UNIVERSITY OF SOUTHAMPTON

Faculty of Social Science
School of Mathematical Sciences

Uni-List Capture-Recapture Approaches, with Uncertainty Quantification, Performance Analysis and a Meta-Analytic Application

by

Layna Charlie Dennett
ORCiD: 0000-0002-3357-3636

A thesis for the degree of
Doctor of Philosophy

October 2025

University of Southampton

Abstract

Faculty of Social Science
School of Mathematical Sciences

Doctor of Philosophy

Uni-List Capture-Recapture Approaches, with Uncertainty Quantification, Performance Analysis and a Meta-Analytic Application

by Layna Charlie Dennett

Meta-analysis is a powerful tool for evaluating numerous studies focused on the same or similar research question and integrating the results to identify a common parameter. This well-established methodology is prone to bias, so this thesis proposes the use of model-based meta-analytic and uni-list capture-recapture approaches, to compute more reliable estimates, with a focus on count data systematically missing zero counts.

For the meta-analytic approach, traditional methodologies do not adequately address zero-truncated count data. This thesis develops a model-based approach with zero-truncated count models which appropriately account for the missing zeroes and an exposure variable if applicable. From these models, a maximum likelihood approach is taken with the expectation-maximisation algorithm, used to compute less biased parameter estimates. Following these approaches, both observed and unobserved heterogeneity are addressed through covariate modelling and overdispersion modelling respectively.

As for the uni-list capture-recapture approach, the Horvitz-Thompson, generalised Chao's and generalised Zelterman's estimators are used for population size estimation, allowing for the inclusion of covariate information and an exposure variable. Also explored is the uncertainty that arises from these estimation methods, with both approximation-based variance estimation methods and the bootstrap algorithm addressed. Various approaches to the bootstrap algorithm and methods for accounting for model uncertainty are developed, in addition to alternative methods of confidence interval construction. The last focus of the thesis addresses the estimators under the presence of one-inflation, and given the poor performance of many of the existing estimators, the generalised-modified Chao's estimator is developed to account for zero-truncation, one-inflation and covariate information.

The methodologies discussed in this thesis are demonstrated through the use of real-life case study data, and assessed through a series of simulation studies.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. Parts of this work have been published as: Dennett and Böhning (2023); Dennett et al. (2025)

Signed:..... Date:.....

Acknowledgements

When I joined the University of Southampton in 2018 as an undergraduate, I'd have never expected that seven years later, I'd be where I am now. Undertaking this PhD has been life-changing and I am so very grateful to all those who supported me on this journey.

Firstly, I would like to thank my supervisor, Prof. Dankmar Böhning, without whom I would not have had the confidence to start this journey. For the past five years, throughout my undergraduate and masters dissertations, to this PhD, his continued support, guidance and immense wealth of knowledge has been invaluable. I would also like to thank Dr. Antony Overstall and the School of Mathematical Sciences graduate office team, for all their help over the years.

Lastly, I would like to thank my friends and family. Thank you to my dad Andrew, my Auntie Joy, my nan Betty, Masmi, Riyah, Nathan, Laura and lovely little Evelyn for their constant encouragement, it means the world. I am grateful to my partner, Charles, who has been there for me throughout this whole journey, through the highs and the lows and kept me sane. Most importantly, I would like to thank my mum, Julie. Thank you for your many sacrifices over the years, for always being there for me and for your unconditional love. I cannot put my gratitude for you into words, and without you, none of this would have been possible.

To my grandad, Charles Ernest Dennett, who motivates me to keep going and whom I miss dearly.

Definitions and Abbreviations

$E(\cdot)$	Mean/expected value of $(\cdot)$
$L_x(\cdot)$	Likelihood of x given $(\cdot)$
$\ell_x(\cdot)$	Log-likelihood of x given $(\cdot)$
$\text{Var}(\cdot)$	Variance of $(\cdot)$
$p_x(\cdot)$, $P(X = x\|(.))$	Probability of exactly x events, given $(\cdot)$
I	Indicator variable
i	Individual index
f_x	Frequency of exactly x events
η	Log-rate
M	Number of missing individuals
m	Largest observed count of events
μ	Expected count
N	Total number of individuals/(unknown) target population size
n	Number of observed individuals/effective sample size
ρ	Probability of success
τ	Exposure variable (person-years)
X	Count of events
$\boldsymbol{v}$	Vector of covariates

Chapter 1

Introduction

In this chapter, background of the case studies and their motivation are discussed, in addition to the aims and objectives of the thesis. At the end of the chapter, the structure of the thesis is outlined.

1.1 Introduction

When dealing with elusive populations, capture-recapture methods are widely used to estimate the total population size, important information which can be utilised in a plethora of ways dependent on the context of the data. For example, to accurately predict and monitor disease outbreaks through knowing the size of the population at risk in epidemiology, or for efficient planning of government funding and infrastructure through knowledge of population sizes by area. However, when data is systematically missing, challenges arise as the rates used in the capture-recapture methods will typically lead to under- or overestimation of the population sizes if the missing data is not appropriately accounted for. One method of addressing this missing data is model-based meta-analysis. Through utilising zero-truncated regression modelling, adjusted and more accurate rates can be computed in order to find reliable capture-recapture estimates of the population sizes.

This thesis works through applying these methods to a case study, looking at the prevalence of completed suicide following bariatric surgery, where studies without at least one completed suicide were excluded due to the search criteria and hence the number of zero completed suicides is missing. An additional case study focused on the captures of snowshoe hares is also used, providing an alternative dataset with different properties for demonstrative purposes. The final chapter of this thesis introduces the concept of one-inflation, and to demonstrate the one-inflated methods, a case study

exploring the prevalence of heroin drug users in the Chiang Mai province of Thailand is used. More information on this dataset is given in Chapter 7.

1.1.1 Background on the suicide data case study

To gain a deeper understanding of this capture-recapture approach to meta-analysis when zero-truncated count data is involved, a case study approach is adopted. The systematic review by Peterhänsel et al. (2013) forms the basis for this case study, containing meta-analytic data on 27 studies exploring the prevalence and risk of completed suicide after bariatric surgery.

Whilst the case study paper does take the missing zeroes into account, they have adopted a proportional model rather than a rate model for finding their estimates. Utilising a zero-truncated binomial distribution for finding the estimated rate of completed suicide after bariatric surgery, with constant probability of "success", $\exp(\eta)$, and the number of trials given by the person-years, τ_i. Whilst the case study found a reasonable estimated rate of completed suicide, and accounts for the missing studies, taking the person-years as the number of trials, and the rate as the success probability allows for a non-zero probability for observing more counts of completed suicide than people in the study. Given that an individual can complete suicide a maximum of one time, it is a probability that is impossible in practice. The alternative approaches discussed in this thesis prevent this unrealistic probability from occurring.

1.1.1.1 Motivation for the systematic review

The global the population is facing an obesity epidemic with, as of 2016, over 2 billion people being classified as either overweight or obese (Shekar and Popkin, 2020), with predictions that 17.5% of the world's population will be classified as obese by 2030 (Lobstein et al., 2022).

Figure 1.1 demonstrates this obesity epidemic, with a clear increasing trend in obesity rates for both men and women in England, from the years of 1993 to 2019. The percentage of women classified as obese is consistently larger than the percentage of men classified as obese. However, for both genders, and therefore the total population, the number of individuals classified as obese is growing each year, with a more rapid increase in more recent years.

As a result of growing obese populations, healthcare systems are experiencing increasing pressures to reduce the prevalence of obesity and the comorbidities of being overweight which are further burdening healthcare systems. These comorbidities, both physiological and psychological, include but are not limited to, type-II diabetes, cardiovascular disease, obesity related cancers, depression, and anxiety disorders (Dixon, 2010).

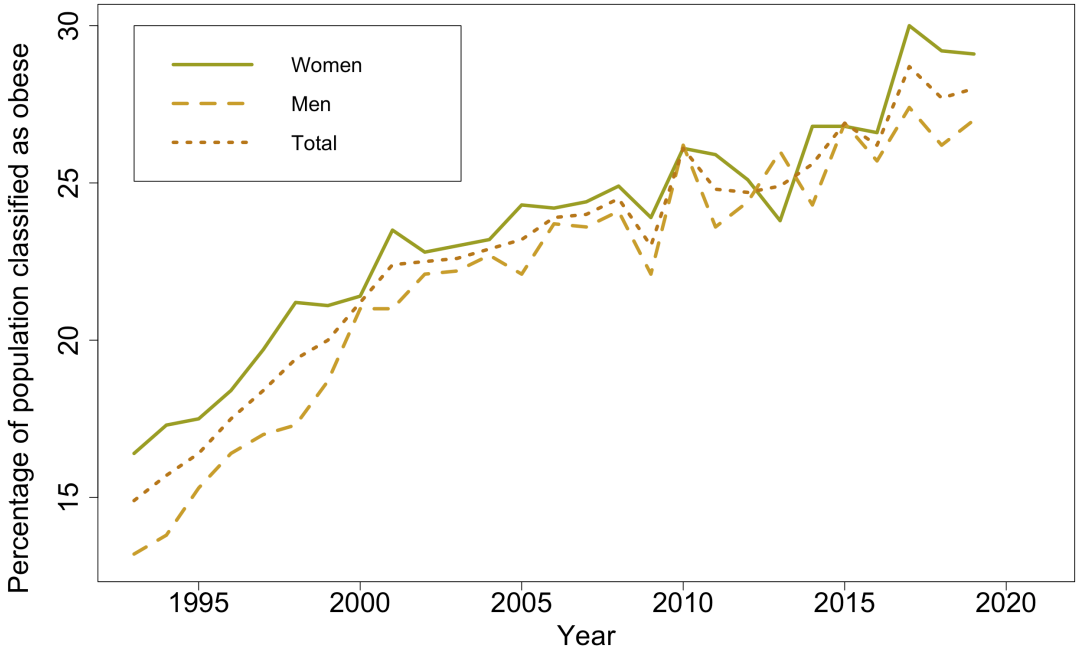

FIGURE 1.1: Percentage of the population in England classified as obese for years 1993 to 2019 for the total population and sub-populations of men and women, using data from NatCen Social Research, University College London (2005-2023).

In an attempt to reduce obesity, with its known success for long term weight loss, many individuals turn to bariatric surgery as a method for 'treating' obesity, where bariatric surgery is an umbrella term for weight loss surgeries. These medical procedures work by restricting food intake, reducing nutrient absorption and/or reducing production of the hormone ghrelin which regulates appetite (Khwaja and Bonanomi, 2010). Examples of these surgeries include gastric band, gastric bypass and sleeve gastrectomy. These surgeries are also used for the reduction of the severity of many obesity-related health conditions (Choban et al., 2002). Whilst an effective treatment, the risk of various long-term side effects mean that bariatric surgery is not a suitable option for everyone. To minimise the damaging consequences, certain measures can be taken by the health clinics performing the procedures. For example, appropriate screening of patients should always be conducted, highlighting potential patients who may be predisposed to mental health issues which increase the risk of suicide, as well as providing ample support afterwards, for those patients who do have the surgery.

Independent of bariatric surgery, completed suicide is one of the leading causes of death and is on the rise. According to the World Health Organisation, more than 700,000 individuals die by suicide annually, leading to an age-standardised rate of 9.0 per 100,000 population (World Health Organisation, 2021), with many more unsuccessful attempts. There are many reasons as to why people to turn to suicide as a solution to their problems, including financial struggles, job insecurity, health issues and family problems.

After bariatric surgery, the apparent risk of completed suicide increases comparative to those who have not undergone the surgery (Peterhänsel et al., 2013) as a result of additional possible triggers for mental health issues. Additionally, complications during and after surgery are relatively common, burdening the patient with further health problems and possible financial strain, particularly in countries such as the USA, where there is no universal healthcare. These complications can lead to the contemplation of suicide for many. The dramatic change in lifestyle required can also impact the mental health of the patients. Many individuals struggle after surgery with the restriction of their diet required, particularly with the lack of the typical comfort foods which lead many individuals to requiring bariatric surgery in the first place. The negative impact of these changes can increase the pressure on the patients mentally, and often leads to suicidal ideologies.

Globally, the age-standardised rate of completed suicide is 2.3 times higher for males (12.6 per 100,000 population) than for females (5.4 per 100,000 population) (World Health Organization, 2019; Nock et al., 2008). This variation is often as a result of the societal pressures placed upon men to be strong, both physically and mentally, worsened by suicide being perceived as taboo and the resulting lack of mental health support surrounding it (Nock et al., 2008). Consequently, investigating whether these rates are reflected within the population of individuals post-bariatric surgery is vital to aid in ensuring there is not the same neglect from general society upon men's well-being as there is in the pre- and post-surgery support. Investigating whether the proportion of women in each study, hence also the proportion of men, has a significant impact on the rate of completed suicide is a simple method to explore this. If the studies with a lower proportion of women have a significantly higher rate of completed suicide, there is reason to believe that the rate of completed suicide of men is higher after bariatric surgery also. However, whilst it is important to explore whether the suicide rate in men after bariatric surgery reflects that of the general population, it is also important that (if the proportion of women has a significant impact on the rate of completed suicide) it is explored whether the risk of suicide by gender after bariatric surgery is different from the baseline. If the rate of suicide for women after bariatric surgery is greater than that of men, it should be investigated further as to why there is this difference from the general population in order to reduce the rate.

Exploring whether the country of origin of a study has a significant impact upon the rate of completed suicide is of great importance as rates vary considerably between countries, from less than 5/100,000 person-years (including but not limited to Greece, Egypt, Brunei and Honduras) to over 30/100,000 person-years (Guyana) in 2016 (World Health Organization, 2019). Addressing the varying rates by country is important, particularly as they can be caused by cultural, economical and environmental differences. These include things like an increase in completed suicide during the polar night in parts of Sweden, Finland and Canada, where seasonal affective disorder is widespread, or in

the USA where financial issues particularly surrounding the lack of public healthcare and the opioid epidemic cause an increase in suicide. These differences in causes by country can impact upon the treatment received by patients of bariatric surgery, so any significant findings could provide an insight as to where additional support or changes are required, in order to reduce the rates. Possible interactions with gender should be explored, with the treatment of men and women varying by country and the rates of completed suicide potentially varying by gender.

1.1.1.2 Suicide data

Table 1.1 displays the 27 observed studies which are included in the systematic review from Peterhänsel et al. (2013), with Table 1.2 summarising the frequencies of the counts of completed suicides. Study 21, Kral 1993, reports its country of origin to be both USA and Sweden. For the purposes of modelling and inference, this is changed to be solely USA, given that it is both listed first and the mode country of origin. Also, study 24, Smith 2004, does not report a proportion of women. In order to compare models which include proportion of women as a covariate to models without, there are two possible options for dealing with this. Firstly, the study can simply be removed from the dataset. However, the more studies included in the dataset, the more accurate and reliable the model inferences are likely to be. Consequently, the second option, being to impute the missing proportion of women from the observed data, is the favourable option. A linear imputation model is fitted to the dataset using backwards stepwise model selection by Bayesian information criterion (BIC) (Davison, 2003), where the preferred model has main effects for person-years and the country of origin as well as their interaction. The resulting imputed value for the proportion of women is $v_{24,1} = 0.823$. It is worth noting that a sensitivity analysis was conducted for this imputation, investigating whether the imputed value affected the results compared to models without Study 24. It was found that the imputed value did not significantly change the results, and is beneficial to include Study 24 with its imputed value for proportion of women in the analysis given that the sample size is already small. Additionally, it is found that proportion of women as a covariate is not significant in the model, so this imputed value is not utilised, but the models can use all 27 observed studies for the modelling and estimation, rather than only 26. Both this change to the proportion of women and the change to country of origin for study 21 are given in Table 1.1, with the changes indicated by italics.

The main issue with the case study data, is the lack of studies which experience a total of zero completed suicides due to the search criteria for papers facing selection bias, requiring for at least one completed suicide to be observed in the study and Peterhänsel et al. (2013) noting that ”studies that reported suicide attempts or suicidal ideation by bariatric patients were not included.” If not taken into account, the missing data can lead to the rate of completed suicide after bariatric surgery being overestimated,

TABLE 1.1: Meta-analytic data from $n = 27$ observed studies from Peterhänsel et al. (2013) numbered and ordered by decreasing size of person-years, with corresponding number of total patients, proportion of women, country of origin of study and count of completed suicides. Study 24. Smith 2004 is missing the proportion of women but the unknown value is imputed to be 0.823. The country of origin for 21. Kral 1993 is reported as "USA/Sweden" but changed to USA for model fitting and analysis. Both changes seen distinguished by italics.

Study	Person-years	Number of total patients	Proportion of women	Country of origin	Count of completed suicide
1. Adams 2007	77,602	9,949	0.860	USA	21
2. Marceau 2007	10,388	1,423	0.720	Canada	6
3. Marsk 2010	8,877	1,216	0.000	Sweden	4
4. Pories 1995	8,316	594	0.832	USA	3
5. Carelli 2010	6,057	2,909	0.684	USA	1
6. Busetto 2007	4,598	821	0.753	Italy	1
7. Smith 1995	3,882	1,762	0.889	USA	2
8. Peeters 2007	3,478	966	0.770	Australia	1
9. Christou 2006	2,599	228	0.820	Canada	2
10. Günther 2006	2,244	98	0.837	Germany	1
11. Capella 1996	2,237	888	0.822	USA	3
12. Suter 2011	2,152	379	0.744	Switzerland	3
13. Suter 2006	1,639	311	0.865	Switzerland	1
14. Van de Weijgert 1999	1634	200	0.870	Netherlands	1
15. Cadière 2011	1,362	470	0.834	Belgium	1
16. Mitchell 2001	1,121	85	0.847	USA	1
17. Himpens 2011	1,066	82	0.902	Belgium	1
18. Näslund 1994	799	85	0.812	Sweden	2
19. Forsell 1999	761	326	0.761	Sweden	1
20. Powers 1997	747	131	0.847	USA	1
21. Kral 1993	477	69	0.812	*USA*	1
22. Näslund 1995	457	142	0.592	Sweden	1
23. Powers 1992	395	100	0.850	USA	1
24. Smith 2004	354	779	*0.823*	USA	1
25. Nocca 2008	228	133	0.677	France	1
26. Svenheden 1997	166	91	0.791	Sweden	1
27. Pekkarinen 1994	146	27	0.704	Finland	1

TABLE 1.2: Frequency table for number of suicides from Peterhänsel et al. (2013) where frequencies for number of suicides between 7 and 20 are zero.

Count of completed suicides, x	0	1	2	3	4	5	6	7,$\cdots$,20	21	22+
Frequency, f_x	-	18	3	3	1	0	1	0	1	0

increasing the apparent risk which may impact upon both the medical professionals and the patients themselves.

If the risk of completed suicide is perceived to be higher, it is likely that many prospective patients will be unnecessarily rejected from having the bariatric surgery, potentially detrimentally affecting their lives and mental state, resulting in further damage. Additionally, the perceived higher risk means that the estimated rate of completed suicide is higher than the true value. If this is the case, it can be assumed that many medical centres unknowingly have an average rate of completed suicide after bariatric surgery higher than the true (average) rate of completed suicide, but under the estimated rate of completed suicide when the missing zero counts are ignored. With this, it would appear that they are below average, when they are not in fact, and should be working to reduce their rates further. Consequently, there is the risk of these medical centres unknowingly neglecting to perform the required pre-screenings, and providing post surgery support to an adequate level. However, adjusting the estimated rate of completed suicide for overcount using capture-recapture methods would allow for these centres to be brought to attention and possibly reduce their rates through exploration of what causes the increase, saving lives as a result.

1.1.2 Background on the snowshoe hares case study data

Snowshoe hares (*Lepus americanus*) are a species of nocturnal hare that reside in North America, characterised by their large hind feet, reminiscent of snowshoes, allowing them to hop and walk on the snow in their climate without sinking. Snowshoe hares have adapted to their climate through camouflage, having white coats to blend in with the snow in winter and a rusty brown coat in summer for when the snow clears.

Runways are trails (or routes) that snowshoe hares create and use all year round, meticulously clearing any stems and leaves that block the trail as they grow.

Live trapping is the process of trapping or capturing live animals without injuring them, and is often an important method of sourcing information on populations, such as the structure, size and distribution of the animal populations.

1.1.2.1 Hares data

Keith and Meslow (1968) provide live trapping data on snowshoe hares between the years of 1962 and 1967 with the number of captures and recaptures included from three different seasons, (mid)winter, spring and summer, and six different study areas. The distribution of total captures of the adult snowshoe hares is given in Table 1.3 with the overall frequency distribution of captures of snowshoe hares given in Table 1.4.

TABLE 1.3: Distribution of total captures of adult snowshoe hares within different yearly trapping periods between 1962 and 1967 and different study areas.

	Square mile study area			Five small study areas		
f_x	Midwinter	Spring	Summer	Midwinter	Spring	Summer
0	-	-	-	-	-	-
1	72	109	184	53	67	168
2	19	45	55	23	26	42
3	2	19	14	6	18	16
4	1	5	4	10	7	1
5	1	3	4	2	4	0
6	0	0	0	0	3	0
7+	0	0	0	0	0	0
Total	95	181	261	94	125	227

TABLE 1.4: Frequency table of the captures of snowshoe hares in Keith and Meslow (1968).

Count of captures, x	0	1	2	3	4	5	6	7+
Frequency, f_x	-	653	210	75	28	14	3	0

1.2 Aims and objectives

The main aim of this thesis is to address the existence of systematically missing studies in count data through developing meta-analytic models and capture-recapture estimators which appropriately account for the zero-truncation of the data. In order to achieve this, the following objectives are required.

1. Examine the count data and its origin to investigate whether the data is in fact zero-truncated.
2. Motivate the study through exploring the use of traditional meta-analytic approaches such as the inverse-variance method and regression modelling and demonstrate their poor performance in the case of zero-truncated count data.
3. Develop zero-truncated count models and investigate the covariate effects, addressing observed heterogeneity, as well as unobserved heterogeneity through overdispersion modelling with the negative-binomial model.
4. Explore the comparative suitability of the models using information criterion and assess goodness-of-fit with fitted frequencies and ratio plots.
5. Demonstrate the unsuitability of the zero-truncated binomial distribution model proposed by the case study through a simulation study.
6. Develop an Expectation-Maximisation algorithm that accounts for exposure and zero-truncation.

7. Develop population size estimators which account for exposure, zero-truncation and covariates, and allow both whole and sub-population sizes to be computed.
8. Explore the performance of each capture-recapture population size estimator via a simulation study.
9. Investigate uncertainty by developing approximation-based variance estimation methods for each of the estimators used for estimating the event rate and population size.
10. Develop non-parametric, semi-parametric and parametric bootstrap algorithms which account for model uncertainty in different ways, and use percentile confidence intervals to investigate the uncertainty.
11. Explore alternative methods of constructing confidence intervals to correct for bias resulting from the bootstrap algorithms and compare the performance of the different methods.
12. Demonstrate the impact of the existence of one-inflation in zero-truncated count data on the estimation methods, and develop an estimator which accounts for zero-truncation, one-inflation and covariate information to deal with this.
13. Explore the performance of this novel population size estimator for zero-truncated and one-inflated and compare it to the existing estimators via a simulation study.

1.3 Thesis outline

This thesis comprises eight chapters, with the first chapter introducing the research, providing background on the case studies, and outlining the aims and objectives of the thesis.

A literature review of the principal methodologies that form the foundations of this research is provided in Chapter 2. In particular, the background, traditional approaches and their limitations are discussed for meta-analysis, capture-recapture and regression modelling. Additionally, some applications to case study data are included for the different approaches.

In Chapter 3, the construction of zero-truncated count distributions is detailed, with model evaluation and comparison methods discussed. Count distributions, including the Poisson, negative-binomial, geometric and binomial distributions, are given with applications to the case study datasets to illustrate the methods given in the chapter. A simulation study is conducted in the final section of this chapter for demonstrative purposes, illustrating that the binomial distribution is not an appropriate model in the given circumstances.

Chapter 4 explores various estimation methods, with the Expectation-Maximisation algorithm developed for the zero-truncated Poisson and geometric models to find the maximum likelihood estimate of the rate parameter. Additionally, capture-recapture population size estimators are explored, specifically the Horvitz-Thompson, Chao's and Zelterman's estimators, with generalised versions of Chao's and Zelterman's estimators to account for covariate information and exposure. Further development of the estimators is included to allow for the assumption of a geometric mixture kernel, instead of the default Poisson mixture kernel. Each of the estimation methods discussed are implemented using case study data, with both whole and sub-population sizes and rates calculated for the Horvitz-Thompson estimator, allowing for a more in depth analysis. A simulation study is utilised in this chapter to explore the comparative performance of the capture-recapture estimators in different data scenarios for understanding which estimator should be used for best and most reliable results.

Uncertainty is unavoidable when estimating parameters and population sizes. There are two main approaches to quantifying this uncertainty available and discussed in Chapters 5 and 6. Firstly, the (normal) approximation-based variance approach developed in Chapter 5 is used for a Wald-type interval for the rate parameter estimation, and the use of conditioning for the population size estimators.

Chapter 6 develops approaches to uncertainty quantification that utilise resampling in order to estimate the variance and confidence intervals. Non-parametric, semi-parametric and parametric bootstrap algorithms are developed for the estimated rate parameter and each of the capture-recapture population size estimators with sub-population specific confidence intervals provided for the rate estimates and Horvitz-Thompson estimates. Within each bootstrap approach, three different methods of accounting for model uncertainty are developed, exploring various ways of fitting models and resampling in order to quantify the uncertainty within the estimation methods. Additionally, alternative confidence interval construction methods are explored in Chapter 6 given that the standard percentile confidence interval construction method is prone to bias and skewness. Bias-corrected percentile intervals, bias-corrected and accelerated percentile intervals and median absolute deviation confidence intervals are developed to reduce this bias and skewness, and increase the reliability of the resulting conclusions. Finally, this chapter includes a series of simulation studies, investigating the performance, namely the precision, coverage, and robustness of the bootstrap algorithms and the different confidence interval construction methods.

Various methods covered in this thesis are reintroduced in Chapter 7, but with one-inflation present in the dataset. For this, a new dataset with excess singletons present is introduced and utilised to demonstrate the ineffectiveness of the methods when the data is one-inflated. To circumvent the issues that arise from using these methods, the modified Chao's estimator is explored, with a generalised-modified Chao's estimator developed to allow for the inclusion and accountability for zero-truncation,

covariate information and one-inflation, with the development of the corresponding approximation-based variance estimation methods. This chapter ends with a simulation study that compares the performance of various capture-recapture estimators included in this work, under the presence of one-inflation.

Lastly, Chapter 8 provides concluding remarks and potential work for future research.

Chapter 2

Overview of Key Methodologies

This chapter provides background information on meta-analysis, regression modelling and capture-recapture. Section 2.1 describes the methods traditionally used for meta-analysis and the corresponding limitations of these methods which make them unsuitable when data is zero-truncated and certain assumptions are not met. As a model-based meta-analysis approach is taken, regression modelling is discussed in Section 2.2, specifically regarding count data with covariate information available to account for observed heterogeneity and the exposure variable. The Poisson, negative-binomial, geometric and binomial distributions are provided given the type of data, followed by the application of these models by example of the case studies. Additionally, the limitations and problems associated with this modelling are provided. To end, Section 2.3 explores the history and background of capture-recapture approaches and the required assumptions. The different data structure types are also explained, and a brief description of the estimators which are applicable for estimating the population size of the zero-truncated count data is included.

2.1 Meta-analysis

Meta-analysis (see Borenstein et al., 2021, for more information) can be simply described as a methodology for evaluating multiple independent studies with a focus on the same or similar research question, combining the results in order to find an overall statistic and trend. The resulting overall statistic can be described as a weighted average, with larger studies possessing more influence for the statistic than studies with smaller sample sizes. The process for meta-analysis can be described in 5 steps:

1. Identify the research question of interest and propose a hypothesis.
2. Compose the systematic review. It is crucial to include quality studies which are relevant to the research question of interest and to consider a range of studies

in order to reduce the risk of selection bias. Additionally, it is important not to automatically disregard any studies which show negative findings in order to reduce the risk of bias. Composing a systematic review can be done by the following steps:

(a) Develop the protocol which will state the objectives and methods of the research.

(b) Conduct a search for any literature which will be applicable to the research question through titles and abstracts.

(c) Narrow the list of possible literature down and select the studies which are appropriate through reviewing the full texts of the studies.

(d) Assess the quality of the chosen studies.

3. From the chosen studies in the systematic review, extract the appropriate data and compute summary measures of each study, typically the effect sizes.
4. Analyse the extracted data by compiling the individual effects and creating a pooled, weighted estimate for the effect of interest from the research question.
5. Interpret the results from analysis.

The inverse-variance method is the conventional approach for fixed effects meta-analysis, where the study-specific weighted averages are chosen to be the inverses of the variance for each estimate (Borenstein et al., 2021; Egger et al., 2008; Cooper et al., 2019; Stangl and Berry, 2000). Studies with a larger population size typically have smaller standard errors and therefore larger inverse-variances. As a result of this, the larger studies are given more weight in the summary measure.

The weighted average summary measure on the log-scale is then given as

$$\frac{\sum_{i=1}^{n} w_i \log\left(\frac{X_i}{\tau_i}\right)}{\sum_{i=1}^{n} w_i},$$

where w_i are study-specific weights, X_i are study-specific counts and τ_i are study-specific exposure variables.

Using the Poisson assumption $\text{Var}(X_i) = E(X_i)$, the expected value is then estimated as X_i. The study specific inverse-variances of the log-rates can be used as weights and calculated as

$$\frac{1}{\text{Var}\left(\log\left(\frac{X_i}{\tau_i}\right)\right)} = \frac{1}{\text{Var}(\log(X_i))} \approx \frac{E(X_i)^2}{\text{Var}(X_i)} = E(X_i) \approx X_i,$$

leading to the summary measure on the log scale

$$\hat{\eta} = \frac{\sum_{i=1}^{n} X_i \log\left(\frac{X_i}{\tau_i}\right)}{\sum_{i=1}^{n} X_i}.$$

Alternatively, an approach on the rate scale can be used, still following the Poisson assumption above (Barendregt et al., 2013). Assuming a homogeneous rate, $\exp(\eta)$, for each of the independent studies to obtain the inverse-variance for each study, the inverse-variances are equal to the corresponding weights, calculated as

$$w_i = \frac{1}{\text{Var}\left(\frac{X_i}{\tau_i}\right)} = \frac{\tau_i^2}{\text{Var}(X_i)} = \frac{\tau_i^2}{\exp(\eta)\tau_i} = \frac{\tau_i}{\exp(\eta)},$$

such that the summary measure on the rate scale is

$$\exp(\hat{\eta}) = \frac{\sum_{i=1}^{n} w_i \frac{X_i}{\tau_i}}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} \tau_i}.$$

These common approaches have their issues, including the reliance on asymptotic normality and the Poisson assumption seen in the variance calculations. However, the most prominent issue arises in the case of taking zero counts into account.

For the suicide case study data, the two approaches respectively result in rates of completed suicide per 100,000 person-years of 45 and 60.

In the case of rare events, meta-analysis can be particularly useful as it is common for studies with rare outcomes to be under-powered. Using meta-analysis to collaborate the studies means a smaller sample is needed, leading to both a reduction of costs of the studies themselves and a pooled statistic with greater overall power comparative to the individual studies. Options for meta-analysis methods are limited for situations with rare outcomes due to the reliance on large sample sizes for approximations for many methods. In order to avoid the computational errors and use the methods available, correction methods can be utilised, but often results in unwanted bias and misleading conclusions.

Meta-analysis and systematic reviews are prone to selection bias (e.g. Kulinskaya et al., 2008, Chapter 15) with the exclusion of studies with no result or zero counts common practice, although this exclusion removes the need for correction methods. Through excluding studies however, problems involving missing data arise with it assumed the given data is complete for meta-analysis. There are 4 main methods for dealing with missing data as follows:

1. Ignore the missing data and analyse only that which is available. This is common for systematic reviews and typically results in inaccurate estimates.

2. Use the observed data to impute the missing data. Another common approach, but results in confidence intervals which are narrow as a result of not considering uncertainty.
3. Accounting for uncertainty in the imputation of the missing data. This is less used due to increased difficulty, but provides more reliable estimates.
4. Use the available data to make assumptions about the relationship of interest and use regression modelling to account for the missing data.

Whilst it is common for systematic reviews to analyse only the observed studies, if the missing studies are those which have zero counts, then the estimates are likely to be grossly overestimated. These misleading results can have detrimental impacts, particularly in medical situations with patients lives at risk, highlighting the importance on ensuring the estimates are as accurate as possible. Whilst traditional modelling provides more accurate estimates, it is not applicable for when the data is zero-truncated as the results will face the same issue of overestimation.

Assuming constant risk of outcome among the different studies is also cause for problems to arise. In practice, this assumption is often not met for an array of reasons, including the unpredictability of human behaviour and health, difference in demographics and difference in trials. The statistical methods available for rate data are also less developed comparable to other data types, leaving limited methods for analysis if problems arise. Consequently, it is not always appropriate to use traditional meta-analytic methods for rate data and alternative methods need to be used. Poisson, negative-binomial and geometric regression can be used for count and rate data, allowing for the consideration of varying time at risk for individuals for each study through the addition of an exposure variable.

2.2 Regression modelling

A generalised linear model (see Dobson and Barnett, 2018, for more information) can be described as a regression model comprised of a random component, a systematic component and a link function. For the index $i = 1, 2, \cdots, n$, where n is the total number of observed individuals, the random component is the response variable, μ_i. Additionally, the linear predictor, $\eta_i = \mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta}$, is the systematic component, connected to the random component by the link function $g(\mu_i)$. Here, $\mathbf{h}(\cdot)$ is a regression function for implementing the different covariates $\mathbf{v}_i$ and the associated parameter vector, $\boldsymbol{\beta}$. Within regression models, covariates are measured, uncontrolled variables which can be included to increase the accuracy of the response variable. Whilst typically not of direct interest in the study, in cases where there is unexplained variance, inclusion of information from covariates in the regression model can reduce both the error and

the variance of the results. For the suicide case study data, the covariates that will be considered are the proportion of women, and the country of origin of the study. Covariate information for age is available in the dataset but it of poor quality, and is therefore not included in the regression modelling. For completeness, analysis of this covariate information is included in Appendix A.1. For the hares case study data, the covariates that will be considered are the season in which the captures took place and the type of study area.

Dependent on the situation and model, different link functions exist. However, as count data is the focus, there is the requirement to be non-negative, so the link function chosen is necessitated to ensure that the fitted values are non-negative. Therefore, there are two commonly used options, either to use a square root link function as follows

$$g(\mu_i) = \sqrt{\mu_i} = \sqrt{\tau_i} \times \mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta},$$

where τ_i is the exposure variable, or more commonly, to use a log-link function, which can be seen in Equation 2.1 and will be used going forward.

$$g(\mu_i) = \log(\mu_i) = \log(\tau_i) + \mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta}, \tag{2.1}$$

where τ_i is the exposure variable. Equivalently, Equation 2.1 can be written as:

$$\mu_i = \tau_i \exp(\eta_i),$$

where $\eta_i = \mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta}$.

A log-link function ensures that the fitted values are non-negative, which is a necessity for count data, and is the link function used in this work.

2.2.1 Distributions

2.2.1.1 Poisson

The Poisson distribution is commonly used for count data due to being integer-valued. Additionally, it can be used as an approximation of the binomial distribution when there is a large number of observations and rare events (small probability), or as a special case of the negative-binomial distribution when the number of successes is large. (See Appendix A.2)

The density of the standard Poisson distribution is

$$p_x(\mu_i) = P(X = x|\mu_i) = \exp(-\mu_i)\frac{\mu_i^x}{x!}, \tag{2.2}$$

where $\mu_i = \tau_i \exp(\eta_i)$ for $i = 1, 2, \cdots, n$.

2.2.1.2 Negative-binomial

The negative-binomial distribution, in this case derived from the Poisson-gamma mixture, is often used as an alternative to the Poisson distribution. This is often the case when the Poisson assumption is not met, meaning that the mean is significantly greater or smaller than the variance, known as under- or overdispersion (Xekalaki, 2014). The negative-binomial distribution addresses the under- or overdispersion with the addition of a dispersion parameter θ, allowing for more flexibility with the parameters. The density of the standard negative-binomial distribution is

$$p_x(\mu_i, \theta) = P(X = x|\mu_i, \theta) = \frac{\Gamma(\theta + x)}{\Gamma(x+1)\Gamma(\theta)} \left(\frac{\mu_i}{\mu_i + \theta}\right)^x \left(\frac{\theta}{\mu_i + \theta}\right)^\theta,$$

where $\mu_i = \tau_i \exp(\eta_i)$.

2.2.1.3 Geometric

The geometric distribution is a special case of the negative-binomial distribution when $\theta = 1$, comprised of the number of x Bernoulli trials required for the first success to occur. The density of the standard geometric distribution is

$$p_x(\mu_i) = P(X = x|\mu_i) = \left(1 - \frac{1}{1+\mu_i}\right)^x \frac{1}{1+\mu_i},$$

where $\mu_i = \tau_i \exp(\eta_i)$.

2.2.1.4 Binomial

When there are only two possible outcomes, success or failure, with respective probabilities ρ_i and $1 - \rho_i$, the binomial distribution describes the probability of x successes (events), for a given number of Bernoulli trials. For the standard binomial distribution, in this application the number of Bernoulli trials is given by the population size. However, to account for the varying exposure in the suicide case study, the number of Bernoulli trials is given by the exposure variable, τ_i, leading to a pseudo-binomial distribution.

This pseudo-binomial distribution has density

$$p_x(\rho_i, \tau_i) = P(X = x|\mu_i, \tau_i) = \binom{\tau_i}{x} \rho_i^x (1 - \rho_i)^{\tau_i - x},$$

where $\rho_i = [1 + \exp(-\eta_i)]^{-1}$.

In the suicide case study, Peterhänsel et al. (2013) assume a constant success probability ρ, not accounting for each of the sub-populations, but we extend the binomial approach to account for covariates, giving a variable success probability ρ_i for each of the studies.

Application: Suicide data

For the case study data, the expected response, $\exp(\eta_i)$, is rate of completed suicide, the response variable, μ_i, is expected count of completed suicide and the exposure variable, τ_i, is the number of person-years, for each study $i = 1, 2, \cdots, 27$. Additionally, $\mathbf{v} = (v_{i1}, v_{i2})^T$ is the vector of covariates considered where v_{i1} is the proportion of women and v_{i2} is an indicator variable for the country of origin, where

$$v_{i2} = \begin{cases} 1 & \text{if country of origin is USA,} \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \cdots, 27$.

The different linear predictors under consideration for this data are seen in Table 2.1.

TABLE 2.1: Linear predictors under consideration with corresponding regression functions for the suicide data.

Linear predictor j	Proportion of women v_1	Country of origin v_2	Interaction v_1v_2	Regression function $\mathbf{h}_j(\mathbf{v})$
1	No	No	No	$\mathbf{h}_1(\mathbf{v}) = 1$
2	Yes	No	No	$\mathbf{h}_2(\mathbf{v}) = (1, v_1)^T$
3	No	Yes	No	$\mathbf{h}_3(\mathbf{v}) = (1, v_2)^T$
4	Yes	Yes	No	$\mathbf{h}_4(\mathbf{v}) = (1, v_1, v_2)^T$
5	Yes	Yes	Yes	$\mathbf{h}_5(\mathbf{v}) = (1, v_1, v_2, v_1v_2)^T$

Results from five generalised linear models assuming a Poisson distribution, using each of the linear predictors in Table 2.1 are seen in Table 2.2. The same models assuming a negative-binomial distribution can be computed, but the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), discussed in Section 3.2, do not improve upon the Poisson models providing no evidence of overdispersion. Therefore, the negative-binomial models approximate the Poisson with an additional parameter to penalise. The linear predictor which minimises both the AIC and BIC is linear predictor 3, suggesting that the country of origin of the study influences the rate of completed suicide. The sub-populations of USA and Other have corresponding estimates of the rate of completed suicide of 34.6 and 68.1 respectively (both per 100,000 person-years), showing a notable difference when the country of origin is USA, having a rate of almost half of the other countries collectively. Whilst not the best fitting model, the intercept-only model with linear predictor 1 produces an estimated rate of completed suicide

of 47.9 per 100,000 person-years which is comparable to that following the method proposed by Barendregt et al. (2013) using standard meta-analysis.

TABLE 2.2: Values for the suicide data of AIC and BIC for each linear predictor used in a Poisson generalised linear model, with the minimum AIC and BIC values and corresponding linear predictor in bold.

Linear predictor	AIC	BIC
1	104.8	106.1
2	106.5	109.1
3	**99.79**	**102.4**
4	101.0	104.9
5	103.0	108.2

However, these models fitted do not take the missing studies with zero counts into consideration, leading to an overestimation of the rate of completed suicide. This leads to the requirement for alternative approaches to be considered. One such approach is capture-recapture, allowing for the excluded studies to be considered, with zero-truncated modelling producing more accurate estimates. Additionally, this method enables the computation of population and sub-population size estimation, hence the number of unobserved studies can be estimated.

Application: Hares data

For the hares case study data, the expected response, $\exp(\eta_i)$, is the rate of capture and $\boldsymbol{v} = (v_{i1}, v_{i2})^T$ is the vector of covariates considered where v_{i1} is a categorical variable for the season with levels

$$v_{i1} = \begin{cases} 1 & \text{captured in Midwinter,} \\ 2 & \text{captured in Spring,} \\ 3 & \text{captured in Summer,} \end{cases}$$

and v_{i2} is a dummy variable for the study area size, where

$$v_{i2} = \begin{cases} 1 & \text{if five smaller study areas,} \\ 0 & \text{if study area is one square mile,} \end{cases}$$

for $i = 1, 2, \cdots, 983$. For the hares case study data, there is no exposure variable, τ_i. The different linear predictors under consideration for this data are seen in Table 2.1.

Given that the data is count data, Poisson, negative-binomial and geometric distributions are under consideration. Covariate effects are accounted for though testing models with each of the linear predictors given in Table 2.3 for each of the given distributions.

TABLE 2.3: Linear predictors under consideration with corresponding regression functions for the hares data.

Linear predictor j	Season v_1	Study area v_2	Interaction v_1v_2	Regression function $\mathbf{h}_j(\mathbf{v})$
1	No	No	No	$\mathbf{h}_1(\mathbf{v}) = 1$
2	Yes	No	No	$\mathbf{h}_2(\mathbf{v}) = (1, v_1)^T$
3	No	Yes	No	$\mathbf{h}_3(\mathbf{v}) = (1, v_2)^T$
4	Yes	Yes	No	$\mathbf{h}_4(\mathbf{v}) = (1, v_1, v_2)^T$
5	Yes	Yes	Yes	$\mathbf{h}_5(\mathbf{v}) = (1, v_1, v_2, v_1v_2)^T$

As with the suicide case study data, fitting these linear predictors with standard regression modelling won't take the missing zero counts into consideration and therefore other approaches are required.

2.3 Capture-recapture

2.3.1 Introduction to capture-recapture

Traditionally used for elusive wildlife populations in ecology, the capture-recapture approach was developed as a better, alternative method to the census approach, which does not accurately measure the total target population size if a complete census cannot be obtained. The applications of capture-recapture methods are evolving, with it being used in numerous fields including epidemiology for estimating prevalence of diseases, social science for estimating the size of drug user populations and criminology for estimating the number of people violating laws among other applications (see Böhning et al., 2018, for further examples).

The general idea of capture-recapture is that on one occasion a sample of a population is captured, the individuals are counted and identified, then released to reintegrate with the population as a whole. On a second occasion, another sample is taken, wherein the individuals captured are counted and identified again, and individuals who have previously being identified are noted. In the basic example, there are two trapping occasions, but can be repeated for m occasions. The number of times individuals are identified at each occasion, f_x, where $x = 1, \cdots, m$, form the observed frequencies (and the unobserved frequency of zero) seen in Table 2.4.

TABLE 2.4: Frequency distribution.

x	0	1	2	3	...	m
f_x	f_0	f_1	f_2	f_3	...	f_m

The total number of times individuals are observed is

$$X = \sum_{x=0}^{m} x f_x,$$

where f_0 is unobserved and hence requires estimation. Also using the frequencies in Table 2.4, the target population size is found through summing each of the frequencies as

$$\begin{aligned} N &= \underbrace{f_0}_{\text{unobserved}} + \underbrace{f_1 + f_2 + f_3 + \cdots + f_m}_{\text{observed}}, \\ &= f_0 + n, \end{aligned} \tag{2.3}$$

where the observed population size is $\sum_{x=1}^{m} f_x = n$.

2.3.2 Assumptions

- Closed system: the target population remains constant, there are no births or deaths.
- Independence between captures: the captures in the different samples do not affect one another.
- Independence between subjects: there is no dependence between subjects, the capture of one subject does not impact or is impacted by another subject.
- Homogeneity of capture probability: each subject has equal probability of being captured.
- Subjects are correctly identified and recorded.
- Once identified, subjects do not lose their record of identification.

2.3.3 Data structure

There are two main types of capture-recapture data, one where there are repeated captures (multiple mark) and another where there are different data sources which can overlap (different sources).

Multiple mark

Table 2.5 displays the structure of data obtained by repeated captures. For occasions $x = 1, 2, \cdots, m$ and individuals $i = 1, 2, \cdots, N$, I_{ix} is an indicator variable which identifies whether individual i was observed on occasion x, where

$$I_{ix} = \begin{cases} 1 & \text{individual } i \text{ was observed on Occasion } x, \\ 0 & \text{otherwise.} \end{cases}$$

TABLE 2.5: Capture-recapture multiple-mark data structure for repeated captures.

	Occasion, x					
Individual, i	1	2	3	...	m	Total
1	$I_{1,1}$	$I_{1,2}$	$I_{1,3}$	...	$I_{1,m}$	X_1
2	$I_{2,1}$	$I_{2,2}$	$I_{2,3}$	...	$I_{2m,}$	X_2
3	$I_{3,1}$	$I_{3,2}$	$I_{3,3}$	...	$I_{3,m}$	X_3
$\vdots$			$\vdots$			$\vdots$
n	$I_{n,1}$	$I_{n,2}$	$I_{n,3}$	...	$I_{n,m}$	X_n
$n+1$	$I_{n+1,1}$	$I_{n+1,2}$	$I_{n+1,3}$	...	$I_{n+1,m}$	X_{n+1}
$\vdots$			$\vdots$			$\vdots$
N	$I_{N,1}$	$I_{N,2}$	$I_{N,3}$	...	$I_{N,m}$	X_N

For each individual, the sum of the indicator variables is the total number of times each individual, i, is identified, also known as the marginal frequency count, $X_i = \sum_{x=1}^{m} I_{ix}$. Given that individuals $i = 1, 2, \cdots, n$ are observed and $i = n+1, \cdots, N$ are unobserved, $I_{ix} = X_i = 0$ for $i = n+1, \cdots, N$.

Different sources

Different trapping occasions that may overlap can be treated as different sources of the data. In this structure, the composition of the data can get more complex the more occasions/sources that exist. The simplest structure of this kind is that of the Lincoln-Petersen (Alpizar-Jara and Pollock, 1996), where there are only two occasions for sourcing data. Over the two occasions, individuals are categorised as identified on both occasions, only one occasion (specifying which one) or neither occasion. Those who were not identified on either occasion are not observed cannot be counted so the total population size is found through dual systems estimation.

TABLE 2.6: Contingency table for two occasions.

		Occasion 2		
		1	0	
Occasion 1	1	f_{11}	f_{10}	n_1
	0	f_{01}	f_{00}	
		n_2		N

Figure 2.1 illustrates the union of these two occasions, with the same information provided as a contingency table in Table 2.6, where

- f_{00} denotes the frequency of unobserved individuals.
- f_{10} denotes the frequency of individuals identified only once and at occasion 1.
- f_{01} denotes the frequency of individuals identified only once and at occasion 2.
- f_{11} denotes the frequency of individuals identified at both occasions.

- n_x denotes the total number of individuals identified at occasion x, where $x = 1, 2$.

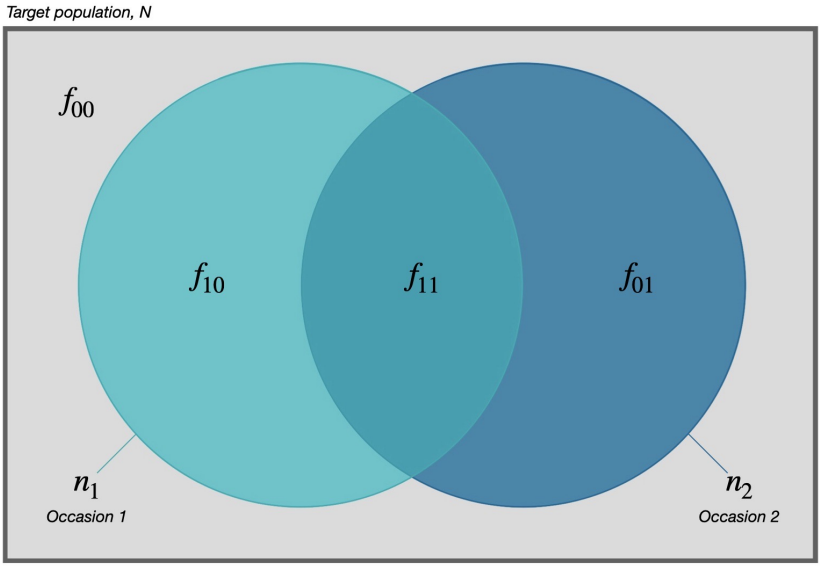

FIGURE 2.1: Venn diagram for two sources of data, Occasion 1 and Occasion 2, with frequencies of identification and total population sizes.

