.111	**0.976**	1.246	**1.095**
	0.3	0.4	89.397	79.680	95.995	84.595	1.074	**0.946**	1.205	**1.062**

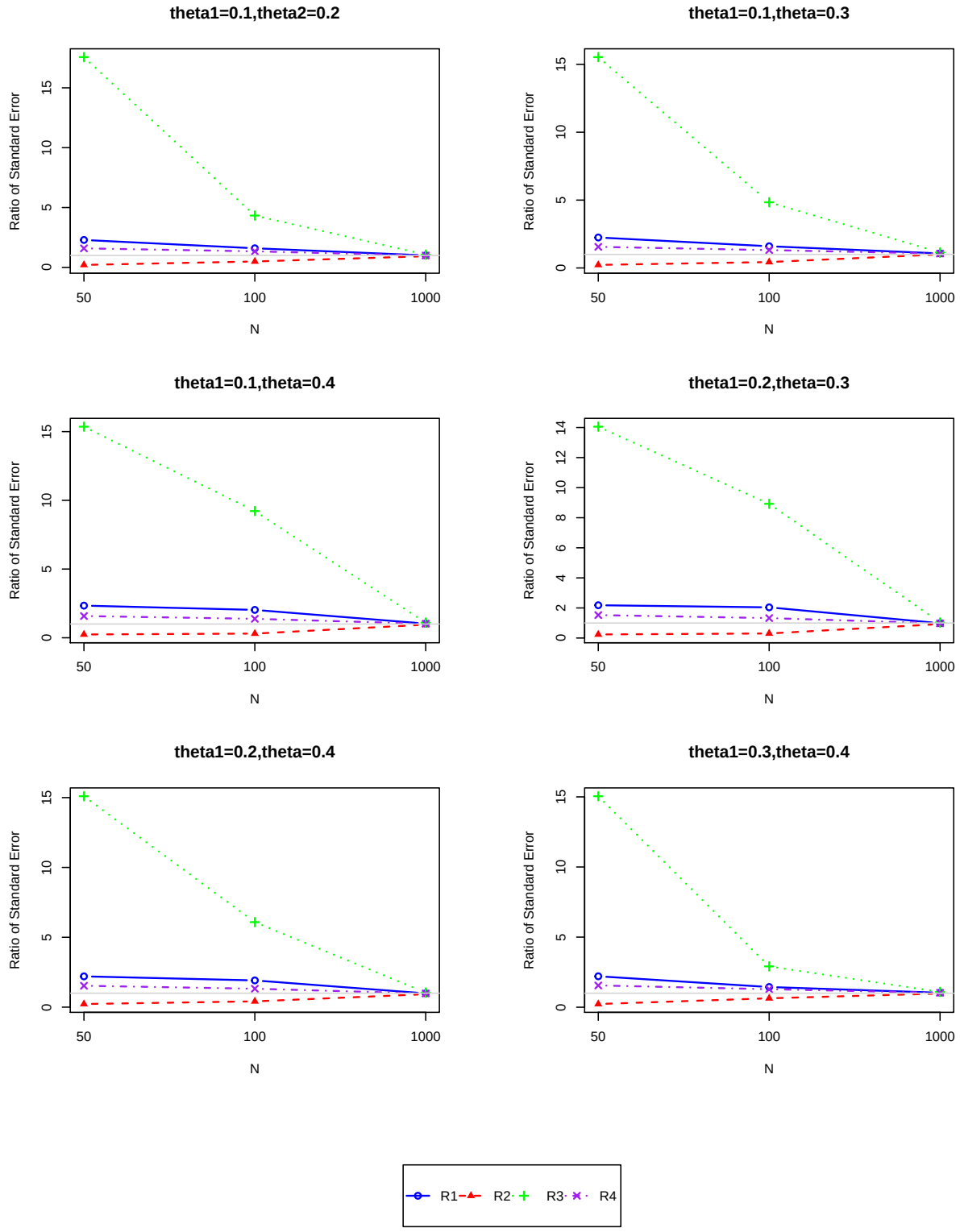

Figure 6.2: *Ratio of standard errors from two formulas (V1 and V2) to the true standard errors of MC and MC3 when data are generated from the mixture of geometric*(θ_1, θ_2)

