Thailand Statistician
October 2023; 21(4): 725-744.
http://statassoc.or.th
Contributed paper

Estimating the Unknown Size of a Population based upon Re-parameterized Geometric Distribution

Rattana Lerdsuwansri, Parawan Pijitrattana* and Suramase Hashim
Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Pathum Thani, Thailand
*Corresponding author; e-mail: parawan@mathstat.sci.tu.ac.th

Received: 25 October 2022
Revised: 25 May 2023
Accepted: 6 June 2023

Abstract

Estimating the unknown size of a partially observed population is challenging particularly when most of observed subjects are captured once. Geometric distribution, one of the most well-utilized discrete distributions in a capture-recapture setting, is re-parameterized corresponding to the uniform Poisson Ailamujia, a flexible model for data set with excesses of ones. The maximum likelihood and Generalized Turing estimators using uniform Poisson Ailamujia distribution were proposed. We address achieving variance estimates of population size estimators by using conditioning approach. In simulation studies, potential of the proposed estimators as well as the confidence interval are investigated and compared to conventional estimators developed on the basis of geometric distribution. All estimators behaved similarly and the presented confidence intervals can improve the estimation. As an application, two real data examples are examined using the proposed estimators.

Keywords: Capture-recapture data, heterogeneity, one-inflation, variance estimation.

1. Introduction

In many fields, capture-recapture (CR) methods have been widely used to estimate the size of hidden populations. For example, CR methods are used to estimate the number of heroin users in northern Thailand (Pijitrattana 2018), the number of European pond turtles in a Venice Lagoon wetland area (Liuzzo et al. 2021), the number of farms in Southeast Asia that experienced foot and mouth disease outbreaks (Sansamur et al. 2021), the number of female sex workers in Vietnam (Nguyen et al. 2021), and the completeness of contact tracing for COVID-19 during the first wave pandemic in Thailand (Lerdsuwansri et al. 2022; Böhning et al. 2023).

A target population is assumed to be closed, which means that no births, deaths, or migration occur during the sampling period. Samples are obtained through the use of identification mechanisms such as registration or trapping systems. Let X_j be the number of times that unit j^{th} is identified during the observational period, for $j = 1, 2, \ldots, N$. The frequency of units identified exactly $0, 1, \ldots, m$ times is denoted by $f_0, f_1, \ldots, f_m$ respectively, where m is the largest observed count. Since all units in the

population have not been observed, the number of units identified zero times, f_0, is unknown. The observed sample size identified at least once is $n = f_1 + f_2 + \cdots + f_m$. The number of unobserved subjects must be estimated to obtain an estimate for the population size $N = n + f_0$.

Assume that $p_x = P(X = x)$ is the probability of identifying a unit x times and p_0 is the probability of a unit not being identified. Since unobserved f_0 can be replaced by the expected value Np_0, estimation of p_0 is required, resulting in the well-known Horvitz-Thompson estimator (Van Der Heijden et al. 2003)

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

Count data are typically represented by a Poisson distribution with parameter λ. In practice, count data modeled by an identical λ, homogeneous Poisson model do not hold due to diversity of subjects in the population. Heterogeneous Poisson model is more flexible and the probability function is given as

$$p_x = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} g(\lambda) d\lambda,$$

where $g(\lambda)$ is mixing distribution. For example, if $g(\lambda)$ is exponential distribution with parameter θ, a geometric distribution with parameter $p = \frac{1}{1+\theta}$ arises. That is a special case of mixing the Poisson and gamma distributions, resulting in a negative binomial distribution. Geometric distributions are commonly used for overdispersion capture-recapture data, and several estimators have been developed. Niwitpong et al. (2013) proposed three estimators, $\hat{N}_{MLEgeo} = \frac{nS}{S-n}$, where $S = \sum_{x=1}^{m} x f_x$, $\hat{N}_{Chao} = n + \frac{f_1^2}{f_2}$, and $\hat{N}_{Censored} = \frac{n^2}{n-f_1}$, based on geometric distribution using three approaches. A classical maximum likelihood estimation for the zero-truncated geometric is considered leading to $\hat{N}_{MLEgeo}$. The second one $(\hat{N}_{Chao})$ is developed in spirit of Chao estimator by keeping the mixing distribution unspecified and applying nonparametric inference based on the Cauchy-Schwarz inequality. Another estimator, $\hat{N}_{Censored}$, is suggested by using all available information but censors counts larger than 1. Anan et al. (2019) proposed a generalized Turing estimator, $\hat{N}_{GTgeo} = \frac{n}{1-\sqrt{\frac{f_1}{S}}}$, for geometric distribution. A modification of Chao's lower bound estimator, $\hat{N}_{MC} = n + \frac{f_2^3}{f_3^2}$, in the case of one-inflation geometric model has been proposed (Böhning et al. 2019). Böhning and Ogden (2021) recently demonstrated that one-inflation models can be fitted by truncating the count of ones and proposed a geometric distribution estimator $\hat{N}_{BO} = n + (n - f_1)\frac{\hat{\theta}}{1-\hat{\theta}-\hat{\theta}(1-\hat{\theta})}$,where $\hat{\theta} = \frac{n-f_1}{n-f_1+S^*}$, $S^* = \sum_{x=0}^{m-2} x f_{x+2}$.

Aljohani et al. (2021) proposed a uniform Poisson-Ailamujia distribution (UPA) as a flexible discrete model for datasets with excesses of ones by combining the uniform and Poisson-Ailamujia distribution. The UPA is a heavy-tailed distribution that can be used to model overdispersed count data. The probability function of UPA distribution is

$$p_x = P(X = x) = \frac{2\alpha}{(1+2\alpha)^{x+1}} \tag{1}$$

for $x = 0, 1, 2, ...$ with parameter $\alpha > 0$. The mean and variance are $\frac{1}{2\alpha}$ and $\frac{2\alpha+1}{4\alpha^2}$, respectively. As can be seen in Figure 1 the frequency of 1's depends on the value of α. The larger the parameter α is, the higher frequency of one counts is. Before we go on, we illustrate the situation at hand with a real data example. The frequency distribution of the number of daily COVID-19 deaths in Switzerland

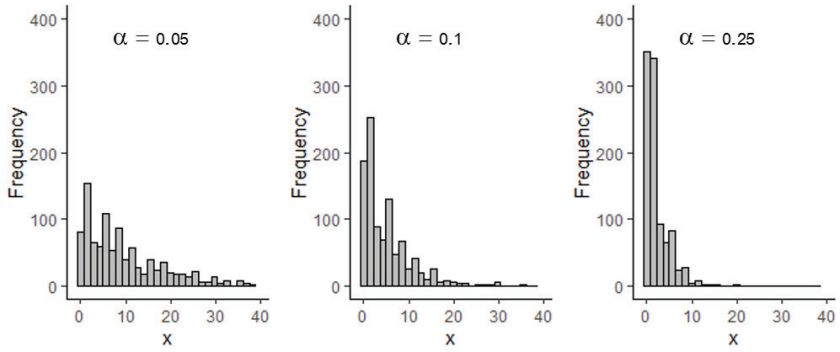

Figure 1 Graph of the UPA distribution

from 1 March to 30 June 2021 (Aljohani et al. 2021) is provided in Table 1.To estimate the UPA and Poisson distribution parameters from the data, the maximum likelihood method was used. The maximum likelihood estimator for the UPA and Poisson distribution parameters is $\hat{\alpha} = \frac{n}{2(S-n)}$ and $\hat{\lambda} = \frac{S}{n}$, respectively. We also see in Figure 2 that the UPA provides a much better fit than the Poisson.

Table 1 The daily numbers of COVID-19 deaths in Switzerland from 1 March to 30 June 2021

Number of deaths	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	15	11	12	8	9	4	4	7	9	6	3	2	1	6	3
Number of deaths	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Frequency	1	3	3	1	1	2	1	1	0	1	0	3	0	1	3

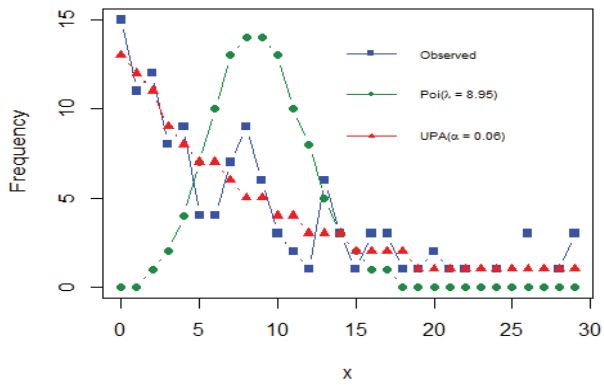

Figure 2 The frequency distribution of COVID-19 deaths in Switzerland from 1 March and 30 June 2021

Tajuddin and Ismail (2022) pointed out that the UPA distribution is a re-parameterized geometric distribution. The probability function in Eqn. (1) can be re-parametrized by changing α through a

function θ, such that

$$\theta(\alpha) = \frac{2\alpha}{1+2\alpha}.$$

Therefore, $$p_x = P(X = x) = (1-\theta)^x\theta, \tag{2}$$

$x = 0,1,2,...,0 < \theta < 1$, which corresponds to the geometric distribution with parameter θ. The parameter θ in the geometric distribution represents the probability of success, whereas the parameter α is a scale that either stretches or shrinks the trend of the probability values. According to the distribution graph (Figure 1), as α increases, the data becomes more inflated with one, implying that estimation of α would be helpful to elucidate one-inflation.

