**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF HUMAN, SOCIAL AND MATHEMATICAL SCIENCES

Academy Unit of Mathematics

# Development of Capture-Recapture Estimators in Closed Populations Including Individual Covariate Information

by

**Alberto Vidal-Diez**

Thesis for the degree of Doctor of Philosophy

December 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF HUMAN, SOCIAL AND MATHEMATICAL SCIENCES
Academy Unit of Mathematics

Doctor of Philosophy

# DEVELOPMENT OF CAPTURE-RECAPTURE ESTIMATORS IN CLOSED POPULATIONS INCLUDING INDIVIDUAL COVARIATE INFORMATION

by Alberto Vidal-Diez

Capture-recapture is an area in statistics which aims at the estimation of the size of an elusive target population using the number of times individuals have been identified within a time period or in several capture occasions. Closed populations models in capture-recapture assume no migration, deaths or births during the study period. Many approaches can be found in the literature that differ in the assumptions about the distribution and sources of heterogeneity of the capture probability based on the observed individuals. There are three main sources of heterogeneity: individual characteristics, the conditions at the time of capture and the behavioural response after being captured and released. Chao's lower bound estimator is a well-known estimator that uses only individuals capture once or twice to estimate the hidden population and it can produce robust estimates assuming a Poisson mixture distribution without specifying directly the mixing distribution.

We develop a framework based on zero- and right-truncated models to extend Chao's lower bound estimator (Chao, 1987) to use individual covariate information for modelling heterogeneity of the capture probability. An initial estimator for continuous-time experiments is presented based on a truncated Poisson distribution with only counts of ones and twos non-truncated. The calculation of this estimator can be easily done with standard statistical software. The methodology is then extended in two ways. For one, Chao's estimator is extended to any member of the power series distribution by using a truncated likelihood using only counts of ones and twos. For two, the framework is extended to include more general cut-off values larger than 2. A statistical test to select the optimal truncation cut-off point is developed and model-averaged estimates are also suggested to combine estimates with different truncation cut-off points.

We also extend an estimator based on a geometric distribution with censoring (Niwitpong et al., 2012) to use individual covariate information. Similarly to the methodology presented based on truncation we generalise estimates to use different censoring cut-off points. All estimates are assessed using simulations and practical guidance and case studies are provided to facilitate the reader the understanding and application of the proposed methods.

# Contents

# List of Figures

# List of Tables

# Declaration of Authorship

I, Alberto Vidal-Diez , declare that the thesis entitled *Development of Capture-Recapture Estimators in Closed Populations Including Individual Covariate Information* 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., 2013b)

Signed:.....

Date:.....

## Acknowledgements

I would like to thank my wonderful wife Reiko for her constant support and love all these years. During this PhD we got married and we had our first son, Hugo. She has been always supporting me and giving me energy and strength to complete this thesis. The path has been long and difficult, but she has been always there, specially in the bad moments, helping me on a daily basis and taking care of Hugo when I needed to have some extra time to progress in the PhD. I will always owe her all those hours I could not spend with her to work in the PhD.

I have to thank my son Hugo, because his arrival was the most amazing experience it has happened in my life, and although he stole so many hours of my sleep those difficult first months, making the PhD even more challenging; his smile and hearing him every morning calling me "papa" make me feel complete and boost my adrenaline to make the most of each day.

I would like to thank my parents who always believe in me. They have always sacrificed everything they had to give to their children the best education and opportunities. They always gave me the freedom to choose my future, they always support my decisions and they are always there when I need any advice.

I also have to thank my sister Sonia, although we have different lives I know she will be always there for me as we will always have each other. A big thank you to my parents-in-law to their constant support during all these PhD years.

I have to specially thank my supervisor Professor Dankmar Böhning, for his patience, constant guidance, his humbleness and his big heart. We started our friendship working together in 2006 and I have been always learning from him since then. I always wished to do a PhD under his supervision until and I was very lucky to have such opportunity at the same time that I could work full time. It is an incredible experience to work with him.

I would like to thank my friend Dr Mark Arnold, another important influence in my career. He helped me to get the funding to start the PhD and he has always supported me in all these years.

Finally I would like to thank Dr Alan Karthikesalingam for supporting me financially in the last years, giving me time to work in the PhD in the key moments and make me feel happy to go to work every day.

Me gustaría darle gracias a mi mujer Reiko for su apoyo y amor constante durante todos estos años. Nos casamos y tuvimos nuestro primer hijo Hugo durante el transcurso de este doctorado. Siempre ha estado animándome y dandome la energía y la fuerza necesaria para acabar la tesis. El camino ha salido largo y difícil, pero ella siempre

estuvo allí, ayudándome cada día, especialmente en los malos momentos, y cuidando a Hugo cuando necesitaba tener algo de tiempo para progresar en el doctorado. Siempre le deberé esas horas que no pudimos pasar juntos por tenerlas que pasar trabajando en el doctorado.

Tengo que darle gracias a Hugo, porque su llegada es la experiencia más increible que ha ocurrido en mi vida, y aunque me robó muchas horas de sueño los primeros meses, e hizo el doctorado incluso más complicado;su sonrisa y escucharle cada mañana llamarme papá, me hace sentir completo y me llena de energía para aprovechar cada día al máximo.

Tengo que dar las gracias a mis padres que siempre creyeron en mi. Han sacrificado siempre todo para dar a sus hijos las mejores oportunidades y la mejor educación. Siempre me dieron la libertad de elegir mi futuro, apoyaron mis decisiones y se que siempre están ahí cuando necesito algún consejo.

Tambien quiero darle gracias a mi hermana Sonia, aunque tenemos vidas diferentes, se que ella siempre estará ahí por mi y siempre nos tendremos el uno al otro. Agradezco enormemente el apoyo constante de mis durante los años del doctorado.

Tengo que darle gracias especialmente a mi supervisor el catedrático Dankmar Böhning, por su paciencia, guía constante, su humildad y su gran corazón. Comenzamos nuestra amistad en el 2006 trabajando juntos, y desde entonces siempre he aprendido de él. Siempre deseaba hacer un doctorado supervisado por él, y tuve muchísimima suerte de tener esa oportunidad al mismo tiempo que podía trabajar a tiempo completo. Es una experiencia increible trabajar con él.

Me gustaría dar las gracias a mi amigo Dr Mark Arnold, otro persona importante en mi carrera profesional. Me ayudó a conseguir la financiación inicial para comenzar el doctorado y siempre me ha ayudado todos estos años.

Finalmente me gustaría darle las gracias a Dr Alan Karthikesalingam por financiarme el doctorado en los últimos años, darme tiempo para completar el doctorado en los momentos claves y hacer que siempre vaya contento a mi trabajo.

# Nomenclature

| | |
|---|---|
| $n$ | number of observed individuals. |
| $m$ | number of trapping occasions or maximum number of captures observed. |
| $f_y$ | number of individuals captured exactly $y$ times. |
| $Z_i$ | vector of covariates values for individuals with the $i$-th unique combination of covariates. |
| $e_y$ | expected value of $f_y$. |
| $N$ | true population size. |
| $J$ | is the number of non-truncated counts. |
| $M_J$ | is the number of covariate combinations when $J$ non-truncated counts are considered. |
| $c$ | is the number of non-censored counts. |
| RMSE | relative mean squared error. |
| RBias | relative bias. |
| GC | Generalised Chao's estimator. |
| ZTP | Zero-truncated Poisson estimator. |
| ZNB | Zero-truncated negative binomial estimator. |
| WLR | Weighted linear regression model estimator. |

# Chapter 1

# Introduction to capture-recapture analysis in closed populations

## 1.1 Overview

Capture-recapture is a statistical area which aims at estimating the size of hidden or elusive populations. In a classic setting where we are interested in estimating $N$ and $p$ for a binomial distribution $Y \sim Bin(N, p)$, capture-recapture intelligently uses a different sampling design to obtain estimates for both parameters. Originally, we have a discrete-time experiment where there is a fixed number of sampling occasions and the population of interest is sampled. At every occasion, the units captured are marked or identified somehow, so a complete capture-recapture history is recorded for each captured unit. The idea is to use the capture history of those units to estimate the size of the population that we have not observed.

There are two main types of models: closed and open population models. A closed population model assumes a constant population without births, deaths or migration. Although in real life most populations should be considered open, the assumption of a closed population is not severely violated when the observational period is small or studying small areas. For that reason, the research in closed populations is still active. Wildlife populations are usually open and there is an interest to include births, deaths and migration to estimate the changes in the population, survival rates and the number of new individuals in the population between sample times. The scope of this thesis is only to develop estimators for closed populations, so open populations are not discussed any further.

Although capture-recapture is a popular analysis in ecology for animal abundance estimation (Boyce et al., 2001; Karanth, 1995; Keating et al., 2002), we find examples in the literature where capture-recapture methods have been applied to other areas: in

software engineering to estimate the number of errors in a computer software (Duran and Wiorkowski, 1981; Nayak, 1988), in public health to estimate the number of drug users in a city or a country or to assess the completeness of medical registries (Böhning et al., 2004; Farcomeni and Scacciatelli, 2013; McDonald et al., 2014; Xu et al., 2014; Bailly et al., 2015; Hay et al., 2009), in demography to estimate the US census undercount (Fienberg, 1972), in veterinary medicine to investigate the number of hidden scrapie population in Great Britain (Böhning and Del Rio Vilas, 2008; Böhning, 2011), in sociology to infer the number of victims in armed conflicts based on multiple registries (Lum et al., 2010; Mitchell et al., 2013), in criminology to estimate the number of illegal immigrants living in the Netherlands coming from some Middle East countries (Van der Heijden et al., 2012).

In this thesis we are interested in two types of capture-recapture settings: discrete and continuous-time experiments:

- A discrete-time experiment involves a fixed number of trapping occasions $m$. So each individual $i$ will have a capture history $x_i = (X_{i1}, X_{i2}, ..., X_{im})$ with $X_{ij}$ being a binary variable indicating whether the subject $i$ was registered at occasion $j$, with $i = 1, \ldots, N$ and $j = 1, \ldots, m$.

  This type of experiment also arises with multiple sources, lists or diagnostic tests where each list is considered as a trapping occasion. Table 1.1 contains an example of the raw data format that we find in discrete-time experiments. There are $m = 4$ trapping occasions. The table is ordered to show the capture history of the $n$ observed individuals in the first rows. All captures for individuals $n+1$ to $N$ are 0 as they were not observed. For instance, the table shows that the first individual was captured in the first and third occasion and the second individual was observed in the first three occasions.

  The last column in table 1.1 reports the number of times individual $i$ has been captured. Some analyses use only this information, ignoring the order of the capture history which implies the assumption of homogeneity and independence among capture occasions. If there is correlation between sample occasions, the estimations can be positively or negatively biased depending on the direction of the correlation (Chao, 2001).

- A continuous-time experiment arises when the captures happen in a fixed period of time $[0, T]$ and the time of each capture is recorded. A sample of counts $Y_1, Y_2, \ldots, Y_N$ can occur in multiple ways. $N$ is the total number of individuals in the population and the quantity of interest. $Y_i$ represents the number of times that individual $i$ has been captured in the study period. We can have this setting for example in the identification of drug users or homeless in a city, estimating the number of big mammals or other animals using fixed cameras to take photos in

Table 1.1: Example of the raw data format for a discrete-time capture-recapture experiment

| Individuals | Sample 1 | Sample 2 | Sample 3 | Sample 4 | $Y$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 2 |
| 2 | 1 | 1 | 1 | 0 | 3 |
| 3 | 0 | 0 | 1 | 1 | 2 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| $n$ | 0 | 0 | 0 | 1 | 1 |
| $n+1$ | 0 | 0 | 0 | 0 | 0 |
| . | . | . | . | . | . |
| $N$ | 0 | 0 | 0 | 0 | 0 |

a geographical area. In these examples captures can occur at any time within a given period previously established.

In this case, the cluster is the individual, and we study the number of repeated identifications in the time interval $[0, T]$. However, the cluster could be a grouping variable like farms, villages or households. A recapture event happens when a second individual from the same cluster is identified. For example, the estimation of the prevalence of a disease at farm level that is based on a passive surveillance database. The unit is the farm and the recapture occurs when a second sick animal from the same farm is found, and the time and location of the farm of origin is recorded. Another example is the estimation of households infected within an outbreak in a village. The cluster is the village in this case.

For some analyses where the data is seen as a continuous-time counting process (Becker, 1984), the time of the capture will be used. However, other models work with a simplified format as a frequency of frequencies (Table 1.2) where $y$ represents the number of captures and $f_y$ is the number of individuals captured exactly $y$ times in the period of the study. This format assumes equal probability of being captured across the period of the study for each individual.

The species richness problem (Chao and Bunge, 2002) is also included under the umbrella of these experiments.

Table 1.2: Frequency of frequencies format in a continuous-time setting

| $y$ | $f_y$ |
|---|---|
| 1 | 20 |
| 2 | 9 |
| 3 | 5 |
| 4 | 2 |

## 1.2 Heterogeneity explained with covariates in closed population models

Pierre Simon Laplace was one of the first people who applied a kind of capture-recapture approach using ratios to estimate the size of the population in France in 1802 based on the number of live births per year and census in several communities (Cochran, 1978). John Graunt is also mentioned (Hald, 1990) in the literature to have used capture-recapture concepts in the XVII century to estimate the population in London based on the number of buried people in a year and the families that had someone dead in the family that year.

We already mentioned the importance of the application of capture-recapture methods in ecology to develop new methodology. The estimation of animal abundance is one of the main applications in the area. One of the first estimators was developed thanks to Petersen (1896), Dahl (1917) and Lincoln's work (1930) (Le Cren, 1965) where Petersen and Dahl studied fish populations and Lincoln estimated the number of birds based on the return of bands. Chapman (Chapman, 1951) also published an adjustment to reduce the bias of the Lincoln-Petersen estimator when there is a small number of marked individuals in the second sample. The Lincoln-Petersen estimator assumes two trapping occasions. Darroch (Darroch, 1958, 1959) and Schnabel (Schnabel, 1938) developed estimates for multiple capture occasions assuming changes in the probability in each occasion, but the same capture probability for all individuals, that is $p_{ij} = p_j$.

We can also highlight the early application of capture-recapture in the demography framework by Seker and Deming (Seker and Deming, 1947) who estimated the birth and death rates in an area in India. A more detailed description of the early history of capture-recapture can be found in Seber's book (Seber, 1982).

All these early developments could not deal with the heterogeneity that real life applications often involve. Carothers (1973) carried out a real study where he tried to isolate conditions of equal and unequal catchability capturing taxi cabs in Edinburgh and concluded that it is almost impossible to have equal catchability in natural populations. He compared several estimates that assumed unequal capture probability (Tanaka, 1956; Marten, 1970). Pollock (1976) started developing models that considered different capture probabilities, he referred his main discoveries to his unpublished thesis. Otis (Otis et al., 1978) published a set of models $M_0, M_t, M_b, M_h, M_{hb}, M_{ht}, M_{bt}, M_{hbt}$ for a fixed number of capture occasions that accounted for time effects ($M_t$) like temperature, individual behaviour ($M_b$) like trap shyness/happiness and individual characteristics ($M_h$) like age, ethnicity, sex, etc... There are models combining these three sources of heterogeneity in the capture probabilities depending on the nature of the study. For more details about the early attempts to deal with heterogeneity in closed populations we refer to Chao's review (Chao, 2001) and Amstrup (Amstrup et al., 2005). Chao's review

distinguished between discrete and continuous time models. Chao provided key references of the different approaches for each of the 8 models (including the homogeneous case, "$M_0$")) presented by Otis (see Table 1 on Chao's paper).

This thesis focuses on the development of methodology to include individual covariate information in order to explain the capture probability of each individual ($M_h$). Pollock (Pollock et al., 1984) tried to incorporated covariates using a full likelihood approach in a classical discrete-time experiment. The problem of working with the full likelihood is that we do not have the covariate information for the individuals who are not captured (Sanathanan, 1972). Pollock developed an *ad hoc* method dividing the population in subgroups. However, the approaches of Huggins (Huggins, 1989) and Ahlo (Alho, 1990) (independently suggested) based on a conditional likelihood have became standard methodology, see also Coull and Agresti (1999); Mao and Lindsay (2002); Böhning and Schön (2005). The method consists in writing the likelihood as a product of two terms, one term depends only on the parameter of the distribution ($L_1$) and the other term depends also on the parameter of interest $N$ ($L_2$). Maximising $L_1$, not involving the parameter of interest $N$, they obtained an estimate for the probability of being captured that can be used in the generalised Horvitz-Thompson estimator to obtain the population size $N$. Huggins method can include all sources of heterogeneity ($M_{thb}$). Borchers (Borchers et al., 1998) proposed a method to use the full likelihood approach specifying the distribution of the covariates.

Huggins and Hwang (2011) reviewed and updated the framework and application of the conditional likelihood in capture-recapture. The structure of the paper is a perfect summary of the classic and modern methodologies developed to solve the problem of the heterogeneity. They divided the paper in the following sections: classic approaches and classic log-linear Poisson models, mixture models, covariate models, sample coverage and other non-parametric methods, and they complemented their work presenting a link between the conditional likelihood and the GLM formulation for an $M_h$ model. The classic log-linear Poisson models are used in multiple registries/lists problems (Fienberg, 1972; Cormack, 1989; Coull and Agresti, 1999). Finite mixture models were believed to be flexible to model the capture-recapture distribution (Norris and Pollock, 1996; Böhning et al., 2004; Böhning and Schön, 2005; Pledger and Phillpot, 2008), but the problems with identifiability pointed out by Link (Link, 2003), Dorazio and Royle (Dorazio and Royle, 2003) and Pledger (Pledger, 2005) have revived the interest in other approaches like the inclusion of covariates to determine the probability of capturing an individual. The advantage of these models is that we can interpret the importance of the association of each covariate with the capture probability and standard methods can be used to choose the best models. There are also other non-parametric approaches like the concept of sample coverage by Good and Turing (Good, 1953), Chao's sample coverage estimator with heterogeneous capture probabilities (Chao and Lee, 1992), Chao's lower bound estimator (Chao, 1987), Jacknife estimator (Burnham and Overton, 1978,

1979) or Huggins and Chao's use of martingale estimating functions (Huggins and Chao, 2002).

The friendliest method to incorporate covariates related to captured probabilities is using a log-linear Poisson model (Tilling and Sterne, 1999; Van der Heijden et al., 2003b,a; Böhning and Del Rio Vilas, 2009). Cruyff and Van der Heijden (2008) used also a zero-truncated negative binomial assuming that the overdispersion parameter of the Poisson distribution follows a gamma distribution. However, the standard form to include covariate information as applied in generalised linear models (GLM) has been criticised in the capture-recapture framework as it can be a non-adequate model for some real life applications (Borchers et al., 1998). Another highlighted issue appears when all covariate information might not be enough to explain the whole heterogeneity which can make the model not identifiable (Pollock, 2002) or in the case of multiple lists could mean that the assumption of independence between lists is violated (Zwane et al., 2004). Zwane and Van der Heijden (2004) proposed semiparametric models using generalised additive models (GAM) with smooth functions that allow to relax the assumption of independence between lists and to include non-linear relationships between the covariates and the probability of being captured. They also proposed a graphical tool to assess the goodness of fit of the models with auxiliary variables. Chen and Lloyd (2000) presented a full non-parametric approach as an alternative to the model of Huggins (1989) and Alho (1990). Huggins and Hwang (Huggins and Hwang, 2011) clearly explained the relation between the conditional likelihood and the link function, showing the methodology to relate the approach to generalised linear models. That link-function does not need to be standard so it could be a P-spline (Stoloksa and Huggins, 2012), a smooth function (Huggins and Hwang, 2007) or any other form like a combination of parametric and non parametric forms (Hwang and Huggins, 2007). The package VGAM in the R statistical software allows to fit all those models. Another approach uses a partial likelihood (Stoloksa et al., 2011) where they condition on the first capture, but they recognised that some efficiency is lost compared to Huggins' approach. However, this approach facilitates the use of generalised additive models (GAM) or generalised linear mixed models (GLMM) to model the capture probabilities.

In the context of multiple lists or registries with covariates, methodology has been developed to study the problem of having missing data in some covariates or covariates which are not available in all lists (Baker, 1990; Zwane et al., 2004; Zwane and Van der Heijden, 2007, 2008; Van der Heijden et al., 2009; Sutherland et al., 2007; Xi et al., 2009). Zwane and Van der Heijden combined capture-recapture with multiple imputation techniques (Little and Rubin, 1987) to make use of all information available rather than following a näive approach removing part of the information.

There is not as much research in the continuous-time setting as there is in the discrete-time experiments. Wilson and Collins (1993) reviewed the performance of several robust estimators like Chao's, Zelterman's and Darroch's estimators under heterogeneity in

continuous-time problems. Other classic references for continuous-time models with heterogeneity are Fisher et al. (1943); Tanton (1965); Boyce et al. (2001); Keating et al. (2002); Chao and Bunge (2002). Becker (1984) was the pioneer in formulating the problem as a counting process. Yip et al. (1996) developed a partial likelihood including time-independent covariates to model the individual characteristics. Later, Hwang and Chao (2002a) extended Yip's model to include the behavioural effect, covariates for individuals characteristics and auxiliary variables related to the time of the capture. Farcomeni and Scacciatelli (2013) combined the approaches of Hwang and Chao (2002a) and Xi et al. (2007) extending Hwang's model to include a frailty to represent the unobserved heterogeneity and generalising the behavioural effect to consider a delayed onset and a finite time memory in the behavioural response to change behaviour for a time period and go back to the original behavioural response.

The presence of covariates adds some other challenges. Measurement errors in the covariate information has been proven a source of bias (Creel et al., 2003; Carroll et al., 2006; Hwang and Huang, 2003). Yip et al. (2005) developed a semiparametric method to take into account measurement error over time. Link et al. (2010) presented a Bayesian method to deal with misidentification in natural existing features like DNA fingerprints. They presented a list of studies where they have incomplete observed covariates. They provided a solution for incomplete and inexact subject-specific random covariate data in capture-recapture studies.

## 1.3 Thesis outline

Our initial motivation to start this work is the extension of Chao's lower bound estimator (Chao, 1987) to include auxiliary variables related to individual characteristics of the observed subjects. In chapter 2, we introduce the classic Chao estimator and its properties; we use a Poisson model and include covariate information in the setting of continuous-time experiments under an approach based on a conditional likelihood. Similarly to the original Chao's estimator our model uses only individuals observed once or twice.

In chapter 3, we extend the estimators with and without covariates presented in the previous chapter based on a truncated distribution with two non-truncated counts to include individuals observed $J$ or less times.

Chapter 4 shows the generalisation of the estimators when the counts are assumed to come from a power series distribution, with examples using the binomial distribution in discrete-time experiments.

Chapter 5 presents a goodness of fit test to choose the optimal truncation cut-off point and explores the idea of model averaging for models with different truncation cut-off points.

In chapter 6, we extend an estimator that uses the concept of censoring for a geometric distribution (Niwitpong et al., 2012) to use covariate information for the observed individuals. We compare the developed estimators based on censoring with the estimators based on truncation shown in the previous chapters.

Analytical variance formulae are calculated for all estimators, simulations from several scenarios are provided in all chapters to evaluate the performance of the proposed estimators, case studies and practical guidance are also included in the thesis. The thesis closes with a chapter on open problems and future work to set the thesis results in perspective.

# Chapter 2

# Generalised Chao estimator for the Poisson case

## 2.1 Motivation

In this chapter, we look at the setting of capturing and recapturing units in a closed population framework during a fixed time period. At the end of the period we would have a sample of counts $y_i$, $i \in \{1, 2, \ldots, n, n+1, \ldots, N\}$, which represents the number of times unit $i$ has been captured within the study period. $N$ is the population size and our variable of interest. $n$ is the total number of captured units. There are $N - n$ unobserved units, with $y_j = 0$, $j \in \{n+1, \ldots, N\}$ when the population is sorted to have the captured subpopulation in the first place. Another common notation is using the sum of frequency of frequencies $n = \sum_{y=1}^{m} f_y$ , where $f_y$ represents the number of units captured exactly $y$ times and $m$ is the largest number of recaptures within the period of interest. The outcome of interest is $f_0$, the number of unobserved units, since $\hat{N} = n + \hat{f}_0$ holds.

This scenario occurs in several ways. We find it in populations that are difficult to be completely observed, like a homeless population, a wildlife population or populations with a disease. The units of these populations can be captured and identified using traps, photos, registers, etc.... Each unit $i$ is identified at time $t$ where $t$ can be a random time point within the study period or a fixed capture occasion. In this case, clustering occurs as we have repeated identification of the same units across the study period. A different setting occurs when the recapture comes from identifying units from the same grouping variable, like a farm, household or villages. An example from the literature is the cholera-outbreak in a community in India (McKendrick, 1926) where the clusters were the households in a village.

The distribution of counts defined above could be modelled based on a mixture probability density function

$$p_y = \int_0^\infty p(y|\lambda)q(\lambda)d\lambda \tag{2.1}$$

where the mixing density $q(\lambda)$ is unspecified and the mixture kernel $p(y|\lambda)$ comes from the Poisson family $p(y|\lambda) = Po(y|\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^y}{y!}$. Mixture models appear to be a common methodology applied in the framework of capture-recapture to account for heterogeneity, see Pledger (2005) for the discrete mixture model approach and Dorazio and Royle (2003) for the continuous mixture model approach. Simple models like $p(y|\lambda)$ lack flexibility to explain the individual heterogeneity. However, the identifiability of mixture models in the capture-recapture area has been questioned recently (Link, 2003, 2006; Holzmann et al., 2006). On the other hand, the nonparametric maximum likelihood estimate (NPMLE)

$$\hat{N} = \frac{n}{1 - \int_0^\infty \mathrm{e}^{-\lambda}\hat{q}(\lambda)d\lambda}.$$

where $\hat{q}_\lambda$ is the NPMLE of the mixing density $q_\lambda$ has been reported to produce occasionally high estimated values of the real population (Wang and Lindsay, 2005, 2008).

Therefore, there is a revived interest in the classic Chao's lower bound estimator and its properties (Chao, 1987). See (Mao, 2008) for further developments.

The advantage of using Chao's lower bound estimator when using the mixture density in (2.1) is that the mixing density $q(\lambda)$ does not need to be estimated; in contrast to other approaches like Cruyff and Van der Heijden (2008) where the overdispersion parameter of the Poisson distribution is assumed to be a gamma distribution, which leads to the use of a zero-truncated negative binomial regression. Here, we briefly present the deduction of Chao's lower bound estimator by applying the Cauchy-Schwarz inequality. Böhning et al. (2006) published a generalization of Chao's inequality for power series distribution following the same reasoning.

The Cauchy-Schwarz inequality states that $|E(XY)|^2 \leq E(X^2)E(Y^2)$. In our case $X = \sqrt{\mathrm{e}^{-\lambda}}$ and $Y = \lambda\sqrt{\mathrm{e}^{-\lambda}}$ which leads to:

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

This is equivalent to $p_1^2 \leq p_0(2p_2)$. To estimate $p_i$, we use the estimates $\hat{p}_i = f_i/N$ with $i = 0, 1, 2$ which leads to Chao's estimate:

$$\hat{f}_0 = f_1^2/(2f_2)$$

and consequently,

$$\hat{N}_c = n + f_1^2/(2f_2) \tag{2.2}$$

We have carried out a simple simulation experiment to show how Chao's lower bound estimator can cope with some level of heterogeneity because it does not need to estimate $q(\lambda)$. We generated 1000 samples from a Poisson mixture distribution $Y \sim 0.5*\lambda_1+0.5*\lambda_2$ where $\lambda_1 = 1$. We can observe in table 2.1 how the accuracy of the point estimator and the confidence intervals decreases when heterogeneity is introduced. The coverage of the confidence intervals suggested by Chao and Burnham in (Chao, 1987) are also presented in the table.

Table 2.1: Chao's estimates and coverage of confidence limits (CL) for different population sizes and levels of heterogeneity based on $\lambda_2$

| $\lambda_1$ | $\lambda_2$ | $N$ | $\hat{N}_{Chao}$ | Chao's Coverage CL | Burnham's Coverage CL |
|---|---|---|---|---|---|
| 1 | 1 | 100 | 103.83 | 94.1 | 94.1 |
| | 2 | | 98.29 | 86.5 | 92.1 |
| | 3 | | 97.14 | 84.8 | 93.0 |
| | 4 | | 97.31 | 84.2 | 97.3 |
| 1 | 1 | 500 | 502.70 | 95.0 | 94.6 |
| | 2 | | 486.98 | 85.0 | 89.8 |
| | 3 | | 480.20 | 72.9 | 82.5 |
| | 4 | | 478.15 | 67.6 | 80.0 |
| 1 | 1 | 1000 | 1001.91 | 94.4 | 94.7 |
| | 2 | | 975.23 | 81.6 | 85.8 |
| | 3 | | 957.18 | 57.0 | 66.4 |
| | 4 | | 955.27 | 54.4 | 63.5 |

Chao provided an analytical variance estimator and confidence intervals and obtained a better coverage probability than the jacknife estimator developed by Burnham and Overton (1978) for sample distributions where the core frequencies are ones and twos. Chao's estimator was proved to be a lower bound estimator of the population size. On the negative side, Chao's estimator did not perform adequately when the mass of the capture frequencies was not in the first two counts, because it only makes use of those frequencies. Chao used a Poisson approximation to a binomial distribution and Jensen's inequality to deduce her estimate $\hat{f}_0 = f_1^2/(2f_2)$.

The chapter is structured as follows. In the next section, we develop a likelihood framework to obtain Chao's lower bound estimator from a truncated Poisson distribution. Sections 2.1.2 and 2.2 extend the methodology to include covariate information to explain heterogeneity at individual level. In section 2.3 we review the results obtained from a simulation study with several scenarios. At the end of the chapter, we also present two case studies to show the applicability of the estimator.

### 2.1.1 Chao's estimator from a truncated Poisson

We assume a truncated Poisson likelihood where only counts of ones and twos are non-truncated to represent the capture-recapture distribution. Therefore, there are only two probabilities to define:

$$\begin{aligned} q_1 &= \frac{p(y=1)}{p(y=1)+p(y=2)} = \frac{\mathrm{e}^{-\lambda}\lambda}{\mathrm{e}^{-\lambda}\lambda + \mathrm{e}^{-\lambda}\lambda^2/2} = \frac{1}{1+\lambda/2} \\ q_2 &= \frac{p(y=2)}{p(y=1)+p(y=2)} = \frac{\mathrm{e}^{-\lambda}\lambda^2/2}{\mathrm{e}^{-\lambda}\lambda + \mathrm{e}^{-\lambda}\lambda^2/2} = \frac{\lambda/2}{1+\lambda/2}. \end{aligned} \tag{2.3}$$

Notice the presence of the denominators to fulfill the property that $q_1 + q_2 = 1$.

This truncated sample leads to a binomial log-likelihood:

$$\begin{aligned} \ell(\lambda) &= f_1 \log(q_1) + f_2 \log(q_2) \\ &= f_1 \log\left(\frac{1}{1+\lambda/2}\right) + f_2 \log\left(\frac{\lambda/2}{1+\lambda/2}\right) \\ &= -f_1 \log(1+\lambda/2) + f_2 \log(\lambda/2) - f_2 \log(1+\lambda/2) \\ &= -(f_1+f_2)\log(1+\lambda/2) + f_2 \log(\lambda/2). \end{aligned} \tag{2.4}$$

We maximise the likelihood by taking the first derivative and considering the score equation or equivalently

$$\frac{d\ell}{d\lambda} = -\frac{f_1+f_2}{2+\lambda} + \frac{f_2}{\lambda} = 0.$$

Solving for $\lambda$, we obtain $\hat{\lambda} = 2f_2/f_1$ and replacing $\hat{\lambda}$ in $\hat{q_1}$ and $\hat{q_2}$, we get

$$\hat{q}_1 = \tfrac{f_1}{f_1+f_2} \text{ and } \hat{q}_2 = \tfrac{f_2}{f_1+f_2}.$$

We recognise the estimated $\hat{\lambda}$ as the family of $\lambda$s presented by Zelterman (1988). Zelterman's estimator used the Horvitz-Thompson estimator, $N = \frac{n}{1-\hat{p_0}}$ where $\hat{p_0} = \mathrm{e}^{-\hat{\lambda}}$ and $\hat{\lambda}_i = (i+1)f_{i+1}/f_i$. Zelterman's estimator with $\hat{\lambda} = \frac{2f_2}{f_1}$ can largely overestimate the true population size if there is heterogeneity. Böhning (2011) pointed out an error in Zelterman's calculation of $E(f_0)$. The correct $E(f_0)$ leads to Chao's lower bound estimator.

*Theorem* 2.1. The expectation of $f_0$ for a truncated Poisson with only non-truncated counts of ones and twos is

$$E(f_0|f_1, f_2; \hat{\lambda}) = \frac{f_1^2}{2f_2}, \text{ for } \hat{\lambda} = 2f_2/f_1$$

*Proof.*
We define the expectation of $y$ as $e_y = Po(y|\lambda)N$, where $N$ is the true unknown population size and $Po(y|\lambda)$ the Poisson probability of having $y$ counts for a known $\lambda$.

$$e_y = E(f_y|\lambda) = Po(y|\lambda)N = Po(y|\lambda)\left(e_0 + f_1 + f_2 + \sum_{j=3}^{\infty} e_j\right) \tag{2.5}$$

$e_0$ and $\sum_{j=3}^{\infty} e_j$ are unknown and need to be estimated as

$$e_0 + e_3^+ = [1 - Po(1|\lambda) - Po(2|\lambda)]\left(e_0 + e_3^+\right) + [1 - Po(1|\lambda) - Po(2|\lambda)]\,(f_1 + f_2)$$

with $e_3^+ = \sum_{j=3}^{\infty} e_j$. Solving for $e_0 + e_3^+$ we obtain,

$$e_0 + e_3^+ = \frac{[1 - Po(1|\lambda) - Po(2|\lambda)]}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) \tag{2.6}$$

Therefore, replacing (2.6) in (2.5) for the case of interest $y = 0$:

$$\begin{aligned} e_0 = E(f_0|\lambda) &= Po(0|\lambda)\left(e_0 + f_1 + f_2 + \sum_{j=3}^{\infty} e_j\right) \\ &= Po(0|\lambda)(f_1 + f_2)\left[1 + \frac{1 - Po(1|\lambda) - Po(2|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}\right] \\ &= \frac{Po(0|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) = \frac{\mathrm{e}^{-\lambda}}{\lambda \mathrm{e}^{-\lambda} + \mathrm{e}^{-\lambda}\lambda^2/2}(f_1 + f_2) \\ &= \frac{f_1 + f_2}{\lambda + \lambda^2/2} \end{aligned} \tag{2.7}$$

Finally, if we substitute $\lambda$ by its maximum likelihood estimator $\hat{\lambda} = 2f_2/f_1$, we obtain Chao's lower bound estimator $\hat{f}_0 = f_1^2/(2f_2)$.

### 2.1.2 Chao's estimator with covariates

In this section we follow Böhning et al. (2013b) and consider a sample where additional information for each captured individual unit $i$ is available: $(Y_1, Z_1), ..., (Y_n, Z_n)$ where $Z_i$ is a p-dimensional vector. The idea is to explain the heterogeneity in the Poisson model with the mixing density (2.1) using the covariate information available to model

different capture probabilities at individual level.
First, a Poisson regression with a log-link function is defined to introduce the covariate information in the likelihood framework:

$$\lambda_i = \mathrm{e}^{\alpha+\beta' Z_i} \tag{2.8}$$

where $\lambda_i$ is the conditional Poisson mean with $P(Y_i = y) = Po(y|\lambda_i)$. $Po(y|\lambda)$ is, as previously defined, a truncated Poisson distribution with only one and two counts. We define the probabilities of the non-truncated counts:

$$P(Y_i = 1) = (1 - q_i) = \frac{\lambda_i \mathrm{e}^{-\lambda_i}}{\lambda_i \mathrm{e}^{-\lambda_i} + \frac{\lambda_i^2}{2}\mathrm{e}^{-\lambda_i}} = \frac{1}{1 + \lambda_i/2}$$

and

$$P(Y_i = 2) = q_i = \frac{\frac{\lambda_i^2}{2}\mathrm{e}^{-\lambda_i}}{\lambda_i \mathrm{e}^{-\lambda_i} + \frac{\lambda_i^2}{2}\mathrm{e}^{-\lambda_i}} = \frac{\lambda_i/2}{1 + \lambda_i/2}. \tag{2.9}$$

Let us assume that there are $M$ different observed covariate combinations or strata with $n_1 + ... + n_M = f_1 + f_2$, where $n_i$ is the frequency of stratum $i$, $n_i = \sum_{j=1}^{2} f_{ij}$ with $f_{ij}$ the number of individuals from strata $i$ captured $j$ times. Continuous covariates could lead to the case where all $n_i$ are equal to one. The truncated Poisson likelihood is defined by

$$\prod_{i=1}^{M} \left(\frac{1}{1 + \lambda_i/2}\right)^{f_{i1}} \times \left(\frac{\lambda_i/2}{1 + \lambda_i/2}\right)^{f_{i2}},$$

replacing $\lambda_i$ from (2.8) we obtain

$$\prod_{i=1}^{M} \left(\frac{1}{1 + \mathrm{e}^{\alpha+\beta' Z_i}/2}\right)^{f_{i1}} \times \left(\frac{\mathrm{e}^{\alpha+\beta' Z_i}/2}{1 + \mathrm{e}^{\alpha+\beta' Z_i}/2}\right)^{f_{i2}} \tag{2.10}$$

where $f_{ij}$ are the frequencies of counts $j$ in the $i$th covariate combination with $j = 1, 2$.

We observe that (2.10) is equal to a binomial logistic likelihood except for the intercept:

$$\prod_{i=1}^{M} (1 - q_i)^{f_{i1}} q_i^{f_{i2}} = \prod_{i=1}^{M} \left(\frac{1}{1 + \mathrm{e}^{\alpha'+\beta' Z_i}}\right)^{f_{i1}} \times \left(\frac{\mathrm{e}^{\alpha'+\beta' Z_i}}{1 + \mathrm{e}^{\alpha'+\beta' Z_i}}\right)^{f_{i2}} \tag{2.11}$$

where $\alpha' = \log(1/2) + \alpha$. Hence a logistic regression model could be fitted to calculate the maximum likehood estimates for the truncated Poisson model. We obtain $\hat{\alpha}'$ and $\hat{\beta}$ by maximising the binomial likelihood. We can successively estimate $\lambda_i$ as

$$\hat{\lambda}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i} = 2\mathrm{e}^{\hat{\alpha}'+\hat{\beta}' Z_i} \qquad \text{for} \qquad i = 1, \ldots, M. \tag{2.12}$$

An estimate of $f_0$ can be obtained as the sum of the estimates for each stratum since $f_0 = \sum_{i=1}^{M} f_{i0}$. The application of theorem 2.1 in each stratum $\hat{f}_{0i}$ leads to (2.7) for each covariate combination:

$$\hat{f}_{i0} = \frac{Po(0|\hat{\lambda}_i)}{Po(1|\hat{\lambda}_i)}(f_{i1} + f_{i2}) = \frac{\mathrm{e}^{-\hat{\lambda}_i}}{\hat{\lambda}_i \mathrm{e}^{-\hat{\lambda}_i} + \hat{\lambda}_i^2 \mathrm{e}^{-\hat{\lambda}_i}/2}(f_{i1} + f_{i2}) = \frac{f_{i1} + f_{i2}}{\hat{\lambda}_i + \hat{\lambda}_i^2/2} \tag{2.13}$$

Finally, the estimator arises summing up over all the covariate combinations.

$$\hat{N}_{GC} = n + \sum_{i=1}^{M} \frac{Po(0|\hat{\lambda}_i)}{Po(1|\hat{\lambda}_i)}(f_{i1} + f_{i2}) = n + \sum_{i=1}^{M} \frac{f_{i1} + f_{i2}}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}. \tag{2.14}$$

*Theorem* 2.2. The generalised Chao's estimator is asymptotically unbiased when the Poisson regression model holds. That means that

$$\frac{E(\hat{N}_{GC})}{N} \underset{N\to\infty}{\longrightarrow} 1$$

*Proof.*
We note that $E(n|\hat{\lambda}_1, ..., \hat{\lambda}_N) = \sum_{i=1}^{N}[1 - Po(0|\hat{\lambda}_i)]$ and $E(\Delta_i|\hat{\lambda}_i) = Po(1|\hat{\lambda}_i) + Po(2|\hat{\lambda}_i)$, where

$$\Delta_i = \begin{cases} 1, & y_i \in \{1, 2\} \\ 0, & otherwise \end{cases}$$

with $y_i$ representing the number of times individual $i$ was captured.

Hence,

$$E(\hat{N}_{GC}|\hat{\lambda}_1, ..., \hat{\lambda}_N) = \sum_{i=1}^{N}[1 - Po(0|\hat{\lambda}_i)] + \sum_{i=1}^{N} \frac{Po(1|\hat{\lambda}_i) + Po(2|\hat{\lambda}_i)}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}$$

which becomes

$$\sum_{i=1}^{N}[1 - Po(0|\hat{\lambda}_i)] + \sum_{i=1}^{N} Po(0|\hat{\lambda}_i)\frac{\hat{\lambda}_i + \hat{\lambda}_i^2/2}{\hat{\lambda}_i + \hat{\lambda}_i^2/2} = N.$$

Then the argument is completed by observing that $\lim_{N\to\infty} E(\hat{\lambda}_i) = \lambda_i$.

## 2.2 Variance estimate of $\hat{\mathbf{N}}_{\mathbf{GC}}$

For the calculation of the variance we apply the technique of conditional moments (Ross, 1985), applied also by Böhning (2008) and Van der Heijden et al. (2003a). The variance can be written as the sum of two terms:

$$Var(\hat{N}_{GC}) = Var\left[E(\hat{N}_{GC}|\Delta_i, i = 1, .., N)\right] + E\left[Var(\hat{N}_{GC}|\Delta_i, i = 1, .., N)\right], \quad (2.15)$$

where

$$\Delta_i = \begin{cases} 1, & y_i \in \{1, 2\} \\ 0, & otherwise \end{cases},$$

The first term estimates the variability coming from the sampling of units. Our generalised Chao's estimator can be written as

$$E(\hat{N}_{GC}|\Delta_i, i = 1, \ldots, N) = E\left(n + \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}\right) = E\left(\sum_{i=1}^{N} \Delta_i + \sum_{i=1}^{N} \gamma_i + \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\lambda}_i + \hat{\lambda}_i^2/2}\right)$$

where

$$\gamma_i = \begin{cases} 1, & y_i \geq 3 \\ 0, & otherwise \end{cases}$$

and

$$\lambda_i = \mathrm{e}^{\alpha + \beta' Z_i}.$$

$\lambda_i$ was described previously as the link between the covariate information and the Poisson parameter.

$E(\hat{N}_{GC}|\Delta_i, i = 1, \ldots, N)$ can be written:

$$E(\hat{N}_{GC}|\Delta_i, i = 1, \ldots, N) \approx \sum_{i=1}^{N} \Delta_i \left(1 + \frac{\mathrm{e}^{-\lambda_i}}{p_i}\right) = \Delta_i \omega_i, \quad (2.16)$$

where $\omega_i = (1 + \frac{\mathrm{e}^{-\lambda_i}}{p_i})$ for simplification in the notation.

$p_i$ is the probability that $\Delta_i = 1$:

$$p_i = p(\Delta_i = 1|\lambda_i) = Po(Y_i = 1|\lambda_i) + Po(Y_i = 2|\lambda_i) = \lambda_i \mathrm{e}^{-\lambda_i} + \lambda_i^2 \mathrm{e}^{-\lambda_i}/2.$$

$\Delta_i$ follows a binomial distribution, hence $E(\Delta_i) = p_i$ and $Var(\Delta i) = p_i(1 - p_i)$. Therefore, we have

$$Var\left(E(\hat{N}|\Delta_1, ..., \Delta_N)\right) \simeq \sum_{i=1}^{N} \omega_i^2 Var(\Delta_i) = \sum_{i=1}^{N} \omega_i^2 p_i(1 - p_i)$$

Finally, this variance can be estimated using the Horvitz-Thompson estimator

$$\widehat{Var}(E(\hat{N}|\Delta_1,...,\Delta_N)) \simeq \sum_{i=1}^{N} \frac{\Delta_i}{\hat{p}_i}\hat{\omega}_i^2\hat{p}_i(1-\hat{p}_i) = \sum_{i=1}^{f_1+f_2}(1-\hat{p}_i)\left(\frac{\hat{p}_i+\mathrm{e}^{-\hat{\lambda}_i}}{\hat{p}_i}\right)^2 \tag{2.17}$$

For the second term, we have to calculate $E\left[Var(\hat{N}_{GC}|\Delta_i, i=1,..,N)\right]$ that reflects the sampling variation in the truncated Poisson distribution conditional on $\Delta_i$. We focus on estimating $Var(\hat{N}_{GC}|\Delta_i, i=1,..,N)$ using the multivariate $\delta$-method and we will then take a moment estimator for the calculation of the expected value:

$$Var(\hat{N}_{GC}|\Delta_i, i=1,..,N) = \widehat{Var}\left(\sum_{i=1}^{N}\frac{\Delta_i}{\hat{\lambda}_i+\hat{\lambda}_i^2/2}\right) = \nabla g(\hat{\alpha}',\hat{\beta})'\widehat{cov}(\hat{\alpha}',\hat{\beta})\nabla g(\hat{\alpha}',\hat{\beta}) \tag{2.18}$$

where $\nabla g(\hat{\alpha}',\hat{\beta}) = \begin{pmatrix} \frac{\partial g}{\partial \alpha'} \\ \frac{\partial g}{\partial \beta_1} \\ \dots \\ \frac{\partial g}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{f_1+f_2}\frac{\hat{\lambda}_i+\hat{\lambda}_i^2}{(\hat{\lambda}_i+\hat{\lambda}_i^2/2)^2} \\ \sum_{i=1}^{f_1+f_2}\frac{\hat{\lambda}_i+\hat{\lambda}_i^2}{(\hat{\lambda}_i+\hat{\lambda}_i^2/2)^2}z_{i1} \\ \dots \\ \sum_{i=1}^{f_1+f_2}\frac{\hat{\lambda}_i+\hat{\lambda}_i^2}{(\hat{\lambda}_i+\hat{\lambda}_i^2/2)^2}z_{ip} \end{pmatrix}$,

and $\lambda_i = 2\mathrm{e}^{\alpha'+\beta' z_i}$ and its estimate $\hat{\lambda}_i = 2\mathrm{e}^{\hat{\alpha}'+\hat{\beta}' z_i}$ is calculated by replacing the parameters $\alpha'$ and $\beta$.

The covariance matrix $\widehat{cov}(\hat{\alpha}',\hat{\beta})$ of the regression parameters estimates is available from the logistic regression as the inverse of the Fisher information matrix. The final variance estimate of $Var(\hat{N}_{GC})$ is the result of the sum of (2.17) and (2.18).

## 2.3 Simulations

In this section we assess the performance of our estimator by running simulations for several scenarios. Whenever possible we compare our Generalised Chao's estimator (GC) with the following estimators:

- **Classic Chao's lower bound estimate** (Chao, 1987):

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}. \tag{2.19}$$

- **Turing estimator**. It provides accurate estimates under homogeneity (Good, 1953):

$$\hat{N}_{Turing} = \frac{n}{1-f_1/S} \tag{2.20}$$

where $S = f_1 + 2f_2 + ... + mf_m$ where $m$ is the maximum number of captures.

The background of the Turing estimator is based on the sample coverage estimator $1 - f_1/S$ (Chao et al., 1992). In the case of equal capture probability for all individuals, the sample coverage is $n/N$ which, if equated to $1 - f_1/S$ leads to $\hat{N}_{Turing}$ estimator (Darroch and Ratcliff, 1980).

- **Zero-Truncated Poisson regression with covariates (ZTP)** (Van der Heijden et al., 2003a,b). The capture probability is modelled based on a zero-truncated Poisson regression model. The population size estimate is calculated using the Horvitz-Thompson estimator:

$$\hat{N}_{ZTP} = \sum_{i=1}^{n} \frac{1}{1 - \mathrm{e}^{-\mathrm{e}^{\hat{\alpha}+\hat{\beta}' Z_i}}}, \tag{2.21}$$

  where $\hat{\alpha} + \hat{\beta}' Z_i$ is the fitted linear predictor of a zero-truncated Poisson regression, and $Z_i$ is a vector of covariates related to the capture-recapture probability. The estimator is asymptotically unbiased and efficient when the assumption of the Poisson distribution is true.

- **Zero-Truncated Negative binomial with covariates (ZNB)** (Cruyff and Van der Heijden, 2008). The heterogeneity in the probability of being captured is modelled using a zero-truncated negative binomial model with covariate information introduced in a similar way as in the ZTP model. It uses a gamma distribution for the parameter of the Poisson model.

### 2.3.1 Simulation 1: All heterogeneity explained by covariate information

#### 2.3.1.1 Description of the simulation

In this simulated case study the data are generated following the next steps:

1. Two vectors $X_1$ and $X_2$ of size $N$ are generated independently following normal distributions with means 5 and 8 respectively and variances of 64 ($X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$).

