*Applied Mathematics*, 2013, 4, 1242-1250
http://dx.doi.org/10.4236/am.2013.49168 Published Online September 2013 (http://www.scirp.org/journal/am)

# Mixture Models for Estimating the Number of Drug Users in Thailand 2005-2007

**Chukiat Viwatwongkasem[1*], Pratana Satitvipawee[1], Suthi Jareinpituk[2], Pichitpong Soontornpipit[1]**
[1]Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand
[2]Department of Epidemiology, Faculty of Public Health, Mahidol University, Bangkok, Thailand
Email: [*]chukiat.viw@mahidol.ac.th

Received June 19, 2013; revised July 19, 2013; accepted July 26, 2013



## ABSTRACT

It is difficult to measure the sizes of illegal drug user populations directly by using the survey method because of many "hidden drug addicts" and the difficulty of receiving a true response. Systematic and routine information on treatment episodes of drug users is adopted to estimate the population size in this study. Mixture models of zero-truncated Poisson distributions using the nonparametric maximum likelihood estimators (NPMLE) by means of capture-recapture repeated count data were used to project the number of drug users. The method was applied to surveillance data of drug users identified by treatment episodes in over 1140 health treatment centers in Thailand from the Bureau of Health Service System Development, Ministry of Public Health. We presented how this mixture model could be utilized to construct the unobserved frequency of drug users with no treatment episode and further estimated the total population size of drug users in the country from 2005 to 2007. The result of simulation was confirmed that mixture model is suitable when population is large. By means of mixture models, the estimations for the number of drug users were fitted with excellent goodness-of-fit values and we were also compared to the conventional Chao estimates. The NPMLE for the total number of drug users in Thailand 2005, 2006, and 2007 were 184,045 (95% CI: 181,297 - 186,793), 230,665 (95% CI: 226,611 - 234,719), 299,670 (95% CI: 294,217 - 305,123), respectively, also 125,265 (95% CI: 123,092 - 127,142), 166,287 (95% CI: 163,222 - 169,352), 228,898 (95% CI: 224,766 - 233,030) for the number of methamphetamine (Yaba) users, and 11,559 (95% CI: 10,234 - 12,884), 11,333 (95% CI: 9276 - 13,390), 8953 (95% CI: 7878 - 10,028) for the number of heroin users, respectively. The numbers of marijuana, kratom-plant, opium, and inhalant users were under-estimated because their symptoms were mild and not severe enough to remedy in health treatment centers which led to the smaller size of the total number of drug users. The well-estimated sizes of heroin and methamphetamine addicts are high reliable because they are based on clearly evident count with a severe addiction problem to health treatment centers. The estimation by means of mixture models can be recommended to monitor drug demand trend and drug health service routinely; it is easy to calculate via the available programs MIXTP based on request.

**Keywords:** Capture-Recapture Count Data; Drug Use in Thailand; Mixture Models of Zero-Truncated Poisson Distributions; Population Size Estimation; Unobserved Zero Count Data

## 1. Introduction

Drug abuse in Thailand has remained a serious health problem; its epidemic is still severe and widespread. Information on the number of illegal drug users is a benefit of the policy and the plan on narcotics control, to implement a reduction strategy, and to allocate resources to the health service. Nevertheless, it is difficult to measure the sizes of drug user populations directly because of many "hidden drug addicts". Surveys, especially on the large national scale, are unlikely to be the most efficient methods due to a huge cost and manpower, the difficulty of receiving a true response, the problems of dealing with a hidden population and ethical issues.

Capture-recapture methods are a classical and useful tool to solve a hidden population problem and to estimate a total population size because it can estimate and adjust for the extent of incomplete ascertainment using information from overlapping lists of cases from two or more distinct sources [1]. Moreover, there are not only the conventional multiple sources methods but also the ap-

[*]Corresponding author.

proaches available based upon one source with repeated counts for each individual. In this study, a single source is considered from a surveillance system counting the number of times that a drug user went to a treatment institution.

There were few studies in Thailand which used the capture-recapture method for estimating the number of drug users. Mastro *et al.* [2] estimated the number of HIV-infected injection drug users in Bangkok under two-sample sources of 18 methadone treatment centers and 72 urine testing police stations. Suppawattanabodee [3] used two sources of health treatment records and police arrestment records for estimating the number of drug users in Bangkok 2001. However, for one source with repeated count data, applications have been few relevant studies in Thailand. Böhning *et al.* [4] estimated the number of drug users in Bangkok 2001 by means of zero-truncated count mixture distributions. Viwatwongkasem, Kuhnert, and Satitvipawee [5] projected the number of heroin users in Bangkok 2002 using the mixture of zero-truncated Poisson models. Note that the zero-truncated Poisson mixture distributions are different from the mixture of zero-truncated Poisson models, at least the mixing distribution in both estimations.

