A COMPARISON OF TRUNCATED POISSON ESTIMATORS OF POPULATION SIZE UNDER MODEL CONTAMINATIONS

NITICHA WANNASIRIKUL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (BIOSTATISTICS)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2005

ISBN 974-04-5847-5
COPYRIGHT OF MAHIDOL UNIVERSITY

Copyright by Mahidol University

Thesis
Entitled

A COMPARISON OF TRUNCATED POISSON ESTIMATORS OF POPULATION SIZE UNDER MODEL CONTAMINATIONS

Niticha Wannasirikul
.....
Miss Niticha Wannasirikul,
Candidate

Chukiat Viwatwongkasem
.....
Assoc. Prof. Chukiat Viwatwongkasem,
M.Sc. (Biostatistics)
Major-Advisor

.....
Prof. Dankmar Böhning,
Ph.D. (Mathematics)
Co-Advisor

Ronny Kuhnert
.....
Dr. Ronny Kuhnert
Dr. rer. med (Non-medical)
External Examiner

Rassmidara Hoonsawat
.....
Assoc. Prof. Rassmidara Hoonsawat,
Ph.D. (Physics)
Dean
Faculty of Graduate Studies

Piangchan Rojanavipart
.....
Assoc. Prof. Piangchan Rojanavipart,
M.H.S. (Biostatistics)
Chair
Master of Science Programme
In Biostatistics
Faculty of Public Health

Copyright by Mahidol University

Thesis
Entitled

A COMPARISON OF TRUNCATED POISSON ESTIMATORS OF POPULATION SIZE UNDER MODEL CONTAMINATIONS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Biostatistics)
on
April 4, 2005

Niticha Wannasirikul
Miss Niticha Wannasirikul,
Candidate

Chukiat Viwatwongkasem.
Assoc. Prof. Chukiat Viwatwongkasem,
M.Sc. (Biostatistics)
Chair

Prof. Dankmar Böhning,
Ph.D. (Mathematics)
Member

Piangchan Rojanavipart
Assoc. Prof. Piangchan Rojanavipart,
M.H.S. (Biostatistics)
Member

Ronny Kuhnert
Dr. Ronny Kuhnert,
Dr. rer. med (Non-medical)
Member

Rassmidara Hoonsawat
Assoc. Prof. Rassmidara Hoonsawat,
Ph.D. (Physics)
Dean
Faculty of Graduate Studies
Mahidol University

C. Chaikittiporn
Assoc. Prof. Chalermchai Chaikittiporn,
Dr. P.H. (Epidemiology)
Dean
Faculty of Public Health
Mahidol University

Copyright by Mahidol University

ACKNOWLEDGEMENT

I would like to express my sincere and deepest appreciate to Assoc. Prof. Chukiat Viwatwongkasem, my major advisor, and Prof. Dankmar Böhning, my co-advisor, for their invaluable suggestion, constant support and warm encouragement. They have sacrificed their time to give advice and had the grace to correct mistakes and make improvements so that this study came out wholly complete. I also acknowledge the opportunity they gave me to pursue this study and to express my idea. I am truly grateful for their helps and thank them profusely.

Thanks to Dr. Ronny Kuhnert, my external examiner, for his expertise advice on FORTRAN programming that this study could not have been completed without.

I am indebted to the teachers of the Department of Biostatistics, Faculty of Public Health, who imparted their knowledge and experiences that really guided me through this study.

Thanks to all the fellow students and friends at the Department of Biostatistics for their help and support throughout this study. Thanks also to the staff of the Department of Biostatistics and every other kind person who has not been mentioned. I will always be grateful for their help.

Finally, I would like to acknowledge my appreciation and gratitude to my parents, my brother and sisters for their continuous support and love through my life.

Niticha Wannasirikul

27th Biostatistics

Copyright by Mahidol University

A COMPARISON OF TRUNCATED POISSON ESTIMATORS OF POPULATION SIZE UNDER MODEL CONTAMINATIONS

NITICHA WANNASIRIKUL 4336106 PHBS/M

M.Sc. (BIOSTATISTICS)

THESIS ADVISORS: CHUKIAT VIWATWONGKASEM, M.Sc., DANKMAR BÖHNING, Ph.D., RONNY KUHNERT, Dr.rer.med.,

ABSTRACT

There are many ways to estimate the size of a population. This study aimed to estimate the population size of truncated Poisson distribution by using the McKendrick's, Mantel-Haenszel's, Zelterman's and the maximum likelihood method via the Horvitz-Thompson approach under homogenous Poisson and under model contaminations. The data were simulated using a truncated Poisson and a truncated Poisson mixture model by means of the FORTRAN programming language. The comparison for finding the best estimator was based on the criteria of bias, variance, and mean squared error for each estimator under consideration.

Results revealed that under a truncated Poisson model, Mantel-Haenszel's method provided the best population size estimator according to its low bias and small variance and mean squared error. For a truncated Poisson mixture model, Zelterman's method gave the best population size estimator. However, the variance of Zelterman's estimator was quite large.

The empirical example on the Cholera data suggested that population size estimators from the four methods had the comparable values since data followed the Poisson distributional assumption. On the other hand, the four methods differed for the drug user data in Bangkok. Standout values of the Zelterman estimator might be taken as diagnostic indication for a violation of the Poisson model.

Study on Mantel-Haenszel's estimator and Zelterman's estimator, especially its variance, can be further pursued.

KEYWORDS:TRUNCATED POISSON DISTRIBUTION / POPULATION SIZE ESTIMATOR / MODEL CONTAMINATIONS

86 P. ISBN 974-04-5847-5

Copyright by Mahidol University

การเปรียบเทียบตัวประมาณค่าที่ตัดปลายแบบพัวซองของขนาดประชากรภายใต้แบบจำลองคอนเทมิเนชั่น (A COMPARISON OF TRUNCATED POISSON ESTIMATORS OF POPULATION SIZE UNDER MODEL CONTAMINATIONS)

นิธิชา วรรณศิริกุล 4336106 PHBS/M
วท.ม. (ชีวสถิติ)

คณะกรรมการควบคุมวิทยานิพนธ์: ชูเกียรติ วิวัฒน์วงศ์เกษม, วท.ม., DANKMAR BÖHNING, Ph.D., RONNY KUHNERT, Dr.rer.med.,

บทคัดย่อ

มีหลายวิธีการที่ใช้ประมาณค่าขนาดประชากร การศึกษานี้ได้เปรียบเทียบตัวประมาณค่าของขนาดประชากรตามวิธีของ Horvitz Thompson จำนวน 4 วิธีคือ McKendrick, Mantel-Haenszel, Zelterman และ maximum likelihood ภายใต้การแจกแจง 2 กรณีคือ ภายใต้การแจกแจงที่ตัดปลายแบบพัวซอง และการแจกแจงที่ตัดปลายแบบพัวซองแบบผสม โดยใช้โปรแกรมภาษา FORTRAN ในการสร้างแบบจำลอง ในการเปรียบเทียบตัวประมาณค่าทั้ง 4 วิธีจะพิจารณาคุณลักษณะของค่าความเอนเอียงของตัวประมาณค่าที่ต่างไปจากค่าจริง (bias), ค่าความแปรปรวนของตัวประมาณค่า (variance), และค่าความแปรปรวนของค่าผิดพลาดระหว่างตัวประมาณค่ากับค่าจริง (mean squared error)

ผลการศึกษาพบว่าตัวประมาณค่าของขนาดประชากรที่ดีที่สุดภายใต้การแจกแจงที่ตัดปลายแบบพัวซองคือวิธีของ Mantel-Haenszel ซึ่งให้ค่าความเอนเอียงน้อยที่สุด และให้ค่าความแปรปรวนและค่าความแปรปรวนของค่าผิดพลาดระหว่างตัวประมาณค่ากับค่าจริงต่ำที่สุด นอกจากนี้ยังพบว่าตัวประมาณค่าของขนาดประชากรที่ดีที่สุดภายใต้การแจกแจงที่ตัดปลายแบบพัวซองแบบผสมคือวิธีของ Zelterman ถึงแม้ว่าจะให้ค่าความแปรปรวนที่ค่อนข้างมาก

จากตัวอย่างข้อมูลอุจจาระร่วงที่ติดเชื้อ Cholera ที่มีลักษณะการแจกแจงที่ตัดปลายแบบพัวซอง พบว่าค่าที่ได้จากการประมาณขนาดประชากรของทั้ง 4 วิธีมีค่าใกล้เคียงกัน ในทางตรงกันข้ามสำหรับข้อมูลผู้ใช้ยาเสพติดในกรุงเทพมหานครที่มีลักษณะไม่เป็นไปตามการแจกแจงแบบพัวซอง พบว่าค่าที่ได้จากการประมาณขนาดประชากรของทั้ง 4 วิธีมีค่าแตกต่างกัน ทั้งนี้ตัวประมาณค่าของวิธี Zelterman น่าจะเป็นตัวประมาณที่ดีที่สุด

ข้อเสนอแนะของการศึกษาต่อไป น่าจะพิจารณาค่าประมาณของขนาดประชากรโดยวิธี Mantel-Haenszel และ Zelterman โดยเฉพาะการประมาณค่าความแปรปรวนของตัวประมาณค่าทั้งสอง

86 หน้า. ISBN 974-04-5847-5

Copyright by Mahidol University

CONTENTS

			Page
ACKNOWLEDGEMENT……			iii
ABSTRACT……..			iv
LIST OF TABLES……...			viii
LIST OF FIGURES…….			ix
CHAPTER			
I	INTRODUCTION……		1
	1.1	Rational and Justification……	1
	1.2	Research Objectives……..	3
	1.3	Scope of Study……...	3
	1.4	Definitions of Term……...	4
II	LITERATURE REVIEW…….		7
	2.1	Horvitz-Thompson approach……	7
	2.2	Truncated Poisson models……	8
	2.3	McKendrick’s estimator……..	10
	2.4	Maximum likelihood estimator……	12
	2.5	Zelterman’s estimator……..	13
	2.6	Mantel-Haenszel’s estimator……	16
	2.7	Reviews of application…….	18
III	METHODOLOGY……..		20
	3.1	The mixture model……..	20
	3.2	The simulation procedures……..	22
		3.2.1 Poisson Model…….	22
		3.2.2 Poisson Mixture Model……	23
	3.3	Statistical Investigation……	25

Copyright by Mahidol University

CONTENT (CONT.)

CHAPTER			Page
	3.4	Empirical Application……	25
IV	RESULTS……..		27
	4.1	Results Based on Simulated Data …….	27
	4.1.1	Truncated Poisson Model……..	27
	4.1.2	Truncated Poisson Mixture Model…….	33
	4.2	Results Emphasizing on Application…….	42
V	DISCUSSION……		50
	5.1	Discussion on the Simulation Results……	50
	5.1.1	Truncated Poisson Model……..	50
	5.1.2	Truncated Poisson Mixture Model…….	51
	5.2	Discussion on the Application Results…….....	52
VI	CONCLUSION AND RECOMMENDATIONS……..		54
	6.1	Conclusion…….	54
	6.2	Recommendations……..	55
REFERENCES…….....			56
APPENDIX…….....			59
BIOGRAPHY……			86

Copyright by Mahidol University

LIST OF TABLES

Table		Page
1	Mean ranking of bias of the four population size estimators (SE)…..	29
2	Mean ranking of variance of the four population size estimators (SE)	31
3	Mean ranking of MSE of the four population size estimators (SE)….	33
4	Mean ranking of bias of the four population size estimators (SE) under truncated Poisson mixture model……...	36
5	Mean ranking of variance of the four population size estimators (SE) under truncated Poisson mixture model……...	39
6	Mean ranking of MSE of the four population size estimators (SE) under truncated Poisson mixture model……...	42
7	Observed and expected counts of the number of cholera data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method…….	43
8	Estimated population sizes for cholera data within 95% confidence interval…….....	44
9	Observed and expected counts of the number of heroin user data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method…….	45
10	Observed and expected counts of the number of methamphetamine user data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method…….	47
11	Estimated number of heroin and methamphetamine users and corresponding 95% confidence interval……...	48

Copyright by Mahidol University

LIST OF FIGURES

Figure		Page
1	Graphs of bias of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method…	28
2	Graphs of variance of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method…	30
3	Graphs of mean square error of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method…….	32
4	Graphs of bias of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model……..	34
5	Graphs of variance of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model……..	37
6	Graphs of mean square error of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model……	40
7	Estimated population sizes for cholera data within 95% confidence interval…….	44
8	Estimated number of heroin users within a 95% confidence interval...	48
9	Estimated number of methamphetamine users within a 95% confidence interval……...	49

Copyright by Mahidol University

CHAPTER I
INTRODUCTION

1.1 Rationale and Justification

There are many ways to estimate the size of a population. The procedure must be chosen carefully. Various techniques are not interchangeable, and an inappropriate choice can lead to serious errors in the final estimate. Each technique applies to a particular situation or type of data and each requires the user to make different assumptions about the population. Before choosing a method, one must carefully consider the quantity to be estimated, the population to which it applies to, the data that are available, and the type of estimate that is sought (1).

The Horvitz-Thompson estimator is an approach to estimate a population size. The advantage of the Horvitz-Thompson method is that its estimator is unbiased and can be computed for any sampling designs. Furthermore, it is often feasible even for very complex sample schemes (2). Hence, Horvitz-Thompson approach is chosen to estimate population size in this study and a capture-recapture data with one source is employed to illustrate the Horvitz-Thompson approach.

Capture-recapture method has been used to estimate the total number of individuals in population. An initial individual sampling is obtained, marked or identified and then sent back to its population. After a period of time, the second sampling is obtained and it will be noted and counted how many the individual in that sampling are already marked and then return them once again to its population. This process has to be done over and over as wished. If the second sample represents the population size as a whole, then the sample proportion of marked individual should be about the same as the population proportion of marked individual. Accordingly, the total number of individuals in the population can be estimated (3).

Copyright by Mahidol University

For data that are counts of the number of observations, a single probability distribution is fitted to the data. A Poisson distribution is most commonly used. Zero-truncated samples occur when observations enter the sample only after the first count occurs (4).

Consider the work presented by Van der Heijden (5) on estimating number of illegal immigrants in the Netherlands. The police registration data are used to derive data on how often each illegal immigrant is caught by the police. Due to the nature of police registration data, a zero count can not be observed, so the data are truncated. These count data can be considered as a special form of capture-recapture data, and traditional capture-recapture methods can be employed to estimate the frequencies of zero counts. Once an estimate of this frequency is obtained, we are then able to have an estimate for the size of population of illegal immigrants which can be obtained by adding the zero-count estimate to the observed number of illegal immigrants apprehended (6).

Capture-recapture method is a count distribution which could be estimated well by means of a truncated Poisson mixture distribution. Suppose that the listing parameter varies in the population according to some unobserved variable, then, the associated marginal distribution will be a mixture of the component density where the component membership is described by the latent variable. Mixture models provided not only a richer class of possible distributions, but also can be well motivated (7).

Truncated Poisson has been widely employed in public health area. For example, in United Kingdom, injecting of drugs and sharing needles among drug injectors is seen as a major public health problem. Gordon Hay and Filip Smit utilized this method to estimate the number of drug injectors from needle exchange data in United Kingdom. According to their study, truncated Poisson is a method of prevalence estimation which, in contrast to the capture-recapture method, can be used when only a single data source is available. The capture-recapture method has disadvantage of requiring three or more distinct data sources with some identifiable overlap (8).

In order to determine the population size according to Horvitz-Thompson approach, one must determine the value of lambda, a parameter of Poisson distribution. In this study, four methods of obtaining lambda: McKendrick’s,

Copyright by Mahidol University

maximum likelihood, Zelterman's, and Mantel-Haenszel's, will be employed with simulated data under various situations. The simulated data is generated by Monte Carlo technique, which provides ability to specify population size, with the aid of the Microsoft FORTRAN programming language. Estimated population sizes computed from the four methods are then compared by considering their bias, variance and mean square error.

Finally, empirical examples of estimating population size will be illustrated based on McKendrick's cholera data in an Indian village and the number of drug user data in Bangkok (9).

1.2 Research Objectives

1.2.1 To estimate parameter of truncated Poisson distribution from McKendrick's, maximum likelihood, Zelterman's, and Mantel-Haenszel's methods.

1.2.2 To estimate the population size by using McKendrick's, maximum likelihood, Zelterman's, and Mantel-Haenszel's methods via Horvitz-Thompson approach under truncated Poisson and truncated Poisson mixture model.

1.2.3 To compare the best estimators of population size by considering their bias, variance, and mean square error.

1.3 Scope of Study

1.3.1 This study strictly deals with the bias, variance, and mean square error of population size estimators from McKendrick's, maximum likelihood, Zelterman's, and Mantel-Haenszel method.

1.3.2 By Monte Carlo method, this study generates data under Poisson model and Poisson mixture model for various conditions

1.3.3 Each condition is replicated from 10,000 times.

Copyright by Mahidol University

1.3.4 Data generated in this study are based on simulation from Microsoft FORTRAN programming language.

1.4 Definitions of Term

Simulation: A quantitative technique that utilizes a computerized mathematical model in order to represent actual making under conditions of uncertainty for evaluating alternative courses of action based upon facts and assumptions (10).

Monte Carlo Technique: The Monte Carlo method was considered to be a technique, using random or pseudorandom numbers, for solution of a model. Random numbers are essentially independent. Random variables at computer centers are arithmetic codes for generating sequences of pseudorandom digits, where each digit (0 through 9) occurs with approximately equal probability (likelihood). Consequently, the sequences can model successive flips of a fair ten-side die. Such codes are called random number generators. Grouped together, these generated digits yield pseudorandom numbers with any required number of elements (11).

FORTRAN: FORTRAN, which stands for FORmula TRANslation, was the first high-level programming language. It made possible to the use of symbolic names to represent mathematical quantities and to write mathematical formulae in a reasonably comprehensible form. The idea of FORTRAN was proposed in the late 1953 by John Backus, in New York, and the first FORTRAN program was run in April 1957 (12).

Bias: Bias is the difference between the expected value of a simple statistic and the population parameter the statistic estimates (13).

Variance: A measure of the spread of scores in a distribution of scores, that is, a measure of dispersion. The larger the variance, the further the individual cases are

Copyright by Mahidol University

from the mean. The smaller the variance, the closer the individual scores are to the mean.

Especially, the population variance is the mean of the sum of squared deviations from the mean score. Taking the square-root of the variance gives the standard deviation (13).

Mean square error: Mean square error is equal to its variance plus the square of its bias. For example, mean square error of $\hat{N}$ for estimating parameter N can be written as:

$$MSE(\hat{N}) = Var(\hat{N}) + \left[E(\hat{N}) - N\right]^2$$

So, for an estimate to have small mean square error, it must simultaneously have small variance and small bias (14).

In this study, the statistics used for investigation and comparison are mean, bias, variance and the mean square error are taken form:

For 10,000 iterations, mean of estimator is defined as

$$Mean(\hat{N}) = E(\hat{N}) = \frac{\sum_{k=1}^{10,000} \hat{N}_k}{10,000}$$

Bias is defined as:

$$Bias(\hat{N}) = E(\hat{N}) - N = \frac{\sum_{k=1}^{10,000} \hat{N}_k}{10,000} - N$$

Variance is defined as:

$$Var(\hat{N}) = E\left[\hat{N} - E(\hat{N})\right]^2 = \frac{\sum_{k=1}^{10,000} \left[\hat{N}_k - E(\hat{N})\right]^2}{10,000}$$

And mean square error is defined as:

$$MSE(\hat{N}) = E\left[\hat{N} - N\right]^2 = \frac{\sum_{k=1}^{10,000} (\hat{N}_k - N)^2}{10,000}$$

Copyright by Mahidol University

Truncated Poisson Distributions: A truncated Poisson distribution is formed from a Poisson distribution by cutting off and ignoring the part lying to the right or left of a fixed variate value. A truncated sample is likewise obtained by ignoring all values greater than or less than fixed value (15). This study emphasizes on a Poisson distribution that is truncated at zero.

Model Contaminations: Contaminated model is one containing observations which are of doubtful accuracy or validity (15).

Copyright by Mahidol University

CHAPTER II
LITERATURE REVIEW

Review of the literature in this chapter is divided into 7 main parts as follow:

2.1 Horvitz-Thompson approach
2.2 Truncated Poisson models
2.3 McKendrick's estimator
2.4 Maximum likelihood estimator
2.5 Zelterman's estimator
2.6 Mantel-Haenszel's estimator
2.7 Reviews of application

2.1 Horvitz-Thompson Approach

The Horvitz-Thompson estimator was originally proposed in 1952 as an estimator of population size that can be constructed for any sampling design, with or without replacement. For a single-stage cluster sampling design, its form is given by (2)

$$\hat{N}_{HTE} = \sum_{i=1}^{n} \frac{I_i}{\pi_i} \qquad (2.1.1)$$

Under homogeneity $\pi_i = \pi$; $\hat{N}_{HTE} = \frac{n}{\pi}$ (2.1.2)

where $I_i =$ Indicator function of unit i^{th}

$\pi_i =$ the probability of the i^{th} unit being selected in the sample

$n =$ the number of distinct units in sample

Copyright by Mahidol University

The importance of the Horvitz-Thompson estimator is that it is an unbiased estimator that can be computed for any sampling design as long as one can assess the probability, π_i, of any unit being included in the sample. This is often feasible even for very complex sample schemes.

