JRHS 2010; 10(2): 110-115

JRHS

Journal of Research in Health Sciences

Original Article

# Application of Mixture Models for Estimating the Prevalence of Cigarette Smoking in Hamadan, Iran

**Shahrbanoo Goli (MSc)[a], Hossein Mahjub (PhD)[b]*, Abbas Moghimbeigi (PhD)[b], Jalal Poorolajal (PhD)[b], Ahmad Heidari Pahlavian (PhD)[c]**

[a] *Department of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[b] *Department of Biostatistics & Epidemiology and Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

[c] *Research Center for Behavior Sciences, Department of Psychology, School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran*

## ARTICLE INFORMATION

***Article history:***

***Received:*** *25 July 2010*

***Revised:*** *25 October 2010*

***Accepted:*** *9 November 2010*

***Available online:*** *24 November 2010*

***Keywords:***

*Smoking*

*Prevalence*

*Finite mixture model*

*Truncated counts distribution*

*Simulation*

**** Correspondence***

*Hossein Mahjub (PhD)*

***Tel:*** *+98 811 8260661*

***E-mail:*** *mahjub@umsha.ac.ir*

## ABSTRACT

**Background**: So far, several studies were conducted to estimate the prevalence of cigarette smoking in Iran, but none of them used a statistical model to deal with unobserved smokers. The present study planned to estimate the accurate prevalence of cigarette smoking using mixture of truncated Poisson distribution.

**Methods**: A cross-sectional study was conducted in Hamadan, west of Iran in 2009, using cluster sampling and 1146 men and women aged ≥18 years were enrolled. The data collection was done by an expert group of psychologists and sociologists. A truncated mixture Poisson distribution was fitted to the daily number of cigarettes smoked by smokers. The number of components of the mixture model and related mean and weight were specified using Bayesian information criteria. Accordingly, the number of cigarette smokers who answered incorrectly to the relevant question was estimated. To investigate the validity of the results, a simulation study was conducted using CAMCR software.

**Results**: Mixture Poisson distribution with four components was the most appropriate model fitted to the count data. After correction for underestimation, the prevalence rate of cigarette smoking in the population was 20.6%, including 36.2% for men and 3.3% for women. According to the simulation study, the bias of estimated prevalence was about zero and the root mean square error was estimated 2.5.

**Conclusion**: The number of unobserved data can be estimated by fitting model to truncated count data. The mixture of truncated Poisson distribution is particularly useful to estimate population size when the main objective of the study is to investigate negative traits to which the participants may answer incorrectly.

**Citation:** Goli S, Mahjub H, Moghimbeigi A, Poorolajal J, Heidari Pahlavian A. Application of mixture models for estimating the prevalence of cigarette smoking in Hamadan, Iran. *JRHS*. 2010;10(2):110-115.

## Introduction

Smoking is the first preventable leading cause of death worldwide with over five million deaths annually [1]. A five-year follow-up study indicated that life expectancy in smokers is 10 years less than nonsmokers [2]. In

Iran, the prevalence of cigarette smoking in people aged ≥15 years is estimated about 11.9% with 13.2 cigarettes per day on average. It is estimated that about 429 million dollars are annually expended for cigarette smoking in the country [3]. Therefore, accurate estimation of the number of smokers is essential for planning health policy. Several studies were conducted to estimate the prevalence of cigarette smoking in Iran [1-8]. However, the number of unobserved smokers was investigated in none of them, although underestimation due to incorrect answer is probable when estimating a negative trait in the general population [9-12]. Poorolajal et al. used a capture-recapture approach to estimate publication bias in meta-analysis. They retrieved studies from three separate sources of electronic databases, reference lists of the studies, and conference databases as well as contact with experts and manufacturers. Capture-recapture method approach was used to estimate the number of missing studies which were detected by none of the sources [13]. In addition, Poisson mixture distribution was used for disease mapping in Iran [14], but not for smoking in Iran.

To estimate the accurate number of people with a negative trait which is usually underestimated, several statistical methods were suggested [10-12, 15, 16]. In the present study, we used mixture of truncated Poisson distribution to specify the number of unobserved smokers and to estimate the accurate prevalence of cigarette smoking in Hamadan County, west of Iran 2009.

## Materials and Methods

A cross-sectional study was conducted in Hamadan County in 2009 using cluster (region) sampling to estimate the prevalence of cigarette smoking among general population. For this purpose, 603 men and 543 women aged ≥18 years were enrolled. The data collection was done by an expert group of psychologists and sociologists using a predefined questionnaire.

