METRON
https://doi.org/10.1007/s40300-025-00300-2

Performance and robustness of single-source capture-recapture population size estimators with covariate information and potential one-inflation

Layna Dennett[1] · Dankmar Böhning[2]

Received: 17 March 2025 / Accepted: 29 November 2025

Abstract
Capture-recapture methods for estimating the total size of elusive populations are widely-used, however, due to the choice of estimator impacting upon the results and conclusions made, the question of performance of each estimator is raised. Motivated by an application of the estimators which allow covariate information to meta-analytic data focused on the prevalence rate of completed suicide after bariatric surgery, where studies with no completed suicides did not occur, this paper explores the performance of the estimators through use of a simulation study. The simulation study addresses the performance of the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators, and develops a novel, generalised, form of the modified Chao estimator to account for both covariate information and one-inflation. In addition, the performance of the analytical approach to variance computation is addressed. Given that the estimators vary in their dependence on distributional assumptions, additional simulations are utilised to address the question of the impact outliers have on performance and inference.

Keywords Horvitz–Thompson · Generalised Chao · Generalised Zelterman · Generalised modified Chao · One-inflation · Simulation study · Performance · Outliers

1 Introduction

Developed for use in ecology, capture-recapture methods are utilised for estimating the total size of elusive populations. An incomplete list of the individuals is used for the estimation, as given the nature of these populations, many individuals remain unobserved. Specifically

Layna Dennett and Dankmar Böhning have contributed equally to this work.

✉ Layna Dennett
l.c.dennett@soton.ac.uk

Dankmar Böhning
d.a.bohning@soton.ac.uk

[1] Southampton Clinical Trials Unit, Southampton General Hospital, University of Southampton, Tremona Road, Southampton, Hampshire SO16 6YD, UK

[2] Southampton Statistical Sciences Research Institute, University of Southampton, University Road, Southampton, Hampshire SO17 1BJ, UK

Published online: 18 December 2025

for animal populations within ecology, traps are placed in a designated study area to capture the animals, where those in the captured sample are uniquely marked and released. On further occasions, additional samples of the animals are taken, recording previously marked individuals and uniquely marking the unmarked individuals. A capture history for each of the individual animals observed at least once, used to estimate the total number of individuals in the population, is achieved by repeating this process a predetermined number of times or within a predetermined time period.

Capture-recapture methods have evolved to have utility in other fields, including those of the application of this paper, medicine and epidemiology. The focus of this paper is on count data with missing zeroes (since single-source capture-recapture studies lead to marginally zero-truncated data), motivated by Peterhänsel et al. [18], while the interest in one-inflation comes from the study by Jongsomjit et al. [14]. In addition, there are now several works that look at the measurement error associated with the covariates involved in the capture-recapture modelling (see [1, 9, 13] for examples). Here, we assume that covariates are measured without error which seems justified in the examples considered.

Firstly, Peterhänsel et al. [18] estimates the prevalence rate of completed suicide after bariatric surgery where studies that do not report any completed suicides are not included due to the search criteria. Studies without completed suicides could not be identified in the meta-analysis of Peterhänsel et al. [18] and correspond to zero counts in the context of single-source capture-recapture that refer to units that equally could not be identified in the capture-recapture method. Hence, and although the meta-analysis at hand is not a strict single-source capture-recapture analysis, both have in common that missing target populations need to be estimated. Single-source capture-recapture methodologies, as discussed in this paper, lead to a similar zero-truncated likelihood and, as such, can be used for this case study. Furthermore, in this setting, independent studies can be viewed as individual units of the target population of a capture-recapture study, and the observation period for each of the studies is comparable to the trapping period. In this sense, the count of completed suicides could be viewed as a proxy for the number of identifications of the study. As the counts are zero-truncated, the number of studies with zero counts remains unknown and therefore it is of interest to estimate. Table 13 in Appendix C contains the data from the 27 observed studies in Peterhänsel et al. [18], including the number of completed suicides, person-years, proportion of women and the country of origin of each study. When we move to the capture-recapture context, each study from the systematic review is treated as an ‘individual’.

The second focus, a capture-recapture study mentioned in Jongsomjit et al. [14], looks at the number of drug (heroin) users by age and gender in the Chiang Mai region of Thailand, with the information obtained through using the records of individuals who contacted the Thanyarak Chiang Mai hospital (and reporting how many times they contacted the hospital). However, given that it would be very unlikely for every single drug user within the region to contact the hospital, it is likely that many drug users will be unidentified by this study. Therefore, the data is zero-truncated as the number of heroin users which did not contact the hospital is unknown and is of interest to be estimated. Tables 14 and 15 in Appendix C contain the data from this study.

To estimate the total number of individuals, and consequently the number of missing individuals, a choice of which estimation method to use is required. However, the capture-recapture estimators approach data differently, resulting in estimates which can differ significantly from one another, impacting the accuracy and reliability of conclusions made.

Motivated by case studies, the aim and (summarised) results of this paper are as follows.

- A a novel capture-recapture estimator which accounts for both covariate information and one-inflation is developed. In particular, the generalised modified Chao estimator is developed in Sect. 4.2, which allows for both covariate information and one-inflation to be accounted for in the estimation process for more reliable and accurate estimates.
- In addition, a simulation study is used to compare the performance of capture-recapture estimators which allow for covariate information and for some violations of the underlying assumptions. The simulation studies are performed in Sect. 6, for both non-inflated and one-inflated datasets. When there are outliers in the data, results indicated that the more robust generalised Chao performs best, and if one-inflation is present in the data, the generalised modified Chao estimator is preferred.

Analysis of the performance of capture-recapture estimators which can incorporate covariate information and cope with additional one-inflation as well as assessing the performance of the variance formula are novel contributions, the conclusions of which can lead to more accurate population size estimates and reliable confidence intervals.

2 Motivating applications

In summary, capture-recapture is the methodology of estimating population sizes when some individuals within the population go unobserved. To illustrate this, we consider two case studies where the interest is on estimating the number of missing units. The first case study considers suicide data containing y_i, the count of completed suicide, with e_i, corresponding person-years, for each study $i = 1, \ldots, n$, where $n = 27$. For the random variable Y_i, it is assumed that $Y_1, \ldots, Y_n$ are independent. Given that the dataset also includes covariates, let $\mathbf{x} = (x_{i1}, x_{i2})^T$ be the vector denoting the covariates for the ith study, where x_{i1} is the proportion of women in study i and x_{i2} is the country of origin for study i, given as

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

Study 24. Smith (2004) does not include a value for the proportion of women, so is imputed as $x_{24,1} = 0.823$ using a linear regression imputation model, where the model is chosen from the full model with backwards stepwise Bayesian information criterion (BIC), resulting in the proportion of women as the response variable and both person-years and country of origin as main effects with their interaction. Additionally, for model fitting purposes, the country of origin of Study 21. Kral (1993) is changed from “USA/Sweden” to “USA” given that “USA” is both listed first, and is the country of origin with highest frequency of occurrence.

The second case study considers heroin users and contains y_i, the count of contact [to the treatment centre] for each individual $i = 1, \ldots, 843$. The heroin data also contains covariate information on both age and gender; however, the covariate information is not at the individual level and therefore they cannot be modelled together, and instead is treated as two separate datasets. Within the age dataset, let x_{i1} be the age range for the individual where

$$x_{i1} = \begin{cases} 1 & \text{if greater than or equal to 40 years old,} \\ 0 & \text{if less than 40 years old.} \end{cases}$$

As for the gender dataset, let x_{i2} be the gender of the individual where

$$x_{i2} = \begin{cases} 1 & \text{if female,} \\ 0 & \text{if male.} \end{cases}$$

Whilst many approaches can be taken within capture-recapture, this paper focuses on the methods of the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators (see [4, 6, 12]). Additionally, for instances where one-inflation is present and covariate information is available, a generalised modified Chao estimator is developed through generalising the modified Chao estimator (see [5]). These methods utilise an expected count value for the missing information computed from a regression model, and as a result, have the benefit of allowing for the inclusion of covariate information. Several assumptions are required in order to use these estimators. Firstly, the population is assumed to be a closed system, for example, no individuals enter or leave a study once started. The second assumption is of independence between individuals, which for studies within a systematic review is reasonable to assume given that to be included in the systematic review, independence between studies is required. Lastly, there is the assumption of independence between captures. For the suicide data, with the studies being treated as the individuals it is reasonable to assume that all three assumptions are met. The assumptions are more complex for the heroin data given that it is from a human population.

3 Review of population size estimators

Here we review recently developed capture-recapture estimators, focusing on those which allow for covariate information, starting with the Horvitz–Thompson estimator.