A mixture model is a flexible approach to cope with long-tailed, skewed, and/or contaminated count distributions in a natural way. The mixing idea corresponds to a mixture representing the presence of sub-populations within an overall population. Formally, a mixture model can cope with not only two or more distributions (heterogeneity) but also includes the case of one distribution (homogeneous population) [6-8]. Böhning and Schön [9] proposed the nonparametric maximum likelihood estimators (NPMLE) of population size based on the counting distribution. Böhning and Kuhnert [10] showed the equivalence of the zero-truncated count mixture distributions and the mixture of zero-truncated count distributions. They stated that for any mixing distribution of the truncated mixture, a usually different mixing distribution of the mixture of truncated counts could be found so that the likelihood surfaces for both models agreed; consequently, for estimating population size, two estimators associated with two models had equal values. Punyacharoensin and Viwatwongkasem [11] predicted HIV incidence in Thailand utilizing the backcalculation of mixture of the past AIDS incidence and AIDS incubation period distributions. Viwatwongkasem, Kuhnert, and Satitvipawee [5] compared the performance of population size estimators under the truncated count model with and without allowance for contaminations among Mc-Kendrick's, Mantel-Haenszel's, Zelterman's, Chao's, the maximum likelihood, and their proposed methods of the mixture of zero-truncated count models. The proposed estimator provided the best choice according to its smallest bias and smallest mean square error for a situation of sufficiently large population sizes and it also performed well even for a homogeneous situation.

Although, the mixture model has been used previously in many fields of application, it is still not very common; only few relevant studies were found in Thailand and, in addition, the numerical computation of mixture model estimates has not been directly provided in the existing standard statistical packages. With the motivation of having at present few relevant studies and unavailable statistical packages with the option or focus on estimating the size of a hidden population, we take this opportunity to address the gap by adopting the nonparametric maximum likelihood estimators (NPMLE) for estimating the mixture parameters of zero-truncated Poisson distributions leading to the population size estimate of interest.

## 2. Methods

### 2.1. The Horvitz-Thompson Approach

Suppose that a registration system identifies $n$ observed cases, but not all cases of a population of size $N$, and the system can identify a case with probability $1-p_0$ where $p_0$ is probability of the unidentified cases. This leads to the expected equation of the population size, $N = Np_0 + N(1-p_0)$ where $N(1-p_0)$ is the expected number of cases identified by the system which simply can be estimated by $n$, number of identified (observed) cases. It leads to the estimating equation

$$N = Np_0 + n, \tag{1}$$

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 (1) can easily be solved for $N$ to provide the Horvitz-Thompson estimator

$$\hat{N} = n/(1-p_0) \tag{2}$$

and $\hat{n}_0 = \hat{N} - n$. The Horvitz-Thompson approach seems easy, but the unknown $p_0$ probability of unobserved cases must be estimated and this is quite differently accomplished in the various methods of estimation.

### 2.2. Data Sources

The surveillance data on the drug addicts undergoing treatment and rehabilitation in the country over 1140 health treatment centers (1144 centers in 2005 and 2006, 1258 centers in 2007) collected by the Bureau of Health Service System Development (BHSSD), Ministry of Public Health, were adopted during 2005 to 2007. Each anonymous record of treatment episodes in database was linked to the same patient with matching keywords, such as age, gender, date of birth, district and city of birth, present address, hospital number and name, date of re-

 

ceiving treatment episodes. This study was approved by the Ethics Committee on Human Rights of the Faculty of Public Health, Mahidol University, with the approved number 105/2011.

## 2.3. Statistical Methods

Suppose that $Y$ is the number of treatment episodes in a case; obviously, $Y$ has the values ranged from 1 to $m$ (without zero value) where $m$ is the largest number of treatment episodes in a case. Now data $Y$ are tallied into a frequency table like **Table 1**. We let $i$ be the number of treatment episodes in a case, $n_i$ be the number (frequency) of cases identified with $i$ episodes where $i = 1, 2, \cdots, m$ and a sample size $n = n_1 + n_2 + \cdots + n_m$ is the total number of observed cases. In **Table 1**, the observed frequencies of treatment episodes for heroin users in Thailand 2005 are $n_1 = 3057$, $n_2 = 791$, $n_3 = 351$, $n_4 = 107$, $n_5 = 80$, $n_6 = 59$, $n_{7+} = 22$.

To estimate the population size $N$ and the size of zero treatment episode $n_0$, we let $p_1, \cdots, p_m$ be probabilities of cases identified $1, \cdots, m$ times. Under homogeneity, the density function $p_i$ is assumed to be a zero-truncated Poisson since zero identification does not occur in the sample; that is,

$$p_i = f_+(i,\lambda) = \frac{f(i,\lambda)}{1 - f(0,\lambda)} = \frac{\exp(-\lambda)\lambda^i / i!}{(1 - \exp(-\lambda))}$$

where $i = 1, 2, \cdots$. However, frequently the homogeneous model is not appropriate in real situations to fit an adequate model. Mixture models allowing for heterogeneity are more flexible and we consider a discrete mixture of truncated Poisson densities of the form

$$p_i = f_+(i,Q) = \sum_{j=1}^{k} q_j f_+(i,\lambda_j) \tag{3}$$

where the mixing distribution $Q = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ q_1 & q_2 & \cdots & q_k \end{pmatrix}$ gives weights $q_j \geq 0$ to parameters $\lambda_j$ for $j = 1, 2, \cdots, k$, $k$ is the number of components in the mixture and $\sum_{j=1}^{k} q_j = 1$. Then, the log-likelihood for the mixture of zero-truncated count densities is

$$\begin{aligned} \log L(Q) &= \sum_{i=1}^{m} n_i \log f_+(i,Q) \\ &= \sum_{i=1}^{m} n_i \log\left[\sum_{j=1}^{k} q_j f_+(i,\lambda_j)\right] \end{aligned} \tag{4}$$