2. Then, the capture-recapture distribution is generated following a Poisson distribution $Y_i \sim Po(\lambda_i)$ where $\lambda_i$ is calculated from a log linear model:

$$\lambda_i = \mathrm{e}^{-0.02X_{1i}+0.03X_{2i}} \qquad \text{with} \quad i = 1, \ldots, N.$$

3. Units which are not captured ($Y_i = 0$) are removed to obtain the sample of captured units.

4. Point and variance estimates are calculated including both covariates into the regression models. Therefore, unbiased estimates are expected from models with covariate information as the entire heterogeneity is explained.

5. Steps 1-4 are repeated 5000 times and statistical measurements such as mean, standard deviation, relative mean squared error and relative bias across samples are used to summarise the results.

#### 2.3.1.2 Results

Table 2.3 shows the point estimates, the standard errors, the *relative mean squared error* (RMSE) and *relative bias* (RBias) respectively. In order to compare the performance of the estimators across multiple population sizes we use $RMSE = \frac{E(\hat{N}-N)^2}{N^2}$ and $RBias = \frac{E(\hat{N}-N)}{N}$.

Chao and Turing's estimators are biased because they are based upon Poisson homogeneity. Classic Chao's estimator performed better under heterogeneity than Turing's estimator. Both estimators ZTP and GC obtain the true population size as expected because the regression models of both estimators were fitted with the same covariates used to generate the heterogeneity of the capture probability (Figure 2.3.1.2).

In this case, the ratio between the standard errors of the GC model over the ZTP estimate is approximately 1.4 for large samples. There is a trade-off for the models with covariates between the accuracy of the estimates and variability. This conclusion becomes clear examining the RMSE and RBias results (Table 2.3). The ZTP-RMSE presents better results than GC because of their differences in variability with similar point estimates. On the other hand the GC-RMSE is similar to classic Chao's estimator because of the larger variance of the GC estimate in spite of its average point estimation being a closer approximation to the true value of the population size. However, the GC-RMSE becomes better than Chao's RMSE when the population size increases. All estimators reduce their RMSE when the population size increases, however the relative biases of Turing's and Chao's estimators remain constant in contrast to GC and ZTP that show a reduction in the relative bias when the population size increases.
Therefore, ZTP is the recommended estimator in a scenario where all assumptions hold.

The last column of table 2.2 shows the ratio between the standard errors calculated from the analytical variance developed in section 2.2 and the estimated true variance calculated as $\frac{1}{R}\sum_{r=1}^{R}(\hat{N}_{GC}^{r} - \bar{N}_{GC})^2$. $\hat{N}_{GC}^{r}$ is the GC estimate in the $r$th simulation run and $R$ is the total number of replications. The ratio was 1.01 for population size larger than 1000, which indicated a good approximation to the true variance.

Table 2.2: Comparison of the empirical and analytical standard errors from the estimates for the sample generated from $Y_i \sim Po(\mathrm{e}^{-0.02X_{1i}+0.03X_{2i}})$ model with covariates $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ and both auxiliary variables used in the model fitting.

| $N$ | Empirical $SE(\hat{N}_{GC})$ | Analytical $SE(\hat{N}_{GC})$ | $\frac{AnalyticalSE_{GC}}{EmpSE_{GC}}$ |
|---|---|---|---|
| 200 | 27.36 | 28.87 | 1.055 |
| 500 | 35.92 | 37.59 | 1.047 |
| 1000 | 49.20 | 49.79 | 1.012 |
| 2000 | 68.20 | 68.91 | 1.010 |
| 5000 | 106.21 | 107.45 | 1.011 |

Table 2.3: Point estimates and standard errors for the sample generated from $Y_i \sim Po(\mathrm{e}^{-0.02X_{1i}+0.03X_{2i}})$ model with covariates $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ independent. Both covariates are included in the estimation process.

| $N$ | GC | ZTP | Turing | Chao |
|---|---|---|---|---|
| | | $\hat{N}$ (SE) | | |
| 200 | 209.47 (28.87) | 203.98 (16.94) | 192.99 (13.70) | 198.12 (1.06) |
| 500 | 509.71 (37.59) | 503.34 (25.21) | 480.19 (21.37) | 490.29 (30.15) |
| 1000 | 1007.64 (49.79) | 1003.11 (34.99) | 960.75 (30.06) | 979.43 (42.22) |
| 2000 | 2008.46 (68.91) | 2002.74 (49.11) | 1919.07 (43.23) | 1954.05 (59.03) |
| 5000 | 5007.80 (107.45) | 5002.69 (77.29) | 4798.39 (66.97) | 4886.51 (93.22) |
| | | RMSE (x 100) | | |
| 200 | 2.1788 | 0.7329 | 0.6027 | 1.0144 |
| 500 | 0.5705 | 0.2615 | 0.3375 | 0.4032 |
| 1000 | 0.2594 | 0.1278 | 0.2534 | 0.2277 |
| 2000 | 0.1188 | 0.0624 | 0.2086 | 0.1404 |
| 5000 | 0.0464 | 0.0244 | 0.1809 | 0.0876 |
| | | RBias | | |
| 200 | 0.0474 | 0.0199 | -0.0351 | -0.0094 |
| 500 | 0.0194 | 0.0067 | -0.0396 | -0.0194 |
| 1000 | 0.0076 | 0.0031 | -0.0393 | -0.0206 |
| 2000 | 0.0042 | 0.0014 | -0.0405 | -0.0230 |
| 5000 | 0.0016 | 0.0005 | -0.0403 | -0.0227 |

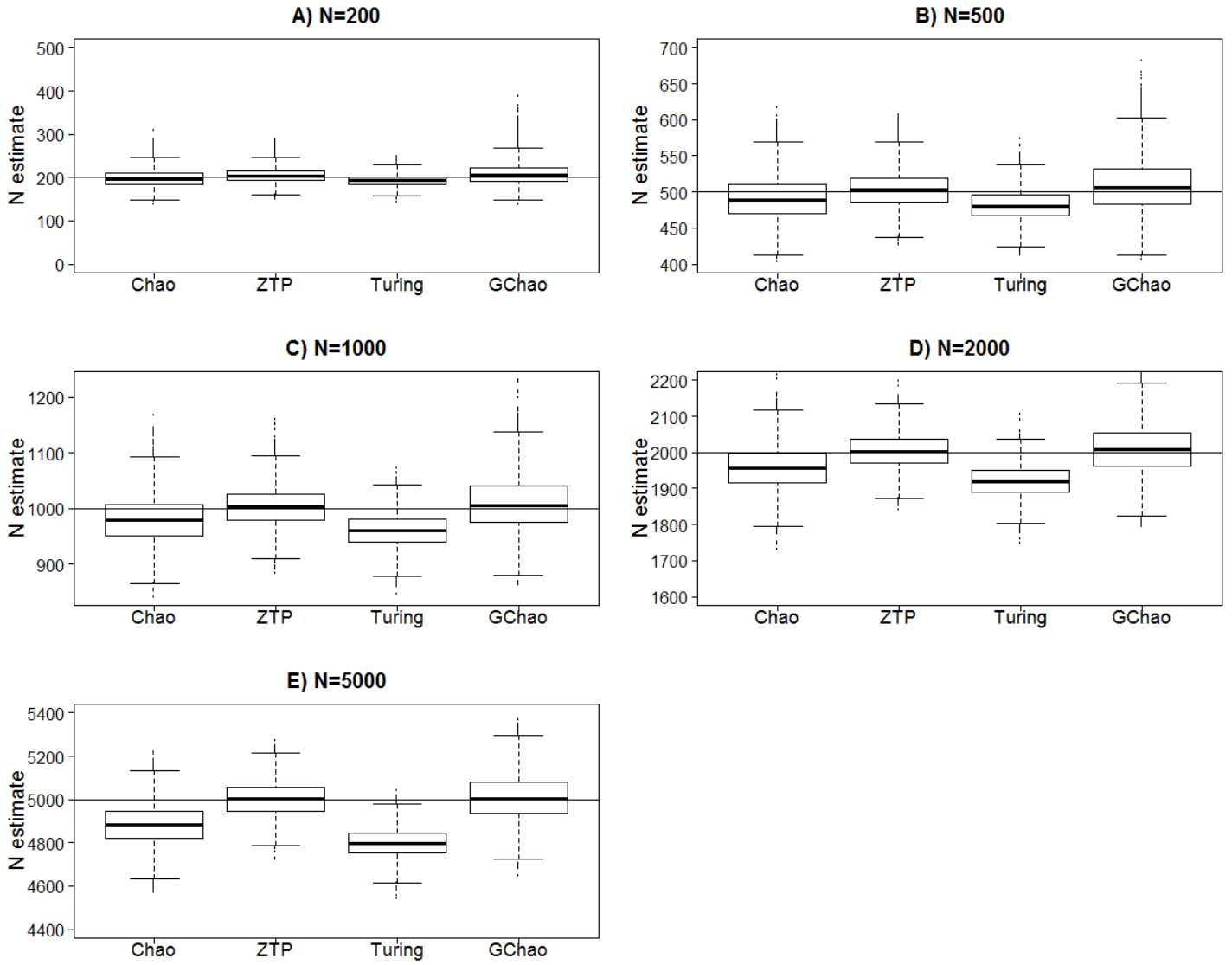


Figure 2.1: Boxplots of $\hat{N}$ for the scenario with $Y \sim Po(\mathrm{e}^{-0.02X_1+0.03X_2})$ with covariates $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ and both covariates used in the estimation process. A) $N = 200$, B) $N = 500$, C) $N = 1000$, D) $N = 2000$, E) $N = 5000$.

### 2.3.2 Simulation 2: Including unexplained heterogeneity

The data are generated applying the same model with the two covariates presented in the previous section 2.3.1.1. However, the capture-recapture estimates are calculated from a model where $X_1$ is the only covariate involved in the estimation. We are interested in assessing the impact of unexplained heterogeneity as found in real life problems. 5000 repetitions of a population are generated for each scenario of interest and are summarised calculating the mean of all $R$ repetitions ($\hat{N} = \frac{1}{R}\sum_{j=1}^{R}\hat{N}_j$).

Once the assumptions for the ZTP estimator do not hold, the estimator underperforms compared to classical Chao's and GC estimators. Turing, as previously seen, underestimates severely because its assumption of homogeneity is not fulfilled. GC estimator presents a good performance although it underestimates the population size because of the additional heterogeneity coming from $X_2$. GC estimator becomes the best estimator for sufficiently large samples when looking at the relative mean squared error (table 2.5). Chao's estimator also presents smaller RMSE than ZTP. We observe that the bias of the estimator increase asymptotically, leading to an increase in the relative bias; in contrast to a decrease of the variance which causes the RMSE values to decrease asymptotically.

The standard errors of our estimator were about 1.5 and 1.05 times the standard errors of the ZTP estimator and Chao's estimator. When comparing Chao's and ZTP standard errors, Chao's estimates present higher variability because it uses less individuals due to the right-truncation. We observe the same result in the comparison between ZTP and GC estimators.

Table 2.4: Standard error estimates for the model presented in section 2.3.1.1 using only $X_1$ as covariate

| $N$ | Empirical $SE(\hat{N}_{GC})$ | Analytical $SE(\hat{N}_{GC})$ | $\frac{AnalyticalSE_{GC}}{EmpSE_{GC}}$ |
|---|---|---|---|
| 200 | 21.84 | 22.94 | 1.050 |
| 500 | 31.34 | 32.53 | 1.038 |
| 1000 | 44.54 | 44.71 | 1.004 |
| 2000 | 61.72 | 62.44 | 1.012 |
| 5000 | 97.50 | 97.81 | 1.003 |

Table 2.5: Point estimates of $\hat{N}$ for the scenario with $Y_i \sim Po(\mathrm{e}^{-0.02X_{1i}+0.03X_{2i}})$ with covariates $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ and the estimation process based only in $X_1$.

| $N$ | GC | ZTP | Turing | Chao |
|---|---|---|---|---|
| | | $\hat{N}$ (SE) | | |
| 200 | 202.00 (22.94) | 195.60 (14.05) | 193.03 (13.70) | 197.91 (19.95) |
| 500 | 496.11 (32.53) | 484.06 (21.27) | 480.82 (21.53) | 490.86 (30.22) |
| 1000 | 987.08 (44.71) | 966.40 (29.77) | 960.74 (30.04) | 979.93 (42.22) |
| 2000 | 1970.71 (62.44) | 1930.65 (41.90) | 1919.92 (43.23) | 1955.70 (59.21) |
| 5000 | 4919.52 (97.81) | 4822.45 (65.96) | 4798.77 (67.75) | 4877.22 (93.25) |
| | | RMSE (x 100) | | |
| 200 | 1.183 | 0.547 | 0.591 | 0.973 |
| 500 | 0.415 | 0.293 | 0.338 | 0.402 |
| 1000 | 0.205 | 0.204 | 0.243 | 0.214 |
| 2000 | 0.115 | 0.167 | 0.207 | 0.137 |
| 5000 | 0.064 | 0.144 | 0.181 | 0.087 |
| | | RBias | | |
| 200 | 0.010 | -0.022 | -0.035 | -0.012 |
| 500 | -0.008 | -0.032 | -0.038 | -0.018 |
| 1000 | -0.013 | -0.034 | -0.039 | -0.020 |
| 2000 | -0.015 | -0.035 | -0.040 | -0.022 |
| 5000 | -0.016 | -0.036 | -0.040 | -0.025 |

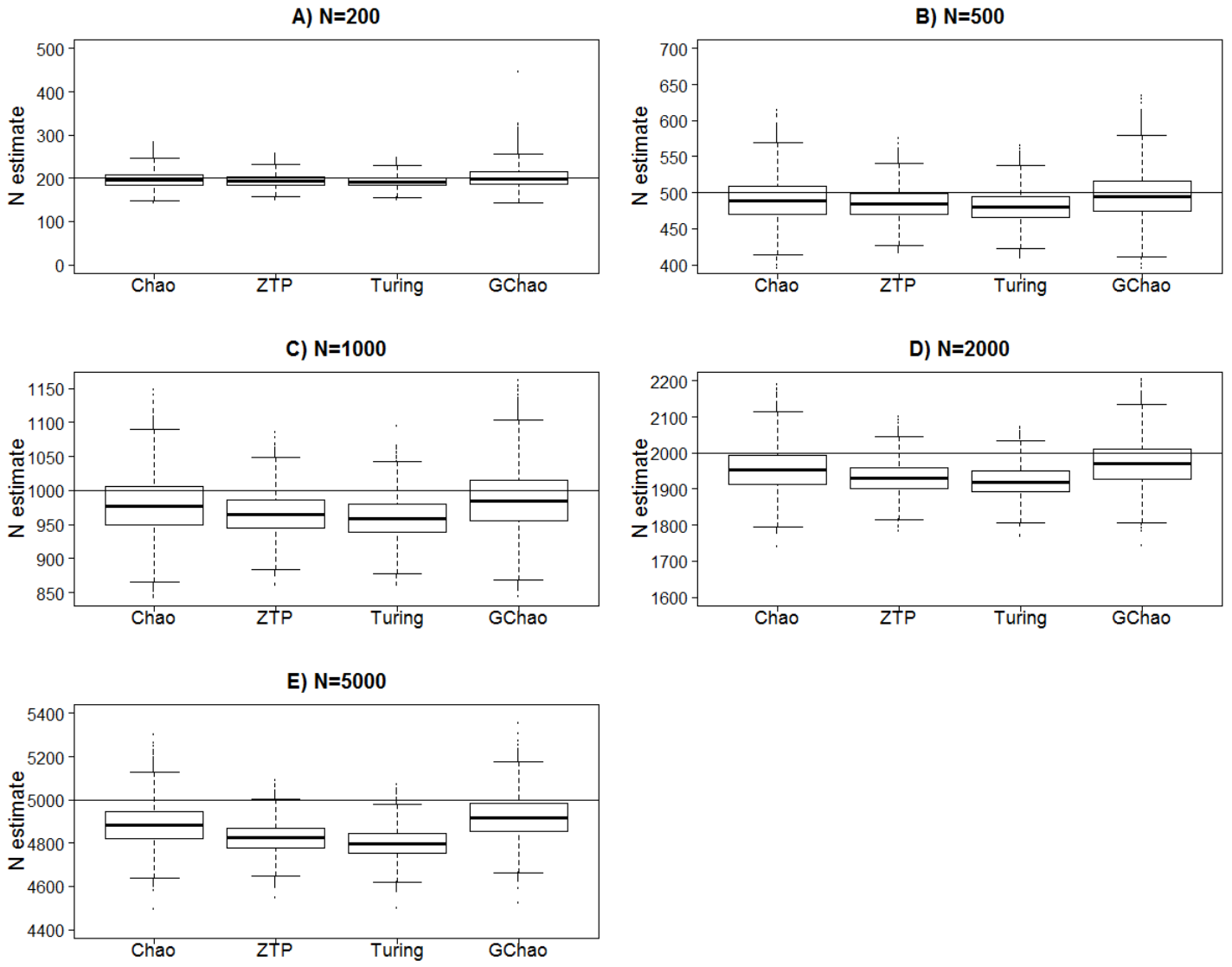


Figure 2.2: Boxplots of $\hat{N}$ for the scenario with $Y_i \sim Po(\mathrm{e}^{-0.02X_{1i}+0.03X_{2i}})$ with covariates $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ and $X_1$ used only in the estimation process. A) $N = 200$, B) $N = 500$, C) $N = 1000$, D) $N = 2000$, E) $N = 5000$.

### 2.3.3 Simulation 3: Poisson with contamination

A binary auxiliary variable $Z_i$ is used to split the data in two populations. The capture distribution is simulated from a Poisson distribution $Y_i \sim Po(\mathrm{e}^{\alpha+\beta' Z_i})$. The normal population is defined with $Z_i = 0$ and $\mathrm{e}^{\alpha} = 0.5$. A contaminated part is included with $Z_i = 1$ and $\mathrm{e}^{\alpha+\beta} = 3$. Two proportions are chosen for the contaminated part, $0.5 \times N$ and $0.1 \times N$, with true population sizes 500, 1000, 2000 and 5000. Each scenario is repeated 5000 times and average estimates are calculated to obtain the final summary.

Chao's and Turing's estimator underestimate the true value $N$, although Chao's estimation, as seen before, copes with some heterogeneity and provides a relatively good lower bound estimation for the scenario with a small contamination (10%) (Figure 2.3.3). All estimators using the complete covariate information obtain accurate estimates as expected. The ZTP estimator is again the best based on the RMSE because of its smaller variance. The ZNB estimator also presents slightly better RMSE than the GC estimator although it only converges in half of the repetitions. The conclusion that truncation leads to higher variability is highlighted again because of the smaller number of individuals used. However, ZTP and GC are similar based on the relative bias and both are slightly better than ZNB estimator because of the convergence problems of this estimator (Table 2.6).

Table 2.6: Point estimates, RMSE and relative bias for a capture-recapture Poisson distribution with contamination: $Y_i \sim Po(\mathrm{e}^{\alpha+\beta' Z_i})$ with $i = 1, .., N$. $Y_i \sim Po(0.5)$ for $Z_i = 0$ and $Y_i \sim Po(3)$ for $Z_i = 1$. Two scenarios based on the probability of $P(Z_i = 1) = 0.5$ and $P(Z_i = 1) = 0.1$

| $N$ | $\hat{N}_{GC}(SE_{\hat{N}_{GC}})$ | $\hat{N}_{ZTP}(SE_{\hat{N}_{ZTP}})$ | $\hat{N}_{Turing}(SE_{\hat{N}_{Turing}})$ | $\hat{N}_{Chao}(SE_{\hat{N}_{Chao}})$ | $\hat{N}_{ZNB}(SE_{\hat{N}_{ZNB}})$ |
|---|---|---|---|---|---|
| | | | $P(Z_i = 1) = 0.5$ | | |
| 500 | 510.099 (59.661) | 507.454 (46.205) | 385.937 (12.675) | 422.968 (21.705) | 515.793 (55.294) |
| 1000 | 1010.196 (78.077) | 1006.764 (61.464) | 771.462 (17.962) | 844.037 (30.185) | 1018.610 (64.035) |
| 2000 | 2011.925 (106.301) | 2010.498 (84.958) | 1544.227 (25.841) | 1688.222 (42.396) | 2024.900 (87.977) |
| 5000 | 5011.728 (164.506) | 5008.882 (131.813) | 3858.160 (39.951) | 4215.017 (66.613) | 5034.720 (136.663) |
| | | | $P(Z_i = 1) = 0.1$ | | |
| 500 | 512.404 (59.661) | 510.188 (33.208) | 484.554 (7.461) | 489.561 (10.437) | 514.584 (33.799) |
| 1000 | 1013.657 (64.157) | 1009.105 (40.474) | 951.486 (10.487) | 959.956 (14.682) | 1015.120 (39.539) |
| 2000 | 2011.473 (58.954) | 2007.910 (43.088) | 1903.058 (15.008) | 1919.414 (20.855) | 2015.960 (45.315) |
| 5000 | 5009.335 (82.694) | 5006.090 (62.563) | 4757.531 (23.748) | 4797.306 (32.807) | 5018.060 (63.824) |

| $N$ | $\hat{N}_{GC}$ | $\hat{N}_{ZTP}$ | $\hat{N}_{Turing}$ | $\hat{N}_{Chao}$ | $\hat{N}_{ZNB}$ | $\hat{N}_{GC}$ | $\hat{N}_{ZTP}$ | $\hat{N}_{Turing}$ | $\hat{N}_{Chao}$ | $\hat{N}_{ZNB}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P(Z_i = 1) = 0.5$ | | | | | $P(Z_i = 1) = 0.1$ | | | | |
| | RMSE | | | | | | | | | |
| 500 | 1.392 | 0.853 | 5.268 | 2.583 | 1.051 | 0.655 | 0.483 | 0.255 | 0.198 | 0.542 |
| 1000 | 0.630 | 0.385 | 5.255 | 2.535 | 0.430 | 0.331 | 0.172 | 0.246 | 0.182 | 0.179 |
| 2000 | 0.292 | 0.188 | 5.210 | 2.482 | 0.205 | 0.082 | 0.047 | 0.240 | 0.173 | 0.058 |
| 5000 | 0.107 | 0.068 | 5.222 | 2.485 | 0.076 | 0.027 | 0.016 | 0.237 | 0.169 | 0.018 |
| | RBias | | | | | | | | | |
| 500 | 0.020 | 0.015 | -0.228 | -0.154 | 0.032 | 0.025 | 0.020 | -0.031 | -0.021 | 0.029 |
| 1000 | 0.010 | 0.007 | -0.229 | -0.156 | 0.019 | 0.014 | 0.009 | -0.049 | -0.040 | 0.015 |
| 2000 | 0.006 | 0.005 | -0.228 | -0.156 | 0.013 | 0.006 | 0.004 | -0.048 | -0.040 | 0.008 |
| 5000 | 0.002 | 0.002 | -0.228 | -0.157 | 0.007 | 0.002 | 0.001 | -0.048 | -0.041 | 0.004 |

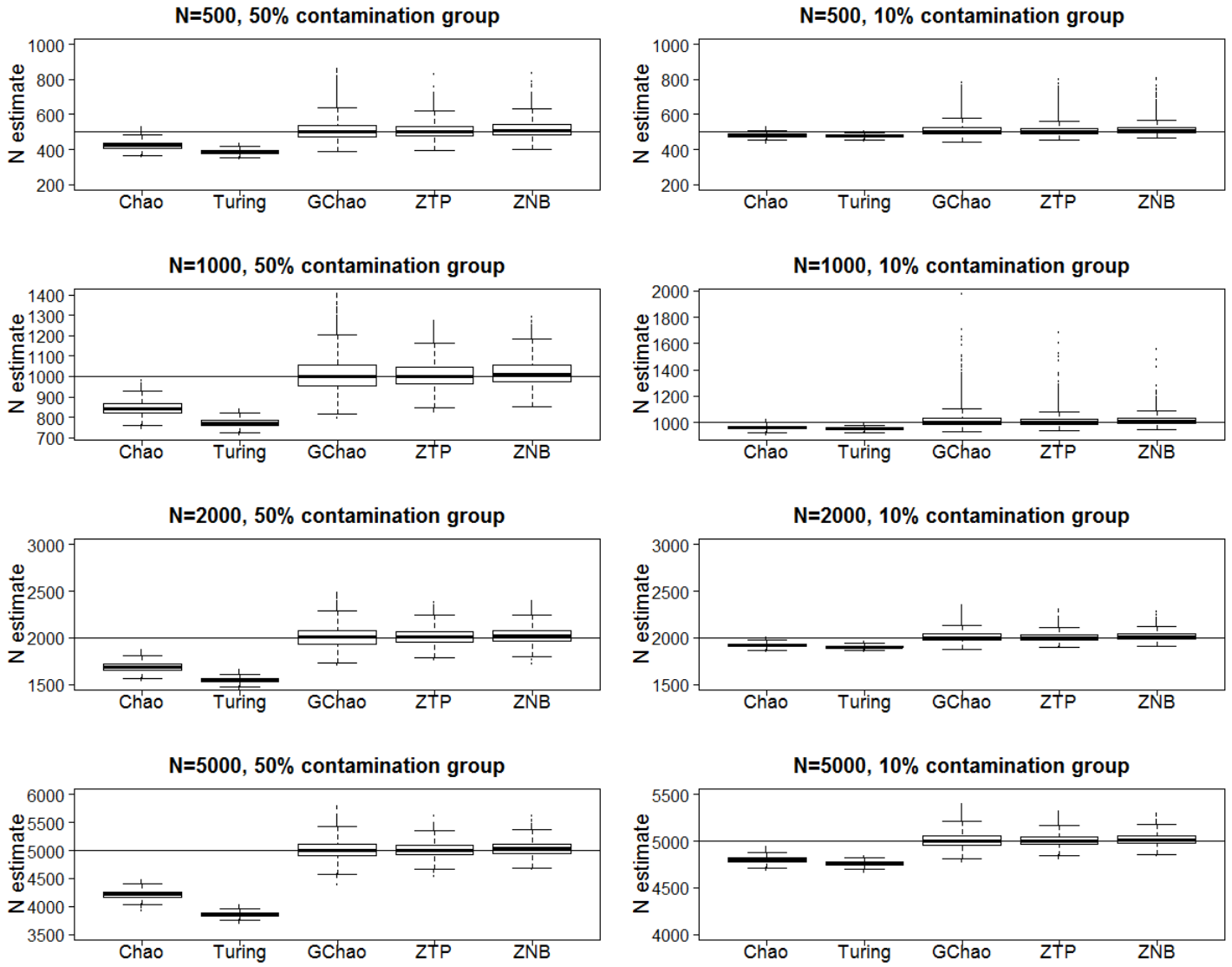


Figure 2.3: Boxplots for the simulated capture-recapture Poisson distribution with contamination. Horizontal line indicates the true population size of the scenario. Two scenarios: size of contamination group 50% (left side) or 10% (right side)

### 2.3.4 Simulation 4: Model with misclassification

In this section we evaluate a situation where we misclassify individuals. We aim to assess the impact of having wrong information in our covariates.

For this simulation we use the model in the third experiment $Y_i \sim Po(\mathrm{e}^{\alpha+\beta Z_i})$ with $Z_i$ as binary covariate. In this case the size of the first group is defined as $0.45 \times N$ at the time of generating the data, however the calculations of the estimates are made assuming that 10% and 20% of the individuals of the total population are misclassified being considered from the first component rather than the second component. Each scenario is repeated 5000 times and average estimates are calculated.

The results (table 2.7) show that all estimators underestimate the true population size, but GC was the least biased. Chao's and Turing's estimates are not affected by the misclassification because they do not use covariate information. In fact the impact of the misclassification in the ZNB and ZTP estimators makes them inferior to Chao's estimator in this scenario. The ZNB model converges on this occasion and its seem to produce slightly better estimates than the ZTP estimator

Our estimator appears to be robust despite introducing wrong information into the logistic model. Part of the distribution of GC estimates contains the true value (Figure 2.3.4) and there is a clear negative effect when the proportion of misclassified individuals increases. On the basis of RMSE and relative bias GC is superior in this particular scenario and it is less sensitive to contamination in the covariate information than the other estimators.

Table 2.7: Point estimates, RMSE and relative bias for a fitted model with misclassified observations $Y_i \sim Po(\mathrm{e}^{\alpha+\beta Z_i})$ with $i = 1, .., N$. $Y_i \sim Po(0.5)$ for $Z_i = 0$ and $Y_i \sim Po(3)$ for $Z_i = 1$. The probability of $P(Z_i = 1) = 0.45$. Two scenarios were generated with 10% and 20% of the population misclassified.

| $N$ | $\hat{N}_{GC}(SE_{\hat{N}_{GC}})$ | $\hat{N}_{ZTP}(SE_{\hat{N}_{ZTP}})$ | $\hat{N}_{Turing}(SE_{\hat{N}_{Turing}})$ | $\hat{N}_{Chao}(SE_{\hat{N}_{Chao}})$ | $\hat{N}_{ZNB}(SE_{\hat{N}_{ZNB}})$ |
|---|---|---|---|---|---|
| 10% misclassified individuals | | | | | |
| 500 | 468.219 (35.837) | 400.913 (14.163) | 396.000 (12.030) | 427.099 (20.627) | 415.66 (17.193) |
| 1000 | 929.164 (47.908) | 800.777 (19.968) | 791.596 (17.089) | 852.123 (29.326) | 829.499 (23.826) |
| 2000 | 1855.799 (66.441) | 1602.074 (28.150) | 1584.467 (24.115) | 1705.119 (41.237) | 1657.81 (33.59) |
| 20% misclassified individuals | | | | | |
| 500 | 449.485 (28.010) | 390.430 (11.967) | 395.851 (11.989) | 426.839 (20.942) | 412.547 (17.097) |
| 1000 | 895.159 (38.207) | 780.624 (16.747) | 791.885 (16.713) | 852.591 (28.798) | 824.161 (23.756) |
| 2000 | 1787.892 (53.248) | 1561.768 (23.733) | 1584.36 (23.931) | 1704.258 (41.626) | 1647.230 (34.296) |

| $N$ | $\hat{N}_{GC}$ | $\hat{N}_{ZTP}$ | $\hat{N}_{Turing}$ | $\hat{N}_{Chao}$ | $\hat{N}_{ZNB}$ | $\hat{N}_{GC}$ | $\hat{N}_{ZTP}$ | $\hat{N}_{Turing}$ | $\hat{N}_{Chao}$ | $\hat{N}_{ZNB}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 10% misclassified individuals | | | | | 20% misclassified individuals | | | | |
| RMSE | | | | | | | | | | |
| 500 | 0.893 | 4.008 | 4.384 | 2.296 | 2.963 | 1.339 | 4.859 | 4.396 | 2.316 | 3.176 |
| 1000 | 0.732 | 4.009 | 4.372 | 2.273 | 2.964 | 1.245 | 4.841 | 4.359 | 2.256 | 3.148 |
| 2000 | 0.632 | 3.978 | 4.331 | 2.216 | 2.956 | 1.200 | 4.815 | 4.333 | 2.230 | 3.140 |
| RBias | | | | | | | | | | |
| 500 | -0.064 | -0.198 | -0.208 | -0.146 | -0.169 | -0.101 | -0.219 | -0.208 | -0.146 | -0.175 |
| 1000 | -0.071 | -0.199 | -0.208 | -0.148 | -0.171 | -0.105 | -0.219 | -0.208 | -0.147 | -0.176 |
| 2000 | -0.072 | -0.199 | -0.208 | -0.147 | -0.171 | -0.106 | -0.219 | -0.208 | -0.148 | -0.176 |

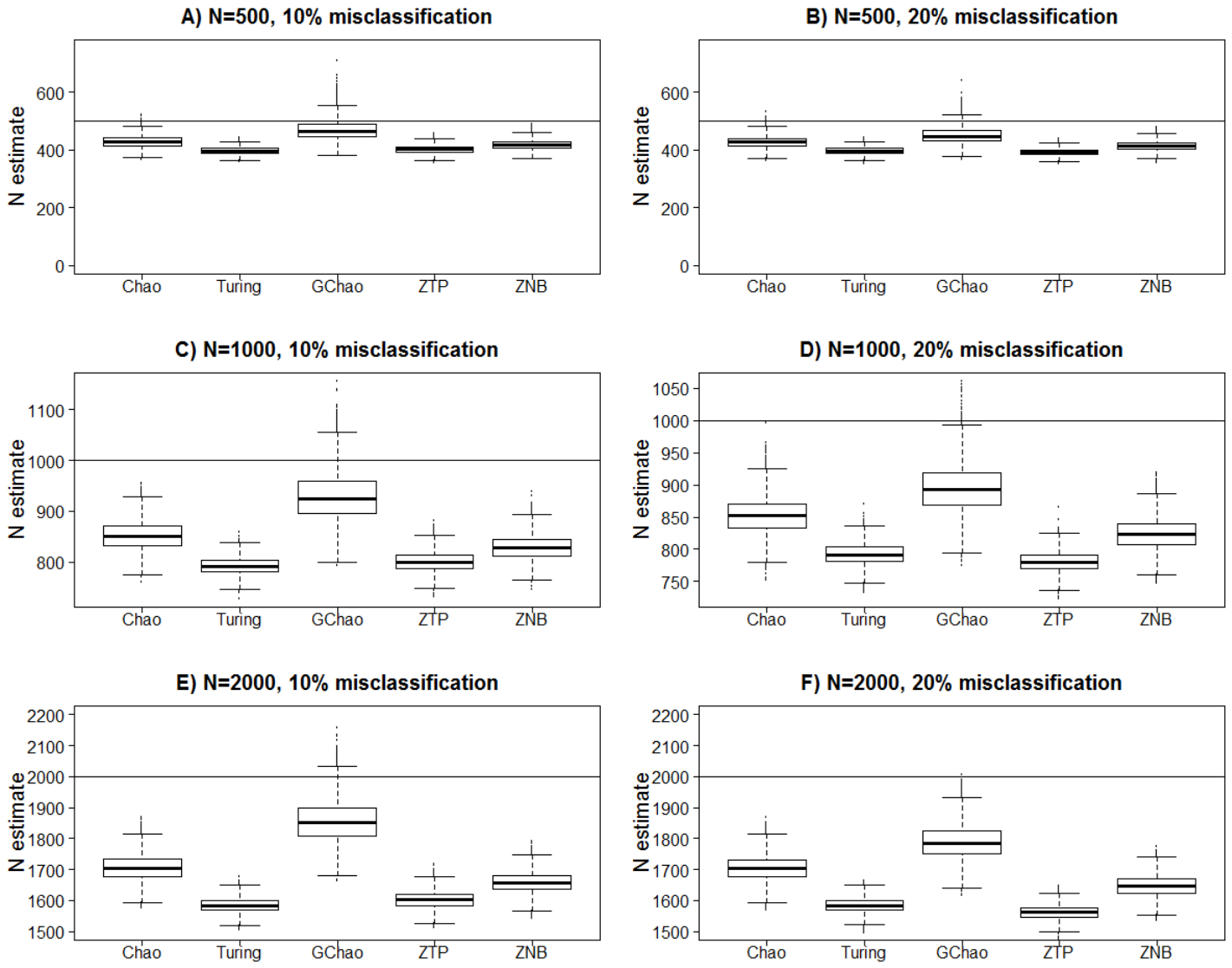


Figure 2.4: Two main scenarios with 10% and 20% misclassified individuals in the population. A) $N = 500$ and 10% misclassified individuals, B) $N = 500$ and 20% misclassified individuals, C) $N = 1000$ and 10% misclassified individuals, D) $N = 1000$ and 20% misclassified individuals, E) $N = 2000$ and 10% misclassified individuals, F) $N = 2000$ and 20% misclassified individuals

### 2.3.5 Simulation 5: Data generated from a negative binomial distribution

We evaluate the performance of our estimator under heterogeneity generated by a negative binomial distribution. The capture-recapture distribution $Y_i|Z_i \sim NB(\mu_i, \theta)$ with $\mu_i = \mathrm{e}^{\alpha+\beta' Z_i}$, $\alpha = 0, \beta = 0.02, \theta = 3$ and $Z_i \sim N(8, 25)$. Zero counts are removed and the remain counts represent the sampling distribution used to calculate the estimates.

The ZNB estimator presents the best relative bias because of its assumptions hold for this experiment (table 2.8). GC is moderately underestimating (figure 2.3.5) although its standard error is about 0.56 times the standard error of the ZNB estimate. The RMSE and relative bias show that ZNB estimator is asymptotically unbiased in comparison to the GC estimator where $E(\hat{N})/N \approx 0.9$ for all population sizes included in the simulation. The classical Chao's estimator performs similarly to the GC estimator in spite of not using any covariate information. The ZTP estimator severely underestimates the true value, performing even worse than Turing estimator.

ZNB estimator works well under the assumption that the capture distribution is a negative binomial, but in other circumstances the zero-truncated negative binomial model tends to have convergence problems like in the first experiment where the ZNB estimator could not be reported.

Table 2.8: Comparison of capture-recapture estimates for captures following a Negative Binomial distribution $Y_i|Z_i \sim NB(\mu_i, \theta)$ with $\mu_i = \mathrm{e}^{0.02Z_i}$, $Z_i \sim N(8, 25)$ and $\theta = 3$

| $N$ | GC | ZTP | Turing | Chao | ZNB |
|---|---|---|---|---|---|
| | | | $\hat{N}$ (SE) | | |
| 200 | 187.02 (22.29) | 167.96 (12.47) | 171.42 (12.93) | 183.50 (20.01) | 208.80 (46.45) |
| 500 | 459.40 (31.80) | 417.73 (19.22) | 428.34 (20.03) | 455.95 (30.18) | 504.98 (58.02) |
| 1000 | 913.06 (43.84) | 833.78 (27.09) | 855.85 (28.41) | 909.04 (41.98) | 1005.80 (78.72) |
| 2000 | 1819.01 (60.35) | 1664.40 (37.55) | 1709.59 (39.31) | 1813.60 (58.75) | 2003.66 (107.31) |
| 5000 | 4542.05 (94.85) | 4159.92 (58.87) | 4274.65 (61.95) | 4532.13 (92.56) | 5005.08 (168.75) |
| | | | RMSE (x 100) | | |
| 200 | 1.644 | 2.881 | 2.397 | 1.674 | 5.587 |
| 500 | 1.061 | 2.845 | 2.210 | 1.158 | 1.356 |
| 1000 | 0.953 | 2.834 | 2.158 | 1.017 | 0.623 |
| 2000 | 0.916 | 2.850 | 2.148 | 0.963 | 0.288 |
| 5000 | 0.881 | 2.851 | 2.129 | 0.915 | 0.114 |
| | | | RBias | | |
| 200 | -0.065 | -0.160 | -0.143 | -0.083 | 0.044 |
| 500 | 0.099 | -0.165 | -0.143 | -0.088 | 0.010 |
| 1000 | -0.087 | -0.166 | -0.144 | -0.091 | 0.006 |
| 2000 | -0.091 | -0.168 | -0.145 | -0.093 | 0.002 |
| 5000 | -0.092 | -0.168 | -0.145 | -0.094 | 0.001 |

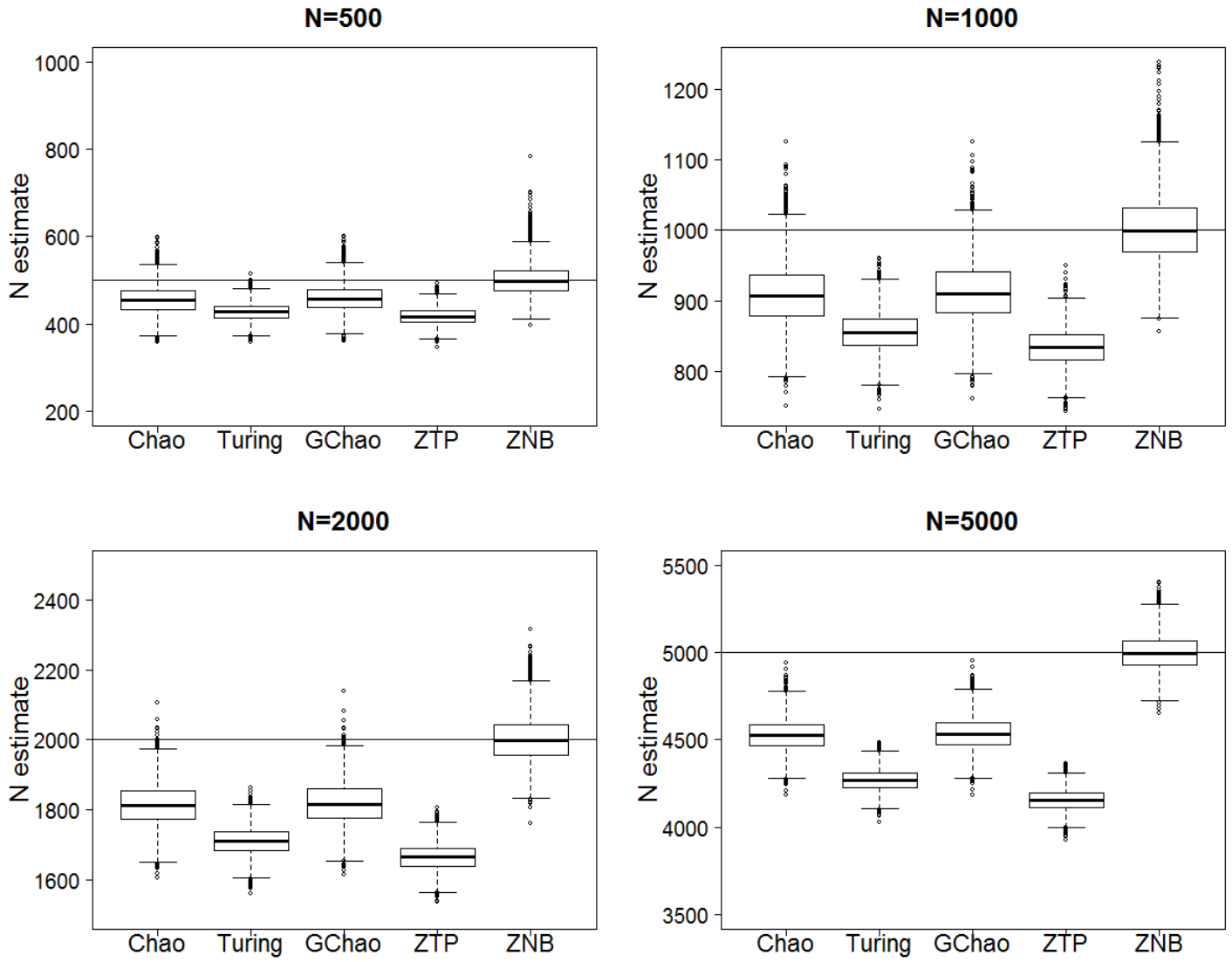


Figure 2.5: Simulation based on a negative binomial $Y_i|Z_i \sim NB(\mu_i, \theta)$ with $\mu_i = \mathrm{e}^{0.02Z_i}$, $Z_i \sim N(8, 25)$ and $\theta = 3$. Horizontal line indicates the true population size of the scenario. A) $N = 500$ B) $N = 1000$ C) $N = 2000$ D) $N = 5000$.

## 2.4 Case studies

### 2.4.1 Carcass submission from animal farms in Great Britain

Private veterinary surgeons (PVS) regularly send animal submissions to the Animal Health and Veterinary Laboratories Agency (AHVLA) to determine the cause of death based on a post-mortem test, to test an animal sample to confirm a disease or to find out whether an animal needs further testing. The PVS might choose to submit or not submit a sample depending on the disease. Only notifiable diseases are compulsory to investigate and report to the authorities. AHVLA could miss submissions for several reasons, for instance, a PVS might have facilities to run some diagnostic tests, he/she might not submit a sample because there is history of a confirmed disease in the farm and the animal presents similar symptoms. The cost is also an important factor, as farmers might not even call a PVS when they believe that the disease is not going to spread to other animals. In fact, the Department of Food and Rural Affairs (DEFRA) used to subsidise some diagnostic tests but the current economic climate is leading to move all costs to farmers.

Our objective is to evaluate the completeness of the farm submissions in Great Britain to understand which proportion of the general picture is being explained. In 2009, the number of farms with cattle was estimated to be 60,571 farms. 48,535 of those farms did not have any submissions that year. From those 12,036 farms that submitted we aim to estimate the total number of farms with unknown disease that did not submit.

Three risk factors related to animal submissions were identified in previous studies carried out at AHVLA: holding type (beef or dairy), holding size and distance to the regional labs. Large holdings are expected to have a larger submission rate because of the potential costs involved if the disease spreads within the farm and their financial resources. The distance from the farm to the closest regional lab is also specially important for carcass samples, because farmers are obliged to cover delivery costs to the regional lab. On the positive side, a carcass sample has higher probability of identifying the disease. In this problem the re-capture comes from the second or more submissions from the same farm, so the dependent variable is the number of submissions from each farm.

The total number of carcass submissions and the total number of submissions including other types of samples (like blood or faecal samples) are the primary endpoints. Table 2.9 contains the data in the format of frequency of frequencies. A ratio plot (Rocchetti et al., 2011; Böhning et al., 2013a) is produced to evaluate the existence of heterogeneity in the probability of submitting animal samples and to identify the right statistical distribution to model the capture-recapture probability. Table 2.10 contains the ratios and their 95% confidence limits. In our case, Figure 2.4.1 presents a structural heterogeneity, which

questions the use of an homogeneous Poisson and suggests the use of a heterogeneous distribution, such as the negative binomial distribution.

Table 2.9: Frequency distribution of number of farms submitting any type of samples (first row) and number of farms submitting carcass samples (second row) to AHVLA regional laboratories in 2009.

| $f_0$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10}$ | $f_{11+}$ | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 48535 | 6340 | 2520 | 1149 | 709 | 380 | 249 | 173 | 135 | 94 | 80 | 207 | 60571 |
| 58713 | 1532 | 231 | 51 | 27 | 6 | 5 | 2 | 1 | 3 | 0 | 0 | 60571 |

Table 2.10: Ratios ($r(x) = (x+1)f_{x+1}/f_x$) and confidence bands for the ratio plot (Figure 2.4.1).

| | Any sample | | Carcass | |
|---|---|---|---|---|
| ratio | $\hat{r}_x$ | $\hat{r}_x$ 95% CL | $\hat{r}_x$ | $\hat{r}_x$ 95% CL |
| $r_1$ | 0.26 | (0.25-0.27) | 0.05 | (0.05-0.05) |
| $r_2$ | 1.19 | (1.14-1.25) | 0.45 | (0.39-0.52) |
| $r_3$ | 1.82 | (1.70-1.96) | 0.88 | (0.65-1.20) |
| $r_4$ | 3.09 | (2.81-3.39) | 2.65 | (1.66-4.22) |
| $r_5$ | 3.22 | (2.84-3.64) | 1.33 | (0.55-3.23) |
| $r_6$ | 4.59 | (3.91-5.38) | 5.83 | (1.78-19.11) |
| $r_7$ | 5.56 | (4.58-6.75) | 3.20 | (0.62-16.49) |
| $r_8$ | 7.03 | (5.61-8.8) | 4.50 | (0.41-49.63) |
| $r_9$ | 6.96 | (5.35-9.06) | 30.00 | (3.12-288.42) |
| $r_{10}$ | 9.36 | (6.95-12.61) | | |
| $r_{11}$ | 31.05 | (23.99-40.19) | | |

The probability of submitting any type of animal samples is found significantly related to the holding size (log-scale) and the type of the farm (dairy or beef) (Table 2.11). In contrast, the total number of carcass submissions depends on the distance and the holding size (log-scale). Non-carcass samples (blood, faecal, etc...) represent a majority of the total samples which do not depend that much on the distance to the regional lab because they are normally sent by post. However, as mentioned previously, farmers need to cover the expensive cost of sending carcass samples to the regional labs.

There are large differences between estimates. The zero truncated Poisson model with covariates provides an estimate which is lower to the one provided by the conventional Chao's estimator that is proven to be a lower bound estimator. The Good-Turing estimator also underestimates due to the non-homogeneous captured probability. GC estimator is significantly larger than Chao's lower bound estimator. The percentage of farms detected with all type of submissions based on generalised Chao's estimate is between $\frac{12{,}036\times100}{22{,}429}(53.7\%)$ and $\frac{12{,}036\times100}{20{,}885}(57.6\%)$. The completeness of carcass submissions is between 21% to 28.5%, which suggests that a further investigation should be

carried out to find out the main causes of missing submissions to establish new policies for increasing submission rates.