3.1 Horvitz–Thompson estimator

Proposed by Horvitz and Thompson [12], the Horvitz–Thompson estimator is a popular capture-recapture population size estimator (see [7], Chapter 11 and [16], Chapter 3). For a given regression function, **h** (see Table 11 in Appendix C for more information), with corresponding coefficients, $\boldsymbol{\beta}$, the total number of studies, N, is given by

$$\widehat{N}^{(HT)} = \sum_{i=1}^{N} \frac{I_i}{1 - P(Y = 0|\hat{\mu}_i)}, \tag{1}$$

where $\hat{\mu}_i = e_i \exp\left[\mathbf{h}(\mathbf{x}_i)^T \hat{\beta}\right]$ is the expected count of study i,[1] $\hat{\beta}$ is estimated from the model and I_i is an indicator variable defined as

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

In our context, for $y = 0, 1, \ldots$, the probability in (1) arises from either a Poisson model

$$P(Y = y) = \exp(-\mu)\frac{\mu^y}{y!},$$

or negative-binomial model

$$P(Y = y) = \frac{\Gamma(y + \alpha)}{\Gamma(y + 1)\Gamma(\alpha)} \left(\frac{\alpha}{\mu + \alpha}\right)^{\alpha} \left(\frac{\mu}{\mu + \alpha}\right)^{y}.$$

[1] Note that $\log(e_i)$ is an offset in the terminology of generalised linear models and e_i corresponds to the person-years of study i in the first example.

For assessing uncertainty, the analytical variance can be computed using the conditional approach proposed by van der Heijden et al. (see [11, page 314]), where the theoretical formula is as follows.

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = E[\text{Var}(\widehat{N}^{(HT)}|I_i)] + \text{Var}(E[\widehat{N}^{(HT)}|I_i]). \tag{2}$$

An approximation of the analytical variance is then given as

$$\widehat{\text{Var}}(\widehat{N}^{(HT)}) = \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \text{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) + \sum_{i=1}^{n} \frac{\exp(-\hat{\mu}_i)}{(1-\exp(-\hat{\mu}_i))^2}, \tag{3}$$

where

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

Whilst widely-used, the Horvitz–Thompson estimator relies heavily on the entire data following the given distributional assumption. As a result, if the counts do not strictly follow the distribution, for example, if the data contain outliers as is often the case for real life data, the accuracy of the resulting population size estimate and precision of confidence intervals can be negatively affected. In addition, larger count values are more susceptible to deviation from the given distribution. Therefore, it is beneficial to explore alternative population size estimators which do not experience these issues and are more resilient to outliers.

3.2 Generalised Chao estimator

Chao's lower bound [8] can be used as an alternative to the Horvitz–Thompson estimator, focusing on estimating the lower bound of the population size. It was developed as a method that incorporates unobserved heterogeneity with the flexible mixture probability density given by

$$k_y(\mu) = \int_0^\infty p_y(\mu)q(\mu)d\mu,$$

where $p_y(\mu) = \exp(-\mu)\mu^y/y!$ is the Poisson mixture kernel and $q(\mu)$ is the mixing density [6]. Whilst a Poisson mixture kernel is used here, a geometric mixture kernel can also be used, where $p_y(\mu) = (1-\mu)^y\mu$. The latter has been developed in Niwitpong et al. [17] and the extension to covariates in Alaimo di Loro et al. [15].

Through using the Cauchy–Schwarz inequality, theoretical probabilities and the corresponding sample estimates (f_y/N), Chao's lower bound estimator for the frequency of zero counts can be found as

$$\hat{f}_0 \geq \frac{f_1^2}{2f_2}.$$

This leads to the conventional Chao's estimator of $N^{(C)} = n + \frac{f_1^2}{2f_2}$, where f_y is the frequency of exactly $Y = y$ counts. However, this estimator does not allow for the inclusion of covariate information, and therefore does not allow for the incorporation of an exposure variable either. Böhning et al. [6] adapted this conventional Chao's estimator to allow for the inclusion of covariate information. The resulting generalised Chao estimator does so through regression modelling, resulting in more representative estimates.

Comparative to the Horvitz–Thompson estimator, this approach has a more relaxed distributional assumption requiring only two consecutive counts to follow the given distribution, rendering it more resilient to outliers. For zero-truncated data, the consecutive counts are typically assumed to be $Y = 1$ and $Y = 2$, with remaining count values truncated. The resulting truncated likelihood is equal to the standard binomial logistic likelihood, the maximum likelihood estimates of the expected counts as follows

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

where $\hat{q}_i$ are the fitted values of the logistic regression model for $i = 1, \ldots, M$, where $M = (f_1 + f_2)$ (please see details in Appendix A).

Using this maximum likelihood estimate, the generalised Chao population size estimate is given as

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

where f_{iy} is the frequency of exactly $Y = y$ counts for individual i.

If no covariate information is available, the generalised Chao estimator becomes the conventional Chao estimator (Chao's lower bound) as $\hat{\mu} = \frac{2f_2}{f_1}$ leading to

$$\begin{aligned}\widehat{N}^{(GC)} &= \frac{f_1 + f_2}{\frac{2f_2}{f_1} + (\frac{2f_2}{f_1})^2/2} \\ &= \frac{f_1 + f_2}{\frac{2f_2}{f_1} + \frac{2f_2^2}{f_1^2}} \\ &= \frac{f_1^2(f_1 + f_2)}{2f_2(f_1 + f_2)} \\ &= \frac{f_1^2}{2f_2} = \widehat{N}^{(C)}.\end{aligned}$$

As with the Horvitz–Thompson estimator, the theoretical formula in (2) proposed by van der Heijden et al. (see [11, page 314]) can be used to find the analytical variance. Böhning et al. [6] approximates this variance for the generalised Chao estimator to be

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{N}^{(GC)}) = &\left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_1+f_2} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) \\ &+ \sum_{i=1}^{f_1+f_2} (1 - \hat{q}_i) \left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{q}_i}\right)^2,\end{aligned} \tag{4}$$

where

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

3.3 Generalised Zelterman estimator

As with the conventional Chao's estimator, the conventional Zelterman estimator, developed by Zelterman [20], assumes that only a small range of count values follow the given distribution, improving its resilience to outliers. The expected count for each study is estimated as

$$\hat{\mu} = \frac{(k+1) f_{k+1}}{f_k}.$$

For zero-truncated data, typically $k = 1$ is assumed, since the missing frequency f_0 is close to the frequencies f_1 and f_2. In this case, all counts besides $Y = 1$ and $Y = 2$ are truncated, leading to the population size estimator $N^{(Z)} = n/(1-\exp(-\hat{\mu}))$. However, the conventional approach also does not allow for covariate information, motivating the truncated maximum likelihood estimate approach of the generalised Zelterman estimator developed by Böhning and van der Heijden [4]. As with the generalised Chao estimator, if no covariates are included in the modelling, the generalised Zelterman estimator is equal to the conventional Zelterman estimator.

Using the same binomial logistic likelihood as in Sect. 3.2, the binary outcome probability, q_i, is connected via a logit link to the linear predictor from the regression model and the expected count parameter respectively as

$$q_i = \frac{e_i \exp(\mathbf{h}(\mathbf{x}_i)^T \boldsymbol{\beta})}{1 + e_i \exp(\mathbf{h}(\mathbf{x}_i)^T \boldsymbol{\beta})}, \text{ and } q_i = \frac{\mu_i/2}{1+\mu_i/2}.$$

Therefore, the expected count can be estimated as $\hat{\mu}_i = 2e_i \exp(\mathbf{h}(\mathbf{x}_i)^T \hat{\boldsymbol{\beta}})$, for $i = 1, \ldots, n$.

Using this value of the parameter in the conventional Zelterman estimator, accounting for covariate information leads to the generalised Zelterman estimator, given formally as

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

for $i = 1, \ldots, n$.

The conditioning approach by van der Heijden et al. (see [11, page 314]) can also be applied to the generalised Zelterman estimator. Following the work of Böhning and van der Heijden [4], the analytical variance is approximated as

$$\begin{aligned} \widehat{\mathrm{Var}}(\widehat{N}^{(GZ)}) = & \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{n} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) \\ & + \sum_{i=1}^{n} \frac{\exp(-\mu_i)}{(1-\exp(-\mu_i))^2}, \end{aligned} \tag{5}$$

where

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

4 Novel estimator for one-inflation: the generalised modified Chao estimator

In some datasets, it can be seen that there are a high frequency of ones, or singletons, relative to the chosen base distribution or mixing kernel. When this occurs, the data is described as one-inflated, and these excess singletons need to be appropriately accounted for in the modelling and estimation processes to avoid grossly over-estimating the total population size. There is a growing focus on estimation with one-inflated data in this field (see the work of Tajuddin et al. [19] which looks at a new Horvitz–Thompson estimator, for example).

In this section, to estimate the total population size when there is one-inflation present, the singletons are truncated from the estimation processes. Whilst there is a possibility that ignoring the singletons in an estimation process can lead to reduced efficiency, including the excess singletons can lead to serious bias [5]. Additionally, in the case of one-inflation which is modelled by including a separate weight parameter for the singletons [2], there is an equivalence between one-inflation and one-truncation in the sense that maximum likelihood estimators for the non-inflated arm of the one-inflation model and the one-truncated model agree, as do their standard errors.

