A Comparison of Population Size Estimators under the Truncated Count Model With and Without Allowance for Contaminations

Chukiat Viwatwongkasem¹,  Ronny Kuhnert², and Pratana Satitvipawee³

¹ Department of Biostatistics, Faculty of Public Health, Mahidol University, 420/1 Rajvithi Road, Phayathai District, Bangkok 10400, Thailand
² Department of Epidemiology and Health, Robert Koch-Institute, Seestraße 10, 13353 Berlin, Germany
³ Department of Biostatistics, Faculty of Public Health, Mahidol University, 420/1 Rajvithi Road, Phayathai District, Bangkok 10400, Thailand

Received 28 January 2008, revised 15 March 2008, accepted 11 July 2008

Summary

The purpose of the study is to estimate the population size under a homogeneous truncated count model and under model contaminations via the Horvitz-Thompson approach on the basis of a count capture-recapture experiment. The proposed estimator is based on a mixture of zero-truncated Poisson distributions. The benefit of using the proposed model is statistical inference of the long-tailed or skewed distributions and the concavity of the likelihood function with strong results available on the nonparametric maximum likelihood estimator (NPMLE). The results of comparisons, for finding the appropriate estimator among McKendrick’s, Mantel-Haenszel’s, Zelterman’s, Chao’s, the maximum likelihood, and the proposed methods in a simulation study, reveal that under model contaminations the proposed estimator provides the best choice according to its smallest bias and smallest mean square error for a situation of sufficiently large population sizes and the further results show that the proposed estimator performs well even for a homogeneous situation. The empirical examples, containing the cholera epidemic in India based on homogeneity and the heroin user data in Bangkok 2002 based on heterogeneity, are fitted with an excellent goodness-of-fit of the models and the confidence interval estimations may also be of considerable interest.

Key words: Capture-Recapture; Population Size Estimation; Truncated Count Model; Population Heterogeneity; Mixture of Truncated Poisson Distributions; Model Contaminations.

1 Introduction

A common method used for estimating a population size $N$ is the capture-recapture experiment. Suppose that a homogeneous registration system identifies $n$ observed cases of a particular condition of interest and the system can also identify a case with probability $1 - p_0$ where $p_0$ is probability of the unidentified cases, so that $N = Np_0 + N(1 - p_0)$. Here, $N(1 - p_0)$ is the expected number of cases identified by the registry that can be estimated by $n$. This leads to the estimating equation $N = Np_0 + n$, which in other words can be stated that the population size, $N$, is the sum of both the unobserved and the observed cases ($n$). The equation can easily be solved for $N$ through the Horvitz-Thompson estimator

$$\hat{N} = n/(1 - p_0),$$  \hfill (1)

* Corresponding author: e-mail: phcvw@mahidol.ac.th, Phone: +66 (0)2 3548530, Fax: +66 (0)2 3548534

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
which is the number of observed cases adjusted for the probability of being included in the registration system. The Horvitz-Thompson approach seems easy, but the unknown probability $p_0$ must be estimated and can be vastly different in the various methods of estimation.

To estimate $p_0$, we need to have an identification mechanism. Suppose a unit can be identified several times, but the number of unidentified units is not known. For example, a surveillance system counts the number of times that a heroin user went to a treatment institution; a police record counts the number of arrests for a drunk driver. This type is called capture-recapture using repeated counting data.

Let $f(y, \lambda)$ denote a density for a count variate $Y$, where $\lambda$ is an unknown parameter. A standard count density is often provided by the Poisson density as

$$f(y, \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \ldots$$

We need to look at zero truncation since zeros (units which are not identified) are not observed in the registration system. Therefore, the associated density with zero truncation is

$$f_+(y, \lambda) = \frac{f(y, \lambda)}{1 - f(0, \lambda)} \cdot \quad y = 1, 2, \ldots$$

In Table 1, let $i$ be the number of times in a case, $n_i$ be the number of cases identified $i$ times, where $i = 1, 2, \ldots, m$ and $m$ is the largest number of times (counted) in a case. If $n$ is the total number of observed cases, then $n = n_1 + n_2 + \ldots + n_m$. Frequency of zero time, $n_0$, is not seen in the population and needs to be estimated.

For frequency data as written in Table 1, the likelihood function for zero-truncated count densities is

$$\prod_{i=1}^{m} (f_+(i, \lambda))^{n_i} = \prod_{i=1}^{m} \left( \frac{f(i, \lambda)}{1 - f(0, \lambda)} \right)^{n_i}, \quad i = 1, 2, \ldots, m$$

Under a zero-truncated Poisson, the log-likelihood function for frequency data is

$$l(\lambda) = \sum_{i=1}^{m} n_i \log \left[ \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{1}{1 - e^{-\lambda}} \right].$$

If an estimate $\hat{\lambda}$ is obtained to estimate the truncated count parameter $\lambda$, then (1) can be constructed according to

$$\hat{N} = \frac{n}{1 - f(0, \hat{\lambda})},$$

where $\hat{p}_0 = f(0, \hat{\lambda})$ is the probability of an individual not being observed. The frequency of zero time (unobserved cases) can also be estimated as

$$n_0 = \hat{N} - n = \frac{n}{1 - f(0, \hat{\lambda})} - n = n \left( \frac{f(0, \hat{\lambda})}{1 - f(0, \lambda)} \right).$$

In the next section we will consider various truncated count estimators, such as the McKendrick’s (McK), Mantel-Haenszel’s (MH), Maximum Likelihood (MLE), Zelterman’s (ZEL), and Chao’s (CHAO) estimators. In Section 3 we will propose the mixture of truncated count models (MIX) that can extend to cope with the presence of heterogeneity. Mixture models have a flexibility of modeling various forms of heterogeneity including homogeneity as a special case. The simulation studies and

Table 1  Observed counts in terms of cell frequency data.

