ENHANCED POPULATION SIZE ESTIMATION USING ZELTERMAN-TYPE ZERO-TRUNCATED DISCRETE LINDLEY DISTRIBUTION UNDER ONE-INFLATED POISSON COUNT DATA

Agog Nathan Samuel[1] , Danjuma Jibasen[2], Sauta S. Abdulkadir [3]

nathan.agog@kasu.edu.ng [1], djibasen@mau.edu.ng [2], ssabdulkadir@mautech.edu.ng[3]

[1]Department of Mathematical Sciences, Faculty of Physical Sciences, Kaduna State University, Nigeria

[1,2,3]Department of Statistics, Faculty of Physical Sciences, Modibbo Adama University, Yola, Adamawa, Nigeria

Abstract

Capture-recapture analysis is widely used for population size estimation in various fields, including ecology, biology, social sciences and medicine. The Zelterman Poisson estimator obtained from Poisson distribution is commonly used for estimating population size in capture-recapture but it tends to underestimate counts when dealing with overdispersed data. To address this limitation, this paper proposes the Zelterman-type estimator (Zelterman-DLD) developed under Zero-Truncated Discrete Lindley (ZTDL) distribution for improved population size estimation. The paper evaluates the performance of two estimators; Zelterman-DLD and Zelterman Poisson estimator (Zelterman-POIS) using count data derived from a one-inflated Poisson distribution. These estimators were assessed across different population sizes under varying levels of one-inflation (10% and 20%). Variance estimation was performed using the conditioning technique, while relative bias (RBIAS) and relative root mean square error (RRMSE) were used to measure the performance of the estimators. Simulation results and application to real-life data shows that the Zelterman-DLD consistently outperforms the Zelterman-POIS, exhibiting lower RBIAS and RRMSE across all scenarios.

Keywords: Zelterman-DLD, Zelterman-POIS, One-inflated Poisson distribution, Variance estimation

1. Introduction

Capture-recapture (CR) is a statistical method used to estimate the population size of a particular species or a group of individuals based on multiple sampling events. It is commonly employed in situations where it is impractical to directly count every member of the population of interest (Seber, 1982). CR methods were originally applied to animal populations in which sequence of samples were taken from a well-defined population (Jibasen, 2016). CR analysis involves repeatedly sampling a population and using recapture data to estimate the number of individuals that were not captured (Anan *et al*., 2017). Population size estimation under the capture-recapture method has been used to model hidden data such as the number of unseen species, the number of hidden drug users, the number of drink-driving offenders, and the number of illegal immigrants (Chao & Bunge, 2002; Heraud-Bousquet *et al.* 2012; Van der Heijden *et al*. 2003).

According to Piatek and Bohning (2024), the zero-truncated (ZT) Poisson distribution is often used as a starting point for modeling the frequencies of positive count data. The widespread of zero-truncated Poisson (ZTP) distribution has prompted the development of mixed-Poisson models for population size estimation in capture. Some of the mixed-Poisson models in capture-recapture includes the work of Rocchetti *et al*., (2011) and Bohning *et al*., (2013) who proposed population size estimators under a zero-truncated Poisson-Gamma mixture model. Pijitrattana (2018) developed a mixture of Poisson-Normal distributions to model the heterogeneity of an unobserved population. Wongprachan (2020) used a mixture of Poisson and Lindley distributions to estimate population size. Also, Wongprachan (2022) improved the Horvitz-Thompson estimator based on the zero-truncated Poisson-Shankar model for modeling hidden population size. However, the ZTP distribution assumes that the mean and variance are equal, which is frequently not the case due to unequal capture probabilities among sample units (heterogeneity) leading to overdispersion or under dispersion.

Rama and Simon (2018) proposed zero-truncated discrete Lindley (ZTDL) distribution and obtained some characteristics of the distribution. The application of ZTDL distribution to three datasets from the field of biology and demography demonstrated its suitability over competing models such as ZTP distribution and ZTPL distribution. This paper explores the issue of population size estimation by utilizing the Horvitz-Thompson (HT) estimator along with the zero-truncated discrete Lindley distribution to develop a Zelterman-type estimator. The effectiveness of this estimator is evaluated through simulations and real data applications, comparing its performance with the existing Zelterman Poisson estimator.

2. Methodology

2.1 Discrete Lindley distribution

The discrete Lindley distribution has been used to model frequency data in biological, ecological, health and epidemiological studies (Abebe & Shanker, 2018). The benefit of discrete Lindley model is that it is more flexible and have the same number of parameters with Poisson distribution as compared to negative binomial and mixed-Poisson. Considering Y, as a random variable which follows a discrete Lindley distribution with the parameter θ . The probability mass function of Y having the discrete Lindley distribution is defined by;

$$p_y(y;\theta) = \frac{\left(e^{\theta}-1\right)^2(1+y)e^{-\theta y}}{e^{2\theta}} \tag{1}$$

for $y = 0,1,2 \dots,;\ \theta > 0$

2.2 Zero-Truncated Discrete Lindley (ZTDL) Distribution

The zero-truncated distribution of Y obtained from equation (1) represents unidentified individuals from a target population having probability p_0.

$$\Pr(Y = y|y > 0) = \frac{\Pr(Y=y)}{1-\Pr(Y=0)}$$

It can be written as $P_y^+ = \frac{P_y}{1-p_0},\ y = 1,2,\ldots$ (2)