The target population size, N, is given as $N = f_{00} + f_{10} + f_{01} + f_{11}$. However, as f_{00} is unobserved, the total population size requires estimation. Assuming independence between occasions, the frequency of individuals unobserved during both occasions can be estimated given that the odds ratio is approximately equal to one (Brittain and Böhning, 2009). Therefore,

$$\frac{f_{11}f_{00}}{f_{10}f_{01}} \approx 1,$$

so

$$\hat{f}_{00} = \frac{f_{10}f_{01}}{f_{11}},$$

which can be used to estimate the target population size, leading to the Lincoln-Petersen estimator as follows.

$$\widehat{N}^{(LP)} = \frac{n_1 n_2}{f_{11}}. \tag{2.4}$$

Equivalently, this estimator can be derived using the assumption of independence between individuals, leading to Occasion 1, $\frac{n_1}{N}$, equalling the proportion of individuals identified at Occasion 2, $\frac{f_{11}}{n_2}$. However, the estimator also requires the assumption that the occasions have no overlap, which is not always the case. If no overlap exists, $f_{11} = 0$, a modification to the Lincoln Petersen estimator developed by Chapman (1951) and shown by Wittes (1972) to have reduced bias can be used to estimate the population size (Brittain and Böhning, 2009) can be used. Using Chapman's estimator, the unobserved frequency is

$$\hat{f}_{00} = \frac{f_{10}f_{01}}{f_{11} + 1},$$

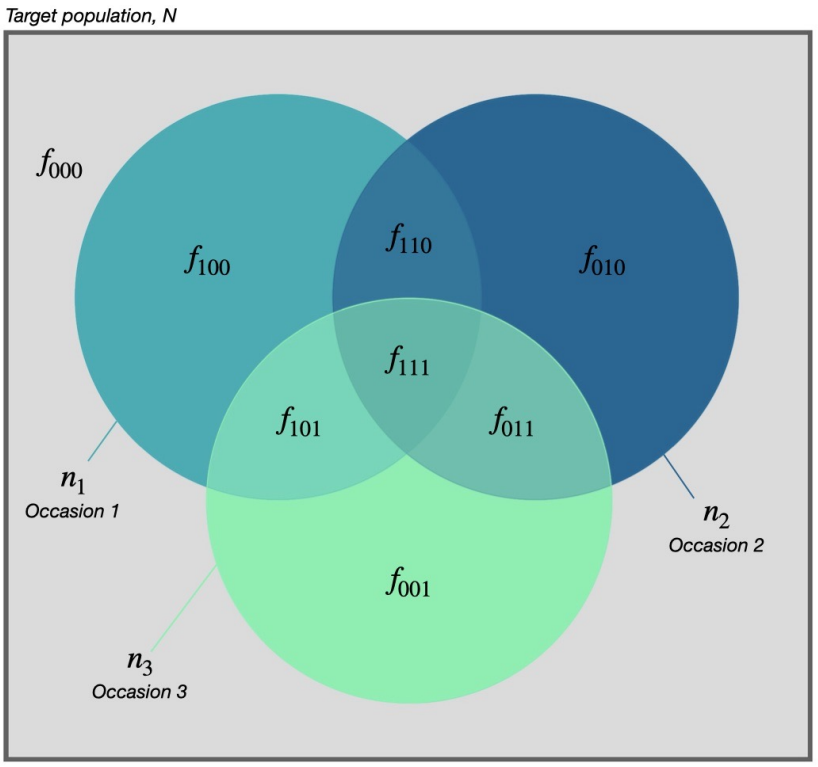

FIGURE 2.2: Venn diagram for three sources of data, Occasion 1, Occasion 2 and Occasion 3, with frequencies of identification and total population sizes.

with corresponding target population size estimate

$$\widehat{N}^{(CPM)} = \frac{(n_1+1)(n_2+1)}{f_{11}+1} - 1.$$

Whilst often not met in application, the estimates rely on the assumption of independence between occasions. Dependence between occasions can be due to either heterogeneity between the individuals within the population, or local dependence. The latter being prevalent in ecology where, for example, an animal may get comfortable or traumatised with the capture process and respectively, can get trap-happy or trap-shy, creating dependence between occasions due to the animal either wanting to be recaptured or doing everything possible to avoid recapture. An increase in dependence between occasions directly correlates to an increase in the bias of the Lincoln-Petersen and Chapman estimators (Chao, 2001; Braeye et al., 2016), leading to the need for alternative methods for when independence cannot be assumed. One of these alternative methods is to use log-linear Poisson models, first proposed by Fienberg (1972).

For the case where there are three occasions, the data structure takes the form of that seen in Figure 2.2 and Table 2.7, where

- $f_0 = f_{000}$ denotes the frequency of unobserved individuals.
- $f_1 = f_{100} + f_{010} + f_{001}$ denotes the frequency of individuals identified only once.

- $f_2 = f_{110} + f_{101} + f_{011}$ denotes the frequency of individuals identified twice.
- $f_3 = f_{111}$ denotes the frequency of individuals identified at each occasion.
- $n = f_1 + f_2 + f_3$ denotes the total number of observed individuals.

TABLE 2.7: Contingency table for three occasions.

Occasion			Frequency
1	2	3	f_{123}
1	1	1	f_{111}
1	1	0	f_{110}
1	0	1	f_{101}
0	1	1	f_{011}
1	0	0	f_{100}
0	1	0	f_{010}
0	0	1	f_{001}
0	0	0	f_{000}

The target population size is then given as

$$\begin{aligned} N = & \underbrace{f_{000}}_{\text{unobserved}} + \underbrace{f_{100} + f_{010} + f_{001} + f_{110} + f_{101} + f_{011} + f_{111}}_{\text{observed}}, \\ & = f_0 + f_1 + f_2 + f_3 \\ & = f_0 + n, \end{aligned}$$

which is equivalent to that in Equation 2.3, where alternative estimation methods are required such as the Horvitz-Thompson estimator seen in Section 4.2, and can be used for data with more sources than 3.

2.3.4 Estimators

To estimate the size of the elusive target populations, various methods are available for different situations. For the case of two source capture-recapture data, the Lincoln-Petersen (Alpizar-Jara and Pollock, 1996) and Chapman (Chapman, 1951) estimators can be used, but for more complex data, in particular when there are covariates present resulting in heterogeneity of the capture probabilities of different sub-populations. Maximum likelihood estimation, the Expectation-Maximisation algorithm and the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) are used in Chapter 4 for estimation of the target population size. Additionally, Chao's lower bound (Chao, 1987) and Zelterman's estimator (Zelterman, 1988) are provided in Chapter 4, alongside the generalised forms (see Böhning et al., 2013b; Böhning and van der Heijden, 2009) which allow for covariates. In instances where there are excess counts of one in the data, the modified Chao's estimator (Böhning et al., 2019) is explored, with the development of the novel

generalised-modified Chao's estimator to account for both covariate information and one-inflation in Chapter 7.

Uncertainty arises with estimation, with the amount of uncertainty quantified through finding standard errors and confidence intervals for the figures calculated. Smaller errors and narrower confidence intervals correspond to estimates with less uncertainty and estimates that are therefore considered more precise. There is importance in quantifying the uncertainty in order to find more robust, reliable and accurate estimates of the parameters of interest. This is explored more in Chapters 5 and 6.

Chapter 3

Zero-Truncated Modelling

Given that typical regression modelling does not accurately represent data with missing zeroes, this chapter discusses alternative methodology that can be utilised to better suit the data. Section 3.1 explores how the proposed distributions can be modified in order to account for the missing zero counts, with Section 3.2 providing methods for model comparison to find the best fitting model, and goodness-of-fit analysis for the selected model with application to the case study data. Lastly, in cases where it is realistically impossible for more counts of events to occur than the number of participants, the binomial distribution is not applicable as it gives a non-zero probability of this occurring. Section 3.3 demonstrates this through use of a simulation study.

3.1 Zero-truncated distribution

With elusive animal populations in particular, it is common for some individuals to go unobserved, so whilst those individuals exist, they lead to unobserved counts of zero, and are hence missing as seen in the hares case study data. However, this missing data does not just occur within ecology. Within meta-analysis, selection bias is common. Often studies with no result are not included within the analysis and lead to missing data, as seen in the suicide case study data where there is a lack of counts of zero counts, leading to zero-truncated data.

Given a chosen baseline model to represent the data, adjustment is required in order to take the lack of zeroes into consideration and avoid overestimating the rate. Using the zero-truncated distribution (Böhning and Friedl, 2021) in Equation 3.1, the chosen baseline distribution can be adjusted, given that $p_x(\mu_i)$ is the density of the baseline distribution.

$$p_x(\mu_i)^+ = \begin{cases} \frac{p_x(\mu_i)}{1-p_0(\mu_i)} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \tag{3.1}$$

where $\mu_i = \tau_i \exp(\eta_i)$ is the given parameter for index i, and $\eta_i = \mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta}$.

3.1.1 Poisson

Applying the baseline formula for zero-truncation from Equation 3.1 to the Poisson distribution gives the density

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

where $\mu_i = \tau_i \exp(\eta_i)$.

3.1.2 Negative-binomial

Zero-truncating the negative-binomial distribution using Equation 3.1 leads to the density

$$p_x(\mu_i, \theta)^+ = \frac{\Gamma(\theta + x) \left(\frac{\mu_i}{\theta+\mu_i}\right)^x \left(\frac{\theta}{\theta+\mu_i}\right)^\theta}{\Gamma(x+1)\Gamma(\theta) \left(1 - \left(\frac{\mu_i}{\theta+\mu_i}\right)^\theta\right)},$$

where $\mu_i = \tau_i \exp(\eta_i)$.

3.1.3 Geometric

Zero-truncating the geometric distribution using Equation 3.1 leads to the following density.

$$p_x(\mu_i)^+ = \left(1 - \frac{1}{1+\mu_i}\right)^{x-1} \frac{1}{1+\mu_i}, \tag{3.2}$$

where $\mu_i = \tau_i \exp(\eta_i)$.

3.1.4 Binomial

The zero-truncated binomial density using Equation 3.1 is

$$p_x(\rho_i, \tau_i)^+ = \frac{\binom{\tau_i}{x} \rho_i^x (1-\rho_i)^{(\tau_i - x)}}{1 - (1-\rho_i)^{\tau_i}},$$

where $\rho_i = [1 + \exp(-\eta_i)]^{-1}$ is the success probability and τ_i is the exposure variable for index i.

The suicide case study utilises a zero-truncated binomial distribution to model the data and estimate the rate of completed suicide after bariatric surgery. Whilst this

approach accounts for the missing data, it allows for a non-zero probability of the count of completed suicide, X_i, being greater than the number of patients in study i for $i = 1, 2, \cdots, n$. In practice, this is impossible, as individuals can only complete suicide one time. Consequently, the binomial model is not appropriate for this situation, however, it is compared to the Poisson, negative-binomial and geometric models for completeness.

3.2 Model evaluation

3.2.1 Likelihood function

When estimating unknown parameters or evaluating the fit of different models, the likelihood function, $L_x(\mu_i)$, is a key component. For example, to compute the maximum likelihood estimate (MLE) of a parameter, which can be used in the Expectation-Maximisation (EM) algorithm or to compare models using the information criterion. For a given distribution's density, $p_x(\mu_i)$, the observed data likelihood is calculated as follows.

$$L_x(\mu_i) = \prod_{i=1}^{n} p_x(\mu_i).$$

Depending on the distribution, manipulating the likelihood to find parameter estimates can become challenging. As a result, it is typically easier to work with the log-likelihood, as seen in Equation 3.3.

$$\ell_x(\mu_i) = \sum_{i=1}^{n} \log(p_x(\mu_i)). \tag{3.3}$$

The likelihood and log-likelihood functions require adjustment for applications within capture-recapture to account for the frequency of counts, f_{ix}, from the data. For estimation of missing studies, the untruncated, complete data (log-)likelihood is considered (see Böhning et al., 2018; McCrea and Morgan, 2014; Borchers et al., 2002).

In the more general case, when the data comes from only one source or there are no unique individuals, so $X_i = X$ and $f_{ix} = f_x$, the complete data likelihood is

$$\prod_{x=0}^{m} p_x(\mu)^{f_x},$$

with corresponding log-likelihood

$$\sum_{x=0}^{m} f_x \log(p_x(\mu)).$$

However, if there are multiple sources of data (such as in the suicide data) or if there are multiple unique individuals (such as in the hares data), the complete data likelihood is

$$\prod_{i=1}^{n}\prod_{x=0}^{m} p_x(\mu_i)^{f_{ix}},$$

with corresponding log-likelihood

$$\sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\log(p_x(\mu_i)). \tag{3.4}$$

The 4 models under consideration are the Poisson, negative-binomial, geometric and the binomial distribution. Using Equation 3.4, the complete data log-likelihoods are as follows.

$$\ell_x(\mu_i) \propto \sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\left[x\log(\mu_i) - \mu_i\right], \tag{3.5}$$

for the Poisson distribution,

$$\ell_x(\mu_i,\theta) \propto \sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\left[x\log\left(\frac{\mu_i}{\theta+\mu_i}\right) + \theta\log\left(\frac{\theta}{\theta+\mu_i}\right)\right], \tag{3.6}$$

for the negative-binomial distribution,

$$\ell_x(\mu_i) \propto \sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\left[\log(\mu_i) + x\log(1-\mu_i)\right], \tag{3.7}$$

for the geometric distribution, and

$$\ell_x(\rho_i,\tau_i) \propto \sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\left[x\log(\rho_i) + (\tau_i - x)\log(1-\rho_i)\right], \tag{3.8}$$

for the binomial distribution.

3.2.1.1 Maximum likelihood estimation (MLE)

Through maximising the (log-)likelihood function, the value of the unknown parameter within a regression model which maximises the probability of obtaining the observed data, and therefore best fits the data, can be found. Maximisation occurs by differentiating the (log-)likelihood with respect to the unknown parameter, setting equal to zero and re-arranging to make the unknown parameter the subject. The resulting parameter estimate is the maximum likelihood estimate (Scholz, 1985). However, the standard approach assumes completeness of the data which is not the case when the data is zero-truncated. The same concept can be used but with using zero-truncated regression

models instead of the standard regression models in order to account for this missing data. Following this approach with the use of a statistical programming language, the maximum likelihood values can be found very efficiently. Alternative methods such as the EM algorithm can also be used to account for the missing data and achieves the same results. Section 4.1 explains the EM algorithm further, however, the MLE approach using the zero-truncated models is typically simpler, more flexible, and more efficient to use, especially in the case where covariates are included.

3.2.2 Information criterion

Information criterion estimate the quality of each model proportionally to the other models fitted to the same dataset. For this method, the models are not required to be nested, provided the data modelled itself is the same, and can be calculated by taking both the general fit of the model to the data, and its complexity into consideration, where its complexity can be described as a penalty term derived from the number of parameters in each model. Factoring both these elements into the calculation leads to a bias-variance trade-off, minimising total error to find the best fitting model. Whilst increasing the number of parameters improves a model's overall fit through reducing the bias with smaller differences between true and estimated values, there exists a possibility of over-fitting, leading to a larger variance. On the other hand, a model with fewer parameters will be prone to under-fitting and smaller variance, but greater bias comparatively. Through inclusion of the penalty term, additional parameters are allowed for improving the fit whilst discouraging over-fitting, balancing the bias and variance to minimise the total error of the model.
The generalised information criterion is given by

$$IC = -2\hat{\ell} + \text{penalty term},$$

where $\hat{\ell}$ is the log-likelihood for a model and the model with the smallest IC is preferred.

3.2.2.1 Akaike information criterion (AIC)

The Akaike Information Criterion (AIC) is a prediction based criterion, assessing the quality of predictions, selecting the preferred model by identifying which model has the lowest AIC value out of those under consideration. As the number of observations increase, the power of the test increases and type II error rate decreases. Type I error rate for AIC is consistent however (Dziak et al., 2020). Therefore, for small numbers of observations, it is important to consider simpler models which can achieve more accurate estimates of the model parameters. The AIC can be calculated by

$$AIC = -2\hat{\ell} + 2d,$$

where d is the number of parameters in the model.

3.2.2.2 Bayesian Information Criterion (BIC)

The Bayesian information criterion (BIC) penalises the model for the number of parameters more than AIC for $n > 8$, due to the logarithm term. This means that, in particular for small n, the BIC is more likely to under-fit. As n increases, whilst at a slower rate than with AIC, the rate of type II errors decreases, and unlike with the AIC, the type I error rate also decreases. Therefore, often the calculated risk of under-fitting the model in order to increase the chances of choosing the true model, defined as being the smallest, correct model, is worthwhile (Dziak et al., 2020). The BIC can be calculated by

$$BIC = -2\hat{\ell} + d\log(n),$$

where d is the number of parameters in the model and n is the number of observed individuals (or studies).

When using the BIC statistic with meta-analytic data, there is a possibility of over-penalising the value of n, which is typically the number of studies. For instance, if one of the studies from the systematic review is split into two, the value of n increases by 1, and the BIC value also increases as a result. However, given that no data has been changed, the overall statistics computed from the data, such as the prevalence rate, would remain the same. Whilst the BIC statistic can be used with meta-analytic data, the choice of n should be done carefully and it can be used alongside the AIC statistic, given that the AIC is computed independently of the choice of n.

3.2.3 Likelihood ratio testing

If two models are hierarchically nested, the likelihood ratio test can be used to assess the goodness-of-fit of the models in order to aid the decision of which model best describes the data. Using the likelihoods of the two models, an asymptotically $\chi^2_{d_1-d_2}$ test statistic can be calculated as

$$2 \times (\hat{\ell}(\text{model 2}) - \hat{\ell}(\text{model 1})),$$

where model 1 is a special case of (nested in) model 2, where some parameters are set to 0. Respectively, model 1 and model 2 have degrees of freedom d_1 and d_2, where $d_1 > d_2$.

The null hypothesis (H_0) is that the nested model fits the data at least as well as the larger model, and is tested against an alternative hypothesis. If the corresponding p-value for the test statistic is less than some pre-defined significance level (typically 0.05), then there is not significant evidence to reject H_0, and the result is statistically significant. Otherwise, the null hypothesis is rejected.

3.2.4 Fitted frequencies

Comparison of fitted frequencies given a chosen model to observed frequencies is a method for assessing goodness-of-fit, with the fitted frequencies calculated using the work of Holling et al. (2016) in Equation 3.9.

$$\hat{f}_x = \sum_{i=1}^{n} p_x(\tau_i \exp(\hat{\eta}))^+. \tag{3.9}$$

Informally, the smaller the difference between the fitted and observed frequencies, the better the fit of the chosen model. Formally, goodness-of-fit can be assessed using a χ^2 test with the test statistic found using Equation 3.10.

$$\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \hat{f}_x)^2}{\hat{f}_x}. \tag{3.10}$$

3.2.5 Ratio plots

Statistical graphs can be a useful tool for analysing data, with ratio plots developed in Böhning et al. (2013a) used as a diagnostic tool for exploring the validity of distributional assumptions. Through multiplying ratios of the neighbouring probabilities by the inverse ratios of their coefficients, the following ratio used in the ratio plots can be found.

$$r_x = \frac{a_x}{a_{x+1}} \frac{p_{x+1}}{p_x},$$

where for different distributional assumptions, the associated coefficient is given as

$$a_x = \begin{cases} \frac{1}{x!} \text{ if Poisson assumed, or} \\ 1 \text{ if geometric assumed.} \end{cases}$$

The theoretical quantity p_x can be approximated by $\frac{f_x}{N}$, given that whilst N is unknown, when substituted into the ratio, the unknown value of N cancels itself out. This then leads to the following estimate of the probability ratio.

$$\hat{r}_x = \frac{a_x}{a_{x+1}} \frac{f_{x+1}}{f_x}.$$

If the ratios in the plot follow a horizontal line, it can be assumed that the given distributional assumption is valid, with ratios not following the pattern of a horizontal line giving evidence for the distributional assumption being invalid.

Application: Suicide data

The zero-truncated Poisson, negative-binomial, geometric and binomial models are all under consideration as intercept-only models. However, as there are the covariates of proportion of women and country of origin available in the case study data, main effects and interaction models are also considered for each distribution. To account for these possible covariate effects, 5 linear predictors are under consideration whose forms are seen in Table 2.1, where linear predictor 1 is the intercept-only model. Country of origin as a categorical variable has many levels where only few occur more than once. To aid in modelling and analysis, country of origin is collapsed to have only two levels; 'USA' and 'Other'.

The values of the maximised log-likelihoods, number of parameters, AIC and BIC for each of these models are shown in Table 3.1, with BIC weights given for the Poisson and negative-binomial distributions. The BIC weights (Wagenmakers and Farrell, 2004) are treated as conditional probabilities for model selection in Chapter 6 and are calculated as

$$w_l(BIC) = \frac{\exp\left[-\frac{1}{2}\Delta_l(BIC)\right]}{\sum_{k=1}^{K}\exp\left[-\frac{1}{2}\Delta_k(BIC)\right]},$$

where $\Delta_l(BIC) = BIC_l - \min(BIC)$ is the difference in BIC value for each model and the best candidate model. This is further developed in Section 6.4.

The difference between the Poisson, negative-binomial and binomial log-likelihoods is negligible, showing little preference of a preferred model. Whilst the differences between the Poisson and binomial models are also negligible for the information criterion, the Poisson is favoured due to the non-zero probability of more suicides occurring than total participants in the study seen in the binomial models. The geometric distribution has AIC and BIC values that are much larger than the other distributions for the respective linear predictors, indicating that it the geometric distribution is not a good fit for the data. Therefore, BIC weights are not given for the geometric distribution. As for the Poisson and negative-binomial models, the value of the dispersion parameter θ is estimated to be very large, resulting in the variance in the negative-binomial models approximately equalling the mean. This leads to the negative-binomial models approximating the Poisson distribution. However, there are differences in the information criterion, with those from the Poisson models being slightly smaller due to one less parameter, hence the Poisson distribution is favoured.

Using the information criterion in Table 3.1, there is evidence that the intercept-only model is favoured for each of the distributions, and whilst not necessary, likelihood ratio testing can be used to provide additional evidence to support this. Table 3.2 shows test statistics and corresponding p-values for comparing the models with covariates (main effects and interaction models) against the intercept-only model for the Poisson distribution, since the information criterion have already shown this to be the best fitting

TABLE 3.1: Values of the maximised log-likelihood, number of parameters, AIC and BIC for the models under consideration for the suicide case study data. Values of BIC weights are included for the Poisson and negative-binomial models under consideration. The geometric distribution fits the data poorly so BIC weights are not given. The binomial distribution is not suitable for this situation so BIC weights are not given.

Distribution	Linear predictor	Maximised log-likelihood	Number of parameters	AIC	BIC	BIC weights
Poisson	**1**	**-23.73**	**1**	**49.45**	**50.75**	**0.4813**
	2	-23.37	2	50.74	53.44	0.1251
	3	-23.03	2	50.05	52.64	0.1863
	4	-22.97	3	51.93	55.92	0.0362
	5	-22.65	4	53.29	58.56	0.0097
Negative-binomial	1	-23.73	2	51.45	54.04	0.0926
	2	-23.37	3	52.74	56.74	0.0241
	3	-23.03	3	52.05	55.94	0.0359
	4	-22.97	4	53.93	59.22	0.0070
	5	-22.65	5	55.29	61.86	0.0019
Geometric	1	-26.03	1	54.07	55.37	-
	2	-25.97	2	55.94	58.54	-
	3	-26.03	2	56.06	58.65	-
	4	-25.97	3	57.94	61.83	-
	5	-25.52	4	59.04	64.23	-
Binomial	1	-23.72	1	49.45	50.75	-
	2	-23.37	2	50.74	53.25	-
	3	-23.03	2	50.05	52.64	-
	4	-22.96	3	51.93	55.70	-
	5	-22.64	4	53.29	58.32	-

TABLE 3.2: Values for likelihood ratio testing for each of the zero-truncated Poisson models with covariates compared to the nested intercept-only model for the suicide case study data.

Linear predictor compared to $j = 1$	d_1	d_2	Likelihood ratio test statistic	p-value
2	26	25	0.6017657	0.2598355
3	26	25	1.3975182	0.5027981
4	26	24	1.4172376	0.2985008
5	26	23	2.0732480	0.2777114

distribution. The null hypothesis of *"H_0 : the nested model fits the data at least as well as the larger model"*, should not be rejected as there are no p-values less than the pre-defined level of significance of 0.05, so there are no statistically significant results. Therefore, it can be assumed that the best fitting model is the intercept-only zero-truncated Poisson regression model. This model leads to an estimated rate of 31.8 completed suicides per 100,000 person-years, which to the same degree of accuracy, is the same estimate from the intercept-only models from the negative-binomial and binomial distributions.

Given that there is a clear preference for the Poisson distribution using the information criterion, and only a small number of counts observed within the suicide case study data, the use of a ratio plot is not the most suitable method for assessing distributional

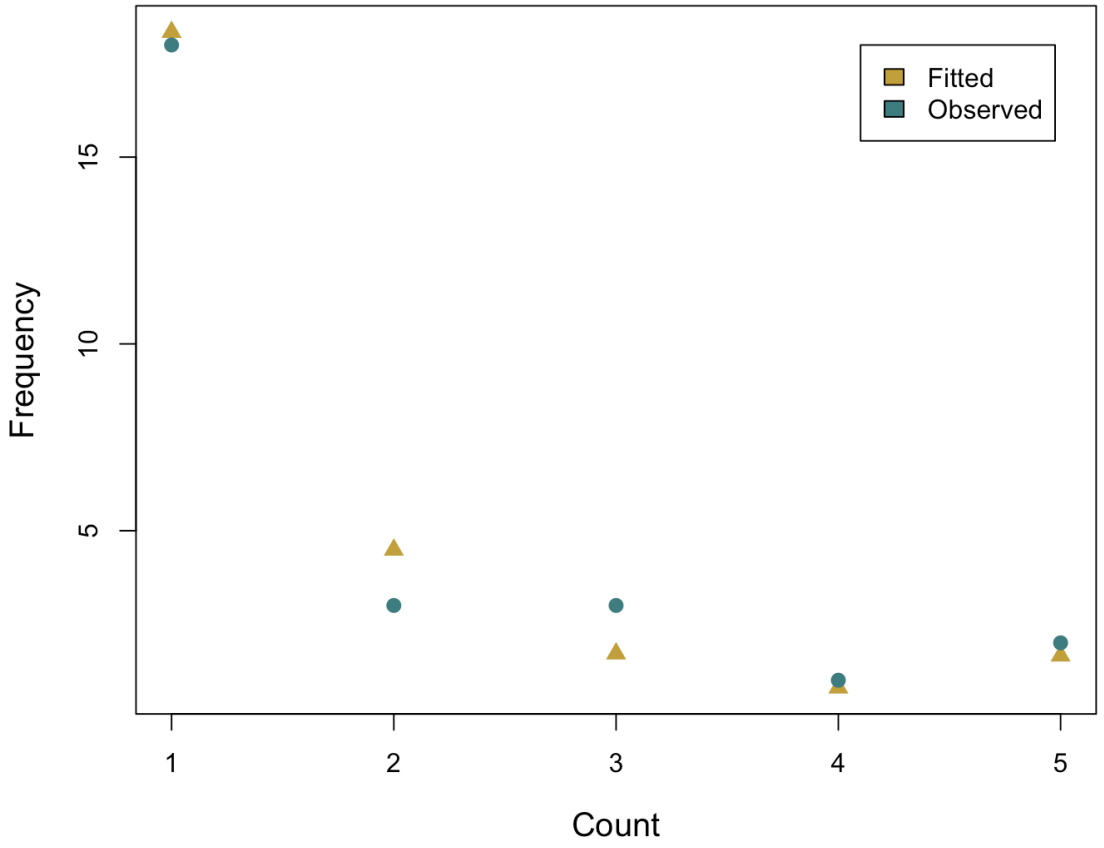

FIGURE 3.1: Plot of the observed frequencies and the fitted frequencies using the Poisson distribution assuming the intercept-only model for the suicide case study data.

assumptions. However, for assessing the goodness-of-fit assuming the Poisson distribution for this case study data with linear predictor 1 can be done using fitted frequencies. Therefore, the adequacy of the intercept-only zero-truncated Poisson regression model is considered by calculating the fitted frequencies and comparing to the observed frequencies of counts of completed suicide where the fitted frequencies are calculated as

$$\hat{f}_x = \sum_{i=1}^{n} p_x(\tau_i \exp(\hat{\eta}))^+ = \sum_{i=1}^{n} \frac{(\tau_i \exp(\hat{\eta}))^x}{(\tau_i \exp(\hat{\eta})) - 1)x!}.$$

Table 3.3 displays the fitted and observed frequencies of counts of completed suicide, where those with counts equal to 5 and above are grouped into one category.

TABLE 3.3: Frequency distribution for observed and fitted count of completed suicide, with the frequencies of more than or equal to 5 counts grouped into one category.

	Count of completed suicide, x					
Frequency type	0	1	2	3	4	5+
Observed, f_x	-	18	3	3	1	2
Fitted, $\hat{f}_x$	-	18.35	4.49	1.71	0.80	1.65

It can be seen that the fitted frequencies of the counts are very close and approximately equal to the observed, in particular for the frequency of one count of completed suicide. For formally assessing the goodness-of-fit of the model, a χ^2 test statistic with 2 degrees of freedom can be computed using Equation 3.10, where $m = 5$. For the null hypothesis, *"H_0 : the chosen model is adequate for the data"*, there is no evidence to suggest that the model is inadequate for the data as the test statistic of 1.593945 is larger than the corresponding critical value of 0.4506914, so the null hypothesis should not be rejected.

Sensitivity analysis

It is important to note that the suicide data follows an exponential pattern, seen in the plot of person-years vs the rate of completed suicide per 100,000 person-years in Figure 3.2 (left). A log-linear regression model transforms the relationship between the explanatory and response variables into one which is more linear, seen in Figure 3.2 (right), likely leading to an increase in the reliability of predictions.

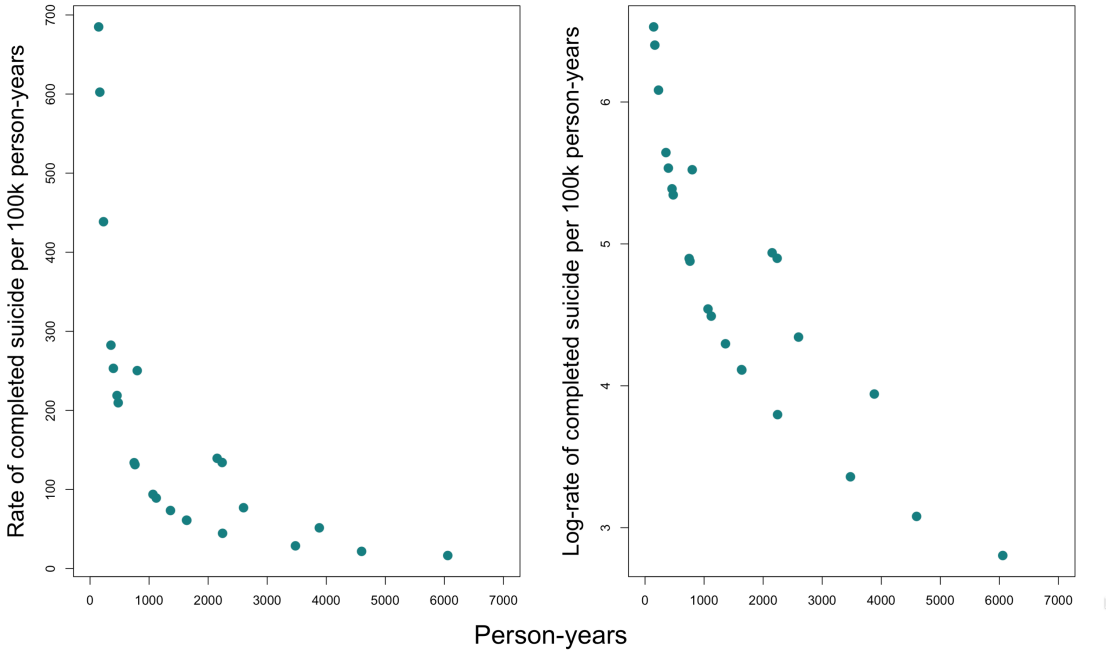

FIGURE 3.2: Side-by-side scatter-plot of rates and log-rates of completed suicide per 100,000 person-years in ascending order of person-years (Peterhänsel et al., 2013). An x-axis limit has been added to exclude values which are very large for better visualisation of the trends.

Setting an exposure variable for the logarithm of person-years results in the implication that the rate of completed suicide is constant over time, implying a linear relationship not the exponential relationship seen. To assess the validity of these underlying assumptions, a sensitivity analysis can be conducted (the same sensitivity analysis is not required for

the hares data, given that there is no exposure variable included in the zero-truncated models for this case study data).

Given the difference in the person-years for each of the studies, it needs to be accounted for. However, given that including the logarithm of person-years implies that there is a linear relationship which is not seen in Figure 3.2, the first approach of the sensitivity analysis is to include person-years in the model as a binary variable, v_4, rather than as an exposure variable. If the resulting model fits the data better than the model with the exposure variable, then there is evidence to suggest that person-years shouldn't be included as an offset. If the model with the exposure variable is preferred however, then there is evidence to support the underlying assumption of constant rate of suicide over time, and the logarithm of person-years should still be included in the final model as an exposure variable.

Including person-years in the model as a binary variable can be done through categorising the person-times values into either "low" or "high" categories, using the median value to decide the cut off for the two groups. The full model (main effects for person-years, proportion of women and country of origin with all interaction terms) can be fitted with the best fitting model selected using backwards stepwise regression. The resulting preferred model takes the following form, assuming a negative-binomial distribution.

$$X_i \sim negbin(\mu_i, \alpha),$$

where α is the overdispersion parameter and

$$\mu_i = \tau_i \exp(\beta_0 + \beta_1 v_1 + \beta_2 v_2 + \beta_3 v_4 + \beta_4 x_1 x_2).$$

The AIC and BIC statistics corresponding to this preferred model are 68.3 and 76.1 respectively, values which are greater than those of the zero-truncated intercept-only Poisson model with person-years as an exposure variable seen in Table 3.1. Therefore, there is evidence that the original model is preferred and person-years should be included as an exposure variable.

An alternative approach to, the sensitivity analysis exploring whether the rate is constant across the population is to utilise a zero-truncated J-component Poisson model (Böhning et al., 2011), where it is assumed that there are at least two sub-populations ($J \geq 2$) with differing rates of completed suicide (if $J = 1$ then there is no heterogeneity and the model is the same as the original preferred model). This approach allows for the exploration of possible unobserved heterogeneity coming from any underlying sub-populations within the dataset. The J-component model can then be given as follows.

$$p_x^J(\mu_i)^+ = \sum_{j=1}^{J} w_j p_x(\mu_j)^+,$$

where the weight, w_j, corresponds to the size of the cluster and

$$p_x^+(\mu_j) = \frac{\exp(-\mu_j)\mu_j^{x_i}}{x_i!}.$$

Applying this two-component model with $J = 2$ to the suicide case study data results in AIC and BIC values of 53.4 and 57.3 respectively. As with the first approach to the sensitivity analysis, these information criterion values are greater than those of the preferred zero-truncated intercept-only Poisson model with person-years included as an exposure variable and no heterogeneity (AIC of 49.5 and BIC of 50.7). Therefore, there is evidence that the original model with no heterogeneity is preferred, and it is reasonable to assume that the rate of completed suicide is constant over time.

Application: Hares data

TABLE 3.4: Values of the maximised log-likelihood, number of parameters, AIC and BIC for the models under consideration for the hares case study data.

Distribution	Linear Predictor	Maximised log-likelihood	Number of parameters	AIC	BIC
Poisson	1	-985.21	1	1972.42	1977.32
	2	-966.43	2	1938.86	1953.53
	3	-982.03	2	1968.06	1977.84
	4	-962.46	3	1932.93	1952.49
	5	-958.16	4	1924.32	1943.88
Negative-binomial	1	-963.80	2	1931.60	1941.39
	2	-950.14	3	1908.27	1927.84
	3	-961.53	3	1929.06	1943.73
	4	-947.72	4	1905.43	1929.89
	5	-944.23	5	1898.46	1922.91
Geometric	1	-963.99	1	1929.98	1934.87
	2	-950.17	2	1906.33	**1921.00**
	3	-961.64	2	1927.28	1937.06
	4	-956.56	3	1903.67	1923.24
	5	**-940.45**	**4**	**1892.91**	1922.25

Table 3.4 provides the values of the maximised log-likelihoods, number of parameters, AIC and BIC values for each of the linear predictor and distribution combinations under consideration for the hares case study data. The negative maximised log-likelihood, AIC and BIC values are notably larger for the Poisson distribution for the respective linear predictors compared to the other distributions, indicating that it is not the preferred distribution. Whilst there is little difference between the log-likelihood values for the negative-binomial and geometric distribution, the AIC and BIC values are larger for

the respective linear predictors for the negative-binomial model, indicating that the negative-binomial distribution is also not preferred, and hence there is evidence that the geometric distribution is the best fitting distribution for the data. The BIC statistic values for the geometric models with linear predictors 2 and 5 have very little difference, with only a slight preference for the full model. However, the log-likelihood and AIC values have more notable differences, with the values indicating a stronger preference for the geometric model with linear predictor 5. Therefore, there is evidence that the full geometric model fits the data best.

TABLE 3.5: Values of the estimated probability ratios for the Poisson and geometric distributions for the hares case study data.

		Count of captures of hares, x				
Distribution		1	2	3	4	5
Poisson	$\hat{r}_x$	6.22	8.40	10.71	10.00	28.00
Geometric		3.11	2.80	2.68	2.00	4.67

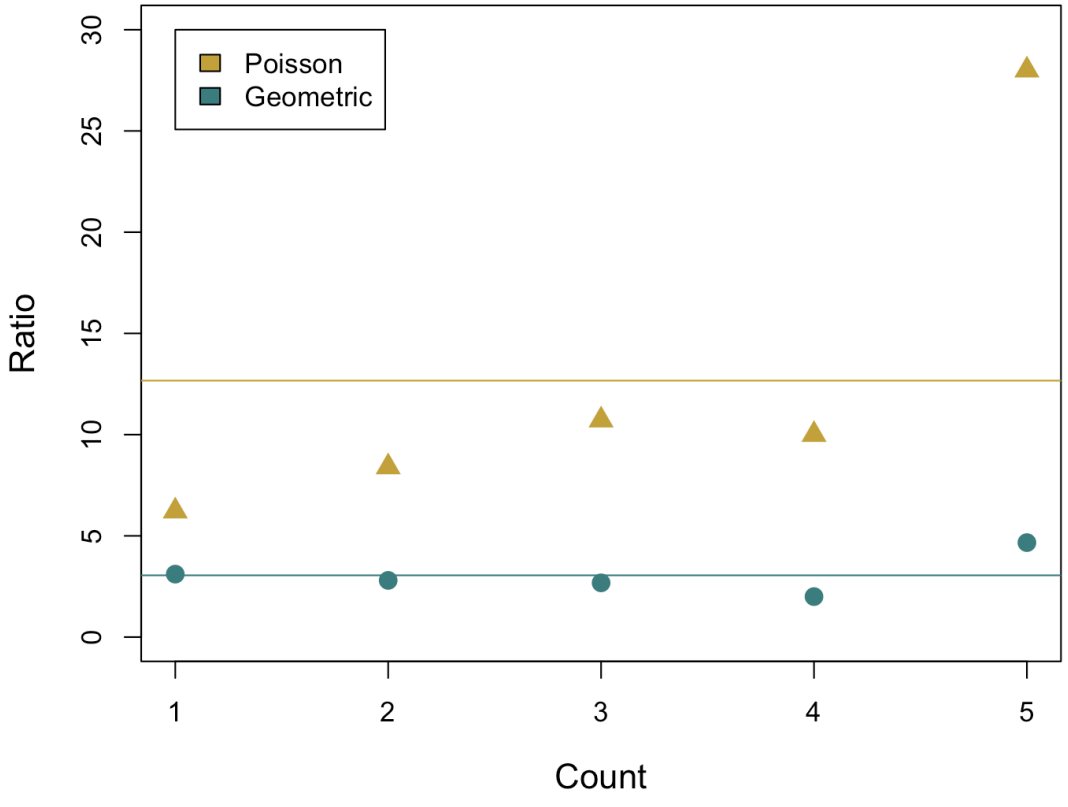

FIGURE 3.3: Ratio plot for both Poisson and geometric distributions for the hares case study data.

Table 3.5 gives the probability ratios for the snowshoe hares capture data for both the Poisson and geometric distributions. The corresponding ratio plots are given in Figure 3.3, supporting the conclusions from the likelihoods and information criterion

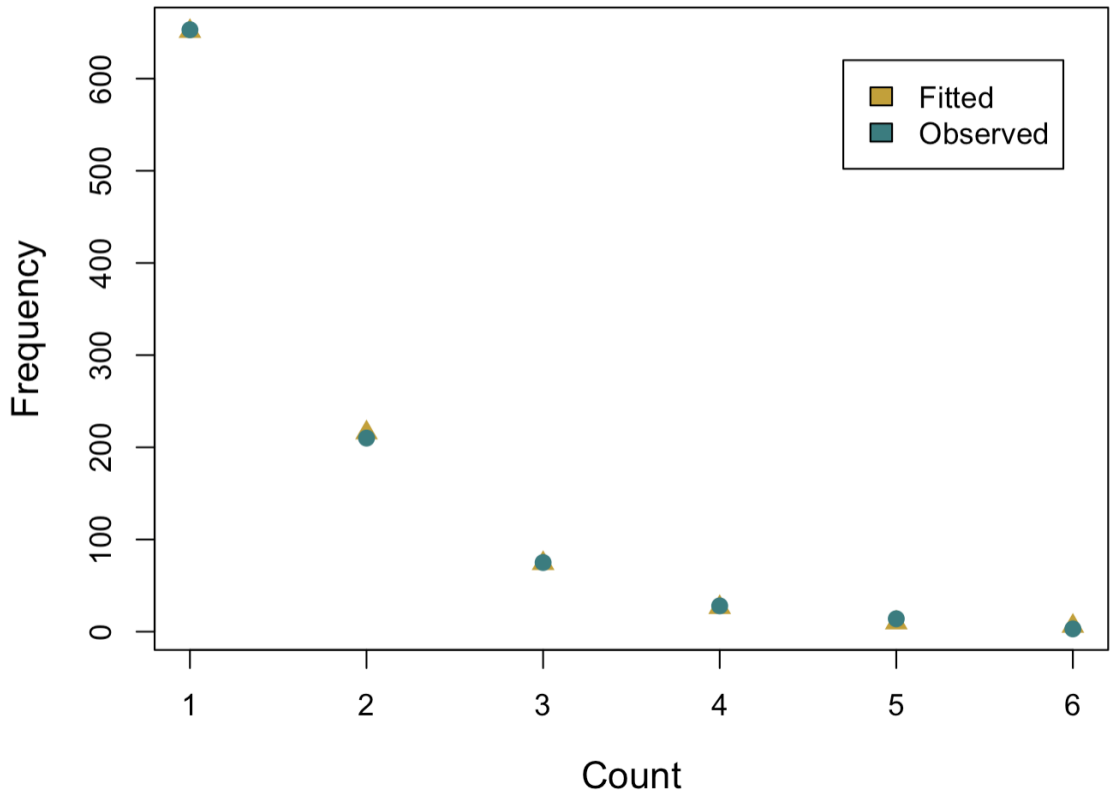

FIGURE 3.4: Plot of the observed frequencies and the fitted frequencies using the geometric distribution assuming the full model for the hares case study data.

that the geometric distribution fits the data better than the Poisson, with the probability ratios for the geometric distribution following a more horizontal pattern.

TABLE 3.6: Values of the observed frequencies and the fitted frequencies using the geometric distribution assuming the full model for the hares case study data.

	Count of captures of hares, x						
Frequency type	0	1	2	3	4	5	6+
Observed, f_x	-	653	210	75	28	14	3
Expected, $\hat{f}_x$	-	651	216	74	26	10	6

Table 3.6 and Figure 3.4 provide the values of the observed frequencies of snowshoe hare captures, in addition to the fitted frequencies using the geometric model with linear predictor 5 (the full model). It can be seen in both the table and the figure that the fitted frequencies closely follow the pattern of the observed frequencies, suggesting that the full model is indeed a good fit for the snowshoe hares dataset.

3.3 Simulation study

For the suicide case study data, the zero-truncated Poisson distribution approximates the zero-truncated binomial distribution, leading to approximately equal parameter estimates. However, with there being a non-zero probability of count of completed suicides being greater than the number of participants, this is not always going to be the case. As a result, the binomial distribution is not suitable for this application, demonstrated by the following simulation study, which shows that the Poisson is the better choice of model.

For this simulation study, there are 3 cases. The first case assumes that the observation period for each study is 1 unit, hence person-times is equal to the number of patients observed, whereas both Cases 2 and 3 assume that the observation period varies between studies.

For each case, given the number of studies, N, and the mean number of total patients, $\bar{t}$, the size of each study is randomly sampled from the Poisson distribution, $t_i \sim Poisson(\bar{t})$. Assuming a constant rate of λ, and observation period for each study, O_i, the count of events is randomly sampled from a binomial distribution, $X_i \sim Binomial(N, \tau_i)$, where $\tau_i = \lambda \times O_i$. Sampling in this way prevents the simulated data from being biased to either the Poisson or binomial distributions, which would give misleading and unreliable results if biased. Counts of zero are truncated from this sampled data to make a zero-truncated dataset, with updated size of each study, t_i^*, observational periods, O_i^*, and counts, X_i^*. From this updated dataset, the intercept-only zero-truncated Poisson and binomial models are applied as in Section 3.1, to calculate the MLE for the rate of events. This process is repeated S times. The summary statistics are calculated using the mean of the rates over each of the S simulated samples. For each of the cases, there are 3 models used to calculate the MLE of the rate,

i the zero-truncated Poisson intercept-only model with parameter $t_i^* O_i^*$,

ii the zero-truncated binomial intercept-only model with parameter t_i^*, and

iii the zero-truncated binomial intercept-only model with parameter $t_i^* O_i^*$,

where (ii) is the model which Peterhänsel et al. (2013) assumes.

Case 1: Assume the observational period is equal to one unit for all studies, so $O_i = 1$ for $i = 1, 2, \cdots, N$.

Case 2: The observational period for each study, O_i, is randomly sampled from a Bernoulli distribution where

$$O_i = \begin{cases} 1 & \text{with probability 0.5,} \\ 5 & \text{with probability 0.5,} \end{cases}$$

for $i = 1, 2, \cdots, N$.

Case 3: In application, it would be unlikely that the observation period for each study is discrete as seen in Case 2, so for Case 3, the observational period is assumed to be log-normally distributed. For each simulated sample, the observational period is sampled from a log-normal distribution, $O_i \sim LogNormal(\gamma, \sigma)$. For the simulated study, to produce a dataset reflective of the suicide case study data, $\gamma = 1.5$ and $\sigma = 0.8$, leading to the results seen in Table 3.7.

For Case 1, as the observational period is assumed to be equal to 1 for each study, models (ii) and (iii) are equal, so only (ii) is performed.

TABLE 3.7: Average (mean) estimated rate of event occurring per 100,000 person-years from the simulation study, where the true rate is 100 occurrences per 100,000 person-years, with 95% percentile confidence intervals given in (brackets). Values given for each of the models under consideration for Cases 1, 2 and 3, assuming $S = 1000$, $\lambda = 0.001$, $N = 150$ and $\bar{t} = 1000$.

	Model					
Case	(i)		(ii)		(iii)	
1	99.80	(74.48, 125.13)	99.74	(74.42, 125.05)	-	-
2	99.90	(90.14, 109.66)	358.37	(311.12, 405.63)	99.90	(90.14, 109.66)
3	99.99	(93.39, 106.59)	657.05	(69.74, 1244.35)	99.99	(90.23, 109.75)

The aim of this simulation study is to determine which combinations of model and case produce estimated rate of events close to the true rate, with reasonable confidence intervals, where the true rate of event occurring is set to 100 events per 100,000 person-years. Table 3.7 shows the results from the simulation study. For Case 1, the rates and corresponding confidence intervals have negligible difference for both model (i) and (ii), with the estimated rates approximately equalling the true rate of 100 per 100,000 person-years. This is expected given that the observation period is simply 1 for each study, making the parameters for both models equal. However, for Cases 2 and 3, these two models do not result in the same estimate. Model (i) produces an estimate of the rate which is approximately equal to the true rate, but model (ii), which Peterhänsel et al. (2013) proposes, produces an estimate of the rate 3 times the true value for Case 2 and 6 times the true value for Case 3, demonstrating the model's lack of suitability and accuracy for this situation. For comparison, model (iii) is also a binomial model, but assumes the same parameter as the Poisson model, taking into consideration the observation period. Whilst Section 3.1.4 explains why the binomial model is not an appropriate choice, models (i) and (iii) produce approximately equal results with negligible difference, demonstrating the importance of taking the observation period into consideration which Peterhänsel et al. (2013) does not do adequately.

Chapter 4

Estimation Methods

Estimation methods are required to find more accurate rates in cases where counts of zero events are systematically missing, which can then be used to approximate the true population size. Adjusted rate values can be estimated through using the Expectation-Maximisation (EM) algorithm in Section 4.1, which also estimates the number of missing studies. However, the EM algorithm is not always used to find the MLE, so alternative population size estimators are required in situations where the population size isn't estimated as a by-product of the rate estimation. These alternative population size estimation methods developed include the Horvitz-Thompson, Chao's and Zelterman's estimators, with the corresponding generalised versions for the latter two, covered in Sections 4.2 to 4.6. Finally, Section 4.7 explores the performance of the capture-recapture estimators in different data scenarios via use of a simulation study.

4.1 Expectation-Maximisation algorithm

One approach to account for missing data when finding the local maximum likelihood parameters is the Expectation-Maximisation (EM) algorithm, first discussed by Orchard and Woodbury (1972) and named by Dempster (1977). Each iteration consists of two main steps, the E-step (expectation step) imputes the missing data, then the the M-step (maximisation step) estimates the parameter. These two steps are repeated and iterated between until convergence if two consecutive log-likelihoods. For capture-recapture specifically, iterate the steps in Algorithm 1 to find the maximum likelihood estimator, using the works of Böhning et al. (2005).

