School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

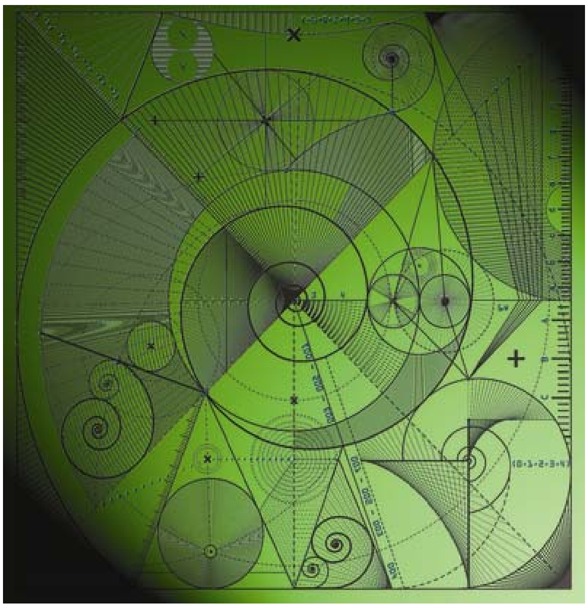
Preprint MPS_2010-26

22 June 2010

An Extension of Chao's Estimator of Population Size Based on the First Three Capture Frequency Counts

by

K. Lanumteang and D. Böhning



An Extension of Chao's Estimator of Population Size Based on the First Three Capture Frequency Counts

K. Lanumteang[a], D. Böhning[a]

[a]*Applied Statistics, School of Biological Sciences, The University of Reading, Philip Lyle Building, Reading, RG6 6BX, UK*

Abstract

A new estimator for estimating the size of an elusive target population is presented using frequency counts from capture-recapture sampling. The proposed estimator is developed by extending the idea of Chao's estimator using monotonicity of ratios of neighbouring frequency counts under a specific Poisson mixture sampling framework, the Poisson-Gamma mixture or negative binomial. The new estimator is achieved using a simple linear model on the basis of the log-ratio of neighbouring frequency counts as dependent variable which is valid under the Poisson-Gamma mixture. A simulation study is provided to study the performance of the proposed estimator under a variety of heterogeneous Poisson capture probabilities. Confidence interval estimation is done by means of an approximating normal approach and a modified bootstrap method, and was found to perform well. A variety of real data sets were also examined in order to illustrate the use of the proposed method.

Keywords:

Email addresses: k.lanumteang@reading.ac.uk (K. Lanumteang), d.a.w.bohning@reading.ac.uk (D. Böhning)

Estimation of population size, Capture-recapture methods, Horvitz–Thompson estimator, Chao's estimator, Homogeneous and heterogeneous Poisson models, negative binomial

1. Introduction

Estimation of the size of an elusive target population is of considerable interest in several fields. For example, ecologists commonly consider how to estimate the number of species in a wildlife population. In social sciences, there is major concern about certain social problems and determining its amount in a target population such as illicit drug users, violators of a law or the number of illegal immigrants. In medicine, there is wide interest in estimating the hidden disease occurrence, the unobserved part of the *disease iceberg* (Woodward, 1999). In public health and epidemiology, there is the frequent problem of determining the completeness of a disease registry (e.g. Corrao et al., 2000; Gallay et al., 2000; Hook and Regal, 1995; Nardone et al., 2003).

Capture-recapture models have been ordinarily used to estimate animal abundance or population size in the ecological sciences (see, for a review, Chao and Bunge, 2002; Darroch, 1958; Eberhardt, 1969; Edwards and Eberhardt, 1967; McDonald and Palanacki, 1989; North, 1981; Pollock, 2000). The origin of capture-recapture modelling goes back to Petersen and Lincoln (Seber, 2002), who used the independent information of two identifying sources or lists to construct an estimator of population size.

Capture-recapture models currently tend to be generally applied in a variety of applications including estimation of the size of a human target population, usually defined by a specific disease experiencing potential severe undercount (e.g. Böhning et al., 2004; Corrao et al., 2000; Gallay et al., 2000; Hay et al., 2009; Hook and Regal, 1995; Nardone et al., 2003; Smit et al., 2002; van Hest et al., 2008), as well as estimation of an elusive target population in the social sciences such as illegal gun owners or car drivers without license (e.g. Carothers, 1973; Chang et al., 1999; Hay, 1997; Hope et al., 2005; van der Heijden et al., 2003a,b).

Several estimators have been proposed to estimate the size of a target population when several identifications of the same unit are available. These include maximum likelihood methods, Zelterman's estimator (Zelterman, 1988), and Chao's lower bound estimators (Chao, 1987). However, several aspects of these estimators are critical. The maximum likelihood estimator is usually efficient only under Poisson homogeneity, whereas Chao's lower bound estimator - although developed under Poisson heterogeneity - uses only part of the available information and, hence, suffers under a lack of efficiency. To be more precise, let f_1 denote the frequencies of individuals which have been identified exactly once in the capture-recapture study, f_2 the number of individuals with exactly two identifications, and so forth, with m being the largest number of re-identifications. Then, $n = f_1 + f_2 + ... + f_m$ is the size of the observed sample. Chao's estimator is given as $\hat{N} = n + f_1^2/(2f_2)$ and it is clear from it's form that it uses only part of the available information, namely the proportion $(f_1 + f_2)/n$.

In this paper we propose a modification of this estimator, namely $\hat{N} = n + (3f_1f_3)/(2f_2^2) \times f_1^2/(2f_2)$ which extends the estimator of Chao by incorporating the adjustment factor $\hat{\gamma} = (3f_1f_3)/(2f_2^2)$. The central point of the paper is to show that this adjustment improves bias and efficiency of Chao's estimator under a wide class of models allowing heterogeneity.

2. The Proposed Estimator

The purpose of a capture-recapture model is to provide an estimator of the population size N or, equivalently, of the frequency of unobserved individuals f_0. From the individual capture-recapture history we can determine the count X of repeated identifications per individual. Let $f_1, f_2, f_3, ..., f_m$ denote the frequencies of distinct individuals identified exactly $1, 2, 3, ..., m$ times during the period of study, and f_0 is the frequency of individuals that were never identified in the study period and hence remain unobserved. Consequently, the total number of population size N can be written as $N = f_0 + f_1 + f_2 + ... + f_m = f_0 + n$, where $n = \sum_{j=1}^{m} f_j$ is the total number of distinct individuals observed. Furthermore, let p_0 be the probability that an individual remains unobserved, so that $E(f_0) = Np_0$. Therefore, we can also write the expected population size as $N = Np_0 + N(1 - p_0)$. Estimating $N(1 - p_0)$ with n leads to $\hat{N} = \frac{n}{1-p_0}$, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). The key issue is to estimate p_0.

Let p_j denote the probability for identifying an individual exactly j times, $j = 0, 1, 2, ..., m$. Under the Poisson distribution these probabilities are given as $p_0 = e^{-\lambda}, p_1 = e^{-\lambda}\lambda, p_2 = \frac{e^{-\lambda}\lambda^2}{2!}, ..., p_m = \frac{e^{-\lambda}\lambda^m}{m!}$ and $\frac{p_1}{p_0} = 2\frac{p_2}{p_1}$. Replacing the unknown Poisson probabilities by observed frequencies provides $f_1/f_0 = 2f_2/f_1$ as an estimating equation for f_0 and Chao's estimator $\hat{f}_0 = f_1^2/(2f_2)$ follows. However, Poisson homogeneity is rarely met in practice and it is more appropriate to incorporate heterogeneity of the identifying probability it is more reasonable to assume that the actual target population may consist of a variety of subgroups. This leads to a Poisson mixture model of the form

$$p_j = \int_0^\infty \frac{e^{-\lambda}\lambda^j}{j!} f(\lambda) d\lambda, \tag{1}$$

where $f(\lambda)$ represents the heterogeneity distribution of the model parameter in the population. A prominent example for $f(\lambda)$ is the Gamma-distribution $f(\lambda) = \theta'\lambda^{k-1}\exp(-\lambda/\theta')/\Gamma(k)$ with parameters $\theta', k > 0$, so that p_j is Poisson-Gamma mixture, or if the marginal is worked out, the *Negative Binomial* distribution. Let $r_j = \frac{jp_j}{p_{j-1}}$, where $p_j = \frac{\Gamma(k+j)}{\Gamma(j+1)\Gamma(k)}\theta^k(1-\theta)^j$ with $\theta' = (1-\theta)/\theta$, then we achieve $r_j = (k+j-1)(1-\theta)$. This clearly implies that there is a linear relationship $r_j = (k-1)(1-\theta) + (1-\theta)\ j$ between r_j and j. Plotting r_j against j leads to the *ratio plot* (see Figure 1 for an illustration), and specific patterns indicate a certain distribution, such as linearity indicates a negative binomial, a horizontal line means the presence of a Poisson distribution and a line passing through the origin indicates a geometric distribution.