To obtain the value of p_0, let $y = 0$ in (1), thus, the expression for the unknown probability becomes;

$$p_0 = \frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}} \tag{3}$$

Thus, equation (2) which is the Zero-truncated Discrete Lindley distribution by Rama and Simon (2018) becomes;

$$P_y^+ = \frac{\left(e^{\theta}-1\right)^2(1+y)e^{-\theta y}}{2e^{\theta}-1} \tag{4}$$

2.3 Zero-truncated Count of Capture-Recapture Data

Zero-truncated count distribution modeling has been a longstanding method for estimating population size using CR data. This approach utilizes aggregate data on the number of sample units captured exactly once (f_1), twice (f_2), three times (f_3) till the last term (f_t), across multiple capture occasions. These counts are then summarized into a frequency distribution of discrete values. The observed population (n) consists of sample units captured at least once, while (f_0) represents the number of uncaptured sample units. The summation of the observed and unobserved units gives the total target population size as $N = n + f_0$.

2.4 Horvitz-Thompson Estimator

Horvitz-Thompson estimator as cited in Kaskasamkul, (2018) introduced a fundamental technique for estimating finite population using various sampling designs, whether with or without replacement. This method is commonly applied in capture-recapture studies to estimate the size N of the target population. Let $Y_i = 1$ as the identifying indicator variable for the i^{th} unit in the population. It takes a value of 1 if the i^{th} individual is identified, and 0 otherwise. The sum of Y_i from $i = 1$ to N gives the number of observed units. Each unit is observed independently with an identical probability of $1 - p_0$, which leads to a Binomial distribution for the probability of observing exactly n units. Additionally, $N(1 - p_0)$ represents the expected number of observed cases, which can be approximated by n, that is, $E(n) = N(1 - p_0)$.Thus, the estimate of the population size N is given by equation:

$$N = N(1 - p_0) + Np_0 \approx n + Np_0 \tag{5}$$

The equation can be rearranged and solved to estimate N, resulting in the Horvitz-Thompson estimator:

$$\widehat{N} = \frac{n}{1-p_0} \tag{6}$$

By using the Horvitz-Thompson approach and substituting p_0 , the population size estimator based on ZTDL distribution is;

$$\widehat{N} = \frac{ne^{2\hat{\theta}}}{2e^{\hat{\theta}}-1} \tag{7}$$

2.4 Proposed Zelterman-type Population Size Estimator

The Zelterman population estimator utilizes the ratio of the neighbouring probabilities $P_y^+(y;\theta)$ and $P_y^+(y+1;\theta)$ of the truncated count to estimate the parameter θ. The ratio of the neigbouring probabilities of the ZTDL probability is given;

$$\hat{\theta} = \frac{P_y^+(y+1;\theta)}{P_y^+(y;\theta)} \tag{8}$$

$$\frac{P_y^+(y+1;\theta)}{P_y^+(y;\theta)} = \frac{\left(e^{\theta}-1\right)^2(2+y)e^{-\theta(y+1)}}{2e^{\theta}-1} \times \frac{2e^{\theta}-1}{\left(e^{\theta}-1\right)^2(1+y)e^{-\theta y}}$$

$$\frac{P_y^+(y+1;\theta)}{P_y^+(y;\theta)} = \frac{(2+y)e^{-\theta y}e^{-\theta}}{(1+y)e^{-\theta y}}$$

Substituting the probabilities on the left-hand side with their corresponding relative frequencies, we have

$$\frac{\frac{f_{y+1}}{N}}{\frac{f_y}{N}} = \frac{(2+y)}{e^{\theta}(1+y)}$$

which can be expressed in terms of e^{θ} to obtain the unknown parameter θ

$$e^{\theta} = \frac{(2+y)f_y}{(1+y)f_{y+1}} \tag{9}$$

Taking the ln of both side

$$\hat{\theta} = ln\left(\frac{(2+y)f_y}{(1+y)f_{y+1}}\right) \tag{10}$$

Kuhnert and Böhning (2009) endorsed using $y = 1$ for two reasons: it provides the closest vicinal frequencies f_1 and f_2 to estimate f_0 , and most counts typically fall into the ones and twos in many applications. Zelterman as cited by Kaskasamkul (2018), asserted that individuals who are never seen should be more similar to those who are rarely seen, suggesting that $y = 1$.

Letting $y = 1$, the parameter of Zelterman-type estimator (Zelterman-DLD) is:

$$\hat{\theta} = ln\left(\frac{3f_1}{2f_2}\right) \tag{11}$$

substitute $\hat{\theta}$ into the population size estimator based on zero-truncated Discrete Lindley distribution in equation (7) to obtain the Zelterman-type estimator;

$$\widehat{N} = \frac{n\left(\frac{3f_1}{2f_2}\right)^2}{2\left(\frac{3f_1}{2f_2}\right)-1}$$

$$\widehat{N} = \frac{n(3f_1)^2}{2f_2(6f_1-2f_2)} \tag{12}$$

2.5 Variance Estimation of Zelterman-DLD Under Conditional Technique

The following conditional technique can be utilized to produce the variance of the Zelterman-DLD. The variance of the Zelterman-DLD comprises of two sources of variation arising from the random sample n, and the other as a result of the predictive value $\hat{p}_0$ based on the observed individuals n.

$$Var(\widehat{N}) = Var_n\{E(\widehat{N}/n)\} + E_n\{Var(\widehat{N}/n)\} \tag{13}$$

