Sains Malaysiana 52(12)(2023): 3577-3587
http://doi.org/10.17576/jsm-2023-5212-18

Estimation of Population Size Based on One-Inflated, Zero-Truncated Count Distribution with Covariate Information

(Anggaran Saiz Populasi Berdasarkan Taburan Kiraan Satu-Lambung, Sifar-Pemangkasan dengan Maklumat Kovariat)

TITA JONGSOMJIT* & RATTANA LERDSUWANSRI

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand

Received: 19 August 2023/Accepted: 12 December 2023

ABSTRACT

In order to estimate the unknown size of the population that is difficult or hidden to enumerate, the capture-recapture method is widely used for this purpose. We propose the one-inflated, zero-truncated geometric (OIZTG) model to deal with three important characteristics of some capture–recapture data: zero-truncation, one-inflation, and observed heterogeneity. The OIZTG model is generated by two distinct processes, one from a zero-truncated geometric (ZTG) process, and the other one-count producing process. To explain heterogeneity at an individual level, the OIZTG model provides a simple way to link the covariate information. The new estimator was proposed based on the OIZTG distributions through the modified Horvitz-Thomson approach, and the parameters of the OIZTG distributions are estimated by using a maximum likelihood estimator (MLE). With regard to making inferences about the unknown size of the population, confidence interval estimations are proposed where variance estimate of population size estimator is achieved by using conditional expectation technique. All of these are assessed through simulation studies. The real data sets are provided for understanding the methodologies.

Keywords: Capture-recapture; geometric regression; observed heterogeneity

ABSTRAK

Dalam proses untuk menganggarkan saiz populasi yang sukar atau tersembunyi untuk dihitung, kaedah tangkap-tangkap semula digunakan secara meluas untuk tujuan ini. Kami mencadangkan model geometrik satu-lambung, geometrik sifar-pemangkasan (OIZTG) untuk menangani tiga ciri penting bagi beberapa data tangkap-tangkap semula: sifar-pemangkasan, satu-inflasi dan heterogeniti yang diperhatikan. Model OIZTG dijana oleh dua proses yang berbeza, satu daripada proses geometri terpangkas sifar (ZTG) dan satu lagi proses menghasilkan satu kiraan. Untuk menerangkan heterogeniti pada peringkat individu, model OIZTG menyediakan cara mudah untuk memautkan maklumat kovariat. Penganggar baharu telah dicadangkan berdasarkan taburan OIZTG melalui pendekatan Horvitz-Thomson yang diubah suai dan parameter taburan OIZTG dianggarkan dengan menggunakan penganggar kemungkinan maksimum (MLE). Berkenaan dengan membuat inferens tentang saiz populasi yang tidak diketahui, anggaran selang keyakinan dicadangkan dengan anggaran varians penganggar saiz populasi dicapai dengan menggunakan teknik jangkaan bersyarat. Kesemua ini dinilai melalui kajian simulasi. Set data sebenar disediakan untuk memahami metodologi.

Kata kunci: Kepelbagaian yang diperhatikan; regresi geometri; tangkap-tangkap semula

INTRODUCTION

Capture-recapture techniques are widely used to estimate the size of hidden population. This population might be a wildlife population or a population of drug addicts. Traditionally, capture-recapture methods are used in the field of wildlife biology/ecology (estimating the number of female grizzly bears (Chao & Huggins 2006). Currently, the methods are applied in a variety of area including social science (estimating the number of illicit drug users (McDonald et al. 2014), public health and epidemiology (investigating the completeness of contact tracing for COVID-19 (Lerdsuwansri et al. 2022))

as well as quantitative criminology (estimating the hidden population size of criminals (Tajuddin, Ismail & Ibrahim 2022)). A closed population (with no birth, death, or migration) of N units is assumed meaning that the population size remains constant during the study period and n distinct units are identified through some mechanisms (traps, lists, & registers). The number of units identified exactly y times is denoted by f_y. However, the number of unobserved units, f_0, remains unknown as some units are not observed at all. The total number of observed units is $n = f_1 + f_2 + f_3 + \cdots + f_m$ where m is the largest counts. Since the unknown size of population $N = n + f_0$, estimating f_0 is necessary to estimate N.

A common estimation approach is to model the number of times a unit has been identified through a counting distribution (Poisson, negative binomial, & geometric). Under homogeneity model, probability of each unit being identified exactly y times is an equal chance (Bunge & Fitzpatrick 1993; Good 1953). However, the homogeneous assumption is unrealistic as individual characteristics (gender, age, social status, & behavior) can lead to variations in capture probabilities known as heterogeneity (Chao 1987; Niwitpong et al. 2013; Zelterman 1988), whether observed or unobserved. Ignoring heterogeneity can lead to an underestimation of the true population size. To get accurate estimates N , covariate information may be used to account for the population heterogeneity.

Before we go on, we illustrate the situation at hand with a real data example. Provided in Table 1 is the frequency distribution of the number of times that a heroin user contacted a hospital and a health treatment center in Chiang Mai, Thailand, from 2013 to 2018. Additionally, individual information such as gender is also collected. A total of 843 observed heroin users consisted of 754 men and 89 women. Among these, 537 had treatment once with 482 men and 55 women. Of the 152 users with treatment twice, 134 were males and 18 were females. Clearly, f_0, the number of hidden heroin users is unobserved and there is a large number of f_1. More details of the data source are provided in Panyalert and Lanamtaeng (2020).