Please insert Figure 1 here

To derive our estimator we consider a Taylor expansion of $\log r_j$ around $(k-1)$ so that

$$\log r_j = \log(k+j-1) + \log(1-\theta) \approx \underbrace{\log(1-\theta) + \log(k-1)}_{\alpha} + \overbrace{\frac{1}{k-1}}^{\beta}\ j. \tag{2}$$

The motivation for the approximation (2) is as follows. Using a logarithmic transformation will guarantee that our population size estimate is feasible (which is not necessarily so when working on the r_j scale). Now, for $j = 2$ or $j = 3$ in (2) we get $\log(r_2) = \log(\frac{2f_2}{f_1}) = \alpha + 2\beta$ and $\log(r_3) = \log(\frac{3f_3}{f_2}) = \alpha + 3\beta$. Solving these equations in α and β can easily be achieved as $\hat{\alpha} = 3\log(\frac{2f_2}{f_1}) - 2\log(\frac{3f_3}{f_2})$ and $\hat{\beta} = \log(\frac{3f_3}{f_2}) - \log(\frac{2f_2}{f_1})$. Then, plugging $\hat{\alpha}$ and $\hat{\beta}$ into (2) and using $j = 1$, (2) provides $\log(r_1) = \log(\frac{f_1}{f_0}) = \alpha + \beta$, or

$$\log(\frac{f_1}{f_0}) = 3\log(\frac{2f_2}{f_1}) - 2\log(\frac{3f_3}{f_2}) + \log(\frac{3f_3}{f_2}) - \log(\frac{2f_2}{f_1}) = 2\log(\frac{2f_2}{f_1}) - \log(\frac{3f_3}{f_2}).$$

Finally, we achieve that $\log(f_0) = \log(f_1) - \log(\frac{4f_2^2}{f_1^2}) + \log(\frac{3f_3}{f_2}) = \log(\frac{3f_1^3 f_3}{4f_2^3})$. Hence, our estimator for f_0 and N, respectively, is

$$\hat{f}_{0New} = \frac{3f_1^3 f_3}{4f_2^3} \text{ and } \hat{N}_{New} = n + \frac{3f_1^3 f_3}{4f_2^3}. \tag{3}$$

3. Properties of the Proposed Estimator

In this section we summarize some properties of the new estimator. Firstly, it should be noted that (3) is closely associated with Chao's estimator $\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}$ (Chao, 1987) in that we can think of (3) as an adjusted Chao estimator of the form $\hat{N}_{New} = n + \frac{f_1^2}{2f_2}\hat{\gamma}$; where $\hat{\gamma} = \frac{3f_1 f_3}{2f_2^2}$. Hence, we investigate the effect of this adjustment factor.

Theorem 1. *Under arbitrary mixing in (1) we have that*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{New})}{N} \geq \lim_{N\to\infty} \frac{E(\hat{N}_{Chao})}{N}$$

and

$$\lim_{N\to\infty} E(\hat{\gamma}) = \lim_{N\to\infty} E\left(\frac{3f_1 f_3}{2f_2^2}\right) = \frac{3p_1 p_3}{2p_2^2} \geq 1.$$

Proof.

As a consequence of the Cauchy-Schwarz inequality we have for arbitrary mixing that the ratios of neighbouring count probabilities experience a monotonicity property as follows (see Chao, 1987)

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq \frac{4p_4}{p_3} \leq ...,$$

so that in particular $\frac{2p_2}{p_1} \leq \frac{3p_3}{p_2}$. Now, $E(\hat{f}_{0New})/N = E(\frac{3f_1^3 f_3}{4f_2^3})/N \to \frac{3}{4}(\frac{p_1^3 p_3}{p_2^3}) = \frac{3}{2}(\frac{p_1 p_3}{p_2^2})(\frac{p_1^2}{2p_2})$ and $E(\hat{f}_{0Chao})/N \to \frac{p_1^2}{2p_2}$ for $N \to \infty$. It remains to show that $\frac{3}{2}\frac{p_1 p_3}{p_2^2} \geq 1$. The latter follows from $\frac{2p_2}{p_1} \leq \frac{3p_3}{p_2}$ which also implies the second part of the theorem, and this ends the proof.

Chao's estimator is a lower bound estimator in the sense that $E(\hat{N}_{Chao})/N \leq 1$ for $N \to \infty$ using that $\frac{p_1}{p_0} \leq \frac{2p_2}{p_1}$. Hence typically Chao's estimator will underestimate the population size. The property in Theorem 1 is remarkable since it guarantees that the asymptotically expected value of (3) is larger than that of Chao's estimator – under fairly general conditions. Next we show that (3) is asymptotically unbiased under Poisson homogeneity – as is Chao's estimator.

Theorem 2. *Under Poisson homogeneity* $p_j = e^{-\lambda}\lambda^j/j!$ *we have that*

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

Proof.

$E(f_j/N)$ converges with $N \to \infty$ to p_j. Hence, $E(\frac{\hat{f}_{0New}}{N}) = E(\frac{3(f_1/N)^3(f_3/N)}{4(f_2/N)^3})$ converges to $\frac{3p_1^3 p_3}{4p_2^3} = e^{-\lambda}$. Finally, $E(\hat{N}_{New}/N) = E(n + \hat{f}_{0New})/N$ converges to $(1 - e^{-\lambda}) + e^{-\lambda} = 1$ and ends the proof.

The next result compares the asymptotic biases for the new and Chao's estimator.

Theorem 3. *Under Poisson heterogeneity according to a Gamma distribution, e.g. (1) is the negative binomial* $p_j = \frac{\Gamma(k+j)}{\Gamma(j+1)\Gamma(k)}\theta^k(1-\theta)^j$ *for* $j = 0, 1, 2, ...$ *we have that*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{New})}{N} = 1 - \frac{\theta^k}{(k+1)^2}$$

and

$$\lim_{N\to\infty} \frac{E(\hat{N}_{Chao})}{N} = 1 - \frac{\theta^k}{k+1},$$

with $1 - \frac{1}{k+1} \leq 1 - \frac{1}{(k+1)^2} \leq 1$.

Proof.

We have for $N \to \infty$ that

$$\begin{aligned} E(\hat{f}_{0New})/N &= E(\frac{3f_1^3 f_3}{4f_2^3})/N \\ &\to \frac{3}{4}(\frac{\frac{k!^3}{(k-1)!^3}\theta^{3k}(1-\theta)^3\frac{(k+2)!^3}{3!(k-1)!}\theta^k(1-\theta)^3}{\frac{(k+1)!^3}{2!^3(k-1)!^3}\theta^{3k}(1-\theta)^6}) \\ &= \frac{k(k+2)}{(k+1)^2}\theta^k, \end{aligned}$$

so that $E(\hat{N}_{New})/N \to (1-\theta^k) + \frac{k(k+2)}{(k+1)^2}\theta^k = (1-\theta^k + \frac{k(k+2)}{(k+1)^2}\theta^k) = 1-\theta^k/(k+1)^2$. On the other hand,

$$\begin{aligned} E(\hat{f}_{0Chao})/N &= E(\frac{f_1^2}{2f_2})/N \\ &\to \frac{1}{2}(\frac{\frac{k!^2}{(k-1)!^2}\theta^{2k}(1-\theta)^2}{\frac{(k+1)!}{2!(k-1)!}\theta^k(1-\theta)^2}) \\ &= \frac{k}{k+1}\theta^k. \end{aligned}$$

and $\hat{N}_{Chao}/N \to (1-\theta^k) + \frac{k}{k+1}\theta^k = (1-\theta^k + \frac{k}{k+1}\theta^k) = 1-\theta^k/(k+1)$.

The result in Theorem 3 indicates the large potential of reducing bias with the new estimator. To explore this a bit further we consider exponential mixing in (1).

Corollary 1. *Let the mixing density $f(\lambda)$ in (1) be the exponential, $k = 1$, so that the marginal (1) is the geometric. Then:*

$$\lim_{N\to\infty} \frac{E(\hat{N}_{New})}{N} = 1 - \frac{\theta}{4} \text{ and } \lim_{N\to\infty} \frac{E(\hat{N}_{Chao})}{N} = 1 - \frac{\theta}{2}.$$

The condition in corollary 1 might appear difficult to be checked. However, exponential mixing means that the shape parameter k equals one which implies that the line in the ratio plot passes through the origin. This can be simply diagnosed and formally tested. An asymptotic unbiased Chao-type estimator for this case ($k = 1$) is provided as $n + f_1^2/f_2$ and an asymptotic unbiased estimator incorporating the first three capture frequency counts is also available as $n + f_1^3 f_3/f_2^3$.

Note that (3) is only well-defined as long as f_2 is positive. Therefore, we suggest to use a modification of (3) which allows $f_2 = 0$, as follows

$$\hat{N}_{New_{Mo}} = n + \frac{3}{4}\frac{f_1(f_1-1)(f_1-2)f_3}{(f_2+1)(f_2+2)(f_2+3)}. \quad (4)$$

In addition, we consider the following truncated version of $\hat{N}_{New}$ to improve its variance. It can be seen from Theorem 3 (and by replacing f_j by their theoretical value p_j) that the expected value of $\hat{\gamma} = \frac{3f_1f_3}{2f_2^2}$ approaches

$$\frac{3\Gamma(k+1)\Gamma(k+3)\Gamma(3)^2\Gamma(k)^3}{2\Gamma(2)\Gamma(k)\Gamma(4)\Gamma(k)\Gamma(k+2)^2} = \frac{k+2}{k+1}$$

for N becoming large assuming the negative binomial distribution for the count probabilities p_j, $j = 0, ..., m$. Note that

$$1 \leq \frac{k+2}{k+1} \leq 2$$

for $0 \leq k \leq \infty$. Hence, truncation at the upper and lower bound of the asymptotically expected value of the multiplier $\hat{\gamma}$ appears reasonably and leads to an adjusted form $\hat{N}_{New}$ as follows:

$$\hat{N}_{NewAdj} = \begin{cases} n + \frac{f_1^2}{2f_2}, & \text{if} \quad \frac{3f_1 f_3}{2f_2^2} \leq 1 \\ n + \frac{f_1^2}{2f_2}(\frac{3f_1 f_3}{2f_2^2}), & \text{if} \quad 1 < \frac{3f_1 f_3}{2f_2^2} < 2 \\ n + \frac{f_1^2}{f_2}, & \text{if} \quad \frac{3f_1 f_3}{2f_2^2} \geq 2. \end{cases} \tag{5}$$

The adjusted form (5) can be expected to show an improved performance in terms of reducing the variance while retaining the reduction in bias, which will be seen in the simulation study section.