Bishop *et al.*, (2007) shows that $E(\widehat{N}/n) \approx \frac{n}{1-p_0}$, which is the same as Horvitz-Thompson expression in equation (6), thus solving the first term by the delta method becomes,

$$Var_n\{E(\widehat{N}/n)\} = Var_n\left\{\frac{n}{1-p_0}\right\} = \frac{1}{(1-p_0)^2}var(n) = \frac{Np_0(1-p_0)}{(1-p_0)^2} \tag{14}$$

Since $E(n) \approx N(1-p_0)$, the expression for the variance can be estimated thus,

$Var_n\{E(\widehat{N}/n)\} = \frac{np_0}{(1-p_0)^2}$. But from equation (3) and (6), p_0 can be estimated as:

$$\hat{p}_0 = \frac{\left(\left(\frac{3f_1}{2f_2}\right)-1\right)^2}{\left(\frac{3f_1}{2f_2}\right)^2} = \frac{(3f_1-2f_2)^2}{(3f_1)^2} \tag{15}$$

thus,

$$Var_n\{E(\hat{N}/n)\} = \frac{n(3f_1)^2(3f_1-2f_2)^2}{\left(12f_1f_2-4f_2^2\right)^2} \tag{16}$$

The second term on the right-hand side of equation (13)

$$Var(\hat{N}/n) = var\left(\frac{n\left(\frac{3f_1}{2f_2}\right)^2}{2\left(\frac{3f_1}{2f_2}\right)-1}\right) \tag{17}$$

where $w = \frac{\left(\frac{3f_1}{2f_2}\right)^2}{2\left(\frac{3f_1}{2f_2}\right)-1} = \left(\frac{9}{4}\right)\frac{f_1^2}{f_2(3f_1-f_2)}$ (18)

$$Var(\hat{N}/n) = var(nw) \approx n^2 var(w) \tag{19}$$

Hence, by applying the bivariate delta method on the expression $var(w)$, the approximation is obtained as;

$$var(w) \approx \nabla\varphi(f_1, f_2)^T cov(f_1, f_2)\nabla\varphi(f_1, f_2) \tag{20}$$

where,

$\nabla\varphi(f_1, f_2) = \begin{bmatrix} \left(\frac{\delta(w)}{\delta f_1}\right) \\ \left(\frac{\delta(w)}{\delta f_2}\right) \end{bmatrix}$ and $\nabla\varphi(f_1, f_2)^T = \begin{bmatrix} \left(\frac{\delta(w)}{\delta f_1}\right) & \left(\frac{\delta(w)}{\delta f_2}\right) \end{bmatrix}$

$$\frac{\delta(w)}{\delta f_1} = \left[\frac{27f_1^2f_2-18f_1f_2^2}{\left(2f_2(3f_1-f_2)\right)^2}\right] \tag{21}$$

Also, the derivative of w with respect to f_2 is obtained as follows;

$$\frac{\delta(w)}{\delta f_2} = \frac{18f_1^2f_2-27f_1^3}{\left(2f_2(3f_1-f_2)\right)^2} \tag{22}$$

The covariance matrix of the multinomial distribution conditional on n for frequencies one and two provided as

$$\widehat{cov}(f_1, f_2) = \begin{pmatrix} f_1\left(1-\frac{f_1}{n}\right) & \frac{-f_1f_2}{n} \\ \frac{-f_1f_2}{n} & f_2\left(1-\frac{f_2}{n}\right) \end{pmatrix} \tag{23}$$

Hence equation (20) becomes:

$$var(w) = \begin{pmatrix} \frac{27f_1^2 f_2 - 18 f_1 f_2^2}{\left(2f_2(3f_1 - f_2)\right)^2} & \frac{18 f_1^2 f_2 - 27\, f_1^3}{\left(2f_2(3f_1 - f_2)\right)^2} \end{pmatrix} \begin{pmatrix} f_1\left(1 - \frac{f_1}{n}\right) & \frac{-f_1 f_2}{n} \\ \frac{-f_1 f_2}{n} & f_2\left(1 - \frac{f_2}{n}\right) \end{pmatrix} \begin{pmatrix} \frac{27f_1^2 f_2 - 18 f_1 f_2^2}{\left(2f_2(3f_1 - f_2)\right)^2} \\ \frac{18 f_1^2 f_2 - 27\, f_1^3}{\left(2f_2(3f_1 - f_2)\right)^2} \end{pmatrix} \quad (24)$$

which can further be simplified as;

$$var(w) = \frac{324 f_1^3}{16[3f_1 - f_2]^4} - \frac{648 f_1^4}{16 f_2 [3f_1 - f_2]^4} - \frac{243 f_1^5}{16 f_2^2 [3f_1 - f_2]^4} + \frac{729 f_1^6}{16 f_2^3 [3f_1 - f_2]^4} \quad (25)$$

Hence, equation (19) becomes:

$$Var\left(\hat{N}/n\right) = n^2 var(w) = n^2 \left(\frac{324 f_1^3}{16[3f_1 - f_2]^4} - \frac{648 f_1^4}{16 f_2 [3f_1 - f_2]^4} - \frac{243 f_1^5}{16 f_2^2 [3f_1 - f_2]^4} + \frac{729 f_1^6}{16 f_2^3 [3f_1 - f_2]^4}\right) \quad (26)$$

Hence, adding the variance estimate in equation (16) and (26) to obtain equation (27)