Algorithm 1 Expectation-Maximisation algorithm

Step 1: Let $s = 0$ and choose a start value for μ_i.
Step 2: For the E-step, evaluate the conditional expectation for the log-likelihood of the complete data, $Q(\mu_i)$.
The total number of individuals for each covariate combination is given by

$$\begin{aligned} N_i &= f_{i0} + f_{i1} + f_{i2} + \cdots + f_{in} \\ &= f_{i0} + \sum_{x=1}^{m} f_x \\ &= f_{i0} + n_i, \end{aligned}$$

and estimated as

$$\widehat{N}_i = n_i + \hat{f}_{i0},$$

where the unobserved frequency is estimated as

$$\begin{aligned} \hat{f}_{i0} &= \mathrm{E}[f_{i0}|f_{i1}, f_{i2}, \cdots, f_{in}; \mu_i] \\ \hat{f}_{i0} &= N_i p_0(\mu_i) \\ \hat{f}_{i0} &= (n_i + \hat{f}_{i0}) p_0(\mu_i) \\ \hat{f}_{i0}(1 - p_0(\mu_i)) &= n_i p_0(\mu_i) \\ \hat{f}_{i0} &= n_i \frac{p_0(\mu_i)}{1 - p_0(\mu_i)}. \end{aligned} \tag{4.1}$$

The value of this estimated unobserved frequency can then be substituted into $Q(\mu_i)$.
Step 3: For the M-step, maximise $Q(\mu_i)$ from the E-step in Step 2 to find the maximum likelihood estimate (MLE) of μ_i by making $\frac{dQ(\mu_i)}{d\mu_i}$ equal to zero and make $\exp(\eta_i)$ the subject.
Step 4: Return to Step 2, let $s = s + 1$ and update the parameter estimate.
Step 5: Repeat Steps 2 to 4 until convergence of the zero-truncated log-likelihood, hence

$$\ell(\mu_i^{(s)}; x_1, \cdots, x_n) - \ell(\mu_i^{(s-1)}; x_1, \cdots, x_n) < \epsilon,$$

where the convergence level, ϵ, is suitably small.

If a Poisson distribution is assumed, the expectation and maximisation steps of Algorithm 1 can be altered to reflect this assumption as follows.

Step 2: For the E-step, the conditional expectation of the log-likelihood of the complete data is evaluated. Applying this to the zero-truncated Poisson distribution, using the complete data log-likelihood in Equation 3.5 as follows

$$\begin{aligned} Q(\mu_i) &\propto \sum_{i=1}^{n} \sum_{x=0}^{m} f_{ix} \left[x \log(\mu_i) - \mu_i \right] \\ &= -\sum_{i=1}^{n} f_{i0}\mu_i + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} \left[x \log(\mu_i) - \mu_i \right]. \end{aligned} \tag{4.2}$$

In the case where only one individual is observed for each covariate combination, $n_i = 1$ for $i = 1, 2, \cdots, n$, and $p_0(\mu_i) = \exp(-\mu_i)$ is substituted into Equation 4.1, leading to

$$\hat{f}_{i0} = n_i \frac{\exp(-\hat{\mu}_i)}{1 - \exp(-\hat{\mu}_i)}, \tag{4.3}$$

which can be substituted into Equation 4.2 as

$$Q(\mu_i) \propto -\sum_{i=1}^{n} \hat{f}_{i0}\mu_i + \sum_{i=1}^{n}\sum_{x=1}^{m} f_{ix}\left[x\log(\hat{\mu}_i) - \hat{\mu}_i\right]. \tag{4.4}$$

Step 3: For the M-step, estimating the maximum likelihood estimate (MLE) of μ_i requires maximisation of $Q(\mu_i)$ given in Equation 4.4, where the maximum is found through equating $\frac{dQ(\mu_i)}{d\mu_i}$ to zero and making $\exp(\hat{\eta})$ the subject, where $\mu_i = \tau_i \exp(\eta_i)$, as follows. It is important to note that given that each study only reports one value for the count of completed suicides, $\sum_{x=1}^{m} f_{ix} = 1$, for $i = 1, 2, \cdots, n$.

$$\begin{aligned}
\frac{dQ(\mu_i)}{d\mu_i} &= -\sum_{i=1}^{n} \hat{f}_{i0} - \sum_{i=1}^{n}\sum_{x}^{m} f_{ix} + \sum_{i=1}^{n}\sum_{x=1}^{m} \frac{x f_{ix}}{\mu_i} \\
\Rightarrow 0 &= -\sum_{i=1}^{n} \hat{f}_{i0} - \sum_{i=1}^{n}\sum_{x=1}^{m} f_{ix} + \sum_{i=1}^{n}\sum_{x=1}^{m} \frac{x f_{ix}}{\hat{\mu}_i} \\
\Rightarrow 0 &= -\sum_{i=1}^{n} \left(\hat{f}_{i0} + 1\right) + \sum_{i=1}^{n}\sum_{x=1}^{m} \frac{x f_{ix}}{\tau_i \exp(\hat{\eta}_i)} \\
\Rightarrow \sum_{i=1}^{n} \tau_i \left(\hat{f}_{i0} + 1\right) &= \frac{\sum_{i=1}^{n}\sum_{x=1}^{m} x f_{ix}}{\exp(\hat{\eta})} \\
\Rightarrow \exp(\hat{\eta}) &= \frac{\sum_{i=1}^{n}\sum_{x=1}^{m} x f_{ix}}{\sum_{i=1}^{n} \tau_i \left(\hat{f}_{i0} + 1\right)}.
\end{aligned}$$

Here, the numerator does not depend on f_{i0} and hence does not change with each iteration of the algorithm, but the denominator updates each iteration.

Alternatively, if a geometric distribution is assumed, the EM algorithm can be developed to reflect this through altering the expectation and maximisation steps of Algorithm 1 as follows.

Step 2: For the E-step, the conditional expectation of the log-likelihood of the complete data is evaluated. Applying this to the zero-truncated geometric distribution, using the complete data log-likelihood in Equation 3.7 as follows

$$\begin{aligned}
Q(\mu_i) &\propto \sum_{i=1}^{n}\sum_{x=0}^{m} f_{ix}\left[\log(\mu_i) + x\log(1-\mu_i)\right] \\
&= \sum_{i=1}^{n} f_{i0}\log(\mu_i) + \sum_{i=1}^{n}\sum_{x=1}^{m} f_{ix}\left[\log(\mu_i) + \log(1-\mu_i)\right].
\end{aligned} \tag{4.5}$$

In the case where only one individual is observed for each covariate combination, $n_i = 1$ for $i = 1, 2, \cdots, n$, and $p_0(\mu_i) = \frac{\mu_i}{1+\mu_i}$ is substituted into Equation 4.1, leading to

$$\hat{f}_{i0} = n_i \frac{\frac{\hat{\mu}_i}{1+\hat{\mu}_i}}{1 - \frac{\hat{\mu}_i}{1+\hat{\mu}_i}}, \tag{4.6}$$

which can be substituted into Equation 4.2 as

$$Q(\mu_i) \propto \sum_{i=1}^{n} \hat{f}_{i0} \log(\mu_i) + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} \left[\log(1-\mu_i) + x \log(\mu_i)\right]. \tag{4.7}$$

Step 3: For the M-step, estimating the maximum likelihood estimate (MLE) of μ_i requires maximisation of $Q(\mu_i)$ given in Equation 4.7, where the maximum is found through equating $\frac{dQ(\mu_i)}{d\mu_i}$ to zero and making $\exp(\hat{\eta})$ the subject, where $\mu_i = \tau_i \exp(\eta_i)$, as follows

$$\begin{aligned}
&\frac{dQ(\mu_i)}{d\mu_i} = \sum_{i=1}^{n} \frac{\hat{f}_{i0}}{\mu_i} + \sum_{i=1}^{n} \sum_{x=1}^{m} \frac{f_{ix}}{\mu_i} - \sum_{i=1}^{n} \sum_{x=1}^{m} \frac{x f_{ix}}{1-\mu_i} \\
&\Rightarrow 0 = \sum_{i=1}^{n} \hat{f}_{i0}(1-\hat{\mu}_i) + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}(1-\hat{\mu}_i) - \sum_{i=1}^{n} \sum_{x=1}^{m} x f_{ix} \hat{\mu}_i \\
&\Rightarrow 0 = \sum_{i=1}^{n} \hat{f}_{i0} - \sum_{i=1}^{n} \hat{f}_{i0}\hat{\mu}_i + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} - \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}\hat{\mu}_i - \sum_{i=1}^{n} \sum_{x=1}^{m} x f_{ix} \hat{\mu}_i \\
&\Rightarrow \sum_{i=1}^{n} \hat{f}_{i0}\hat{\mu}_i + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}\hat{\mu}_i + \sum_{i=1}^{n} \sum_{x=1}^{m} x f_{ix} \hat{\mu}_i = \sum_{i=1}^{n} \hat{f}_{i0} + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} \\
&\Rightarrow \sum_{i=1}^{n} \hat{f}_{i0}\tau_i \exp(\hat{\eta}_i) + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}\tau_i \exp(\hat{\eta}_i) + \sum_{i=1}^{n} \sum_{x=1}^{m} x f_{ix} \tau_i \exp(\hat{\eta}_i) = \sum_{i=1}^{n} \hat{f}_{i0} + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} \\
&\Rightarrow \exp(\hat{\eta}) \left[\sum_{i=1}^{n} \hat{f}_{i0}\tau_i + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}\tau_i + \sum_{i=1}^{n} \sum_{x=1}^{m} x f_{ix} \tau_i\right] = \sum_{i=1}^{n} \hat{f}_{i0} + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix} \\
&\Rightarrow \exp(\hat{\eta}) = \frac{\sum_{i=1}^{n} \hat{f}_{i0} + \sum_{i=1}^{n} \sum_{x=1}^{m} f_{ix}}{\sum_{i=1}^{n} \hat{f}_{i0}\tau_i + \sum_{i=1}^{n} \sum_{x=1}^{m} (x+1) f_{ix} \tau_i}
\end{aligned}$$

Here, the numerator does depend on f_{i0} so does change with each iteration of the algorithm along with the denominator.

Application: Suicide data

For convergence level set to 10e-9, it takes 12 iterations for the EM algorithm reach convergence for the case study data assuming a Poisson distribution, resulting in an estimated rate of completed suicide, $\exp(\hat{\eta})$, of 0.00031752. This is the same rate computed using the function `zerotrunc` from the package `countreg` in `R` which models the zero-truncated Poisson regression model with person-years as an exposure variable

seen in Section 3.2.5. In context, this can be described as a rate of completed suicide of 31.8/100,000 person-years, which is considerably lower than the rates computed in Section 2.1. Given that the rate computed here is the same as found through using the more flexible and efficient zero-truncated regression models to find the MLE, the EM algorithm is not necessarily the best approach for finding the prevalence rate, especially if the preferred model were to have covariate information included. Since the EM algorithm is not helpful in this situation, taking a long time to code, inflexible for covariate information and less efficient compared to other approaches, it is not recommended to use the EM algorithm for for estimating the prevalence rate, and instead use a likelihood approach such as those taken in the `zerotrunc` and `vglm` functions in R, which are more flexible and less computationally intensive.

Application: Hares data

Unlike with the suicide case study, the prevalence rate is not the focus for the snowshoe hares dataset, instead the total number of snowshoe hares both observed and unobserved is the quantity of interest to estimate. The focus of the EM algorithm is to compute the maximum likelihood estimate value for the parameter $\exp(\eta)$, but given that the unobserved frequency of zero counts is estimated in the E-step of the algorithm, the total number of snowshoe hares can be estimated through using this method.

Assuming a geometric distribution with a convergence level of $10e-9$, the EM algorithm takes 15 iterations to reach convergence, resulting in an estimated total number of snowshoe hares of 2077. However, given that the focus of the EM algorithm is not the total population size, and that there is more flexibility in using estimation methods which relying on modelling that can fit the data better, alternative methods such as the Horvitz-Thompson estimator below are better for estimating the population size.

4.2 Horvitz-Thompson estimator

The estimated number of studies with zero counts of completed suicide for each covariate combination is calculated from the EM algorithm (in the E-step), as a by-product, but is of interest itself. To find this value, alternative methods of finding the MLE to the EM algorithm can be used. One possible method for estimating the total number of studies, N, that can be used was proposed by Horvitz and Thompson (1952). Given an indicator variable, I_i, for $i = 1, 2, \cdots, N$, where

$$I_i = \begin{cases} 1 & \text{study } i \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

the observed population size can be written as $n = \sum_{i=1}^{N} I_i$, with expectation $E\left(\sum_{i=1}^{N} I_i\right) = E(n) = n$. Additionally, the probability of a study being unobserved is $p_0(\hat{\mu}_i)$. Therefore, the probability of a study being observed (the inclusion probability) is $1 - p_0(\hat{\mu}_i)$, so, $E(I_i) = (1 - p_0(\hat{\mu}_i))$.

Following the work of Horvitz and Thompson (1952), the population total, $T = \sum_{i=1}^{N} y_i$, where y_i is a measurement associated with element i for $i = 1, 2, \ldots, N$, can be estimated as

$$\widehat{T} = \sum_{i=1}^{N} I_i w_i y_i,$$

where w_i is a constant that acts as a weight whenever the ith element is observed.

$\widehat{T}$ is required to be an unbiased estimator, so

$$E(\widehat{T}) = T,$$

leading to

$$\sum_{i=1}^{N} (1 - p_0(\hat{\mu}_i)) w_i y_i = \sum_{i=1}^{N} y_i,$$

hence, $w_i = (1 - p_0(\hat{\mu}_i))^{-1}$ as $E(I_i) = (1 - p_0(\hat{\mu}_i))$.

Therefore,

$$\begin{aligned} \widehat{T} &= \sum_{i=1}^{N} \frac{I_i y_i}{1 - p_0(\hat{\mu}_i)} \\ &= \sum_{i=1}^{n} \frac{y_i}{1 - p_0(\hat{\mu}_i)}, \end{aligned} \tag{4.8}$$

is the Horvitz-Thompson estimator for estimating the population size.

The population size which is estimated depends on the value of y_i, where there are 3 cases, seen in Overton and Stehman (1995), for $i = 1, 2, \cdots, N$ as follows:

Case 1: $\widehat{T} = N$, the total number of units when, $y_i = 1$. Note that if the expected count value is constant across all i, the estimated total number of elements is calculated as

$$\widehat{N}^{(HT)} = \frac{n}{1 - p_0(\hat{\mu})}.$$

Case 2: $\widehat{T} = N_A$, the number of units in sub-population A when,

$$y_i = \begin{cases} 1 & \text{for } i \in A, \\ 0 & \text{for } i \notin A. \end{cases}$$

Case 3: $\widehat{T} = N_{y_A}$, the total of y in sub-population A when,

$$y_i = \begin{cases} y_i & \text{for } i \in A, \\ 0 & \text{for } i \notin A. \end{cases}$$

Therefore, for estimating the total number of missing individuals (units), Case 1 is applied, leading to the Horvitz-Thompson estimator seen in McCrea and Morgan (2014, Chapter 3) and Borchers et al. (2002, Chapter 11) as follows

$$\widehat{N}^{(HT)} = \sum_{i=1}^{n} \frac{1}{1 - p_0(\hat{\mu}_i)},$$

for the Poisson and geometric distributions, and

$$\widehat{N}^{(HT)} = \sum_{i=1}^{n} \frac{1}{1 - p_0(\hat{\mu}_i, \theta)},$$

for the negative-binomial distribution, where $\hat{\mu}_i = \tau_i \exp\left(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}\right)$ and θ is the over-dispersion parameter.

Assuming a Poisson distribution, where $p_0(\hat{\mu}_i) = \exp(-\hat{\mu}_i)$ and $\hat{\mu}_i = \tau_i \exp\left(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}\right)$, the applied Horvitz-Thompson estimator (see van der Heijden et al., 2003, for more information) is

$$\widehat{N}^{(HT)} = \sum_{i=1}^{n} \frac{1}{1 - \exp(-\hat{\mu}_i)}.$$

Whilst not preferred for the either the suicide or hares case study data, the negative-binomial model can be assumed for the Horvitz-Thompson estimator, where $p_0(\hat{\mu}_i, \theta) = \left(\frac{\theta}{\hat{\mu}_i + \theta}\right)^{\theta}$ and $\hat{\mu}_i = \tau_i \exp\left(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}\right)$ (see Cruyff and van der Heijden, 2008, for more information), leading to

$$\widehat{N}^{(HT)} = \sum_{i=1}^{n} \frac{1}{1 - \left(\frac{\theta}{\hat{\mu}_i + \theta}\right)^{\theta}}.$$

If a geometric distribution is assumed for the hares case study data, the Horvitz-Thompson estimator is given as

$$\widehat{N}^{HT} = \sum_{i=1}^{n} \frac{1}{1 - \frac{1}{1+\hat{\mu}_i}}$$

Given that $\widehat{N} = n + \hat{f}_0$, and let $\widehat{M} = \hat{f}_0$, the number of missing studies, $\widehat{M}$, is estimated as

$$\widehat{M} = \widehat{N} - n.$$

Sup-population specific population sizes can also be estimated given that the Horvitz-Thompson estimator allows for covariates to be included using Case 2. Therefore, the estimated size of sub-population A, given linear predictor j, is calculated as

$$\widehat{N}_A^{(HT)} = \sum_{i \in A} \frac{1}{1 - p_0\left(\tau_i \exp\left(\mathbf{h}_j(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_j\right)\right)},$$

for the Poisson or geometric distributions, and

$$\widehat{N}_A^{(HT)} = \sum_{i \in A} \frac{1}{1 - p_0\left(\tau_i \exp\left(\mathbf{h}_j(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_j\right), \theta\right)},$$

for the negative-binomial distribution, where θ is the dispersion parameter.

Additionally, the number of missing studies for each sub-population is estimated by

$$\widehat{M}_A = \widehat{N}_A - n_A,$$

where n_A is the observed population size for sub-population A.

Application: Suicide data

For the case study data, the number of missing studies found using the Horvitz-Thompson estimator assuming the Poisson distribution is estimated to be $\widehat{M}^{HT} = 107$ as $\widehat{N}^{(HT)} = 134$ and $n = 27$.

Additionally, for the suicide case study data, eight sub-populations are considered, defined by the combination of country of origin being either USA or other and the proportion of women being in $[0, 0.75)$, $[0.75, 0.80)$, $[0.80, 0.85)$ or $[0.85, 1]$, where these intervals are defined by the approximate quantiles of the variable.

TABLE 4.1: Estimates for the number of missing studies in the eight sub-populations with the number of observed studies shown in [square brackets] for the suicide case study data.

	Proportion of women				
Country of origin	$[0, 0.75)$	$[0.75, 0.80)$	$[0.80, 0.85)$	$[0.85, 1]$	Total
USA	0 [1]	0 [0]	22 [6]	8 [3]	30 [10]
Other	42 [6]	23 [4]	7 [4]	5 [3]	77 [17]
Total	42 [7]	23 [4]	29 [10]	13 [6]	107 [27]

Table 4.1 displays the estimated number of missing studies overall as well as the number of missing studies for each of the eight sub-populations, assuming the preferred intercept-only zero-truncated Poisson model. Sub-population specific estimates of the number of missing studies show that it is unlikely that many of the missing studies

which have a proportion of women less than 80% originate from the USA. This is reasonable given that only 10% of observed studies from the USA have a proportion of women less than 80% compared to 59% in observed studies originating from outside the USA. Consequently, it is also reasonable that it is more likely for missing studies that originate from outside the USA to have a proportion of women less than 80%.

Application: Hares data

For the hares case study data, the number of missing snowshoe hares, is estimated using the Horvitz-Thompson estimator assuming the geometric distribution to be $\widehat{M}^{HT} = 2140$ as $\widehat{N}^{(HT)} = 3123$ and $n = 983$.

TABLE 4.2: Estimates using the Horvitz-Thompson estimator for the number of missing snowshoe hares in the sub-populations with the observed number of snowshoe hares trapped shown in [square brackets].

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	301 [95]	298 [181]	614 [261]	1212 [537]
Five small areas	121 [94]	137 [125]	669 [227]	927 [446]
Total	422 [189]	435 [306]	1283 [488]	2140 [983]

Table 4.2 displays the estimated number of missing snowshoe hares overall as well as the number of observed hares for each of the six sub-populations, assuming the preferred geometric model with both covariate effects and an interaction effect.

4.3 Chao's Lower Bound estimator

Capture-recapture methods have been in common use for estimating the population size and investigating the dynamics of biological populations for a long time. Prior to Chao (1987) however, many of the previous works relied on the equal-probability-of-capture assumption, which is often deemed as unattainable in biological populations (Carothers, 1973). To deal with this unobserved heterogeneity arising from unobserved differences in factors such as the mental health of an individual or the proficiency of a medical professional, Chao's Lower Bound estimator is a capture-recapture approach for estimating the lower bound of a target population's size (Chao, 1987, 1989).

In application, unpredictability exists, which leads to variation in the capture probabilities, typically rendering the simple model, $p_x(\mu)$, insufficient given that it is not flexible enough to adequately represent this variation. Instead, a more flexible mixture probability density can be utilised to model the population heterogeneity as follows.

$$k_x(\mu) = \int_0^\infty p_x(\mu) q(\mu) d\mu, \tag{4.9}$$

where $q(\mu)$ is the mixing density, $p_x(\mu) = \frac{\exp(-\mu)\mu^x}{x!}$ is the Poisson mixture kernel (Böhning et al., 2013b), and $p_x(\mu) = (1-\mu)^x\mu$ is the geometric mixture kernel (Niwitpong et al., 2013). Here, the mixing density q is unknown so the lower bound approach is taken through estimation as without knowing the value of the density, the mixture probability density cannot directly be used in modelling to estimate p_0 to use in the Horvitz-Thompson (or other) estimator for population size estimation.

Using the mixture from Equation 4.9, the ratios of neighbouring probabilities, q_x, and the corresponding (known) coefficients, a_x, have the following relationship.

$$r_x = \frac{a_x}{a_{x+1}}\frac{q_{x+1}}{q_x} = \mu, \tag{4.10}$$

with the following monotonicity

$$r_x \leq r_{x+1}. \tag{4.11}$$

The ratio in Equation 4.10 can be adjusted to provide the ratio of the mixtures as

$$r_x = \frac{a_x}{a_{x+1}}\frac{k_{x+1}}{k_x},$$

which can be approximated as

$$r_x = \frac{a_x}{a_{x+1}}\frac{f_{x+1}}{f_x}, \tag{4.12}$$

by substituting the sample estimates $\frac{f_x}{N}$ for the theoretical quantities of the mixtures, k_x.

Applying Equations 4.12 and 4.11 leads to the following inequality.

$$\begin{aligned}\frac{a_0}{a_1}\frac{f_1}{f_0} &\leq \frac{a_1}{a_2}\frac{f_2}{f_1}\\ f_0 &\geq \frac{a_0 a_2}{a_1^2}\frac{f_1^2}{f_2},\end{aligned}$$

hence the lower bound of the frequency of zeros can be estimated as

$$\hat{f}_0 = \frac{a_0 a_2}{a_1^2}\frac{f_1^2}{f_2}, \tag{4.13}$$

where the value of the corresponding coefficients, a_x are dependent on the distribution and given as

$$a_x = \begin{cases}\frac{1}{x!} \text{ if Poisson, or}\\ 1 \text{ if geometric.}\end{cases} \tag{4.14}$$

Alternatively, an estimate for the lower bound of the frequency of zero counts can be found using the Cauchy-Schwarz inequality, leading to

$$\left(\int_0^\infty \exp(-\mu)\mu q(\mu)d\mu\right)^2 \leq \int_0^\infty \exp(-\mu)q(\mu)d\mu \times \int_0^\infty \exp(-\mu)\mu^2 q(\mu)d\mu. \quad (4.15)$$

in the case of a Poisson mixture kernel and

$$\left(\int_0^\infty (1-\mu)\mu q(\mu)d\mu\right)^2 \leq \int_0^\infty \mu q(\mu)d\mu \times \int_0^\infty (1-\mu)^2\mu q(\mu)d\mu. \quad (4.16)$$

in the case of a geometric mixture kernel.

Therefore, the inequality in Equation 4.15 (Poisson) can be given by $k_1(\mu)^2 \leq k_0(\mu) \times 2k_2(\mu)$ and Equation 4.16 (geometric) can be given by $k_1(\mu)^2 \leq k_0(\mu) \times k_2(\mu)$ using theoretical probabilities. Given that p_x has the sample estimate $\frac{f_x}{N}$, Equations 4.15 and 4.16 can also be given in terms of frequencies by

$$\hat{f}_0 \geq \begin{cases} \frac{f_1^2}{2f_2} \text{ if Poisson mixture kernel, or} \\ \frac{f_1^2}{f_2} \text{ if geometric mixture kernel,} \end{cases} \quad (4.17)$$

providing an estimated lower bound for the frequency of zero counts, where f_x is the frequency of exactly x counts and supporting the values of the corresponding coefficients given in Equation 4.14.

Given that $\widehat{N} = n + \hat{f}_0$, the estimated total population size is found by

$$\widehat{N}^{(C)} = \begin{cases} n + \frac{f_1^2}{2f_2} \text{ if Poisson mixture kernel, or} \\ n + \frac{f_1^2}{f_2} \text{ if geometric mixture kernel.} \end{cases} \quad (4.18)$$

Application: Suicide data

Using the case study data and Equation 4.18 assuming the Poisson mixture kernel, the total population size can then be estimated as

$$\widehat{N}^{(C)} = 27 + \frac{18^2}{2 \times 3} = 81,$$

which is much smaller than the estimate provided by the Horvitz-Thompson estimator in Section 4.2 as this conventional method does not allow for the inclusion of covariates and that Chao's estimator is a lower bound. For example, in this case study application, person-years is not considered which leads to an underestimation of the population size.

This exclusion of covariates also increases the bias of the estimator and does not allow for sub-population sizes to be estimated.

Application: Hares data

Using the hares case study data, the total number of snowshoe hares can then be estimated using the conventional Chao's estimator assuming a geometric mixture kernel as

$$\widehat{N}^{(C)} = 983 + \frac{653^2}{210} = 3013.519 \approx 3014,$$

which is slightly lower than the total estimate of $\widehat{N}^{(HT)} = 3123$ provided by the Horvitz-Thompson estimator. This finding is understandable, given that the conventional Chao's estimator is a lower bound estimator.

4.4 Generalised Chao's estimator

Böhning et al. (2013b) developed a generalised form of Chao's estimator to allow for covariate information. It is a generalised form as if there are no covariates available, then it is identical to the conventional form of Chao's estimator.

Assuming the Poisson regression with $\mu_i = \tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \boldsymbol{\beta})$, the population heterogeneity modelled in Equation 4.9 can be accounted for. Truncating all counts besides $X = 1$ and $X = 2$ leads to the associated truncated Poisson model as follows.

$$p_1(\mu_i) = 1 - q_i \text{ and } p_2(\mu_i) = q_i, \tag{4.19}$$

Given that $q_1 = (1 - q)$, $q_2 = q$ and $a_x = \frac{1}{x!}$ for the Poisson mixture kernel, Equation 4.10 becomes

$$\mu = \frac{a_1}{a_2}\frac{q_2}{q_1} = 2\frac{q}{1-q},$$

Replacing $p_2(\mu) = q$ with its sample estimate $\frac{f_2}{N}$ where $N = (f_1 + f_2)$ makes μ equivalent to

$$\mu = 2\frac{f_2/(f_1+f_2)}{1 - f_2/(f_1+f_2)} = 2\frac{f_2}{f_1},$$

verifying that the ratios in Equations 4.10 and 4.12 are equal.

Rearranging for $\hat{q}$ results in

$$\hat{q} = \frac{\hat{\mu}}{2 + \hat{\mu}},$$

making the probabilities in Equation 4.19 equivalent to

$$p_1(\mu_i) = \frac{2}{2+\mu_i} \text{ and } p_2(\mu_i) = \frac{\mu_i}{2+\mu_i}.$$

The associated truncated Poisson likelihood is

$$L = \prod_{i=1}^{f_1+f_2} \left(\frac{2}{2+\mu_i}\right)^{f_{i1}} \times \left(\frac{\mu_i}{2+\mu_i}\right)^{f_{i2}},$$

equal to the standard binomial logistic likelihood

$$L = \prod_{i=1}^{f_1+f_2} (1-q_i)^{f_{i1}} \times (q_i)^{f_{i2}}.$$

Given that the likelihood of the truncated Poisson model is equal to that of the normal binomial logistic, a logistic regression model can be utilised to find the associated maximum likelihood estimates. Hence,

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, (f_1 + f_2)$, where $\hat{q}_i$ are the fitted values of the logistic regression model.

The estimated frequency of zero counts can be found using the maximum likelihood estimates as follows.

$$\hat{f}_0 = \sum_{i=1}^{f_1+f_2} \frac{f_{i1}+f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}. \tag{4.20}$$

Given that $\widehat{N} = n + \hat{f}_0$, the estimated total population size is found by

$$\widehat{N}^{(GC)} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1}+f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}, \tag{4.21}$$

where $\widehat{N}_{GC} \geq \widehat{N}_C$.

However, assuming the geometric mixture kernel with $\mu_i = \exp(\boldsymbol{h}(\boldsymbol{v}_i)^T\boldsymbol{\beta})$, the population heterogeneity modelled in Equation 4.9 can be accounted for following the same approach as assuming the Poisson mixture kernel.

Truncating all counts except the singletons and doubletons leads to the following associated truncated geometric model.

$$p_1(\mu_i) = 1 - q_i \text{ and } p_2(\mu_i) = q_i, \tag{4.22}$$

Given that $q_1 = (1-q)$, $q_2 = q$ and $a_x = 1$ for the geometric mixture kernel, Equation 4.10 becomes

$$\mu = \frac{a_1}{a_2}\frac{q_2}{q_1} = \frac{q}{1-q},$$

Replacing $p_2(\mu) = q$ with its sample estimate $\frac{f_2}{N}$ where $N = (f_1 + f_2)$ makes μ equivalent to

$$\mu = \frac{f_2/(f_1+f_2)}{1 - f_2/(f_1+f_2)} = \frac{f_2}{f_1},$$

verifying that the ratios in Equations 4.10 and 4.12 are equal.

Rearranging for $\hat{q}$ results in

$$\hat{q} = \frac{\mu}{1+\mu},$$

making the probabilities in Equation 4.22 equivalent to

$$p_1(\mu_i) = \frac{1}{1+\mu_i} \text{ and } p_2(\mu_i) = \frac{\mu_i}{1+\mu_i}.$$

The associated truncated geometric likelihood is

$$L = \prod_{i=1}^{f_1+f_2} \left(\frac{1}{1+\mu_i}\right)^{f_{i1}} \times \left(\frac{\mu_i}{1+\mu_i}\right)^{f_{i2}},$$

equal to the standard binomial logistic likelihood

$$L = \prod_{i=1}^{f_1+f_2} (1-q_i)^{f_{i1}} \times (q_i)^{f_{i2}}.$$

Given that the likelihood of the truncated geometric model is equal to that of the normal binomial logistic, a logistic regression model can be utilised to find the associated maximum likelihood estimates. Hence,

$$\hat{\mu}_i = \frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, (f_1 + f_2)$, where $\hat{q}_i$ are the fitted values of the logistic regression model.

The estimated frequency of zero counts can be found using the maximum likelihood estimates as follows.

$$\begin{aligned}
\hat{f}_0 = \sum_{i=1}^{f_1+f_2} E[f_{i0}|f_{i2}, f_{i3}, q] &= \sum_{i=1}^{f_1+f_2} \frac{p_0(\hat{\mu}_i)}{p_1(\hat{\mu}_i) + p_2(\hat{\mu}_i)}(f_{i1}+f_{i2}) \\
&= \frac{f_{i1}+f_{i2}}{\left(1 - \frac{1}{1+\hat{\mu}_i}\right) + \left(1 - \frac{1}{1+\hat{\mu}_i}\right)^2} \\
&= \frac{f_{i1}+f_{i2}}{\left(1 - \frac{1}{1+\hat{\mu}_i}\right)\left(2 - \frac{1}{1+\hat{\mu}_i}\right)}.
\end{aligned} \tag{4.23}$$

Given that $\widehat{N} = n + \hat{f}_0$, the estimated total population size is found by

$$\widehat{N}^{(GC)} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{(1-\hat{\zeta}_i)(2-\hat{\zeta}_i)}, \tag{4.24}$$

where $\hat{\zeta}_i = \frac{1}{1+\hat{\mu}_i} = 1 - \hat{q}_i$ and $\widehat{N}_{GC} \geq \widehat{N}_C$.

Application: Suicide data

Applying the generalised Chao's estimator assuming the Poisson mixture kernel using Equation 4.24 to the suicide case study data leads to the estimated population size

$$\widehat{N}^{(GC)} = 27 + \sum_{i=1}^{21} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2} = 172.659 \approx 173,$$

which is much larger than the estimate using the conventional form of 81 total studies. This is expected given that the conventional Chao's estimator is a lower bound estimator and doesn't account for the covariate information available. However, this is also much larger than the estimate found from the Horvitz-Thompson estimator in Section 4.2 of 134. This difference may be as a result of the small sample size of the suicide data which is utilised in the generalised Chao's estimator, leading to a possible reduction in accuracy. This difference motivates the use of the simulation study later in this chapter, comparing the performance of the different capture-recapture estimators.

Application: Hares data

Applying the generalised Chao's estimator to the hares case study data assuming the geometric mixture kernel using Equation 4.24 leads to the following estimated population size.

$$\widehat{N}^{(GC)} = 983 + \sum_{i=1}^{863} \frac{f_{i1} + f_{i2}}{\hat{\zeta}_i + \hat{\zeta}_/2} = 3890.141 \approx 3890,$$

which is considerably larger than that found through the conventional approach, an expected result given that covariate information is now included and the conventional approach being a lower bound estimator. As with the suicide data, the generalised Chao's estimate is also larger than the Horvitz-Thompson estimate ($\widehat{N}^{(HT)} = 3123$). This difference and common trend motivates the use of the simulation study in Section 4.7 to assess the performance of the different estimators and check the accuracy of the Horvitz-Thompson estimator to ensure that it does not underestimate the total population size.

4.5 Zelterman's estimator

In many situations, the assumption that the the entire range of observed counts follow the chosen distribution cannot be met, leading to unrepresentative and unreliable estimates. For the Horvitz-Thompson estimator, the chosen distribution is assumed to be the Poisson distribution, where large count values in particular are more susceptible to deviation. An alternative approach developed by Zelterman (1988) relaxes this assumption by instead assuming that only a small range of count values follow the Poisson distribution, similarly to Chao's estimator. Two consecutive frequencies (f_x and f_{x+1}) are used to estimate the count value as

$$\hat{\mu}_x = \frac{(x+1)f_{x+1}}{f_x}.$$

Typically $x = 1$, as the frequencies f_1 and f_2 are in close proximity to f_0, the target value needing predicting. Additionally, in the majority of cases, f_1 and f_2 have the highest frequencies for the data (Böhning and van der Heijden, 2009). Setting $x = 1$ leads to the case where all counts are truncated except $X = 1$ and $X = 2$ with the corresponding truncated Poisson distribution probabilities

$$q_1 = \frac{2}{2+\mu} \text{ and } q_2 = \frac{\mu}{2+\mu}.$$

The truncated Poisson distribution leads to a binomial likelihood (Böhning et al., 2013b), as seen in Section 4.4, with log-likelihood

$$\ell = f_1 \log(q_1) + f_2 \log(q_2),$$

which is maximised as

$$\hat{q}_2 (= 1 - \hat{q}_1) = \frac{\hat{\mu}}{2+\hat{\mu}} = \frac{f_2}{f_1 + f_2}.$$

Therefore, Zelterman (1988) proposes that the estimated count is calculated as

$$\hat{\mu} = \frac{2f_2}{f_1},$$

and used in the Horvitz-Thompson estimator given in Section 4.2 as

$$\widehat{N}^{(Z)} = \frac{n}{1 - p_0(\hat{\mu})} = \frac{n}{1 - \exp(-\hat{\mu})}.$$

The standard Zelterman's estimator relies on the assumption that a small range of consecutive counts follow the Poisson distribution. To adjust Zelterman's estimator to reflect the assumption that a small range of consecutive counts follow a geometric

instead, the ratio of neighbouring zero-truncated geometric probabilities can be used as follows, given the zero-truncated geometric density in Equation 3.2.

$$\begin{aligned}\frac{p_{x+1}(\mu)}{p_x(\mu)} &= \frac{(1-\mu_i)^x\mu_i}{(1-\mu_i)^{x-1}\mu_i}\\ &= (1-\mu_i).\end{aligned}$$

The probabilities can be replaced with the corresponding sample estimates $\frac{f_x}{N}$, leading to the following estimate.

$$\frac{f_{x+1}}{f_x} = 1-\hat{\mu}_i,$$

where $\mu_i = p_0(\mu_i)$.

Therefore, the estimated probability of zero given μ_i is given as

$$p_0(\hat{\mu}_i) = 1-\frac{f_{x+1}}{f_x}.$$

As with the standard Zelterman's estimator, given that the frequencies f_1 and f_2 are in close proximity to f_0, these consecutive frequencies are chosen to estimate the total probabilities. The estimate of the probability of zero counts is then given by

$$p_0(\hat{\mu}_i) = 1-\frac{f_2}{f_1}.$$

Therefore, using the Horvitz-Thompson estimator as with the standard Zelterman's estimator, the adjusted Zelterman's estimator based on the geometric distribution is given as follows.

$$\begin{aligned}\widehat{N}^{(Z)} &= \frac{n}{1-p_0(\hat{\mu}_i)}\\ &= \frac{n}{1-\left(1-\frac{f_2}{f_1}\right)}\\ &= \frac{nf_1}{f_2}.\end{aligned}$$

Application: Suicide data

For the suicide case study data, assuming a Poisson distribution, the estimated count value is calculated as

$$\hat{\mu} = \frac{2\times 3}{18} = \frac{1}{3},$$

leading to an estimated total number of studies of

$$\widehat{N}^{(Z)} = \frac{27}{1-\exp(-\frac{1}{3})} = 95.24861 \approx 95.$$

This is much smaller than the estimated target population size provided by the Horvitz-Thompson estimator in Section 4.2. However, as with the conventional Chao's Lower Bound estimator, the conventional Zelterman's estimator does not account for person-years or covariates, which can lead to an inaccurate estimated population size.

Application: Hares data

For the hares case study data, the estimated total number of snowshoe hares using Zelterman's estimator assuming the geometric distribution is given as

$$\widehat{N}^{(Z)} = \frac{983 \times 653}{420} = 3056.662 \approx 3057.$$

This is notably larger than the Horvitz-Thompson estimate, however, only slightly larger than the conventional Chao's estimator.

4.6 Generalised Zelterman's estimator

The generalised approach to Zelterman's estimator developed by Böhning and van der Heijden (2009) accounts for covariates, in order to estimate the target population size more accurately. Through some modification to the model specification, the generalised Zelterman's estimator can also account for an exposure variable.

As in Section 4.4, truncating all counts besides $X = 1$ and $X = 2$ leads to the truncated Poisson likelihood being equivalent to the binomial logistic likelihood as follows.

$$\prod_{i=1}^{f_1+f_2} \left(\frac{2}{2+\mu_i}\right)^{f_{i1}} \times \left(\frac{\mu_i}{2+\mu_i}\right)^{f_{i2}} = \prod_{i=1}^{f_1+f_2} (1-q_i)^{f_{i1}} \times (q_i)^{f_{i2}},$$

where

$$q_i = \frac{\mu_i}{2+\mu_i} = \frac{\mu_i/2}{1+\mu_i/2} \text{ and } q_i = \frac{\tau_i \exp(\eta_i)}{1+\tau_i \exp(\eta_i)}.$$

The maximum likelihood estimates are found using a logistic regression model, resulting in the generalised Zelterman's estimator for the MLE as follows.

$$\hat{\mu}_i = 2\tau_i \exp(\hat{\eta}_i), \tag{4.25}$$

for $i = 1, 2, \cdots, n$.

Given that $\hat{\eta}_i = \mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}$, Equation 4.25 allows for both an exposure variable and covariates to be accounted for in the calculation for the expected count values. As with the conventional Zelterman's estimator, the Horvitz-Thompson estimator seen in

Section 4.2 is used with the generalised Zelterman's estimator to estimate the target population size as follows.

$$\begin{aligned}\widehat{N}^{(GZ)} &= \sum_{i=1}^{n} \frac{1}{1-\exp(-\hat{\mu}_i)} \\ &= \sum_{i=1}^{n} \frac{1}{1-\exp(-2\tau_i \exp(\hat{\eta}_i))}.\end{aligned}$$

The existing generalised Zelterman's estimator can be altered to allow for the assumption of the geometric distribution, instead of the Poisson distribution as above, through following the same key steps.

As seen in Section 4.4, truncating all counts besides the singletons and the doubletons leads to the following associated truncated geometric model.

$$p_1(\mu_i) = (1-\mu_i)\mu_i = 1 - q_i \text{ and } p_2(\mu_i) = (1-\mu_i)^2\mu_i = q_i.$$

The resulting truncated geometric likelihood is then equal to the binomial logistic likelihood as follows.

$$\prod_{i=1}^{f_1+f_2} \left(\frac{1}{1+\mu_i}\right)^{f_{i1}} \times \left(\frac{\mu_i}{1+\mu_i}\right)^{f_{i2}} = \prod_{i=1}^{f_1+f_2} (1-q_i)^{f_{i1}} \times (q_i)^{f_{i2}},$$

where

$$q_i = \frac{\mu_i}{1+\mu_i} \text{ and } q_i = \frac{\tau_i \exp(\eta_i)}{1+\tau_i \exp(\eta_i)}.$$

Using a logistic regression model to find the maximum likelihood estimates, the generalised Zelterman's estimator is given as

$$\hat{\mu}_i = \tau_i \exp(\hat{\eta}_i), \tag{4.26}$$

for $i = 1, 2, \cdots, n$.

As with the approach for assuming the Poisson distribution, given that $\hat{\eta}_i = \mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}}$, Equation 4.26 allows for both an exposure variable and covariates to be accounted for in the calculation for the expected count values, unlike with the conventional approach to Zelterman's estimator. To estimate the total population size, the Horvitz-Thompson estimator seen in Section 4.2 can be used used with the generalised Zelterman's estimator

to estimate the target population size as follows.

$$\begin{aligned}\widehat{N}^{(GZ)} &= \sum_{i=1}^{n} \frac{1}{1-\frac{1}{1+\hat{\mu}_i}} \\ &= \sum_{i=1}^{n} \frac{1+\hat{\mu}_i}{\hat{\mu}_i} \\ &= \sum_{i=1}^{n} \frac{1}{\hat{\mu}_i} + 1 \\ &= \sum_{i=1}^{n} \frac{1}{\tau_i \exp(\hat{\eta}_i)} + 1.\end{aligned}$$

Alternatively, the generalised Zelterman's estimator, assuming a geometric distribution, can be written as

$$\widehat{N}^{(GZ)} = \sum_{i=1}^{n} \frac{1}{1-\hat{\zeta}_i},$$

where $\hat{\zeta}_i = \dfrac{1}{1+\hat{\mu}_i}$ for $\hat{\mu}_i = \tau_i \exp(\hat{\eta}_i)$.

Application: Suicide data

Applying the generalised Zelterman's estimator to the suicide case study data, assuming a Poisson distribution, leads to an estimated total number of studies of

$$\widehat{N}^{(GZ)} = \sum_{i=1}^{27} \frac{1}{1-\exp(-2\tau_i \exp(\hat{\eta}_i))} = 175.1877 \approx 175.$$

The resulting estimate is comparable to the results from the generalised Chao's estimator but much higher than the conventional Zelterman's estimator without covariates, and the estimate using the Horvitz-Thompson estimator in Section 4.2 which assumes that all counts follow the Poisson distribution.

Application: Hares data

Applying the generalised Zelterman's estimator to the hares case study data, assuming a geometric distribution, leads to an estimated total number of snowshoe hares of

$$\widehat{N}^{(GZ)} = \sum_{i=1}^{983} \frac{1+\hat{\mu}_i}{\hat{\mu}_i} = 3601.666 \approx 3602.$$

Whilst the difference between the generalised Chao's and generalised Zelterman's estimates is not large, the generalised Zelterman's estimate is slightly smaller than the generalised Chao's estimate. However, it is notably larger than the results from the

Horvitz-Thompson estimator as seen with the suicide case study data. This difference between the generalised Zelterman's and Horvitz-Thompson estimators is due to the more relaxed distributional assumption seen in the development of the generalised Zelterman's estimator.

4.7 Simulation study

4.7.1 Definitions

It is important to consider the performance of each population size estimators to understand which estimator to utilise in different situations, where the performance measures are accuracy, precision, coverage and robustness, given formally in Definitions 4.1, 4.2, 4.3 and 4.4.

Definition 4.1 (Accuracy)**.** A type of observational error computed as

$$median(|\widehat{\boldsymbol{N}} - N|),$$

where N is the true population size and $\widehat{\boldsymbol{N}} = (\widehat{N}_1, \ldots, \widehat{N}_S)$ are the estimated population sizes for each iteration $s = 1, \ldots, S$ of the simulation study.

Definition 4.2 (Precision)**.** The degree of random error affiliated with the population size estimate computed as

$$median(\boldsymbol{CI}_U - \boldsymbol{CI}_L),$$

where $\boldsymbol{CI}_L = (CI_{L,1}, \ldots, CI_{L,S})$ and $\boldsymbol{CI}_U = (CI_{U,1}, \ldots, CI_{U,S})$ respectively are the lower and upper confidence intervals for each iteration $s = 1, \ldots, S$ of the simulation study. For this work, 95% percentile confidence intervals are used to compute the upper and lower confidence interval limits.

Definition 4.3 (Coverage)**.** The probability that the true population size, N, is contained within the confidence interval computed as

$$\frac{1}{S} \times \sum_{s=1}^{S} I_s \times 100\%,$$

where I_s is an indicator variable for $s = 1, \ldots, S$ defined as

$$I_s = \begin{cases} 1 & \text{if } CI_{L,s} \leq N \leq CI_{U,s}, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 4.4 (Robustness)**.** Tukey (1960), Huber (1964) and Hampel (1971) formed the foundations of robust statistics, where robustness is defined as the resilience of an

estimator to outliers. In this case determined by the comparison of values of accuracy, precision and coverage of estimators applied to data with and without outliers.

For computing the performance measures, the median is used in favour of the mean. Through preliminary work of performing simulation studies on a smaller scale, it was found that there was not a difference in the overall results when comparing using the mean versus the median. The median is favoured in this work for giving less weight to the extreme values, whilst still capturing the variation in results. For example, in Section 4.7.3, the Figures illustrate the large variation in results for the Horvitz-Thompson estimator, where the median is used for the computation of the accuracy and precision values. In the preliminary work, the mean was also used for these same measures, and produced the same results, but with wider confidence interval ranges due to the number of extreme values in the Horvitz-Thomson estimator's results, having the same impact on the overall conclusions. Therefore, it was not seen that using the median as the chosen measure of centrality would be an issue for the final results.

4.7.2 Methodology

As in Section 3.3, the following variables are required to simulate the data set that mimics the characteristics of the suicide case study data.

- N: total number of studies.
- $\bar{t}$: mean number of individuals per study.
- λ^C: constant rate of event.
- γ: logarithm of the mean for the observation period.
- σ: logarithm of the standard deviation for the observation period.

Section 3.3 does not include covariates in the simulation study, so to simulate covariates reflective of those in the suicide case study data additional information is required. The covariates of the case study data are numeric for the proportion of women, and binary for the country of origin. Therefore, the additional information required is as follows.

- α: shape parameter for the beta distribution to simulate proportion variable.
- β: shape parameter for the beta distribution to simulate proportion variable.
- ρ: success probability for the Bernoulli distribution to simulate binary variable.

Given the above variables, the data can be simulated as follows.

- Size of each study is sampled from the Poisson distribution, $t_i \sim Poisson(\bar{t})$.

- Observation period of each study is sampled from the log-normal distribution, $O_i \sim lognormal(\gamma, \sigma)$.
- Count of events for each study is sampled from the binomial distribution, $X_i \sim binomial(\tau_i, \lambda^C)$, where $\tau_i = t_i \times O_i$.
- Covariate for proportion for each study is sampled from the beta distribution, $v_{i1} \sim beta(\alpha, \beta)$.
- Binary covariate for each study is sampled from the Bernoulli distribution, $v_{i2} \sim Bernoulli(\rho)$.

To create a zero-truncated data set, counts of zero are truncated from the simulated data, from which the Horvitz-Thompson, generalised Chao's and generalised Zelterman's population size estimates can be calculated. The respective analytical variances for each of the population size estimators, used in constructing the confidence intervals, can be computed utilising a normal approximation approach, discussed further in Sections 5.3, 5.4 and 5.5. This process is repeated S times, with the resulting estimates and corresponding variances enabling the assessment of the various reliability measures. To compute the averages for accuracy and precision, the median is used as it gives less weight to outliers compared with the mean.

Given that the number of completed suicides is equal to the rate of completed suicide multiplied by the exposure variable for each study, outlier values of the counts can be simulated using outlier rates and sampled person-years. Outlier rates are sampled from the uniform distribution to allows for variability in rates across studies as seen in real data scenarios. Thus, in order to sample the outlier rates, a range from which to sample from requires specification, where the observed data is used to define what counts are classified as outliers. The lower bound is given by

$$\lambda^L = Q3 + 3 \times IQR, \tag{4.27}$$

where $Q3$ is the third quartile of the observed rates, IQR is the inter-quartile range and the IQR is multiplied by 3 to ensure the rates are clearly defined outliers. The upper limit is defined as

$$\lambda^U = 1.2 \times \lambda^L,$$

such that the range of rates is wide enough to allow for variation but not to provide unrealistic or highly improbable outliers.

To convert the outlier rates into outlier counts, the sampled rates are multiplied by sampled person-years, which are not sampled as outliers.