Suppose that the identification mechanism is able to identify a case with probability $1-p_0$, where p_0 is probability of the unidentifiable case, so that $N = Np_0 + N(1-p_0)$. Note that $n(1-p_o)$ is the expected number of cases identified by the mechanism which simply can be estimated by n. This leads to the simple estimating equation for N

$$N = Np_0 + n \qquad (2.1.3)$$

,which in other words states that the population size (N) is the sum of the unobserved (Np_0) and observed cases (n). The equation can easily be solved for N to provide the Horvitz-Thompson estimator

$$\hat{N}_{HTE} = \frac{n}{(1-p_0)} \qquad (2.1.4)$$

, which is the number of observed cases adjusted for the probability of being identified by the mechanism (16). p_0 is unknown for most applications and must be estimated.

2.2 Truncated Poisson Models

For the time being we consider the sample as untruncated. A simple distributional model for count $y = 0,1,2,...$ is given by the Poisson. Let y be a Poisson random variable, the probability mass function is obtained as:

$$\Pr(y = j \mid \lambda) = \frac{e^{-\lambda}\lambda^j}{j!} \quad , j = 0,1,2,... \qquad (2.2.1)$$

, where $j = 0,1,2,...$ are the possible values of y and $E(y) = \lambda$

The probabilities of zero and positive counts are

$$\Pr(y = 0 \mid \lambda) = e^{-\lambda} \qquad (2.2.2)$$

Copyright by Mahidol University

$$\Pr(y > 0 \mid \lambda) = 1 - e^{-\lambda} \qquad (2.2.3)$$

The conditional probability of observing y events given that $y > 0$ zero is computed with the law of conditional probability:

$$\Pr(y = j \mid y > 0) = \frac{\Pr(y = j \cap y > 0)}{\Pr(y > 0)} \qquad (2.2.4)$$

We transform to y as x such that x is truncated at zero:

$$x = \begin{cases} y & , y > 0 \\ undefined & , y = 0 \end{cases}$$

So

$$\Pr(x = i \mid x > 0, \lambda) = \frac{\Pr(x \mid \lambda)}{\Pr(x > 0 \mid \lambda)}$$

$$= \frac{e^{-\lambda}\lambda^{i}}{i![1 - e^{-\lambda}]} \qquad , i = 1,2,...,n \qquad (2.2.5)$$

, where there are $i = 1,2,...,n$ observations with $\Pr(x_i > 0 \mid \lambda) = 1 - e^{-\lambda}$ .Each probability is increased by the factor $\left[1 - e^{-\lambda}\right]^{-1}$, which distributes the probability of a zero count across all positive counts in the truncated sample. This forces the truncated probability mass function to sum to 1 (17).

Grogger and Carson (18) provided the mean and variance of the truncated variable $x \mid x \rangle 0$. Since zero counts are excluded, the expected value is increased by the inverse of the probability of a positive count:

$$E(x \mid x > 0, \lambda) = \frac{\lambda}{\Pr(x > 0 \mid \lambda)} = \frac{\lambda}{1 - e^{-\lambda}} \qquad (2.2.6)$$

The expected value for the zero truncated count converges to the expected value of y without truncated as the probability of a zero count approaches 0. With zero counts excluded, the variance is less than that of the untruncated Poisson distribution and equals.

$$Var(x \mid x > 0, \lambda) = E(x \mid x > 0, \lambda)\left[1 - \Pr(x = 0 \mid \lambda)E(x \mid x > 0, \lambda)\right]$$

$$= \frac{\lambda}{1 - e^{-\lambda}}\left[1 - \frac{\lambda}{e^{-\lambda} - 1}\right] \qquad (2.2.7)$$

Copyright by Mahidol University

The zero-truncated Poisson distribution is defined by a probability function conditional on $x > 0$ that is equation (2.2.5) with $\Pr(x > 0 \mid \lambda) = \left[1 - e^{-\lambda}\right], i = 1,2,...,n$. An estimate $\hat{\lambda}$ for λ can be obtained by fitting the truncated Poisson distribution using McKendrick, maximum likelihood, Zelterman and Mantel-Haenszel methods that will be discussed in the following sections. This estimate can be used to find the probability of an individual not being observed, $\hat{p}_o = e^{-\hat{\lambda}}$ (6)

It follows that the Horvitz-Thompson estimator ($\hat{N}_{HTE}$) can be constructed according to

$$\hat{N}_{HTE} = \frac{n}{1 - \hat{p}_0}$$

from which a new estimate of N with $\hat{p}_o = e^{-\hat{\lambda}}$ as:

$$\hat{N}_{HTE} = \frac{n}{1 - e^{-\hat{\lambda}}}$$

λ is the truncated Poisson parameter to be estimated.

2.3 McKendrick's Estimator

Xiao-Li Meng (19) examined link of McKendrick's work to the EM algorithm due to an improvement made by Irwin on a novel method McKendrick used for estimating an infection rate when the observed data do not distinguish between those individuals who are not susceptible to the infection and those who are susceptible, but do not develop symptoms. Meng illustrated the central ideas underlying the EM algorithm as well as its properties. The examination also suggested a profiling strategy for speeding up EM. McKendrick's data on an epidemic of cholera were used for illustration and compared EM with Irwin's method as well as the Newton-Raphson algorithm.

According to Meng, McKendrick considered the simplest movement of an infection process (i.e. one-dimensional and irreversible), for which he derived, by solving a differential equation, the negative binomial as a distribution of the number

Copyright by Mahidol University

of individuals who have experienced x attacks, which the Poisson distribution as a limiting case. He then applied the Poisson model to several data sets. He found the method is well fitted for a Poisson model, especially for his data on the Cholera.

To some, fitting a Poisson model to this data set means to estimate the Poisson mean, λ, by the sample average

$$\hat{\lambda} = \frac{\sum_{j=0}^{m} jf_j}{\sum_{j=0}^{m} f_j} \quad , j = 0,1,2,...,m$$

where f_j is frequency of j and m is the largest observed counts of j.

The fitted counts are then calculated as

$$\frac{\left(\sum_{j=0}^{N} f_j\right)\hat{\lambda}^j e^{-\hat{\lambda}}}{j!} \quad , j = 0,1,2,...,m$$

It is obvious that data that has too many zero cases is inappropriate for the Poisson distribution. Hence, McKendrick used a different fitting procedure which at first might seem to be mysterious.

He first let

$$S = 0.f_0 + 1.f_1 + 2.f_2 + ... + m.f_m = \sum_{j=0}^{m} jf_j$$

$$S^2 = 0^2.f_0 + 1^2.f_1 + 2^2.f_2 + ... + m^2.f_m = \sum_{j=1}^{m} j^2 f_j$$

, where $S =$ Sum of frequency x count

and $S^2 =$ Sum of frequency x count square

Expected value of Poisson distribution, λ, and variance can be respectively written as

$$E(X) = \lambda = \frac{S}{N} \qquad (2.3.1)$$

$$E(X^2) - (EX)^2 = \lambda \qquad (2.3.2)$$

Copyright by Mahidol University

By solving equation (2.3.1) and (2.3.2), arrives the McKendrick's estimator, $\hat{\lambda}_{Mc}$;

$$\hat{\lambda}_{Mc} = \frac{S^2 - S}{S} \tag{2.3.3}$$

McKendrick also showed that the population size estimator, $\hat{N}$, can be written in the form

$$\hat{N} = \frac{(S)(S)}{S^2 - S} \tag{2.3.4}$$

Technically speaking, McKendrick's model is a zero-truncated Poisson model. As Irwin noted, the method, with appropriate modification of the moment-equation underlying (2.3.4), can be applied for fitting other discrete distributions with zero-class missing.

Irwin also made a key improvement on McKendrick's method with the help of iteration. Meng pointed out that Irwin's iteration in fact converges to the maximum likelihood estimate (MLE) of λ under the zero-truncated Poisson model.

By substituting $\hat{\lambda}_{McKendrick}$ in population size equation from Horvitz-Thompson method, we now can write population size estimator according to McKendrick's method as

$$\hat{N}_{McKendrick} = \frac{n}{1 - e^{-\hat{\lambda}_{Mc}}} \tag{2.3.5}$$

2.4 Maximum Likelihood Estimator

For Maximum likelihood estimation, the desirable properties of consistency, normality, and efficiency are asymptotic. This means that these properties have been proven to hold as the sample size approaches ∞ (4).

For Truncated Poisson

$$\Pr(x_i = i \mid x_i \rangle 0, \lambda) = \frac{e^{-\lambda}\lambda^i}{i![1 - e^{-\lambda}]} \quad , \; i = 1,2,...,n$$

Its likelihood function is

Copyright by Mahidol University

$$L(\lambda) = \prod_{i=1}^{m} P_i^{f_i}$$

$$= \prod_{i=1}^{m} \left[\frac{e^{-\lambda}\lambda^i}{i!} \frac{1}{1-e^{-\lambda}} \right]^{f_i}$$

, where m is the largest observed counts i and f_i is frequency of i.

Then the log-likelihood function is

$$LogL(\lambda) = \sum_{i=1}^{m} f_i \left[-\lambda + i\log\lambda - \log(i!) - \log\left(1-e^{-\lambda}\right) \right]$$

$$= -\lambda n + \log\lambda \sum_{i=1}^{m} i f_i - \sum_{i=1}^{m} f_i \log(i!) - n\log\left(1-e^{-\lambda}\right)$$

Differentiate with respect to λ

$$\frac{dL}{d\lambda} = -n + \frac{\bar{x}}{\lambda} n - 0 - \frac{ne^{-\lambda}}{\left(1-e^{-\lambda}\right)}$$

Finding the maximum likelihood estimator $\hat{\lambda}$ by solving $\frac{dL}{d\lambda} = 0$ yields

$$\lambda = \bar{x}\left(1-e^{-\lambda}\right) \qquad (2.4.1)$$

However, this is not the 'closed-form' solution. Iterative methods must be used until $\hat{\lambda}$ converges to a constant with acceptable error of less than 0.00001. $\hat{\lambda}_{Mc}$ is used as initial value in this study.

$$\lambda \xrightarrow{iterative} \hat{\lambda}_{MLE}$$

Hence, the population size estimator for maximum likelihood method is

$$\hat{N}_{MLE} = \frac{n}{1-e^{-\hat{\lambda}_{MLE}}} \qquad (2.4.2)$$

2.5 Zelterman's Estimator

Daniel Zelterman (20) proposed for the relative frequency of the unobservable zero class in a truncated Poisson distribution. A comparison on estimator between maximum likelihood, $\hat{\psi}$, and his proposed estimator under truncated Poisson sampling, Q_i, was made.

Copyright by Mahidol University

$$Q_i = \exp\left[-(i+1)\frac{P_{i+1}}{P_i}\right] \qquad ,i = 1,2,...,m$$

Replace parameter P with statistic f ,

$$Q_i = \exp\left[-(i+1)\frac{f_{i+1}}{f_i}\right] \qquad ,i = 1,2,...,m$$

In practice i is chosen to be one or two,

$$Q_1 = \exp\left[-2\frac{f_2}{f_1}\right]$$

Q_1 generally has a smaller asymptotic MSE than does $\hat{\psi}$ (MLE) when sampling under a truncated Poisson distribution.

$$\frac{MSE(\hat{\psi})}{MSE(Q_1)} \geq 1$$

As can be seen, this ratio is always greater than unity. While maximum likelihood techniques work well in certain situations, the sampling model is known and correct.

In this study, Q_i is the probability of an individual not to be observed,

$$\hat{p}_o = \exp(-\hat{\lambda}) = Q_i = \exp\left[-(i+1)\frac{f_{i+1}}{f_i}\right] \qquad (2.5.1)$$

The population size of the Horvitz-Thompson estimator can be written as

$$\hat{N}_{Zel} = \frac{n}{1-e^{-(i+1)\frac{f_{i+1}}{f_i}}}$$

The value of i is usually chosen to be 1 or 2

$$\hat{N}_{Zel} = \frac{n}{1-e^{-2\frac{f_2}{f_1}}}$$

It should be noted that the estimator is primarily based on the lower frequencies (f_1 and f_2). This emphasis on the lower frequency classes makes sense. People seen once or twice are likely to bear a greater resemblance to persons never seen, than what would have been the case with persons seen very often. In additions, the emphasis on the lower frequency class makes the estimates robust in the presence of ‘heterogeneity’, e.g. persons seen very often may form a different subgroup as

Copyright by Mahidol University

compared to persons rarely seen. This method is particularly suitable for a mixing distributions and it can be readily obtained without using statistical packages.

Gordon Hay and Filip Smit (8) presented a study on the size of drug injecting population from needle exchange data. The truncated Poisson estimator is a method of prevalence estimation which is contrast to the capture-recapture method. The truncated Poisson estimators is the number of attendances by individual drug users, which can be extracted form drug treatment agencies' recording systems, following a recognized statistical distribution and distribution can be used to obtain an estimate of the size of the hidden drug using population which is not in contact with that agency. A truncated Poisson estimator proposed by Zelterman (1988) and related estimators can be applied in data generated by counts of individuals who have been seen once, twice and so on. Zelterman's estimator assumes that the observed series of frequencies follows a distribution, similar to the Poisson distribution, which is truncated below one. Zelterman's estimator of the unknown population size is given by

$$\hat{N} = \frac{n}{\left[1 - \exp\left(-2\frac{f_2}{f_1}\right)\right]}$$

,where n represents the number of observations.

The result is 647 individuals attended the needle exchange in 1997, the total number of drug injectors was estimated to be 1041 within a 95% confidence interval of 960 – 1137.

As with other drug misuse prevalence estimation methods, there are several assumptions related to the use of this method such as closure, homogeneity, perfect identification, and constant recapture probability. These assumptions can be checked and handled data-analytically.

However, according to Gordon Hay, there may be occasional underestimate of the true population size in the presence of heterogeneity despite Zelterman's mathematical proof of the robustness of his estimator in presence of heterogeneity.

Copyright by Mahidol University

2.6 Mantel-Haenszel's Estimator

From Zelterman's equation (2.5.1) we have a truncated Poisson estimator by Zelterman's method as follow

$$\hat{\lambda} = (i+1)\frac{P_{i+1}}{P_i}$$

Replace parameter P with statistic f ,

$$\hat{\lambda} = (i+1)\frac{f_{i+1}}{f_i} \qquad (2.6.1)$$

We propose the truncated Poisson estimator by using the Mantel-Haenszel method by taking the weighted mean into equation (2.6.1)

$$\hat{\lambda} = \frac{\sum_{i=1}^{m-1} W_i (i+1)\frac{f_{i+1}}{f_i}}{\sum_{i=1}^{m-1} W_i}$$

, where $W_i = f_i$

$$\hat{\lambda}_{MH} = \frac{\sum_{i=1}^{m-1} (i+1) f_{i+1}}{\sum_{i=1}^{m-1} f_i} \qquad (2.6.2)$$

or we can write

$$\hat{\lambda}_{MH} = \frac{S - f_1}{n - f_m} \qquad (2.6.3)$$

Hence, the population size estimator by Horvitz-Thompson method is

$$\hat{N}_{MH} = \frac{n}{1 - e^{-\hat{\lambda}_{MH}}} \qquad (2.6.4)$$

P.G. Moore (21, 22) proposed the function that is readily calculated for any number of cell frequencies and gave a direct estimate of the Poisson parameter without the need for subsidiary tables.

Copyright by Mahidol University

No. of emissions per interval	0	1	2	3	…	m	…	Total
Frequency of intervals	f_0	f_1	f_2	f_3	…	f_m	…	N

The notation that he used is given in the table. If λ is the Poisson parameter for the population, he wrote, in series, as

$$0.e^{-\lambda} + 1.\frac{\lambda e^{-\lambda}}{1!} + 2.\frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + m.\frac{\lambda^m e^{-\lambda}}{m!} = \lambda\left[e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^{m-1} e^{-\lambda}}{m-1!}\right]$$

$$\lambda = \frac{\sum_{j=0}^{m} j\frac{\lambda^j e^{-\lambda}}{j!}}{\sum_{j=0}^{m-1}\frac{\lambda^j e^{-\lambda}}{j!}}$$

Note that this method does not truncate at the frequency f_0 but rather at the frequency above f_m. However, P.G.Moore's estimator can be adjusted to look similar to Mantel-Haenszel's estimator, which truncate at zero, as

$$\lambda = \frac{\sum_{j=0}^{m} jP_j}{\sum_{j=o}^{m-1} P_j}$$

This suggests that in order to estimate the Poisson parameter from zero-truncated series

$$\lambda = \frac{\sum_{i=1}^{m} iP_i}{\sum_{i=1}^{m-1} P_i}$$

and we can write similar form as equation (2.6.2)

$$\lambda = \frac{\sum_{i=1}^{m-1}(i+1)f_{i+1}}{\sum_{i=1}^{m-1} f_i}$$

Note that even though the estimators from both methods look some what similar but they are obtained from different ideas.

Copyright by Mahidol University

2.7 Reviews of Application

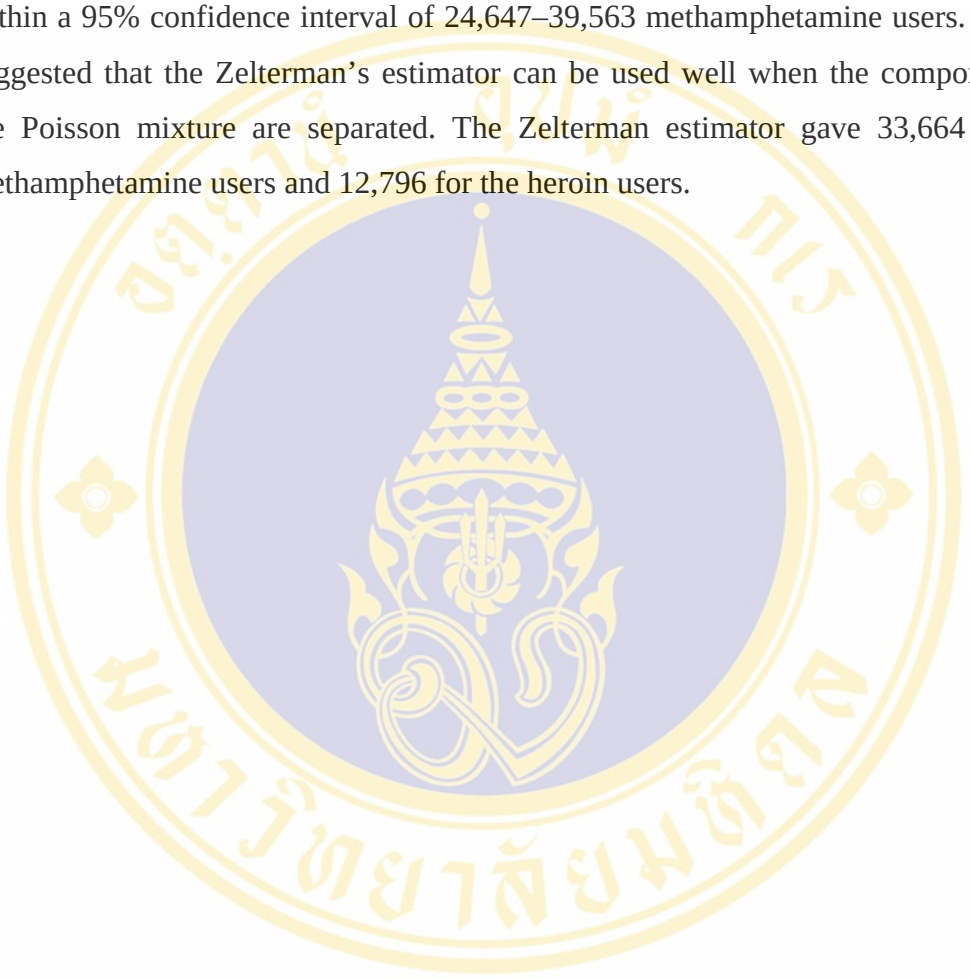
McKendrick, 1926 (following Meng (19)) applied the Poisson model to an epidemic of cholera in an Indian village. The data available were the counts of cholera cases in each house form which McKendrick constructed an estimate of the number of cholera affected households with zero cases of cholera. The cholera epidemic has affected a village in India with 223 households. Note that a household might be affected but no cases have been observed. And n_i be the number of households with exactly i cases. The data are: $n_1 = 32$, $n_2 = 16$, $n_3 = 6$, $n_4 = 1$, so that $n_{obs} = 55$. However, n_0 is not observed and need to be found. He found, by using his method, $n_0 = 37$, which is close with the result from using maximum likelihood method ($n_0 = 34$). The numbers of household affected by the cholera, $\hat{N}$, is then equal to 92 comparing to the value from the maximum likelihood of 88.

