BENIN JOURNAL OF PHYSICAL SCIENCES
BJPS Vol. 2(1), June, pg. 213-232 (2025)
ISSN 3043-6931(Print)
3043-694X(Online)
www.bjps.org.ng

Some Population Size Estimators Based on Zero-Truncated Discrete Lindley Distribution with Applications to Capture-Recapture Problems

N. S. AGOG[*,1], D. JIBASEN[2], S. S. ABDULKADIR[3], and B. Z. REUBEN[4]

[1]*Department of Mathematical Sciences, Faculty of Physical Sciences, Kaduna State University, Nigeria*
[1,2,3,4]*Department of Statistics, Faculty of Physical Sciences, Modibbo Adama University, Yola, Adamawa, Nigeria*

Received: 06/06/2025 ***Accepted:*** 29/06/2025

Abstract

Capture-recapture methods are essential for estimating hidden populations in fields such as public health and the social sciences. Traditional estimators based on the Poisson distribution often underestimate population sizes when the data exhibit overdispersion. To address this limitation, this study introduced the Zelterman-type and Mantel-Haenszel-type estimators within the zero-truncated discrete Lindley distribution. The conditional technique was utilized for variance estimation. The performance of the proposed estimators was assessed through simulation studies using one-inflated Poisson data across varying population sizes and inflation levels. The results showed that the proposed estimators outperformed traditional Poisson and geometric-based methods, yielding lower relative bias (RBias) and relative root mean square error (RRMSE). Real-life applications further validated these findings, where the Zelterman-type estimator produced near-exact results for golf tees data, ($\widehat{N}_{ZT}$=250.42, SE=22.90) compared to estimators from Poisson and geometric. The Mantel-Haenszel-type estimator proved effective in estimating undetected cases, particularly where traditional models were ineffective. Further findings show that the Zelterman-type estimator performs best on data that are highly right-skewed, slightly platykurtic, or leptokurtic, while the Mantel-Haenszel-type estimator performs better on moderately right-skewed data. In conclusion, the proposed estimator can serve as an alternative estimator in situations where the traditional estimators have failed.

Keywords: *Zelterman-type, Mantel-Haenszel-type estimator, One-inflated Poisson distribution, Variance estimation.*

*Corresponding author e-mail: nathan.agog@kasu.edu.ng

1. INTRODUCTION

Population size estimation under the capture-recapture method has been used to model hidden data. For instance, capture-recapture has been employed to estimate the number of heroin users in Northern Thailand (Pijitrattana, 2018), assess the population size of female sex workers in Vietnam (Nguyen *et al*., 2021), assess the population of European pond turtles in the wetlands of the Venice Lagoon (Liuzzo *et al*., 2021), and determine the number of farms affected by foot-and-mouth disease outbreaks in Southeast Asia (Sansamur *et al*., 2021), as well as to evaluate the completeness of COVID-19 contact tracing during the first wave of the pandemic in Thailand (Lerdsuwansri *et al*., 2022; Böhning *et al*., 2023).

Zero-truncated count distribution modeling has been a longstanding method for estimating population size using CR data. This approach uses 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. 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 $N = n + f_0$. According to Piatek and Bohning (2024), the zero-truncated Poisson distribution is often used as a starting point for modeling the frequencies of positive count data. However, the zero-truncated Poisson 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 underdispersion. The geometric distribution is mostly applied in modeling overdispersed capture-recapture data, leading to the development of various estimators. Anan *et al.* (2019) introduced a new Turing-type estimator for unknown population size in a CRC technique in which the count of identifications follows a geometric distribution. This is said to be a Poisson count adjusted for exponentially distributed heterogeneity. This estimator is suitable in CRC situations where the geometric distribution is better compared to the Poisson distribution.

Böhning *et al.* (2016) examined the ratios of neighbouring count probabilities. These ratios were estimated using observed frequencies, regardless of whether the distribution was zero-truncated or untruncated. The researchers explored a broader range of regression models, specifically those based on fractional polynomials, and showed that this approach yields a valid count distribution. They applied the proposed methodology to analyze different empirical applications and also conducted a simulation study to validate their findings.

In this paper, we are interested in exploring the discrete Lindley distribution in the capture-recapture technique by utilizing the Horvitz-Thompson (HT) method. The subsections in materials and methods section includes; the proposed

Zelterman-type estimator, the proposed Mantel-Haenszel-type estimator, variance estimation, while the subsection for the result and discussion includes; the performance of the proposed estimators through simulations and real data applications for population size estimation and finally the summary and conclusion.

2. MATERIALS AND METHOD

Horvitz-Thompson estimator (as cited in Kaskasamkul, 2018) introduced a fundamental technique for estimating a finite population using various sampling designs, with or without replacement. The equation of the Horvitz-Thompson estimator is given as:

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

Where n = observed individuals.

p_0= probability of unobserved individuals.