In certain capture-recapture studies, one-inflation in the count distribution can be observed due to difficulties in recapturing individuals and behavioral responses. Ignoring one-inflation may lead to significant overestimation of the population size. Several estimators have been developed to address this issue. 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. Godwin (2017) proposed the one-inflated, zero-truncated negative binomial (OIZTNB) model to estimate population size. What they have in common are one-inflation parameter and covariate information incorporating into truncated regression model. Although covariates can help to improve the fit of the model, OIZTNB model have the boundary problem. Böhning, Kaskasamkul and van der Heijden (2019) suggested modification of Chao's lower bound estimator (Chao 1987) to avoid overestimation caused by one-inflation, but it does not account for heterogeneity. Böhning and Friedl (2021) proposed a population size estimation for sparse count data using a zero-truncated, one-inflated model, but it does not consider observed heterogeneity.

In this study, we are interested in estimating population size using zero-truncated, one-inflated capture-recapture count data with observed heterogeneity. We propose the one-inflated, zero-truncated geometric (OIZTG) model for the Horvitz-Thompson estimator (Horvitz & Thompson 1952) $\widehat{N}$ of the population size. Additionally, confidence interval estimations for the unknown population size N are provided using the conditional expectation technique. Simulation studies and real datasets are utilized to assess the effectiveness of these methods and enhance their understanding.

TABLE 1. Frequency distribution of heroin user contacts in Chiang Mai, Thailand, from 2013 to 2018

Gender	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	n
Male	-	482	134	73	30	13	7	5	7	0	1	1	0	0	1	754
Female	-	55	18	7	4	2	1	1	1	0	0	0	0	0	0	89
Total	-	537	152	80	34	15	8	6	8	0	1	1	0	0	1	843

Materials and Methods

The one-inflated, zero-truncated geometric model

The one-inflated, zero-truncated geometric (OIZTG) model employs two components that correspond to two generating processes. The first process is from the zero-truncated geometric (ZTG) model that generates counts, some of which may be one. The probability mass function for a random variable Y that follows a ZTG distribution is

$$P^{ZTG}(Y = y) = (1-\theta)^{y-1}\theta \quad ; y = 1,2,\ldots,$$

where θ is the geometric parameter, $0 < \theta < 1$. The second process is from one-count producing that generates structural ones. The combination of these two processes is called the OIZTG distribution which can be expressed as:

$$P^{OIZTG}(Y = y) = \begin{cases} w + (1-w)\theta & ; y = 1 \\ (1-w)(1-\theta)^{y-1}\theta & ; y = 2,3,\ldots \end{cases},$$

where is the one-inflation parameter, or the additional probability of a 1 count occurring, $0 \leq w \leq 1$. The mean of the OIZTG distribution is $w + (1-w)\left(\frac{1}{\theta}\right)$.

Incorporating covariates in the OIZTG model

In standard models, covariate affecting the mean of the distribution can be incorporated into the model by using reparameterization. For the case of OIZTG distribution, there is no simple way to directly express the mean as a function of covariates. In the cases of OIPP model (Godwin & Böhning 2017) and OIZTNB model (Godwin 2017), it has been usual to include covariates into the models in the same way as they are included in their equivalent untruncated models. This approach assumes that the effects of covariates are on the conditional distribution and the inflation parameter are the same for both truncated and untruncated models. The OIZTG model follows this groundwork as well.

In the original geometric model, the mean is $\mu = 1\theta$. Alternatively, we can use this relationship to express θ in terms of μ:

$$\theta = \frac{1}{\mu}. \qquad (1)$$

To explain heterogeneity at an individual level, a reparameterization is necessary to link the mean of the distribution to covariates. Using the natural logarithm link function, we have $ln(\mu_i) = \beta^T x_i$ and we get $\mu_i = e^{\beta^T x_i}$. Substituting the re-parameterized value of μ_i into the equation (1), we obtain

$$\theta_i = \frac{1}{\mu_i} = \frac{1}{e^{\beta^T x_i}} = e^{-\beta^T x_i},$$

where β is a vector of coefficients; $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$; p; is the number of covariates; and x_i is a vector of covariate values for subject i or $x_i = (1, x_i1, \ldots, x_{ip})^T$. Therefore, the OIZTG distribution can be rewritten as follows:

$$P^{OIZTG}(Y_i = y_i) = \begin{cases} w + (1-w)(\theta_i) & ; y_i = 1 \\ (1-w)(1-\theta_i)^{y_i-1}(\theta_i) & ; y_i = 2,3,\ldots \end{cases},$$

where $\theta_i = exp\,(-\beta^T x_i)$. Additionally, we can use a logit link function for the one-inflation parameter assuming that equally across subject for all subject i in the population as $w = \frac{1}{1+e^{-\varphi}}$ where φ may be estimated without limited range, ensuring that $0 \leq w \leq 1$.

Maximum likelihood estimation of the OIZTG model

Let Y_i be the count variable that represents the number of times a i^{th} unit was identified during the study period. The log-likelihood function for independent sampling from the OIZTG distribution can be written as:

$$l(w,\beta) = \sum_{i=1}^{n}\{I_{y=1}log[w + (1-w)(\theta_i)] + (1 - I_{y=1}) [log(1-w) + (y_i - 1)log(1-\theta_i) + log\theta_i]\},$$

where $I_{y=1} = 1$ if $y_i = 1$ and $I_{y=1} = 0$ if $y_i > 1$. Noted that the estimated value of β is required for the estimation of θ_i. The first derivatives of the log-likelihood are given by

$$\frac{\partial l(w,\beta)}{\partial w} = \sum_{i=1}^{n}\left\{I_{y=1}\left[\frac{1-\theta_i}{w+(1-w)(\theta_i)}\right] + (1 - I_{y=1})\left[\frac{-1}{1-w}\right]\right\},$$

$$\frac{\partial l(w,\beta)}{\partial \beta_r} = \sum_{i=1}^{n}\left\{ I_{y=1}\left[\frac{(1-w)(-X_{ir}\theta_i)}{w+(1-w)(\theta_i)}\right] + \left(1 - I_{y=1}\right) \left[\frac{(y_i-1)(X_{ir}\theta_i)}{1-\theta_i} - X_{ir}\right]\right\}; r = 0,1,2,\ldots,p.$$

Due to the fact that we do not have any closed form expressions for $\widehat{w}$ and $\widehat{\beta}$ and the complexity of

the derivatives involved, maximizing the log-likelihood function directly can be a difficult undertaking. However, the *maxLik* function in the R program provides a useful tool for estimating the parameters of a given probability distribution.