One-inflation (in addition to the zero-truncation) is accounted for in the modelling through utilising the one-inflated zero-truncated baseline distribution given below.

$$p(y;\mu)^{+1} = \begin{cases} (1-\omega) + \omega p(y;\mu)^{+} & \text{if } y = 1, \text{ and} \\ \omega p(y;\mu)^{+} & \text{if } y = 2, 3, \ldots, \end{cases}$$

where $\omega \in [0, 1]$ is the inflation parameter, or weight, and $p(y;\mu)^{+}$ is the zero-truncated baseline distribution [3], which in this case, is the zero-truncated Poisson distribution given below.

$$\begin{aligned} p_P(y;\mu)^{+} &= \frac{p_P(y;\mu)}{1 - p_P(0;\mu)} \\ &= \frac{\exp(-\mu)\mu^{y}}{y!(1-\exp(-\mu))}. \end{aligned}$$

4.1 Modified Chao estimator

The modified Chao estimator is an extension of the standard Chao's lower bound estimator, developed by Böhning et al. [5] to account for one-inflation within a dataset. This modification accounts for the one-inflation through using substitution to remove the singletons from the estimator, relying on only the doubletons and tripletons for estimating the total population size instead of the singletons and doubletons. Whilst the modified Chao estimator avoids the use of the frequency of counts of one, f_1, it still gives a lower bound for the frequency of zero counts, f_0 (the expected value). The estimated frequency of zero counts is given as

$$\hat{f}_0^{(MC)} = \frac{a_0 a_3^2}{a_2^3} \frac{f_2^3}{f_3^2},$$

where $a_x = \frac{1}{x!}$ if a Poisson distribution is assumed, or $a_x = 1$ if a geometric distribution is assumed. The resulting estimated total population size is then given as

$$\widehat{N}^{(MC)} = \begin{cases} n + \frac{2}{9}\frac{f_2^3}{f_3^2} & \text{if Poisson,} \\ n + \frac{f_2^3}{f_3^2} & \text{if geometric.} \end{cases}$$

4.2 Novel generalised-modified Chao estimator

Here we develop a novel capture-recapture estimator. Where the generalised Chao estimator deals with covariate information, we develop the generalised modified Chao estimator which accounts for covariate information and one-inflation. As with the standard Chao's lower bound estimator, the modified Chao estimator only utilises frequencies and therefore does not allow for the inclusion of covariate information. The generalisation of this estimator to account for covariate information is done through using the work of Böhning et al. [6], first by truncating all counts except the doubletons and tripletons, leading the the associated truncated Poisson model below:

$$p_2(\mu_i) = \frac{\exp(-\mu_i)\mu_i^2}{2} = (1 - q_i) \quad \text{and} \quad p_3(\mu_i) = \frac{\exp(-\mu_i)\mu_i^3}{6} = q_i. \tag{6}$$

The ratio of neighbouring probabilities, q_x, and the corresponding known coefficients, a_x, have the below relationship.

$$r_x = \frac{a_x}{a_{x+1}} \frac{q_{x+1}}{q_x} = \mu, \tag{7}$$

where $r_x \leq r_{x+1}$. Given that $q_2 = (1 - q)$ and $q_3 = q$, (7) becomes

$$\mu = \frac{a_2}{a_3} \frac{q}{1 - q} = 3 \frac{q}{1 - q},$$

which can be rearranged for $\hat{q}$ as follows.

$$\hat{q} = \frac{\mu}{3 + \mu},$$

if a Poisson distribution is assumed, leading to the probabilities in (6) being equal to

$$p_2(\mu_i) = \frac{3}{3 + \mu_i} = \text{ and } p_3(\mu_i) = \frac{\mu_i}{3 + \mu_i}.$$

The associated truncated Poisson likelihood is

$$L = \prod_{i=1}^{f_2+f_3} \left(\frac{3}{3 + \mu_i} \right)^{f_{i2}} \left(\frac{\mu_i}{3 + \mu_i} \right)^{f_{i3}},$$

which is identical to the standard binomial logistic likelihood

$$L = \prod_{i=1}^{f_2+f_3} (1 - q_i)^{f_{i2}} (q_i)^{f_{i3}},$$

and therefore, logistic regression analysis can be utilised to compute the maximum likelihood estimates for the truncated Poisson model.

The estimated frequency of zero counts for the ith covariate combination can be found with the following expectation.

$$\begin{aligned} \hat{f}_{i0} = E[f_{i0} | f_{i2}, f_{i3}, q] &= \frac{p_0(\hat{\mu}_i)}{p_2(\hat{\mu}_i) + p_3(\hat{\mu}_i)} (f_{i2} + f_{i3}) \\ &= \frac{1}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6} (f_{i2} + f_{i3}). \end{aligned}$$

The generalised modified Chao estimator is then the sum of the estimated frequency of zero counts across each of the covariate combinations, with the observed counts as follows:

$$\widehat{N}^{(GMC)} = n + \sum_{i=1}^{f_2+f_3} \frac{1}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6}.$$

It is worth noting that in the absence of covariate information in the modelling, as with the generalised Chao and classical Chao estimators, the generalised modified Chao will become the modified Chao estimator.

As with the other estimators, the conditioning approach van der Heijden et al. (see [11, page 314]) can be utilised for uncertainty quantification with the generalised modified Chao. Following the same method as Böhning et al. [6], the variance is given by

$$\widehat{\mathrm{Var}}(\widehat{N}^{(GMC)}) = E[\mathrm{Var}(\widehat{N}^{(GMC)}|I_i)] + \mathrm{Var}(E[\widehat{N}^{(GMC)}|I_i]), \tag{8}$$

where the first term is estimated to be

$$\widehat{\mathrm{Var}}(\widehat{N}^{(GMC)}|I_i) = \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right),$$

when a Poisson distribution is assumed, where for $\mu_i = \exp(\boldsymbol{h}(\boldsymbol{x}_i)^T \hat{\boldsymbol{\beta}}) e_i$,

$$\nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}}) = \frac{\hat{\mu}_i^2 + \hat{\mu}_i^3/2}{(\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6)^2} \boldsymbol{h}(\boldsymbol{x}_i)^T.$$

The second term in (8), the expectation can be given as

$$E[\widehat{N}^{(GMC)}|I_i] = E\left[n + \sum_{i=1}^{N} \frac{I_i}{\hat{\mu}_i^2/2 + \hat{\mu}_i^3/6} |I_i\right] \approx \sum_{i=1}^{N} I_i w_i,$$

where $w_i = 1 + p_0(\mu_i)/p_i$ 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 $E(I_i) = p_i$ and $\mathrm{Var}(I_i) = p_i(1 - p_i)$, hence

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

which can be estimated as

$$\begin{aligned}\widehat{\mathrm{Var}}(E[\widehat{N}^{(GMC)}|I_i]) &= \sum_{i=1}^{N} \frac{I_i}{p_i} p_i(1 - p_i) w_i^2 \\ &= \sum_{i=1}^{f_2+f_3} (1 - \hat{p}_i)\left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2.\end{aligned}$$

Therefore, (8) is given as

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{N}^{(GMC)}) = &\left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right)^T \mathrm{Cov}(\hat{\boldsymbol{\beta}}) \left(\sum_{i=1}^{f_2+f_3} \nabla G(\hat{\mu}_i|\hat{\boldsymbol{\beta}})\right) \\ &+ \sum_{i=1}^{f_2+f_3} (1 - \hat{p}_i)\left(1 + \frac{\exp(-\hat{\mu}_i)}{\hat{p}_i}\right)^2.\end{aligned} \tag{9}$$

Table 1 Values of the BIC statistic for models under consideration, where the Poisson and negative-binomial distributions model the full data and the binomial distribution models the truncated data

Distribution	Linear predictor	Log-likelihood	BIC
Poisson (full data)	**1**	**− 23.7**	**50.7**
	2	− 23.4	53.4
	3	−23.0	52.6
	4	− 23.0	55.9
	5	− 22.7	58.6
Negative-binomial (full data)	1	− 23.7	54.0
	2	− 23.4	56.7
	3	− 23.0	55.9
	4	− 23.0	59.2
	5	− 23.7	61.9
Binomial (truncated data)	**1**	**− 7.8**	**18.6**
	2	− 7.0	20.2
	3	− 7.8	21.6
	4	− 7.0	23.2
	5	− 5.7	23.5

Preferred models are indicated in bold text

5 Applications

5.1 Suicide data

Prior to the population size estimation with the capture-recapture estimators, it is important to explore whether there is one-inflation present given the large number of singletons in the dataset. If there is one-inflation present, the generalised modified Chao estimator should be used (the modified Chao estimator can also be used, but it is better to include covariate information when available and significant). Otherwise, either the Horvitz–Thompson, generalised Chao or generalised Zelterman estimators should be used.