4. Variance Estimator and Confidence interval

4.1. Variance Estimator

In order to investigate the variance of the proposed estimator we simply derive it by conditioning. It can be noted that the variation of $\hat{N}_{New} = n + \frac{3f_1^3 f_3}{4f_2^3}$ is arising from two sources, the random variation of sampling n individuals from N and the random variation with respect to estimation of $\hat{\lambda_0}$ where $\hat{\lambda_0} = \frac{3f_1^3 f_3}{4f_2^3}$. Böhning (2008) provided a simple formula for variance computation of population size which can be also applied to derived the variance approximation of the new proposed estimator as follows:

$$Var_{\hat{\lambda}_0|n}(n + \hat{\lambda_0}) = \underbrace{E_n\{Var_{\hat{\lambda}_0|n}(n + \hat{\lambda_0})\}}_{[1]} + \underbrace{Var_n\{E_{\hat{\lambda}_0|n}(n + \hat{\lambda_0})\}}_{[2]}, \tag{6}$$

where E_n and Var_n refer to the marginal distribution of n and $\hat{\lambda_0} = \frac{3}{4}\frac{f_1^3 f_3}{f_2^3}$. Assuming that $E_{\hat{\lambda_0}|n}(n + \hat{\lambda_0})$ in the second term [2] of (6) can be estimated by $n + \hat{\lambda_0}$ we have that

$$Var_n\{E_{\hat{\lambda_0}|n}(n + \hat{\lambda_0})\} = \widehat{Var_n}\{n + \hat{\lambda_0}\} = Var_n\{n\} = Np_0(1 - p_0). \quad (7)$$

Since $\hat{p_0} = \frac{\hat{f_0}}{n+\hat{f_0}}$ and $N\widehat{(1 - p_0)} = n$, (7) can be estimated by

$$\widehat{Var_n}\{E_{\hat{\lambda_0}|n}(n + \hat{\lambda_0})\} = \frac{\frac{3n}{4} f_1^3 f_3}{n f_2^3 + \frac{3}{4} f_1^3 f_3}. \quad (8)$$

Now, consider the first term in (6), $E_n\{Var_{\hat{\lambda_0}|n}(n + \hat{\lambda_0})\}$, and assume again that $E_n\{Var_{\hat{\lambda_0}|n}(n+\hat{\lambda_0})\}$ can be estimated by $Var_{\hat{\lambda_0}|n}(n+\hat{\lambda_0}) = Var_{\hat{\lambda_0}|n}(\frac{3}{4}\frac{f_1^3 f_3}{f_2^3})$. Using the multivariate delta-method (see Bishop et al., 1975) we are able to achieve that

$$Var_{\hat{\lambda_0}|n} = \nabla g \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}^T Cov \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \nabla g \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}, \quad (9)$$

where $g(f_1, f_2, f_3) = \frac{3}{4}\frac{f_1^3 f_3}{f_2^3}$ and $\nabla_i g(f_1, f_2, f_3) = \frac{\partial}{\partial f_i} g(f_1, f_2, f_3)$. It is easy to see that

$$\nabla g \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} \frac{9}{4}\frac{f_1^2 f_3}{f_2^3} & -\frac{9}{4}\frac{f_1^3 f_3}{f_2^4} & \frac{3}{4}\frac{f_1^3}{f_2^3} \end{pmatrix}^T. \quad (10)$$

Recall that the covariance matrix of the multinomial vector $(f_1, f_2, f_3)^T$ is estimated by

$$\hat{Cov} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} f_1(1 - \frac{f_1}{n}) & -\frac{f_1 f_2}{n} & -\frac{f_1 f_3}{n} \\ -\frac{f_1 f_2}{n} & f_2(1 - \frac{f_2}{n}) & -\frac{f_2 f_3}{n} \\ -\frac{f_1 f_3}{n} & -\frac{f_2 f_3}{n} & f_3(1 - \frac{f_3}{n}) \end{pmatrix}. \quad (11)$$

Hence (9) becomes ultimately

$$Var_{\hat{\lambda}_0|n}(\frac{3}{4}\frac{f_1^3 f_3}{f_2^3}) = (\frac{9}{4})^2 \frac{f_1^5 f_3^2}{f_2^6}\{\frac{f_1}{f_2}+1\} + (\frac{3}{4})^2 \frac{f_1^6 f_3}{f_2^6}\{1-\frac{f_3}{n}\}. \tag{12}$$

Substituting (8) and (12) into (6), we finally have that

$$Var_{\hat{\lambda}_0|n}(n+\frac{3}{4}\frac{f_1^3 f_3}{f_2^3}) = (\frac{9}{4})^2 \frac{f_1^5 f_3^2}{f_2^6}\{\frac{f_1}{f_2}+1\} + (\frac{3}{4})^2 \frac{f_1^6 f_3}{f_2^6}\{1-\frac{f_3}{n}\} + \frac{\frac{3n}{4} f_1^3 f_3}{n f_2^3 + \frac{3}{4} f_1^3 f_3}. \tag{13}$$

It is seen from (13) that the first term $(\frac{9}{4})^2 \frac{f_1^5 f_3^2}{f_2^6}\{\frac{f_1}{f_2}+1\} + (\frac{3}{4})^2 \frac{f_1^6 f_3}{f_2^6}\{1-\frac{f_3}{n}\}$ is estimating the random variation stemming from sampling n units from the target population and the second term $\frac{\frac{3n}{4} f_1^3 f_3}{n f_2^3 + \frac{3}{4} f_1^3 f_3}$ is the approximating the random variation due to estimating the number of unobserved cases f_0.

4.2. Confidence interval

Once we have provided an estimator of the variance of the estimator of interest, a confidence interval of the population size N can be constructed using the normal approximation as $\hat{N} \pm 1.96 Se(\hat{N})$, where $Se(\hat{N})$ is the estimated standard error of $\hat{N}$. Alternatively, we can also investigate the confidence interval by using the bootstrap method. The main benefit of using bootstrap method is that it does not require a formula for a variance estimator and might be preferable for small sizes. We focus here on the the percentile bootstrap method. The procedure for constructing 95% confidence intervals using the percentile bootstrap method is as follows:

1) A sample of size $\hat{N}$ is drawn with replacement from the data set which contains both observed individuals (n counts of $1, 2, 3, ..., m$ with associated frequencies $f_1, f_2, ..., f_m$) and estimated unobserved frequency $\hat{f}_0$ of individuals with zero-counts. $\hat{N}$ is determined according to the

estimator under investigation. We do not only bootstrap from the observed sample of size n because the variance of estimating N arises from two sources, the random variation due to drawing n individuals from the target population of size N and the random variation from estimating the parameter of interest from the observed n units, as just mentioned in subsection 4.1 (see, for a review, van der Heijden et al., 2003a; Böhning, 2008).

2) Then, the resampled zero counts of individuals never identified are omitted. Then, using only the new sample of size n^* a new estimate $\hat{N}^*$ is calculated.

3) Step 1) and 2) are repeated B times. This provides $\hat{N}_1^*, \hat{N}_2^*, \hat{N}_3^*, \ldots, \hat{N}_B^*$.

4) The lower and upper bound of the 95% confidence interval are calculated from $P_{2.5}$ and $P_{97.5}$, the 2.5^{th} and 97.5^{th} percentile of the data set obtained in 3), respectively.

5) The standard error of estimate is now found from the sample $\hat{N}_1^*, \hat{N}_2^*, \hat{N}_3^*, \ldots, \hat{N}_B^*$.