ESTIMATING THE SIZE OF AN UNKNOWN POPULATION USING THE OIZTG MODEL

The unknown size N of a closed population consists of the number of observed and unobserved units, $N = n + f_0$. Capture-recapture methods use information on f_y ,$y = 1, 2, ..., m$ to estimate f_0 leading to the estimate of N as $\widehat{N} = n + \widehat{f}_0$. The Horvitz-Thompson estimator for estimating f_0 is $\widehat{f}_0 = \sum_{i=1}^{n} \frac{P(Y_i=0)}{1-P(Y_i=0)}$ where $P(Y_i = 0)$ represents the probability of i^{th} unit being unobserved. If one-inflation exists, the conventional Horvitz-Thompson estimator needs to be modified as suggested by Böhning and Friedl (2021). We base our inference on the zero-one-truncated probability function. Hence, the modified Horvitz-Thompson estimator for estimating f_0 is given by

$$\widehat{f}_0 = \sum_{i=1}^{n} I_{y>1} \frac{P(Y_i=0)}{1-P(Y_i=0)-P(Y_i=1)},$$

where $I_{y>1} = 1$ if i^{th} unit is identified more than once and $I_{y>1} = 0$, otherwise. Accordingly, the estimated population size using the OIZTG model can be obtained as follows

$$\widehat{N}^{OIZTG} = n + \sum_{i=1}^{n} I_{y>1} \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i-\widehat{\theta}_i(1-\widehat{\theta}_i)}.$$

where $\widehat{\theta}_i$ is the MLEs of the OIZTG model.

VARIANCE OF THE POPULATION SIZE ESTIMATOR UNDER THE OIZTG MODEL

If the sample is large enough for the normal approximation to hold, the Wald approach is a valid method to construct a confidence interval for the population size N. The (1-α)100% confidence interval for the size N of a population is given as $\widehat{N}^{OIZTG} \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\widehat{N}^{OIZTG})$ where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)^{th}$ is the percentile of the standard normal distribution and $\widehat{SE}(\widehat{N}^{OIZTG})$ is the estimated standard error of $\widehat{N}^{OIZTG}$.

According to aforementioned outcomes, $\widehat{N}^{OIZTG}$ is derived. Here, we will develop variance of $\widehat{N}^{OIZTG}$ which can be written as

$$Var\left(\widehat{N}^{OIZTG}\right) = Var(n) + Var\left(\sum_{i=1}^{n} I_{y>1} \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i-\widehat{\theta}_i(1-\widehat{\theta}_i)}\right). \quad (2)$$

For the first term of Equation (2), the distribution of n is binomial with sample size parameter N and success parameter 1 - $P(Y=0)$. Hence, $Var(n) = N(1-P(Y = 0))P(Y=0)$ which can be estimated by $n\frac{\hat{f}_0}{\widehat{N}}$. Using the OIZTG model, we obtain an estimate of the variance of the sample size:

$$\widehat{Var}(n) = \left(\frac{n}{\widehat{N}^{OIZTG}}\right) \sum_{i=1}^{n} I_{y>1} \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i-\widehat{\theta}_i(1-\widehat{\theta}_i)}, \quad (3)$$

where n is the number of observed units; $\widehat{\theta}_i$ is obtained from fitting the OIZTG model, and indicator $I_{y>1} = 1$ if $y_i > 1$ and $I_{y>1} = 0$, elsewhere.

For the second term of equation (2), the variance of the number of unobserved units which can be estimated using the conditional expectation approach taken by Böhning (2008). Let us now consider $X = \hat{f}_0 = \sum_{i=1}^{n} I_{y>1} \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i-\widehat{\theta}_i(1-\widehat{\theta}_i)}$ and $Y = I_{y>1}$. That is

$$Var\left(\hat{f}_0\right) = E_{I_{y>1}}\left[Var\left(\hat{f}_0 \middle| I_{y>1}\right)\right] + Var_{I_{y>1}}\left[E\left(\hat{f}_0 \middle| I_{y>1}\right)\right]. \quad (4)$$

The first term of equation (4) represents the variation in estimating $\widehat{f}_0$ based on $I_{y>1}$ data. This term can be estimated by $Var\left(\hat{f}_0 \middle| I_{y>1}\right)$ using the δ-method. We consider $X = \widehat{\theta}_i$ and $f(X) = f\left(\widehat{\theta}_i\right) = \sum_{i=1}^{n} I_{y>1} \frac{\widehat{\theta}_i}{1-\widehat{\theta}_i-\widehat{\theta}_i(1-\widehat{\theta}_i)} = \hat{f}_0$. For convenience of notation define the vector with derivatives

$$\tau(w,\beta) = \begin{pmatrix} \frac{d}{dw} f_0 \\ \frac{d}{d\beta_r} f_0 \end{pmatrix} = \begin{pmatrix} \frac{d}{dw} \sum_{i=1}^{n} I_{y>1} \frac{\theta_i}{1-\theta_i-\theta_i(1-\theta_i)} \\ \frac{d}{d\beta_r} \sum_{i=1}^{n} I_{y>1} \frac{\theta_i}{1-\theta_i-\theta_i(1-\theta_i)} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ \sum_{i=1}^{n} I_{y>1} \frac{\theta_i X_{ir}(\theta_i^2-1)}{(1-\theta_i-\theta_i(1-\theta_i))^2} \end{pmatrix},$$