Abebe and Shanker (2018) introduced the discrete Lindley distribution and applied it in modeling frequency data in biological, ecological, health and epidemiological studies. The probability mass function of the discrete Lindley distribution is defined by;

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

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

Rama and Simon (2018) proposed the zero-truncated discrete Lindley (ZTDL) distribution, which is given by;

$$P_y^+ = \frac{\left(e^{\theta}-1\right)^2(1+y)e^{-\theta y}}{2e^{\theta}-1},\ y = 1,2, \ldots \tag{3}$$

The expression of the zero-truncated distribution is given as;

$$P_y^+ = \frac{P_y}{1-p_0},\ y = 1,2, \ldots \tag{4}$$

The unknown probability p_0, obtained from equation (2), is given as;

$$p_0 = \frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}} \tag{5}$$

Substituting p_0 into the Horvitz-Thompson estimator in equation (1) gives the capture-recapture model based on the zero-truncated discrete Lindley distribution as:

$$\widehat{N}_Z = \frac{n}{1-\left\{\frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}}\right\}} \tag{6}$$

2.1 Proposed Zelterman-type Population Size Estimator based on Zero-truncated discrete Lindley distribution.

The Zelterman population estimator utilises 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 neighbouring probabilities of equation (2) is given as;

$$\hat{\theta} = \frac{P_y^+(y+1;\theta)}{P_y^+(y;\theta)} \tag{7}$$

$$\frac{P_y^+(y+1;\theta)}{P_y^+(y;\theta)} = \frac{(2+y)e^{-\theta y}e^{-\theta}}{(1+y)e^{-\theta y}} \tag{8}$$

Substituting the probabilities with their relative frequencies into (8), 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} as follows;

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

Thus,

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

Kuhnert and Böhning (2009) endorsed using $y = 1$ because it provides the closest frequencies to estimate f_0 and most counts fall into f_1 and f_2 in most applications. Zelterman (as cited in 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) in equation (10) becomes:

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

substitute $\hat{\theta}$ into the population size estimator based on zero-truncated Discrete Lindley distribution in equation (6) to obtain the Zelterman-type estimator;

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

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

2.2 Proposed Mantel-Haenszel-type Population Size Estimator based on Zero-truncated discrete Lindley distribution.

The main idea in this method is that weight is added to the Zelterman-type estimator based on the zero-truncated discrete Lindley distribution. Thus, the parameter of the Zelterman-type estimator in equation (9) is utilized:

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

Let the weight be $(1+y)W_{y+1}$. By adding weight to the parameter of the Zelterman estimator based on the zero-truncated discrete Lindley distribution, it becomes;

$$e^{\widehat{\theta}_{MT}} = \frac{\sum_{y=1}^{k}(1+y)W_{y+1}\left(\frac{(2+y)f_y}{(1+y)f_{y+1}}\right)}{\sum_{y=1}^{k}(1+y)W_{y+1}}$$

Assuming that $W_{y+1} = f_{y+1}$, indicating that the weights represent frequency for each class y.

$$e^{\widehat{\theta}_{MT}} = \frac{\sum_{y=1}^{k}(1+y)f_{y+1}\left(\frac{(2+y)f_y}{(1+y)f_{y+1}}\right)}{\sum_{y=1}^{k}(1+y)f_{y+1}} \tag{13}$$

Thus, equation (13) becomes;

$$\hat{\theta}_{MT} = \ln\left(\frac{\sum_{y=1}^{k}(2+y)f_y}{\sum_{y=1}^{k}(1+y)f_{y+1}}\right) \tag{14}$$

Thus, the parameter of the Mantel-Haenszel estimator based on ZTDLD is:

$$\hat{\theta}_{MT} = \ln\left(\frac{3f_1+4f_2+5f_3+6f_4+7f_5+8f_6+\cdots+(2+k)f_k}{2f_2+3f_3+4f_4+5f_5+6f_6+\cdots+(1+k)f_{(1+k)}}\right) \tag{15}$$

$$\hat{\theta}_{MT} = \ln\left(\frac{S}{m}\right) \tag{16}$$

Where $S = 3f_1 + 4f_2 + 5f_3 + 6f_4 + 7f_5 + 8f_6 + \cdots + (2+k)f_k = \sum_{y=1}^{k}(2+y)f_y$

Where $m = 2f_2 + 3f_3 + 4f_4 + 5f_5 + 6f_6 + \cdots + (1+k)f_{(1+k)} = \sum_{y=1}^{k-1}(1+y)f_{y+1}$

One of the good attributes of the Mantel-Haenszel-type estimator is that it involves a lot of information about the frequency counts. Thus, substituting $\hat{\theta}_{MT}$ as θ into equation (5), it becomes:

$$\hat{p}_0 = \left(\frac{S-m}{S}\right)^2 \tag{17}$$

Hence, substituting $\hat{p}_0$ into equation (6) to obtain the Mantel-Haenszel-type population size estimator given as:

$$\widehat{N}_{MT} = \frac{n}{1-\left(\frac{S-m}{S}\right)^2} \tag{18}$$

2.3 Variance Estimation

The conditional technique by Bohning (2008) was utilized for deriving the variance of $\widehat{N}_{ZT}$. This variance comprises of two sources of variation arising from the random sample, and the other as a result of the predictive value, $\hat{p}_0$ based on the observed individuals n (full details is in Appendix A.1). The variance of $\widehat{N}_{MT}$ was derived based on one source of variation as utilized by Anan *et al.* (2019), where n is assumed to be fixed (full details is in Appendix A.2). The estimated variances of the Zelterman-type estimator and the Mantel-Haenszel-type estimator are given as;

$$Var(\widehat{N}_{ZT}) = \frac{n(3f_1)^2(3f_1-2f_2)^2}{(12f_1f_2-4f_2^2)^2} + n^2\left(\frac{324f_1^3}{16[3f_1-f_2]^4} - \frac{648f_1^4}{16f_2[3f_1-f_2]^4} - \frac{243f_1^5}{16f_2^2[3f_1-f_2]^4} + \frac{729f_1^6}{16f_2^3[3f_1-f_2]^4}\right) \quad (19)$$