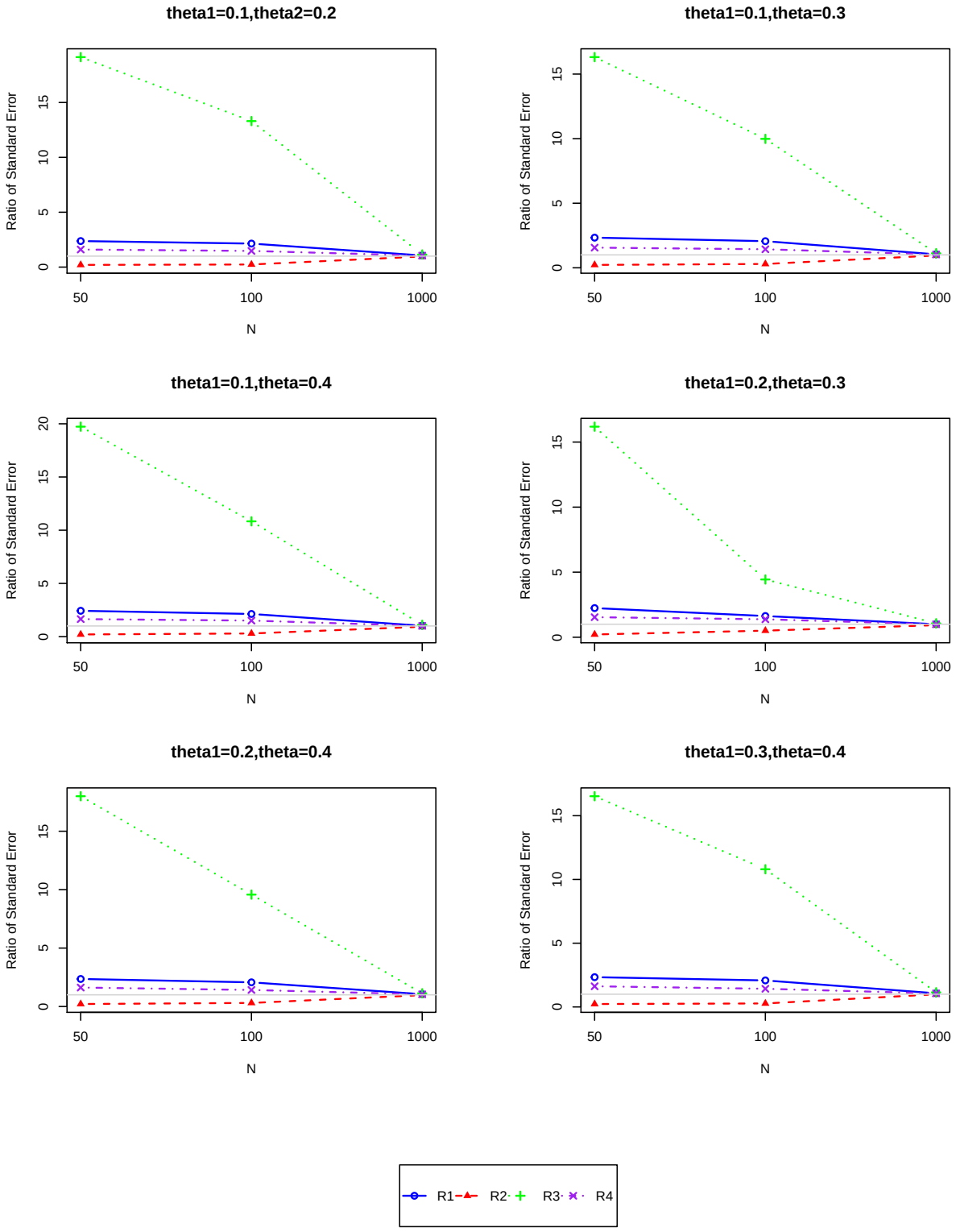

Figure 6.3: *Ratio of standard errors from two formulas (V1 and V2) to the true standard errors of MC and MC3 when data are generated from the mixture of geometric*(θ_1, θ_2) *with 20% one-inflation*