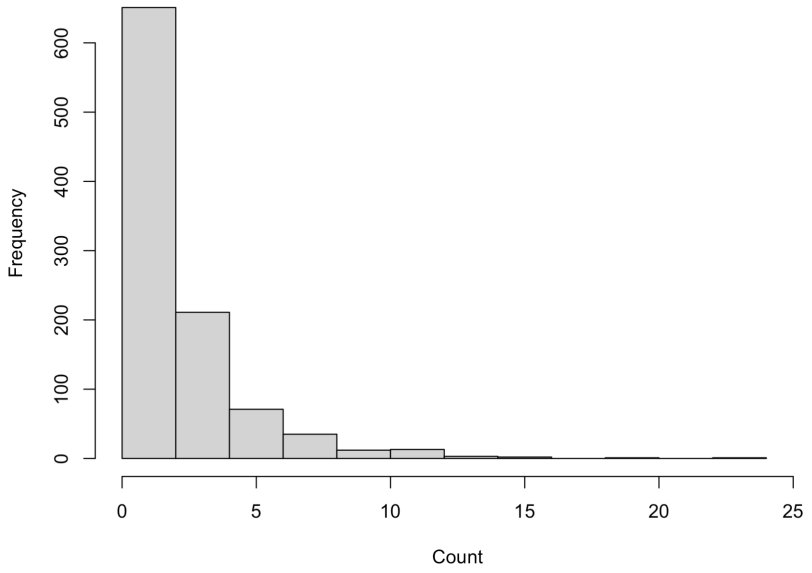

FIGURE 4.1: Histogram of simulated counts with N=1000 and 0.1% outliers

4.7.3 Results

A histogram of simulated counts with 0.1% outliers can be seen in Figure 4.1, illustrating that the outliers are not necessarily obvious, and that they can look like they fit the trend of the non-outlier counts.

TABLE 4.3: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's and generalised Zelterman's, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0071$, $\lambda^U \approx 0.0085$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	Horvitz-Thompson	16	26	133	287	845	2.56e+07
	Generalised Chao's	25	28	27	27	26	26
	Generalised Zelterman's	29	32	31	32	32	52
Precision	Horvitz-Thompson	95	99	118	149	345	6.7e+08
	Generalised Chao's	162	163	163	162	162	154
	Generalised Zelterman's	181	181	185	184	187	205
Coverage	Horvitz-Thompson	95.5%	76.9%	3.3%	0.0%	0.0%	99.4%
	Generalised Chao's	96.4%	96.0%	96.4%	96.7%	95.7%	96.0%
	Generalised Zelterman's	95.7%	94.7%	95.8%	96.7%	94.8%	89.5%

Table 4.3 gives the values of accuracy, precision and coverage for the Horvitz-Thompson, generalised Chao's and generalised Zelterman's estimators when $N = 1000$ for proportions of outlier counts varying from 0% to 10%.

It can be seen that when all counts follow the distributional assumption, there is negligible difference between the values for coverage for each estimator, and each value being preferable at over 95%. However, there are notable differences in the values for accuracy

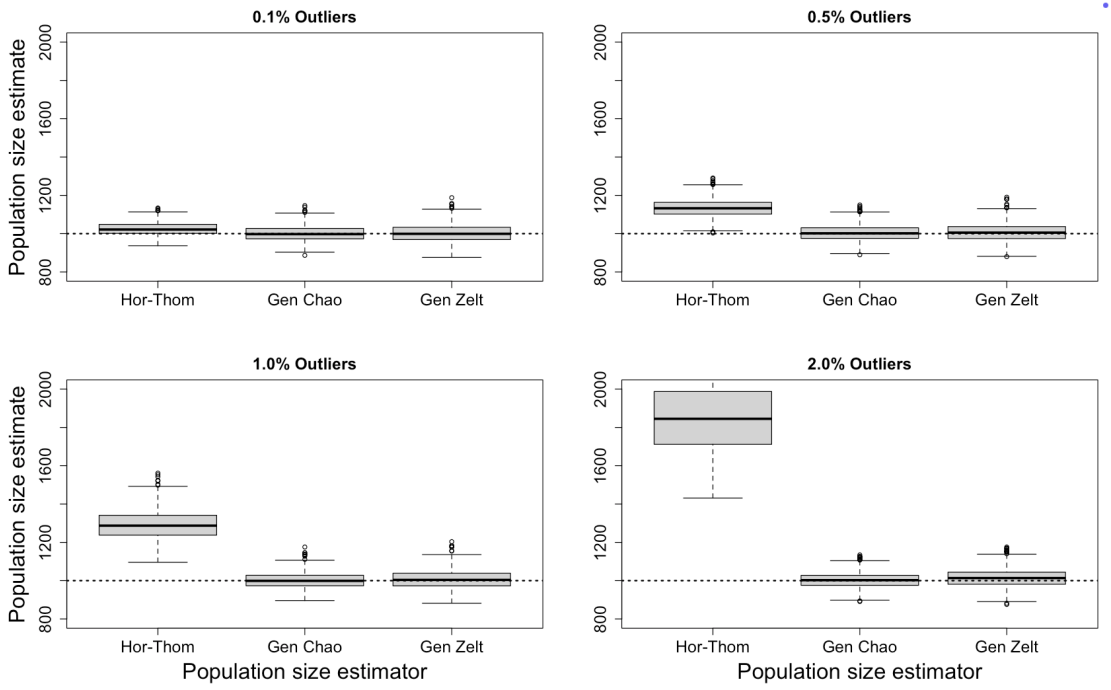

FIGURE 4.2: Box plots with the (individual) population size estimates from the simulation study, with a dotted line illustrating where the true value lies for illustrating the **accuracy** of the different capture-recapture estimators for different proportions of outliers when $N = 1000$ and $\lambda_L = Q3 + 3 \times IQR$.

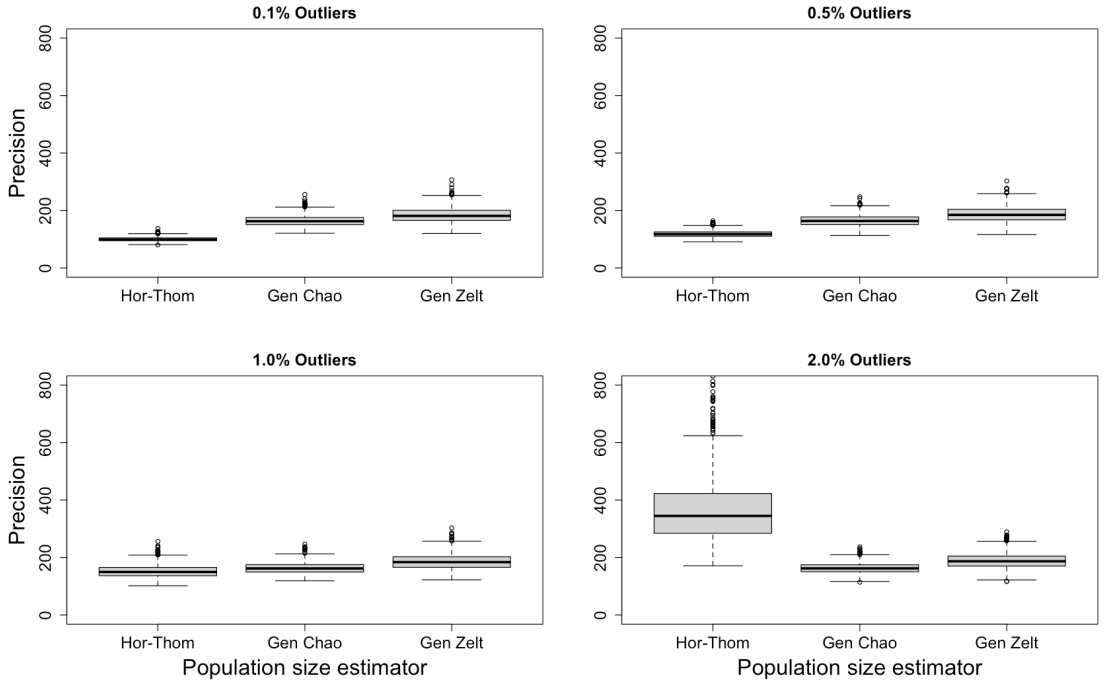

FIGURE 4.3: Box plots showing the **precision** of the confidence intervals for the capture-recapture estimators for different proportions of outliers when $N = 1000$ and $\lambda_L = Q3 + 3 \times IQR$.

and precision, with the Horvitz-Thompson estimator having the narrowest confidence intervals and estimates closest to the true values on average.

The preference for the Horvitz-Thompson estimator shifts once outliers are included in the data. At only 0.1% of counts being outliers, whilst the precision remains the lowest, the coverage is reduced to around 77% and accuracy reduced (the higher the value, the lower the accuracy itself) to have little difference from the other estimators. As more outliers are included in the data, the performance of the Horvitz-Thompson estimator worsens, with the distance between the estimated values and the true values growing further apart. Whist the width of the confidence intervals do not increase much overall until 2% outliers, given the inaccuracy of the Horvitz-Thompson estimator with outliers, the coverage decreases to 0%, meaning that by % outliers, none of the resulting confidence intervals contain the true value. It is important to note that whilst the coverage of the Horvitz-Thompson estimator is very high when there are 10% outliers, it is only due to the fact that the confidence intervals are very wide making it very unlikely that the resulting confidence interval wouldn't contain the true value. Therefore, this coverage value is misleading as the estimates are incredibly inaccurate and confidence intervals so wide that no useful conclusions could be made from them.

Throughout the varying proportions of outliers, the generalised Chao's and generalised Zelterman's estimators both perform well consistently. For the smaller proportions of outliers, and no outliers, there is very little difference between the performance of these two estimators, however, past 2% outliers, there is a clear preference for the generalised Chao's estimator. This preference is easy to see when there are 10% outliers, where the generalised Chao's estimator performs no differently to the other proportions of outliers, whereas the generalised Zelterman's performance declines slightly. At 10% outliers, the generalised Zelterman's estimator still outperforms the Horvitz-Thompson estimator, however, the resulting estimates are further from the true value, with wider confidence intervals and worse coverage. The findings reflect those of Böhning (2010), where it as found that whilst the conventional Chao's estimator and standard Zelterman's estimator perform similarly, there is a preference for the conventional Chao's estimator.

These results, looking at the impact of the varying proportions of outliers in the accuracy and precision of the capture-recapture estimators are given visually in Figures 4.2 and 4.3 respectively. In Figure 4.2 can be seen that there is no visual difference in the median accuracy of the generalised Chao's and generalised Zelterman's estimators, and the spread of the accuracy values remain consistent. However, as the proportion of outliers increase for the Horvitz-Thompson estimator, the median value grows further from the true value, and the spread of the accuracy values also increase, indicating that it is not a robust estimator in terms of accuracy. Similar results are seen in Figure 4.3, where the median precision values, and the spread of these values, are very consistent for the generalised Chao's and generalised Zelterman's estimators across the different proportions of outliers. In contrast, the Horvitz-Thompson estimator is not resilient to

outliers, as when the proportion of outliers increases, the median width of the confidence intervals and the spread of these values increase.

TABLE 4.4: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's and generalised Zelterman's, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0071$, $\lambda^U \approx 0.0085$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers. Number of outliers required to be integers so values for the proportion of 0.1% outliers are not given.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	Horvitz-Thompson	11	-	48	141	398	9.77e+06
	Generalised Chao's	19	-	19	18	19	18
	Generalised Zelterman's	21	-	22	21	22	28
Precision	Horvitz-Thompson	67	-	79	107	232	3.22e+08
	Generalised Chao's	116	-	116	115	113	109
	Generalised Zelterman's	130	-	131	130	130	143
Coverage	Horvitz-Thompson	94.8%	-	35.0%	0.5%	0.0%	100.0%
	Generalised Chao's	96.9%	-	96.3%	96.7%	95.5%	95.7%
	Generalised Zelterman's	94.6%	-	95.9%	95.7%	95.0%	95.0%

Given that there is an importance in assessing the performance of the capture-recapture population size estimators in different data scenarios, a simulation study is also conducted when the total population size is chosen to be $N = 500$. These results are given in Table 4.4. As for when $N = 1000$, the values in Table 4.4 show that the Horvitz-Thompson estimator is the preferred method when there are no outlier counts in the data, but performance of the estimator declines once outliers are included, even when the proportion of outliers is very small. Additionally, the generalised Chao's and generalised Zelterman's estimators demonstrate their resilience to outliers with no notable changes in performance between the proportions of outliers of 0% to 2%. As with the previous simulation study for when $N = 1000$, the generalised Chao's estimator remains resilient to outliers, even at 10% outliers, however, the performance of the generalised Zelterman's estimator does decline at 10% outliers, indicating a preference for the generalised Chao's estimator.

Whilst the outlier counts generated using the definition in Equation 4.27 are not necessarily obvious in comparison to the rest of the simulated counts, there is interest in exploring whether these conclusions hold under a smaller lower bound for the outlier rate, and therefore resulting in more subtle outlier counts. To do this, the following formula for the lower bound of outlier rates can be utilised,

$$\lambda_L = Q3 + 1.5 \times IQR,$$

where the upper bound for the outlier rate is still given by $\lambda_U = 1.2 \times \lambda_L$, but with this updated value of λ_L.

TABLE 4.5: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's and generalised Zelterman's, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0046$, $\lambda^U \approx 0.0055$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	Horvitz-Thompson	16	20	69	133	294	3693
	Generalised Chao's	25	28	27	27	26	26
	Generalised Zelterman's	29	32	31	32	32	51
Precision	Horvitz-Thompson	95	97	106	115	147	2192
	Generalised Chao's	162	163	163	162	162	155
	Generalised Zelterman's	181	181	185	184	187	204
Coverage	Horvitz-Thompson	95.5%	90.2%	29.9%	1.6%	0.0%	3.9%
	Generalised Chao's	96.4%	96.0%	96.4%	96.7%	95.8%	96.3%
	Generalised Zelterman's	95.7%	94.7%	95.8%	96.6%	94.8%	90.1%

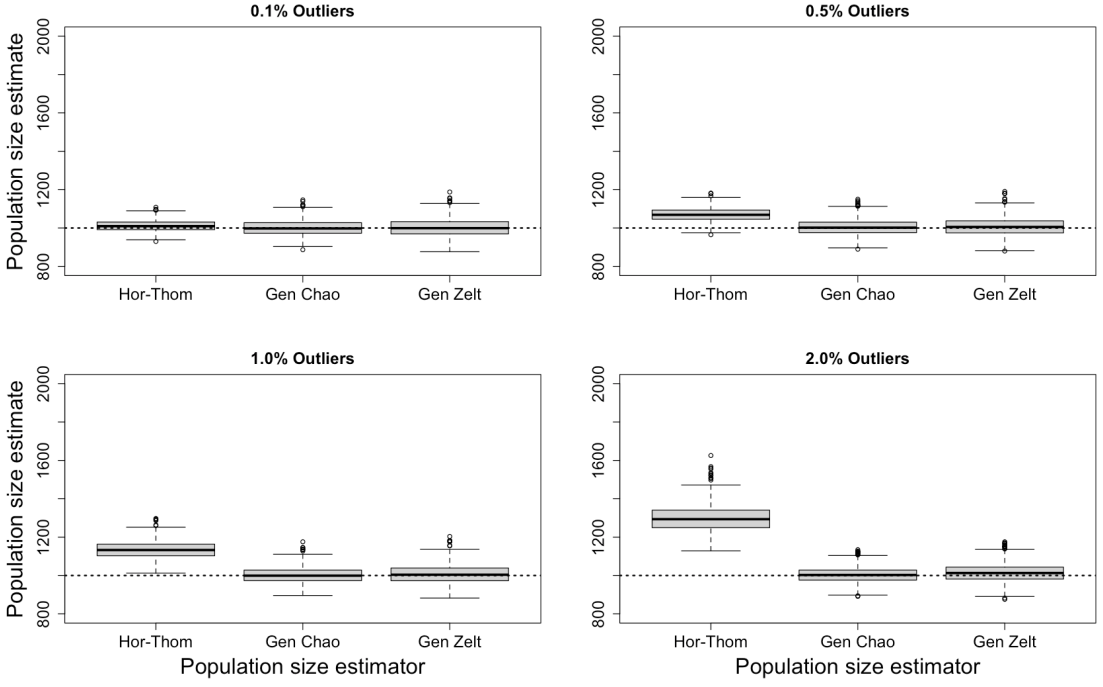

FIGURE 4.4: Box plots with the (individual) population size estimates from the simulation study, with a dotted line illustrating where the true value lies for illustrating the **accuracy** of the different capture-recapture estimators for different proportions of outliers when $N = 1000$ and $\lambda_L = Q3 + 1.5 \times IQR$.

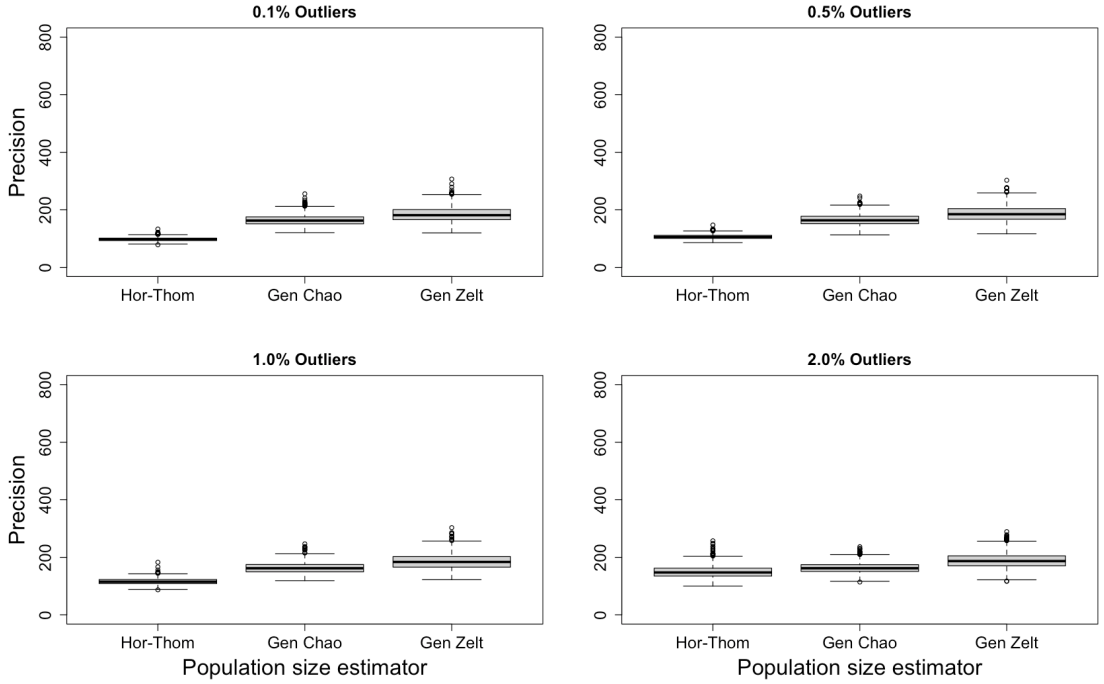

FIGURE 4.5: Box plots showing the **precision** of the confidence intervals for the capture-recapture estimators for different proportions of outliers when $N = 1000$ and $\lambda_L = Q3 + 1.5 \times IQR$.

TABLE 4.6: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's and generalised Zelterman's, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0046$, $\lambda^U \approx 0.0055$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers. Number of outliers required to be integers so values for the proportion of 0.1% outliers are not given.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	Horvitz-Thompson	11	-	25	65	142	1873
	Generalised Chao's	19	-	19	18	19	19
	Generalised Zelterman's	21	-	22	21	22	28
Precision	Horvitz-Thompson	67	-	72	82	103	1337
	Generalised Chao's	116	-	116	115	113	109
	Generalised Zelterman's	130	-	131	130	130	143
Coverage	Horvitz-Thompson	94.8%	-	69.9%	16.5%	0.2%	12.9%
	Generalised Chao's	96.9%	-	96.3%	96.7%	95.6%	95.9%
	Generalised Zelterman's	94.6%	-	95.9%	95.6%	95.0%	95.2%

Tables 7.10 and 4.6 provide the results from this simulation study for $N = 1000$ and $N = 500$ respectively, where it can be seen that despite the outlier counts being more subtle within the dataset, the results are very similar. Up to 2% outliers, the generalised Zelterman's estimator performs well, with adequate performance at 10% outliers. However, the generalised Chao's estimator performs consistently well for all proportions of outliers tested, with no notable change in performance across the different proportions of outliers.

Similarly, the results for the Horvitz-Thompson estimator follow the same trends as for when the outlier counts are generated using Equation 4.27, though the decline in performance is less dramatic. As the proportion of outlier counts in the data increase, the median distance between the true value and the estimated values grows, with the median width of the confidence intervals also growing. What makes the Horvitz-Thompson estimator particularly poor in this data scenario when there are outlier counts involved however, is the coverage. Even for a small number of subtle outliers, the coverage rapidly falls to where for when there are only 5 outlier counts in the dataset, only (approximately) 70% of the resulting confidence intervals contain the true value, decreasing to around 17% by 10 outliers. making the majority of the confidence intervals ineffectual.

The results for both accuracy and precision for when $N = 1000$ are given visually via box plots in Figures 4.4 and 4.5 respectively. In these plots, there is very little noticeable difference for the generalised Chao's and generalised Zelterman's estimators across the different proportions of outliers, testifying to their robustness. However, for the Horvitz-Thompson estimator, the plots for accuracy illustrate how sensitive the estimator is to outliers, with not only the spread of the values growing as the proportion of outliers grows, but the estimated values also get further from the true value. By 1% outliers, the minimum estimated population size from the simulation study is greater than the true value. The results in the precision box plots do not look as damaging for the Horvitz-Thompson estimator, since the precision values even at 2% outliers are still smaller than those for the generalised Chao's and generalised Zelterman's estimators. That being said, these plots also illustrate the lack of robustness of the Horvitz-Thompson, as both the spread of the precision values and the precision values themselves are getting larger as the proportion of outliers grow in the dataset.

Overall, these simulation studies indicate that for this type of data, if there are guaranteed no outliers included in the data, the Horvitz-Thompson estimator is a very good estimator. In real life situations, however, data can be unpredictable with outliers being a common occurrence. Given this, the generalised Chao's and generalised Zelterman's estimators are preferred as they not only perform well when outliers are included in the data, but also when they are not, with the generalised Chao's estimator performing marginally better overall, particularly when there are a higher proportion of outliers present in the data.

Chapter 5

Uncertainty Quantification: Approximation-Based Variances

To some degree, error is inevitable when using estimation methods, leading to uncertainty. In this chapter, approximation-based methods for quantifying this uncertainty are discussed for both the prevalence rate estimates and population size estimates, specifically the Wald-type interval and variance by conditioning.

5.1 Introduction

Mathematician John Allen Paulos describes uncertainty as "the only certainty there is" (Paulos, 2007). In the context of capture-recapture, uncertainty is the error and variability that arises from the various estimation methods. Given the nature of capture-recapture, the issue of uncertainty is very prevalent, but is often overlooked. Quantifying the level of uncertainty is crucial for model credibility in order to make reliable inferences from parameter estimates in addition to assessing the reliability of population estimates from capture-recapture methods. Estimates with higher amounts of uncertainty may be deemed untrustworthy, leading to the possibility of misinformed conclusions which can have negative implications, including financial and physical implications, especially in medical settings.

5.2 Wald-type interval for rate estimation

One of the most basic forms of uncertainty quantification is the Wald-type interval, a widely accepted confidence interval in practice based on the Wald statistic and relies on a normal approximation of the binomial distribution. The end-points of the $(1-\alpha)100\%$

interval for the rate are calculated as

$$\exp\left[\hat{\eta} \pm z_{1-\alpha/2}\text{s.e.}(\hat{\eta})\right],$$

where $\hat{\eta} = \mathbf{h}(\mathbf{v})^T\hat{\boldsymbol{\beta}}$, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution and s.e.$(\hat{\eta})$ is the standard error of $\hat{\eta}$ (Brown et al., 2001).

Application: Suicide data

Applying this approach to the suicide case study data, under the Poisson intercept-only model, a 95% confidence interval for the rate of completed suicide of $(23.3, 43.2)$ per 100,000 person-years is obtained. For context, the global rate of completed suicide, independent of whether the individuals have had bariatric surgery or not, as of 2019 is 9 per 100,000 person-years (World Health Organization, 2019). This rate is much lower than the estimated rate of completed suicide after bariatric surgery of 31.2 per 100,000 person-years but is also much smaller than the lower bound of the 95% confidence interval. Given that the global rate of completed suicide is not contained within the confidence interval of completed suicide rates after bariatric surgery, there is evidence to suggest that compared to the general population, individuals who have had bariatric surgery are at a notably higher risk of completed suicide, and is something that should be looked into further to reduce the difference between the two rates.

Similarly, neither of the rates computed through using the traditional meta-analytical approaches of 45 and 60 per 100,000 person-years are included in the 95% confidence interval above for the rate of completed suicide after bariatric surgery. This highlights the importance of accounting for the missing zero counts when computing the prevalence rate from a meta-analysis.

The Wald test statistic from the intercept-only zero-truncated Poisson model fails to monotonically increase as a function of its distance from $\beta = 0$, the null value, an aberration first observed by Hauck and Donner (1977), known as the Hauck-Donner effect (Yee, 2022). This effect can lead to the inference from the Wald test statistic, and hence the Wald confidence interval, being unreliable, in addition to the approach not accounting for model uncertainty, hence the interval is likely to underestimate the uncertainty. To take this uncertainty into consideration, resampling methods like the bootstrap algorithm can be used.

5.3 Horvitz-Thompson estimator variance by conditioning

Quantifying the uncertainty of population estimates can be done through calculating the variance of the corresponding estimator. Böhning (2008) finds the estimated variance of

the Horvitz-Thompson estimator to be

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = n\frac{p_0(\mu)}{(1-p_0(\mu))^2}. \tag{5.1}$$

However, for zero-truncated data, the probability of observing a count of zero events is unknown. This probability then requires estimation which leads to additional variance which will not be accounted for making Equation 5.1 unsuitable. To solve this problem, variance estimation by conditioning can be used, with the theoretical formula proposed by van der Heijden et al. (2003, page 314) given below.

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = E[\text{Var}(\widehat{N}^{(HT)}|I_i)] + \text{Var}(E[\widehat{N}^{(HT)}|I_i]), \tag{5.2}$$

where

$$I_i = \begin{cases} 1 & \text{study } i \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Whilst this method typically assumes that the data follows a Poisson distribution, the same approximation methods used for this assumption can be adjusted to develop a variance for data assumed to follow a geometric distribution.

The first term demonstrates the sampling variance in the zero-truncated Poisson distribution, and the second term demonstrates the variance in the observed sample.

To approximate each of the terms in Equation 5.2, the δ-method (see Powell, 2007; Oehlert, 1992, for more detail) can be used (e.g. Bishop et al., 2007, page 481). Following the work of van der Heijden et al. (2003), the first term is approximated by

$$E[\text{Var}(\widehat{N}^{(HT)}|I_i)] \approx \text{Var}(\widehat{N}^{(HT)}|I_i),$$

where $\text{Var}(\widehat{N}^{(HT)}|I_i)$ is also estimated using the δ-method as

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}|I_i) = \left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial\hat{\boldsymbol{\beta}}} \frac{1}{1-p_0(\hat{\mu}_i)}\right)^T (W(\hat{\boldsymbol{\beta}}))^{-1} \left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial\hat{\boldsymbol{\beta}}} \frac{1}{1-p_0(\hat{\mu}_i)}\right),$$

where $(W(\hat{\boldsymbol{\beta}}))$ is the the observed information matrix

$$\begin{aligned} W(\hat{\boldsymbol{\beta}}) &= \left[-\frac{\partial^2 \ell(\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}^T}\right] \\ &\approx \text{Cov}(\hat{\boldsymbol{\beta}})^{-1}, \end{aligned}$$

and let

$$\frac{1}{1-p_0(\hat{\mu}_i)} = G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}).$$

For the Poisson distribution,

$$p_0(\hat{\mu}_i) = \exp(-\hat{\mu}_i) = \exp(-\exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i),$$

hence

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}\frac{1}{1-p_0(\hat{\mu}_i)} = -\frac{\exp(\log(\hat{\mu}_i)-\hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2}\mathbf{h}(\mathbf{v}_i)^T = \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}).$$

Alternatively, for the geometric distribution,

$$p_0(\hat{\mu}_i) = \hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i,$$

hence

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}\frac{1}{1-p_0(\hat{\mu}_i)} = -\frac{\hat{\mu}_i}{(1-\hat{\mu}_i))^2}\mathbf{h}(\mathbf{v}_i)^T = \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}).$$

Therefore,

$$\widehat{\mathrm{Var}}(\widehat{N}^{(HT)}|I_i) = \left(\sum_{i=1}^{n}\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \mathrm{Cov}(\hat{\boldsymbol{\beta}})\left(\sum_{i=1}^{n}\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right). \tag{5.3}$$

As for the second term, using the δ-method, the expectation can be approximated as

$$E[\widehat{N}^{(HT)}|I_i] \approx \sum_{i=1}^{N}\frac{I_i}{1-p_0(\hat{\mu}_i)}.$$

Under the assumption of independence, the variance of this expectation is as follows.

$$\begin{aligned}
\mathrm{Var}(E[\widehat{N}^{(HT)}|I_i]) &\approx \mathrm{Var}\left(\sum_{i=1}^{N}\frac{I_i}{1-p_0(\hat{\mu}_i)}\right)\\
&= \sum_{i=1}^{N}\left[\frac{1}{(1-p_0(\hat{\mu}_i))^2}\mathrm{Var}(I_i)\right]\\
&= \sum_{i=1}^{N}\frac{p_0(\hat{\mu}_i)(1-p_0(\hat{\mu}_i))}{(1-p_0(\hat{\mu}_i))^2}\\
&= \sum_{i=1}^{N}\frac{p_0(\hat{\mu}_i)}{1-p_0(\hat{\mu}_i)}.
\end{aligned}$$

Given that N is unobserved, the above variance requires estimation. Using only the observed studies, an unbiased estimator for $\mathrm{Var}(E[\widehat{N}^{(HT)}|I_i])$ is given by

$$\begin{aligned}
\widehat{\mathrm{Var}}(E[\widehat{N}^{(HT)}|I_i]) &= \sum_{i=1}^{N}I_i\frac{p_0(\hat{\mu}_i)}{(1-p_0(\hat{\mu}_i))^2}\\
&= \sum_{i=1}^{n}\frac{p_0(\hat{\mu}_i)}{(1-p_0(\hat{\mu}_i))^2}.
\end{aligned} \tag{5.4}$$

If the Poisson distribution is assumed

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = \sum_{i=1}^{n} \frac{\exp(-\hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2},$$

therefore, Equation 5.2 is then given by the sum of Equations 5.3 and 5.4, given as

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{n} \frac{\exp(-\hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2}, \tag{5.5}$$

where

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = -\frac{\exp(\log(\hat{\mu}_i) - \hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2}\mathbf{h}(\mathbf{v}_i)^T.$$

Alternatively, if a geometric distribution is assumed,

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = \sum_{i=1}^{n} \frac{\hat{\mu}_i}{(1-\hat{\mu}_i)^2},$$

and Equation 5.2 is given as

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{n} \frac{\hat{\zeta}_i}{(1-\hat{\zeta}_i)^2}, \tag{5.6}$$

where

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = -\frac{\hat{\zeta}_i}{(1-\hat{\zeta}_i)^2}\mathbf{h}(\mathbf{v}_i)^T$$

for $\hat{\zeta}_i = \frac{1}{1+\hat{\mu}_i}$ and $\hat{\mu}_i = \tau_i \exp(\hat{\eta})$.

Application: Suicide data

Applying Equation 5.5 to the suicide case study data results in a variance of 1676.53 with a corresponding standard error of 40.95 leading to a 95% confidence interval for $\widehat{N}^{(HT)}$ of

$$\begin{aligned}\widehat{N}^{(HT)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(HT)})} &= 134.03 \pm 1.96 \times \sqrt{1676.53} \\ &= (54.78, 214.28) \\ &\approx (54, 214).\end{aligned}$$

In context, this confidence interval means that there should have likely been between 54 and 214 studies included in the systematic review total, and between 27 and 187 of those studies should report zero counts of completed suicide.

Application: Hares data

Applying Equation 5.6 to the case study data results in a variance of 249655.6, with a corresponding standard error of 499.66, leading to a 95% confidence interval for $\widehat{N}^{(HT)}$ of

$$\begin{aligned}\widehat{N}^{(HT)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(HT)})} &= 3122.67 \pm 1.96 \times \sqrt{249655.6} \\ &= (2143.37, 4101.98) \\ &\approx (2143, 4102).\end{aligned}$$

5.4 Generalised Chao's estimator variance by conditioning

The conditioning technique proposed by van der Heijden et al. (2003, page 314), seen in Equation 5.7, with the addition of the δ-method (Powell, 2007; Oehlert, 1992), can be used to find the standard error of the generalised Chao's estimator, following the work of Böhning et al. (2013b).

$$\widehat{\text{Var}}(\widehat{N}^{(GC)}) = E[\text{Var}(\widehat{N}^{(GC)}|I_i)] + \text{Var}(E[\widehat{N}^{(GC)}|I_i]). \tag{5.7}$$

For the first term, an estimator for the variance is developed as

$$\begin{aligned}\text{Var}(\widehat{N}^{(GC)}|I_i) &\approx \text{Var}\left(n + \sum_{i=1}^{f_1+f_2} \frac{1}{\hat{\mu}_i + \hat{\mu}_i^2/2}\right) \\ &= \text{Var}\left(\sum_{i=1}^{f_1+f_2} \frac{1}{\hat{\mu}_i + \hat{\mu}_i^2/2}\right),\end{aligned}$$

where $\dfrac{1}{\hat{\mu}_i + \hat{\mu}_i^2/2} = G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})$.

Using the multivariate δ-method, the variance is estimated as

$$\widehat{\text{Var}}(\widehat{N}^{(GC)}|I_i) = \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),$$

where for $\mu_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = \frac{\hat{\mu}_i + \hat{\mu}_i^2}{(\hat{\mu}_i + \hat{\mu}_i^2/2)^2}\mathbf{h}(\mathbf{v}_i)^T.$$

For the second term, the expectation is

$$E[\widehat{N}^{(GC)}|I_i] = E\left[n + \sum_{i=1}^{N} \frac{I_i}{\hat{\mu}_i + \hat{\mu}_i^2/2}|I_i\right] \approx \sum_{i=1}^{N} I_i w_i,$$

where $w_i = 1 + p_0(\mu_i)/p_i$ with $p_0(\mu_i) = \exp(-\mu_i)$ and $p_i = p_1(\mu_i) + p_2(\mu_i) = \exp(-\mu_i)\mu_i + \exp(-\mu_i)\mu_i^2/2$.

The indicator variable I_i is binary with expectation

$$E[I_i] = p_i,$$

and variance

$$\text{Var}(I_i) = p_i(1 - p_i).$$

Therefore, the second term of Equation 5.7 is

$$\text{Var}\left(\sum_{i=1}^{N} I_i w_i\right) = \sum_{i=1}^{N} p_i(1 - p_i)w_i^2,$$

and estimated by

$$\widehat{\text{Var}}(E[\widehat{N}^{(GC)}|I_i]) = \sum_{i=1}^{N} \frac{I_i}{p_i} p_i(1 - p_i)w_i^2 = \sum_{i=1}^{f_1+f_2} (1 - \hat{p}_i)\left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2.$$

Therefore, Equation 5.7 is given by

$$\widehat{\text{Var}}(\widehat{N}^{(GC)}) = \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{f_1+f_2} (1 - \hat{p}_i)\left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2.$$

It is important to note that this variance formula is for when a Poisson mixture kernel is assumed. In the case where a geometric mixture kernel is assumed, the variance formula is altered as follows. The variance in the first term is given as

$$\text{Var}(\widehat{N}^{(GC)}|I_i) = \text{Var}\left(n + \sum_{i=1}^{f_1+f_2} \frac{1}{(1 - \zeta_i)(2 - \zeta_i)}\right) = \text{Var}\left(\sum_{i=1}^{f_1+f_2} \frac{1}{(1 - \zeta_i)(2 - \zeta_i)}\right),$$

where $\zeta_i = \dfrac{1}{1+\mu_i}$ and $\dfrac{1}{(1-\zeta_i)(2-\zeta_i)} = G(\mu_i|\hat{\boldsymbol{\beta}})$.

Using the multivariate $\delta-$method, the above variance can be estimated as

$$\widehat{\text{Var}}(\widehat{N}^{(GC)}|I_i) = \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),$$

where for $\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = \left(\frac{1}{\hat{\mu}_i + 2\hat{\mu}_i^2} - \frac{(1+\hat{\mu}_i)(1+4\hat{\mu}_i)}{(\hat{\mu}_i+2\hat{\mu}_i^2)^2}\right).$$

As for the second term, the expectation is given as

$$\begin{aligned} E[\widehat{N}^{(GC)}|I_i] &= E\left[n + \sum_{i=1}^{N} \frac{I_i}{(1-\zeta_i)(2-\zeta_i)}|I_i\right] \\ &\approx \sum_{i=1}^{N} I_i w_i, \end{aligned}$$

where $w_i = 1 + \dfrac{p_0(\mu_i)}{p_i}$ with $p_0(\mu_i) = \dfrac{1}{1+\mu_i}$ and

$$p_i = p_1(\mu_i) + p_2(\mu_i) = \frac{1}{1+\mu_i}\left(1 - \frac{1}{1+\mu_i}\right)\left(2 - \frac{1}{1+\mu_i}\right).$$

As in the Poisson case, I_i is a binary indicator variable with expectation $E[I_i] = p_i$ and variance $\text{Var}(I_i) = p_i(1-p_i)$.

Therefore, the second term is equal to

$$\text{Var}\left(\sum_{i=1}^{N} I_i w_i\right) = \sum_{i=1}^{N} p_i(1-p_i)w_i^2,$$

which is estimated as

$$\begin{aligned} \widehat{\text{Var}}(E[\widehat{N}^{(GC)}|I_i]) &= \sum_{i=1}^{N} \frac{I_i}{p_i}(1-p_i)w_i^2 \\ &= \sum_{i=1}^{f_1+f_2} (1-\hat{p}_i)\left(1 + \frac{1}{(1-\hat{\zeta}_i)(2-\hat{\zeta}_i)}\right)^2, \end{aligned} \tag{5.8}$$

where $\hat{\zeta}_i = \dfrac{1}{1-\hat{\mu}_i}$.

Therefore, Equation 5.7 is given by

$$\widehat{\text{Var}}(\widehat{N}^{(GC)}) = \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{f_1+f_2} (1-\hat{p}_i)\left(1+\frac{1}{(1-\hat{\zeta}_i)(2-\hat{\zeta}_i)}\right)^2. \tag{5.9}$$

Application: Suicide data

Assuming a Poisson mixture kernel and applying Equation 5.4 to the suicide case study data leads to a variance of 12707.05 with a corresponding standard error of 112.73 and 95% confidence interval for $\widehat{N}^{(GC)}$ of

$$\begin{aligned}\widehat{N}^{(GC)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(GC)})} &= 172.66 \pm 1.96 \times \sqrt{12707.05} \\ &= (-48, 394) \\ &\approx (27, 394).\end{aligned}$$

The lower limit for this confidence interval shows that this is not an appropriate interval, given that the total population size cannot be negative and the total population size also has a minimum value of 27, given that 27 studies were observed. The suicide case study data itself is a small dataset, but even smaller for the generalised Chao's estimator given that only frequencies of ones and twos are not truncated, this is likely the cause of the large variance computed using the analytical approach which leads to the inappropriate confidence interval. The simulation study in Section 3.3 utilises this analytical approach to variance computation for constructing the confidence intervals, where the results indicate that when the (total) population size is larger (for example, $N = 500$ and $N = 1000$), this approach leads to reasonably wide confidence intervals, that when centred around the total population size, do not result in negative lower limits. Additionally, the resulting intervals have excellent coverage, suggesting that if the observed population size is relatively large, then this approach is suitable.

As the problems with the analytical approach arise as a result of the small population size of the data, there is incentive to use the bootstrap algorithm discussed in Section 6, since this algorithm does not rely on the size of the observed sample and instead iteratively resamples in order to achieve a large number of samples to compute intervals from.

Application: Hares data

Applying Equation 5.9 and assuming a geometric mixture kernel to the hares case study data leads to a variance of 204151.3 with a corresponding standard error of 451.83,

leading to a 95% confidence interval for $\widehat{N}^{(GC)}$ of

$$\begin{aligned}\widehat{N}^{(GC)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(GC)})} &= 3890.14 \pm 1.96 \times \sqrt{204151.3} \\ &= (3004.57, 4775.71) \\ &\approx (3005, 4776).\end{aligned}$$

Unlike with the suicide case study data, the lower limit for this confidence interval is not only greater than 0, but also greater than the observed number of hares, leading to a confidence interval which is appropriate for the data. This is likely due to the larger sample size, given that the observed number of hares ($n = 983$) is much greater than the observed number of studies in the suicide data ($n = 27$), truncating all counts besides the singletons and doubletons does not leave only a small number of individuals to model and compute a confidence interval from. This is supported by the simulation study in Section 3.3, where when the (total) population size is large, the resulting confidence intervals constructed using this analytical approach are of reasonable width (that does not lead to a negative lower limit when centred around the total population size) with excellent coverage.

5.5 Generalised Zelterman's estimator variance by conditioning

As with Sections 5.3 and 5.3, the conditioning technique in Equation 5.10 proposed by van der Heijden et al. (2003), can be used to find the standard error of the generalised Zelterman's estimator, following the work of Böhning and van der Heijden (2009).

$$\widehat{\text{Var}}(\widehat{N}^{(GZ)}) = E[\text{Var}(\widehat{N}^{(GZ)}|I_i)] + \text{Var}(E[\widehat{N}^{(GZ)}|I_i]). \tag{5.10}$$

Given that the generalised Zelterman's estimator utilises the Horvitz-Thompson estimator, the steps and methods used in Section 5.3 to find the variance of the Horvitz-Thompson estimator are used here to find the variance of the generalised Zelterman's estimator. Therefore, when a Poisson distribution is assumed, the first term is as follows.

$$\begin{aligned}E[\text{Var}(\widehat{N}^{(GZ)}|I_i)] &\approx \text{Var}(\widehat{N}^{(GZ)}) \\ &= \text{Var}\left(\sum_{i=1}^{N} I_i \frac{1}{1-\exp(-\hat{\mu}_i)}\right) \\ &= \left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \frac{1}{1-\exp(-\hat{\mu}_i)}\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \frac{1}{1-\exp(-\hat{\mu}_i)}\right) \\ &= \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),\end{aligned}$$

where

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = -\frac{\exp(\log(\hat{\mu}_i) - \hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2}\mathbf{h}(\mathbf{v}_i)^T.$$

The second term of Equation 5.10 is then estimated as

$$E[\widehat{N}^{(GZ)}|I_i] \approx \sum_{i=1}^{N} \frac{I_i}{1-\exp(-\mu_i)},$$

where using the independence assumption, the variance is calculated as

$$\text{Var}\left(\sum_{i=1}^{N} \frac{I_i}{1-\exp(-\mu_i)}\right) = \sum_{i=1}^{N} \frac{\exp(-\mu_i)}{1-\exp(-\mu_i)},$$

with the unbiased estimator

$$\begin{aligned}\widehat{\text{Var}}\left(\sum_{i=1}^{N} \frac{I_i}{1-\exp(-\mu_i)}\right) &= \sum_{i=1}^{N} I_i \frac{\exp(-\mu_i)}{(1-\exp(-\mu_i))^2} \\ &= \sum_{i=1}^{n} \frac{\exp(-\mu_i)}{(1-\exp(-\mu_i))^2}.\end{aligned}$$

The variance of the generalised Zelterman's estimator seen in Equation 5.10 is as follows.

$$\begin{aligned}\widehat{\text{Var}}(\widehat{N}^{(GZ)}) = &\left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \\ &\sum_{i=1}^{n} \frac{\exp(-\mu_i)}{(1-\exp(-\mu_i))^2}.\end{aligned} \tag{5.11}$$

When a geometric distribution is assumed, the variance formula is different, where the first term is given as follows.

$$\begin{aligned}E[\text{Var}(\widehat{N}^{(GZ)}|I_i] &\approx \text{Var}(\widehat{N}^{(GZ)}) \\ &= \text{Var}\left(\sum_{i=1}^{n} \frac{1+\hat{\mu}_i}{\hat{\mu}_i}\right),\end{aligned}$$

where $\dfrac{1+\hat{\mu}_i}{\hat{\mu}_i} = G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})$.

Using the multivariate δ-method, the variance is estimated as

$$\widehat{\text{Var}}(\widehat{N}^{(GZ)}|I_i) = \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right), \tag{5.12}$$

where for $\mu_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}}$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = -\frac{1}{\hat{\mu}_i}\mathbf{h}(\mathbf{v}_i)^T.$$

For the second term, the expectation is given as

$$\begin{aligned} E[\widehat{N}^{(GZ)}|I_i] &= E\left[\sum_{i=1}^{n}\frac{1+\hat{\mu}_i}{\hat{\mu}_i}\right] \\ &\approx \sum_{i=1}^{N} I_i\frac{1+\hat{\mu}_i}{\hat{\mu}_i}, \end{aligned}$$

where using the independence assumption, the variance is calculated as

$$\text{Var}\left(\sum_{i=1}^{N} I_i\frac{1+\hat{\mu}_i}{\hat{\mu}_i}\right) = \sum_{i=1}^{N}\frac{1/(1+\hat{\mu}_i)}{1-1/(1+\hat{\mu}_i)},$$

with the unbiased estimator

$$\begin{aligned} \widehat{\text{Var}}\left(\sum_{i=1}^{N} I_i\frac{1+\hat{\mu}_i}{\hat{\mu}_i}\right) &= \sum_{i=1}^{N} I_i\frac{1/(1+\hat{\mu}_i)}{(1-1/(1+\hat{\mu}_i))^2} \\ &= \sum_{i=1}^{n}\frac{1/(1+\hat{\mu}_i)}{(1-1/(1+\hat{\mu}_i))^2} \\ &= \sum_{i=1}^{n}\frac{1/(1+\hat{\mu}_i)}{\hat{\mu}_i^2} \\ &= \sum_{i=1}^{n}\frac{1}{\hat{\mu}_i^2(1+\hat{\mu}_i)}. \end{aligned} \tag{5.13}$$

Therefore, summing Equations 5.12 and 5.13 leads to the estimated variance of the generalised Zelterman's estimator when a geometric distribution is assumed as follows.

$$\begin{aligned} \widehat{\text{Var}}(\widehat{N}^{(GZ)}) = &\left(\sum_{i=1}^{n}\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}})\left(\sum_{i=1}^{n}\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \\ &\sum_{i=1}^{n}\frac{1}{\hat{\mu}_i^2(1+\hat{\mu}_i)}. \end{aligned} \tag{5.14}$$

Application: Suicide data

Applying Equation 5.11 to the suicide case study data results in a variance of 13425.49 with corresponding standard error 115.87 and 95% confidence interval

$$\begin{aligned}\widehat{N}^{(GZ)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(GZ)})} &= 175.19 \pm 1.96 \times \sqrt{13425.49} \\ &= (-52, 402) \\ &\approx (27, 402).\end{aligned}$$

As with the confidence interval for the generalised Chao's estimator in Section 5.4, this confidence interval is not appropriate for the data given that the lower limit is negative and the population size must be both non-negative and has a minimum of 27 given the number of observed studies. Similarly, these results motivate the use of both the bootstrap algorithm for computing the variance and confidence intervals instead.

Application: Hares data

Applying Equation 5.14 to the hares case study data results in a variance of 57003.52, with corresponding standard error 238.75, resulting in the 95% confidence interval

$$\begin{aligned}\widehat{N}^{(GZ)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(GZ)})} &= 3601.67 \pm 1.96 \times \sqrt{57003.52} \\ &= (3133.72, 4069.62) \\ &\approx (3134, 4070).\end{aligned}$$

As with the interval for the generalised Chao's estimator, this interval is of reasonable width, with the lower limit being both greater than zero and greater than the observed number of hares in the dataset. This differs to when the analytical approach is applied to the suicide case study data, where the observed dataset is much smaller in size, leaving a very small number of singletons and doubletons to use in the modelling and estimation processes. These results, supported by the simulation study in Section 3.3, suggest that the analytical approach is appropriate to use when there is a large sample size available, and if there is not, then the bootstrap algorithm approach may be more suitable.

Chapter 6

Uncertainty Quantification: Bootstrap Algorithms

6.1 Introduction

The bootstrap algorithm (see Efron, 1979, 1981a, 1985; Efron and Tibshirani, 1993) is a resampling method utilised for uncertainty quantification. Data is sampled and resampled iteratively with replacement in to create a sampled dataset to estimate the rate and target population size from and calculate the corresponding variance of the bootstrap samples. Using capture-recapture methods, the level of uncertainty of the estimated number of studies can be computed (see Buckland and Garthwaite, 1991; Zwane and van der Heijden, 2003; Efron, 1981a), using bootstrapping. The standard approach for the bootstrap algorithm is to take a non-parametric approach, where each bootstrap sample has size equal to that of the observed data (e.g. Nock et al., 2008, Method 1), seen in Section 6.2. However, the non-parametric approach risks underestimating the variance as it does not account for the uncertainty in sampling n out of N, which the semi-parametric, imputed approach developed in Section 6.3 does account for through randomly sampling the elements from the complete dataset made up of both the observed and unobserved elements. Both the non-parametric and the semi-parametric approaches randomly sample the studies to use for modelling risking creating a sample which is unrepresentative of the data, especially with the small observed sample size. This is a particular issue when accounting for model uncertainty, where correlation between covariate combinations has the potential to give inflated results. A parametric approach to bootstrapping can help avoid this, where instead of resampling the studies themselves, the response variable is sampled from a given distribution to create a more reliable sampled dataset. This approach is seen in Section 6.4.

A summary of the different bootstrapping approaches discussed in this section are as follows:

Approach 1: Non-parametric bootstrap algorithm - sample from only the observed data at random with replacement and model the sampled data using a chosen model. Compute desired statistics from chosen model.

Approach 2: Semi-parametric bootstrap algorithm - sample from the imputed complete data at random with replacement and model the sampled data using a chosen model. Compute desired statistics from chosen model.