$$Var(\widehat{N}_{MT}) = \frac{nS^2(S-m)^2}{(2mS-m^2)^2} \quad (20)$$

The 95% confidence interval of the Zelterman-type and the Mantel-Haenszel-type Population size estimator are constructed using the normal approximation approach based on the population size estimator and the estimated variance given as;

$$\widehat{N}_{ZT} \pm Z_{0.975}\sqrt{Var(\widehat{N}_{ZT})} \quad (21)$$

$$\widehat{N}_{MT} \pm Z_{0.975}\sqrt{Var(\widehat{N}_{MT})} \quad (22)$$

2.4 Simulation Study

Monte Carlo simulation was carried out in the R environment to investigate the performance of the proposed estimators and compared with some existing estimators. The data was generated for N= 60, 100, 1000, and 5000 population sizes repeated 1000 times. The mean of the Poisson distribution was considered at ($\lambda = 1.0, 1.05, 1.10$), and 10% one-inflation. The performance of each of the estimators was measured in terms of relative bias (RBias) and relative root mean square error (RRMSE) given as:

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

The proposed Zelterman-type estimator is compared with the Zelterman estimator introduced by Zelterman (as cited in Anan, 2016) under the Poisson distribution, and the extension of the Zelterman estimator under the geometric distribution by Anan (2016):

$$\widehat{N}_{ZP} = \frac{n}{1-exp\left(-\frac{2f_2}{f_1}\right)} \quad (22)$$

$$\widehat{Var}(\widehat{N}_{ZP}) = nG(\hat{\lambda})\left[1 + nG(\hat{\lambda})\hat{\lambda}^2\left(\frac{1}{f_1} + \frac{1}{f_2}\right)\right] \quad (23)$$

where $G(\hat{\lambda}) = \frac{exp(-\hat{\lambda})}{(1-exp(-\lambda))^2}$ and $\hat{\lambda} = \frac{2f_2}{f_1}$

$$\widehat{N}_{ZG} = \frac{nf_1}{f_2} \quad (24)$$

$$\widehat{Var}(\widehat{N}_{ZG}) = \frac{nf_1(f_1-f_2)}{{f_2}^2} + n^2\left(\frac{f_1}{f_2^2} + \frac{f_1^2}{f_2^3}\right) \quad (25)$$

The proposed Mantel-Haenszel-type estimator is compared with the Mantel-Haenszel estimator developed by Wannasirikul (as cited in Anan, 2016) under the Poisson distribution is given as:

$$\widehat{N}_{MH} = \frac{n}{1-\left\{exp\left(-\frac{S-f_1}{n-f_m}\right)\right\}} \quad (26)$$

Where $S = \sum_{x=1}^{m} xf_x$ and $n = f_1 + f_2 + \cdots + f_m$.

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-type, Zelterman-Poisson and Zelterman-Geometric estimators using 10% one-inflated Poisson count data, simulated 1000 times.

Table 1: *Comparing Population size estimates for Zelterman-type, Zelterman-Poisson, and Zelterman-Geometric for one-inflated Poisson distribution at 10% one-inflation, simulated 1000 times.*

	Zelterman-type			Zelterman-Poisson			Zelterman-Geometric		
Lambda	$\widehat{N}_{ZT}$	RBias	RRMSE	$\widehat{N}_{ZP}$	RBias	RRMSE	$\widehat{N}_{ZG}$	RBias	RRMSE
				$N = 60, n = 25$					
1.00	60.133	0.0022	0.4157	49.491	-0.1752	0.4886	70.261	0.1710	0.8207
1.05	60.501	0.0084	0.4557	49.696	-0.1717	0.5053	70.576	0.1763	0.8522
1.10	56.803	-0.0533	0.4068	47.325	-0.2113	0.4805	65.684	0.0947	0.7732
				$N = 100, n = 50$					
1.00	111.198	0.1120	0.3030	92.925	-0.0708	0.3196	128.675	0.2868	0.6297
1.05	107.434	0.0743	0.2968	90.495	-0.0951	0.3202	123.493	0.2349	0.5991

1.10	103.362	0.0336	0.2696	87.838	-0.1216	0.3071	117.953	0.1795	0.5515
				$N = 1000, n = 500$					
1.00	1046.22	0.0462	0.0918	885.17	-0.1148	0.1388	1201.06	0.2011	0.2494
1.05	1008.33	0.0083	0.0848	860.72	-0.1393	0.1570	1149.05	0.1491	0.2107
1.10	976.20	-0.0238	0.0863	840.03	-0.1600	0.1739	1104.83	0.1048	0.1835
				$N = 5000, n = 2500$					
1.00	5209.55	0.0419	0.0525	4411.30	-0.1177	0.1215	5977.33	0.1955	0.2046
1.05	5011.40	0.0023	0.0366	4283.42	-0.1433	0.1460	5705.59	0.1411	0.1530
1.10	4846.56	-0.0307	0.0450	4177.33	-0.1645	0.1667	5478.80	0.0958	0.1130