Table 2.11: Results from the logistic regressions to obtain the Generalised Chao estimates. Chao, Zero-truncated Poisson and Turing estimates are also reported for total number of farms submitting any sample and total number of farms submitting carcass samples.

| Logistic regression analysis TOTAL NUMBER OF SUBMISSIONS | | | | |
|---|---|---|---|---|
| covariate | coef | SE-coef | Z | p-value |
| log-size | 0.33 | 0.03 | 12.5 | 0.00 |
| type(1=dairy 0=beef) | 0.29 | 0.05 | 5.55 | 0.00 |
| log-distance | -0.01 | 0.04 | -0.10 | 0.92 |
| Estimated farms submitting any animal sample (95% CI): | | | | |
| [based on TOTAL NUMBER OF SUBMISSIONS] | | | | |
| $n$ | G-Chao | Chao | Turing | ZTP |
| 12036 | 21657 | 20011 | 15532 | 18346 |
| | (20885,22429) | (17932,18760) | (15349,15716) | (19993,20029) |
| Logistic regression analysis NUMBER OF CARCASS SUBMISSIONS | | | | |
| covariate | coef | SE-coef | Z | p-value |
| log-size | 0.32 | 0.08 | 4.10 | 0.00 |
| type(1=dairy 0=beef) | 0.05 | 0.16 | 0.38 | 0.71 |
| log-distance | -0.15 | 0.09 | -1.66 | 0.09 |
| Estimated farms submitting carcasses (95% CI): | | | | |
| [based on NUMBER OF CARCASS SUBMISSIONS ONLY] | | | | |
| $n$ | G-Chao | Chao | Turing | ZTP |
| 1858 | 7688 | 6938 | 5279 | 6008 |
| | (6523, 8853) | (6868,7009) | (4645,5913) | (5293, 6723) |

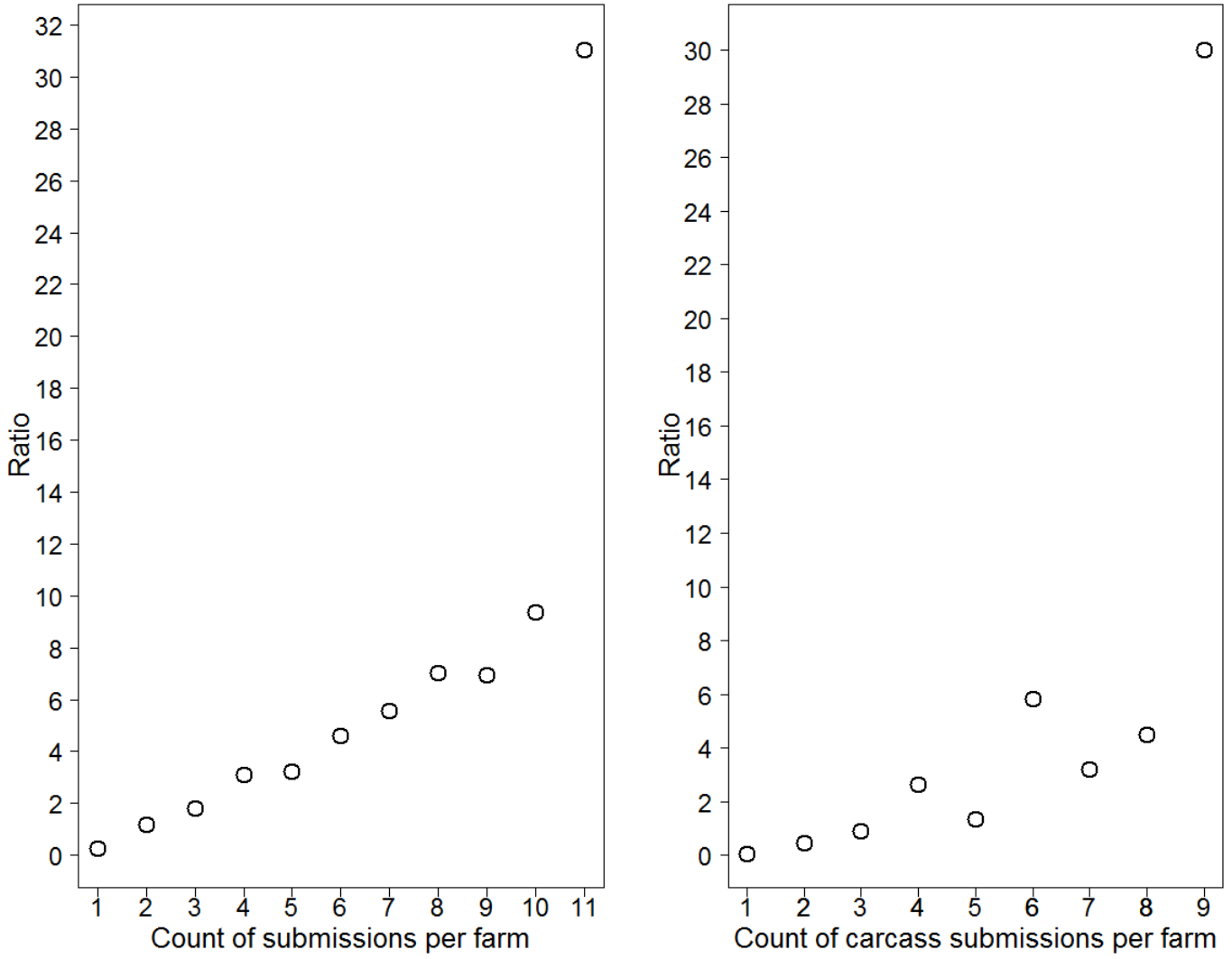


Figure 2.6: Ratio plot to investigate the presence of heterogeneity in the number of animal submissions and carcass submissions respectively. $r(x) = (x+1)f_{x+1}/f_x$

### 2.4.2 Drug users in Bangkok

Böhning estimated the size of the hidden population of heroin and methamphetamine drug users in Bangkok in 2001 and 2002 (Böhning et al., 2004). Ratio plots (Böhning et al., 2013a; Rocchetti et al., 2011) are firstly calculated to determine whether there is heterogeneity in which case the ratio plot also provides information about the type of heterogeneity (Figure 2.4.2). All graphs suggest a structural heterogeneity. Age and gender are chosen as covariates related to the probability of being identified in the registries. Both years are analysed independently to fulfil the assumption of closed populations. Individual logistic models are fitted for each drug and year with the outcome variable being the probability of appearing in the registries twice.

Both covariates are found to be significant in all models with the exception of age in the heroin model in 2002 (table 2.12). GC point estimates are consistently larger than the other estimates. However, its confidence intervals for both drugs in 2002 overlap with the confidence limits of classic Chao's estimator. Although Chao's confidence interval seems to be quite narrow in comparison to the confidence intervals of the other estimates. The zero-truncated Poisson estimator clearly underestimates in all scenarios with similar results to Turing's estimator.

Table 2.12: Results from the logistic regression models for the calculation of GC estimates, for both drugs and years

| Heroin drug users 2001 | | | | |
|---|---|---|---|---|
| covariate | coef | SE-coef | Z | p-value |
| Gender | -1.095 | 0.136 | -8.040 | <0.001 |
| Age | 0.014 | 0.004 | 3.720 | < 0.001 |
| **Heroin drug users 2002** | | | | |
| covariate | coef | SE-coef | Z | p-value |
| Gender | -0.839 | 0.141 | -5.964 | < 0.001 |
| Age | 0.005 | 0.005 | 1.136 | 0.256 |
| **Methamphetamine drug users 2001** | | | | |
| covariate | coef | SE-coef | Z | p-value |
| Gender | -1.369 | 0.314 | -4.356 | < 0.001 |
| Age | 0.038 | 0.007 | 5.694 | < 0.001 |
| **Methamphetamine drug users 2002** | | | | |
| covariate | coef | SE-coef | Z | p-value |
| Gender | -0.855 | 0.296 | -2.889 | 0.004 |
| Age | 0.029 | 0.008 | 3.682 | < 0.001 |

Table 2.13: Point estimates and asymptotic confidence limits of the number of heroin and metamphetamine drug users in Bangkok

| Case | $\hat{N}_{GC}$ (95% CL) | $\hat{N}_{ZTP}$ (95% CL) | $\hat{N}_{Turing}$ (95% CL) | $\hat{N}_{Chao}$ (95% CL) |
|---|---|---|---|---|
| Heroin 2001 | 10661 (10001,11320) | 7605 (7398,7813) | 7802 (7639,7965) | 9825 (9805,9845) |
| Heroin 2002 | 8121 (7677,8566) | 5885 (5760,6009) | 6202 (6075,6330) | 7805 (7786,7825) |
| Methamphetamine 2001 | 14539 (11205,17873) | 10990 (9397,12583) | 9483 (8123,10844) | 10967 (10879,11056) |
| Methamphetamine 2002 | 8390 (7002,9779) | 5808 (5248,6369) | 5944 (5159,6730) | 7498 (7421,7576) |

Table 2.14: Ratios and 95% confidence limits for the Bangkok drug users case study

| Case | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_6$ | $r_7$ | $r_8$ | $r_9$ |
|---|---|---|---|---|---|---|---|---|---|
| Heroin 2001 | 1.45 (1.37-1.53) | 2.1 (1.97-2.24) | 3.18 (2.96-3.43) | 3.68 (3.38-4.01) | 4.94 (4.49-5.44) | 4.82 (4.32-5.39) | 1.46 (1.18-1.82) | 5.06 (3.63-7.07) | 0 |
| Heroin 2002 | 1.5 (1.42-1.6) | 2.65 (2.48-2.83) | 3.39 (3.15-3.64) | 3.44 (3.16-3.73) | 3.52 (3.17-3.91) | 3.91 (3.4-4.49) | 4.57 (3.8-5.5) | 1.84 (1.29-2.63) | 2.78 (1.38-5.6) |
| Metha 2001 | 0.53 (0.48-0.58) | 0.85 (0.72-1.01) | 1.16 (0.84-1.6) | 1.56 (0.88-2.79) | 4.8 (2.25-10.25) | | | | |
| Metha 2002 | 0.41 (0.37-0.46) | 0.95 (0.8-1.12) | 1.84 (1.41-2.41) | 4.17 (2.84-6.11) | | | | | |

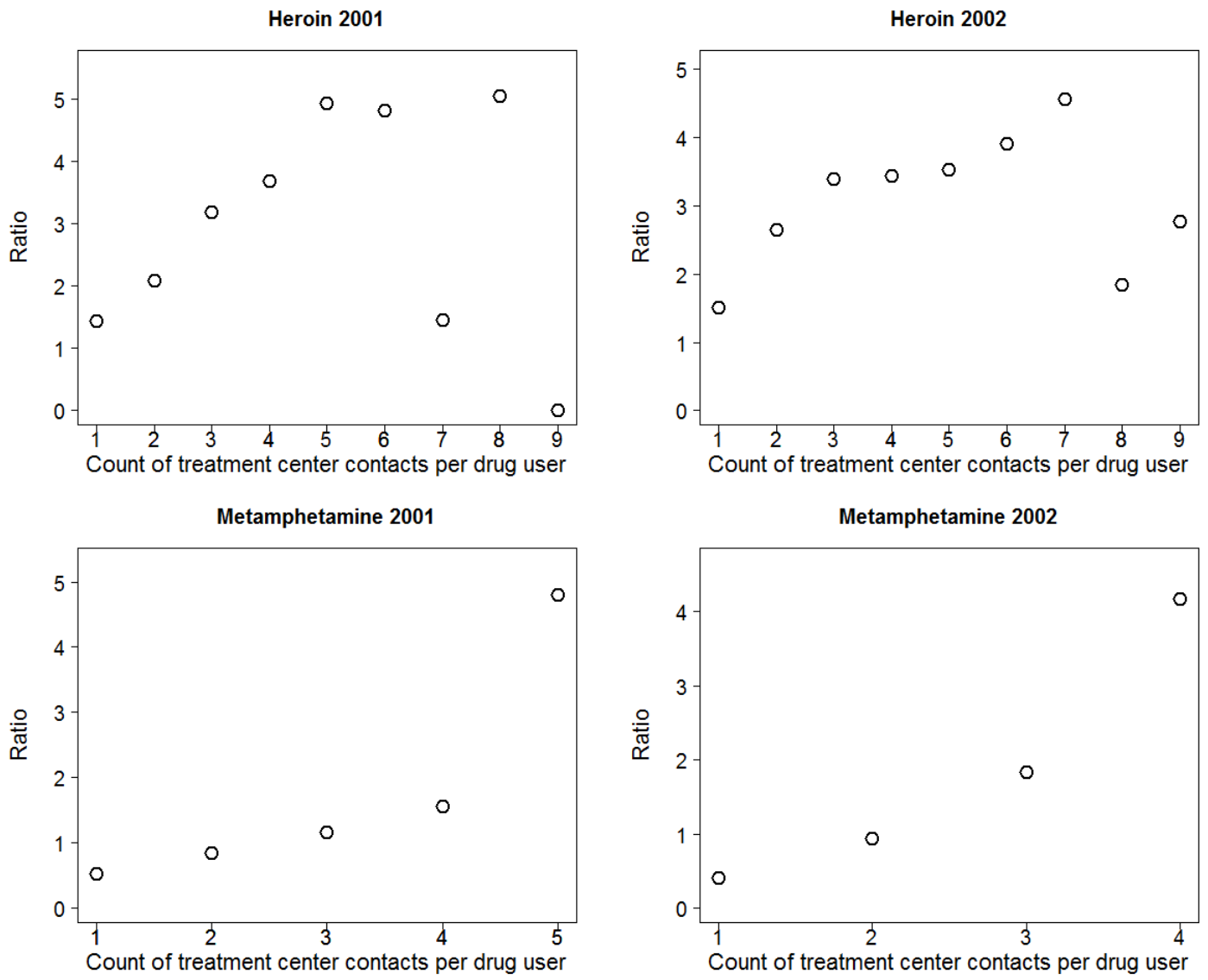


Figure 2.7: Ratio plots for Bangkok drug users case study. From upper left to lower right: Heroin users 2001, heroin users 2002, methamphetamine users 2001 and methamphetamine users 2002.

## 2.5 Conclusions

In this chapter, we have developed a framework to extend Chao's lower bound estimator to include covariate information of the observed units in order to model the heterogeneity of the capture-recapture probability and obtain less biased estimates. A key finding in the chapter was the proportionality between the likelihood of a truncated Poisson model with non-truncated counts of ones and twos and the likelihood of a binomial logistic regression model. Such relation provided a straightforward process to calculate our estimates with a standard statistical software. We also provided analytical formulae for the variance of the estimator that was proven to be very close to the empirical variance in our simulations.

The properties of our estimator were investigated by running simulations and comparing the generalised Chao estimator with other popular estimators. We initially showed that providing all covariates involved in the simulated heterogeneity of a Poisson distribution the zero-truncated Poisson estimator and the generalised Chao estimator obtained accurate estimations but the former estimator was better because its variability was smaller. Chao's estimator provided robust estimation for scenarios where the heterogeneity is not large. Turing's estimator presented the smallest variability because it is the closest to the maximum likelihood estimator assuming homogeneity. The second experiment, where contamination was introduced as a binary covariate, revealed a good performance of the generalised Chao estimator, although, on the basis of the relative mean squared error, the estimators from the zero-truncated Poisson and zero-truncated negative binomial model were slightly better due to their small variability.

Potential bias to test our estimator was introduced in three-fold: 1) Using part of the covariate information available when fitting the regression models to obtain our estimates of the population size; 2) generating a wrong covariate misclassifying part of the population; 3) generating data based on a negative binomial distribution. The zero-truncated Poisson estimator underperformed when the assumptions did not hold, performing sometimes even worse than the Turing estimator. The generalised Chao's estimator was robust across all scenarios and we highlighted the trade-off between the variance and the accuracy when using covariates and truncation. The zero-truncated negative binomial model could not be applied in all scenarios and it showed convergence problems in many cases.

In the following chapter, we develop the framework to consider different truncation cut-off points.

# Chapter 3

# Generalised Chao estimator considering all frequency counts

The generalised Chao estimator presented in the previous chapter was developed from a truncated Poisson distribution with only counts of ones and twos non-truncated. In this chapter we aim to extend the generalised Chao estimator by increasing the number of non-truncated counts to assess whether more information leads to more efficient estimators.

We initially develop the simplest case with 3 non-truncated counts and no covariates. Then we extend the estimate to $J$ counts without covariates ($J \geq 2$) and later we follow the same procedure to obtain the final estimate with $J$ counts and covariates. We also deduce an analytical formula for the variance and simulations are conducted to evaluate the impact of increasing the number of non-truncated counts.

## 3.1 Extension of Chao's estimator without covariate information

Two methods are applied to obtain estimates: a) The first method consists in using the EM algorithm (Dempster et al., 1977) based on a complete likelihood and latent variables. b) The second method is based on directly maximising the truncated likelihood using numerical optimisation algorithms like Nelder-Mead (Nelder and Mead, 1965) or BFGS (Broyden-Fletcher-Goldfarb-Shanno) (Broyden, 1969; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) algorithms which are provided under the R internal command *optim*.

### 3.1.1 Complete likelihood

The EM algorithm (Dempster et al., 1977) assumes that observed data represent only a part of so called complete data, where missing information should be considered as well. The EM algorithm consists of two stages, expectation and maximisation. In order to maximise the likelihood function, the posterior expectations of the complete data likelihood need to be estimated, but in order to estimate those expectations we need to obtain the estimation of the parameter of interest from the likelihood. Initial values of the parameter of interest are normally provided to start the iterative algorithm.

In our case, the complete likelihood for a Poisson distribution for $J$ non-truncated counts with $m$ being the maximum number of captures can be written as

$$\mathcal{L}(\lambda) = \prod_{j=0}^{m} p_j^{f_j}$$

where

$$p_j = \mathrm{e}^{-\lambda}\lambda^j/j!$$

is the probability of being captured $j$ times, while $f_j$ is the number of units captured exactly $j$ times. Therefore, the expected complete log-likelihood is defined as

$$\ell(\lambda) = e_0 \log(p_0) + f_1 \log(p_1) + ... + f_J \log(p_J) + e_{J+1} \log(p_{J+1}) + ... + e_m \log(p_m) \quad (3.1)$$

where $f_1, \ldots, f_J$ represent the observed frequencies of the non-truncated counts considered to obtain our estimate. Hence, the rest of the frequencies of counts are assumed to be unobserved and their expectations are used ($e_k = E(f_k|\lambda)$), $k \in \{0, J+1, \ldots, m\}$).

Replacing the probabilities in the likelihood we obtain

$$\begin{aligned}
\ell(\lambda) &= e_0 \log(\mathrm{e}^{-\lambda}) + f_1 \log(\mathrm{e}^{-\lambda}\lambda) + ... + f_J \log(\mathrm{e}^{-\lambda}\lambda^J/J!) + e_{J+1} \log(\mathrm{e}^{-\lambda}\lambda^{J+1}/(J+1)!) \\
&\quad + ... + e_m \log(\mathrm{e}^{-\lambda}\lambda^m/m!) \\
&= -\lambda(e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m) \\
&\quad + \log(\lambda)(f_1 + 2f_2 + ... + Jf_J + (J+1)e_{J+1} + ... + me_m) \\
&\quad - (f_2 \log(2!) + ... + f_J \log(J!) + e_{J+1} \log(J+1!) + ... + e_m \log(m!)) \qquad (3.2)
\end{aligned}$$

#### 3.1.1.1 M Step

The likelihood can be maximised calculating the first derivative and solving the score equation $\frac{d\ell(\lambda)}{d\lambda} = 0$.

$$\frac{d\ell(\lambda)}{d\lambda} = (e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m) + \frac{(f_1 + 2f_2 + ... + Jf_J + (J+1)e_{J+1} + ... + me_m)}{\lambda} = 0$$

leading to

$$\hat{\lambda} = \frac{(f_1 + 2f_2 + ... + Jf_J + (J+1)e_{J+1} + ... + me_m)}{(e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m)}.$$

The EM algorithm leads to an updated parameter estimates:

$$\hat{\lambda} = \frac{(f_1 + 2f_2 + ... + Jf_J + (J+1)e_{J+1} + ... + me_m)}{(e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m)}. \tag{3.3}$$

#### 3.1.1.2 E step

Estimates for $e_0, e_{J+1}, \ldots, e_m$ are necessary in order to calculate $\hat{\lambda}$. We write:

$$E(f_y|f_1, .., f_J; \lambda) = Po(y|\lambda)N = Po(y|\lambda)(e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m) \tag{3.4}$$

where $J$ is the number of non-truncated counts, $m$ is the maximum number of captures and $Po(y|\lambda)$ is the probability of being captured $y$ times from a Poisson distribution.

The next step is the estimation of the latent variables $e_0$ and $\sum_{j=J+1}^{m} e_j$:

$$e_0 + \sum_{j=J+1}^{m} e_j = \left(1 - \sum_{i=1}^{J} Po(y|\lambda)\right)(f_1 + ... + f_J) + \left(1 - \sum_{i=1}^{J} Po(y|\lambda)\right)(e_0 + \sum_{j=J+1}^{m} e_j).$$

Therefore, solving for $e_0 + \sum_{j=J+1}^{m} e_j$

$$e_0 + \sum_{j=J+1}^{m} e_j = \frac{\left(1 - \sum_{i=1}^{J} Po(y|\lambda)\right)(f_1 + ... + f_J)}{\sum_{i=1}^{J} Po(y|\lambda)}. \tag{3.5}$$

We substitute (3.5) in the calculation of (3.4) to obtain:

$$\begin{aligned} E(f_y|f_1, .., f_J; \lambda) &= Po(y|\lambda)(e_0 + f_1 + ... + f_J + e_{J+1} + ... + e_m) \\ &= Po(y|\lambda)(f_1 + ... + f_J) + Po(y|\lambda)\frac{1 - \sum_{x'=1}^{J} Po(y'|\lambda)}{\sum_{y'=1}^{J} Po(y'|\lambda)}[f_1 + ... + f_J] \\ &= \frac{Po(y|\lambda)}{\sum_{y'=1}^{J} Po(y'|\lambda)}[f_1 + ... + f_J] = \frac{\lambda^y/y!}{\sum_{y'=1}^{J} \lambda^{y'}/y'!}[f_1 + ... + f_J]. \end{aligned} \tag{3.6}$$

We are particularly interested in $e_0 = E(f_0|f_1, .., f_J; \lambda)$:

$$E(f_0|f_1, .., f_J; \lambda) = \frac{1}{\sum_{y'=1}^{J} \lambda^{y'}/y'!}[f_1 + ... + f_J]. \tag{3.7}$$

An initial $\hat{\lambda}$ estimate value $\lambda_0$ is firstly chosen . Then $\lambda_0$ is used in the expectations formulae, which is needed in the likelihood to obtain a new maximum likelihood estimate $\hat{\lambda} = \lambda_1$. The process is repeated recursively until the difference $\mid \lambda_{k+1} - \lambda_k \mid$ or $\mid \ell_{r+1} - \ell_r \mid$ is smaller than a chosen tolerance threshold.

### 3.1.2 Truncated likelihood

#### 3.1.2.1 3-counts

In this section a truncated likelihood is initially developed for a Poisson distribution with 3 non-truncated counts $f_1, f_2, f_3$ and no covariate information included. Therefore three probabilities $p_1, p_2, p_3$ are calculated

$$p_y = P(Y = y|\lambda) = \frac{\frac{\mathrm{e}^{-\lambda}\lambda^y}{y!}}{\sum_{j=1}^{3} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}}, \qquad \text{for} \quad y \in \{1, 2, 3\}.$$

The Poisson truncated likelihood is defined as

$$\mathcal{L}(\lambda|f_1, f_2, f_3) = p_1^{f_1} \cdot p_2^{f_2} \cdot p_3^{f_3} = \left( \frac{\mathrm{e}^{-\lambda}\lambda}{\sum_{j=1}^{3} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}} \right)^{f_1} \left( \frac{\mathrm{e}^{-\lambda}\lambda^2/2!}{\sum_{j=1}^{3} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}} \right)^{f_2} \left( \frac{\mathrm{e}^{-\lambda}\lambda^3/3!}{\sum_{j=1}^{3} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}} \right)^{f_3}. \tag{3.8}$$

Consequently, the log-likelihood is:

$$\begin{aligned} \ell(\lambda|f_1, f_2, f_3) &= f_1 \times \log(p_1) + f_2 \times \log(p_2) + f_3 \times \log(p_3) \\ &= (f_2 + 2f_3)\log(\lambda) - (f_1 + f_2 + f_3)\log(6 + 3\lambda + \lambda^2) + f_1 \log(6) + f_2 \log(3) \end{aligned}$$

and it is maximised solving the score equation $\frac{d\ell(\lambda)}{d\lambda}$=0. We obtain

$$\frac{d\ell(\lambda)}{d\lambda} = -(f_1 + f_2 + f_3)\frac{2\lambda + 3}{6 + 3\lambda + \lambda^2} + \frac{f_2 + 2f_3}{\lambda}$$

$$= -\frac{3f_1\lambda + 2f_1\lambda^2 - 6f_2 + f_2\lambda^2 - 12f_3 - 3f_3\lambda}{\lambda(6 + 3\lambda + \lambda^2)} = 0$$

Solving for $\lambda$ results in:

$$\hat{\lambda} = \frac{-3(f_1 - f_3) + \sqrt{9(f_1 - f_3)^2 + 24f_2(2f_1 + f_2)}}{2(2f_1 + f_2)} \tag{3.9}$$

The other possible solution

$$\hat{\lambda} = \frac{-3(f_1 - f_3) - \sqrt{9(f_1 - f_3)^2 + 24f_2(2f_1 + f_2)}}{2(2f_1 + f_2)}$$

from the quadratic equation is not feasible. This is clear if $f_1 \geq f_3$. Otherwise $9(f_1 - f_3)^2 + 24f_2(f_1 + f_2) > 9(f_1 - f_3)^2$. Hence the second solution is always negative.

An analytical solution is still available for the case of three non-truncated counts, but next section shows that a numerical optimisation algorithm is necessary in order to maximise the likelihood when more non-truncated counts are considered.

#### 3.1.2.2 J-counts

The same process applied in the previous section (3.1.2.1), is followed to extend the approach to use a Poisson distribution with $J$ non-truncated counts, where $J \geq 2$. The truncated counts are $\Omega = \{0, J+1, \ldots, m\}$, with $m$ the maximum number of captures observed.

The likelihood is defined as

$$\mathcal{L}(\lambda | f_1, \ldots, f_J) = \sum_{j=1}^{J} p_j^{f_j},$$

where

$$p_y = \frac{\frac{e^{-\lambda}\lambda^y}{y!}}{\sum_{j=1}^{J} \frac{e^{-\lambda}\lambda^j}{j!}}, \qquad \text{for} \quad y \in \{1, \ldots, J\}.$$

The notation $\omega = \sum_{j=1}^{J} \frac{\lambda^j}{j!}$ is used for simplification in the following expressions.

The log-likelihood results in

$$\begin{aligned}
\ell(\lambda | f_1, \ldots, f_J) &= f_1 \times \log(p_1) + \ldots + f_J \times \log(p_J) \\
&= f_1 \log\left(\frac{\lambda}{\omega}\right) + \ldots + f_J \log\left(\frac{\lambda^J / J!}{\omega}\right) \\
&= f_1 \log(\lambda) - f_1 \log(\omega) + \ldots + f_J J \log(\lambda) - f_J \log(\omega) - log(J!) \\
&= \log(\lambda)(f_1 + 2f_2 + \ldots + Jf_J) - \log(\omega)(f_1 + f_2 + \ldots + f_J) \\
&- f_2 \log(2) + \ldots - f_J \log(J!)
\end{aligned} \tag{3.10}$$

The maximum likelihood estimate $\hat{\lambda}$ can be directly obtained from numerical optimisation algorithms like Nelder-Mead or BFGS as mentioned previously.

The next step is the estimation of $E(f_0 | f_1, \ldots, f_J; \lambda)$. The same reasoning as presented in section 3.1.1.2 is followed and $\hat{f}_0$ is provided.

$$E(f_0 | f_1, \ldots, f_j; \lambda) = \frac{1}{\omega}[f_1 + ... + f_J] = \frac{1}{\sum_{y'=1}^{J} \hat{\lambda}^{y'} / y'!}[f_1 + ... + f_J]. \tag{3.11}$$

where $\hat{\lambda}$ was previously estimated.

## 3.2 Generalised Chao estimator with covariates using $J$ counts

### 3.2.1 Complete likelihood

The EM algorithm based on the complete likelihood is applied as shown in section 3.1.1, but this time covariate information is included in the model.

The covariate information is log-linked with $\lambda_i$.

$$\lambda_i = \mathrm{e}^{\alpha+\beta' Z_i}, \tag{3.12}$$

where $Z_i$ represent a vector of $p$ covariates, $i$ is the index of the different covariate combinations, $i = 1, \ldots, M_J$ where $M_J$ is the total number of strata when $J$ counts are non-truncated. $n_i = \sum_{j=1}^{J} f_{ij}$ is the number of units observed for the $i$th stratum, where $f_{ij}$ is the number of units captured $j$ times with the $i$th set of characteristics.

The Poisson complete likelihood is defined as

$$\mathcal{L}(\lambda_i) = \prod_{i=1}^{M_J} \prod_{j=0}^{m} p_{ij}^{f_{ij}} \tag{3.13}$$

where $M_J$ is the number of covariate combinations, $m$ is the maximum number of captures available and

$$p_{ij} = \mathrm{e}^{-\lambda_i} \lambda_i^j / j! \ , \tag{3.14}$$

the probability of being captured $j$ times for units in the $i$th covariate combination.

The expected complete log-likelihood is presented here because it is computationally more efficient and easier to work with.

$$\ell(\lambda_i) = \sum_{i=1}^{M_J} e_{i0} \log(p_{i0}) + f_{i1} \log(p_{i1}) + \ldots + f_{iJ} \log(p_{iJ}) + e_{i(J+1)} \log(p_{i(J+1)}) + \ldots + e_{im} \log(p_{im}), \tag{3.15}$$

where $J$ is the number of counts included for the estimation.

$$\begin{aligned}\ell(\lambda) = &\sum_{i=1}^{M_J}(e_{i0}\log(\mathrm{e}^{-\lambda_i}) + f_{i1}\log(\mathrm{e}^{-\lambda_i}\lambda_i) + ... + f_{iJ}\log(\mathrm{e}^{-\lambda_i}\lambda_i^J/J!)\\ &+ e_{i(J+1)}\log(\mathrm{e}^{-\lambda_i}\lambda_i^{J+1}/(J+1)!) + ... + e_{im}\log(\mathrm{e}^{-\lambda_i}\lambda_i^m/m!))\\ &= \sum_{i=1}^{M_J} -\lambda_i(e_{i0} + f_{i1} + ... + f_{iJ} + e_{i(J+1)} + ... + e_{im})\\ &+ \log(\lambda_i)(f_{i1} + 2f_{i2} + ... + Jf_{iJ} + (J+1)e_{i(J+1)} + ... + me_{im})\\ &- f_2\log(2!) + ... + f_J\log(J!) + f_{J+1}\log(J+1!) + ... + f_m\log(m!)\end{aligned}$$

The log-likelihood can be written with respect to $\alpha$ and $\beta$.

$$\begin{aligned}\ell(\alpha,\beta) = &\sum_{i=1}^{M_J} -\mathrm{e}^{\alpha+\beta' Z_i}\left(\sum_{j=1}^{J} f_{ij} + e_{i0} + \sum_{s=(J+1)}^{m} e_{is}\right) + \sum_{i=1}^{M_J}(\alpha+\beta' Z_i)\left(\sum_{j=1}^{J} jf_{ij} + \sum_{s=(J+1)}^{m} se_{is}\right)\\ &- \sum_{j=2}^{m} f_j\log(j!)\end{aligned} \tag{3.16}$$

An optimisation algorithm could be used to estimate $\alpha$ and $\beta$ once estimates for the expectations of the frequencies are obtained and replaced into 3.16.

The expectations are computed following the technique in section 3.1.1.2.

**E step**

$$e_{iy} = E(f_{iy}|f_{i1},..,f_{iJ};\lambda_i) = Po(x|\lambda_i)(e_{i0} + f_{i1} + ... + f_{iJ} + e_{i(J+1)} + ... + e_{im}) \tag{3.17}$$

$e_{i0}$ and $\sum_{j=J+1}^{m} e_{ij}$ are unknown. Now,

$$e_{0i} + \sum_{j=J+1}^{m} e_{ij} = \left(1 - \sum_{i=1}^{J} Po(y|\lambda_i)\right)(f_{i1}+...+f_{iJ}) + \left(1 - \sum_{i=1}^{J} Po(y|\lambda_i)\right)\left(e_{i0} + \sum_{j=J+1}^{m} e_{ij}\right)$$

Therefore, solving for $e_{i0} + \sum_{j=J+1}^{m} e_{ij}$

$$e_{i0} + \sum_{j=J+1}^{m} e_{ij} = \frac{\left(1 - \sum_{i=1}^{J} Po(x|\lambda_i)\right)(f_{i1} + ... + f_{iJ})}{\sum_{i=1}^{J} Po(x|\lambda_i)} \tag{3.18}$$

(3.17) can be calculated by replacing (3.18) to obtain:

$$\begin{aligned} e_{iy} = E(f_{iy}|f_{i1},..,f_{iJ};\lambda_i) &= Po(x|\lambda_i)\left(e_{i0}+f_{i1}+...+f_{iJ}+e_{i(J+1)}+...+eim\right) \\ &= Po(y|\lambda_i)(f_{i1}+...+f_{iJ}) + Po(y|\lambda_i)\frac{\left(1-\sum_{y'=1}^{J} Po(y'|\lambda_i)\right)}{\sum_{y'=1}^{J} Po(y'|\lambda_i)}[f_{i1}+...+f_{iJ}] \\ &= \frac{Po(y|\lambda_i)}{\sum_{y'=1}^{J} Po(y'|\lambda_i)}[f_{i1}+...+f_{iJ}] \\ &= \frac{\lambda_i^y/y!}{\sum_{y'=1}^{J}\lambda_i^{y'}/y'!}[f_{i1}+...+f_{iJ}]. \end{aligned} \tag{3.19}$$

Finally, our parameter of interest $f_0$ can be estimated as $\hat{f}_0 = \sum_{i=1}^{M_J} \hat{f}_{i0}$.

$$\hat{f}_0 = \sum_{i=1}^{M_J} \frac{1}{\sum_{y'=1}^{J} \mathrm{e}^{(\hat{\alpha}+\hat{\beta}'z_i)y'/y'!}}[f_{i1}+...+f_{iJ}] \tag{3.20}$$

The implementation of the EM algorithm can be summarised as follows:

- Choose initial values for $\alpha$ and $\beta$
- Replace those values in the expectations formula (3.20)
- Replace the calculated expectations in the log likelihood and maximise it to obtain $\hat{\alpha}_1$ and $\hat{\beta}_1$
- Repeat the procedure until the maximum difference between the parameters corresponding to step $k+1$ and $k$ is less than a tolerance $\tau > 0$.

$$max(\mid \hat{\alpha}_{k+1} - \hat{\alpha}_k \mid, \mid \hat{\beta}_{k+1} - \hat{\beta}_k \mid) < \tau.$$

### 3.2.2 Truncated likelihood

In this section, we extend the methodology presented in section 3.1.2.2 to include covariate information working directly with a truncated Poisson likelihood rather than with the complete Poisson likelihood developed in the previous section 3.2.1.

$J$ counts are considered to be used and $m$ is defined as the maximum number of counts in the capture distribution. Covariate information is also available and linked with $\lambda_i$ as in previous sections. Let

$$\lambda_i = \mathrm{e}^{\alpha+\beta' Z_i} \qquad \text{for} \quad i = 1,\ldots,M_J,$$

where $M_J$ is the total number of covariate combinations when $J$ counts are non-truncated, and $Z_i$ is a vector of covariates.

In this case, a Poisson likelihood truncating the counts $0, J+1, \ldots, m$ is defined as

$$\mathcal{L}(\lambda_i | f_1, \ldots, f_J) = \prod_{i=1}^{M_J} \prod_{j=1}^{J} p_{ij}^{f_{ij}}$$

where

$$p_{iy} = \frac{\frac{\mathrm{e}^{-\lambda_i}\lambda_i^y}{y!}}{\sum_{j=1}^{J} \frac{\mathrm{e}^{-\lambda_i}\lambda_i^j}{j!}}, \qquad \text{for} \quad y \in \{1, \ldots, J\} \tag{3.21}$$

is the probability of being captured $y$ times for units with covariate combination $i$.

Therefore, the log-likelihood becomes

$$\ell(\lambda_i | f_1, \ldots, f_J) = \sum_{i=1}^{M_J} [f_{i1} \times \log(p_{i1}) + \ldots + f_{iJ} \times \log(p_{iJ})] \tag{3.22}$$

For simplification we assign $\omega_i = \sum_{j=1}^{J} \frac{\lambda_i^j}{j!}$. Hence, the log-likelihood after replacing the capture probabilities from (3.21) in (3.22) is

$$\begin{aligned}
\ell(\lambda_i | f_{i1}, \ldots, f_{iJ}) &= \sum_{i=1}^{M_J} f_{i1} \log\left(\frac{\lambda_i}{\omega_i}\right) + \ldots + f_{iJ} \log\left(\frac{\lambda_i^J / J!}{\omega_i}\right) \\
&= \sum_{i=1}^{M_J} f_{iJ} log\lambda_i - f_{i1} \log(\omega_i) + \ldots + f_{iJ} J \log(\lambda_i) - f_{iJ} \log(\omega_i) \qquad (3.23) \\
&\quad - f_{i2} \log(2) - \ldots - f_{iJ} \log J! \\
&= \sum_{i=1}^{M_J} \left(\sum_{j=1}^{J} j f_{ij}\right) \log(\lambda_i) - \left(\sum_{j=1}^{J} f_{ij}\right) \log(\omega_i) - \sum_{j=2}^{J} f_{ij} \log(J!) \qquad (3.24)
\end{aligned}$$

Finally, the log-likelihood with respect to $\alpha$ and $\beta$ is calculated replacing $\lambda_i$ with the linear predictor. Firstly, we see that

$$\log(\omega_i) = \sum_{j=1}^{J} \log\left(\frac{\mathrm{e}^{(\alpha+\beta' z_i)j}}{j!}\right) = \sum_{j=1}^{J} (\alpha + \beta' z_i) j - \log(j!), \tag{3.25}$$

and therefore,

$$\begin{aligned}
\ell(\alpha, \beta | f_{i1}, \ldots, f_{iJ}) = \sum_{i=1}^{M_J} \Bigg[ & (\alpha + \beta' Z_i) \left(\sum_{j=1}^{J} j f_{ij}\right) - \sum_{j=2}^{J} f_{ij} \log(J!) \qquad (3.26) \\
& - \left(\sum_{j=1}^{J} f_{ij}\right) \sum_{k=1}^{J} (\alpha + \beta' Z_i) k - \log(k!) \Bigg] \qquad (3.27)
\end{aligned}$$

At this stage, an optimisation algorithm can be used to maximise the likelihood and obtain estimates for $\alpha$ and $\beta$.

The calculation of $E(f_0|f_1,\ldots,f_j;\lambda_i)$ is identical to the E step in section 3.2.1. $\hat{\alpha}$ and $\hat{\beta}$ are obtained by maximising the log-likelihood.

Finally, we find

$$\begin{aligned} e_0 = E(f_0|f_1,\ldots,f_j;\lambda_i) &= \sum_{i=1}^{M_J} \frac{1}{\sum_{y'=1}^{J} Po(y'|\lambda_i)}[f_{i1}+\ldots+f_{iJ}] = \sum_{i=1}^{M_J} \frac{1}{\omega_i}[f_{i1}+\ldots+f_{iJ}] \\ &= \sum_{i=1}^{M_J} \frac{1}{\sum_{j=1}^{J}\left(\frac{e^{(\hat{\alpha}+\hat{\beta}'Z_i)j}}{j!}\right)}[f_{i1}+\ldots+f_{iJ}] \end{aligned} \tag{3.28}$$

## 3.3 Variance estimator for $N_{GC}$ with J non-truncated counts and covariates

We use the same conditioning technique applied to the case $J=2$ to obtain an analytical variance estimate (section 2.2):

$$Var(\hat{N}_{GC}) = Var\left[E(\hat{N}_{GC}|\Delta_i, i=1,..,N)\right] + E\left[Var(\hat{N}_{GC}|\Delta_i, i=1,..,N)\right], \tag{3.29}$$

where

$$\Delta_i = \begin{cases} 1, & y_i \in \{1,..,J\} \\ 0, & otherwise \end{cases}$$

The first part relates to the sampling variance and the second part represents the variance coming from the estimate itself.

Our estimate $\hat{N}_{GC}$ when using $J$ non-truncated counts and covariates can be written as

$$\begin{aligned} E(\hat{N}_{GC}|\Delta_i, i=1,\ldots,N) &= E\left(n + \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\lambda}_i + \hat{\lambda}_i^2/2 + \ldots + \hat{\lambda}_i^J/J!}\right) \\ &= E\left(\sum_{i=1}^{N}\Delta_i + \sum_{i=1}^{N}\gamma_i + \sum_{i=1}^{N} \frac{\Delta_i}{\hat{\lambda}_i + \hat{\lambda}_i^2/2 + \ldots + \hat{\lambda}_i^J/J!}\right), \end{aligned}$$

where

$$\gamma_i = \begin{cases} 1, & y_i \geq J+1 \\ 0, & otherwise \end{cases},$$

and

$$\lambda_i = e^{\alpha+\beta'Z_i}.$$

$\lambda_i$ links the covariate information with the Poisson parameter.

We can also write

$$E(\hat{N}|\Delta_i, i = 1, \ldots, N) \approx \sum_{i=1}^{N} \Delta_i \left( \frac{\hat{p}_i + \mathrm{e}^{\lambda_i}}{p_i} \right) = \sum_{i=1}^{N} \Delta_i \omega_i.$$

with $\omega_i = 1 + \frac{\mathrm{e}(\lambda_i)}{p_i}$ for simplification.

$p_i$ is defined as the probability that $\Delta_i = 1$:

$$p_i = p(\Delta_i = 1|\lambda_i) = \lambda_i \mathrm{e}^{-\lambda_i} + \lambda_i^2 \mathrm{e}^{-\lambda_i}/2 + ... + \lambda_i^J \mathrm{e}^{-\lambda_i}/J!,$$

The $E(\Delta_i) = p_i$ and $Var(\Delta_i) = p_i(1 - p_i)$ because $\Delta_i$ follows a binomial distribution. Ultimately, we achieve

$$Var\left( E(\hat{N}|\Delta_i, i = 1, \ldots, N) \right) \simeq \sum_{i=1}^{N} Var(\Delta_i \omega_i) \simeq \sum_{i=1}^{N} p_i(1 - p_i) w_i^2.$$

The Horvitz-Thompson estimator is applied to estimate the variability:

$$\widehat{Var}(E(\hat{N}|\Delta_i, i = 1, \ldots, N))) \simeq \sum_{i=1}^{N} \frac{\Delta i}{\hat{p}_i} \hat{p}_i(1 - \hat{p}_i)\hat{\omega}_i^2 = \sum_{i=1}^{f_1+f_2+...+f_J} (1 - \hat{p}_i) \left[ \frac{\hat{p}_i + \mathrm{e}^{-\hat{\lambda}_i}}{\hat{p}_i} \right]^2. \tag{3.30}$$

The multivariate Delta method is used for calculating the second term:

$$E[Var(N_{GC}|\Delta_i, i = 1, .., N)] = \nabla g(\hat{\alpha}, \hat{\beta})^T cov(\hat{\alpha}, \hat{\beta}) \nabla g(\hat{\alpha}, \hat{\beta}) \tag{3.31}$$

where

$$\nabla g(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} \frac{\partial g}{\partial \alpha} \\ \frac{\partial g}{\partial \beta_1} \\ ... \\ \frac{\partial g}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-\sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j-1!}}{\left( \sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j!} \right)^2} \\ \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-\sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j-1!}}{\left( \sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j!} \right)^2} z_{i1} \\ ... \\ \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-\sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j-1!}}{\left( \sum_{j=1}^{J} \frac{\hat{\lambda}_i^j}{j!} \right)^2} z_{ip} \end{pmatrix}.$$

$\nabla g(\alpha, \beta)$ can be also expressed in terms of $\hat{\alpha}$ and $\hat{\beta}$ that have been obtained in the maximisation of the likelihood.

The covariance matrix $cov(\hat{\alpha}, \hat{\beta})$ is calculated as the inverse of the observed Fisher information (or the inverse of the Hessian of the negative log likelihood).

$$cov(\hat{\alpha}, \hat{\beta}) = -\left(\frac{\partial}{\partial\alpha\partial\beta}\ell(\alpha, \beta)\right)^{-1}$$

$$cov(\hat{\alpha}, \hat{\beta}) = -\begin{pmatrix} \frac{\partial^2\ell(\alpha,\beta)}{\partial\alpha^2} & \frac{\partial^2\ell(\alpha,\beta)}{\partial\alpha\beta_1} & \cdots & \frac{\partial^2\ell(\alpha,\beta)}{\partial\alpha\beta_p} \\ \frac{\partial^2\ell(\alpha,\beta)}{\partial\alpha\beta_1} & \frac{\partial^2\ell(\alpha,\beta)}{\partial\beta_1^2} & \cdots & \frac{\partial^2\ell(\alpha,\beta)}{\partial\beta_1\beta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2\ell(\alpha,\beta)}{\partial\alpha\beta_p} & \frac{\partial^2\ell(\alpha,\beta)}{\partial\beta_p\beta_1} & \cdots & \frac{\partial^2\ell(\alpha,\beta)}{\partial\beta_p^2} \end{pmatrix}^{-1}$$

The partial derivatives are presented here although an approximation of the covariance matrix is commonly produced by the optimisation function of the statistical software.

$$\frac{\partial^2\ell(\alpha, \beta)}{\partial\alpha^2} = -\sum_{j=1}^{J} f_j \left( \frac{\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j^2}{(j+1)!} \sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!} - \left(\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j}{(j+1)!}\right)^2}{\left(\sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!}\right)^2} \right)$$

$$\frac{\partial^2\ell(\alpha, \beta)}{\partial\alpha\partial\beta_j} = -z_j \sum_{j=1}^{J} f_j \left( \frac{\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j^2}{(j+1)!} \sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!} - \left(\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j}{(j+1)!}\right)^2}{\left(\sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!}\right)^2} \right)$$

$$\frac{\partial^2\ell(\alpha, \beta)}{\partial\beta_j^2} = -z_j^2 \sum_{j=1}^{J} f_j \left( \frac{\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j^2}{(j+1)!} \sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!} - \left(\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j}{(j+1)!}\right)^2}{\left(\sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!}\right)^2} \right)$$

$$\frac{\partial^2\ell(\alpha, \beta)}{\partial\beta_j\partial\beta_k} = -z_j z_k \sum_{j=1}^{J} f_j \left( \frac{\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j^2}{(j+1)!} \sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!} - \left(\sum_{j=1}^{J-1} \frac{\hat{\lambda}_i^j j}{(j+1)!}\right)^2}{\left(\sum_{j=0}^{J-1} \frac{\hat{\lambda}_i^j}{(j+1)!}\right)^2} \right)$$

## 3.4 Simulation Results

### 3.4.1 Heterogeneity without covariate information

In this simulation the number of captures $Y_i$ is generated as $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$, a mixture of two Poissons, for $i = 1, ..., N$ and $\lambda \in \{2, .., 7\}$. Multiple scenarios were

conducted varying the population size, the number of non-truncated counts and the level of heterogeneity (Tables 3.1 and 3.2, figures 3.1, 3.2, 3.3 and 3.4).

In the absence of auxiliary variables related to the generated heterogeneity, we showed in section 2.1 that the classic Chao estimator using two counts could still provide robust estimates. Here we aim to evaluate whether using more information could improve Chao's estimator. We observe in figure 3.1 that for all levels of heterogeneity defined by $\lambda$, two counts provided the most accurate estimates and an increase in non-truncated counts decreases the accuracy of the estimates. We also notice a bathtub effect for $\lambda \geq 6$, despite stronger heterogeneity than other scenarios, less biased estimates are obtained.

However, figure 3.2 shows a decrease in the standard deviation when the number of non-truncated counts increases. There is a tradeoff between accuracy and variability. The relative mean squared error and the relative bias are good measures to assess the estimates accounting for point estimation and variability and they also allow us to compare estimates across different population sizes (figures 3.3, 3.4). In this particular example, the models with two non-truncated counts present the best performance based on the RMSE criteria, although the model with three non-truncated counts provide similar RMSE values. The RMSE and relative bias remained constant across populations of $N \geq 500$ (Table 3.1). It is clear from the simulations that for large $N$ the bias becomes dominating, leading to the result that $J = 2$ (classical Chao) performs best.