Corollary 1. Consider the estimator $\hat{N} = \frac{n(3f_1)^2}{2f_2(6f_1 - 2f_2)}$ .Then:

$$Var\left(\hat{N}\right) = \frac{n(3f_1)^2(3f_1 - 2f_2)^2}{\left(12 f_1 f_2 - 4 f_2^2\right)^2} + n^2 \left(\frac{324 f_1^3}{16[3f_1 - f_2]^4} - \frac{648 f_1^4}{16 f_2 [3f_1 - f_2]^4} - \frac{243 f_1^5}{16 f_2^2 [3f_1 - f_2]^4} + \frac{729 f_1^6}{16 f_2^3 [3f_1 - f_2]^4}\right) \quad (27)$$

2.6 Simulation Study

The simulation was conducted to investigate the performance of the proposed estimator (Zelterman-DLD) and to compare it with Zelterman-POIS. The data were generated using R program and each condition was generated 1000 times. The population size was set at ($N = 100\ and\ 500$), under varying levels of one-inflation (10% and 20%) with parameter ($\lambda = 1.0, 1.05, 1.10, 1.15, 1.30, 1.35, 1.40, 1.45$). Zero counts from the generated population are replaced it with ones. Randomly sample 50% of the observations from the generated population, this yields zero-truncated one-inflated Poisson count data.

The performance of each of the estimators is measured in terms of relative bias (RBias) and relative root mean square error (RRMSE) given as:

$$RBias\left(\hat{N}\right) = \frac{1}{N}\left[E\left(\hat{N}\right) - N\right] \text{ and } RRMSE\left(\hat{N}\right) = \frac{1}{N}\sqrt{var\left(\hat{N}\right) + \left\{bias\left(\hat{N}\right)\right\}^2}$$

3. RESULT AND DISCUSSION

3.1 Results of the Simulation Study on the Estimated Population Size on One-inflated

Table 1 presents a comparison of RBias and RRMSE for the Zelterman-DLD and Zelterman-POIS estimators using one-inflated Poisson count data, simulated 1000 times. Figures 1 and 2 illustrate the estimated population size for a smaller population (N = 100) at λ = 1.15 and λ = 1.40, while Figures 3 and 4 depict the estimates for a larger population (N = 500) at λ = 1.05 and λ = 1.40. The results indicate that the Zelterman-DLD consistently provides more accurate and precise estimates compared to the Zelterman-POIS.

Table 1: Comparing RBias with RRMSE of Zelterman-DLD and Zelterman-POIS for one-inflated Poisson Count data, simulated 1000 times.

		Zelterman-DLD (N=100, n=50)			Zelterman-POIS (N=100, n=50)		
% one inflation	lambda	$\widehat{N}$	RBias	RRMSE	$\widehat{N}$	RBias	RRMSE
10%	1.00	111.198	0.112	0.3030	92.925	-0.0708	0.3196
	1.05	107.434	0.0743	0.2968	90.495	-0.0951	0.3202
	1.10	103.362	0.0336	0.2696	87.838	-0.1216	0.3071
	1.15	98.986	-0.0101	0.2662	85.021	-0.1498	0.3096
20%	1.30	112.664	0.1266	0.3259	93.901	-0.0610	0.3387
	1.35	109.449	0.0945	0.3022	91.796	-0.0820	0.3255
	1.40	103.914	0.0391	0.2837	88.212	-0.1179	0.3187
	1.45	104.711	0.0471	0.2916	88.738	-0.1126	0.3252
		Zelterman-DLD (N=500, n=250)			Zelterman-POIS (N=500, n=250)		
% one inflation	lambda	$\widehat{N}$	RBias	RRMSE	$\widehat{N}$	RBias	RRMSE
10%	1.00	526.191	0.0524	0.1294	444.658	-0.1107	0.1629
	1.05	506.901	0.0138	0.1202	432.211	-0.1356	0.1766
	1.10	490.409	-0.0192	0.1182	421.582	-0.1568	0.1893
	1.15	474.258	-0.0515	0.1199	411.190	-0.1776	0.2018
20%	1.30	520.989	0.0420	0.1295	441.317	-0.1174	0.1706
	1.35	508.562	0.0171	0.1204	433.278	-0.1334	0.1754
	1.40	497.850	-0.0043	0.1213	426.383	-0.1472	0.1846
	1.45	484.590	-0.0308	0.1207	420.085	-0.1598	0.1918

The performance of the two estimators, Zelterman-DLD and Zelterman-POIS, was evaluated using one-inflated Poisson count data for population sizes $N = 100$ and $N = 500$. The assessment focused on relative bias (RBias) and relative root mean square error (RRMSE) to determine the estimators' accuracy and precision.

Across all scenarios, the Zelterman-DLD consistently outperforms Zelterman-POIS by exhibiting lower RBias and RRMSE values. This indicates that Zelterman-DLD provides more accurate and precise population size estimates under one-inflated Poisson count data conditions.

For a population size of $N = 100$ with $n = 50$ the Zelterman-DLD produced the most accurate estimates, particularly at 10% one-inflation, where at $\lambda = 1.15$, the estimated population size was $\widehat{N} = 98.986$ with $RBias = -0.0101$ and $RRMSE = 0.2662$. At 20% one-inflation, the best estimate was $\widehat{N} = 103.914$ at $\lambda = 1.40$, with $RBias = -0.0391$ and $RRMSE = 0.2837$. These values closely approximate the true population size ($N = 100$), reinforcing the estimator's reliability and efficiency.

For a larger population size of $N = 500$ with $n = 250$, Zelterman-DLD continued to demonstrate superior performance. At $\lambda = 1.05$, the estimated population size was $\widehat{N} = 506.901$ with $RBias = 0.0138$ and $RRMSE = 0.1202$. Similarly, at $\lambda = 1.40$, the estimator yielded $\widehat{N} = 497.850$ with $RBias = -0.0043$ and $RRMSE = 0.1213$. These results highlight the robustness of the Zelterman-DLD, particularly in larger sample sizes, confirming its effectiveness in population size estimation under one-inflated Poisson models.