To explore whether one-inflation is present in the data, both zero-truncated models and zero-truncated one-inflated models are fitted to the dataset, considering the various linear predictors given in Table 11 in Appendix C to account for the available covariate information. The best fitting model is selected using the BIC and if the selected model is a zero-truncated (non-one-inflated) model, then it is reasonable to assume that there is no one-inflation present in the data. For the Horvitz–Thompson estimator, the choice of distribution is between the Poisson and negative-binomial distributions, given the nature of count data. Both the generalised Chao and generalised Zelterman estimators do not have a choice of distribution, instead both require a binomial logistic regression model to be fitted.

Table 1 gives the log-likelihood and BIC statistic values for each of the linear predictors and distribution combinations under consideration for the zero-truncated (non-one-inflated) models. The Horvitz–Thompson estimator utilises the entire data available, and hence the Poisson and negative-binomial distributions model with the full data. However, the generalised Chao and generalised Zelterman estimators use the truncated data, containing only counts of $Y = 1$ and $Y = 2$, so the binomial distribution models using the truncated data. Since the data used for the binomial models differs from the other models, the results cannot

Table 2 Values of the BIC statistic for the zero-truncated one-inflated models under consideration

Distribution	Linear predictor	BIC
Poisson (full data)	**1**	**54.0**
	2	60.0
	3	59.2
	4	63.9
	5	67.8

The preferred model is indicated in bold text

be directly compared. Between the models within each distribution, there is little change in the log-likelihood values, and negligible difference between the Poisson and the negative-binomial distributions. Therefore, utilising BIC statistics for model selection, the preferred model to be used in the Horvitz–Thompson estimator is the intercept-only Poisson model, and for the generalised Chao and generalised Zelterman estimators, the intercept-only binomial model is preferred.

Given that the Poisson model is preferred for the full data in Table 1, for comparative purposes, the corresponding zero-truncated one-inflated Poisson models are fitted with the respective BIC values given in Table 2.

The intercept-only model is also preferred under the assumption of one-inflation, however, the BIC statistic is greater than that of the intercept-only zero-truncated Poisson model, and therefore there is not sufficient evidence to suggest that there is one-inflation present in the data and the generalised modified Chao estimator should not be utilised.

Utilising the preferred zero-truncated models in the application of the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators respectively, leads to the estimated total number of studies of $\widehat{N}^{(HT)} = 134$, $\widehat{N}^{(GC)} = 173$ and $\widehat{N}^{(GZ)} = 175$. It is to be expected that the generalised Chao is comparable to, but slightly lower than, the generalised Zelterman estimate given the similarity of the methods and that the generalised Chao is a lower bound estimator. However, they both differ largely from the Horvitz–Thompson estimate which is much smaller as a result of the difference in distributional assumptions and models used. The difference in population size estimates can lead to differing conclusions, and hence the estimator chosen can have an impact on the reliability of any conclusions made.

As for the uncertainty assessment for the estimators, using (3), (4) and (5), the variances for the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators are $\widehat{\text{Var}}(\widehat{N}^{(HT)}) = 1677$, $\widehat{\text{Var}}(\widehat{N}^{(GC)}) = 12707$ and $\widehat{\text{Var}}(\widehat{N}^{(GZ)}) = 13425$ respectively. The 95% confidence interval for each estimator is computed as

$$CI = \widehat{N} \pm 1.96\sqrt{\widehat{\text{Var}}(\widehat{N})}, \tag{10}$$

leading to the corresponding confidence intervals $CI^{(HT)} = (51, 214)$, $CI^{(GC)} = (27, 394)$ and $CI^{(GZ)} = (27, 402)$.

The latter two confidence intervals are twice the width of the interval from the Horvitz–Thompson estimator, indicating that there is considerably more uncertainty associated with using the generalised Chao and generalised Zelterman. This increased uncertainty is expected given that the observed number of studies is already small to estimate from, and the generalised Chao and generalised Zelterman estimators truncate the data further, estimating from an even smaller sample size; the more data available, the less uncertainty there is when estimating. Since the Horvitz–Thompson estimator utilises the entire data available, the uncertainty is reduced, leading to a smaller variance and narrower confidence interval which has a lower

Table 3 Simulated data with values of the number of completed suicides, person-years, proportion of women and country of origin of the studies, where the number of completed suicides are sampled from an alternative distribution to be outliers

Study	Number of completed suicides	Person-years	Proportion of women	Country of origin
i	y_i	e_i	x_{i1}	x_{i2}
28	14	1862	0.8371998	Other
29	17	2410	0.8087218	USA
30	16	1951	0.8430489	Other

limit greater than the lower bound for the total number of studies, making it a more reliable confidence interval to draw conclusions from.

However, the question of what happens if the model and distributional assumptions are not met remains, for example, how do outliers affect the results and inference. To address the question of the impact outliers have on the estimators and corresponding conclusions, outliers can be added to the case study data. The observed rates for each study, computed by dividing the number of completed suicides by the person-years, are used to find the lower bound for which rates are classed as outliers. Formally, the lower bound for a rate to be classified as an outlier is computed as

$$\lambda^L = Q3 + 3 \times IQR, \tag{11}$$

where $Q3$ is the third quartile and IQR is the inter-quartile range for the observed rate. To mimic the variability of rates between studies experienced in real life data, the outlier rates are sampled from a uniform distribution with a lower bound of λ^L and an upper bound, λ^U, computed as

$$\lambda^U = 1.2 \times \lambda^L. \tag{12}$$

To convert the outlier rates into counts, person-years is required, found by sampling the number of participants in each study from the Poisson distribution and observational period from the log-normal distribution and multiplying the respective values for each study. The sampled person-years multiplied by the outlier rates produces counts of completed suicide which are classified as outliers for the data. With only 27 observed studies, the addition of 3 studies with counts that are outliers leads to a proportion of 10% of the observed data being outliers, and a proportion of approximately 2% of the total data. Table 3 displays the values of the number of completed suicides, person-years, proportion of women and country of origin of the 3 additional studies. Utilising the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators respectively, the estimated total number of studies are $\widehat{N}^{(HT)} = 479717$, $\widehat{N}^{(GC)} = 176$ and $\widehat{N}^{(GZ)} = 180$, and corresponding 95% confidence intervals are $CI^{(HT)} = (30, 47285488)$, $CI^{(GC)} = (30, 397)$ and $CI^{(GZ)} = (30, 411)$.

There is no notable impact from outliers on the generalised Chao and generalised Zelterman estimators, with the estimates after outliers differing only slightly from the number of outliers studies included and the number of studies estimated from the data without outliers combined. However, the number of total studies found using the Horvitz–Thompson estimator is increased significantly after the inclusion of outliers to a number of studies which is inaccurate, with the corresponding confidence interval indicating a large quantity of uncertainty with its width.

Table 4 BIC values for various linear predictors for the zero-truncated and the zero-truncated one-inflated Poisson model

One-inflation	Linear predictor	Number of parameters	BIC
No (full data)	1	1	2234.07
	2	2	2206.69
	3	2	2240.71
Yes (Full data)	1	1	2057.16
	2	**2**	**2049.95**
	3	2	2070.46

As the results differ by such a large margin, it is important to know which of the estimators produces the most reliable results for the given data and hence the best conclusions. Given that the Horvitz–Thompson estimator is one of the more commonly used capture-recapture estimators, there is motivation to compare its performance in the presence of outliers to the more robust estimators. The generalised Chao and generalised Zelterman estimators do not depend on the higher order counts, so a comparison of these robust estimators to the bias that is shown in the Horvitz–Thompson estimators may encourage the use of more robust estimators. This shows the importance of looking at outlier information and will be done more systematically in Sect. 6.1 with a simulation study to compare the performance of the estimators. However, first we have our second case study application.

5.2 Heroin data

As with the Suicide data, there are also a lot of singletons in the Heroin data, and it is therefore important to explore whether there is one-inflation present.

Here we look at the Heroin data, where covariate information is given separately for age and gender. Given the formatting of this data, we model the covariate information separately for the two covariates. Therefore, only 3 possible linear predictors can be used, shown in Table 12 in Appendix C. To assess for possible one-inflation, both zero-truncated and zero-truncated one-inflated Poisson models with each of the three linear predictors will be fitted to the Heroin data, with the resulting BIC values compared in the same manner as for the Suicide data.

Given the results in Table 4, where the BIC values are lower for the models where one-inflation is assumed, there is evidence to suggest that there is one-inflation present in the Heroin dataset. Therefore, the generalised modified Chao estimator should be utilised in order to account for the additional singletons.

To utilise the generalised modified Chao estimator, first all counts besides the doubletons and tripletons should be truncated, then the binomial logistic regression models with each linear predictor can be fitted. Table 5 provides the BIC values for each of the competing models, whilst there is little difference between the BIC values, the value for the intercept-only model is slightly lower than the others and therefore, the intercept-only model is preferred. Although the preference for the intercept-only model means that the covariate information available does not improve the fit of the model and the resulting estimate, it cannot be known whether the covariate information is significant or not before analysing the data. Therefore, it is always recommended to use the covariate information available, since there is a possibility that it could lead to an improvement in the modelling.