Recently, the presence of one-inflation in capture-recapture-type data has attracted some attention. Godwin and Böhning (2017) added an excess probability of observing one counts in the positive Poisson (PP) distribution and propose the one-inflated positive Poisson (OIPP) distribution. To allow unobserved heterogeneity, Godwin (2017) proposed the one-inflated zero-truncated negative binomial (OIZTNB) model. Jongsomjit et al. (2022) develop the one-inflated, zero-truncated geometric (OIZTG) model using reparameterization for the mean of the OIZTG distribution. What they have in common are one-inflation parameter and covariates incorporating into truncated regression model. Although covariates can help to improve the fit of the model, they are unavailable in many cases. Böhning et al. (2019) proposed a modification of Chao estimator to avoid overestimation due to one-inflation by incorporating counts of two and three. However, the modified Chao estimator produces a relatively large variance if compared to the maximum likelihood estimator. Böhning and Friedl (2021) addressed the population size estimator based on the unconditional likelihood with one-inflation model. The estimator via the profile likelihood performs well but the suggested approach is computationally intensive due to using grid of f_0 values and semi-parametric bootstrap. Intuitively, it is not easy for practitioners to use.

As an alternative model, we are interested in raising the UPA distribution in the capture-recapture context. The rest of the paper is organized as follow: the maximum likelihood and generalized Turing estimators for the zero-truncated uniform Poisson-Ailamujia are proposed in Section 2 and Section 3, respectively. The variance of proposed estimators for confidence interval estimation is derived in Section 4. Performance of the proposed estimators as well as the confidence interval are evaluated by simulation studies in Section 5. Two real datasets are used to demonstrate the proposed estimators in Section 6. The final section summarizes major results and remarks some points of work.

2. Population Size Estimator via Maximum Likelihood Estimation under the Uniform Poisson-Ailamujia Distribution

Since the count data from capture-recapture is non-zero, it can be modeled as a zero-truncated UPA distribution: $p_x^+ = \frac{2\alpha}{(1+2\alpha)^x}$. The likelihood function of count data X is

$$L(\alpha) = \prod_{x=1}^{m}\left(\frac{2\alpha}{(1+2\alpha)^x}\right)^{f_x}.$$

The log-likelihood function is

$$\log L(\alpha) = \sum_{x=1}^{m} f_x \log 2\alpha - \sum_{x=1}^{m} x f_x \log(1+2\alpha)$$

$$= n\log 2\alpha - S\log(1+2\alpha), \tag{3}$$

where $n = \sum_{x=1}^{m} f_x$ and $S = \sum_{x=1}^{m} xf_x$. The maximum likelihood estimator of α can be solved by maximizing (3) leads to the score-equation

$$\frac{n}{\alpha} = \frac{2S}{1+2\alpha},$$

which provided $\hat{\alpha}_{MLE} = \frac{n}{2(S-n)}$ and $\hat{p}_0 = \frac{n}{S}$. The maximum likelihood estimator under the UPA distribution using the Horvitz-Thomson approach is

$$\hat{N}_{MLEupa} = \frac{n}{1-\frac{n}{S}} = \frac{nS}{S-n}, \tag{4}$$

which is the same as the maximum likelihood estimator base on geometric distribution ($\hat{N}_{MLEgeo}$)(Niwitpong et al. 2013). An estimate variance for $\hat{N}_{MLEupa}$ is given in Appendix.

Theorem 2.1 *The $\hat{N}_{MLEupa}$ is asymptotically unbiased under the UPA distribution*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{MLEupa})}{N} \to 1.$$

Proof: If X has the UPA distribution, $E(X) = E(\frac{S}{N}) = \frac{1}{2\alpha}$ and $E(\frac{n}{N}) = 1 - p_0 = \frac{1}{1+2\alpha}$, $E(\frac{n}{S}) = E(\frac{n/N}{S/N}) = \frac{2\alpha}{1+2\alpha}$. $E(\frac{\hat{N}_{MLEupa}}{N}) = E(\frac{\frac{nS}{S-n}}{N}) = E(\frac{S}{N}\frac{\frac{n}{S}}{1-\frac{n}{S}}) \xrightarrow[\mathrm{N}\to\infty]{} (\frac{1}{2\alpha})(\frac{\frac{2\alpha}{1+2\alpha}}{1-\frac{2\alpha}{1+2\alpha}}) = 1$. This proves that $\hat{N}_{MLEupa}$ is asymptotically unbiased under the UPA distribution.

3. A generalized Turing estimator under the uniform Poisson-Ailamujia distribution

Let X denote the number of times a unit was identified during the research period. Assume that X has the UPA distribution, so $p_0 = \frac{2\alpha}{1+2\alpha}$, $p_1 = \frac{2\alpha}{(1+2\alpha)^2}$, and $E(X) = \frac{1}{2\alpha}$. Consider $p_0 = p_1(1+2\alpha) = p_1(1+\frac{1}{E(X)})$ that could be estimated by $\frac{f_1}{N}(1+\frac{N}{S}) = \frac{f_1}{N} + \frac{f_1}{S}$. Since $N = \frac{n}{1-p_0}$,

$$\begin{aligned}
p_0 &= \frac{f_1}{\frac{n}{1-p_0}} + \frac{f_1}{S} \\
p_0 - \frac{f_1(1-p_0)}{n} &= \frac{f_1}{S} \\
p_0 + \frac{f_1 p_0}{n} &= \frac{f_1}{S} + \frac{f_1}{n} \\
p_0(1+\frac{f_1}{n}) &= \frac{f_1}{S} + \frac{f_1}{n} \\
\hat{p}_0 &= \frac{nf_1 + Sf_1}{Sn + Sf_1}.
\end{aligned} \tag{5}$$

The generalized Turing estimator for the UPA distribution using the Horvitz-Thomson approach is

$$\hat{N}_{GTupa} = \frac{n}{1-\frac{nf_1+Sf_1}{nS+Sf_1}} = \frac{S(n+f_1)}{S-f_1}. \tag{6}$$

Theorem 3.1 *The $\hat{N}_{GTupa}$ is asymptotically unbiased under the UPA distribution*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{GTupa})}{N} \to 1.$$

Proof: Consider $E(X) = E(\frac{S}{N}) = \frac{1}{2\alpha}, E(\frac{f_1}{N}) = p_1 = \frac{2\alpha}{(1+2\alpha)^2}$, $E(\frac{n}{N}) = 1 - p_0 = \frac{1}{1+2\alpha}$, so that $E(\frac{f_1}{S}) = E(\frac{f_1/N}{S/N}) = \frac{4\alpha^2}{(1+2\alpha)^2}$ and $E(\frac{n}{S}) = E(\frac{n/N}{S/N}) = \frac{2\alpha}{1+2\alpha}$. Then, $E(\frac{\hat{N}_{GTupa}}{N}) = E(\frac{\frac{S(n+f_1)}{S-f_1}}{N}) = E(\frac{S(n+f_1)}{N(S-f_1)}) = E(\frac{S}{N}\frac{\frac{n}{S}+\frac{f_1}{S}}{1-\frac{f_1}{S}}) \xrightarrow[\mathrm{N}\to\infty]{} (\frac{1}{2\alpha})(\frac{2\alpha+8\alpha^2}{1+4\alpha}) = 1$. This proves that $\hat{N}_{GTupa}$ is asymptotically unbiased under the UPA distribution.

4. Variance of Estimates

To form the confidence interval (CI) for the true population size N, quantification of uncertainty surrounding the estimates of N is important. It is critical to assess the accuracy of the proposed estimators. Despite a precise estimate of N, if the associated estimation of variance is poor, then coverage of the associated CI may falsely indicate poor estimation by the point estimator. One might conclude that the point estimator results in a poor coverage rate.