Approach 3: Parametric bootstrap algorithm - use AIC or BIC weights to choose a distribution from the models under consideration to sample counts from, create a new dataset with the observed explanatory variables. Model the sampled data using a chosen model. Compute desired statistics from chosen model.

To model the sampled data and account (or not account) for model uncertainty, the chosen model can be selected using one of 3 methods:

Method 1: (Full) In each bootstrap iteration, fit each of the competing models to the sampled data and select the preferred model using AIC or BIC statistics (see Silverman et al., 2024, for similar work). This method (fully) accounts for model uncertainty but risks high correlation between the covariate combinations as a result of models with covariates having the lowest AIC or BIC value being selected but leading to inflated estimates and wide confidence intervals.

Method 2: (Partial) Use an additional bootstrap algorithm, fitting the models to the sampled data and recording the frequency of each model being preferred according to the AIC or BIC statistic. Use the linear predictor and distribution combination which is preferred the majority of the times in the main bootstrap algorithm. This method (partially) accounts for model uncertainty.

Method 3: (None) Use the same linear predictor and distribution combination that is preferred for the observed data. This method does not account for model uncertainty, risking underestimation of the variance.

Constructing confidence intervals aid in the quantification of the level of uncertainty of an estimate from the bootstrap data. Introduced by Efron (1979), the widely accepted percentile method is one way of constructing confidence intervals. For a given value of α, the $(100-\alpha)\%$ confidence interval limits are computed by finding the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ percentiles of the bootstrap statistics ordered numerically lowest to highest. Therefore, the $(100-\alpha)\%$ percentile confidence intervals for the bootstrap estimates of the rate of completed suicide and population sizes are as follows.

$$\left[\exp\left(\hat{\eta}\right)_{lower}, \exp\left(\hat{\eta}\right)_{upper}\right] = \left[\exp\left(\hat{\eta}^*_{(\frac{\alpha}{2}\times B)}\right), \exp\left(\hat{\eta}^*_{((1-\frac{\alpha}{2})\times B)}\right)\right],$$
$$\left[\widehat{N}_{lower}, \widehat{N}_{upper}\right] = \left[\widehat{N}^*_{(\frac{\alpha}{2}\times B)}, \widehat{N}^*_{((1-\frac{\alpha}{2})\times B)}\right],$$

where B is the number of bootstrap samples.

6.2 Approach 1: Non-parametric

The non-parametric approach randomly resamples with replacement from the observed data only, resulting in each bootstrap sample being of size n. Given a chosen combination of linear predictor and distribution, the sampled data is modelled and the desired statistics computed from the results. This process is repeated B times.

6.2.1 Method 1: Full

Formally, the non-parametric bootstrap algorithm using Method 1 for model selection is as follows in Algorithm 2.

Algorithm 2 Bootstrap Approach 1, Method 1

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_n^*, \mathbf{v}_n^*, x_n^*)\}$, through drawing n observations from the original dataset, $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$, at random and with replacement.
Step 3: Fit the competing models to the bootstrapped data and select the preferred model using either the AIC or BIC. Estimate the rate or total population size using the methods discussed in Chapter 4.
Step 4: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Application: Suicide data

For quantifying the rate uncertainty, six sub-populations are defined by the combination of covariates observed. Specifically, the combinations of the country of origin of the study being either USA or other, and the proportion of women in the study being either 0.75, 0.8 or 0.85, where the proportions are approximately the three quartiles of the observed data. Additionally, eight sub-populations are considered for assessing the level of uncertainty when estimating the target population size. These sub-populations are defined by the various combinations of the country of origin of each study and the respective proportion of women. Country of origin is either USA or Other, and the proportion of women for each study lies in one of the following intervals, $[0, 0.75)$, $[0.75, 0.80)$, $[0.80, 0.85)$ and $[0.85, 1]$, where the cut-off points are three approximate quartiles for the observed data.

To formulate the non-parametric bootstrap algorithm to compute the rate and Horvitz-Thompson population size estimates, Step 3 of Algorithm 2 can be modified as follows.

Step 3: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor and $\tilde{D} \in \{\text{Poisson (P), negative-binomial (NB)}\}$ be the distribution that minimises the BIC with corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}$ of $\boldsymbol{\beta}_j$ for the respective model. If $\tilde{D} = (NB)$, let $\hat{\theta}$ be the estimate of the dispersion parameter.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.8, 1)^T$, $\bar{\mathbf{v}}_4 = (0.8, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$, the sub-population specific estimated rate of completed suicide is calculated as

$$\exp(\eta_{bk}^*) = \exp\left[\mathbf{h}_{\tilde{j}}(\bar{\mathbf{v}}_k)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(P)}\right]\right)} & \text{if } \tilde{D} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(NB)}\right], \hat{\theta}\right)} & \text{if } \tilde{D} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of studies is $\widehat{N}_b^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.1: Sub-population specific 95% percentile confidence intervals for the rate of completed suicide (per 100,000 person-years) using the **non-parametric** bootstrap samples.

	Proportion of women		
Country of origin	0.75	0.80	0.85
USA	(9.9, 54.8)	(13.7, 51.9)	(17.7, 50.1)
Other	(19.6, 58.4)	(15.3, 57.1)	(15.8, 57.0)

The 95% percentile confidence intervals can be calculated from the bootstrap samples for both the rate of completed suicide and population sizes. The confidence intervals for the rate of each of the six sub-populations can be seen in Table 6.1, where each of the intervals are approximately centred at the estimated rate of 31.8 completed suicides per 100,000 person-years. Comparing the widths of the confidence intervals to assess the degree of uncertainty show that for studies originating from the USA, the rate estimates

are less uncertain for higher proportions of women. Given that the more data used for estimation, the more accurate the estimate typically is and hence has less uncertainty, this conclusion is plausible. For studies originating from the USA, only 10% of studies have a proportion of women lower than 0.80, so the sub-population with a higher proportion of women has more data and hence a more accurate estimate. However, this trend of the higher proportion of women leading to lower uncertainty does not continue for studies originating from countries besides the USA. With only 18% of studies from outside the USA having a proportion of women greater than 0.85, 41% greater than 0.80 and 65% greater than 0.75, the widths of the intervals don't differ in size greatly. Therefore, for countries outside of the USA, the level of uncertainty is relatively constant for different proportions of women included in the study.

TABLE 6.2: Values of 95% percentile confidence intervals for the suicide case study data using the Horvitz-Thompson population size estimates from the **non-parametric** bootstrap samples for the eight sub-populations, marginal totals and overall total.

	Proportion of women				
Country of origin	$[0, 0.75)$	$[0.75, 0.80)$	$[0.8, 0.85)$	$[0.85, 1]$	Total
USA	$(1, 3)$	$(0, 0)$	$(19, 52)$	$(8, 18)$	$(28, 75)$
Other	$(23, 70)$	$(16, 51)$	$(7, 20)$	$(5, 19)$	$(57, 230)$
Total	$(24, 1842)$	$(16, 51)$	$(27, 115)$	$(14, 134)$	$(91, 7059)$

Table 6.2 provides 95% percentile confidence intervals using the Horvitz-Thompson population size estimates from the non-parametric bootstrap algorithm. For each of the sup-populations, the estimates of the total number of studies from the bootstrap data are highly correlated, leading to inflated upper limits of the confidence intervals. This can be seen in the confidence interval of the marginal totals, where the upper limits are exceptionally large in comparison to the summation of the individual upper limits for the sub-populations. In particular, the confidence interval for the total number of studies indicates that there is a lot of uncertainty with a very wide interval, which may not be the case if model uncertainty is taken into consideration using Method 2.

For the generalised Chao's and generalised Zelterman's estimators, Step 3 of the non-parametric bootstrap algorithm can be modified as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for linear predictors $j = 1, \cdots, 4$ in Table 2.1. Let $\tilde{j} = 1, \cdots, 4$ be the linear predictor which minimises the BIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

For the generalised Chao's and generalised Zelterman's estimators, in the bootstrap algorithms only linear predictors 1, 2, 3 and 4 are fitted due to the interaction term in linear predictor 5 leading to predicted values of either 0 or 1 when the binomial logistic regression model is fitted to the truncated data, suggesting that there are issues with the model.

As with the Horvitz-Thompson estimator, the total population size estimates confidence intervals for the generalised Chao's and generalised Zelterman's population size estimators are very wide with a large upper limit. Respectively, the 95% percentile confidence intervals for the two estimators are (67, 10870), and (67, 9625864000). Similarly, this inflated upper limit of the confidence interval is as a result of the high correlation between covariate combinations, leading to large population size estimates and confidence intervals which provide no useful information. Utilising an alternative approach for accounting for model uncertainty within the non-parametric bootstrap algorithm with a lower risk of this high correlation may lead to a more appropriate confidence interval.

Application: Hares data

The sub-populations for the hares case study dataset are defined by the combination of covariates observed, specifically the different seasons and study areas. Given that the AIC is used for model selection for the observed data, the AIC is used in the bootstrap algorithm for consistency. Additionally, only the geometric distribution is assumed in the bootstrap algorithm due to the large disparity in the AIC, BIC and likelihood statistics for the geometric models to the Poisson and negative-binomial models.

To formulate the non-parametric bootstrap algorithm to compute the Horvitz-Thompson population size estimates, Step 3 of Algorithm 2 can be modified as follows.

Step 3: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor that minimises the AIC with corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ of $\boldsymbol{\beta}_j$ for the respective model.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}}_{\tilde{j}}\right]\right)},$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i\in A}\widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n}\widehat{N}_{bi}^{(HT)*}$.

TABLE 6.3: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **non-parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	$(279, 339)$	$(533, 643)$	$(770, 925)$	$(1319, 1575)$
Five small areas	$(275, 336)$	$(368, 445)$	$(670, 806)$	$(1589, 1898)$
Total	$(557, 671)$	$(904, 1083)$	$(1443, 1724)$	$(2910, 3467)$

Table 6.3 provides the 95% percentile confidence intervals for each sub-population, marginal totals and the overall total estimated population size, using the Horvitz-Thompson estimator from the non-parametric bootstrap algorithm. Given that the upper limits of the confidence intervals are not inflated, like with the suicide case study data, estimates of the number of hares for each sub-population are not highly correlated. Additionally, the 95% percentile confidence interval for the overall number of hares is approximately centred around the corresponding Horvitz-Thompson population size estimate for the observed data, however, the sub-population intervals are not all centred around their corresponding estimates for the observed data. Overall, the intervals are relatively narrow, indicating that there is not a high quantity of uncertainty, which is to be expected in comparison to the suicide data given that the size of the data is much larger and the more data available the less uncertainty with estimation.

For the generalised Chao's and generalised Zelterman's estimators, Step 3 of the non-parametric bootstrap algorithm can be modified as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for linear predictors $j = 1, \cdots, 5$ in Table 2.1. Let $\tilde{j} =$

$1, \cdots, 5$ be the linear predictor which minimises the AIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for hares with the same covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

The resulting 95% percentile confidence interval for the overall number of snowshoe hares using the generalised Chao's estimator is (2994, 3954). As for the generalised Zelterman's estimator, the resulting 95% percentile confidence interval for the overall number of snowshoe hares is (3218, 4268). Both intervals contain the respective population size estimates of the observed data and are approximately centred at these estimates.

6.2.2 Method 2: Partial

Given that accounting for model uncertainty through using Method 1, sampling from the best fitting competing model for each bootstrap iteration, leads to high correlation between the covariate combinations and hence very wide confidence intervals, an alternative approach should be considered. This alternative of Method 2 requires an additional bootstrap algorithm to find out which of the competing models fits the data best the majority of times. Formally, this additional bootstrap is as follows in Algorithm 3.

From the results from the additional bootstrap, compute the proportion that each combination of $(\tilde{j}, \tilde{D})$ is selected as the best. Use this linear predictor and distribution combination to model the bootstrap data in the non-parametric bootstrap algorithm for

Algorithm 3 Bootstrap Approach 1, Method 2, Additional Bootstrap

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \cdots, (\tau_n^*, \mathbf{v}_n^*, x_n^*)\}$, through drawing n observations from the original dataset, $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$, at random and with replacement.
Step 3: Fit the competing models to the bootstrap data. Let $(\tilde{j}_b, \tilde{D}_b)$ be the linear predictor and distribution combination that minimises the AIC or BIC.
Step 4: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

computing the desired estimates. Formally, the non-parametric bootstrap algorithm is then as follows in Algorithm 4.

Algorithm 4 Bootstrap Approach 1, Method 2

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_n^*, \mathbf{v}_n^*, x_n^*)\}$, through drawing n observations from the original dataset, $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$, at random and with replacement.
Step 3: Fit the model with linear predictor and distribution combination $(\tilde{j}, \tilde{D})$ found using Algorithm 3 and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 4: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Application: Suicide data

Step 3 of Algorithm 3 can be modified as follows to find the proportion of times each linear predictor and distribution combination is preferred for the rate and Horvitz-Thompson population size estimates to use in the non-parametric bootstrap algorithm.

Step 3: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j}_b = 1, \cdots, 5$ be the linear predictor and $\tilde{D}_b \in \{\text{Poisson (P), negative-binomial (NB)}\}$ be the distribution that minimises the BIC. Record the combination $(\tilde{j}_b, \tilde{D}_b)$.

Additionally, Step 3 of Algorithm 4 can be modified as follows to estimate the rate and Horvitz-Thompson population size estimates.

Step 3: Fit the zero-truncated model with linear predictor and distribution combination $(\hat{\tilde{j}}, \hat{\tilde{D}})$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(\hat{\tilde{D}})}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$ and $\hat{\theta}$ be the estimate of the dispersion parameter if $\hat{\tilde{D}} = (NB)$.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.80, 1)^T$, $\bar{\mathbf{v}}_4 = (0.80, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$, the sub-population specific estimated rate of

completed suicide is calculated as

$$\exp(\eta^*_{bk}) = \exp\left[\mathbf{h}_{\hat{\hat{j}}}(\bar{\mathbf{v}}_k)^T \hat{\boldsymbol{\beta}}^{(\hat{\hat{D}})}_{\hat{\hat{j}}}\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}^{(HT)*}_{bi} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}^{(P)}_{\hat{\hat{j}}}\right]\right)} & \text{if } \hat{\hat{D}} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}^{(NB)}_{\hat{\hat{j}}}\right], \hat{\theta}\right)} & \text{if } \hat{\hat{D}} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}^{(HT)*}_{bA} = \sum_{i \in A} \widehat{N}^{(HT)*}_{bi}$ and the estimated total number of studies is $\widehat{N}^{(HT)*}_{b} = \sum_{i=1}^{n} \widehat{N}^{(HT)*}_{bi}$.

TABLE 6.4: Proportion of times each linear predictor and distribution combination has lowest BIC statistic from the **non-parametric** bootstrap algorithm for computing the rate and Horvitz-Thompson estimator for the suicide case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Poisson	78.2%	15.3%	4.5%	0.3%	1.7%
Negative-binomial	0.0%	0.0%	0.0%	0.0%	0.0%

Table 6.4 displays the proportions that each linear predictor and distribution combination has the lowest BIC statistic from the non-parametric bootstrap algorithm. Given that the intercept-only zero-truncated Poisson model is chosen as the best fitting model for approximately 80% of the bootstrap iterations, the linear predictor and distribution combination to use in the bootstrap algorithm for computing the 95% confidence intervals for the rate of completed suicide and Horvitz-Thompson population size estimator is $(\hat{\hat{j}}, \hat{\hat{D}}) = (1, P)$. The 95% percentile confidence intervals for the rate and population size utilising this linear predictor and distribution combination are (8.1, 66.0) per 100,000 person-years and (74, 411) respectively. The interval for the rate is notably wider than that of the Wald-type interval, likely due to the fact that the Hauck-Donner effect results in the uncertainty being underestimated. Given that there are no covariates included in the favoured model, the confidence intervals are for the total population rather than for sub-populations as with the alternative approach to accounting for model uncertainty. However, the confidence interval for the rate of completed suicide computed using only the intercept-only zero-truncated Poisson model is narrower than the sub-population confidence intervals seen in Table 6.1. Similarly, the confidence interval for the Horvitz-Thompson population size is narrower than the confidence interval for the

total population size seen in Table 6.2 as a result of the reduction in bias from not fitting each of the competing models to each iteration.

As the generalised Chao's and generalised Zelterman's estimators utilise a binomial logistic regression model on a truncated dataset rather than the Poisson or negative-binomial distributions that the Horvitz-Thompson estimator utilises, Step 3 in the additional bootstrap algorithm, Algorithm 3, requires further modification. This is done as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for linear predictors $j = 1, \cdots, 4$ in Table 2.1. Record the linear predictor which minimises the BIC, $\tilde{j}_b$.

The linear predictor with the highest proportion of times selected, $\hat{\tilde{j}}$, is used in the bootstrap algorithm to compute the generalised Chao's and generalised Zelterman's population size estimates by modifying Step 3 as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the binomial logistic regression model with linear predictor $\hat{\tilde{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

TABLE 6.5: Proportion of times each linear predictor has lowest BIC statistic from the **non-parametric** bootstrap algorithm for computing the generalised Chao's and generalised Zelterman's estimators for the suicide case study data.

	Linear Predictor			
Distribution	1	2	3	4
Binomial	70.2%	6.1%	18.0%	5.7%

Table 6.5 displays the proportions for the linear predictors being selected as the best fitting model for the non-parametric bootstrap algorithm. Utilising the intercept-only binomial logistic regression model for computing the generalised Chao's and generalised Zelterman's population size estimators in the bootstrap algorithm, given that linear predictor 1 has the highest proportion at 70%, leads to 95% percentile confidence intervals of (64, 697) and (63, 712) respectively.

These intervals are comparable to one another, but are wider than the interval computed using the Horvitz-Thompson estimator, suggesting that for at least smaller sample sizes, the generalised Chao's and generalised Zelterman's estimators have a higher level of uncertainty, and that the total number of studies estimate of 134 using the Horvitz-Thompson estimator is less uncertain than the estimates of 173 and 175 computed by the generalised Chao's and generalised Zelterman's estimators respectively.

Using this approach for accounting for model uncertainty in the bootstrap algorithm provides much less biased and skewed confidence intervals which are also narrower, enabling more inferences to be made. Given that for each bootstrap, the model preferred the highest proportion of times is the same model as preferred for the observed dataset, Method 2 of model selection is the same as Method 3, not accounting for model uncertainty at all and simply using the model preferred for the observed data so this additional bootstrap algorithm is not required. Additionally, for each of the population size estimators, the linear predictor (and distribution) preferred the highest proportion of times is the same as for the observed data, meaning these confidence intervals are the same as for if no model uncertainty was considered.

Application: Hares data

Step 3 of Algorithm 3 can be modified as follows to find the proportion of times each linear predictor is preferred for the Horvitz-Thompson population size estimates to use in the non-parametric bootstrap algorithm.

Step 3: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\tilde{j}_b = 1, \cdots, 5$ be the linear predictor that minimises the AIC. Record the value of $\tilde{j}_b$.

Additionally, Step 3 of Algorithm 4 can be modified as follows to estimate the Horvitz-Thompson population size estimates.

Step 3: Fit the zero-truncated geometric model with linear predictor $\hat{\hat{j}}$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}\right]\right)},$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.6: Proportion of times each linear predictor and distribution combination has lowest AIC statistic from the **non-parametric** bootstrap algorithm for computing the Horvitz-Thompson estimator for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Geometric	0%	1%	0%	2%	97%

Table 6.6 provides the proportions that each linear predictor has the lowest AIC statistic from the non-parametric bootstrap algorithm. The full model is selected as the best fitting model 97% of times, therefore, the zero-truncated geometric model with linear predictor $\hat{\hat{j}} = 5$ is the model used in the bootstrap algorithm to compute the 95% percentile confidence intervals for the Horvitz-Thompson population size estimates.

Additionally, the preferred model for the observed data, the full model, is the same model that is preferred the highest proportion of times in the simulated data. As a result of this, Methods 2 and 3 of accounting for model uncertainty for the snowshoe hares dataset are equal. Therefore, the confidence intervals computed using Method 2 are the same as for if no model uncertainty was considered.

TABLE 6.7: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **non-parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	(278, 341)	(534, 644)	(771, 925)	(1320, 1579)
Five small areas	(275, 336)	(367, 445)	(669, 805)	(1592, 1898)
Total	(557, 671)	(906, 1084)	(1445, 1725)	(2916, 3469)

Table 6.7 provides the 95% percentile confidence intervals of the Horvitz-Thompson population size estimates using the non-parametric bootstrap algorithm, only fitting the full model in each iteration. The results from using Method 2, fitting only the full model each iteration, and the results from using Method 1, fitting all competing models each iteration, are highly comparable, with only small changes in the intervals. This small difference is to be expected given that the full model is preferred 97% of the time, indicating that in Method 1, the full model is fitted in the vast majority of iterations. Therefore, for the hares case study data, there is little benefit in accounting for the additional model uncertainty through using Method 1.

Given that the generalised Chao's and generalised Zelterman's estimators use a different regression model to the Horvitz-Thompson estimator, Step 3 in Algorithm 3 requires additional modification as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for linear predictors $j = 1, \cdots, 5$ in Table 2.3. Record the linear predictor which minimises the AIC, $\tilde{j}_b$.

The linear predictor with the highest proportion of times selected, $\hat{\tilde{j}}$, is used in the bootstrap algorithm to compute the generalised Chao's and generalised Zelterman's population size estimates by modifying Step 3 as follows.

Step 3: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the binomial logistic regression model with linear predictor $\hat{\tilde{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

TABLE 6.8: Proportion of times each linear predictor and distribution combination has lowest AIC statistic from the **non-parametric** bootstrap algorithm for computing the generalised Chao's and generalised Zelterman's estimators for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Binomial	27.1%	33.4%	4.9%	5.3%	29.3%

Table 6.8 provides the proportion of times that each linear predictor is preferred using the AIC statistic for the non-parametric bootstrap algorithm. The model with linear predictor 2 was preferred more times than the other linear predictors, but unlike with the proportions given in Table 6.6 for the geometric regression models, there is no clear preference for one linear predictor over another with the proportions for linear predictors 1, 2 and 5 being very close to each other. The resulting 95% percentile confidence intervals for the generalised Chao's and generalised Zelterman's estimators respectively are (2979, 3873) and (3194, 4196), however, given that there was no clear preference for one linear predictor, it can be assumed that Method 2 of the non-parametric bootstrap algorithm is not the most suitable method.

6.2.3 Method 3: None

Method 3 for the bootstrap algorithm does not account for model uncertainty, instead it fits only the preferred model for the observed data. Formally, this bootstrap is given in Algorithm 5.

Algorithm 5 Bootstrap Approach 1, Method 3

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_n^*, \mathbf{v}_n^*, x_n^*)\}$, through drawing n observations from the original dataset, $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$, at random and with replacement.
Step 3: Fit the preferred model for the observed data to the bootstrapped data and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 4: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Application: Suicide data

Given that for Method 2, the linear predictor and distribution combination preferred the majority of times for the bootstrap data is the same as preferred for the observed data, the results from the non-parametric bootstrap algorithm using Method 2 are the same as the results from using Method 3.

Application: Hares data

Similarly to the suicide case study data, the preferred model for the bootstrap data for the Horvitz-Thompson estimator is the same as for the observed data. Therefore, the results from Methods 2 and 3 for the non-parametric bootstrap algorithms are equal. However, for the generalised Chao's and generalised Zelterman's estimators, the preferred linear predictor for the observed data, $j = 1$, is not the same as the preferred linear predictor for the bootstrapped data, $j = 2$. Therefore, the results from Methods 2 and 3 will vary. To perform Method 3 of the non-parametric bootstrap algorithm, Step 3 of Algorithm 5 can be modified as follows.

Step 3: Let $\tilde{j}$ be the linear predictor that is preferred for the observed data, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

Using Method 3 for the non-parametric bootstrap algorithm results in the 95% percentile confidence intervals for the total number of snowshoe hares (2926, 3750) and (3153,

4092) for the generalised Chao's and generalised Zelterman's estimators respectively. These estimates are comparable to those found using Method 1 for accounting for model uncertainty, although slightly narrower due to the the model uncertainty not being accounted for, and hence less variation was taken into consideration. Given that it is seen in Table 6.8 that there is no clear preference for one model, and that the preferred model for the observed data is only selected approximately one third of the time, there is evidence that Method 3 is not appropriate for this data when using the generalised Chao's and generalised Zelterman's estimators.

6.3 Approach 2: Semi-parametric

Given that a non-parametric approach samples only from the observed data, there is a risk of underestimating the variability of the target population size, as mentioned by Norris and Pollock (1996). This risk occurs as the approach relies on the data being observed and doesn't consider the variance comes from the sampling of n from N and estimating f_0 from n. To avoid this, *the parametric bootstrap approach* by Zwane and van der Heijden (2003) and *Method 3* by Norris and Pollock (1996) can be modified to form an semi-parametric bootstrap algorithm. This approach samples from the total population, made up of the observed data used in the non-parametric approach and the unobserved data which has length equal to the imputed number of missing studies. Through sampling from the complete data, the number of observed studies, n, is treated as a random quantity, and the population size is estimated using a capture-recapture estimator and rounded to the nearest integer.

6.3.1 Method 1: Full

Formally, the semi-parametric bootstrap algorithm accounting for model uncertainty using Method 1 through fitting the competing models in each iteration using Method 1, is as follows in Algorithm 6.

Application: Suicide data

To formulate the semi-parametric bootstrap algorithm to compute the rate and Horvitz-Thompson population size estimates, Step 4 of Algorithm 6 can be modified as follows.

Step 4: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor and $\tilde{D} \in \{\text{Poisson (P)}, \text{negative-binomial (NB)}\}$ be the distribution that minimises the BIC with corresponding maximum likelihood

Algorithm 6 Bootstrap Approach 2, Method 1

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\left\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_{\widehat{N}}^*, \mathbf{v}_{\widehat{N}}^*, x_{\widehat{N}}^*)\right\}$, through drawing $\widehat{N}$ observations from the complete data, $\left\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{\widehat{N}}, \mathbf{v}_{\widehat{N}}, x_{\widehat{N}})\right\}$, at random and with replacement. Given that for the complete data, studies $i = n + 1, \cdots, \widehat{N}$ are unobserved, τ_i and $\mathbf{v}_i$ are unknown, but $x_i = 0$ as the unobserved data are the missing counts of zero events.
Step 3: Truncate studies with a count of events of zero, $x_i^* = 0$ for $i = 1, 2, \cdots, \widehat{N}$. As these studies are truncated, the unknown values of the covariates, $\mathbf{v}_i^*$, and person-years, τ_i^*, are also truncated, so it is unimportant that there is missing covariate information. The sampled observed number of studies is then denoted by n_b^*.
Step 4: Fit the competing models and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 5: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

estimates $\hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}$ of $\boldsymbol{\beta}_j$ for the respective model. If $\tilde{D} = (NB)$, let $\hat{\theta}$ be the estimate of the dispersion parameter.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.80, 1)^T$, $\bar{\mathbf{v}}_4 = (0.80, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$, the sub-population specific estimated rate of completed suicide is calculated as

$$\exp(\eta_{bk}^*) = \exp\left[\left(\mathbf{h}_{\tilde{j}}\left(\bar{\mathbf{v}}_k\right)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}\right]\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(P)}\right]\right)} & \text{if } \tilde{D} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(NB)}\right], \hat{\theta}\right)} & \text{if } \tilde{D} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of studies is $\widehat{N}_b^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.9: Sub-population specific 95% percentile confidence intervals for the rate of completed suicide (per 100,000 person-years) using the **semi-parametric** bootstrap samples.

	Proportion of women		
Country of origin	0.75	0.80	0.85
USA	(9.6, 56.0)	(13.2, 52.8)	(17.7, 51.2)
Other	(17.5, 59.2)	(12.5, 57.7)	(10.6, 57.7)

Table 6.9 displays the 95% percentile confidence intervals for the estimated rate of completed suicide from the semi-parametric bootstrap data for the six sub-populations. The width of these intervals show that there is less uncertainty for studies which originate in the USA and for studies which have a higher proportion of women. These conclusions are expected given that there are more studies from the USA, hence more information to estimate from, and more studies which have a higher proportion of women, in particular for those also originating from the USA. The intervals in Table 6.1 from the non-parametric bootstrap approach are very close to the semi-parametric intervals and also support these conclusions. Given that the intervals are comparable, it suggests that the variance that comes from sampling n from N is not important to the rate estimation and hence either the non-parametric or semi-parametric approaches to the bootstrap algorithm are appropriate for the data.

TABLE 6.10: Values of 95% percentile confidence intervals for the suicide case study data using the Horvitz-Thompson population size estimates from the **semi-parametric** bootstrap samples for the eight sub-populations, marginal totals and overall total.

	Proportion of women				
Country of origin	$[0, 0.75)$	$[0.75, 0.80)$	$[0.8, 0.85)$	$[0.85, 1]$	Total
USA	$(1, 4)$	$(0, 0)$	$(18, 54)$	$(8, 18)$	$(27, 79)$
Other	$(22, 78)$	$(16, 60)$	$(7, 26)$	$(5, 24)$	$(56, 312)$
Total	$(24, 10215)$	$(16, 61)$	$(27, 135)$	$(14, 182)$	$(90, 28913)$

As with the non-parametric bootstrapping approach, high correlation in the bootstrap estimates of the number of studies for the individual sub-populations from randomly resampling the studies to create the new datasets, gives inflated upper limits for the 95% percentile confidence intervals. As a result of this, any interpretations made from these confidence intervals would not be reliable. The additional uncertainty that comes from sampling n from N that is accounted for in the semi-parametric approach makes this correlation increase, with these intervals being much wider than those seen in Table 6.2.

To formulate the semi-parametric bootstrap algorithm for the generalised Chao's and generalised Zelterman's estimators, Step 4 can be modified as follows.

Step 4: If $\sum_{i=1}^{n_b^*}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n_b^*}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for the linear predictors in Table 2.1. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor which minimises the BIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

Similarly to the results from the Horvitz-Thompson estimator, the estimated population size confidence intervals for the generalised Chao's and generalised Zelterman's estimators are very wide through experiencing bias from the high correlation between covariate combinations. For the generalised Chao's estimator, $\widehat{N} = 173$, leading to the 95% percentile confidence interval of (61, 11872). This is also slightly wider than than the respective interval from the non-parametric bootstrap, given the additional uncertainty from treating the observed number of studies as a random variable. The change in width for the 95% percentile confidence interval from the generalised Zelterman's estimator is even larger. Using $\widehat{N} = 175$, the resulting interval is (57, 7483018000), illustrating that there is a high level of uncertainty with the generalised Zelterman's using this bootstrap approach, and hence may not be the best approach to take.

Application: Hares data

To formulate the semi-parametric bootstrap algorithm to compute the Horvitz-Thompson population size estimates, Step 4 of Algorithm 6 can be modified as follows.

Step 4: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor that minimises the AIC with corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ of $\boldsymbol{\beta}_j$ for the respective model.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}\right]\right)}$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.11: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **semi-parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	(279, 478)	(527, 641)	(763, 920)	(1278, 1564)
Five small areas	(255, 333)	(337, 438)	(664, 804)	(1581, 1970)
Total	(542, 790)	(871, 1074)	(1432, 1720)	(2865, 3510)

Table 6.11 provides the 95% percentile confidence intervals for each sub-population, marginal totals and the overall total estimated population size, using the Horvitz-Thompson estimator for the semi-parametric bootstrap algorithm. Comparatively to the results from the non-parametric bootstrap algorithm (Approach 1), the intervals for the semi-parametric bootstrap algorithm are more centred around the corresponding Horvitz-Thompson estimates for the observed data. Similarly to the non-parametric approach however, the intervals are relatively narrow, indicating that there is minimal correlation between the covariates and a small amount of uncertainty, likely due to the large size of the data.

To formulate the semi-parametric bootstrap algorithm for the generalised Chao's and generalised Zelterman's estimators, Step 4 can be modified as follows.

Step 4: If $\sum_{i=1}^{n_b^*}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n_b^*}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for the linear predictors in Table 2.3. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor which minimises the AIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_{\tilde{j}}$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

Using Method 1 for the semi-parametric bootstrap algorithm results in the 95% percentile confidence intervals for the total number of snowshoe hares (2957, 3988) and (3184, 4304) for the generalised Chao's and generalised Zelterman's estimators respectively. These intervals are comparable to the non-parametric intervals also using Method 1, indicating that there is not a notable difference in using Approaches 1 or 2 to the bootstrap algorithm. This result is different to the suicide dataset, where the intervals for the different approaches vary considerably, likely due to the large difference in size of the overall observed dataset.

6.3.2 Method 2: Partial

Alternatively, model uncertainty can be accounted for using Method 2, through an additional bootstrap algorithm to assess which model is preferred the majority of the time. Formally, the additional bootstrap algorithm for comparing the competing models is as follows in Algorithm 7.

Algorithm 7 Bootstrap Approach 2, Method 2, Additional Bootstrap

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\widehat{N}$ studies $\left\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_{\widehat{N}}^*, \mathbf{v}_{\widehat{N}}^*, x_{\widehat{N}}^*)\right\}$, from the complete data, $\left\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{\widehat{N}}, \mathbf{v}_{\widehat{N}}, x_{\widehat{N}})\right\}$, at random and with replacement. Given that for the complete data, studies $i = n+1, \cdots, \widehat{N}$ are unobserved, τ_i and $\mathbf{v}_i$ are unknown, but $x_i = 0$ as the unobserved data are the missing counts of zero events.
Step 3: Truncate studies with a count of events of zero, $x_i^* = 0$ for $i = 1, 2, \cdots, \widehat{N}$. As these studies are truncated, the unknown values of the covariates, $\mathbf{v}_i^*$, and person-years, τ_i^* are also truncated, so it is unimportant that there is missing covariate information. The sampled observed number of studies is then denoted by n_b^*.
Step 4: Fit the competing models to the bootstrap data. Let $(\tilde{j}_b, \tilde{D}_b)$ be the linear predictor and distribution combination that minimises the AIC or BIC.
Step 5: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Compute the proportion that each combination of $(\tilde{j}, \tilde{D})$ is selected as the best using the results from the additional bootstrap. Use this linear predictor and distribution combination in the semi-parametric bootstrap algorithm to compute the rate and population size estimates. Formally, the semi-parametric bootstrap algorithm is then as follows in Algorithm 8.

Algorithm 8 Bootstrap Approach 2, Method 2

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\left\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_{\widehat{N}}^*, \mathbf{v}_{\widehat{N}}^*, x_{\widehat{N}}^*)\right\}$, through drawing $\widehat{N}$ observations from the complete data, $\left\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{\widehat{N}}, \mathbf{v}_{\widehat{N}}, x_{\widehat{N}})\right\}$, at random and with replacement. Given that for the complete data, studies $i = n + 1, \cdots, \widehat{N}$ are unobserved, τ_i and $\mathbf{v}_i$ are unknown, but $x_i = 0$ as the unobserved data are the missing counts of zero events.
Step 3: Truncate studies with a count of events of zero, $x_i^* = 0$ for $i = 1, 2, \cdots, \widehat{N}$. As these studies are truncated, the unknown values of the covariates, $\mathbf{v}_i^*$, and person-years, τ_i^* are also truncated, so it is unimportant that there is missing covariate information. The sampled observed number of studies is then denoted by n_b^*.
Step 4: Fit the model with linear predictor and distribution combination $(\tilde{j}, \tilde{D})$ found using Algorithm 7 and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 5: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Application: Suicide data

Step 4 of Algorithm 7 can be modified as follows to find the proportion of times each linear predictor and distribution combination is preferred for the rate and Horvitz-Thompson population size estimates to use in the semi-parametric bootstrap algorithm.

Step 4: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j}_b = 1, \cdots, 5$ be the linear predictor and $\tilde{D}_b \in \{\text{Poisson (P)}, \text{negative-binomial (NB)}\}$ be the distribution that minimises the BIC. Record the combination $(\tilde{j}_b, \tilde{D}_b)$.

Additionally, Step 4 of Algorithm 8 can be modified as follows to estimate the rate and Horvitz-Thompson population size estimates.

Step 4: Fit the zero-truncated model with linear predictor and distribution combination $(\hat{\tilde{j}}, \hat{\tilde{D}})$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(\hat{\tilde{D}})}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$ and $\hat{\theta}$ be the estimate of the dispersion parameter if $\hat{\tilde{D}} = (NB)$.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.80, 1)^T$, $\bar{\mathbf{v}}_4 = (0.80, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$, the sub-population specific estimated rate of completed suicide is calculated as

$$\exp(\eta_{bk}^*) = \exp\left[\left(\mathbf{h}_{\hat{\tilde{j}}}(\bar{\mathbf{v}}_k)\right)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(\hat{\tilde{D}})}\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}^{(P)}\right]\right)} & \text{if } \hat{\hat{D}} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}^{(NB)}\right], \hat{\theta}\right)} & \text{if } \hat{\hat{D}} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of studies is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.12: Proportion of times each linear predictor and distribution combination has lowest BIC statistic from the **semi-parametric** bootstrap algorithm for computing the rate and Horvitz-Thompson estimator for the suicide case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Poisson	78.2%	15.3%	3.4%	1.3%	1.7%
Negative-binomial	0.1%	0.0%	0.0%	0.0%	0.0%

Table 6.12 displays the proportions that each linear predictor and distribution combination has the lowest BIC statistic from the semi-parametric bootstrap algorithm. As with the non-parametric bootstrap approach, the intercept-only zero-truncated Poisson model is preferred for approximately 80% of the iterations. Therefore, using the linear predictor and distribution combination $(\hat{\hat{j}}, \hat{\hat{D}}) = (1, P)$, the 95% percentile confidence intervals for the rate and total population size from the Horvitz-Thompson estimator are (24.9, 50.2) and (91,166) respectively. The interval for the rate is similar to that using the original approach, but the interval for the population size is much narrower as a result of the reduction of bias from the correlation between covariate combinations.

Whilst the Horvitz-Thompson estimator above uses either the Poisson or negative-binomial distribution, the generalised Chao's and generalised Zelterman's estimators use binomial logistic regression, so Step 4 of Algorithm 7 can be modified as follows.

Step 4: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 2. Otherwise, fit competing binomial logistic regression models for the linear predictors in Table 2.1. Record the linear predictor, $\tilde{j}_b$, which minimises the BIC.

Additionally, Step 4 of Algorithm 8 can be modified as follows to estimate the generalised Chao's and generalised Zelterman's population size estimates.

Step 4: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the binomial logistic

regression model with linear predictor $\hat{\hat{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

TABLE 6.13: Proportion of times each linear predictor has lowest BIC statistic from the **semi-parametric** bootstrap algorithm for computing the generalised Chao's estimator for the suicide case study data.

	Linear Predictor			
Distribution	1	2	3	4
Binomial	67.5%	7.6%	20.2%	4.6%

TABLE 6.14: Proportion of times each linear predictor has lowest BIC statistic from the **semi-parametric** bootstrap algorithm for computing the generalised Zelterman's estimator for the suicide case study data.

	Linear Predictor			
Distribution	1	2	3	4
Binomial	67.5%	8.3%	20.1%	4.2%

Proportions that each linear predictor is preferred for the binomial logistic regression models for the generalised Chao's and generalised Zelterman's from the semi-parametric bootstrap algorithm are displayed in Tables 6.13 and 6.14 respectively, where the intercept-only model is preferred the majority of times. Therefore, linear predictor 1 is utilised for computing the estimates for each bootstrap iteration in order to find the 95% confidence intervals of (59, 727) and (54, 750) respectively. As with the non-parametric approach, these intervals are comparable to one another, but are much

narrower than the respective intervals computed through fitting each of the competing models for each iteration

Application: Hares data

Step 4 of Algorithm 7 can be modified as follows to find the proportion of times each linear predictor is preferred for the Horvitz-Thompson population size estimates to use in the non-parametric bootstrap algorithm.

Step 4: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\tilde{j}_b = 1, \cdots, 5$ be the linear predictor that minimises the AIC. Record the value of $\tilde{j}_b$.

Additionally, Step 4 of Algorithm 8 can be modified as follows to estimate the Horvitz-Thompson population size estimates.

Step 4: Fit the zero-truncated geometric model with linear predictor $\hat{\tilde{j}}$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\hat{\tilde{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}\right]\right)},$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.15: Proportion of times each linear predictor and distribution combination has lowest AIC statistic from the **semi-parametric** bootstrap algorithm for computing the Horvitz-Thompson estimator for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Geometric	0.0%	0.6%	0.0%	2.1%	97.3%

Table 6.24 provides the proportions that each linear predictor has the lowest AIC statistic from the semi-parametric bootstrap algorithm. As with the non-parametric bootstrap algorithm, the full model is preferred 97% of the time, and therefore the zero-truncated geometric model with linear predictor $\hat{\tilde{j}} = 5$ is used in the bootstrap algorithm to compute the 95% percentile confidence intervals for the Horvitz-Thompson population size estimates.

TABLE 6.16: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **semi-parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	(279, 475)	(529, 640)	(763, 922)	(1275, 1561)
Five small areas	(253, 334)	(336, 437)	(665, 804)	(1584, 1965)
Total	(541, 786)	(872, 1072)	(1432, 1722)	(2865, 3509)

The 95% percentile confidence intervals for the Horvitz-Thompson population size estimates using the semi-parametric bootstrap algorithm are provided in Table 6.16. Similarly to the non-parametric bootstrap Method 2, the results in Table 6.16 are comparable to the results using the semi-parametric bootstrap with Method 1 for accounting for model uncertainty, due to the full model being preferred such a high proportion of times. Given the similarity between methods for model uncertainty accountability, it is reasonable to say that there is little benefit in fitting each of the competing models and the increase in computational time associated with the method.

Additionally, the preferred model for the observed data (the full model) is the same model that is preferred the highest proportion of times in the simulated data. As a result of this, Methods 2 and 3 of accounting for model uncertainty for the hares case study data are equal. Therefore, the confidence intervals computed using Method 2 are the same as for if no model uncertainty was considered.

Given that the generalised Chao's and generalised Zelterman's estimators use a different regression model to the Horvitz-Thompson estimator, Step 4 in Algorithm 7 requires additional modification as follows.

Step 4: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 2. Otherwise, fit competing binomial logistic regression models for the linear predictors in Table 2.1. Record the linear predictor, $\tilde{j}_b$, which minimises the BIC.

The linear predictor with the highest proportion of times selected, $\hat{\tilde{j}}$, is used in the bootstrap algorithm to compute the generalised Chao's and generalised Zelterman's population size estimates by modifying Step 4 of Algorithm 8 as follows.

Step 4: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the binomial logistic regression model with linear predictor $\hat{\tilde{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

TABLE 6.17: Proportion of times each linear predictor has lowest AIC statistic from the **semi-parametric** bootstrap algorithm for computing the generalised Chao's estimator for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Binomial	23.7%	35.0%	4.9%	6.5%	29.9%

TABLE 6.18: Proportion of times each linear predictor has lowest AIC statistic from the **semi-parametric** bootstrap algorithm for computing the generalised Zelterman's estimator for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Binomial	25.1%	34.4%	5.0%	6.0%	29.5%

As with the results from Method 2 of Approach 1, the binomial logistic regression model with linear predictor 2 is preferred the most amount of times for both the generalised Chao's and generalised Zelterman's estimators. Another similarity is that there is no clear preference for a singular linear predictor using this method, with linear predictors 2 and 5 being preferred a similar amount of times. Using the semi-parametric bootstrap algorithm with only the model with linear predictor 2 fitted in each iteration leads to the 95% percentile confidence intervals (2937, 3892) and (3154, 4215) for the generalised Chao's and generalised Zelterman's estimators respectively. However, whilst the results are comparable to the intervals from alternative approaches and methods, given that there is no clear preference for a singular linear predictor, there is evidence that Method 2 is not the most appropriate method for the snowshoe hares dataset.

6.3.3 Method 3: None

Method 3 for the bootstrap algorithm fits only the preferred model for the observed data to each iteration and consequently does not account for model uncertainty. Formally, this bootstrap is given in Algorithm 9.

Algorithm 9 Bootstrap Approach 2, Method 3

Step 1: Let $b = 1$.
Step 2: Build a bootstrap sampled dataset, $\{(\tau_1^*, \mathbf{v}_1^*, x_1^*), \ldots, (\tau_n^*, \mathbf{v}_n^*, x_n^*)\}$, through drawing n observations from the complete data, $\left\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{\widehat{N}}, \mathbf{v}_{\widehat{N}}, x_{\widehat{N}})\right\}$, at random and with replacement. Given that for the complete data, studies $i = n + 1, \cdots, \widehat{N}$ are unobserved, τ_i and $\mathbf{v}_i$ are unknown, but $x_i = 0$ as the observed data are missing the counts of zero events.
Step 3: Truncate studies with a count of events of zero, $x_i^* = 0$ for $i = 1, 2, \cdots, \widehat{N}$. As these studies are truncated, the unknown values of the covariates, $\mathbf{v}_i^*$, and person-years, τ_i^* are also truncated, so it is unimportant that there is missing covariate information. The sampled observed number of studies is then denoted by n_b^*.
Step 4: Fit the preferred model for the observed data to the bootstrapped data and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 5: If $b = B$, stop. Otherwise, return to Step 2 with $b := b + 1$.

Application: Suicide data

As with Approach 1, the preferred model for the observed data is the same model found to be preferred the majority of times for use in Method 2. Therefore, the linear predictor and distribution combination used in the semi-parametric bootstrap algorithm with Method 2 is the same that would be used for Method 3 and the results will be identical.

Application: Hares data

Similarly to the suicide case study data, the linear predictor preferred the majority of times for the semi-parametric bootstrap algorithm with the Horvitz-Thompson estimator is the same as is preferred for the observed data. Therefore, the results from Methods 2 and 3 for the semi-parametric bootstrap algorithms are equal. The preferred models for the generalised Chao's and generalised Zelterman's estimators are not however the same for the observed data as found in the bootstrap in Method 2. As a result, Step 4 of Algorithm 9 can be modified as follows.

Step 4: Let $\tilde{j}$ be the linear predictor that is preferred for the observed data, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

Using Method 3 for the semi-parametric bootstrap algorithm results the 95% percentile confidence intervals for the total number of snowshoe hares of (2890, 3785) and (3113, 4128) for the generalised Chao's and generalised Zelterman's estimators respectively, intervals which are comparative to those found through both Approaches 1 and 2 and different methods for accounting for model uncertainty. However, whilst the intervals are comparable to others found through alternative methods, given that there is no clear preference for a single model as found in the bootstrap in Method 2, there is evidence that Method 3 is not appropriate for this case study.

6.4 Approach 3: Parametric

For the parametric approach, response variable values are sampled from a given distribution in order to create a dataset that can be used for modelling as with the other methods. Fitting the ten competing models to the observed data and calculating the corresponding AIC or BIC weights of each of the models gives the probability that each model is selected as the best candidate model (Wagenmakers and Farrell, 2004), where the AIC or BIC weights are calculated respectively as

$$w_l(AIC) = \frac{\exp\left[-\frac{1}{2}\Delta_l(AIC)\right]}{\sum_{k=1}^{K}\exp\left[-\frac{1}{2}\Delta_k(AIC)\right]} \text{ and } w_l(BIC) = \frac{\exp\left[-\frac{1}{2}\Delta_l(BIC)\right]}{\sum_{k=1}^{K}\exp\left[-\frac{1}{2}\Delta_k(BIC)\right]}, \quad (6.1)$$

where $\Delta_l(AIC) = AIC_l - \min(AIC)$ and $\Delta_l(BIC) = BIC_l - \min(BIC)$ are the difference between the either the AIC or BIC value for each model respectively, and the best

candidate model for $l = 1, 2, \cdots, J$, where J is the total number of linear predictors under consideration. Sampling l^* from $\{1, 2, \cdots, J\}$, where each value has corresponding probability $\{w_1, w_2, \cdots, w_J\}$, gives the linear predictor and distribution pair (j^*, D^*) used to sample the response variables.

6.4.1 Method 1: Full

Formally, the parametric bootstrap algorithm using Method 1 for model selection is as follows in Algorithm 10.

Algorithm 10 Bootstrap Approach 3, Method 1

Step 1: Fit the competing models to the observed data.
Step 2: For each model, compute the AIC or BIC weights, w_j, using Equation 6.1, where the AIC or BIC weight of each model, l, can be seen as the probability of model l being selected as the best candidate model (Wagenmakers and Farrell, 2004).
Step 3: Let $b = 1$.
Step 4: Sample l^* from $\{1, 2, \cdots, J\}$, where J is the number of models under consideration and the values have respective probabilities $\{w_1, w_2, \cdots, w_J\}$ of being sampled.
Step 5: Sample n counts from the model given by the linear predictor distribution pair (j^*, D^*). Use these sampled counts x_i^* to create a sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1^*), \ldots, (\tau_n, \mathbf{v}_n, x_n^*)\}$, where τ_i and $\mathbf{v}_i$ are the observed person-years and covariates respectively for $i = 1, 2, \cdots, n$. Sample x_i^* from the distribution given by (j^*, D^*) with probability function

$$\begin{cases} p_x^+\left(\tau_i \exp\left[\mathbf{h}_{j^*}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{j^*}^{(D^*)}\right]\right) & \text{if } D^* = (P), \\ p_x^+\left(\tau_i \exp\left[\mathbf{h}_{j^*}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{j^*}^{(D^*)}\right], \hat{\theta}_{j^*}^{(D^*)}\right) & \text{if } D^* = (NB), \end{cases}$$

where if $D^* = (NB)$, $\hat{\theta}_{j^*}^{D^*}$ is the estimated dispersion parameter.
Step 6: Fit the competing models and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 7: If $b = B$, stop. Otherwise, return to Step 4 with $b := b + 1$.

Application: Suicide data

Using the same 6 sub-populations for the covariate combinations defined in Section 6.2, the parametric bootstrap algorithm for computing the rate and Horvitz-Thompson estimator, accounting for model uncertainty using Method 1 through fitting the competing models in each iteration, is given by modifying Step 6 of Algorithm 10 as follows.