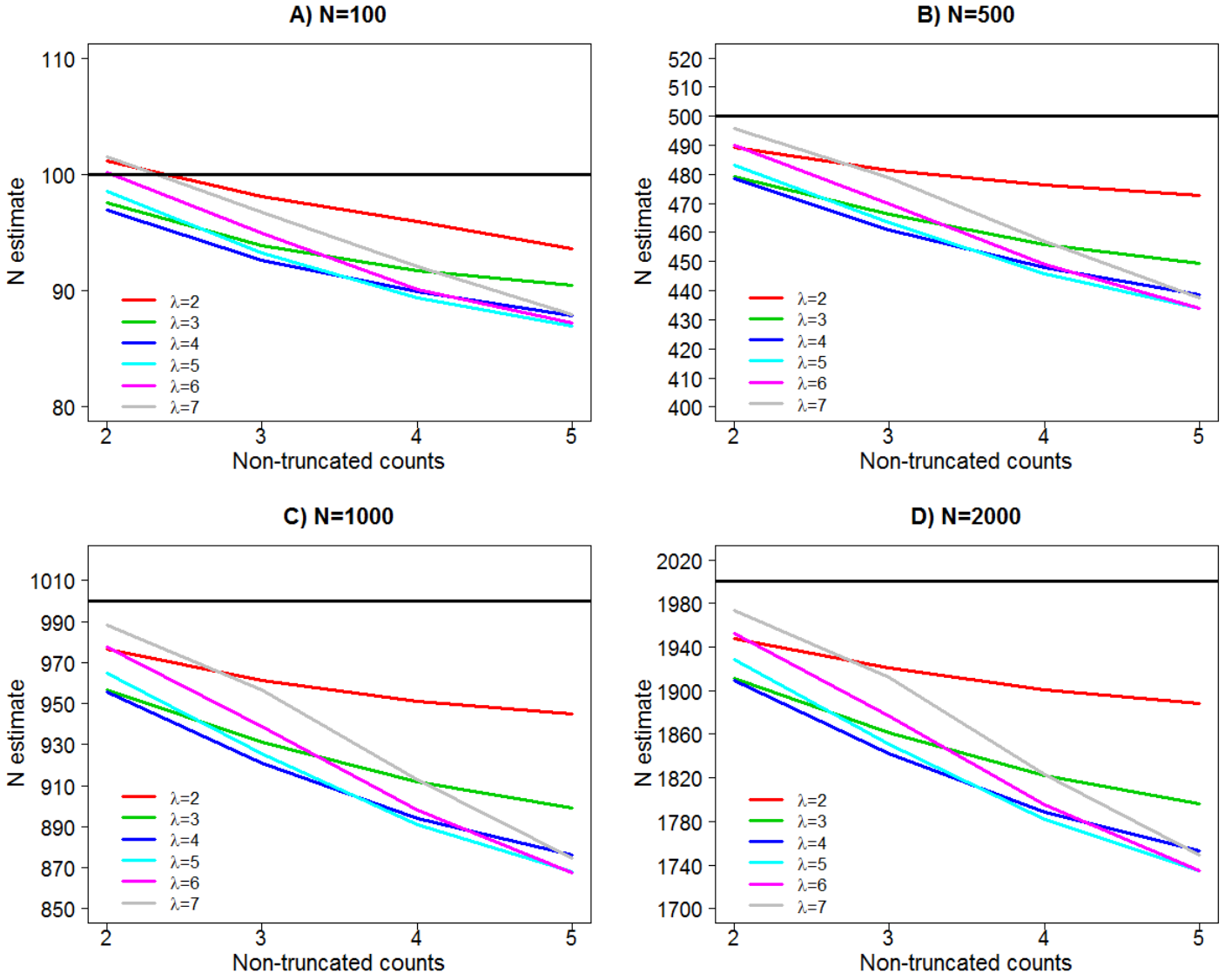


Figure 3.1: Population estimates for the model $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$ for $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$. A) $N = 100$ B) $N = 500$ C) $N = 1000$ D) $N = 2000$

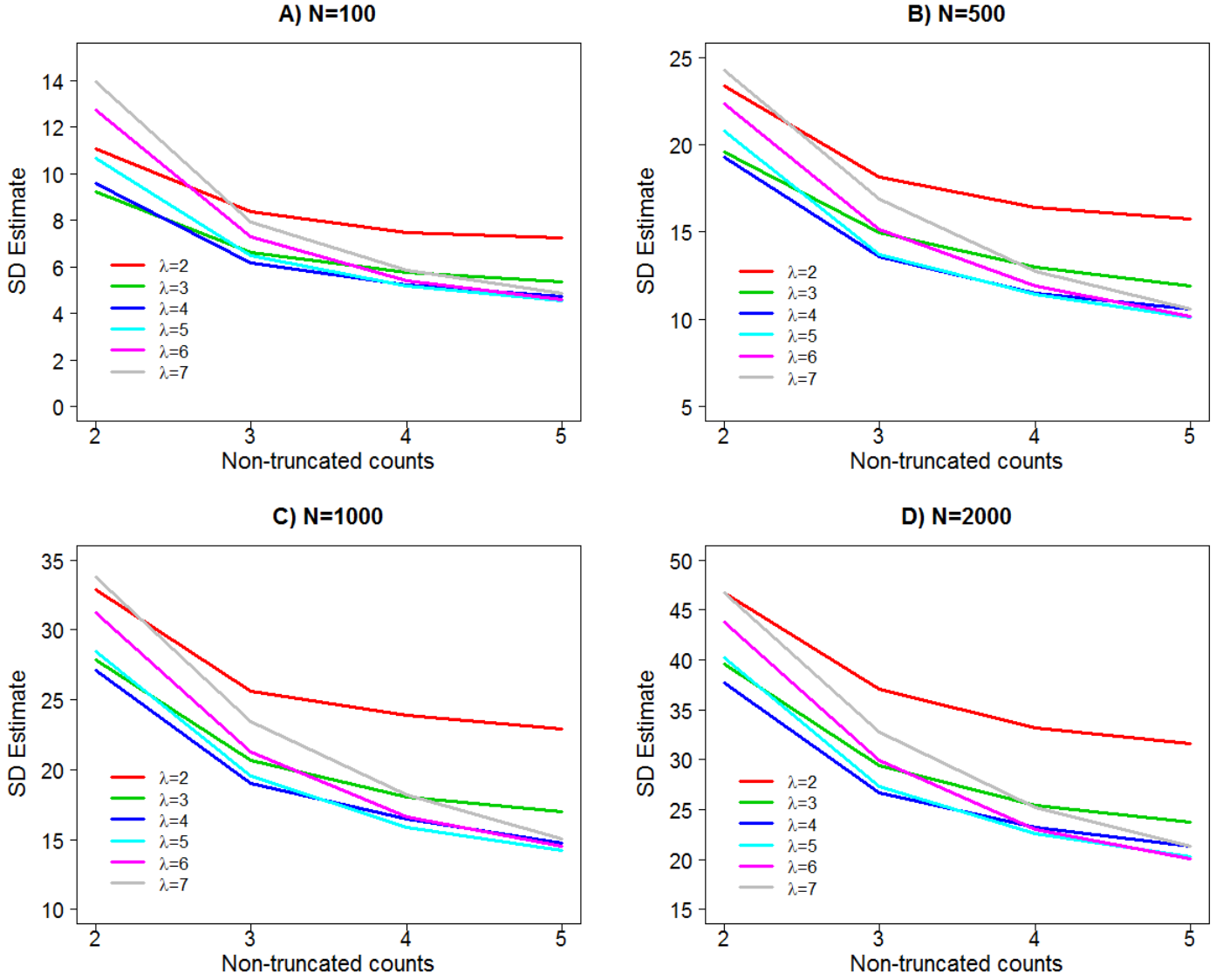


Figure 3.2: SD estimates for the model $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$ with $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$. A) $N = 100$ B) $N = 500$ C) $N = 1000$ D) $N = 2000$

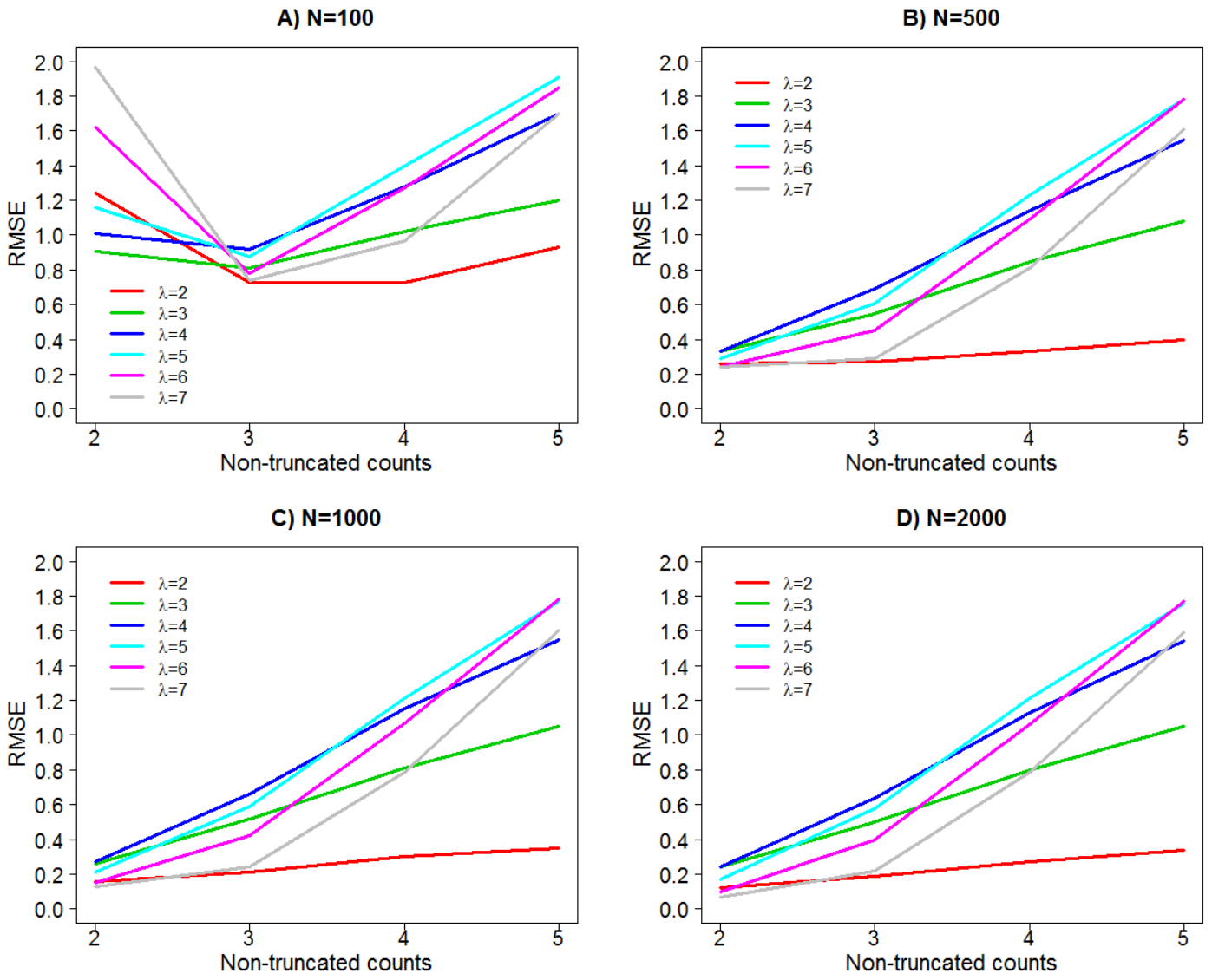


Figure 3.3: RMSE (x100) estimates for the model $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$ with $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$. A) $N = 100$ B) $N = 500$ C) $N = 1000$ D) $N = 2000$

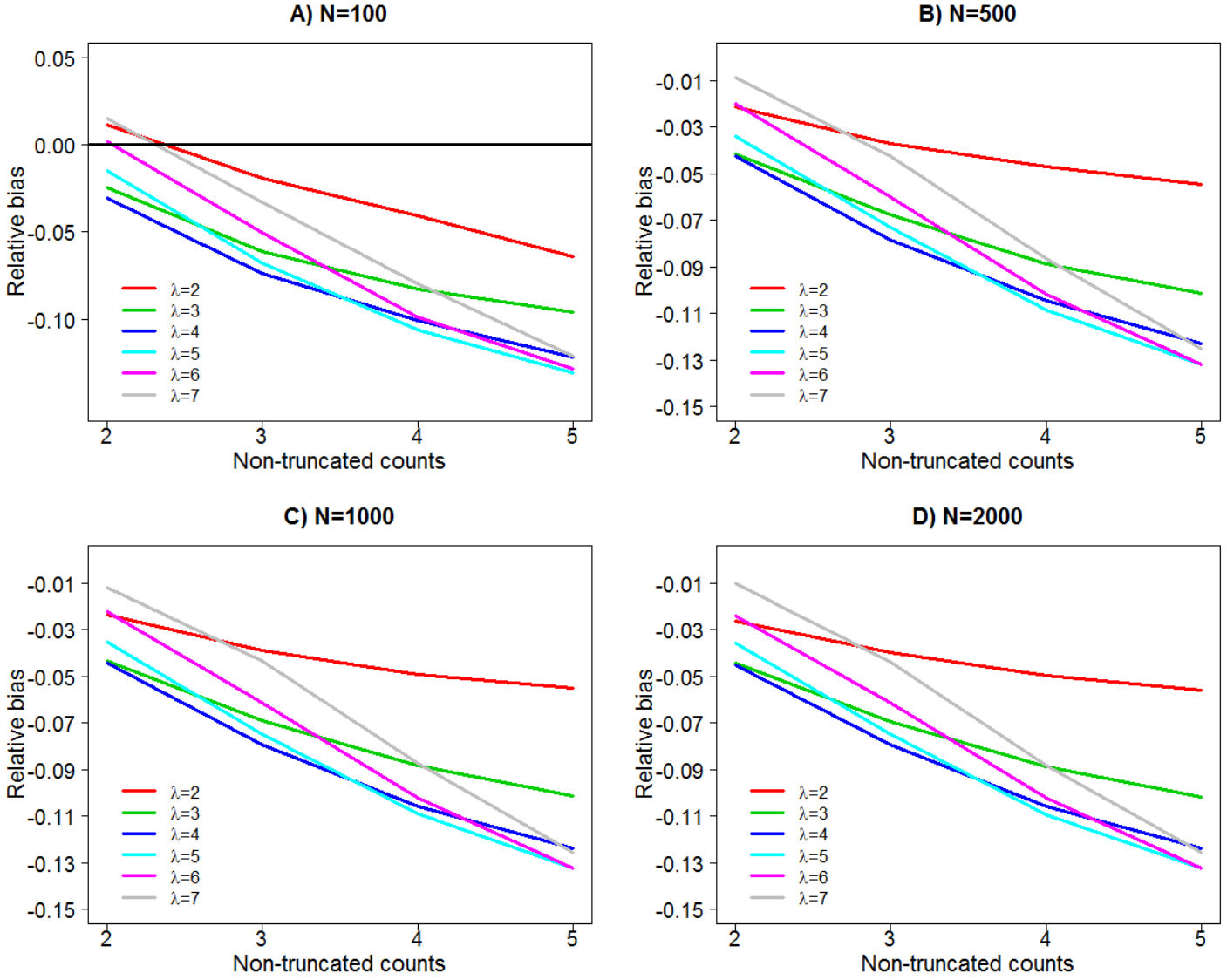


Figure 3.4: Relative bias estimates for the model $Y_i \sim 0.5Po(1)+0.5Po(\lambda)$ with $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$. A) $N = 100$ B) $N = 500$ C) $N = 1000$ D) $N = 2000$

Table 3.1: Point estimates and SD estimates for the model $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$ with $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$. Italics are only use for visual purposes.

| $N$ | # NTC | $\lambda$ | $\hat{N}$ | $SD_{Emp}$ | $\lambda$ | $\hat{N}$ | $SD_{Emp}$ |
|---|---|---|---|---|---|---|---|
| 100 | 2 | 2 | 101.16 | 11.09 | 3 | 97.53 | 9.21 |
| | 3 | | 98.11 | 8.36 | | 93.87 | 6.61 |
| | 4 | | 95.91 | 7.47 | | 91.75 | 5.79 |
| | 5 | | 93.62 | 7.25 | | 90.44 | 5.35 |
| 500 | 2 | | 489.41 | 23.36 | | 479.14 | 19.59 |
| | 3 | | 481.45 | 18.15 | | 466.15 | 14.97 |
| | 4 | | 476.51 | 16.42 | | 455.74 | 12.98 |
| | 5 | | 472.63 | 15.76 | | 449.38 | 11.89 |
| 1000 | 2 | | 976.64 | 32.91 | | 956.89 | 27.89 |
| | 3 | | 961.36 | 25.61 | | 930.99 | 20.66 |
| | 4 | | 951.02 | 23.87 | | 911.84 | 18.01 |
| | 5 | | 945.12 | 22.89 | | 898.85 | 16.96 |
| 2000 | 2 | | 1947.78 | 46.7 | | 1911.41 | 39.58 |
| | 3 | | 1920.81 | 37.11 | | 1861.41 | 29.37 |
| | 4 | | 1901.16 | 33.19 | | 1822.37 | 25.41 |
| | 5 | | 1888.35 | 31.59 | | 1796.4 | 23.77 |
| 100 | 2 | 4 | 96.95 | 9.57 | 5 | 98.51 | 10.65 |
| | 3 | | 92.65 | 6.16 | | 93.25 | 6.49 |
| | 4 | | 89.94 | 5.21 | | 89.38 | 5.20 |
| | 5 | | 87.84 | 4.75 | | 86.96 | 4.54 |
| 500 | *2* | | 478.68 | 19.29 | | 483.14 | 20.79 |
| | *3* | | 460.79 | 13.62 | | 463.44 | 13.69 |
| | *4* | | 447.78 | 11.47 | | 445.72 | 11.46 |
| | *5* | | 438.60 | 10.62 | | 434.06 | 10.14 |
| 1000 | 2 | | 955.71 | 27.08 | | 964.74 | 28.48 |
| | 3 | | 920.85 | 19.00 | | 925.43 | 19.56 |
| | 4 | | 894.2 | 16.49 | | 890.99 | 15.89 |
| | 5 | | 876.35 | 14.78 | | 867.85 | 14.19 |
| 2000 | 2 | | 1909.70 | 37.64 | | 1928.45 | 40.24 |
| | 3 | | 1841.83 | 26.67 | | 1850.78 | 27.31 |
| | 4 | | 1788.25 | 23.17 | | 1781.39 | 22.57 |
| | 5 | | 1752.71 | 21.28 | | 1735.15 | 20.32 |
| 100 | 2 | 6 | 100.17 | 12.72 | 7 | 101.51 | 13.94 |
| | 3 | | 94.99 | 7.31 | | 96.71 | 7.94 |
| | 4 | | 90.13 | 5.41 | | 92.06 | 5.87 |
| | 5 | | 87.19 | 4.61 | | 87.91 | 4.85 |
| 500 | 2 | | 490.11 | 22.37 | | 495.71 | 24.26 |
| | 3 | | 470.02 | 15.17 | | 478.84 | 16.86 |
| | 4 | | 449.05 | 11.89 | | 456.87 | 12.74 |
| | 5 | | 434.02 | 10.20 | | 437.54 | 10.57 |
| 1000 | 2 | | 977.57 | 31.22 | | 988.01 | 33.76 |
| | 3 | | 938.81 | 21.29 | | 956.71 | 23.45 |
| | 4 | | 897.98 | 16.65 | | 912.76 | 18.21 |
| | 5 | | 867.51 | 14.54 | | 874.57 | 15.01 |
| 2000 | 2 | | 1952.46 | 43.76 | | 1973.34 | 46.73 |
| | 3 | | 1877.14 | 29.92 | | 1912.26 | 32.79 |
| | 4 | | 1795.11 | 22.96 | | 1823.50 | 25.18 |
| | 5 | | 1734.92 | 20.07 | | 1748.81 | 21.38 |

Table 3.2: RMSE and relative bias for the model $Y_i \sim 0.5Po(1)+0.5Po(\lambda)$ with $i = 1, \ldots, N$ and $\lambda = \{2, .., 7\}$

.

| $N$ | # NTC | $\lambda$ | RMSE | Rbias | $\lambda$ | RMSE | RBias |
|---|---|---|---|---|---|---|---|
| 100 | 2 | 2 | 1.24 | 0.0116 | 3 | 0.91 | -0.0247 |
| | 3 | | 0.73 | -0.0189 | | 0.81 | -0.0613 |
| | 4 | | 0.73 | -0.0409 | | 1.02 | -0.0825 |
| | 5 | | 0.93 | -0.0638 | | 1.2 | -0.0956 |
| 500 | *2* | | 0.26 | -0.0212 | | 0.33 | -0.0417 |
| | *3* | | 0.27 | -0.0371 | | 0.55 | -0.0677 |
| | *4* | | 0.33 | -0.0470 | | 0.85 | -0.0885 |
| | *5* | | 0.40 | -0.0547 | | 1.08 | -0.1012 |
| 1000 | 2 | | 0.16 | -0.0234 | | 0.26 | -0.0431 |
| | 3 | | 0.21 | -0.0386 | | 0.52 | -0.069 |
| | 4 | | 0.30 | -0.0490 | | 0.81 | -0.0882 |
| | 5 | | 0.35 | -0.0549 | | 1.05 | -0.1011 |
| 2000 | 2 | | 0.12 | -0.0261 | | 0.24 | -0.0443 |
| | 3 | | 0.19 | -0.0396 | | 0.50 | -0.0693 |
| | 4 | | 0.27 | -0.0494 | | 0.80 | -0.0888 |
| | 5 | | 0.34 | -0.0558 | | 1.05 | -0.1018 |
| 100 | 2 | 4 | 1.01 | -0.0305 | 5 | 1.16 | -0.0149 |
| | 3 | | 0.92 | -0.0735 | | 0.88 | -0.0675 |
| | 4 | | 1.28 | -0.1006 | | 1.40 | -0.1062 |
| | 5 | | 1.70 | -0.1216 | | 1.91 | -0.1304 |
| 500 | 2 | | 0.33 | -0.0426 | | 0.29 | -0.0337 |
| | 3 | | 0.69 | -0.0784 | | 0.61 | -0.0731 |
| | 4 | | 1.14 | -0.1044 | | 1.23 | -0.1086 |
| | 5 | | 1.55 | -0.1228 | | 1.78 | -0.1319 |
| 1000 | 2 | | 0.27 | -0.0443 | | 0.21 | -0.0353 |
| | 3 | | 0.66 | -0.0792 | | 0.59 | -0.0746 |
| | 4 | | 1.15 | -0.1058 | | 1.21 | -0.109 |
| | 5 | | 1.55 | -0.1237 | | 1.77 | -0.1321 |
| 2000 | *2* | | 0.24 | -0.0452 | | 0.17 | -0.0358 |
| | *3* | | 0.64 | -0.0791 | | 0.58 | -0.0746 |
| | *4* | | 1.13 | -0.1059 | | 1.21 | -0.1093 |
| | *5* | | 1.54 | -0.1236 | | 1.76 | -0.1324 |
| 100 | 2 | 6 | 1.62 | 0.0017 | 7 | 1.97 | 0.0151 |
| | 3 | | 0.78 | -0.0501 | | 0.74 | -0.0329 |
| | 4 | | 1.27 | -0.0987 | | 0.97 | -0.0794 |
| | 5 | | 1.85 | -0.1281 | | 1.70 | -0.1209 |
| 500 | 2 | | 0.002 | -0.0198 | | 0.24 | -0.0086 |
| | 3 | | 0.005 | -0.0600 | | 0.29 | -0.0423 |
| | 4 | | 0.012 | -0.1019 | | 0.81 | -0.0863 |
| | 5 | | 0.02 | -0.1320 | | 1.61 | -0.1249 |
| 1000 | 2 | | 0.15 | -0.0224 | | 0.13 | -0.0120 |
| | 3 | | 0.42 | -0.0612 | | 0.24 | -0.0433 |
| | 4 | | 1.07 | -0.1020 | | 0.79 | -0.0872 |
| | 5 | | 1.78 | -0.1325 | | 1.60 | -0.1254 |
| 2000 | 2 | | 0.10 | -0.0238 | | 0.07 | -0.0133 |
| | 3 | | 0.40 | -0.0614 | | 0.22 | -0.0439 |
| | 4 | | 1.06 | -0.1024 | | 0.79 | -0.0882 |
| | 5 | | 1.77 | -0.1325 | | 1.59 | -0.1256 |

### 3.4.2 Generalised Chao's estimate using covariates and $J$ non-truncated counts

#### 3.4.2.1 One covariate

In this section, our purpose is to investigate a scenario with heterogeneity described by a continuous covariate. The estimates are calculated also including that covariate in the model for the estimation. Therefore, we expect to obtain accurate estimates of the true population size when considering the available covariate. The capture-recapture distribution is originated from a Poisson distribution with parameter $\lambda_i$ ($Y_i \sim Po(\lambda_i)$), where $\lambda_i$ is obtained as

$$\log(\lambda_i) = 0.04X_{1i},$$

where $X_1$ follows a normal distribution with mean 20 and variance 225 ($X_1 \sim N(20, 225)$). Once the capture-recapture distribution is created, zeros are excluded to analyse only the observed units.

Similar estimates are found across all scenarios with different population sizes which indicates that high number of counts do not need to be used to obtain reliable estimates. For scenarios with the same population size $N$, there is a slightly decreasing trend of the standard deviation when the number of non-truncated counts increases. For scenarios with the same number of non-truncated counts, the standard deviation increases asymptotically with respect to $N$, in contrast to the RMSE that decreases (Table 3.3). This simulation is also useful to validate the method and the programming code.

#### 3.4.2.2 Two covariates with unexplained heterogeneity

This simulation was previously presented in 2.3.1.1. Two covariates are generated independently to build the capture-recapture distribution but only one covariate is considered in the model fitting for the estimation of the population size. In this way, we assess the effect of fitting the partially specified model with unknown information missing.

The capture-recapture distribution $Y$ follows a Poisson with $\lambda_i$ parameter ($Y \sim Po(\lambda_i)$), where $\lambda_i$ is defined as

$$\lambda_i = \mathrm{e}^{-0.02X_{i1}+0.03X_{i2}},$$

and $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$ are independent.

The results (Table 3.4, figures 3.5, 3.6) show that an increase in the number of non-truncated counts leads to a reduction of the standard deviations and an increase in

Table 3.3: Estimates for the model $Y_i \sim Po(\mathrm{e}^{0.04X_1})$ with $X_1 \sim N(20, 225)$

| $N$ | # non-truncated counts | $\hat{N}$ | Empirical SD | RMSE (x 100) | RBias |
|---|---|---|---|---|---|
| 500 | 2 | 496.99 | 14.37 | 0.09 | -0.0060 |
| | 3 | 498.06 | 13.79 | 0.08 | -0.0039 |
| | 4 | 498.40 | 13.33 | 0.07 | -0.0032 |
| | 5 | 498.80 | 13.12 | 0.06 | -0.0024 |
| | 6 | 498.48 | 13.01 | 0.06 | -0.0030 |
| 1000 | 2 | 996.24 | 21.57 | 0.05 | -0.0038 |
| | 3 | 997.78 | 19.58 | 0.04 | -0.0022 |
| | 4 | 998.06 | 18.82 | 0.04 | -0.0019 |
| | 5 | 998.64 | 18.76 | 0.04 | -0.0014 |
| | 6 | 998.90 | 18.08 | 0.03 | -0.0011 |
| 2000 | 2 | 1998.72 | 30.95 | 0.02 | -0.0006 |
| | 3 | 2002.24 | 28.64 | 0.02 | 0.0011 |
| | 4 | 1999.79 | 28.10 | 0.02 | -0.0001 |
| | 5 | 2000.09 | 27.56 | 0.02 | 0.0001 |
| | 6 | 1999.06 | 26.61 | 0.02 | -0.0005 |
| 5000 | 2 | 5013.78 | 52.74 | 0.01 | 0.0028 |
| | 3 | 5010.00 | 50.02 | 0.01 | 0.0020 |
| | 4 | 5005.54 | 47.30 | 0.01 | 0.0011 |
| | 5 | 5004.63 | 46.59 | 0.01 | 0.0009 |
| | 6 | 5002.30 | 45.16 | 0.01 | 0.0005 |

the bias of the estimates. For this example, the model based on 3 counts present the best RMSE values. The model with 2 non-truncated counts is unstable for small samples but it provides similar RMSE values than the 3 counts model in large samples. The variance follows an asymptotic increase with respect to the population size $N$, in contrast to a decreasing trend of the RMSE estimates (figure 3.6) . The developed analytical standard deviation is close to the empirical standard deviation, calculated as the standard deviation of the estimates from all replications ($R$) of the simulation $\left(SD_{emp} = \sum_{k=1}^{R} \sqrt{\frac{1}{R}(\hat{N}_{GC} - \bar{\hat{N}}_{GC})^2}\right)$ (Table 3.4).

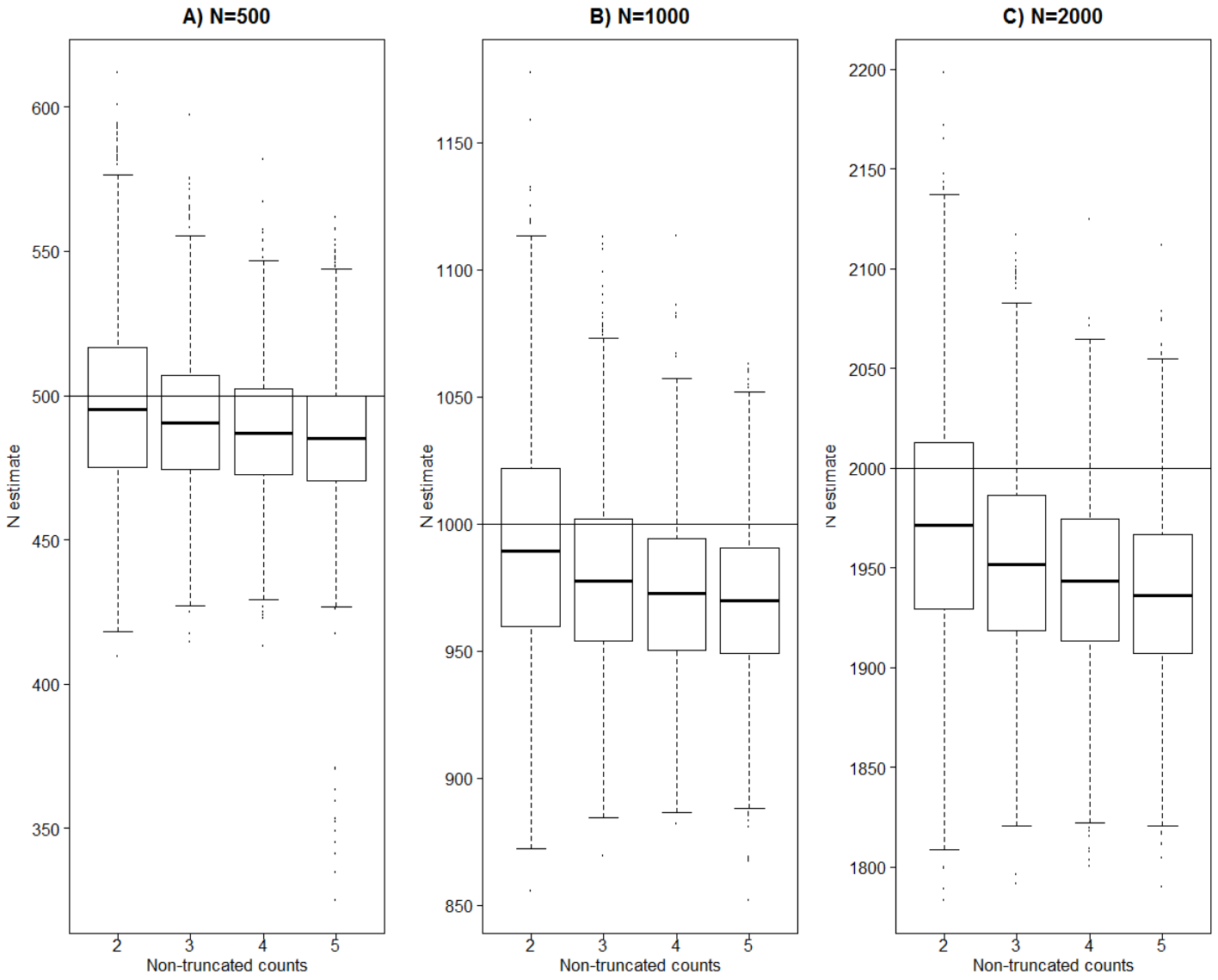


Figure 3.5: Boxplots based on the estimates for the model $Y_i \sim Po(\mathrm{e}^{-0.02X_{i1}+0.03X_{i2}})$ with $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$. Estimates based on models including only $X_1$. A) $N = 500$ B) $N = 1000$ C) $N = 2000$

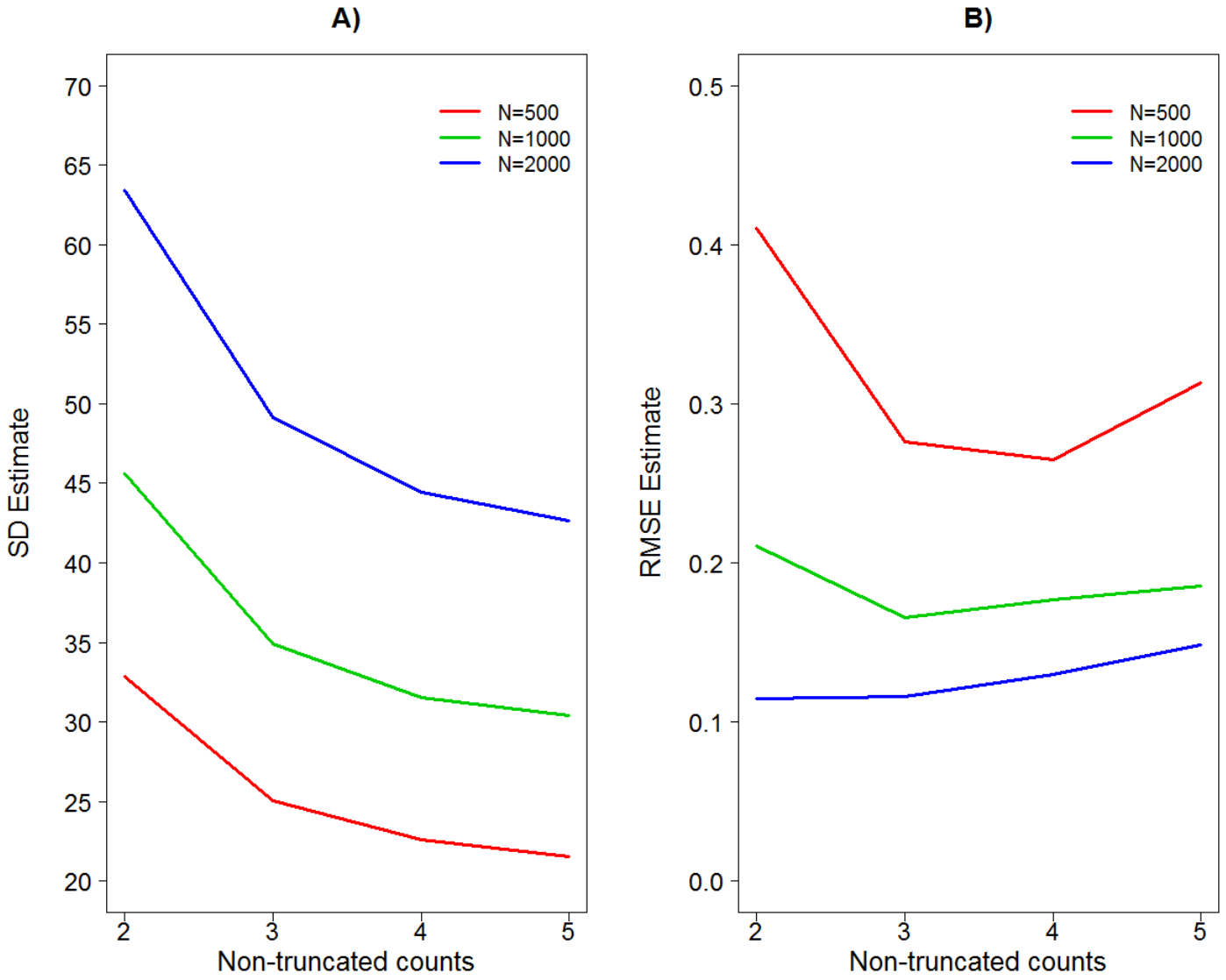


Figure 3.6: SD and RMSE estimates for the model $Y_i \sim Po(\mathrm{e}^{-0.02X_{i1}+0.03X_{i2}})$ with $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$. Estimates based on models including only $X_1$. A) SD estimates B) Relative Mean Squared Error (RMSE) x 100

Table 3.4: Estimates for the model $Y_i \sim Po(\mathrm{e}^{-0.02X_{i1}+0.03X_{i2}})$ with $X_1 \sim N(5, 64)$ and $X_2 \sim N(8, 64)$. Estimates based on models including only $X_1$.

| $N$ | # non-truncated counts | $\hat{N}$ | $SD_{emp}$ | $SD_{ana}$ | RMSE (x 100) | RBias |
|---|---|---|---|---|---|---|
| 500 | 2 | 496.84 | 27.11 | 32.86 | 0.30 | -0.0063 |
| | 3 | 491.56 | 23.89 | 25.05 | 0.26 | -0.0169 |
| | 4 | 488.86 | 22.64 | 22.58 | 0.26 | -0.0223 |
| | 5 | 482.80 | 23.20 | 21.57 | 0.33 | -0.0344 |
| 1000 | 2 | 988.99 | 40.46 | 45.65 | 0.18 | -0.0110 |
| | 3 | 979.51 | 33.89 | 34.91 | 0.16 | -0.0205 |
| | 4 | 974.32 | 31.51 | 31.51 | 0.17 | -0.0257 |
| | 5 | 968.21 | 30.62 | 30.38 | 0.19 | -0.0318 |
| 2000 | 2 | 1957.27 | 52.21 | 63.40 | 0.11 | -0.0214 |
| | 3 | 1955.70 | 47.50 | 49.10 | 0.11 | -0.0222 |
| | 4 | 1912.99 | 38.24 | 44.42 | 0.23 | -0.0435 |
| | 5 | 1935.13 | 42.89 | 42.65 | 0.15 | -0.0324 |

## 3.5 Conclusions

We developed a framework to extend the generalised Chao estimator to include also individuals captured more than twice ($J \geq 2$). The methodology did not present the same appeal seen in the previous chapter where standard statistical software could be employed to obtain the generalised Chao estimator with 2 non-truncated counts. However, we have implemented R functions that will be available online to provide a user-friendly environment to obtain the estimates presented in the current chapter.

Two methods were applied to obtain estimates, the EM algorithm and numerical optimisation algorithms (Nelder-Mead, BFGS, etc...). The EM algorithm should be more efficient, however when covariate information was included, we did not find a solution in the $M$ step, so we use Nelder-Mead to maximise the likelihood and obtain the parameter estimates to replace in the calculation of the expectations.

All simulations conducted concluded that estimates with 2 and 3 non-truncated counts were the most efficient, in the sense of the RMSE criteria. However, other captures distributions might require larger amount of information and a potential test to determine the optimal cut-off would complement the analysis. A $\chi^2$-squared test to find the right amount of truncation will be presented in chapter 5.

The bias/variance tradeoff was observed across the information utilised within each true population size when unexplained heterogeneity was present. There was a negative trend in the variability with respect to the number of counts included in the analysis against

a positive trend of the bias. In terms of the RMSE, the efficiency of the estimator improved asymptotically with respect to the population size.

# Chapter 4

# Power Series

In this chapter, we generalise the framework developed for the Poisson distribution (chapter 3) to the family of power series of distributions. The chapter is structured similarly to previous chapters. We initially present the calculations without using covariate information to describe the simplest case. The methodology is later extended to incorporate variables at individual level related to the probability of being captured. Simulations are carried out to assess the characteristics of our estimators and cases studies are also discussed. Final conclusions are highlighted at the end.

Power series distributions are discrete distributions on $\mathbb{N}$ where the probability density function takes the following form

$$P(X = k|\theta) = \frac{a_k \theta^k}{g(\theta)} \tag{4.1}$$

with $g(\theta) = \sum_{k=0}^{\infty} a_k \theta^k$ with $\theta \in \mathbb{R}$, $k = 0, 1, 2, ...$

We are mainly interested in the binomial, Poisson and geometric distributions, particular cases of the power series family that appear frequently in the capture-recapture area. The parameters for these specific distributions are defined in table 4.1. The $h()$ link function in the table will be described later in the context of estimators including covariate information.

Table 4.1: Parametrization for the Power Series Distributions

| Distribution | Original parameters | $\theta$ | $a_k$ | $g(\theta)$ | $k$ | $h$-function |
|---|---|---|---|---|---|---|
| Binomial | $Bin(m,p)$ | $\frac{p}{1-p}$ | $\binom{m}{k}$ | $(1+\theta)^m$ | $0,1,..,m$ | $\log(\theta_i)$ |
| Poisson | $Po(\lambda)$ | $\lambda$ | $\frac{1}{k!}$ | $\mathrm{e}^{\theta}$ | $0,1,2...$ | $\log(\theta_i)$ |
| Geometric | $G(p)$ | $1-p$ | $1$ | $\frac{1}{1-\theta}$ | $0,1,2,...$ | $\log(\frac{1-\theta_i}{\theta i})$ |

## 4.1 Power series distribution without covariates point estimation

### 4.1.1 2 counts

We initially build the methodology for the simplest scenario where all counts were truncated except counts of one and two ($J = 2$). The only two probabilities to define are

$$q_1 = P(X = 1|\theta) = \frac{a_1\theta/g(\theta)}{\frac{a_1\theta}{g(\theta)} + \frac{a_2\theta^2}{g(\theta)}} = \frac{a_1}{a_1 + a_2\theta}$$
$$q_2 = 1 - q_1 = P(X = 2|\theta) = \frac{a_2\theta^2/g(\theta)}{\frac{a_1\theta}{g(\theta)} + \frac{a_2\theta^2}{g(\theta)}} = \frac{a_2\theta}{a_1 + a_2\theta}.$$

Therefore the likelihood function presents the following form

$$\mathcal{L}(\theta) = q_1^{f_1} q_2^{f_2}$$

and subsequently the log-likelihood is described as

$$\begin{aligned}\ell(\theta) &= f_1 \log\left(\frac{a_1}{a_1 + a_2\theta}\right) + f_2 \log\left(\frac{a_2\theta}{a_1 + a_2\theta}\right) \\ &= f_1 \log(a_1) + f_2 \log(a_2) + f_2 \log(\theta) - (f_1 + f_2)\log(a_1 + a_2\theta).\end{aligned}$$

The log-likelihood can be maximised calculating the first derivative with respect to $\theta$ and solving the score equation $\frac{\ell(\theta)}{d\theta} = 0$

$$\frac{d\ell(\theta)}{d\theta} = f_2\frac{1}{\theta} - (f_1 + f_2)\frac{a_2}{a_1 + a_2\theta} = 0.$$

Equivalently, we can write

$$a_1 f_2 - f_1 a_2\theta = 0.$$

The maximum likelihood estimator takes the form

$$\hat{\theta} = \frac{a_1 f_2}{a_2 f_1}. \tag{4.2}$$

The specific estimates for the binomial, Poisson and geometric case are deduced replacing $\theta, a_1, a_2$ with the parametrisations shown in table 4.1:

- **Binomial**

$$\frac{\hat{p}}{1-\hat{p}} = \frac{\binom{m}{1} f_2}{\binom{m}{2} f_1} \Rightarrow \hat{p} = \frac{2f_2}{(m-1)f_1 + 2f_2}.$$

- **Poisson**

$$\hat{\lambda} = \frac{(1/1!)f_2}{(1/2!)f_1} \Rightarrow \hat{\lambda} = \frac{2f_2}{f_1}.$$

- **Geometric**

$$1 - \hat{p} = \frac{f_2}{f_1} \Rightarrow \hat{p} = \frac{f_1 - f_2}{f_1}.$$

Notice that the Poisson estimate coincides with the calculations shown in previous chapters.

### 4.1.2 J counts

In this section, we extend the formulae to include the first $J$ counts into the likelihood. Two approaches could be considered: 1) the application of the EM algorithm with the likelihood assuming complete data and 2) the likelihood based on $J$ non-truncated counts.

#### 4.1.2.1 Complete likelihood

The EM algorithm can be applied following the reasoning described in section 3.1.1.

- **M step**
  The likelihood and probabilities are calculated based on the assumption of the

data set being complete or missing data being imputed. Therefore, the likelihood is defined as

$$\mathcal{L}(p|\theta) = \prod_{k=0}^{m} p_{k|\theta}^{f_k} ,$$

where

$$p_{k|\theta} = P(X = k|\theta) = a_k\theta^k/g(\theta)$$

is the probability of a unit being captured exactly $k$ times.

Hence, the log-likelihood is written

$$\ell(\theta) = e_0\left(\log(a_0) - \log(g_\theta)\right) + \left(\sum_{k=1}^{J} f_k \log(a_k)\right) + \left(\sum_{k=1}^{J} k f_k\right)\log(\theta) - \left(\sum_{k=1}^{J} f_k \log(g_\theta)\right)$$
$$\left(\sum_{l=J+1}^{m} e_l \log(a_l)\right) + \left(\sum_{l=J+1}^{m} l e_l\right)\log(\theta) - \left(\sum_{l=J+1}^{m} e_l \log(g_\theta)\right) , \tag{4.3}$$

where $e_i = E(f_i|\theta)$, $i \in \{0, J+1, \ldots, m\}$ are unknown parameters that represent the expected number of units captured $0, J+1, \ldots, m$ times. These parameters need to be estimated to obtain estimates for $\theta$.

- **E step**

  The E step aims to provide estimates of the expectations of capture frequencies to obtain a maximum likelihood estimator from (4.3). The expectation of being captured exactly $y$ times is

$$e_y = E(f_y|\theta) = p(X = y|\theta)N = p(X = y|\theta)\left(e_0 + f_1 + f_2 + ... + \sum_{j=J+1}^{\infty} e_j\right) \tag{4.4}$$

  then

$$e_0 + e_{J+1}^{+} = \left[1 - p(X = 1|\theta) - p(X = 2|\theta) - ... - p(X = J|\theta)\right]\left(e_0 + e_{J+1}^{+}\right) +$$
$$\left[1 - p(X = 1|\theta) - p(X = 2|\theta) - ... - p(X = J|\theta)\right](f_1 + f_2 + ... + f_J)$$

  with $e_{J+1}^{+} = \sum_{j=J+1}^{\infty} e_j$. Solving for $e_0 + e_{J+1}^{+}$, we obtain

$$e_0 + e_{J+1}^{+} = \frac{\left[1 - p(X = 1|\theta) - p(X = 2|\theta)... - p(X = J|\theta)\right]}{p(X = 1|\theta) + p(X = 2|\theta) + ... + p(X = J|\theta)}(f_1 + f_2 + ... + f_J) \tag{4.5}$$

Therefore, replacing (4.5) in (4.4)

$$\begin{aligned} e_y &= p(X=y|\theta)\left(e_0 + f_1 + f_2 + ... + f_J + \sum_{j=J+1}^{\infty} e_j\right) \\ &= p(X=y|\theta)(f_1 + f_2 + ...f_J)\left[1 + \frac{[1 - p(X=1|\theta) - ... - p(X=J|\theta)]}{p(X=1|\theta) + ... + p(X=J|\theta)}\right] \\ &= \frac{p(X=y|\theta)}{p(X=1|\theta) + ... + p(X=J|\theta)}(f_1 + f_2 + ... + f_J) \\ &= \frac{a_y\theta^y}{\sum_{j=1}^{J} a_j\theta^j}(f_1 + f_2 + ... + f_J). \end{aligned} \tag{4.6}$$

Consequently, our primary outcome $y = 0$ can be estimated substituting the probabilities in (4.6) as

$$e_0 = \frac{a_0}{\sum_{j=1}^{J} a_j\theta^j}(f_1 + f_2 + ... + f_J). \tag{4.7}$$

Obviously, $\theta$ is unknown and needs to be estimated to be able to calculate $e_0$. The EM algorithm starts from a chosen initial value $\hat{\theta}_0$ that follows the estimation of the unknown expectations of capture frequencies; it similarly leads to a new estimate $\hat{\theta}_1$ coming from the maximisation of the likelihood. This iterative process continues until the difference between two consequent estimates is less than a predefined tolerance value $\tau$, $\mid \hat{\theta}_{t+1} - \hat{\theta}_t \mid < \tau$.

For the simplest case where only 2 counts are considered, $e_0$ can be easily obtained for the three distributions of interest replacing the parameters shown in table 4.1:

- **Binomial distribution**: Here we have that

$$\begin{aligned} e_0 &= \left(\frac{1}{m\dfrac{\hat{p}}{1-\hat{p}} + \dfrac{m(m-1)}{2}\left(\dfrac{\hat{p}}{1-\hat{p}}\right)^2}\right)(f_1 + f_2) \\ &= \left(\frac{2(1-\hat{p})^2}{2m\hat{p}(1-\hat{p}) + m(m-1)\hat{p}^2}\right)(f_1 + f_2) \end{aligned}$$

Replacing $\hat{p} = \dfrac{2f_2}{(m-1)f_1 + 2f_2}$

$$= \frac{\dfrac{2(m-1)^2 f_1^2}{((m-1)f_1 + 2f_2)^2}}{\left(\left(\dfrac{4mf_2}{(m-1)f_1 + 2f_2}\right)\left(\dfrac{(m-1)f_1}{(m-1)f_1 + 2f_2}\right) + m(m-1)\dfrac{4f_2^2}{((m-1)f_1 + 2f_2)^2}\right)}(f_1 + f_2)$$

$$= \frac{2(m-1)^2 f_1^2}{4mf_2(m-1)f_1 + m(m-1)(2f_2)^2}(f_1 + f_2) = \frac{(m-1)f_1^2}{2mf_2}. \tag{4.8}$$

We recognise the classic Chao's estimator for the binomial case where the number of captured occasions is fixed a priori.

- **Poisson distribution**: For the Poisson we find

$$e_0 = \frac{1}{\hat{\lambda} + \frac{\hat{\lambda}^2}{2}}(f_1 + f_2)$$

Replacing $\hat{\lambda} = 2f_2/f_1$ we obtain Chao's lower bound estimator for continuous-time experiments:

$$e_0 = \frac{1}{\frac{2f_2}{f_1} + \frac{2f_2^2}{f_1^2}}(f_1 + f_2) = \frac{1}{\frac{2f_2 f_1 + 2f_2^2}{f_1^2}}(f_1 + f_2) = \frac{f_1^2}{2f_2}. \tag{4.9}$$

- **Geometric distribution:** In the geometric case we find that

$$e_0 = \frac{1}{(1-\hat{p}) + (1-\hat{p})^2}(f_1 + f_2).$$

Inserting for $\hat{p}$ the previously calculated maximum likelihood estimator $\hat{p} = \frac{f_1 - f_2}{f_1}$

$$e_0 = \frac{f_1 + f_2}{\frac{f_2}{f_1} + \frac{f_2^2}{f_1^2}} = \frac{f_1^2}{f_2}. \tag{4.10}$$

This estimator was described in Niwitpong et al. (2012) and the results presented in the paper will be discussed in greater detail in the following chapters.