The average number of cigarettes consumed per day was considered as count data. The probability of incorrect answer to the question related to smoking was indicated by $P_0$. Accordingly, the total number of cigarette smokers based on Horvitz Thompson estimator can be achieved by the following formula [16, 17]:

$$\hat{N} = \frac{n_{obs}}{1 - p_0} \quad (1)$$

where $n_{obs}$ indicates the number of observed cigarette smokers.

If we consider variable Y as the number of cigarettes consumed daily for smokers, Y = 0 would indicate the number of smokers who answered incorrectly to the question related to smoking. In other words, Y = 0 shows the number of smokers who had expressed their daily cigarette consumption is nil. Hence, Y has a zero-truncated distribution. Accordingly, we can estimate total number of cigarette smokers by fitting an appropriate distribution to the variable Y. If $f(Y=0; \lambda)$ considered as an appropriate distribution for variable Y, then number of cigarette smokers would be estimated from the following formula [15, 16].

$$\hat{N}_{HTE} = \frac{n_{obs}}{1 - f(Y = 0; \lambda)} \quad (2)$$

Poisson distribution is being considered as Y distribution and $f(y;\lambda)$ as Poisson density function. Then λ would indicate the average number of daily cigarette consumption. In this setting, mixture distributions deliver better estimate than simple distributions. The following formula shows a K component Poisson mixture model with parameters $(\lambda_j, q_j; j=1,...,k)$ [9, 16].

$$f(y,Q) = \sum_{j=1}^{k} q_j \lambda_j^y e^{-\lambda_j} / y! \quad (3)$$

In this formula, mixture distribution has discrete distribution Q. If we assume $\sum_{j=1}^{k} q_j = 1$, non-negative weight of $q_j$ would be attributed to Poisson with mean of $\lambda_j$. Then the number of unobserved (missing) data will be estimated from formula $\sum_{j=1}^{k} q_j e^{-\lambda_j}$ [16]. The maximum likelihood estimation of parameters was performed using the EM algorithm [16-18]. To select a model with appropriate number of components, a Bayesian Information Criterion (BIC) was used based on

$$BIC = 2L(\hat{Q}_k) - (2k-1)log(n_{obs})$$

where $L(\hat{Q}_k)$ is the maximum likelihood function. According to this criterion, the largest BIC value is considered as the best model [16].

To fit the mixture model to the data and to conduct a simulation study, we used Computer-Assisted Mixture model analysis for Capture-Recapture count data (CAMCR). This software provides a distribution based on global maximum likelihood estimator of the parameters [18]. To conduct a simulation study using R2.11.1 software, the actual population size, ($N_{est}$), was considered as a fixed number. Based on the available data, a mixture Poisson distribution with appropriate parameters was identified. Furthermore, the maximum frequency for detecting a person by the system was specified. In the simulation study, the selected mixture Poisson distribution was used and the process was repeated 500 times ($r$=500), estimating smokers population size in each run ($N_i$; $i$=1,…, $r$). The estimated mean and variance of the estimator was obtained by calculating the mean and variance of the samples obtained from repeated process.

If $N_i$ is considered as population size in *i-th* step, then bias can be obtained using the following formula.

$$Bias = \overline{N} - N_{est}, \quad \overline{N} = \frac{1}{r}\sum_{i=1}^{r} N_i \quad (4)$$

In addition, Root Mean Square Error (RMSE) of the estimator can be obtained, using:

$$RMSE = \sqrt{Bias^2 + Variance} \quad (5)$$

## Results

Totally, 233 individuals consumed one to 60 cigarettes per day. The results of mixture Poisson model fitted to the data are presented in Table 1. In this table, the maximum likelihood estimates of the model parameters were specified for the conditions that the number of components is 1, 2, 3 and 4. The values of log likelihood function and BIC for each distribution with specified components were indicated as well. The results show that the mixture Poisson distribution with four components was the most appropriate model.

This distribution showed that more than a quarter of the smokers consumed three cigarettes per day. Also about one-third consumed eight cigarettes, one-third consumed one packet, and 5% of consumers smoked two packets per day on average.

The number of unobserved (missing) smokers was estimated three persons. Accordingly, the total number of smokers among the population under study was estimated as 236. Hence, the prevalence of cigarette smoking was 20.6%, including 36.2% among men and 3.3% among women.

The validity of the results was investigated using the parametric simulation study with BIC criteria. In the simulation process, 500 samples of 236 were simulated for the selected four-component mixture Poisson distribution with specified parameters in Table 1.

Accordingly, the bias estimator of population size was estimated to be zero and the RMSE was estimated to be 2.5. The result of simulation study was summarized in Table 2.

According to the result of simulated model, a valid smoker population size was obtained based on the mixture Poisson model with four components.