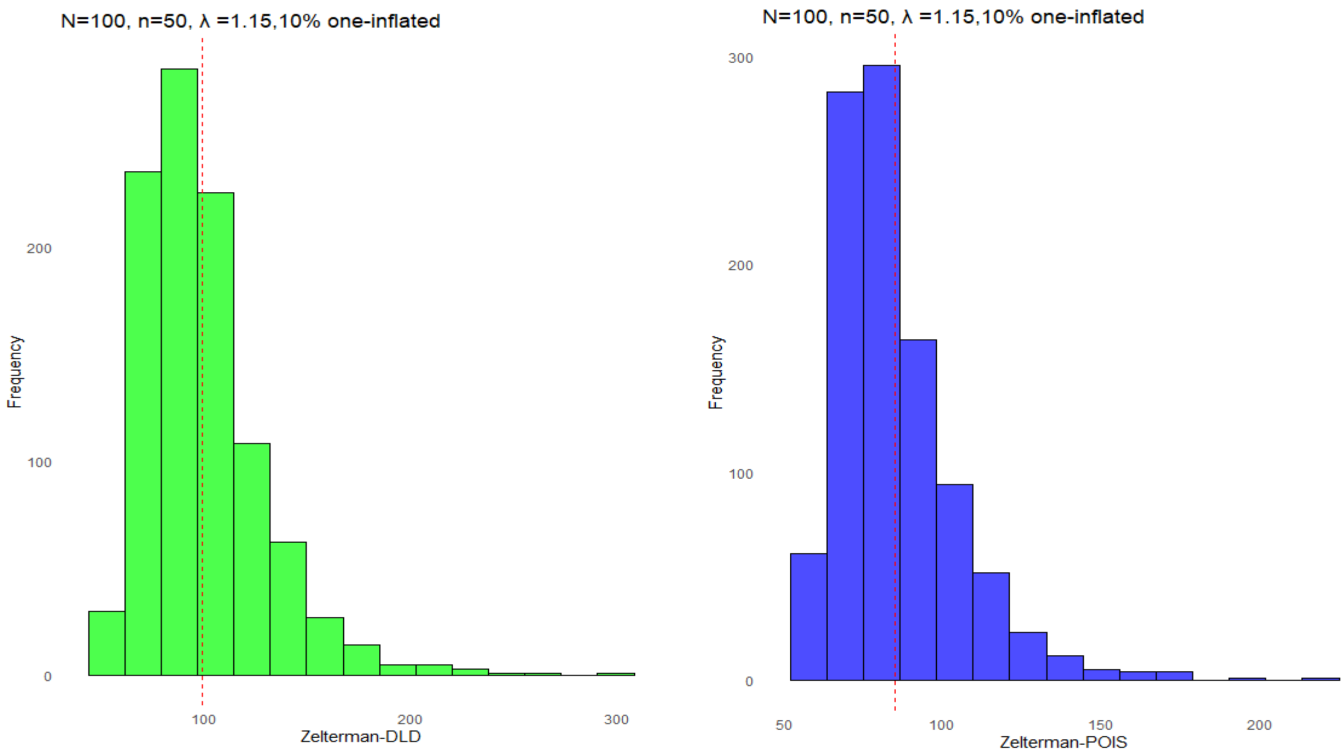

Figure 1: Histogram showing the estimates of Zelterman-DLD and Zelterman-POIS for $N = 100$, $n = 50$, $\lambda = 1.15$ at 10% one-inflated Poisson Count simulated 1000 times