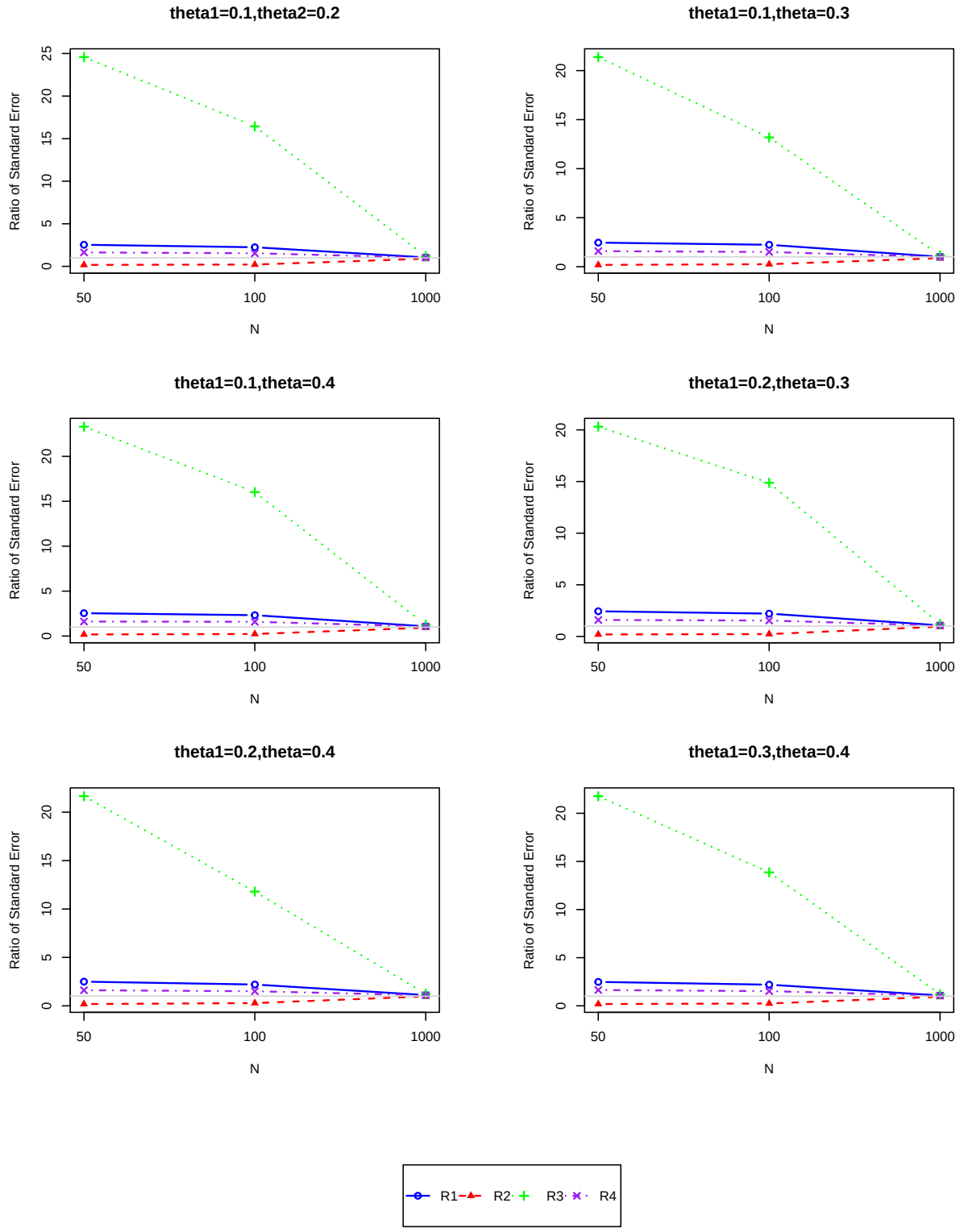

Figure 6.4: *Ratio of standard errors from two formulas (V1 and V2) to the true standard errors of MC and MC3 when data are generated from the mixture of geometric(θ_1, θ_2) with 50% one-inflation*