Table 5 BIC statistics from the binomial logistic regression models fitted for each of the linear predictors under consideration for the generalised modified Chao estimator, where a Poisson mixture kernel is assumed and all counts are truncated *except the doubletons* $(Y = 2)$ *and tripletons* $(Y = 3)$

Linear predictor	BIC
1	**304**
2	310
3	309

Table 6 BIC statistics from the binomial logistic regression models fitted for each of the linear predictors under consideration for the generalised Chao estimator, where a Poisson mixture kernel is assumed and all counts are truncated *except the singletons* $(Y = 1)$ *and doubletons* $(Y = 2)$

Linear predictor	BIC
1	**734**
2	737
3	740

Using the intercept-only binomial logistic regression model to find the value of $\hat{\mu}_i$ for the generalised modified Chao leads to a population size estimate of $\widehat{N}^{(GMC)} = 965$. Given that no covariate information is included in this model, the result is the same as found through using the modified Chao estimator as follows.

$$\widehat{N}^{(MC)} = n + \frac{2}{9}\frac{f_2^3}{f_3^2} = 964.94 \approx 965.$$

As for the uncertainty assessment, the variance formula in (9) leads to $\text{Var}(\widehat{N}^{(GMC)}) = 1888$. Using the formula for constructing a confidence interval in (10), the corresponding 95% confidence interval is $CI^{(GMC)} = (880, 1050)$. This interval is of a reasonable width, with a lower limit greater than the observed number of individuals (843) suggesting that this is an appropriate interval.

To demonstrate the effect of accounting for the excess singletons, the generalised Chao estimator is fitted.

Table 6 give the BIC values for the three binomial logistic regression models under consideration, where it can be seen that there is little difference between each of the models, but the intercept-only model is chosen given that it has the lowest BIC value. Using the intercept-only binomial logistic regression model in the computation of the generalised Chao estimator results in an estimate of $\widehat{N}^{(GC)} = 2975.376 \approx 2975$, which is considerably greater than the population size when the excess singletons are appropriately accounted for, illustrating that if the appropriate estimator is not utilised, the results can be greatly impacted.

6 Simulation study

6.1 Design

Aims. The intention of the simulation work is to illustrate the two main issues raised in the manuscript, the issue of one-inflation and the issue of outliers. Both of these aspects have been studied for non-inflated data with different proportions of outliers as well as for one-inflated data with different proportions of outliers. Additionally, there is a strong focus in this work around the different forms of the Chao's estimator, due to its robustness properties. This includes the conventional form, generalised version for covariates, modified version for one-inflation and generalised modified version for both covariate information and one-inflation. This robustness is important as it means that the estimator only allows for the counts of ones and twos (or twos and threes in the case for one-inflation) to be correct, which in turn allows for deviations of higher order counts that may arise from model invalidity or simply by outliers, the latter being the focus of much of this work.

The BIC is used for model selection, and given that it is widely known to be model-consistent, the performance of the BIC is not the focus of the simulation study.

Non-one-inflated data. To create a data set where for each study $i = 1, \dots, N$, the values for counts, person-years and covariates are simulated to reflect the values in the case study, certain parameters require defining. For simulating person-years, the mean number of participants per study, $\bar{t}$, logarithm of the mean γ and standard deviation σ of the observational period are required. Using the sampled person-years, and a constant rate of event λ^C, the count values can be simulated. As the covariates, α and β are shape parameters for a beta distribution to simulate x_1, and the probability of success for x_2 is given by ρ. Using the predefined parameters, the number of participants in each study is sampled from the Poisson distribution, $t_i \sim Poisson(\bar{t})$, and the observational period for each study is sampled from the log-normal distribution, $O_i \sim lognormal(\gamma, \sigma)$, leading to the sampled exposure variable of person-years, $e_i = t_i \times O_i$. The count of events for each study is then sampled from the binomial distribution as $Y_i \sim binomial(e_i, \lambda^C)$. In this simulation study, we include covariate information indirectly through a varying offset (logarithm of person-years) for each study. Whilst it is not covariate information with an independent parameter to be estimated, as shown in the Peterhänsel et al. [18] case study, including person-years in the estimation plays an enormous role. Clearly, generalisations of the conventional Chao and Zelterman estimator are needed to incorporate this form of covariate information (covariate with fixed parameter). The sampling process is repeated S times, creating a zero-truncated data set for each iteration though removing all studies which have a count of $Y_i = 0$. To this data, the capture-recapture estimators and respective analytical variances discussed in Sect. 3 can be computed, producing population size estimates and respective confidence intervals.

One-inflated data. To simulate the zero-truncated one-inflated dataset, first a zero-truncated dataset is simulated using the same methods as given in Sect. ??. Once the dataset has been simulated, a fifth of the counts are randomly selected and set to equal 1.

Performance measures. To assess the performance of the estimators via the simulation studies, the following measures are looked at.

- Accuracy:
$$median(|\widehat{\boldsymbol{N}} - N|),$$
where $\widehat{\boldsymbol{N}} = (\widehat{N}_1, \dots, \widehat{N}_S)$ are the estimated population sizes from each iteration of the simulation study and N is the true population size.
- Precision:
$$median(\boldsymbol{CI}_U - \boldsymbol{CI}_L),$$
where $\boldsymbol{CI}_L = CI_{L,1}, \dots, CI_{L,S}$ and $\boldsymbol{CI}_U = CI_{U,1}, \dots, CI_{U,S}$ respectively are the lower and upper limits of the 95% confidence intervals for the estimated population size for each iteration of the simulation study.

- Coverage:

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

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

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

- Robustness: Defined as the resilience of the estimator to outliers.
  - In the simulation study, robustness is measured through comparing the values for accuracy, precision and coverage for data without outliers to values for data with outliers. To simulate the outlier counts, the person-years are multiplied by an outlier rate sampled from the uniform distribution, $\lambda_i^O \sim uniform(\lambda^L, \lambda^U)$, where the boundary values are chosen by an approximation of the results from (11) and (12) applied to the data being replicated. Given that the order of the studies does not impact the modelling results, the defined proportion of outliers are included at the end of the data.

6.2 Results

Non-one-inflated data. Summary of simulation study results:

- If no outliers are present, Horvitz–Thompson estimator performs the best overall.
- Horvitz–Thompson estimator is very sensitive to outliers.
- Generalised Chao and generalised Zelterman estimators very resilient to outliers.
- Generalised Chao and generalised Zelterman estimators perform similarly under the same circumstances.
- If outliers present, generalised Chao estimator performs the best overall.

Table 7 displays the values of accuracy, precision and coverage for the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators, and total number of studies is $N = 1000$. The performance measures are given for proportions of outliers varying from 0 to 2% to also assess robustness of each estimator. When the counts follow the distributional assumptions perfectly, the Horvitz–Thompson estimator is both the most accurate and the most precise, illustrated in Fig. 1 with the smallest inter-quartile and total ranges for both measures occurring for the Horvitz–Thompson estimator, and the corresponding plot for precision being closer to zero than the alternative estimators. Whilst the generalised Chao has the highest coverage, for all the estimators coverage is desirable at at least 95%, with negligible difference. However, as outliers are introduced to the data, the preference for the Horvitz–Thompson estimator becomes less obvious. Up to 0.5% of the counts being outliers, the Horvitz–Thompson estimator has the best performance for precision, however, once more outliers are introduced to the data, precision is dramatically reduced. Additionally, with as little as 0.1% outliers, the accuracy of the Horvitz–Thompson estimator is impacted, and coverage is significantly decreased, with only 70% of the confidence intervals containing the true value. As the proportion of outliers increases, the performance of the Horvitz–Thompson estimator worsens, with estimates getting further from the true value and confidence intervals getting wider from increased uncertainty. It appears that past a certain proportion of outliers, the coverage begins to improve, with coverage having an increase of 50% between 1%

Table 7 Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz–Thompson, generalised Chao and generalised Zelterman when there is no one-inflation present, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L = 0.007$, $\lambda^U = 0.009$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers

Measure	Estimator	Proportion of outliers				
		0.0%	0.1%	0.5%	1.0%	2.0%
Accuracy	Horvitz–Thompson	16	30	211	677	2.1e+06
	Generalised Chao	25	27	27	27	26
	Generalised Zelterman	29	32	31	32	32
Precision	Horvitz–Thompson	95	100	136	290	6.7e+07
	Generalised Chao	162	162	163	162	162
	Generalised Zelterman	181	181	185	184	187
Coverage	Horvitz–Thompson	95.5%	69.6%	7.0%	7.4%	60.6%
	Generalised Chao	96.4%	96.0%	96.4%	96.7%	95.6%
	Generalised Zelterman	95.7%	94.7%	95.8%	96.7%	94.8%