In this situation, with the help of gradient functions and the consideration at the boundaries of parameter space, the log-likelihood is concave on the parameter space of all discrete probability densities on which it can be maximized, leading to the nonparametric maximum likelihood estimator (NPMLE) of $Q$. To proceed in the EM context, we need the *complete data log likelihood*, which is given in this case as

$$\begin{aligned} \log L_{CD}(Q) = &\sum_{i=1}^{m} n_i \sum_{j=1}^{k} z_{ij} \log f_+(i,\lambda_j) \\ &+ \sum_{i=1}^{m} n_i \sum_{j=1}^{k} z_{ij} \log q_j \end{aligned} \tag{5}$$

where the unobserved covariate $z_{ij}$ is 1 if $i$ belongs to component $j$ and 0 otherwise. In the E-step, the unobserved indicator variates, $z_{ij}$, are replaced by their expected posterior probabilities, $e_{ij}$, leading to

$$\begin{aligned} e_{ij} &= E(z_{ij} | n_i; q_j, \lambda_j) \\ &= P(z_{ij} = 1 | n_i; q_j, \lambda_j) \\ &= \frac{f_+(i,\lambda_j) q_j}{\sum_{j'=1}^{k} f_+(i,\lambda'_j) q'_j} \end{aligned} \tag{6}$$

In the M-step, the new values $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$, $\hat{q}_1, \cdots, \hat{q}_k$ are found, which maximize the expected version of complete log likelihood (5). The results of the weighting estimates $\hat{q}_1, \cdots, \hat{q}_k$ are obtained by

$$\hat{q}_j = \frac{1}{n} \sum_{i=1}^{m} n_i e_{ij}, \text{ for } j = 1, \cdots, k \tag{7}$$

Similarly, the solution after solving the equations of derivatives with respect to $\hat{\lambda}_j$ is obtained by

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{m} i n_i e_{ij}}{\sum_{i=1}^{m} n_i e_{ij}} \left(1 - \exp\left(-\hat{\lambda}_j\right)\right), \text{ for } j = 1, \cdots, k \tag{8}$$

Note that (8) does not provide a close form solution; the iterative procedure is needed until the desired accuracy is achieved. Having identified the model and the associated parameter estimates, we can estimate the probability of zero treatment episodes $p_0$ as

$$\hat{p}_0 = \sum_{j=1}^{k} \exp\left(-\hat{\lambda}_j\right) \hat{q}_j \tag{9}$$

so that the Horvitz-Thompson approach leads to a population size estimate

$$\hat{N} = n\left(\sum_{j=1}^{k} \frac{\hat{q}_j}{1 - \exp\left(-\hat{\lambda}_j\right)}\right) = \frac{n}{1 - \sum_{j=1}^{k} \exp\left(-\hat{\lambda}_j\right) \hat{q}_j} \tag{10}$$

## 2.4. Model Evaluation

It is crucial to select an appropriate model among various potential models differing in the number of components $k$. The smallest value of the Bayesian Information Criterion (BIC) is considered to choose the best model: the smaller

 

BIC-value, the better model.

$$BIC = -2\log L\left(\hat{Q}_k\right) + \left(2k-1\right)\log\left(n\right) \qquad (11)$$

The BIC adjusts the log-likelihood with the number of parameters $\left(2k-1\right)$ multiplied by the log-sample size; BIC works well as model selection criterion in mixture model since it does not suffer under likelihood irregularities that are typical for mixture models [8,12].

## 3. An Application

For the surveillance data of heroin users 2005 in **Table 1**, the observed frequencies were $n_1$ = 3057, $n_2$ = 791, $n_3$ = 351, $n_4$ = 107, $n_5$ = 80, $n_6$ = 59, $n_{7+}$ = 22. **Table 2** showed that the mixture of two-components of zero-truncated Poisson model was the best fitting with the smallest BIC value. The results produced $\hat{n}_0 = 7092$ for the unobserved number of heroin users without any treatment episodes and $\hat{N} = 11{,}559$ for the total number of heroin users whereas a well-established alternative Chao's [13] estimator $\hat{N}_{Chao} = n + n_1^2/\left(2n_2\right)$ yielded $\hat{N}_{Chao} = 10{,}374$ which was close to the appropriate NPMLE model with $k$ = 2 components in the mixture. Likewise, **Figure 1** compared frequency distributions of treatment episodes among the observed frequencies, single Poisson with zero-truncation, and the mixture of zero-truncated Poisson with two components. The Poisson mixture provided an excellent goodness-of-fit to the observed frequencies whereas the simple Poisson was not adequate; it was clearly evident with the smallest BIC value.