The performance of the three estimators, $\widehat{N}_{ZT}$, $\widehat{N}_{ZP}$, and $\widehat{N}_{ZG}$, were evaluated based on 10% one-inflated Poisson count data. The result shows that ($\widehat{N}_{ZT}$) produced estimates that were close to the true population size with relatively small Rbias and RRMSE as compared to $\widehat{N}_{ZP}$, and $\widehat{N}_{ZG}$. For $N = 60$, the best estimate was $\widehat{N}_{ZT} = 60.133$, $RBias = 0.0022$ and $RRMSE = 0.4157$, particularly at $\lambda = 1.00$. For $N = 100$, the best estimate was $\widehat{N}_{ZT} = 103.362$, $RBias = 0.0336$ and $RRMSE = 0.2696$, particularly at $\lambda = 1.10$. For a large population size $N = 1000$, the Zelterman-type estimator had the most accurate estimates, particularly at $\lambda = 1.05$, where $\widehat{N}_{ZT} = 1008.33$ with $RBias = 0.0083$ and $RRMSE = 0.0848$. The Zelterman-type estimator continues to demonstrate better performance by having estimates, $\widehat{N}_{ZT} = 5011.40$, $RBias = 0.0023$ and $RRMSE = 0.0366$ at $\lambda = 1.05$. The estimates of $\widehat{N}_{ZP}$, highly underestimate the true population size while $\widehat{N}_{ZG}$ consistently overestimate the true population size. Thus, $\widehat{N}_{ZT}$ is the most suitable estimator since it provides the smallest Rbias and RRMSE.

Table 2: *Simulation results of 95% confidence interval of Zelterman estimators*

	Zelterman-type		Zelterman-Poisson		Zelterman-Geometric	
Lambda	$\widehat{N}_{ZT}$	95% CI	$\widehat{N}_{ZP}$	95% CI	$\widehat{N}_{ZG}$	95% CI
			$N = 60, n = 25$			
1.00	60.133	39-81	49.491	12-87	70.261	-6-146
1.05	60.501	36-85	49.696	12-88	70.576	-7-148
1.10	56.803	36-77	47.325	13-82	65.684	-5-136
			$N = 100, n = 50$			
1.00	111.198	85-138	92.925	48-138	128.675	37-220
1.05	107.434	82-133	90.495	48-133	123.493	36-211
1.10	103.362	78-128	87.838	47-128	117.953	35-201
			$N = 1000, n = 500$			
1.00	1046.22	969-1122	885.17	758-1012	1201.06	942-1460
1.05	1008.33	934-1082	860.72	740-981	1149.05	902-1396
1.10	976.20	904-1048	840.03	725-955	1104.83	868-1341
			$N = 5000, n = 500$			
1.00	5209.55	5036-5375	4411.3	4129-4693	5977.33	5402-6552
1.05	5011.40	4848-5175	4283.42	4017-4550	5705.59	5160-6252
1.10	4846.56	4688-5006	4177.33	3923-4432	5478.80	4956-6001

Table 2 presents the 95% CI of the Zelterman estimators for different population sizes. The 95% CI of $\widehat{N}_{ZT}$ is the narrowest across all the population sizes, suggesting higher precision compared to $\widehat{N}_{ZP}$ and $\widehat{N}_{ZG}$.

In Table 3, the result shows comparison between the Mantel-Haenszel-type estimator and the Mantel-Haenszel Poisson estimator using one-inflated Poisson count data at 10% one-inflated.

Table 3: *Comparing Population size estimates for Mantel-Haenszel-type and Mantel-Haenszel-Poisson for one-inflated Poisson distribution at 10% one-inflation, simulated 1000 times.*

		Mantel-type			Mantel-Poisson	
Lambda	$\widehat{N}_{MT}$	RBias	RRMSE	95% CI	$\widehat{N}_{MP}$	Rbias
			$N = 60, n = 25$			
1.00	60.189	0.0032	0.2543	42-78	44.705	-0.2549
1.05	57.645	-0.0393	0.2393	41-75	43.020	-0.2830
1.10	55.497	-0.0751	0.2318	39-72	41.588	-0.3069

			$N = 100, n = 50$			
1.00	115.663	0.1566	0.2266	92-140	85.754	-0.1425
1.05	111.189	0.1119	0.1983	88-134	82.689	-0.1731
1.10	107.682	0.0768	0.1734	86-129	80.514	-0.1949
			$N = 1000, n = 500$			
1.00	1128.06	0.1281	0.1341	1054-1202	764.260	-0.2357
1.05	1087.01	0.0870	0.0959	1017-1156	730.096	-0.2699
1.10	1054.74	0.0547	0.0698	988-1122	705.491	-0.2945
			$N = 5000, n = 2500$			
1.00	5615.61	0.1231	0.1243	5452-5780	2951.84	-0.4096
1.05	5423.90	0.0848	0.0863	5268-5580	2825.27	-0.4349
1.10	5255.52	0.0511	0.0537	5106-5405	2699.77	-0.4600

For a very small size population, say ($N = 60$), the Mantel-Haenszel estimator under the zero-truncated discrete Lindley distribution (Mantel-Haenszel-type) produced an estimate of $\widehat{N}_{MT} = 60.189$ with an RBias of 0.0032 at $\lambda = 1.00$. For a small population size ($N = 100$), the Mantel-Haenszel-type estimator yields, $\widehat{N}_{MT} = 107.682$ and RBias = 0.0768 at $\lambda = 1.10$. Also, considering a large population $N = 1000$, the Mantel-Haenszel estimator under the zero-truncated distribution (Mantel-Haenszel-type estimator) had the best estimates, particularly at $\lambda = 1.10$, the population size estimate, $\widehat{N}_{MT} = 1054.74$ with $RBias = 0.0547$ performs better. For a very large population size $N = 5000$, the Mantel-Haenszel-type estimator continued to demonstrate better performance by having estimates with the least RBias as compared to Mantel-Poisson. The Mantel-Haenszel-type estimator produced estimates, $\widehat{N}_{MT} = 5255.52$, $RBias = 0.0537$ at $\lambda = 1.10$ performs better. This result shows that the Mantel-Haenszel-type estimator consistently performs better than the Mantel-Haenszel estimator under the Poisson distribution and can serve as an alternative estimator.