Suppawattanabodee, B., 2003 (9) estimated the number of drug user in Bangkok and investigated the percentage of under-reporting number of drug users from two sources. The data were taken from two sources: 61 treatment centers and 88 police stations in Bangkok during the period of 3 months from October 1 to December 31, 2001. A capture-recapture method was applied to estimate the total number of drug users. Matching type was considered to estimate drug users in Bangkok. However, the variance of the estimated was too large. Alternatively, she also estimated with mixture distributions of the truncated Poisson models which use only one source of the health treatment centers data. An estimated number of drug users found was 44,380.

Böhning D., 2004 (23) used the data from Suppawattanabodee B., 2003 under a Poisson mixture model to estimating the number of drug users in Bangkok 2001 by using the EM algorithm to find the maximum likelihood estimates. According from the data from 61 health treatment centers in Bangkok, of 11,222 drugs users found during the period, 7,063 (62.9%) were heroin users, 3.346 (29.8%) were methamphetamine users, and the remaining 813 (7.3%) distributed under 15 drug

Copyright by Mahidol University

categories, none above 1%. By using the truncated Poisson mixture model that showed excellent goodness-of-fit, an estimate of the unobserved frequency of drug users with zero treatment episodes could be constructed leading to an estimate of 11,296 within a 95% confidence interval of 8,964–13,628 heroin users and 32,105 within a 95% confidence interval of 24,647–39,563 methamphetamine users. He also suggested that the Zelterman's estimator can be used well when the components of the Poisson mixture are separated. The Zelterman estimator gave 33,664 for the methamphetamine users and 12,796 for the heroin users.



Copyright by Mahidol University

CHAPTER III
METHODOLOGY

The methodology of this study is mainly concerned with a simulation technique under the truncated Poisson mixture model using the Microsoft FORTRAN programming language. Estimators of population size are compared by considering bias, variance and mean square error. This chapter emphasizes on the important issues as follow:

3.1 The mixture model

3.2 The simulation procedure

3.3 Statistical investigation

3.4 Empirical Application

3.1 The Mixture Model

The importance of mixture distribution, enormous developments and frequent applications over the recent years are due to the fact that mixture models offer natural models for unobserved population heterogeneity. These models are too strict to capture the variation of the parameter over a diversity of subpopulations. In the case that the population consists of various subpopulations, denoted by $\lambda_1, \lambda_2, ..., \lambda_k$ where k denotes the number (possibly unknown) of subpopulations, we call this situation the heterogeneous case. Mixture models are especially used when we observed only the sample $x_1, x_2, ..., x_k$ from the marginal density of X, the mixture density, but the subpopulation belonging to the observation is not observed. Therefore, we speak of unobserved heterogeneity or extra-population heterogeneity.

Let a latent variable Z describe the population membership. Then the joint density can be written as:

Copyright by Mahidol University

$$f(x,z) = f(x \mid z)p(z) = f(x;\lambda_z)p_z$$

, where $f(x \mid z) = f(x;\lambda_z)$ is the conditional density on membership in subpopulation z, the unconditional density $f(x)$ is the marginal density of mixture distribution.

The Mixture density can be defined as follow:

$$f(x,P) = \sum_{z=1}^{k} f(x \mid z)p(z) = \sum_{j=1}^{k} f(x;\lambda_j)p_j$$

Note: p_j is the probability of belonging the j^{th} subpopulation having parameter λ_j.

These probabilities p_j have to meet the constraints that

$$p_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{k} p_j = 1$$

If there are k components in the mixture, it will be called a k-component finite mixture model. In a general finite mixture model, there is a set of $2k$ parameters.

The mixture distribution is $$P = \begin{bmatrix} \lambda_1,...,\lambda_k \\ p_1,...,p_k \end{bmatrix}$$

Each column corresponds to a component.

In this study, we are interested for a 2-component mixture model such that the mixtures of Poisson distribution can be defined as:

$$f(x,P) = \sum_{j=1}^{k=2} p_j \frac{e^{-\lambda_j}\lambda_j^x}{x!}$$

$$= f(x,\lambda_1)p_1 + f(x,\lambda_2)p_2 \qquad (3.1.1)$$

Hence, the mixture distribution for a 2-component is

$$P = \begin{bmatrix} \lambda_1, \lambda_2 \\ p_1, p_2 \end{bmatrix}$$

In this simulation we let $\lambda_1 = \lambda, \lambda_2 = \mu$. Also, let $p_1 = p$ and $p_2 = (1-p)$. So, (3.1.1) becomes

$$f(x,P) = f(x,\lambda)p + f(x,\mu)(1-p) \qquad (3.1.2)$$

For $p = 1$, $f(x,P)$ is a Poisson distribution, otherwise $f(x,P)$ is called a misspecified model.

Copyright by Mahidol University

Many applications are of the following type: under standard assumptions the population is homogeneous, leading to a simple, one-parametric and natural density. Examples include the binomial, the Poisson, the geometric, the exponential, and the normal distribution (with additional variance parameter). If these standard assumptions are violated because of population heterogeneity, mixture models can easily capture these additional complexities (24, 25).

3.2 The Simulation Procedure

The simulation is divided into two cases: one is simulating under a Poisson model and the other under the Poisson mixture model. The population sizes for each model are 100, 200, 500 and 1000.

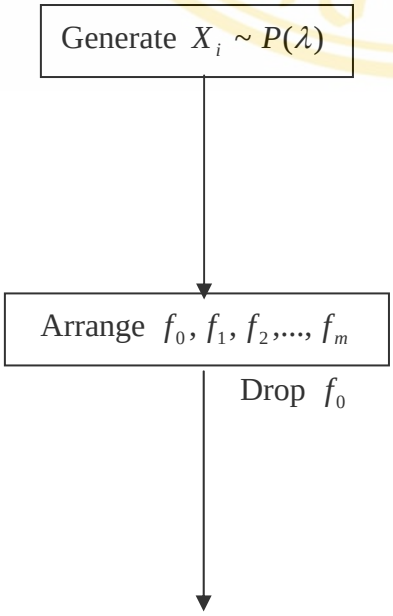
3.2.1 Poisson model

The Poisson model data is generated under the following criteria:

$p = 1.0$

$\lambda = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$

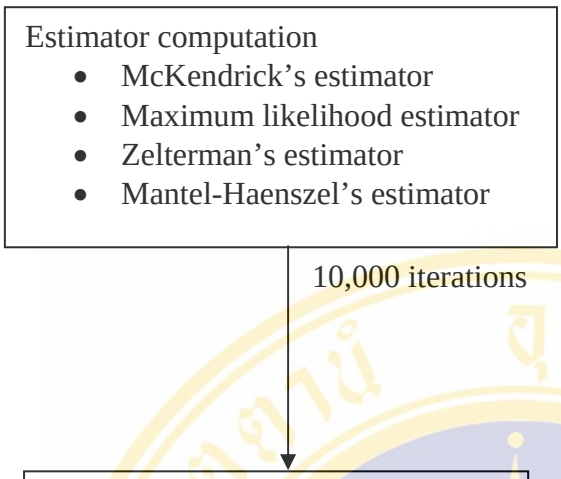
The simulation procedures are as follow:



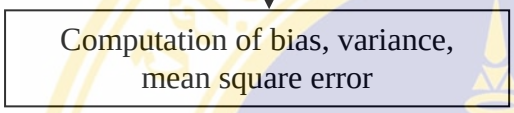
Generate random numbers for X_i from Poisson distribution for each value of λ.

Assort frequencies of each value of X_i: f_0 being the frequency of $X_i = 0$ and f_m being the frequency of maximum value of X_i. f_0 is then truncated before proceeding to estimator computation.

Copyright by Mahidol University



Estimate the parameter of truncated Poisson distribution, $\hat{\lambda}$, and population size estimator, $\hat{N}$, by McKendrick's, maximum likelihood, Zelterman's, and Mantel-Haenszel's method. Repeat the estimation for all 10,000 iterations.



Compute bias, variance, and mean square error of population size estimator from each method.

3.2.2 Poisson mixture model

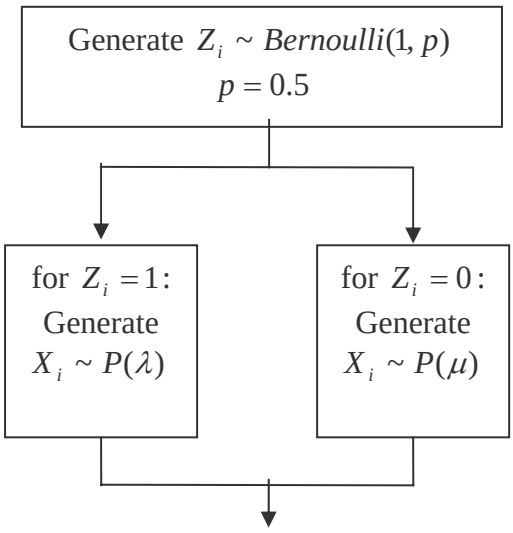
The Poisson mixture model data is generated under the following criteria:

$p = 0.5$

$\lambda = 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$

$\mu = 1.5, 1.6, 1.7, 1.8, 1.9, 2.0$

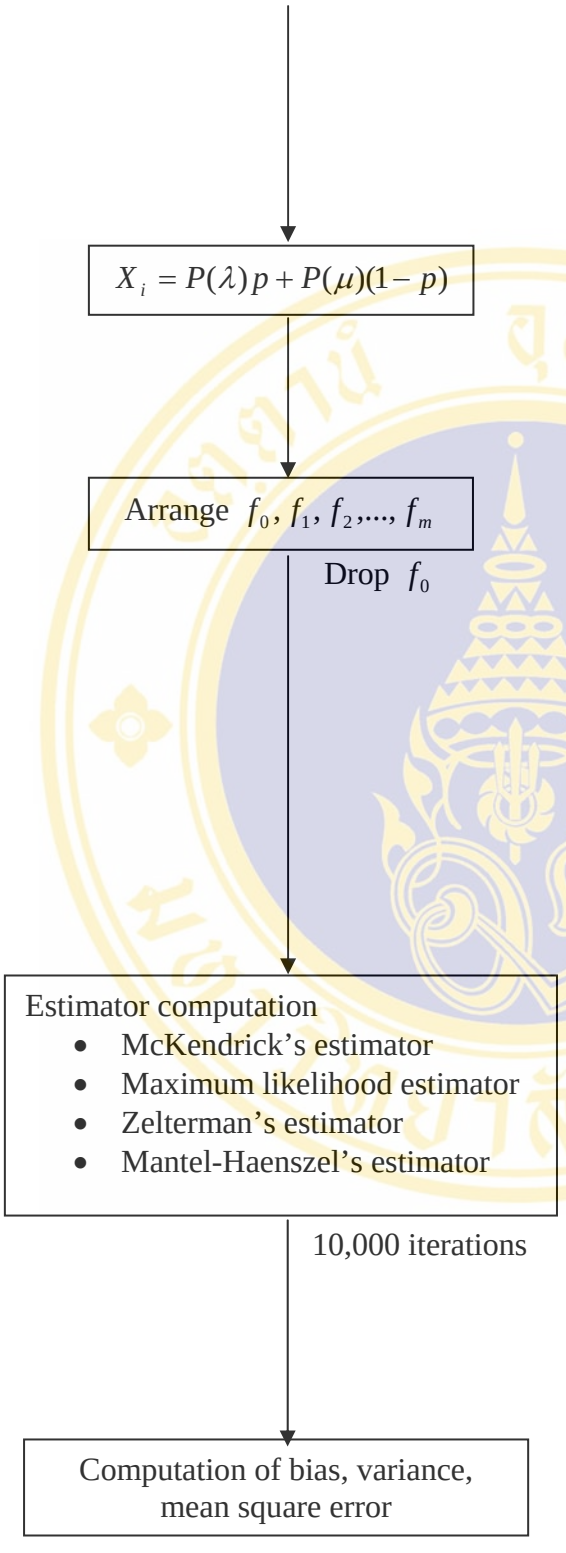
The simulation procedures are as follow:



Generate random numbers for Z_i from Bernoulli distribution given $p = 0.5$

- For $Z_i = 1$: Generate random numbers for X_i from Poisson distribution for each value of λ.
- For $Z_i = 0$: Generate random numbers

Copyright by Mahidol University



for X_i from Poisson distribution for each value of μ.

All values of X_i generated form a 2-component mixture model.

Assort frequencies of each value of X_i in the mixture model: f_0 being the frequency of $X_i = 0$ and f_m being the frequency of maximum value of X_i. f_0 is then truncated before proceeding to estimator computation.

Estimate the parameter of truncated Poisson distribution, $\hat{\lambda}$, and population size estimator, $\hat{N}$, by McKendrick's, maximum likelihood, Zelterman's, and Mantel-Haenszel's method. Repeat the estimation for all 10,000 iterations.

Compute bias, variance, and mean square error of population size estimator from each method.

Copyright by Mahidol University

3.3 Statistical Investigation

Each of the four population size estimators from both truncated Poisson model and truncated Poisson mixture model will be investigated through the following,

- Bias investigation
- Variance Investigation
- Mean-square error comparison

3.4 Empirical Application

McKendrick's cholera data in an Indian village is used to demonstrate data distribution of the truncated Poisson type. And drug user in Bangkok metropolitan data (9) is used to illustrate an example of a truncated Poisson mixture model. To confirm the type of distribution that each data set represents, chi-square goodness of fit test (26) is employed under the following hypothesis:

H_0 : The data follow a truncated Poisson distribution.

H_a : The data do not follow a truncated Poisson distribution.

The chi-square goodness of fit is given by $\chi^2 = \sum_{j=1}^{m} \frac{\left(f_j - n\hat{p}_j\right)^2}{n\hat{p}_j}$

, where probability of the truncated Poisson distribution $\hat{p}_j$ is $\frac{e^{-\hat{\lambda}}\hat{\lambda}^j}{j![1-e^{-\hat{\lambda}}]}$ and the critical value according to the level of significance α and the degree of freedom $m-2$ is $\chi^2_{\alpha,m-2}$

Then each set of data is input into the program that is run for the simulation data. Then population size estimator from each of the four methods is obtained. The population size estimator is illustrated within a 95% confidence interval, which is calculated from corresponding variance.

According to Van der Heijden et al., the variance of population size estimator from McKendrick's, maximum likelihood, and Mantel-Haenszel's method is obtained

Copyright by Mahidol University

by fitting the zero-truncated homogenous Poisson distribution (5) and can be expressed as follow:

$$\text{var}(\hat{N}) = \left(n_{obs} \frac{e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2} \right)^2 \left(\sum_{i=1}^{n_{obs}} y_i \lambda^{-2} - \frac{n_{obs} e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2} \right)^{-1} + n_{obs} \frac{e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2}$$

Then a 95% confidence interval for N is written as:

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

Following the work of Zelterman (20), proposed an estimator of truncated Poisson distribution, Q_j, where

$$n_{obs} Var(Q_1) = e^{-\hat{\lambda}_{MLE}} (1-e^{-\hat{\lambda}_{MLE}})(\hat{\lambda}_{MLE} + 2)$$

Then the standard deviation is

$$S.D. = \sqrt{\frac{e^{-\hat{\lambda}_{MLE}} (1-e^{-\hat{\lambda}_{MLE}})(\hat{\lambda}_{MLE} + 2)}{n_{obs}}}$$

Hence, a 95% confidence interval for the Zelterman's estimator, replace N with n_{obs} is written as:

$$\frac{n_{obs}}{(1-e^{-\hat{\lambda}_{Zel}}) \pm (1.96)(S.D.)}$$

Copyright by Mahidol University

CHAPTER IV
RESULTS

The main issues in this chapter are the results of comparing four estimators of population size under the truncated Poisson model and truncated Poisson mixture model based on their bias, variance and mean squared error.

4.1 Results based on simulated data

4.2 Results emphasizing on application

4.1 Results Based on Simulated Data

4.1.1 Truncated Poisson model

The truncated Poisson model is the situation that probability is equal to 1.0.

a.) Investigation of bias

Formally, biasness is the difference between the expected value of the estimator and the population parameter. To determine which estimator provides unbiased property, one should determine which estimator has small bias, in other words, which estimator has value of bias close to zero. The values of bias of population size for each value of lambda are demonstrated in Figure 1. The graphs are represented for each population size.

With the simulation results, Mantel-Haenszel method gives the least values of bias for all population sizes. Hence, according to Figure 1, the graphs of bias from Mantel-Haenszel method are closest to zero following by maximum likelihood methods and, in most cases, McKendrick's method and Zelterman's method. However, for the population size of 1,000 the values of bias from Zelterman's method are slightly less than McKendrick's method for several lesser values of lambda.

Copyright by Mahidol University

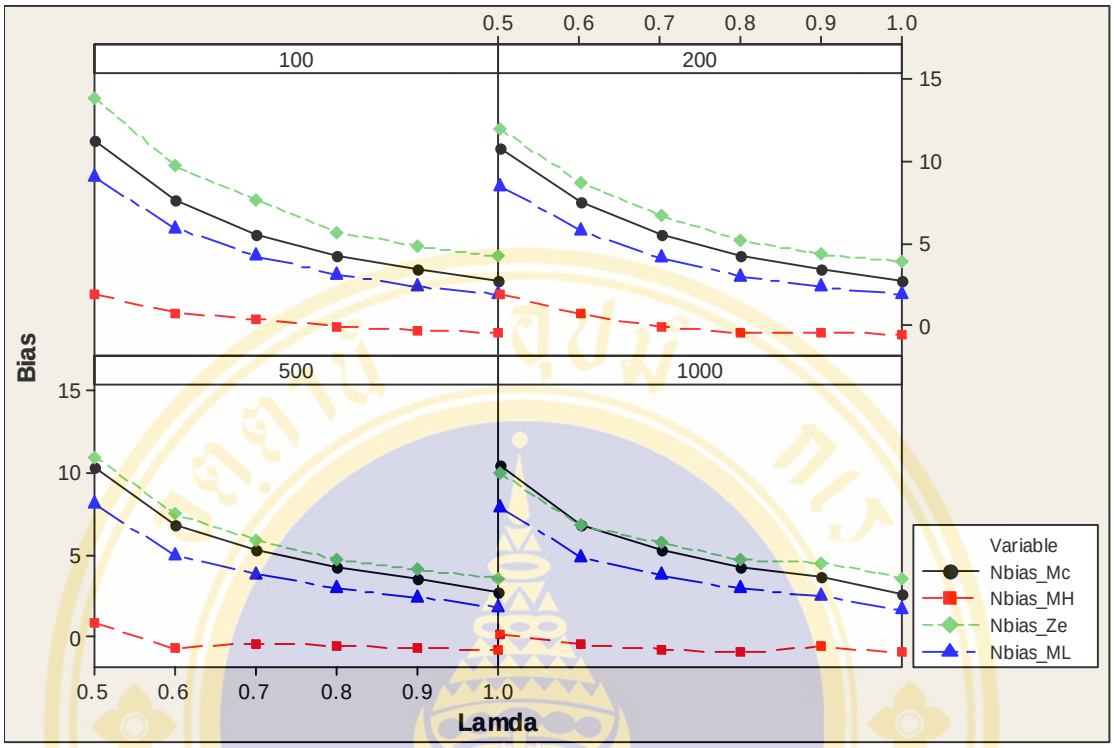


Figure 1. Graphs of bias of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method

Data are manipulated for better comprehension by assigning each value of bias a ranking. 1 will be assigned to the bias that is closest to zero and 4 will be assigned to the worst bias. According to the ranking, the method that gives the most bias closed to zero will has smallest mean ranking of the bias.

Table 1 shows mean ranking of bias, with its respective standard error in parenthesis, from McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood method for each population size.

According to Table 1, Mantel-Haenszel's method has mean ranking of bias of 1 with zero standard error since its gives the bias closest to zero for all cases. Similarly, maximum likelihood method gives the bias second closest to zero for all cases. McKendrick's and Zelterman's methods have mean ranking of bias of 3.08 with 0.058 standard error and 3.92 with 0.058 standard error respectively. This is not unanimous result because of some bias ranking variations in population size of 1,000.

Copyright by Mahidol University

Table 1. Mean ranking of bias of the four population size estimators (SE)

$p = 1.0$	Mean ranking of bias of Estimator (SE)			
N	McK	MH	Zel	MLE
100	3 (0)	1 (0)	4 (0)	2 (0)
200	3 (0)	1 (0)	4 (0)	2 (0)
500	3 (0)	1 (0)	4 (0)	2 (0)
1000	3.33 (.211)	1 (0)	3.67 (.211)	2 (0)
Total	3.08 (.058)	1 (0)	3.92 (.058)	2 (0)

b.) Investigation of variance

Variance is the difference between individual value of estimator and mean value. A high variance implies that more individual values tend to spread away from their mean. On the other hand, small variance suggests that more individual values of estimator are closer to their mean.