where $r = 0, 1, ..., p$ and the observed information matrix represented in the following matrix:

$$O(w,\beta) = \begin{pmatrix} O_{ww} & O_{w\beta} \\ O_{\beta w} & O_{\beta\beta} \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 l(w,\beta)}{\partial w^2} & -\frac{\partial^2 l(w,\beta)}{\partial w \partial \beta_r} \\ -\frac{\partial^2 l(w,\beta)}{\partial \beta_s \partial w} & -\frac{\partial^2 l(w,\beta)}{\partial \beta_s \partial \beta_r} \end{pmatrix},$$

where $\frac{\partial^2 l(w,\beta)}{\partial w^2} = \sum_{i=1}^{n} \left\{ I_{y=1} \left[\frac{(-1)(1-\theta_i)^2}{(w+(1-w)(\theta_i))^2} \right] + (1 - I_{y=1}) \left[\frac{-1}{(1-w)^2} \right] \right\},$

$$\frac{\partial^2 l(w,\beta)}{\partial w \partial \beta_r} = \frac{\partial^2 l(w,\beta)}{\partial \beta_s \partial w} = -\sum_{i=1}^{n}\left\{I_{y=1}\left[\frac{X_{ir}\theta_i}{(w+(1-w)(\theta_i))^2}\right]\right\},$$

$$r = 0,1,\ldots,p,$$

$$\frac{\partial^2 l(w,\beta)}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n}\left\{I_{y=1}\left[\frac{X_{ir}X_{is}\theta_i w(1-w)}{(w+(1-w)(\theta_i))^2}\right] + (1-I_{y=1})\left[\frac{-X_{ir}X_{is}\theta_i(y_i-1)}{(1-\theta_i)^2}\right]\right\}, r,s = 0,1,\ldots,p.$$

Therefore, the first term of Equation (4) can be estimated as follows:

$$E_{I_{y>1}}[Var(\hat{f}_0|I_{y>1})] \approx \tau^T(\hat{w},\hat{\beta})O^{-1}(\hat{w},\hat{\beta})\tau(\hat{w},\hat{\beta}) \quad (5)$$

where $\hat{w}$ and $\hat{\beta}$ are obtained from fitting the OIZTG model.

The second term of Equation (4) shows variation in the obtained sample. If the number of observations is not too small, we can safely assume that $E(\hat{f}_0|I_{y>1}) \approx \sum_{i=1}^{n} I_{y>1}\frac{\hat{\theta}_i}{1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)}$. Hence,

$$Var_{I_{y>1}}[E(\hat{f}_0|I_{y>1})] \approx \sum_{i=1}^{n}\left(\frac{\hat{\theta}_i}{1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)}\right)^2 Var_{I_{y>1}}(I_{y>1}). \quad (6)$$

Since the identification that has been observed more than once occurs independently for each population unit with probability 1-$P(Y_i = 0)$ - 1-$P(Y_i = 1)$, the variance number of each observed case is $Var_{I(y>1)}(I_{y>1}) = (1\text{-}P(Y_i = 0) - P(Y_i = 1))(P(Y_i = 0) + P(Y_i = 1))$. Using the OIZTG model the variance of $I_{y>1}$ can readily be estimated as

$$\widehat{Var}_{I_{y>1}}(I_{y>1}) = \left(1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)\right)\left(\hat{\theta}_i+\hat{\theta}_i(1-\hat{\theta}_i)\right). \quad (7)$$

Substituting Equation (7) in Equation (6),

$$Var_{I_{y>1}}[E(\hat{f}_0|I_{y>1})] \approx \sum_{i=1}^{n}\left(\frac{\hat{\theta}_i}{1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)}\right)^2 \left(1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)\right)\left(\hat{\theta}_i+\hat{\theta}_i(1-\hat{\theta}_i)\right),$$

which is estimated from the observed data by

$$\sum_{i=1}^{n} I_{y>1}\left(\frac{\hat{\theta}_i}{1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)}\right)^2\left(\hat{\theta}_i+\hat{\theta}_i(1-\hat{\theta}_i)\right) = \sum_{i=1}^{n} I_{y>1}\frac{(2-\hat{\theta}_i)\hat{\theta}_i^3}{(1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i))^2}.$$

Consequently, the second term of Equation (4) can be estimated as follows:

$$Var_{I_{y>1}}[E(\hat{f}_0|I_{y>1})] \approx \sum_{i=1}^{n} I_{y>1}\frac{(2-\hat{\theta}_i)\hat{\theta}_i^3}{\left(1-\hat{\theta}_i-\hat{\theta}_i(1-\hat{\theta}_i)\right)^2}, \quad (8)$$

where $\hat{\theta}_i$ is obtained from fitting the OIZTG model, and indicator $I_{y>1} = 1$ if $y_i > 1$ and $I_{y>1} = 0$, elsewhere. Therefore, the estimated variance of $\hat{N}^{OIZTG}$ can be written as

$$\widehat{Var}\left(\hat{N}^{OIZTG}\right) = \widehat{Var}(n) + \widehat{Var}(\hat{f}_0),$$

where we obtained this result by adding the expressions (3), (5), and (8).