3.2 Applications to Real-life Datasets

This section provides three well-known datasets in the field of capture-recapture. Also, one new dataset on recidivism of offenders in Nigerian correctional Centre is introduced to enhance the applicability of the proposed estimators and for comparison with existing estimators.

Data on Golf tees (Borchers *et al.* 2002)

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. 162 groups of golf tees were

found, while the remaining were missed, and their total number needs to be estimated. Table 4 presents the frequency of the Golf tees data.

Table 4: *The frequency of Golf tees*

y	0	1	2	3	4	5	6	7	8
fy	88	46	28	21	13	23	14	6	11

To enhance comparison with previous studies, this research is considered alongside other notable estimators. Niwitpong *et al.* (2003) introduced the Chao estimator under the geometric distribution, which produced an estimated number of golf tees at 230, with the 95% CI ranging from 207 to 253. Rocchetti *et al*. (2014) applied a regression estimator under the beta-binomial distribution, yielding a lower estimate of 216 golf tees. This estimate was accompanied by a 95% CI of 193–238 under the asymptotic approximation, and 188–247 under a nonparametric approach.

Table 5: *Results for Golf tees data with standard errors and confidence interval (n=162, N=250)*

Estimator	$\hat{f}_0$	$\widehat{N}$	Bias	$\widehat{SE}(\widehat{N})$	95% CI	C.I length
Zelterman-Poisson	68.11	230.11	-19.89	29.14	172-289	117
Zelterman-Geometric	104.14	266.14	16.14	65.12	139-394	255
Zelterman-type	88.42	250.42	0.42	22.90	207-294	87

Anan (2016) applied the Zelterman estimator under the zero-truncated Poisson distribution for population size on the popular Golf tees data. The number of golf tees was estimated to be 231 with a 95% CI of 171–289 and a standard error of 29.9. Also, Anan *et al.* (2017) proposed the linear regression estimation based on the Conway-Maxwell-Poisson distribution (LCMP), this method reported the population of golf tees data as 223 with a standard error of 33.09 and a 95% C.I (159-288). Revisiting the performance of the Zelterman estimator under the Poisson distribution (Zelterman-Poisson), the result agrees with the findings of Anan (2016). The Zelterman estimator under the geometric distribution (Zelterman-Geometric) produced a total population estimate of 266.14, which overestimates the true population size of 250. The Standard error indicates less precision of the estimate, while the 95% CI (139–394) has a wide C.I length, suggesting that the estimate is unstable. The proposed Zelterman estimator under the discrete Lindley distribution (Zelterman-type) estimated a total population of 250.42, which is very close to the known true population. The Standard error

(22.90) is the least among the competing estimators, with the 95% CI (207-294), which has the shortest C.I length. The results indicate that the Zelterman-type estimator provides the best option for estimating the population size of the golf tees.

Data on Illegal immigrants in the Netherlands (Van der Heijden *et al.* (2003)

Another area of interest is the data on illegal immigrants recorded in the Netherlands. The record shows that a total of 1880 individuals were not effectively expelled, but some were caught more than once. The number of times they were apprehended was recorded as $(f_1, \dots, f_6) = (1645,183,37,13,1,1)$. To enhance this research, results from previous studies that have estimated the population size of illegal immigrants in the Netherlands were reviewed. Previous studies have estimated the population size of illegal immigrants not effectively expelled from the Netherlands using different statistical methods. Van der Heijden *et al.* (2003) applied various zero-truncated Poisson regression models on illegal immigrants not effectively expelled from the Netherlands. Their objective was to assess the extent of unobserved immigrant populations who were not effectively expelled. Among the models evaluated, the null model produced the lowest estimate of the total population size, yielding $\widehat{N}$=7080 with a 95% C.I (6363–7797). In contrast, the full model, which incorporated additional covariates, yielded the highest estimate $\widehat{N}$=12691 with a much wider C.I (7185–18198). However, model diagnostics based on Pearson residuals indicated a lack of fit, suggesting the presence of unobserved heterogeneity not adequately captured by the fitted models. The authors suggested that even the highest estimate of $\widehat{N}$=12691 is likely an underestimate of the true number of illegal immigrants residing in the Netherlands. Similarly, Wongprachan (2020) used the Zero-Truncated Poisson-Lindley (ZTPL) model, estimating the population size at 13334 with a 95% CI of 12073–14595 and a standard error of 643.15. Table 6 shows the estimates of illegal immigrant who were not effectively expelled from the Netherlands using the Zelterman estimators.

Table 6: *Results for illegal immigrants with standard errors and confidence intervals*

Estimator	$\hat{f}_0$	$\widehat{N}$	$\widehat{SE}(\widehat{N})$	95% CI	C.I length
Zelterman-Poisson	7544.56	9424.56	683.97	8084-10765	2681
Zelterman-Geometric	15,019.45	16899.45	1367.20	14220-19579	5359
Zelterman-type	11,282.69	13162.69	282.17	12610-13716	1106

Using the Zelterman-Poisson estimator, the number of unobserved illegal immigrants not effectively expelled was 7544.56, leading to a total of 9424.56 (8084-10765) and a standard error (683.97). The Zelterman-Geometric estimated the

number of unobserved illegal immigrants not effectively expelled as 15019.45, leading to a total population of illegal immigrants not effectively expelled as 16899.45 (14220-19579) with a standard error (1367.20). On the other hand, the Zelterman-type produced a moderate estimate for the number of unobserved illegal immigrant as 11282.69, resulting in a total population estimate of 13162.69. This method had the lowest standard error (282.17), indicating a more precise estimate. Additionally, the 95% C.I (12610–13716) has the shortest C.I length, suggesting that this estimator provides a more stable and reliable population estimate. Thus, the Zelterman-type estimate aligns more closely with the results obtained by Wongprachan (2020). Given the stability and improved precision of the Zelterman-type estimator, it appears to be the most reliable approach for estimating the hidden population of the illegal immigrants not effectively expelled from the Netherlands.