Step 6: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor and $\tilde{D} \in \{\text{Poisson (P), negative-binomial (NB)}\}$ be

the distribution that minimises the BIC with corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}$ of $\boldsymbol{\beta}_j$ for the respective model. If $\tilde{D} = (NB)$, let $\hat{\theta}$ be the estimate of the dispersion parameter.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.80, 1)^T$, $\bar{\mathbf{v}}_4 = (0.80, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$ are the covariates for each of the 6 sub-populations, the sub-population specific estimated rate of completed suicide is calculated as

$$\exp(\eta_{bk}^*) = \exp\left[\mathbf{h}_{\tilde{j}}(\bar{\mathbf{v}}_k)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(\tilde{D})}\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(P)}\right]\right)} & \text{if } \tilde{D} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\tilde{j}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}^{(NB)}\right], \hat{\theta}\right)} & \text{if } \tilde{D} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of studies is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.19: Sub-population specific 95% percentile confidence intervals for the rate of completed suicide (per 100,000 person-years) using the **parametric** bootstrap samples.

	Proportion of women		
Country of origin	0.75	0.80	0.85
USA	(17.5, 42.6)	(18.9, 42.6)	(19.7, 42.6)
Other	(15.2, 57.7)	(16.3, 59.2)	(16.4, 59.2)

Using this parametric bootstrap, 95% percentile intervals for the six sub-populations can be calculated as seen in Table 6.19. Each of the intervals are approximately centred at the estimated rate of 31.8 completed suicides per 100,000 person-years, calculated using regression modelling and analytical methods in the previous chapter. As in Section 6.2, the confidence intervals for USA are narrower for higher proportions of women given that there is more data available for these covariate combinations, reducing the level of uncertainty. Additionally, whilst there is some decrease in uncertainty as the proportion of women increases for studies originating outside of the USA with the corresponding confidence intervals narrowing slightly, it is not as significant of a change given that in studies originating outside of the USA, the proportion of women varies more comparatively to the inside the USA. There is also increased uncertainty for studies originating outside of the USA with a proportion of women of 0.85 given that there are more studies observed from the USA with this covariate combination. The reverse

is true, where for covariate combinations with reduced proportion of women, there is more uncertainty with wider confidence intervals for those originating in the USA as there are fewer corresponding studies observed.

TABLE 6.20: Values of 95% percentile confidence intervals for the suicide case study data using the Horvitz-Thompson population size estimates from the **parametric** bootstrap samples for the eight sub-populations, marginal totals and overall total.

	Proportion of women				
Country of origin	$[0, 0.75)$	$[0.75, 0.80)$	$[0.8, 0.85)$	$[0.85, 1]$	Total
USA	$(1, 2)$	$(0, 0)$	$(22, 42)$	$(9, 16)$	$(32, 60)$
Other	$(29, 124)$	$(16, 50)$	$(7, 19)$	$(5, 15)$	$(58, 223)$
Total	$(30, 156)$	$(16, 50)$	$(31, 59)$	$(15, 29)$	$(101, 280)$

The 95% percentile confidence intervals are also calculated for the estimated population sizes, as seen in Table 6.20. Model uncertainty is accounted for through allowing for various linear predictors in the model selection of the resampled data. However, unlike in Sections 6.2 and 6.3, the resampled dataset is not created at random but instead using a distributional model representative for the observed data, creating a more reliable dataset to model from. This reduces the unexpected correlation between the covariates seen in Tables 6.2 and 6.10, giving much more reliable confidence intervals that are better centred around the corresponding population size estimates from the observed data, suggesting that a parametric approach to the bootstrap algorithm is most suitable.

To formulate the semi-parametric bootstrap algorithm for the generalised Chao's and generalised Zelterman's estimators, Step 6 of Algorithm 10 can be modified as follows.

Step 6: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 4. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for the linear predictors in Table 2.1. Let $\tilde{j} = 1, \cdots, 4$ be the linear predictor which minimises the BIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

As for the generalised Chao's and generalised Zelterman's estimators, the respective 95% percentile confidence intervals are (62, 1081) and (61, 9846). Whilst these intervals are much narrower than the non-parametric and semi-parametric bootstrap approaches, hence having less uncertainty, the intervals still much wider than the interval for the total population size using the Horvitz-Thompson estimator and the parametric bootstrap. This is a trend seen for each of the bootstrap approaches, and is to be expected given that the generalised Chao's and generalised Zelterman's estimators utilise only the counts of one and two from the dataset. Given that the number of observed studies is already small at 27, leading to more uncertainty compared to if there was a higher number of observed studies, once truncated for use of the generalised Chao's and generalised Zelterman's estimators, this observed number of studies gets even smaller leading to more uncertainty.

Application: Hares data

To formulate the semi-parametric bootstrap algorithm to compute the Horvitz-Thompson population size estimates, Step 6 of Algorithm 10 can be modified as follows.

Step 6: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\hat{\hat{j}} = 1, \cdots, 5$ be the linear predictor that minimises the AIC with corresponding maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}$ of $\boldsymbol{\beta}_j$ for the respective model.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}\right]\right)},$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_b^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

Table 6.21 contains the 95% percentile confidence intervals from the parametric bootstrap algorithm for the sub-populations, sub-totals and total number of snowshoe hares.

TABLE 6.21: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	(235, 525)	(391, 537)	(789, 1134)	(1204, 1569)
Five small areas	(194, 319)	(242, 333)	(686, 1045)	(1521, 1969)
Total	(472, 740)	(666, 821)	(1575, 2030)	(2829, 3405)

The interval is comparable to the results from Method 1 Approach 1 to the bootstrap algorithm, and slightly narrower than Method 1 Approach 2 to the bootstrap algorithm. Given the comparability to the other approaches, and that the interval is approximately centred at the estimated total number of snowshoe hares from the observed data, there is evidence that this 95% percentile confidence interval is appropriate and reliable for the given data.

To formulate the parametric bootstrap algorithm for the generalised Chao's and generalised Zelterman's estimators, Step 6 can be modified as follows.

Step 6: If $\sum_{i=1}^{n_b^*}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n_b^*}(x_i^* = 2) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for the linear predictors in Table 2.3. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor which minimises the AIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

Using Method 1 for the parametric bootstrap algorithm results in the 95% percentile confidence intervals for the total number of snowshoe hares of (3015, 4027) and (3224, 4324) for the generalised Chao's and generalised Zelterman's estimators respectively. These intervals are comparable in width to the corresponding intervals for the non-parametric and semi-parametric bootstrap algorithms, however, both the upper and lower limits of each are higher.

6.4.2 Method 2: Partial

Similarly to the other bootstrap approaches, using an alternative approach to account for the model uncertainty may reduce the level of uncertainty seen in the confidence intervals. The additional bootstrap algorithm for computing the proportions each model is preferred is given as follows in Algorithm 11

Algorithm 11 Bootstrap Approach 3, Method 2, Additional Bootstrap

Step 1: Fit the competing models to the observed data.
Step 2: For each model, compute the AIC or BIC weights, w_j, using Equation 6.1, where the AIC or BIC weight of each model, l, can be seen as the probability of model l being selected as the best candidate model (Wagenmakers and Farrell, 2004).
Step 3: Let $b = 1$.
Step 4: Sample l^* from $\{1, 2, \cdots, J\}$, where J is the number of models under consideration and the values have respective probabilities $\{w_1, w_2, \cdots, w_J\}$ of being sampled.
Step 5: Sample n counts from the model given by the linear predictor and distribution pair, (j^*, D^*), given by the sampled value of l^*. Use these sampled counts, x_i^*, to create a sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1^*), \ldots, (\tau_n, \mathbf{v}_n, x_n^*)\}$, where τ_i and $\mathbf{v}_i$ are the observed person-years and covariates respectively for $i = 1, 2, \cdots, n$.
Step 6: Fit the competing models to the bootstrap data. Let $(\tilde{j}, \tilde{D})$ be the linear predictor and distribution combination that minimises the AIC or BIC.
Step 7: If $b = B$, stop. Otherwise, return to Step 4 with $b := b + 1$.

Compute the proportion that each combination of $(\tilde{j}, \tilde{D})$ is selected as the best using the results from the additional bootstrap. Use this linear predictor and distribution combination in the parametric bootstrap algorithm to compute the rate and population size estimates. Formally, the parametric bootstrap algorithm is then as follows in Algorithm 11.

Application: Suicide data

Step 6 of Algorithm 11 can be modified as follows to find the proportion of times each linear predictor and distribution combination is preferred for the rate and Horvitz-Thompson population size estimates to use in the parametric bootstrap algorithm.

Algorithm 12 Bootstrap Approach 3, Method 2

Step 1: Fit the competing models to the observed data.
Step 2: For each model, compute the AIC or BIC weights, w_j, using Equation 6.1, where the AIC or BIC weight of each model, l, can be seen as the probability of model l being selected as the best candidate model (Wagenmakers and Farrell, 2004).
Step 3: Let $b = 1$.
Step 4: Sample l^* from $\{1, 2, \cdots, J\}$, where J is the number of models under consideration and the values have respective probabilities $\{w_1, w_2, \cdots, w_J\}$ of being sampled.
Step 5: Sample n counts from the model given by the linear predictor distribution pair, (j^*, D^*). Use these sampled counts x_i^* to create a sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1^*), \ldots, (\tau_n, \mathbf{v}_n, x_n^*)\}$, where τ_i and $\mathbf{v}_i$ are the observed person-years and covariates respectively for $i = 1, 2, \cdots, n$.
Step 6: Fit the model with linear predictor and distribution combination $(\tilde{j}, \tilde{D})$ found using Algorithm 11 and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 7: If $b = B$, stop. Otherwise, return to Step 4 with $b := b + 1$.

Step 6: Fit the ten competing models given by each of the linear predictors in Table 2.1 for both the zero-truncated Poisson and negative-binomial distributions. Let $\tilde{j}_b = 1, \cdots, 5$ be the linear predictor and $\tilde{D}_b \in \{\text{Poisson (P), negative-binomial (NB)}\}$ be the distribution that minimises the BIC. Record the combination $(\tilde{j}, \tilde{D})$.

Additionally, Step 6 of Algorithm 12 can be modified as follows to estimate the rate and Horvitz-Thompson population size estimates.

Step 6: Fit the zero-truncated model with linear predictor and distribution combination $(\hat{\tilde{j}}, \hat{\tilde{D}})$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(\hat{\tilde{D}})}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$ and $\hat{\theta}$ be the estimate of the dispersion parameter if $\hat{\tilde{D}} = (NB)$.

Rate: Given $\bar{\mathbf{v}}_1 = (0.75, 1)^T$, $\bar{\mathbf{v}}_2 = (0.75, 0)^T$, $\bar{\mathbf{v}}_3 = (0.80, 1)^T$, $\bar{\mathbf{v}}_4 = (0.80, 0)^T$, $\bar{\mathbf{v}}_5 = (0.85, 1)^T$, and $\bar{\mathbf{v}}_6 = (0.85, 0)^T$, the sub-population specific estimated rate of completed suicide is calculated as

$$\exp(\eta_{bk}^*) = \exp\left[(\mathbf{h}_{\hat{\tilde{j}}}(\bar{\mathbf{v}}_k)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(\hat{\tilde{D}})}\right],$$

for $k = 1, \ldots, 6$.

Horvitz-Thompson: The estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \begin{cases} \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\tilde{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(P)}\right]\right)} & \text{if } \hat{\tilde{D}} = P, \\ \dfrac{1}{1 - p_0\left(\tau_i \exp\left[\mathbf{h}_{\hat{\tilde{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}^{(NB)}\right], \hat{\theta}\right)} & \text{if } \hat{\tilde{D}} = NB, \end{cases}$$

where the number of studies in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of studies is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.22: Proportion of times each linear predictor and distribution combination has lowest BIC statistic from the **semi-parametric** bootstrap algorithm for computing the rate and Horvitz-Thompson estimator for the suicide case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Poisson	78.0%	9.9%	8.2%	2.1%	1.0%
Negative-binomial	0.7%	0.1%	0.0%	0.0%	0.0%

The proportions of each linear predictor and distribution combination having the lowest BIC statistic from the parametric bootstrap algorithm are given in Table 6.22. For over 80% of the iterations, the intercept-only zero-truncated Poisson model is preferred, as with both the non-parametric and semi-parametric approaches. Using the linear predictor and distribution combination $(\hat{\hat{j}}, \hat{\hat{D}}) = (1, P)$ to compute the 95% percentile confidence intervals for the rate and total population size from the Horvitz-Thompson estimator leads to the intervals of (22.2, 41.9) and (106, 185) respectively. As with the other bootstrap approaches, the rate interval is not notably impacted by the change in approach of accounting for model uncertainty, but there is a notable decrease in the width of the population size confidence interval from the reduction in bias.

Step 6 of the parametric bootstrap algorithm is modified as follows to reflect Method 2 to account for model uncertainty using the generalised Chao's and generalised Zelterman's estimators is.

Step 6: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 4. Otherwise, truncate the bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit competing binomial logistic regression models for the linear predictors in Table 2.1. Let $\tilde{j}_b = 1, \cdots, 4$ be the linear predictor which minimises the BIC.

Additionally, Step 6 of Algorithm 12 can be modified as follows to estimate the generalised Chao's and generalised Zelterman's population size estimates.

Step 6: If $\sum_{i=1}^{n}(x_i^* = 1) = 0$ or $\sum_{i=1}^{n}(x_i^* = 2) = 0$, return to Step 4. Otherwise, truncate bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the logistic regression model with linear predictor $\hat{\hat{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = 2\tau_i \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for studies with the same person-years and covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

TABLE 6.23: Proportion of times each linear predictor has lowest BIC statistic from the **semi-parametric** bootstrap algorithm for computing the generalised Chao's and generalised Zelterman's estimators for the suicide case study data.

	Linear Predictor			
Distribution	1	2	3	4
Binomial	81.0%	9.0%	8.2%	1.8%

Table 6.23 provides the proportions that each linear predictor is preferred for the parametric bootstrap algorithm. As with the non-parametric and semi-parametric approaches, the intercept-only binomial logistic model is preferred the majority of the times (over 80%). Using this model in the parametric algorithm then leads to the 95% percentile confidence intervals for the generalised Chao's and generalised Zelterman's estimators of (61, 573) and (59, 580) respectively, where both intervals are narrower than the respective intervals found using the alternative approach of accounting for model uncertainty of Method 1.

Application: Hares data

Step 6 of Algorithm 11 can be modified as follows to find the proportion of times each linear predictor and distribution combination is preferred for the Horvitz-Thompson population size estimates to use in the parametric bootstrap algorithm.

Step 6: Fit the five competing models given by each of the linear predictors in Table 2.3 for the zero-truncated geometric distribution. Let $\hat{\tilde{j}}_b = 1, \cdots, 5$ be the linear predictor that minimises the AIC. Record the value of $\tilde{j}_b$.

Additionally, Step 6 of Algorithm 12 can be modified as follows to estimate the Horvitz-Thompson population size estimates.

Step 6: Fit the zero-truncated geometric model with linear predictor $\hat{\hat{j}}$ to the sampled dataset. Let $\hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Horvitz-Thompson: The estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(HT)*} = \frac{1}{1 - p_0\left(\exp\left[\mathbf{h}_{\hat{\hat{j}}}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\hat{\hat{j}}}\right]\right)},$$

where the number of snowshoe hares in sub-population A is $\widehat{N}_{bA}^{(HT)*} = \sum_{i \in A} \widehat{N}_{bi}^{(HT)*}$ and the estimated total number of snowshoe hares is $\widehat{N}_{b}^{(HT)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(HT)*}$.

TABLE 6.24: Proportion of times each linear predictor has lowest AIC statistic from the **parametric** bootstrap algorithm for computing the Horvitz-Thompson estimator for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Geometric	0.0%	1.3%	0.0%	2.0%	96.7%

Table 6.24 provides the proportion of times each linear predictor is preferred for the parametric bootstrap algorithm for the Horvitz-Thompson estimator. Linear predictor 5, the full model, is preferred 97% of the time, the vast majority of occasions and therefore is the linear predictor to be used in the modelling within each iteration for the parametric bootstrap algorithm.

TABLE 6.25: Values of 95% percentile confidence intervals for the hares case study data using the Horvitz-Thompson population size estimates from the **parametric** bootstrap samples for the six sub-populations, marginal totals and overall total.

	Season			
Study area	Midwinter	Spring	Summer	Total
Square mile area	(255, 383)	(515, 648)	(716, 938)	(1173, 1474)
Five small areas	(168, 218)	(294, 358)	(683, 934)	(1535, 1896)
Total	(439, 577)	(831, 977)	(1460, 1791)	(2791, 3253)

Table 6.25 contains the values of the 95% percentile confidence intervals for the Horvitz-Thompson estimates using the parametric bootstrap algorithm. The results from this bootstrap algorithm are comparable to the alternative approaches, however, the interval corresponding to the total number of studies is not centred around the estimated value using the Horvitz-Thompson estimator for the observed data. Possibly indicating that

this method and approach combination is not as suitable for this dataset as Method 1 for accounting for model uncertainty.

As with Approaches 1 and 2 to the bootstrap algorithm with Method 2 for accounting for model uncertainty, the preferred model for the observed data, the full data, is the same model that is preferred the highest proportion of times in the simulated data. Therefore, Methods 2 and 3 for accounting for model uncertainty with the parametric bootstrap algorithm will produce identical results.

Given that the generalised Chao's and generalised Zelterman's estimators use a binomial regression model compared to the geometric regression model used for the Horvitz-Thompson estimator, Step 6 in Algorithm 11 requires additional modification as follows.

Step 6: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 4. Otherwise, truncate the bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the competing binomial logistic regression models for the linear predictors in Table 2.3. Let $\tilde{j} = 1, \cdots, 5$ be the linear predictor which minimises the BIC.

The linear predictor that is preferred the highest proportion of times, $\hat{\tilde{j}}$, is used in the bootstrap algorithm to compute the generalised Chao's and generalised Zelterman's population size estimates by modifying Step 6 of Algorithm 12 as follows.

Step 6: If $\sum_{i=1}^{\widehat{N}}(x_i^* = 1) = 0$ or $\sum_{i=1}^{\widehat{N}}(x_i^* = 2) = 0$, return to Step 4. Otherwise, truncate the bootstrap dataset for all counts except $X = 1$ and $X = 2$. Fit the binomial logistic regression model with linear predictor $\hat{\tilde{j}}$ to the sampled truncated dataset. Let $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\hat{\tilde{j}}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as study i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1-\exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$

TABLE 6.26: Proportion of times each linear predictor has lowest AIC statistic from the **parametric** bootstrap algorithm for computing the generalised Chao's and generalised Zelterman's estimators for the hares case study data.

	Linear Predictor				
Distribution	1	2	3	4	5
Geometric	0.0%	99.1%	0.0%	0.0%	0.9%

Table 6.26 provides the proportions that each linear predictor has the lowest AIC statistic from the parametric bootstrap algorithm. As with the other approaches to the bootstrap algorithm with Method 2 for accounting for model uncertainty, linear predictor 2 is preferred the majority of times, however, in this case it is preferred 99% of times, a clear majority that is not seen in the alternative approaches.

Fitting only models with linear predictor 2 in the parametric bootstrap algorithm results in the 95% percentile confidence intervals of (2952, 3797) and (3166, 4115) for the generalised Chao's and generalised Zelterman's estimators respectively. These intervals are both comparable to the intervals computed using alternative methods and approaches and approximately centred at the corresponding estimates of the total number of snowshoe hares from the observed data. Given this and that there is a clear preference for linear predictor 2, there is evidence that there is little benefit in the additional computational time that is required for Method 1 of accounting for model uncertainty for the snowshoe hares dataset, and Method 2 can be used instead.

6.4.3 Method 3: None

As with the other approaches to the bootstrap algorithm, method 3 only fits the preferred model for the observed data to each iteration and therefore does not account for model uncertainty. Formally, this bootstrap is given in Algorithm 13.

Application: Suicide data

As with the other approaches, the model preferred the majority of times using Method 2 is the same model as is preferred for the observed data. Therefore, Methods 2 and 3 for accounting for model uncertainty with the parametric bootstrap algorithm will return the same results.

Algorithm 13 Bootstrap Approach 3, Method 3

Step 1: Fit the competing models to the observed data.
Step 2: For each model, compute the AIC or BIC weights, w_i, using Equation 6.1, where the AIC or BIC of each model, l, can be seen as the probability of model l being selected as the best candidate model (Wagenmakers and Farrell, 2004).
Step 3: Let $b = 1$.
Step 4: Sample l^* from $\{1, 2, \cdots, J\}$, where J is the number of models under consideration and the values have respective probabilities $\{w_1, w_2, \cdots, w_J\}$ of being sampled.
Step 5: Sample n counts from the model given by the linear predictor distribution pair, (j^*, D^*). Use these sampled counts x_i^* to create a sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1^*), \ldots, (\tau_n, \mathbf{v}_n, x_n^*)\}$, where τ_i and $\mathbf{v}_i$ are the observed person-years and covariates respectively for $i = 1, 2, \cdots, n$.
Step 6: Fit the model that is preferred for the observed data and estimate the rate or total population size using the methods discussed in Chapter 4.
Step 7: If $b = B$, stop. Otherwise, return to Step 4 with $b := b + 1$.

Application: Hares data

Similarly to the non-parametric and semi-parametric bootstrap algorithms, the preferred model for the Horvitz-Thompson estimator using Method 2 for accounting for model uncertainty is the same model that is preferred for the observed data, and therefore Methods 2 and 3 will return the same results. However, linear predictor 2 is preferred the majority of times using Method 2 for the generalised Chao's and generalised Zelterman's estimators. Therefore, Step 6 of Algorithm 13 can be modified as follows.

Step 6: let $\tilde{j}$ be the linear predictor that is preferred for the observed data, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_j$.

Generalised Chao's: Let

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1 - \hat{q}_i},$$

for $i = 1, 2, \cdots, f_1 + f_2$, then the estimated target population size is calculated as

$$\widehat{N}_b^{(GC)*} = n + \sum_{i=1}^{f_1+f_2} \frac{f_{i1} + f_{i2}}{\hat{\mu}_i + \hat{\mu}_i^2/2}.$$

Generalised Zelterman's: Let

$$\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{\tilde{j}}),$$

where $i = 1, 2, \cdots, n$, then the estimated target population size for snowshoe hares with the same covariates as hare i for $i = 1, 2, \cdots, n$ is calculated as

$$\widehat{N}_{bi}^{(GZ)*} = \frac{1}{1 - \exp(-\hat{\mu}_i)},$$

where $\widehat{N}_b^{(GZ)*} = \sum_{i=1}^{n} \widehat{N}_{bi}^{(GZ)*}$.

The 95% percentile confidence intervals from the parametric bootstrap algorithm using Method 3 for accounting for model uncertainty for the generalised Chao's and generalised Zelterman's estimators respectively are (2997, 3886) and (3203, 4203). These results are comparable to the intervals found using the alternative approaches, and are approximately centred at the estimated total number of hares from the observed data. However, whilst the intervals are comparable, model 5, the preferred model for the observed data, is very rarely preferred as seen in Table 6.26, so in this case, Method 3 may not be the most appropriate method to use.

6.5 Alternative methods of constructing confidence intervals

Section 6 uses the standard percentile method to construct the 95% confidence intervals used to quantify the level of uncertainty of the estimated rate of completed suicide and the estimated population sizes from the bootstrap data. However, the percentile method does not perform well in every case, particularly in cases with small sample sizes, as discussed in Hall and Martin (1989). This is also seen in Section 6, with the inflated upper limits of the estimated population size confidence intervals as a result of the biased and skewed bootstrapped data.

6.5.1 Bias-corrected and accelerated percentile method

The poor performance of the standard percentile method in certain situations is discussed by Efron (1981a; 1981b; 1982, Chapter 10; 1987) and Hall (1988), and they provide alternative approaches including the improved methods of bias-corrected (BC) percentile and bias-corrected and accelerated (BC_a) percentile methods. These improved methods both correct for bias through using the estimated proportion of bootstrap parameter estimates that are less than the original parameter estimate, also known as the bias-correction factor, $\hat{z}_0$. This bias-correction factor for the estimated population size is calculated as follows.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\sum_{b=1}^{B}(\widehat{N}_b^{*} < \widehat{N})}{B}\right),$$

where Φ is the standard normal cumulative distribution function.

For the bias-corrected percentile method, the bias-corrected significance level values for the adjusted confidence interval limits are then

$$\alpha_1^* = \Phi\left(2\hat{z}_0 + z_{(\frac{\alpha}{2})}\right)$$
$$\alpha_2^* = \Phi\left(2\hat{z}_0 + z_{(1-\frac{\alpha}{2})}\right),$$

where $z_{(\frac{\alpha}{2})}$ is the $100 \times \frac{\alpha}{2}$ percentile of a standard normal distribution.

The $(100-\alpha)\%$ bias-corrected (BC) percentile confidence interval is then

$$\left[\widehat{N}_{lower}, \widehat{N}_{upper}\right] = \left[\widehat{N}^*_{(\alpha_1^* \times B)}, \widehat{N}^*_{(\alpha_2^* \times B)}\right].$$

Using the jackknife resampling algorithm developed by Quenouille (1949), estimated population sizes $\widehat{N}_i^{(jack)}$, where $i = 1, 2, \cdots, n$, are calculated for the bias-corrected and accelerated percentile method. Whist the simulation study in Section 4.7 suggests that the generalised Chao's estimator is the superior capture-recapture estimator for the data types covered in this thesis, the jackknife resampling algorithm is formally provided below for the Horvitz-Thompson estimator in Algorithm 14, followed by the jackknife resampling algorithm for both the generalised Chao's and generalised Zelterman's estimators in Algorithm 15.

Algorithm 14 Jackknife resampling: Horvitz-Thompson estimator

Step 1: Set $i = 1$.
Step 2: Remove the ith row from the observed dataset to create the jackknife sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{i-1}, \mathbf{v}_{i-1}, x_{i-1}), (\tau_{i+1}, \mathbf{v}_{i+1}, x_{i+1}), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$.
Step 3: Fit the chosen regression model to the sampled dataset. Let $\hat{\beta}^{(jack)}$ be the maximum likelihood estimate of β.
Step 4: The estimated target population size is calculated as
Horvitz-Thompson:

$$\widehat{N}_i^{(jack)} = \sum_{k \in J} \frac{1}{1 - \exp\left[-\tau_k \exp\left(\hat{\beta}^{(jack)}\right)\right]},$$

where $J = 1, \cdots, i-1, i+1, \cdots, n$.
Step 5: If $i = n$, stop. Otherwise, return to Step 2 with $i := i + 1$.

These estimated population sizes allow for the computation of an acceleration constant, $\hat{a}$, used to correct for skewness as it is proportional to the skewness of the data (Efron, 1987), calculated as

$$\hat{a} = \frac{r_3}{6r_2^{3/2}},$$

where

$$r_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{N}_i^{(jack)} - \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{N}_i^{(jack)}\right)\right)^3,$$

Algorithm 15 Jackknife resampling: Generalised Chao's and generalised Zelterman's estimators

Step 1: Set $i = 1$.

Step 2: Remove the ith row from the observed dataset to create the jackknife sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1), \ldots, (\tau_{i-1}, \mathbf{v}_{i-1}, x_{i-1}), (\tau_{i+1}, \mathbf{v}_{i+1}, x_{i+1}), \ldots, (\tau_n, \mathbf{v}_n, x_n)\}$.

Step 3: Truncate the jackknife sampled dataset for all counts except $X = 1$ and $X = 2$. Fit the competing logistic regression models and select the model with the lowest AIC value. Let $\hat{q}$ be the corresponding fitted values and $\hat{\beta}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of β_j from the chosen regression model.

Step 4: Compute the estimated target population size as

Generalised Chao:

$$\widehat{N}_i^{(jack)} = n + \sum_{k=1}^{f_1+f_2} \frac{f_{k1} + f_{k2}}{\hat{\mu}_k + \hat{\mu}_k^2/2},$$

where

$$\hat{\mu}_k = \begin{cases} 2\dfrac{\hat{q}_k}{1-\hat{q}_k} & \text{if Poisson distribution assumed,} \\ \dfrac{\hat{q}_k}{1-\hat{q}_k} & \text{if geometric distribution assumed,} \end{cases}$$

for $k = 1, \cdots, i-1, i+1, \cdots, f_1 + f_2$.

Generalised Zelterman:

$$\widehat{N}_i^{(jack)} = \sum_{k=1}^{n} \frac{1}{1 - \exp(-\hat{\mu}_k)},$$

where

$$\hat{\mu}_k = \begin{cases} 2\tau_k \exp(\hat{\eta}_k) & \text{if Poisson distribution assumed,} \\ \tau_k \exp(\hat{\eta}_k) & \text{if geometric distribution assumed,} \end{cases}$$

for $k = 1, \cdots, i-1, i+1, \cdots, n$.

Step 5: If $i = n$, stop. Otherwise, return to Step 2 with $i := i + 1$.

and

$$r_2 = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{N}_i^{(jack)} - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{N}_i^{(jack)} \right) \right)^2.$$

The bias-corrected and accelerated significance level values for the adjusted confidence interval limits are then calculated as

$$\alpha_1^* = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(\frac{\alpha}{2})}}{1 - \hat{a}\left(\hat{z}_0 + z_{\frac{\alpha}{2}}\right)} \right)$$

$$\alpha_2^* = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\frac{\alpha}{2})}}{1 - \hat{a}\left(\hat{z}_0 + z_{(1-\frac{\alpha}{2})}\right)} \right),$$

leading to the $(100 - \alpha^*)\%$ bias-corrected and accelerated (BC_a) percentile confidence interval

$$\left[\widehat{N}_{lower}, \widehat{N}_{upper}\right] = \left[\widehat{N}^*_{(\alpha_1^* \times B)}, \widehat{N}^*_{(\alpha_2^* \times B)}\right].$$

Note that if $\hat{a} = 0$ then this confidence interval is equal to that of the bias-corrected percentile method.

Application: Suicide data

Table 6.27 provides the 95% confidence intervals for Horvitz-Thompson population size estimates for the three approaches to the bootstrap algorithm discussed in Section 6 using the standard percentile approach in addition to the BC and BC_a approaches. It is clear to see that for the non-parametric and semi-parametric approaches, correcting for the bias arising from the high correlation between the bootstrap estimated population sizes for the various sub-populations produces much more appropriate confidence intervals that are more comparable to the confidence interval from the parametric bootstrap which observes less bias. Additionally, bias correcting and accelerating improves intervals from each of the bootstrap approaches with the upper limits being less inflated and more appropriate for the given data. These results are depicted visually in Figure 6.1.

TABLE 6.27: Values for 95% confidence intervals for the Horvitz-Thompson population size estimates for the non-parametric, semi-parametric and parametric bootstrap algorithms accounting for uncertainty by comparing models each iteration, using the standard, bias-corrected and bias-corrected and accelerated percentile methods applied to the suicide case study data.

		Percentile interval method		
Approach	Method	Standard	BC	BC_a
A1: Non-parametric	M1: Full	(91, 7059)	(89, 230)	(76, 157)
	M2: Partial	(74, 441)	(77, 488)	(66, 372)
	M3: None	(74, 441)	(77, 488)	(66, 372)
A2: Semi-parametric	M1: Full	(91, 28913)	(88, 303)	(75, 157)
	M2: Partial	(91, 166)	(91, 166)	(80, 155)
	M3: None	(91, 166)	(91, 166)	(80, 155)
A3: Parametric	M1: Full	(101, 280)	(100, 260)	(88, 185)
	M2: Partial	(106, 185)	(109, 191)	(100, 173)
	M3: None	(106, 185)	(109, 191)	(100, 173)

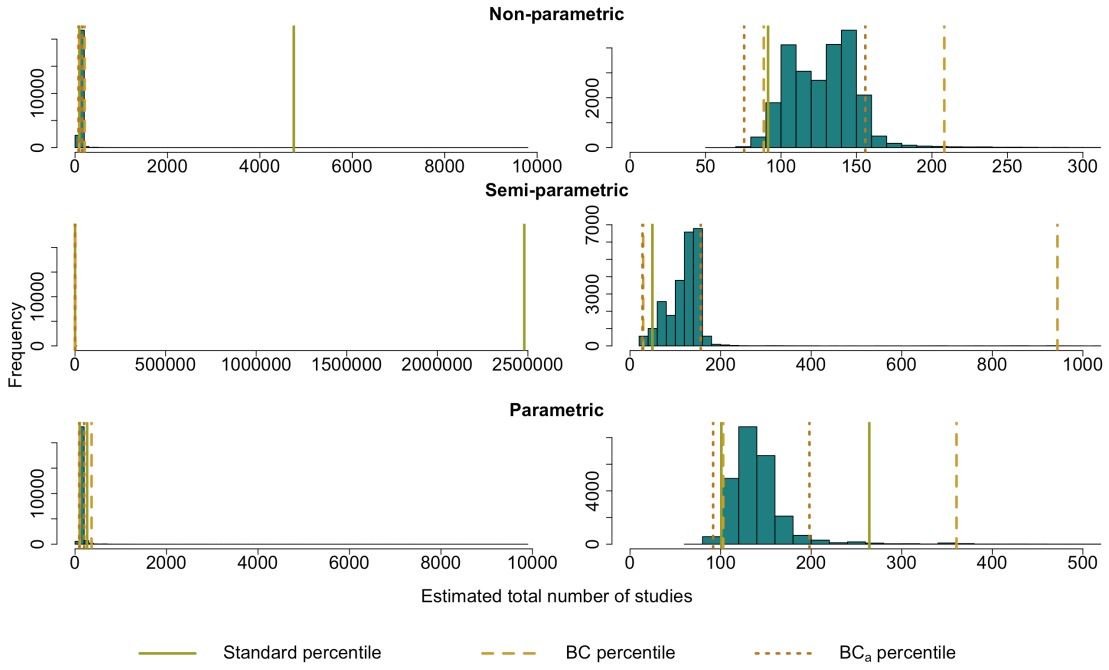

FIGURE 6.1: Histograms of the results from the three bootstrap approaches with method 1 discussed in Section 6 for the Horvitz-Thompson estimator applied to the suicide case study data, with standard percentile confidence intervals in addition to BC and BC_a percentile confidence intervals, each with 95% significance. Each approach has two plots, one to display the bootstrap data itself and another with smaller x-axis limits to better display the histogram given the large range of the data.

Application: Hares data

TABLE 6.28: Values for 95% confidence intervals for the Horvitz-Thompson population size estimates for the non-parametric, semi-parametric and parametric bootstrap algorithms accounting for uncertainty by comparing models each iteration, using the standard, bias-corrected and bias-corrected and accelerated percentile methods applied to the hares case study data.

		Percentile interval method		
Approach	Method	Standard	BC	BC_a
A1: Non-parametric	M1: Full	(2994, 3954)	(3580, 4653)	(3602, 4654)
	M2: Partial	(2978, 3797)	(3479, 4644)	(3503, 4654)
	M3: None	(2926, 3750)	(3243, 4261)	(3286, 4328)
A2: Semi-parametric	M1: Full	(2967, 3973)	(3584, 4918)	(3584, 4919)
	M2: Partial	(2937, 3892)	(3443, 4661)	(3474, 4701)
	M3: None	(2890, 3785)	(3221, 4213)	(3272, 4516)
A3: Parametric	M1: Full	(3015, 4027)	(3683, 4962)	(3699, 4934)
	M2: Partial	(2952, 3797)	(3358, 4311)	(3390, 4335)
	M3: None	(2997, 3886)	(3545, 4499)	(3565, 4499)

6.5.2 Median absolute deviation

Figure 6.1 shows that the bootstrapped data is heavy tailed and asymmetric, leading to percentile intervals that are biased and not centred around the median of the data. The BC and BC_a approaches to the percentile method correct for this bias and skewness to an extent, but an alternative approach of constructing confidence intervals which reduces bias should be explored. Median absolute deviation (MAD), popularised by Hampel (1974), utilises the median of the data to give less weight to outliers and heavy tails, has the highest breakdown point possible (50%) (Rousseeuw and Croux, 1993), and includes a consistency constant that enables the MAD to be an unbiased and robust estimator of the standard distribution.

The median absolute deviation is given by Huber (1981) as

$$\text{MAD}_B = C_B \times \text{median}\{|\widehat{N}_b^* - \text{median}\{\widehat{N}_b^*\}|\},$$

where $b = 1, \cdots, B$ and C_B is the consistency constant.

The 95% confidence interval for the estimated population size is then calculated as

$$\left[\widehat{N} - 1.96 \times \text{MAD}, \widehat{N} + 1.96 \times \text{MAD}\right].$$

Typically, the data is assumed to follow a normal distribution, disregarding abnormalities arising from outliers, leading to the consistency constant of $C_B = 1.4826$. If the underlying distribution is not the normal distribution, the consistency constant can be found as $C_B = [\Phi^{-1}(\frac{3}{4})]^{-1}$, where $\Phi^{-1}(\frac{3}{4})$ is the $\frac{3}{4}$th quantile of the underlying distribution Leys et al. (2013).

Application: Suicide data

Table 6.29 provides the 95% confidence intervals for the total population size using the Horvitz-Thompson estimator, constructed with the median absolute deviation for each combination of approach and method for the bootstrap algorithm, applied to the suicide case study data. For comparison purposes, the 95% (standard) percentile confidence intervals are also provided in the table.

It can be seen that for each approach, the bias seen in the intervals associated with Method 1 of accounting for model uncertainty is notably reduced, particularly for Approaches 1 and 2 where the upper limits for the standard percentile confidence intervals are very inflated. However, for Methods 2 and 3 (which for the suicide data produce identical results) there is less of a difference between the standard percentile method and the median absolute deviation.

TABLE 6.29: Values of the 95% confidence intervals constructed with the standard percentile method using the Horvitz-Thompson estimator and the median absolute deviation for each bootstrap approach and method combination applied to the suicide case study data.

Approach	Method	Standard Percentile	MAD
A1: Non-parametric	M1: Full	(91, 7059)	(81, 182)
	M2: Partial	(74, 441)	(15, 265)
	M3: None	(74, 441)	(15, 265)
A2: Semi-parametric	M1: Full	(91, 28913)	(81, 182)
	M2: Partial	(91, 166)	(99, 168)
	M3: None	(91, 166)	(99, 168)
A3: Parametric	M1: Full	(101, 280)	(92, 176)
	M2: Partial	(106, 185)	(97, 171)
	M3: None	(106, 185)	(97, 171)

Application: Hares data

Table 6.30 provides the 95% median absolute deviation confidence intervals for the total population size using the generalised Chao's estimator for each combination of approach and method for the bootstrap algorithm, applied to the hares case study data. For comparison purposes, the 95% (standard) percentile confidence intervals are also provided in the table.

TABLE 6.30: Values of the 95% confidence intervals constructed with the standard percentile method using the generalised Chao's estimator and the median absolute deviation for each bootstrap approach and method combination applied to the hares case study data.

Approach	Method	Standard Percentile	MAD
A1: Non-parametric	M1: Full	(2994, 3954)	(2931, 3887)
	M2: Partial	(2978, 3797)	(2928, 3817)
	M3: None	(2926, 3750)	(2884, 3701)
A2: Semi-parametric	M1: Full	(2967, 3973)	(2984, 3909)
	M2: Partial	(2937, 3892)	(2899, 3851)
	M3: None	(2890, 3785)	(2855, 3743)
A3: Parametric	M1: Full	(3015, 4027)	(2966, 3914)
	M2: Partial	(2952, 3797)	(2922, 3743)
	M3: None	(2997, 3886)	(2957, 3824)

Unlike with the suicide case study data, there is little difference in the corresponding intervals using the two confidence interval construction methods due to the reduced bias and therefore reduced extreme values present in the Hares data, which the MAD approach aids in correcting. The width of the respective intervals are approximately the same, with the upper and lower bounds of the MAD intervals being slightly reduced, due to any extreme values present being given less weight. Given the small different in the results and the lack of difference in computational burden of the two approaches,

to decide which method returns the best results and hence which approach should be used for confidence interval construction, a simulation study will be conducted in Section 6.6.3.

6.6 Simulation study

For each of the simulation studies conducted in this section, to ensure that the results are comparable, the same methods are used to simulate the data itself. These methods are also the same as those utilised in Section 3.3, starting by specifying the values of the following variables.

- N: total number of studies.
- $\bar{t}$: mean number of individuals per study.
- λ^C: constant rate of event.
- γ: logarithm of the mean for the observation period.
- σ: logarithm of the standard deviation for the observation period.
- α: shape parameter for the beta distribution to simulate proportions.
- β: shape parameter for the beta distribution to simulate proportions.
- ρ: success probability for the Bernoulli distribution to simulate binary variable.

Once the above variables are specified, the simulation study can be conducted, starting with simulating the size of each study, the observation period of each study, the count of events for each study, a proportion covariate for each study and a binary covariate for each study, each simulated from the same distributions as given in Section 3.3. The resulting data is then zero-truncated through removing the studies with a count of zero and the chosen bootstrap algorithm is conducted, treating the simulated zero-truncated dataset as the "original" dataset, recording the width of the resulting confidence intervals and whether they contain the true value of N. This process is repeated S times to conduct the simulation study and obtain the results of performance.

For the simulation studies below, the values of the variables are given as follows.

- $S = 1000$
- $N = 1000$
- $\bar{t} = 900$
- $\lambda^C = 0.0004$

- $\gamma = 1.5$
- $\sigma = 0.8$
- $\alpha = 36$
- $\beta = 8.5$
- $\rho = 0.4$

6.6.1 Bootstrap algorithm

Whilst the results from the bootstrap algorithms suggest that Approach 3, the parametric approach, is the preferred approach due to the reduction in bias as a result of the issue of multicollinearity, it is important to properly assess whether the performance of this approach is actually better than the alternatives. Similarly, it is important to explore the performance of the different methods for accounting for, or not accounting for, model uncertainty within the algorithms. To adequately assess this performance, simulation studies can be conducted in a similar way to as in Sections 3.3 and 4.7. Accuracy is not a criteria since the purpose of the bootstrap algorithms is to quantify the uncertainty through computing confidence intervals, rather than producing an estimate for the parameter of interest. Therefore, the only criteria for the performance of each bootstrap algorithm is the precision and coverage of the resulting confidence intervals.

For completeness, to properly assess the performance of each of the methods and approaches, a simulation study should be conducted for each combination, resulting in 9 simulation studies in total. Additionally, the generalised Chao's estimator is used for estimating the total population size within the bootstrap algorithms, given that the simulation study in Section 4.7 shows that it is the best capture-recapture population size estimator overall out of those tested.

TABLE 6.31: Simulation study results for the performance of each combination of approach and method with the bootstrap algorithm with the generalised Chao's estimator used for population size estimation, where $N = 1000$.

Approach	Method	Precision	Coverage
A1: Non-parametric	M1: Full	142.47	89.4%
	M2: Partial	136.40	89.2%
	M3: None	136.24	89.1%
A2: Semi-parametric	M1: Full	154.85	93.9%
	M2: Partial	153.40	94.0%
	M3: None	150.19	93.9%
A3: Parametric	**M1: Full**	**126.69**	**98.6%**
	M2: Partial	126.97	98.5%
	M3: None	126.77	98.5%

Table 6.31 provides the precision and coverage results from the simulation studies exploring the performance of each combination of approach and method for the bootstrap

algorithms. It can be seen that there is very little difference between the different methods of accounting for, or not accounting for, model uncertainty, for both the precision and the coverage. This suggests that there is very little benefit, if any, of the additional computational time required for Method 1 over the other methods. Whilst there is little difference in the coverage and precision between Method 2 and Method 3, as a result of the small difference in computational burden between these methods, Method 2 is the preferred method as it better accounts for the differences between datasets and how they are modelled.

The differences between the performance of each of the approaches is much larger. Whilst the coverage of the non-parametric approach is almost 90%, and with the coverage of the semi-parametric approach being above 90%, but below 95%, there is a clear preference for the parametric approach, where approximately 98.5% of the confidence intervals found with these bootstraps contain the true value. There is also a benefit with Approach 3 that the confidence intervals are slightly narrower and therefore more precise, potentially aiding in making more accurate and informative conclusions from the resulting intervals. These conclusions support the findings from the bootstrap algorithms themselves in this chapter, where there was a preference for the parametric approach to the bootstrap algorithm given that it produced samples with reduced bias from the lack of a multicollinearity issue, unlike with Approaches 1 and 2.

Taking these results into consideration, it is recommended to use the (fully) parametric bootstrap algorithm for quantifying uncertainty, and if there is a clear preference for a single model, utilising Method 2 (partial) for accounting for model uncertainty. If there is no clear preference, then Method 1 (full) should be used. Computationally, the non-parametric bootstrap algorithm is less intense than the other two approaches. However, at 10,000 bootstrap iterations, there was negligible difference between the different approaches to the bootstrap algorithm (when the same method of accounting for model uncertainty was utilised. Method 1 (full) of accounting for model uncertainty with 10 competing models, across all bootstrap approaches, took between 10 and 30 minutes on average, depending on the computer's processing speed and number of cores utilised (working in parallel takes the time down to 10 to 15 minutes on average). However, the final bootstrap algorithm when Method 2 (partial) of accounting for model uncertainty, or Method 1 (none) is used takes less time with the algorithm running in 5 to 15 minutes on average (on the lower end when parallelisation is used in the code). Taking these timings into consideration, the combination of Method 2 (partial) of accounting for model uncertainty and the (fully) parametric bootstrap algorithm (or Method 1 if there is no clear preference for a single model) is recommended for use.

For demonstrative purposes, the same simulations are repeated, however, using the Horvitz-Thompson estimator for estimating the population size within each bootstrap algorithm instead. The results for these simulations are given in Table 6.32, where it can be seen that overall, the performance of the bootstrap algorithms is very poor. There is

TABLE 6.32: Simulation study results for the performance of each combination of approach and method with the bootstrap algorithm with the Horvitz-Thompson estimator used for population size estimation.

Approach	Method	Precision	Coverage
A1: Non-parametric	M1: Full	61.66	77.4%
	M2: Partial	62.10	77.7%
	M3: None	61.66	77.8%
A2: Semi-parametric	M1: Full	90.96	94.0%
	M2: Partial	**91.10**	**94.4%**
	M3: None	90.95	94.0%
A3: Parametric	M1: Full	36.12	54.8%
	M2: Partial	36.16	55.1%
	M3: None	36.08	54.6%

still very little difference in performance between each of the methods, and the coverage of the intervals found using Approach 2 are comparable to when the generalised Chao's estimator is used, however, the coverage for the other approaches is no longer ideal. The precision for both Approach 1 and Approach 2 is smaller, particularly for the parametric approach. The narrower intervals likely contribute to the reduced coverage values, wherein for the parametric approach, only just over half of all intervals contain the true value, a result which is far less desirable than the corresponding 98% of intervals found when the generalised Chao's is used. Taking the performance values into consideration, along with the results from the simulation study in Section 4.7, there is considerable evidence to suggest that the generalised Chao's estimator should be used in estimating the total (or missing) population size instead of the more commonly used Horvitz-Thompson estimator.

6.6.2 Bias-corrected and bias-corrected and accelerated

As with the standard bootstrap approaches, it is important to explore the performance of the bias-corrected and the bias-corrected and accelerated percentile confidence intervals. To do this, similar simulation studies to those conducted in Sections 6.6.1 and 6.6.3, are conducted but using the BC and BC_a approaches to confidence interval construction.

Table 6.33 provides the results from the simulation study exploring the performance of the bias-corrected percentile confidence intervals when the generalised Chao's estimator is used for estimating the total population size. It can be seen throughout that the results are much less consistent within each approach, with greater differences between methods, particularly with Approach 3. Additionally, each of the coverage values are lower compared with those of the standard percentile confidence intervals (in the standard bootstrap algorithms), indicating that the bias-corrected percentile method is inferior to the standard approach, especially with the additional computational burden required to compute the BC intervals.

TABLE 6.33: Simulation study results for the performance of the bias-corrected percentile confidence intervals for each combination of approach and method with the bootstrap algorithm with the generalised Chao's estimator used for population size estimation.

Approach	Method	Precision	Coverage
A1: Non-parametric	M1: Full	141.97	87.8%
	M2: Partial	135.14	88.2%
	M3: None	134.27	88.0%
A2: Semi-parametric	M1: Full	155.09	91.7%
	M2: Partial	135.15	88.2%
	M3: None	**148.89**	**92.3%**
A3: Parametric	M1: Full	141.97	87.8%
	M2: Partial	103.67	65.9%
	M3: None	103.67	65.7%

Despite the bias-corrected percentile confidence intervals performing poorly in comparison to the other approaches, for completeness, the bias-corrected and accelerated interval approach will also be used in a simulation study in order to assess its performance.

TABLE 6.34: Simulation study results for the performance of the bias-corrected and accelerated percentile confidence intervals for each combination of approach and method with the bootstrap algorithm with the generalised Chao's estimator used for population size estimation.

Approach	Method	Precision	Coverage
A1: Non-parametric	M1: Full	140.52	86.6%
	M2: Partial	135.15	88.2%
	M3: None	132.30	86.8%
A2: Semi-parametric	M1: Full	153.49	89.9%
	M2: Partial	135.15	88.2%
	M3: None	**147.06**	**90.6**%
A3: Parametric	M1: Full	140.52	86.6%
	M2: Partial	103.47	66.6%
	M3: None	103.47	66.4%

6.6.3 Median absolute deviation

Similarly to as with the bias-corrected and bias-corrected and accelerated intervals, it is important that the performance of the median absolute deviation confidence intervals is explored. To do this, the same methods are utilised as for the simulation study in Section 6.6.1, however, instead of constructing the standard percentile confidence intervals after each bootstrap algorithm, the median absolute deviation is used for constructing the confidence intervals.

Given that the generalised Chao's is the favoured capture-recapture estimator out of those considered, it will be used in the simulation study for testing the performance of the median absolute deviation.

TABLE 6.35: Simulation study results for the performance of the Median Absolute Deviation for each combination of approach and method with the bootstrap algorithm with the generalised Chao's estimator used for population size estimation.