In this section, the variance of $\hat{N}_{MLEupa}$ and $\hat{N}_{GTupa}$ are derived using a conditional technique (Böhning 2008) (for full details see Appendix). The estimated variance for $\hat{N}_{MLEupa}$ and $\hat{N}_{GTupa}$ are provided as follows:

$$\widehat{Var}(\hat{N}_{MLEupa}) = \frac{Sn^2}{(S-n)^2} + \frac{S^2n^3}{(S-n)^4} \tag{7}$$

$$\widehat{Var}(\hat{N}_{GTupa}) = \frac{Sf_1(n+f_1)(n+S)}{n(S-f_1)^2} + \frac{Sf_1(n+S)^2(Sn+f_1^2)}{(n+f_1)(S-f_1)^4} + \frac{Sf_1^2(n+f_1)(n+S)}{(S-f_1)^4}. \tag{8}$$

5. Simulation Study

Simulation was used to investigate the performance of the proposed estimators and confidence interval estimations. The proposed estimators are compared to the existing estimator developed using the geometric distribution as follows:

$$\hat{N}_{MLEgeo} = \frac{nS}{S-n} \tag{9}$$

$$\widehat{Var}(\hat{N}_{MLEgeo}) = \frac{S^2n^2}{(S-n)^3} \tag{10}$$

$$\hat{N}_{Chao} = n + \frac{f_1^2}{f_2} \tag{11}$$

$$\widehat{Var}(\hat{N}_{Chao}) = \frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2} \tag{12}$$

$$\hat{N}_{Censored} = \frac{n^2}{n-f_1} \tag{13}$$

$$\widehat{Var}(\hat{N}_{Censored}) = \frac{f_1}{(1-f_1/n)^2}\frac{2n-f_1}{n-f_1} \tag{14}$$

$$\hat{N}_{GTgeo} = \frac{n}{1-\sqrt{\frac{f_1}{S}}} \tag{15}$$

$$\widehat{Var}(\hat{N}_{GTgeo}) = \frac{n\sqrt{\frac{f_1}{S}}}{(1-\sqrt{\frac{f_1}{S}})^2} + n^2\left(\frac{S+f_1}{4S^2(1-\sqrt{\frac{f_1}{S}})^4}\right) \quad (16)$$

$$\hat{N}_{MC} = n + \frac{f_2^3}{f_3^2} \quad (17)$$

$$\widehat{Var}(\hat{N}_{MC}) = \left(\frac{\hat{f}_{0MC}^2}{f_2+f_3}\right)\left(1+\frac{(2f_2+3f_3)^2}{f_2f_3}\right); \hat{f}_{0MC} = \frac{f_2^3}{f_3^2} \quad (18)$$

$$\hat{N}_{BO} = n + (n-f_1)\frac{\hat{\theta}}{1-\hat{\theta}-\hat{\theta}(1-\hat{\theta})}; \hat{\theta} = \frac{n_1}{n_1+S^*},\ n_1 = n-f_1,\ S^* = \sum_{x=0}^{m-2} xf_{x+2} \quad (19)$$

$$\widehat{Var}(\hat{N}_{BO}) = n_1^2\frac{(1+\hat{\theta})^2}{(1-\hat{\theta})^6}\left(\frac{n_1}{\hat{\theta}^2}+\frac{S^*}{(1-\hat{\theta})^2}\right)^{-1} + \frac{n_1\hat{\theta}^3(2-\hat{\theta})}{(1-\hat{\theta}-\hat{\theta}(1-\hat{\theta}))^2} + n\hat{\theta}. \quad (20)$$

The population sizes were set at $N = 100,500$ for small populations, $N = 1,000$ for medium populations, and $N = 5,000,\ 10,000$ for large populations. Data were generated from the UPA distribution to assess the estimator's efficiency. Also, we are interested in a setting where data did not follow UPA. Data from the Conway-Maxwell Poisson (CMP) and Poisson-Lindley (PL) distributions were generated to investigate the estimator's effectiveness under misspecification. The two-parameter CMP distribution is defined by $p_x = \frac{\lambda^x}{(x!)^\nu}\frac{1}{Z(\lambda,\nu)}$ where $Z(\lambda,\nu) = \sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}$ is a normalizing constant. The CMP distribution includes Poisson $(\nu = 1)$, Bernoulli $(\nu \to 1)$ and geometric $(\nu = 0)$ distributions as special cases. The PL distribution representing long-tailed count data is given by $p_x = \frac{\theta^2(x+\theta+2)}{(\theta+1)^{x+3}}, \theta \geq 0$. Both CMP and PL distributions allow for underdispersion and overdispersion. Therefore, the data were generate using three settings as follows:

- The UPA distribution with $\alpha = 0.05, 0.10$ for mild one inflation and $\alpha = 0.25$ for strong one inflation.
- The CMP distribution with mean parameter $\lambda = 0.5$ and dispersion parameter $\nu = 0.1, 0.5, 0.8$.
- The PL distribution $\theta = 0.3, 0.6, 1.0$ with 5% one-inflation.

5.1. Simulation result of estimators

To assess the performance of the proposed estimators, the relative bias (Rbias) and relative root mean squared error (RRMSE) were calculated for each scenario:

$$\text{Rbias} = \frac{E(\hat{N})-N}{N}$$

$$\text{RRMSE} = \frac{1}{N}\sqrt{Var(\hat{N})+(E(\hat{N})-N)^2}.$$

Here, $E(\hat{N}) = \frac{1}{10,000}\sum_{t=1}^{10,000}\hat{N}_{(t)}$, $Var(\hat{N}) = \frac{1}{10,000}\sum_{t=1}^{10,000}\{\hat{N}_{(t)}-E(\hat{N})\}^2$, and $\hat{N}_{(t)}$ denotes the estimated values of the population size at replication t.

Table 2 Rbias of estimators following the UPA distribution

α	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.05	0.0009	0.0279	0.0004	-0.0010	0.1695	-0.0164	0.0009	-0.0004
0.10	0.0026	0.0294	0.0021	0.0006	0.1738	-0.0554	0.0026	0.0007
0.25	0.0074	0.0394	0.0061	0.0043	0.2823	-0.2117	0.0074	0.0037
				$N = 500$				
0.05	0.0002	0.0056	0.0005	0.0001	0.0149	-0.0168	0.0002	0.0003
0.10	0.0008	0.0054	0.0006	0.0003	0.0196	-0.0564	0.0008	0.0003
0.25	0.0017	0.0083	0.0017	0.0012	0.0333	-0.2124	0.0017	0.0011
				$N = 1,000$				
0.05	0.0000	0.0022	0.0000	-0.0001	0.0081	-0.0169	0.0000	-0.0001
0.10	0.0002	0.0025	0.0002	0.0001	0.0104	-0.0567	0.0002	0.0001
0.25	0.0006	0.0040	0.0007	0.0004	0.0160	-0.2126	0.0006	0.0004
				$N = 5,000$				
0.05	0.0000	0.0004	0.0000	0.0000	0.0012	-0.0169	0.0000	0.0000
0.10	0.0000	0.0004	0.0000	0.0000	0.0018	-0.0568	0.0000	-0.0001
0.25	0.0065	0.0034	0.0022	0.0042	0.0062	-0.2082	0.0065	0.0032
				$N = 10,000$				
0.05	0.0000	0.0001	0.0000	0.0000	0.0007	-0.0170	0.0000	0.0000
0.10	0.0001	0.0004	0.0001	0.0001	0.0010	-0.0568	0.0001	0.0001
0.25	0.0000	0.0004	0.0000	0.0000	0.0016	-0.2126	0.0000	0.0000

- In the UPA distribution case, all estimators are asymptotically unbiased except $\hat{N}_{BO}$ (see Table 2). All estimators except $\hat{N}_{BO}$ overestimate for small population sizes and become less biased as population sizes increase. $\hat{N}_{BO}$ is underestimated, especially for strong one inflation. Table 3 shows the RRMSE, which is used to compare the performance of all estimators. The results show that $\hat{N}_{MLEgeo}$ and $\hat{N}_{MLEupa}$ are likely to be the best choices, as they provide the smallest RRMSE for all cases. The RRMSE of $\hat{N}_{Censored}$, $\hat{N}_{GTgeo}$and $\hat{N}_{GTupa}$ are all relatively small, and $\hat{N}_{GTupa}$ is a reasonable compromise between $\hat{N}_{Censored}$ and $\hat{N}_{GTgeo}$.

- In the CMP distribution case, all estimators except $\hat{N}_{BO}$ overestimate. $\hat{N}_{BO}$ is underestimated (see Table 4). The RRMSE is shown in Table 5 and is used to compare the performance of all estimators. For a small population, $\hat{N}_{GTupa}$ appears to be an appropriate choice. $\hat{N}_{Chao}$ is a good option for medium and large populations. Rbias and RRMSE decrease as ν decreases, because the CMP distribution contains the geometric or UPA distribution as a special boundary case. All estimators based on geometric or UPA distributions perform better for small ν in CMP distribution.