<table>
<thead>
<tr>
<th>Times of repeated identifications (Count)</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency $n_i$</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>\ldots</td>
<td>$n_m$</td>
</tr>
</tbody>
</table>
their results for comparing performance among these truncated count estimators are presented in Section 4. In section 5, our approach provides some applications, and is followed by the confidence interval estimation in section 6. The discussion will appear in Section 7.

2 Truncated Count Estimators

Six estimators for estimating population sizes via the Horvitz-Thompson’ s approach contain a set of three estimators (McK, MLE, MH) based upon the homogeneous bases, a set of two estimators [ZEL, CHAO] derived from the heterogeneous bases, and a proposed estimator [MIX] derived by using either homogeneity or heterogeneity, depending upon one or more component sizes.

2.1 McKendrick’s Estimator (McK)

McKendrick (1926) presented the original zero-truncated Poisson estimator in his presentation at the Edinburgh Mathematical Society. He was confronted with the data of the cholera epidemic during the period of his service in India. A cholera epidemic affected a village with 223 houses. Let \( n_i \) be the number of houses with exactly \( i \) cases. The frequencies of the affected houses were \( n_1 = 32, n_2 = 16, n_3 = 6, \) and \( n_4 = 1 \). It is interesting to note that there is also a number, \( n_0 = 168 \), which reported the frequency of houses with no cholera cases. McKendrick knew that some unknown-cholera-affected houses were also included in these 168 houses, though no cases were observed. McKendrick ignored the 168 houses with zero cases in his analysis since they were not helpful in determining the number of affected houses, and he developed a moment estimator for the number \( n_0 \) of affected houses with no cases from the zero-truncated Poisson distribution.

If \( S_0 = \sum_{i=1}^{m} i^2 n_i = n, S_1 = \sum_{i=1}^{m} i n_i, \) and \( S_2 = \sum_{i=1}^{m} i^2 n_i, \) we have \( E(Y) = \lambda = E(S_1)/N \) and \( E(Y^2) = \text{var}(Y) + (E(Y))^2 = \lambda + \lambda^2 = E(S_2)/N. \) McKendrick’s moment estimators are provided by equating \( E(Y) \) and \( S_1/N \) as well \( E(Y^2) \) and \( S_2/N \) leading to \( \hat{\lambda}_{\text{MCK}} = (S_2 - S_1)/S_1 \) and \( \tilde{N}_{\text{MCK}} = S_1^2/(S_2 - S_1). \) By Taylor’s series theorem for \( \exp(-\hat{\lambda}) = 1 - \hat{\lambda} + \hat{\lambda}^2/2 - \ldots, \) we have \( 1 - \exp(-\hat{\lambda}) \approx \hat{\lambda}. \) An alternative to estimate \( N \) according to the Horvitz-Thompson’s method is to substitute \( \hat{\lambda}_{\text{MCK}} \) as

\[
\tilde{N}_{\text{HT}} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MCK}})} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MCK}})} \approx \frac{n}{\hat{\lambda}_{\text{MCK}}} = \frac{nS_1}{S_2 - S_1}. \quad (7)
\]

We can now state an interesting result concerning \( \tilde{N}_{\text{HT}} \) and \( \tilde{N}_{\text{MCK}} \) on the following:

**Theorem 1** Let \( Y_1, Y_2, \ldots, Y_n \) be a sample from a zero-truncated Poisson density as given in Eq. (2) and this sample is arranged in a form of frequencies \( n_1, n_2, \ldots, n_m \) with respect to \( 1, 2, \ldots, m \) counts where \( m \) is the largest count. Frequency of zero count, \( n_0 \), is unobserved. If \( S_0 = \sum_{i=1}^{m} i^2 n_i = n, S_1 = \sum_{i=1}^{m} i n_i, S_2 = \sum_{i=1}^{m} i^2 n_i, \) then \( 0 < S_0 \leq S_1 \leq S_2 \) and McKendrick’s moment estimator is provided as \( \hat{\lambda}_{\text{MCK}} = (S_2 - S_1)/S_1 \) and \( \tilde{N}_{\text{MCK}} = S_1^2/(S_2 - S_1). \) The Horvitz-Thompson estimate based upon \( \hat{\lambda}_{\text{MCK}} \) to estimate \( N \) is related to \( \tilde{N}_{\text{MCK}} \) via the inequality

\[
\tilde{N}_{\text{HT}} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MCK}})} \approx \frac{nS_1}{S_2 - S_1} \leq \tilde{N}_{\text{MCK}} = \frac{S_1^2}{S_2 - S_1}. \quad (8)
\]

**Proof** By the first order of Taylor’s series theorem for \( 1 - \exp(-\hat{\lambda}) \approx \hat{\lambda}, \) we have \( \tilde{N}_{\text{HT}} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MCK}})} \approx \frac{n}{\hat{\lambda}_{\text{MCK}}}. \) Substituting \( \hat{\lambda}_{\text{MCK}} = (S_2 - S_1)/S_1 \) and using \( 0 < S_0 \leq S_1 \leq S_2, \) we obtain

\[
\tilde{N}_{\text{HT}} \approx \frac{nS_1}{S_2 - S_1} = \frac{S_0S_1}{S_2 - S_1} \leq \frac{S_1^2}{S_2 - S_1} = \tilde{N}_{\text{MCK}}.
\]

Thus, (8) must hold with equality only for \( m = 1. \)
In this paper, we choose McKendrick’s estimator according to the Horvitz-Thompson’s approach, which is \( \hat{N}_{HT} = \frac{nS_1}{S_2 - S_1} \). Indeed, the cholera epidemic data are still referred to frequently. Dahiy and Gross (1973) investigated asymptotic properties of the McKendrick estimator. Irwin (1963) looked at maximum likelihood estimation to improve McKendrick’s method with the help of iteration. Meng (1997) used to illustrate the expectation-maximization (EM) algorithm. Blumenthal et al. (1978) investigated maximum likelihood estimation for this case. Mao and Lindsay (2003) showed that McKendrick estimator and the improved maximum likelihood solution provided an excellent fit to the data.

