Conference Proceedings

ICAS2014

International Conference on Applied Statistics

"Connecting Local to Global with Statistics"

Khon Kaen University
50 Years of Social Devotion

สมาคมสถิติแห่งประเทศไทย
THAI STATISTICAL ASSOCIATION

KMUTT
King Mongkut's University of Technology Thonburi

CRN
เครือข่ายวิจัยสถิติศาสตร์

May 21 – 24, 2014

Pullman Khon Kaen Raja Orchid Hotel
Khon Kaen, Thailand

www.icas2014.stat.in.th

Supported by

Thai Statistical Association (TSA)
Statistics Cooperative Research Network (CRN)
Khon Kaen University
King Mongkut's University of Technology Thonburi
Faculty of Science, Khon Kaen University

Sponsors

Solution Center Co., Ltd
Booknet Co., Ltd

Organized by

Department of Statistics, Faculty of Science
Khon Kaen University

Department of Mathematics, Faculty of Science
King Mongkut's University of Technology Thonburi

Closing Remark

Ladies and gentlemen. This is the final day of the ICAS 2014 conference. On behalf of the organizing committee, I would like to say that it has been our privilege and great honors to have all of you at this conference. The theme for this year conference is “connecting local to global with statistics”. I strongly believe that during the past three days, you got a chance to see the connection between Statistics and other fields through both formal and informal presentations. Statistics has been and always be an important key of research in many fields, economic and financial, management, biostatistics, mathematics, to name a few. We got an opportunity to welcome many international participants from many countries and also have a privilege to welcome many famous statisticians as our keynote and invited speakers, including the CRN session. I believe that this conference has achieved its goal as to connect statisticians in Thailand and abroad. For the conference, ICAS stands for International Conference on Applied Statistics. After today, however, in my point of view, ICAS may also stands for “I Can Analyze more with Statistics” which is also our goal for the conference. I would like to take this opportunity to thank everyone who makes this conference possible. The organizing committee and staff who worked so hard preparing for the conference. The reviewers who put their effort to ensure the quality of every papers presented at this conference. I’d also like to thank King Mongkut's University of Technology, Thonburi, Thai Statistical Association for hosting this conference and our keynote and every invited speakers. And last but not least, every participant of the ICAS 2014. I hope you have enjoyed our hospitality and had a good time in Khon Kaen. If there is any mistake or inconvenience, I deeply apologize and promise that we will take your suggestion and improve it in our next conference. For those who signed up for the post-conference excursion, I wish you have a memorable moment in Lao. And for every one of you, I wish you have a safe trip home. Finally, I wish to see you all again in ICAS 2015 at King Mongkut’s University of Technology Thonburi next year. Thank you and good bye.

Asst. Prof. Pramote Krongyuth
Chair of Organizing Committee

COMMITTEES

Academic Committees

Chair

Assoc. Prof. Adisak Pongpullponsak	King Mongkut's University of Technology Thonburi, Thailand

Assistant to Chair

Asst. Prof. Dr. Sukuman Sarikavanij	King Mongkut's University of Technology Thonburi, Thailand
Assoc. Prof. Dr. Supunnee Ungpansattawong	Khon Kaen University, Thailand
Asst. Prof. Dr. Tidadeaw Mayureesawan	Khon Kaen University, Thailand
Dr. Wuttichai Srisodaphol	Khon Kaen University, Thailand

Organizing Committees

Chair

Asst. Prof. Promote Krongyuth	Khon Kaen University, Thailand

Assistant to Chair for Registration and Financial Affairs

Ms. Yupaporn Tongprasit	Khon Kaen University, Thailand
Asst. Prof. Sukanya Ruangsuwan	Khon Kaen University, Thailand

Assistant to Chair for Public Relations

Asst. Prof. Dr. Kunlaya Pattanagul	Khon Kaen University, Thailand
Asst. Prof. Dr. Wattana Pattanagul	Khon Kaen University, Thailand