The values of variance of population size for each lambda are drawn in Figure 2. Each graph represents variance for each population size.

From the simulation results, Mantel-Haenszel's and maximum likelihood method give similar values of variance. According to Figure 2, Mantel-Haenszel's method seems to give smaller variance for smaller population sizes and maximum likelihood method give better variance for larger population sizes, while variances from McKendrick's and Zelterman's method are greater than those from the previous two methods in all cases.

Copyright by Mahidol University

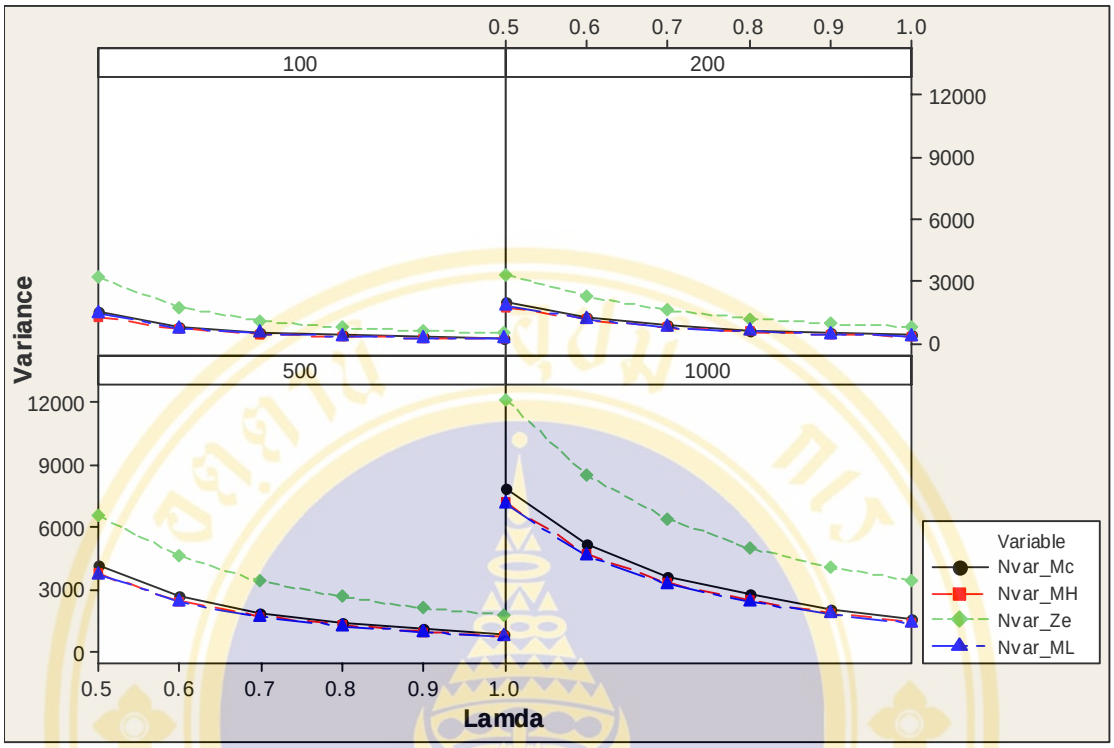


Figure 2. Graphs of variance of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method

Variance data are manipulated in the same way as for bias data previously mentioned. The better variance will have smaller ranking value.

Table 2 demonstrates mean of variance ranking along with standard error in parenthesis.

Similar to what was pointed out previously from graphical demonstration, variances from Mantel-Haenszel's method are all ranked first for smaller population size (N=100) and maximum likelihood method are ranked first for larger population size (N=500, 1000). For N=200, mean of variance ranking from each method are combination of the first and the second ranking, hence means are 1.17 and 1.83 with standard error of 0.167 for both. Meanwhile, variances from McKendrick's and Zelterman's method are ranked third and forth respectively in all cases.

Copyright by Mahidol University

Table 2. Mean ranking of variance of the four population size estimators (SE)

$p = 1.0$	Mean ranking of variance of Estimator (SE)			
N	McK	MH	Zel	MLE
100	3 (0)	1 (0)	4 (0)	2 (0)
200	3 (0)	1.17 (.167)	4 (0)	1.83 (.167)
500	3 (0)	2 (0)	4 (0)	1 (0)
1000	3 (0)	2 (0)	4 (0)	1 (0)
Total	3 (0)	1.54 (.104)	4 (0)	1.46 (.104)

c.) Investigation of mean square error

Mean square error measures the difference between individual value of estimator and the expected value. Similar to variance, less mean square error indicates that individual values are closer to the expected value, and higher mean square error suggests wider spread of individual values. Figure 3 presents four graphs of lambda versus mean square error for various population sizes.

Simulation results show that Mantel-Haenszel's and maximum likelihood method give less mean square error than those from McKendrick's and Zelterman's methods. This can be clearly demonstrated by Figure 3. Mean square errors from the first two methods plot very close lines and closer to zero than those from the latter two. Furthermore, for larger population sizes Mantel-Haenszel's and maximum likelihood method draw nearly the same line of mean square error vs. lambda.

Copyright by Mahidol University

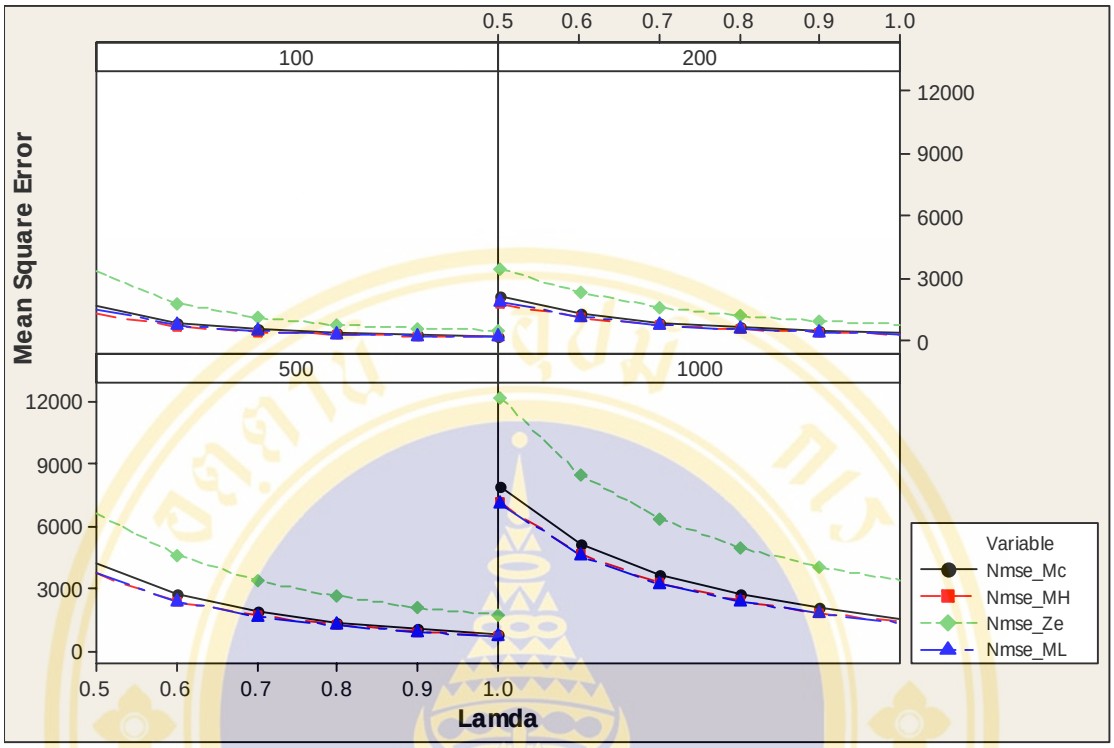


Figure 3. Graphs of mean square error of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method

Like bias and variance, mean square error data are also assigned to ranking. The mean ranking for mean square error are shown in Table 3.

According to Table 3, mean square errors from Mantel Haenszel method are all ranked first for population sizes of 100 and 500, while mean square errors from maximum likelihood method are all ranked first for the population size of 1000. For N=500 mean square errors from the two methods ranks first and second alternately. Mean square errors from McKendrick's and Zelterman's method are still ranked third and forth in Poisson model as in case of bias and variance.

Copyright by Mahidol University

Table 3. Mean ranking of MSE of the four population size estimators (SE)

$p = 1.0$	Mean ranking of MSE of Estimator (SE)			
N	McK	MH	Zel	MLE
100	3 (0)	1 (0)	4 (0)	2 (0)
200	3 (0)	1 (0)	4 (0)	2 (0)
500	3 (0)	1.67 (.211)	4 (0)	1.33 (.211)
1000	3 (0)	2 (0)	4 (0)	1 (0)
Total	3 (0)	1.42 (.103)	4 (0)	1.58 (.103)

4.1.2 Truncated Poisson mixture model

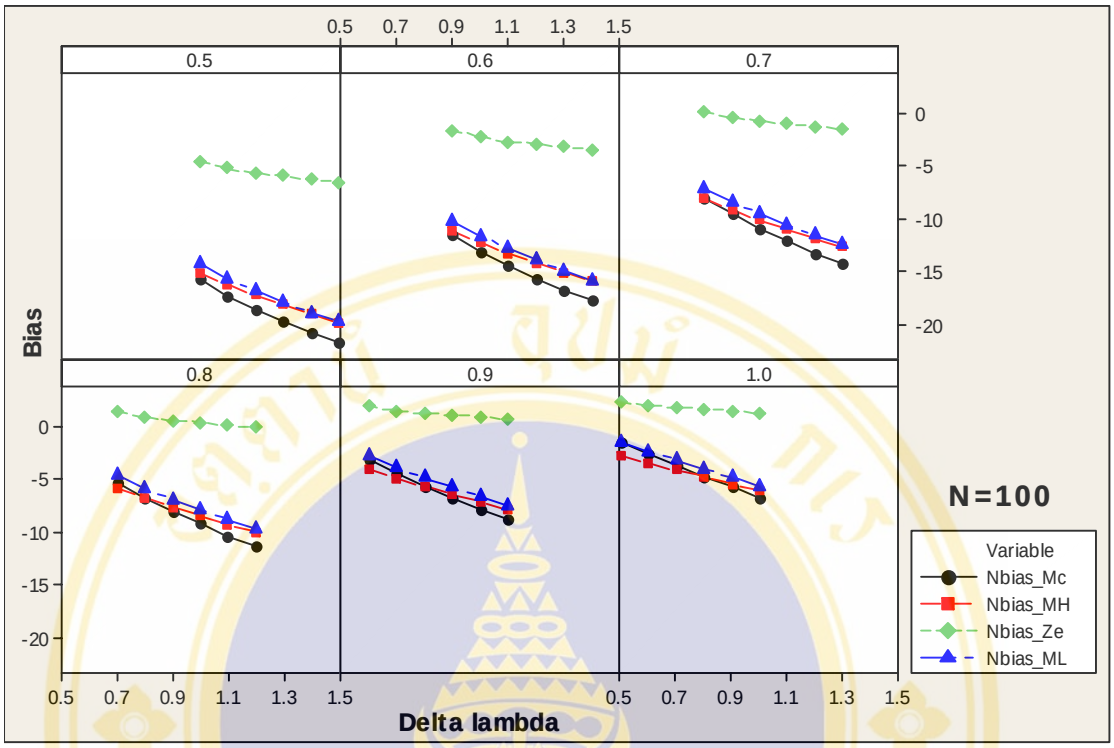
The truncated Poisson mixture model is the situation that probability is equal to 0.5.

a.) Investigation of bias

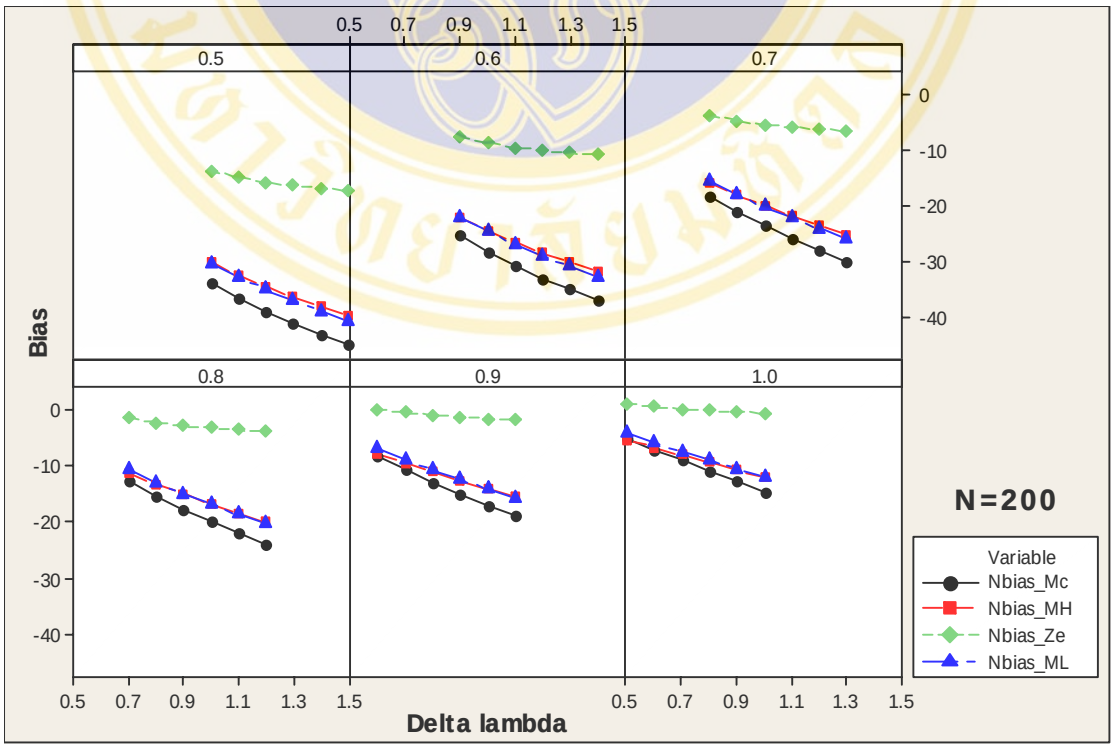
The biases of population size are plotted versus delta lambda for each lambda and population sizes. Note that in this study $\mu = \lambda + \Delta\lambda$. Figure 4 a, b, c, and d demonstrate the graphs for N=100, 200, 500, and 1000 respectively.

According to the figure, Zelterman's method gives the values of bias closer to zeros than those from McKendrick's, Mantel-Haenszel's, and maximum likelihood methods. The differences of bias between Zelterman's method and the other three methods are quite noticeable, while the differences among the three latter methods are relatively small. For higher values of lambda, all the graphs shift closer to zero, hence smaller values of bias. Further more, increasing values of delta lambda result in larger values of bias. Notice that, graphs in each population size still look very similar in shape and tendency.

Copyright by Mahidol University

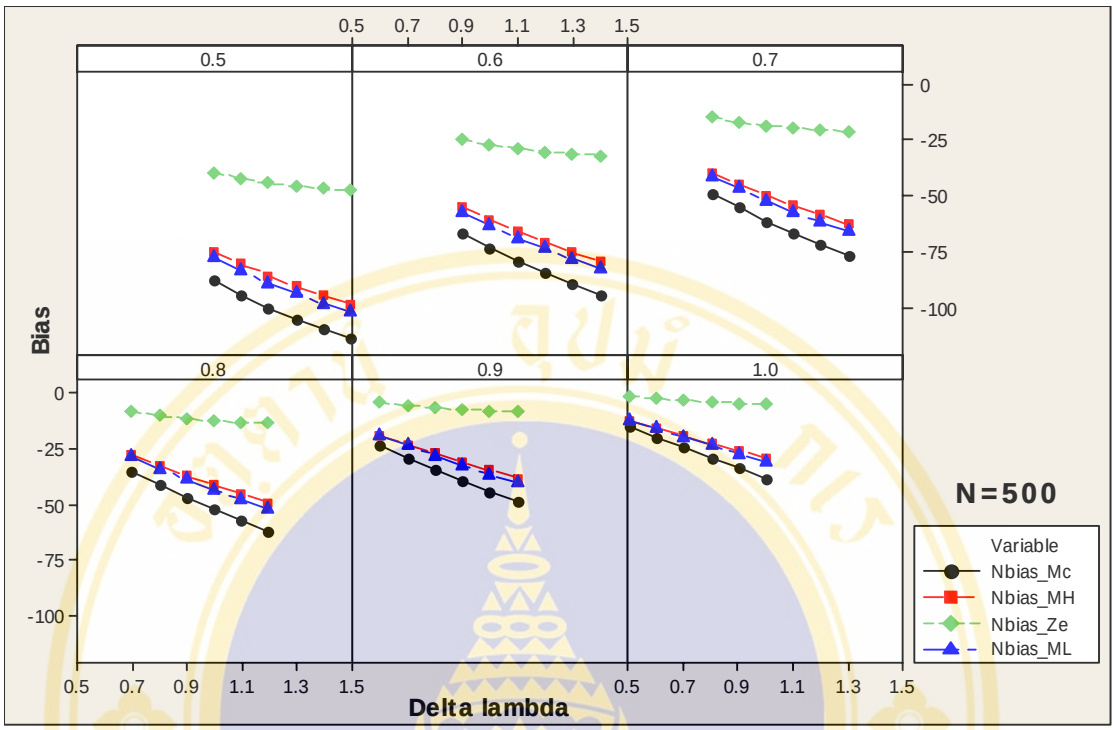


(a)

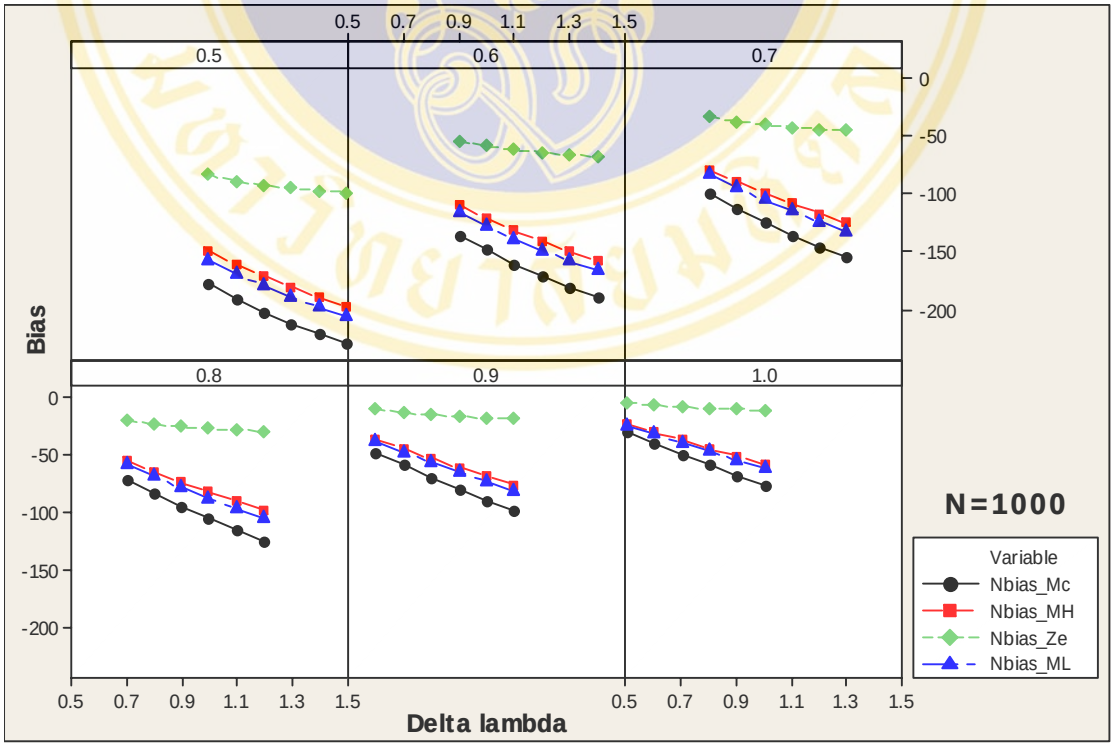


(b)

Copyright by Mahidol University



(c)



(d)

Figure 4 (a – d). Graphs of bias of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model

Copyright by Mahidol University

Ranking then will be assigned to values of bias and the means are calculated for comparison purposes. This is very practical especially for data that are close together and can not be clearly demonstrated by graph. The mean ranking of the bias of the four methods are shown in Table 4. Values of bias from Zelterman's method rank first in all but one case in population size of 100. Maximum likelihood method ranks second, 1.97(0.028), in small population of 100. But for larger population size Mantel-Haenszel's method ranks slightly better, thus smaller values of bias. Values of bias from McKendrick's method rank forth in all cases for truncated Poisson mixture model.