### 2.2 Maximum likelihood estimator (MLE)

From (4), the log-likelihood function under a zero-truncated Poisson can be rewritten as

\[
L(\lambda) = \lambda n \log \lambda - \sum_{i=1}^{m} n_i \log (i!) - n \log (1 - e^{-\lambda})
\]

Differentiation with respect to \( \lambda \) and finding the maximum likelihood estimator by solving \( \frac{dL}{d\lambda} = 0 \), yields \( \hat{\lambda}_{\text{MLE}} = \bar{y}(1 - e^{-\hat{\lambda}_{\text{MLE}}}) \). However, this does not have a closed form solution. Iterative methods must be applied to achieve \( \hat{\lambda}_{\text{MLE}} \) up to an acceptable error. Practically, if we apply the second order of Taylor series theorem for \( \exp(-\hat{\lambda}) \approx 1 - \hat{\lambda} + \hat{\lambda}^2/2 \), then it is not necessary to use an iterative method. The solution of the maximum likelihood method is given by \( \hat{\lambda} = \bar{y}(1 - e^{-\hat{\lambda}}) \approx \bar{y}(\hat{\lambda} - \hat{\lambda}^2/2) \). It yields

\[
\hat{\lambda}_{\text{MLE}} = 2(\bar{y} - 1)/\bar{y}.
\]

Hence, the population-size estimator for the maximum likelihood method and the approximate maximum likelihood is

\[
\hat{N}_{\text{MLE}} = n/(1 - \exp(-\hat{\lambda}_{\text{MLE}})) \quad \text{and} \quad \tilde{N}_{\text{MLE}} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{MLE}})}.
\]

### 2.3 Zelterman’s estimator (ZEL)

Zelterman (1988) proposed a family of estimators of the relative frequency of the unobservable zero class in a truncated Poisson distribution. His attention was restricted to a possibly contaminated Poisson sampling. To find an estimate for \( p_0 = f(0, \bar{\lambda}) = \exp(-\bar{\lambda}) \) under the truncated Poisson sampling, Zelterman (1988) proposed a family of robust estimators as

\[
\hat{Q}_i = \exp(-\hat{\lambda}_{\text{ZEL}}) = \exp\left( -\frac{(i + 1)n_{i+1}}{n_i} \right), \quad i = 1, 2, \ldots, m - 1,
\]

where the truncated Poisson estimator can be found as follow:

\[
\hat{\lambda}_{\text{ZEL}} = \frac{(i + 1)n_{i+1}}{n_i}, \quad i = 1, 2, \ldots, m - 1.
\]

In practice, \( i \) is chosen to be small, such as one or two. We choose \( i = 1 \); consequently, \( \hat{Q}_1 = \exp(-2n_2/n_1) \). The population size of Zelterman’s estimator can be written as

\[
\hat{N}_{\text{ZEL}} = \frac{n}{1 - f(0, \hat{\lambda}_{\text{ZEL}})} = \frac{n}{1 - \exp(-\hat{\lambda}_{\text{ZEL}})}.
\]

### 2.4 Mantel Haenszel’s estimator (MH) as an extension of Zelterman’s estimator

Wannasirikul (2005) proposed an extension of Zelterman’s estimator by using the Mantel-Haenszel method by taking the weighting mean into the Eq. (12).

\[
\hat{\lambda}_{\text{MH}} = \frac{\sum_{i=1}^{m-1} W_i (i + 1)n_{i+1}/n_i}{\sum_{i=1}^{m-1} W_i}.
\]
If we let \( W_i = n_i \), we achieve
\[
\hat{\lambda}_{\text{MH}} = \frac{\sum_{i=1}^{m-1} (i+1) n_{i+1}}{\sum_{i=1}^{m-1} n_i}.
\]
Hence, the population size estimator according to Horvitz-Thompson method is
\[
\hat{N}_{\text{MH}} = \frac{n}{1-f(0, \hat{\lambda}_{\text{MH}})} = \frac{n}{1 - \exp (-\hat{\lambda}_{\text{MH}})}.
\]
Indeed, Moore (1952, 1954) proposed the estimation of the Poisson parameter from a truncated distribution. The result of Moore’s estimator is identical to (14). Even though the truncated estimators, derived from both methods of Mantel-Haenszel and Moore, are identical they are obtained from different ideas.

2.5 Chao’s estimator (CHAO)
Let \( p_0, p_1, p_2, \ldots, p_m \) denote the probabilities of units identified 0, 1, 2, \ldots, \( m \) times and similarly let us look at the observed frequencies \( n_1, n_2, \ldots, n_m \) of units identified 1, 2, \ldots, \( m \) times within the time span of the study period. Under homogeneity, we let the distribution function of \( p_0, p_1, p_2, \ldots, p_m \) be a Poisson, that is, \( p_i = f(i, \lambda) = e^{-\lambda} \lambda^i/i! \), for \( i = 0, 1, \ldots \) However, the homogeneous model is not appropriate in many situations in real life, Chao (1987) has proposed the heterogeneous model with \( p_i = \int_0^{\infty} f(i, \lambda) dG(\lambda) = \int_0^{\infty} e^{-\lambda} \lambda^i/i! g(\lambda) d\lambda \), where \( g(\lambda) \) represents the heterogeneity density of the model parameter in the population. The important lower bound estimator of Chao’s method for estimating the population size is based on the Cauchy-Schwarz inequality. We now consider some forms of the inequality of Cauchy-Schwarz
\[
\left( \int_0^\infty u(\lambda) v(\lambda) g(\lambda) d\lambda \right)^2 \leq \left( \int_0^\infty u(\lambda)^2 g(\lambda) d\lambda \right) \left( \int_0^\infty v(\lambda)^2 g(\lambda) d\lambda \right),
\]
where \( u(\lambda) = \sqrt{\exp (-\lambda)} \) and \( v(\lambda) = \sqrt{\exp (-\lambda)} \). This can be written as \( p_i^2 \leq p_0^2(2p_2) \). Hence, the lower bound for the unknown probability of a zero is \( p_0 \geq p_i^2/(2p_2) \). Replacing unknown probabilities by their associated frequencies, we achieve Chao’s lower bound estimate
\[
\hat{n}_0 \geq n_i^2/(2n_2).
\]
The population size estimator according Chao’s method is
\[
\hat{N}_{\text{CHAO}} = n_i^2/(2n_2) + n.
\]