5. Real Data Examples and Empirical Applications

There exist a variety of published studies applying the ideas of capture-recapture to the estimation of the total number (adjusted for hidden cases) of patients with infectious diseases such as tuberculosis, HIV/AIDS, legionnaires disease, or malaria (e.g. Gallay et al., 2000; Nardone et al., 2003; van Hest et al., 2008). However, most studies use frequency data from multiple sources with problems of matching and potentially different target areas. Here, we illustrate the use of our proposed estimator in particular data sets with repeated identifications from only one source which is the underlying assumption to apply the model in (1). To achieve a better judgment of the proposed estimator we include the following estimators in the comparison:

- Chao: $\hat{N}_{Chao} = n + f_1^2/(2f_2)$,
  $\widehat{Var(\hat{N}_{Chao})} = \frac{1}{4}\frac{f_1^4}{f_2^3} + \frac{f_1^3}{f_2^2} + \frac{1}{2}\frac{f_1^2}{f_2} - \frac{1}{4}\frac{f_1^4}{nf_2^2} - \frac{1}{2}\frac{f_1^4}{f_2(2nf_2+f_1^2)}$

- MLE: $\hat{N}_{MLE} = \frac{n}{1-\exp(-\hat{\lambda})}$ where $\hat{\lambda}$ is the maximum likelihood estimator under Poisson homogeneity,
  $\widehat{Var(\hat{N}_{MLE})} = \frac{\hat{N}}{(\exp(\frac{\sum jf_j}{\hat{N}}) - \frac{\sum jf_j}{\hat{N}} - 1)}$

- Zel: $\hat{N}_{Zel} = \frac{n}{1-\exp(-2f_2/f_1)}$,
  $\widehat{Var(\hat{N}_{Zel})} = n(\frac{\exp(-\frac{2f_2}{f_1})}{(1-\exp(-\frac{2f_2}{f_1}))^2})\{1 + n(\frac{\exp(-\frac{2f_2}{f_1})}{(1-\exp(-\frac{2f_2}{f_1}))^2})(\frac{2f_2}{f_1})^2(\frac{1}{f_1} + \frac{1}{f_2})\}$

the latter being suggested by Zelterman (1988). We apply these estimators to studies from illicit drug use and biodiversity. We have also computed confidence intervals according to both, the approximate normal and Bootstrap method, outlined in the previous section.

5.1. Drug Use

5.1.1. Scottish Drug Users

Hay and Smit (2003) provide data on drug user contact to a Scottish needle exchange programme in 1997. As the authors say

> Data were collated on individuals who have visited a Scottish needle exchange in the year 1997. We prefer not to explicitly state

the needle exchange from which we obtained these data; however the data were collated during a programme of drug misuse prevalence research in Scotland and was the only one operating in that area at that time. The needle exchange assigns a unique identifier number to each individual accessing the service, thus enabling it to produce statistics on the number of people who had contacted the service over a given period.

The system provided a record of the number of individuals accessing the service over the time period from January to December 1997. The number of visited drug users over this 12 months was 647 and the frequency count of contacting this treatment center is shown in Table 1, with a maximum number of contacts of 105. Here, only the frequency count up to 28 is shown. We are able to compute all estimators of the total estimate of drug users for this data set. The result is shown in Table 2.

Please insert Table 1-2 here

5.1.2. Bangkok Heroin Users

We are interested in estimating the total number of heroin users in Bangkok (Thailand) in 2002. The data was collected by the Office of the Narcotics Control Board (ONCB), Ministry of the Prime Minister, in cooperation with the Drug Abuse Prevention and Treatment Division, Health Department and Medical Service Department, Bangkok Metropolitan Administration. The database recorded all replicated treatment contacts of drug users from the 61 health treatment centers in Bangkok metropolis. The treatment episodes for heroin users are shown in Table 3 (Source: Viwatwongkasem et al., 2008).

We use this frequency table as basis for all estimators considered as provided in Table 4.

Please insert Table 3-4 here

As can be seen in Table 4, the estimated number of heroin users from our method is between the estimators obtained using Chao's and Zelterman's methods. However, similar to the previous application, the proposed estimator shows larger variation.

5.2. Butterfly Data

The Malayan butterfly data go back to Fisher *et al.* (1943) (see Chao and Bunge, 2002) and have been frequently serving as test data for estimators under development. The frequency count of identifying distinct species is shown in Table 5. There are in 620 observed distinct species. Table 6 shows the result of estimated numbers of Malayan butterfly species.

Please insert Table 5-6 here

In this example, the overall impression is that all estimators show similar results in terms of estimated numbers of species for both point and interval estimations. As expected, the MLE provides not only the smallest estimate (underestimation bias), but also gives the least variation. In contrast, our proposed estimator and Zelterman's method yields a larger estimate and variation. In addition, the new estimator provides a similar estimate for this data as the Poisson-Gamma-based estimator suggested by Chao and Bunge (2002) who also used the Malayan butterfly data to illustrate their estimator.

6. Simulation Study

6.1. Simulation Scenarios

A simulation experiment is undertaken to study the performance of the proposed estimator and some competitors such as maximum likelihood, Zelterman's and Chao's estimator. The scope of study covers a variety of situations of heterogeneity in the capture probabilities. Counts have been sampled from the following distributions: Negative Binomial$(k, \theta); k = 2, 3, \theta = 0.4, 0.5, 0.6$ and $k = 4, 5, \theta = 0.6, 0.7, 0.8$, Geometric$(\theta); \theta = 0.3, 0.4, 0.5, 0.6$ and two-component Poisson Mixture; $0.5Poi(\lambda) + 0.5Poi(\mu), \lambda = 0.5, 1, \mu = 2, 3, 4, 5, 6$. We used a population size of 100, 1,000, 10,000 and 100,000, respectively, and each scenario is repeated 10,000 times. To evaluate performance of estimation, we look at relative bias ($RBias = \frac{E(\hat{N})-N}{N}$) and relative mean square error ($RMSE = \frac{E(\hat{N}-N)^2}{N^2}$). Furthermore, both simulated approximation variance of the new proposed estimator and bootstrap percentile method (using a resample size of 1,000) is investigated.

6.2. Simulation Results

We start with an illustration and show the results of estimating population size *from one sample* from a population with a known capture probability and a known population size. The artificial data set of frequency counts of identifications of distinct individuals was generated from $f_j = E(f_j) = Np_j$. We assume that $N = 1,000$ and p_j corresponds to a Negative Binomial $NB(4, \theta = 0.6, 0.7, 0.8)$, a Geometric $Geo(\theta = 0.3, 0.4, 0.5)$ and a two-component Poisson mixture $0.5Poi(0.5) + 0.5Poi(\mu; \mu = 1, 2, 3)$. To illustrate the behaviour of the estimators a sample was generated from each of the above three scenarios, for example, for $N = 1,000$ and $NB(4, 0.7)$, we got $f_0 = 240, f_1 = 288, f_2 = 216, f_3 = 130, f_4 = 68, f_5 = 33, f_6 = 15, f_7 = 6, f_8 = 3$ and $f_9 = 1$. Then, f_0 was omitted and only the remaining zero-truncated frequencies $f_1, f_2, ..., f_m$ with $n = \sum_{j=1}^{m} f_j$ were used to estimate f_0 and N. The results for all estimators are shown in Table 7. It is clear that if heterogeneity becomes more pronounced our proposed estimator noticeably provides the most accurate results. However, these are the results from only one simulated sample. We now undertake a more profound simulation investigation.

Please insert Table 7 here

6.2.1. Heterogeneity in Identification

As a summary result it can be said that under a Negative-Binomial the MLE and Chao's estimator show a clear underestimation of population size, whereas Zelterman, the new estimator and the adjusted form tend to overestimate for a small population size, see Table 8. It also can be seen from Table 8– 9 that the proposed estimator and it's adjusted form perform similarly in cases of large population size. In addition, the adjusted form also significantly reduces the variance if compared with it's original form, in particular for small size. Furthermore, the proposed estimators show a good performance for estimating population size as does Chao's and Zelterman's estimator, in particular they provide smallest *RBias* and *RMSE* for the larger sample size. In summary, it is reasonable to state that the proposed estimators (in particular the adjusted versions) are suitable under the Negative Binomial distributional model.

Please insert Table 8-9 here

For the case that the identification probabilities arise from the Geometric distribution, the new estimator generally shows a good performance in terms of accuracy as it gives on average the smallest *RBias* in almost all cases, see Table 10. According to *RMSE*, although the new estimator seems to be of lack of precision for the small population size, it shows excellent performance against the other methods for larger size, see Table 11.

Please insert Table 10-11 here

Similar to the results under a Negative Binomial distribution, Zelterman and the new estimator seem to provide overestimation of population size, in contrast to the MLE and Chao's estimator which always show underestimation for all two-component Poisson Mixture scenarios, see Table 12. If large population sizes are considered under a discrete Poisson mixture, our proposed estimators not only shows high performance in terms of accuracy, but it also performs similar to the other methods in terms of precision. However, the new proposed estimator is less satisfactory for smaller size as well as it shows high variance, see Table 12-13.

Please insert Table 12-13 here

6.2.2. Variance Approximation

This section presents the results on variance approximation of the new estimator. We compared the variance approximation of the new estimator in equation (13) with estimating variance using the bootstrap and simulation methods. To define the investigated statistics in Table 14-16, $Se(\hat{N})$ denotes standard error of the new estimator computing based 10,000 repeated simulation samples, whereas $Mean(\widehat{Se(\hat{N})})$ and $Mean(\widehat{Se(\hat{N})_{Bt}})$ represent an average estimated standard error from equation (13) and the bootstrap percentile method, respectively. It is seen from Table 14-16 that, $Se(\hat{N})$, $Mean(\widehat{Se(\hat{N})})$ and $Mean(\widehat{Se(\hat{N})_{Bt}})$ are quite similar in their values. $Mean(\widehat{Se(\hat{N})})$ is slightly smaller than $Se(\hat{N})$ and $Mean(\widehat{Se(\hat{N})_{Bt}})$. As a result, it is reasonable to state that the variance approximation of the new estimator in equation (13) can be utilized to represent the true variance.