- In the PL distribution with 5% one-inflation case, almost all estimators overestimate. $\hat{N}_{BO}$ is underestimated (see Table 6) for $\theta = 0.6, 1.0$. Table 7 shows the RRMSE, which is used to compare the performance of all estimators. $\hat{N}_{BO}$ appears to be a good choice for small and medium populations with $\theta = 0.3, 0.6$. For a large population, $\hat{N}_{MC}$ appears to be a good option. Except for $\hat{N}_{Chao}$, the other estimators produced similar results with low RRMSE. $\hat{N}_{GTupa}$ represents a reasonable compromise between $\hat{N}_{Censored}$ and $\hat{N}_{GTgeo}$. All estimators' Rbias and RRMSE values increase as θ increases.

Table 3 RRMSE of estimators following the UPA distribution

α	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.05	0.0328	0.1194	0.0459	0.0366	0.9370	0.0357	0.0328	0.0438
0.10	0.0491	0.1451	0.0666	0.0540	1.0412	0.0707	0.0491	0.0621
0.25	0.0878	0.2123	0.1129	0.0942	1.1850	0.2196	0.0878	0.1018
				$N = 500$				
0.05	0.0150	0.0373	0.0206	0.0166	0.0717	0.0222	0.0150	0.0198
0.10	0.0219	0.0511	0.0295	0.0240	0.0987	0.0597	0.0219	0.0276
0.25	0.0390	0.0811	0.0501	0.0420	0.1795	0.2140	0.0390	0.0455
				$N = 1,000$				
0.05	0.0105	0.0257	0.0145	0.0117	0.0452	0.0198	0.0105	0.0140
0.10	0.0155	0.0354	0.0211	0.0171	0.0652	0.0584	0.0155	0.0197
0.25	0.0278	0.0564	0.0357	0.0300	0.1165	0.2134	0.0278	0.0325
				$N = 5,000$				
0.05	0.0047	0.0110	0.0065	0.0052	0.0177	0.0175	0.0047	0.0063
0.10	0.0069	0.0155	0.0093	0.0076	0.0269	0.0571	0.0069	0.0088
0.25	0.0109	0.0234	0.0133	0.0096	0.0486	0.2082	0.0109	0.0111
				$N = 10,000$				
0.05	0.0033	0.0078	0.0046	0.0037	0.0125	0.0173	0.0033	0.0044
0.10	0.0050	0.0110	0.0067	0.0054	0.0188	0.0569	0.0050	0.0063
0.25	0.0086	0.0177	0.0111	0.0093	0.0342	0.2127	0.0086	0.0101

Table 4 Rbias of estimators following the CMP distribution $\lambda = 0.5$

ν	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.1	0.1023	0.1130	0.0780	0.0855	1.3318	-0.4514	0.1023	0.0813
0.5	0.4380	0.3693	0.3615	0.3929	3.2313	-0.5320	0.4380	0.3848
0.8	0.7007	0.6614	0.6228	0.6540	3.3008	-0.5649	0.7007	0.6467
				$N = 500$				
0.1	0.0814	0.0532	0.0588	0.0691	0.1852	-0.4525	0.0814	0.0657
0.5	0.3974	0.2626	0.3158	0.3540	1.0881	-0.5340	0.3974	0.3462
0.8	0.6684	0.4753	0.5562	0.6087	2.8192	-0.5697	0.6684	0.6005
				$N = 1,000$				
0.1	0.0777	0.0440	0.0550	0.0658	0.1367	-0.4531	0.0777	0.0624
0.5	0.3883	0.2479	0.3061	0.3451	0.8262	-0.5343	0.3883	0.3373
0.8	0.6572	0.4581	0.5449	0.5981	1.8952	-0.5701	0.6572	0.5899
				$N = 5,000$				
0.1	0.0755	0.0382	0.0526	0.0638	0.0963	-0.4531	0.0755	0.0605
0.5	0.3831	0.2393	0.3014	0.3406	0.6742	-0.5344	0.3831	0.3329
0.8	0.6482	0.4440	0.5353	0.5893	1.4754	-0.5701	0.6482	0.5811
				$N = 10,000$				
0.1	0.0758	0.0381	0.0529	0.0641	0.0899	-0.4531	0.0758	0.0608
0.5	0.3829	0.2389	0.3013	0.3405	0.6576	-0.5344	0.3829	0.3328
0.8	0.6479	0.4439	0.5354	0.5893	1.4252	-0.5701	0.6479	0.5811

Table 5 RRMSE of estimators following the CMP distribution $\lambda = 0.5$

ν	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.1	0.2072	0.3766	0.2280	0.2067	5.2190	0.4555	0.2072	0.2102
0.5	0.5439	0.6570	0.5144	0.5166	7.9771	0.5349	0.5439	0.5138
0.8	0.8345	1.1330	0.8295	0.8168	6.7757	0.5673	0.8345	0.8145
				$N = 500$				
0.1	0.1115	0.1397	0.1074	0.1055	0.4690	0.4533	0.1115	0.1053
0.5	0.4196	0.3222	0.3486	0.3799	2.2961	0.5346	0.4196	0.3734
0.8	0.6960	0.5379	0.5941	0.6401	8.0861	0.5702	0.6960	0.6329
				$N = 1,000$				
0.1	0.0942	0.1009	0.0841	0.0865	0.3023	0.4535	0.0942	0.0853
0.5	0.3991	0.2802	0.3230	0.3580	1.1268	0.5346	0.3991	0.3510
0.8	0.6705	0.4882	0.5630	0.6132	2.6686	0.5703	0.6705	0.6055
				$N = 5,000$				
0.1	0.0792	0.0550	0.0597	0.0686	0.1453	0.4532	0.0792	0.0658
0.5	0.3853	0.2461	0.3048	0.3433	0.7259	0.5345	0.3853	0.3357
0.8	0.6509	0.4502	0.5390	0.5923	1.5785	0.5701	0.6509	0.5842
				$N = 10,000$				
0.1	0.0776	0.0474	0.0565	0.0665	0.1167	0.4532	0.0776	0.0635
0.5	0.3840	0.2424	0.3030	0.3418	0.6847	0.5344	0.3840	0.3342
0.8	0.6492	0.4470	0.5373	0.5908	1.4736	0.5701	0.6492	0.5827

Table 6 Rbias of estimators following the PL distribution with 5% one-inflation

θ	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.3	0.0879	0.1741	0.0908	0.0881	0.1474	0.0285	0.0879	0.0893
0.6	0.1199	0.2050	0.1300	0.1229	0.2189	-0.0805	0.1199	0.1254
1.0	0.1550	0.2657	0.1777	0.1630	0.4311	-0.2490	0.1550	0.1668
				$N = 500$				
0.3	0.0859	0.1415	0.0888	0.0871	0.0340	0.0273	0.0859	0.0882
0.6	0.1169	0.1678	0.1253	0.1207	0.0664	-0.0804	0.1169	0.1229
1.0	0.1452	0.2119	0.1658	0.1547	0.0997	-0.2496	0.1452	0.1586
				$N = 1,000$				
0.3	0.0852	0.1377	0.0885	0.0867	0.0276	0.0268	0.0852	0.0879
0.6	0.1164	0.1639	0.1251	0.1206	0.0546	-0.0803	0.1164	0.1229
1.0	0.1435	0.2067	0.1650	0.1538	0.0793	-0.2495	0.1435	0.1579
				$N = 5,000$				
0.3	0.0854	0.1346	0.0883	0.0868	0.0220	0.0270	0.0854	0.0879
0.6	0.1162	0.1609	0.1248	0.1204	0.0446	-0.0802	0.1162	0.1228
1.0	0.1426	0.2019	0.1637	0.1529	0.0603	-0.2495	0.1426	0.1570
				$N = 10,000$				
0.3	0.0853	0.1342	0.0883	0.0868	0.0214	0.0269	0.0853	0.0879
0.6	0.1161	0.1604	0.1246	0.1203	0.0429	-0.0803	0.1161	0.1226
1.0	0.1424	0.2014	0.1635	0.1527	0.0578	-0.2495	0.1424	0.1568

5.2. Simulation result of confidence interval estimations

The 95% confidence interval estimations of N are constructed using the normal approximation approach based on the population size estimators and the estimated variances as: $\hat{N} \pm 1.96\sqrt{\widehat{Var}(\hat{N})}$. Estimation performance is measured using coverage probability (CP) and average length (AL), which