Chapter 7

Conclusion and Future Work

This chapter provides the discussion and conclusion of this thesis. Some further works are also examined in the last section in order to extend and develop the research in the future.

7.1 Discussion and conclusion

Capture-recapture technique is an important topic in many research areas. It is used to estimate the target elusive population size N. The problem consists in predicting a value for the number of units that have been missed by using the information collected from the captured units during a study period. The predicted value of missing units depends on the model for the capture of units based on a zero-truncated count distribution. The typical model is the Poisson or the binomial. The various models and estimators were developed and proposed to improve inferences in capture-recapture studies which always rely on certain assumptions but might be violated in real situations due to time effect, heterogeneity or behavioural response among others. Heterogeneity in the capture probabilities is a common occurrence. A simple model is not flexible enough to capture the variation in the recapture probability for the distinct units of real-life population so the mixture is considered as a natural model for modelling a heterogeneous population. Additionally, some capture-recapture data show some sort of *one-inflation* in the count distribution. Some portion of the population is mostly captured only once. This may be a consequence of many factors such as trap avoidance, low probability of recapturing the same individual in large cities/areas within a short period of survey, misclassification and so on. As a result of one-inflation being present in the count data, some fitting models suffer from a boundary problem and some estimators provide extreme overestimation of the population size, particularly, Chao's lower bound estimator. The aim of this thesis is to develop the estimators and the models specifically designed to estimate the size of a population for one-inflated capture-recapture count data allowing

for heterogeneity. These models are based upon the geometric distribution as it is a remarkably simple distribution with memoryless property and also provides a flexible model for some heterogeneity in the count data. The estimators are developed under three concepts.

The first concept is suggested in Chapter 3. It is based on a modification by truncating singletons and applying the conventional Turing and maximum likelihood estimation approach to the one-truncated geometric data for estimating the parameter p_0. These estimators $\hat{p}_0$ are applied to the Horvitz-Thompson approach for the new population size estimators T_OT and MLE_OT which are denoted as modified Turing and modified MLE estimators, respectively. The simulation results provide evidence that the T_OT and MLE_OT can solve the problem of one-inflation. They show a good performance of accuracy and perform best with the smallest mean square error for all conditions of study when the comparison is done among existing conventional estimators. Although both proposed estimators can perform very well and reasonable under the one-inflation, T_OT gives superior results compared to MLE_OT.

The second concept is the model-based approach. It focuses on developing a statistical model that describes the mechanism to generate the extra of count ones as shown in Chapter 4. The new estimator MLE_ZTOI is developed from a maximum likelihood approach by using the nested EM algorithm based upon the zero-truncated one-inflated geometric distribution. Maximum likelihood approach is interesting and indicated to use for developing an estimator as it has desirable mathematical and optimality properties, particularly, it becomes minimum variance unbiased estimator as the sample size increase. These properties can be confirmed by the simulation studies in Section 4.6. The simulation results also show that the new estimator MLE_ZTOI can cope with the problem of one-inflation by reducing an overestimation and perform best among all proposed estimators, T_OT, MLE_OT and MLE_ZTOI and existing conventional estimators, Chao, Turing and MLE.