Data on recidivism of offenders in the Nigerian correctional centre (Oyewo, 2023)

The Nigerian Correctional Service has a history of having a high recidivism rate, this is seen in a study conducted by Oyewo (2023), that out of the 11930 prisoners found guilty in Nigerian jails, 6447 were convicted once, 2951 were convicted twice, 1469 were convicted three times, 536 were convicted four times, 295 were convicted five times, 232 were convicted six times or more. Table 8 shows the estimates of recidivism of offenders in the Nigerian correctional centres. The aim in this example is to estimate the number of offenders who missed rearrest, either due to limited resources, inadequate infrastructure, or not seen.

Table 8: *Estimates of Recidivism in Nigerian Jails with standard errors and confidence intervals for Zelterman Estimators*

Estimator	$\hat{f}_0$	$\widehat{N}$	$\widehat{SE}(\widehat{N})$	95% CI	C.I length
Zelterman-Poisson	7964.30	19894.30	293.78	19318-20470	1152
Zelterman-Geometric	14133.27	26063.27	605.34	24877-27250	2373
Zelterman-type	11136.95	23066.95	179.54	22715-23419	704

In this study, the Zelterman-Poisson estimated the number of offenders who missed rearrest, either due to limited resources, inadequate infrastructure, or not being seen completely as $(\hat{f}_0 = 7964.30)$, leading to a total population of recidivism among offenders as $(\widehat{N}_{ZP} = 19894.30, SE = 293.78\ )$ and the 95% CI ranged from 19318 to 20470. Zelterman-Geometric estimated the number of offenders who missed rearrest, either due to limited resources, inadequate infrastructure, or not being seen completely as $(\hat{f}_0 = 14133.27)$ leading to a total population size of recidivism as $(\widehat{N}_{ZG} = 26063.27,\ \ SE = 605.34)$, and the 95% CI

ranged from 24877 to 27250. On the other hand, the Zelterman-type produced a moderate estimate for the number of offenders who missed rearrest, either due to limited resources, inadequate infrastructure, or not being seen completely as $(\hat{f}_0 = 11136.95)$, resulting in a total population estimate of $(\widehat{N}_{ZT} = 23066.95,\ SE = 179.54)$. This method had the lowest standard error, indicating a more precise estimate. Additionally, the 95% (22715–23419) has the shortest C.I length, suggesting that this estimator provides a more stable and reliable population size estimate. Thus, the Zelterman-type estimator demonstrates a superior performance over both the Zelterman-Poisson and Zelterman-Geometric estimators in terms of precision, as reflected by its lower standard error and the short C.I length. This suggests that the Zelterman-type estimator is the most reliable method for estimating the number of recidivisms among offenders

Data on Bowel Cancer Patients (Lloyd and Frommer, 2004)

Lloyd and Frommer (2004) conducted a study at St. Vincent's Hospital, Sydney, involving 122 patients with confirmed bowel cancer status. Each patient underwent a sequence of binary diagnostic tests over six successive days, with the presence of blood in faeces recorded on each occasion. Patients who received negative results on all six tests were not further examined, leaving their true disease status unknown. Conversely, individuals with at least one positive result were confirmed to have the disease. The frequency distribution of the test result counts was reported as follows: The frequency was reported as $(f_0, \dots, f_6) = (22,8,12,16,21,12,31)$. Table 9 presents the results of the population size estimates of the Bowel cancer patients.

Table 9: *Results for Bowel Cancer with standard errors and confidence intervals*

Estimator	$\hat{f}_0$	$\widehat{N}$	Bias	$\widehat{SE}(\widehat{N})$	95% CI	C.I length
Zelterman-Poisson	5.24	105.24	-16.76	7.91	90-121	31
Zelterman-Geometric	-	66.67	-55.33	30.06	8-126	118
Zelterman-type	-	100	-22.00	70.71	-39-239	278
Mantel-Haenszel-type	13.16	113.16	-8.84	3.86	106-121	15
Mantel-Poisson	-	100	-22.00	-	-	

The Zelterman estimator based on the Poisson distribution estimated the total number of bowel cancer patients as 105.24 (90-121), out of which 5.24 represented the number of patients who received negative results on all six tests. The Zelterman estimators under the geometric, zero-discrete Lindley, and Mantel-Haenszel estimators under the Poisson distribution couldn't estimate the number

of patients who received negative results on all six tests. The Mantel-Haenszel estimator under the zero-truncated discrete Lindley distribution (Mantel-Haenszel-type) estimated the number of patients who received negative results on all six tests, $\hat{f}_0 = 13.16$, resulting in a total population of 113 bowel cancer patients. The proposed Mantel-Haenszel-type estimator produced a more reliable estimate, which is close to the true population of the bowel cancer patients. Table 10 provide a descriptive analysis of the real-life dataset to aid understanding of the data.