Table 7 RRMSE of estimators following the PL distribution with 5% one-inflation

θ	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.3	0.0958	0.2457	0.1056	0.0976	0.6032	0.0446	0.0958	0.1022
0.6	0.1393	0.2954	0.1618	0.1451	0.7214	0.0955	0.1393	0.1518
1.0	0.1951	0.4026	0.2347	0.2067	1.7595	0.2560	0.1951	0.2141
				$N = 500$				
0.3	0.0875	0.1526	0.0919	0.0891	0.0764	0.0312	0.0875	0.0909
0.6	0.1212	0.1863	0.1324	0.1256	0.1475	0.0836	0.1212	0.1287
1.0	0.1540	0.2405	0.1785	0.1642	0.2455	0.2510	0.1540	0.1690
				$N = 1,000$				
0.3	0.0861	0.1434	0.0901	0.0878	0.0534	0.0289	0.0861	0.0894
0.6	0.1186	0.1731	0.1286	0.1229	0.1026	0.0819	0.1186	0.1257
1.0	0.1480	0.2215	0.1715	0.1586	0.1668	0.2502	0.1480	0.1631
				$N = 5,000$				
0.3	0.0855	0.1357	0.0886	0.0870	0.0291	0.0274	0.0855	0.0882
0.6	0.1166	0.1628	0.1255	0.1209	0.0577	0.0806	0.1166	0.1233
1.0	0.1435	0.2049	0.1650	0.1539	0.0865	0.2497	0.1435	0.1581
				$N = 10,000$				
0.3	0.0854	0.1348	0.0885	0.0869	0.0252	0.0271	0.0854	0.0880
0.6	0.1163	0.1614	0.1250	0.1206	0.0501	0.0804	0.1163	0.1229
1.0	0.1429	0.2029	0.1641	0.1532	0.0722	0.2495	0.1429	0.1574

are defined as follows:

$$\text{CP} = \frac{\sum_{t=1}^{10,000} C_{(t)}}{10,000},$$

where $C_{(t)}$ equals 1 if the true population size N is within the confidence interval and 0 otherwise.

$$\text{AL} = \frac{\sum_{t=1}^{10,000} (\hat{N}_{U_{(t)}} - \hat{N}_{L_{(t)}})}{10,000},$$

where $\hat{N}_{U_{(t)}}$ and $\hat{N}_{L_{(t)}}$ are the upper and lower estimates of N at replication t, respectively.

- In the UPA distribution case (see Table 8), the coverage probabilities of almost all estimators, except $\hat{N}_{MC}$ and $\hat{N}_{BO}$, close to the nominal level. The coverage probabilities of $\hat{N}_{MC}$ are lower than the nominal level and converge to nominal level as N increases, and $\hat{N}_{MC}$ provides the longest average length. Because the average lengths of $\hat{N}_{MLEupa}$ are slightly longer than those of $\hat{N}_{MLEgeo}$, $\hat{N}_{MLEupa}$ has higher coverage probabilities than $\hat{N}_{MLEgeo}$. Similarly, the average lengths of $\hat{N}_{GTupa}$ are slightly longer than those of $\hat{N}_{GTgeo}$, so the coverage probabilities of $\hat{N}_{GTupa}$ are greater than those of $\hat{N}_{GTgeo}$.

- With the exception of $\hat{N}_{BO}$, almost all estimators' coverage probabilities are close to the nominal level for $N = 100$ in the CMP distribution (see Table 9). For $N = 500$ and $1,000$, $\hat{N}_{MC}$ has the highest coverage probabilities and the longest average lengths. $\hat{N}_{MLEgeo}$ provides relatively low coverage compared to $\hat{N}_{MLEupa}$ for small and medium population, just as $\hat{N}_{GTgeo}$ provides relatively low coverage compared to $\hat{N}_{GTupa}$. As a result, the proposed confidence intervals can improve estimates. Except for $\hat{N}_{Chao}$ and $\hat{N}_{MC}$ when $\nu = 0.1$, all estimators provide relatively

poor coverage probabilities for large population.

- Coverage probabilities of $\hat{N}_{Chao}$ greater than the nominal level in the PL distribution with 5% one-inflation for $N = 100$. For $N = 100$ and $\theta = 0.3$, coverage probabilities of $\hat{N}_{BO}$ greater than the nominal level. Coverage probabilities of $\hat{N}_{Censored}$, $\hat{N}_{MLEupa}$, and $\hat{N}_{GTupa}$ close to the nominal level for $N = 100$ and $\theta = 1.0$. Coverage probabilities of $\hat{N}_{MC}$ close to the nominal level for $N = 500$ and $1,000$, but the confidence intervals of $\hat{N}_{MC}$ are the widest, followed by $\hat{N}_{Chao}$. The average lengths of the other estimators are quite short.

As the UPA distribution is a re-parameterized geometric distribution, the simulation results show that $\hat{N}_{MLEupa}$ and $\hat{N}_{MLEgeo}$ provide the same Rbias and RRMSE. However, it was found that $\hat{N}_{MLEupa}$ has a higher variance than $\hat{N}_{MLEgeo}$, concluding that the coverage probability is greater in all situations.

Table 8 Coverage probability (Average length) of estimators following the UPA distribution

ν	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.05	0.95(13.08)	0.89(39.15)	0.93(17.81)	0.94(14.08)	0.83(155.83)	0.93(13.79)	0.95(13.15)	0.93(16.98)
0.10	0.95(19.37)	0.91(50.06)	0.94(26.04)	0.94(20.54)	0.85(161.11)	0.76(17.67)	0.96(19.71)	0.94(24.15)
0.25	0.95(34.66)	0.92(77.07)	0.94(44.50)	0.94(35.86)	0.86(265.84)	0.04(20.49)	0.97(37.66)	0.94(39.92)
				$N = 500$				
0.05	0.95(29.10)	0.94(71.44)	0.95(40.18)	0.94(31.66)	0.90(123.75)	0.83(30.70)	0.95(29.24)	0.95(38.50)
0.10	0.95(43.05)	0.94(98.90)	0.95(58.19)	0.94(45.92)	0.91(178.09)	0.20(39.37)	0.95(43.77)	0.95(54.24)
0.25	0.95(76.26)	0.94(157.54)	0.95(98.39)	0.94(79.29)	0.91(324.92)	0.00(45.64)	0.96(82.47)	0.95(88.58)
				$N = 1,000$				
0.05	0.95(41.14)	0.94(98.11)	0.95(56.76)	0.95(44.76)	0.92(163.79)	0.68(43.42)	0.95(41.33)	0.95(54.44)
0.10	0.95(60.78)	0.95(137.87)	0.95(82.26)	0.94(64.90)	0.93(239.61)	0.02(55.61)	0.95(61.79)	0.95(76.71)
0.25	0.95(107.52)	0.95(220.18)	0.95(138.82)	0.94(111.88)	0.93(436.76)	0.00(64.49)	0.96(116.20)	0.95(125.04)
				$N = 5,000$				
0.05	0.95(91.94)	0.95(215.67)	0.95(126.97)	0.94(100.11)	0.94(340.95)	0.06(97.05)	0.95(92.35)	0.95(121.82)
0.10	0.95(135.82)	0.95(305.20)	0.95(183.84)	0.94(145.09)	0.94(509.74)	0.00(124.32)	0.96(138.07)	0.95(171.55)
0.25	0.98(241.97)	0.97(487.63)	0.98(309.17)	0.99(250.66)	0.95(939.75)	0.00(145.27)	0.99(261.61)	0.99(279.62)
				$N = 10,000$				
0.05	0.95(130.02)	0.95(304.27)	0.95(179.59)	0.95(141.59)	0.94(479.63)	0.00(137.25)	0.95(130.61)	0.95(172.32)
0.10	0.95(192.09)	0.95(431.41)	0.95(260.12)	0.94(205.24)	0.94(717.44)	0.00(175.81)	0.95(195.27)	0.95(242.72)
0.25	0.95(339.52)	0.95(689.10)	0.95(438.29)	0.94(353.36)	0.95(1,321.58)	0.00(203.92)	0.97(366.74)	0.95(395.05)

Table 9 Coverage probability (Average length) of estimators following the CMP distribution $\lambda = 0.5$

