Identifying one-inflation using ratio regression in capture-recapture problems

Rattana Lerdsuwansri, Parawan Pijitrattana, Patarawan Sangnawakij
Thammasat University, Pathum Thani, Thailand
Krisana Lanumteang
Maejo University, Chiang Mai, Thailand
Antonello Maruotti
LUMSA, Rome, Italy
Herwig Friedl
Technical University Graz, Graz, Austria
and
Dankmar Böhning *University of Southampton, Southampton, UK*

June 18, 2025

Abstract

Ratio regression has been developed as a flexible instrument to allow for a wide class of count data distributions. In particular, it turns out to be useful in zero-truncated count distributions as they typically arise in capture-recapture settings. One-inflation describes the occurrence of extra-ones relative to a base count distribution and is a phenomenon frequently occurring in capture-recapture studies caused, for example, by behavioral change. The work presented here shows how one-inflation can be incorporated into ratio regression modeling and how one-inflation can be assessed in ratio regression modeling. Population size estimation on the basis of ratio regression is discussed and applied to a case study on heroin users in Chiang Mai (Thailand). For all model-based estimators computational inference is developed by means of the bootstrap. As several versions of the bootstrap are possible, a simulation study is included

comparing the different approaches. One of the main results shows that integrating the model selection into the population size inference leads to favourable properties such as good coverage probabilities.

Key words: Zero-truncated count distribution, Count inflation, Semi-parametric estimator of population size, Nonparametric bootstrap method, Model selection

1 Introduction

Ratio regression is a powerful tool to allow a broader modeling approach to count data distributions. It has been developed in series of works motivated by zero-truncated count data as they typically arise in marginal capture-recapture sampling frames (Böhning *et al.* 2013, Böhning 2016, Böhning *et al.* 2016, Böhning *et al.* 2023). For a general introduction into the capture-recapture methodology, see McCrea and Morgan (2015) or Böhning, Bunge, and van der Heijden (2018). Recently, one-inflation has found considerable interest in the area of count modeling (Böhning *et al.* 2019, Böhning and Ogden 2021, Böhning and Friedl 2024, Chiu and Chao 2016, Godwin 2017, Godwin and Böhning 2017, Godwin 2019, Tuoto *et al.* 2022), and this paper is about incorporating potential one-inflation into the ratio regression modeling.

Before we start introducing the basic methodology, we illustrate in elementary terms which problems arise in the case of one-inflation for zero-truncated count data. As a simple and hypothetical example consider 500 counts sampled from a Poisson distribution with mean 2 and 500 extra counts of 1 so that the total population size is $N = 1000$. The entire frequency distribution is shown in Table 1, where f_x denotes the frequency of count x. Suppose we ignore

Table 1: One-inflated Poisson data

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
74	626	130	108	33	19	8	1	1	926

knowledge of f_0, in other words, we are treating the sample as zero truncated with observed sample size $n = 926$. We can find an estimate of N using the EM algorithm (Dempster, Laird and Rubin 1977) of $\hat{N} = 1434$ which is about 1.5 times the true size of $N = 1000$. This overestimation issue becomes even more pronounced when using the estimator of Chao (1987) (see also Chao and Colwell 2017), given by $\hat{f}_0 = f_1^2/(2f_2)$, which has been developed to adjust for population heterogeneity in the Poisson parameter. The Chao estimator has a

lower bound property in the sense that the estimate provides a lower bound in expectation if the Poisson parameter follows an arbitrary random effects distribution. However, one-inflation models do not belong to this class of unobserved heterogeneity models and the estimator loses its lower bound property. In the example, it can be seen that $\hat{f}_0 = 1507$, so that $\hat{N}_C = n + \hat{f}_0 = 2433$, considerably overestimating the true N of 1000. Hence, one-inflation carries the risk of overestimation the true population size, potentially quite strongly.

The current work has focused on one-inflation and ratio regression and has the following key points of novelty:

- It shows how one-inflation can be incorporated into ratio regression modeling.
- It demonstrates how one-inflation can be assessed in the general context of ratio regression.
- It shows how population size estimation can be achieved using ratio regression with and without one-inflation.
- It illustrates the concepts with a case study on heroin users in Chiang Mai (Thailand).
- It discusses how model selection can be built into the process of population size inference.
- It also provides a new semi-parametric estimator of population size which is built on the ratio regression modeling with and without one-inflation.
- A simulation study is provided comparing these proposed approaches.

In the next section, we describe ratio regression and illustrate which risks are involved in ratio regression when one-inflation is ignored.

2 Count distribution and ratio regression modeling

We consider a count random variable X taking values $x \in \{0, 1, \cdots, m\}$. Here, m is a positive integer or $m = \infty$, depending on the setting. Let p_x denote the associated probability mass function $P(X = x) = p_x$ for which we seek

an appropriate model. A key idea is that it is frequently easier to develop an appropriate model for p_x if we consider ratios of neighboring probabilities

$$R_x = \frac{p_{x+1}}{p_x} \tag{1}$$

for $x = 0, 1, \cdots, m-1$. If $p_x = \exp(-\theta)\theta^x/x!$ is the Poisson distribution (where $\theta > 0$) then $R_x = \theta/(x+1)$. If $p_x = \theta(1-\theta)^x$ (where $\theta \in (0,1)$) is the geometric distribution then $R_x = 1-\theta$. Note that the ratio of successive probabilities is particularly suitable for zero-truncated distributions as the zero-truncated ratio

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

and the untruncated ratio (1) are identical. It is also suitable for distributional families where the normalizing constant is more difficult to compute as it cancels out in the ratio. To illustrate as an example, we consider the two-parameter Conway-Maxwell-Poisson (COM) distribution with the probability mass function as

$$p_x = \frac{\mu^x/(x!)^\lambda}{c(\mu,\lambda)}, \tag{2}$$

where $c(\theta)$ is the normalizing constant defined by $c(\theta) = c(\mu,\lambda) = \sum_{j=0}^{\infty} \mu^j/(j!)^\lambda$ for both μ and λ are positive, or $\mu \in (0,1)$ for $\lambda = 0$. The COM distribution contains some well-known discrete distributions. For $\lambda = 1$ the COM distribution simply reduces to the Poisson(μ) and for $\lambda = 0$ it is the geometric distribution. More details on the COM distribution including an illustration of its flexibility are given in Sellers (2023). The corresponding ratios for the COM distribution are given by

$$R_x = \frac{\mu}{(x+1)^\lambda}. \tag{3}$$

Here we see a benefit of moving to ratios as we reach a simplified model where the normalizing constant has cancelled out. Taking logarithms on both sides of (3), we achieve

$$\log R_x = \log\mu - \lambda\log(x+1) = \beta_0 + \beta_1\log(x+1). \tag{4}$$

It is convenient to think of (4) as regression of R_x on $\log(x+1)$ using a log-link function so-called ratio regression. Then, $\log\mu$ and λ correspond to the intercept and the slope, respectively. The geometric distribution is characterized by a slope of zero, whereas the Poisson distribution has a fixed negative slope

of -1. The COM distribution has an arbitrary intercept and arbitrary negative slope. From (4), we have $\mu = \exp(\beta_0)$ and no restriction on β_0 as $\mu > 0$ implies $\beta_0 \in (-\infty, \infty)$. However, we must constrain $\beta_1 < 0$ due to $\lambda > 0$ if we would like to retain a valid COM distribution. The fundamentals of ratio regression have been developed in Böhning, Baksh, Lerdsuwansri, and Gallagher (2013) as well as in Böhning (2016).