3 The Proposed Estimator Based on the Mixture of Truncated Count Distributions (MIX)
The contaminating idea is extended to a mixture of two or more distributions. Böhning et al. (2005) proposed the truncated count mixture models for capture-recapture count data with missing zero cases. Böhning and Kuhnert (2006) showed the equivalence of the truncated count mixture distributions and the mixtures of truncated count distributions. They stated that while the former model was more natural to formulate and interpret, the latter model was theoretically easier to treat. The results can be found that for any mixing distribution of the truncated mixture, a (usually different) mixing distribution of the mixture of truncated counts can be found so that the two sets of mixture distributions are identical. This implies that the likelihood surfaces for both situations agree, and both models are equivalent. Consequently, for estimating population size, the two Horvitz-Thompson estimators, associated with the two models, have equal value. In this article, we are interested in achieving strong results for mixtures of truncated Poisson densities, including reliable, global construction of the un-
ique NPMLE (nonparametric maximum likelihood estimator) of the mixing distribution implying a unique estimator for the population size.

Suppose that a contamination is composed of $k$ distributions with the zero-truncated count densities $f_+(y, \lambda_1), f_+(y, \lambda_2), \ldots, f_+(y, \lambda_k)$ with parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$. The mixture of zero-truncated count densities is provided as

$$f_+(y, Q) = \sum_{j=1}^{k} q_j f_+(y, \lambda_j), \quad y = 1, 2, \ldots,$$

where $f_+(y, \lambda_j) = \frac{f(y, \lambda_j)}{1 - f(0, \lambda_j)}$ and $q_1, q_2, \ldots, q_k$ represent the mixing proportions where $q_j \geq 0$ and $\sum_{j=1}^{k} q_j = 1$. The mixing distribution $Q$, giving weight $q_j$ with respect to $\lambda_j$ for $j = 1, 2, \ldots, k$, is denoted as

$$Q = \left( \begin{array}{cccc} \lambda_1 & \lambda_2 & \ldots & \lambda_k \\ q_1 & q_2 & \ldots & q_k \end{array} \right).$$

For a relative frequency data, let $n_1, n_2, \ldots, n_m$ be the frequencies of exactly $i$ time values, where $i = 1, 2, \ldots, m$ and $m$ is the largest time. The incomplete-data log-likelihood in those terms for mixtures of zero-truncated count densities is

$$l(Q) = \sum_{i=1}^{m} n_i \log f_+(i, Q) = \sum_{i=1}^{m} n_i \log \left( \sum_{j=1}^{k} q_j f_+(i, \lambda_j) \right).$$

The gradient function as a normalized version of the directionally partial derivatives with respect to $q_j$ is given by

$$d(\lambda, Q) = \frac{1}{n} \sum_{i=1}^{m} n_i \frac{f_+(i, \lambda)}{f_+(i, Q)}.$$

With the help of the gradient function, the nonparametric maximum likelihood estimators (NPMLE) of $Q$ can be characterized. If the log-likelihood is concave overall parameter space, the NPMLE provides the global maximum. Local maximum likelihood solutions can exist only for component sizes that are smaller than the component size of the NPMLE. The general mixture likelihood theorem (Laird, 1978; Lindsay, 1983; Böning, 2000) states that $\hat{Q}$ is NPMLE $\Leftrightarrow d(\lambda, \hat{Q}) \leq 1$ for all $\lambda > 0$.

(i) EM-Algorithm

A variety of numerical algorithms exists to find the global maximum likelihood estimator, the NPMLE, if it exists. It has become very popular to use the EM algorithm (Dempster, Laird, and Rubin, 1977; Böning, 2000) in connection with mixture models (McLachlan and Krishnan, 1997; McLachlan and Peel, 2000). The NPMLE is frequently computed by the EM algorithm, by fixing the component size $k$, and by gradually increasing stepwise from 1 to the maximum component size of the NPMLE. To proceed in the EM context, we need the complete data log likelihood, which is given in this case as

$$l_{CD}(Q) = \sum_{i=1}^{m} n_i \sum_{j=1}^{k} z_{ij} \log f_+(i, \lambda_j) + \log q_j,$$

where the unobserved covariate $z_{ij}$ is 1 if $i$ belongs to component $j$ and 0 otherwise. The EM algorithm, containing Expectation and Maximization steps repeatedly, increases the conditional expectation...
of the complete-data log-likelihood (23). In the E-step, the unobserved indicator variates $z_{ij}$ are replaced by their expected values, $e_{ij}$, conditional upon the observed data and the current values of $\lambda_j$ and $q_j$ for $j = 1, 2, \ldots, k$, leading to

$$e_{ij} = E(z_{ij} \mid n_{ij}; q_j, \lambda_j) = P(z_{ij} = 1 \mid n_{ij}; q_j, \lambda_j) = \frac{f_x(i, \lambda_j) q_j}{\sum_{j=1}^{k} f_x(i, \lambda_j') q_j}.$$  \hfill (24)

In order to improve the parameter estimates, in the M-step, the new values, $\hat{\lambda}_1, \ldots, \hat{\lambda}_k, \hat{q}_1, \ldots, \hat{q}_k$ are found, which maximize the expected value of complete log likelihood (23),

$$\sum_{i=1}^{m} n_i \sum_{j=1}^{k} e_{ij} \log f_x(i, \lambda_j) + \sum_{i=1}^{m} n_i \sum_{j=1}^{k} e_{ij} \log q_j.$$  \hfill (25)