Table 4. Mean ranking of bias of the four population size estimators (SE) under truncated Poisson mixture model

$p = 0.5$	Mean ranking of bias of Estimator (SE)			
N	McK	MH	Zel	MLE
100	3.69 (.087)	3.28 (.076)	1.06 (.056)	1.97 (.028)
200	3.97 (.028)	2.47 (0.93)	1 (0)	2.56 (.084)
500	4 (0)	2.03 (.028)	1 (0)	2.97 (.028)
1000	4 (0)	2 (0)	1 (0)	3 (0)
Total	3.92 (.025)	2.44 (.053)	1.01 (.014)	2.62 (.042)

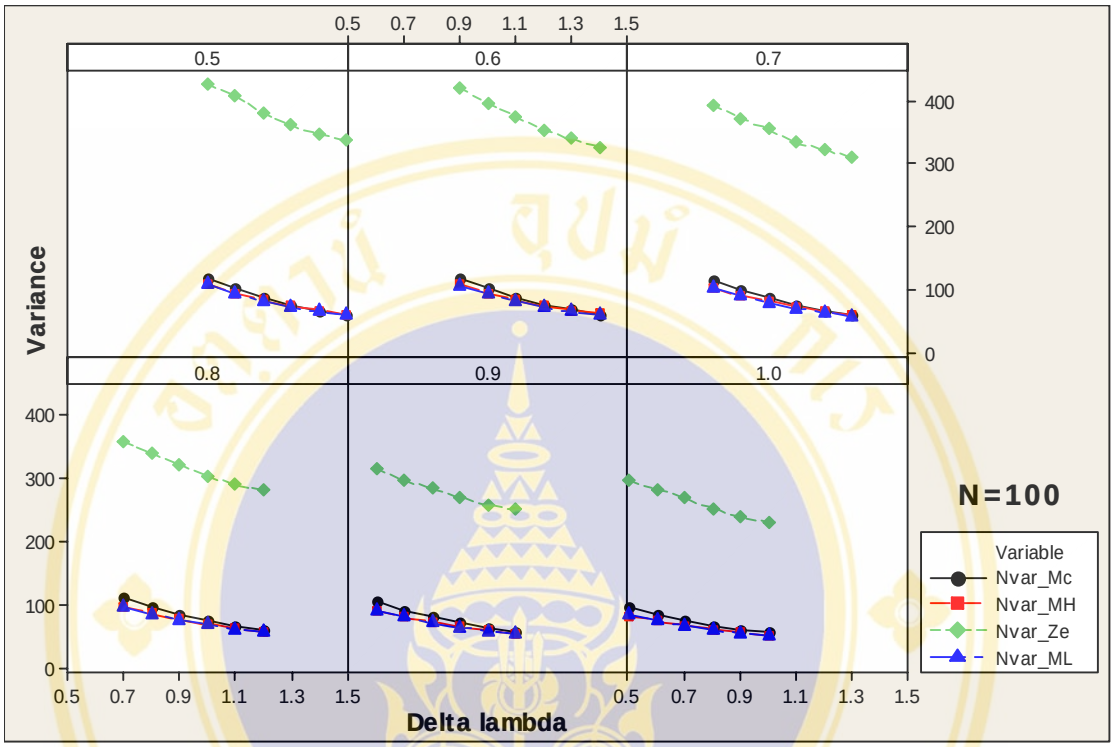
b.) Investigation of variance

The values of variance are plotted versus delta lambda for each lambda and population sizes in Figure 5 a, b, c, and d.

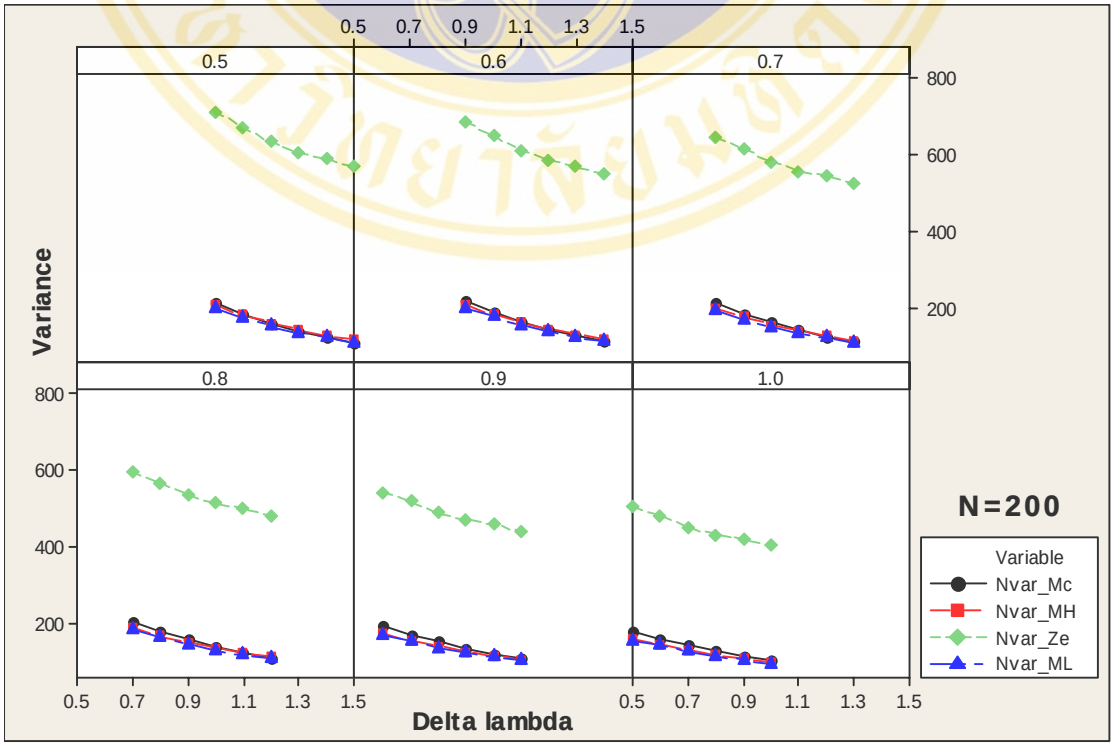
According to Figure 5, variances from Zelterman's method are clearly higher than variances from McKendrick's, Mantel-Haenszel's and maximum likelihood methods. Variances from the latter three methods are very close together. The differences should be clearer when we assign each value a ranking later on. For a specific population size and lambda, the larger values of delta lambda, the smaller variances of population size. For higher values of lambda, variances from Zelterman's method decrease gradually, while variances from McKendrick's, Mantel-Haenszel's

Copyright by Mahidol University

and maximum likelihood methods do not differ much. Finally, all graphs are still in the same shape and tendency for different population sizes

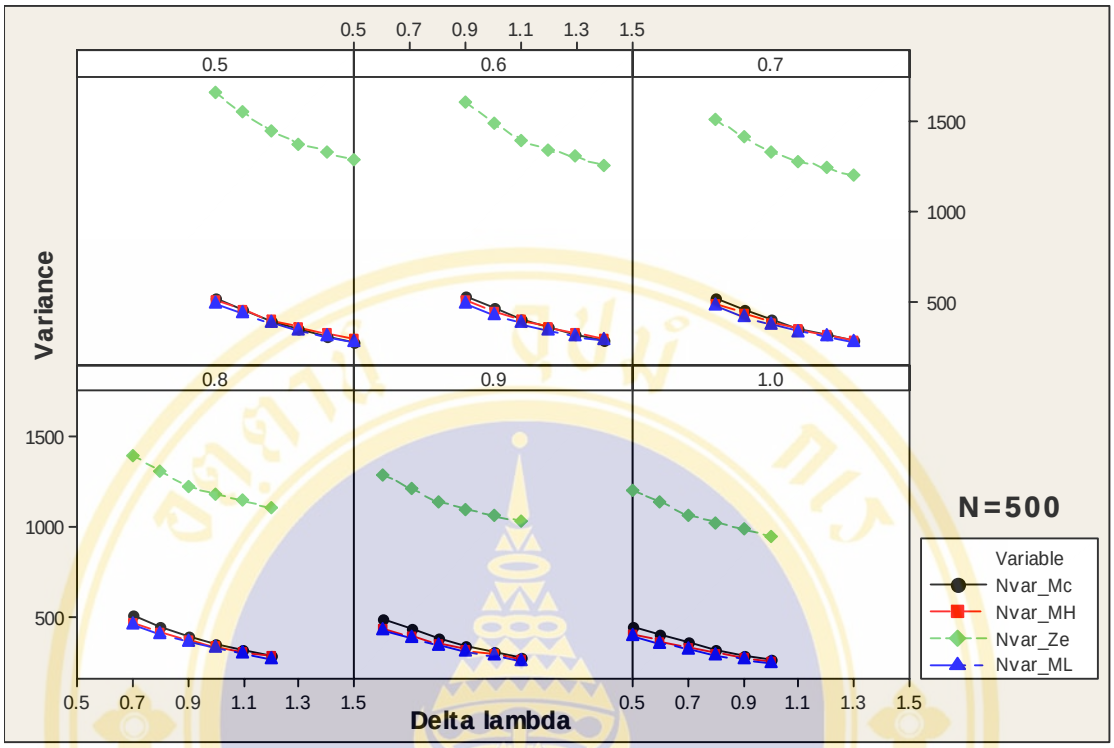


(a)

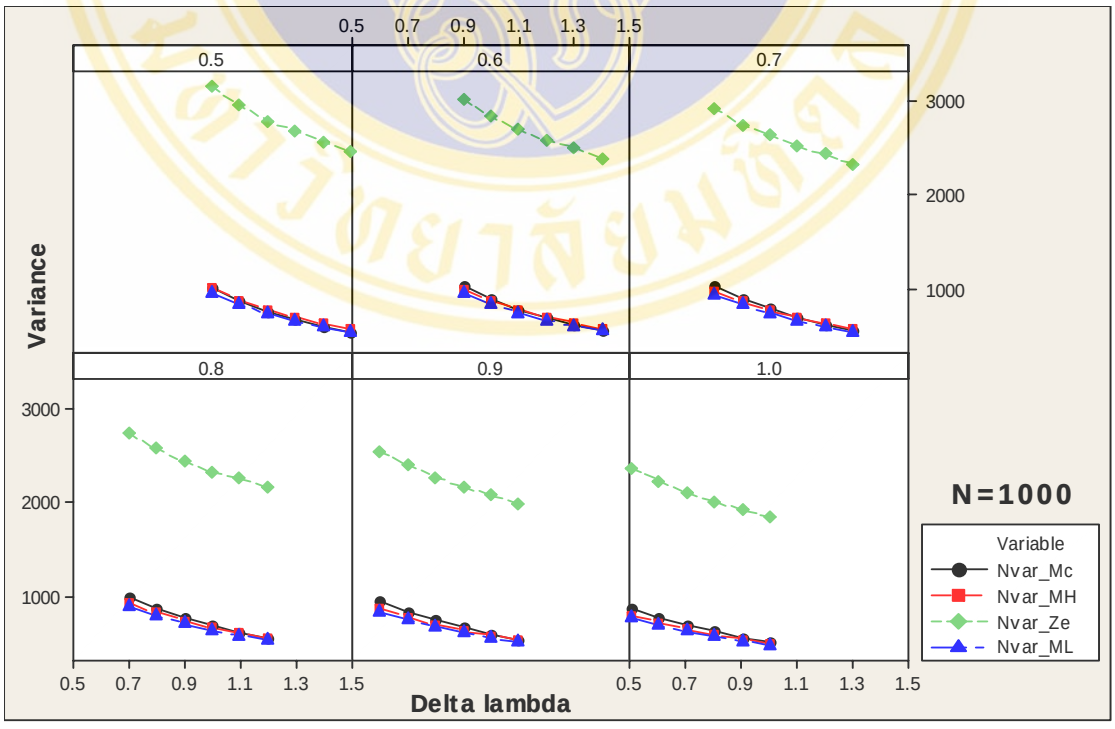


(b)

Copyright by Mahidol University



(c)



(d)

Figure 5 (a - d). Graphs of variance of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model

Copyright by Mahidol University

Ranking system is applied to the variance as in the preceding cases. Mean ranking of variance from each method are shown in Table 5. Variances from maximum likelihood dominate in all population sizes with the mean ranking value of 1.06(0.019) following by variances from Mantel-Haenszel's and McKendrick's methods. As seen previously from the graphical demonstrations, variances from Zelterman's method rank forth in all cases and all population sizes.

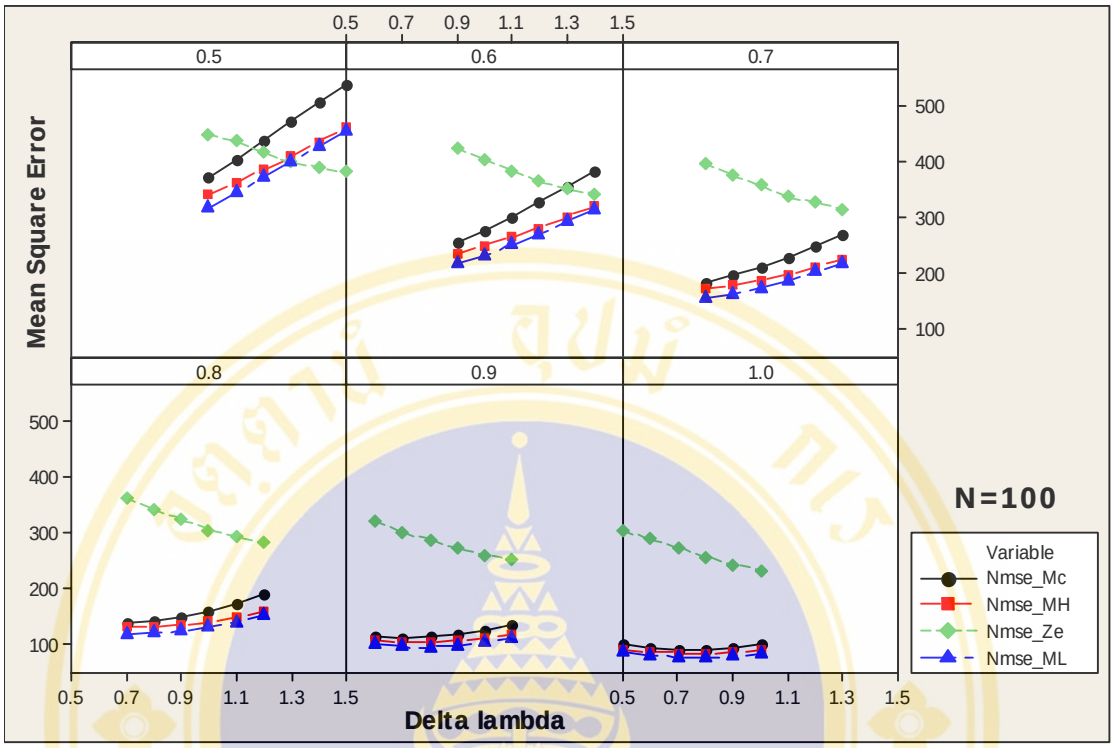
Table 5. Mean ranking of variance of the four population size estimators (SE) under truncated Poisson mixture model

$p = 0.5$	Mean ranking of variance of Estimator (SE)			
N	McK	MH	Zel	MLE
100	2.92 (.047)	2 (.069)	4 (0)	1.08 (.047)
200	2.67 (.098)	2.28 (.076)	4 (0)	1.06 (.039)
500	2.69 (.087)	2.28 (.076)	4 (0)	1.03 (.028)
1000	2.61 (.100)	2.33 (.080)	4 (0)	1.06 (.039)
Total	2.72 (.043)	2.22 (.039)	4 (0)	1.06 (.019)

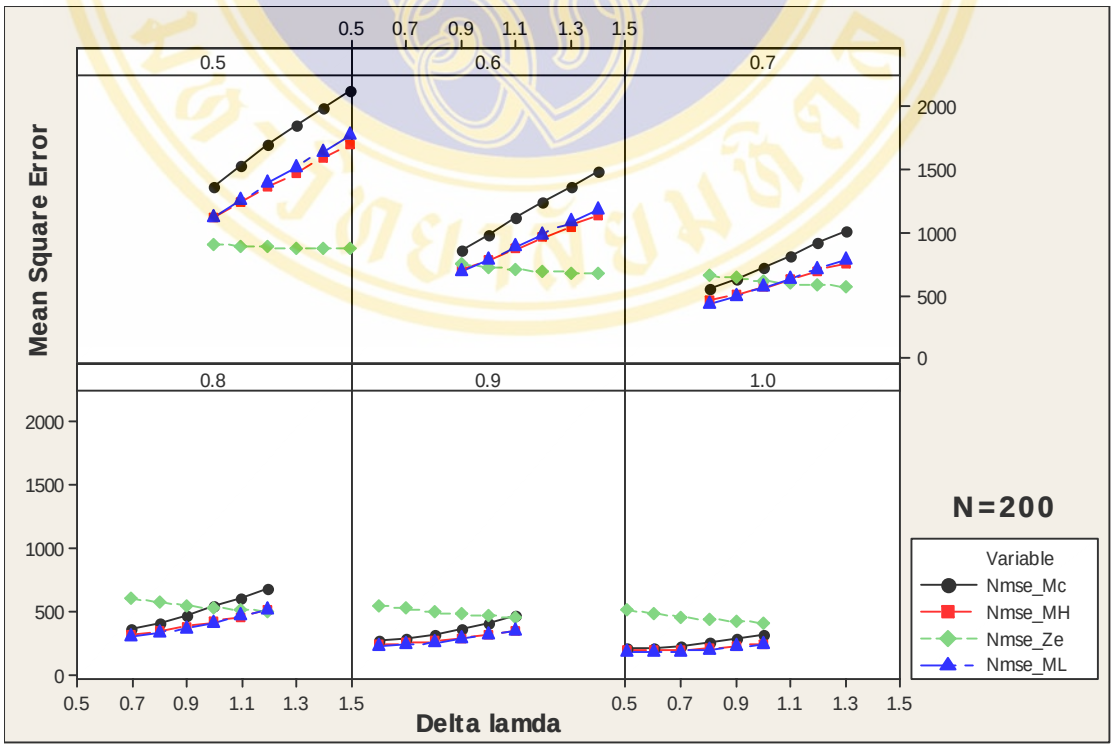
c.) Investigation of mean square error

The mean square errors from each method are plotted versus delta lambda for each lambda and population sizes in Figure 6 a, b, c, d. Mean square errors from Zelterman's method are relatively higher than those from the other three methods at small population sizes. However, for larger population sizes, mean square errors from Zelterman's method become relatively smaller. For each population size, when delta lambdas increase, mean square errors from McKendrick's, Mantel-Haenszel's and maximum likelihood are all increasing significantly, but the mean square errors from Zelterman's method are decreasing, especially for smaller population sizes. For higher values of lambda, mean square errors from each of the four methods decrease and the values from all methods are closer together, especially for larger population sizes.