Given a sample $X_1, X_2, \cdots, X_n$ of size n, we can estimate R_x by $r_x = f_{x+1}/f_x$, where f_y is the frequency of sample elements X_i equal to y. However, the aforementioned concept is far more general as we are able to consider more general models

$$\log r_x = \beta_0 + \beta_1 g_1(\log(x+1)) + \cdots + \beta_p g_p(\log(x+1)) + \epsilon_x, \tag{5}$$

where $g_j(.)$ are known functions for $j = 1, 2, \cdots, p$ and ϵ_x is a random error.

To estimate the regression coefficients β_0 and β_j for $j = 1, 2, \cdots, p$, we use weighted least square with weights as the inverse variance of $\log r_x$ which is the inverse of $1/f_{x+1} + 1/f_x$. Strictly speaking, fitting the model $E(\log r_x) = \beta_0 + \beta_1 g_1(\log(x+1)) + \cdots + \beta_p g_p(\log(x+1))$ will lead to fits $\widehat{\log r_x}$, but it seems reasonable to take anti-logs of the fits (which we denote as $\hat{r}_x$) to achieve

$$\hat{r}_x = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 g_1(\log(x+1)) + \cdots + \hat{\beta}_p g_p(\log(x+1))\right].$$

We can then use the recursive relationship $\hat{p}_{x+1} = \hat{r}_x \hat{p}_x$ to find

$$\hat{p}_{x+1} = \hat{p}_0 \prod_{j=0}^{x} \hat{r}_j \tag{6}$$

for $x = 0, 1, \cdots, m-1$. Finally, we can determine $\hat{p}_0$ as the inverse of

$$1 + \sum_{j=0}^{m-1} \prod_{x=0}^{j} \hat{r}_x. \tag{7}$$

For more details, see Böhning, Lerdsuwansri and Sangnawakij (2023).

Hence, any regression model of the type given in (5) can be related to a unique discrete probability distribution. Moreover, we see the importance of the link function as it guarantees that all fitted ratios are positive.

3 Ratio regression and one-inflation

We are interested in investigating the effect of one-inflation on the ratio regression. Given a base distribution p_x, one-inflation is defined by means of an extra-weight α leading to

$$p'_x = \begin{cases} \alpha p_x, \text{ if } x \neq 1 \\ (1-\alpha) + \alpha p_x, \text{ if } x = 1. \end{cases} \tag{8}$$

Figure 1 shows a ratio regression from a Poisson with parameter 3 and 70% one-inflation. Note that only the ratios for $x = 0$ and $x = 1$ are affected by the extra-inflation as the weight parameter cancels out for $x > 1$. In fact, p'_1/p'_0 is increased and p'_2/p'_1 is decreased. It appears reasonable to assume that both effects balance out as they go in opposite directions.

The situation is different when there is zero-truncation. In the case of a zero-truncated base distribution (8) takes the form

$$p'_x = \begin{cases} \alpha \frac{p_x}{1-p_0}, \text{ if } x \neq 1 \\ (1-\alpha) + \alpha \frac{p_x}{1-p_0}, \text{ if } x = 1. \end{cases} \tag{9}$$

An illustration of the ratio regression from a zero-truncated Poisson with parameter 3 and 70% one-inflation is given in Figure 2. Clearly, here a less balancing situation is occurring as the ratio p'_1/p'_0 is missing. Any regression will have a slope biased to take a more positive value. Hence, we need ratio regression models coping with one-inflation. Model (10) is the basic regression model, here for the case of the COM distribution

$$\text{model } M_0 : \log r_x = \beta_0 + \beta_1 \log(x+1). \tag{10}$$

To allow for one-inflation we can simply include an extra-term in (10) leading to

$$\text{model } M_1 : \log r_x = \beta_0 + \beta_1 \log(x+1) + \beta_2 I_1(x) \tag{11}$$

where $I_1(.)$ is the indicator function defined as $I_1(x) = 1$ if $x = 1$ and 0 otherwise. We note that model M_0 is nested in model M_1 which would allow a likelihood ratio test to determine if these two models are significantly different, i.e., if there is evidence of one-inflation. However, we are interested in putting the assessment in a broader context. For this purpose, we utilize the Akaike-

and Bayesian Information Criteria, defined as

$$AIC = -2\log L + 2p \tag{12}$$

and

$$BIC = -2\log L + p\log(n), \tag{13}$$

respectively. Here, $\log L$ is the log-likelihood, evaluated at the maximum likelihood estimates for the model under consideration, p is the number of model parameters, and n (only relevant for (13)) is the number of different, observed x-values. More details on weighted regression and computation of the likelihood as well as the information criteria are given in the Appendix section.

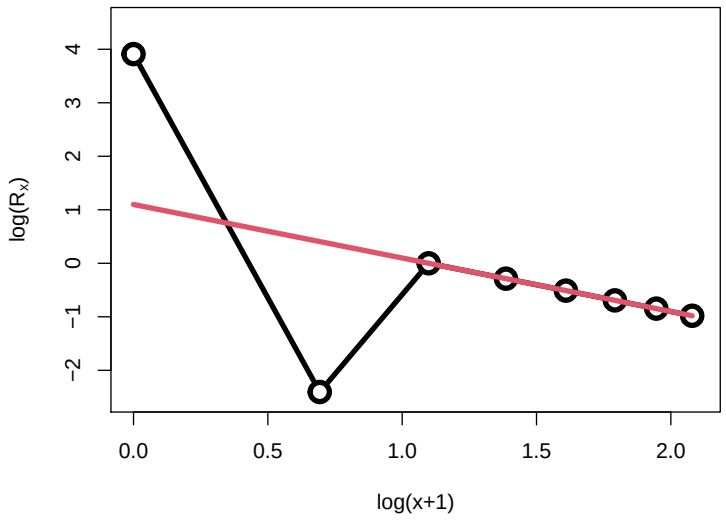


Figure 1: Log ratios for counts from Poisson distribution with parameter 3 and 70% one-inflation. The solid red line represents no inflation.

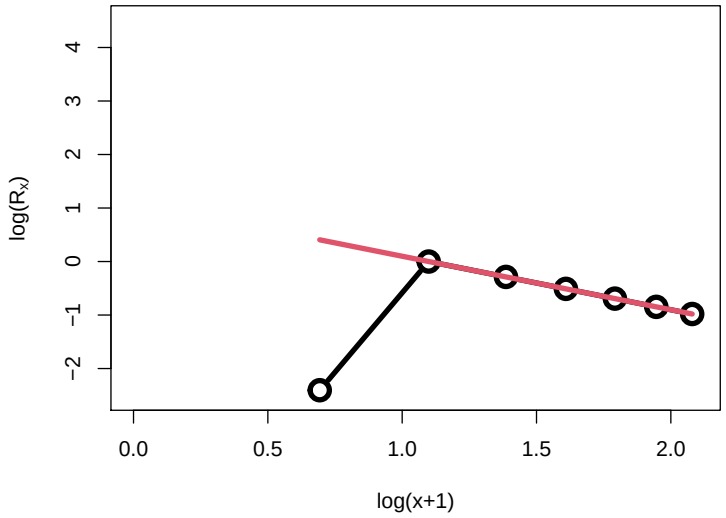


Figure 2: Log ratios for a zero-truncated Poisson distribution with parameter 3 and 70% one-inflation. The solid red line represents no inflation.