SIMULATION STUDY

We conduct a simulation study to investigate the performance of the proposed estimator $\hat{N}^{OIZTG}$ and confidence interval estimations compared with the existing estimator developed using the geometric distribution as follows: the Chao's lower bound estimator, $\hat{N}^{C} = n + \frac{f_1^2}{f_2}$ with $\widehat{Var}\left(\hat{N}^{C}\right) = \frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2}$, the modified Chao's lower bound estimator, $\hat{N}^{MC} = n + \frac{f_2^3}{f_3^2}$ with $\widehat{Var}\left(\hat{N}^{MC}\right) = \frac{\left(\frac{f_2^3}{f_3^2}\right)^2\left(1+\frac{(2f_2+3f_3)^2}{f_2 f_3}\right)}{f_2+f_3}$, the Horvitz-Thomson (HT) estimator under the ZTG model, $\hat{N}^{ZTG} = \sum_{i=1}^{n}\frac{1}{1-\hat{\theta}_i}$ where $\hat{\theta}_i$ is the MLEs of the OIZTG model when $w = 0$. Moreover, the proposed estimator is also compared with the other estimator as follows: the HT estimator under the OIPP model, $\hat{N}^{OIPP} = \sum_{i=1}^{n}\frac{1}{1-e^{-\hat{\lambda}_i}}$ where $\hat{\lambda}_i$ is the MLEs of the OIPP model and the HT estimator under the OIZTNB model, $\hat{N}^{OIZTNB} = \sum_{i=1}^{n}\frac{1}{1-\left(1+\frac{\hat{\lambda}_i}{\hat{\alpha}}\right)^{-\hat{\alpha}}}$ where $\hat{\alpha}$ and $\hat{\lambda}_i$ are the MLEs of the OIZTNB model.

The simulation experiment is conducted under specific conditions: (1) The population size (N) varied between 500, 1,000, and 2,000. (2) A binary covariate variable X was generated from a Bernoulli distribution with $p = 0.5$. (3) The response variable Y was generated from the OIZTG distribution with parameters $w \in \{0.1,0.3,0.5\}$. and $\theta_i \in \{0.1, 0.3, 0.5\}$.

To demonstrate the data generating process, assuming a population size of $N = 500$, with $w = 0.1$ extra-ones, $N_{Dist} = 450$ follows a geometric distribution with parameter ($\theta_{x=0}$), $\theta_{x=1}$) = (0.1, 0.3) for binary covariate variable (X) which $\theta = 0.1$ if $X = 0$ and $\theta = 0.3$ if X = 1. Additionally, there are $N_{OI} = 50$ extra ones in the population that the total population size is $N = N_{Dist} + N_{OI}$. A simple algorithm to generate data are as follows:

Step 1 Generate a binary covariate variable X of size N = 500 from a Bernoulli distribution with parameter $p = 0.5$.

Step 2 Construct the parameter θ_i for $i = 1, 2, ..., 450$ by setting $\theta_i = 0.1$ if $X_i = 0$ and $\theta_i = 0.3$ if $X_i = 1$.

Step 3 Generate the one-inflated geometric data by randomly generating Y_i from a Geometric distribution with parameter θ_i for i = 1, 2, ..., 450, and setting Y_i = 1 for i = 451,...,500.

Step 4 Remove the zero counts from the sample, resulting in a new sample of size n. This step yields the one-inflated zero-truncated geometric data.

This algorithm is repeated M = 10,000 times. Average percentage of relative bias (%RBias) and average percentage of relative mean square error (%RMSE) are calculated. Additionally, 95% CI, coverage probability (CP) and average lengths (AL) are computed.

Results and Discussion

Table 2 provides %RBias and %RMSE for various estimators. With the increasing N, the %RBias as well as %RMSE of all estimators $\widehat{N}$ consistently decrease. When the parameter w increases, the %RBias of $\widehat{N}^{C}$, $\widehat{N}^{MC}$, $\widehat{N}^{ZTG}$ and $\widehat{N}^{OIZTG}$ tend to increase, while the %RBias of $\widehat{N}^{OIPP}$ and $\widehat{N}^{OIZTNB}$ decrease. Similarly, when w increases, the %RMSE of , and , tend to increase, but the %RMSE of $\widehat{N}^{MC}$, $\widehat{N}^{OIPP}$, and $\widehat{N}^{OIZTNB}$ decrease. Moreover, $\widehat{N}^{C}$ and $\widehat{N}^{ZTG}$ have a consistent tendency to overestimate the true value, with $\widehat{N}^{OIZTNB}$ exhibiting a more severe bias towards overestimation. Conversely, the estimators $\widehat{N}^{MC}$ and $\widehat{N}^{OIPP}$ consistently underestimate the true value. However, among these estimators, $\widehat{N}^{MC}$ and $\widehat{N}^{OIZTG}$ demonstrate the least bias.

CP and AL of the 95% confidence intervals using the OIZTG model as the data generation process were presented in Table 3. As can be seen, the CIs of $\widehat{N}^{C}$ and $\widehat{N}^{ZTG}$ do not cover the true population size in almost all cases, as they tend to overestimate. On the other hand, the CIs of $\widehat{N}^{OIPP}$ and showed low CP. However, the CI of exhibited a high CP and narrow range. Additionally, it was observed that increasing the values of N and w led to an increase in the performance of the CI of $\widehat{N}^{OIZTNB}$.

TABLE 2. %RBias (and %RMSE below) of the proposed estimator and comparators Chao, modified Chao, MLE under zero-truncated geometric, MLE under OIPP model and MLE under OIZTNB model when the OIZTG is the data-generating process