Assistant to Chair for Utilities Administration

Mr. Mathee Pongkitiwitoon	Khon Kaen University, Thailand
Mr. Prem Jansawang	Khon Kaen University, Thailand

Assistant to Chair for Conference Evaluation

Assoc. Prof. Wichuda Chaisiwamongkol	Khon Kaen University, Thailand

REVIEWERS

Prof. Dr. Andrei I. Volodin	University of Regina, Canada
Prof. Dr. Bimal K. Sinha	University of Maryland, Baltimore County, USA
Prof. Dr. Dankmar Böhning	University of Southampton, UK
Prof. Dr. Hung T. Nguyen	New Mexico State University, USA
Prof. Dr. Malinee Laopaiboon	Khon Kaen University, Thailand
Prof. Dr. Thomas Mathew	University of Maryland, Baltimore County, USA
Prof. Dr. Timothy E. O' Brien	Loyola University, USA
Prof. Dr. Vladik Kreinovich	University of Texas at El Paso, USA
Assoc. Prof. Adisak Pongpullponsak	King Mongkut's University of Technology Thonburi, Thailand
Assoc. Prof. Dr. Guido Knapp	Clausthal University of Technology, Germany
Assoc. Prof. Dr. Kamon Budsaba	Thammasat University, Thailand
Assoc. Prof. Dr. Montip Tiensuwan	Mahidol University, Thailand
Assoc Prof. Dr. Pongsa Pornchaiwiseskul	Chulalongkorn University, Thailand
Assoc Prof. Prasit Payakkapong	Kasetsart University, Thailand
Assoc Prof. Puchong Praekhaow	King Mongkut's University of Technology Thonburi, Thailand
Assoc Prof. Dr. Sa-aat Niwitpong	King Mongkut's University of Technology North Bangkok, Thailand
Assoc Prof. Dr. Settapat Chinviriyasit	King Mongkut's University of Technology Thonburi, Thailand
Assoc. Prof. Dr. Supunnee Ungpansattawong	Khon Kaen University, Thailand
Assoc. Prof. Dr. Veeranan Pongsapakdee	Silpakorn University, Thailand
Assoc. Prof. Dr. Wirawan Chinviriyasit	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Anamai Na-udom	Naresuan University, Thailand
Asst. Prof. Dr. Autcha Araveeporn	King Mongkut's Institute of Technology Ladkrabang, Thailand
Asst. Prof. Dr. Bungon Kumphon	Mahasarakham University, Thailand
Asst. Prof. Dr. Chaiya Dumkum	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Chukiat Worasucheep	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Chunchom Pongchavalit	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Katechan Jampachaisri	Naresuan University, Thailand
Asst. Prof. Dr. Kunlaya Pattanagul	Khon Kaen University, Thailand
Asst. Prof. Dr. Kusaya Plungpongpun	Silpakorn University, Thailand
Asst. Prof. Dr. Naratip Jansakul	Prince of Songkla University, Thailand
Asst. Prof. Dr. NipapornChutiman	Mahasarakham University, Thailand
Asst. Prof. Dr. Saowanit Sukparungsee	King Mongkut's University of Technology North Bangkok, Thailand
Asst. Prof. Siriluck Jermjitpornchai	Mahasarakham University, Thailand
Asst. Prof. Dr. Sujitta Suraphee	Mahasarakham University, Thailand
Asst. Prof. Dr. Suksan Prombanpong	King Mongkut's University of Technology Thonburi, Thailand