4 A case study on heroin users in the Chiang Mai province (Thailand)

Drug abuse has become a serious problem in Thailand. The legal system has shifted from punishment to treatment and rehabilitation for people who use drugs. Here we consider heroin user data from the Chiang Mai province in 2013 – 2018. The list of the surveillance system is from Chiang Mai Thanyarak Hospital serving as a treatment facility. The information is constructed on the basis of frequencies of both outpatient and inpatient treatment episodes.

Shown in Table 2 is the number of heroin users who visited the treatment facility in 2013-2018. There were 537 drug addicts who contacted the hospital exactly once, 152 were treated twice, 80 were treated three times, and so forth. A total size of observed heroin users were 843 patients with a list of 1481 records receiving treatment for heroin. We use only the data with a count of repeated visits not more than 6 as the last frequencies are very low (see also Jongsomjit and Lerdsuwansri (2023) for further details). Clearly, heroin addicts who never visited the treatment facility do not appear in the register and hence there are

no zeros observed. In addition, a relatively large number of ones are there.

Table 2: Frequencies f_x of the number of times x a heroin user visited a treatment center in Chiang Mai province in 2013-2018

x	1	2	3	4	5	6	n
f_x	537	152	80	34	15	8	843

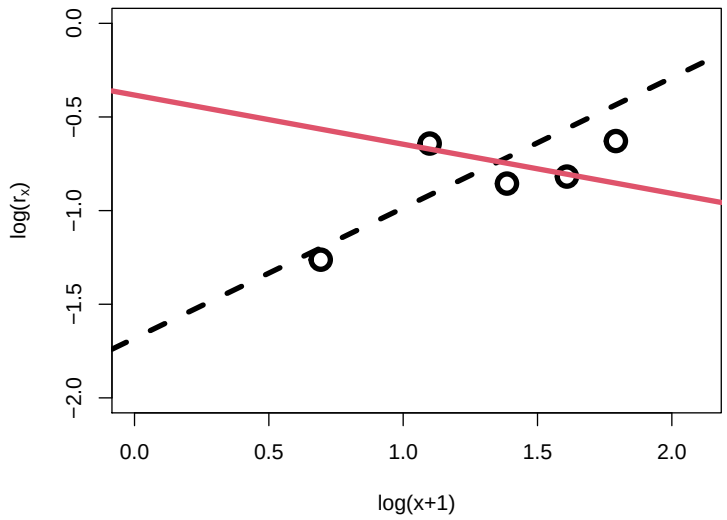


Figure 3: Log ratios for the heroin user data set from Chiang Mai. The circles represent the observed log ratios, M_1 is the solid red line, and M_0 is the dashed line. Note that only 5 points are used in the plot as the last frequencies are very low.

Table 3: Estimated regression coefficients and associated model selection criteria for the case of heroin users

Model	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	Log-L	AIC	BIC
M_0	-1.68 (0.30)	0.70 (0.30)	-	0.49	5.01	3.84
M_1	-0.38 (0.33)	-0.26 (0.26)	-0.70 (0.17)	6.18	-4.37	-5.93

Table 3 shows the regression coefficients with their associated standard errors

as well as the model selection criteria. We can see that for model M_0 the slope estimate is much more positive and the intercept considerably more negative than the respective terms for model M_1. Hence, model M_1 has a clear debiasing effect on the relevant parameter estimates. This point is also illustrated in Figure 3. Note that in Figure 3 only the non-inflated part of model M_1 is shown, as it is the one used for prediction at $x = 0$.

According to the information criteria, AIC and BIC, model M_1 gives a better fit than model M_0. Using the fitted value from M_1 (ignoring the one-inflation term) $\tilde{r}_x = \exp[\hat{\beta}_0 + \hat{\beta}_1 \log(x+1)]$, following the details given already in (6) and (7), we can then use the recursive relationship $\tilde{p}_{x+1} = \tilde{r}_x \tilde{p}_x$ to find

$$\tilde{p}_0 = \big(1 + \sum_{j=0}^{m-1} \prod_{x=0}^{j} \tilde{r}_x\big)^{-1}. \tag{14}$$

Hence, from the fitted model for $\tilde{r}_x$ we have unique probabilities $\tilde{p}_0, \tilde{p}_1, \cdots, \tilde{p}_m$ readily available, constructed as previously.

We now apply these ideas to the heroin user data set. According to the previous results, we arrive at model M_1 as a suitable model. As there is a one-to-one relationship, any ratio regression model can be associated with a corresponding, unique probability model. In the case here, the fitted values $\tilde{r}_x = \exp[-0.38 - 0.26 \log(x+1)]$ lead to $\tilde{p}_0 = 0.41, \tilde{p}_1 = 0.28, \tilde{p}_2 = 0.16, \tilde{p}_3 = 0.08, \tilde{p}_4 = 0.04, \tilde{p}_5 = 0.02$, and $\tilde{p}_6 = 0.01$. Remarkably, once the ratio regression model has been estimated, a valid count distribution can be simply derived as just illustrated above.

5 Simulation for model-based assessment

We are interested in not only incorporating one-inflation into ratio regression model but also using the log-likelihood-based selection criteria, AIC and BIC, to select the best-fitted one. As the associated ratio regression model leads to unique probabilities for the count random variable, a question arises whether the suggested information criteria correctly select the model. We therefore provide a simulation study for the model-based assessment to illustrate how this concept works. The simulation settings and procedure are given in the following. The population size N is set at 100 and 1000. We consider three distributions: (1) Poisson distributions with means $\theta = 2$ and 3, (2) geometric distributions with probability parameters $\theta = 0.2$ and 0.3, and (3) Conway-Maxwell-Poisson

distributions with parameters $(\mu, \lambda) = (2,\ 0.9)$ and $(3,\ 0.9)$. We study three setting: no, 10% and 50% extra-ones. N counts are generated under the respective settings and zero-counts are ignored. We fit models M_0 and M_1, as well as compute the AIC and BIC values for each model. At this stage, the model with the smallest AIC and BIC is chosen. Each scenario in the simulation is conducted in `R` (R Core Team 2024) and repeated 10,000 times. Finally, the average percentage of selecting model M_0 is evaluated.

The simulation results are presented in Table 4. It is clearly seen that the highest percentages of selecting model M_0 appear under the situation of no one-inflation for all population scenarios. The percentages of selecting model M_0 always decrease as the amount of one-inflation increase indicating that AIC and BIC do a good job of selecting the true model. The results behave the same way for all distributions in the study. The simulations show that performance of the model selection improves for a larger population size. Evidently, the information criteria provide a reliable way of selecting between the ratio regression model with and without one-inflation.