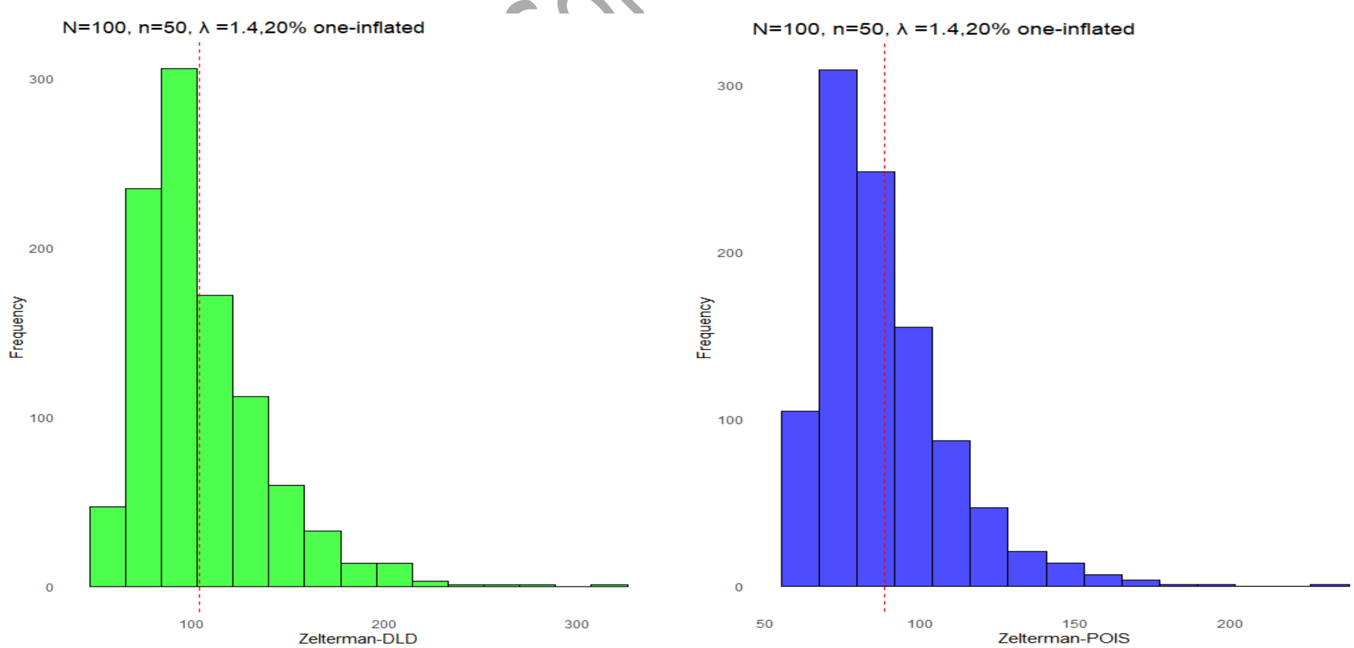

Figure 2: Histogram showing the estimates of Zelterman-DLD and Zelterman-POIS for $N = 100$, $n = 50$, $\lambda = 1.40$ at 20% one-inflated Poisson Count data simulated 1000 times.

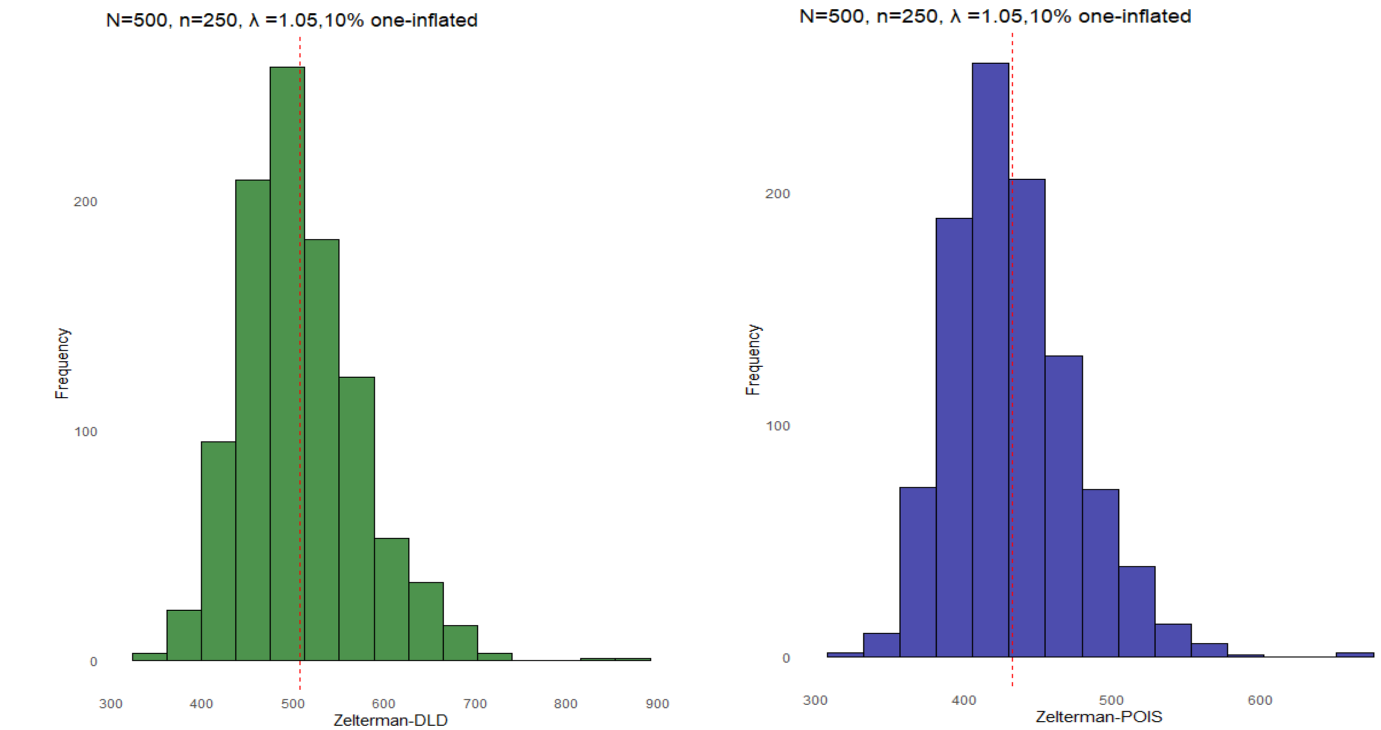

Figure 3: Histogram showing the estimates of Zelterman-DLD and Zelterman-POIS for $N = 500$, $n = 250$, $\lambda = 1.05$ at 20% one-inflated Poisson Count simulated 1000 times