## Discussion

There are several studies, which estimated number of unobserved cases using models fitted to count data [9-12, 15, 16]. Most of these studies used simple Poisson models or Zelterman's estimator, based on a simple Poisson distribution. Van der Heijden et al. estimated the number of illegal immigrants to Netherland using a simple Poisson model. The count data was based on the number of times that each non-legal immigrant was arrested by police. They used a truncated Poisson regression model to estimate population size [12]. Bouchard et al. estimated the number of offending farmers who cultivated marijuana in an eastern province in Canada using count data including the information of the farmers who were arrested once or twice for marijuana

cultivation. Contrary to the present study, they used Zelterman's estimator to estimate population size [10]. Vilas and Böhning used count data to estimate the total number of holdings affected with classical scrapie in Great Britain using Zelterman's estimator [15, 19].

Mixture models were used by Böhning et al in 2005 to estimate population size in human societies. A Poisson model requires certain assumptions such as independence of observations and homogeneity of the parameters while mixture model is much more flexible. If the value of a specified parameter in the population under study changes due to changes of some latent variables, the mixture distribution can explain over dispersion among the data considering the effect of latent variables [16].

**Table 1**: Results of mixture Poisson distribution fitted to the count data related to cigarette smoking

| Number of components | Mean of components | Weight of components | Values of log likelihood function | Values of BIC criterion | Number of unobserved smokers | Total number of smokers |
|---|---|---|---|---|---|---|
| 1 | 12.12 | 1.000 | -1189 | -2381 | 0 | 233 |
| 2 | 5.996 | 0.615 | -797.9 | -1606 | 0 | 233 |
| | 21.88 | 0.385 | | | | |
| 3 | 5.442 | 0.549 | -757.4 | -1533 | 1 | 234 |
| | 17.66 | 0.389 | | | | |
| | 36.53 | 0.062 | | | | |
| 4 | 3.225 | 0.285 | -737.5 | -1500 | 3 | 236 |
| | 8.404 | 0.329 | | | | |
| | 19.17 | 0.331 | | | | |
| | 37.37 | 0.055 | | | | |

**Table 2**: Number of unobserved smokers and total number of smokers estimated by mixture Poisson model and simulation study

| Model | Total number of smokers | Unobserved number of smokers | Bias | RMSE[a] |
|---|---|---|---|---|
| Mixture Poisson model | 236 | 3 | - | - |
| Simulation study | 236 | 3 | 0 | 2.5 |

[a] RMSE: Root mean square of error

The present study showed that the fitted mixture Poisson distribution was appropriate for cigarette smokers. Böhning et al. conducted a similar study, using mixture Poisson model, to estimate the prevalence of heroin consumption in Bangkok. They developed the CAMust software for data analysis [9]. However, the present study had two advantages compared to the Böhning's study. First, CAMCR software which was used in the present study could estimate global maximum likelihood values of the parameters while CAMust software could not [18]. Furthermore, the validity of the results of the present study was investigated by a simulation study [20].

A cross-sectional study, conducted in Tehran in 2003, reported the prevalence of smoking 37.4% in men and 4.2% in women [5] which is similar to the results of present study. However, in some studies, the reported prevalence of cigarette smoking was less than what reported in this study. A cross-sectional study, conducted in Bandar-Abbas City in 2007, reported the prevalence cigarette smoking 11.7% among general population aged ≥15 years, 22.7% in men and 0.9% in women [7]. Another study, performed in Yazd City in 2000, estimated the prevalence of smoking 31.2% in men [8]. A third study, conducted in Tehran in 2001, reported the prevalence of cigarette smoking 10.6% among general population, 22% in men and 2.1% in women [4]. A further study, conducted in Semnan Province in 2009, estimated the prevalence of smoking

2% in women [2]. The difference in the prevalence rates of cigarette smoking estimated by previous studies and the present study may be due to the diversity of age distribution among the study populations.

The estimated prevalence of cigarette smoking was 20.6%, 36.2% among men, and 3.3% among women. These findings revealed that the mixture Poisson model fitted well to the data. An advantage of this study compared with the previous studies, which were conducted for estimation of the prevalence of smoking in Iran, is that the number of unobserved smokers was investigated using mixture Poisson model. The number of unobserved cigarette smokers was three. The small value of this estimate indicates that data collection was done properly. This finding is confirmed by this reality that data were collected by a well-trained team including expert psychologists and sociologists.

## Conclusion

The evaluation number of unobserved data can be estimated by fitting model to truncated count data. The truncated mixture Poisson model is particularly useful to estimate population size when the main objective of the study is to investigate negative traits to which the participants may answer incorrectly. In this study, we found that there were three unobserved person. With this finding, we ensure to data collection method.