### Confidence Intervals

*Bootstrap resampling technique* was applied to compute the variance of mixtures of truncated count data since the direct computation via the information matrix was usually difficult. For each nonparametric bootstrap, frequencies $n_1^*, n_2^*, \cdots, n_m^*$ were sampled from a multinomial distribution with size parameter $n$ and categorical probability parameters $n_i/n$; then $\hat{N}^*$ was constructed using the BIC-selected mixture model. Suppose that there were $B$ samples of size $n$ each, population size estimates $\hat{N}_1^*, \cdots, \hat{N}_B^*$ were available. These resampled data were used to compute variances and confidence intervals as asymptotic normal intervals.

The appropriate NPMLE $\hat{Q}$ for $k$ = 2 components of mixture yielded $\hat{N} = 11{,}559$. The associated 95% confidence interval for heroin users in Thailand 2005 was established and lied between (10,234, 12,884). As a general trend, the estimated size $\hat{N}$ was about 3 times higher than the observed data $\left(n\right)$. Also, it was usually important to provide an estimate of *completeness* (of the surveillance stream) given as $n/\hat{N} \times 100\%$, which was for heroin users 38.6% (95% CI: 34.7% - 43.6%).

## 4. Results

Thailand Narcotics Annual Report 2006 of the Office of

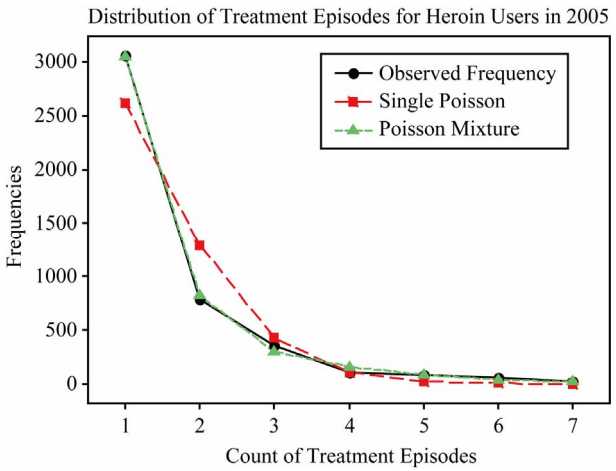


**Figure 1. Frequency distribution of treatment episodes among the observed counts, single Poisson, and mixture of two truncated Poisson.**

**Table 1. Observed frequencies of treatment episodes of heroin users in Thailand 2005: $n_1$ = 3057, $n_2$ = 791, $n_3$ = 351, $n_4$ = 107, $n_5$ = 80, $n_6$ = 59, $n_{7+}$ = 22.**

| Number of treatment episodes in a case ($i$) | 0 | 1 | 2 | … | $m$ = 7+ | Total |
|---|---|---|---|---|---|---|
| Number of cases (Frequency $n_i$) | - | 3057 | 791 | … | 22 | $n$ = 4467 |

**Table 2. The $k$-components mixture of zero-truncated Poisson models for estimating the size of heroin users in Thailand 2005.**

| $k$ | $\hat{\lambda}_j$ | $\hat{q}_j$ | log-likelihood | BIC | $\hat{n}_0$ | $\hat{N}$ |
|---|---|---|---|---|---|---|
| 1 | 0.9869 | 1.000 | −4893 | 9796 | 2654 | 7121 |
| **1**<br>**2** | **0.4094**<br>**2.7772** | **0.7965**<br>**0.2035** | **−4543** | **9111** | **7092** | **11,559** |
| 1<br>2<br>3 | 0.0419<br>0.8546<br>3.2260 | 0.3346<br>0.5361<br>0.1293 | −4538 | 9119 | 36,757 | 41,224 |

 

the Narcotics Control Board [14] showed that the number of drug addicts undergoing treatment and rehabilitation had increased from 41,564 patients in year 2005 to 43,156 patients in 2006. Most of them were the drug patients who underwent treatment for the first time (about 80%). Adolescents aged between 15 and 24 years old were the biggest group (49%). Among them, 83% were new drug patients while 17% were the relapsing patients. 33% of the total drug patients were unemployed while 19% were laborers and 13% were students. Drug epidemics were mostly found in Bangkok (50%) and 30% were located in the central region of Thailand while the rest were in the North, the South, and the Northeast, respectively. Methamphetamine (Yaba) addicts were still the biggest group of drug patients in all treatment centers (79%) because the ingredients were not hardly available and the price was not too high in comparing purity and severity. The second biggest group was cannabis (marijuana) addicts (11%); most marijuana was spread in many urban and rural areas; however, the price of marijuana was still cheaper compared with other illicit drugs. Heroin epidemic was still important though it has a high price but the injuries were quite severe. Kratom plants were abused in many areas of the country side; farmers and peasants used kratom plants for working in the rice fields to work longer. The number of club drugs like ecstasy, ketamine, cocaine, and crystallized methamphetamine had an increasing trend in big cities and the rich persons.