Asst. Prof. Dr. Sukuman Sarikavanij	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Sumlearng Chunrungsikul	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Taweesak Siripornpibul	Naresuan University, Thailand
Asst. Prof. Dr. Wararit Panichkitkosolkul	Thammasat University, Thailand
Asst. Prof. Dr. Winai Bodhisuwan	Kasetsart University, Thailand
Asst. Prof. Dr. Woranut Koetsinchai	King Mongkut's University of Technology Thonburi, Thailand
Asst. Prof. Dr. Yupaporn Areepong	King Mongkut's University of Technology North Bangkok, Thailand
Dr. Ampai Thongteeraparp	Kasetsart University, Thailand
Dr. Anuwat Sae-tang	King Mongkut's University of Technology Thonburi, Thailand
Dr. Dawud Thongtha	King Mongkut's University of Technology Thonburi, Thailand
Dr. Dusadee Sukawat	King Mongkut's University of Technology Thonburi, Thailand
Dr. Krisana Lanumteang	Maejo University, Thailand
Dr. Monchaya Chiangpradit	Mahasarakham University, Thailand
Dr. Naowarut Meejun	Silpakorn University, Thailand
Dr. Ngamphol Soonthronworasiri	Mahidol University, Thailand
Dr. NuengruithaiTharawatcharasart	Mahidol University Kanchanaburi Campus, Thailand
Dr. Pariwate Varnakovida	King Mongkut's University of Technology Thonburi, Thailand
Dr. Teerapol Saleewong	King Mongkut's University of Technology Thonburi, Thailand
Dr. Vanida Pongsakchat	Burapha University, Thailand
Dr. Wibulsak Wattayu	King Mongkut's University of Technology Thonburi, Thailand
Dr. Wuttichai Srisodaphol	Khon Kaen University, Thailand

CONTENTS

Page

Page

Poster Presentation

Some aspects of data analysis under confidentiality protection

Bimal Sinha
Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA, sinha@umbc.edu

Abstract

Statisticians working in most federal agencies are often faced with two conflicting objectives: (1) collect and publish useful datasets for designing public policies and building scientific theories, and (2) protect confidentiality of data respondents which is essential to uphold public trust, leading to better response rates and data accuracy. In this talk I will provide a survey of two statistical methods currently used at the U.S. Census Bureau: synthetic data and noise perturbed data.

*Corresponding Author
E-mail Address: sinha@umbc.edu

Goodness of fit test for time series models

Sangyeol Lee

Department of Statistics, Seoul National University, Seoul, Korea, jpgrslee@yahoo.com

Abstract

In this talk, the empirical process and entropy based goodness of fit tests for time series models are discussed. The time series models under consideration are unstable autoregressive models with unit roots, GARCH models, and ACD models. It is well known that the parameter estimation affects the asymptotic behavior of the goodness of fit tests. Special attention is paid to how to overcome this problem.

*Corresponding Author
E-mail Address: jpgrslee@yahoo.com

Roles of statisticians in research synthesis: experience in Cochrane systematic reviews

Malinee Laopaiboon
Department of Biostatistics and Demography, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand, malinee@kku.ac.th

Abstract

Research synthesis is the application, in practice, of the principle that science is cumulative. Methods for research synthesis, including systematic review and meta-analysis, are systematically extracting and integrating data from a variety of sources in order to draw more reliable conclusions about a given question or topic. Statistician is one of the research team who has an important role in producing a high quality systematic review. In my talk, I will present the experience as a statistician in conducting Cochrane systematic reviews. I will also introduce the Cochrane Collaboration, a global independent network of various professionals including statisticians, responding to the challenge of making the vast amounts of evidence generated through research useful for informing decisions about health.

*Corresponding Author
E-mail Address: malinee@kku.ac.th

Ratio plot for the power series distribution, generalized Turing estimation and estimating the size of elusive populations

Dankmar Böhning[1*], Peter van der Heijden[2] and Heinz Holling[3]
[1]*Southampton Statistical Sciences Research Institute & Mathematical Sciences, University of Southampton, Southampton, UK, D.A.Bohning@soton.ac.uk*
[2]*Universities of Utrecht, NL and Southampton, UK*
[3]*Department of Psychology and Sport Science, University of Muenster, Germany, holling@uni-muenster.de*

Abstract

The Turing estimator for population size estimation based upon zero-truncated capture-recapture counts has been developed in the Poisson distributional context. However, capture-recapture count data do rarely follow a Poisson distribution. The talk presents to generalizations. For one, it generalizes the Turing estimator to the Power series distribution. For two, it presents a version of the Turing estimator which is more robust to contaminating observations. By means of the ratio plot, it is decided which member of the power series works best and which proportion of uncontaminated data are used in the Turing estimate. The concept is applied to determine the size of the homeless population in Utrecht (The Netherlands).