Copyright by Mahidol University

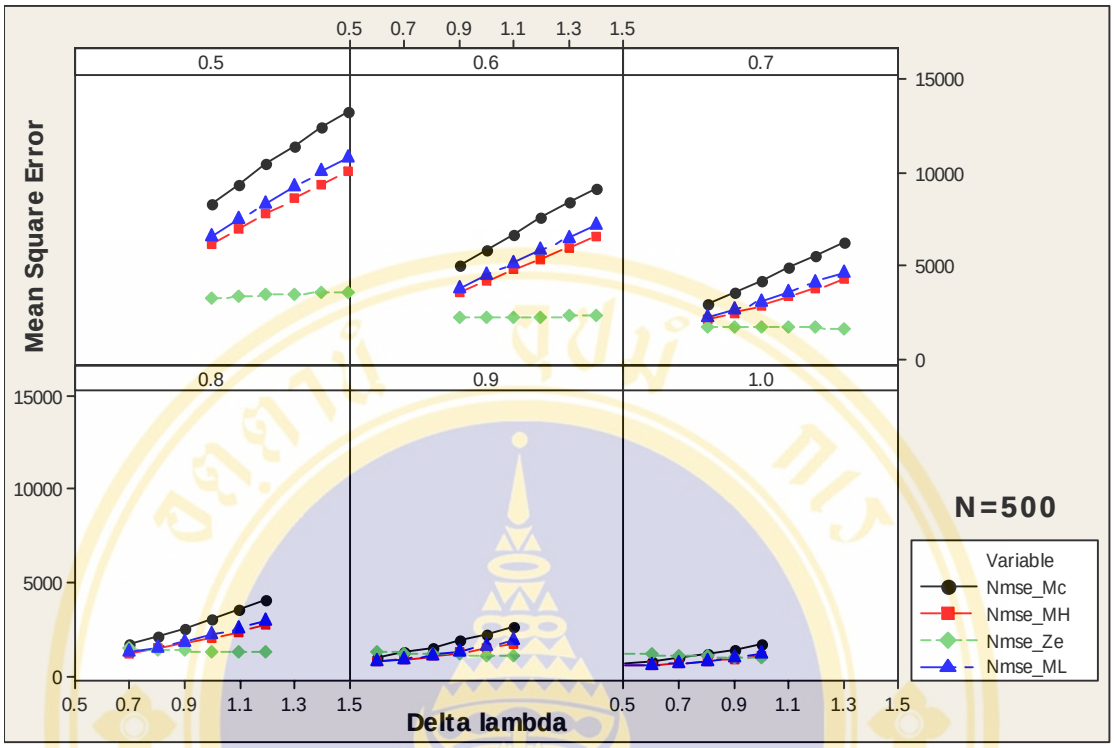


(a)

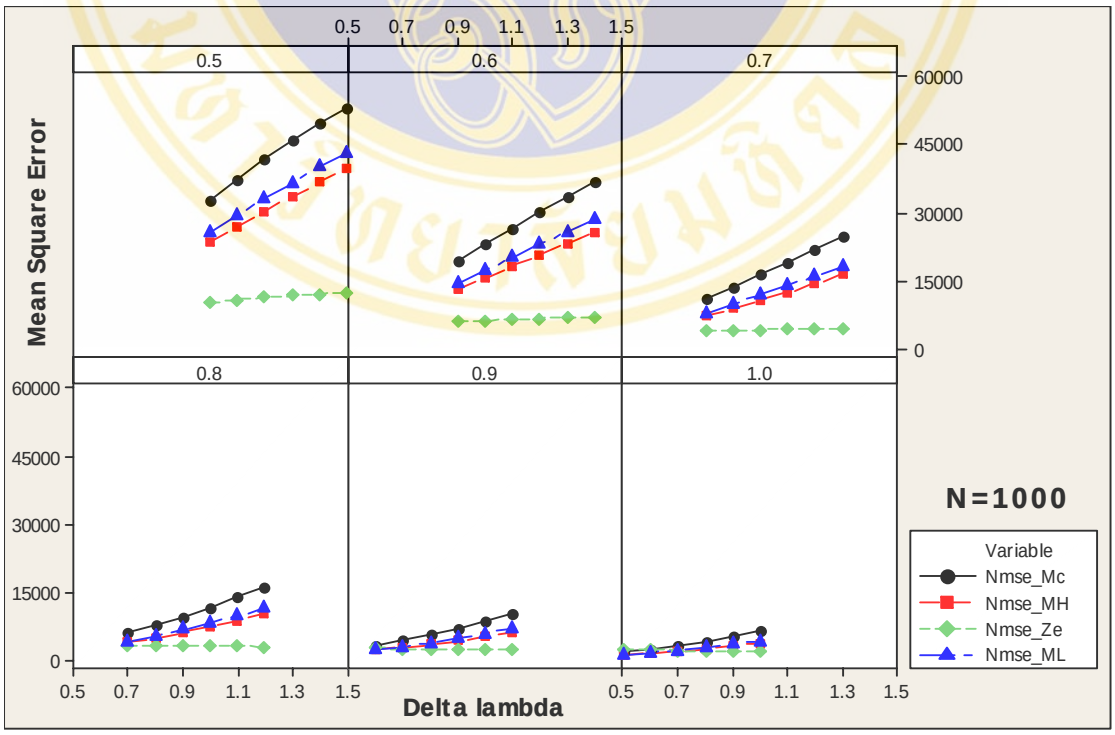


(b)

Copyright by Mahidol University



(c)



(d)

Figure 6 (a – d). Graphs of mean square error of population size estimator from McKendrick's, Mantel-Haenszel's, Zelterman's and maximum likelihood method under truncated Poisson mixture model

Copyright by Mahidol University

Mean square errors from each of the four methods are then ranked and assigned ranking values accordingly. The mean ranking of mean square error are displayed in Table 6. Maximum likelihood method gives the best mean ranking of mean square errors for small population sizes (N=100 and 200) following by Mantel-Haenszel's, McKendrick's, and Zelterman's methods. For larger population size, mean ranking from Zelterman's method become smaller relative to the other three methods, and mean square errors from McKendrick's method maintain highest mean ranking values.

Table 6. Mean ranking of MSE of the four population size estimators (SE) under truncated Poisson mixture model

$p = 0.5$	Mean ranking of MSE of Estimator (SE)			
N	McK	MH	Zel	MLE
100	3.17 (.063)	2.08 (.047)	3.69 (.131)	1.06 (.039)
200	3.56 (.084)	1.97 (.049)	2.61 (.237)	1.86 (.160)
500	3.89 (.053)	1.83 (.063)	1.61 (.184)	2.67 (.105)
1000	3.97 (.028)	1.89 (.053)	1.22 (.113)	2.92 (.047)
Total	3.65 (.040)	1.94 (.027)	2.28 (.117)	2.13 (.079)

4.2 Results Emphasizing on Application

This section emphasizes on some application results from two sets of data, one represents truncated Poisson model and the other represents truncated Poisson Mixture model.

The cholera data was chosen to illustrate an example for truncated Poisson model. McKendrick (1926) applied the Poisson model to an epidemic of cholera in an Indian village. The data are given in Table 7, where j represents the number of cases an individual house has experienced and $f(j)$ is the corresponding observed number of such house. Population sizes are estimated by McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood method. Note that for $j = 0$, a household might

Copyright by Mahidol University

be infected but did not develop symptom, hence to estimate the size of population that was infected, $f(0)$ must be truncated.

Table 7 (a) and (b) show the chi-square goodness of fit test for cholera data. The chi-square goodness of fit is 0.03368, with 1 degrees of freedom and p-value is 0.8544, hence the null hypothesis is accepted at $\alpha = 0.05$. This implies that the cholera data is truncated Poisson distribution.

Table 7.
(a) Observed and expected counts of the number of cholera data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method

j	f_j	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	32	0.591462	32.530404
2	16	0.287500	15.812509
3	6	0.093166	5.124138
4	1	0.022643	1.245381
≥ 5	0	0.004403	0.242144
N	55		

(b) After combining the last three classes to make each expected frequency greater than or equal to 5

j	f_j	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	32	0.591462	32.530404
2	16	0.287500	15.812509
3	7	0.120212	6.611663
n	55		$\chi^2 = 0.03368$

$p - value = 0.8544$

Table 8 and Figure 7 show estimated population sizes within a 95% confidence interval for each estimating method. Estimated population sizes from each

Copyright by Mahidol University

method are close together and correspond 95% confidence intervals are comparable, except for that from Zelterman's method which is quite large.

Table 8. Estimated population sizes for cholera data within 95% confidence interval

Estimators	Estimated population size	95% CI
McKendrick's	91	(67,115)
Mantel-Haenszel's	88	(67,110)
Zelterman's	88	(65,134)
Maximum Likelihood	89	(67,112)

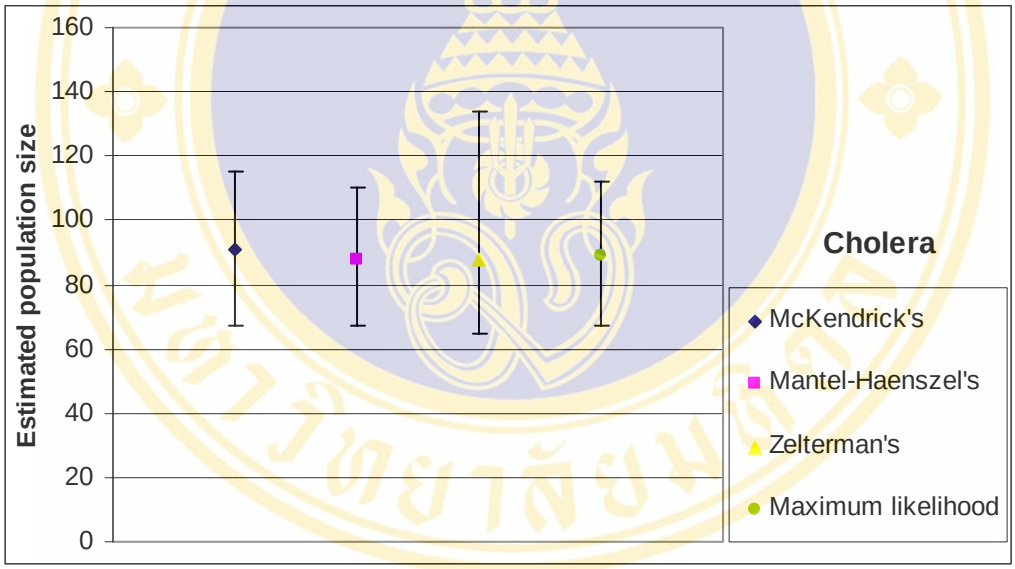


Figure 7. Estimated population sizes for cholera data within 95% confidence interval

The data of "drug users in Bangkok metropolitan region" are chosen to illustrate an example of application for truncated Poisson mixture model. A population-based study was conducted utilizing all data on treatment episodes of drug users from all 61 health treatment centers in the Bangkok metropolitan region to estimate the size of drug users. The data source covered the period from October 1 to December 31, 2001. Numbers of unobserved drug users are estimated by McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood methods.

Copyright by Mahidol University

Based upon the frequency of treatment episodes of each patient, a count distribution arose which could be well modeled by means of a Poisson mixture model (23).

Table 9 (a) and (b) display count data at each frequency of treatment of heroin user and chi-square goodness of fit test for its distribution. The chi-square goodness of fit is 8502.69077, with 7 degrees of freedom and p-value is less than 0.00001, hence the null hypothesis is rejected at $\alpha = 0.05$. So the data of heroin user are not a truncated Poisson distribution.

Table 9.
(a) Observed and expected counts of the number of heroin user data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method

Frequency of treatment	Count frequency	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	2955	0.187630637366	1325.047561081
2	1186	0.258117745090	1822.827515828
3	803	0.236723139553	1671.738811524
4	611	0.162826401514	1149.880047490
5	416	0.089598125743	632.741963996
6	338	0.041085802947	290.147940413
7	278	0.016148710285	114.042192033
8	180	0.005553822051	39.221091323
9	125	0.001697827468	11.990057576
10	74	0.000467129893	3.298871302
11	38	0.000116839336	0.825119388
12	20	0.000026788718	0.189181929
13	14	0.000005669604	0.040038740
14	11	0.000001114214	0.007868582
15	4	0.000000204372	0.001443276
16	1	0.000000035144	0.000248184
17	3	0.000000005688	0.000040167
18	4	0.000000000869	0.000006140
23	1	0.000000005959	0.000042086
	7062		

Copyright by Mahidol University

(b) After combining the last eleven classes to make each expected frequency greater than or equal to 5

Frequency of treatment	Count frequency	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	2955	0.187630637366	1325.047561081
2	1186	0.258117745090	1822.827515828
3	803	0.236723139553	1671.738811524
4	611	0.162826401514	1149.880047490
5	416	0.089598125743	632.741963996
6	338	0.041085802947	290.147940413
7	278	0.016148710285	114.042192033
8	180	0.005553822051	39.221091323
9	295	0.002315621264	16.352917369
	7062		$\chi^2 = 8502.69077$

$p - value < 0.00001$

Table 10 (a) and (b) show count data at each frequency of treatment of methamphetamine user and chi-square goodness of fit test for its distribution. The chi-square goodness of fit is 163.19773, with 1 degrees of freedom and p-value is less than 0.00001, hence the null hypothesis is rejected at $\alpha = 0.05$. Therefore the data of methamphetamine user are not a truncated Poisson distribution.

Copyright by Mahidol University

Table 10. Observed and expected counts of the number of methamphetamine user data assuming a truncated Poisson distribution with λ estimated from data by using maximum likelihood method

Frequency of treatment	Count frequency	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	3114	0.8818589831644840	2950.700157668360
2	163	0.1086405292450500	363.511210853938
3	23	0.0089226394241940	29.855151513353
4	20	0.0005496118357998	1.839001202586
5	9	0.0000270837500601	0.090622227701
6	3	0.0000011121932934	0.003721398760
7	3	0.0000000391475833	0.000130987814
8	3	0.0000000012056957	0.000004034258
9	4	0.0000000000330079	0.000000110444
10	3	0.0000000000008133	0.000000002721
12	1	0.0000000000000041	0.000000000014
	3346		

(b) After combining the last nine classes to make each expected frequency greater than or equal to 5

Frequency of treatment	Count frequency	$\hat{p}_j = \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{(1-e^{-\hat{\lambda}})\, j!}$	$n\hat{p}_j$
1	3114	0.8818589831644840	2950.700157668360
2	163	0.1086405292450500	363.511210853938
3	69	0.0095004875904516	31.788631477651
	3346		$\chi^2 = 163.19773$

$p - value < 0.00001$

The results of estimated number of drug user are shown in Table 11, Figure 8, and Figure 9. The estimated population sizes are displayed separately for heroin and methamphetamine users. Estimated population sizes from McKendrick's, Mantel-Haenszel's, and maximum likelihood method are considerably close together with

Copyright by Mahidol University

narrow 95% confidence intervals. However, Zelterman's method gives significantly higher estimated population size and wider 95% confidence intervals.

Table 11. Estimated number of heroin and methamphetamine users and corresponding 95% confidence interval

Estimators	Estimated population size			
	Heroin users	95% CI	Methamphetamine users	95% CI
McKendrick's	7182	(7160,7205)	8269	(7825,8713)
Mantel-Haenszel's	7680	(7623,7738)	18661	(16808,20515)
Zelterman's	12797	(12515,13091)	33664	(27797,42672)
Maximum Likelihood	7544	(7495,7594)	15322	(13987,16658)

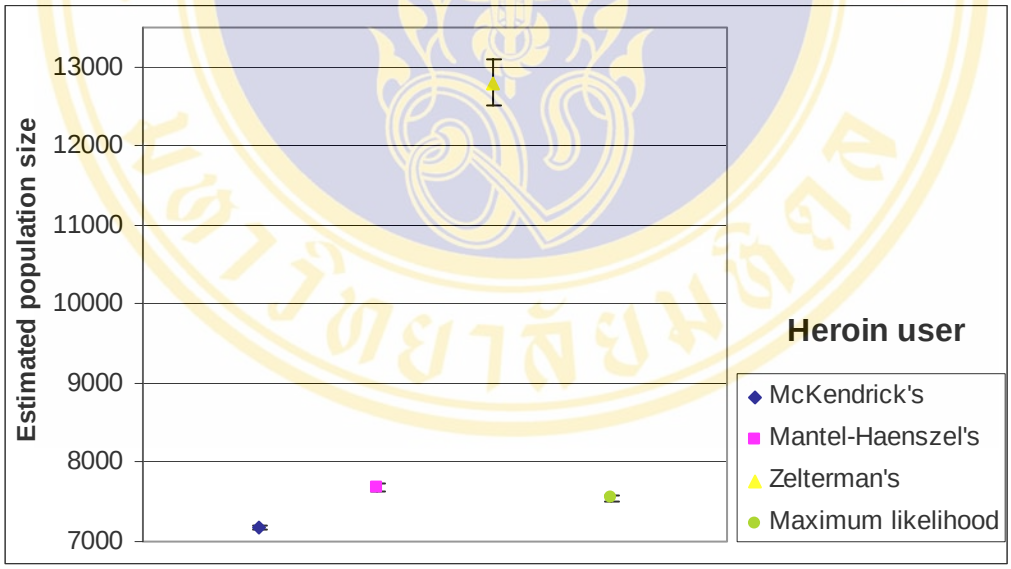


Figure 8. Estimated number of heroin users within a 95% confidence interval

Copyright by Mahidol University

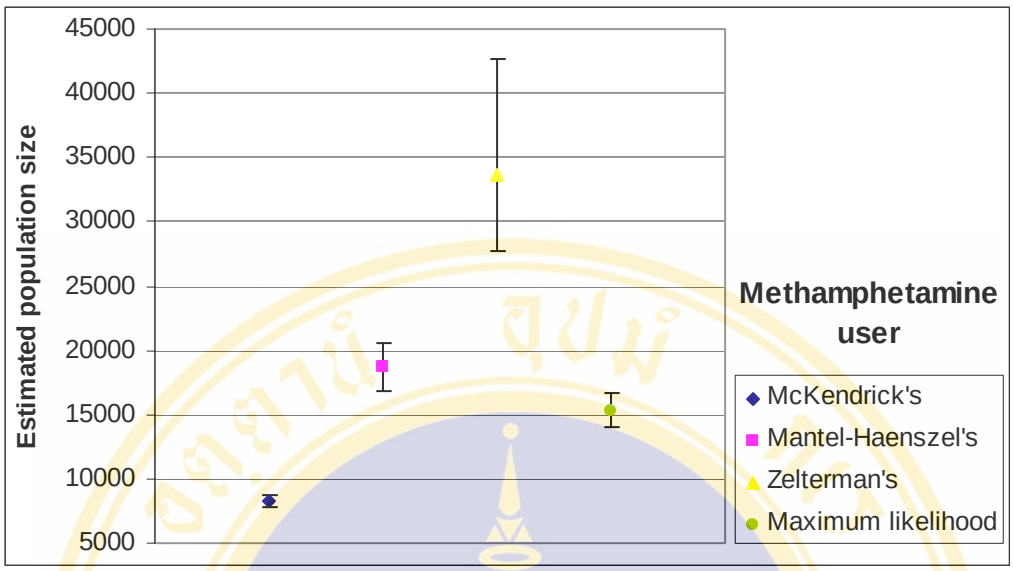


Figure 9. Estimated number of methamphetamine users within a 95% confidence interval

Copyright by Mahidol University

CHAPTER V
DISCUSSION

The discussion for this study is divided into two parts. The first part is concentrating on the simulation results, and the second part deals with the results from applications.

5.1 Discussion on the Simulation Results

The simulation generates two sets of data, each set represents each type of distribution. For $p = 1$, simulated data has the distribution of the Poisson type and for $p = 0.5$ it becomes Poisson mixture model. Results are discussed for each type of distribution separately. In each distribution, population sizes estimators from McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood method are compared and discussed according to the criteria mentioned in Chapter III. Mean rankings of bias, variance, and mean square error are considered to determine the best method for estimate population size that fits Poisson model and Poisson mixture model.

5.1.1 Truncated Poisson model

According to the mean ranking of bias of population size estimators in Table 1, Mantel-Haenszel's method impressively provides the best bias for estimators, following by maximum likelihood method, while Zelterman's method has the worst bias among all four methods. This claim is valid for all population sizes.

From Table 2, Mantel-Haenszel's and maximum likelihood methods provide comparable variance. However, if we look closely at each population size, Mantel-

Copyright by Mahidol University

Haenszel's method gives better variance at smaller population sizes while maximum likelihood method does better at the larger ones. This is surprising because of the knowledge that the maximum likelihood estimator is sufficient statistic. McKendrick's and Zelterman's method still last third and fourth in variance regardless of variation in population sizes.

Mean ranking of the mean square error interprets very similarly to that from mean ranking of variance. Mantel-Haenszel's method gives outstanding mean square error in small population sizes while maximum likelihood does the same for larger population sizes. It is evident that Mantel-Haenszel's method is the most powerful method to estimate population size of the population that fits the Poisson model. However, maximum likelihood method is an acceptable alternative if large population sizes are to be expected. This being said, Zelterman's method is most inappropriate to estimate the population sizes of the Poisson distribution type.

5.1.2 Truncated Poisson mixture model

According to Table 4, which demonstrates mean ranking of bias of population size estimator from each method, Zelterman's method significantly has the best bias among the four methods. Hence, Zelterman's method provides the estimated population size closest to the true value of population size in all cases. Maximum likelihood method follows in second place for small population size, while Mantel-Haenszel's method gives better estimate for large population sizes. McKendrick's method gives the worst estimator.

From Table 5, maximum likelihood shows minimum variance for all population sizes, followed by Mantel-Haenszel's and McKendrick's methods. Though not, Zelterman's method has maximum variance considering its ability to estimate good population size mentioned earlier. This may be because population size estimated by Zelterman's method for each of 10,000 replications contributes to high spread in value, hence high variance.

Mantel-Haenszel's, Zelterman's, and maximum likelihood methods have comparable mean square errors according to Table 6, while McKendrick's method has the highest value of mean square error.

Copyright by Mahidol University

Despite the variance, Zelterman's method has the best performance of estimating population size for population that fits Poisson mixture model, since it has magnificent small bias. This result is in agreement with Zelterman (21) who claimed that this method is suitable for sampling under a truncated Poisson mixture distribution.

5.2 Discussion on the Application Results

From previous discussion, we already know which estimator is most suitable for each type of population distribution. This section we look forward to discuss some examples demonstrated by this knowledge and how to be able to choose one method that suit each application.