Table 10: *Descriptive statistics of the real-life data*

	Sum	Mean	Std. Dev	Variance	Skewness	Kurtosis
Golf tees	250	27.78	25.47	648.94	1.98	4.11
Immigrant	1880	313.33	656.05	430397.47	2.39	5.76
Recidivism	11930	1988.33	2414.02	5827499.87	1.62	2.34
Bowel	122	17.43	7.83	61.29	0.72	0.09

Data on golf tees, illegal immigrants not effectively expelled from the Netherlands, and recidivism of offenders in the Nigerian correctional centres were positively skewed with a sharper peak and heavier tails, reflecting the presence of more extreme values than a normal distribution. Also, the bowel cancer data was moderately skewed (skewness < 1) and less dispersed (kurtosis <1). The Zelterman-type estimator provided a good estimate for dataset that are strongly skewed to the right, while the Mantel-Haenszel-type estimator provides a good estimate for datasets that are less dispersed and moderately skewed to the right.

4. CONCLUSION

By introducing and validating the Zelterman-type and the Mantel-Haenszel-type estimators based on the zero-truncated discrete Lindley distribution, the research provides robust alternatives that significantly improve estimation precision. Simulation results confirmed that these estimators exhibit lower relative bias and error rates across varying conditions compared to their competitors. Applications to real-world datasets ranging from golf tees data, illegal immigrant populations in the Netherlands to recidivism of offenders in the Nigerian correctional centres, consistently showed that the Zelterman-type estimator performed better. The Mantel-Haenszel-type also proved essential in contexts where other models fail to account for hidden subpopulations such as the Bowel cancer patient data. The proposed estimators offer valuable tools for researchers and policymakers seeking more dependable population size estimates in complex and hidden populations. This study focused primarily on single-source capture-recapture data. Future

research should investigate the applicability of the proposed estimators to multi-list capture-recapture scenarios.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Abebe, B., & Shanker, R. A. (2018). Discrete Lindley distribution with applications in biological sciences. *Biometrics and Biostatistics International Journal*, *7*(1), 48-52.

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

[3] Anan, O., Böhning, D., & Maruotti, A. (2017). Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Statistical Methods & Applications*, *26*, 49-79. http://dx.doi.org/10.1007/s10260-016-0358-7

[4] Anan, O., Kanjanasamranwong, P., & Chantarangsri, W. (2019, December). Comparison of Confidence Intervals for the TG Estimator in Capture-recapture Data. In *Journal of Physics: Conference Series* (Vol. 1417, No. 1, p. 012017). IOP Publishing.

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

[6] Böhning, D., Lerdsuwansri, R., & Sangnawakij, P. (2023). Modeling COVID-19 contact-tracing using the ratio regression capture–recapture approach. *Biometrics*, *79*(4), 3818-3830.

[7] Böhning, D., Rocchetti, I., Alfó, M., & Holling, H. (2016). A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics*, *72*(3), 697-706.

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

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

[10] 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.

[11] Lerdsuwansri, R., Sangnawakij, P., Böhning, D., Sansilapin, C., Chaifoo, W., Polonsky, J. A., & Del Rio Vilas, V. J. (2022). Sensitivity of contact-tracing for

COVID-19 in Thailand: a capture-recapture application. *BMC Infectious Diseases*, *22*(1), 101.

[12] Lloyd, C. J., & Frommer, D. (2004). Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Australian & New Zealand Journal of Statistics*, *46*(4), 531-542.

[13] Liuzzo, M., Borella, S., Ottonello, D., Arizza, V., & Malavasi, S. (2021). Population abundance, structure and movements of the European pond turtle, Emys orbicularis (Linnaeus 1758) based on capture-recapture data in a Venice Lagoon wetland area, Italy. *Ethology Ecology & Evolution*, *33*(6), 561-575.

[14] Nguyen, L. T., Patel, S., Nguyen, N. T., Gia, H. H., Raymond, H. F., Hoang, V. T. H., & Abdul-Quader, A. S. (2021). Population Size Estimation of Female Sex Workers in Hai Phong, Vietnam: Use of Three Source Capture-Recapture Method. *Journal of Epidemiology and Global Health*, *11*(2), 194-199.

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

[16] Oyewo, O. O. (2023). Inmate Rehabilitation Idea and the Challenges of Policy Implementation in the Nigerian Correctional Centres. AKSU Journal of Administration and Corporate Governance, 3(3), 228-241

[17] 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.

[18] 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).

[19] 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.

[20] Rocchetti, I., Alfó, M., & Böhning, D. (2014). A regression estimator for mixed binomial capture–recapture data. *Journal of Statistical Planning and Inference*, *145*, 165-178. http://dx.doi.org/10.1016/j.jspi.2013.08.010

[21] Sansamur, C., Wiratsudakul, A., Charoenpanyanet, A., & Punyapornwithaya, V. (2021). Estimating the number of farms experienced foot and mouth disease outbreaks using capture-recapture methods. *Tropical Animal Health and Production*, *53*(1), 1-9.

[22] 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.

[23] 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.

APPENDIX

$$Var(\widehat{N}) = Var_n\{E(\widehat{N}/n)\} + E_n\{Var(\widehat{N}/n)\} \quad (27)$$

A.1 Variance Estimation of Zelterman-type estimator

The variance of the Zelterman-type 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. Note that $E(\widehat{N}/n) \approx \frac{n}{1-p_0}$, thus solving the first term by the delta method becomes,

$$Var_n\{E(\widehat{N}/n)\} = Var_n\left\{\frac{n}{1-p_0}\right\} = \frac{1}{(1-p_0)^2} var(n) = \frac{Np_0(1-p_0)}{(1-p_0)^2} \quad (28)$$

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

$Var_n\{E(\widehat{N}/n)\} = \frac{np_0}{(1-p_0)^2}$.