#### 4.1.2.2 Truncated likelihood

Another approach consists in working directly with the truncated distribution; the probabilities for the non-truncated counts are defined as

$$p_k = P(X = k|\theta) = \frac{a_k\theta^k/g(\theta)}{\sum_{j=1}^{J}\frac{a_j\theta^j}{g(\theta)}} = \frac{a_k\theta^k}{\sum_{j=1}^{J} a_j\theta^j}. \tag{4.11}$$

The likelihood function is written as

$$\mathcal{L}(p) = \prod_{j=1}^{J} p_j^{f_j}$$

accordingly the log-likelihood is

$$\ell(\theta) = \left(\sum_{k=1}^{J} f_k \log(a_k)\right) + \left(\sum_{k=1}^{J} k f_k\right)\log(\theta) - \left(\sum_{k=1}^{J} f_k\right)\log\left(\sum_{j=1}^{J} a_j\theta^j\right). \tag{4.12}$$

Term $\theta$ can be estimated maximising the likelihood with a numerical algorithm and be used later in the calculation of the expectations. The formula for the estimation of the expectation of $f_0$ is equal to the formula presented in (4.7).

Tables 4.2 and 4.3 contain the formulae for the estimation of $N$ and $f_0$ for the binomial and geometric distribution with $J$ counts with and without covariate information.

## 4.2 Power series distribution with covariates point estimation

In the situation when the probability of being captured is not the same for all study units, we propose to include covariate information for each captured unit. These variables are linked to $\theta$ by a link function $h$. 4.1 contains the canonical link functions for the distributions of interest, but other link functions could be used.

$$h(\theta_i) = \alpha + \beta' Z_i, \tag{4.13}$$

where $\beta$ is a vector of coefficients with length equal the number of covariates. In this case we only consider the approach with the truncated likelihood using $J$ non-truncated counts as the EM algorithm cannot find a solution for $\alpha$ and $\beta$ in the M-step and we

need a numerical algorithm. Therefore, the probabilities with respect to $\alpha$ and $\beta$ are

$$P(X = k|\theta_i) = \frac{a_k\theta_i^k/g(\theta_i)}{\sum_{j=1}^{J} \frac{a_j\theta_i^j}{g(\theta_i)}} = \frac{a_k\left(h^{-1}(\alpha+\beta' Z_i)\right)^k}{\sum_{j=1}^{J} a_j\left(h^{-1}(\alpha+\beta' Z_i)\right)^j}, \tag{4.14}$$

as $g(\theta_i)$ cancels out. From the log-likelihood presented in the previous section (4.1.2.2), we replace $\theta_i$ by the its inverse link to the linear predictor defined in (4.13).

Let $M_J$ be the number of covariate combinations when $J$ non-truncated counts are used, and $f_{ik}$ be the number of individuals captured $k$ times with the $i$th covariate combination.

We have the following log-likelihoods:

$$\ell(\theta_i) = \sum_{i=1}^{M_J}\sum_{k=1}^{J} f_{ik}\log(a_k) + \sum_{i=1}^{M_J}\left(\sum_{k=1}^{J} kf_{ik}\right)\log(\theta_i) - \sum_{i=1}^{M_J}\left(\left(\sum_{k=1}^{J} f_{ik}\right)\log\left(\sum_{j=1}^{J} a_j\theta_i^j\right)\right) \tag{4.15}$$

$$\begin{aligned}\ell(\alpha,\beta) = &\sum_{i=1}^{M_J}\sum_{k=1}^{J} f_{ik}\log(a_k) + \sum_{i=1}^{M_J}\left(\sum_{k=1}^{J} kf_{ik}\right)\log(h^{-1}(\alpha+\beta' Z_i)) \\ &- \sum_{i=1}^{M_J}\left(\left(\sum_{k=1}^{J} f_{ik}\right)\log\left(\sum_{j=1}^{J} a_j\left(h^{-1}(\alpha+\beta' Z_i)\right)^j\right)\right).\end{aligned} \tag{4.16}$$

The log-likelihood with respect to $\alpha$ and $\beta$ for the binomial and the geometric distribution can be found in tables 4.2 and 4.3.

The expected number of units captured $y$ times is calculated as

$$e_y = E(f_y|\alpha,\beta) = \sum_{i=1}^{M_J} e_{iy},$$

where $e_{iy} = E(f_{iy}|\alpha,\beta)$ is the expected number of units captured $y$ times for the $i$-th covariate combination. Hence $e_y$ can be estimated as the sum of the expected number

of units captured $y$ times across all covariates combinations:

$$e_{iy} = p(X = y|\alpha, \beta)N_i = p(X = y|\alpha, \beta)\left(e_{i0} + f_{i1} + f_{i2} + ... + \sum_{j=J+1}^{\infty} e_{ij}\right) . \quad (4.17)$$

$e_{i0} + \sum_{j=J+1}^{\infty} e_{ij}$ is unknown and need to be calculated. We have the following:

$$e_{i0} + e^{+}_{i(J+1)} = [1 - p(X = 1|\alpha, \beta) - ... - p(X = J|\alpha, \beta)]\left(e_{i0} + e^{+}_{i(J+1)}\right) + [1 - p(X = 1|\alpha, \beta) - ... - p(X = J|\alpha, \beta)]\left(f_{i1} + f_{i2} + ... + f_{iJ}\right)$$

with $e^{+}_{i(J+1)} = \sum_{j=J+1}^{\infty} e_{ij}$.

Solving for $e_{i0} + e^{+}_{i(J+1)}$, we yield

$$e_{i0} + e^{+}_{i(J+1)} = \frac{[1 - p(X = 1|\alpha, \beta)... - p(X = J|\alpha, \beta)]}{p(X = 1|\alpha, \beta)... + p(X = J|\alpha, \beta)}(f_{i1} + f_{i2} + ... + f_{iJ}). \quad (4.18)$$

We replace $e_{i0} + e^{+}_{i(J+1)}$ on (4.17) and achieve

$$\begin{aligned}
e_{iy} &= p(X = y|\alpha, \beta)\left(e_{i0} + f_{i1} + f_{i2} + ... + f_{iJ} + \sum_{j=J+1}^{\infty} e_{ij}\right) \\
&= p(X = y|\alpha, \beta)(f_{i1} + f_{i2} + ...f_{iJ})\left[1 + \frac{[1 - p(X = 1|\alpha, \beta)... - p(X = J|\alpha, \beta)]}{p(X = 1|\alpha, \beta)... + p(X = J|\alpha, \beta)}\right] \\
&= \frac{p(X = y|\alpha, \beta)}{p(X = 1|\alpha, \beta) + \ldots + p(X = J|\alpha, \beta)}(f_{i1} + f_{i2} + ... + f_{iJ}) \\
&= \frac{\frac{a_y\left(h^{-1}(\alpha + \beta' Z_i)\right)^y}{g(h^{-1}(\alpha + \beta' Z_i))}}{\sum_{j=1}^{J} a_j \frac{\left(h^{-1}(\alpha + \beta' Z_i)\right)^j}{g(h^{-1}(\alpha + \beta' Z_i))}}(f_{i1} + f_{i2} + ... + f_{iJ}) \\
&= \frac{a_y\left(h^{-1}(\alpha + \beta' Z_i)\right)^y}{\sum_{j=1}^{J} a_j\left(h^{-1}(\alpha + \beta' Z_i)\right)^j}(f_{i1} + f_{i2} + ... + f_{iJ}).
\end{aligned}$$

$E(f_y|\alpha, \beta')$ will be the sum over all covariate combinations $M_J$

$$e_y = \sum_{i=1}^{M_J} e_{iy} = \sum_{i=1}^{M_J} \frac{a_y\left(h^{-1}(\hat{\alpha} + \hat{\beta}' Z_i)\right)^y}{\sum_{j=1}^{J} a_j\left(h^{-1}(\hat{\alpha} + \hat{\beta}' Z_i)\right)^j}(f_{i1} + f_{i2} + ... + f_{iJ}). \quad (4.19)$$

To calculate the expected number of non-captured units, $e_0 = E(f_0|\alpha, \beta')$, we use (4.19):

$$e_0 = \sum_{i=1}^{M_J} \frac{a_0}{\sum_{j=1}^{J} a_j\left(h^{-1}(\alpha + \beta' Z_i)\right)^j}(f_{i1} + f_{i2} + ... + f_{iJ}). \quad (4.20)$$

The estimators for the binomial and the geometric case assuming the canonical link as link function are obtained by replacing in (4.20) the parameters presented in table 4.1:

- **Binomial**

$$e_{0_{BIN}} = \sum_{i=1}^{M_J} \frac{1}{\sum_{j=1}^{J} \binom{m}{j} \left(\mathrm{e}^{\alpha+\beta' Z_i}\right)^j} (f_{i1} + f_{i2} + ... + f_{iJ})$$

- **Geometric**

$$e_{0_{GEO}} = \sum_{i=1}^{M_J} \frac{1}{\sum_{j=1}^{J} \left(\frac{1}{1+\mathrm{e}^{\alpha+\beta' Z_i}}\right)^j} (f_{i1} + f_{i2} + ... + f_{iJ}).$$

| Binomial Bin(p,m) | |
|---|---|
| 2 counts without covariates | |
| $\ell(p\|m)$ | $f_1 * \log(m) + f_2 \log(m(m-1)/2) + f_2 \log(\frac{p}{1-p}) - (f_1+f_2)\log(m(1+(m-1)\frac{p}{1-p}))$ |
| $E(f_y)$ | $\frac{\binom{m}{y}\left(\frac{p}{1-p}\right)^y}{m\left(\frac{p}{1-p}\right)+\frac{m(m-1)}{2}\left(\frac{p}{1-p}\right)^2}(f_1+f_2)$ |
| $\hat{N}$ | $n + \frac{1}{m\frac{p}{1-p}+\frac{m(m-1)}{2}(\frac{p}{1-p})^2}(f_1+f_2)$ |
| $J$ counts without covariates | |
| $\ell(p\|m)$ | $\sum_{j=1}^{J} f_j \log(\binom{m}{j}) + \left(\sum_{j=1}^{J} j f_j\right)\log\left(\frac{p}{1-p}\right) - \left(\sum_{i=1}^{J} f_j\right)\log\left(\sum_{j=1}^{J}\binom{m}{j}\left(\frac{p}{1-p}\right)^j\right)$ |
| $E(f_y)$ | $\frac{\binom{m}{y}\left(\frac{\hat{p}}{1-\hat{p}}\right)^y}{\sum_{j=1}^{J}\binom{m}{j}\left(\frac{\hat{p}}{1-\hat{p}}\right)^j}(f_1+...+f_J)$ |
| $\hat{N}$ | $n + \frac{1}{\sum_{j=1}^{J}\binom{m}{j}\left(\frac{\hat{p}}{1-\hat{p}}\right)^j}(f_1+...+f_J)$ |
| $J$ counts and covariates | |
| $\ell(\alpha,\beta\|J)$ | $\sum_{i=1}^{M_J}\sum_{k=1}^{J} f_{ik}\log(\binom{m}{k}) + \sum_{i=1}^{M_J}\left(\sum_{k=1}^{J} k f_{ik}\right)(\alpha+\beta' Z_i) - \sum_{i=1}^{M_J}\left(\left(\sum_{k=1}^{J} f_{ik}\right)\log\left(\sum_{j=1}^{J}\binom{m}{j}\left(\mathrm{e}^{\alpha+\beta' Z_i}\right)^j\right)\right)$ |
| $E(f_y)$ | $\left(\sum_{i=1}^{M_J}\frac{\binom{m}{y}\mathrm{e}^{y*(\hat{\alpha}+\hat{\beta}' Z_i)}}{\sum_{j=1}^{J}\binom{m}{j}\mathrm{e}^{j*(\hat{\alpha}+\hat{\beta}' Z_i)}}\right)(f_1+...+f_J)$ |
| $\hat{N}$ | $n + \left(\sum_{i=1}^{M_J}\frac{1}{\sum_{j=1}^{J}\binom{m}{j}\mathrm{e}^{j*(\hat{\alpha}+\hat{\beta}' Z_i)}}\right)(f_1+...+f_J)$ |

Table 4.2: Estimates for the case of a capture-recapture binomial distribution

| Geometric G(p) | |
|---|---|
| 2 counts without covariates | |
| $\ell(p)$ | $f_2 \log(1-p) - (f_1+f_2)\log(2-p)$ |
| $E(f_y)$ | $\frac{(1-p)^y}{(1-\hat{p})+(1-\hat{p})^2}(f_1+f_2)$ |
| $\hat{N}$ | $\mathrm{n} + \frac{f_1+f_2}{(1-\hat{p})+(1-\hat{p})^2}$ |
| J counts without covariates | |
| $\ell(p\|m)$ | $(\sum_{j=1}^{J} j f_j)\log(1-p) - (\sum_{i=1}^{J} f_j)\log\left(\sum_{j=1}^{J}(1-p)^j\right)$ |
| $E(f_y)$ | $\frac{(1-p)^y}{\sum_{j=1}^{J}(1-\hat{p})^j}(f_1+...+f_J)$ |
| $\hat{N}$ | $n + \frac{1}{\sum_{j=1}^{J}(1-\hat{p})^j}(f_1+...+f_J)$ |
| J counts with covariates | |
| $\ell(\alpha,\beta\|J)$ | $\sum_{i=1}^{M_J}\left(\sum_{k=1}^{J} k f_{ik}\right)\log(\frac{1}{1+\mathrm{e}^{\alpha+\beta' Z_i}}) - \sum_{i=1}^{M_J}\left(\left(\sum_{k=1}^{J} f_{ik}\right)\log\left(\sum_{j=1}^{J}\left(\frac{1}{1+\mathrm{e}^{\alpha+\beta' Z_i}}\right)^j\right)\right)$ |
| $E(f_y)$ | $\sum_{i=1}^{M_J}\frac{1/\left(1+\mathrm{e}^{y*(\hat{\alpha}+\hat{\beta}' Z_i)}\right)}{\sum_{j=1}^{J} 1/\left(1+\mathrm{e}^{j*(\hat{\alpha}+\hat{\beta}' Z_i)}\right)}(f_1+...+f_J)$ |
| $\hat{N}$ | $n + \sum_{i=1}^{M_J}\frac{1}{\sum_{j=1}^{J} 1/\left(1+\mathrm{e}^{j*(\hat{\alpha}+\hat{\beta}' Z_i)}\right)}(f_1+...+f_J)$ |

Table 4.3: Estimates for the case of a capture-recapture geometric distribution

## 4.3 Analytical variance: the case with covariates

We follow the conditioning moments methodology (Ross, 1985) as shown in previous chapters (2.2).

$$Var(\hat{N}_{GC}) = Var[E(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)] + E[Var(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)]. \quad (4.21)$$

The variability from the sampling is represented by the first term of the right hand side of the equation. For its estimation we initially calculate

$$E(\hat{N}_{GC}|\Delta_i, i = 1, ..., N) = E\left(n + \sum_{i=1}^{N} \frac{a_0 \Delta_i}{\sum_{j=1}^{J} a_j \hat{\theta}_i^j} |\Delta_i, i = 1, ...N\right) \approx \sum_{i=1}^{N} \Delta_i \omega_i \ ,$$

where $\Delta_i$ is a binary variable defined as

$$\Delta_i = \begin{cases} 1, & y_i \in \{1, 2, ..., J\} \\ 0, & otherwise \end{cases}$$

and

$$\omega_i = 1 + \frac{a_0}{p_i g(\theta_i)} \quad (4.22)$$

and $p_i$ is the probability of being captured between 1 and $J$ times ($p(\Delta_i = 1)$):

$$p_i = p(\Delta_i = 1) = P(Y_i \in \{1, 2, ..., J\}) = P(Y_i = 1|\theta_i) + ...P(Y_i = J|\theta_i) = \sum_{j=1}^{J} a_j \theta_i^j / g(\theta_i).$$

The expectation and variance of $\Delta_i$ is

$$E(\Delta_i) = p_i \text{ and } Var(\Delta_i) = p_i(1 - p_i).$$

Therefore, the variance of the conditional expected value $E(\hat{N}_{GC})$ is

$$Var[E(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)] = Var\left(\sum_{i=1}^{N} \Delta_i \omega_i\right) = \sum_{i=1}^{N} p_i(1 - p_i)\omega_i^2.$$

This variance can be estimated using the Horvitz-Thompson estimator which leads to:

$$\widehat{Var}[E(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)] = \sum_{i=1}^{N} \frac{\Delta_i}{\hat{p}_i} \hat{p}_i(1 - \hat{p}_i)\hat{\omega}_i^2 = \sum_{i=1}^{f_1+...+f_J} (1 - \hat{p}_i) \left(1 + \frac{a_0}{\hat{p}_i g(\hat{\theta}_i)}\right)^2. \quad (4.23)$$

The second term in (4.21) relates to the variability from the estimate itself. $Var(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)$ is calculated and considered in itself as a moment estimator for the expected value. The multivariate $\delta$-method is applied:

$$E[Var(\hat{N}_{GC}|\Delta_i, i = 1, ..., N)] \approx \nabla v(\hat{\alpha}, \hat{\beta})' c\hat{o}v(\hat{\alpha}, \hat{\beta}) \nabla v(\hat{\alpha}, \hat{\beta}) \ , \tag{4.24}$$

where

$$v(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{f_1+...+f_J} \frac{a_0}{\sum_{j=1}^{J} a_j \theta_i^j} = \sum_{i=1}^{f_1+...+f_J} \frac{a_0}{\sum_{j=1}^{J} a_j \left(h^{-1}(\alpha + \beta' Z_i)\right)^j}$$

$$\text{and } \nabla v(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} \frac{\partial v}{\partial \alpha} \\ \frac{\partial v}{\partial \beta_1} \\ ... \\ \frac{\partial v}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-a_0 \sum_{j=1}^{J} j a_j \theta_i^{j-1} \partial \theta_i / \partial \alpha}{\left(\sum_{j=1}^{J} a_j \theta_i^j\right)^2} \\ \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-a_0 \sum_{j=1}^{J} j a_j \theta_i^{j-1} \partial \theta_i / \partial \beta_1}{\left(\sum_{j=1}^{J} a_j \theta_i^j\right)^2} Z_{i1} \\ ... \\ \sum_{i=1}^{f_1+f_2+...+f_J} \frac{-a_0 \sum_{j=1}^{J} j a_j \theta_i^{j-1} \partial \theta_i / \partial \beta_p}{\left(\sum_{j=1}^{J} a_j \theta_i^j\right)^2} Z_{iJ} \end{pmatrix}$$

Replacing $\theta_i = h^{-1}(\alpha + \beta' Z_i)$, we obtain:

$$\nabla v(\hat{\alpha},\hat{\beta}) = \begin{pmatrix} \sum_{i=1}^{f_1+f_2+\ldots+f_J} \dfrac{-a_0 \sum_{j=1}^{J} j a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j-1} \partial h^{-1}(\alpha+\beta' Z_i)/\partial\alpha}{\left(\sum_{j=1}^{J} a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j}\right)^2} \\ \sum_{i=1}^{f_1+f_2+\ldots+f_J} \dfrac{-a_0 \sum_{j=1}^{J} j a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j-1} \partial h^{-1}(\alpha+\beta' Z_i)/\partial\beta_1}{\left(\sum_{j=1}^{J} a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j}\right)^2} Z_{i1} \\ \ldots \\ \sum_{i=1}^{f_1+f_2+\ldots+f_J} \dfrac{-a_0 \sum_{j=1}^{J} j a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j} \partial h^{-1}(\alpha+\beta' Z_i)/\partial\beta_p}{\left(\sum_{j=1}^{J} a_j \left(h^{-1}(\alpha+\beta' Z_i)\right)^{j}\right)^2} Z_{iJ} \end{pmatrix}$$

where $\beta$ is the vector of coefficients defined in (4.13). The covariance matrix $cov(\hat{\alpha},\hat{\beta})$ is the inverse of the observed Fisher information, that can also be estimated as part of the numerical algorithm used to maximise the likelihood. The likelihood of the model with respect to parameters $\alpha$ and $\beta$ was described in (4.15).

## 4.4 Simulations

In this section, several scenarios are simulated with focus only on the binomial distribution. Simulations for the geometric distribution will appear in a later chapter where it will be compared to other new estimators suitable for a capture probability based on a geometric distribution.

### 4.4.1 Estimators for comparison

Our estimator is compared to two other estimators that assume a binomial distribution and constant capture probability over individuals.

#### 4.4.1.1 Maximum likelihood estimator for the binomial distribution

We now deduce the maximum likelihood estimator under homogeneity for the binomial distribution using the EM algorithm. We assume we have complete data or any missing data can be imputed. For the maximisation step (M step) we calculate the likelihood

function.
The capture probabilities are defined by

$$q_j = P(Y = j) = \binom{m}{j} p^j (1-p)^{m-j} \ ,$$

where $m$ is the number of capture occasions and $p$ is the probability of being captured, the same for all individuals because homogeneity is assumed. The likelihood takes this form

$$\mathcal{L}(p|m) = \prod_{j=0}^{m} q_j^{f_j}.$$

and the log-likelihood is

$$\begin{aligned}\ell(p|m) &= \left(\sum_{j=1}^{m} f_j \binom{m}{j}\right) + \left(\sum_{j=1}^{m} j f_j\right) \log(p) + \left(\sum_{j=0}^{m} (m-j) \times f_j\right) \log(1-p) \\ &= \left(\sum_{j=1}^{m} f_j \binom{m}{j}\right) + \left(\sum_{j=1}^{m} j f_j\right) \log(p) + \left(m e_0 + \sum_{j=1}^{m} (m-j) \times f_j\right) \log(1-p),\end{aligned}$$

where $e_0 = E(f_0|p)$ is the expected number of units that have been never captured.
An estimate for $p$ can be obtained by maximising the log-likelihood, or calculating the roots of the likelihood equation:

$$\frac{d\ell(p|m)}{dp} = \frac{\left(\sum_{j=1}^{m} j f_j\right)}{p} - \frac{\left(\sum_{j=0}^{m} (m-j) \times f_j\right)}{1-p} = 0.$$

We obtain

$$\hat{p} = \frac{\sum_{j=1}^{m} j f_j}{m \times (n + e_0)} \ , \tag{4.25}$$

but $e_0$ is unknown. It can be estimated in the E step as

$$E(f_0|\hat{p}) = e_0 = N * q_0 = (n + e_0)(1-\hat{p})^m \ ,$$

from where

$$\hat{e}_0 = \frac{n(1-\hat{p})^m}{1-(1-\hat{p})^m}. \tag{4.26}$$

follows.

The EM algorithm starts with an initial value at stage 0 $\hat{p}_0$ to be used in the calculation

of $\hat{e}_0$, which is consequently applied for the estimation of $p$ at stage 1, $\hat{p}_1$. The process repeats recursively until the difference of the estimates between two consecutive steps are smaller than a chosen tolerance $\tau$, $\mid p_{k+1} - p_k \mid < \tau$.

#### 4.4.1.2 Generalised Turing estimator for power series

In this section, the Turing estimator is obtained for the power series and the particular cases of the Poisson, the binomial and the geometric distribution.

First, we present Turing's reasoning to estimate $p_0 = P(Y = 0|\lambda)$ for the Poisson case. We have that

$$P(Y = 0|\lambda) = \mathrm{e}^{-\lambda} = \frac{\lambda \mathrm{e}^{-\lambda}}{\lambda} = \frac{p_1}{E(Y)},$$

which can be estimated as $\frac{f_1/N}{S/N}$ where $S = \sum_{i=1}^{m} i f_i$, being $m$ the maximum number of times a unit was captured. This leads to the Turing estimator $\hat{N} = \frac{n}{1-f_1/S}$.

As we assumed at the beginning of the chapter,

$$P(Y = k|\theta) = \frac{a_k \theta^k}{g(\theta)}$$

where $g(\theta)$ is the normalising constant to make the probabilities sum up to 1. It is assumed that

$$g(\theta) = \sum_{k=0}^{\infty} a_k \theta^k, \text{with } \theta \in \mathbb{R}.$$

We calculate the probabilities for $Y = 0$ and $Y = 1$ to follow the same induction process shown above for the Poisson case:

$$P(Y = 0|\theta) = \frac{a_0}{g(\theta)}$$
$$P(Y = 1|\theta) = \frac{a_1 \theta}{g(\theta)}.$$

On the other hand,

$$E(Y|\theta) = \theta \frac{g\prime(\theta)}{g(\theta)}.$$

Then, we can apply the ratio that Good (1953) showed $\frac{p_1}{E(X)}$

$$\frac{p_1}{E(Y|\theta)} = \frac{\frac{a_1\theta}{g(\theta)}}{\frac{\theta g\prime(\theta)}{g(\theta)}} = \frac{a_1}{g\prime(\theta)} \Rightarrow \frac{a_0}{a_1}\frac{p_1}{E(Y|\theta)} = \frac{a_0}{g\prime(\theta)}.$$

We know that $p_0 = \frac{a_0}{g_\theta} = \frac{a_0}{g(g'^{-1}(g'(\theta))}$ leading to

$$\frac{a_0}{g\left(g\prime^{-1}(g\prime(\theta))\right)} = p_0.$$

Consequently, the estimate for $p_0$ for the power series is

$$\hat{p}_0 = \frac{a_0}{g\left(g\prime^{-1}\left(\frac{a_1 S}{f_1}\right)\right)}.$$

Specific estimates for distributions of interest are obtained replacing the power series parameters to the common parameters of those distributions (4.1).

- **Poisson case**
  We know $a_x$ and $g(\theta)$:

$$a_x = 1/x!$$
$$g(\theta) = \mathrm{e}^{\theta}.$$

  Firstly we calculate $g\prime(\theta)$, the derivative of $g(\theta)$:

$$g\prime(\theta) = \lambda = \mathrm{e}^{\theta}.$$

  Therefore, the inverse of $g\prime(\theta)$ is

$$g\prime^{-1}(\theta) = \log(\lambda)$$
$$\hat{p}_0 = \frac{1}{\mathrm{e}^{\log\left(\frac{S}{f_1}\right)}} = \frac{f_1}{S}.$$

  If $\hat{p_0}$ is replaced in the Horvitz-Thompson estimator, $\hat{N}$ is estimated as

$$\hat{N}_{Turing-Poi} = \frac{n}{1 - \frac{f_1}{S}}. \tag{4.27}$$

- **Binomial case**

  The known parameters for the binomial case are

$$a_x = \binom{m}{x}$$
$$g(\theta) = (1+\theta)^m .$$

  We calculate the first derivative of $g(\theta)$ and its inverse function

$$\theta = g\prime(\theta) = m\,(1+\theta)^{m-1} = \lambda$$
$$g\prime^{-1}(\lambda) = \lambda = \left(\frac{\lambda}{m}\right)^{1/(m-1)} - 1,$$

  so that

$$\hat{p}_0 = \left(\frac{f_1}{S}\right)^{\frac{m}{m-1}}.$$

  The total population can be estimated using $\hat{p}_0$ in the Horvitz-Thompson estimator

$$\hat{N}_{Turing-Bin} = \frac{n}{1 - (\frac{S}{f_1})^{\frac{m}{m-1}}}. \tag{4.28}$$

- **Geometric case**

  Finally, we apply the same process to the geometric case. The parameters of the power series for the geometric distribution are

$$a_x = 1$$
$$g(\theta) = \frac{1}{1-\theta}.$$

  We can obtain an estimate of $p_0$ once we have an estimate of $g\prime^{-1}(\lambda)$:

$$g\prime(\theta) = \frac{1}{(1-\theta)^2} = \lambda$$
$$g\prime^{-1}(\lambda) = 1 - \frac{1}{\sqrt{\lambda}}$$

so that

$$\hat{p}_0 = \frac{1}{\frac{1}{1-\left(1-\frac{1}{\sqrt{\frac{S}{f_1}}}\right)}} = \sqrt{\frac{f_1}{S}}.$$

Therefore, the estimate for the population size is

$$\hat{N}_{Turing-Geo} = \frac{n}{1-\sqrt{\frac{f_1}{S}}}. \tag{4.29}$$

### 4.4.2 Results

2000 samples have been simulated from the same population and their capture-recapture distribution are simulated. The capture-recapture distribution is generated from a binomial distribution $Y_i \sim Bin(p_i, m)$. Two scenarios with the same characteristics are proposed differing only in the number of captured occasions $m = 10$ and $m = 20$ for true population sizes 500,1000,2000. $p_i$ is based on two covariates

$$logit(p_i) = -0.05X_1 + 0.035X_2 \ ,$$

where $i$ represents the index for a generic individual in the population, $X_1$ follows a normal distribution with mean 40 and variance 144 ($X_1 \sim N(40, 144)$) and $X_2 \sim N(8, 64)$. $X_1$ and $X_2$ are independent.

Estimates are calculated fitting a model using only $X_1$ as independent variable to study the impact of having unexplained heterogeneity (Tables 4.4 and 4.5, and figures 4.1, 4.2, 4.3 and 4.4).

Firstly we observe for scenarios with the same number of capture occasions and true population size, that an increase in the number of non-truncated counts leads to a larger biased estimates but a reduction in the standard deviation. We see again the trade-off between variance and precision. The maximum likelihood estimator and Turing's estimator are largely biased because they assume homogeneity and their variances are the smallest, as expected.

The estimates improve considerably when the number of captured occasions are doubled. The ratio between standard deviations having 10 or 20 capture times is equal or greater than 2 across all population sizes and estimators.

The relative mean squared error (RMSE) and the relative bias (RBias) are also calculated to study the asymptotic behaviour with respect to the population size (Table 4.5 and figure 4.4). When the true population size increases, the RMSE of the generalised Chao estimator decreases across all scenarios, Turing's RMSE also decreases slightly and the RMSE of the maximum likelihood estimate (MLE) remains constant. The relative bias of the generalised Chao estimators increases with the increase of the true population size, in contrast to the relative bias of Turing and MLE that do not present any trend. The RMSE and the relative bias are negatively correlated to the number of capture occasions.

The optimal number of non-truncated counts varies depending on the scenario and our criteria, whether we consider the RMSE or the relative bias. We also have to notice that the scale of those measures is reported multiplied by 100, so differences are very small in this example. The boxplots presented in figures 4.1, 4.2, 4.3 can be also helpful to decide about the best estimator for our objective.

In this chapter, we also developed an analytical formula for the variance of the generalised Chao estimator with $J$ counts (section 4.3). We observe that the analytical standard deviation is very close to the empirical standard deviation for $m = 10$, but the difference increases for $m = 20$ (table 4.4).

Table 4.4: Population and SD estimates from a model assuming $Y \sim Bin(m, p_i)$ with $logit(p_i) = -0.05X_1$ with the true model based on $logit(p_i) = -0.05X_1 + 0.035X_2$. $X_1 \sim N(40, 144)$ and $X_2 \sim N(8, 64)$, independently.

| $N$ | $m$ | NT counts | $\hat{N}_{GC}$ | $\hat{N}_{Turing}$ | $\hat{N}_{MLE}$ | Emp $SD_{GC}$ | Analytical $SD_{GC}$ | Emp $SD_{Turing}$ | Emp $SD_{MLE}$ |
|---|---|---|---|---|---|---|---|---|---|
| 500 | 10 | 2 | 496.115 | 450.187 | 436.929 | 28.146 | 28.799 | 12.994 | 11.968 |
| | | 3 | 490.873 | | | 21.171 | 21.887 | | |
| | | 4 | 487.754 | | | 19.525 | 19.400 | | |
| | | 5 | 485.187 | | | 19.026 | 18.296 | | |
| | | 6 | 485.714 | | | 18.387 | 17.972 | | |
| 500 | 20 | 2 | 500.012 | 478.403 | 468.460 | 14.073 | 16.226 | 6.967 | 6.478 |
| | | 3 | 496.618 | | | 9.997 | 13.040 | | |
| | | 4 | 494.577 | | | 9.150 | 11.453 | | |
| | | 5 | 493.761 | | | 8.401 | 10.248 | | |
| | | 6 | 491.968 | | | 8.221 | 9.250 | | |
| 1000 | 10 | 2 | 987.272 | 899.542 | 872.574 | 38.768 | 39.436 | 18.602 | 17.094 |
| | | 3 | 979.416 | | | 30.764 | 30.469 | | |
| | | 4 | 973.508 | | | 27.310 | 27.113 | | |
| | | 5 | 969.374 | | | 27.080 | 25.676 | | |
| | | 6 | 968.814 | | | 25.620 | 25.198 | | |
| 1000 | 20 | 2 | 996.090 | 956.220 | 936.298 | 17.748 | 22.024 | 9.766 | 9.181 |
| | | 3 | 992.499 | | | 13.633 | 18.280 | | |
| | | 4 | 988.849 | | | 12.417 | 16.129 | | |
| | | 5 | 985.912 | | | 12.145 | 14.394 | | |
| | | 6 | 983.942 | | | 11.965 | 13.010 | | |
| 2000 | 10 | 2 | 1972.951 | 1800.926 | 1746.601 | 53.984 | 55.131 | 26.762 | 24.457 |
| | | 3 | 1955.646 | | | 41.542 | 42.851 | | |
| | | 4 | 1944.895 | | | 38.680 | 38.210 | | |
| | | 5 | 1940.762 | | | 38.160 | 36.329 | | |
| | | 6 | 1936.480 | | | 36.927 | 35.586 | | |
| 2000 | 20 | 2 | 1991.078 | 1912.380 | 1872.078 | 25.695 | 30.800 | 13.542 | 12.616 |
| | | 3 | 1982.500 | | | 19.994 | 25.647 | | |
| | | 4 | 1976.044 | | | 17.996 | 22.665 | | |
| | | 5 | 1971.069 | | | 17.352 | 20.269 | | |
| | | 6 | 1966.666 | | | 16.745 | 18.340 | | |

Table 4.5: RMSE (x100) and relative bias (x100) values from a model assuming $Y \sim Bin(m, p_i)$ with $logit(p_i) = -0.05X_1$ with the true model based on $logit(p_i) = -0.05X_1 + 0.035X_2$. $X_1 \sim N(40, 144)$ and $X_2 \sim N(8, 64)$, independently.

| $N$ | $m$ | Non truncated counts | $RMSE_{GC}(x100)$ | $RMSE_{Turing}(x100)$ | $RMSE_{MLE}(x100)$ | $RBias_{GC}(x100)$ | $RBias_{Turing}(x100)$ | $Rbias_{MLE}(x100)$ |
|---|---|---|---|---|---|---|---|---|
| 500 | 10 | 2 | 0.3228 | 1.0600 | 1.6485 | -0.777 | -9.963 | -12.614 |
| | | 3 | 0.2125 | | | -1.825 | | |
| | | 4 | 0.2124 | | | -2.449 | | |
| | | 5 | 0.2325 | | | -2.963 | | |
| | | 6 | 0.2168 | | | -2.857 | | |
| 500 | 20 | 2 | 0.0792 | 0.2060 | 0.4147 | 0.002 | -4.319 | -6.308 |
| | | 3 | 0.0445 | | | -0.677 | | |
| | | 4 | 0.0452 | | | -1.085 | | |
| | | 5 | 0.0438 | | | -1.248 | | |
| | | 6 | 0.0528 | | | -1.606 | | |
| 1000 | 10 | 2 | 0.1664 | 1.0438 | 1.6529 | -1.273 | -10.046 | -12.743 |
| | | 3 | 0.1370 | | | -2.058 | | |
| | | 4 | 0.1447 | | | -2.649 | | |
| | | 5 | 0.1671 | | | -3.063 | | |
| | | 6 | 0.1629 | | | -3.119 | | |
| 1000 | 20 | 2 | 0.033 | 0.2012 | 0.4142 | -0.391 | -4.378 | -6.370 |
| | | 3 | 0.0242 | | | -0.750 | | |
| | | 4 | 0.0278 | | | -1.115 | | |
| | | 5 | 0.0346 | | | -1.409 | | |
| | | 6 | 0.0401 | | | -1.606 | | |
| 2000 | 10 | 2 | 0.0911 | 1.0087 | 1.6202 | -1.352 | -9.954 | -12.670 |
| | | 3 | 0.0923 | | | -2.218 | | |
| | | 4 | 0.1133 | | | -2.755 | | |
| | | 5 | 0.1241 | | | -2.962 | | |
| | | 6 | 0.1349 | | | -3.176 | | |
| 2000 | 20 | 2 | 0.0185 | 0.1965 | 0.4131 | -0.446 | -4.381 | -6.396 |
| | | 3 | 0.0176 | | | -0.875 | | |
| | | 4 | 0.0224 | | | -1.198 | | |
| | | 5 | 0.0284 | | | -1.447 | | |
| | | 6 | 0.0348 | | | -1.667 | | |

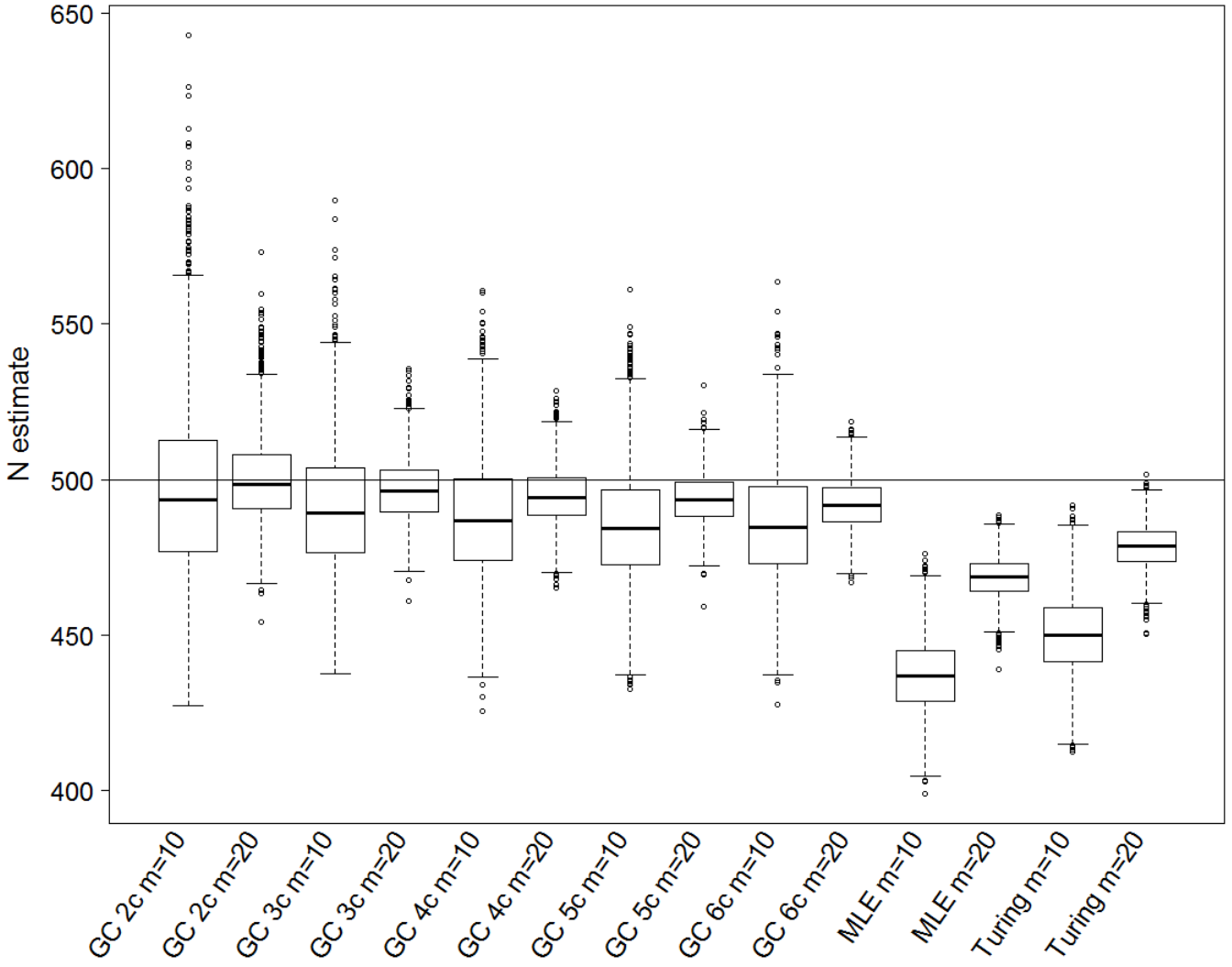


Figure 4.1: Boxplot for $\hat{N}$ with $N = 500$. Data generated by a model with $p_i = -0.05X_1 + 0.035X_2$ with independent $X_1 \sim N(40, 12)$ and $X_2 \sim N(8, 8)$ and model fitting using only $X_1$.

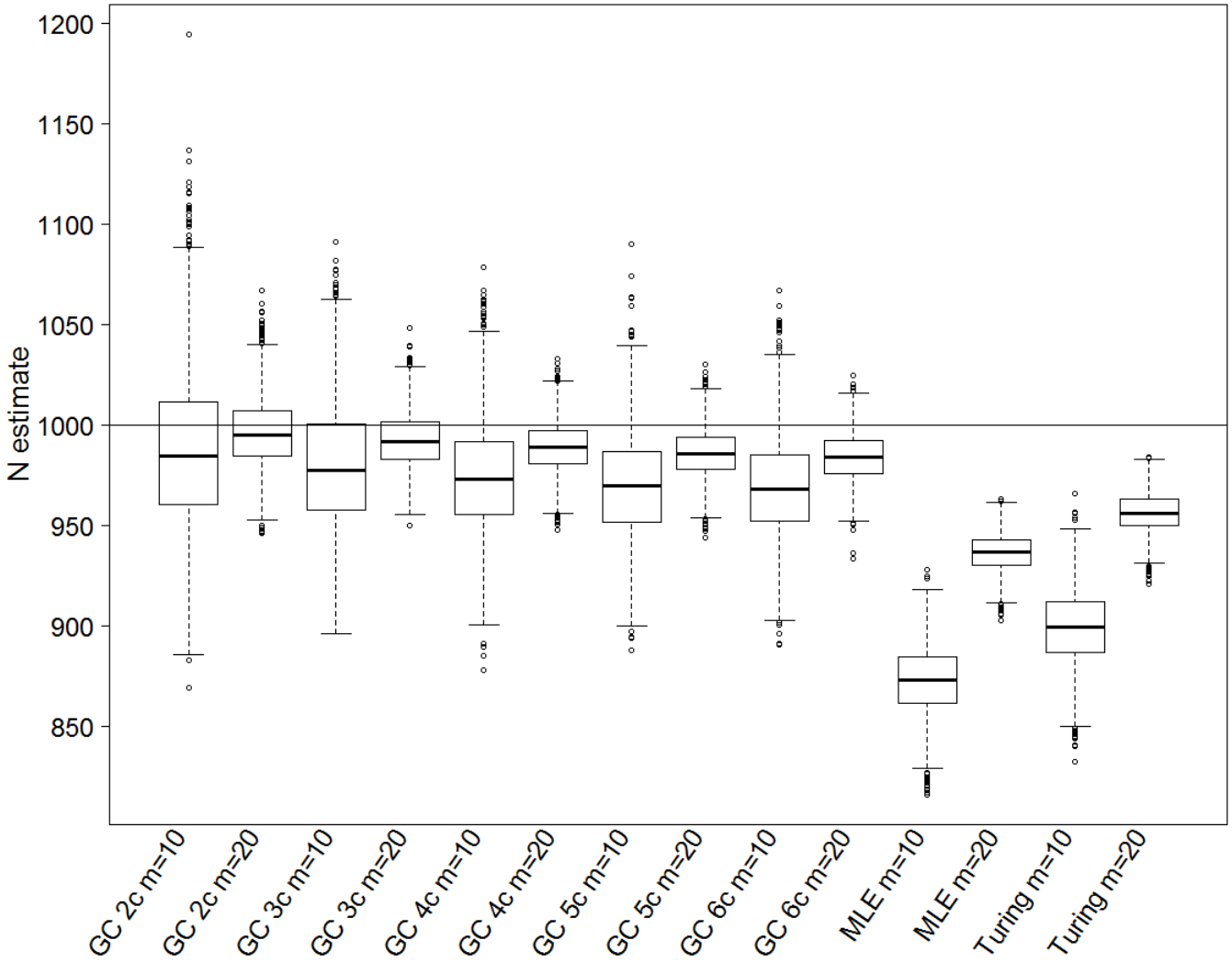


Figure 4.2: Boxplot for $\hat{N}$ with $N = 1000$. Data generated by a model with $p_i = -0.05X_1 + 0.035X_2$ with independent $X_1 \sim N(40, 144)$ and $X_2 \sim N(8, 64)$ and model fitting using only $X_1$.

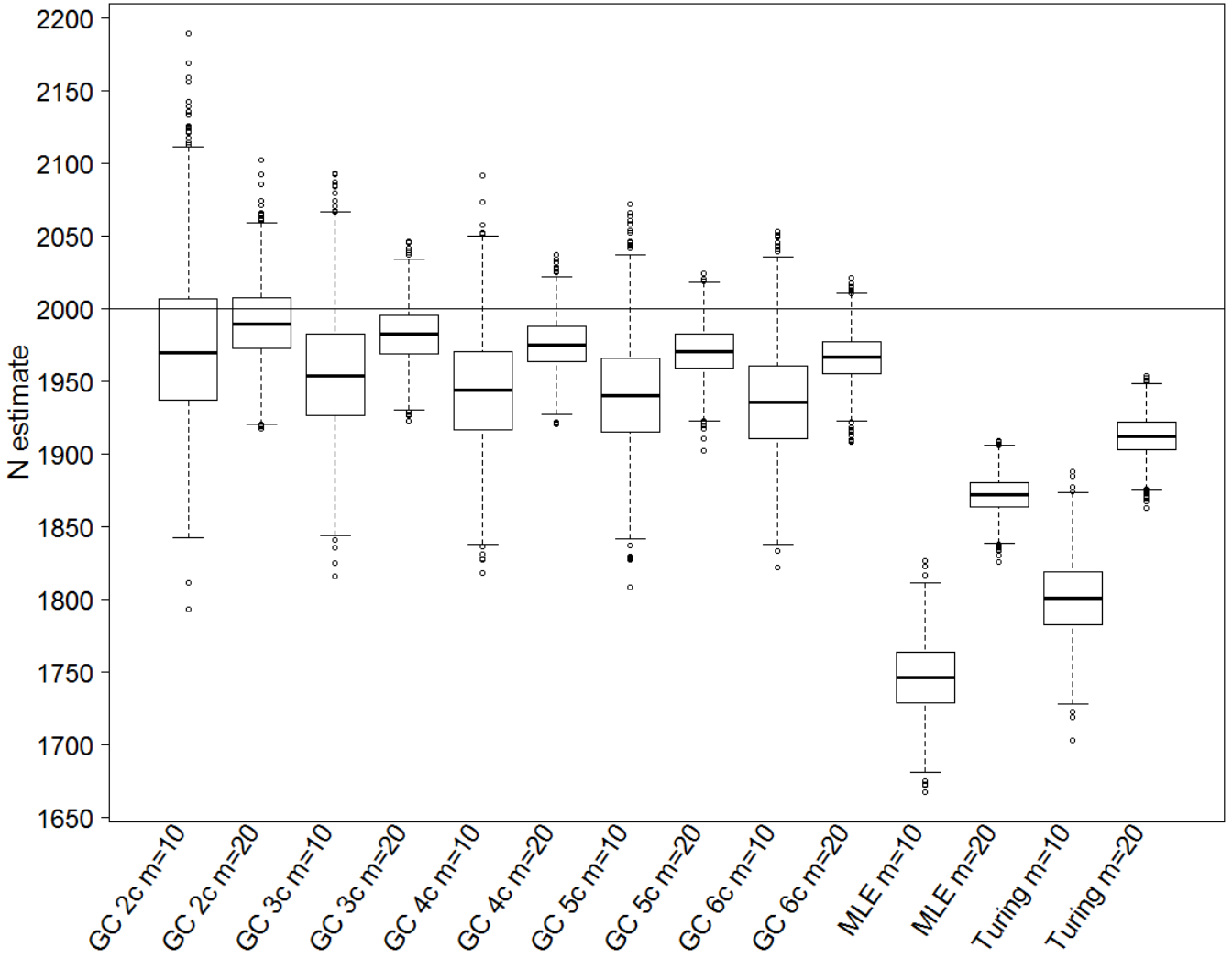


Figure 4.3: Boxplot for $\hat{N}$ with $N = 2000$. Data generated by a model with $p_i = -0.05X_1 + 0.035X_2$ with independent $X_1 \sim N(40, 144)$ and $X_2 \sim N(8, 64)$ and model fitting using only $X_1$.