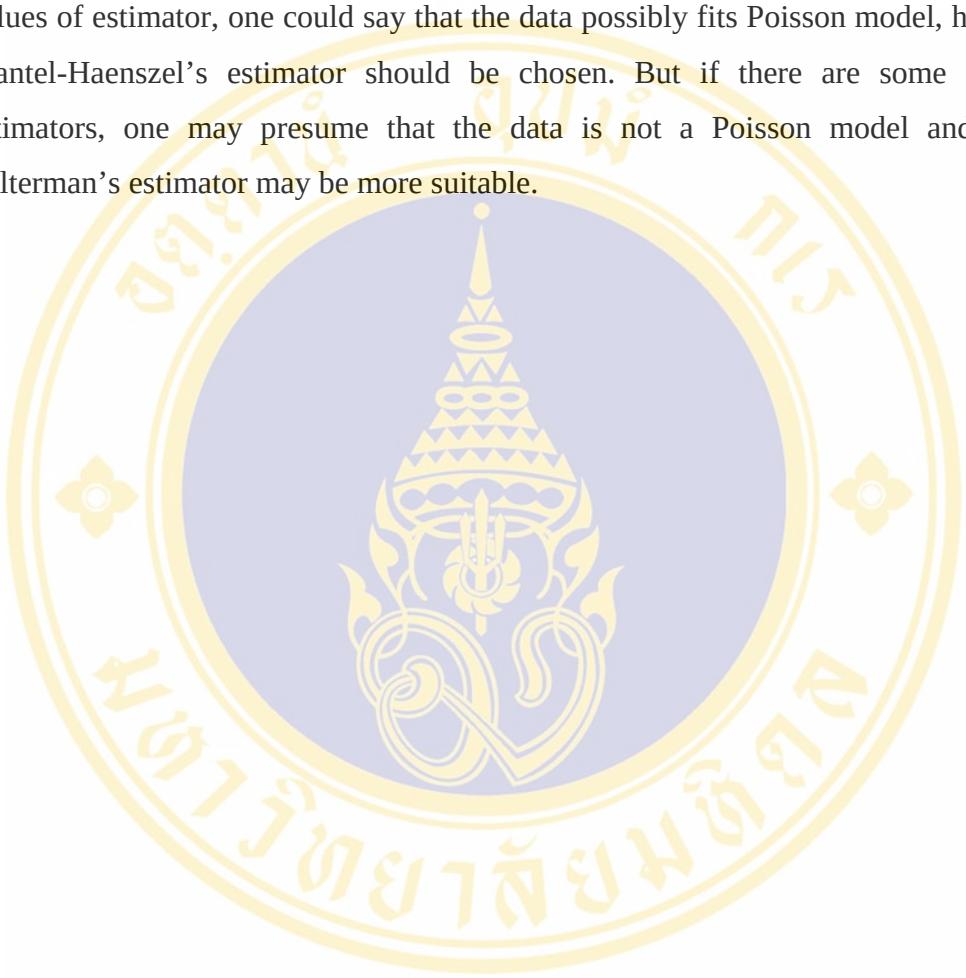
Table 8 shows results of estimated population size from the four methods. According to McKendrick (20), the cholera data observed in an Indian village is well demonstrated by a truncated Poisson distribution, and chi-square goodness of fit test proves so (Table 7 (a, b)). Hence, from the discussion on the simulation results, we can assume that Mantel-Haenszel's method is the most suitable estimator for this data. Therefore, the estimated population size for the cholera data is 88 within a 95% confidence interval of 67 and 110, which is narrowest among the others.

Table 11 demonstrates population size estimators for heroin and methamphetamine users in Bangkok in 2001 respectively. The drug user data were used by Böhning (24) to be best fitted by a Poisson mixture model, which agrees with chi-square goodness of fit test (Table 9 (a, b) and Table 10 (a, b)). Hence, from the knowledge from simulation results, we may safely declare that Zelterman's method is most suitable to estimate population size for this data. This result agrees with Böhning's work for well separated components of the Poisson mixture model, which has $\hat{N} = 33{,}664$ for the methamphetamine users and $\hat{N} = 12{,}796$ for heroin users.

From these two applications, one may notice that McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood method yield similar population size estimators for data that fit Poisson model. On the other hand, for data that fits

Copyright by Mahidol University

Poisson mixture model, there may be one or more estimators that stand out among the others. This is very useful for estimating population size of data which distribution is yet not known. For example, after calculating estimated population sizes of a population employing the previous four methods, if all four methods have similar values of estimator, one could say that the data possibly fits Poisson model, hence the Mantel-Haenszel's estimator should be chosen. But if there are some standout estimators, one may presume that the data is not a Poisson model and hence, Zelterman's estimator may be more suitable.



Copyright by Mahidol University

CHAPTER VI
CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

This study focused on comparing population size estimators via Horvitz-Thompson's approach from McKendrick's, Mantel-Haenszel's, Zelterman's, and maximum likelihood methods. The four methods are used to estimate truncated Poisson estimator of population size of data simulated by the aid of Microsoft FORTRAN programming. The data are simulated under Poisson model and Poisson mixture model. The conclusions of result are:

- Under the truncated Poisson model, Mantel-Haenszel's method has the most accurate estimated population size providing its minimum bias, variance, and mean square error.
- Under the truncated Poisson mixture model, Zelterman's method provides the best estimated population size according to the minimum bias and acceptable variance and mean square error.
- Population size estimators from the four methods are found to be close together for the data of a truncated Poisson distribution type. This was well illustrated by the cholera data.
- We also found that for a truncated Poisson mixture model, there can be one or two population size estimators, especially for that of Zelterman's, which have values further apart from the others.

Note that the results of this study are based on the scope of study mentioned at the beginning.

Copyright by Mahidol University

6.2 Recommendations

The recommendations for further study:

- This study focused on the zero-truncated Poisson distribution. However, truncation on the left and right hand of a distribution can lead to a practical study.
- Further study on population size estimator for truncated Negative Binomial distribution can be pursued.
- This study can be further extended for higher components of Poisson mixture model.
- Extended study on Mantel-Haenszel's estimator is suggested.
- Further study on the large variance obtained from Zelterman's estimator should be considered.

Copyright by Mahidol University

REFERENCES

1. Wickens, T.D. Quantitative methods for estimating the size of a drug-using population. The Journal of Drug Issues 1993; 23(2): 185-216.
2. Levy, P.S. and Lemeshow, S. **Sampling of Populations Methods and Applications.** 3rd ed. New York: John Wiley & Sons. 1999.
3. Heang, W. Comparative estimating population size when using capture-recapture method [M.S. Thesis in Statistics]. Bangkok: Faculty of Graduate Studies, Chulalongkorn University, 1995.
4. Long, J.S. **Advance qualitative techniques in the social science: Regression models for categorical and limited dependent variables.** 7th ed. London: International education and professional, 1977.
5. Van der Heijden, P.G.M. Estimating the size of a criminal population from police record using the truncated Poisson regression model. Statistica Neerlandica, 2003; 57(3): 289-304.
6. Van der Heijden, P.G.M., Bustami, R., Cruijff, M., Engbersen, G., and Van Houwelingen, H. Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modeling 2003; 3(4): 305-322.
7. Böhning, D., Kuhnert, R., Dietz, E., Schön, D. Mixture models for capture-recapture data based upon the counting sources distribution. Statistical Methods and Applications, 2005. [to appear]
8. Hay, G. and Smit, F. Estimating the number of drug injectors from needle exchange data. Addiction Research and Theory 2003; 11(4): 235-243.
9. Suppawattanabodee, B. Estimating the number of drug users in Bangkok: A capture-recapture method. [M.S. Thesis in Biostatistics]. Bangkok: Faculty of Graduate Studies, Mahidol University, 2003.
10. Thieraut, R.J., Klekamp, R.C. **Decision making through operations research.** 2nd ed. Singapore: John Wiley & Sons, 1975.

Copyright by Mahidol University

11. Rubinstein, R.Y. **Simulation and the Monte Carlo Method.** New York: John Wiley & Sons, 1981.
12. Hahn, B. **FORTRAN 90 for Scientists and Engineers.** Cambridge: University Press, 1994.
13. Vogt, W.P. **Dictionary of Statistics and Methodology: A non-technical guide for the social sciences.** California: Sage, 1993.
14. Birkes, D. and Dodge, Y. **Alternative Methods of Regression.** New York: John Wiley & Sons, 1993.
15. Kendall, M.G. and Buckland, W.R. **A Dictionary of Statistical Terms.** New York: Hafner Publishing Company. 1971.
16. Böhning, D. and Schön, D. A Monotonicity property for the nonparametric maximum likelihood estimator of the population size based upon the counting distribution. Journal of the Royal Statistical Society, Series C (Applied Statistics), 2005. [to appear]
17. Hogg, R.V. and Klugman, S.A. **Loss Distribution.** New York: John Wiley & Sons, 1984.
18. Grogger, J.T. and Carson, R.T. Model for Truncated Count. Journal of Applied Econometrics 1991; 6(3): 225-238.
19. Meng, X.L. The EM algorithm and medical studies: a historical link. Statistical Methods in Medical Research 1997; 6: 3-23.
20. Zelterman, D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. Journal of Statistical Planning and Inference 1988; 18: 225-237.
21. Moore, P.G. The estimation of the Poisson parameter from a truncated distribution. Biometrika 1952; 39(3/4): 247-251.
22. Moore, P.G. A note on truncated Poisson distribution. Biometrics 1954; 10: 402-406.
23. Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W., and Viwatwongkasem, C. Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. European Journal of Epidemiology, 2004; 19: 1075-1083.

Copyright by Mahidol University

24. Böhning, D. and Seidel, W. Editorial: Recent developments in mixture models. Computational Statistics & Data Analysis 2003; 41: 349-357.
25. Titterington, D.M., Smith, A.F.M., Makov, U.E. **Statistical Analysis of Finite Mixture Distributions.** Chichester: John Wiley & Sons, 1985.
26. Tamhane, A.C. and Dunlop, D.D., **Statistics and Data Analysis from Elementary to Intermediate.** New Jersey: Prentice-Hall. 2000.
27. David, F.N. and Johnson, N.L. The Truncated Poisson. Biometrics 1952; 8: 275-285.
28. Plackett, R.L. The truncated Poisson distribution. Biometrics 1953; 9: 485-488.
29. Rider, P.R. Truncated Poisson distributions. Journal of the American Statistical Association. 1953; 48(264): 826-830.
30. Tate, R.F. and Goen, R.L. Minimum variance unbiased estimation for the truncated Poisson distribution. The Annals of Mathematical Statistics. 1958; 29(3): 755-765.
31. Dietz, E. and Böhning, D. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis. 2000; 34: 441-459.

Copyright by Mahidol University

APPENDIX

Copyright by Mahidol University

APPENDIX
FORTRAN SUBROUTINES USED IN THIS STUDY

1. Main program

```
PROGRAM  main

IMPLICIT     none
INTEGER      a, i, j, n1, n2, m, ncs, k, R, Z, iseed, f(10000), IR(10000), IR1(10000),
             IR0(10000), X(10000)
INTEGER      LvarMcPt, LvarMHPt, LvarZePt, LvarMLPt
INTEGER      LbiaMcPt, LbiaMHPt, LbiaZePt, LbiaMLPt, LmseMcPt, LmseMHPt,
             LmseZePt, LmseMLPt
INTEGER      NvarMcPt, NvarMHPt, NvarZePt, NvarMLPt
INTEGER      NbiaMcPt, NbiaMHPt, NbiaZePt, NbiaMLPt, NmseMcPt,
             NmseMHPt, NmseZePt, NmseMLPt
REAL         P(100), lamda(100), dlamda(100), u(100), explam(100)
REAL         lamda_Mc, Nhat_Mc
REAL         lamda_MH, Nhat_MH
REAL         lamda_Ze, Nhat_Ze
REAL         lamda_ML, Nhat_ML
REAL         Rlam_Mc(10000), RNhat_Mc(10000)
REAL         Rlam_MH(10000), RNhat_MH(10000)
REAL         Rlam_Ze(10000), RNhat_Ze(10000)
REAL         Rlam_ML(10000), RNhat_ML(10000)

REAL         LmeanMc, LbiasMc, LvarMc, LmseMc
REAL         LmeanMH, LbiasMH, LvarMH, LmseMH
REAL         LmeanZe, LbiasZe, LvarZe, LmseZe
```

Copyright by Mahidol University

```
REAL        LmeanML, LbiasML, LvarML, LmseML
REAL        NmeanMc, NbiasMc, NvarMc, NmseMc
REAL        NmeanMH, NbiasMH, NvarMH, NmseMH
REAL        NmeanZe, NbiasZe, NvarZe, NmseZe
REAL        NmeanML, NbiasML, NvarML, NmseML
REAL        LAmeanMc(100), LAbiasMc(100), LAvarMc(100), LAmseMc(100)
REAL        LAmeanMH(100), LAbiasMH(100), LAvarMH(100), LAmseMH(100)
REAL        LAmeanZe(100), LAbiasZe(100), LAvarZe(100), LAmseZe(100)
REAL        LAmeanML(100), LAbiasML(100), LAvarML(100), LAmseML(100)
REAL        NAmeanMc(100), NAbiasMc(100), NAvarMc(100), NAmseMc(100)
REAL        NAmeanMH(100), NAbiasMH(100), NAvarMH(100),
            NAmseMH(100)
REAL        NAmeanZe(100), NAbiasZe(100), NAvarZe(100), NAmseZe(100)
REAL        NAmeanML(100), NAbiasML(100), NAvarML(100), NAmseML(100)
CHARACTER(len=40)       name1,name2,name3,name4
EXTERNAL   RNBIN,RNSET

name1 = 'case2.txt'

print*, 'Enter the value of N:  '
read(*, *) n1
print*, 'Enter the output filename in format *.txt'
read(*, *) name2
print*, 'Enter the filename for ranked Lamda in format *.txt'
read(*, *) name3
print*, 'Enter the filename for ranked N in format *.txt'
read(*, *) name4
open(1, file = name1)
open(2, file = name2)
open(3, file = name3)
open(4, file = name4)
```

Copyright by Mahidol University

```
read(1, *) ncs

do a = 1, ncs
        read(1, *) P(a), lamda(a), dlamda(a)
end do

10      format(A4, 3A8, 36A15)
        write(2, 10)    'N', 'P', 'Lamda', 'u', 'Lmean_Mc', 'Lmean_MH', 'Lmean_Ze',
                        'Lmean_ML', 'Lvar_Mc', 'Lvar_MH', 'Lvar_Ze', 'Lvar_ML',
                        'Lbias_Mc', 'Lbias_MH', 'Lbias_Ze', 'Lbias_ML', 'Lmse_Mc',
                        'Lmse_MH', 'Lmse_Ze', 'Lmse_ML', 'Nmean_Mc',
                        'Nmean_MH', 'Nmean_Ze', 'Nmean_ML', 'Nvar_Mc',
                        'Nvar_MH', 'Nvar_Ze', 'Nvar_ML', 'Nbias_Mc', 'Nbias_MH',
                        'Nbias_Ze', 'Nbias_ML', 'Nmse_Mc', 'Nmse_MH', 'Nmse_Ze',
                        'Nmse_ML'

11      format(A4, 3A8, 12A7)
        write(3, 11)    'N', 'P', 'Lamda', 'u', 'LvMc', 'LvMH', 'LvZe', 'LvML', 'LbMc',
                        'LbMH', 'LbZe', 'LbML', 'LmMc', 'LmMH', 'LmZe', 'LmML'
        write(4, 11)    'N', 'P', 'Lamda', 'u', 'NvMc', 'NvMH', 'NvZe', 'NvML', 'NbMc',
                        'NbMH', 'NbZe', 'NbML', 'NmMc', 'NmMH', 'NmZe', 'NmML'

do a = 1, ncs
        iseed = 9999
        call RNSET(iseed)
        R = 10000

        do k = 1, R
                u(a) = lamda(a) + dlamda(a)
                m = 0
                do i = 1, n1
                        if (P(a).eq.1)  then
```

Copyright by Mahidol University

```
                call RNPOI(1, lamda(a), IR1(1))
                X(i) = IR1(1)
        else if (P(a).ne.1) then
                call RNBIN(1, 1, P(a),IR(1))
                Z = IR(1)
                if (Z.eq.1)     then
                        call RNPOI(1, lamda(a), IR1(1))
                        X(i) = IR1(1)
                else
                        call RNPOI(1, u(a), IR0(1))
                        X(i) = IR0(1)
                end if
        end if
        ! find maximum value of X(i);m
        if (X(i).le.m) then
                m = m
        else
                m = X(i)
        end if
 end do

! sort Xi to Xn1
do j = 0, m
        f(j) = 0
        do i = 1, n1
                if (X(i).eq.j) then
                        f(j) = f(j)+1
                else
                        f(j) = f(j)
                end if
        end do
end do
```

Copyright by Mahidol University

```
        n2 = n1 - f(0)

        ! calculate population size estimators
        call McK(j, m, n2, f, lamda_Mc, Nhat_Mc)
        call MH(j, m, n2, f, lamda_MH, Nhat_MH)
        call Zel(n2, m, f, lamda_Ze, Nhat_Ze)
        call MLE(j, m, n2, f, lamda_Mc, lamda_ML, Nhat_ML)

        Rlam_Mc(k) = lamda_Mc
        RNhat_Mc(k) = Nhat_Mc
        Rlam_MH(k) = lamda_MH
        RNhat_MH(k) = Nhat_MH
        Rlam_Ze(k) = lamda_Ze
        RNhat_Ze(k) = Nhat_Ze
        Rlam_ML(k) = lamda_ML
        RNhat_ML(k) = Nhat_ML
end do
explam(a) = (lamda(a)*P(a) + u(a)*(1 - P(a))) / (1 - (P(a)*exp(-lamda(a))) +
                ((1 – P(a))*exp(-u(a))))

! calculate mean, bias, variance, and MSE for lambda
call    LStat (a, k, R, P, lamda, explam, Rlam_Mc, LmeanMc, LbiasMc,
        LvarMc, LmseMc)
call    LStat (a, k, R, P, lamda, explam, Rlam_MH, LmeanMH, LbiasMH,
        LvarMH, LmseMH)
call    LStat (a, k, R, P, lamda, explam, Rlam_Ze, LmeanZe, LbiasZe,
        LvarZe, LmseZe)
call    LStat (a, k, R, P, lamda, explam, Rlam_ML, LmeanML, LbiasML,
        LvarML, LmseML)

LAmeanM(a) = LmeanMc
LAbiasMc(a) = LbiasMc
```

Copyright by Mahidol University

```
LAvarMc(a)  = LvarMc
LAmseMc(a)  = LmseMc

LAmeanMH(a) = LmeanMH
LAbiasMH(a) = LbiasMH
LAvarMH(a) = LvarMH
LAmseMH(a) = LmseMH

LAmeanZe(a) = LmeanZe
LAbiasZe(a) = LbiasZe
LAvarZe(a) = LvarZe
LAmseZe(a) = LmseZe

LAmeanML(a) = LmeanML
LAbiasML(a) = LbiasML
LAvarML(a) = LvarML
LAmseML(a) = LmseML

! calculate mean, bias, variance, and MSE for N
call    NStat (k, R, n1, RNhat_Mc, NmeanMc, NbiasMc, NvarMc, NmseMc)
call    NStat (k, R, n1, RNhat_MH, NmeanMH, NbiasMH, NvarMH,
        NmseMH)
call    NStat (k, R, n1, RNhat_Ze, NmeanZe, NbiasZe, NvarZe, NmseZe)
call    NStat (k, R, n1, RNhat_ML, NmeanML, NbiasML, NvarML,
        NmseML)

NAmeanMc(a) = NmeanMc
NAbiasMc(a) = NbiasMc
NAvarMc(a) = NvarMc
NAmseMc(a) = NmseMc

NAmeanMH(a) = NmeanMH
```