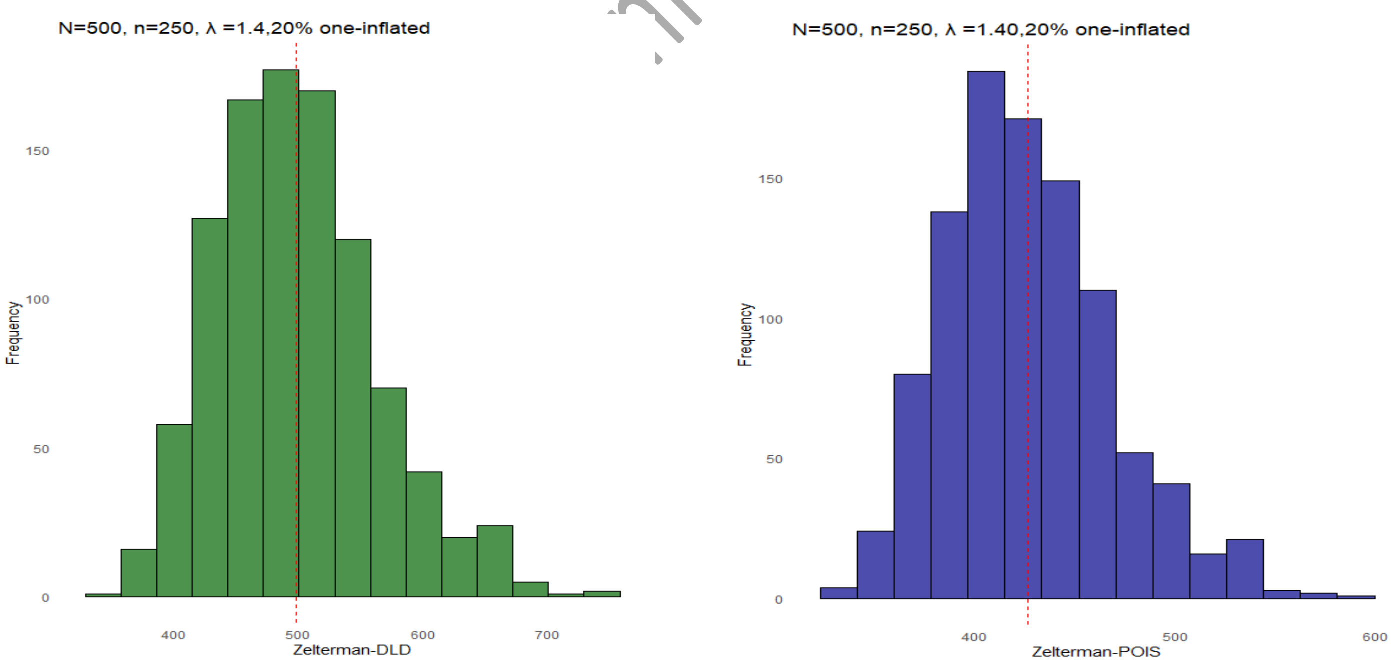

Figure 4: Histogram showing the estimates of Zelterman-DLD and Zelterman-POIS for $N = 500$, $n = 250$, $\lambda = 1.40$ at 20% one-inflated Poisson Count data simulated 1000 times.

3.2 Application with Real-life Datasets

In this section, two well-known datasets are used with the Zelterman-DLD and Zelterman Poisson estimators to demonstrate how the proposed estimator works in real-life situations.

Data 1 (Golf tees data)

In an experiment, 250 groups of golf tees were placed in a study area. Some were left visible above the grass, while others were hidden. Students from the 1999 statistics honors class at the University of St. Andrews (Scotland) recorded their observations (Borchers *et al*. 2004). They found 162 groups, but some were missed, and their total number needs to be estimated. The recorded counts for different group sizes were $(f_0, \ldots, f_8) = (88{,}46{,}28{,}21{,}13{,}23{,}14{,}6{,}11)$. This example is important for testing different estimation methods because the actual total number of groups is known.

Table 2. Results for Golf tees data with standard errors and confidence interval

Estimator	$\hat{f}_0$	$\widehat{N}$	$\widehat{SE}(\widehat{N})$	95% CI
Zelterman-POIS	68.11	230.11	29.14	172-289
Zelterman-DLD	88.42	250.42	22.90	207-294

Anan (2016) applied the Zelterman-POIS for population size using the popular Golf tees data, estimating the population size 231 with a 95% CI of 171–289 and a standard error of 29.9. The revisiting the performance of Zelterman-POIS, the results shows that the estimated number of missed Golf tees by Zelterman-POIS is 68.11, leading to a total population estimate of 230.11. The standard error (SE) is 29.14, indicating moderate variability in the estimate. The 95% confidence interval ranges from 172 to 289, which suggests reasonable uncertainty in the population estimate.

The estimated number of missed Golf tees by Zelterman-DLD is 88.42, leading to a total population estimate of 250.42, which is closest to the known true population size (250). The SE is the lowest among the two estimators (22.90), indicating a more precise estimate. The 95% CI is relatively narrow (207–294), suggesting that this estimator provides a more stable and reliable estimate. Comparing these two estimators, the Zelterman-DLD appears to be the most accurate, as its population estimate 250.42 is closest to the true population size (250). Additionally, it has the lowest standard error and the narrowest confidence interval, making it the most precise and reliable estimator in this case. while the Zelterman-POIS underestimates the total population. These results highlight the importance of selecting an appropriate estimator for population size estimation in capture-recapture studies.

Data 2 (Netherland illegal Immigrants Data)

In 1967, the Dutch police recorded data on illegal immigrants in the Netherlands. A total of 1,880 individuals were expelled, but some were caught more than once. The number of times they were apprehended was recorded as $(f_1, \ldots, f_6) = (1645,183,37,13,1,1)$. This paper revisits the data originally analyzed by Van der Heijden *et al*. (2003) and recently analyzed by Wongprachan (2020) to demonstrate the effectiveness of the proposed Zelterman-DLD method and variance estimation in estimating population size.