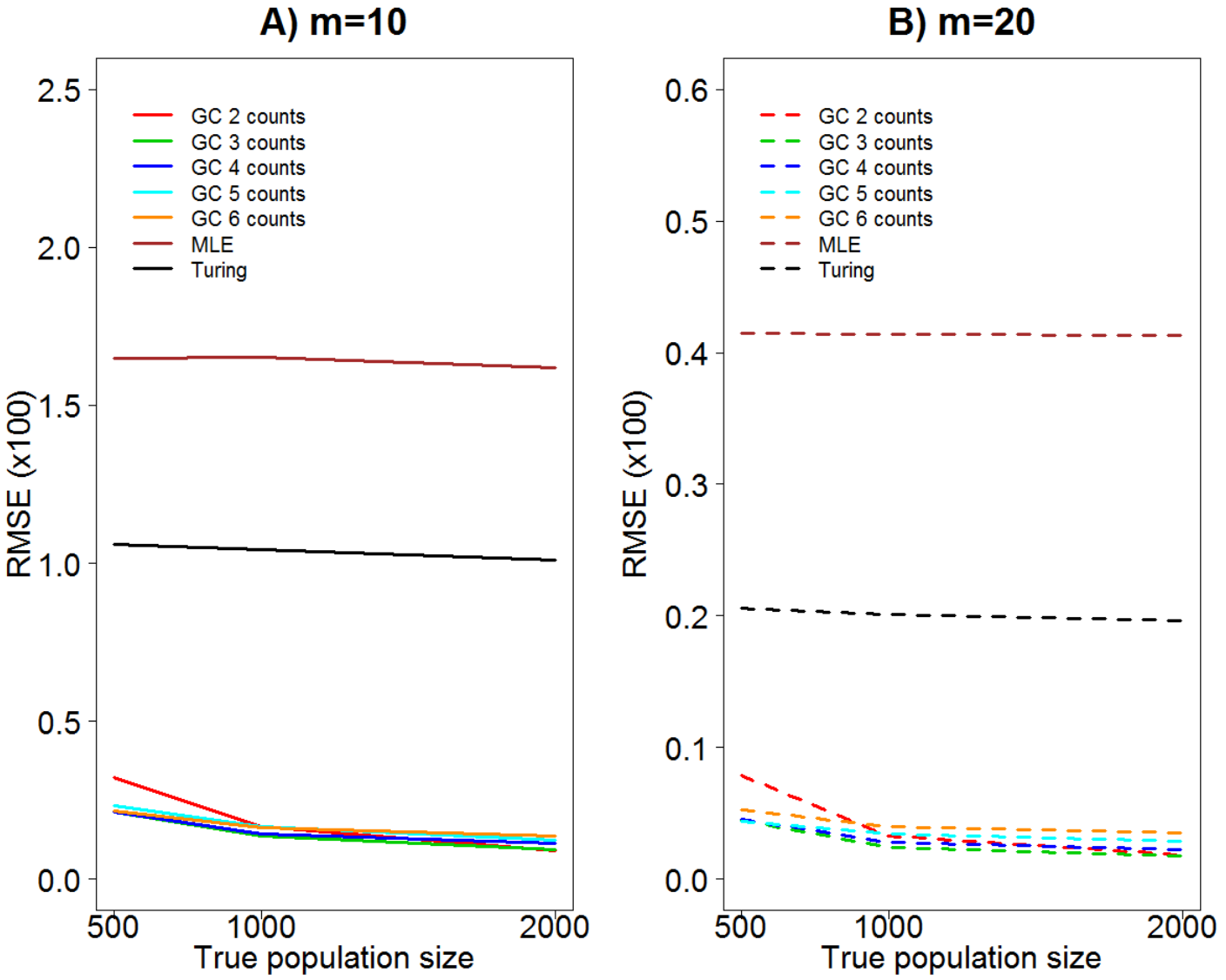


Figure 4.4: Relative mean squared error (RMSE) for binomial cases with the number of occasions A) $m = 10$ and B) $m = 20$

## 4.5 Case study

An example of binomial data with auxiliary variables is presented in Amstrup et al. (2005) (chapters 2 and 4). The data comes from a live-trapping experiment where deer mice were captured during 6 consecutive nights. The weight, age (young or adult) and gender for the captured mice were recorded. Table 4.6 contains all data for the 38 mice observed.

We initially look at the ratio plot to assess the presence of heterogeneity (Figure 4.5). The $y$-axis of the ratio plot is $\hat{r}_x = \frac{(x+1)f_{x+1}}{f_x}$ as an estimation of $r_x = \frac{(x+1)p(Y=x+1)}{p(Y=x)}$ and the $x$-axis is the number of captures ($x = 1, \ldots, m-1$). The confidence limits of the ratio plot are calculated as $exp(log(\hat{r}_x) \pm \sqrt{1/f_{x+1} + 1/f_x})$. We observe in figure 4.5 a potential structural heterogeneity.

Our generalised Chao's estimator will be compared to the weighted linear regression model estimator (WLR)(Rocchetti et al., 2011). This estimator can be applied to the *Katz* family of distributions (Johnson et al., 2005) which include binomial, Poisson and negative binomial distributions. The idea behind the WLR estimator is to fit a linear regression model between the count $x$ and the ratio described above $\hat{r}_x$ or the $\log(\hat{r}_x)$. The WLR estimator does not use any individual covariate information; moreover behavioural and time effects maybe present. Hence, $f_0$ can be predicted based on the model:

$$\log\left(\frac{(x+1)f_{x+1}}{f_x}\right) = \alpha + \beta x + \epsilon_x, \tag{4.30}$$

where $\epsilon_x$ is a random error, $x = 1, \ldots, m-1$. $f_0$ can be estimated as $\hat{f}_0 = f_1 \mathrm{e}^{-\hat{\alpha}}$. The authors suggest the use of weighted least squares with the matrix of weights chosen inversely to the covariance matrix $cov(Y)$. A practical approximation is provided as a diagonal matrix with values

$$w_x = \left(\frac{1}{f_x} + \frac{1}{f_{x+1}}\right)^{-1}. \tag{4.31}$$

A variance estimator for $\hat{N}$ is also deduced and defined as

$$Var(\hat{N}) \approx n\frac{\hat{f}_0}{\hat{N}} + \mathrm{e}^{-2\hat{\alpha}} f_1 \left(Var(\hat{\alpha}) f_1 + 1\right). \tag{4.32}$$

The results are shown in tables 4.7 and 4.8. The generalised Chao's estimator with 2 non-truncated counts does not seem adequate because of the larger standard error than the models with more non-truncated counts and the estimate of $f_0$ as shown in figure 4.6. The point estimators with larger number of non-truncated counts present similar results. The standard errors increase significantly when the variable *weight* is added into the model. The variability decreases when the number of non-truncated counts increases as

Table 4.6: Individual capture history with 3 covariates: sex (0:female,1:male), age (0:adult,1:young) and weight(in grams), $m = 6$ trapping occasions.

| ID | # captures | Sex | Age | Weight |
|---|---|---|---|---|
| 1 | 6 | 1 | 1 | 12 |
| 2 | 5 | 0 | 1 | 15 |
| 3 | 4 | 1 | 1 | 15 |
| 4 | 5 | 1 | 1 | 15 |
| 5 | 6 | 1 | 1 | 13 |
| 6 | 5 | 1 | 0 | 21 |
| 7 | 5 | 1 | 1 | 11 |
| 8 | 4 | 1 | 0 | 15 |
| 9 | 6 | 1 | 1 | 14 |
| 10 | 5 | 1 | 1 | 13 |
| 11 | 5 | 1 | 1 | 14 |
| 12 | 5 | 0 | 0 | 22 |
| 13 | 6 | 1 | 1 | 14 |
| 14 | 4 | 1 | 1 | 11 |
| 15 | 2 | 0 | 1 | 10 |
| 16 | 2 | 0 | 0 | 23 |
| 17 | 3 | 0 | 1 | 7 |
| 18 | 2 | 1 | 1 | 8 |
| 19 | 3 | 1 | 0 | 19 |
| 20 | 3 | 1 | 1 | 13 |
| 21 | 3 | 0 | 1 | 5 |
| 22 | 2 | 0 | 0 | 20 |
| 23 | 3 | 1 | 1 | 12 |
| 24 | 1 | 0 | 1 | 6 |
| 25 | 4 | 0 | 0 | 22 |
| 26 | 3 | 0 | 1 | 10 |
| 27 | 4 | 0 | 1 | 14 |
| 28 | 2 | 0 | 0 | 19 |
| 29 | 1 | 0 | 0 | 19 |
| 30 | 1 | 0 | 0 | 20 |
| 31 | 3 | 1 | 0 | 16 |
| 32 | 2 | 0 | 1 | 11 |
| 33 | 1 | 1 | 1 | 14 |
| 34 | 1 | 0 | 1 | 11 |
| 35 | 1 | 1 | 0 | 24 |
| 36 | 1 | 1 | 0 | 9 |
| 37 | 1 | 1 | 1 | 16 |
| 38 | 1 | 0 | 0 | 19 |

observed in the simulations. The weighted linear regression estimator provides a higher point estimate but a larger confidence interval compared to our estimator.

The expected values of the WLR estimator are close to the observed values with the exception of the mice captured 5 times (figure 4.6). $\hat{r}_4$, the ratio between mice captured

5 and 4 times (figure 4.5) seems to be an outlier or unusual value and the weights are reducing the impact of that point in the regression. The expected values for the generalised Chao's estimator with 3 and 4 non truncated counts seem reasonable with respect to the scale of the graph. However the estimators with 5 and 6 non-truncated counts are clearly negatively affected by the counts in the tail.

In this small dataset with covariates, the covariates did not show a large impact in the results compared to estimates without covariates. The p-values from the likelihood ratio tests found only significant variables for models with 5 and 6 non-truncated counts (table 4.8). Amstrup et al. (2005) calculated a similar estimate (39.9 (SE - 1.7)) based on Huggings $M_h$ model with covariates. The expected values of the WLR estimator were better than the ones of the generalised Chao estimator, however the variability is larger despite not using any auxiliary variables. The possibility of changing the cut-off point of truncation to decide the balance between precision and variability is an advantage of the generalised Chao's estimator.

Table 4.7: Point estimates and 95% asymptotic confidence intervals for the deer mice case study ($m = 6$) with 3 covariates: sex (0:female,1:male), age (0:adult,1:young) and weight(in grams).

| Counts | Model | $\hat{N}$ | $\hat{SE}$ | Asymptotic 95% CI |
|---|---|---|---|---|
| 2 | Sex | 70 | 2.77 | 64-75 |
| | Sex+Age | 69 | 5.52 | 58-79 |
| | Sex+Age+Weight | 51 | 15.33 | 21-81 |
| 3 | Sex | 42 | 0.73 | 41-44 |
| | Sex+Age | 43 | 1.22 | 41-45 |
| | Sex+Age+Weight | 42 | 3.26 | 36-49 |
| 4 | Sex | 41 | 0.49 | 40-42 |
| | Sex+Age | 41 | 0.77 | 39-42 |
| | Sex+Age+Weight | 41 | 2.21 | 36-45 |
| 5 | Sex | 39 | 0.32 | 39-39 |
| | Sex+Age | 39 | 0.5 | 38-40 |
| | Sex+Age+Weight | 40 | 1.76 | 36-43 |
| 6 | Sex | 39 | 0.26 | 39-39 |
| | Sex+Age | 39 | 0.47 | 38-40 |
| | Sex+Age+Weight | 40 | 1.59 | 36-43 |
| 6 | WLR estimate | 44 | 4.26 | 36-52 |

Table 4.8: Likelihood ratio tests for models estimating the number of deer mice.

| Counts | Model | $\chi^2$ | df | P-value |
|---|---|---|---|---|
| 2 | Sex | 1.323 | 1 | 0.2500 |
| | Sex+Age | 0.204 | 1 | 0.6517 |
| | Sex+Age+Weight | 0.058 | 1 | 0.8095 |
| 3 | Sex | 0.214 | 1 | 0.6438 |
| | Sex+Age | 1.367 | 1 | 0.2424 |
| | Sex+Age+Weight | 0.741 | 1 | 0.3894 |
| 4 | Sex | 0.958 | 1 | 0.3277 |
| | Sex+Age | 0.974 | 1 | 0.3237 |
| | Sex+Age+Weight | 0.014 | 1 | 0.9045 |
| 5 | Sex | 4.006 | 1 | 0.045 |
| | Sex+Age | 1.803 | 1 | 0.1794 |
| | Sex+Age+Weight | 3.524 | 1 | 0.061 |
| 6 | Sex | 11.413 | 1 | 0.0007 |
| | Sex+Age | 4.943 | 1 | 0.0262 |
| | Sex+Age+Weight | 5.115 | 1 | 0.0237 |

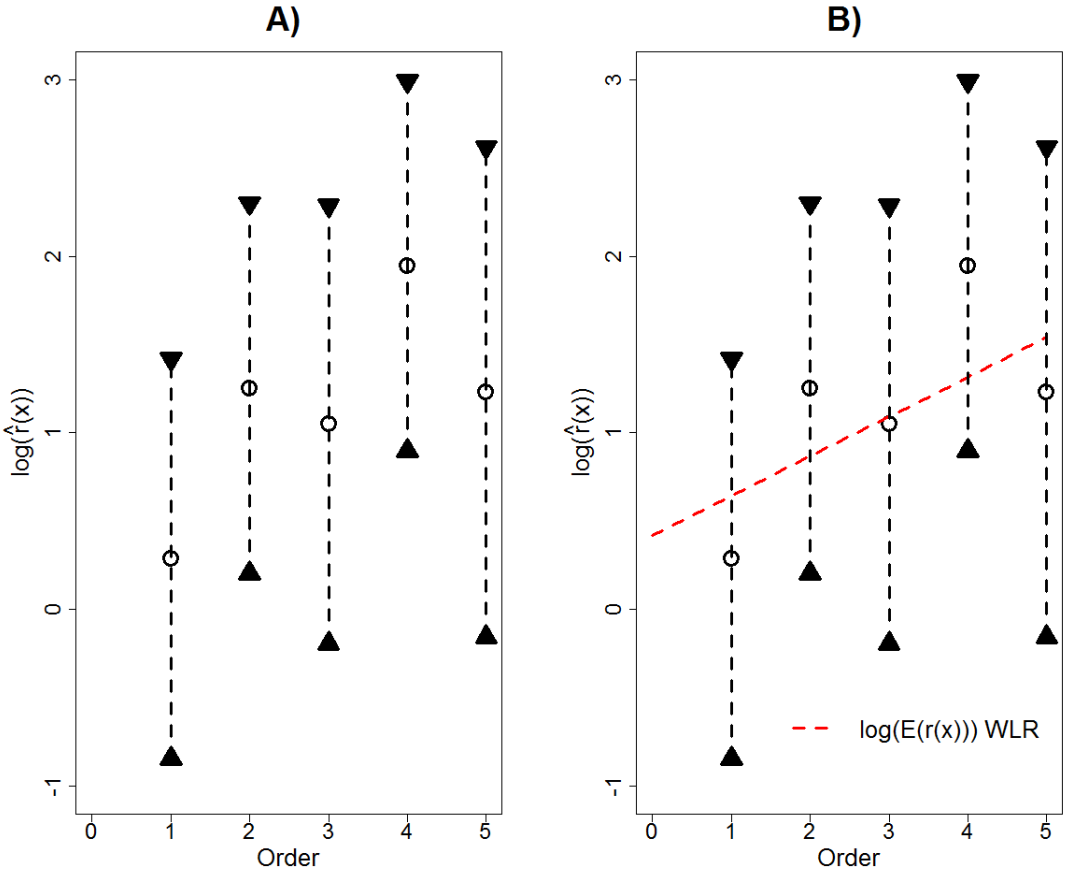


Figure 4.5: A) Ratio plot for the deer mice example.$log(\hat{r}_x) = log\left(\frac{(x+1)f_{x+1}}{f_x}\right)$. B) Fitted ratio values $log(E(r(x)))$ for the WRL estimator

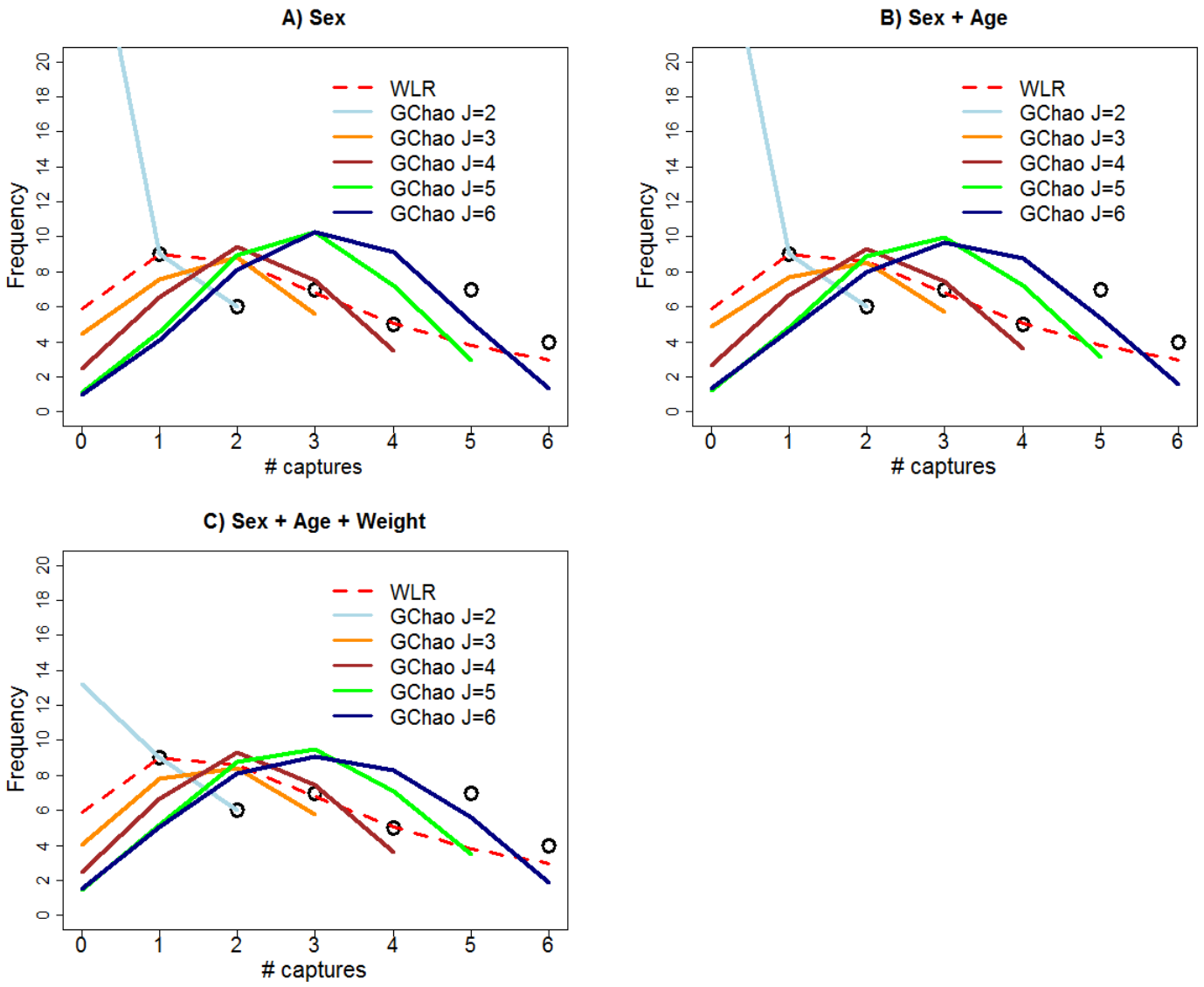


Figure 4.6: Observed vs fitted frequencies for the deer mice experiment. Models with covariates: A) Sex. B) Sex and Age. C) Sex, Age and Weight

## 4.6 Conclusions

In this chapter, we have extended the generalisation of Chao's estimator for the Poisson distribution to any capture-recapture distributions belonging to the power series. We have focused on the Poisson, binomial and geometric distributions because they commonly appear in the capture-recapture area. The Poisson case has been used to validate the general estimators of the power series distributions. Simulations and a case study were only presented for the binomial distribution because an estimator based on the truncated geometric distribution will be part of the comparison with other estimators based on the geometric distribution in a later chapter.

The deduction of the estimators started from the simplest case with 2 non-truncated counts and no covariate information to the most complex case with $J$ non-truncated counts and the addition of auxiliary variables to explain the heterogeneity of the capture-recapture distribution. We also detailed a generalised Turing estimator for the power series and the maximum likelihood estimator for the binomial case to be compared with our estimators.

Our findings for the estimators based on the binomial distribution were similar to the results obtained in previous chapters. The estimates are asymptotically biased and there is a shrinkage effect in the variance when the population size increases. The error cannot be reduced because it decomposes into the sum of the bias squared and the variance and the reduction in bias leads to an increase in the variability and vice-versa.

In the binomial case we have also shown the impact of increasing the number of captured occasions. A duplication in the number of occasions led to a reduction by half the standard deviations and a significant reduction of the RMSE.

The estimators behaviour with respect to the number of non-truncated counts depends on the specific case. However, we observed in the simulations that estimators with small number of non-truncated counts are more efficient based on the relative mean squared error. We obtained a similar conclusion for the Poisson case. In the case study we also concluded that some truncation cut-off values were not adequate based on the fitted values. The question of determining an optimal cut-off point arises and a solution is proposed in the following chapter.

# Chapter 5

# Selecting the "right" cut-off estimate

In this chapter, we describe methods to decide about the best cut-off truncation point for the generalised Chao estimator with and without covariate information. We have seen in previous chapters how an increase in the number of non-truncated counts leads to an increase in the bias but a reduction in the variability, because more information is considered. The ratio plot, also described previously, is a useful tool to assess visually the presence and type of heterogeneity, but it can also be used to find an optimal truncation cut-off point (Böhning et al., 2013a). A formal $\chi^2$ test is provided initially for the Poisson case without covariates and its extension to the power series distributions using covariate information.

Another approach to obtain an optimal estimator based on the methodology of model averaging is also presented. Simulations and case studies were conducted to compare to other estimators.

## 5.1 Goodness of fit

### 5.1.1 Poisson case without covariates

We aim to develop a formal test to find the optimal upper truncation point $J$ to obtain the best estimate. We firstly look at the Poisson case without covariates.

A $\chi^2$ test is defined by

$$\chi^2(J|\hat{\lambda}_J) = \sum_{y=1}^{J} \frac{\left[f_y - E(f_y|J,\hat{\lambda})\right]^2}{E(f_y|J,\hat{\lambda})} = \sum_{y=1}^{J} \frac{\left[f_y - \hat{f}_y\right]^2}{\hat{f}_y}. \tag{5.1}$$

We can replace $\hat{f}_y = E(f_y|J,\hat{\lambda})$ by the estimates shown in (3.6) to obtain

$$\chi^2(J|\hat{\lambda}_J) = \sum_{y=1}^{J} \left[ \frac{\left[ f_y - \left( \sum_{j=1}^{J} f_j \right) \frac{\hat{\lambda}_J^y / y!}{\sum_{k=1}^{J} \hat{\lambda}_J^k / k!} \right]^2}{\left( \sum_{j=1}^{J} f_j \right) \frac{\hat{\lambda}_J^y / y!}{\sum_{k=1}^{J} \hat{\lambda}_J^k / k!}} \right]. \tag{5.2}$$

where $J \geq 2$. Notice that $\hat{\lambda}_J$ is the estimate of $\lambda$ when the first $J$ counts are the only non-truncated counts in the distribution.

Note that $\chi^2(J,\hat{\lambda_J}) = 0$ for $J = 2$ because the perfect fit is achieved. Under the null hypothesis that the model is valid, the statistic follows a $\chi^2$ distribution with $J-2$ degrees of freedom where $J$ is the number of non-truncated counts ($\chi^2(J,\hat{\lambda_J}) \approx \chi^2_{J-2}$). This asymptotic result needs the constrain $f_x \geq 5$. Therefore, the optimal $J$ is defined as the largest truncation point where the test is not significant. We have calculated the kernel density of the $\chi^2$ statistic to confirm the number of degrees of freedom of the $\chi^2$ distribution (figure 5.1). We have carried out a simple simulation exercise where the capture-recapture distribution follows an homogeneous Poisson $Y \sim Po(2)$. Each truncation point is compared to theoretical $\chi^2$ distributions.

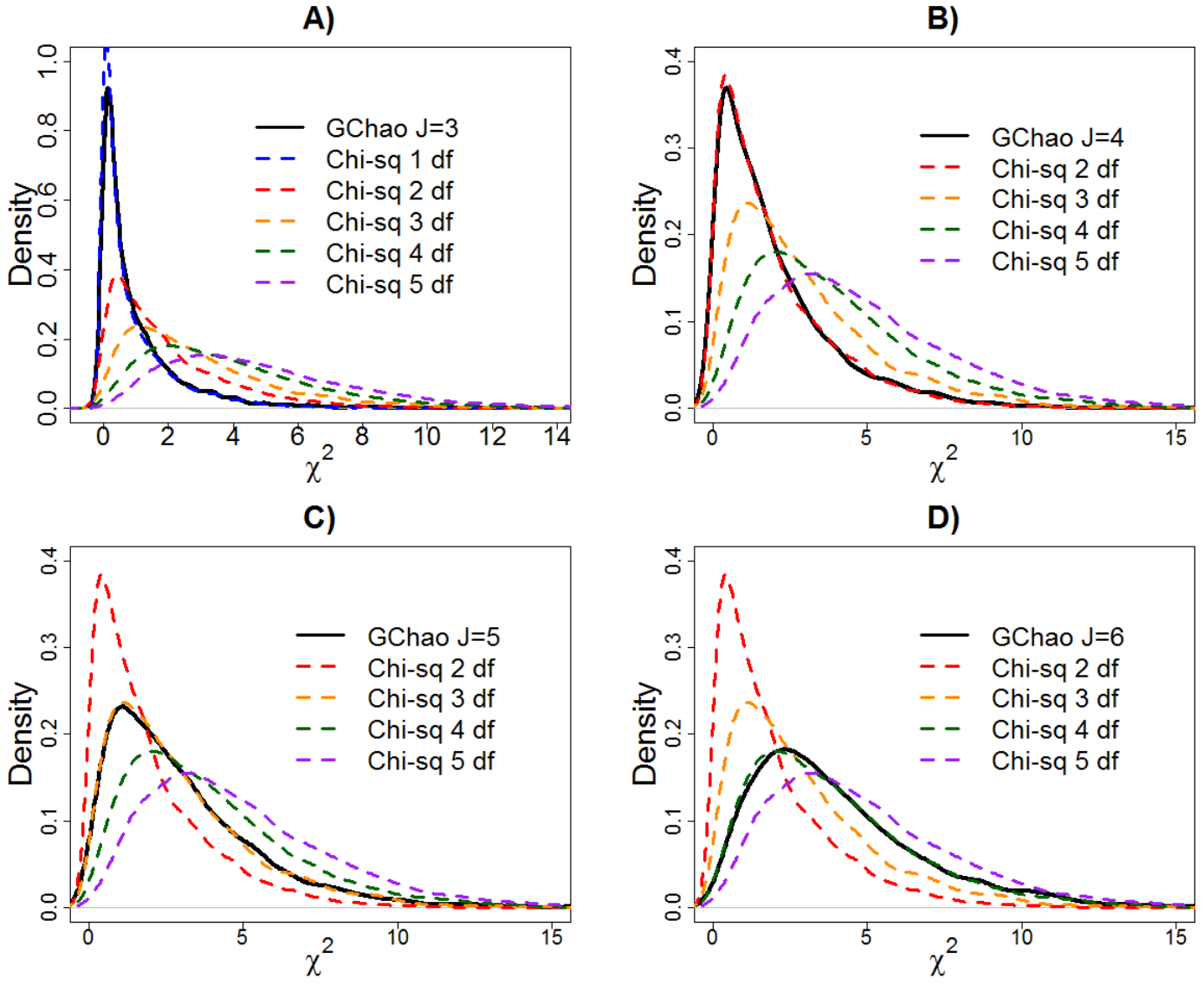


Figure 5.1: Comparison of theoretical $\chi^2$ distributions with the density of the $\chi^2$ statistic for a simulation from a Poisson distribution without covariates. A)$J = 3$ B)$J = 4$ C)$J = 5$ D)$J = 6$

### 5.1.2 Power series distributions without covariates

The test can be extended to the power series distributions applying the estimates for the expectations of the frequencies obtained in (4.6). Hence, the $\chi^2$ statistic is defined by

$$\chi^2(J|\hat{\theta}_J) = \sum_{y=1}^{J} \left[ \frac{\left[ f_y - \left( \sum_{j=1}^{J} f_j \right) \frac{a_y \hat{\theta_J}^y}{\sum_{k=1}^{J} a_k \hat{\theta_J}^k} \right]^2}{\left( \sum_{j=1}^{J} f_j \right) \frac{a_y \hat{\theta_J}^y}{\sum_{k=1}^{J} a_k \hat{\theta_J}^k}} \right]. \tag{5.3}$$

The characteristics of the $\chi^2(J|\hat{\theta}_J)$ statistic are the same as described above. It follows a $\chi^2$ distribution with $J-2$ degrees of freedom ($\chi^2(J|\hat{\theta}_J) \sim \chi^2_{J-2}$). We have conducted another small simulation exercise to confirm the $J-2$ degrees of freedom. 2000 replications are generated from a population with a capture-recapture distribution $Y \sim Bin(0.25, 10)$. The kernel density of the $\chi^2(J|\hat{p})$ is compared graphically to theoretical $\chi^2$ distributions (figure 5.2). We also estimate the degrees of freedom of the distribution based on data using the function *fitdistr* from the R-package MASS.

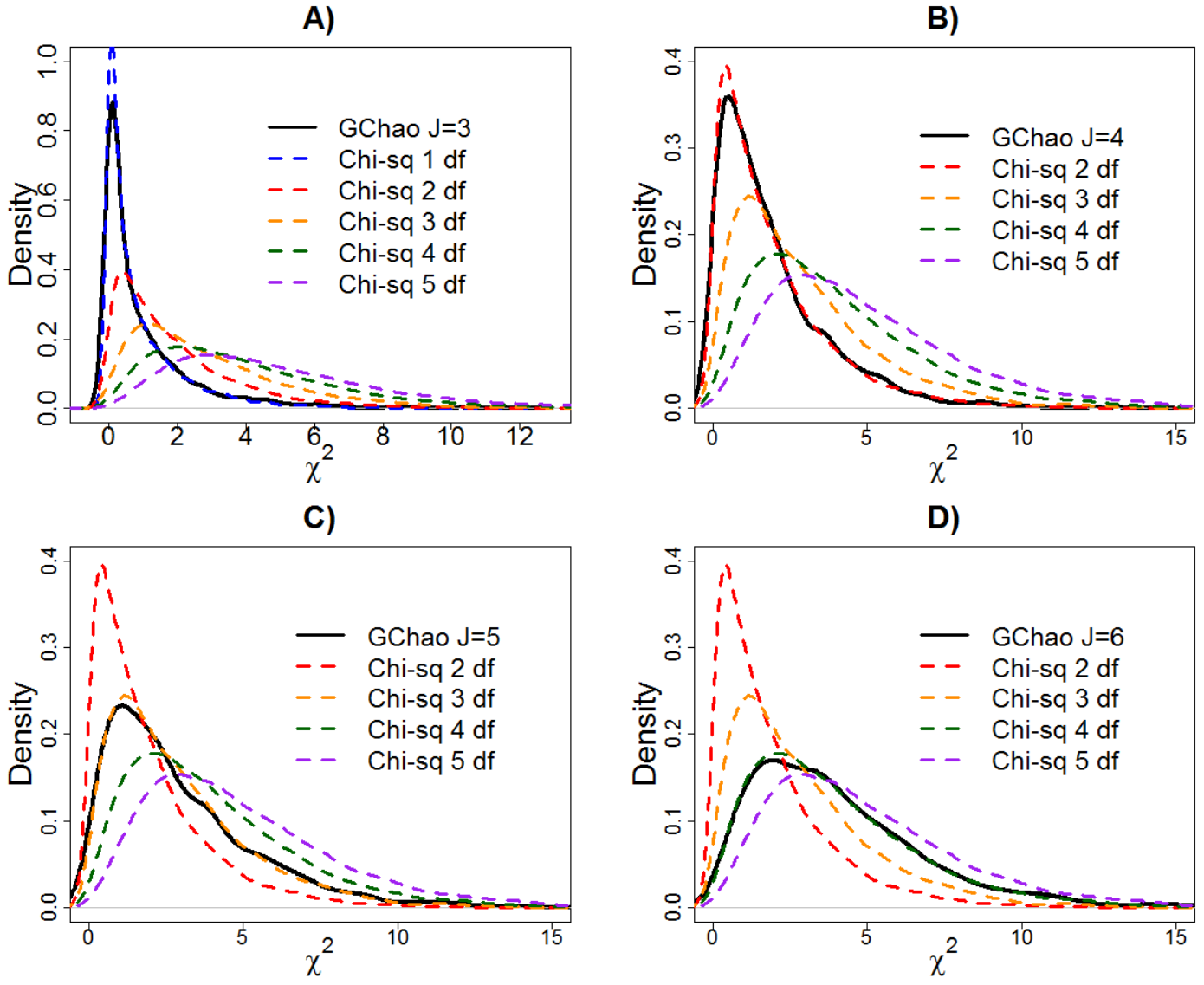


Figure 5.2: Comparison of theoretical $\chi^2$ distributions with the density of the $\chi^2$ statistic for a simulation from a binomial distribution without covariates.A)$J = 3$ B)$J = 4$ C)$J = 5$ D)$J = 6$

### 5.1.3 Power series distributions with covariates

Holling et al. (2013) described a graphical tool to assess the goodness of fit of count distributions when covariate information is available. Several examples with heterogeneity are provided in the paper where the probability $p_y\left(\lambda(\mu_i, \theta_n)\right)$ depends on a vector of known parameters $\mu_i$ and a vector of unknown parameters $\theta_n$.

Our scenario for the Poisson case and auxiliary variables can be described in such a way, where the $\lambda$ function is $\lambda_i = \mathrm{e}^{\alpha+\beta' Z_i}$. $Z_i$ is the known vector and $\alpha$ and $\beta$ formed the unknown part. Holling et al. (2013) applied the same marginal method to calculate the expectation of the frequencies $f_y$ as the sum of the frequencies across all $n$ covariate combinations. Therefore,

$$\hat{f}_y(\hat{\theta}_n) = \sum_{i=1}^{n} p_y\left(\lambda(\mu_i, \hat{\theta}_n)\right) \tag{5.4}$$

which translates in our case and notation into

$$\hat{f}_y = \sum_{i=1}^{M_J} \hat{f}_{iy}, \tag{5.5}$$

where $M_J$ is the number of covariate combinations when $J$ non-truncated counts are considered.

Consequently, we can deduce the $\chi^2$ statistic replacing $\hat{f}_y$ by $E(f_y|\hat{\lambda}_i)$ shown in (3.19).

$$\chi^2(J|\hat{\lambda}_{iJ}) = \sum_{y=1}^{J} \left[ \frac{\left[ f_y - \sum_{i=1}^{M_J} \left( \frac{\left( \sum_{l=1}^{J} f_{il} \right) \hat{\lambda}_{iJ}^y / y!}{\sum_{k=1}^{J} \hat{\lambda}_{iJ}^k / k!} \right) \right]^2}{\sum_{i=1}^{M_J} \left( \frac{\left( \sum_{l=1}^{J} f_{il} \right) \hat{\lambda}_{iJ}^y / y!}{\sum_{k=1}^{J} \hat{\lambda}_{iJ}^k / k!} \right)} \right], \tag{5.6}$$

(5.7)

for $y = 1, \ldots, J$. The index $j$ indicates the assumption that the model used $j$ non-truncated counts. $\hat{\lambda}_{iJ}$ is the $\lambda$ estimate for the $i$th covariate combination with $J$ non-truncated counts. We can also replace $\lambda_{iJ} = \mathrm{e}^{\alpha_J+\beta_J' Z_i}$ with $\alpha_J$ and $\beta_J$ are the estimations

of $\alpha$ and $\beta$ for the model with $J$ non-truncated counts. Hence we obtain

$$\chi^2(J|\hat{\alpha}_J,\hat{\beta}'_J)=\sum_{y=1}^{J}\left[\frac{\left[f_y-\sum_{i=1}^{M_J}\left(\frac{\left(\sum_{l=1}^{J}f_{il}\right)\mathrm{e}^{y(\hat{\alpha}_J+\hat{\beta}'_J z_i)}/y!}{\sum_{k=1}^{J}\mathrm{e}^{k(\hat{\alpha}_J+\hat{\beta}'_J Z_i)}/k!}\right)\right]^2}{\sum_{i=1}^{M_J}\left(\frac{\left(\sum_{l=1}^{J}f_{il}\right)\mathrm{e}^{y(\hat{\alpha}_J+\hat{\beta}'_J Z_i)}/J!}{\sum_{k=1}^{J}\mathrm{e}^{k(\hat{\alpha}_J+\hat{\beta}'_J Z_i)}/k!}\right)}\right]. \tag{5.8}$$

We can generalise this formal test to the power series distributions. In chapter 4, we calculated $e_y$ (4.19) and the $h$ function was defined to link the parameter $\theta$ and the covariate information, $h(\theta_i)=\alpha+\beta' Z_i$. Therefore the test statistic can be written as:

$$\chi^2(J|\hat{\theta}_{iJ})=\sum_{y=1}^{J}\left[\frac{\left[f_y-\sum_{i=1}^{M_J}\left(\frac{a_y\hat{\theta}_{iJ}^{y}\left(\sum_{l=1}^{J}f_{il}\right)}{\sum_{k=1}^{J}a_k\hat{\theta}_{iJ}^{k}}\right)\right]^2}{\sum_{i=1}^{M_J}\left(\frac{a_y\hat{\theta}_{iJ}^{y}\sum_{l=1}^{J}f_{il}}{\sum_{k=1}^{J}a_k\hat{\theta}_{iJ}^{k}}\right)}\right]. \tag{5.9}$$

$$\chi^2(J|\hat{\alpha}_J,\hat{\beta}'_J)=\sum_{y=1}^{J}\left[\frac{\left[f_y-\sum_{i=1}^{M_J}\left(\frac{a_y(h^{-1}(\hat{\alpha}_J,\hat{\beta}'_J))^{y}\sum_{l=1}^{J}f_{il}}{\sum_{k=1}^{J}a_k(h^{-1}(\hat{\alpha}_J,\hat{\beta}'_J))^{k}}\right)\right]^2}{\sum_{i=1}^{M_J}\left(\frac{a_y(h^{-1}(\hat{\alpha}_J,\hat{\beta}'_J))^{y}\sum_{l=1}^{J}f_{il}}{\sum_{k=1}^{J}a_k(h^{-1}(\hat{\alpha}_J,\hat{\beta}'_J))^{k}}\right)}\right]. \tag{5.10}$$

This $\chi^2$ statistic follows a $\chi^2$ distribution but the number of degrees of freedom need to be calculated. Two simulation studies have been conducted to calculate the number of degrees of freedom. 2000 replications from a population were generated following the example presented in (4.4.2). This time we first draw data based only on $X_1$ that

follows a normal distribution with mean 40 and variance 144 ($X_1 \sim N(40, 144)$) with $logit(p_i) = -0.05X_1$ as link function. The capture-recapture distribution follows $Y \sim Bin(p_i, 10)$. The second simulation adds $X_2 \sim N(8, 64)$ to the model as $logit(p_i) = -0.05X_1 + 0.035X_2$. The idea is to evaluate whether an increase in the number of covariates affects the number of degrees of freedom.

The results from the simulation concluded that the number of degrees of freedom only depends on the number of non-truncated counts and is independent of the number of covariates included in the model (Figure 5.3 and 5.4). Therefore, the $\chi^2(J|\hat{\theta}_{iJ})$ statistic follows a $\chi^2$ distribution with $J - 2$ degrees of freedom based on simulated data. The number of covariates in the model does not impact the degrees of freedom, the reason could be that we are taking the average over the various strata defined by the covariates.

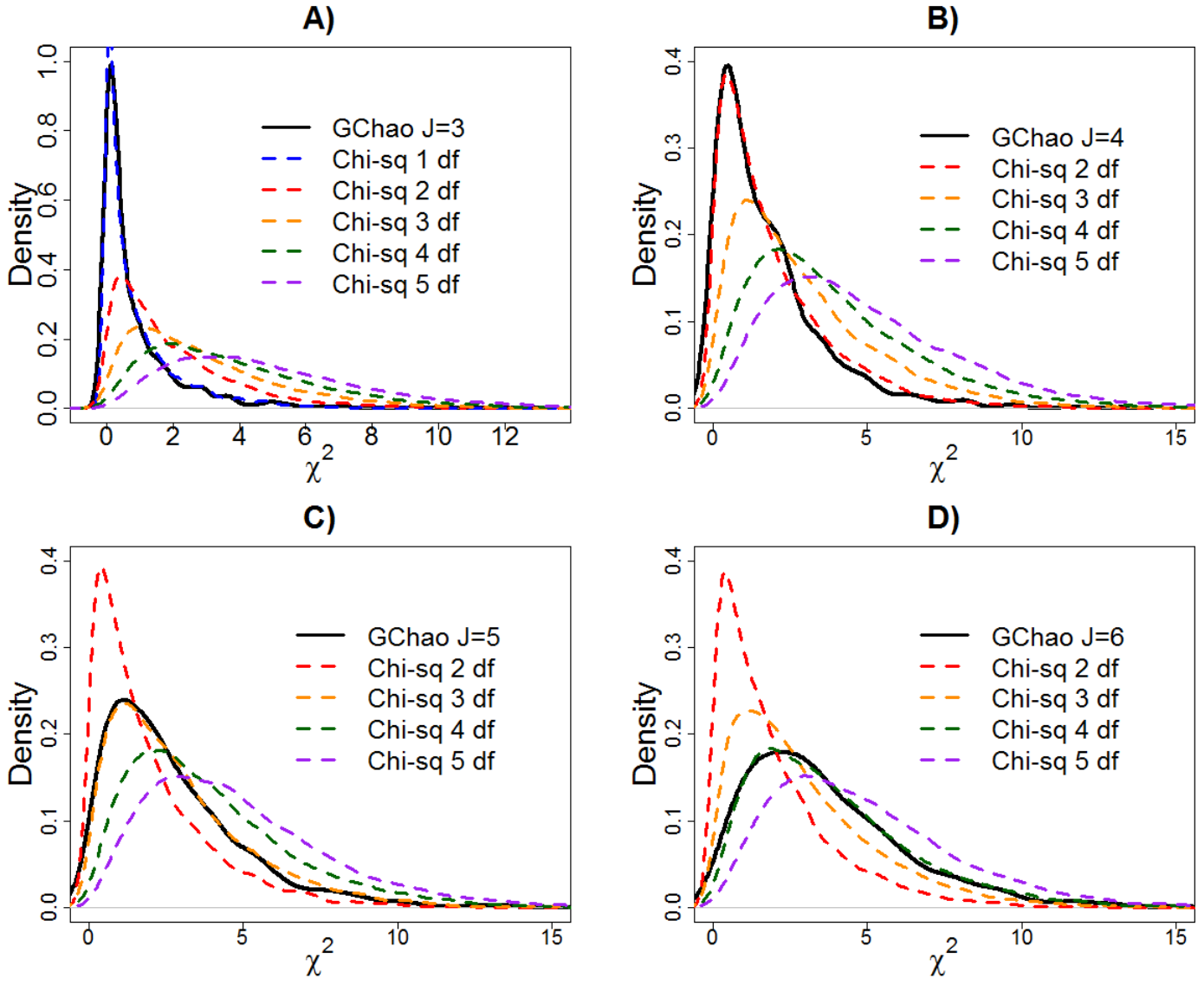


Figure 5.3: Comparison of theoretical $\chi^2$ distributions with the kernel density of the $\chi^2$ statistic from a capture-recapture distribution $Y \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1)$. $J$ indicates the number of non-truncated counts in the model. A)$J = 3$ B)$J = 4$ C)$J = 5$ D)$J = 6$.

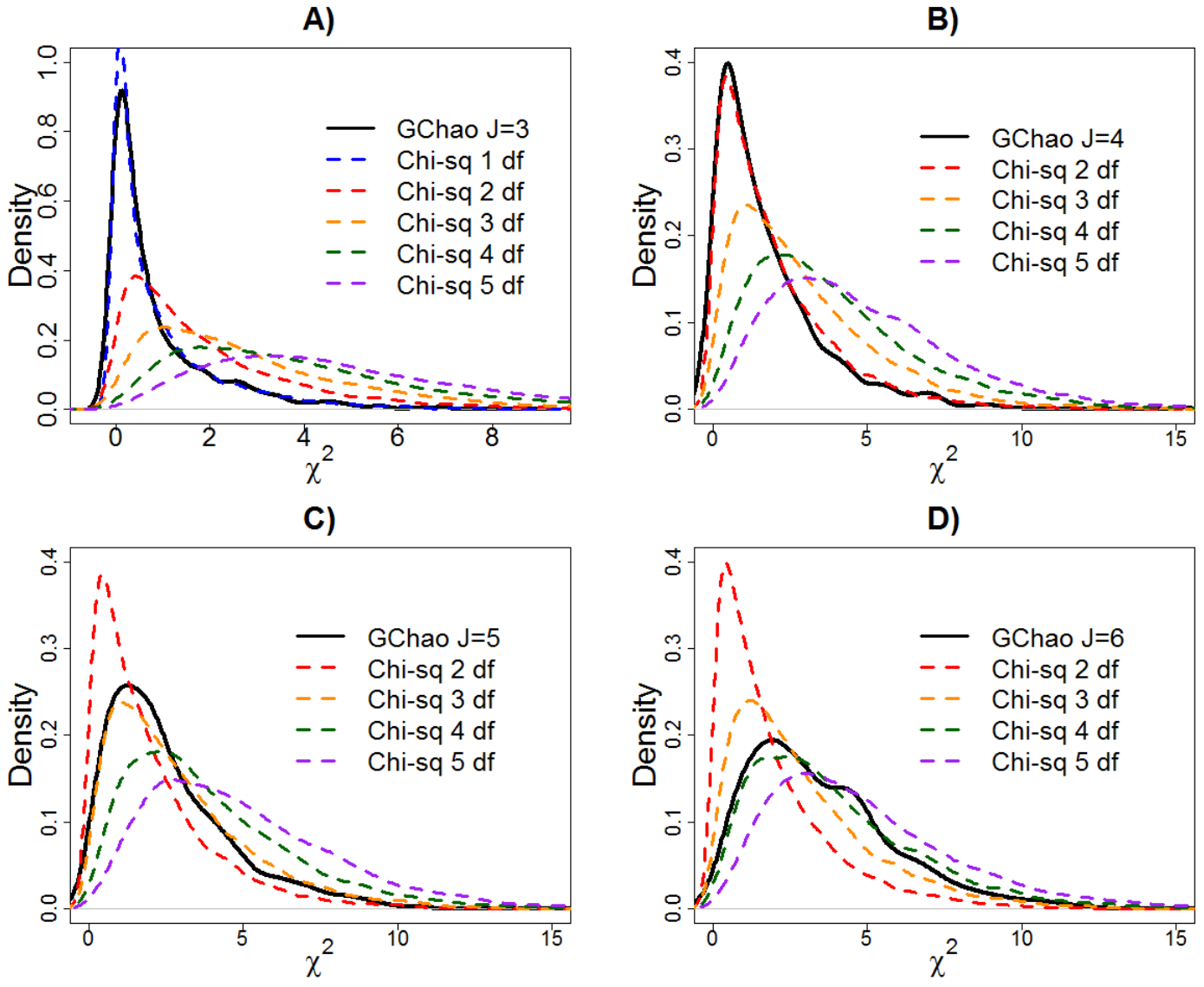


Figure 5.4: Comparison of theoretical $\chi^2$ distributions with the kernel density of the $\chi^2$ statistic from a capture-recapture distribution $Y \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$. $J$ indicates the number of non-truncated counts in the model. A)$J = 3$ B)$J = 4$ C)$J = 5$ D)$J = 6$.

## 5.2 Model averaging

In previous chapters we described the common trade-off between the bias of an estimate and its variance. We observed that when unexplained heterogeneity is present, the closest models to the true value of the population are the models with the smallest number of non-truncated counts $J$. When $J$ increases, the estimates become more biased but their variance decreases.

We have now presented the $\chi^2$ test to find an optimal level of truncation to add to other useful tools like the ratio plot (Rocchetti et al., 2011; Böhning et al., 2013a) and the adjusted frequency plot (Holling et al., 2013). We explore in this section the application of model averaging theory to build an estimate balanced between bias and variance.

Stanley and Burnham (1998) proposed a frequentist approach to calculate model averaging estimates in closed-population capture-recapture framework. Information-theoretic criteria like AIC are used to obtain the weights of each model. However, a key assumption to compare models using AIC or any similar criterion is to have the same initial dataset. In our case the data changes with respect to the level of truncation. To the best of our knowledge we have not found any example in the literature for comparing models with different truncation levels. Therefore, we propose two ad hoc sets of weights and investigate their performance by simulations. The first set is

$$w_1(j) = \frac{f_j}{\sum_{k=1}^{J-1} f_k}, \tag{5.11}$$

with $j = 1, .., J-1$. The second set is

$$w_2(j) = \frac{f_j / \widehat{Var}(\hat{N}_{J=j+1})}{\sum_{k=1}^{J-1} \left( f_k / \widehat{Var}(\hat{N}_{J=k+1}) \right)}, \tag{5.12}$$

with $j = 1, .., J-1$. $\widehat{Var}(\hat{N}_{J=j+1})$ is the variance estimate of the population estimate for the model with $j+1$ non-truncated counts.

