Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 86 (2016) 216 – 219

Procedia Computer Science

2016 International Electrical Engineering Congress, iEECON2016, 2-4 March 2016, Chiang Mai, Thailand

Bootstrapping with R to determine variances of mixture model estimates in predicting confidence intervals for population sizes

Chareena Ujeh[a], Pratana Satitvipawee[a], Jutatip Sillabutra[a], Pichitpong Soontornpipit[a], Prasong Kitidamrongsuk[a], Chukiat Viwatwongkasem[a] *

[a]*Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok 10400, Thailand*

Abstract

It is not easy to find the variance estimates of mixture model via the theoretical derivation directly. Instead, bootstrapping denoting a resampling technique from an original sample dataset with replacement allocation is used to calculate variances of mixture estimates of zero-truncated Poisson distributions in a prediction of population size and its confidence interval. The application is the estimation of the number of drug (opium) users in Thailand 2007 under surveillance data of counts of treatment episodes in a case. The results indicated that there were 3,262 observed opium cases who received treatments, the estimate of the unobserved number of opium users without receiving any treatment was 3,931, leading to total population size estimate of 7,193 opium users. The 95% confidence intervals as a by-product of bootstrapping were 6,674-7,712 under bootstrap normal base, 6,782-7,761 under bootstrap 95% percentiles, and 6,626-7,605 under bootstrap *t* intervals. Bootstrapping algorithm with R program is available here.

Peer-review under responsibility of the Organizing Committee of iEECON2016

Keywords: Bootstrapping with R; Variance Estimation; Mixture Model; Population Size Estimation

1. Introduction

Bootstrapping is a resampling technique with replacement allocation from an original sample dataset in order to estimate the precision of sample statistics, to perform significance tests, and to validate statistical models when theoretical parts are complex and/or the sample size is not met enough conditions in practice. Mixture model

* Corresponding author. Tel.: +66-2-354-8530; fax: +66-2-354-8534.
E-mail address: chukiat.viw@mahidol.ac.th.

Peer-review under responsibility of the Organizing Committee of iEECON2016
doi:10.1016/j.procs.2016.05.096

corresponding to mixture distribution is also based on a complex procedure that theoretical inference is difficult to get the result directly, especially in finding variance estimates. Bootstrap method initiated by Efron in 1979 [1] is placed under pressure of this situation.

2. Methodology and design

The formulas, notations, and data are followed from the work of Viwatwongkasem et al. [2, 3]. Let n_i be the number of drug (opium) users reported with i treatment episodes. The data of opium cases are $n_1 = 2200$, $n_2 = 703$, $n_3 = 197$, $n_4 = 76$, $n_5 = 50$, $n_6 = 33$, $n_7 = 3$. In total, the sample size, $n = \sum_{i=1}^{m} n_m = 3262$, for opium use are observed. Suppose that a mixture of zero-truncated Poisson densities is given as $f_+(i,Q) = \sum_{j=1}^{k} q_j f_+(i,\lambda_j)$ where $f_+(i,\lambda_j) = Po(i,\lambda_j)/(1 - Po(0,\lambda_j))$, $q_1,...,q_k$ and $\lambda_1,...,\lambda_k$ are found from the k -component. The Horvitz-Thompson estimator for estimating the total population size of opium use and the number of opium users without receiving any treatment (zero episode) can be estimated as

$$\hat{N} = n\left(\sum_{j=1}^{k} \frac{\hat{q}_j}{1-\exp(-\hat{\lambda}_j)}\right) \qquad \text{and} \qquad \hat{n}_0 = \hat{N} - n$$

where $\hat{q}_j = \frac{1}{n}\sum_{i=1}^{m} n_i e_{ij}$, $\hat{\lambda}_j = \frac{\sum_{i=1}^{m} i\, n_i e_{ij}}{\sum_{i=1}^{m} n_i e_{ij}}(1-\exp(-\hat{\lambda}_j))$, and $e_{ij} = \frac{f_+(i,\lambda_j)q_j}{\sum_{j'=1}^{k} f_+(i,\lambda'_j)q'_j}$.

EM algorithm for mixtures of zero-truncated Poisson distributions

Step 0 Choose some initial estimates $\hat{q}_j^{(r)}$ and $\hat{\lambda}_j^{(r)}$ at cycle r

Step 1 Compute $\hat{N}^{(r)} = n\left(\sum_{j=1}^{k} \frac{\hat{q}_j^{(r)}}{1-\exp(-\hat{\lambda}_j^{(r)})}\right)$, $\hat{n}_0^{(r)} = \hat{N}^{(r)} - n$, and $e_{ij}^{(r)} = \frac{f_+(i,\hat{\lambda}_j^{(r)})\hat{q}_j^{(r)}}{\sum_{j'=1}^{k} f_+(i,\hat{\lambda}_j'^{(r)})\hat{q}_j'^{(r)}}$.

Step 2 Compute new estimates as $\hat{q}_j^{(r+1)} = \frac{1}{n}\sum_{i=1}^{m} n_i e_{ij}^{(r)}$, $\hat{\lambda}_j^{(r+1)} = \frac{\sum_{i=1}^{m} i\, n_i e_{ij}^{(r)}}{\sum_{i=1}^{m} n_i e_{ij}^{(r)}}(1-\exp(-\hat{\lambda}_j^{(r)}))$

Step 3 Set $r = r+1$ and go back to Step 1. The step 1 and 2 are repeated until convergence.

Bootstrap variance to determine population size

1. Bootstrap frequencies $n_1^*, n_2^*, ..., n_m^*$ are sampled from a multinomial distribution with size parameter n and with nonparametric probability parameter $p_i = n_i / n$
2. For each bootstrap sample $\{n_1^*, n_2^*, ..., n_m^*\}$ of size n , $\hat{N}^*$ is constructed from above EM algorithm
3. If there are R bootstrap samples, then population size estimates $\hat{N}_1^*, \hat{N}_2^*, ..., \hat{N}_R^*$ are available, the bootstrap mean and variance are $\bar{N}^* = \frac{1}{R}\sum_{b=1}^{R} \hat{N}_b^*$, $\hat{\sigma}^2 = \frac{1}{(R-1)}\sum_{b=1}^{R} (\hat{N}_b^* - \bar{N}^*)^2$. Also, the standard error $\hat{\sigma}$ is used to obtain bootstrap intervals.
4. Bootstrap 95% confidence intervals based on normal is $\hat{N} \pm 1.96\hat{\sigma}$, $\left(\hat{N}_{(0.025)}^*, \hat{N}_{(0.975)}^*\right)$ under bootstrap percentile, and $\left(\hat{N} + t_{0.025}\hat{\sigma}, \hat{N} + t_{0.975}\hat{\sigma}\right)$ under bootstrap t_α interval where t_α is the α^{th} ordered value of the statistics $t = (\bar{N}^* - \hat{N})/\hat{\sigma}$.