Both double sums in (25) can be maximized separately since they are not related. To find the expressions for $q_j$ in the second term of (25), we introduce Lagrange’s method with the constraint that $\sum_{j=1}^{k} q_j = 1$, and the results of the weighting estimates $\hat{q}_1, \ldots, \hat{q}_k$ are obtained by

$$\hat{q}_j = \frac{1}{n} \sum_{i=1}^{m} n_i e_{ij}, \quad \text{for} \quad j = 1, \ldots, k.$$  \hfill (26)

Similarly, in the first term of (25), the new estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ will depend on the form of $f_x(i, \lambda_j)$. The solution after solving the equations of derivatives with respect to $\lambda_j$ is obtained by

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{m} m_i e_{ij} (1 - f(0, \hat{\lambda}_j))}{\sum_{i=1}^{m} m_i e_{ij}}, \quad \text{for} \quad j = 1, \ldots, k.$$  \hfill (27)

In the sense of estimating an unknown population size $N$, the Horvitz-Thompson estimate under the mixture of zero-truncated count models is

$$\hat{N}_{\text{MIX}} = n \sum_{j=1}^{k} \hat{q}_j (1 - f(0, \hat{\lambda}_j)).$$ \hfill (28)

The following algorithm is used to compute the population size estimator.

**Step 0** Choose some initial values of $\hat{q}_j^{(0)}$ and $\hat{\lambda}_j^{(0)}$ for $\hat{Q}^{(0)}$, and set $j = 0$.

**Step 1** Compute $\hat{N}^{(j+1)} = n \sum_{j=1}^{k} \hat{q}_j^{(j)} (1 - f(0, \hat{\lambda}_j^{(j)}))$ and $\hat{N}_0^{(j+1)} = \hat{N}^{(j+1)} - n$.

**Step 2** Use the complete frequency table $n_i^{(j+1)}, n_1, \ldots, n_m$ to compute new maximum likelihood estimates of $\hat{q}_j^{(j+1)} = \frac{1}{n} \sum_{i=1}^{m} n_i e_{ij}$ and $\hat{\lambda}_j^{(j+1)} = \frac{\sum_{i=1}^{m} m_i e_{ij} (1 - f(0, \hat{\lambda}_j^{(j)}))}{\sum_{i=1}^{m} m_i e_{ij}}$ for $\hat{Q}^{(j+1)}$, where

$$e_{ij}^{(j)} = \frac{f_x(i, \hat{\lambda}_j^{(j)}) q_j^{(j)}}{\sum_{i=1}^{m} f_x(i, \hat{\lambda}_j^{(j)}) q_j^{(j)}},$$

then set $j = j + 1$ and go back to Step 1.

The steps 1 and 2 are repeated until convergence to a constant with an acceptable error. With this algorithm, $\hat{N}_{\text{MIX}}$ and $\hat{N}_0$ are obtained directly. The nested EM algorithm is discussed in more detail in McLachlan and Krishnan (1997) and Böhning and Kuhnert (2006).

### 4 Performance among Truncated Count Estimators

To compare the performance in terms of bias, variance, and mean square errors (MSE) for a population size, we used FORTRAN90 with IMSL library to develop some procedures and results.
4.1 A simulation plan
The simulation plan is divided into two cases: one under a homogeneous Poisson model and another under a contamination model of two-component Poisson mixture models. Let the Poisson parameter \( \lambda \) under homogeneity be some constants varying from 1 to 5. The pairs of two Poisson mixture parameters \((\lambda_1, \lambda_2)\) under heterogeneity are (1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5) and (2,6) with equal weights. The differences between two Poisson parameters defining the degree of heterogeneity (\(DH\)) are 1 (close to homogeneity), 2 (weak heterogeneity), 3 (moderate heterogeneity), and 4 (strong heterogeneity). The population sizes \(N\) are 100, 200 (for small), 500, 1000, 2000 (for medium), and 5000, 10000 (for large). The Poisson data \(Y\) are generated and arranged in a form of frequencies \(n_0; n_1; \ldots; n_m\) with respect to the counts 0, 1, 2, \ldots, \(m\). The frequency of the zero count, \(n_0\), is dropped before proceeding to computation. All estimates are then computed and the measures of the performance are evaluated from all 5000 replications.

4.2 Results for comparisons among truncated count estimators
The results contain a number of details that we cannot entirely reproduce here; however, the interested reader can request the details directly from the correspondent author or the website of Biometrical Journal. In the following we focus on estimating population size \(N\) under homogeneity and heterogeneity.

![Graph](https://example.com/graph.png)

**Figure 1** Bias for Homogeneity a) Lamda = 2, b) \(N = 200\).
4.2.1 Results for homogeneity

The three estimators MLE, MH, and McK would be expected to be more appropriate since they rely on the assumption of homogeneity. However, the results in Figure 2 illustrate that the appropriate estimators to estimate the population size are MLE, MH, and MIX (containing one component in this case) with the smallest MSE.

Figure 1 shows the performance in terms of bias of six population size estimates under homogeneity. We found that all five estimators MLE, MH, MIX, CHAO, and McK perform well with small values of bias, and the ZEL estimator performs the worst with its largest bias value. In more detail, the MH, MLE, and MIX perform well when population size is small \( (N \leq 200) \). For medium to large population sizes \( (N \geq 2000) \), the McK and MLE methods perform best with smallest bias; however, for some cases of population sizes, the MIX and CHAO methods give the smallest values of bias. The MLE estimator is asymptotically unbiased with respect to \( N \) as are McK and MH. The bias of the ZEL estimator is persistent, regardless of both \( N \) and \( \lambda \).

According to variance and MSE, the patterns of the variance and MSE estimates in all population size estimators are similar in this study. Figure 2 shows the performance in terms of MSE. The MH, MLE, and MIX methods perform well generally with the smallest MSE. The MH method is better for smaller population sizes \( (N \leq 200) \) while the MLE and MIX methods perform best especially for larger population sizes \( (N \geq 2000) \). The methods of ZEL and CHAO are the worst under homogeneity with the largest MSE in all situations.