Table 3. Results for illegal immigrants with standard errors and confidence interval

Estimator	$\hat{f}_0$	$\hat{N}$	$\widehat{SE}(\hat{N})$	95% CI
Zelterman-POIS	7544.56	9424.56	683.97	8084-10765
Zelterman-DLD	11,282.69	13162.69	282.17	12610-137156

Previous studies have estimated the population size of illegal immigrants in the Netherlands using different statistical methods. Van der Heijden *et al.* (2003) applied a zero-truncated Poisson distribution, estimating the population size at 7,080 with a 95% CI of 6,363–7,797 and a standard error of 366. Similarly, Wongprachan (2020) used the Zero-Truncated

Poisson-Lindley (ZTPL) model, estimating the population size at 13,334 with a 95% CI of 12,073–14,595 and a standard error of 643.15.

In this study, the Zelterman-POIS estimated the number of unknown illegal immigrants at 7,544.56, leading to a total population estimate of 9,424.56. The standard error (SE) was 683.97, and the 95% CI ranged from 8,084 to 10,765. On the other hand, the Zelterman-DLD produced a higher estimate for the number of unknown individuals (11,282.69), resulting in a total population estimate of 13,162.69. This method had the lowest standard error (282.17), indicating a more precise estimate. Additionally, the 95% confidence interval was narrower (12,610–13,7156), suggesting that this estimator provides a more stable and reliable population estimate.

By comparison between the two estimators, the result indicates that Zelterman-DLD outperforms Zelterman-POIS in terms of precision, as evidenced by its lower standard error and narrower confidence interval. While the Zelterman-POIS estimate is closer to the findings of Van der Heijden *et al*. (2003), the Zelterman-DLD estimate aligns more closely with Wongprachan (2020). Given its stability and improved precision, the Zelterman-DLD appears to be the most reliable approach for estimating the hidden population of illegal immigrants in this dataset.

4. Conclusion

The findings demonstrates that the Zelterman-DLD is a superior choice for estimating population size from one-inflated Poisson count data, outperforming Zelterman-POIS in terms of bias and accuracy. The paper provides insight into how different values of λ and population sizes impact estimation accuracy. It establishes conditions under which Zelterman-DLD performs optimally, which can guide future research in similar count data settings. Additionally, real-world application was demonstrated and it proves that the Zelterman-DLD can serve as an alternative approach for population size estimation in CR settings.

References

Abebe, B., & Shanker, R. A. (2018). Discrete Lindley distribution with applications in biological sciences. *Biometrics and Biostatistics International Journal*, *7*, 48-52.

Anan, O. (2016). *Capture-recapture modelling for zero-truncated count data allowing for heterogeneity* (Doctoral dissertation, University of Southampton).

Anan, O., Böhning, D., & Maruotti, A. (2017). Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation*, 87(10), 2094-2114.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media.

Böhning, D., Baksh, M. F., Lerdsuwansri, R., & Gallagher, J. (2013). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics*, 22(1), 135-155.

Borchers, D. L., Buckland, S. T., Zucchini, W., & Borchers, D. L. (2002). *Estimating animal abundance: closed populations*. London: Springer

Chao, A., & Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, *58*(3), 531-539.

Héraud-Bousquet, V., Lot, F., Esvan, M., Cazein, F., Laurent, C., Warszawski, J., & Gallay, A. (2012). A three-source capture-recapture estimate of the number of new HIV diagnoses in children in France from 2003–2006 with multiple imputation of a variable of heterogeneous catchability. *BMC infectious diseases*, 12, 1-9.

Jibasen, D. (2016). Capture-Recapture Type Model for estimating Elusive Events with Tag Loss. *FUW Trends in Science & Technology Journal*, 1(2),517 - 521

Kaskasamkul, P. (2018). *Capture-recapture estimation and modelling for one-inflated count data* (Doctoral dissertation, University of Southampton).

Kuhnert, R., & Böhning, D. (2009). CAMCR: computer-assisted mixture model analysis for capture–recapture count data. *AStA Advances in Statistical Analysis*, 93(1), 61-71.

Piatek, M. E., & Böhning, D. (2024). Deriving a zero-truncated modelling methodology to analyse capture–recapture data from self-reported social networks. Metron, 82(2), 135-160.

Pijitrattana, M. P. (2018). *A flexible, discrete and smooth capture-recapture model based upon counts of repeated identifications using validation samples* (Doctoral Dissertation, Thammasat University).

Rama, S., & Simon, S. (2018). A zero-truncated discrete Lindley distribution with applications. *Int. J. of Statistics in Medical and Biological Research*, 2(1), 8-15.

Rocchetti, I., Bunge, J., & Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *The Annals of Applied Statistics*, 5(2B),1512-1533.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan.

Vander Heidjen, P.G., Bustami, R., Cruyff, M.J., Engbersen, G., and Van Houwelingen, H.C (2003). Point and interval estimation of the Population size using the truncated Poisson regression model. *Statistical Modelling*, 3 (4), 305-322.

Wongprachan, R. (2020). Modelling Population Size Using Horvitz-Thompson Approach Based on the Zero-Truncated Poisson Lindley Distribution. In *Numerical Computations: Theory* *and Algorithms: Third International Conference, NUMTA 2019, Crotone, Italy, June 15–21, 2019, Revised Selected Papers, Part II 3* (pp. 239-254). Springer International Publishing.

Wongprachan (2022). Hidden Population Size Estimation Under the Zero Truncated Poisson Shanker Model. Proceedings of the World Congress on Engineering, London, U.K, July 6-8, 2022.