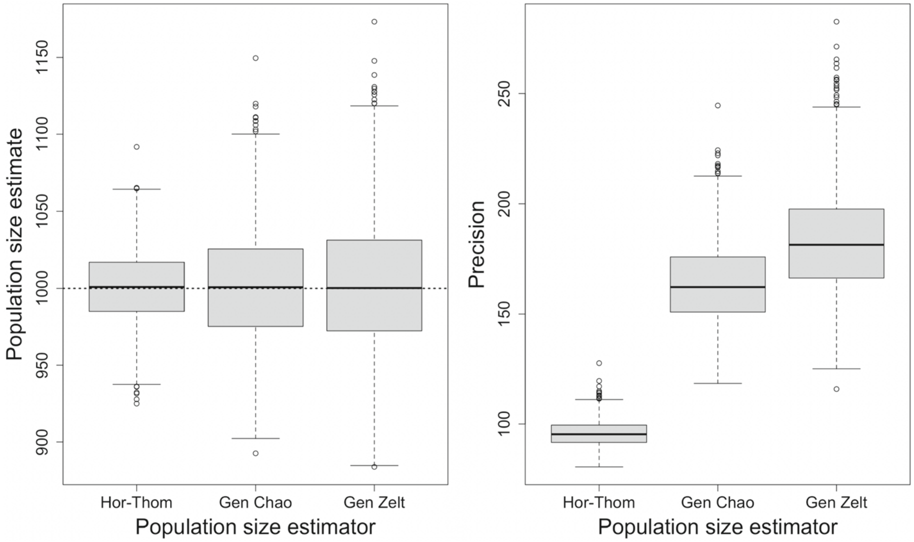

Fig. 1 Box plots showing the values of the population size estimates (left) and the values for the precision of the confidence intervals (right) for the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators when there are no outliers in the data and the dashed line represents the true population size of $N = 1000$

and 2% outliers, however, this is due to the width of the intervals growing, increasing the changes of the interval to contain the true value. Changes in the accuracy and precision of the population size estimates respectively are illustrated in Figs. 2 and 3, where the dispersion of the Horvitz–Thompson values increases as the proportion of outliers increases, and the median values grow farther from either the true population size or a reasonable width of confidence interval.

For completeness, Table 8 demonstrates the effect of outliers on the performance of the estimators when the total number of studies differs, specifically when $N = 500$. As with when

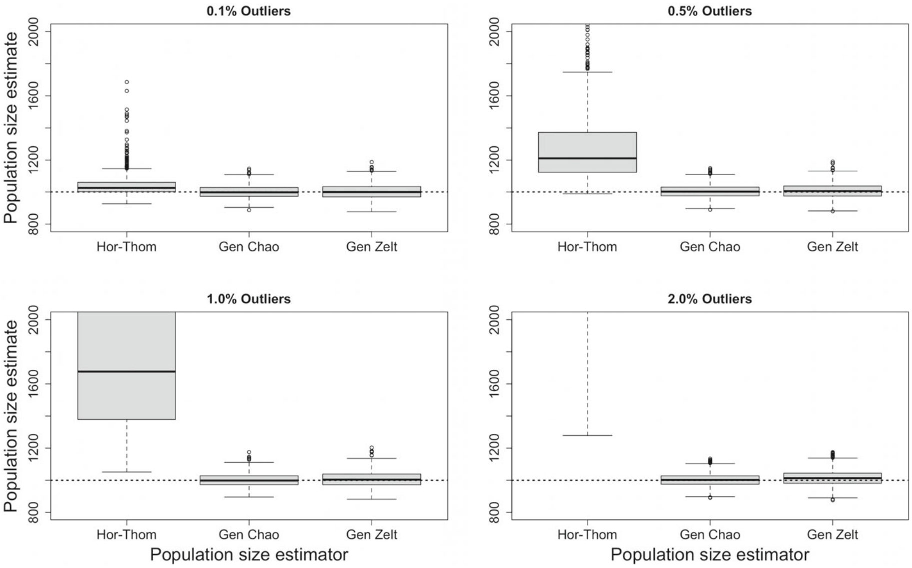

Fig. 2 Box plots showing the values of the population size estimates for the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators and varying proportions of outliers. The dashed line represents the true population size of $N = 1000$

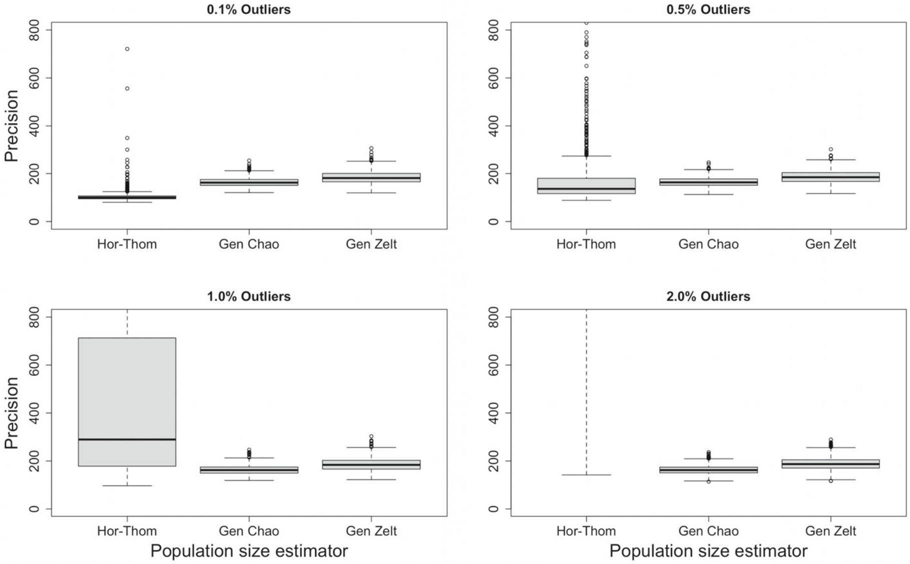

Fig. 3 Box plots showing the values of the precision from the 95% confidence intervals for the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators and varying proportions of outliers when the true population size is $N = 1000$

Table 8 Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz–Thompson, generalised Chao and generalised Zelterman when there is no one-inflation present, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L = 0.0004$, $\lambda^U = 0.0004$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers

Measure	Estimator	Proportion of outliers				
		0.0%	0.1%	0.5%	1.0%	2.0%
Accuracy	Horvitz–Thompson	11	–	62	293	4797
	Generalised Chao	19	–	19	18	19
	Generalised Zelterman	21	–	22	21	22
Precision	Horvitz–Thompson	67	–	83	181	10096
	Generalised Chao	116	–	116	115	113
	Generalised Zelterman	130	–	131	130	130
Coverage	Horvitz–Thompson	94.8%	–	34.8%	14.5%	51.1%
	Generalised Chao	96.9%	–	96.3%	96.7%	95.5%
	Generalised Zelterman	94.6%	–	96.7%	95.7%	95.0%

Number of outliers required to be integers so values for the proportion of 0.1% outliers are not given

$N = 1000$, when all data follows the distribution, the Horvitz–Thompson estimator performs the best, but the preference changes to the generalised Chao and generalised Zelterman estimators once outliers are introduced.

Throughout both tables, there is little difference between the performance for the generalised Chao and generalised Zelterman estimators, for both varying total sizes of data and proportions of outliers. The generalised Zelterman confidence intervals are on average closer to the nominal level, which is often preferred, whereas the generalised Chao estimator is on average more conservative. However, the generalised Chao estimator is consistently more accurate and precise, and given that the difference in coverage between the estimators is small, there is still the preference for the generalised Chao estimator. The main variation is as a result of data sizes, where for larger data sets, the generalised Chao estimator is more precise, and the generalised Zelterman estimator more precise for smaller data sets.

The values in Tables 7 and 8 suggest that it is the number of outliers rather than the proportion of outliers that impact the Horvitz–Thompson estimator's performance. For each proportion of outliers included respectively, comparing the performance of the Horvitz–Thompson estimator for $N = 1000$ and $N = 500$ indicate that the larger study size of $N = 1000$ impacts the estimator more, with a reduction in accuracy, coverage and precision. However, if the number of outliers is used as the comparison measure, rather than the proportion of outliers, the performance of the estimator is much more comparable. For example, when 5 outliers are included in the data, the proportion of outliers is 0.5% for $N = 1000$ and 1.0% for $N = 500$. For these proportions, the values of accuracy, precision and coverage respectively are more comparable and differ less from each other with the different population sizes than when a proportion of 0.5% outliers is used for $N = 500$. Given these results are replicated for the other proportions simulated, it is important for data to follow the distributional assumptions for the Horvitz–Thompson estimator to be used, as even for a very large population size, a very small number of outliers impact its performance.