Copyright by Mahidol University

```
        NAbiasMH(a) = NbiasMH
        NAvarMH(a) = NvarMH
        NAmseMH(a) = NmseMH

        NAmeanZe(a) = NmeanZe
        NAbiasZe(a) = NbiasZe
        NAvarZe(a) = NvarZe
        NAmseZe(a) = NmseZe

        NAmeanML(a) = NmeanML
        NAbiasML(a) = NbiasML
        NAvarML(a) = NvarML
        NAmseML(a) = NmseML
12      format(I4, 3F8.2, 36F15.8)
        write(2, 12)    n1, P(a), lamda(a), u(a), LAmeanMc(a), LAmeanMH(a),
                        LAmeanZe(a), LAmeanML(a), LAvarMc(a), LAvarMH(a),
                        LAvarZe(a), LAvarML(a), LAbiasMc(a), LAbiasMH(a),
                        LAbiasZe(a), LabiasML(a), LAmseMc(a), LAmseMH(a),
                        LAmseZe(a), LAmseML(a), NAmeanMc(a), NAmeanMH(a),
                        NAmeanZe(a), NAmeanML(a), NAvarMc(a), NAvarMH(a),
                        NAvarZe(a), NAvarML(a), NAbiasMc(a), NAbiasMH(a),
                        NAbiasZe(a), NabiasML(a), NAmseMc(a), NAmseMH(a),
                        NAmseZe(a), NAmseML(a)

        ! rank bias, variance, and MSE of lambda from each method
        call    LvarRank (a, LAvarMc, LAvarMH, LAvarZe, LAvarML, LvarMcPt,
                LvarMHPt, LvarZePt, LvarMLPt)
        call    LbiaRank (a, LAbiasMc, LAbiasMH, LAbiasZe, LAbiasML,
                LbiaMcPt, LbiaMHPt, LbiaZePt, LbiaMLPt)
        call    LmseRank (a, LAmseMc, LAmseMH, LAmseZe, LAmseML,
                LmseMcPt, LmseMHPt, LmseZePt, LmseMLPt)
```

Copyright by Mahidol University

```
        ! rank bias, variance, and MSE of N from each method
        call    NvarRank (a, NAvarMc, NAvarMH, NAvarZe, NAvarML, NvarMcPt,
                NvarMHPt, NvarZePt, NvarMLPt)
        call    NbiaRank (a, NAbiasMc, NAbiasMH, NAbiasZe, NAbiasML,
                NbiaMcPt, NbiaMHPt, NbiaZePt, NbiaMLPt)
        call    NmseRank (a, NAmseMc, NAmseMH, NAmseZe, NAmseML,
                NmseMcPt, NmseMHPt, NmseZePt, NmseMLPt)

13      format(I4, 3F8.2, 12I7)
        write(3, 13)    n1, P(a), lamda(a), u(a), LvarMcPt, LvarMHPt, LvarZePt,
                        LvarMLPt, LbiaMcPt, LbiaMHPt, LbiaZePt, LbiaMLPt,
                        LmseMcPt, LmseMHPt, LmseZePt, LmseMLPt
        write(4, 13)    n1, P(a), lamda(a), u(a), NvarMcPt, NvarMHPt, NvarZePt,
                        NvarMLPt, NbiaMcPt, NbiaMHPt, NbiaZePt, NbiaMLPt,
                        NmseMcPt, NmseMHPt, NmseZePt, NmseMLPt

end do
```

END PROGRAM

2. Subroutine for Mckendrick's method

```
Subroutine McK (j, m, n2, f, lamda_Mc, Nhat_Mc)

!i (input): the ith order of the probability of Z(i)
!m (input): the maximum value of generation
!n2(input): the numbers of observed cases (excluded f(0))
!f(i) (input): vector of frequency of i
!lamda_Mc (output): the lamda from McKendrick's method
!Nhat_Mc (output): the parameter of N from McKendrick's method
```

Copyright by Mahidol University

```
IMPLICIT      none
INTEGER       j, n2, m, f(m)
REAL          sumS, sumSsq
REAL          lamda_Mc, Nhat_Mc

sumS = 0.0    ! S: sum of frequency x count
sumSsq = 0.0  ! Ssq: sum of frequency x count square

! 1.compute sumS and sumSsq
do j = 0, m
        sumS = sumS + (j*f(j))
        sumSsq = sumSsq + (f(j)*(j**2))
end do

! 2.compute lamda_Mc
lamda_Mc = (sumSsq - sumS) / sumS

! 3.compute Nhat_Mc
if ((1 - EXP(-lamda_Mc)).ne.0)        then
        Nhat_Mc = n2 / (1 - EXP(-lamda_Mc))
else
        write(2, *)      '***** Divided by Zero *****(Mc)'
end if

Return
End
```

Copyright by Mahidol University

3. Subroutine for Mantel-Haenszel's method

Subroutine MH (j, m, n2, f, lamda_MH, Nhat_MH)

```
!i (input): the ith order of the probability of Z(i)
!m (input): the maximum value of generation
!n2(input): the numbers of observed cases (excluded f(0))
!f(i) (input): vector of frequency of i
!lamda_MH (output): the lamda from Mantel Hanezel's method
!Nhat_MH (output): the parameter of N from Mantel Hanezel's method

IMPLICIT      none
INTEGER       j, n2, m, m2, f(m)
REAL          sum1, sum2
REAL          lamda_MH, Nhat_MH

! 1.compute sum1 and sum2
sum1 = 0.0
sum2 = 0.0
m2 = m - 1
if (m.gt.1)      then
        do j = 1, m2
                sum1 = sum1 + (j+1)*f(j+1)
                sum2 = sum2 + f(j)
        end do
else    ! for m = 1
        write(2, *)      'm=',m
end if

! 2.compute lambda_MH
if (sum2.ne.0) then
        lamda_MH = (sum1 / sum2)
```

Copyright by Mahidol University

```
else
        write(2, *)     'sum2(MH)=',sum2
end if

! 3.compute Nhat_MH
if ((1-EXP(-lamda_MH)).ne.0)        then
        Nhat_MH = n2/(1-EXP(-lamda_MH))
else
        write(2, *)'***** Divided by Zero *****(MH)'
end if

Return
End
```

4. Subroutine for Zelterman's method

```
Subroutine Zel (n2, m, f, lamda_Ze, Nhat_Ze)

!i (input): the ith order of the probability of Z(i)
!m (input): the maximum value of generation
!n2(input): the numbers of observed cases (excluded f(0))
!f(i) (input): vector of frequency of i
!lamda_Ze (output): the lamda from Zelterman's method
!Nhat_Ze (output): the parameter of N from Zelterman's method

IMPLICIT    none
INTEGER     n2, m, f(m)
REAL        lamda_Ze, Nhat_Ze

! 1.compute lamda_Ze
if (f(1).ne.0 .or. f(2).ne.0) then
```

Copyright by Mahidol University

```
        lamda_Ze = (2)*real(f(2)) / real(f(1))
! 2.compute Nhat_Ze
        if ((1-EXP(-lamda_Ze)).ne.0) then
                Nhat_Ze = n2/(1-EXP(-lamda_Ze))
        end if
end if

Return
End
```

5. Subroutine for Maximum likelihood method

```
Subroutine MLE (j, m, n2, f, lamda_Mc, lamda_ML, Nhat_ML)

!i (input): the ith order of the probability of Z(i)
!m (input): the maximum value of generation
!n2(input): the numbers of observed cases (excluded f(0))
!f(i) (input): vector frequency of i
!lamda_Mc (input): the lamda from McKendrick's method
!lamda_ML (output): the lamda from Maximum Likelihood method
!Nhat_ML (output) : the parameter of N from Maximum Likelihood method

IMPLICIT    none
INTEGER     j, n2, m, f(m)
REAL        sumS, x_bar, oldlamda, newlamda
REAL        lamda_Mc, lamda_ML, Nhat_ML

! 1. compute x_bar
sumS = 0.0      ! S: sum of frequency x count
do j = 0, m
        sumS = sumS + (j*f(j))
```

Copyright by Mahidol University

```
end do

x_bar = sumS / n2

! 2. compute lamda_ML
oldlamda = 0
newlamda = lamda_Mc

do while (abs(oldlamda - newlamda).ge.(0.00001))
        oldlamda = newlamda
        newlamda = x_bar*(1-EXP(-oldlamda))
end do

lamda_ML = newlamda

! 3.compute Nhat_ML
if ((1-EXP(-lamda_ML)).ne.0)        then
        Nhat_ML = n2 / (1 - EXP(-lamda_ML))
else
        write(2, *)'***** Divided by Zero *****(MLE)'
end if

Return
End
```

6. Subroutine for computing bias, variance, and mean square error of lambda

Subroutine LStat (a, k, R, P, lamda, explam, Y, mean, bias, var, mse)

!a (input)
!k (input)

Copyright by Mahidol University

```
!R (input)
!lamda(a) (input)
!explam(a) (input)
!mean (output)
!var (output)
!mse (output)

IMPLICIT    none
INTEGER     a, k, R
REAL        sum1, sum2, sum3, lamda(100), explam(100), P(100)
REAL        Y(1), mean, bias, var, mse
            !Y(1) can be Rlam_Mc(1), Rlam_MH(1), Rlam_Ze(1), Rlam_ML(1)

! 1.compute Mean
sum1 = 0
do k = 1, R
        sum1 = sum1+Y(k)
end do
mean = sum1 / real(R)

! 2.compute Bias
if (P(a).eq.1) then
        bias = mean-lamda(a)
else
        bias = mean-explam(a)
end if

! 3.compute Variance
sum2 = 0
do k = 1, R
        sum2 = sum2 + ((Y(k) - mean)**2)
end do
```

Copyright by Mahidol University

```
var = sum2 / real(R)

! 4.compute Mean Square Error
sum3 = 0
if (P(a).eq.1) then
        do k = 1, R
                sum3 = sum3 + ((Y(k) - lamda(a))**2)
        end do
else
        do k = 1, R
                sum3 = sum3 + ((Y(k) - explam(a))**2)
        end do
end if
mse = sum3 / real(R)

Return
End
```

7. Subroutine for computing bias, variance, and mean square error of N

```
Subroutine NStat (k, R, n1, T, mean, bias, var, mse)

!k (input)
!R (input)
!n1(input)
!mean (output)
!var (output)
!mse (output)

IMPLICIT      none
INTEGER       k, R, n1
```

Copyright by Mahidol University

```
REAL        sum1, sum2, sum3
REAL        T(1), mean, bias, var, mse
            !T(1) can be Nhat_Mc(1), Nhat_MH(1), Nhat_Ze(1), Nhat_ML(1)

! 1.compute Mean
sum1 = 0
do k = 1, R
        sum1 = sum1 + T(k)
end do
mean = sum1 / real(R)

! 2.compute Bias
bias = mean - n1

! 3.compute Variance
sum2 = 0
do k = 1, R
         sum2 = sum2 + ((T(k) - mean)**2)
end do
var = sum2 / real(R)

! 4.compute Mean Square Error
sum3 = 0
do k = 1, R
        sum3 = sum3 + ((T(k) - n1)**2)
end do
mse = sum3 / real(R)

Return
End
```

Copyright by Mahidol University

8. Subroutine for ranking bias of lambda

```
Subroutine    LbiaRank (a, LAbiasMc, LAbiasMH, LAbiasZe, LAbiasML,
LbiaMcPt,
              LbiaMHPt, LbiaZePt, LbiaMLPt)

IMPLICIT      none
INTEGER       a, LbiaMcPt, LbiaMHPt, LbiaZePt, LbiaMLPt
REAL          LAbiasMc(1), LAbiasMH(1), LAbiasZe(1), LAbiasML(1)

LbiaMcPt = 1
LbiaMHPt = 1
LbiaZePt = 1
LbiaMLPt = 1
if (abs(LAbiasMc(a)).gt.abs(LAbiasMH(a))) then
        LbiaMcPt = LbiaMcPt + 1
else
        LbiaMHPt = LbiaMHPt + 1
end if

if (abs(LAbiasMc(a)).gt.abs(LAbiasZe(a)))   then
        LbiaMcPt = LbiaMcPt + 1
else
        LbiaZePt = LbiaZePt+1
end if

if (abs(LAbiasMc(a)).gt.abs(LAbiasML(a))) then
        LbiaMcPt = LbiaMcPt+1
else
        LbiaMLPt = LbiaMLPt+1
end if
```

Copyright by Mahidol University

```
if (abs(LAbiasMH(a)).gt.abs(LAbiasZe(a)))  then
        LbiaMHPt = LbiaMHPt+1
else
        LbiaZePt = LbiaZePt+1
end if

if (abs(LAbiasMH(a)).gt.abs(LAbiasML(a)))        then
        LbiaMHPt = LbiaMHPt+1
else
        LbiaMLPt = LbiaMLPt+1
end if
if (abs(LAbiasZe(a)).gt.abs(LAbiasML(a)))  then
        LbiaZePt = LbiaZePt+1
else
        LbiaMLPt = LbiaMLPt+1
end if

Return
End
```

9. Subroutine for ranking variance of lambda

```
Subroutine    LvarRank (a, LAvarMc, LAvarMH, LAvarZe, LAvarML, LvarMcPt,
              LvarMHPt, LvarZePt, LvarMLPt)

IMPLICIT      none
INTEGER       a, LvarMcPt, LvarMHPt, LvarZePt, LvarMLPt
REAL          LAvarMc(1), LAvarMH(1), LAvarZe(1), LAvarML(1)

LvarMcPt = 1
LvarMHPt = 1
```

Copyright by Mahidol University

```
LvarZePt = 1
LvarMLPt = 1

if (LAvarMc(a).gt.LAvarMH(a))       then
        LvarMcPt = LvarMcPt+1
else
        LvarMHPt = LvarMHPt+1
end if
if (LAvarMc(a).gt.LAvarZe(a))       then
        LvarMcPt = LvarMcPt+1
else
        LvarZePt = LvarZePt+1
end if

if (LAvarMc(a).gt.LAvarML(a))       then
        LvarMcPt = LvarMcPt+1
else
        LvarMLPt = LvarMLPt+1
end if

if (LAvarMH(a).gt.LAvarZe(a))       then
        LvarMHPt = LvarMHPt+1
else
        LvarZePt = LvarZePt+1
end if

if (LAvarMH(a).gt.LAvarML(a))       then
        LvarMHPt = LvarMHPt+1
else
        LvarMLPt = LvarMLPt+1
end if
```

Copyright by Mahidol University

```
if (LAvarZe(a).gt.LAvarML(a))       then
        LvarZePt = LvarZePt+1
else
        LvarMLPt = LvarMLPt+1
end if
Return
End
```

10. Subroutine for ranking mean square error of lambda

```
Subroutine    LmseRank (a, LAmseMc, LAmseMH, LAmseZe, LAmseML,
LmseMcPt,
              LmseMHPt, LmseZePt, LmseMLPt)

IMPLICIT      none
INTEGER       a, LmseMcPt, LmseMHPt, LmseZePt, LmseMLPt
REAL          LAmseMc(1),LAmseMH(1),LAmseZe(1),LAmseML(1)

LmseMcPt = 1
LmseMHPt = 1
LmseZePt = 1
LmseMLPt = 1

if (LAmseMc(a).gt.LAmseMH(a))    then
        LmseMcPt = LmseMcPt+1
else
        LmseMHPt = LmseMHPt+1
end if

if (LAmseMc(a).gt.LAmseZe(a))       then
        LmseMcPt = LmseMcPt+1
```

Copyright by Mahidol University

```
else
        LmseZePt = LmseZePt+1
end if
if (LAmseMc(a).gt.LAmseML(a))    then
        LmseMcPt = LmseMcPt+1
else
        LmseMLPt = LmseMLPt+1
end if

if (LAmseMH(a).gt.LAmseZe(a))    then
        LmseMHPt = LmseMHPt+1
else
        LmseZePt = LmseZePt+1
end if

if (LAmseMH(a).gt.LAmseML(a))    then
        LmseMHPt = LmseMHPt+1
else
        LmseMLPt = LmseMLPt+1
end if

if (LAmseZe(a).gt.LAmseML(a))    then
        LmseZePt = LmseZePt+1
else
        LmseMLPt = LmseMLPt+1
end if

Return
End
```

Copyright by Mahidol University

11. Subroutine for ranking bias of N

```
Subroutine    NbiaRank (a, NAbiasMc, NAbiasMH, NAbiasZe, NAbiasML,
NbiaMcPt,
              NbiaMHPt, NbiaZePt, NbiaMLPt)

IMPLICIT      none
INTEGER       a, NbiaMcPt, NbiaMHPt, NbiaZePt, NbiaMLPt
REAL          NAbiasMc(1), NAbiasMH(1), NAbiasZe(1), NAbiasML(1)

NbiaMcPt = 1
NbiaMHPt = 1
NbiaZePt = 1
NbiaMLPt = 1

if (abs(NAbiasMc(a)).gt.abs(NAbiasMH(a)))        then
        NbiaMcPt = NbiaMcPt+1
else
        NbiaMHPt = NbiaMHPt+1
end if

if (abs(NAbiasMc(a)).gt.abs(NAbiasZe(a)))  then
        NbiaMcPt = NbiaMcPt+1
else
        NbiaZePt = NbiaZePt+1
end if

if (abs(NAbiasMc(a)).gt.abs(NAbiasML(a)))then
        NbiaMcPt = NbiaMcPt+1
else
        NbiaMLPt = NbiaMLPt+1
end if
```

Copyright by Mahidol University

```
if (abs(NAbiasMH(a)).gt.abs(NAbiasZe(a))) then
        NbiaMHPt = NbiaMHPt+1
else
        NbiaZePt = NbiaZePt+1
end if

if (abs(NAbiasMH(a)).gt.abs(NAbiasML(a)))       then
        NbiaMHPt = NbiaMHPt+1
else
        NbiaMLPt = NbiaMLPt+1
end if

if (abs(NAbiasZe(a)).gt.abs(NAbiasML(a))) then
        NbiaZePt = NbiaZePt+1
else
        NbiaMLPt = NbiaMLPt+1
end if

Return
End
```

12. Subroutine for ranking variance of N

```
Subroutine   NvarRank (a, NAvarMc, NAvarMH, NAvarZe, NAvarML, NvarMcPt,
             NvarMHPt, NvarZePt, NvarMLPt)
IMPLICIT     none
INTEGER      a, NvarMcPt, NvarMHPt, NvarZePt, NvarMLPt
REAL         NAvarMc(1), NAvarMH(1), NAvarZe(1), NAvarML(1)

NvarMcPt = 1
NvarMHPt = 1
```

Copyright by Mahidol University

```
NvarZePt = 1
NvarMLPt = 1

if (NAvarMc(a).gt.NAvarMH(a))      then
        NvarMcPt = NvarMcPt+1
else
        NvarMHPt = NvarMHPt+1
end if

if (NAvarMc(a).gt.NAvarZe(a))      then
        NvarMcPt = NvarMcPt+1
else
        NvarZePt = NvarZePt+1
end if

if (NAvarMc(a).gt.NAvarML(a))      then
        NvarMcPt = NvarMcPt+1
else
        NvarMLPt = NvarMLPt+1
end if

if (NAvarMH(a).gt.NAvarZe(a))      then
        NvarMHPt = NvarMHPt+1
else
        NvarZePt = NvarZePt+1
end if

if (NAvarMH(a).gt.NAvarML(a))      then
        NvarMHPt = NvarMHPt+1
else
        NvarMLPt = NvarMLPt+1
end if
```

Copyright by Mahidol University

```
if (NAvarZe(a).gt.NAvarML(a))       then
        NvarZePt = NvarZePt+1
else
        NvarMLPt = NvarMLPt+1
end if

Return
End
```

13. Subroutine for ranking mean square error of N

```
Subroutine    NmseRank (a, NAmseMc, NAmseMH, NAmseZe, NAmseML,
              NmseMcPt, NmseMHPt, NmseZePt, NmseMLPt)

IMPLICIT      none
INTEGER       a,NmseMcPt,NmseMHPt,NmseZePt,NmseMLPt
REAL          NAmseMc(1),NAmseMH(1),NAmseZe(1),NAmseML(1)

NmseMcPt = 1
NmseMHPt = 1
NmseZePt = 1
NmseMLPt = 1

if (NAmseMc(a).gt.NAmseMH(a))    then
        NmseMcPt = NmseMcPt+1
else
        NmseMHPt = NmseMHPt+1
end if

if (NAmseMc(a).gt.NAmseZe(a))       then
        NmseMcPt = NmseMcPt+1
```

Copyright by Mahidol University

```
else
        NmseZePt = NmseZePt+1
end if

if (NAmseMc(a).gt.NAmseML(a))    then
        NmseMcPt = NmseMcPt+1
else
        NmseMLPt = NmseMLPt+1
end if

if (NAmseMH(a).gt.NAmseZe(a))    then
        NmseMHPt = NmseMHPt+1
else
        NmseZePt = NmseZePt+1
end if

if (NAmseMH(a).gt.NAmseML(a))   then
        NmseMHPt = NmseMHPt+1
else
        NmseMLPt = NmseMLPt+1
end if

if (NAmseZe(a).gt.NAmseML(a))    then
        NmseZePt = NmseZePt+1
else
        NmseMLPt = NmseMLPt+1
end if

Return
End
```

Copyright by Mahidol University

BIOGRAPHY

NAME	Miss Niticha Wannasirikul
DATE OF BIRTH	28 April 1978
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Kasetsart University, 1996-2000: Bachelor of Science (Statistics) Mahidol University, 2000-2005: Master of Science (Biostatistics)
POSITION & OFFICE	2004-present, Siam Commercial New York Life Insurance Public Company Limited. Bangkok, Thailand.
HOME ADDRESS	43/52 Ladprao 80, Wangthonglang, Bangkok, Thailand.

Copyright by Mahidol University