6 Population size estimation

We now turn to quantify the number of unobserved units and the unknown true size of the population in the context of capture-recapture approach. For a study of a closed population of size N units, we assume that a sample of observed counts of size n is available with f_x as the frequency of units which have been observed exactly x times for $x = 1, 2, \cdots, m$ where m is the largest observed count. Since unseen units do not appear in the sample, the corresponding frequency f_0 is unknown. An estimate for f_0 can be obtained using the Horvitz-Thompson estimator

$$\hat{f}_0 = n\frac{\hat{p}_0}{1-\hat{p}_0}. \tag{15}$$

This leads to the familiar Horvitz-Thompson estimator of population size N,

$$\hat{N} = n + \hat{f}_0 = \frac{n}{1-\hat{p}_0}, \tag{16}$$

where $\hat{p}_0$ is the estimated probability of a zero count. In the current work, $\hat{p}_0$ is found from the ratio regression based on model M_0.

Likewise, we have model M_1, the ratio regression model coping with one-inflation. However, some modifications are required for estimating the pop-

Table 4: Percentages of selecting model M_0 based upon AIC and BIC; $P(\theta)$ stands for Poisson distribution with parameter θ, $G(\theta)$ for geometric with parameter θ, and $COM(\mu, \lambda)$ for Conway-Maxwell-Poisson distribution with parameters (μ, λ)

	$N = 100$		$N = 1000$	
Population	AIC	BIC	AIC	BIC
no one-inflation				
P(2)	0.5454	0.5097	0.6752	0.6615
P(3)	0.6666	0.6498	0.7321	0.7485
G(0.2)	0.7193	0.7704	0.7828	0.8816
G(0.3)	0.6841	0.6946	0.7763	0.8408
COM(2, 0.9)	0.6050	0.5634	0.7018	0.6995
COM(3, 0.9)	0.7028	0.7012	0.7482	0.7787
10% one-inflation				
P(2)	0.5238	0.5003	0.3737	0.3604
P(3)	0.5841	0.5654	0.0637	0.0710
G(0.2)	0.5917	0.6454	0.0569	0.1100
G(0.3)	0.6224	0.6265	0.2097	0.2762
COM(2, 0.9)	0.5787	0.5400	0.3008	0.2989
COM(3, 0.9)	0.5192	0.5187	0.0169	0.0222
50% one-inflation				
P(2)	0.2559	0.2181	0.0008	0.0005
P(3)	0.1031	0.0888	0	0
G(0.2)	0.0890	0.0934	0	0
G(0.3)	0.1978	0.1813	0	0
COM(2, 0.9)	0.2236	0.1889	0.0001	0.0001
COM(3, 0.9)	0.0550	0.0501	0	0

ulation size. In the case that the observed sample will contain one-inflated singletons and non-inflated singletons, it is not known which singleton belongs to the inflated and which belongs to the non-inflated part. Consequently, the singletons are completely removed and estimation is based on the remaining counts of size $n - f_1$. Also, $\tilde{p}_0$ and $\tilde{p}_1$ are the predicted probabilities for the non-inflated part of M_1. This leads to a modified Horvitz-Thompson estimator

$$\hat{f}_0 = (n - f_1)\frac{\tilde{p}_0}{1 - \tilde{p}_0 - \tilde{p}_1}, \tag{17}$$

and the total population size estimator $\hat{N} = n + (n - f_1)\frac{\tilde{p}_0}{1-\tilde{p}_0-\tilde{p}_1}$ follows.

In addition to the Horvitz-Thompson estimator, we have also included the Chao estimator $\hat{N}_C = n + f_1^2/(2f_2)$ (Chao 1989, Chao and Colwell 2017, Chao

and Bunge 2002) and the Chao estimator modified for one inflation $\hat{N}_{MC} = n + (2/9)f_2^3/f_3^2$ (for details see Böhning, Kaskasamkul and van der Heijden 2019). If a geometric distribution as a kernel in the nonparametric Chao mixture deems more appropriate (than a Poisson kernel), the Chao bounds take the form $\hat{N}_{CG} = n + f_1^2/f_2$ and modified for one inflation $\hat{N}_{MCG} = n + f_2^3/f_3^2$. The latter seems reasonable in case the ratio regression has a positive slope which is not feasible for the Conway-Maxwell-Poisson distribution. Projection to the feasible space (arbitrary intercept and non-positive slope) leads to a zero slope which corresponds to the geometric distribution.

7 A semi-parametric estimator of population size for the general ratio regression model

Let us now consider the general case of a ratio regression model M_1 with potential one-inflation

$$\log r_x = \beta_0 + \beta_1 g_1(\log(x+1)) + \cdots + \beta_p g_p(\log(x+1)) + \beta_{p+1} I_1(x) + \epsilon_x, \quad (18)$$

for $x = 0, 1, \cdots, m-1$, where $I_1(x)$ is the indicator function for $x = 1$. Suppose we fit model (18) and use the part without the one-inflation term in model M_1 (this is the black line in Figure 3)

$$\tilde{r}_x = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 g_1(\log(x+1)) + \cdots + \hat{\beta}_p g_p(\log(x+1))\right],$$

we are able to find $\tilde{p}_0$ as the inverse of

$$1 + \sum_{j=0}^{m-1} \prod_{x=0}^{j} \tilde{r}_x. \quad (19)$$

From here the size estimator of missed units $\hat{f}_0 = (n - f_1)\frac{\tilde{p}_0}{1-\tilde{p}_0-\tilde{p}_1}$ follows as developed in the previous section.

Alternatively, we can think of

$$\tilde{p}_x = \tilde{p}_{x+1}/\tilde{r}_x, \quad (20)$$

from where

$$\tilde{p}_0 = \tilde{p}_1/\tilde{r}_0 \quad (21)$$

follows. Unfortunately, replacing $\tilde{p}_1$ by f_1 in (21) is not feasible as f_1 is one-inflated. However, using $\tilde{p}_1 = \tilde{p}_2/\tilde{r}_1$ we can extend (21) by

$$\tilde{p}_0 = \frac{\tilde{p}_2}{\tilde{r}_0\tilde{r}_1}. \tag{22}$$

Now it is possible to replace $\tilde{p}_2$ in (22) by its estimate f_2, which leads to the semi-parametric estimator

$$\hat{f}_{0,sm} = \frac{f_2}{\tilde{r}_0\tilde{r}_1}. \tag{23}$$

Ultimately, we define the semi-parametric estimator of population size

$$\hat{N}_{sm} = n + \frac{f_2}{\tilde{r}_0\tilde{r}_1}. \tag{24}$$

8 Population size inference for the case study

So far, population size estimators are derived using ratio regression with and without one-inflation. Here we have the Horvitz-Thompson (HT) and semi-parametric (SM) estimators. Additionally, we have Chao's estimator and the modified Chao's estimators which do not require any modeling. In this section, we discuss population size inference which can be done by the nonparametric bootstrap method in different ways. These will be illustrated through the case study on heroin users in Chiang Mai (Thailand).