The MIXTP program developed by authors was available to achieve the estimates of population size under the mixture of truncated count models via FORTRAN POWERSTATION and now it is available on the request.

**Table 3** illustrated the sizes of drug users, estimated by the mixture of truncated Poisson models and classified by types of drugs in Thailand 2005-2007. Methamphetamine users were the biggest group and tended to increase from 2005-2007 while heroin users trended to decrease slightly because of its high price. The numbers of marijuana, kratom-plant, and inhalant users were under-estimated because of their mild severities. Trend of marijuana sizes increased from 2005-2007 while trends of kratom-plant, inhalant, and others were difficult to predict.

## 5. A Simulation Study

Although data fitting of mixture of truncated count model was well in the examples, we wish to ensure this in general case via the simulation experiment. Let count variables $Y_i$ be generated from a two-component mixture of Poisson distributions with equal weights attached to the component means $\lambda_1 = 1$ and $\lambda_2$ where $\lambda_2 = 1, 2, \cdots, 5$. That is, $Y_i \sim 0.5Po(1) + 0.5Po(\lambda_2)$ where $i = 1, 2, \cdots, N$. Population sizes $N$ were 200 (for small), 1000 (for medium), 5000 (for large), 10,000 (very large). Each simulated datum $Y_i$ was tallied to get frequencies $n_0, n_1, \cdots, n_m$ with respect to the counts $0, 1, \cdots, m$ where $n_0 + n_1 + \cdots + n_m = N$. Then $n_0$ was dropped and zero-truncated frequencies $n_1, \cdots, n_m$ were used to compute population size estimators of mixture model and Chao. This was done under 5000 replications; mean, standard deviation (SD), and root of mean square error (RMSE) of all estimates were computed and determined from these replications.

The results are found in **Table 4** and we can conclude in the following:

- Under homogeneity $(\lambda_2 = 1)$, mixture model estimator with 1 component in this case performs well with smaller RMSE, regardless of population size; Chao's estimator is worse with larger RMSE under this homogeneity.
- Under heterogeneity $(\lambda_2 > 1)$, Chao's estimator performs better when population size is small to moderate ($N$ = 200, 1000); mixture model estimator is better when population is large to very large ($N \geq 5000$) and degrees of heterogeneity are strong $(\lambda_2 = 4, 5)$, at least $\lambda_2 = 3$.
- Furthermore, we found that if the weak degrees of heterogeneity occur $(\lambda_2 = 2, 3)$ in combination with small to moderate population size ($N$ = 200, 1000), mixture model estimator has a problem of the large excess values of standard deviation.

## 6. Discussion

The NPMLE method provides well-estimated sizes of various drug-user target populations, obtained from the surveillance data on the drug addicts with emphasis on methamphetamine (228,898 cases in 2007) and heroin (8953 cases in 2007) users. It can be expected that these surveillance data provide a high reliability because they are based on clearly evident contact counts of drug addicts with a severe addiction problem to health treatment centers. A comparison by means of a national household survey of ONCB [15] yielded the under-estimated sizes of 66,320 methamphetamine users and 3907 heroin users per year.

In contrast, the estimated sizes of this study using NPMLE of users with kratom-plant (less than 18,720 cases in 2007), marijuana (27,323 cases in 2007), and inhalant (13,362 cases in 2007) are frequently under-estimated because of their low severity of symptoms to cure, leading to the smaller size of total number of drug users (299,670 cases in 2007). This fact is confirmed by a national household survey of the ONCB [15] that reported an estimate of 378,214 kratom-plant users, 57,527

 