N	w	$\theta_{(x=0)}$	$\theta_{(x=1)}$	$\widehat{N}^{C}$	$\widehat{N}^{MC}$	$\widehat{N}^{ZTG}$	$\widehat{N}^{OIPP}$	$\widehat{N}^{OIZTNB}$	$\widehat{N}^{OIZTG}$
500	0.1	0.1	0.3	37.19	1.71	5.96	-17.21	7.87E+06	**0.54**
				15.12	1.33	0.44	2.99	4.37E+13	**0.09**
			0.5	43.77	**-1.44**	13	-20.7	1.11E+07	3.55
				21.35	2.78	2.03	4.36	2.13E+14	**0.44**
		0.3	0.5	42.66	**2.22**	16.59	-30.84	2.17E+01	2.45
				20.46	4.8	3.15	9.58	9.31E+01	**0.58**
	0.3	0.1	0.3	197.07	1.97	21.37	-13.08	5.05E+05	**0.87**
				402.71	1.19	4.71	1.74	6.08E+12	**0.08**
			0.5	229.35	**-0.23**	48.75	-12.45	3.31E+06	6.82
				548.34	2.44	24.96	1.68	5.04E+13	**0.88**
		0.3	0.5	200.36	**2.46**	61.66	-21.47	7.46E+01	3.9
				416.82	4.02	39.23	4.7	3.52E+03	**0.67**
	0.5	0.1	0.3	589.37	2.53	46.45	-9.02	1.43E+02	**0.82**
				3615.36	1.2	22.03	0.83	2.04E+05	**0.06**
			0.5	676.66	**1.33**	113.88	-3.79	3.00E+04	7.76
				4794.94	2.5	135.96	**0.46**	1.49E+10	1.09
		0.3	0.5	552.46	**3.13**	142.19	-11.94	1.17E+02	4.16
				3164.05	3.57	208.22	1.58	1.11E+03	**0.64**

1000	0.1	0.1	0.3	36.52	**0.46**	5.93	-17.19	1.13E+07	0.49
				13.94	0.52	0.39	2.97	6.02E+13	**0.04**
			0.5	42.82	-3.7	12.54	-20.9	8.72E+06	**3.02**
				19.36	1.13	1.74	4.4	4.76E+13	**0.24**
		0.3	0.5	41.81	**-0.57**	16.12	-31.01	1.51E+01	1.79
				18.58	1.72	2.78	9.65	1.13E+01	**0.27**
	0.3	0.1	0.3	193.77	0.72	21.08	-13.15	2.80E+05	**0.69**
				382.1	0.45	4.52	1.74	4.04E+11	**0.04**
			0.5	225.34	**-2.32**	47.68	-12.79	2.13E+06	6.2
				517.9	0.9	23.29	1.7	1.69E+13	**0.57**
		0.3	0.5	197.82	**-0.06**	60.64	-21.76	4.04E+01	3.1
				398.33	1.43	37.35	4.78	7.38E+01	**0.33**
	0.5	0.1	0.3	578.04	0.89	45.82	-9.08	6.57E+01	**0.66**
				3402.98	0.34	21.2	0.83	3.56E+04	**0.03**
			0.5	663.21	**-1.13**	111.18	-4.47	1.16E+04	6.92
				4493.28	0.72	126.39	**0.34**	1.19E+10	0.69
		0.3	0.5	543.12	**0.71**	139.72	-12.43	7.27E+01	3.3
				2999.45	1.17	197.94	1.61	2.64E+02	**0.31**
2000	0.1	0.1	0.3	36.12	**-0.12**	5.84	-17.21	1.44E+07	0.4
				13.34	0.23	0.36	2.97	1.89E+14	**0.02**
			0.5	42.45	-4.57	12.5	-20.9	1.15E+07	**2.96**
				18.54	0.65	1.64	4.39	1.79E+14	**0.16**
		0.3	0.5	41.36	-1.8	15.93	-31.08	1.16E+01	**1.55**
				17.64	0.84	2.63	9.67	4.50E+00	**0.14**
	0.3	0.1	0.3	192.9	**-0.04**	21.05	-13.16	2.00E+05	0.65
				375.26	0.18	4.47	1.74	2.64E+11	**0.02**
			0.5	224.13	**-3.42**	47.44	-12.9	1.67E+06	5.97
				507.18	0.45	22.78	1.69	7.87E+12	**0.45**
		0.3	0.5	196.8	**-1.24**	60.14	-21.91	3.12E+01	2.71
				390.75	0.65	36.45	4.82	1.76E+01	**0.19**
	0.5	0.1	0.3	573.94	**0.21**	45.65	-9.1	2.07E+01	0.61
				3324.63	0.14	20.95	0.83	9.46E+01	**0.02**
			0.5	654.94	**-2.07**	109.7	-4.83	2.56E+02	6.49
				4334.23	0.32	121.63	**0.3**	1.86E+06	0.52
		0.3	0.5	540.44	**-0.51**	138.14	-12.7	5.30E+01	2.82
				2945.09	0.49	192.1	1.64	5.05E+01	**0.18**

The bold number shows the smallest %RBias and %RMSE.

TABLE 3. The coverage probability (CP) and average length (AL) of 95% CI of using the OIZTG model as the data generation process