8.1 Nonparametric bootstrap method without model selection for population size inference

Table 5 shows the estimated number of hidden heroin users and population sizes. The last two columns of table present the 90% and 95% percentile bootstrap confidence intervals for N. We use $B = 10000$ bootstrap samples based on the procedure given in Algorithm 1 further below. The results demonstrate that the estimated population size from the conventional Chao estimator is considerably different from the modified Chao's estimator. They also depend on the choice of the probability base model (Poisson or geometric distribution). In addition, choosing between Chao's and the modified Chao's estimator requires knowledge on the presence of one-inflation. In other words, a distributional analysis is required before population size estimation can be considered. For example, the ratio plot (Böhning et al., 2013; Böhning, 2016) can be used to

explore the pattern of the distribution for capture-recapture data. In fact, it is often easier to find an appropriate ratio regression model than a proper model for the count distribution. As shown above, the regression model approach leads to a valid count distribution. From the results in Table 5, it can be seen that the HT and SM estimates computed under the same regression model do not differ substantially. Both $\hat{f}_0$ and $\hat{N}$ computed based on models M_0 are considerably different from those computed from model M_1. To choose a reasonable regression model, we may use the AIC and BIC as performance criteria. As noted in Table 3, model M_1 provides the smallest AIC and BIC for the heroin addict dataset. This evidence shows that the estimators based on model M_1 could be applied for population size inference.

Table 5: Estimated number of unobserved units, population size estimates (Bootstrap median) and 90% and 95% percentile bootstrap confidence interval (CI) for population size of Chiang Mai heroin user

Method	$\hat{f}_0$	$\hat{N}$	90% CI	95% CI
HT (model M_0)	2950	3793 (3835)	(2987, 5010)	(2841, 5323)
HT (model M_1)	414	1257 (1265)	(938, 2587)	(915, 3117)
SM (model M_0)	2710	3553 (3595)	(2844, 4614)	(2717, 4858)
SM (model M_1)	392	1235 (1245)	(933, 2526)	(912, 3058)
Chao (geometric)	1897	2740 (2742)	(2413, 3149)	(2361, 3235)
modified Chao (geometric)	549	1392 (1397)	(1167, 1796)	(1133, 1911)
original Chao (Poisson)	949	1792 (1792)	(1627, 1999)	(1601, 2040)
modified Chao (Poisson)	122	965 (964)	(913, 1054)	(907, 1078)

Table 6: Estimated population size, standard error of $\hat{N}$, 90% and 95% percentile bootstrap confidence interval (CI) for the number of Chiang Mai heroin users and percentage of selecting model M_1 using the nonparametric bootstrap method with model selection

Method	$\hat{N}$	SE($\hat{N}$)	90% CI	95% CI	% of model M_1
HT (Algorithm 2)	1249	975.07	(938, 3874)	(914, 4213)	83.23
HT (Algorithm 3)	1244	957.32	(933, 3870)	(912, 4220)	83.87
SM (Algorithm 2)	1242	948.72	(934, 3768)	(911, 4096)	83.61
SM (Algorithm 3)	1248	945.91	(936, 3814)	(912, 4172)	84.00

To estimate confidence intervals, we use the nonparametric bootstrap as described in the capture-recapture literature (for details see Anan, Böhning and Maruotti 2017). The major difference to the standard nonparametric bootstrap

is the fact that the following bootstrap takes into account the observed sample size n as part of the inference (each bootstrap sample will generate a different observed n). However, it does not account for model selection.

Algorithm 1: Simple bootstrap with imputation of f_0

1. Find the estimate of N, denoted as $\hat{N}$, and $\hat{f}_0 = \hat{N} - n$ from the sample data (0-truncated sample).
2. Draw a bootstrap sample of size $\hat{N}$ from a multinomial distribution with probabilities $\hat{f}_0/\hat{N}, f_1/\hat{N}, \cdots, f_m/\hat{N}$, where m is the largest count.
3. Truncate zero counts and compute the population size from the bootstrap sample, namely $\hat{N}^*$.
4. Repeat steps 2 and 3 for B times to get estimates $\hat{N}^{*(1)}, \hat{N}^{*(2)}, \cdots, \hat{N}^{*(B)}$.
5. Find the lower and upper limits of the $(1-\alpha)100\%$ confidence interval for N from the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of $\hat{N}^{*(b)}$, for $b = 1, 2, \cdots, B$.

8.2 Nonparametric bootstrap method with model selection for population size inference

As shown in the previous sections, the ratio regression with and without one-inflation can be used to guide population size estimation provided by AIC or BIC. However, basing inferences on a single model may lead to bias towards the model of choice according to AIC or BIC. In other words, there are two realistic models (M_0 and M_1) in our case that could have been selected for inference. Therefore, in the following part, we will introduce two bootstrap methods for the HT and SM model-based estimators and will include model selection in the inference process to deal with the uncertainty that comes with the sampling. The procedures for estimating the population size for each bootstrap method are given in Algorithms 2 and 3. Note that the bootstrapping in Algorithm 2 is different from that in Algorithm 3. The former requires a double-bootstrap sample to estimate the unobserved frequency in order to improve the performance of bootstrap methods for bias correction. Meanwhile, the bootstrap method in Algorithm 3 has no imputation for the missing data.

Algorithm 2: Double bootstrap with incorporating model selection

1. From the zero-truncated sample, draw a bootstrap sample of size n from a multinomial distribution with probabilities $f_1/n, f_2/n, \cdots, f_m/n$.
2. Find the regression coefficients under models M_0 and M_1, calculate AIC for each model
3. Determine if M_0 or M_1 is more appropriate w.r.t. the AIC.
4. Estimate f_0 and N from the model-based estimator, according to the choice of M_0 or M_1, denoted as $\hat{f}_0^*$ and $\hat{N}^*$, respectively.
5. Draw a bootstrap sample from a multinomial distribution with probabilities $\hat{f}_0^*/\hat{N}^*, f_1/\hat{N}^*, f_2/\hat{N}^*, \cdots, f_m/\hat{N}^*$.
6. Truncate zero counts and find the best of models under regression analysis.
7. According to the choice of the best model, estimate N using the model-based estimator, namely $\hat{N}^{**}$.
8. Repeat the steps 1 to 7 for B times to get bootstrap estimates $\hat{N}^{**(1)}$, $\hat{N}^{**(2)}, \cdots, \hat{N}^{**(B)}$.
9. Find the population size estimate as median and the confidence interval for N as percentile intervals of the bootstrap estimates $\hat{N}^{**(1)}$, $\hat{N}^{**(2)}$, $\cdots, \hat{N}^{**(B)}$.

Algorithm 3: Simple bootstrap with incorporating model selection

1. From the zero-truncated sample, draw a bootstrap sample of size n from a multinomial distribution with probabilities $f_1/n, f_2/n, \cdots, f_m/n$.
2. Find the regression coefficients under models M_0 and M_1, calculate AIC from each model.
3. Determine if M_0 or M_1 is more appropriate w.r.t. the AIC.
4. Estimate f_0 and N on the basis of the chosen model M_0 or M_1.
5. Repeat the steps 1 to 4 for B times to get bootstrap estimates $\hat{N}^{*(1)}$, $\hat{N}^{*(2)}, \cdots, \hat{N}^{*(B)}$.