Therefore, the simplest case comprises the estimates for the models with two and three non-truncated counts. The weights in this case are:

$$
\begin{aligned}
w_1(1) &= \frac{f_1}{f_1 + f_2} \\
w_1(2) &= \frac{f_2}{f_1 + f_2} \\
w_2(1) &= \frac{f_1/\widehat{Var}(\hat{N}_{J=2})}{f_1/\widehat{Var}(\hat{N}_{J=2}) + f_2/\widehat{Var}(\hat{N}_{J=3})} \\
w_2(2) &= \frac{f_2/\widehat{Var}(\hat{N}_{J=3})}{f_1/\widehat{Var}(\hat{N}_{J=2}) + f_2/\widehat{Var}(\hat{N}_{J=3})}.
\end{aligned}
$$

The weights $w_1(j)$ are designed to take into account the frequency distribution. In contrast, the alternative weights $w_2(j)$ includes the inverse of the variance to reduce the weights of models with larger uncertainty. Hence, the population estimates can be written as

$$
N_{MA1} = \sum_{j=1}^{J-1} w_1(j)\hat{N}_{J=j+1} \tag{5.13}
$$

$$
N_{MA2} = \sum_{j=1}^{J-1} w_2(j)\hat{N}_{J=j+1}, \tag{5.14}
$$

where $\hat{N}_{J=j+1}$ is the generalised Chao estimator for the model with $j+1$ non-truncated counts.

The efficiency of these new estimators is investigated going back to the simulation based on a binomial distribution with unexplained heterogeneity due to a missing covariate in the model fitting (4.4.2).

Table 5.1 and 5.2 contain the estimates, standard errors, relative mean squared errors and relative bias for the generalised Chao's estimator and the model averaged estimators respectively. Figures 5.5 and 5.6 clearly show how the estimates based on model averaging are a balance between the estimates involved in their calculation. We observe that the model with two non-truncated counts tend to be more unstable presenting a good point estimate but a large variance. The weighted estimates are ideal to obtain a balanced solution with smaller variance and better point estimates. The weighted estimators using $w_1$ weights favour the point estimation. In contrast to the other weighted estimator (5.14) which penalises large variance leading to an estimate with smaller variance but greater bias. Both weighted estimators show smaller RMSE than the original estimators (5.7). The $w_1$ weights provide a slightly better estimate with respect to the RMSE but both weighted estimates are similar in these particular scenarios (table 5.2).

Table 5.1: Generalised Chao estimates based on a capture-recapture distribution $Y_i \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$ where the model fitting included only $X_1$; $N = 1000$.

| Non-truncated counts | $\hat{N}$ | $\hat{SE}_N$ | RMSE x 100 | RBias |
|---|---|---|---|---|
| 2 | 988.12 | 38.89 | 0.158 | -0.011 |
| 3 | 978.36 | 30.21 | 0.130 | -0.021 |
| 4 | 972.14 | 27.63 | 0.145 | -0.027 |
| 5 | 969.28 | 26.55 | 0.157 | -0.030 |
| 6 | 968.55 | 26.30 | 0.160 | -0.031 |

Table 5.2: Model averaging estimates based on a capture-recapture distribution $Y_i \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$ where the model fitting included only $X_1$. MA1 and MA2 are the average models with $w_1$ and $w_2$ weights respectively; $N = 1000$. The standard errors presented are empirical.

| Estimates involved | $\hat{N}_{MA_1}$ | $\hat{N}_{MA_2}$ | $\hat{SE}_{MA_1}$ | $\hat{SE}_{MA_2}$ | $RMSE_{MA1}$ ($x100$) | $RMSE_{MA2}$ ($x100$) | $RBias_{MA1}$ | $RBias_{MA2}$ |
|---|---|---|---|---|---|---|---|---|
| 2-3 | 984.18 | 981.88 | 33.76 | 32.20 | 0.130 | 0.127 | -0.015 | -0.018 |
| 2-4 | 981.61 | 978.45 | 31.81 | 29.94 | 0.126 | 0.127 | -0.018 | -0.021 |
| 2-5 | 980.38 | 976.88 | 31.00 | 29.08 | 0.126 | 0.129 | -0.019 | -0.022 |
| 2-6 | 979.88 | 976.26 | 30.68 | 28.76 | 0.126 | 0.130 | -0.019 | -0.023 |

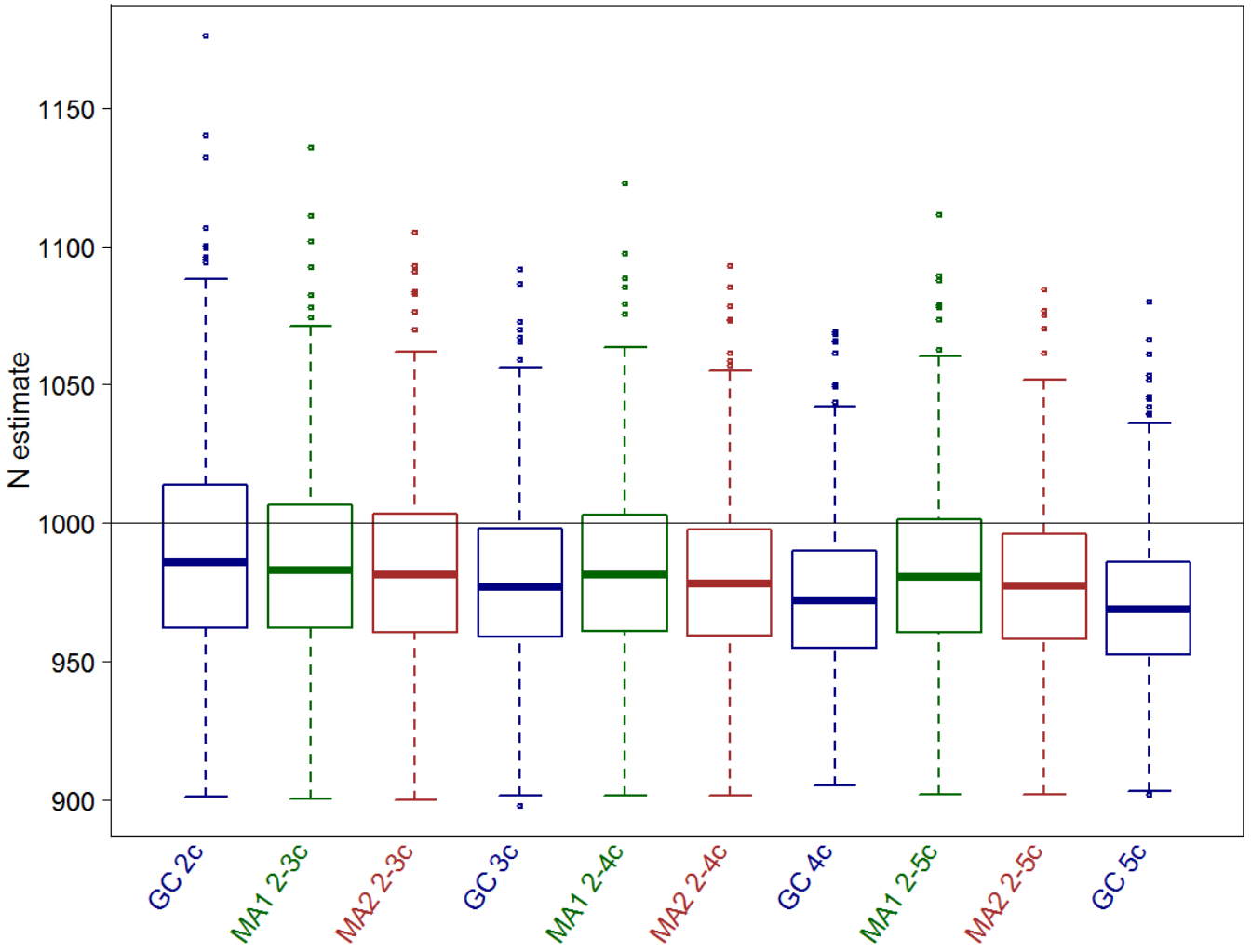


Figure 5.5: Comparison of generalised Chao and model averaging estimates based on a capture-recapture distribution $Y_i \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$ where the model fitting included only $X_1$. MA1 and MA2 are the average models with $w_1$ and $w_2$ weights respectively.

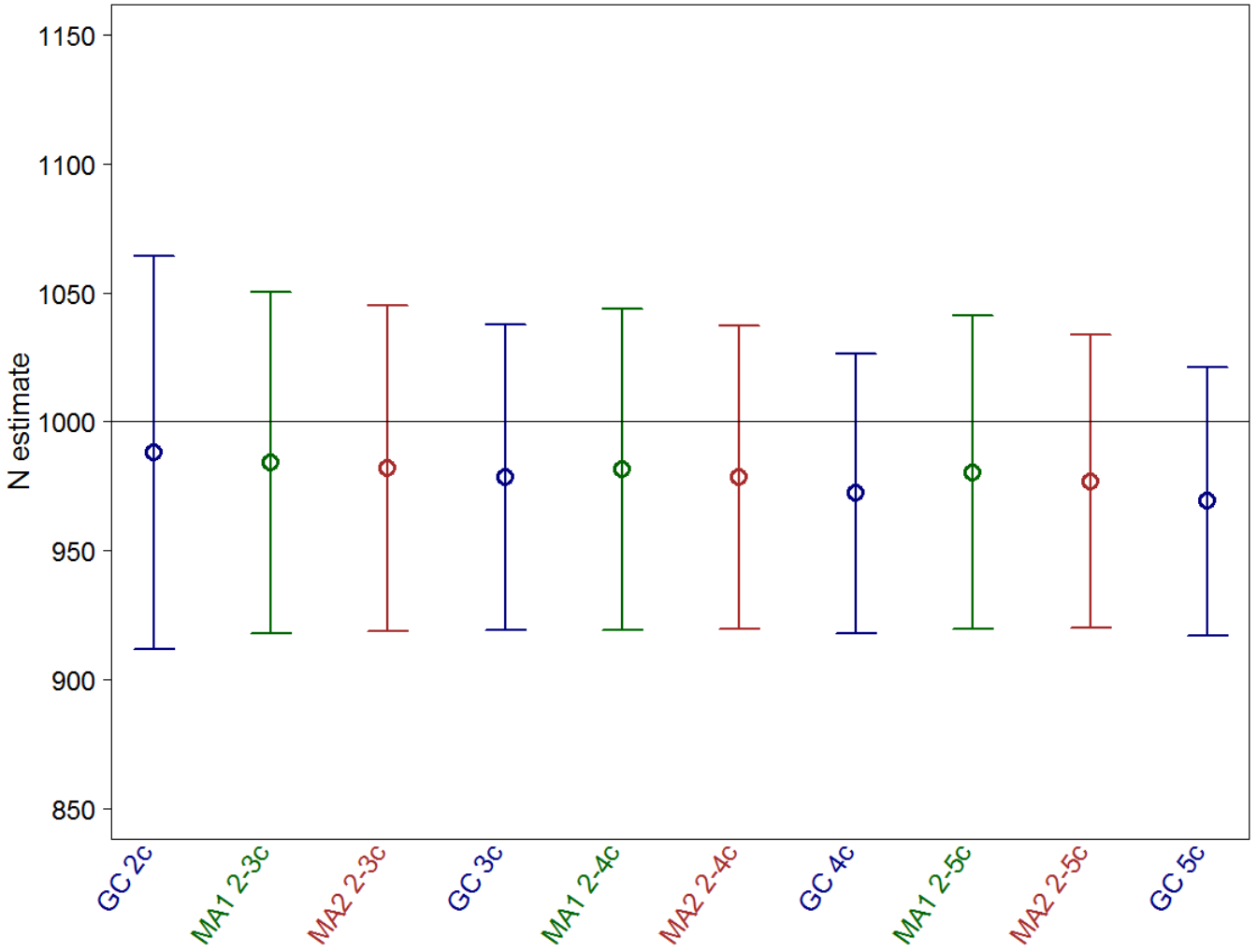


Figure 5.6: Comparison of generalised Chao and model averaging estimates and 95% CI based on a capture-recapture distribution $Y_i \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$ where the model fitting included only $X_1$. MA1 and MA2 are the average models with $w_1$ and $w_2$ weights respectively.

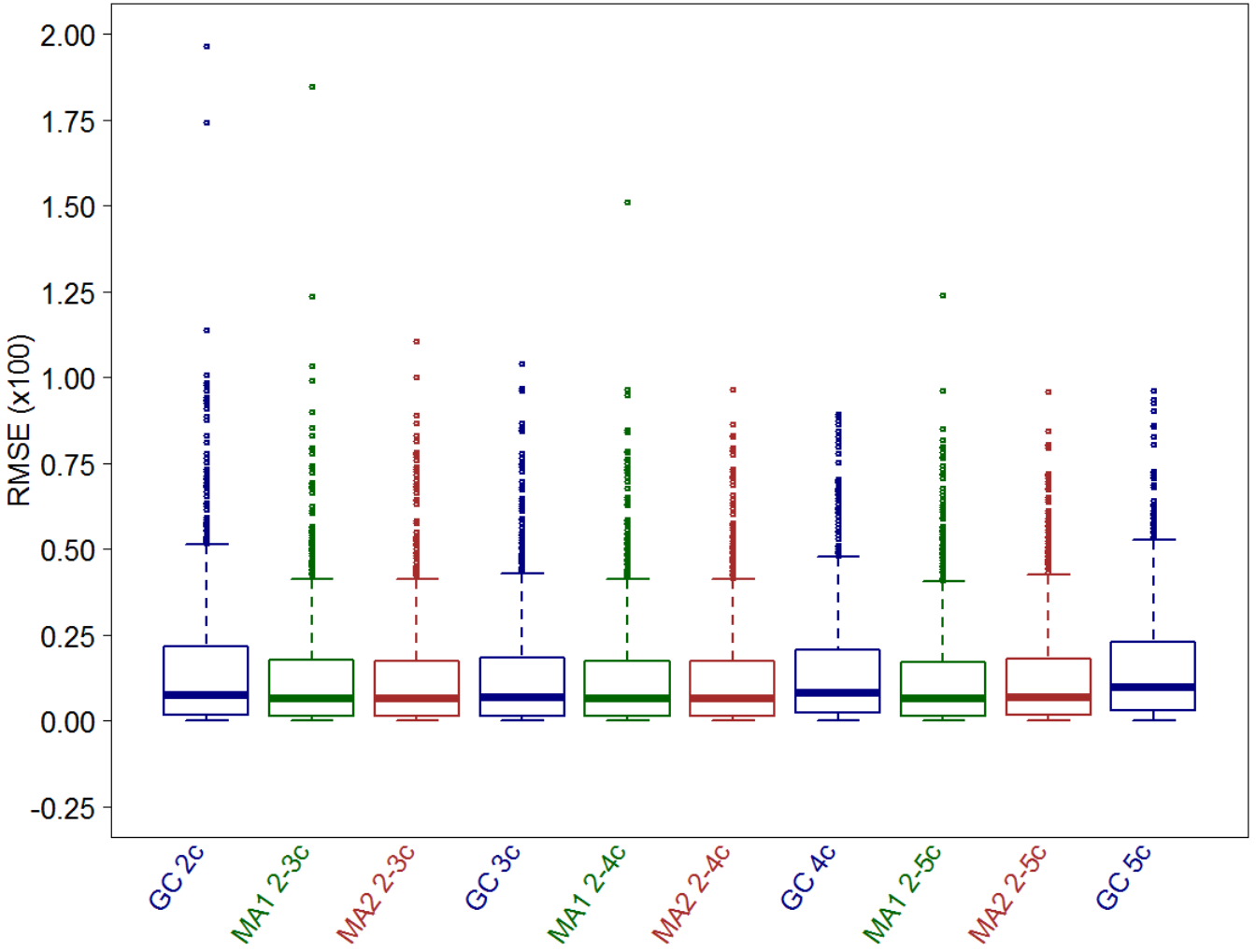


Figure 5.7: RMSE (x100) values for generalised Chao and model averaging estimates based on a capture-recapture distribution $Y_i \sim Bin(p_i, 10)$ with $p_i = expit(-0.05X_1 + 0.035X_2)$ where the model fitting included only $X_1$. MA1 and MA2 are the average models with $w_1$ and $w_2$ weights respectively.

## 5.3 Case study

Section 4.5 described a live-trapping experiment on deer mice. We represented the ratio plot (figure 4.5) and the covariate-adjusted frequency plot (figure 4.6) to choose the best truncation cut-off.

Now we can use the chi-square test and we can also provide average estimates using the weights proposed in the previous section. The chi-square test (table 5.3) suggests to look at the model with $J = 4$. Although in this case the chi-square test for $J = 5$ is borderline significant and for $J = 6$ is borderline non-significant, we could observe more details about these models. The covariate-adjusted frequency plot shows that the expected value of $f_5$ is underestimated under the estimator based on 5 non-truncated counts. On the other hand, the estimator based on all the data included overestimates the observed number of mice captured 4 times. There is no large difference between point estimates of these 3 estimators, but we would recommend to use the estimator that uses 4 non-truncated counts.

We could also report the estimate from the weighted average of the estimators based on 2 to 4 non truncated counts. Both estimates are a good compromise between the variance and the bias. In this example, there were only 36% of the captures in the first 2 counts, an indication that we should be cautious with the model with two non-truncated counts. So the weighted estimates might provide a more reliable solution.

Table 5.3: Generalised Chao's estimates for the deer mice case study: $\chi^2$ statistics and p-values

| Non-truncated counts | $\hat{N}$ | $\hat{SE}$ | $\chi^2(J)$ | df | p-value |
|---|---|---|---|---|---|
| 2 | 51.20 | 15.33 | 0 | | |
| 3 | 42.06 | 3.26 | 1.129 | 1 | 0.288 |
| 4 | 40.48 | 2.213 | 2.534 | 2 | 0.272 |
| 5 | 39.47 | 1.758 | 8.520 | 3 | 0.036 |
| 6 | 39.54 | 1.585 | 8.083 | 4 | 0.089 |

Table 5.4: Model averaging point estimates and standard errors for the deer mice case study.

| Models involved | $\hat{N}_{MA1}$ | $\hat{SE}_{MA1}$ | $\hat{N}_{MA2}$ | $\hat{SE}_{MA2}$ |
|---|---|---|---|---|
| 2-3 counts | 47.54 | 6.90 | 42.64 | 6.53 |
| 2-4 counts | 45.30 | 6.73 | 41.12 | 6.41 |
| 2-5 counts | 44.22 | 6.65 | 40.39 | 6.36 |
| 2-6 counts | 43.25 | 6.58 | 40.02 | 6.33 |

## 5.4 Conclusions

Our initial motivation was to extend Chao's lower bound estimator to incorporate covariate information to model the individual capture probability. Chao's estimator is robust because it only uses individuals observed once or twice. We also extended the initial estimator to include individuals captured more than 2 times. We developed the initial framework under the assumption of a Poisson distribution and we extended it to the power series of distributions.

The results from our simulations showed that models with low number of non-truncated counts obtained the most accurate point estimates but with larger variability. Therefore, the question to solve was to develop a criterion to choose the "right" model based on the data. In this chapter we presented a $\chi^2$ test to find an optimal truncation cut-off point for our estimates with and without covariates. We also provided two weighted estimators to combine models considering varying truncation points to obtain a balanced solution to the trade-off between variance and bias observed in previous chapters. We also found the application of graphical tools like the ratio plot and the adjusted-covariate frequency plot very practical to assess our estimates in more detail.

We finish here exploring the methodology applying truncation and developing in the next chapter new estimates based on the concept of censoring under the assumption of a geometric distribution.

# Chapter 6

# Estimates with censoring and covariates

In this chapter, we explore the concept of censoring applied to capture-recapture estimation based on a geometric distribution. Niwitpong et al. (2012) presented an estimator based on the geometric distribution with different capture probability for individuals captured once and individuals captured more than one time. We extend this estimator to include covariate information and consider other censoring cut-off points. Simulations are also presented assessing the performance of the new estimators and comparing them to the previously developed estimators assuming a left and right truncated geometric distribution. The theory is also applied to a case study at the end of the chapter.

## 6.1 Point estimation for the geometric distribution with censored data

We follow the same structure as in previous chapters. We start extending the estimator described by Niwitpong et al. (2012) but censoring individuals captured 2 or more times. Thereafter, we censor individuals captured more than $c$ times and we finally extend the methodology to include covariate information.

There are different ways to handle censoring in survival analysis. The meaning of censoring individuals captured more than $c$ times in our case is to aggregate all individuals observed $c$ or more times in one category. In contrast, we saw that truncation ignored the individuals captured more than $J$ times for the estimation of $f_0$.

### 6.1.1 Censoring units captured more than 2 times

We initially have a zero-truncated geometric distribution where individuals captured more than 2 times are censored. Three probabilities are considered: the probability of being captured once, being capture twice and being captured more than twice (censored individuals).

$$\begin{aligned}
q_1 &= P(Y=1) = \frac{p(1-p)}{1-p} = p \\
q_2 &= P(Y=2) = \frac{p(1-p)^2}{1-p} = p(1-p) \\
q_3 &= P(Y>2) = 1 - q_1 - q_2 = 1 - p - p(1-p) = (1-p)^2,
\end{aligned}$$

where $p$ is the probability of being captured.
The likelihood has the following form

$$\mathcal{L}(p) = q_1^{f_1} q_2^{f_2} q_3^{n-f_1-f_2}.$$

Hence, the log likelihood is

$$\begin{aligned}
\ell(p) &= (f_1+f_2)\log(p) + f_2\log(1-p) + 2(n-f_1-f_2)\log(1-p) \\
&= (f_1+f_2)\log(p) + (2n-2f_1-f_2)\log(1-p).
\end{aligned}$$

To maximise the log likelihood we solve the score equation $\frac{d\ell(p)}{dp} = 0$, or

$$\frac{d\ell(p)}{dp} = \frac{f_1+f_2}{p} - \frac{(2n-2f_1-f_2)}{1-p} = 0.$$

Hence, the maximum likelihood estimator is

$$\hat{p} = \frac{f_1+f_2}{2n-f_1}. \tag{6.1}$$

### 6.1.2 Censoring units captured $c \geq 2$ times

We look now at the likelihood when we censor all individuals who were captured more than $c$ times. The zero-truncated probabilities are defined as

$$\begin{aligned}
q_k &= P(Y=k) = \frac{p(1-p)^k}{1-p} = p(1-p)^{k-1} \text{ , for } k = 1,..,c \\
q_{c+} &= P(Y>c) = 1 - \sum_{j=1}^{c} q_j = 1 - \sum_{j=0}^{c-1} p(1-p)^j = \sum_{j=c}^{\infty} p(1-p)^j = \sum_{j=0}^{\infty} p(1-p)^{j+c} \\
&= (1-p)^c \sum_{j=0}^{\infty} p(1-p)^j = (1-p)^c.
\end{aligned}$$

Notice the property of the geometric distribution that the zero-truncated geometric is still a geometric distribution.

The likelihood is written as

$$\mathcal{L}(p) = \left(\prod_{j=1}^{c} q_j^{f_j}\right)(1-p)^{c\times(n-\sum_{j=1}^{c} f_j)}.$$

Therefore, the log-likelihood is

$$\ell(p) = \left(\sum_{j=1}^{c} f_j\right)\log(p) + \left(\sum_{j=1}^{c}(j-1)f_j\right)\log(1-p) + c\left(n - \left(\sum_{j=1}^{c} f_j\right)\right)\log(1-p).$$

We solve the score equation to find the maximum likelihood estimate:

$$\frac{d\ell(p)}{dp} = \frac{\left(\sum_{j=1}^{c} f_j\right)}{p} - \frac{\left(\sum_{j=1}^{c}(j-1)*f_j\right)}{1-p} - \frac{c\left(n-\left(\sum_{j=1}^{c} f_j\right)\right)}{1-p} = 0.$$

The maximum likelihood estimate is

$$\hat{p} = \frac{\sum_{j=1}^{c} f_j}{cn + \sum_{j=1}^{c}(j-c)f_j}. \tag{6.2}$$

For the particular case where $c = 2$, we obtain (6.1).

An estimate $\hat{f}_0$ can now be obtained using $\hat{p}$. The expected value $e_0 = E(f_0|\hat{p})$ will be

$$\begin{aligned} e_0 &= Np_0 = (n+e_0)\times\hat{p} \\ e_0 &= \frac{n\hat{p}}{(1-\hat{p})} = \frac{n\sum_{j=1}^{c} f_j}{\left(\sum_{j=1}^{c} f_j(j-(c+1))\right) + nc}. \end{aligned}$$

Consequently, the population estimate is

$$\hat{N}_{G-censor} = n + \frac{n\hat{p}}{(1-\hat{p})} = \frac{n}{1-\hat{p}} = \frac{n}{1 - \dfrac{\sum_{j=1}^{c} f_j}{c*n + \sum_{j=1}^{c}(j-c)f_j}}.$$

$\hat{N}_{G-censor}$ is actually the Horvitz-Thompson estimator, where $\hat{p}$ is the probability of not being captured. Notice that the estimate for $c = 1$ leads to the estimate $\hat{N} = \frac{n}{1-f_1/n}$ (Niwitpong et al., 2012).

## 6.2 Point estimation for the geometric distribution with censored data and covariates

Following the technique applied in previous chapters we are going to introduce covariate information at individual level to explain the heterogeneity in the capture-recapture distribution. $p_i$ is linked to explanatory variables using

$$p_i = \frac{\mathrm{e}^{\alpha+\beta' Z_i}}{1+\mathrm{e}^{\alpha+\beta' Z_i}}, \tag{6.3}$$

where $i = 1, \ldots, M$, $M$ being the total number of different covariate combinations.

As a result of this link, we can easily obtain the probabilities and the likelihood for the case with auxiliary variables, replacing the unique probability $p$ by $p_i$:

$$\begin{aligned} q_{ik} = P(Y_i = k) &= \frac{p_i(1-p_i)^k}{1-p_i} = p_i(1-p_i)^{k-1} \\ q_{ic+} = P(Y_i > c) &= 1 - \sum_{j=1}^{c} q_{ij} = 1 - \sum_{j=0}^{c-1} p_i(1-p_i)^j = \sum_{j=c}^{\infty} p_i(1-p_i)^j \\ &= \sum_{j=0}^{\infty} p_i(1-p_i)^{j+c} = (1-p_i)^c \sum_{j=0}^{\infty} p_i(1-p_i)^j = (1-p_i)^c. \end{aligned}$$

The likelihood and log-likelihood are provided as

$$\begin{aligned} \mathcal{L}(p_i) &= \prod_{i=1}^{M} \left( \prod_{j=1}^{c} q_{ij}^{f_{ij}} q_{ic+}^{(n_i - \sum_{j=1}^{c} f_{ij})} \right) \\ \ell(p_i) &= \sum_{i=1}^{M} \Bigg[ \Big( \sum_{j=1}^{c} f_{ij} \Big) \log(p_i) + \Big( \sum_{j=1}^{c} (j-1) \times f_{ij} \Big) \log(1-p_i) \\ &\quad + c \Big( n_i - \sum_{j=1}^{c} f_{ij} \Big) \log(1-p_i) \Bigg], \end{aligned}$$

where $n_i$ is the number of individuals observed in the $i$th strata. The log-likelihood, replacing $logit(p_i) = \alpha + \beta' Z_i$, becomes

$$\begin{aligned} \ell(\alpha, \beta) = \sum_{i=1}^{M} \Bigg[ &\left( \sum_{j=1}^{c} f_{ij} \right) \log \left( \frac{\mathrm{e}^{\alpha+\beta' Z_i}}{1+\mathrm{e}^{\alpha+\beta' Z_i}} \right) + \left( \sum_{j=1}^{c} (j-1) \times f_{ij} \right) \log \left( \frac{1}{1+\mathrm{e}^{\alpha+\beta' Z_i}} \right) \\ &+ c \left( n_i - \left( \sum_{j=1}^{c} f_{ij} \right) \right) \log \left( \frac{1}{1+\mathrm{e}^{\alpha+\beta' Z_i}} \right) \Bigg]. \end{aligned}$$

We do not have a closed form to estimate $\alpha$ and $\beta$, but they can be obtained maximising the likelihood using a numerical algorithm as shown in previous chapters.

Now, $\hat{f}_0$ is calculated as $E(f_0|\hat{\alpha},\hat{\beta}) = \sum_{i=1}^{M} E(f_{i0})$, we first calculate all $e_{i0} = E(f_{i0})$ for each strata:

$$e_{i0} = N_i\hat{q}_{i0} = (n_i + e_{i0}) \times \hat{q}_{i0} \ ,$$

where $N_i$ is the total number of units with covariate combination $i$ in the population, $n_i$ is the total number of units sampled and $q_{i0}$ is the probability of non-being captured for individuals with covariate combination $i$ :

$$e_{i0} = \frac{n_i\hat{q}_{i0}}{(1-\hat{q}_{i0})}.$$

We can write the expression with respect to $\hat{\alpha}$ and $\hat{\beta}'$ as

$$e_{i0} = n_i \mathrm{e}^{\hat{\alpha}+\hat{\beta}'Z_i}.$$

Hence, the estimate of the population size $N$ is,

$$\hat{N}_{Geo-censored} = n + \sum_{i=1}^{M} n_i \mathrm{e}^{\hat{\alpha}+\hat{\beta}'Z_i}.$$

$\hat{f}_0$ is calculated as the weighted sum of the number of people in each strata. Notice that the weights depend on the individual probability of being captured in that strata. If $q_{i0} \geq 0.5$, we expect to have at least the same number of individuals of that stratum within the hidden population.

## 6.3 Analytical variance for the estimator of population size based upon the geometric distribution with censoring

An estimator was presented for the geometric distribution using the concept of censoring and including covariate information to adjust for heterogeneity in the capture probability:

$$N_{Geo-censor} = n + \sum_{j=1}^{M} n_j \frac{p_j}{1-p_j} = n + \sum_{j=1}^{M} n_j \mathrm{e}^{\alpha+\beta'Z_j} = \sum_{i=1}^{N} \Delta_i(1 + \mathrm{e}^{\alpha+\beta'Z_i}), \qquad (6.4)$$

where $Z_j$ contains the $j$th covariate information, $Z_i$ contains the covariate information for the $i$th individual, $M$ is the number of different covariate combinations and $n_j$ is the

total number of individuals captured with the $j$th covariate combination.

The link between the probability $p_i$ of being captured and the covariate information was provided as $p_i = expit(\alpha+\beta' Z_i)$. The variance can be obtained using the same approach applied in previous chapters where

$$var(\hat{N}_{Geo-censor}) = E\left[Var(\hat{N}_{Geo-censor}|\Delta_i, i = 1, .., N)\right]+Var\left[E(\hat{N}_{Geo-censor}|\Delta_i, i = 1, .., N)\right], \tag{6.5}$$

where

$$\Delta_i = \begin{cases} 1, & y_i \in \{1, .., m\} \\ 0, & otherwise \end{cases},$$

and $m$ is the maximum number of occasions that an individual was observed. We can compute

$$E(\Delta_i) = q_i \text{ and } Var(\Delta_i) = q_i(1 - q_i),$$

where $q_i$ is the probability for the $i$th individual of being captured:

$$q_i = p(\Delta_i = 1|p_i) = \sum_{k=1}^{m} p_i(1 - p_i)^k = \sum_{k=1}^{m} \frac{\mathrm{e}^{\alpha+\beta Z_i'}}{(1 + \mathrm{e}^{\alpha+\beta' Z_i})^{(k+1)}}. \tag{6.6}$$

$E(N_{Geo-censor}|\Delta i, i = 1, \ldots, N)$ can also be written as

$$E(\hat{N}_{Geo-censor|\Delta_i, i=1,..N}) = \sum_{i=1}^{N} \Delta_i \left(1 + \frac{q_i}{\sum_{k=1}^{m}(1 - p_i)^{k+1}}\right) = \sum_{i=1}^{N} \Delta_i \omega_i.$$

We can now calculate the variance as

$$Var(\hat{N}_{Geo-censor|\Delta_i, i=1,..N}) = \sum_{i=1}^{N} Var(\Delta_i \omega_i) = \sum_{i=1}^{N} q_i(1 - q_i)\omega_i^2.$$

We can use the Horvitz-Thompson estimator to calculate an estimate of this variance:

$$\begin{aligned} \widehat{Var}(\hat{N}_{Geo-censor|\Delta_i, i=1,..N}) &= \sum_{i=1}^{N} \frac{\Delta_i}{\hat{q}_i} \hat{q}_i(1 - \hat{q}_i)\hat{\omega}_i^2 \\ &= \sum_{i=1}^{f_1+\ldots+f_m} (1 - \hat{q}_i) \left(1 + \frac{\hat{q}_i}{\sum_{k=1}^{m}(1 - \hat{p}_i)^{k+1}}\right)^2. \end{aligned} \tag{6.7}$$

Ultimately, we use the multivariate Delta method to calculate the second term:

$$\widehat{Var}[E(N_{GC}|\Delta_i, i = 1, .., N)] = \nabla g(\hat{\alpha}, \hat{\beta})' \widehat{cov}(\hat{\alpha}, \hat{\beta}) \nabla g(\hat{\alpha}, \hat{\beta}), \tag{6.8}$$

where $g(\hat{\alpha}, \hat{\beta}) = n + \sum_{j=1}^{f_1+\ldots+f_m} n_j e^{\hat{\alpha}+\hat{\beta}' Z_j}$ and then

$$\nabla g(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} \frac{\partial g}{\partial \hat{\alpha}} \\ \frac{\partial g}{\partial \hat{\beta}_1} \\ \ldots \\ \frac{\partial g}{\partial \hat{\beta}_p} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{f_1+f_2+\ldots+f_m} n_j e^{\hat{\alpha}+\hat{\beta}' Z_j} \\ \sum_{j=1}^{f_1+f_2+\ldots+f_m} n_j Z_1 e^{\hat{\alpha}+\hat{\beta}' Z_j} \\ \ldots \\ \sum_{j=1}^{f_1+f_2+\ldots+f_m} n_j Z_p e^{\hat{\alpha}+\hat{\beta}' Z_j} \end{pmatrix}.$$

$\hat{\beta}$ is the vector of estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$, associated to each auxiliary variable.

An estimate of the covariance matrix ($\widehat{cov}(\hat{\alpha}, \hat{\beta})$) of model parameters is the inverse of the Fisher information matrix, normally provided within the routine of the numerical algorithm. The final estimate is obtained summing (6.7) and (6.8)

## 6.4 Simulations

Following the reasoning of previous chapters we have simulated a population with a capture-recapture distribution depending on 2 auxiliary variables. However, we fit our models only with 1 covariate to assess the estimator performance when there is unmeasured heterogeneity. We simulate 2000 repetitions of capturing a population. The capture-recapture distribution follows a geometric distribution $Y \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1$ following a normal distribution with mean 40 and variance 144 ($X_1 \sim N(40, 144)$), $X_2 \sim N(10, 9)$, and $X_1$ and $X_2$ are independent.

Three estimators are reported, the extension of the geometric estimator based on censoring and including covariate information, the generalised Chao estimator using truncation and covariate information and Turing's estimator for the geometric distribution calculated in (4.29).

The models based on truncation present better point estimates than the models that apply the concept of censoring (table 6.2). However, the standard deviations of the estimators based on censoring are much smaller being close to the standard deviation of Turing's estimator. When we increase the cut-off for censoring, the bias increased slowly and the variability remains fairly constant with a slight decrease (Figures 6.1,6.2,6.3).

The variance estimates for the estimatiors based on censoring and covariates were small. We observe that the part of the variance that increases with respect to the population size estimate, does not have a large impact in the global estimate, in contrast to the estimators based on truncation. However, it is reassuring to see that the analytical formula provided a good approximation to the empirical standard deviation (table 6.1).

The RMSE values reflect an asymptotic improvement in the censored models as we observed for the truncated models in previous chapters. The censored models seem to perform better for small samples with respect to the RMSE, but the models with left and right truncation are superior in medium and large samples. We find an exception for the estimators with two non-truncated and censored counts where the large variability of the truncated models leads to a larger RMSE values.

Table 6.1: Comparison between the analytical and the empirical standard deviation for the estimate assuming censoring and a geometric capture-recapture distribution. The scenario comprises data generated from a geometric distribution $Y_i \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$ independent. Model fitting based on $X_1$ only.

| Counts | Analytical SD | Empirical SD | Analytical SD | Empirical SD | Analytical SD | Empirical SD |
|---|---|---|---|---|---|---|
| | $N = 500$ | | $N = 1000$ | | $N = 2000$ | |
| 2 | 45.93 | 45.31 | 64.39 | 69.79 | 89.83 | 89.64 |
| 3 | 43.81 | 43.58 | 61.26 | 67.65 | 85.55 | 86.56 |
| 4 | 42.91 | 42.90 | 59.96 | 66.40 | 83.74 | 85.43 |
| 5 | 42.49 | 42.60 | 59.37 | 66.10 | 82.89 | 84.91 |
| 6 | 42.21 | 42.26 | 59.11 | 65.93 | 82.45 | 84.84 |

Table 6.2: Estimates from the model with a capture-recapture geometric $Y_i \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$, independently. Model fitting based on $X_1$ only.

| $N$ | Estimate | Counts | $\hat{N}$ | $SD_N$ | RMSE(x100) |
|---|---|---|---|---|---|
| 500 | Censored | 2 | 473.05 | 45.93 | 1.11 |
| | | 3 | 469.52 | 43.81 | 1.13 |
| | | 4 | 467.24 | 42.91 | 1.17 |
| | | 5 | 465.96 | 42.49 | 1.19 |
| | | 6 | 464.89 | 42.21 | 1.21 |
| 500 | Truncated | 2 | 505.5 | 87.09 | 3.18 |
| | | 3 | 487.38 | 61.93 | 1.41 |
| | | 4 | 480.63 | 52.89 | 1.16 |
| | | 5 | 476.21 | 48.59 | 1.12 |
| | | 6 | 472.51 | 46.19 | 1.12 |
| 500 | Turing | | 456.28 | 40.4 | 1.42 |
| 1000 | Censored | 2 | 945.42 | 64.39 | 0.78 |
| | | 3 | 936.37 | 61.26 | 0.86 |
| | | 4 | 931.73 | 59.96 | 0.91 |
| | | 5 | 929.16 | 59.37 | 0.94 |
| | | 6 | 928.09 | 59.11 | 0.95 |
| 1000 | Truncated | 2 | 990.07 | 113.59 | 1.32 |
| | | 3 | 972.20 | 82.03 | 0.80 |
| | | 4 | 958.63 | 71.34 | 0.77 |
| | | 5 | 948.75 | 66.02 | 0.77 |
| | | 6 | 940.80 | 63.07 | 0.83 |
| 1000 | Turing | | 914.75 | 61.50 | 1.11 |
| 2000 | Censored | 2 | 1877.42 | 89.83 | 0.58 |
| | | 3 | 1861.06 | 85.55 | 0.67 |
| | | 4 | 1851.41 | 83.74 | 0.73 |
| | | 5 | 1846.18 | 82.89 | 0.77 |
| | | 6 | 1843.04 | 82.45 | 0.80 |
| 2000 | Truncated | 2 | 1958.56 | 154.8 | 0.64 |
| | | 3 | 1925.23 | 114.67 | 0.45 |
| | | 4 | 1902.42 | 99.80 | 0.48 |
| | | 5 | 1884.17 | 92.43 | 0.55 |
| | | 6 | 1872.11 | 88.54 | 0.61 |
| 2000 | Turing | | 1820.96 | 83.13 | 0.97 |

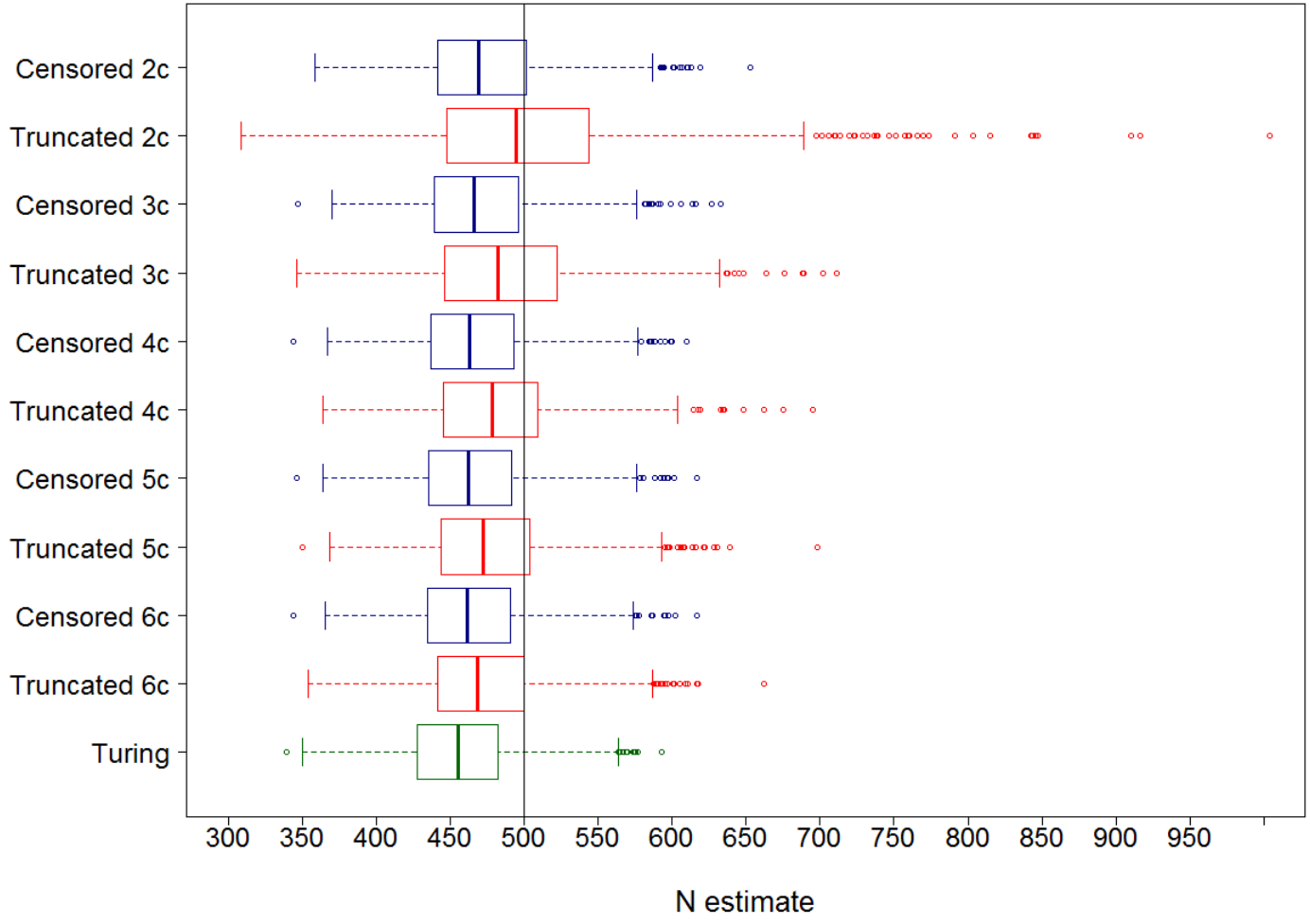


Figure 6.1: Comparison of estimates based on truncation and censoring for $N = 500$. The capture distribution $Y_i \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$, independently. Model fitting based on $X_1$ only.

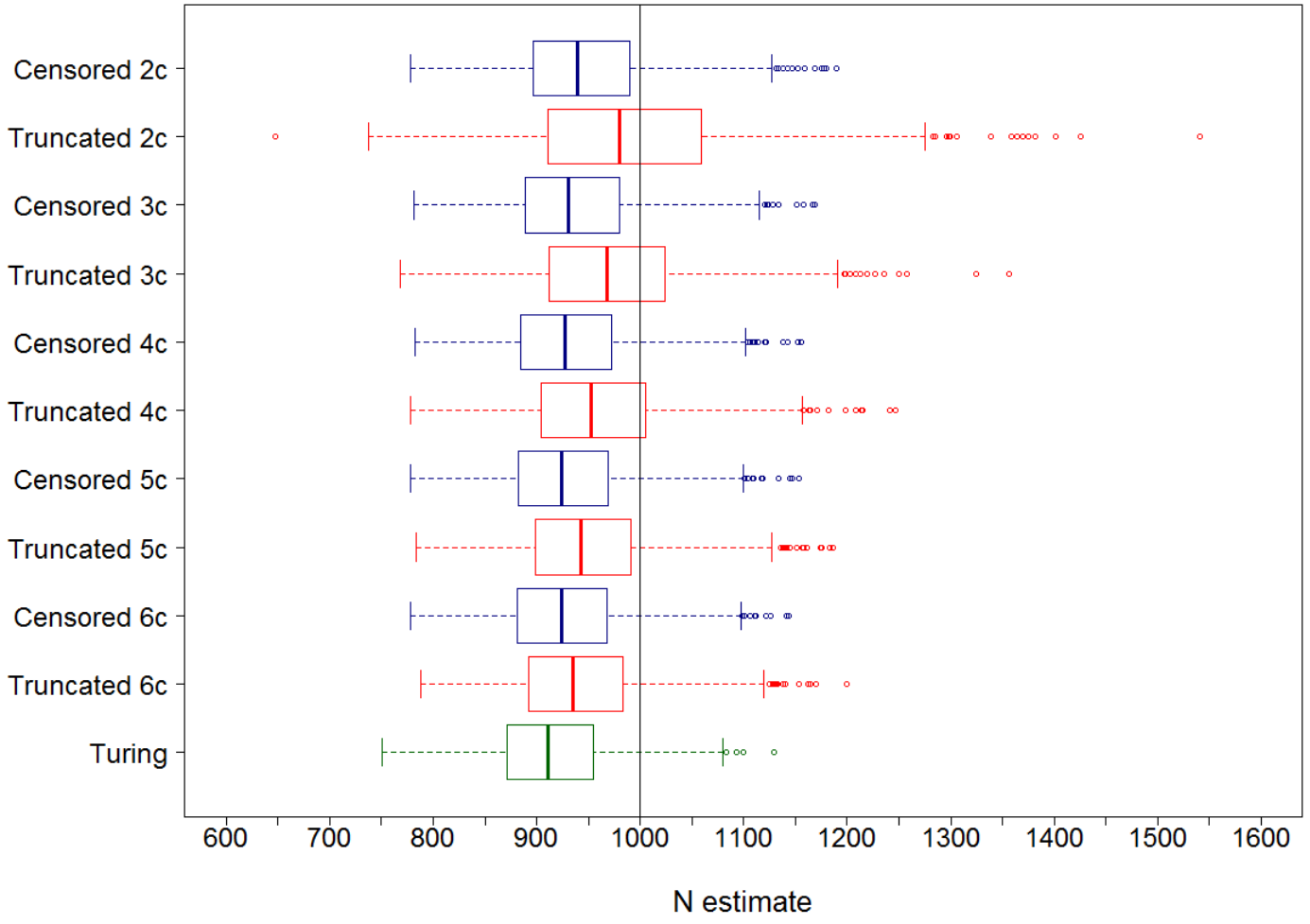


Figure 6.2: Comparison of estimates based on truncation and censoring for $N = 1000$. The capture distribution $Y_i \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$, independently. Model fitting based on $X_1$ only.

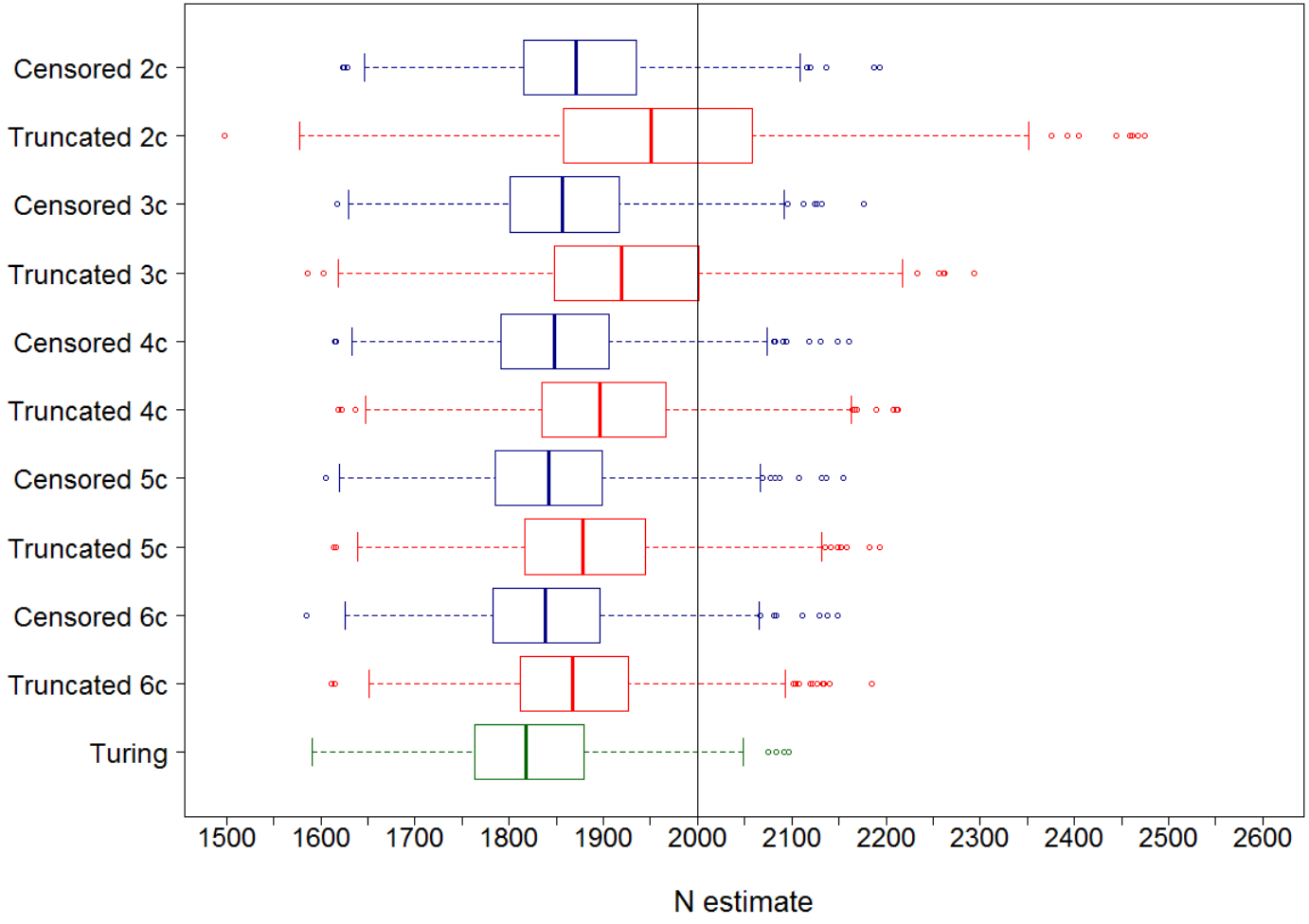


Figure 6.3: Comparison of estimates based on truncation and censoring for $N = 2000$. The capture distribution $Y_i \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$, independently. Model fitting based on $X_1$ only.