*Corresponding Author
E-mail Address: D.A.Bohning@soton.ac.uk

Small sample asymptotics: Methodology and two applications

Thomas Mathew

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA, mathew@umbc.edu

Abstract

Standard likelihood based methods usually used to analyze data arising from a parametric model are typically accurate to the first order. Small sample asymptotic procedures provide major improvements in accuracy, and are available for discrete as well as for continuous data. In the talk, small sample asymptotics will be introduced and explained using two applications. The first application is on the computation of tolerance limits under the logistic regression model for binary data. The data consist of binary responses, and upper tolerance limits are to be constructed for the number of positive responses in future trials corresponding to a fixed level of the covariates. The problem has been motivated by an application of interest to the U.S. Army, dealing with the testing of ballistic armor plates for protecting soldiers from projectiles and shrapnel, where the probability of penetration of the armor plate depends on covariates such as the projectile velocity, size of the armor plate, etc. The second application is on a multivariate bioassay problem: several independent multivariate bioassays are performed at different laboratories or locations, and the problem of interest is to test the homogeneity of the relative potencies, assuming the usual slope-ratio or parallel line assay model. The problem has been investigated in the literature using likelihood based methods, under the assumption of a common covariance matrix across the different studies. This assumption is relaxed in this investigation. Numerical results show that for both of the above applications, usual likelihood based procedures can be inaccurate in terms of providing satisfactory coverage probabilities or type I error probabilities. Furthermore, methodology based on small sample asymptotics results in significantly more accurate results in the small sample scenario. The first application will be illustrated using data from the U.S. Army dealing with the testing of ballistic armor plates. The bioassay application will be illustrated using data from a dental study, where pain intensity scores based on a standard treatment and a test treatment are analyzed.

*Corresponding Author
E-mail Address: mathew@umbc.edu

Statistical process control: Statistics for uniformity

Adisak Pongpullponsak
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand, iadinsak@kmutt.ac.th

Abstract

Consecutive statistical hypothesis testing is the main activity in statistical process control (SPC). Starting from parametric statistics with normality assumption followed by cumulative sum (CUSUM) and exponentially weighted moving average (EWMA), and many techniques until economic design. The development also starting from univariate to multivariate to meet the demand in monitoring several underlying process variables simultaneously. The nonparametric statistics come to SPC arena when the normality assumption cannot be met and the data-free development cycle followed the parametric history. Fuzzy logic has been used with quality variable in linguistic terms. The data reductions like Principal Components Analysis (PCA) and the data depths are the techniques to handle the multivariate scheme.

*Corresponding Author
E-mail Address: iadinsak@kmutt.ac.th

Optimal design for Poisson regression with fixed and random effects

Heinz Holling
Department of Psychology and Sport Science, University of Muenster, Germany, holling@uni-muenster.de

Abstract

Consecutive statistical hypothesis testing is the main activity in statistical process control (SPC). Starting from parametric statistics with normality assumption followed by cumulative sum (CUSUM) and exponentially weighted moving average (EWMA), and many techniques until economic design. The development also starting from univariate to multivariate to meet the demand in monitoring several underlying process variables simultaneously. The nonparametric statistics come to SPC arena when the normality assumption cannot be met and the data-free development cycle followed the parametric history. Fuzzy logic has been used with quality variable in linguistic terms. The data reductions like Principal Components Analysis (PCA) and the data depths are the techniques to handle the multivariate scheme.