6. Find the population size estimate as median and the confidence interval for N as percentile intervals of the bootstrap estimates $\hat{N}^{*(1)}$, $\hat{N}^{*(2)}$, $\cdots$, $\hat{N}^{*(B)}$.

Table 6 presents the estimated population size and confidence interval computed using the nonparametric bootstrap method with model selection in statistical inference. For the heroin dataset, the estimates of N obtained from the two types of bootstrap are fairly close. The population size estimate from the HT estimator is similar to that of the SM estimator. Furthermore, we can see that the model-based estimators provide population size estimates that are closer to the modified Chao estimators for one-inflation than the conventional Chao estimator. Evidence for choosing the correct model is also shown by the percentage of selecting model M_1. It is given in the last column of Table 6. Obviously, model M_1 has a high percentage of selection from the resample process. Since model M_1 corresponds to the extra-one term, the heroin dataset is assumed to be the one-inflation scenario.

The comparison of the confidence intervals for N given in Tables 5 and 6 is displayed also in Figure 4. The bootstrap methods incorporating model selection for population size inference provide a larger interval length than those without model selection. This appears to be in line with the expectation that model selection goes along with an increased variability. However, we will study this in more details in a simulation which will be used to assess the performance of our methods. It is described in the next section.

9 Simulation for population size estimators

A simulation study is undertaken to investigate the performance of the proposed estimators derived from ratio regression modeling and other well-established ones for population size estimation. We evaluate the performance of the methods in terms of frequentist performance measures, including relative bias, relative mean squared error (RMSE), coverage probability of the confidence interval, and interval width. For the study, a specific base model for count data is required to draw the samples. We consider Poisson, geometric, and COM distributions, and their one-inflated versions. The parameter configurations of the probability models and population sizes are the same as those in Section 5. Here, $B =$

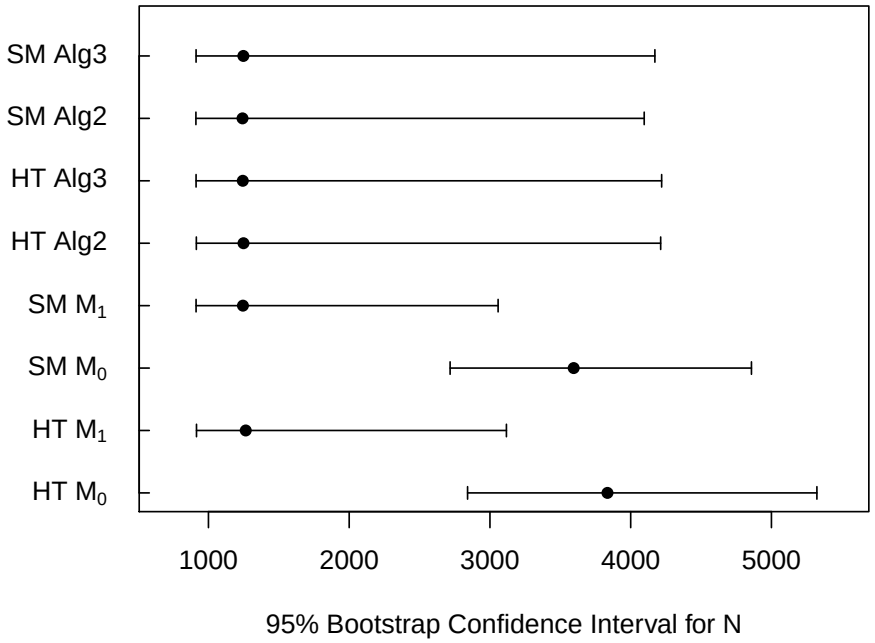


Figure 4: Interval plots of 95% bootstrap confidence intervals for N from the heroin users data computed by the methods with and without model selection.

1000 bootstrap samples are used based on the procedures given in Algorithm 1 for Chao's and modified Chao's estimators and model-based estimators without model selection, as well as Algorithms 2 and 3 for model-based estimators using model selection. Each scenario is repeated $H = 1500$ times. We average the relative bias and RMSE of the estimated population size across all simulation replications. They are given by

$$RB(\hat{N}) = \frac{1}{N}\left(\frac{1}{H}\sum_{h=1}^{H}\hat{N}_h - N\right) = \frac{Bias(\hat{N})}{N}$$

and

$$RMSE(\hat{N}) = \frac{1}{N^2}\left(\frac{1}{H}\sum_{h=1}^{H}(\hat{N}_h - N)^2\right) = \frac{Var(\hat{N}) + [Bias(\hat{N})]^2}{N^2}.$$

For interval estimation, the coverage probability and expected length of the confidence interval for N are computed by

$$CP = \frac{\#(L_h \leq N \leq U_h)}{H}$$

and

$$EL = \frac{1}{H}\sum_{h=1}^{H}(U_h - L_h),$$

respectively, where $\#(L_h \leq N \leq U_h)$ is the number that the true parameter value N lies within the lower and upper limits. As usual, we seek small bias and variance. However, our focus here is on coverage probability as the major performance measure to account for uncertainty assessment. We prefer a confidence interval that has close-to-nominal coverage. Additionally, our interest is in the assessment of including model selection into the estimation of population size, in particular the assessment of identifying one-inflation using ratio regression.

The population size estimates and their performances based on data generated either without or with one-inflation sampling are presented in Figures 5, 6 and 7. The main simulation results are summarized as follows in some key messages. More extensive simulation results are available as web supplement. In Figure 5 we focus on coverage probability of the 95% confidence intervals for the stand-alone estimators (no modeling involved) of Chao and its modified version for one-inflation based on the Poisson kernel (classical Chao) and the geometric kernel. They occur as C, MC, CG and MCG, respectively. We clearly can see that these estimators work fine under the setting they have been developed for, but break down if these settings are not met. Figure 6 shows the coverage probabilities for models M_1 and M_0 without model selection (meaning based on Algorithm 1). We see the break-down of coverage probability for M_0 if one-inflation increases. The coverage probability for M_1 is fine in all cases. However, this should be carefully interpreted as in this case model M_1 includes the Poisson, geometric and COM distribution as well as one-inflation. Hence it covers a wider range of distribution but might be different in other settings. In Figure 7 we see the coverage probabilities for the model-based estimators including model selection. HT stands for the Horvitz-Thompson as developed in Section 6 and SM stands for the semi-parametric estimator developed in Section 7. In a nutshell, the computational expensive Algorithm 2 works fine in

all scenarios, where the simplified and less expensive Algorithm 3 breaks down in several settings, see the left panels in Figure 7. It appears that only Algorithm 2 can be recommended and there are negligible differences between the (more complex) Horvitz-Thompson estimator and the simple semi-parametric estimator.