## Acknowledgments

This article was a part of MSc thesis supported by Hamadan University of Medical Sciences. We would like to thank Deputy of Education as well as Deputy of Research and Technology of Hamadan University of Medical Sciences for financial support of this study. We also wish to thank Research Center for Health Sciences and Faculty of Health for their valuable collaboration with this study.

## Conflict of interest statement

The authors declare that they have no conflicts of interest.

## Funding

This study was funded by the Deputy of Research and Technology of Hamadan University of Medical Sciences.

## References

1. Heydari G, SharifKashani B, Sharifi H, Hosseini M, Masjedi MR. Evaluation of the Cigarettes' Brands in Tehran, 2008. *Hakim.* 2009;12(2):48-53.
2. Rashidy-Pour A, Malek M, Eskandarian R, Ghorbani R. Epidemiology of smoking among adult women population of Semnan province, Iran. *Scientific Journal of Semnan University of Medical Sciences.* 2009;11(2):75-82.
3. Sharif-Rad GR, Hazavei MM, Hasan-Zadeh A, A D-A. The effect of health education based on health belief model on preventive actions of smoking in grade one, middle school students. *Scientific Journal of Arak University of Medical Sciences.* 2007;10(1):250-257.
4. Emami H, Habibian S, Salehi P, Azizi F. Pattern of smoking habit in an urban area in Tehran, 2001: Tehran glucose and Lipid study. *Scientific Journal of the Faculty of Medicine.* 2003;27(1):47-52.
5. Kelishadi R, Hashemipoor M, Sarafzadegan N, et al. Effects of some environmental factors on smoking and the consequences of smoking on major cardiovascular disease (CVD) risk factors in adolescent: Isfahan healthy heart program-heart health promotion from childhood. *Scientific Journal of Guilan University of Medical Sciences.* 2004;13(50):62-74.
6. Noori M, Adili F, Poorebrahimi R, Heshmat R, Fakhrzadeh H. Evaluation of smoking habit and its relation with cardiovascular risk factors in residents covered by population health center of Tehran University of Medical Sciences. *Iranian Journal of Diabetes and Lipid.* 2006;3:91-97.
7. Aghamolayi T, Zare S. Smoking and Water-pipe use model for population over 15 year in Bandar Abbas. *Scientific Journal of Hormozgan University of Medical Sciences.* 2007;11(4):241-246.
8. Hoseini S. The first age, the first place, the first and most important motivation factors recommended smoking in the city of Yazd. *Scientific Journal of Yazd University of Medical Sciences.* 2000;8(1):12-17.
9. Böhning D, Suppawattanabodee B, Kusolvisitkul W, Viwatwongkasem C. Estimating the number of drug users in

Bangkok 2001: A capture–recapture approach using repeated entries in one list. *Eur J Epidemiol.* 2004;19:1075-1083.

**10.** Bouchard M. A Capture-Recapture Model to Estimate the Size of Criminal Populations and the Risks of Detection in a Marijuana Cultivation Industry. *J Quant Criminol.* 2007;23:221-241.

**11.** Hall WD, Ross JE, Lynskey MT, Law MG, Degenhardt LJ. How many dependent heroin users are there in Australia? *Med J Aust.* 2000;173:528-531.

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

**13.** Poorolajal J, Haghdoost A, Mahmoodi M, Majdzadeh R, Nasseri-Moghaddam S, Fotouhi A. Capture-recapture method for assessing publication bias. *JRMS.* 2010;15(2):107-115.

**14.** Gharibzadeh S, Mahjub H, Moghimbeigi A, Sadri GH. Disease Mapping of Acute Flaccid Paralysis in Iran Using Mixture Distributions. *Hakim.* 2010;12(4):11-18.

**15.** Del Rio Vilas VJ, Böhning D. Application of one-list capture–recapture models to scrapie surveillance data in Great Britain. *Prev Vet Med.* 2008;85:253-266.

**16.** Böhning D, Dietz E, Kuhnert R, Schiin D. Mixture models for capture-recapture count data. *Stat Methods Appt.* 2005;14:29-43.

**17.** Böhning D, Kuhnert R. Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions. *Biometrics.* 2006;62:1207-1215.

**18.** Kuhnert R, Böhning D. CAMCR: Computer-Assisted Mixture model analysis for Capture–Recapture count data. *AStA Adv Stat Anal.* 2009;93:61-71.

**19.** Navaratna WCW, del Rio Vilas VJ, Böhning D. Extending Zelterman's Approach for Robust Estimation of Population Size to Zero-truncated Clustered Data. *Biom J.* 2008;50(4):584-596.

**20.** Goli S, Mahjub H, Moghimbeigi A, Poorolajal J. The application of mixture models for estimation prevalence of various diseases using capture-recapture count data [MSc Thesis]. Hamadan, Hamadan University of Medical Sciences; 2010.