As it is shown in Chapter 2 and 3 that a classical Chao's estimator is popular and frequently used in capture-recapture study but it is severely affected by one-inflation due to the fact that its formula relates to a square of singletons, $n + f_1^2/(2f_2)$. Hence the last concept focuses on modifying the classical Chao estimator by avoiding using the frequency of count ones for estimation. Chao's estimator is modified to involve the frequency of counts of twos and threes instead of the frequency of counts of ones and twos (see Chapter 5). The modified Chao estimator (MC) can retain the good properties of a classical Chao estimator. It is asymptotic unbiased estimator for a power series distribution with and without one-inflation. It provides a lower bound estimator under a mixture of power series distributions with and without one-inflation. These good properties can be seen from theoretical, analytic and simulation results. It shows a good performance in simulation studies and it is applicable in real situations. However, the modified Chao estimator is a biased estimator when the sample size is

small, therefore three bias-correction versions of the modified Chao estimator (MC1, MC2 and MC3) have been developed. The frequency of counts is assumed to follow a Poisson distribution which is a classical assumption in analysis of frequency table. The property of equidispersion and the third moment of the Poisson distribution are used to reduce the bias. All bias-reduction versions can reduce the bias considerably as can be seen from simulation studies, especially the best one is the bias reduction version three (MC3).

Furthermore, a variance approximation of the modified Chao estimator (MC) and the modified Chao estimator with bias reduction version 3 (MC3) are examined in Chapter 6. A conditioning technique is used to derive the variance of the estimator MC and MC3. The variation of MC and MC3 estimator arise from two sources. The first source is from the random variation of observed sampling from population (n). The second is from the variation of the predicted estimate ($\hat{f}_0$). As there is nothing uncertain about observed n we focus on constructing the partial variance of prediction, $var(\hat{f}_0)$, for both MC and MC3, denoted by V1 and V2, respectively. The simulation study presents that V2 has the best performance for estimating the variance of MC and MC3 as it provides the closest values to the true variance. V1 gives severe overestimates for small sample size but it is asymptotically unbiased when the sample size is large. Therefore, it can be reasonably stated that the variance estimation V2 approximates well the true variance of MC and MC3 whereas the variance estimation V1 can stand for the true variance of MC and MC3 only in case of a large sample size.

It can be seen clearly that all proposed estimators based on different concepts can cope with the problem of one-inflation. Each concept has a different strength and limitation.

1. The first (T_OT) and third (MC and MC3) concept are simpler whereas the second (MLE_ZTOI) concept is more complex and more computational demanding.

2. The second concept uses a model-based approach to explain the extra-ones whereas the first and third concept ignore the information from count of ones that is the main information of data.

3. The first and second concept are a parametric approach whereas the third concept is completely nonparametric.

4. Although the second concept produces the best estimates among estimators based on the parametric approach, it may experience boundary problems.

5. The third concept has neither an identifiability problem, nor is there need to specify a mixing distribution.

6. The first and second concept are suitable for a heterogeneous population following the geometric distribution with one-inflation whereas the third concept is more

flexible as it can be applied for the population following a power series distribution and the mixture of power series distribution with and without one-inflation.

7. The first and second concept are good for all sizes of population whereas the third concept (MC) is good for the large size of population. However, the bias correction version 3 (MC3) in the third concept can cope with this problem. MC3 is also good for all sizes of population.

8. The first and second concept provide small variances whereas the third concept provides large variances. In other words, the first and second concept are superior to the third concept in terms of precision.

In this thesis, all proposed estimators are developed under one-inflated capture-recapture count data. All procedures and algorithms for calculations have been done by *R programming*. For efficient estimation, it is necessary to check and ensure the validation of all basic assumptions of the considered estimators. Here, the ratio plot is used as preliminary investigating tool for the presence of one-inflation. Alternatively, the likelihood ratio test can also be used for testing the distribution with one-inflation.

7.2 Future work

Although the results of all proposed estimators and models have illustrated the efficiency of dealing with the zero-truncated capture-recapture count data with one-inflation, there are some points that could be further developed and extended.