*Corresponding Author
E-mail Address: holling@uni-muenster.de

Impact of missing values data on the prediction of logistic regression

Puchong Praekhaow
Department of mathematics, King Mongkut's University of the Technology Thonburi, Bangkok, Thailand, puchong.pra@kmutt.ac.th.

Abstract

The purpose of this research is to study the impact of using Binary Logistic Regression to analyze the datasets with missing values of 5%, 10%, 15% and 25%. The samples are three sizes of the missing values datasets which are 60, 90 and 120, respectively. Five hundred using R statistical software with Monte Carlo simulation method. Each dataset is analyzed by using the Binary Logistic Regression technique to find an appropriate model of Logistic Regression. The experiment has evaluated the error of dependent variables prediction against the model. The results show that lost datasets have a significant effect on error predictions. However the sample sizes haves no effect at p > 0.05.

Keywords: Logistic regression, missing values data

Corresponding Author
E-mail Address: puchong.pra@kmutt.ac.th

1. Introduction

Data missing is a common problem in statistical analysis. For example, a sample unit lose, a data that is an outlier, and some incomplete information. In many situations, a statistician has to deal with missing data. Many of these methods were developed in order to deal with missing data in sample surveys. [1]: In general, the methods for missing data can be divided into three categories such as a) Case/Pairwise Deletion, which are the easiest and more commonly applied, b) Parameter estimation, the Maximum likelihood procedures used to estimate the missing data by distribution model of a variable, and c) Imputation techniques, where missing values are replaced with estimated ones based on information available in the data set. The Replacement of the lost data mostly used imputation method by average values. [6]: The main purpose of this paper is to study the impact of missing values data on the dependent variables prediction of Logistic Regression analysis when it has a binary dependent variable.

1.1 Case Deletion

The complete data is a common method in many statistical programs. The elimination of all cases with missing values is one way to complete a data for analysis. A variation of this method consists of determining the extent of missing data on each instance, and deletes the instances with high levels of missing data. Before deleting any attributes, it is necessary to evaluate its relevance to the analysis. The relevant attributes should be kept even with a high degree of missing values. The case deletion is less risky if it involves minimal loss of sample size and if there is no structure or pattern to the missing data. However, the case deletion has been shown to produce more biased estimates than alternative methods. Thus, the case deletion should be applied only in cases in which data are missing completely at random [6]

1.2 Mean Imputation

Mean Imputation is one of the most frequently used methods. It replaces the missing data for a characteristic by the mean of all values in the class. If the value x_i is missing then it will be replaced by

$$\overset{\frown}{x_i} = \frac{\sum_{k=1}^{n_k} x_k}{n_k} \qquad (1)$$

where n_k represents the number of non-missing values in the sample. Although this method is used extensively, it still has some defects, for example, overestimated of sample size, underestimated of variance, correlation is negatively biased, and the distribution of new values is an incorrect representation of the population values due to distortion of the shape of the distribution by adding values equal to the mean. Replacing all missing records with a single value will deflate the variance and increase the significance of any statistical tests based on it. However, mean imputation has given good experimental results in data sets used for classification purposes [4].

1.3 Logistic Regression

Logistic regression is an important statistical tool in applied statistics. The binary logistic regression analysis was referred to as a discriminant analysis used to model the relationship of a binary dependent variable (yes or no) to metric or nonmetric independent variables [8]. However, logistic regression analysis is not similar with most discriminant function analysis where the predictors variables do not have to be normally distributed, linearly related, or of equal variance within each group [7]. Multiple logistic models are based on a linear relationship between the natural logarithm ($\ln$) of odds of an event and multiple independent variables. The form of this relationship is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{2}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{3}$$

$$\ln(o) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{4}$$

$$L = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{5}$$

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}} \tag{6}$$

where Y is binary and represent the event of interest (success/failure), p is the proportion of successes, o is the odds of event, L is the $\ln$ (odds of event), $X_1,\ldots,X_k$ are the independent variables, $\beta_1,\ldots,\beta_k$ are the coefficient of variables, β_0 is the intercept, and ε is the random error. The model coefficients are estimated from data sets (sample size = n) by maximum likelihood estimation (MLE). In that form, the log-likelihood function for binary-logistic model is given as follows:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\mu_i \middle| (1-\mu_i)\right) + \ln(1-\mu_i) \right\} \tag{7}$$