![Figure 2](http://www.biometrical-journal.com)  
Figure 2  MSE for Homogeneity a) Lamda = 2, b) \( N = 5000 \).
Note here that the proposed estimator, MIX, based upon one component of the mixture of the truncated count distributions, is identical to the MLE, so that the results coincide (small numerical differences sometimes occur due to computational issues).

4.2.2 Results for heterogeneity

The results show that the two estimators ZEL and CHAO, derived for the heterogeneous case, have good performance while the MIX estimator with $k = 2$ components in this simulation plan is an excellent candidate for use. Moreover, all three estimators McK, MH, and MLE derived under homogeneity show the worst performance. The results depend on the population size ($N$) and the strength or degree of heterogeneity ($DH$) and it can be summarized as follows.

**Bias** Figure 3 shows that the methods of ZEL and CHAO perform well with the smallest bias under the closeness of homogeneity ($DH = 1$) and under a weak degree of heterogeneity ($DH = 2$) in combination with small to moderate population sizes ($N = 200$ to $2000$). The proposed estimator, MIX, is an excellent estimator providing the smallest values of bias for all large population sizes ($N = 5000, 10000$); also it performs best with the smallest bias when the degrees of heterogeneity are moderate to large ($DH = 3$ to $4$) in combination with smaller population sizes ($N = 200$ to $2000$).

**Variance** In all cases, regardless of $N$ and $DH$, the McK and MLE have the smallest values of the variance, followed by MH. These three variance estimates are not far apart.

**Figure 3** Bias for Heterogeneity a) $N = 500$, b) $N = 10000$. 

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim  
www.biometrical-journal.com
Three estimators, named CHAO, ZEL, and MIX, allowing for heterogeneity, still perform best with the smallest MSE. Figure 4 shows the performance in a term of MSE for all six estimators under heterogeneity. For small to moderate population sizes \(N = 500\), the CHAO estimator looks best. The proposed estimator, MIX, performs best with the smallest MSE for large population sizes \(N = 5000; 10000\), and it also remains best when the degrees of heterogeneity are larger \(DH = 3\) to \(4\) in combination with moderate population sizes \(N = 2000; 5000\). The ZEL estimator has smallest MSE when the degrees of heterogeneity are weak \(DH = 1; 2\) in combination with medium population sizes \(N = 2000; 5000\).

In summary, the proposed estimator, MIX, performs best when population size is large \(N = 5000; 10000\) and when population size is moderate to large \(N = 2000; 5000\) in combination with higher degrees of heterogeneity \(DH = 3, 4\).

5 Empirical Applications

Example 1 (Cholera epidemic in India) The cholera data were chosen to illustrate an example of a homogeneous population of a zero-truncated count model. A cholera epidemic affected a village with 223 houses. Let \(n_i\) be the number of houses with exactly \(i\) cases. The observed frequencies of the affected houses were \(n_1 = 32, n_2 = 16, n_3 = 6, \) and \(n_4 = 1\). A chi-square goodness of fit test under the null hypothesis of a homogeneous zero-truncated Poisson data is computed by \(\chi^2 = \sum_{i=1}^{4} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}\).
with degrees of freedom \( m' = 2 \), where \( \hat{n}_i \) are the expected frequencies of the affected houses with \( i \) cases denoted by \( \hat{n}_i = n f_i (\hat{\lambda}, \hat{\alpha}_{MLE}) \) and \( m' \) is the largest number of cells having expected values larger than five. Cells with fewer than five expected values were collapsed. The conclusion reveals that the cholera data follow homogeneity of a zero-truncated Poisson distribution with \( p \)-value of 0.85.

**Example 2** (Drug users in Bangkok 1999–2002) The data of the counts of the treatment episodes on drug use provide an example of heterogeneity. The list is anonymous information containing all 61 health treatment centers in the Bangkok metropolitan region in the years 1999–2002, which were permitted to treat drug addicts, and arise from the surveillance system of the Office of the Narcotics Control Board (ONCB), Ministry of Prime Minister, in cooperation with the Drug Abuse Prevention and Treatment Division, Health Department and Medical Service Department, Bangkok Metropolitan Administration. Each anonymous record of the treatment episodes in the list of the surveillance system is linked to the same patient with some covariates, such as age, gender, date of birth, district and city of birth, present address, hospital number, hospital name, and date of receiving treatment episodes. From this anonymous source, it is possible to construct information on the frequency of episodes for each patient. Most of the drug users are heroin and methamphetamine. In this study, the application is solely focused on heroin users in 2002. The observed distributions of the counts of the treatment episodes for heroin users are \( n_1 = 2176, n_2 = 1600, n_3 = 1278, n_4 = 976, n_5 = 748, n_6 = 570, n_7 = 455, n_8 = 368, n_9 = 281, n_{10} = 254, n_{11} = 188, n_{12} = 138, n_{13} = 99, n_{14} = 67, n_{15} = 44, n_{16} = 34, n_{17} = 17, n_{18} = 3, n_{19} = 3, n_{20} = 2, \) and \( n_{21} = 1 \). The data of heroin users under the null hypothesis of a homogeneous truncated Poisson have an inadequate goodness of fit with 15718.4 chi-square value, at 10 degrees of freedom and \( p \)-value is less than 0.00001. In addition, these data are fitted with the excellent goodness of fit under a mixture of three components \( (k = 3) \) of truncated Poisson distributions with \( p \)-value of 0.44 at degrees of freedom \( m' - 1 - p = 15 \) where \( m' \) is the largest number of cells having expected values larger than five and \( p \) is number of estimated parameters in the model. For this mixture of truncated Poisson densities, \( p = 2k - 1 = 5 \), and the expected frequency \( \hat{n}_i \) with \( i \) treatment episodes is denoted by \( \hat{n}_i = n f_i (\hat{\lambda}, \hat{\alpha}) = n \left( \sum_{j=1}^{k} \hat{q}_j f_j (\hat{\lambda}_j) \right) \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \hat{\lambda}_j )</th>
<th>( \hat{q}_j )</th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>BIC</th>
<th>( \hat{n}_0 )</th>
<th>( \hat{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cholera epidemic data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>1.00</td>
<td>-54.7</td>
<td>-111.4</td>
<td>-113.4</td>
<td>33</td>
<td>88</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.69</td>
<td>-54.7</td>
<td>-115.4</td>
<td>-125.4</td>
<td>33</td>
<td>88</td>
</tr>
<tr>
<td>2</td>
<td>0.97</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Heroin users in Bangkok 2002</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.13</td>
<td>1.00</td>
<td>-25045</td>
<td>-50092</td>
<td>-50099</td>
<td>151</td>
<td>9453</td>
</tr>
<tr>
<td>1</td>
<td>2.08</td>
<td>0.67</td>
<td>-21515</td>
<td>-43035</td>
<td>-43057</td>
<td>891</td>
<td>10193</td>
</tr>
<tr>
<td>2</td>
<td>7.88</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td><strong>1.04</strong></td>
<td><strong>0.34</strong></td>
<td><strong>-21362</strong></td>
<td><strong>-42733</strong></td>
<td><strong>-42769</strong></td>
<td><strong>1827</strong></td>
<td><strong>11130</strong></td>
</tr>
<tr>
<td>2</td>
<td>3.84</td>
<td>0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>9.22</strong></td>
<td><strong>0.20</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.45</td>
<td>0.15</td>
<td>-21360</td>
<td>-42733</td>
<td>-42782</td>
<td>2879</td>
<td>12182</td>
</tr>
<tr>
<td>2</td>
<td>2.28</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4.99</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9.56</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