Please insert Table 14-16 here

7. Conclusions and Discussion

A diversity of estimators in the capture-recapture field exists such as the estimators of Chao (1987) and Zelterman (1988), being widely applied in many areas of interest, especially in public health and social sciences. Here, we have introduced a new method of estimating the population size under a specific form of heterogeneity for the identification probability of distinct individuals. We have also been able to see how accurate and precise the method is performing when it is compared to other frequently used estimators. Overall, the proposed estimator is more accurate as well as providing small bias in the homogeneous Poisson case which asymptotically disappears. It is also found that the new estimator compares well with Chao's estimator since it's expected value is equal or greater than the one of Chao's estimator. Hence, it improves Chao's estimator which is known to provide a lower bound. In a simulation study, the new estimator tended to overestimate, whereas all the other methods under consideration provided the known underestimation phenomenon in almost all scenarios of heterogeneous identification probabilities. However, although the proposed estimator showed superior performance in terms of accuracy, it evidently gave also the largest variation. Hence, the new method has lack of precision; nonetheless, the variation of the new estimators considerably decreased for large population size (1,000 and more) which is typically the case for situations of interest in surveillance and public health. In addition, the adjusted forms of the new estimator can be utilized for sample sizes below $1,000$ which significantly reduces the variance. We also provided a formula of variance approximation of the new estimator. This variance formula is not only useful to determine the efficiency of estimating, but it can be also used to construct confidence intervals. In short, the new estimator can be an alternative form of population size estimation especially for large populations and heterogeneous capturing probabilities.

Acknowledgements

Both authors would like to thank the reviewers for the very helpful comments which lead to considerable improvement of the paper as well as to interesting and promising modifications of the proposed estiamtor. Both authors are grateful to the Ministry of Science & Technology, the Royal Thai Government for providing PhD-funding for the first author. The authors thank Dr. Chukiat Viwatwongkasem and Busaba Suppawattanabodee for their cooperation in providing public access to the surveillance data on drug users in Bangkok Metropolis.

Bishop YYM.,Firnberg SE.,Holland PW., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge.

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

Böhning D, Suppawattanabodee B, Kusolvisitkul W, Viwatwongkasem C., 2004. Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. European Journal of Epidemiology 19, 1075-1083.

Carothers AD., 1973. Capture-recapture methods applied to a population with known parameters. The Journal of Animal Ecology 42, 125-146.

Chang YF, LaPorte RE, Aaron DJ, Songer TJ.,1999. The importance of source selection and pilot study in the capture-recapture application. Journal of Clinical Epidemiology 52, 927-928.

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

Chao A, Bunge J., 2002. Estimating the number of species in a stochastics abudance model. Biometrics 58, 531-539.

Corrao G, Bagnardi V, Vittadini G, Favilli S., 2000. Capture-recapture methods to size alcohol related problems in a population. Journal of Epidemiology and Community Health 54, 603-610.

Darroch JN.,1958. The multiple-recapture census: I. Estimation of a closed population. Biometrika 45, 343-359.

Eberhardt L., 1969. Population estimation from recapture frequencies. The Journal of Wildlife Management 33, 28-39.

Edwards WR, Eberhardt L.,1967. Estimating cottontail abundance from live-trapping data. The Journal of Wildlife Management 31, 87-96.

Gallay A, Vaillant V, Bouvet P, Grimont P, Desenclos JC., 2000. How many foodborne outbreaks of salmonella infection occurred in France in 1955? : Application of the capture-recapture method to three surveillance systems. American Journal of Epidemiology 152, 171-177.

Hay G., 1997. The selection from multiple data source in epidemiological capture-recapture studies. The Statistician 46, 515-520.

Hay G, Gannon M, MacDougall J, Eastwood C, Williams K, Millar T., 2009. Capture-recapture and anchored prevalence estimation of injecting drug users in England: national and regional estimates. Statitical Methods in Medical Research 18, 323-339.

Hay G, Smit F., 2003. Estimating the number of drug injectors from needle exchange data. Addiction Research and Theory 11, 235-243.

Hook EB, Regal R., 1995. Capture-recapture methods in epidemiology: methods and limitations. Epidemiologic Reviews 17, 243-264.

Hope, VD, Hickman M, Tilling K., 2005. Capturing crack cocaine use: Estimating the prevalence of crack cocaine use in London using capture-recapture with covariates. Addiction 100, 1701-1708.

Horvitz DG, Thompson DJ., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.

McDonald JF, Palanacki D., 1989. Interval estimation of the size of a small population from a mark-recapture experiment. Biometrics 45, 1223-1231.

Nardone A, Decludt B, Jarraud S, Etienne J, Hubert B, Infuso A, Gallay A, Desenclos JC., 2003. Repeat capture-recapture studies as part of the evaluation of the surveillance of legionnaires's disease in France. Epidemiology and Infection 131, 647-654.

North PM., 1981. Models for a simple capture-recapture procedure to estimate population size. Biometrics 37, 661-672.

Pollock KH., 2000. Capture-recapture models. Journal of the American Statistical Association 95, 293-296.

Seber GAF., 2002. The Estimation of Animal Abundance and Related Parameter (2nd edn). Blackburn Press, Caldwell, NJ.

Smit F, Reinking D, Reijerse M., 2002. Estimating the number of people eligible for health service use. Evaluation and Program Planning 25, 101-105.

van der Heijden, PGM, Bustami R, Cruyff MJLF, Engbersen G, van Houwelingen HC., 2003. Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modelling 3, 305-322.

van der Heijden P GM, Cruyff MJLF, van Houwelingen HC., 2003. Estimating the size of a criminal population from police records using the truncated Poisson regression model. Statistica Neerlandica 57, 289-304.

van Hest NAH, Grant AD, Smit F, Story A, Richardus JH., 2008. Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data. Epidemiology and Infection 136, 14-22.

Viwatwongkasem C, Kuhnert R, Satitvipawee P., 2008. A comparison of population size estimators under the truncated count model with and without allowance for contaminations. Biometrical Journal 50, 1006-1021.

Woodward M., 1999. Epidemiology: Study Design and Data Analysis. Chapman & Hall, CRC: London, Boca Raton, FL.

Zelterman D., 1988. Robust Estimation in truncated discrete distributions with application to capture-recapture experiments. Journal of Statistical Planning and Inference 18, 225-237.

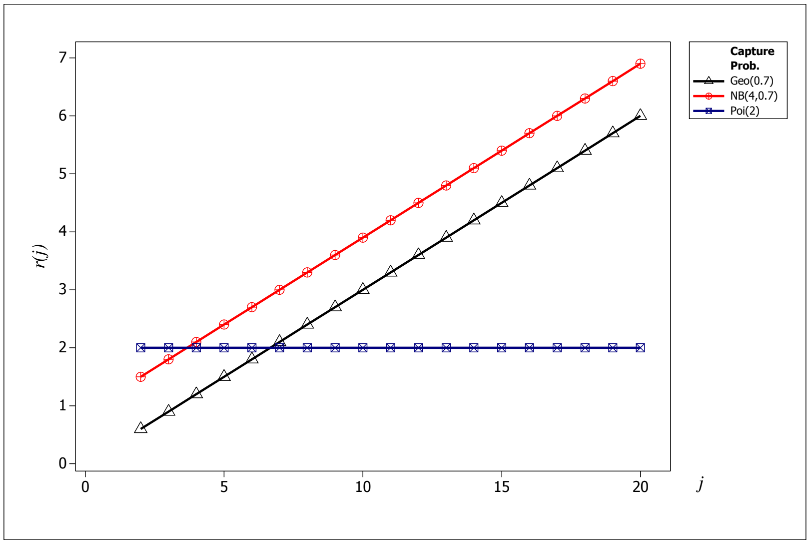


Figure 1: r_j as a function of j; $p_j \sim NB(4, 0.7)$, $Geo(0.7)$, and $Poi(2)$

Table 1: The frequency of individual contacts at Scottish needle exchange, in 1997; $n = 647$

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
f_j	175	85	50	47	37	38	32	16	17	17	15	11	9	12
j	15	16	17	18	19	20	21	22	23	24	25	26	27	28+
f_j	13	7	6	2	3	5	8	2	6	1	2	3	3	25

Table 2: Estimated total number of Scottish drug injectors in 1997

Method	$\hat{N}$	$\widehat{Se(\hat{N})}$	95% CI (Approximate Normal)	$Se(\hat{N})_{BT}$	95% CI (Bootstrap Percentile)
MLE	648	0.67	645 - 649	1.00	646 - 649
Chao	828	34.85	759 - 896	36.91	763 - 907
New	975	137.99	704 - 1,245	150.94	788 - 1,379
NewAdj	975	-	-	103.76	779 - 1,169
NewMo	948	-	-	145.78	774 - 1,326
Zel	1,042	85.25	874 - 1,209	87.44	909 - 1,246

Table 3: The frequency count of times that heroin users contacted health treatment centers in Bangkok, Thailand in 2002; $n = 9,302$

j	1	2	3	4	5	6	7	8	9	10	11
f_j	2,176	1,600	1,278	976	748	570	455	368	281	254	188
j	12	13	14	15	16	17	18	19	20	21	
f_j	138	99	67	44	34	17	3	3	2	1	