ν	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.1	0.97(71.87)	0.94(127.89)	0.96(82.98)	0.97(77.41)	0.87(1,167.28)	0.00(17.18)	0.99(96.35)	0.97(80.77)
0.5	0.95(134.62)	0.98(192.69)	0.99(141.12)	1.00(163.51)	0.92(2,901.07)	0.00(14.08)	1.00(226.93)	1.00(165.92)
0.8	0.94(190.02)	0.99(272.09)	0.99(198.71)	1.00(262.00)	0.93(3,171.35)	0.00(12.31)	1.00(367.21)	1.00(264.20)
				$N = 500$				
0.1	0.88(153.92)	0.96(253.93)	0.94(178.25)	0.93(164.23)	0.94(768.09)	0.00(38.29)	0.98(201.37)	0.94(171.94)
0.5	0.12(283.47)	0.84(368.55)	0.44(294.20)	0.40(330.49)	0.99(2,389.71)	0.00(31.33)	0.82(457.96)	0.47(335.84)
0.8	0.01(408.08)	0.56(483.40)	0.13(408.21)	0.17(523.35)	1.00(6,787.60)	0.00(27.17)	0.89(755.28)	0.21(527.43)
				$N = 1,000$				
0.1	0.75(216.48)	0.95(353.36)	0.90(250.63)	0.85(230.68)	0.97(1,000.75)	0.00(54.12)	0.93(282.45)	0.88(241.60)
0.5	0.00(396.15)	0.56(510.94)	0.10(410.69)	0.04(459.24)	1.00(2,684.47)	0.00(44.34)	0.11(635.85)	0.06(466.80)
0.8	0.00(569.77)	0.14(669.96)	0.00(569.46)	0.00(724.98)	1.00(5,998.95)	0.00(38.40)	0.02(1,045.89)	0.00(730.75)
				$N = 5,000$				
0.1	0.12(481.89)	0.87(781.46)	0.55(557.81)	0.30(512.96)	0.93(2,083.02)	0.00(121.03)	0.29(627.34)	0.40(537.40)
0.5	0.00(879.10)	0.00(1,128.15)	0.00(911.55)	0.00(1,015.78)	0.14(5,219.32)	0.00(99.12)	0.00(1,404.73)	0.00(1,032.75)
0.8	0.00(1,260.35)	0.00(1,472.95)	0.00(1,258.35)	0.00(1,593.50)	0.01(10,550.22)	0.00(85.92)	0.00(2,298.62)	0.00(1,606.37)
				$N = 10,000$				
0.1	0.00(681.67)	0.75(1,104.75)	0.24(789.02)	0.05(725.60)	0.84(2,912.40)	0.00(171.15)	0.02(887.36)	0.09(760.18)
0.5	0.00(1,242.59)	0.00(1,594.11)	0.00(1,288.63)	0.00(1,435.44)	0.00(7,264.65)	0.00(140.16)	0.00(1,984.71)	0.00(1,459.46)
0.8	0.00(1,781.10)	0.00(2,082.00)	0.00(1,779.01)	0.00(2,251.30)	0.00(14,473.64)	0.00(121.48)	0.00(3,246.32)	0.00(2,269.56)

Table 10 Coverage probability (Average length) of estimators following the PL distribution with 5% one-inflation

θ	$\hat{N}_{MLEgeo}$	$\hat{N}_{Chao}$	$\hat{N}_{Censored}$	$\hat{N}_{GTgeo}$	$\hat{N}_{MC}$	$\hat{N}_{BO}$	$\hat{N}_{MLEupa}$	$\hat{N}_{GTupa}$
				$N = 100$				
0.3	0.61(19.67)	1.00(67.16)	0.87(26.86)	0.68(21.04)	0.88(114.4897)	0.98(18.1171)	0.63(19.99)	0.82(24.93)
0.6	0.77(32.63)	1.00(84.99)	0.91(43.44)	0.78(34.26)	0.89(174.7998)	0.68(21.3377)	0.82(34.75)	0.85(38.79)
1.0	0.87(50.44)	0.99(116.90)	0.94(65.42)	0.88(53.68)	0.88(359.1516)	0.01(20.2592)	0.96(59.01)	0.91(58.28)
				$N = 500$				
0.3	0.00(43.76)	0.37(130.30)	0.08(59.92)	0.01(46.98)	0.95(124.6446)	0.81(40.3912)	0.00(44.44)	0.05(55.82)
0.6	0.06(72.32)	0.57(172.57)	0.21(96.06)	0.08(76.05)	0.96(242.9618)	0.09(47.6781)	0.08(76.83)	0.14(86.23)
1.0	0.22(110.56)	0.65(236.40)	0.35(143.09)	0.22(117.48)	0.95(413.7706)	0.00(45.2807)	0.35(128.20)	0.28(127.84)
				$N = 1{,}000$				
0.3	0.00(61.80)	0.03(181.39)	0.00(84.75)	0.00(66.41)	0.97(167.6844)	0.56(57.0557)	0.00(62.75)	0.00(78.96)
0.6	0.00(102.11)	0.14(241.39)	0.02(135.73)	0.00(107.44)	0.97(328.6457)	0.00(67.3927)	0.00(108.44)	0.01(121.88)
1.0	0.02(155.74)	0.25(330.76)	0.05(201.89)	0.02(165.59)	0.97(556.7300)	0.00(63.9985)	0.05(180.34)	0.03(180.29)
				$N = 5{,}000$				
0.3	0.00(138.17)	0.00(400.03)	0.00(189.33)	0.00(148.45)	0.84(356.9896)	0.00(127.6087)	0.00(140.29)	0.00(176.50)
0.6	0.00(228.13)	0.00(535.54)	0.00(303.22)	0.00(240.10)	0.83(706.9182)	0.00(150.6790)	0.00(242.20)	0.00(272.40)
1.0	0.00(347.57)	0.00(733.11)	0.00(450.38)	0.00(369.50)	0.89(1187.2977)	0.00(143.1100)	0.00(402.16)	0.00(402.38)
				$N = 10{,}000$				
0.3	0.00(195.37)	0.00(564.78)	0.00(267.79)	0.00(209.95)	0.66(502.3754)	0.00(180.4411)	0.00(198.37)	0.00(249.63)
0.6	0.00(322.56)	0.00(756.50)	0.00(428.67)	0.00(339.47)	0.64(993.4746)	0.00(213.0924)	0.00(342.44)	0.00(385.15)
1.0	0.00(491.29)	0.00(1035.78)	0.00(636.62)	0.00(522.27)	0.78(1668.3064)	0.00(202.3949)	0.00(568.35)	0.00(568.77)

6. Applications

In this section, two real datasets are used to demonstrate the suitability of the proposed estimator in comparison to existing estimators. The daily numbers of COVID-19 deaths in Switzerland between 1 March to 30 June 2021, where the number of zeros is known. The total number of COVID-19 deaths is 121. As seen in Table 1, there are zero daily deaths for 15 days, one daily death for 11 days, two daily deaths for 12 days, and so on. The maximum likelihood estimation was used to estimate the parameters under the UPA and geometric distributions, where $\hat{\alpha} = \frac{n}{2(S-n)}$ and $\hat{p} = \frac{n}{S}$. The total number of deaths estimated is presented in Table 11. $\hat{N}_{MLEgeo}$, $\hat{N}_{GTgeo}$, $\hat{N}_{MLEupa}$, and $\hat{N}_{GTupa}$ provide estimations close to total number. Moreover, the MLE of the UPA distribution parameter is 0.06, which is relatively low and indicates the absence of one-inflation, that is consistent with the observed data. The confidence intervals for all estimators include the true population size. However, $\hat{N}_{Chao}$ and $\hat{N}_{MC}$ are inappropriate because the lower bounds are less than the observed count.

Table 11 COVID-19 deaths data: population size estimates ($n = 106, N = 121$)

Estimator	$\hat{N}$	$\widehat{SE(\hat{N})}$	95%CI
MLEgeo($\hat{\theta} = 0.11$)	119	4.11	(111 ,127)
Chao	116	7.45	(101 , 131)
Censored	118	5.38	(108 , 129)
GTgeo	119	4.37	(110 , 127)
MC	133	30.19	(74 , 192)
BO	116	4.21	(108 , 125)
MLEupa($\hat{\alpha} = 0.06$)	119	5.17	(108, 129)
GTupa	119	4.14	(111 , 127)

Another area of interest is determining the size of a population with addiction problems. Cruyff MJ and Van Der Heijden (2008) provide data on the number of applications for a methadone treatment programme made by opiate users in Rotterdam. Frequency distribution of opiate users is provided in

Table 12. The maximum likelihood method was used to estimate the UPA and geometric distribution parameters, as was done with the COVID-19 deaths data. The population size of opiate users in Rotterdam is estimated in Table 13. The Chi-square test provided $\chi^2 = 5.4412, df = 6$ ($p-value = 0.4886$) for the zero-truncated UPA distribution with $\hat{\alpha} = 0.71$. At a 0.05 level of significance, there was sufficient evidence to conclude that the distribution of the opiate users in Rotterdam was not different from the zero-truncated UPA distribution with $\hat{\alpha} = 0.71$. Clearly, $\hat{\alpha}$ is rather high indicating that the data has one inflation. This corresponds to the observed data, with one count accounting for approximately 60% of the total observed data.