**Table 3. Estimating number of drug users classified by type of drugs in Thailand 2005-2007.**

| $n_i$ | Types of illegal drug users 2007 | | | | | | |
|---|---|---|---|---|---|---|---|
| | All type | Methamphetamine | Heroin | Marijuana | Opium | Inhalants | Others |
| 1 | 66,991 | 50,485 | 2299 | 6054 | 2200 | 3027 | 2926 |
| 2 | 10,931 | 7925 | 477 | 991 | 703 | 477 | 358 |
| 3 | 1716 | 1042 | 219 | 76 | 197 | 78 | 104 |
| 4 | 439 | 154 | 121 | 31 | 76 | 25 | 32 |
| 5 | 174 | 54 | 34 | 17 | 50 | - | 19 |
| 6 | 86 | 6 | 47 | - | 33 | - | - |
| 7 | 22 | 1 | 10 | - | 3 | 8 | - |
| 8 | 28 | 10 | 18 | - | - | - | - |
| 9 | 23 | - | 20 | - | - | - | 3 |
| 10 | 23 | 11 | 9 | 1 | - | - | 2 |
| 14 | 14 | 14 | - | - | - | - | - |
| $n$ | 80,447 | 59,702 | 3254 | 7170 | 3262 | 3615 | 3444 |
| $k$ | 3 | 3 | 2 | 2 | 2 | 2 | 2 |
| $\hat{N}$ | 299,670 | 228,898 | 8953 | 27,323 | 7193 | 13,362 | 18,720 |
| $\hat{N}_{Chao}$ | 285,725 | 220,505 | 8794 | 25,662 | 6704 | 13,220 | 15,401 |
| $SD(\hat{N})$ | 2782.3 | 2108.0 | 548.6 | 731.4 | 264.9 | 1597.2 | 1859.8 |
| 95% lower | 294,217 | 224,766 | 7878 | 25,889 | 6674 | 10,231 | 15,075 |
| 95% upper | 305,123 | 233,030 | 10,028 | 28,757 | 7712 | 16,493 | 22,365 |
| $n_i$ | Types of illegal drug users 2006 | | | | | | |
| | All type | Methamphetamine | Heroin | Marijuana | Opium | Inhalants | Others |
| 1 | 58,578 | 42,372 | 2690 | 5445 | 2418 | 2924 | 2729 |
| 2 | 11,239 | 7897 | 775 | 1025 | 559 | 578 | 405 |
| 3 | 2018 | 1240 | 320 | 158 | 98 | 101 | 101 |
| 4 | 439 | 182 | 157 | 21 | 35 | 28 | 16 |
| 5 | 138 | 61 | 27 | 1 | 31 | 5 | 13 |
| 6 | 132 | 29 | 49 | - | 33 | 4 | 17 |
| 7 | 14 | 14 | - | - | - | - | - |
| 8 | - | - | - | - | - | - | - |
| 9 | 9 | 9 | - | - | - | - | - |
| 13 | 14 | 14 | - | - | - | - | - |
| 14 | 29 | 29 | - | - | - | - | - |
| $n$ | 72,610 | 51,847 | 4018 | 6650 | 3174 | 3640 | 3281 |
| $k$ | 3 | 3 | 2 | 1 | 2 | 2 | 2 |
| $\hat{N}$ | 230,665 | 166,287 | 11,333 | 20,283 | 8973 | 11,773 | 13,186 |
| $\hat{N}_{Chao}$ | 225,265 | 165,522 | 8686 | 21,112 | 8404 | 11,036 | 12,475 |
| $SD(\hat{N})$ | 2068.2 | 1563.7 | 1049.5 | 421.8 | 306.9 | 768.8 | 1184.4 |
| 95% lower | 226,611 | 163,222 | 9276 | 19,456 | 8371 | 10,266 | 10,865 |
| 95% upper | 234,719 | 169,352 | 13,390 | 21,110 | 9575 | 13,280 | 15,507 |

 

**Continued**

| $n_i$ | Types of illegal drug users 2005 | | | | | | |
|---|---|---|---|---|---|---|---|
| | All type | Methamphetamine | Heroin | Marijuana | Opium | Inhalants | Others |
| 1 | 50,370 | 35,422 | 3057 | 4073 | 2478 | 2446 | 2894 |
| 2 | 10,760 | 7785 | 791 | 786 | 520 | 393 | 485 |
| 3 | 1959 | 1262 | 351 | 97 | 97 | 53 | 99 |
| 4 | 299 | 111 | 107 | 6 | 21 | 28 | 26 |
| 5 | 202 | 68 | 80 | 16 | 7 | 10 | 21 |
| 6 | 135 | 10 | 59 | 3 | 55 | - | 8 |
| 7 | 27 | 13 | - | 1 | - | 13 | - |
| 8 | 8 | 8 | - | - | - | - | - |
| 9 | 16 | - | 15 | - | - | - | 1 |
| 12 | 19 | - | 7 | - | - | - | 12 |
| $n$ | 63,795 | 44,679 | 4467 | 4982 | 3178 | 2943 | 3546 |
| $k$ | 4 | 2 | 2 | 2 | 2 | 2 | 3 |
| $\hat{N}$ | 184,045 | 125,117 | 11,559 | 16,041 | 9297 | 10,989 | 13,201 |
| $\hat{N}_{Chao}$ | 181,692 | 125,265 | 10,374 | 15,535 | 9082 | 10,555 | 12,180 |
| $SD(\hat{N})$ | 1401.8 | 1033.0 | 676.0 | 418.9 | 311.2 | 562.1 | 571.6 |
| 95% lower | 181,297 | 123,092 | 10,234 | 15,220 | 8687 | 9887 | 12,081 |
| 95% upper | 186,793 | 127,142 | 12,884 | 16,862 | 9907 | 12,091 | 14,321 |