The generalized modified Chao estimator of population size for capture-recapture study might be another aspect for further work if covariates are available. The modified Chao estimator provides a lower bound of the population size under one-inflated unobserved heterogeneity as shown in Chapter 5. If heterogeneity is observed and available in form of covariates, this information can be used to reduce the bias of the modified Chao estimator for one-inflation (see Böhning et al. (2013b) for motivation).

The modified Chao estimator, using likelihood framework in Section 6.2, is extended to include covariate information working directly with a truncated power series likelihood rather than with the complete power series likelihood, truncating all counts except counts of twos and threes. Let $(X_1, z_1), ..., (X_N, z_N)$ be a sample with covariate information where z_i is a p-dimensional vector additional information on unit i. It can be assumed that the heterogeneity can be captured by mean of a power series regression model with log-link function

$$\theta_i = \exp(\alpha + \beta^T z_i) \tag{7.1}$$

for $i = 1, ..., M$ where M is the total number of covariate combinations with $n_1 + n_2 + ... + n_M = n$ and n_i is the frequency of covariate combination i. The associated truncated

power series model with all counts truncated except $X_i = 2$ and $X_i = 3$ is

$$P(X_i = 2) = (1 - p_i) = \frac{1}{1 + (b_3/b_2)\theta_i}$$

and

$$P(X_i = 3) = p_i = \frac{(b_3/b_2)\theta_i}{1 + (b_3/b_2)\theta_i}.$$

The truncated power series likelihood is given as

$$\begin{aligned} &\prod_{i=1}^{M} \left(\frac{1}{1 + (b_3/b_2)\theta_i} \right)^{f_{i2}} \times \left(\frac{(b_3/b_2)\theta_i}{1 + (b_3/b_2)\theta_i} \right)^{f_{i3}} \\ &= \prod_{i=1}^{M} \left(\frac{1}{1 + (b_3/b_2)\exp(\alpha + \beta^T z_i)} \right)^{f_{i2}} \times \left(\frac{(b_3/b_2)\exp(\alpha + \beta^T z_i)}{1 + (b_3/b_2)\exp(\alpha + \beta^T z_i)} \right)^{f_{i3}} \end{aligned} \tag{7.2}$$

where f_{ij} is the frequencies of count j in the covariate combination i when $j = 2 \quad \text{or} \quad j = 3$. The likelihood in (7.2) can be written in another form as

$$\prod_{i=1}^{M} (1 - p_i)^{f_{i2}} p_i^{f_{i3}} = \prod_{i=1}^{M} \left(\frac{1}{1 + \exp(\alpha' + \beta^T z_i)} \right)^{f_{i2}} \times \left(\frac{\exp(\alpha' + \beta^T z_i)}{1 + \exp(\alpha' + \beta^T z_i)} \right)^{f_{i3}} \tag{7.3}$$

where $\alpha' = \log(b_3/b_2) + \alpha$. Therefore, the log-likelihood becomes

$$l(\theta_i \mid f_2, f_3) = \sum_{i=1}^{M} f_{i3} \log[(b_3/b_2)\theta_i] - \sum_{i=1}^{M} (f_{i2} + f_{i3}) \log[1 + (b_3/b_2)\theta_i] \tag{7.4}$$

Hence, the idea is to use this likelihood to estimate α and β in the model (7.1) and then develop from here inference for a generalized, modified and covariate-adjusted Chao estimator.

Other ideas for future work include on improved diagnostic methodology beyond the ratio plot for diagnosis of one-inflation. Another interesting area of future work is investigating count inflation other than those of ones.

In summary, several approaches for modelling and estimating in capture-recapture study have been developed. A new crucial problem in many fields is presence of one-inflation in capture-recapture count data. It affects the efficiency of inference. The new proposed models and estimators for one-inflated heterogeneous population under three concepts in this thesis have shown the good performance to cope with this situation. It can be expected that the knowledge gained from this thesis will lead to considerable impact in theoretical and practical research in capture-recapture methods based on counting distribution.