N	w	$\theta_{(x=0)}$	$\theta_{(x=1)}$	CP					AL				
				C	MC	ZTG	OIPP	OIZTG	C	MC	ZTG	OIPP	OIZTG
500	0.1	0.1	0.3	0.0148	0.8940	0.5253	0	0.9457	163.93	218.28	58.84	NaN	54.44
			0.5	0.0498	0.8223	0.4047	0	0.9874	217.82	327.82	105.83	NaN	131.24
		0.3	0.5	0.0882	0.8793	0.2262	0	0.9675	228.68	405.15	114.76	NaN	145.18
	0.3	0.1	0.3	0	0.8878	0	0	0.9471	NaN	203.01	NaN	NaN	50.68
			0.5	0	0.8239	0	0.0744	0.9832	NaN	306.75	NaN	81.35	143.63
		0.3	0.5	0	0.8765	0	0.0012	0.9738	NaN	380.12	NaN	106.09	142.88
	0.5	0.1	0.3	0	0.8774	0	0	0.9523	NaN	190.81	NaN	NaN	44.74
			0.5	0	0.8129	0	0.6174	0.9914	NaN	299.46	NaN	88.55	147.99
		0.3	0.5	0	0.8739	0	0.1302	0.9793	NaN	359.17	NaN	98.03	136.36
1000	0.1	0.1	0.3	0	0.9084	0.2018	0	0.9403	NaN	281.85	81.91	NaN	76.36
			0.5	0.0001	0.8075	0.0833	0	0.9746	243.01	409.83	143.33	NaN	180.99
		0.3	0.5	0.0011	0.8970	0.0221	0	0.9632	290.04	524.37	155.87	NaN	199.00
	0.3	0.1	0.3	0	0.9013	0	0	0.9396	NaN	256.98	NaN	NaN	70.74
			0.5	0	0.8186	0	0.0068	0.8977	NaN	373.53	NaN	118.00	191.99
		0.3	0.5	0	0.8949	0	0	0.9618	NaN	475.09	NaN	NaN	194.41
	0.5	0.1	0.3	0	0.8990	0	0	0.9420	NaN	227.00	NaN	NaN	61.96
			0.5	0	0.8219	0	0.4918	0.8913	NaN	334.96	NaN	120.28	190.09
		0.3	0.5	0	0.8881	0	0.0218	0.9649	NaN	421.72	NaN	136.89	178.72
2000	0.1	0.1	0.3	0	0.9190	0.0160	0	0.9402	NaN	379.19	114.12	NaN	107.60
			0.5	0	0.7733	0.0025	0	0.9308	NaN	549.79	196.69	NaN	251.69
		0.3	0.5	0	0.8920	0.0001	0	0.9572	NaN	706.38	226.94	NaN	277.79
	0.3	0.1	0.3	0	0.9144	0	0	0.9254	NaN	339.62	NaN	NaN	99.55
			0.5	0	0.7907	0	0	0.6731	NaN	492.21	NaN	NaN	259.70
		0.3	0.5	0	0.8981	0	0	0.9280	NaN	630.03	NaN	NaN	268.76
	0.5	0.1	0.3	0	0.9104	0	0	0.9212	NaN	294.35	NaN	NaN	86.65
			0.5	0	0.8000	0	0.3068	0.5896	NaN	429.12	NaN	167.03	247.58
		0.3	0.5	0	0.8981	0	0.0003	0.9237	NaN	548.37	NaN	222.17	243.10

If the confidence interval does not contain the true parameter N, then the length of the confidence interval will be undefined or NaN

AN APPLICATION TO A DRUG USE POPULATION IN CHIANG MAI, THAILAND

We demonstrate the proposed estimators by an application to estimate the size of a drug use population in Chiang Mai, Thailand. The data was collected by a hospital and a health treatment center, which recorded information on heroin users, including their gender and how many times they were treated (Table 1).

We examine the distributions providing the best fit to the observed counts. The associated distributions are presented in Figure 1 and show clear evidence that the OIZTG distribution provides a better fit compared to the other distributions.

In Table 4, various statistics are provided including the estimated number of unobserved drug users, estimates for the number of heroin users, standard error, and 95% CI for *N*. Among the comparators, if $\widehat{N}^{C}$ and $\widehat{N}^{MC}$ are close, this indicates lack of evidence for one-inflation. Evidently, we are dealing with the situation of one-inflation as the difference between $\widehat{N}^{C}$ and $\widehat{N}^{MC}$ is quite substantial. Therefore, $\widehat{N}^{OIPP}$, $\widehat{N}^{OIZTNB}$ and $\widehat{N}^{OIZTG}$ are the candidates for use. As can be seen, $\widehat{N}^{OIZTG}$ is not only providing the better fit to the observed count distribution, but also compromising on the estimates between $\widehat{N}^{OIPP}$ and $\widehat{N}^{OIZTNB}$. We conclude that the total number of heroin users is 1385 with a 95% CI of (1246, 1525).

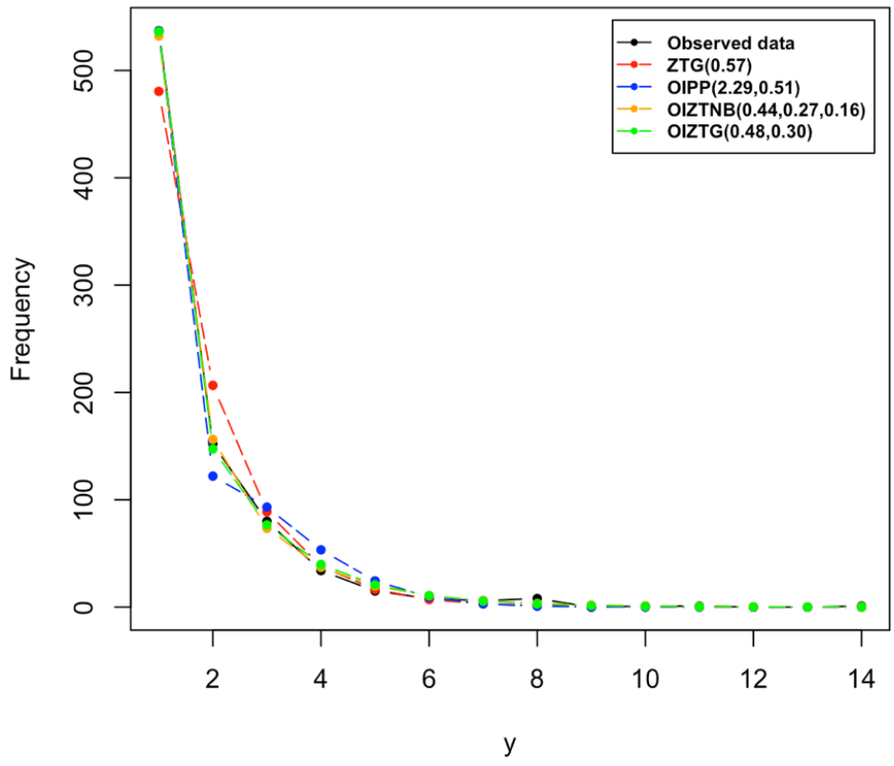

FIGURE 1. Frequency distribution of heroin users with gender as a covariate among the observed counts, ZTG, OIZTG, OIPP and OIZTNB distribution