**Table 4. Mean, SD, RMSE of population size estimators of Chao and mixture model.**

| $N$ | $\lambda_2$ | Chao | | | Mixture | | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | RMSE | Mean | SD | RMSE |
| 10000 | 1 | 10000.4 | 159.3 | 159.3 | 999.4 | 117.7 | **117.7** |
| | 2 | 9727.2 | 105.6 | **292.6** | 9983.4 | 387.4 | 387.8 |
| | 3 | 9551.4 | 86.9 | 456.9 | 9997.6 | 185.8 | **185.9** |
| | 4 | 9541.1 | 84.3 | 466.6 | 10002.4 | 143.4 | **143.5** |
| | 5 | 9630.2 | 90.3 | 380.7 | 10001.5 | 122.1 | **122.1** |
| 5000 | 1 | 5000.8 | 113.6 | 113.6 | 4999.5 | 83.9 | **83.9** |
| | 2 | 4865.5 | 74.9 | **153.9** | 5082.4 | 678.5 | 683.5 |
| | 3 | 4776.8 | 60.9 | 231.4 | 5006.6 | 133.4 | **133.6** |
| | 4 | 4771.7 | 59.6 | 235.4 | 5005.0 | 103.1 | **103.3** |
| | 5 | 4816.8 | 63.4 | 193.8 | 5003.6 | 87.1 | **87.2** |
| 1000 | 1 | 1002.6 | 50.7 | 50.8 | 1001.4 | 37.4 | **37.4** |
| | 2 | 974.0 | 33.3 | **42.3** | 1221.8 | 743.1 | 775.5 |
| | 3 | 956.1 | 27.6 | **51.8** | 1051.2 | 293.4 | 297.8 |
| | 4 | 955.0 | 27.2 | **52.6** | 1008.8 | 54.8 | 55.5 |
| | 5 | 964.3 | 28.9 | 45.9 | 1004.7 | 41.5 | **41.8** |
| 200 | 1 | 203.0 | 23.9 | 24.1 | 201.3 | 17.0 | **17.1** |
| | 2 | 195.7 | 15.5 | **16.1** | 348.3 | 436.1 | 460.6 |
| | 3 | 191.9 | 12.8 | **15.2** | 351.3 | 448.3 | 473.2 |
| | 4 | 191.8 | 12.5 | **14.9** | 258.4 | 284.7 | 290.6 |
| | 5 | 194.0 | 13.7 | **15.0** | 213.1 | 134.2 | 134.8 |

 

marijuana users, and 48,849 inhalant users in year 2007, leading to 575,312 cases for total number of drug users.

The huge difference in values between two methods mentioned, stem mainly from the severity of symptoms of drug use. With this point of view, the estimated sizes of methamphetamine and heroin users from the surveillance data of this NPMLE study seem to be more useful than those from the national survey, in particular, if viewed from the perspective of a benefit of allocating resources on health service, monitoring drug epidemics, and planning policy on narcotics control. In general, the estimated sizes from this study are at least three times higher than the observed data. Hence, the completeness of identification is about 30% - 40%.

Due to the result of simulation that mixture model estimator behaves well when population size is large, there is no reason to reject the use of mixture model to estimate the hidden population size and the total population size for each type of drugs since the observed total number of drug users is large enough. However, there is something called a boundary problem: extremely large observations in some samples. This could explain the overestimation effect seen in the simulation for $N = 200$. Kuhnert *et al.* [16] used the median for a series of estimates of population size in their simulation to avoid highly influential size estimates. Basically, the mixture model is a flexible approach to cope with homogeneity and heterogeneity, including long-tailed, skewed, and/or contaminated distributions in a natural way.

Recently, there has been an increased interest in zero-truncated count models. These models can be applied in many areas such as illegal immigrants, illegal gun owners, HIV epidemic, scrapie disease on sheep, or criminal persons. This article has shown how the mixture models allowing for heterogeneity can be applied to estimate the unobserved population size of drug users with zero treatment episodes and then estimate the total population size of illegal drug addicts. Indeed, there are not only the estimators available based upon mixture models but also there are the Mantel-Haenszel's [17], Zelterman's [18], Chao's [13], and maximum likelihood methods available in estimating population sizes. Viwatwongkasem, Kuhnert, and Satitvipawee [5] found that the mixture of zero-truncated count model and Chao's model provided the best choice among the above estimators, according to its smallest bias and smallest mean square error, especially for a situation of sufficiently large population sizes; furthermore, the mixture itself also performed well even for a homogeneous situation. Although the mixture model provides a nice estimate, its variance estimate is usually difficult to find. *Bootstrap resampling technique* was applied to compute the variance of mixture of truncated count data, instead of the direct computation via the information matrix. Further study should focus on the estimation of the variance of mixture models. But this is a challenging task as stated by Chao [19], Cormack [20], and Böhning and Schön [9]. Other parametric models such as the binomial model, the hypergeometric model, and the inverse sampling of the negative binomial model should be considered in any future research.

Fortunately, the appropriate NPMLE models for these surveillance data of drug users in 2005-2007 do not face a spurious value of overestimation. However, in few occurrences, the NPMLE of mixture may provide an overestimation. The occurrence of overestimates is due to the boundary problem of an estimate which is evaluated at the boundary of parameter space. The improvement in reducing overestimation bias should be investigated in any further study.