Table 12 Frequencies of opiate users in Rotterdam

i	0	1	2	3	4	5	6	7	8	9	10	n
f_i	-	1,206	474	198	95	29	19	5	2	0	1	2,029

Table 13 Estimates of the population size of opiate users in Rotterdam

Estimator	$\hat{N}$	$\widehat{SE(\hat{N})}$	95%CI
MLEgeo($\hat{\theta} = 0.59$)	4,930	130.88	(4,674 , 5,187)
Chao	5,097	232.72	(4,641 , 5,554)
Censored	5,002	159.38	(4,690 , 5,315)
GTgeo	4,966	147.12	(4,678 , 5,254)
MC	4,745	537.76	(3,691 , 5,799)
BO	2,272	26.64	(2,220 , 2,325)
MLEupa($\hat{\alpha} = 0.71$)	4,930	177.60	(4,582 , 5,278)
GTupa	4,975	153.74	(4,674 , 5,276)

7. Discussion and Conclusion

This study introduced two population size estimators based on a re-parameterized geometric distribution known as the UPA distribution that is well-suited for modeling over-dispersed count data. The first estimator presented was the maximum likelihood estimator for the UPA distribution. The generalized Turing estimator for the UPA distribution was the second. We evaluate the proposed estimators' accuracy and precision by comparing them to existing estimators. In the UPA scenario, the proposed estimators appear to be accurate as well as having low bias and mean square error. The proposed estimators also perform well despite misspecification. For example, the case of CMP with small ν indicating overdispersion. In addition, the proposed estimators provide reasonable estimates based on PL distribution with 5% one-inflation, which are not conducive to estimators developed under UPA or geometric distributions. We also provide the variance approximation formula for the proposed estimators to construct confidence intervals. Although the proposed estimators have a slightly higher variance than the maximum likelihood and generalized Turing based on geometric distribution, the presented confidence intervals can improve the coverage probabilities.

The results from real data examples need some comments. According to daily counts of COVID-19 deaths in Switzerland with known f_0. The UPA parameter estimate $\hat{\alpha}$ was relatively low indicating that the data was not one inflation. Furthermore, using data of opiate users in Rotterdam with unknown f_0, it was discovered that the estimated parameter $\hat{\alpha}$ was quite high. This indicates that the

data has one inflation corresponding to the observed data, and that f_1 was found to be greater than half the number of observed data. In conclusion, the proposed estimators are an appropriate choice for estimating the population size of overdispersion data. It has the advantage of being simple to calculate and can point out how much of the data is one-inflated based on the estimated parameters $\hat{\alpha}$.

Acknowledgments

The authors gratefully acknowledge the editor and referees for their valuable comments and suggestions which greatly improve this paper. The authors gratefully acknowledge the financial support provided by Faculty of Science and Technology Thammasat University, Contract No SciGR 9/2565.

References

Aljohani HM, Akdoğan Y, Cordeiro GM, Afify AZ. The uniform Poisson– Ailamujia distribution: Actuarial measures and applications in biological science. Symmetry. 2021; 13(7):1258.

Anan O, Böhning D, Maruotti A. On the Turing estimator in capture–recapture count data under the geometric distribution. Metrika. 2019; 82(2): 149-172.

Böhning D. A simple variance formula for population size estimators by conditioning. Stat Methodol. 2008; 5(5): 410-423.

Böhning D, Friedl H. Population size estimation based upon zero-truncated, one-inflated and sparse count data. Stat Methods Appt. 2021; 30(4): 1197-217.

Böhning D, Kaskasamkul P, van der Heijden PG. A modification of Chao's lower bound estimator in the case of oneinflation. Metrika. 2019; 82(3): 361-384.

Böhning D, Lerdsuwansri R, Sangnawakij P. Modelling Covid-19 contact-tracing using the ratio regression capture-recapture approach. Biometrics. 2023: 1-13.

Böhning D, Ogden HE. General flation models for count data. Metrika. 2021; 84(2): 245-261.

Cruyff MJ, Van Der Heijden PG. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. Biom J. 2008; 50(6): 1035-1050.

Godwin RT. One-inflation and unobserved heterogeneity in population size estimation. Biom J. 2017; 59(1): 79-93.

Godwin RT, Böhning D. Estimation of the population size by using the one-inflated positive Poisson model. J R Stat Soc Ser C Appl Stat. 2017; 66(2): 425-448.

Jongsomjit T, Lerdsuwansri R, Lanumteang K. Estimation of population size based on zero-truncated,one-inflated and covariate information. In: Proceedings of the 2nd International Conference on Science Technology and Innovation Maejo University; 2022 Mar 18; Thailand. 2022. pp. 29-35.

Lerdsuwansri R, Sangnawakij P, Böhning D, Sansilapin C, Chaifoo W, Polonsky JA, et al. Sensitivity of contact-tracing for COVID-19 in Thailand: a capturere capture application. BMC Infect Dis. 2022; 22(1): 1-10.

Liuzzo M, Borella S, Ottonello D, Arizza V, Malavasi S. Population abundance,structure and movements of the European pond turtle, Emysorbicularis (Linnaeus1758) based on capture-recapture data in a Venice Lagoon wetland area, Italy. Ethol Ecol Evol. 2021; 33(6): 561-75.

Nguyen LT, Patel S, Nguyen NT, Gia HH, Raymond HF, Abdul-Quader AS, et al. Population Size Estimation of Female Sex Workers in Hai Phong, Vietnam:Use of Three Source Capture–RecaptureMethod. J Epidemiol Glob Health. 2021; 11(2): 194.

Niwitpong Sa, Böhning D, van der Heijden PG, Holling H. Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity. Metrika. 2013; 76(4): 495-519.

Pijitrattana P. A flexible, discrete and smooth capture-recapture model based upon counts of repeated identifications using validation samples. PhD [dissertation]. Thammasat University; 2018.

Sansamur C, Wiratsudakul A, Charoenpanyanet A, Punyapornwithaya V. Estimating the number of farms experienced foot and mouth disease outbreaks using capture-recapture methods. Trop Anim Health Prod. 2021; 53(1): 1-9.

Tajuddin RRM, Ismail N. Comment on Aljohani et al. The Uniform Poisson– Ailamujia Distribution: Actuarial Measures and Applications in Biological Science. Symmetry 2021, 13, 1258. Symmetry. 2022; 14(1): 121.

Van Der Heijden PG, Bustami R, Cruyff MJ, Engbersen G, Van Houwelingen HC. Point and interval estimation of the population size using the truncated Poisson regression model. Stat Modelling. 2003; 3(4): 305-322.

A. Appendix

According to the conditional technique (Böhning 2008)

$$Var\hat{N} = \underbrace{Var_n E(\hat{N}|n)}_{(1)} + \underbrace{E_n Var(\hat{N}|n)}_{(2)}, \tag{21}$$

where E_n and Var_n are referred to the distribution of n which follows the binomial distribution with parameter N and $1-p_0$.

A.1. Variance of maximum likelihood estimator

For the first term of (21), since $E(\hat{N}|n) \approx \frac{n}{1-p_0}$,

$$Var_n E(\hat{N}_{MLEupa}|n) \approx Var_n(\frac{n}{1-p_0}) = \frac{1}{(1-p_0)^2} Var(n) = \frac{1}{(1-p_0)^2} N p_0(1-p_0). \tag{22}$$

Since $E(n) = N(1-p_0)$ and $\hat{p}_0 = \frac{n}{S}$, the first term of (21) can be estimated as

$$\begin{aligned} \widehat{Var_n} E(\hat{N}_{MLEupa}|n) &= \frac{n\hat{p}_0}{(1-\hat{p}_0)^2} \\ &= \frac{n\frac{n}{S}}{(1-\frac{n}{S})^2} \\ &= \frac{Sn^2}{(S-n)^2}. \end{aligned} \tag{23}$$

The second term of (21) can be determined as

$$\begin{aligned} \hat{E}_n Var(\hat{N}_{MLEupa}|n) &= Var\left(\frac{n}{1-\frac{n}{S}}|n\right) \\ &= n^2 Var\left(\frac{1}{1-\frac{n}{S}}\right). \end{aligned} \tag{24}$$

$Var\left(\frac{1}{1-\frac{n}{S}}\right)$ can be estimated by the delta method. Assume $y=\frac{n}{S}$, $g(y)=\frac{1}{1-y}$, and $g'(y)=\frac{1}{(1-y)^2}$, so

$$\begin{aligned} Var\left(\frac{1}{1-\frac{n}{S}}\right) &= \frac{1}{(1-\frac{n}{S})^4}Var\left(\frac{n}{S}\right) \\ &= \frac{n^2}{(1-\frac{n}{S})^4}Var\left(\frac{1}{S}\right). \end{aligned} \tag{25}$$