References

Allen, H., Looft, T., Bayles, D., Humphrey, S., Levine, U., Alt, D., and Stanton, T. (2011). Antibiotics in feed induce prophages in swine fecal microbiomes. *MBio*, 2(6):1–9.

Amstrup, S., McDonald, T., and Manly, B. (2005). *Handbook of Capture-Recapture Analysis*. Princeton: Princeton University Press.

Anan, O. (2016). *Capture-recapture modelling for zero-truncated count data allowing for heterogeneity*. PhD thesis, University of Southampton.

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis, Theory and Practice*. New York: McGraw-Hill.

Böhning, D. (2008). A simple variance formular for population size estimators by conditioning. *Statistical Methodology*, 5:410–423.

Böhning, D. (2010). Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scandinavian Journal of Statistics*, 37:221–236.

Böhning, D. (2015). Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *Metron*, 73:201–216.

Böhning, D., Baksh, M., Lerdsuwansri, R., and Gallaghher, 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 Schon, D. (2005). Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14:29–43.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388.

Böhning, D. and Kuhnert, R. (2006). Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics*, 62:1207–1215.

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*, 54:721–737.

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*, 19:1075–1083.

Böhning, D., Van der Heijden, P., and Bunge, J. (2017). *Capture-Recapture Methods for the Social and Medical Sciences*. Boca Raton: Chapman & Hall/CRC.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. (2013b). A generalization of Chao's estimator for covariate information. *Biometrics*, 64(4):1033–1042.

Böhning, D. and Vilas, V. D. R. (2008). Estimating the hidden number of scrapie affected holding in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics*, 13:1–22.

Brittain, S. and Böhning, D. (2009). Estimators in capture-recapture studies with two sources. *AStA Advances in Statistical Analysis*, 93:23–47.

Brookmeyer, R. and Gail, M. (1988). A method for obtaining short term projections and lower bound on the size of the aids epidemic. *Journal of the American Statistical Association*, 83(402):301–308.

Bruno, G., RaPorte, R., Merletti, E., and et al. (1994). National diabetes programs: application of capture-recapture to "count" diabetes. *Diabetes Care*, 17:548–556.

Bunge, J., Böhning, D., Allen, H., and Foster, J. (2012). Estimating population diversity with unreliable low frequency counts. *Biocomputing 2012 : Proceeding of the pacific symposium, Kohala Coast, Hawaii, USA, 2-6 January 2012. World Scientific Publishing*.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, 88:364–373.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data, 2nd edition*. Econometric Society Monograph No.53, Cambridge University Press.

Casella, G. and Berger, R. (2008). *Statistical Inference, 2nd edition*. California: Brooks/Cole Publishing Company.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783–791.

Chao, A. (1989). Estimating the 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*, 58:531–539.

Chao, A. and Lee, S. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87:210–217.

Dorazio, R. and Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59:351–364.

Edwards, W. and Eberhardt, L. (1967). Estimating cottontail abundance from live trapping data. *The Journal of Wildlife Management*, 31(1):87–96.

Farcomeni, A. (2017). Fully general chao and zelterman estimators with application to a whale shark population. *Journal of the Royal Statistical Society (Series C), Applied Statistics*, 67:217–229.

Farcomeni, A. and Scacciatelli, D. (2013). Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy. *Annals of Applied Statistics*, 7:2293–2314.

Gallay, A., Vaillant, V., Bouvet, P., Grimont, P., and Desenclos, J. (2000). How many foodborne outbreaks of salmonella infection occurred in france in 1995? Application of the capture-recapture method to three surveillance systems. *American Journal of Epidemiology*, 152:171–177.

Godwin, R. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, 59:79–93.

Grogger, J. and Carson, R. (1991). Models for truncated counts. *Journal of Applied Econometrics*, 6:225–238.

Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory*, 11:235–243.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Hser, Y. (2001). Population estimation of illicit drug users in Los Angeles county. *The Journal of Drug Issues*, 23(2):323–334.

Huggins, R. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statistics and Probability Letters*, 54:147–152.