Note that for ZTDLD,

$$\hat{p}_0 = \frac{\left(\left(\frac{3f_1}{2f_2}\right)-1\right)^2}{\left(\frac{3f_1}{2f_2}\right)^2} = \frac{(3f_1-2f_2)^2}{(3f_1)^2} \quad (29)$$

thus,

$$Var_n\{E(\widehat{N}/n)\} = \frac{n(3f_1)^2(3f_1-2f_2)^2}{(12f_1f_2-4f_2^2)^2} \quad (30)$$

The second term on the right-hand side of equation (27)

$$Var(\widehat{N}/n) = var\left(\frac{n\left(\frac{3f_1}{2f_2}\right)^2}{2\left(\frac{3f_1}{2f_2}\right)-1}\right) \quad (31)$$

where $w = \frac{\left(\frac{3f_1}{2f_2}\right)^2}{2\left(\frac{3f_1}{2f_2}\right)-1} = \left(\frac{9}{4}\right)\frac{f_1^2}{f_2(3f_1-f_2)}$ (32)

$$Var(\widehat{N}/n) = var(nw) \approx n^2 var(w) \quad (33)$$

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

$$var(w) \approx \nabla\varphi(f_1, f_2)^T cov(f_1, f_2) \nabla\varphi(f_1, f_2) \tag{34}$$

where,

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

$$\frac{\delta(w)}{\delta f_1} = \left[\frac{27 f_1^2 f_2 - 18 f_1 f_2^2}{\left(2 f_2 (3 f_1 - f_2)\right)^2}\right] \tag{35}$$

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

$$\frac{\delta(w)}{\delta f_2} = \frac{18 f_1^2 f_2 - 27\, f_1^3}{\left(2 f_2 (3 f_1 - f_2)\right)^2} \tag{36}$$

Hence equation (34) becomes:

$$var(w) = \begin{pmatrix} \frac{27 f_1^2 f_2 - 18 f_1 f_2^2}{\left(2 f_2 (3 f_1 - f_2)\right)^2} & \frac{18 f_1^2 f_2 - 27\, f_1^3}{\left(2 f_2 (3 f_1 - f_2)\right)^2} \end{pmatrix} \begin{pmatrix} f_1\left(1 - \frac{f_1}{n}\right) & \frac{-f_1 f_2}{n} \\ \frac{-f_1 f_2}{n} & f_2\left(1 - \frac{f_2}{n}\right) \end{pmatrix} \begin{pmatrix} \frac{27 f_1^2 f_2 - 18 f_1 f_2^2}{\left(2 f_2 (3 f_1 - f_2)\right)^2} \\ \frac{18 f_1^2 f_2 - 27\, f_1^3}{\left(2 f_2 (3 f_1 - f_2)\right)^2} \end{pmatrix} \tag{37}$$

which can further be simplified as;

$$var(w) = \frac{324 f_1^3}{16[3 f_1 - f_2]^4} - \frac{648 f_1^4}{16 f_2 [3 f_1 - f_2]^4} - \frac{243 f_1^5}{16 f_2^2 [3 f_1 - f_2]^4} + \frac{729 f_1^6}{16 f_2^3 [3 f_1 - f_2]^4} \tag{38}$$

Hence, equation (33) becomes:

$$Var\left(\widehat{N}/n\right) = n^2 var(w) = n^2 \left(\frac{324 f_1^3}{16[3 f_1 - f_2]^4} - \frac{648 f_1^4}{16 f_2 [3 f_1 - f_2]^4} - \frac{243 f_1^5}{16 f_2^2 [3 f_1 - f_2]^4} + \frac{729 f_1^6}{16 f_2^3 [3 f_1 - f_2]^4}\right) \tag{39}$$

A.2 Variance estimation for the Mantel-Haenszel-type population size estimator

Here, we assume that the variation comes from only one source, and thus the variance is computed by the first term of the conditioning technique of variance estimation by Bohning (2008). This same approach was utilized by Anan *et al.* (2019), where n is assumed to be fixed.

$$Var\left(\widehat{N}_{MT}\right) = Var_n\left\{E\left(\widehat{N}/n\right)\right\} \tag{40}$$

Considering the parameter of the Mantel-Haenszel-type estimator $\hat{p}_0 = \left(\frac{S-m}{S}\right)^2$ and the associated population size estimator $\widehat{N}_{MT} = \frac{n}{1-\left(\frac{S-m}{S}\right)^2}$. Starting from the right-hand side of equation (41), where $E\left(\widehat{N}/n\right) \approx \frac{n}{1-p_0}$,

thus $Var_n\left(\frac{n}{1-p_0}\right) = \left(\frac{1}{1-p_0}\right)^2 Var_n(n)$ (41)

$Var_n\left(\frac{n}{1-p_0}\right) = \left(\frac{1}{1-p_0}\right)^2 Np_0(1-p_0)$ (42)

where $E(n) \approx N(1-p_0)$and $\hat{p}_0 = \left(\frac{S-m}{S}\right)^2$

hence, $Var_n\left(\frac{n}{1-p_0}\right) = \left(\frac{1}{1-p_0}\right)^2 np_0$ s

Thus, the estimated variance is;

$Var(\widehat{N}_{MT}) = Var_n\left(\frac{n}{1-p_0}\right) = \frac{nS^2(S-m)^2}{(2mS-m^2)^2}$ (43)