Table 4: Estimated total number of heroin users in Bangkok, Thailand 2000

Method	$\hat{N}$	$\widehat{Se(\hat{N})}$	95% CI (Approximate Normal)	$Se(\hat{N})_{BT}$	95% CI (Bootstrap Percentile)
MLE	9,454	12.84	9,430 - 9,479	13.40	9,518 - 9,573
Chao	10,782	80.21	10,625 - 10,940	85.71	10,625 - 10,945
New	11,714	250.16	11,224 - 12,205	265.07	11,256 - 12,279
NewAdj	11,714	-	-	249.39	11,257 - 12,241
NewMo	11,701	-	-	255.71	11,250 - 12,216
Zel	12,078	184.54	11,717 - 12,440	188.45	11,728 - 12,476

Table 5: Malayan butterfly data (Fisher *et al.*, 1943)

j	1	2	3	4	5	6	7	8	9	10	11	12	13
f_j	118	74	44	24	29	22	20	19	20	15	12	14	6
j	14	15	16	17	18	19	20	21	22	23	24	24+	n
f_j	12	6	9	9	6	10	10	11	5	3	3	119	620

Table 6: Estimated total number of Malayan butterfly species

Method	$\hat{N}$	$\widehat{Se(\hat{N})}$	95% CI (Approximate Normal)	$Se(\hat{N})_{BT}$	95% CI (Bootstrap Percentile)
MLE	624	1.80	604 - 645	2.23	616 - 624
Chao	715	22.07	671 - 756	24.78	672 - 766
New	754	63.49	630 - 879	85.33	659 - 1,017
NewAdj	754	-	-	66.17	670 - 919
NewMo	741	-	-	62.80	664 - 898
Zel	868	67.04	736 - 999	64.61	711 - 973

Table 7: Estimated population size based upon one sample and four different estimators, true $N = 1,000$

Capture Probability(p_j)	MLE	Chao	New	NewAdj	NewMo	Zel
$NB(4, 0.6)$	921	973	994	994	989	1,005
$NB(4, 0.7)$	892	953	993	993	984	980
$NB(4, 0.8)$	860	919	987	987	970	935
$Geo(0.3)$	732	850	926	926	914	930
$Geo(0.4)$	675	800	899	899	883	859
$Geo(0.5)$	635	750	878	878	857	791
$0.5Poi(0.5) + 0.5Poi(1.0)$	923	948	993	993	967	953
$0.5Poi(0.5) + 0.5Poi(2.0)$	796	869	963	963	948	900
$0.5Poi(0.5) + 0.5Poi(3.0)$	743	842	974	974	959	914
$0.5Poi(0.5) + 0.5Poi(4.0)$	719	847	1066	1006	1041	992

Table 8: *RBias* of population size estimators for counts drawn from $NB(k, \theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 100$			
$NB(2, 0.4)$	-0.1261	-0.0349	0.1023	0.0324	-0.0113	0.0459
$NB(2, 0.5)$	-0.1685	-0.0591	0.1328	0.0433	-0.0167	0.0053
$NB(2, 0.6)$	-0.2058	-0.0892	0.1886	0.0560	-0.0344	-0.0417
$NB(3, 0.4)$	-0.0510	-0.0014	0.0817	0.0317	0.0031	0.0758
$NB(3, 0.5)$	-0.0851	-0.0154	0.0823	0.0389	-0.0027	0.0468
$NB(3, 0.6)$	-0.1203	-0.0335	0.1073	0.0559	-0.0084	0.0161
$NB(4, 0.6)$	-0.0716	-0.0092	0.0814	0.0463	0.0020	0.0396
$NB(4, 0.7)$	-0.1013	-0.0270	0.1157	0.0696	-0.0061	0.0094
$NB(4, 0.8)$	-0.1220	-0.0445	0.2471	0.1232	-0.0233	-0.0219
$NB(5, 0.6)$	-0.0426	0.0014	0.0612	0.0372	0.0034	0.0479
$NB(5, 0.7)$	-0.0700	-0.0108	0.0837	0.0569	-0.0009	0.0256
$NB(5, 0.8)$	-0.0917	-0.0219	0.1871	0.1132	0.0018	0.0034
			$N = 1,000$			
$NB(2, 0.4)$	-0.1321	-0.0519	-0.0106	-0.0106	-0.0180	0.0093
$NB(2, 0.5)$	-0.1776	-0.0815	-0.0167	-0.0167	-0.0268	-0.0310
$NB(2, 0.6)$	-0.2164	-0.1173	-0.0249	-0.0248	-0.0398	-0.0806
$NB(3, 0.4)$	-0.0554	-0.0147	0.0009	0.0013	-0.0034	0.0335
$NB(3, 0.5)$	-0.0911	-0.0301	-0.0026	-0.0020	-0.0084	0.0144
$NB(3, 0.6)$	-0.1269	-0.0515	-0.0032	-0.0024	-0.0117	-0.0136
$NB(4, 0.6)$	-0.0779	-0.0241	0.0014	0.0024	-0.0042	0.0100
$NB(4, 0.7)$	-0.1093	-0.0455	-0.0001	0.0019	-0.0091	-0.0184
$NB(4, 0.8)$	-0.1407	-0.0790	0.0014	0.0066	-0.0172	-0.0625
$NB(5, 0.6)$	-0.0486	-0.0116	0.0023	0.0033	-0.0017	0.0173
$NB(5, 0.7)$	-0.0770	-0.0264	0.0020	0.0038	-0.0045	-0.0006
$NB(5, 0.8)$	-0.1068	-0.0522	0.0037	0.0083	-0.0093	-0.0348

Table 8(con.): *RBias* of population size estimators for counts drawn from $NB(k, \theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 10,000$			
$NB(2, 0.4)$	-0.1327	-0.0531	-0.0169	-0.0169	-0.0177	0.0068
$NB(2, 0.5)$	-0.1777	-0.0832	-0.0269	-0.0269	-0.0278	-0.0343
$NB(2, 0.6)$	-0.2176	-0.1197	-0.0386	-0.0386	-0.0401	-0.0838
$NB(3, 0.4)$	-0.0559	-0.0159	-0.0035	-0.0035	-0.0039	0.0299
$NB(3, 0.5)$	-0.0915	-0.0310	-0.0070	-0.0070	-0.0076	0.0124
$NB(3, 0.6)$	-0.1279	-0.0537	-0.0125	-0.0125	-0.0133	-0.0173
$NB(4, 0.6)$	-0.0785	-0.0258	-0.0047	-0.0047	-0.0052	0.0069
$NB(4, 0.7)$	-0.1105	-0.0477	-0.0087	-0.0087	-0.0096	-0.0214
$NB(4, 0.8)$	-0.1420	-0.0815	-0.0147	-0.0147	-0.0165	-0.0656
$NB(5, 0.6)$	-0.0491	-0.0128	-0.0018	-0.0017	-0.0021	0.0145
$NB(5, 0.7)$	-0.0777	-0.0278	-0.0040	-0.0040	-0.0046	-0.0030
$NB(5, 0.8)$	-0.1080	-0.0544	-0.0081	-0.0081	-0.0094	-0.0377
			$N = 100,000$			
$NB(2, 0.4)$	-0.1327	-0.0533	-0.0177	-0.0177	-0.0178	0.0064
$NB(2, 0.5)$	-0.1778	-0.0833	-0.0277	-0.0277	-0.0278	-0.0346
$NB(2, 0.6)$	-0.2177	-0.1200	-0.0398	-0.0398	-0.0399	-0.0841
$NB(3, 0.4)$	-0.0560	-0.0160	-0.0040	-0.0040	-0.0040	0.0294
$NB(3, 0.5)$	-0.0915	-0.0312	-0.0077	-0.0077	-0.0077	0.0120
$NB(3, 0.6)$	-0.1280	-0.0540	-0.0134	-0.0134	-0.0135	-0.0176
$NB(4, 0.6)$	-0.0786	-0.0259	-0.0051	-0.0051	-0.0052	0.0067
$NB(4, 0.7)$	-0.1107	-0.0480	-0.0095	-0.0095	-0.0096	-0.0218
$NB(4, 0.8)$	-0.1423	-0.0818	-0.0162	-0.0162	-0.0163	-0.0659
$NB(5, 0.6)$	-0.0492	-0.0129	-0.0021	-0.0021	-0.0021	0.0143
$NB(5, 0.7)$	-0.0777	-0.0280	-0.0046	-0.0046	-0.0046	-0.0033
$NB(5, 0.8)$	-0.1080	-0.0546	-0.0090	-0.0090	-0.0091	-0.0379