10 Discussion

The current work has presented a flexible tool for finding an appropriate count density which is quite suitable for population size estimation. In the past, capture-recapture estimation has been dominated by stand-alone estimators. A prominent representative is the lower bound estimator developed by Chao (Chao 1989, Chao and Colwell 2017) which performs well in many settings. However, it fails to do so when there is one-inflation where it loses its lower bound property and can experience serious overestimation bias. Modification of Chao's estimator has been suggested in the case of one-inflation (Böhning *et al.* 2019, Chiu and Chao 2016) that performs well under the occurrence of extra-ones and retains the lower bound property. The general disadvantage of stand-alone estimators, however, can be seen in the fact that they might perform well in situations for which they have been developed whereas they might fail to do so in others.

Here we present a model-based approach which selects models first on the basis of some selection criterion, such as the AIC or BIC, then, after selecting the best model, constructs the population size estimator on the basis of the chosen model. Although this procedure seems quite reasonable, it has the drawback that it will be biased towards the selected model. In other words, the available sample is only one of many possible and if another sample of the same type would be available, another model might have been chosen. This is particular the case in situations where several models under competition have information criteria values which are quite close to each other. For this reason, a bootstrap procedure has been developed and shown to perform well which allows the integration of model selection into the estimation of population size, an example of computational statistical inference. It is applied here to models with and without incorporating one-inflation, however, the principle is far more general applicable. A similar approach has been suggested by Silverman *et al.* (2024) with the difference that the suggested bootstrap only focuses on the ob-

served data (similar to our Algorithm 3) and ignores the uncertainty involved in estimating the missing cell.

Evidently, bootstrapping for population size estimation can be done in different ways. In Algorithm 1, model selection is ignored and the expectation here is that the involved uncertainty is underestimated. Algorithm 2 is suggesting a double bootstrap. In the first bootstrap model selection is accounted for, in the second bootstrap sample estimation of the missing cell count is adjusted for. Algorithm 3 is a simplified version of Algorithm 2 in which the bootstrap for the model selection part is retained, but the bootstrap for the imputation part is dropped. These have been compared in a simulation study to shed some light on their performance. As a major result we can summarize that including model selection into the inference is essential to yield valid inference and the simplified version (Algorithm 3) seems not sufficient to accomplish this. It seems necessary to account for uncertainty in the imputation step which is delivered with Algorithm 2.

11 Appendix: Likelihood and information criteria in weighted regression

The weighted regression is handled slightly different in the literature and, consequently, in the associated computer packages. We follow the approach that `R` (R Core Team 2024) is taking. Consider a linear regression model

$$y_i = x_i^T \beta + \epsilon_i, \tag{25}$$

where the independent errors are $\epsilon_i \sim N(0, \sigma^2/w_i)$. Here σ^2 is an unknown variance parameter and the positive weights w_i are assumed to be known, often proportional to some pre-specified variances. In total, we write $\sigma_i^2 = \sigma^2/w_i$. The log-likelihood is provided as

$$\log L(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_i \log(\sigma_i^2) - \frac{1}{2}\sum_i \frac{(y_i - x_i^T\beta)^2}{\sigma_i^2}. \tag{26}$$

The β coefficients are estimated as weighted least squares $\hat{\beta}$ which are also the maximum likelihood estimates under normality. Care needs to be taken when estimating the variance parameter σ^2. Typically, the unbiased estimator $SSE/(n-p)$ is used where $SSE = \sum_i w_i(y_i - \hat{y}_i)^2$ with $\hat{y}_i = x_i^T\hat{\beta}$ for $i =$

$1, 2, \cdots, n$ and p is the number of β-parameters. An important point is that the maximum likelihood estimator of σ^2 is given by $\hat{\sigma}^2 = SSE/n$ so that the maximum log-likelihood is given as $\log L(\hat{\beta}, \hat{\sigma}^2)$. As there are p parameters in β and one additional variance parameter, AIC and BIC are given as

$$-2 \log L(\hat{\beta}, \hat{\sigma}^2) + 2(p+1)$$

and

$$-2 \log L(\hat{\beta}, \hat{\sigma}^2) + (p+1) \log(n),$$

respectively. Note that n here is referred to as the number of observed x-values.

References

Anan, O., Böhning, D. and Maruotti, A. (2017) Uncertainty estimation in heterogeneous capture-recapture count data. *Journal of Statistical Computation and Simulation*, **10**, 2094–2114.

Böhning, D. (2016) Ratio plot and ratio regression with applications to social and medical sciences. *Statistical Science*, **31**, 205–218.

Böhning, D., Rocchetti, I. Alfó, M. and Holling, H. (2016) A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics*, **72**, 697–706.

Böhning, D. and Ogden, H. (2021) General flation models for count data. *Metrika*, **84**, 245–261.

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

Böhning, D., Bunge, J. and van der Heijden, P.G.M. (2018) *Capture-Recapture Methods for the Social and Medical Sciences.* Boca Raton: Chapman & Hall/CRC.

Böhning, D., Kaskasamkul, P. and van der Heijden, P.G.M. (2019) A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, **82**, 361–384.

Böhning, D., Lerdsuwansri, R. and Sangnawakij, P. (2023) Modeling covid-19 contact-tracing using the ratio regression capture-recapture approach. *Biometrics*, **79**, 1–13.

Böhning, D. and Friedl, H. (2024) One-inflation and zero-truncation count data modelling revisited with a view on Horvitz-Thompson estimation of population size. *International Statistical Review*, (to appear).

Chao, A. (1989) Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438.

Chao, A. and Bunge, J. (2002) Estimating the number of species in a stochastic abundance model. *Biometrics*, **58**, 531–539.

Chao, A., Colwell, R.K. (2017) Thirty years of progeny from Chaos inequality: estimating and comparing richness with incidence data and incomplete sampling. *SORT* **41**, 3–54

Chiu, C.-H., Chao, A. (2016) Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ* **4**, e1634.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, **39**, 1–22.

Godwin, R. (2017) One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, **59**, 79–93.

Godwin, R. (2019) The one-inflated positive Poisson mixture model for use in population size estimation. *Biometrical Journal*, **61**, 1541–1556.

Godwin, R., and Böhning D. (2017) Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society Series C, Applied Statistics*, **66**, 425–448.

Jongsomjit, T. and Lerdsuwansri, R. (2023) Estimation of population size based on one-inflated, zero-truncated count distribution with covariate information. *Sains Malaysiana*, **52**, 3867–3877.

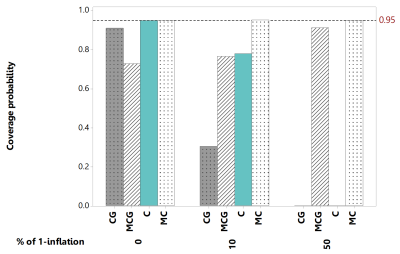
McCrea, R.S. and Morgan, B.J.T. (2015) *Analysis of Capture-Recapture Data*. Boca Raton: Chapman & Hall/CRC.

Sellers, K.F. (2023) *The Conway–Maxwell–Poisson distribution.* Cambridge University Press.

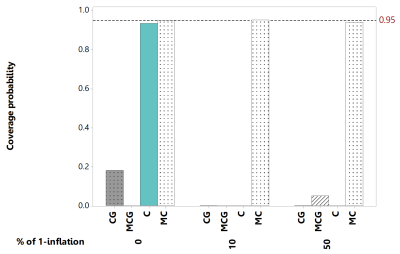
Silverman, B. W., Chan, L. and Vincent, K. (2024) Bootstrapping multiple systems estimates to account for model selection. *Statistics and Computing*, **34**, 44.