Overall, the Horvitz–Thompson estimator performs better than the alternative estimators when the data follows the distributional assumption given. However, this is more often than not the case as a result of unpredictability within real life populations. Therefore, assumptions

Table 9 Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz–Thompson, generalised Chao, generalised Zelterman and generalised modified Chao when there is one-inflation present, where $S = 1000$, $N = 1000$, $\bar{t} = 900$, $\lambda^{C} = 0.0004$, $\lambda^{L} = 0.007$, $\lambda^{U} = 0.009$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers

Measure	Estimator	Proportion of outliers				
		0.0%	0.1%	0.5%	1.0%	2.0%
Accuracy	HT	370	81187662	95449602	115553849	130555650
	GC	662	797	789	794	784
	GZ	749	1217	1218	1226	1234
	GMC	46	226	233	214	222
Precision	HT	189	5328655340	5843597127	6373528861	6125530967
	GC	464	542	536	542	538
	GZ	538	773	789	777	781
	GMC	279	621	633	611	608
Coverage	HT	0%	99.0%	99.9%	100.0%	99.4%
	GC	0%	0%	0%	0%	0%
	GZ	0%	0%	0%	0%	0%
	GMC	93.8%	93.3%	93.5%	93.9%	94.2%

are not always met and in the presence of outliers, the generalised Chao and generalised Zelterman estimators are the preferred estimator, given they are more robust.

One-inflated data. Summary of simulation study results:

- Generalised modified Chao estimator always performs the best when the data is one-inflated, whether outliers are present or not.
- Generalised modified Chao estimator is resilient to outliers.
- Generalised Chao and generalised Zelterman estimators remain unaffected by outliers, but estimates are not accurate and very poor coverage.
- If outliers are present, Horvitz–Thompson has very inaccurate estimates and misleading coverage results due to very wide confidence intervals.

As with the simulation study for non-one-inflated data, Table 9 provides the values of the performance measures, with robustness of the estimators demonstrated by the results also given for varying proportions of outliers in the simulated datasets. For this simulation study, given that the data is one-inflated (with approximately 1/5th of the counts being excess singletons), the generalised modified Chao estimator is utilised, in addition to the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators for comparative purposes.

For 0% outliers, each of the estimators are appropriate given the accordance to the desired levels of accuracy and precision, the coverage of the Horvitz–Thompson, generalised Chao and generalised Zelterman estimators is at 0%, meaning that none of the confidence intervals constructed contain the true values and therefore are not appropriate. As with the non-one-inflated data, the generalised Chao and generalised Zelterman estimators perform consistently throughout each of the varying proportion of outliers. However, whilst these estimators are consistent, they are not appropriate estimators for one-inflated data, given the very low coverage of the corresponding confidence intervals. Additionally, the Horvitz–Thompson estimator is not an appropriate estimator in the case of one-inflation. Similarly to the simulation study with non-one-inflated data, outliers are introduced to the data, the coverage of the

Table 10 Values for the reliability measures of accuracy, precision and coverage for the capture-recapture population size estimators of Horvitz–Thompson, generalised Chao, generalised Zelterman and generalised-modified Chao when there is one-inflation present, where $S = 1000$, $N = 500$, $\bar{t} = 900$, $\lambda^C = 0.0004$, $\lambda^L = 0.007$, $\lambda^U = 0.009$, $\gamma = 1.5$, $\sigma = 0.8$, $\alpha = 36$, $\beta = 8.5$ and $\rho = 0.4$ for various proportions of outliers

Measure	Estimator	Proportion of outliers				
		0.0%	0.1%	0.5%	1.0%	2.0%
Accuracy	HT	184	–	27784376	29809669	31229695
	GC	333	–	395	394	389
	GZ	377	–	609	614	618
	GMC	33	–	123	114	115
Precision	HT	134	–	2004076150	1814267321	1367398644
	GC	330	–	381	382	379
	GZ	382	–	546	550	553
	GMC	194	–	446	436	433
Coverage	HT	0%	–	99.0%	99.1%	99.9%
	GC	0%	–	0%	0%	0%
	GZ	0%	–	0%	0%	0%
	GMC	93.3%	–	99.8%	99.6%	99.8%

Horvitz–Thompson estimator gets very close to 100%. This result is misleading given that the population size estimates are very inaccurate and the only reason the coverage is so high, is due to the confidence intervals being very wide (poor precision).

Given that the generalised modified Chao estimator has similar relaxed distributional assumptions to the generalised Chao and generalised Zelterman estimators, throughout each of the varying proportions of outliers, it has relatively consistent performance, leading to it being a robust estimator. The generalised modified Chao estimator performs well across each of the measures, in particular the coverage, which is over 90% for each of the proportions of outliers. The results from this simulation study indicate that the generalised modified Chao not only performs well in the case where one-inflation is present, but performs better than the other capture-recapture estimators. This outcome is expected since the other capture-recapture estimators explored in the simulation study are not able to appropriately account for excess singletons in a dataset.

For completeness, Table 10 demonstrates the effect of the varying size of the population on the performance of the estimators, by changing the total population size from $N = 1000$ to $N = 500$. The results of this simulation study follow the same trends as in Table 9, where the Horvitz–Thompson estimator is not resilient to outliers and performs poorly with the one-inflated data. Additionally, whilst the generalised Chao and generalised Zelterman estimators are robust, with reasonable widths of confidence intervals, they are not accurate and have poor coverage of the resulting confidence intervals. Whilst an accuracy measure of 333 (for the generalised Chao estimator with no outliers) may not seem overly poor in a dataset with a large population size, when the total number of studies are only $N = 500$, being approximately 333 studies incorrect either way results in very incorrect estimates. The conclusion of this simulation study is the same as for when the population size is larger, that the generalised modified Chao estimator not only performs well in the instance where there are excess singletons in the data, but it also performs considerably better than the other capture-recapture estimators explored.

7 Discussion

This paper develops a capture-recapture estimator that accounts for both one-inflation and covariate information and explores the performance of different capture-recapture population size estimators when dealing with zero-truncated meta-analytic count data in the case of one-inflated data and non-one-inflated data, through utilising simulation studies. A benefit of this approach is the flexibility enabled when creating the data sets, allowing for different data scenarios to be applied and covariate information included to test the performance of the estimators more thoroughly.

For the simulation studies in this manuscript, outliers were inserted at the higher order counts. Neither the Chao or modified Chao estimators are affected by these higher order count outliers, but all typical full data models are. An explanation behind this choice of outlier counts is given in Appendix B.

The results from the simulation study for non-one-inflated data indicate a preference for the Horvitz–Thompson estimator only if the data does not contain outliers. Given that within real life data, it is a common occurrence for outliers to be included, and even if it is only a small proportion, the Horvitz–Thompson estimator is not the most reliable. Between the generalised Chao and generalised Zelterman estimators, there is very little difference in performance, with the reliability measures unaffected by outliers, demonstrated by the consistent desirable coverage in addition to appropriate accuracy and precision irrespective of the proportion of outliers. The negligible difference in performance means that either estimator is appropriate and would return reliable results, but specifically for larger data, the generalised Chao is favoured, and the generalised Zelterman favoured for smaller data.

As for the simulation study for one-inflated data, the results indicate that whether there are outliers in the data or not, the generalised modified Chao estimator is preferable, given that the alternative estimators are inaccurate with poor coverage of the resulting confidence intervals. This preference for the generalised modified Chao is expected given that the other capture-recapture estimators are not able to appropriately account for these excess singletons, leading to over-estimation of the population size and confidence intervals which do not contain the true value.

For future work, additional data structures could be explored, such as data with different covariate variable types or sampling distributions assumed, to examine the estimators' performance in a wider range of scenarios. It could also prove beneficial to explore their performance using alternative confidence interval construction methods like the bootstrap algorithm and the percentile method, given that the analytical approach taken in this paper does not produce appropriate intervals for the small number of studies from the case study used with the generalised Chao and generalised Zelterman estimators. Lastly, the estimators discussed in this paper are not the only capture-recapture estimators available so the performance of a wider range of estimators could be explored. Examples of further estimators include the Turing estimator [10] and conventional Chao and Zelterman estimators discussed in Sect. 3. These estimators are not appropriate for the work in this paper since they don't allow for the inclusion of covariate information, however, for data without covariate information or exposure variables, there may be value in exploring their performance.

Appendix A: The truncated Poisson likelihood

This is largely following [6]. For $i = 1, \ldots, M$ where $M = f_1 + f_2$ is the number of studies with one or two events, let

$$\mu_i = e_i \exp(\mathbf{h}^T(\mathbf{x}_i)\boldsymbol{\beta}) = e_i \exp(\beta_0 + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)$$

where e_i is the value of person-years corresponding to study i, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^*)$ and $\mathbf{h}^*$ is equal to $\mathbf{h}$ without the intercept.

The Poisson likelihood truncated for all counts except ones and twos is

$$\begin{aligned}
&\prod_{i=1}^{M} \left(\frac{1}{1+\mu_i/2}\right)^{f_{i1}} \left(\frac{\mu_i/2}{1+\mu_i/2}\right)^{f_{i2}}, \\
&= \prod_{i=1}^{M} \left(\frac{1}{1+e_i \exp(\beta_0 + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}\right)^{f_{i1}} \left(\frac{e_i \exp(\beta_0 + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}{1+e_i \exp(\beta_0 + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}\right)^{f_{i2}}, \\
&= \prod_{i=1}^{M} (1-q_i)^{f_{i1}} q_i^{f_{i2}},
\end{aligned}$$

hence, $q_i = \dfrac{1}{1+\mu_i/2}$.