Using the delta method, we achieved

$$\begin{aligned} Var\left(\frac{1}{S}\right) &= \frac{1}{S^4}Var(S) \\ &= \frac{1}{S^4}Var(N\bar{X}) \\ &= \frac{1}{S^4}N^2\frac{Var(X)}{N} \\ &= \frac{1}{S^4}NVar(X). \end{aligned} \tag{26}$$

Note that X has the UPA distribution with $E(X)=\frac{1}{2\alpha}$ and $Var(X)=\frac{2\alpha+1}{4\alpha^2}$. Since $E(X)=\frac{S}{N}$, $Var(X)=\frac{S}{N}+\left(\frac{S}{N}\right)^2$, and $N=\frac{nS}{S-n}$,

$$\begin{aligned} Var(\frac{1}{S}) &= \frac{1}{S^4}N\left(\frac{S}{N}+\frac{S^2}{N^2}\right) \\ &= \frac{1}{S^3}\left(1+\frac{S}{N}\right) \\ &= \frac{1}{S^3}\left(1+\frac{S-n}{n}\right) \\ &= \frac{1}{S^2n}. \end{aligned} \tag{27}$$

Therefore, $Var\left(\frac{1}{1-\frac{n}{S}}\right)=\frac{n^2}{(1-\frac{n}{S})^4}\frac{1}{S^2n}=\frac{nS^2}{(S-n)^4}$. The second term of (21) can be determined as

$$\begin{aligned} \hat{E}_nVar(\hat{N}_{MLEupa}|n) &= n^2Var\left(\frac{1}{1-\frac{n}{S}}\right) \\ &= \frac{S^2n^3}{(S-n)^4}. \end{aligned} \tag{28}$$

The variance of maximum likelihood estimator is

$$\widehat{Var}(\hat{N}_{MLEupa})=\frac{Sn^2}{(S-n)^2}+\frac{S^2n^3}{(S-n)^4}. \tag{29}$$

A.2. Variance of generalized Turing estimator

$$\text{Since} \quad E(\hat{N}|n) \approx \frac{n}{1-p_0}, \quad Var_n E(\hat{N}_{GTupa}|n) \approx Var_n(\frac{n}{1-p_0}) = \frac{1}{(1-p_0)^2} Var(n) = \frac{1}{(1-p_0)^2} N p_0(1-p_0).$$

Since $E(n) = N(1-p_0)$ and $\hat{p}_0 = \frac{nf_1+Sf_1}{Sn+Sf_1}$ (5), the first term of (21) can be estimated as

$$\widehat{Var}_n E(\hat{N}_{GTupa}|n) = \frac{n\hat{p}_0}{(1-\hat{p}_0)^2} = \frac{Sf_1(n+f_1)(n+S)}{n(S-f_1)^2}. \tag{30}$$

For the second term of (21),

$$E_n Var(\hat{N}_{GTupa}|n) = Var\left(\frac{n}{1-\frac{nf_1+Sf_1}{Sn+Sf_1}}|n\right) = n^2 Var\left(\frac{1}{1-\frac{nf_1+Sf_1}{Sn+Sf_1}}\right).$$

If $y = \frac{nf_1+Sf_1}{Sn+Sf_1}$ and $g(y) = \frac{1}{1-y}$, then $g'(y) = \frac{1}{(1-y)^2}$. By the delta method, $Var\left(\frac{1}{1-\frac{nf_1+Sf_1}{Sn+Sf_1}}\right)$ can be approximated as

$$Var\left(\frac{1}{1-\frac{nf_1+Sf_1}{Sn+Sf_1}}\right) = \frac{1}{\left(1-\frac{nf_1+Sf_1}{Sn+Sf_1}\right)^4} Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right).$$

The second term of (21) can be estimated as

$$\hat{E}_n Var(\hat{N}_{GTupa}|n) = n^2 Var\left(\frac{1}{1-\frac{nf_1+Sf_1}{Sn+Sf_1}}\right) = \frac{n^2}{\left(1-\frac{nf_1+Sf_1}{Sn+Sf_1}\right)^4} Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right). \tag{31}$$

The conditional technique is used to estimate $Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right)$,

$$Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right) = Var_{f_1} E\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) + E_{f_1} Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right). \tag{32}$$

Here, E_{f_1} and Var_{f_1} are referred to the distribution of f_1 which follows the binomial distribution with

parameter N and p_1. Since $E\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) \approx \frac{nf_1+Sf_1}{Sn+Sf_1}$,

$$\begin{aligned} Var_{f_1}E\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) &\approx Var_{f_1}\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right) \\ &= Var_{f_1}\left(\frac{f_1(n+S)}{S(n+f_1)}\right) \\ &= \left(\frac{n+S}{S}\right)^2 Var_{f_1}\left(\frac{1}{1+\frac{n}{f_1}}\right). \end{aligned}$$

Assume $y = f_1$ and $g(y) = \frac{1}{1+\frac{n}{y}}$, then $g'(y) = \frac{\frac{n}{y^2}}{(1+\frac{n}{y})^2}$. By the delta method, $Var_{f_1}\left(\frac{1}{1+\frac{n}{f_1}}\right)$ can be approximated as

$$\begin{aligned} Var_{f_1}\left(\frac{1}{1+\frac{n}{f_1}}\right) &= \left(\frac{\frac{n}{f_1^2}}{(1+\frac{n}{f_1})^2}\right)^2 Var(f_1) \\ &= \frac{n^2}{f_1^4(1+\frac{n}{f_1})^4} Np_1(1-p_1) \\ &= \frac{n^2}{f_1^4(1+\frac{n}{f_1})^4} f_1\left(1-\frac{f_1}{N}\right) \\ &= \frac{n^2}{f_1^3(1+\frac{n}{f_1})^4}\left(1-\frac{f_1}{N}\right). \end{aligned}$$

This leads to

$$\widehat{Var}_{f_1}E\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) = \left(\frac{n+S}{S}\right)^2 \frac{n^2}{f_1^3(1+\frac{n}{f_1})^4}\left(1-\frac{f_1}{N}\right). \tag{33}$$

The expected value $E_{f_1}Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right)$ can be estimated as

$$\begin{aligned} \hat{E}_{f_1}Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) &\approx Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}|f_1\right) \\ &= Var\left(\frac{f_1(n+S)}{S(n+f_1)}|f_1\right) \\ &= \left(\frac{f_1}{n+f_1}\right)^2 Var\left(1+\frac{n}{S}|f_1\right). \end{aligned}$$

Assume $y = S$ and $g(y) = 1+\frac{n}{y}$, so $g'(y) = -\frac{n}{y^2}$. Using the delta method $Var\left(1+\frac{n}{S}|f_1\right)$ can be

estimated as

$$\begin{aligned} Var\left(1+\frac{n}{S}|f_1\right) &= \frac{n^2}{S^4}Var(S) \\ &= \frac{n^2}{S^4}Var(N\bar{X}) \\ &= \frac{n^2}{S^4}N^2\frac{Var(X)}{N} \\ &= \frac{n^2}{S^4}NVar(X). \end{aligned}$$

Since X has the UPA distribution with $E(X)=\frac{S}{N}$ and $Var(X)=\frac{S}{N}+\left(\frac{S}{N}\right)^2$, $Var\left(1+\frac{n}{S}|f_1\right)=\frac{n^2}{S^4}\left(\frac{S}{N}+\frac{S^2}{N^2}\right)=\frac{n^2}{S^3}\left(1+\frac{S}{N}\right)$. Therefore,

$$\hat{E}_{f_1}Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}f_1\right)=\left(\frac{f_1}{n+f_1}\right)^2\frac{n^2}{S^3}\left(1+\frac{S}{N}\right). \tag{34}$$

Substituting (33) and (34) into (32), this leads to

$$Var\left(\frac{nf_1+Sf_1}{Sn+Sf_1}\right)=\left(\frac{n+S}{S}\right)^2\frac{n^2}{f_1^3(1+\frac{n}{f_1})^4}(1-\frac{f_1}{N})+\left(\frac{f_1}{n+f_1}\right)^2\frac{n^2}{S^3}(1+\frac{S}{N}). \tag{35}$$

Substituting (35) into (31) leads to the second term of (21) as

$$\hat{E}_nVar(\hat{N}_{GTupa}|n)=\frac{Sf_1(n+S)^2(Sn+f_1^2)}{(n+f_1)(S-f_1)^4}+\frac{Sf_1^2(n+f_1)(n+S)}{(S-f_1)}^4.$$

Therefore, the variance of the generalized Turing estimator is

$$\widehat{Var}(\hat{N}_{GTupa})=\frac{Sf_1(n+f_1)(n+S)}{n(S-f_1)^2}+\frac{Sf_1(n+S)^2(Sn+f_1^2)}{(n+f_1)(S-f_1)^4}+\frac{Sf_1^2(n+f_1)(n+S)}{(S-f_1)^4}. \tag{36}$$