Table 9: *RMSE* of population size estimators for counts drawn from $NB(k, \theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 100$			
$NB(2, 0.4)$	0.017436	0.007103	0.322176	0.023931	0.055283	0.026119
$NB(2, 0.5)$	0.031113	0.014611	0.444714	0.049028	0.103335	0.027807
$NB(2, 0.6)$	0.047743	0.027054	0.738501	0.088966	0.169868	0.036023
$NB(3, 0.4)$	0.003222	0.002308	0.220680	0.009114	0.017575	0.026815
$NB(3, 0.5)$	0.008521	0.004579	0.152294	0.018016	0.031767	0.019818
$NB(3, 0.6)$	0.017133	0.010355	0.187029	0.041730	0.058681	0.021858
$NB(4, 0.6)$	0.006593	0.004972	0.131836	0.020873	0.032070	0.017851
$NB(4, 0.7)$	0.013899	0.011858	0.239724	0.052637	0.076174	0.022364
$NB(4, 0.8)$	0.026576	0.032748	1.084269	0.157403	0.242154	0.044130
$NB(5, 0.6)$	0.002662	0.002600	0.052894	0.010521	0.012551	0.014369
$NB(5, 0.7)$	0.007154	0.006949	0.108087	0.030457	0.037630	0.017049
$NB(5, 0.8)$	0.015763	0.021379	0.540063	0.107888	0.156400	0.032862
			$N = 1,000$			
$NB(2, 0.4)$	0.017599	0.003176	0.002812	0.002621	0.002759	0.001757
$NB(2, 0.5)$	0.031791	0.007479	0.005924	0.005529	0.005857	0.002936
$NB(2, 0.6)$	0.047332	0.015324	0.012544	0.011473	0.012379	0.009179
$NB(3, 0.4)$	0.003133	0.000381	0.000728	0.000671	0.000650	0.002206
$NB(3, 0.5)$	0.008418	0.001267	0.001744	0.001646	0.001649	0.001445
$NB(3, 0.6)$	0.016367	0.003402	0.004228	0.003998	0.004013	0.001853
$NB(4, 0.6)$	0.006209	0.000977	0.001817	0.001702	0.001683	0.001245
$NB(4, 0.7)$	0.012299	0.003003	0.005032	0.004666	0.004726	0.002101
$NB(4, 0.8)$	0.020837	0.008663	0.017714	0.015629	0.016264	0.007297
$NB(5, 0.6)$	0.002441	0.000346	0.000809	0.000753	0.000733	0.001098
$NB(5, 0.7)$	0.006148	0.001275	0.002629	0.002418	0.002449	0.001296
$NB(5, 0.8)$	0.012102	0.004395	0.010177	0.009095	0.009400	0.003764

Table9(con.): $RMSE$ of population size estimators for counts drawn from $NB(k,\theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 10,000$			
$NB(2,0.4)$	0.017620	0.002869	0.000518	0.000518	0.000541	0.000205
$NB(2,0.5)$	0.031610	0.006997	0.001197	0.001197	0.001246	0.001365
$NB(2,0.6)$	0.047396	0.014489	0.002517	0.002517	0.002622	0.007285
$NB(3,0.4)$	0.003133	0.000268	0.000073	0.000073	0.000076	0.000994
$NB(3,0.5)$	0.008385	0.000999	0.000207	0.000207	0.000213	0.000278
$NB(3,0.6)$	0.016390	0.002965	0.000543	0.000543	0.000561	0.000467
$NB(4,0.6)$	0.006177	0.000703	0.000181	0.000181	0.000185	0.000160
$NB(4,0.7)$	0.012236	0.002363	0.000524	0.000524	0.000537	0.000631
$NB(4,0.8)$	0.020280	0.006889	0.001799	0.001795	0.001840	0.004636
$NB(5,0.6)$	0.002419	0.000185	0.000075	0.000075	0.000076	0.000289
$NB(5,0.7)$	0.006060	0.000830	0.000250	0.000249	0.000253	0.000133
$NB(5,0.8)$	0.011729	0.003126	0.000952	0.000949	0.000967	0.001665
			$N = 100,000$			
$NB(2,0.4)$	0.017610	0.002846	0.000336	0.000336	0.000338	0.000057
$NB(2,0.5)$	0.031612	0.006950	0.000813	0.000813	0.000818	0.001213
$NB(2,0.6)$	0.047401	0.014404	0.001682	0.001682	0.001694	0.007096
$NB(3,0.4)$	0.003133	0.000268	0.000073	0.000073	0.000076	0.000994
$NB(3,0.5)$	0.008385	0.000999	0.000207	0.000207	0.000213	0.000278
$NB(3,0.6)$	0.016390	0.002965	0.000543	0.000543	0.000561	0.000467
$NB(4,0.6)$	0.006176	0.000674	0.000042	0.000042	0.000042	0.000055
$NB(4,0.7)$	0.012264	0.002312	0.000135	0.000135	0.000137	0.000493
$NB(4,0.8)$	0.020251	0.006720	0.000414	0.000414	0.000420	0.004376
$NB(5,0.6)$	0.002424	0.000170	0.000012	0.000012	0.000012	0.000212
$NB(5,0.7)$	0.006047	0.000788	0.000045	0.000045	0.000045	0.000023
$NB(5,0.8)$	0.011672	0.002996	0.000169	0.000169	0.000171	0.001459

Table 10: *RBias* of population size estimators for counts drawn from $Geo(\theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 100$			
$Geo(0.3)$	-0.2622	-0.1244	0.1640	-0.0172	-0.0677	-0.0231
$Geo(0.4)$	-0.3172	-0.1704	0.1695	-0.0331	-0.1032	-0.0970
$Geo(0.5)$	-0.3581	-0.2112	0.3865	-0.0365	-0.1242	-0.1575
$Geo(0.6)$	-0.3811	-0.2395	0.7796	-0.0182	-0.1585	-0.2017
			$N = 1,000$			
$Geo(0.3)$	-0.2692	-0.1475	-0.0610	-0.0642	-0.0739	-0.0663
$Geo(0.4)$	-0.3271	-0.1980	-0.0844	-0.0889	-0.1013	-0.1382
$Geo(0.5)$	-0.3710	-0.2467	-0.1021	-0.1096	-0.1256	-0.2047
$Geo(0.6)$	-0.4060	-0.2947	-0.1117	-0.1272	-0.1479	-0.2672
			$N = 10,000$			
$Geo(0.3)$	-0.2701	-0.1497	-0.0733	-0.0733	-0.0745	-0.0703
$Geo(0.4)$	-0.3277	-0.1996	-0.0984	-0.0984	-0.1000	-0.1409
$Geo(0.5)$	-0.3724	-0.2497	-0.1229	-0.1229	-0.1251	-0.2086
$Geo(0.6)$	-0.4078	-0.2993	-0.1457	-0.1457	-0.1491	-0.2728
			$N = 100,000$			
$Geo(0.3)$	-0.2702	-0.1500	-0.0749	-0.0749	-0.0750	-0.0709
$Geo(0.4)$	-0.3278	-0.2000	-0.1000	-0.1000	-0.1001	-0.1414
$Geo(0.5)$	-0.3725	-0.2500	-0.1249	-0.1249	-0.1251	-0.2090
$Geo(0.6)$	-0.4081	-0.3000	-0.1496	-0.1496	-0.1500	-0.2736

Table 11: *RMSE* of population size estimators for counts drawn from $Geo(\theta)$

p_j	MLE	Chao	New	NewAdj	NewMo	Zel
			$N = 100$			
$Geo(0.3)$	0.0712	0.0270	6.3955	0.0469	0.2105	0.0385
$Geo(0.4)$	0.1044	0.0460	1.8536	0.0736	0.2393	0.0468
$Geo(0.5)$	0.1347	0.0744	59.5715	0.1324	0.5010	0.0761
$Geo(0.6)$	0.1577	0.1091	81.3850	0.2317	0.7598	0.1149
			$N = 1,000$			
$Geo(0.3)$	0.0727	0.0226	0.0110	0.0100	0.0119	0.0070
$Geo(0.4)$	0.1074	0.0405	0.0197	0.0178	0.0214	0.0219
$Geo(0.5)$	0.1383	0.0628	0.0322	0.0279	0.0347	0.0451
$Geo(0.6)$	0.1659	0.0900	0.0557	0.0432	0.0580	0.0757
			$N = 10,000$			
$Geo(0.3)$	0.0730	0.0225	0.0060	0.0060	0.0062	0.0052
$Geo(0.4)$	0.1074	0.0400	0.0107	0.0107	0.0110	0.0201
$Geo(0.5)$	0.1388	0.0625	0.0169	0.0169	0.0175	0.0438
$Geo(0.6)$	0.1664	0.0899	0.0246	0.0246	0.0255	0.0748
			$N = 100,000$			
$Geo(0.3)$	0.0730	0.0225	0.0057	0.0057	0.0057	0.0050
$Geo(0.4)$	0.1075	0.0400	0.0101	0.0101	0.0101	0.0200
$Geo(0.5)$	0.1388	0.0625	0.0158	0.0158	0.0158	0.0437
$Geo(0.6)$	0.1665	0.0900	0.0227	0.0227	0.0228	0.0749

Table 12: *RBias* of population size estimators for $p_j \sim 0.5Poi(\lambda) + 0.5Poi(\mu)$