TABLE 4. Estimated total number of heroin users with gender as a covariate

Estimator	$\hat{f}_0$	$\hat{N}$	$\widehat{SE}(\hat{N})$	95% CI of
Chao	1897	2740	216.34	(2316, 3164)
Modified Chao	549	1392	181.33	(1036, 1747)
ZTG ($\hat{\theta}$ *=0.57*)	1114	1957	77.50	(1805, 2109)
OIPP ($\hat{\lambda}$ *=2.29,*$\hat{w}$ *=0.51*)	95	938	15.54	(908, 969)
OIZTNB ($\hat{\lambda}$ *=0.44,*$\hat{\alpha}$ *=0.27,*$\hat{w}$ *=0.16*)	2831	3674	-	-
OIZTG ($\hat{\theta}$ *=0.48,* $\hat{w}$ *=0.30)*	542	1385	71.25	(1246, 1525)

Noted that the estimated SE and CI for the are not included in this research

CONCLUSIONS

Capture-recapture is a useful method for estimating the size of an elusive target population. During the capture-recapture sampling process, frequency count data is collected over the observational period. In addition to the frequency counts, data on various characteristics such as gender, or other relevant factors may also be collected. However, there are cases where certain individuals remain unobserved because they have never been identified, leading to missing zero-count data. Estimating the number of unobserved cases is typically necessary in such situations. In some capture-recapture studies, the observed data shows the presence of one-inflation in the count distribution, indicating that a portion of the population is primarily captured only once. Ignoring this one-inflation phenomenon can result in a significant overestimation of the population size. Additionally, it is important to consider variations in capture probability due to heterogeneity. By incorporating a heterogeneous Poisson model, which accounts for this heterogeneity, a more realistic estimation of the true population size can be achieved. The commonly used negative binomial distribution has been applied as a model for capture-recapture data. However, many studies have demonstrated the failure of accurately estimating the dispersion parameter in the negative binomial distribution, leading to spurious population size estimates *N*. As an alternative approach, this study proposes the use of the geometric distribution to overcome these limitations.

We proposed the one-inflated, zero-truncated geometric (OIZTG) model, which is designed to handle three crucial aspects often observed in capture-recapture data: zero-truncation, one-inflation, and observed heterogeneity. The OIZTG model also includes covariates that link the mean of the model to the covariates through a log link function. A new estimator $\hat{N}^{OIZTG}$ is proposed based on the OIZTG distribution through the modified Horvitz-Thomson approach. The simulation results show that $\hat{N}^{OIZTG}$ is an asymptotic estimator under the OIZTG distributions. In addition, we employed the OIZTG model to construct confidence intervals (CI) for the population size *N* using the Wald approach. The estimation of the variance of the proposed estimator was based on the conditional expectation technique. Simulation results confirm that the proposed CI is a suitable choice for estimating the CI of the population size *N* based on the OIZTG distribution.

ACKNOWLEDGEMENTS

The authors are grateful to the Science Achievement Scholarship of Thailand (SAST) and the king Prajadhipok and queen Rambhaibarni memorial foundation for providing funding for Tita Jongsomjit. The authors would also thank referees for thoughtful and helpful comments.

REFERENCES

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

Böhning, D. & Friedl, H. 2021. Population size estimation based upon zero-truncated, one-inflated and sparse count data. *Stat. Methods Appl.* 30: 1197-1217.

Böhning, D., Kaskasamkul, P. & 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.

Bunge, J. & Fitzpatrick, M. 1993. Estimating the number of species: A review. *Journal of the American Statistical Association* 88: 364-373.

Chao, A. & Huggins, R.M. 2006. Four. Modern closed-population capture-recapture models. In *Handbook of Capture-Recapture Analysis*, edited by Amstrup, S.C., McDonald, T.L. & Manly, B.F.J. Princeton: Princeton University Press. pp. 58-87.

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43(4): 783-791.

Godwin, R.T. 2017. One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal* 59(1): 79-93.

Godwin, R.T. & 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(2): 425-448.

Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40(3-4): 237-264.

Horvitz, D.G. & Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260): 663-685.

Lerdsuwansri, R., Sangnawakij, P., Böhning, D., Sansilapin, C., Chaifoo, W., Polonsky, J. & Del Rio Vilas, V. 2022. Sensitivity of contact-tracing for COVID-19 in Thailand: A capture-recapture application. *BMC Infectious Diseases* 22: 101.

McDonald, S., Hutchinson, S., Schnier, C., McLeod, A. & Goldberg, D. 2014. Estimating the number of injecting drug users in Scotland's HCV-diagnosed population using capture–recapture methods. *Epidemiology & Infection* 142(1): 200-207.

Niwitpong, S., Böhning, D., Van Der Heijden, P.G.M. & Holling, H. 2013. Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika* 76(4): 495-519.

Panyalert, K. & Lanamtaeng, K. 2020. Factors influencing drug rehabilitation attendance at Thunyarak Chiangmai Hospital for substance addiction. *Proceedings of the 18th Scientific and Technological Conference,* 440-451. Maejo University, Thailand, 28 February 2020.

Tajuddin, R.R.M., Ismail, N. & Ibrahim, K. 2022. Estimating population size of criminals: A new Horvitz–Thompson estimator under one-inflated positive Poisson-Lindley Model. *Crime & Delinquency* 68(6-7): 1004-1034.

Zelterman, D. 1988. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference* 18(2): 225-237.

*Corresponding author; email: tita.jong@dome.tu.ac.th