\( \sum_{j=1}^{k} \hat{q}_j f_j (\hat{\lambda}_j) \) for these data are fitted with the excellent goodness of fit under a mixture of three components. 

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

www.biometrical-journal.com
For two data sets under homogeneity and heterogeneity, the nonparametric maximum likelihood estimations (NPMLE) based on the mixture of truncated count models are illustrated in Table 2. We consider the Akaike’s Information Criterion (AIC), the Bayesian Information Criterion (BIC), the log-likelihood values, and the above chi-square goodness of fit of the mixture, in order to evaluate and select the best model in the context of mixture models. The first data set (cholera epidemic in India) with no heterogeneity is fitted by the NPMLE with one component solution and the estimate of the affected houses with no cholera case is found as $\hat{n}_0 = 33$, leading to the total number of the affected houses $\hat{N} = 88$. For the second data set (heroin users in Bangkok 2002) with a long-tailed distribution, the mixture of three-components of zero-truncated Poisson models is the best fitted and this produces that $\hat{n}_0 = 1827$ for the unobserved number of heroin users and $\hat{N} = 11130$ for the total number of heroin users.

### 6 Confidence Interval Estimation

Before computing the confidence intervals, the variances of the estimates need to be derived.

#### 6.1 Variances

Under homogeneity of a zero-truncated Poisson model, Van der Heijden et al. (2003) proposed the estimated variance of $\hat{N}$. In this paper their methods are used for three homogeneous estimators of McKendrick, Maximum likelihood, and Mantel-Haenszel as follows:

$$
\text{var}(\hat{N}) = \left( n - \frac{\exp(-\hat{\lambda})}{1 - \exp(-\hat{\lambda})} \right)^2 \left( \frac{n[1 - \exp(-\hat{\lambda}) - \hat{\lambda}\exp(-\hat{\lambda})]}{\hat{\lambda}(1 - \exp(-\hat{\lambda}))^2} \right)^{-1} + \left( \frac{n\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \right),
$$

where the estimated values of $\hat{\lambda}$ are replaced by $\hat{\lambda}_{M,K}$, $\hat{\lambda}_{MLE}$, and $\hat{\lambda}_{MH}$ for McKendrick’s, Maximum likelihood, and Mantel-Haenszel’s methods, respectively. Then a 95% confidence interval for $N$ is written as:

$$
\hat{N} \pm 1.96 \sqrt{\text{var}(\hat{N})}.
$$

According to Zelterman (1988), the estimated variance of $\hat{Q}_1$ can be obtained as

$$
\text{var}(\hat{Q}_1) = \frac{\exp(-\hat{\lambda}_{MLE})(1 - \exp(-\hat{\lambda}_{MLE}))}{n} \left( \hat{\lambda}_{MLE} + 2 \right).
$$

Hence, a 95% confidence interval for $N$ of Zelterman’s method is written as

$$
\frac{n}{(1 - \exp(-\hat{\lambda}_{ZEL}))} \pm 1.96 \sqrt{\text{var}(\hat{Q}_1)}.
$$

Chao (1987) proposed for her lower bound estimator of the population size a variance estimator calculated by a standard asymptotic approach given as

$$
\text{var}(\hat{N}_{CHAO}) = \frac{n_2(0.25(n_1/n_2)^4 + (n_1/n_2)^3 + 0.5(n_1/n_2)^2)}.
$$