Approach	Method	Precision	Coverage
A1: Non-parametric	M1: Full	142.06	91.3%
	M2: Partial	137.64	90.8%
	M3: None	136.99	90.6%
A2: Semi-parametric	M1: Full	159.12	94.5%
	M2: Partial	157.14	94.6%
	M3: None	153.47	94.8%
A3: Parametric	**M1: Full**	**129.04**	**98.8%**
	M2: Partial	129.34	98.7%
	M3: None	127.67	98.6%

Table 6.35 provides the findings from the simulation study for exploring the performance of the median absolute deviation as a method for constructing confidence intervals in the different bootstrap algorithms when the generalised Chao's estimator is used for estimating the total population size. The results are very similar to when the standard percentile interval is used for confidence interval construction for the simulations of the bootstrap algorithms in Section 6.6.1. However, across each of the bootstrap algorithms, the coverage is higher than when the standard percentile approach is used. The coverage increases by approximately 1.5 for the non-parametric bootstrap algorithm, an increase of approximately 1 for the semi-parametric bootstrap algorithm and an increase of approximately or the semi-parametric the increase is approximately 1 and for the for the parametric bootstrap algorithm. The coverage is considerably higher for each of the approaches compared to when either the bias-corrected or bias-corrected and accelerated percentile confidence intervals are used.

Similarly, the difference between the precision values for when the standard percentile method is used compared to the median absolute deviation is very small, with a slight increase in width of confidence intervals for when the median absolute deviation is used.

Given that using the median absolute deviation does not add to the computational burden of performing the bootstrap algorithms comparatively to using the standard percentile method, the results from the simulation studies indicate a preference for the MAD. Additionally, despite the difference in coverage being small, since the complexity of the methods are comparable, it is logical to use the MAD when the bootstrap data is normally distributed and be able to obtain slightly better intervals when possible. If the bootstrap data is not normally distributed, an alternative consistency constant should be used.

Chapter 7

Methods under presence of one-inflation

This chapter discusses the methods covered and developed in previous chapters with an application to zero-truncated one-inflated count data. In many situations, there is a presence of excess singletons in count data, often to individuals getting trap-shy with the consequences act as a deterrent. For example, in arrest data, there is often excess singletons with a high number of individuals only getting arrested or going to jail once, due to many people not wanting to experience it again. Another cause for one-inflation is that if a 'treatment' proves successful, for example, in the heroin case study data used in this chapter, there may be an increase in the number of singletons due to people not needing to return, with the treatment for drug addiction being successful. Nonetheless, no matter what the reason for excess singletons, it is important to use estimation methods which appropriately account for one-inflation to avoid having inflated population size estimates which in turn, may affect conclusions and decisions made based on those estimates. To do so, this chapter develops the generalised-modified Chao's estimator to appropriately account for the excess singletons, with application to a case study to demonstrate the impact one-inflation can have on population size estimation. Finally, a simulation study is developed to explore the performance of the generalised-modified Chao's estimator to ensure that it performs to an appropriate level, and to demonstrate why the existing methods are not appropriate to use in this situation.

7.1 Background

Heroin, also known as *diacetylmorphine* or *diamorphine*, is a semi-synthetic, highly addictive, illegal drug used for the euphoric effect it has when taken (National Institute on Drug Abuse website, 2021). The recreational drug is synthesised from morphine, an opiate used as an analgesic (painkiller), extracted from opium poppies. Use of pure

heroin typically takes the form of smoking or snorting, but impure heroin that is created from crude processing methods is typically administered intravenously (injected into the veins), intramuscular (injected into the muscles) or subcutaneously (injected under the skin) after being dissolved and diluted into a liquid (National Institute on Drug Abuse website, 2021).

As with many illegal drugs, heroin is ”cut” with various additives for different uses. Some additives such as starch and powdered milk are added to the heroin for the purpose of increasing the weight of the heroin before selling, cutting costs by reducing the percentage of actual heroin in the drug. However, other additives such as fentanyl, another illicit drug, are added to the heroin to make the drug more potent. This contamination of the heroin can make it much more dangerous, particularly if the user does not know what the heroin is contaminated with given that fentanyl is much stronger than heroin and takes only a small amount of the drug to cause an overdose.

Many factors can impact the effect heroin has on an individual, including the dosage taken, age of the user, frequency of use, duration of use, method of drug use, environment, pre-existing medical conditions (physical and mental) and whether any other drugs were also consumed (for example, alcohol, herbal, over-the-counter or recreational) (The Centre for Addiction and Mental Heatlh). The short-term effects of the drug include the ”rush” of euphoria that many users enjoy, along with nausea, dry mouth, drowsiness, clouded mental function, slowed heart rate and slowed breathing. Sometimes, the slowed breathing can be so severe that it is life-threatening, potentially causing brain damage or resulting in a coma (National Institute on Drug Abuse website, 2021). Other more serious risks include the risk of overdose, made worse when the user does not know the dosage they are taking, with varying levels of impurities making it hard to know what is a ’safe’ dosage along with unknown contaminants increasing the potency of the drug. Additionally, there is the risk of infection when the drug is injected, particularly since many drug users do not reliably have access to clean and new needles for each injection. This has historically been a very large issue as drug users who inject themselves have increased risks of HIV, tuberculosis (TB), and viral hepatitis B and C (HBV and HCV) (World Health Organisation).

Thailand has had a long and complex history with its population and drug abuse. For several decades, it was one of the major sources of illicit opium production (Windle, 2015), making heroin a popular drug of choice for a long time. However, in the late 1990s and early 2000s, the production of illicit opium was suppressed in the country through intervention with the opium farmers. As a result of this and other factors, whilst it is still a widespread problem with many individuals still abusing the drug, heroin is no longer the most popular drug in Thailand, ’yaa baa’ (methamphetamine) is instead. In 2003, the then prime minister started a ’war on drugs’ that lasted for three months (Vongchak et al., 2005)with an aim to shift focus from imprisonment of drug abusers to rehabilitation, and to stop the problem of drug trafficking in the country, but

instead thousands were killed, with over half of the 2800 killed having no connection to drugs. In addition to the killings, thousands more were forced into 'treatment' for their addiction, wherein if they did not go voluntarily, they were forced into detention centres run by the military (Windle, 2015).

Panyalert and Lanamteng (2020) provide information on the number of heroin users by both age and gender that contacted the Thanyarak Chiang Mai hospital, Thailand, for treatment for their drug addiction between 2013 and 2018. In this data, the number of heroin users in the area that did not seek treatment help at least once is unknown, leading to zero-truncated data. In this instance, it is important to know the total (estimated) number of heroin users in the province for proper and effective resource allocation among other reasons.

TABLE 7.1: Distribution of counts of heroin users in Chiang Mai, Thailand by age.

Age	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
< 40	-	309	100	53	24	11	7	5	7	0	1	1	0	0	1
≥ 40	-	228	52	27	10	4	1	1	1	0	0	0	0	0	0
Total	-	537	152	80	34	15	8	6	8	0	1	1	0	0	1

TABLE 7.2: Distribution of counts of heroin users in Chiang Mai, Thailand by gender.

Gender	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
Male	-	482	134	73	30	13	7	5	7	0	1	1	0	0	1
Female	-	55	18	7	4	2	1	1	1	0	0	0	0	0	0
Total	-	537	152	80	34	15	8	6	8	0	1	1	0	0	1

7.1.1 Modelling

For the heroin case study data, the expected response, $\exp(\eta_i)$, is the rate of contact to the treatment centre. Whilst there is information available on two covariates, age and gender, the information is not at the individual level. Therefore, both the age and gender of each individual is not known, only the overall distribution of people by age and the overall distribution of people by gender. As a result of this, the covariates cannot be modelled together. Instead, two datasets will be explored where the overall counts for each dataset is the same. The first dataset explores the counts of contact with treatment centres with age as a covariate, with the second exploring the counts of contact with treatment centres with gender as a covariate. In the age dataset, v_{i1} is the covariate for the binary variable indicating the age range of the individual where for $i = 1, 2, \cdots, 843$,

$$v_{i1} = \begin{cases} 0 & \text{if below 40 years old, and} \\ 1 & \text{if greater than or equal to 40 years old.} \end{cases}$$

For the other dataset, the gender dataset, v_{i2} is the covariate for the binary variable indicating the gender of the individual where

$$v_{i2} = \begin{cases} 0 & \text{if male, and} \\ 1 & \text{if female,} \end{cases}$$

for $i = 1, 2, \cdots, 843$.

TABLE 7.3: Linear predictors under consideration with corresponding regression functions.

Linear predictor j	Age v_1	Gender v_2	Regression function $\mathbf{h}_j(\mathbf{v})$
1	No	No	$\mathbf{h}_1(\mathbf{v}) = 1$
2	Yes	No	$\mathbf{h}_2(\mathbf{v}) = (1, v_1)^T$
3	No	Yes	$\mathbf{h}_3(\mathbf{v}) = (1, v_2)^T$

As with the hares case study data, there is no exposure variable, τ_i, for the heroin case study data. Since there is only one covariate available for each dataset, only two models are considered for each. For both datasets, the intercept-only model is identical, overall there are only three unique models, the intercept-only model, a main effect model for age and a main effect model for gender. The linear predictors for these different models are given in Table 7.3. Given that the heroin case study data is count data, the Poisson, negative-binomial and geometric models are all under consideration. Table 7.4 provides the values of the maximised log-likelihoods, number of parameters, AIC and BIC values for each of the linear predictor and distribution combinations under consideration from the zero-truncated regression modelling.

TABLE 7.4: Values of the maximised log-likelihood, number of parameters, AIC and BIC for the models under consideration.

Distribution	Linear Predictor	Maximised log-likelihood	Number of parameters	AIC	BIC
Poisson	1	-1113.67	1	2229.34	2234.07
	2	-1096.91	2	2197.21	2206.69
	3	-1113.62	2	2231.24	2240.71
Negative-binomial	1	-993.54	2	1990.90	2000.38
	2	**-985.18**	**3**	**1976.36**	**1990.57**
	3	-993.43	3	1992.86	2007.07
Geometric	1	-1012.32	1	2026.64	2031.37
	2	**-1000.82**	**2**	**2005.65**	**2015.12**
	3	-1012.29	2	2028.57	2038.04

The results from Table 7.4 indicate that the addition of the age covariate improves the fit of the model given that for each of the three distributions under consideration, the addition of age reduces the values of the negative log-likelihood, AIC and BIC statistics. The results also indicate that there is a preference for the negative-binomial distribution.

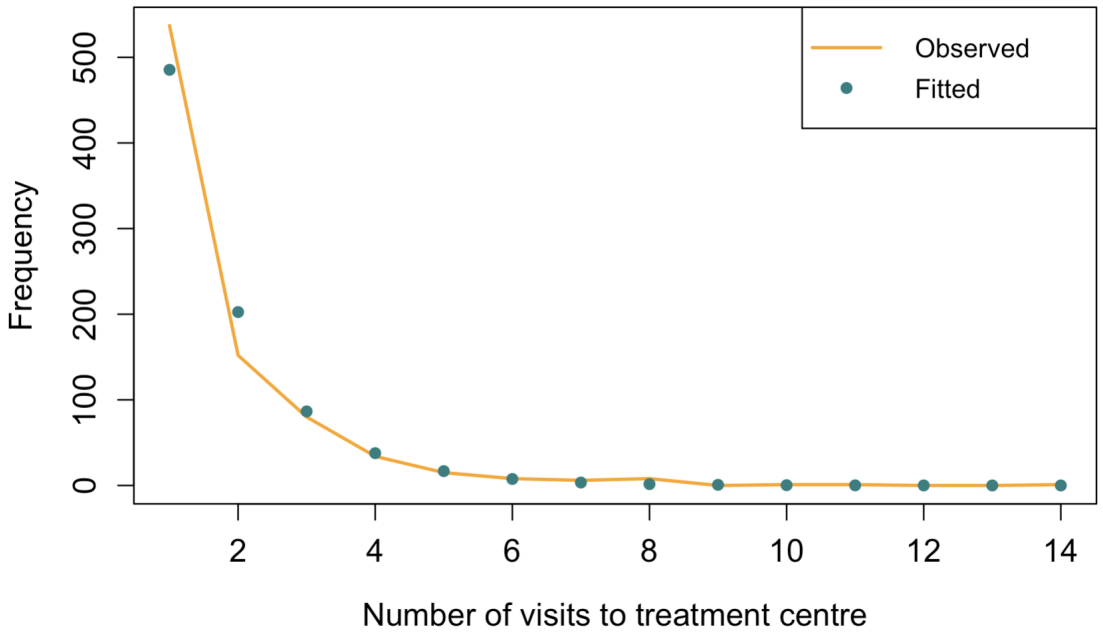

FIGURE 7.1: Observed and fitted frequencies for the heroin dataset with the geometric model with age as a covariate assumed.

7.1.2 Fitted frequencies

Table 7.5 provides the values of the observed and fitted frequencies of counts of heroin users contacting the treatment centre in Chiang Mai, Thailand, assuming the geometric model with age as a covariate.

TABLE 7.5: Frequency distribution of the captures of heroin users

	Count of heroin users, x									
Frequency type	0	1	2	3	4	5	6	7	8	9 +
Observed, f_x	-	537	152	80	34	15	8	6	8	3
Fitted, $\hat{f}_x$	-	485	203	87	38	17	8	3	2	3

7.1.3 Estimation

Whilst the negative-binomial main effects for age model is the best fitting for the data out of the models under consideration, when using this model in the Horvitz-Thompson estimator, the resulting total population size estimate of $\widehat{N}^{(HT)} = 4,048,078$ is very large and spurious. In fact, this estimated total number of heroin users in Chiang Mai is approximately 3 times greater than the estimated population size of the city itself (1.3 million) (NASA Earth Observatory, 2024). This poor result is in part due to the commonly seen boundary problem demonstrated in Böhning (2015). Therefore, despite having the lowest information criterion values, the negative-binomial distribution is not a suitable choice for this dataset.

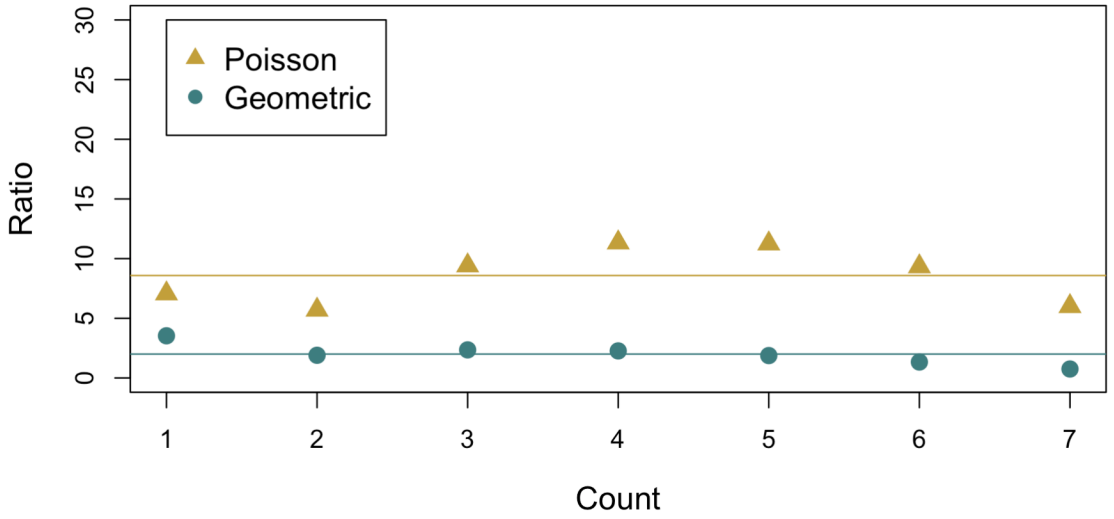

FIGURE 7.2: Ratio plot comparing the validity of the Poisson and geometric mixture kernel assumptions

Instead, the geometric distribution will be used, given that the fit of the geometric model with a main effect for age is a better fit than the Poisson, with it still being a good fit for the data as seen in the ratio plot in Figure 7.2.

7.1.3.1 Horvitz-Thompson estimator

For the heroin case study data when age is included as a covariate and the zero-truncated geometric distribution is used, the total population size using the Horvitz-Thompson estimator is given as $\widehat{N}^{(HT)} = 2036$. Given the observed population size, the total number of heroin users who didn't seek treatment in the Chiang Mai province is found to be $\widehat{M} = 1193$.

7.1.3.2 Chao's lower bound estimator

Using the heroin case study data, the total number of heroin users can then be estimated using the conventional Chao's estimator assuming a geometric mixture kernel as

$$\begin{aligned}\widehat{N}^{(C)} &= n + \frac{f_1^2}{f_2} \\ &= 843 + \frac{537^2}{152} \\ &= 2740,\end{aligned}$$

which is much larger than the total estimate of $\widehat{N}^{(HT)} = 2036$ provided by the Horvitz-Thompson estimator.

It is important to note that this estimate may be higher than expected given that there is a possibility for one-inflation, and since Chao's estimator relies completely on the singletons and doubletons in the data, having excess singletons in the data would lead to an inflated population size estimate. This will also be true for the other estimators (including the Horvitz-Thompson estimator) that do not appropriately account for excess singletons, however, given that the Horvitz-Thompson relies on other counts besides just the singletons and doubletons like Chao's and Zelterman's estimators do (and their generalised counterparts), it is likely that the Horvitz-Thompson estimator will be slightly more robust to one-inflation.

7.1.3.3 Generalised Chao's estimator

Table 7.6 provides the BIC values and the generalised Chao's estimates for each of the three models under consideration, assuming a geometric mixture kernel.

TABLE 7.6: Estimated total number of heroin users in Chiang Mai, Thailand for each linear predictor under consideration using the generalised Chao's estimator, assuming a geometric mixture kernel, with corresponding AIC and BIC values for each linear predictor.

Linear predictor	AIC	BIC	$\widehat{N}^{(GC)}$
1	729	734	3402
2	**728**	737	**3457**
3	731	740	3406

It can be seen that the preferred binomial logistic regression model is that of the model with age as a covariate (using the AIC as the chosen information criterion), with a corresponding total population size estimate of $\widehat{N}^{(GC)} = 3457$, an estimate that is considerably larger than the existing estimates using the Horvitz-Thompson and conventional Chao's estimators.

7.1.3.4 Conventional Zelterman's estimator

Assuming a geometric distribution, the conventional Zelterman's estimator for the heroin case study data is given as follows.

$$\begin{aligned}\widehat{N}^{(Z)} &= \frac{nf_1}{f_2} \\ &= \frac{843 \times 537}{152} \\ &= 2978,\end{aligned}$$

which is much larger than the Horvitz-Thompson and conventional Chao's estimator but approximately equal to the generalised Chao's estimate.

7.1.3.5 Generalised Zelterman's estimator

The same binomial logistic regression models are fitted to the heroin case study data to estimate the total population size using the generalised Zelterman's estimator as with the generalised Chao's estimator. Therefore, the same intercept-only binomial logistic regression model is preferred, leading to an estimate of the total number of heroin users of $\widehat{N}^{(GZ)} = 3420$. This estimate is larger than all of the existing estimates, however, it is possible that the large difference is due to one-inflation, given that the generalised Zelterman's estimate relies heavily on the singletons in the dataset, and assumes the counts of 1 and 2 follow the given distributional assumption. Whilst a robust estimator, if there are excess singletons, the generalised Zelterman's estimator is no longer robust given that the important assumption is not met. To test whether this is the case, a likelihood ratio test can be used to explore whether there is in fact excess singletons in this dataset that are impacting the capture-recapture estimates.

7.2 Likelihood ratio testing

The existence of one-inflation can be explored by testing if the number of singletons is compatible with the baseline model chosen. This is done via a likelihood ratio test, testing the null hypothesis model of no one-inflation against the alternative hypothesis model of one-inflation. The likelihood ratio statistic (Böhning and van der Heijden, 2019) is given as

$$\begin{aligned}\lambda &= 2[\log L_{+1} - \log L_{+}] \\ &= 2[f_1 \log(f_1/n) + n_1 \log(1 - f_1/n) + \log L_{++} - \log L_{+}].\end{aligned}$$

Application: Heroin data

Assuming a geometric baseline distribution, a likelihood ratio test can be conducted to identify whether there is excess singletons in the heroin dataset. First, the zero-truncated density is required for the null hypothesis model, given as

$$p_x(\mu)^{+} = \left(1 - \frac{1}{1+\mu_i}\right)^{x-1} \frac{1}{1+\mu_i},$$

for $x = 1, 2, \cdots$, with the corresponding log-likelihood

$$\log L_{+} = S_1 \log\left(1 - \frac{1}{1+\mu}\right) + n \log\left(\frac{1}{1+\mu}\right),$$

with the maximum likelihood estimate $\hat{\mu}_0 = \frac{1}{S_1/n+1}$ where $S_1 = \sum_{x=1}^{m} f_x(x-1)$.

The zero-one-truncated density is then required for the alternative hypothesis model, given as

$$p_x(\mu)^{++} = \frac{\left(1-\frac{1}{1+\mu}\right)^x \frac{1}{1+\mu}}{1-\frac{1}{1+\mu}-\left(1-\frac{1}{1+\mu}\right)\frac{1}{1+\mu}},$$

for $x = 2, 3, \cdots$, with the corresponding zero-one-truncated likelihood given as

$$\log L_{++} = S_2 \log\left(1-\frac{1}{1+\mu}\right) + n_1 \log\left(\frac{1}{1+\mu}\right),$$

with the maximum likelihood estimate $\hat{\mu} = \frac{1}{S_2/n_1+1}$ where $S_2 = \sum_{x=2}^{m} f_x(x-2)$.

The likelihood ratio test statistic can then be given as follows.

$$\begin{aligned}\lambda &= 2[537\log(537/843) + 306\log(1-537/843)\\ &+ S_2\log\left(1-\frac{1}{1+\mu}\right) + 306\log\left(\frac{1}{1+\mu}\right)\\ &- S_1\log\left(1-\frac{1}{1+\mu}\right) + 843\log\left(\frac{1}{1+\mu}\right)]\\ &= 36.71,\end{aligned}$$

with a corresponding χ^2_1 p-value of $1.37e-9 \approx 0$, providing strong evidence that the heroin case study data is one-inflated. Therefore, under the assumption that the data is one-inflated, an alternative approach to the existing capture-recapture estimators is required in order to appropriately account for the excess singletons in the estimation of the total number of heroin users in Chiang Mai, Thailand.

7.3 Modified Chao's estimator

As previously mention, an excess of often singletons occur in count data, leading to one-inflation. An example of this occurring can be seen in ecology when individuals become trap-shy and avoid being trapped again after the first occasion. Another example of cause for one-inflation is where there is a small probability of recurrence, such as readmission to hospital for for cancer in epidemiological data or rearrest in criminology data.

Böhning et al. (2019) developed a modified version of the standard Chao's lower bound estimator to account for this inflation in count data, using methods of substitution to eliminate the need for singletons in the calculations. Instead, doubletons and tripletons are utilised for estimating the total population size. However, the estimated frequency of zero counts given in Equation 4.13 relies on the singletons following the given distribution. To circumvent this, additional ratios can be constructed to estimate the

frequency of singletons for substitution as follows.

$$\begin{aligned}\frac{a_1}{a_2}\frac{f_2}{f_1} &\leq \frac{a_2}{a_3}\frac{f_3}{f_2}\\ \Rightarrow \hat{f}_1 &= \frac{a_1 a_3}{a_2^2}\frac{f_2^2}{f_3}.\end{aligned}$$

Substituting $\hat{f}_1$ into Equation 4.13 gives

$$\begin{aligned}\hat{f}_0^{(MC)} &= \frac{a_0 a_2}{a_1^2}\frac{1}{f_2}\left(\frac{a_1 a_3}{a_2^2}\frac{f_2^2}{f_3}\right)^2\\ &= \frac{a_0 a_2}{a_1^2}\frac{1}{f_2}\left(\frac{a_1^2 a_3^2}{a_2^4}\frac{f_2^4}{f_3^2}\right)\\ &= \frac{a_0 a_3^2}{a_2^3}\frac{f_2^3}{f_3^2}\end{aligned}$$

If a Poisson mixture kernel is assumed, $a_0 = 1/1! = 1$, $a_2 = 1/2! = 1/2$ and $a_3 = 1/3! = 1/6$, and the estimated frequency of zero counts is

$$\begin{aligned}\hat{f}_0^{(MC)} &= \frac{a_0 a_3^2}{a_2^3}\frac{f_2^3}{f_3^2}\\ &= \frac{1 \times 1/36}{1/8}\frac{f_2^3}{f_3^2}\\ &= \frac{2}{9}\frac{f_2^3}{f_3^2},\end{aligned}$$

and modified Chao's estimator is then computed as

$$\widehat{N}^{(MC)} = n + \frac{2}{9}\frac{f_2^3}{f_3^2}.$$

Alternatively, if a geometric mixture kernel is assumed, $a_x = 1$ and the frequency of zero counts is then estimated as

$$\begin{aligned}\hat{f}_0^{(MC)} &= \frac{1 \times 1^2}{1^3}\frac{f_2^3}{f_3^2}\\ &= \frac{f_2^3}{f_3^2},\end{aligned}$$

leading to the estimated total population size

$$\widehat{N}^{(MC)} = n + \frac{f_2^3}{f_3^2}.$$

Application: Heroin data

Assuming a geometric mixture kernel, the total population size using the modified Chao's estimator for the heroin case study data can then be estimated as follows.

$$\begin{aligned}\widehat{N}^{(MC)} &= n + \frac{f_2^3}{f_3^2} \\ &= 843 + \frac{152^3}{80^2} \\ &= 1392.\end{aligned}$$

This estimate is considerably smaller than the estimates computed from the capture-recapture estimators when the excess singletons are not accounted for. This is to be expected given that the one-inflation present in the data leads to higher than expected estimates for the total population size.

7.4 Generalised-modified Chao's estimator

Given that only the frequency of counts is utilised in the computation for the total population size for the modified Chao's estimator, any available covariates, including exposure variables, are not accounted for. Using the work of Böhning et al. (2013b), the modified Chao's estimator can be generalised, first through truncating all counts besides $X_i = 2$ and $X_i = 3$, which leads to the following associated truncated Poisson model.

$$p_2(\mu_i) = \frac{\exp(-\mu_i)\mu_i^2}{2} = (1 - q_i) \text{ and } p_3(\mu_i) = \frac{\exp(-\mu_i)\mu_i^3}{6} = q_i. \tag{7.1}$$

Given that $q_2 = (1 - q)$ and $q_3 = q$, Equation 4.10 becomes

$$\mu = \frac{a_2}{a_3}\frac{q_3}{q_2} = 3\frac{q}{1-q}. \tag{7.2}$$

Replacing the value of $p_3(\mu) = q$ with its sample estimate $\frac{f_3}{N}$ where $N = (f_2 + f_3)$, makes μ equivalent to

$$\mu = 3\frac{f_3/(f_2 + f_3)}{1 - f_3/(f_2 + f_3)} = 3\frac{f_3}{f_2},$$

verifying that the ratios in Equations 4.10 and 4.12 are equal.

Rearranging Equation 7.2 for $\hat{q}$ results in

$$\hat{q} = \frac{\mu}{3 + \mu},$$

making the probabilities in Equation 7.1 equivalent to

$$p_2(\mu_i) = \frac{3}{3+\mu_i} = \text{ and } p_3(\mu_i) = \frac{\mu_i}{3+\mu_i}.$$

Therefore, the associated truncated Poisson likelihood is

$$L = \prod_{i=1}^{f_2+f_3} \left(\frac{3}{3+\mu_i}\right)^{f_{i2}} \left(\frac{\mu_i}{3+\mu_i}\right)^{f_{i3}},$$

equal to the standard binomial logistic likelihood

$$L = \prod_{i=1}^{f_2+f_3} (1-q_i)^{f_{i2}} (q_i)^{f_{i3}},$$

resulting in the standard binomial logistic log-likelihood

$$\ell = f_2 \log(1-q) + f_3 \log(q).$$

The maximum likelihood of q can be found as

$$\begin{aligned}
\frac{d\ell}{dq} &= \frac{f_2}{1-q} - \frac{f_3}{q} \\
\Rightarrow 0 &= \frac{f_2}{1-\hat{q}} - \frac{f_3}{\hat{q}} \\
\Rightarrow 0 &= f_2\hat{q} - f_3(1-\hat{q}) \\
\Rightarrow 0 &= f_2\hat{q} - f_3 + f_3\hat{q} \\
\Rightarrow \hat{q} &= \frac{f_3}{f_2+f_3}.
\end{aligned}$$

Following Böhning et al. (2013b), the estimate for the frequency of zero counts can be computed using the conditional expectation as

$$\begin{aligned}
\hat{f}_0^{(GMC)} = E[f_0|f_2,f_3,q] &= \frac{p_0(\hat{\mu})}{p_2(\hat{\mu})+p_3(\hat{\mu})}(f_2+f_3) \\
&= \frac{a_0}{a_2\hat{\mu}^2+a_3\hat{\mu}^3}(f_2+f_3) \\
&= \frac{a_0}{a_2\left(3\frac{f_3}{f_2}\right)^2+a_3\left(3\frac{f_3}{f_2}\right)^3}(f_2+f_3) \\
&= \frac{a_0}{a_2\left(9\frac{f_3^2}{f_2^2}\right)+a_3\left(27\frac{f_3^3}{f_2^3}\right)}(f_2+f_3)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{f_2 + f_3}{\frac{9}{2}\frac{f_3^2}{f_2^2} + \frac{27}{6}\frac{f_3^3}{f_2^3}} \\
&= \frac{2f_2^3(f_2 + f_3)}{9f_3^2\left((f_2 + f_3)\right)} \\
&= \frac{2f_2^3}{9f_3^2} = \hat{f}_0^{(MC)}.
\end{aligned}
$$

however, this does not allow for covariates to be accounted for, hence it is equal to the estimated frequency of zeroes for the modified Chao. The frequency of zero counts can instead be computed through using the fitted values from the binomial logistic regression model to estimate the value of μ with Equation 7.2, leading to the following estimate for the frequency of zero counts.

$$
\begin{aligned}
\hat{f}_{i0} = E[f_{i0}|f_{i1}, f_{13}, \mu_i] &= \frac{p_0(\hat{\mu}_i)}{p_2(\hat{\mu}_i) + p_3(\hat{\mu}_i)}(f_{i2} + f_{i3}) \\
&= \frac{\exp(-\hat{\mu}_i)}{\exp(-\hat{\mu}_i)\hat{\mu}_i^2/2! + \exp(-\hat{\mu}_i)\hat{\mu}_i^3/3!}(f_{i2} + f_{i3}) \\
&= \frac{1}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6}(f_{i2} + f_{i3}),
\end{aligned}
$$

therefore, the generalised and modified Chao's estimator is given as

$$
\begin{aligned}
\widehat{N}^{(GMC)} &= n + \sum_{i=1}^{N} \hat{f}_{i0} \\
&= n + \sum_{i=1}^{f_2+f_3} \frac{1}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6}.
\end{aligned}
\tag{7.3}
$$

However, if a geometric mixture kernel is assumed, the estimated frequency of zero counts is given as

$$
\begin{aligned}
\hat{f}_{i0} = E[f_{i0}|f_{i2}, f_{i3}, \mu_i] &= \frac{p_0(\hat{\mu}_i)}{p_2(\hat{\mu}_i) + p_3(\hat{\mu}_i)}(f_{i2} + f_{i3}) \\
&= \frac{\hat{\zeta}_i}{\hat{\zeta}_i(1-\hat{\zeta}_i)^2 + \hat{\zeta}_i(1-\hat{\zeta}_i)^3}(f_{i2} + f_{i3}) \\
&= \frac{1}{(1-\hat{\zeta}_i)^2(2-\hat{\zeta}_i)}(f_{i2} + f_{i3}).
\end{aligned}
$$

Therefore, the generalised and modified Chao's estimator assuming a geometric mixture kernel is given as

$$
\begin{aligned}
\widehat{N}^{(GMC)} &= n + \sum_{i=1}^{N} \hat{f}_{i0} \\
&= n + \sum_{i=1}^{f_2+f_3} \frac{1}{(1-\hat{\zeta}_i)^2(2-\hat{\zeta}_i)},
\end{aligned}
\tag{7.4}
$$

where $\hat{\zeta}_i = \frac{1}{1+\hat{\mu}_i}$.

Application: Heroin data

Once all counts besides those of the doubletons and tripletons are truncated from the heroin case study data, a binomial logistic regression model can be fitted.

TABLE 7.7: Estimated total number of heroin users in Chiang Mai, Thailand for each linear predictor under consideration using the generalised-modified Chao's estimator, assuming a geometric mixture kernel, with corresponding AIC and BIC values for each model

Linear predictor	AIC	BIC	$\widehat{N}^{(GMC)}$
1	**301**	**304**	**1227**
2	303	310	1227
3	302	309	1231

Table 7.7 provides the AIC and BIC values for the models under consideration, where there is very little difference between the AIC values for the intercept-only model and the models with covariates. Given that the difference is very small, it is not surprising that there is very little difference between each of the total population size estimates for the different models. However, as expected, the estimates are much smaller than those found through using the other capture-recapture estimators which do not account for the excess singletons appropriately.

7.5 Uncertainty quantification

7.5.1 Variance by conditioning

The theoretical formula proposed by van der Heijden et al. (2003, page 314) for variance estimation by conditioning is altered to be used for the generalised-modified Chao's estimator, given in Equation 7.5,

$$\widehat{\text{Var}}(\widehat{N}^{(GMC)}) = E[\text{Var}(\widehat{N}^{(GMC)}|I_i)] + \text{Var}(E[\widehat{N}^{(GMC)}|I_i]). \tag{7.5}$$

An estimator for the variance in the first term is developed as

$$\begin{aligned}\text{Var}(\widehat{N}^{(GMC)}|I_i) &= \text{Var}\left(n + \sum_{i=1}^{f_2+f_3} \frac{1}{\mu_i^2/2 + \mu_i^3/6}\right) \\ &= \text{Var}\left(\sum_{i=1}^{f_2+f_3} \frac{1}{\mu_i^2/2 + \mu_i^3/6}\right),\end{aligned}$$

where $\frac{1}{\mu_i^2/2+\mu_i^3/6} = G(\mu_i|\hat{\boldsymbol{\beta}})$.

The above variance can be estimated using the multivariate- δ-method as

$$\widehat{\text{Var}}(\widehat{N}^{(GMC)}|I_i) = \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),$$

where for $\mu_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = \frac{\hat{\mu}_i^2 + \hat{\mu}_i^3/2}{(\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6)^2}\mathbf{h}(\mathbf{v}_i)^T.$$

The expectation for the second term is given as

$$\begin{aligned} E[\widehat{N}^{(GMC)}|I_i] &= E\left[n + \sum_{i=1}^{N} \frac{I_i}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6}|I_i\right] \\ &\approx \sum_{i=1}^{N} I_i w_i, \end{aligned}$$

where $w_i = 1 + \frac{p_0(\mu_i)}{p_i)}$ with $p_0(\mu_i) = \exp(-\mu_i)$ and

$$p_i = p_2(\mu_i) + p_3(\mu_i) = \exp(-\mu_i)\mu_i^2/2 + \exp(-\mu_i)\mu_i^3/6.$$

The indicator variable I_i is binary with expectation

$$E[I_i] = p_i,$$

and variance

$$\text{Var}(I_i) = p_i(1-p_i).$$

Hence the second term of Equation 7.5 is given by

$$\text{Var}\left(\sum_{i=1}^{N} I_i w_i\right) = \sum_{I_i}^{N} p_i(1-p_i)w_i^2,$$

which is estimated as

$$\begin{aligned} \widehat{\text{Var}}(E[\widehat{N}^{(GMC)}|I_i]) &= \sum_{i=1}^{N} \frac{I_i}{p_i} p_i(1-p_i)w_i^2 \\ &= \sum_{i=1}^{f_2+f_3} (1-\hat{p}_i)\left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2. \end{aligned}$$

Therefore, Equation 7.5 is given by

$$\widehat{\text{Var}}(\widehat{N}^{(GMC)}) = \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{f_2+f_3} (1-\hat{p}_i)\left(1+\frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2. \tag{7.6}$$

It is important to note that this variance formula is when a Poisson mixture kernel is assumed. In the case where a geometric mixture kernel is assumed, the variance formula is altered as follows.

The variance in the first term is given as

$$\begin{aligned}\text{Var}(\widehat{N}^{(GMC)}|I_i) &= \text{Var}\left(n + \sum_{i=1}^{f_2+f_3} \frac{1}{(1-\zeta_i)^2(2-\zeta_i)}\right) \\ &= \text{Var}\left(\sum_{i=1}^{f_2+f_3} \frac{1}{(1-\zeta_i)^2(2-\zeta_i)}\right),\end{aligned}$$

where $\zeta_i = \dfrac{1}{1+\mu_i}$ and $\dfrac{1}{(1-\zeta_i)^2(2-\zeta_i)} = G(\mu_i|\hat{\boldsymbol{\beta}})$.

Using the multivariate $\delta-$method, the above variance can be estimated as

$$\widehat{\text{Var}}(\widehat{N}^{(GMC)}|I_i) = \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),$$

where for $\hat{\mu}_i = \exp(\mathbf{h}(\mathbf{v}_i)^T\hat{\boldsymbol{\beta}})\tau_i$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = \left(\frac{3(1-\hat{\mu}_i)^2}{1+\hat{\mu}_i+\hat{\mu}_i^2+2\hat{\mu}_i^3} - \frac{(1+\hat{\mu}_i)^3(1+2\hat{\mu}_i+6\hat{\mu}_i^2)}{(1+\hat{\mu}_i+\hat{\mu}_i^2+2\hat{\mu}_i^3)^2}\right)\mathbf{h}(\mathbf{v}_i)^T.$$

As for the second term in Equation 7.5, the expectation is given as

$$\begin{aligned}E[\widehat{N}^{(GMC)}|I_i] &= E\left[n + \sum_{i=1}^{N} \frac{I_i}{(1-\zeta_i)^2(2-\zeta_i)}|I_i\right] \\ &\approx \sum_{i=1}^{N} I_i w_i,\end{aligned}$$

where $w_i = 1 + \dfrac{p_0(\mu_i)}{p_i}$ with $p_0 = \dfrac{1}{1+\mu_i}$ and

$$p_i = p_2(\mu_i) + p_3(\mu_i) = \left(1-\frac{1}{1+\mu_i}\right)^2\left(2-\frac{1}{1+\mu_i}\right).$$

As in the Poisson case, I_i is a binary indicator variable with expectation $E[I_i] = p_i$ and variance $\text{Var}(I_i) = p_i(1-p_i)$.

Therefore, the second term in Equation 7.5 is

$$\text{Var}\left(\sum_{i=1}^{N} I_i w_i\right) = \sum_{i=1}^{N} p_i(1-p_i)w_i^2,$$

which is estimated as

$$\begin{aligned}\widehat{\text{Var}}(E[\widehat{N}^{(GMC)}|I_i]) &= \sum_{i=1}^{N} \frac{I_i}{p_i} p_i(1-p_i)w_i^2 \\ &= \sum_{i=1}^{f_2+f_3} (1-\hat{p}_i)\left(1 + \frac{1}{(1-\hat{\zeta}_i)^2(2-\hat{\zeta}_i)}\right)^2,\end{aligned}$$

where $\hat{\zeta}_i = \dfrac{1}{1+\hat{\mu}_i}$.

Therefore, Equation 7.5 is given by

$$\begin{aligned}\widehat{\text{Var}}(\widehat{N}^{(GMC)}) = &\left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \\ &\sum_{i=1}^{f_2+f_3} (1-\hat{p}_i)\left(1 + \frac{1}{(1-\hat{\zeta}_i)^2(2-\hat{\zeta}_i)}\right)^2.\end{aligned} \tag{7.7}$$

Application: Heroin data

Assuming the binomial logistic regression model with gender as a covariate, Equation 7.6 leads to an estimated variance of 21406 with a corresponding standard error estimate of 146 and 95% confidence interval for $\widehat{N}^{(GMC)}$ of

$$\begin{aligned}\widehat{N}^{(GMC)} \pm 1.96 \times \sqrt{\widehat{\text{Var}}(\widehat{N}^{(GMC)})} &= 1227 \pm 1.96 \times \sqrt{21406} \\ &= (940, 1514).\end{aligned}$$

Unlike some of the intervals constructed in earlier chapters using conditioning for variance estimation with the suicide case study data, this interval appears reasonable given that it is of adequate width such that useful conclusions may be made. Additionally, the lower limit is not only greater than 0, but it is also greater than the observed population size, which was not the case for the generalised Chao's and generalised Zelterman's intervals for the suicide case study data in Chapter 5.

7.5.2 Bootstrap algorithm

Whilst the simulation study in Section 6.6.1 suggests that the parametric bootstrap algorithm should be used for optimal results, the semi-parametric bootstrap algorithm still performs well, with over 90% coverage. In the case of count data that is both zero-truncated and one-inflated, the parametric bootstrap algorithm becomes much more computationally intensive, and it was shown in Section 6.6.1 that there was little benefit of the parametric bootstrap algorithm over the semi-parametric bootstrap algorithm. Furthermore, if the median absolute deviation is utilised for constructing the confidence intervals, the coverage for the semi-parametric bootstrap algorithm increases to approximately 95%, as seen in Section 6.6.3, and is therefore used here (with Method 1 for accounting for model uncertainty) for quantifying the uncertainty of the estimates found using the generalised-modified Chao's estimator.

Application: Heroin data

The semi-parametric bootstrap algorithm with Method 1 for accounting for model uncertainty (fully accounting for it) is given in Algorithm 6. To apply this algorithm to the Heroin data with the generalised-modified Chao's estimator, Step 4 of the algorithm can be adjusted as follows.

Step 4: If $\sum_{i=1}^{n_b^*}(x_i^* = 2) = 0$ or $\sum_{i=1}^{n_b^*}(x_i^* = 3) = 0$, return to Step 2. Otherwise, truncate bootstrap dataset for all counts except $X = 2$ and $X = 3$. Fit competing binomial logistic regression models for the linear predictors in Table 7.3. Let $\tilde{j} = 1, \cdots, 3$ be the linear predictor which minimises the AIC, $\hat{q}$ be the corresponding fitted values and $\hat{\boldsymbol{\beta}}_{\tilde{j}}$ be the corresponding maximum likelihood estimates of $\boldsymbol{\beta}_{\tilde{j}}$.

Generalised-modified Chao's: Let

$$\hat{\mu}_i = 3\frac{\hat{q}_i}{1-\hat{q}_i},$$

for $i = 1, 2, \cdots, f_2 + f_3$, then, assuming a geometric mixture kernel, the estimated target population size is calculated as

$$\widehat{N}_b^{(GMC)*} = n + \sum_{i=1}^{f_2+f_3} \frac{f_{i2} + f_{i3}}{(1-\hat{\zeta}_i)^2(2-\hat{\zeta}_i)},$$

where $\hat{\zeta}_i = \frac{1}{1+\hat{\mu}_i}$.

This bootstrap algorithm results in a 95% confidence interval of $(1144, 1461)$, which is narrower than the interval computed through using conditioning in Section 7.5.1, and still contains the estimated total population size of $\widehat{N} = 1227$. For the heroin case study

data, both the variance by conditioning approach and the bootstrap algorithm approach work well for quantifying the uncertainty of the population size estimate, likely due in part to the observed population being reasonably large. However, it is worth noting that the variance by conditioning approach does not always perform well, seen in Chapter 5, where for the suicide case study data, the lower limits on multiple confidence intervals were negative. In this situation, the bootstrap algorithm is preferable given that it reduced the bias and resulted in confidence intervals of reasonable widths that contained the estimated value and had a lower limit greater than the observed population size. The variance by conditioning approach is less computationally intensive than the bootstrap algorithm, but returns wider confidence intervals in this scenario, and should the resulting interval be inappropriate (such as a negative lower limit for the estimated total population size when there is a positive observed population size), then the bootstrap algorithm should be used for the uncertainty quantification.

7.6 Simulation study

The results from the generalised-modified Chao's estimator appear reasonable, for example the total population size is notably greater than the observed population size with the missing zero counts but not as large as the alternative capture-recapture estimators which result in a very large number of zero counts. Additionally, the resulting estimated variance with corresponding confidence interval appears reasonable since the lower bound of the interval is both greater than zero and the observed population size. However, given that the true total population size is unknown, through applying the estimator to this dataset, the actual performance of the estimator is not known. Therefore, there is motivation for a simulation study to test whether the estimator not only performs well but is a better choice than the alternative, existing capture-recapture estimators in the case of one-inflated data. The simulation study will follow the same structure as used for testing the performance of the existing capture-recapture estimators, exploring the accuracy, precision, coverage and robustness of the generalised-modified Chao's estimator.

Table 7.8 displays the results from the simulation study when the true population size is $N = 1000$. The proportion of outliers in the simulated data ranges from 0% to 10%, with approximately $1/5$ of the counts being additional singletons. It is clear to see that when there are no outliers, for accuracy and coverage in particular, the generalised-modified Chao's is preferred. Whilst the precision for each of the estimators is not very large, for the Horvitz-Thompson, generalised Chao's and generalised Zelterman's estimators, that their accuracy is poor resulting in a coverage value of 0%. This means that none of the simulated confidence intervals contain the true population size.

TABLE 7.8: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's, generalised Zelterman's and generalised-modified Chao's when the counts are one-inflated, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0071$, $\lambda^U \approx 0.0085$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

Measure	Estimator	Proportion of Outliers					
		0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	HT	370	7.36e+07	8.44e+07	9.37e+07	1.06e+08	1.06e+08
	GC	663	800	801	786	781	793
	GZ	751	1226	1223	1220	1232	1254
	GMC	49	231	216	220	216	218
Precision	HT	189	4.96e+09	5.39e+09	5.66e+09	5.70e+09	5.67e+09
	GC	464	542	544	534	537	541
	GZ	539	770	774	770	780	789
	GMC	278	628	609	612	602	606
Coverage	HT	0%	98.3%	99.8%	100.0%	99.9%	99.8%
	GC	0%	0%	0%	0%	0%	0%
	GZ	0%	0%	0%	0%	0%	0%
	GMC	94.9%	94.4%	94.3%	95.1%	93.5%	95.1%

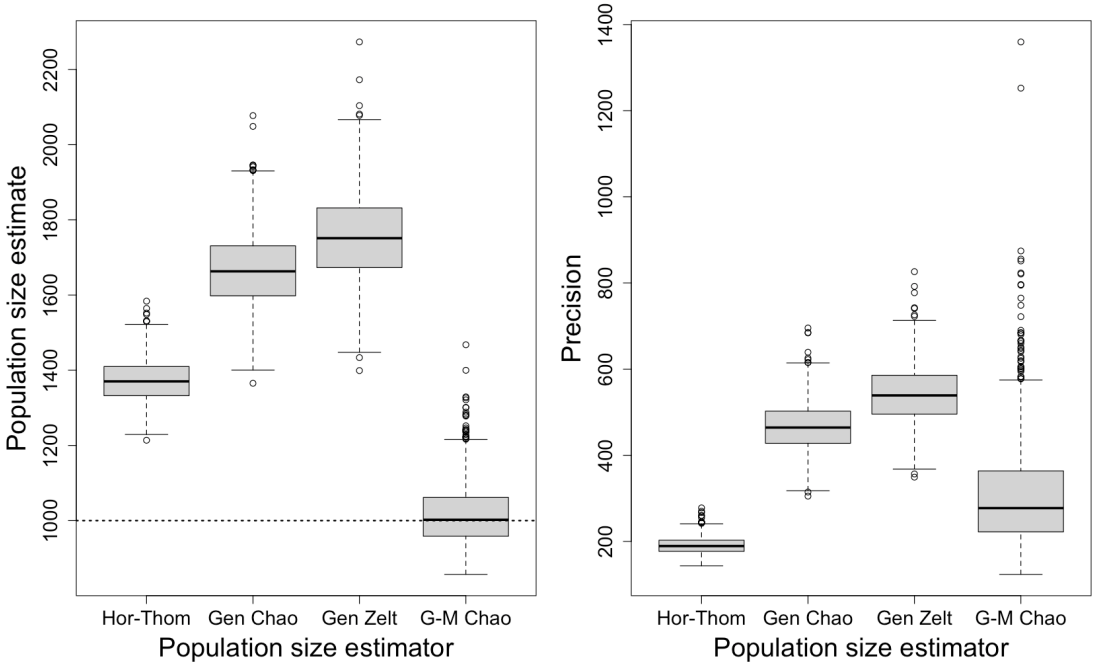

FIGURE 7.3: Box plots showing the accuracy of the population size estimates (left) and the precision of the resulting confidence intervals (right) for the capture-recapture population size estimators when the data is one-inflated, $\lambda_L = Q3 + 3 \times IQR$ and $N = 1000$.

These results are illustrated in Figure 7.3, where the box plot on the left provides the population size estimates for each capture-recapture estimator (accuracy), and the box plot on the right provides the precision values for the resulting confidence intervals for each capture-recapture estimator. It can be seen that the estimated population sizes for the generalised-modified Chao's estimator are not only centred around the true value ($N = 1000$), but they are also much closer than the estimates for the other three capture-recapture estimators, demonstrating how much more accurate the generalised-modified Chao's estimator is for one-inflated data. The box plot for the precision values looks very similar to Figures 4.3 and 4.5, where the Horvitz-Thompson estimator has the most precise confidence intervals, and there isn't a large difference in the performance of the generalised Chao's and generalised Zelterman's estimators. However, given that these estimators are inaccurate, the relatively narrow confidence intervals result in poor coverage and are therefore ineffective intervals for inference. As for the generalised-modified Chao's estimator, the precision values are centred at a reasonable width, between the Horvitz-Thompson estimator and the other two estimators, however, the spread of the precision values is much greater. This indicates that the results are quite inconsistent, however, given that the coverage is good, and the majority of the precision values result in narrow confidence intervals, useful conclusions can still be made from these intervals.

Once outliers are introduced, the preference for the generalised-modified Chao's estimator becomes even clearer. The accuracy and precision values for the Horvitz-Thompson estimator get very large with the coverage value also going to near 100%. This coverage value does not indicate that the Horvitz-Thompson estimator is a good choice as it is only so high as a result of the very wide confidence intervals. As for the generalised Chao's and generalised Zelterman's estimators, for the simulated data with outliers, the performance values are very consistent as with the simulation for non-one-inflated data in Chapter 4. The coverage and accuracy differ greatly, with 0% coverage and very inaccurate estimates (with N=1000 and accuracy of around 800 and 1200 for the estimators respectively, these estimates are approximately double that of the true value). These estimators are therefore not appropriate in the case where the counts are one-inflated.