## 7. Acknowledgements

We are grateful to Kanya Boonthongtho, our M.Sc. (Biostatistics) student as well as the Bureau of Health Service System Development (BHSSD), Ministry of Public Health, for providing the surveillance dataset. We would like to thank the referees and the editors for comments which greatly improved this paper. This study was partially supported for publication by the China Medical Board (CMB), Faculty of Public Health, Mahidol University, Bangkok, Thailand.

## REFERENCES

[1] E. B. Hook and R. R. Regal, "Capture-Recapture Methods in Epidemiology: Methods and Limitations," *Epidemiologic Reviews*, Vol. 17, No. 2, 1995, pp. 243-264.

[2] T. D. Mastro, D. Kitayaporn, B. G. Weniger, *et al.*, "Estimating the Number of HIV-Infected Injection Drug Users in Bangkok: A Capture-Recapture Method," *American Journal of Public Health*, Vol. 84, No. 7, 1994, pp. 1094-1099. doi:10.2105/AJPH.84.7.1094

[3] B. Suppawattanabodee, "Estimating the Number of Drug Users in Bangkok: A Capture-Recapture Method," Master of Sciences Thesis, Mahidol University, Bangkok, 2003.

[4] D. Böhning, B. Suppawattanabodee, W. Kusolvisitkul, and C. Viwatwongkasem, "Estimating the Number of Drug Users in Bangkok 2001: A Capture-Recapture Approach Using Repeated Entries in One List," *European Journal of Epidemiology*, Vol. 19, No. 12, 2004, pp. 1075-1083. doi:10.1007/s10654-004-3006-8

[5] C. Viwatwongkasem, R. Kuhnert and P. Satitvipawee, "A Comparison of Population Size Estimators under the Truncated Count Model with and without Allowance for Contaminations," *Biometrical Journal*, Vol. 50, No. 6, 2008, pp. 1006-1021. doi:10.1002/bimj.200810484

[6] D. Böhning, "Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others," Chapman & Hall/CRC, Boca Raton, 2000.

[7] B. G. Lindsay, "The Geometry of Mixture Likelihoods

 

Part I: A General Theory," *Annals of statistics*, Vol. 11, No. 3, 1983, pp. 783-792. doi:10.1214/aos/1176346245

[8] G. McLachlan and D. Peel, "Finite Mixture Models," Wiley, New York, 2000. doi:10.1002/0471721182

[9] D. Böhning and D. Schön, "Nonparametric Maximum Likelihood Estimation of Population Size Based on the Counting Distribution," *Journal of the Royal Statistical Society*: *Series C* (*Applied Statistics*), Vol. 54, No. 4, 2005, pp. 721-737. doi:10.1111/j.1467-9876.2005.05324.x

[10] D. Böhning and R. Kuhnert, "Equivalence of Truncated Count Mixture Distributions and Mixture of Truncated Count Distributions," *Biometrics*, Vol. 62, No. 4, 2006, pp. 1207-1215. doi:10.1111/j.1541-0420.2006.00565.x

[11] N. Punyacharoensin and C. Viwatwongkasem, "Trends in Three Decades of HIV/AIDS Epidemic in Thailand by Nonparametric Backcalculation Method," *AIDS*, Vol. 23, No. 9, 2009, pp. 1143-1152. doi:10.1097/QAD.0b013e32832baa1c

[12] N. M. Laird, "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, Vol. 73, No. 364, 1978, pp. 805-811. doi:10.1080/01621459.1978.10480103

[13] A. Chao, "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability," *Biometrics*, Vol. 43, No. 4, 1987, pp. 783-791. doi:10.2307/2531532

[14] Office of the Narcotics Control Board (ONCB), "Thailand Narcotics Annual Report," Aroon Printing Co., Ltd., Bangkok, 2006.

[15] Office of the Narcotics Control Board (ONCB), Academic Network Organization Board on Substance Abuse, "Estimation of the Number of Drug Addicts in Thailand 2007," Union Ultra Violet Co., Ltd., Bangkok, 2007.

[16] R. Kuhnert, V. J. Del Rio Vilas, J. Gallagher and D. Böhning, "A Bagging-Based Correction for the Mixture Model Estimator of Population Size," *Biometrical Journal*, Vol. 50, No. 6, 2008, pp. 993-1005. doi:10.1002/bimj.200810485

[17] N. Wannasirikul, "A Comparison of Truncated Poisson Estimators of Population Size under Model Contaminations," Master of Sciences Thesis, Mahidol University, Bangkok, 2005.

[18] D. Zelterman, "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments," *Journal of Statistical Planning and Inference*, Vol. 18, No. 2, 1988, pp. 225-237. doi:10.1016/0378-3758(88)90007-9

[19] A. Chao, "Estimating Population Size for Sparse Data in Capture-Recapture Experiments," *Biometrics*, Vol. 45, No. 2, 1989, pp. 427-438. doi:10.2307/2531487

[20] R. M. Cormack, "Interval Estimation for Mark-Recapture Studies of Closed Populations," *Biometrics*, Vol. 48, No. 2, 1992, pp. 567-576. doi:10.2307/2532310

 