Assuming the normality of $\hat{N}_{CHAO}$, the approximate 95% confidence interval is

$$
\hat{N}_{CHAO} \pm 1.96 \sqrt{\text{var}(\hat{N}_{CHAO})}.
$$

For the mixture of truncated count distributions, its variance is not a by-product of the EM-procedures and the direct computation via the information matrix is usually difficult; therefore, the
bootstrap resampling technique is applied in this case. Firstly, $n^{(b)}$ is sampled from a binomial distribution with a success parameter $p = n/N$ and a sample size parameter $N$, as well as the parameter in the mixture of truncated count distributions, is estimated from the origin data set. Secondly, frequencies $n^{(b)}_1, n^{(b)}_2, \ldots, n^{(b)}_m$ are sampled from the mixture of truncated count models with parameters as estimated in the original data set. Note that

$$n^{(b)}_i = \# \{ Y^{(b)}_i = i \}$$

where $Y^{(b)}_i \sim f_{\hat{\Phi}}(y; \hat{\Theta}) = \sum_{j=1}^{k} \hat{q}_j f_{\hat{\theta}}(y, \hat{\lambda}_j)$ and $\sum_{i=1}^{m} n^{(b)}_i = n^{(b)}$. Alternatively, $n^{(b)}_1, n^{(b)}_2, \ldots, n^{(b)}_m$ can also be sampled from a multinomial distribution, that is,

$$f(n^{(b)}_1, n^{(b)}_2, \ldots, n^{(b)}_m) = \left( \frac{n^{(b)}}{n^{(b)}_1, n^{(b)}_2, \ldots, n^{(b)}_m} \right) \prod_{i=1}^{m} [f_{\hat{\Phi}}(i; \hat{\Theta})]^{n^{(b)}_i},$$

where $p_i = f_{\hat{\Phi}}(i; \hat{\Theta})$ denotes the probability of cases identified $i$ times from the mixture under the parametric approach. For the nonparametric approach, $p_i = n_i/n$ can also be estimated from the original data set. In this paper, $p_i = n_i/n$ is used. For each of these $B$ resamples, $n^{(b)}_i$ is estimated by using the EM algorithm for the mixture of truncated count models; these resample data are used to compute variances and confidence intervals. It was found that the statistics of interest stabilized beyond $B = 1000$, so $B = 5000$ was considered to be sufficient in all the bootstrap calculations. The confidence intervals were calculated as asymptotic normal intervals.

### 6.2 Confidence interval estimations in empirical applications

Table 3 presents the various estimates of the total number and their associated 95% confidence intervals. For the cholera epidemic data evidence has been provided for homogeneity. Accordingly, estimates do not differ much; neither do their confidence intervals with the exception of Zelterman which has a large confidence interval. For the anonymous data of the counts of the treatment episodes on heroin users in Bangkok 2002 heterogeneity has been established. Hence, there is a first group of estimators which use false homogeneity and hence underestimate. The second group incorporates heterogeneity and produces considerably higher estimates. It is interesting to see that the mixture model produces the smallest variance estimate (hence shortest confidence interval) and, as expected, Chao’s interval is smaller than that for Zelterman. It is clear from the previous analysis that the second group is the more reliable estimators.

<table>
<thead>
<tr>
<th>Estimators</th>
<th>$N$</th>
<th>95% Confidence Interval for $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholera epidemic data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McKendrick’s</td>
<td>91</td>
<td>67–115</td>
</tr>
<tr>
<td>Mantel-Haenszel’s</td>
<td>88</td>
<td>67–110</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>89</td>
<td>67–112</td>
</tr>
<tr>
<td>Zelterman’s</td>
<td>88</td>
<td>65–134</td>
</tr>
<tr>
<td>Chao’s</td>
<td>87</td>
<td>58–116</td>
</tr>
<tr>
<td>Mixture of Truncated Count</td>
<td>88</td>
<td>82–94</td>
</tr>
<tr>
<td>Heroin users in Bangkok 2002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McKendrick’s</td>
<td>9328</td>
<td>9318–9339</td>
</tr>
<tr>
<td>Mantel-Haenszel’s</td>
<td>9481</td>
<td>9454–9509</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>9453</td>
<td>9428–9479</td>
</tr>
<tr>
<td>Zelterman’s</td>
<td>12077</td>
<td>11979–12177</td>
</tr>
<tr>
<td>Chao’s</td>
<td>10782</td>
<td>10620–10944</td>
</tr>
<tr>
<td>Mixture of Truncated Count</td>
<td>11130</td>
<td>11014–11246</td>
</tr>
</tbody>
</table>
7 Discussion

Recently, there has been an increased interest in zero-truncated count models. These models can be applied in social science areas such as illegal immigrants, illegal gun owners, illicit drug users, car drivers without license, or criminals to provide the estimates of the size of the unobserved population. This article has shown the performance of the mixture model for truncated counts compared with those of other models to estimate population size. The conclusions are that the proposed model based on the mixture of the truncated Poisson models provides the best estimated population size under heterogeneity and performs well even under homogeneity. This finding fits into a variety of scientific research work since Everitt (1996) and Hogg and Klugman (1984) had stated the benefit of mixture modeling in that it can be used to naturally model long-tailed and/or skewed distributions and also can cope with data from populations known or suspected to contain a number of separate subpopulations. Schork, Allison, and Thiel (1996) also stated the extremely flexible statistical modeling devices of mixture distributions. In addition, under heterogeneity, Chao’s estimator provides a lower bound value for estimating population size while Zelterman’s estimator provides an upper bound, so the mixture of truncated count models seems to provide the appropriate compromise value lying between Chao’s and Zelterman’s estimators.

Further study could focus on the estimation of the variance of the population size estimators and investigate especially the direct estimation of the variance for the mixture of the truncated count model. This is not an easy task as stated by Chao (1989), Cormack (1992), and Böhning et al. (2005), so that in this study we use the bootstrap resampling technique to compute the variance for the population size based upon the mixture model. Other parametric models such as the binomial model, the hypergeometric model, or the inverse sampling of the negative binomial model should be considered in any future research.

Acknowledgements

The work of Chukiat Viwatwongkasem was supported by the National Research Council of Thailand. The authors wish to express their thanks to Busaba Suppawattanabodee as well as the Office of the Narcotics Control Board, the Drug Abuse Prevention and Treatment Division, Health Department and Medical Service Department, Bangkok Metropolitan Administration, for the anonymous drug user data in Bangkok. The authors are also grateful to the Editor, the Associate Editor, and Reviewers for their helpful and clarifying comments.

Conflict of Interests Statement

The authors have declared no conflict of interest.

References