λ	μ	MLE	Chao	New	NewAdj	NewMo	Zel
				$N = 100$			
0.5	2.0	-0.1947	-0.1048	0.1884	0.0423	-0.0395	-0.0644
0.5	3.0	-0.2501	-0.1310	0.2557	0.0081	-0.0190	-0.0394
0.5	4.0	-0.2737	-0.1187	0.6396	0.0393	0.0654	0.0664
0.5	5.0	-0.2858	-0.0829	2.1944	0.1110	0.2091	0.2540
1.0	2.0	-0.0517	-0.0035	0.1331	0.0963	0.0077	0.0187
1.0	3.0	-0.1084	-0.0230	0.1473	0.0739	0.0174	0.0331
1.0	4.0	-0.1407	-0.0226	0.2229	0.0850	0.0405	0.0852
1.0	5.0	-0.1578	-0.0096	0.3462	0.1102	0.0693	0.1643
				$N = 1,000$			
0.5	2.0	-0.2058	-0.1309	-0.0264	-0.0276	-0.0417	-0.0994
0.5	3.0	-0.2573	-0.1547	-0.0047	-0.0214	-0.0217	-0.0803
0.5	4.0	-0.2803	-0.1496	0.0983	0.0062	0.0700	-0.0003
0.5	5.0	-0.2912	-0.1247	0.2858	0.0548	0.2319	0.1419
1.0	2.0	-0.0610	-0.0250	0.0094	0.0143	0.0002	-0.0111
1.0	3.0	-0.1165	-0.0430	0.0186	0.0188	0.0094	0.0001
1.0	4.0	-0.1472	-0.0439	0.0483	0.0447	0.0364	0.0433
1.0	5.0	-0.1636	-0.0345	0.0831	0.0721	0.0672	0.1081

Table12(con.): *RBias* of population size estimators for

$p_j \sim 0.5Poi(\lambda) + 0.5Poi(\mu)$

λ	μ	MLE	Chao	New	NewAdj	NewMo	Zel
				$N = 10,000$			
0.5	2.0	-0.2068	-0.1329	-0.0400	-0.0400	-0.0415	-0.1021
0.5	3.0	-0.2580	-0.1572	-0.0213	-0.0217	-0.0229	-0.0846
0.5	4.0	-0.2810	-0.1529	0.0687	0.0066	0.0661	-0.0075
0.5	5.0	-0.2921	-0.1289	0.2346	0.0488	0.2296	0.1318
1.0	2.0	-0.0620	-0.0272	0.0004	0.0005	-0.0005	-0.0140
1.0	3.0	-0.1171	-0.0448	0.0099	0.0099	0.0090	-0.0028
1.0	4.0	-0.1477	-0.0457	0.0370	0.0370	0.0359	0.0396
1.0	5.0	-0.1642	-0.0368	0.0680	0.0680	0.0665	0.1030
				$N = 100,000$			
0.5	2.0	-0.2069	-0.1332	-0.0416	-0.0416	-0.0417	-0.1025
0.5	3.0	-0.2579	-0.1573	-0.0227	-0.0227	-0.0228	-0.0848
0.5	4.0	-0.2810	-0.1530	0.0667	0.0064	0.0664	-0.0078
0.5	5.0	-0.2921	-0.1293	0.2291	0.0480	0.2286	0.1306
1.0	2.0	-0.0621	-0.0273	-0.0002	-0.0002	-0.0003	-0.0142
1.0	3.0	-0.1172	-0.0449	0.0094	0.0094	0.0093	-0.0029
1.0	4.0	-0.1478	-0.0459	0.0359	0.0359	0.0358	0.0392
1.0	5.0	-0.1642	-0.0370	0.0668	0.0668	0.0666	0.1027

Table 13: *RMSE* of population size estimators for $p_j \sim 0.5Poi(\lambda) + 0.5Poi(\mu)$

λ	μ	MLE	Chao	New	NewAdj	NewMo	Zel
				$N = 100$			
0.5	2.0	0.0436	0.0294	1.2641	0.0848	0.2344	0.0364
0.5	3.0	0.0656	0.0310	1.4570	0.0573	0.2156	0.0397
0.5	4.0	0.0774	0.0302	9.0444	0.0656	0.4079	0.0806
0.5	5.0	0.0839	0.0374	375.7771	0.1352	2.1824	0.3013
1.0	2.0	0.0077	0.0136	0.2933	0.0367	0.1752	0.0957
1.0	3.0	0.0143	0.0100	0.3123	0.0233	0.1694	0.0829
1.0	4.0	0.0217	0.0095	0.6376	0.0222	0.2736	0.1088
1.0	5.0	0.0266	0.0115	1.8505	0.0297	0.5621	0.1796
				$N = 1,000$			
0.5	2.0	0.0429	0.0186	0.0135	0.0121	0.0134	0.0124
0.5	3.0	0.0665	0.0250	0.0123	0.0073	0.0114	0.0090
0.5	4.0	0.0788	0.0235	0.0362	0.0046	0.0275	0.0045
0.5	5.0	0.0850	0.0171	0.1565	0.0086	0.1134	0.0299
1.0	2.0	0.0042	0.0017	0.0057	0.0045	0.0055	0.0052
1.0	3.0	0.0138	0.0026	0.0054	0.0032	0.0051	0.0047
1.0	4.0	0.0219	0.0026	0.0088	0.0026	0.0082	0.0072
1.0	5.0	0.0269	0.0020	0.0165	0.0034	0.0150	0.0129

Table13(con.): *RMSE* of population size estimators for

$p_j \sim 0.5Poi(\lambda) + 0.5Poi(\mu)$

λ	μ	MLE	Chao	New	NewAdj	NewMo	Zel
				$N = 10,000$			
0.5	2.0	0.0428	0.0178	0.0027	0.0027	0.0028	0.0107
0.5	3.0	0.0666	0.0248	0.0015	0.0015	0.0016	0.0074
0.5	4.0	0.0790	0.0235	0.0068	0.0004	0.0064	0.0005
0.5	5.0	0.0853	0.0168	0.0604	0.0029	0.0579	0.0183
1.0	2.0	0.0039	0.0009	0.0005	0.0005	0.0005	0.0005
1.0	3.0	0.0137	0.0021	0.0005	0.0005	0.0005	0.0005
1.0	4.0	0.0218	0.0022	0.0019	0.0009	0.0019	0.0018
1.0	5.0	0.0270	0.0014	0.0055	0.0016	0.0054	0.0052
				$N = 100,000$			
0.5	2.0	0.0428	0.0177	0.0018	0.0018	0.0019	0.0105
0.5	3.0	0.0665	0.0248	0.0006	0.0006	0.0006	0.0072
0.5	4.0	0.0790	0.0234	0.0047	0.0001	0.0046	0.0001
0.5	5.0	0.0853	0.0167	0.0530	0.0024	0.0528	0.0172
1.0	2.0	0.0039	0.0008	0.0001	0.0001	0.0001	0.0001
1.0	3.0	0.0137	0.0020	0.0001	0.0001	0.0001	0.0001
1.0	4.0	0.0218	0.0021	0.0013	0.0008	0.0013	0.0013
1.0	5.0	0.0270	0.0014	0.0045	0.0015	0.0045	0.0045

Table 14: Estimated standard error of estimating population size of the proposed estimator for $p_j \sim NB(k,\theta)$

p_j	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$
	$N = 1,000$			$N = 10,000$		
$NB(2,0.4)$	50.30	49.06	53.72	150.52	149.70	151.58
$NB(2,0.5)$	72.56	71.36	77.69	217.81	216.52	220.56
$NB(2,0.6)$	105.72	103.88	113.78	316.97	313.04	317.81
$NB(3,0.4)$	25.84	26.03	29.07	78.98	77.41	78.28
$NB(3,0.5)$	42.62	40.91	44.51	123.43	122.79	124.85
$NB(3,0.6)$	64.61	63.82	69.09	195.33	190.23	193.38
$NB(4,0.6)$	42.44	40.66	43.96	126.03	124.10	126.32
$NB(4,0.7)$	69.43	69.50	74.99	210.53	210.49	215.11
$NB(4,0.8)$	142.34	132.55	146.22	396.69	390.72	398.01
$NB(5,0.6)$	30.30	27.84	30.29	85.05	84.33	85.33
$NB(5,0.7)$	52.03	50.93	54.79	160.10	152.19	154.35
$NB(5,0.8)$	100.13	96.98	105.49	310.10	293.28	299.67

Table 15: Estimated standard error of estimating population size of the proposed estimator for $p_j \sim Geo(\theta)$

p_j	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$
	$N = 1,000$			$N = 10,000$		
$Geo(0.3)$	85.50	82.66	91.82	250.67	243.54	248.26
$Geo(0.4)$	110.63	103.79	115.74	310.10	319.81	324.46
$Geo(0.5)$	154.84	141.98	160.69	420.39	422.33	431.44
$Geo(0.6)$	201.55	192.32	225.04	576.47	567.02	576.62

Table 16: Estimated standard error of estimating population size of the proposed estimator for $p_j \sim 0.5Poi(\lambda) + 0.5Poi(\mu)$

λ	μ	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$	$Se(\hat{N})$	Mean $\widehat{Se(\hat{N})}$	Mean $\widehat{Se(\hat{N})}_{Bt}$
		$N = 1,000$			$N = 10,000$		
0.5	2	110.08	105.71	115.84	332.84	320.88	326.83
0.5	3	119.28	107.60	119.29	331.09	320.18	327.13
0.5	4	162.09	150.91	172.62	459.62	449.79	456.78
0.5	5	270.54	254.87	307.85	740.94	726.54	738.81
1.0	2	74.82	71.95	77.48	227.56	222.43	227.57
1.0	3	70.91	68.29	73.80	200.07	205.28	208.55
1.0	4	77.88	76.74	84.36	233.39	236.98	239.65
1.0	5	97.34	94.79	106.12	291.49	284.27	289.72