This is a conventional binomial logistic likelihood and can be further written as

$$\prod_{i=1}^{M} \left(\frac{1}{1+e_i \exp(\beta_0^{'} + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}\right)^{f_{i1}} \left(\frac{e_i \exp(\beta_0^{'} + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}{1+e_i \exp(\beta_0^{'} + \mathbf{h}^{*T}(\mathbf{x}_i)\boldsymbol{\beta}^*)}\right)^{f_{i2}},$$

with $\beta_0^{'} = \log(1/2) + \beta_0$

Once the binomial logistic likelihood has been fitted one can compute

$$\hat{\mu}_i = 2\frac{\hat{q}_i}{1-\hat{q}_i} = 2e_i \exp(\hat{\beta}_0^{'} + \mathbf{h}^{*T}(\mathbf{x}_i)\hat{\boldsymbol{\beta}}^*).$$

Note that $f_{i1} + f_{i2} = 1$ in our case as each study i has either a count of one or a count of two, given it is a truncated study where all counts are truncated except ones and twos.

Appendix B: Simulation study outliers

For the simulation studies in this manuscript, outliers are only at the higher order counts. If outliers were added at the count of one, then the data becomes one-inflated and the modified Chao or generalised modified Chao estimators can be utilised. There is a rationale for the existence of one inflation, namely, that there is a behavioural change after the first identification. There is no such obvious rationale for inflated counts of two and three, however, for completeness, the potential for inflation at these counts is explored below.

Assume that the counts of 2 and 3 are inflated y some common factor c. Given that the probability of a 2 is $p_2 = c\exp(-\lambda)\lambda^2/2$ and the probability of a 3 is $p_3 = c\exp(-\lambda)\lambda^3/6$, we can estimate that $\lambda = 3p_3/2p_2$ since the normalising constant drops out. Hence, the modified Chao estimator would not be affected by this kind of inflation.

Appendix C: Tables

See Tables 11, 12, 13, 14, and 15.

Table 11 Linear predictors for models under consideration for the Suicide data

Linear predictor	Proportion of women, x_1	Country of origin, x_2	Interaction x_1x_2	$\boldsymbol{h}(\boldsymbol{x})$
1	No	No	No	$\boldsymbol{h}_1(\boldsymbol{x}) = 1$
2	Yes	No	No	$\boldsymbol{h}_2(\boldsymbol{x}) = (1, x_1)^T$
3	No	Yes	No	$\boldsymbol{h}_3(\boldsymbol{x}) = (1, x_2)^T$
4	Yes	Yes	No	$\boldsymbol{h}_4(\boldsymbol{x}) = (1, x_1, x_2)^T$
5	Yes	Yes	Yes	$\boldsymbol{h}_5(\boldsymbol{x}) = (1, x_1, x_2, x_1x_2)^T$

Table 12 Linear predictors for models under consideration for the Heroin data

Linear predictor	Age, x_1	Gender, x_2	$\boldsymbol{h}(\boldsymbol{x})$
1	No	No	$\mathbf{h}_1(\mathbf{v}) = 1$
2	Yes	No	$\mathbf{h}_2(\mathbf{v}) = (1, v_1)^T$
3	No	Yes	$\mathbf{h}_3(\mathbf{v}) = (1, v_2)^T$

Table 13 Meta-analytic data from Peterhänsel et al. [18], numbered and ordered by decreasing size of person-years

Study	Number of completed suicides	Person-years	Proportion of women	Country of origin
i	y_i	e_i	x_{i1}	x_{i2}
1. Adams 2007	21	77602	0.860	USA
2. Marceau 2007	6	10388	0.720	Canada
3. Marsk 2010	4	8877	0.000	Sweden
4. Pories 1995	3	8316	0.832	USA
5. Carelli 2010	1	6057	0.684	USA
6. Busetto 2007	1	4598	0.753	Italy
7. Smith 1995	2	3882	0.889	USA
8. Peeters 2007	1	3478	0.770	Australia
9. Christou 2006	2	2599	0.820	Canada
10. Günther 2006	1	2244	0.837	Germany
11. Capella 1996	3	2237	0.822	USA
12. Suter 2011	3	2152	0.744	Switzerland
13. Suter 2006	1	1639	0.865	Switzerland

Table 13 continued

Study i	Number of completed suicides y_i	Person-years e_i	Proportion of women x_{i1}	Country of origin x_{i2}
14. Van de Weijgert 1999	1	1634	0.870	Netherlands
15. Cadière 2011	1	1362	0.834	Belgium
16. Mitchell 2001	1	1121	0.847	USA
17. Himpens 2011	1	1066	0.902	Belgium
18. Näslund 1994	2	799	0.812	Sweden
19. Forsell 1999	1	761	0.761	Sweden
20. Powers 1997	1	747	0.847	USA
21. Kral 1993	1	477	0.812	*USA*
22. Näslund 1995	1	457	0.592	Sweden
23. Powers 1992	1	395	0.850	USA
24. Smith 2004	1	354	*0.823*	USA
25. Nocca 2008	1	228	0.677	France
26. Svenheden 1997	1	166	0.791	Sweden
27. Pekkarinen 1994	1	146	0.704	Finland

The table includes the number of person-years, the proportion of women, the country of origin 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

Table 14 Capture-recapture data in Jongsomjit et al. [14] of the 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 15 Capture-recapture data in Jongsomjit et al. [14] of the distribution of counts of heroin users in Chiang Mai, Thailand by gender

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}
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

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40300-025-00300-2.

Declarations

Conflict of interest Not applicable.

References

1. Alaimo Di Loro, P., Maruotti, A.: A semi-parametric maximum-likelihood analysis of measurement error in population size estimation. J. R. Stat. Soc. Ser. C Appl. Stat. **73**(5), 1310–1332 (2024)
2. Böhning, D.: On the equivalence of one-inflated zero-truncated and zero-truncated one-inflated count data likelihoods. Biom. J. **65**(2), 2100343 (2023)
3. Böhning, D., 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. Stat. Methods Appl. **30**(4), 1197–1217 (2021)
4. Böhning, D., Heijden, P.G.: A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. Ann. Appl. Stat. **3**(2), 595–610 (2009)
5. Böhning, D., Kaskasamkul, P., Heijden, P.G.: A modification of Chao's lower bound estimator in the case of one-inflation. Metrika **82**(3), 361–384 (2019)
6. Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., Arnold, M.: A generalization of Chao's estimator for covariate information. Biometrics **69**(4), 1033–1042 (2013)
7. Borchers, D.L., Buckland, S.T., Zucchini, W., Borchers, D.: Estimating animal abundance: closed populations (2002)
8. Chao, A.: Estimating the population size for capture-recapture data with unequal catchability. Biometrics 783–791 (1987)
9. Dotto, F., Farcomeni, A.: A generalized Chao estimator with measurement error and external information. Environ. Ecol. Stat. **25**(1), 53–69 (2018)
10. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**(3–4), 237–264 (1953)
11. Heijden, P.G.M., Bustami, R., Cruyff, M.J.L.F., Engbersen, G., Van Houwelingen, H.C.: Point and interval estimation of the population size using the truncated Poisson regression model. Stat. Model. **3**(4), 305–322 (2003)
12. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**(260), 663–685 (1952)
13. Hwang, W.-H., Stoklosa, J., Wang, C.-Y.: Population size estimation using zero-truncated Poisson regression with measurement error. J. Agric. Biol. Environ. Stat. **27**(2), 303–320 (2022)
14. Jongsomjit, T., Lerdsuwansri, R., et al.: Estimation of population size based on one-inflated, zero-truncated count distribution with covariate information. PhD thesis, Thammasat University (2023)
15. Loro, P., Dotto, F., Maruotti, A.: On the robustness of the Chao's estimator with covariate information and measurement error for population size estimation. Statistics 1–16 (2025)
16. McCrea, R.S., Morgan, B.J.T.: Analysis of capture-recapture data (2014)
17. Niwitpong, S., Böhning, D., Heijden, P.G., Holling, H.: Capture-recapture estimation based upon the geometric distribution allowing for heterogeneity. Metrika **76**(4), 495–519 (2013)
18. Peterhänsel, C., Petroff, D., Klinitzke, G., Kersting, A., Wagner, B.: Risk of completed suicide after bariatric surgery: a systematic review. Obes. Rev. **14**(5), 369–382 (2013)
19. Tajuddin, R.R.M., Ismail, N., Ibrahim, K.: Estimating population size of criminals: a new Horvitz–Thompson estimator under one-inflated positive Poisson–Lindley model. Crime Delinq. **68**(6–7), 1004–1034 (2022)
20. Zelterman, D.: Robust estimation in truncated discrete distributions with application to capture-recapture experiments. J. Stat. Plan. Inference **18**(2), 225–237 (1988)