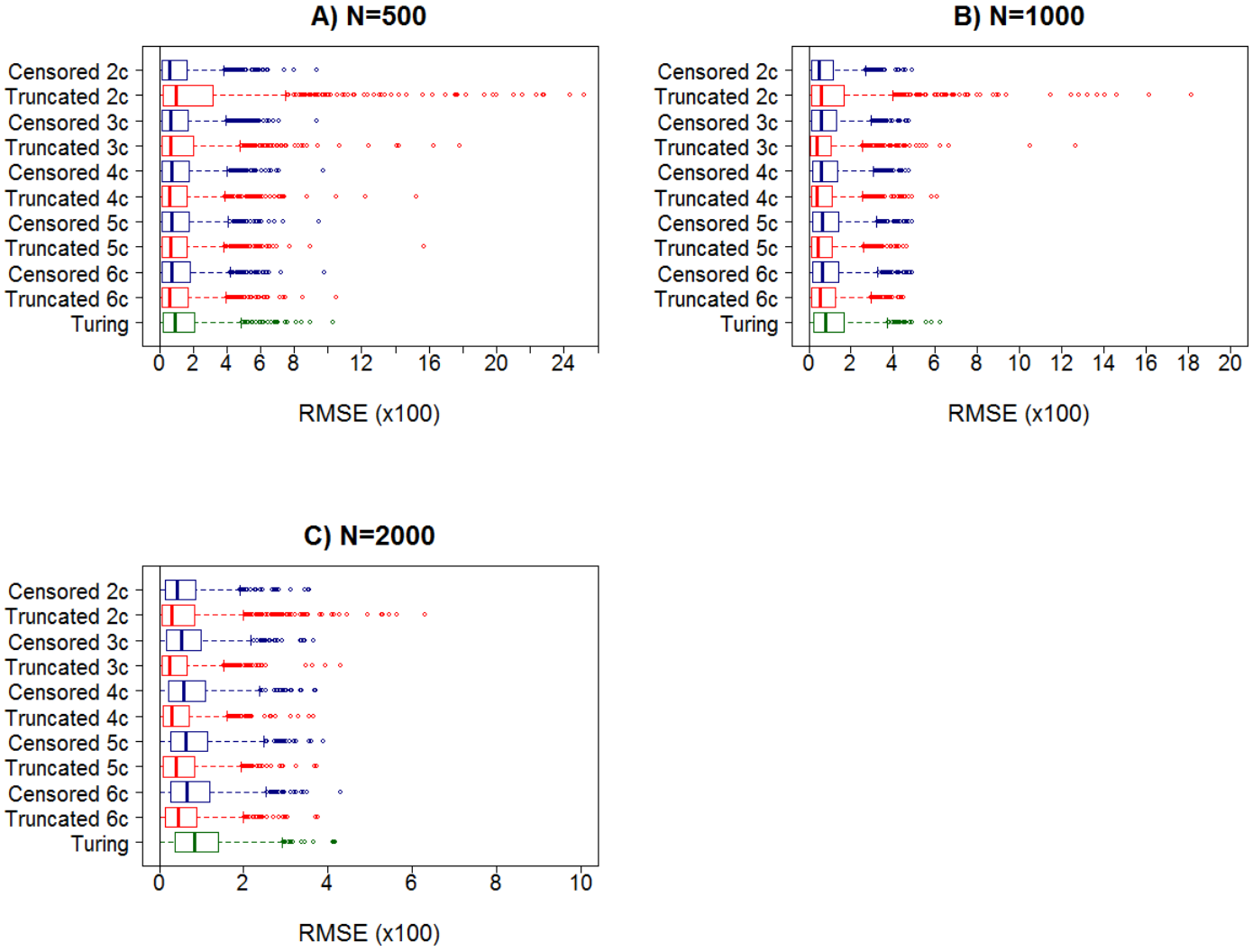


Figure 6.4: Comparison of RMSE(x100) values. The capture distribution $Y \sim G(q_i)$, where $logit(q_i) = -0.02X_1 + 0.13X_2$ with $X_1 \sim N(40, 144)$ and $X_2 \sim N(10, 9)$, independently. Model fitting based on $X_1$ only. A)$N = 500$, B)$N = 1000$, C)$N = 2000$

## 6.5 Case study

We aim at estimating the number of heroin users in Bangkok. The data were previously described in detail (Böhning et al., 2004). It is based on a study conducted at the end of 2001 in Bangkok. They collected information of treatment episodes of drug users from 61 health centres. The data consist on one entry list with repeated entries. We consider three covariates: gender, age group and marital status.

The ratio plot, based on $r(x) = \frac{(x+1)f_{x+1}}{f_x}$ (figure 6.5 and table 6.3), suggests the present of structural heterogeneity. A potential distribution to fit the data can be the geometric distribution, as a mixture between a Poisson and an exponential distribution. We compare two estimators based on a geometric distribution: 1) The estimator developed in section 4.2 using a zero- and righ-truncation. 2) the estimator based on zero-truncation and censoring (section 6.2). Table 6.4 shows the point estimates and the standard errors for all models varying the number of censored/truncated counts and the covariate structure.

The standard errors follow the same pattern observed in the simulations. The estimator based on a geometric censored distribution presents smaller standard errors than the zero- and right-truncated estimator. The standard errors decrease slightly when decreasing the number of censored counts compared to the faster decrease when the number of non-truncated counts increased. In the case of the estimator based on zero- and right-truncation, this decrease is caused by the inclusion of more data and the increase in the bias that leads to smaller population estimates.

The point estimates from the estimators using censoring are fairly similar across the different covariate structures and the number of counts considered. In contrast, the estimators with truncation obtained larger population sizes. The estimates decrease when more information is considered. The estimator based on two non-truncated counts is usually less stable, although in this example frequencies of counts 1 and 2 represent 58.61% of the observed population.

Likelihood ratio tests are calculated to determine the impact of each covariate in the probability of being captured (table 6.5). Age group and marital status are significantly associated to the capture probability in the models using censoring. On the other hand gender and marital status are significant across all models based on truncation. Age group is only found significant in three of the truncated models.

The expected log-ratio plot and covariate-frequency plots were obtained to assess the performance of the estimators (Figures 6.6, 6.7, 6.8, 6.9). The lines of the estimators using less counts are at the front overlapped with the estimators with more counts for easier visualization. The estimators using a censored geometric distribution severely underestimate the number of individuals captured once (figure 6.8) which have a large

impact in the estimations. They also underestimate other frequencies. The reader should notice the scale of the graph to realise of the size of the underestimation. The zero and right-truncated models present a better fit. The increase of non-truncated counts leads to an increase of the underestimation of $f_1$. The problem could be caused by the fact that the capture-recapture distribution is one-inflated. These models also overestimate severely $f_2$ (figure 6.9).

The $\chi^2$ test described in the previous chapter is significant for all combinations, an indication of the lack of fit of the models. The geometric distribution has not produced the results expected and other distribution should be tested to find a better fit to the capture-recapture distribution.

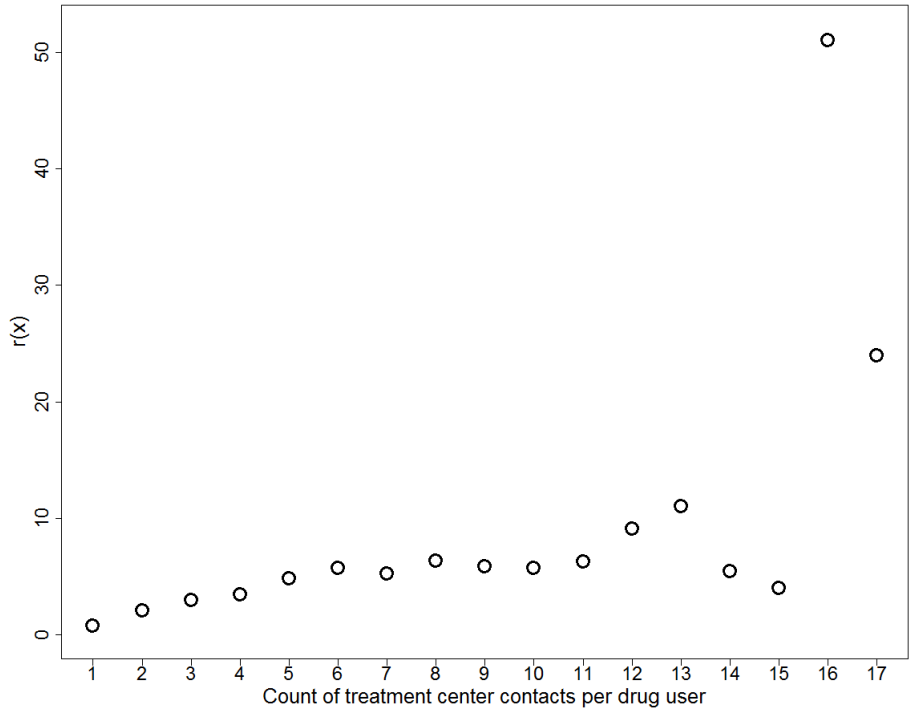


Figure 6.5: Ratio plot for the case study on the hidden number of heroin users in Bangkok

Table 6.3: Ratios $\hat{r}_x = (x+1)f_{x+1}/f_x$ and 95% confidence limits for the heroin drug users in Bangkok.

| ratio | $\hat{r}_x$ | $\hat{r}_x$ 95% CL |
|---|---|---|
| $r_1$ | 0.80 | (0.74 - 0.85) |
| $r_2$ | 2.05 | (1.88 - 2.25) |
| $r_3$ | 3.00 | (2.70 - 3.34) |
| $r_4$ | 3.45 | (3.04 - 3.91) |
| $r_5$ | 4.86 | (4.20 - 5.62) |
| $r_6$ | 5.73 | (4.89 - 6.73) |
| $r_7$ | 5.24 | (4.33 - 6.32) |
| $r_8$ | 6.32 | (5.03 - 7.94) |
| $r_9$ | 5.84 | (4.38 - 7.79) |
| $r_{10}$ | 5.73 | (3.87 - 8.47) |
| $r_{11}$ | 6.32 | (3.68 - 10.85) |
| $r_{12}$ | 9.10 | (4.6 - 18.02) |
| $r_{13}$ | 11.00 | (4.99 - 24.23) |
| $r_{14}$ | 5.45 | (1.74 - 17.13) |
| $r_{15}$ | 4.00 | (0.45 - 35.79) |
| $r_{16}$ | 51.00 | (5.3 - 490.31) |
| $r_{17}$ | 24.00 | (5.37 - 107.24) |

Table 6.4: Case study: Number of heroin users in Bangkok. Point estimates and standard errors for all models varying the number of non-truncated/non-censored counts and the covariates (gender, marital status (MS) and age group).

| Counts | Model | $\hat{N}_{cens}$ | $\hat{SE}_{N_{cens}}$ | $\hat{N}_{trunc}$ | $\hat{SE}_{N_{trunc}}$ |
|---|---|---|---|---|---|
| 2 | Sex | 10941 | 114.17 | 19438 | 357.81 |
| | Sex+Marital Status | 11057 | 115.03 | 19605 | 368.35 |
| | Sex+Marital Status+Age group | 11070 | 115.62 | 19619 | 370.88 |
| 3 | Sex | 10880 | 101.25 | 15051 | 209.18 |
| | Sex+Marital Status | 10718 | 101.91 | 15101 | 212.11 |
| | Sex+Marital Status+Age group | 10729 | 102.33 | 15160 | 215.17 |
| 4 | Sex | 10835 | 95.65 | 13075 | 156.88 |
| | Sex+Marital Status | 10578 | 96.18 | 13121 | 158.82 |
| | Sex+Marital Status+Age group | 10589 | 96.57 | 13135 | 159.67 |
| 5 | Sex | 10808 | 92.27 | 12162 | 133.72 |
| | Sex+Marital Status | 10486 | 92.79 | 12196 | 135.01 |
| | Sex+Marital Status+Age group | 10498 | 93.15 | 12203 | 135.47 |
| 6 | Sex | 10465 | 90.76 | 11495 | 118.11 |
| | Sex+Marital Status | 10458 | 91.18 | 11529 | 119.35 |
| | Sex+Marital Status+Age group | 10468 | 91.51 | 11544 | 120.01 |
| 7 | Sex | 10482 | 90.51 | 11006 | 107.13 |
| | Sex+Marital Status | 10476 | 90.88 | 11026 | 107.85 |
| | Sex+Marital Status+Age group | 10487 | 91.22 | 11036 | 108.23 |

Table 6.5: Case study: Number of heroin users in Bangkok. Likelihood ratio tests for all models varying the number of non-truncated/non-censored counts and the covariates (gender, marital status (MS) and age group).

| NC/NT counts | Model | df | $\chi^2_{cens}$ | $p-value_{cens}$ | $\chi^2_{trunc}$ | $p-value_{trunc}$ |
|---|---|---|---|---|---|---|
| 2 | Sex | 1 | 0.054 | 0.8162 | 7.078 | 0.0078 |
| | Sex+MS | 3 | 24.254 | $< .0001$ | 17.312 | 0.0006 |
| | Sex+MS+Age group | 4 | 15.174 | 0.0043 | 2.534 | 0.6386 |
| 3 | Sex | 1 | 0.200 | 0.6547 | 8.368 | 0.0038 |
| | Sex+MS | 3 | 21.148 | $< .0001$ | 17.754 | 0.0005 |
| | Sex+MS+Age group | 4 | 14.36 | 0.0062 | 12.564 | 0.0136 |
| 4 | Sex | 1 | 0.010 | 0.9203 | 30.224 | $< .0001$ |
| | Sex+MS | 3 | 20.124 | 0.0002 | 24.068 | $< .0001$ |
| | Sex+MS+Age group | 4 | 15.352 | 0.0040 | 7.082 | 0.0991 |
| 5 | Sex | 1 | 0.002 | 0.9643 | 29.926 | $< .0001$ |
| | Sex+MS | 3 | 20.356 | 0.0001 | 24.182 | $< .0001$ |
| | Sex+MS+Age group | 4 | 16.326 | 0.0026 | 6.764 | 0.1489 |
| 6 | Sex | 1 | -0.006 | 1.0000 | 36.698 | $< .0001$ |
| | Sex+MS | 3 | 15.318 | 0.0016 | 30.368 | $< .0001$ |
| | Sex+MS+Age group | 4 | 14.532 | 0.0058 | 6.764 | 0.0135 |
| 7 | Sex | 1 | 0.012 | 0.9128 | 65.682 | $< .0001$ |
| | Sex+MS | 3 | 15.734 | 0.0013 | 22.628 | $< .0001$ |
| | Sex+MS+Age group | 4 | 15.024 | 0.0047 | 11.470 | 0.0218 |

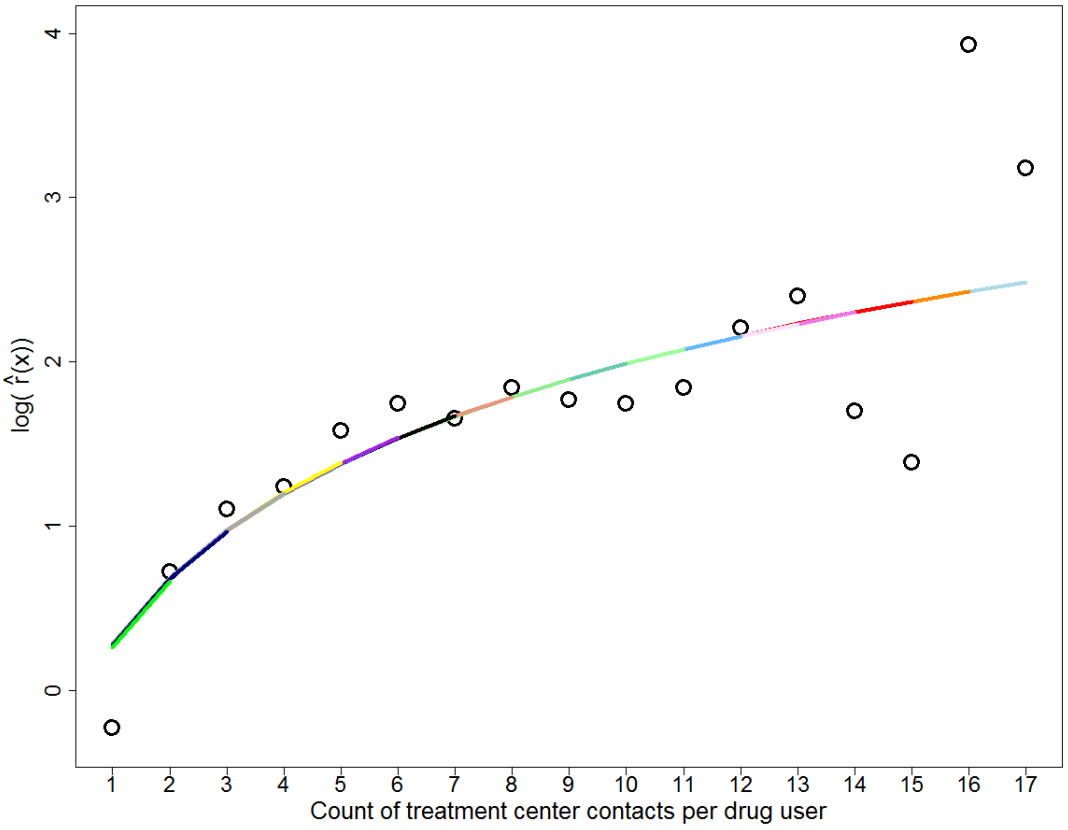


Figure 6.6: Case study: number of heroin users in Bangkok. Observed and fitted log ratio plot for the model based on a zero-truncated and censored geometric distribution.

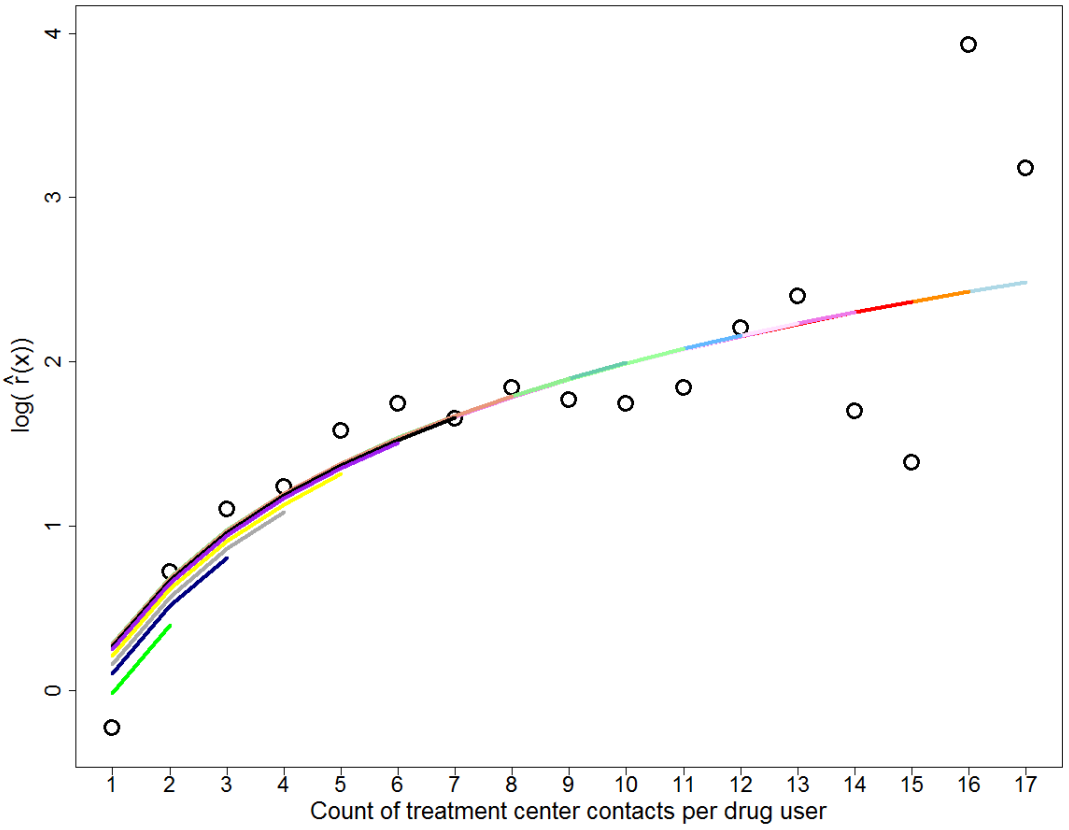


Figure 6.7: Case study: number of heroin users in Bangkok. Observed and fitted log ratio plot for the model based on a zero and right truncated geometric distribution.

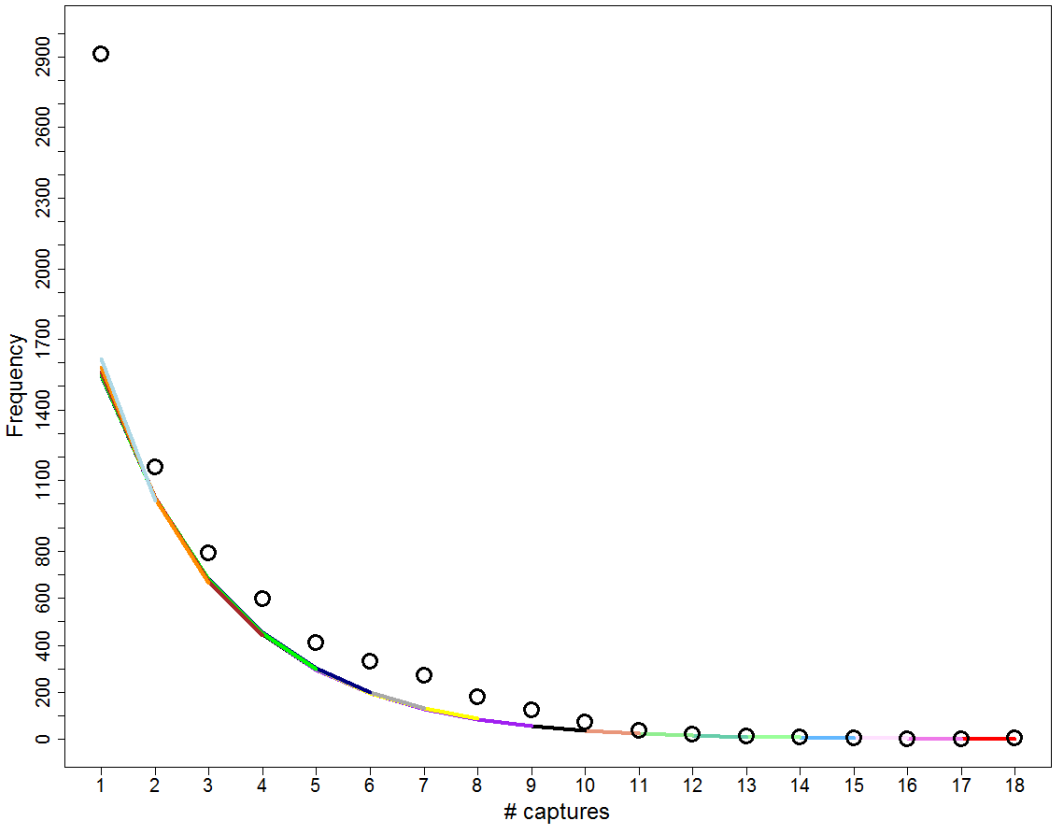


Figure 6.8: Case study number of heroin users in Bangkok. Observed and fitted covariate-adjusted frequency plot for the model based on a zero-truncated and censored geometric distribution.

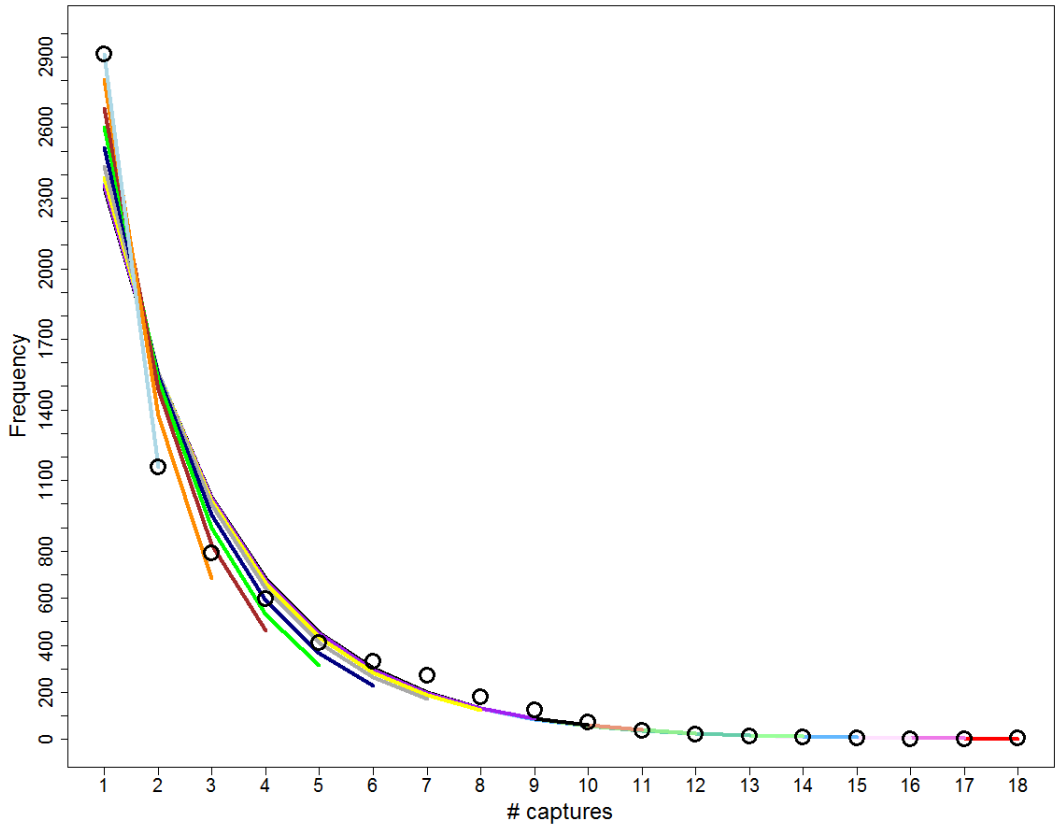


Figure 6.9: Case study number of heroin users in Bangkok. Observed and fitted covariate-adjusted frequency plot for the model based on a zero and right-truncated geometric distribution.

## 6.6 Conclusions

In this chapter, we explored the idea of applying censoring rather than truncation for situations where the geometric distribution is appropriate. We were motivated by the fact that censoring does not ignore the censored counts compared to the concept of truncation. We started with the estimator suggested by Niwitpong et al. (2012) and we extended it to choose a censoring cut-off and to include covariate information to model the individual capture-recapture probability.

We compared the new estimators with the estimators from a zero and right truncated geometric distribution. The estimators based on truncations were calculated applying the general formulae described in chapter 4. These "truncated" estimators were found superior in large samples to the estimators based on censoring on the basis of the RMSE criterion. However, "censored" estimators with two counts presented better RMSE than the estimators based on two non-truncated counts because of their large standard errors. The estimates from the models using a geometric distribution with censoring presented larger bias but smaller variability. An increase in the number of non-censored counts did not show a large impact in the estimates of the population and the standard errors.

A case study was also included where both models could not explain all the heterogeneity, but the models with censoring underestimated largely the number of individuals captured once that accounted 41.9% of the capture distribution. The geometric distribution was not the right distribution in this problem and other distributions could be tested.

We do not investigate further this route of estimation after the results obtained. The application of right-truncation has given evidence to provide more realistic and accurate results.

# Chapter 7

# General conclusions and discussion

Our initial motivation was to extend Chao's lower bound estimator to include auxiliary variables measured on the captured individuals. We chose Chao's estimator because of its robustness even in the presence of some heterogeneity. We showed that a model with right-truncation associated with the multinomial likelihood was the appropriate way to incorporate covariate information. The chapter 2 considered a model assuming a Poisson distribution with all counts truncated except individuals captured one or two times. A link between the likelihood of the truncated Poisson with two non-truncated counts and a logistic regression likelihood was exploited to provide an easy way to obtain an estimate of the population with any standard statistical package.

The simulated scenarios with unexplained heterogeneity showed a good performance of the generalised Chao estimator (GC) compared to the estimator based on a zero-truncated Poisson model (ZTP) (Van der Heijden et al., 2003a). The generalised Chao estimator obtained better relative mean squared error in populations larger than 1000 individuals. The point estimates were better than the ZTP estimator but the variance was larger. Bigger differences between GC and ZTP were found in the scenarios with part of the observed individuals misclassified. GC presented robust estimates and considerably smaller RMSE values. The robustness of the GC estimate was also observed in the simulated scenario generated from a negative binomial distribution.

The natural following step was to develop a framework where the level of truncation changes. Chapter 3 continued assuming a Poisson distribution but the number of non-truncated counts is a parameter $J$, with $J \geq 2$. The methodology led to the application of numerical algorithms to obtain maximum likelihood estimators, contrary to the easy calculation of the GC estimates with two non-truncated counts. The simulations carried out in the chapter concluded that an increase in the number of non-truncated counts increased the bias of the estimate but reduces its variability. The models with 2, 3 or

4 non-truncated counts presented the best RMSE values; an indication that the tail of the capture-recapture distribution can impact negatively the accuracy of the estimates.

In the following chapter 4, we relaxed the assumption of the capture-recapture distribution. We generalised the framework for the power series distributions. We focused on the Poisson, binomial and geometric distribution as typical distributions found in the field of capture-recapture. An analytical variance was also provided. The simulations were based on a binomial distribution and conclusions similar to those observed in the Poisson case were obtained. We also showed the impact of having more captured occasions in the particular case of a binomial distribution.

Following the reasoning of any statistical modelling, we were interested in assessing the performance of our models in real life and the establishment of a decision rule to determine the optimal cut-off truncation point. A $\chi^2$ test was developed in the chapter 5. We found that our proposed $\chi^2$ statistic followed a $\chi^2$ distribution with $J-2$ degrees of freedom independently of how many covariates we include in the analysis. The chapter was also complemented with a section proposing two ad-hoc model-averaging estimators. These weighted estimators provided a balanced solution combining models with different truncation cut-off points. They are specially adequate to reduce the large variability we normally obtain using only two non-truncated counts.

The model with two non-truncated counts can sometimes produce larger estimates variance compared to the models with more non-truncated counts. Although the $\chi^2$ test should determine the optimal number of non-truncated counts, we saw that it is useful to use other visual tools like the ratio plot and the covariate-adjusted frequency plot to decide the best model.

The models inherit the problems of the models based on the conditional approach introduced by Huggins and Alho (Huggins, 1989; Alho, 1990; Farcomeni and Tardella, 2012). The covariate information used to infer results comes only from the observed individuals. Our models also make the common assumption in modelling that there is not unobserved heterogeneity. Although the modelling and the application of the EM algorithm were designed to cope with observed and some unobserved heterogeneity. We showed in our simulations that in the case of unexplained heterogeneity our estimators underestimate the true population size but performed better than other estimators with covariates like the zero-truncated Poisson estimator (Van der Heijden et al., 2003b).

Another limitation of our models is the assumption of equal probability between sample occasions and the omission of behavioural effects. It is essential to use graphical tools and the developed $\chi^2$ test to assess the validity of the model as other authors have argued that capture-recapture inference requires more complex models like mixture models or generalised additive models. For instance, we observed for the case study of the number of heroin users in Bangkok that our models assuming a geometric distribution did not achieve a good fit to the data.

Chapter 6 was motivated by the idea of exploring a different approach to work with the tails of capture-recapture distributions. We extended an estimator based on the geometric distribution (Niwitpong et al., 2012) that censored all individuals captured more than once. Firstly we changed the censoring cut-off point to add covariate information later on. The new estimates were compared with the estimators based on truncated distributions. The standard errors of the estimators based on censoring were smaller than using truncation but the point estimates were less accurate. The RMSE values concluded that censoring could only outperform the truncation approach in small populations. It is interesting to note that the expectation that censoring could be a valuable way to use the entire information turned out to be illusive for the case of the geometric distribution.

Across the thesis we included several case studies that provided a practical guide to choose and validate the performance of the right estimator. All algorithms implemented along the duration of the PhD will be joined into an R library to facilitate the calculation of our estimates and to provide statistical and graphical tools to assess its efficiency and lack of fit.

## 7.1 Future Work

Our estimators are based on the conditional likelihood approach (Huggins, 1989) under the umbrella of the $M_h$ models that use individual covariate information of the captured subjects. Behavioural effects and covariates related to the time of capture could be added in the future to complete the set of models that consider the other two potential sources of heterogeneity based on Otis' classification (Otis et al., 1978). The first problem is to determine the best way to introduce the effects into the model. In a discrete experiment we could follow Huggins (1989) and use a log-linear model. However, the standard approaches for continuous-time experiments (Hwang and Chao, 2002a; Farcomeni and Scacciatelli, 2013) involve the use of Cox-type models for multiple events that take into account time-dependent covariates.

The model linking the parameters of the likelihood and the covariates could also be extended to use more complex structures like splines or generalised additive models. The challenge is to use those frameworks considering truncation. Generalised additive models for location, scale and shape (GAMLSS) could be a solution.

Our framework and methodology could be applied to other potential useful distributions in the area of capture-recapture that are not included under the power series distributions. The use of truncation add another difficulty to the application of those distributions. If we deviate from the power series distributional family, the major difficulty arises that the Chao estimator is no longer a lower bound as the argumentation involved in the Cauchy-Schwarz inequality (Böhning et al., 2006), presented also in chapter 2, is

no longer valid. Here, again, simulation work could help to investigate the validity of the truncated likelihood approach.

The techniques developed in the capture-recapture area to work with missing data and measurement errors in individuals' covariates could be also explored in our estimators. The sensitivity of our models to missing values in some covariates could be investigated in detail and the current techniques could be applied in our specific framework. An extension of our EM algorithm could be developed to impute all missing information. The use of the EM algorithm and the maximum likelihood estimation has been suggested when only categorical covariates are used (Van der Heijden et al., 2009). The other option is to use the multiple imputation by chained equations method (mice) that is simpler when there is a mixture of categorical and continuous covariates (Zwane and van der Heijden, 2008).

The impact on our estimates of extreme values or outliers in the covariate information need to be assessed and graphical tools could be developed to identify the individuals causing confounded estimates.

# References


Alho, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46(3):623–635.

Amstrup, S., McDonald, T., and Manly, B. (2005). *Handbook of capture-recapture analysis*. Princeton University Press.

Bailly, L., Daures, J., Dunais, B., and Pradier, C. (2015). Bayesian estimation of a cancer population by capture-recapture with individual capture heterogeneity and small sample. *BMC Medical Research Methodology*, 15(39).

Baker, S. (1990). A simple EM algorithm for capture-recapture data with categorical covariates. *Biometrics*, 46:1193–2000.

Becker, N. (1984). Estimating population size in capture-recapture experiments in continuous time. *Australian Journal of Statistics*, 26:1–7.

Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology*, 5:410–423.

Böhning, D. (2011). Capture-recapture estimation by means of empirical Bayesian smoothing with an application to the geographical distribution of hidden scrapie in Great Britain. *Applied Statistics*, 60(part 5):723–741.

Böhning, D., Baksh, M., Lerdsuwansri, R., and Gallagher, J. (2013a). Use of the ratio plot in capture-recapture estimation. *Journal of Computational and Graphical Statistics*, 22(1):135–155.

Böhning, D. and Del Rio Vilas, V. (2008). Estimating the hidden number of scrapie affected holdings in great britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological, and Enviromental Statistics*, 13:1–22.

Böhning, D. and Del Rio Vilas, V. (2009). On the question of proportionality of the count of observed scrapie cases and the size of holding. *BMC Veterinary Research*, 5:17.

Böhning, D., Holling, H., Böhning, W., and Viwatwongkasem, C. (2006). A generalization of Chao's inequality for population size estimation. *Laos Journal Applied Science*, 1:466–470.

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Applied Statistics*, 54(Part 4):721–737.

Böhning, D., Suppawattanabodee, B., and Kusolvisitkul, W. (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., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. (2013b). A generalization of Chao's estimator for covariate information. *Biometrics*, 69(4):1033–42.

Borchers, D., Zucchini, W., and Fewster, R. (1998). Mark-recapture models for line transect surveys. *Biometrics*, 54:1207–1220.

Boyce, M., Mackenzie, D., Manly, B., Horoldson, M., and Moddy, D. (2001). Negative binomial models for abundance estimation of multiple closed populations. *Journal of Wildlife Management*, 65:498–509.

Broyden, C. (1969). A new double rank minimization algorithm. *Notices American Mathematical Society*, 16:670.

Burnham, K. and Overton, W. (1978). Estimation of the size of a closed population when capture probabilities when capture probabilities vary among animals. *Biometrika*, 65(3):625–633.

Burnham, K. and Overton, W. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5):927–936.

Carothers, A. (1973). Capture-recapture methods applied to a population with known parameters. *Journal of Animal Ecology*, 42:125–146.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement errors in non linear models: A modern perspective*. Champman and Hall.London.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, 45:427–438.

Chao, A. and Lee, S. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87:210–217.

Chao, A., Lee, S., and Jeng, S. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48(1):201–216.

Chapman, D. (1951). Some properties of the hypergeometric distribution with application to zoological censuses. *University California Public Statistics*, 1:131–160.

Chen, S. and Lloyd, C. (2000). A non-parametric approach to the analysis of two-stage mark-recapture experiments. *Biometrika*, 87(3):663–649.

Cochran, W. (1978). Laplace ratio estimates. *Contributions to survey sampling and applied statistics.Academic Press*, pages 3–10.

Cormack, R. (1989). Log-linear models for capture-recapture. *Biometrics*, 45:395–413.

Coull, B. and Agresti, A. (1999). The use of mixed logit models to reflect heterogenity in capture-recapture studies. *Biometrics*, 55:294–301.

Creel, S., Spong, G., Sands, J., Rotella, J., Zeigle, J., Joe, L., Murphy, K., and Smith, D. (2003). Population size estimation i yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, 12:2003–2009.

Cruyff, M. and Van der Heijden, P. (2008). Point and interval estimation of the population size using a zero truncated negative binomial regression model. *Biometrical Journal*, 50:1035–1050.

Darroch, J. (1958). The multiple recapture census, I: estimation of a closed population. *Biometrika*, 45:343–359.

Darroch, J. (1959). The multiple recapture census, II: estimation when there is immigration or death. *Biometrika*, 46:336–351.

Darroch, J. and Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics*, 36:149–153.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39:1–38.

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.

Duran, J. and Wiorkowski, J. (1981). Capture-recapture sampling for estimating software error content. *IEEE Transactions on Software Engineering*, 7(1):147–148.

Farcomeni, A. and Scacciatelli, D. (2013). Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy. *The Annals of Applied Statistics*, 7(4):2293–2314.

Farcomeni, A. and Tardella, L. (2012). Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics*, 6:2602–2626.

Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete $2^k$ contingency tables. *Biometrika*, 59(3):591–603.

Fisher, R., Corbet, A., and Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12:42–58.

Fletcher, R. (1970). A new approach to variable metric methods. *Computer Journal*, 13:317–322.

Goldfarb, D. (1970). A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24:23–26.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Hald, A. (1990). *A History of Probability and Statistics and Their Applications Before 1750*. Wiley.New York.

Hay, G., Gannon, M., MacDougall, J., Eastwood, C., Williams, K., and Millar, T. (2009). Capture-recapture and anchored prevalence estimation of injecting drug users in england: national and regional estimates. *Statistical Methods in Medical Research*, 18:323–339.

Holling, H., Böhning, W., Böhning, D., and Formann, A. (2013). The covariate-adjusted frequency plot. *Statistical Methods in Medical Research*.

Holzmann, H., Munk, A., and Zucchini, W. (2006). On identifiability in capture-recapture models. *Biometrics*, 62:934–939.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

Huggins, R. and Chao, A. (2002). Asymptotic properties of an optimal estimating function approach to the analysis of mark recapture data. *Communication in Statistics*, 31:575–597.

Huggins, R. and Hwang, W. (2007). Non parametric estimation of population size from capture-recapture when the capture probability depends on a covariate. *Journal Royal Statistical Society Series C*, 56:429–443.

Huggins, R. and Hwang, W. (2011). A review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review*, 79(3):385–400.

Hwang, W. and Chao, A. (2002a). Continuous-time capture-recapture models with covariates. *Statistica Sinica*, 12:1115–1131.

Hwang, W. and Chao, A. (2002b). Continuous-time capture-recapture with covariates. *Statistica Sinica*, 12:1115–1131.

Hwang, W. and Huang, S. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrics*, 59:1113–1122.

Hwang, W. and Huggins, R. (2007). Application of semiparametric regression models in the analysis of capture recapture experiments. *Australia and New Zealand Journal of Statistics*, 49:191–202.

Johnson, N., Kemp, A., and Kotz, S. (2005). *Univariate Discrete Distributions*. Wiley.

Karanth, K. (1995). Estimating tiger panthera tigris populations from camera-trap data using capture-recapture models. *Biological Conservation*, 71(3):333–338.

Keating, K., Schwartz, C., Haroldson, M., and Moody, D. (2002). Estimating numbers of females with cubs-of-the-year in the yellowstone grizzly bear population. *Ursus*, 13:161–174.

Le Cren, E. (1965). A note on the history of mark-recapture populatione estimates. *Journal of Animal Ecology*, 34:453–454.

Link, W. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous probabilities. *Biometrics*, 59:1123–1130.

Link, W. (2006). Response to a paper by Holzmann,Munk and Zucchini. *Biometrics*, 62:936–939.

Link, W., Yoshizaki, J., Bailey, L., and Pollock, K. (2010). Uncovering a latent multinomial: Analysis of mark-recapture data with misindentification. *Biometrics*, 66:178–185.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley.New York.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010). Measuring elusive populations with Bayesian model averaging for multiple systems estimator: A case study of lethal violations in Casanare. *Statistics, Politics and Policy*, 1.

Mao, C. (2008). Computing an NPMLE for a mixing distribution in two closed heterogeneous population size models. *Biometrical Journal*, 50(6):983–992.

Mao, C. and Lindsay, B. (2002). Diagnostic for the homogeneity of inclusion probabilities in a bernoulli census. *Sankhya Series A*, 64:626–639.

Marten, G. (1970). A regression method for mark-recapture estimation for population size with unequal catchability. *Ecology*, 51:291–295.

McDonald, S., Hutchinson, S., Schnier, C., McLeod, A., and Goldberg, D. (2014). Estimating the number of injecting drug users in Scotland's HIV-diagnosed population using capture-recapture methods. *Epidemiology and Infection*, 142(1):200–207.

McKendrick, A. (1926). Application of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:99–130.

Mitchell, S., Ozonoff, A., Zaslavsky, A., Hedt-Gauthier, B., Lum, K., and Coull, B. (2013). A comparison of marginal and conditional models for capture-recapture data with application to human rights violations data. *Biometrics*, 69:1022–1032.

Nayak, T. (1988). Estimating population size by recapture sampling. *Biometrika*, 75:113–120.

Nelder, J. and Mead, R. (1965). A simplex method for computer minimization. *Computer journal*, 7:308–313.

Niwitpong, S., Böhning, D., van der Heijden, P., and Holling, H. (2012). Capture-recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika*, DOI 10.1007/s00184-012-0401-0.

Norris, J. and Pollock, K. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52:639–649.

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.

Pledger, S. and Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, 50(6):1022–1034.

Pollock, K. (1976). Building models of capture-recapture experiments. *The Statistician*, 25:253–260.

Pollock, K. (2002). The use of auxiliary variables in capture recapture modelling: An overview. *Journal of Applied Statistics*, 29:85–102.

Pollock, K., Hines, J., and Nichols, J. (1984). The use of auxiliary variables in capture–recapture and removal experiments. *Biometrics*, 40:329–340.

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.

Ross, S. (1985). *Introduction to Probability Models*. Academic Press.Orlando.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43:142–152.

Schnabel, Z. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45:348–352.

Seber, G. (1982). *The Estimation of Animal Abundance (2nd ed.)*. London:Griffin.

Seker, C. and Deming, W. (1947). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44:101–115.

Shanno, D. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematic of Computation*, 24:647–657.

Stanley, T. and Burnham, K. (1998). Information-theoretic model selection and model averaging for closed-population capture-recapture studies. *Biometrical Journal*, 40:475–494.

Stoloksa, J. and Huggins, R. (2012). A robust p-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics and Data Analysis*, 56:408–417.

Stoloksa, J., Hwang, W., Huggins, R., and Wu, S. (2011). Capture-recapture models with covariates: a partial likelihood approach. *Biometrics*, 67:1659–1665.

Sutherland, J., Schwarz, C., and Rivest, L. (2007). Multilist population estimation with incomplete and partial stratification. *Biometrics*, 63:910–916.

Tanaka, R. (1956). On differential response to life traps of marked and unmarked populations. *Annals of Zoology Japan*, 29:44–51.

Tanton, M. (1965). Problems of live-trapping and population estimation for the wood mouse (*Apodemus sylvaticus*). *Journal of Animal Ecology*, 34:1–22.

Tilling, K. and Sterne, J. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology*, 149(4):392–400.

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 Van Houwelingen, H. (2003b). Estimating the size of criminal population from police records using the truncated poisson regression model. *Statistica Neerlandica*, 57:1–16.

Van der Heijden, P., Whittaker, J., Cruyff, M., Bakker, B., and Van der Vliet, R. (2012). People born in the middle east but residing in the Netherlands: Invariant population size estimates and the tole of active and passive covariates. *The Annals of Applied Statistics*, 6(3):831–852.

Van der Heijden, P., Zwane, E., and Hessen, E. (2009). Structurally missing data problems in multiple list capture-recapture data. *AStA Advances in Statistical Analysis*, 93:5–21.

Wang, J. and Lindsay, B. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, 100:942–959.

Wang, J. and Lindsay, B. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology*, 5:30–45.

Wilson, K. and Collins, M. (1993). Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, 79:543–553.

Xi, L., Watson, R., Wang, J., and Yip, P. (2009). Estimation in capture-recapture models when covariates are subject to measurement errors and missing data. *Canadian Journal of Statistics*, 37:645–658.

Xi, L., Yip, P., and Watson, R. (2007). A unified likelihood-based approach for estimating population size in continuous-time capture-recapture experiments with frailty. *Biometrics*, 63:228–236.

Xu, Y., Fyfe, M., Walker, L., and Cowen, L. (2014). Estimating the number of injection drug users in greater victoria, canada using capture-recapture methods. *Harm Reduction Journal*, 11(9).

Yip, P., Huggings, R., and Lin, D. (1996). Inference for capture-recapture experiments in continuous time with variable capture rates. *Biometrika*, 83(2):477–483.

Yip, P., Lin, H., and Xi, L. (2005). A semiparametric method for estimating population size for capture-recapture experiment with random covariates in continuous time. *Biometrics*, 61:1085–1093.

Zelterman, D. (1988). Robust estimation in truncated discrete distributions with applications to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18:225–237.

Zwane, E. and Van der Heijden, P. (2004). Semiparametric models for capture-recapture studies with covariates. *Computational Statistics and Data Analysis*, 47:729–743.

Zwane, E. and Van der Heijden, P. (2007). Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. *Statistics in Medicine*, 26(5):1069–1089.

Zwane, E. and Van der Heijden, P. (2008). Capture-recapture studies with incomplete mixed categorical and continuous covariates. *Journal of Data Science*, 6:557–572.

Zwane, E. and van der Heijden, P. (2008). Capture-recapture studies with incomplete mixed categorical and continuous covariates. *Journal of Data Science*, 6:557–572.

Zwane, E., Van der Pal, K., and Van der Heijden, P. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, 23:2267–2281.