When applying the generalised-modified Chao's estimator to the simulated data with both one-inflation and outliers, whilst the results are not as good as for the case with no outliers, the results are very consistent and still much better outcomes than with the alternative estimators. The coverage remains around 93-95% which is a very desirable amount of coverage, and the median distance from the true value is much smaller than for the alternative estimators. Lastly, whilst the median width of the confidence intervals does approximately double once outliers are introduced into the data, the width remains constant for all proportions of outliers and is a reasonable width comparatively to the intervals constructed with the Horvitz-Thompson estimator.

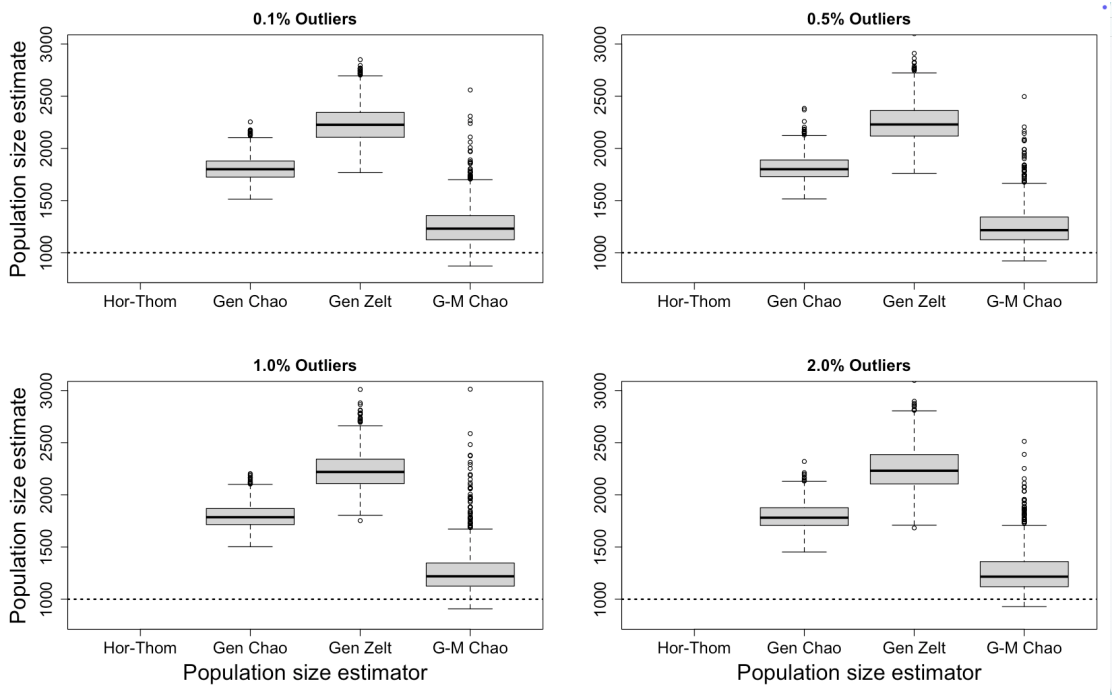

FIGURE 7.4: Box plots showing the accuracy of the population size estimates for the capture-recapture estimators when the data is one-inflated for different proportions of outliers when $\lambda_L = Q3 + 1.5 \times IQR$ and $N = 1000$.

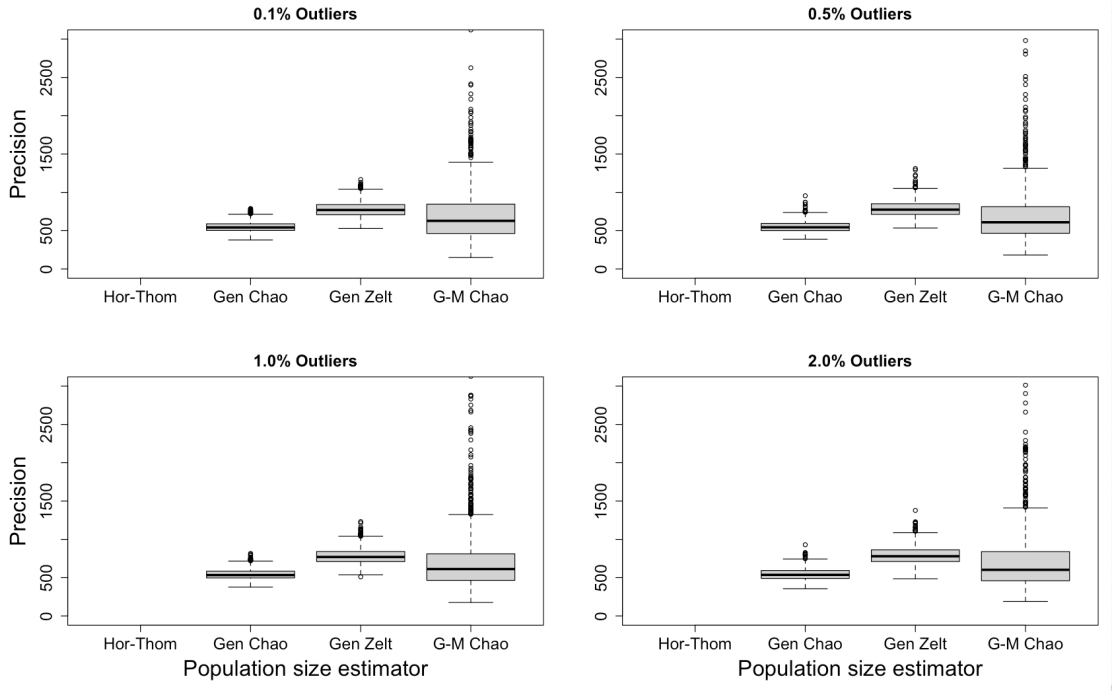

FIGURE 7.5: Box plots showing the precision of the confidence intervals for the capture-recapture estimators when the data is one-inflated for different proportions of outliers when $\lambda_L = Q3 + 1.5 \times IQR$ and $N = 1000$.

Figures 7.4 and 7.5 provide box plots show the accuracy and precision of the capture-recapture estimators respectively, visually illustrating the results from Table 7.8. A y-axis limit has been put in place due to very large Horvitz-Thompson estimator values, resulting in these values not being visible in the plots.

As for the other estimators, the generalised Chao's, generalised Zelterman's and generalised-modified Chao's estimated population sizes are all very consistent no matter the proportion of outliers, further providing evidence to their robustness. The generalised Chao's and generalised Zelterman's estimated population sizes are all very far from the true value, with many of the resulting estimates being almost double that of the true value, testifying that these estimators should not be used when there is evidence of one-inflation present in the data. The generalised-modified Chao's estimator performs well on the other hand. Whilst it is not as accurate as when there are no outliers included in the data, the estimated values are only about 200 units away from the true value, which is a notable improvement over the other estimators available. As seen previously in Figure 7.3, the precision values for the generalised-modified Chao's estimator are very spread out, but given the substantial improvement in both the coverage and the accuracy in comparison to the alternative estimators, and that the majority of the precision values are below around 700, it is still a good estimator to use and useful inference may be made from the resulting conclusions.

TABLE 7.9: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's, generalised Zelterman's and generalised-modified Chao's when the counts are one-inflated, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0071$, $\lambda^U \approx 0.0085$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	HT	184	-	2.14e+07	2.83e+07	3.21e+07	3.21e+07
	GC	333	-	402	396	396	396
	GZ	377	-	625	618	620	626
	GMC	33	-	120	108	117	105
Precision	HT	134	-	1.86e+09	1.99e+09	1.99e+09	1.98e+09
	GC	330	-	387	384	383	383
	GZ	382	-	554	552	558	556
	GMC	194	-	445	425	433	416
Coverage	HT	0%	-	97.5%	99.2%	99.6%	99.9%
	GC	0%	-	0%	0%	0%	0%
	GZ	0%	-	0%	0%	0%	0%
	GMC	93.3%	-	99.7%	99.9%	99.8%	99.6%

Table 7.9 provides the same information as Table 7.8 but for a true population size of $N = 500$ instead of $N = 1000$. It is important to note that for the proportion of outliers of 0.1%, there are no results given that 0.1% of 500 is 0.5 which is not an integer and rounds to either 0 or 1, where the results for 0 outliers is already provided in the table. The results for this smaller total population size follow the same trends as already discussed. Since the trends are consistent across both smaller and larger population, there is

evidence to suggest that in the presence of one-inflation, the generalised-modified Chao's estimator is a reliable estimator that performs notably better than the Horvitz-Thompson, generalised Chao's or the generalised Zelterman's estimators.

As discussed in Section 4.7, using the outlier rate lower bound defined in Equation 4.27 ($\lambda_L = Q3 + 3 \times IQR$) to determine the lower bound of the outlier rate can create clear outliers. However, outliers are not always clear and obvious in a dataset and therefore there is interest in using a less obvious outlier rate. The same simulation study as above is conducted, but using a smaller lower (and upper) bound for the outlier rate, determined by the following.

$$\lambda_L = Q3 + 1.5 \times IQR,$$

where the upper bound is still given by $\lambda_U = 1.2 \times \lambda_L$, but with this updated value of λ_L.

TABLE 7.10: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's, generalised Zelterman's and generalised-modified Chao's when the counts are one-inflated, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0046$, $\lambda^U \approx 0.0056$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	HT	370	7.42e+07	7.73e+07	8.04e+07	8.58e+07	8.64e+07
	GC	663	802	797	794	787	800
	GZ	751	1227	1227	1226	1237	1243
	GMC	49	223	228	231	219	230
Precision	HT	189	4.94e+09	5e+09	5.89e+09	4.98e+09	5.04e+09
	GC	464	546	540	538	540	543
	GZ	539	776	772	774	782	785
	GMC	278	628	620	618	615	626
Coverage	HT	0%	97.6%	98.7%	99.6%	99.7%	99.9%
	GC	0%	0%	0%	0%	0%	0%
	GZ	0%	0%	0%	0%	0%	0%
	GMC	94.9%	93.8%	93.7%	94.4%	94.2%	94.6%

The results from this simulation study are given in Tables 4.5 and 4.6 for total population sizes of $N = 1000$ and $N = 500$ respectively. Despite the outliers being smaller and therefore more subtle in the datasets, the results remain the same. The generalised Chao's and generalised Zelterman's estimators have reasonably wide confidence intervals, but due to very poor accuracy, they have 0% coverage consistently, and therefore should not be used when one-inflation is present in the data. Similarly, the Horvitz-Thompson estimator is not appropriate when there are excess singletons in the data, as whilst the coverage is above 95% when outliers are introduced, this is only due to the very wide confidence intervals, from which no reasonable conclusions can be made. For all proportions of outliers, the generalised-modified Chao's estimator performs well, with coverage over 90%, reasonably wide confidence intervals and good accuracy, and the

TABLE 7.11: Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz-Thompson, generalised Chao's, generalised Zelterman's and generalised-modified Chao's when the counts are one-inflated, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L \approx 0.0046$, $\lambda^U \approx 0.0056$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers.

		Proportion of Outliers					
Measure	Estimator	0.0%	0.1%	0.5%	1.0%	2.0%	10.0%
Accuracy	HT	184	-	2.32e+07	2.39e+07	2.68e+07	2.70e+07
	GC	333	-	403	394	392	393
	GZ	377	-	618	612	609	618
	GMC	33	-	115	117	120	110
Precision	HT	134	-	1.82e+09	1.76e+09	1.82e+09	1.84e+09
	GC	330	-	385	378	379	382
	GZ	382	-	550	546	546	551
	GMC	194	-	431	439	445	416
Coverage	HT	0%	-	97.6%	98.2%	99.9%	99.3%
	GC	0%	-	0%	0%	0%	0%
	GZ	0%	-	0%	0%	0%	0%
	GMC	93.3%	-	99.9%	100.0%	99.9%	99.8%

majority of confidence intervals contain the true value. Therefore, if there is evidence that the data is one-inflated, the generalised-modified Chao's estimator should be used.

Chapter 8

Conclusion and Future Work

This chapter provides the conclusions of the thesis and a discussion of potential future work.

8.1 Conclusion

Meta-analytic methods are widely accepted for computing an overall weighted-average from the findings of numerous independent studies focused on the same or similar research question. Traditional meta-analysis and regression modelling are not always appropriate however, such as in the case for this thesis where the focus was on counts of zero systematically missing from the data. In this instance, the traditional methods do not address the missing zero counts and typically result in an overestimation of the prevalence rate parameter and inaccurate total population sizes.

This thesis proposed a novel model-based approach to meta-analysis in Chapter 3, utilising zero-truncated count modelling to better estimate the prevalence, and more accurately model the data available. Using this approach, observed heterogeneity is addressed through investigation into covariate effects, enabling the prevalence rates to be stratified by the covariate information. Unobserved heterogeneity was also addressed, through overdispersion modelling, however, neither the suicide case study data or the hares case study data were not overdispersed, and for the suicide case study data, the zero-truncated intercept-only Poisson regression model was preferred, leading to an overall rate of completed suicide and no evidence of residual heterogeneity.

The model-based approach developed in Chapter 3 enables the estimation of both total and sub-population sizes in Chapter 4, using the Expectation Maximisation algorithm and the Horvitz-Thompson estimator. Alternative capture-recapture estimators were also explored in Chapter 4, the generalised Chao's and generalised Zelterman's estimators, providing more robust estimates of the total population sizes through their relaxed

distributional assumptions comparative to those of the Horvitz-Thompson estimator. The existing estimators were adjusted to allow for the inclusion of an exposure variable, the person-years, for the suicide case study data, for more accurate estimation. Additionally, each of the estimators were developed to allow for the geometric distribution, rather than the default Poisson distribution. This additional development allowed for more reliable application of all three capture-recapture estimators to the hares case study data.

The resulting population size is dependent on the capture-recapture estimator used, and this difference in results motivated the development of a simulation study to assess various performance measures for each estimator. The Horvitz-Thompson estimator was first developed multiple decades before the generalised Chao's and generalised Zelterman's estimators, so it is unsurprising that this estimator is typically the most used method. Although it is used more frequently, the alternative estimators performed notably better in certain situations, particularly the generalised Chao's estimator. The simulation study revealed that if there are no outliers in the data and the data therefore perfectly follows the given distributional assumption, the Horvitz-Thompson estimator is the superior estimator, providing more accurate results with more precise confidence intervals. However, with real data, this assumption is often not met with outliers being common. Once outlier counts were introduced to the simulated data, in various proportions and degrees, the accuracy, precision and coverage of the Horvitz-Thompson estimator all decrease, whereas the performance measures for the generalised Chao's and generalised Zelterman's estimators remain consistent and overall, perform well, and hence are more robust estimators for real data. However, given that the generalised Chao's and generalised Zelterman's estimators both truncate the data further, the sample size of the data should be taken into consideration when choosing which estimator to use, as using a smaller dataset in estimation may lead to less accurate results. If there is adequate data available, then the results from the simulation study indicate that the generalised Chao's estimator is the preferred method of estimating the total population size.

Chapters 5 and 6 both investigate methods of uncertainty quantification for the prevalence rate and population size estimation. Approximation-based variance estimation approaches were developed in Chapter 5 and applied to both the suicide and hares case study data. Through the application of these methods to the data, it was found that the analytical approach to uncertainty quantification is not always appropriate, possibly due to the limited data size of the suicide data. In the application to the suicide case study data, the intervals for the generalised Chao's and generalised Zelterman's estimators contained negative lower limits, and given that in this instance, you cannot have a negative total population size, nor a total population size lower than what is observed (27), these intervals were not appropriate. The analytical approach to uncertainty quantification returned appropriate confidence intervals for the hares case study data, which

is notably larger in size than the suicide case study data. Taking the potential for the resulting confidence intervals into consideration, alternative approaches to quantifying uncertainty are required. Additionally, the analytical approaches risk underestimating the amount of uncertainty given that model uncertainty is not accounted for, motivating the development of various bootstrap algorithms in Chapter 6.

Various designs of the bootstrap algorithm were developed in Chapter 6, each with a different combination of approach, being either non-parametric, semi-parametric or parametric, and method for accounting for model uncertainty. Each approach and method combination of the bootstrap algorithm was applied to both the suicide case study data and the hares case study data, which aided in identifying which combination to use.

When applying the bootstrap algorithms to the hares case study data, it could be seen that there was little difference between each of the resulting intervals. However, when the semi-parametric approach was utilised, the width of the intervals was increased, a trend that was seen in the intervals for each of the population size estimators. Additionally, the intervals found through using Method 1 for accounting for model uncertainty were slightly wider than the intervals found using the other methods. This difference is to be expected, given that Method 1 accounts for the most model uncertainty out of the three methods, investigating each of the competing models in each iteration, rather than assuming the model preferred the majority of times for the observed data was suitable for each iteration or not accounting for model uncertainty at all.

These trends were much more visible when each of the bootstrap algorithms were applied to the suicide case study data, where the available sample sizes are much smaller. For this application, the resulting intervals for Method 1 were very wide, no matter which population size estimator was used. These wide confidence intervals were particularly notable for the non-parametric and semi-parametric approaches, where high correlation between covariate combinations led to bias and intervals which are not useful for inference.

Out of the three approaches, the parametric approach, Approach 3, resulted in the most appropriate intervals overall. These findings were seen in both datasets, but on a much less notable scale for the hares case study data.

Overall, these results indicate that, particularly for smaller datasets, the parametric approach to the bootstrap algorithm should be used, namely due to the reduction in bias in the intervals compared to the other approaches. As for accounting for model uncertainty, provided that there is a clear preference for a single model, Method 2 returns the most appropriate intervals with less bias and computational time compared to Method 1. However, if there is not a clear preference for a single model, Method 1 is preferred, given the importance for accounting for model uncertainty to not underestimate the variation. This was the case for the hares case study data in Approaches 1 and 2 to

the bootstrap algorithms when the generalised Chao's and generalised Zelterman's estimates were used.

These conclusions are supported by the simulation study in Section 6.6, where the intervals found using Approach 3 was both the most precise and have the highest coverage at 98%. Whilst the simulation study indicates that Method 1 is the preferred method, based on the results alone, the difference between the three methods is negligible (0.28 difference in the precision and 0.1% difference in the coverage between Methods 1 and 2), so taking the computational time into account, if there is a clear preference for a single model with the observed data, Method 2 with the parametric bootstrap algorithm should be used.

Alternative methods of confidence interval construction explored revealed that whilst the standard percentile method performs very well, it is susceptible to bias and skewness and therefore it is not always the best option. The bias-corrected and bias-corrected and accelerated percentile methods returned less biased and skewed intervals, but the simulation study in Section 6.6 revealed that the coverage reduces to around 92% and 90% respectively, despite the width of the intervals increasing. Therefore, unless there is significant bias, as with Approach 2 of the bootstrap algorithm with the suicide case study data, the standard percentile method is preferred.

Median absolute deviation was also investigated, where it was found to be a very slight improvement over the standard percentile method, with similar precision values and a slight increase in coverage (from 98.6% to 98.8% coverage). Given that there is no notable difference in the computational intensity of the two methods, and that the median absolute deviation is more resilient to bias and skewness since it gives less weight to extreme values, provided that the results from the bootstrap are approximately normally distributed. the median absolute deviation should be used for confidence interval construction.

The final chapter of this thesis introduced the concept of one-inflation and explored the failings of the current methodologies. The methods discussed previously in this thesis fail to account for the excess singletons that are present in one-inflated data, leading to grossly overestimated population sizes, and the modified Chao's estimator does not allow for covariate information which can limit the accuracy of the resulting population size estimates. The novel generalised-modified Chao's estimator was developed to account for both the available covariate information and excess singletons. Application of these methods to the newly introduced heroin case study data illustrated the impact of not appropriately accounting for one-inflation and a simulation study conducted at the end of Chapter 7 provided evidence that the generalised-modified Chao's estimator performs not only better than the existing capture-recapture estimators but also performs well overall, including when outliers are included in the data. Therefore, if there is

evidence that there is one-inflation present in a dataset, and covariate information available, the generalised-modified Chao's estimator should be used.

8.2 Future work

This section outlines some areas which can be expanded or worked on in the future to expand the methodologies covered in this thesis.

- Expand the simulation study work to include other data scenarios with different parameters, number of covariates and population sizes.
- Explore a Bayesian approach to capture-recapture parameter estimation (see Lee and Chen, 1998; King et al., 2009, for examples of Bayesian capture-recapture approaches) and compare the reliability of this approach to the reliability of the frequentist approach taken in this thesis so far.
- Explore the iterated bootstrap algorithm and the percentile-t method.
- Find more meta-analytic data to apply the methodologies explored in this thesis to, particularly another one-inflated dataset with covariate information.
- Explore the Turing estimator and investigate whether it can also be generalised to allow for covariate information.
- Expand on the application of the bootstrap algorithm application to one-inflated data further.
- Explore the impact of the alternative methods of confidence interval construction (namely the MAD approach) on one-inflated data via a simulation study.

8.2.1 Iterated bootstrap algorithm and the percentile-t method

As discussed in Section 6.5.1, the traditional percentile method for confidence interval construction is prone to bias. An alternative approach that can be utilised is the percentile-t method, proposed by Efron (1979), a more computationally complex approach compared to the standard approach but leads to intervals with comparatively less bias.

Given $\widehat{N}$, the estimated value of the total population size N, and its standard error, $s.e.(\widehat{N})$, the percentile-t method utilises the pivotal statistic in Equation 8.1 to construct the interval.

$$T = \frac{\widehat{N} - N}{s.e.(\widehat{N})} \tag{8.1}$$

The estimated total population size and the corresponding standard error are computed using the bootstrap algorithms seen in Section 6. However, a problem arises with the pivotal statistic in that the standard bootstrap approaches discussed prior, only one pivotal statistic can be computed. Using the work of Efron (1979), the percentile-t interval is given as

$$(\widehat{N} - \hat{t}^{(1-\alpha)} s.e.(\widehat{N}), \widehat{N} - \hat{t}^{(\alpha)} s.e.(\widehat{N})),$$

where $\hat{t}^{\alpha}$ is the $100\alpha th$ percentile of t. Therefore, if only one value of t is to be computed, no percentiles can be extracted and the interval cannot be constructed.

To circumvent the issue of only one value of t being computed and incorporate the percentile-t method into the bootstrap algorithm, Hall et al. (1989) suggested use of the iterated bootstrap first introduced by Hall (1986) and Beran (1987). This iterated, or 'double', bootstrap results in percentile-t intervals with high coverage, stable lengths and accurate endpoints in both complex situations and in situations where the sample size is small (DiCiccio et al., 1992).

Following the same steps as developed in the bootstrap algorithms in Section 6, the iterated can be constructed with the addition of an internal bootstrap from which the parameter of interest is estimated and corresponding standard error computed.

Given these estimates, the number of bootstrap samples of the external bootstrap, $b_1 = 1, \cdots, B_1$, and the number of bootstrap samples of the internal bootstrap, $b_2 = 1, \cdots, B_2$, the standard error of the estimated population size is calculated as

$$s.e.(\widehat{N}^*_{b_1}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{B_2 - 1} \sum_{b_2=1}^{B_2} \left(\widehat{N}^{**}_{b_2} - \frac{1}{B2} \sum_{b_2=1}^{B_2} \widehat{N}^{**}_{b_2} \right)^2}.$$

Using this standard error, an estimated pivotal statistic can be computed as

$$\hat{t}_{b_1} = \frac{\widehat{N}^*_{b_1} - \widehat{N}}{s.e.(\widehat{N}^*_{b_1})},$$

where $\widehat{N}^*_{b_1}$ is the estimated population size from bootstrap sample b_1, $\widehat{N}$ is the estimated population size using the original data and $s.e.(\widehat{N}^*_{b_1})$ is the standard error of the population size from bootstrap sample b_1.

The resulting percentile-t confidence interval is then given by

$$\left(\widehat{N} - \hat{t}^{(1-\alpha)} s.e.(\widehat{N}), \widehat{N} - \hat{t}^{(\alpha)} s.e.(\widehat{N}) \right),$$

where $\hat{t}^{(\alpha)}$ is the $100\alpha th$ percentile of $\hat{t}$ and

$$s.e.(\widehat{N}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{B_1 - 1} \sum_{b_1=1}^{B_1} \left(\widehat{N}^*_{b_1} - \frac{1}{B1} \sum_{b_1=1}^{B_1} \widehat{N}^*_{b_1} \right)^2}.$$

In Section 6, comparison of the non-parametric, semi-parametric and parametric bootstrap algorithm approaches indicates that the parametric approach is favoured, and hence this approach is utilised for the iterative bootstrap method.

For the capture-recapture population size estimator, E, where

$$E = \begin{cases} HT & \text{for Horvitz-Thompson,} \\ GC & \text{for generalised Chao's, or} \\ GZ & \text{for generalised Zelterman's,} \end{cases}$$

the iterated bootstrap algorithm is given formally as follows in Algorithm 16.

The choice of B_2 impacts the heavy computational burden that is a disadvantage for the percentile-t approach, with the larger the number of internal bootstrap replications the larger the computational burden. Given that the internal bootstrap is used to find the standard error, the burden can be minimised by setting B_2 to be considerably smaller than the number of bootstrap replications required to give reliable confidence intervals. Efron and Tibshirani (1993, page 52) suggests that for computing the standard error, $B_2 = 50$ is typically large enough to produce a good estimation, with it being a rare occurrence that $B_2 \geq 200$ is required for standard error estimation. From this, letting $50 \leq B_2 < 200$ will aid in reducing the computational burden whilst still giving acceptable results.

Section 8.2.1 contains some application of these methods to the suicide case study used throughout this thesis. Future work for this section includes expanding the application further to include more datasets, such as the hares case study data, which have a larger sample size and investigate whether this has a positive impact on the results.

Additionally, a simulation study could be conducted to explore the performance of the approach to confidence interval construction, as is done in Section 6.6, looking at the precision and coverage of the confidence intervals. Preliminary investigations into this performance indicated that the coverage was very low, approximately 50-70% depending on which approach and method for the bootstrap algorithm was used. Given the increase in computational intensity of this method, along with a likely decrease in coverage, future work could also include working on the iterated bootstrap algorithm to improve its performance.

Algorithm 16 Iterated Bootstrap

Step 1: Fit the J competing models to the observed data.
Step 2: For each model, calculate the BIC weights using Equation 6.1, where the BIC weight of model l can be seen as the probability of model l being selected as the best candidate model (Wagenmakers and Farrell, 2004).
Step 3: Let $b_1 = 1$.
Step 4: Sample l^* from $\{1, 2, \cdots, J\}$, where J is the number of competing models and the values have respective probabilities $\{w_1, w_2, \cdots, w_J\}$ of being sampled.
Step 5: Sample n counts from the model given by the linear predictor distribution pair, (j^*, D^*). Use these sampled counts x_i^* to create a sampled dataset $\{(\tau_1, \mathbf{v}_1, x_1^*), \ldots, (\tau_n, \mathbf{v}_n, x_n^*)\}$, where τ_i and $\mathbf{v}_i$ are the observed person-years and covariates respectively for $i = 1, 2, \cdots, n$. Sample x_i^* from the distribution given by (j^*, D^*) with probability function

$$\begin{cases} p_x\left(\tau_i \exp\left[\mathbf{h}_{j^*}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{j^*}^{(D^*)}\right]\right) & \text{if } D^* = (P), \\ p_x\left(\tau_i \exp\left[\mathbf{h}_{j^*}(\mathbf{v}_i)^T \hat{\boldsymbol{\beta}}_{j^*}^{(D^*)}\right], \hat{\theta}_{j^*}^{(D^*)}\right) & \text{if } D^* = (NB), \end{cases}$$

where if $D^* = (NB)$, $\hat{\theta}_{j^*}^{D^*}$ is the estimated dispersion parameter.

Step 6: Estimate the total population size $\widehat{N}_b^{(E)} *$ for a given capture-recapture estimator E using the methods discussed in Chapter 4.
Step 7: Let $b_2 = 1$.
Step 8: Create an internal bootstrap dataset by repeating steps 4-6 $B2$ times using the dataset created in Step 5, computing $N_{b_2}^{(E)**}$ for $b_2 = 1, \cdots, B2$.
Step 9: Compute the pivotal statistic as

$$\hat{t}_{b_1}^{(E)} = \frac{\widehat{N}_{b_1}^{(E)*} - \widehat{N}}{s.e.(\widehat{N}_{b_1}^{(E)*})},$$

where

$$s.e.(\widehat{N}_{b_1}^{(E)*}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{B_2 - 1} \sum_{b_2=1}^{B_2} \left(\widehat{N}_{b_2}^{(E)**} - \frac{1}{B2} \sum_{b_2=1}^{B_2} \widehat{N}_{b_2}^{(E)**}\right)^2}.$$

Step 10: If $b_1 = B1$, stop. Otherwise, return to step 4 with $b_1 := b_1 + 1$.

Application: Suicide data

For $B_1 = 100$ external bootstrap replications and $B_2 = 50$ internal bootstrap replications, the iterated bootstrap results in a 95% percentile-t confidence interval for the Horvitz-Thompson estimator of (95, 156), which is narrower than that of the standard percentile approach (for Approach 1, Method 2) seen in Section 6 of (106, 185). This percentile-t interval is also narrower than both the bias-corrected and the bias-corrected and accelerated intervals seen in Section 6.5.1, suggesting that the percentile-t approach may lead to a larger reduction in bias. The interval is slightly narrower than the same interval constructed using the median absolute deviation approach of (97, 171), suggesting a further reduction in bias.

For any larger of a value of $B1$, the bias gets higher and the confidence interval gets very wide. Similarly for the other capture-recapture estimators, the confidence interval is too wide to make any reliable conclusions from the results, with some very large outlier estimates of the total population size. For the other estimators however, this poor result occurs for any number of replications.

Given that the inclusion of model uncertainty through model testing and selection processes in each bootstrap iteration leads to high correlation between the sub-populations in addition to high bias, resulting in wide and ineffectual confidence intervals, model uncertainty can be accounted for through testing which model and distribution combination is favoured the majority of times, where only the resulting combination is used in the iterated bootstrap algorithms. In Section 6, testing of which model and distribution combinations indicates that for each bootstrap algorithm method and estimator, the intercept-only model is preferred, with the Poisson distribution for the Horvitz-Thompson estimator and binomial logistic for the generalised Chao's and generalised Zelterman's estimators. The results from using these model and distribution combinations in the iterated bootstrap algorithm are given in Table 8.1.

TABLE 8.1: Values of the 95% percentile-t confidence intervals for the suicide data using the non-parametric, semi-parametric and parametric approaches to the iterated bootstrap algorithm using Method 2 to account for model uncertainty for the Horvitz-Thompson, generalised Chao's and generalised Zelterman's estimators when $B_1 = 500$ and $B_2 = 200$.

	Capture-recapture estimator		
Bootstrap algorithm	Horvitz-Thompson	Generalised Chao's	Generalised Zelterman's
Non-parametric	(129, 138)	(72, 198)	(70, 202)
Semi-parametric	(120, 146)	(116, 185)	(63, 198)
Parametric	(123, 140)	(114, 193)	(116, 197)

The values in Table 8.1 indicate that there is little reduction in bias, if any, when using the percentile-t approach to confidence interval construction compared to the median absolute deviation. Given that the percentile-t approach is very computationally intensive, and the differences between the percentile-t intervals and the MAD intervals are very small, there may not be a benefit in using the percentile-t approach, especially if there is a reduction in coverage. Preliminary exploration of coverage through using similar simulation studies as in Section 6.6.1 indicate that the coverage of the percentile-t approach is between 50 and 75%, much lower than that of the other confidence interval construction methods, particularly MAD where the coverage is at almost 99%. This can be explored further in the future, with formal simulation study testing, and possibly development of the percentile-t approach to improve its coverage.

Appendix A

Regression modelling

A.1 Suicide data: age as a covariate

TABLE A.1: Meta-analytic data from $n = 27$ observed studies from Peterhänsel et al. (2013) numbered and ordered by decreasing size of person-years. The table includes the number of person-years, the proportion of women, the country of origin, the average age of participants at the start of study and the number of completed suicides for each study. The proportion of women for 24. Smith 2004 is unknown but is imputed to be 0.823. The country of origin for 21. Kral 1993 is reported as "USA/Sweden" but changed to USA for model fitting.

Study	Person-years	Proportion of women	Country of origin	Age	Number of completed suicides
i	τ_i	v_{i1}	v_{i2}	v_{i3}	x_i
1. Adams 2007	77602	0.860	USA	39.5	21
2. Marceau 2007	10388	0.720	Canada	40.1	6
3. Marsk 2010	8877	0.000	Sweden	-	4
4. Pories 1995	8316	0.832	USA	37.3	3
5. Carelli 2010	6057	0.684	USA	44.63	1
6. Busetto 2007	4598	0.753	Italy	38	1
7. Smith 1995	3882	0.889	USA	-	2
8. Peeters 2007	3478	0.770	Australia	47.1	1
9. Christou 2006	2599	0.820	Canada	42	2
10. Günther 2006	2244	0.837	Germany	32	1
11. Capella 1996	2237	0.822	USA	37	3
12. Suter 2011	2152	0.744	Switzerland	39.4	3
13. Suter 2006	1639	0.865	Switzerland	38	1
14. Van de Weijgert 1999	1634	0.870	Netherlands	34	1
15. Cadière 2011	1362	0.834	Belgium	41	1
16. Mitchell 2001	1121	0.847	USA	56.8	1
17. Himpens 2011	1066	0.902	Belgium	50	1
18. Näslund 1994	799	0.812	Sweden	37	2

(To be continued)

Study	Person-years	Proportion of women	Country of origin	Age	Number of completed suicides
i	τ_i	v_{i1}	v_{i2}	v_{i3}	x_i
19. Forsell 1999	761	0.761	Sweden	40	1
20. Powers 1997	747	0.847	USA	39.4	1
21. Kral 1993	477	0.812	*USA*	38	1
22. Näslund 1995	457	0.592	Sweden	39.3	1
23. Powers 1992	395	0.850	USA	38.8	1
24. Smith 2004	354	*0.823*	USA	39.5	1
25. Nocca 2008	228	0.677	France	41.57	1
26. Svenheden 1997	166	0.791	Sweden	-	1
27. Pekkarinen 1994	146	0.704	Finland	36	1

Peterhänsel et al. (2013) includes information on the average age of participants in the studies, also provided in Table A.1, however, this covariate information has issues with its quality. Firstly, not all of the studies included in the systematic review report an average age, only 24 of the 27 studies report this information. The suicide case study data is already small, so reducing its size by another three studies in order to include age as a covariate may have a detrimental effect on the accuracy of the results. Of these 24 studies, three studies are split into two groups, meaning that they report two different average ages. Additionally, the measure of centrality chosen for the 24 studies which report an average age varies, with two studies reporting the median age, six studies reporting the mean age and 16 studies that do not report which measure of centrality they utilise. Finally, the observational or follow-up periods of the 24 studies varies from several months to multiple decades. This can lead to problems with interpreting the results from a model with the age covariate included as the average age of that study may no longer be relevant. For example, for a study which reports the median age at the start of their study to be 40, but their suicides occurred 20 years later, then the reported age of 40 would no longer be reliable and any inference made would also be inaccurate.

Despite this poor quality of the covariate information, the ages that are reported will be explored as a covariate in the modelling process for completeness. Where studies give two average ages, the midpoint of the two ages given will be used for the overall age of the study, and the three studies which did not report an average age will be excluded from the model. It is important to note that since the dataset has now changed, with 24 not 27 observed studies, any AIC or BIC values resulting from the models cannot be compared to the existing models in Section 3.2, and therefore those models are refitted with the remaining 24 data points.

Each of the linear predictors with the explored covariate combinations are given in Table A.2, where v_1 refers to the proportion of women, v_2 refers to the country of origin and v_3 refers to the average age of the participants. The results here support the conclusions made in Section 3.2 with the complete dataset, where the zero-truncated

TABLE A.2: Linear predictors used to fit the zero-truncated Poisson and negative-binomial (NB) models and the corresponding BIC values, where Y indicates that the main effect or interaction is included in the model and N otherwise.

Linear Predictor	v_1	v_2	v_3	v_1x_2	v_1v_3	v_2v_3	$v_1v_2v_3$	Poisson BIC	NB BIC
1	**N**	**N**	**N**	**N**	**N**	**N**	**N**	**44.6**	47.8
2	Y	N	N	N	N	N	N	47.1	50.3
3	N	Y	N	N	N	N	N	46.7	49.9
4	N	N	Y	N	N	N	N	46.7	49.9
5	Y	Y	N	N	N	N	N	49.8	53.0
6	Y	N	Y	N	N	N	N	48.5	51.6
7	N	Y	Y	N	N	N	N	48.9	52.0
8	Y	Y	Y	N	N	N	N	51.6	54.8
9	Y	Y	N	Y	N	N	N	51.5	54.7
10	Y	N	Y	N	N	N	N	50.1	53.3
11	N	Y	Y	N	N	Y	N	49.5	52.7
12	Y	Y	Y	Y	Y	N	N	53.4	56.6
13	Y	Y	Y	Y	N	Y	N	53.2	56.4
14	Y	Y	Y	N	Y	Y	N	51.2	55.3
15	Y	Y	Y	Y	Y	Y	N	54.9	58.0
16	Y	Y	Y	Y	Y	Y	Y	60.3	63.5

intercept-only Poisson model is preferred, and including age as a covariate in the model only increases the BIC value. Therefore, there is evidence that the age covariate should not be included in the modelling.

A.2 The Poisson as a limiting case of the negative-binomial

For mean of negative-binomial distribution: $\mu = \frac{\alpha(1-\theta)}{\theta}$, and variance: $\sigma^2 = \frac{\alpha(1-\theta)}{\theta^2}$. The mean can be rearranged in terms of θ

$$\theta = \frac{\alpha}{\alpha + \mu}$$

Substitute the value of θ into the density of the negative-binomial distribution

$$\begin{aligned} p_x(\theta) &= \frac{\Gamma(\alpha + x)}{\Gamma(x+1)\Gamma(\alpha)}(1-\theta)^x\theta^\alpha \\ \Rightarrow p_x(\mu) &= \frac{\Gamma(\alpha + x)}{\Gamma(x+1)\Gamma(\alpha)}\frac{\mu^x}{(\mu+\alpha)^x}\left(1+\frac{\mu}{\alpha}\right)^{-\alpha} \\ p_x(\mu)_{\alpha\to\infty} &= \exp(-\mu)\frac{\mu^x}{x!} \end{aligned}$$

which can be described as $X \sim Poisson(\mu)$. The value of α determines the deviation of the negative-binomial from the Poisson distribution so for large α, using the negative-binomial distribution as an approximation of Poisson is robust.

References

Alpizar-Jara, R. and Pollock, K. H. A combination line transect and capture-recapture sampling model for multiple observers in aerial surveys. *Environmental and Ecological Statistics*, 3(4):311–327, 1996.

Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., and Vos, T. Meta-analysis of prevalence. *J Epidemiol Community Health*, 67(11):974–978, 2013.

Beran, R. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media, 2007.

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

Böhning, D. Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scandinavian Journal of Statistics*, 37(2):221–236, 2010.

Böhning, D. Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *Metron*, 73(2):201–216, 2015.

Böhning, D. and Friedl, H. Population size estimation based upon zero-truncated, one-inflated and sparse count data: Estimating the number of dice snakes in Graz and flare stars in the Pleiades. *Statistical Methods & Applications*, 30(4):1197–1217, 2021.

Böhning, D. and van der Heijden, P. G. M. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, 3(2):595–610, 2009.

Böhning, D. and van der Heijden, P. G. M. The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. *The Annals of Applied Statistics*, 2019.

Böhning, D., Dietz, E., Kuhnert, R., and Schön, D. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14:29–43, 2005.

Böhning, D., Kuhnert, R., and Vilas, V. D. R. Capture-recapture estimation by means of empirical Bayesian smoothing with an application to the geographical distribution of hidden scrapie in Great Britain. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 60(5):723–741, 2011.

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

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., and Arnold, M. A generalization of Chao's estimator for covariate information. *Biometrics*, 69(4): 1033–1042, 2013b.

Böhning, D., Bunge, J., and van der Heijden, P. G. M. *Capture-Recapture Methods for the Social and Medical Sciences*. Boca Raton: CRC Press, 2018.

Böhning, D., Kaskasamkul, P., and van der Heijden, P. G. M. A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, 82(3):361–384, 2019.

Borchers, D. L., Buckland, S. T., Zucchini, W., and Borchers, D. L. *Estimating Animal Abundance: Closed Populations*. London: Springer, 2002.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. *Introduction to Meta-Analysis*. John Wiley & Sons, 2021.

Braeye, T., Verhaegen, J., Mignon, A., Flipse, W., Pierard, D., Huygen, K., Schirvel, C., and Hens, N. Capture-recapture estimators in epidemiology with applications to pertussis and pneumococcal invasive disease surveillance. *PloS One*, 11(8):e0159832, 2016.

Brittain, S. and Böhning, D. Estimators in capture–recapture studies with two sources. *AStA Advances in Statistical Analysis*, 93(1):23–47, 2009.

Brown, L. D., Cai, T. T., and DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.

Buckland, S. T. and Garthwaite, P. H. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, pages 255–268, 1991.

Carothers, A. D. Capture-recapture methods applied to a population with known parameters. *The Journal of Animal Ecology*, pages 125–146, 1973.

Chao, A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791, 1987.

Chao, A. Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, pages 427–438, 1989.

Chao, A. An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, 2001.

Chapman, D. G. Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Stat.*, 1:131–160, 1951.

Choban, P. S., Jackson, B., Poplawski, S., and Bistolarides, P. Bariatric surgery for morbid obesity: Why, who, when, how, where, and then what? *Cleveland Clinic Journal of Medicine*, 69(11):897–903, 2002.

Cooper, H., Hedges, L. V., and Valentine, J. C. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, 2019.

Cruyff, M. J. L. F. and van der Heijden, P. G. M. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, 50(6):1035–1050, 2008.

Davison, A. C. *Statistical Models*. Cambridge University Press, 2003.

Dempster, A. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

Dennett, L. C. and Böhning, D. Performance of capture-recapture population size estimators under covariate information. *arXiv preprint arXiv:2312.08391*, 2023.

Dennett, L. C., Overstall, A., and Böhning, D. Zero-Truncated Modelling in a Meta-Analysis on Suicide Data after Bariatric Surgery. *The American Statistician*, pages 1–14, 2025. URL `https://doi.org/10.1080/00031305.2025.2507380`.

DiCiccio, T. J., Martin, M. A., and Young, G. A. Analytical approximations for iterated bootstrap confidence intervals. *Statistics and Computing*, 2:161–171, 1992.

Dixon, J. B. The effect of obesity on health outcomes. *Molecular and Cellular Endocrinology*, 316(2):104–108, 2010.

Dobson, A. J. and Barnett, A. G. *An Introduction to Generalized Linear Models*. CRC press, 2018.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., and Jermiin, L. S. Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2):553–565, 2020.

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1 – 26, 1979.

Efron, B. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319, 1981a.

Efron, B. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158, 1981b.

Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, 1982.

Efron, B. Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1):45–58, 1985.

Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC, 1993.

Egger, M., Smith, G. D., and Altman, D. *Systematic Reviews in Health Care: Meta-Analysis in Context*. John Wiley & Sons, 2008.

Fienberg, S. E. The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59(3):591–603, 1972.

Hall, P. On the bootstrap and confidence intervals. *The Annals of Statistics*, pages 1431–1452, 1986.

Hall, P. Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics*, pages 927–953, 1988.

Hall, P. and Martin, M. A. A note on the accuracy of bootstrap percentile method confidence intervals for a quantile. *Statistics & Probability Letters*, 8(3):197–200, 1989.

Hall, P., Martin, M. A., and Schucany, W. R. Better nonparametric bootstrap confidence intervals for the correlation coefficient. *Journal of Statistical Computation and Simulation*, 33(3):161–172, 1989.

Hampel, F. R. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Hauck, W. W. and Donner, A. Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, 72:851–853, 1977.

Holling, H., Böhning, W., Böhning, D., and Formann, A. K. The covariate-adjusted frequency plot. *Statistical Methods in Medical Research*, 25(2):902–916, 2016.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. .

Huber, P. J. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 1981.

Keith, L. B. and Meslow, E. C. Trap response by snowshoe hares. *The Journal of Wildlife Management*, pages 795–801, 1968.

Khwaja, H. A. and Bonanomi, G. Bariatric surgery: techniques, outcomes and complications. *Current Anaesthesia & Critical Care*, 21(1):31–38, 2010.

King, R., Bird, S. M., Hay, G., and Hutchinson, S. J. Estimating current injectors in Scotland and their drug-related death rate by sex, region and age-group via Bayesian capture—recapture methods. *Statistical Methods in Medical Research*, 18(4):341–359, 2009.

Kulinskaya, E., Morgenthaler, S., and Staudte, R. G. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. John Wiley & Sons, 2008.

Lee, S.-M. and Chen, C. W. S. Bayesian inference of population size for behavioral response models. *Statistica Sinica*, pages 1233–1247, 1998.

Leys, C., Klein, O., Bernard, P., and Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

Lobstein, T., Brinsden, H., and Neveux, M. World Obesity Atlas 2022, 2022.

McCrea, R. S. and Morgan, B. J. T. *Analysis of Capture-Recapture Data*. Boca Raton: CRC Press, 2014.

NASA Earth Observatory. Hazy Skies in a Growing City, 2024. URL `https://earthobservatory.nasa.gov/images/152635/hazy-skies-in-a-growing-city#:~:text=Chiang%20Mai's%20population%20has%20grown,to%20United%20Nations%20population%20data`. Visited on 31/07/2024.

NatCen Social Research, University College London. Health Survey for England. *London: NHS Digital*, 2005-2023. URL `https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england`.

National Institute on Drug Abuse website. What is heroin and how is it used?, 2021. URL `https://nida.nih.gov/publications/research-reports/heroin/what-heroin`. Visited on 17/06/2024.

Niwitpong, S., Böhning, D., van der Heijden, P. G. M., and Holling, H. Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika*, 76:495–519, 2013.

Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., and Lee, S. Suicide and suicidal behavior. *Epidemiologic Reviews*, 30(1):133–154, 2008.

Norris, J. L. and Pollock, K. H. Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, 3:235–244, 1996.

Oehlert, G. W. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Orchard, T. and Woodbury, M. A. A missing information principle: theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume Volume 1: Theory of Statistics, pages 697–716. University of California Press, 1972.

Overton, W. S. and Stehman, S. V. The Horvitz-Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling. *The American Statistician*, 49(3):261–268, 1995.

Panyalert and Lanamteng. Factors influencing drug rehabilitation attendance at Thanyarak Chiang Mai hospital for substance addiction - proceedings of the 18th scientific and technological conference, 2020. Maejo University, Thailand.

Paulos, J. A. *A Mathematician Plays the Stock Market*. Basic Books, 2007.

Peterhänsel, C., Petroff, D., Klinitzke, G., Kersting, A., and Wagner, B. Risk of completed suicide after bariatric surgery: A systematic review. *Obesity Reviews*, 14(5):369–382, 2013.

Powell, L. A. Approximating variance of demographic parameters using the delta method: A reference for avian biologists. *The Condor*, 109(4):949–954, 2007.

Quenouille, M. H. Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45, pages 483–484. Cambridge University Press, 1949.

Rousseeuw, P. J. and Croux, C. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

Scholz, F. W. Maximum likelihood estimation. *Encyclopedia of Statistical Sciences*, 1985.

Shekar, M. and Popkin, B. *Obesity: Health and Economic Consequences of an Impending Global Challenge*. World Bank Publications, 2020.

Silverman, B. W., Chan, L., and Vincent, K. Bootstrapping multiple systems estimates to account for model selection. *Statistics and Computing*, 34(1):44, 2024.

Stangl, D. and Berry, D. A. *Meta-Analysis in Medicine and Health Policy*. CRC Press, 2000.

The Centre for Addiction and Mental Heatlh. Heroin. URL `https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/heroin`. Visited on 17/06/2024.

Tukey, J. W. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485, 1960.

van der Heijden, P. G. M., Bustami, R., Cruyff, M. J. L. F., Engbersen, G., and van Houwelingen, H. C. Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, 3(4):305–322, 2003.

Vongchak, T., Kawichai, S., Sherman, S., Celentano, D. D., Sirisanthana, T., Latkin, C., Wiboonnatakul, K., Srirak, N., Jittiwutikarn, J., and Aramrattana, A. The influence of Thailand's 2003 'war on drugs' policy on self-reported drug use among injection drug users in Chiang Mai, Thailand. *International Journal of Drug Policy*, 16 (2):115–121, 2005. . URL `https://www.sciencedirect.com/science/article/pii/S095539590400132X`.

Wagenmakers, E.-J. and Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11:192–196, 2004.

Windle, J. Drugs and Drug Policy in Thailand, Improving Global Drug Policy: Comparative Perspectives and UNGASS 2016. *Washington: Brookings Institute*, 2015.

Wittes, J. T. 331. Note: On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate. *Biometrics*, pages 592–597, 1972.

World Health Organisation. People who inject drugs. URL `https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/populations/people-who-inject-drugs`. Visited on 17/06/2024.

World Health Organisation. Suicide, 2021. URL `https://www.who.int/news-room/fact-sheets/detail/suicide`. Visited on 24/11/2021.

World Health Organization. Suicide Worldwide in 2019: Global Health Estimates. Technical report, World Health Organization, 2019.

Xekalaki, E. Under- and overdispersion. *Wiley StatsRef: Statistics Reference Online*, 2014.

Yee, T. W. On the Hauck–Donner Effect in Wald Tests: Detection, Tipping Points, and Parameter Space Characterization. *Journal of the American Statistical Association*, 117: 1763–177, 2022.

Zelterman, D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18(2): 225–237, 1988.

Zwane, E. N. and van der Heijden, P. G. M. Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & Probability Letters*, 65(2):121–125, 2003.