Jouanjus, E., Pourcel, L., Saivin, S., Molinier, L., and Lapeyre, M. (2012). Use of multiple sources and capture-recapture method to estimate the frequency of hospitalizations related to drug abuse. *Pharmacoepidemiology and Drug Safety*, 21(7):733–741.

Keith, L. and Meslow, E. (1968). Trap response by snowshoe hares. *The Journal of Wildlife Management*, 32(4):795–801.

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.

Link, W. (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. *EASE '15 Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, 15.

McCrea, R. and Morgan, B. (2014). *Analysis of Capture-Recapture Data*. Boca Raton: Chapman & Hall/CRC.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

Meng, X.-L. (1997). The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research*, 6:3–23.

Navaratna, W., del Rio Vilas, V., and Böhning, D. (2008). Extending Zeltermans approach for robust estimation of population size to zero-truncated clustered data. *Biometrical Journal*, 50(4):584–596.

Neubauer, G., Djuras, G., and Friedl, H. (2011). Models for underreporting:a Bernoulli sampling approach for reported counts. *Austrian Journal of Statistics*, 40:85–92.

Neubauer, G. and Friedl, H. (2006). Modelling sample sizes of frequencies.

Niwitpong, S., Böhning, D., Van der Heijden, P., and Holling, H. (2013). Capture-recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika*, 76:495–519.

Norris, J. and Pollock, K. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, 3(3):235–244.

Otis, D., Burnham, K., White, G., and Anderson, D. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62:1–135.

Pledger, S. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, 61:868–876.

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.

Seber, G. (2002). *The Estimation of Animal Abundance and Related Parameter (2nd ed)*. London: Griffin.

Sekar, C. and Deming, E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44:101–115.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Stock, A., Jürgens, K., Bunge, J., and Stoeck, T. (2009). Protistan diversity in the suboxic and anoxic waters of the gotland deep (Baltic Sea) as revealed by 18s rrna clone libraries. *Aquatic Microbial Ecology*, 55:267–284.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E., Ellisman, M., Grethe, J., and Wooley, J. (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: the camera resource. *Nucleic Acids Research*, 39:546–551.

Van der Heijden, P., Bustami, R., Cruyff, M., Engbersen, G., and van Houwelingen, H. (2003a). Point and interval estimation of the population size using the truncated poisson regression model. *Statistical Modelling*, 3:305–322.

Van der Heijden, P., Cruyff, M., and Böhning, D. (2014). Capture-recapture to estimate crime populations. in: G.J.N. Bruinsma and D.L. Weisburd (eds.). *Encyclepedia of Criminology and Criminal Justice. Berlin: Springer*, pages 267–278.

Van der Heijden, P., Cruyff, M., and van Houwelingen, H. (2003b). Estimating the size of criminal population from police record using the truncated Poisson regression model. *Statistica Neerlandica*, 57:289–304.

Vergne, T., Calavas, D., Cazeau, G., Dufour, B., and Grosbois, V. (2012). A Bayesian zero-truncated approach for analysis capture-recapture count data from classical scrapie surveillance in France. *Preventive Veterinary Medicine*, 105:127–135.

Vergne, T., Paul, M., Chaengprachak, W., Durand, B., Gilbert, M., Dufour, B., Roger, F., Kasemsuwan, S., and Grosbois, V. (2014). Zero-inflated model for identifying disease risk factors when case detection is imperfect: Application to highly pathogenic avain influenza h5n1 in Thailand. *Preventive Veterinary Medicine*, 114:28–36.

Viwatwongkasem, C., Kuhnert, R., and Satitvipawee, P. (2008). A comparison of population size estimators under the truncated count model with and without allowance for contamination. *Biometrical Journal*, 50(6):1006–1021.

Wang, J.-P. and Lindsay, B. (2005). A penalized nonparametric maximum likelihood appraoch to species richness estimation. *Journal of the American Statistical Association*, 100:942–959.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Heidelberg: Springer.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with applications to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18:225–237.