When logistic regression is estimated using a Newton - Raphson type of MLE algorithm, the estimated fit is then determined by taking the first derivative of the log-likelihood function with respect to β, setting it to zero, and solving, The first derivative of the log-likelihood function is commonly referred to as the gradient [3], or score function. The second derivative of log-likelihood with respect to β produces the hessian matrix [3], from which the standard errors of the predictor parameter estimates are derived. The logistic gradient and hessian functions are given as

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) x_i \tag{8}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \left\{ x_i x_i' \mu_i (1-\mu_i) \right\} \tag{9}$$

where $y_i = 0,1,2,\ldots,n$ is the success number, μ_i is the mean values, and x_i is the independent values.

2. Methodology

Supposed that X_1 and X_2 are independent variables which are normally distributed and are simulated with Monte Carlo method, and where Y is the dependent variable calculated by equation (10). If $p \geq 0.5$ then Y = 1, otherwise Y = 0, and where logistic regression is given by the equation (11).

$$p = \frac{e^{0.2+0.3X_1+0.4X_2}}{1+e^{0.2+0.3X_1+0.4X_2}} \tag{10}$$

$$\ln(o) = 0.2 + 0.3X_1 + 0.4X_2 \tag{11}$$

The R programming version R.2.11.1 is used to generate each dataset. The sample sizes of datasets are 60, 90 and 120 respectively. The missing values are applied to these datasets about of 5%, 10%, 15% and 25% with random method. The missing values are replaced by the mean of all values in the residue of data. Each dataset is used in finding logistic regression model. These models are used to predict the dependent variable. After that, the errors of predictions in the each case are calculated in percentage. The experiments are classified into 12 cases in this research. All cases are generated by about 500 rounds of simulations. In addition, there is a hypothesis testing that the prediction of Logistic Regression is impacted by sample size and percentage of missing values. These variables are tested by the two-way analysis of variance (ANOVA) at significant level of 0.05. ANOVA is a statistical models used to analyze the differences between group means and their associated procedures, [5], and the two-way ANOVA measures the significant effect of two independent variables. The model of test is given as:

$$\mathrm{y} = \mu + \tau + \gamma + \varepsilon \tag{12}$$

where y is the corrected percentage of prediction by logistic regression model, μ is the overall mean, τ is the missing values, γ is the sample size, and ε is the random error.

3. Results and Discussion

The results of experiment are repeated 500 times, as shown in Table 1. The forecast value is compared by the actual value of the data. Then, the correct number is calculated in percentage.

Table 1: Percentage of the correct number

Accurate Percentage		Sample sizes		
		60	90	120
Missing values	5%	99.9%	99.8%	99.8%
	10%	86.4%	91.0%	86.8%
	15%	70.0%	68.0%	55.4%
	25%	32.2%	23.2%	13.2%

In Table 1, we found that the current prediction is decreased as the missing values increased from 5% to 10%, 20%, and 25%, respectively. The accuracy is changed with no direction when the sample size was increased from 60 to 90, and 120, respectively.

Therefore, the missing percent of data may have impact to the accurate prediction by logistic regression model, as shown in the Figure 1.