Tuoto, T., Di Cecco, D., and Tancredi, A. (2022) Bayesian analysis of one-inflated models for elusive population size estimation. *Biometrical Journal*, **64**, 912–933.

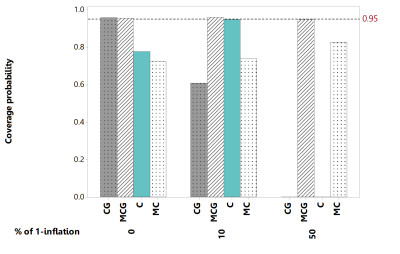
R Core Team (2024) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.



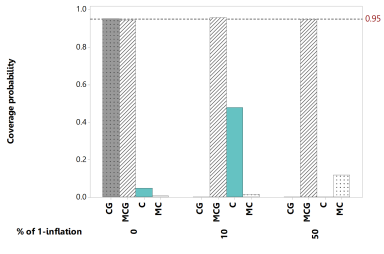
(a) P(3) and $N = 100$



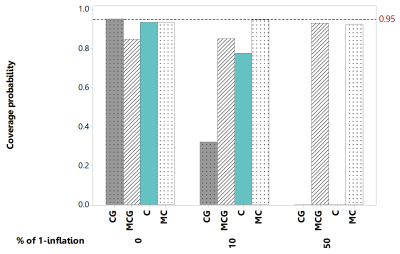
(b) P(3) and $N = 1000$



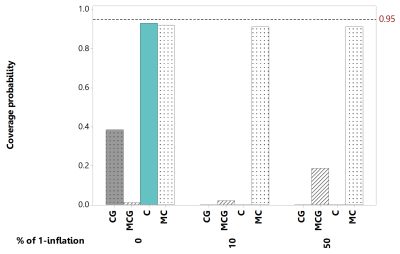
(c) G(0.2) and $N = 100$



(d) G(0.2) and $N = 1000$

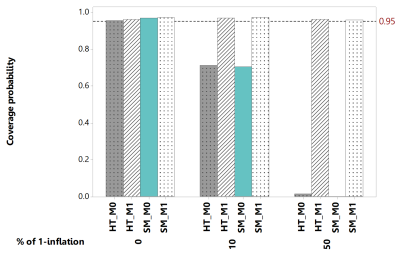


(e) COM(3, 0.9) and $N = 100$

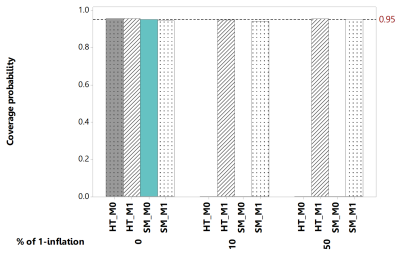


(f) COM(3, 0.9) and $N = 1000$

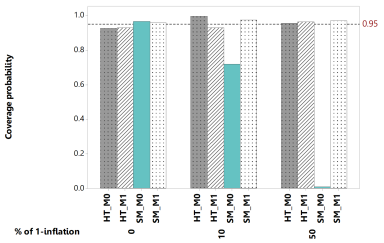
Figure 5: Coverage probabilities of 95% confidence intervals for N under Chao's and modified Chao's estimators with bootstrap simulations.



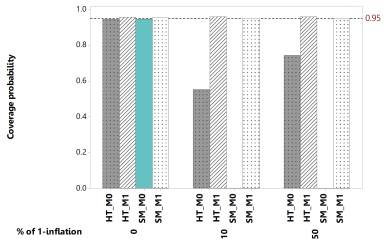
(a) P(3) and $N = 100$



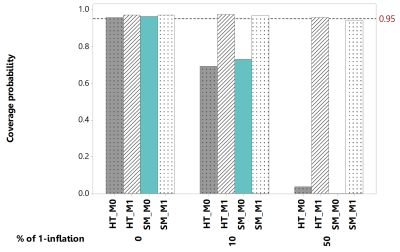
(b) P(3) and $N = 1000$



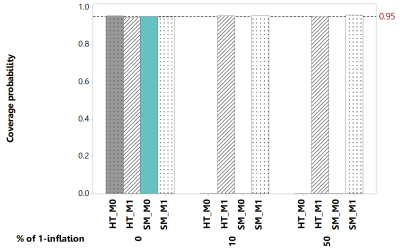
(c) G(0.2) and $N = 100$



(d) G(0.2) and $N = 1000$

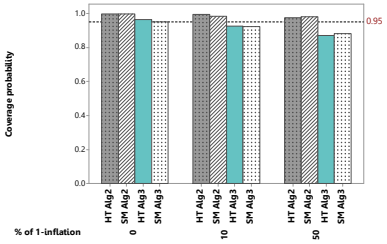


(e) COM(3, 0.9) and $N = 100$

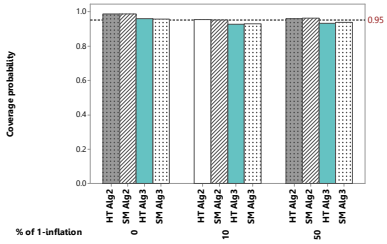


(f) COM(3, 0.9) and $N = 1000$

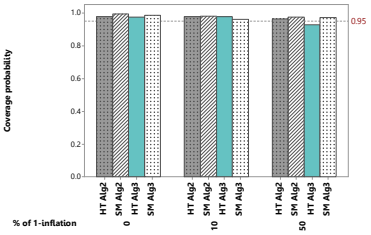
Figure 6: Coverage probabilities of 95% confidence intervals for N under simulations with bootstrap Algorithm 1.



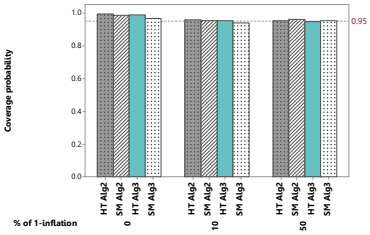
(a) P(3) and $N = 100$



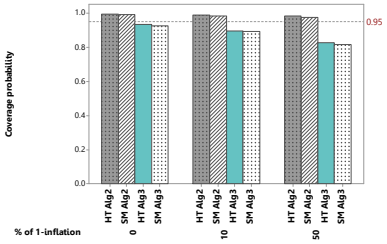
(b) P(3) and $N = 1000$



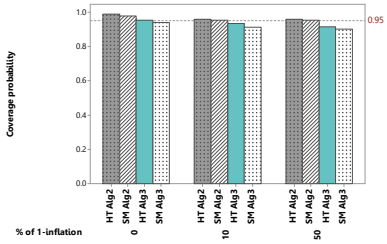
(c) G(0.2) and $N = 100$



(d) G(0.2) and $N = 1000$



(e) COM(3, 0.9) and $N = 100$



(f) COM(3, 0.9) and $N = 1000$

Figure 7: Coverage probabilities of 95% confidence intervals for N under simulations with bootstrap Algorithms 2 and 3.