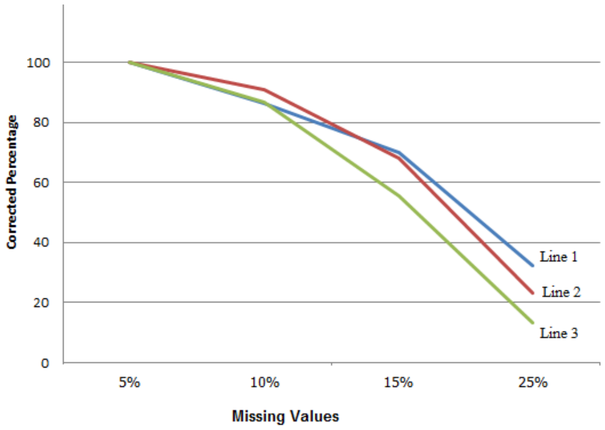


Figure 1: The correlation between missing values versus corrected percentage

In Figure 1, the line 1, line 2, and line 3 curves represent the sample sizes of 60, 90, and 120, respectively. These curves are changed on the same direction, for example, missing values decreases the corrected percentage. ANOVA is used to calculate the effect of the lost data.

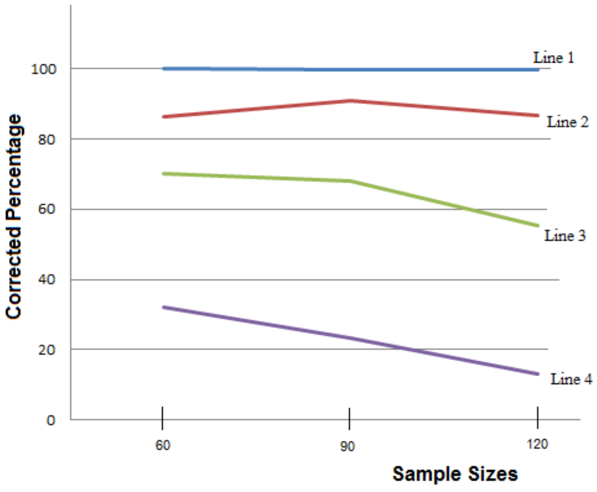


Figure 2: The correlation between sample sizes versus corrected percentage

In Figure 2, the line 1, line 2, line 3, and line 4 curves represent the missing values of 5%, 10%, 15%, and 20%, respectively. These curves show that the changes have no direction, i.e., the corrected percentage is not depended on the sample sizes. Interestingly, a few loss of data (5%) is not affected in the analysis. Before, the hypothesis testing is evaluated by ANOVA method. The data of each group is tested by Shapiro Wilk method [9] showing an approximately normal distribution, (refer to Table 2 for sample size group and Table 3 for missing values group).

Table 2: Tests of normality for sample size group

	Shapiro Wilk		
Sample Sizes	Statistic	df	Sig
60	0.943	4	0.671
90	0.905	4	0.458
120	0.941	4	0.658

Table 2 shows that every sample size accepts the hypothesis of normal distribution at 0.05 significance levels, (p-values > 0.05). It means that each group of the sample size has an approximately normal. Thus, the use of ANOVA is sufficient in the analysis.

Table 3: Tests of normality for missing value group

	Shapiro Wilk		
Missing Values	Statistic	df	Sig
5%	0.750	3	<0.01
10%	0.815	3	0.150
15%	0.851	3	0.242
20%	0.999	3	0.942

Table 3 shows the missing values of 10%, 15%, and 20% accept the hypothesis of normal distribution at 0.05 significance levels, (p-values > 0.05). It means that the missing values group of 10%, 15%, and 20% have distributed the approximately normal distribution. That is, the Missing Values Data has impacted on the Prediction of Logistic Regression. The missing values group of 5% reject the hypothesis of normal distribution at 0.05 significance levels, (p-value < 0.05). The percentage of 5% missing values in every group is a constant values (refer to Table 1). It means that 5% missing value does not have an impact on the Prediction of Logistic Regression. The model can now be predicted with a high corrected percentage.

Table 4: Levene’s Test of Equality of Error Variances

F	df1	df2	Sig
3.210	3	8	0.083

Table 4 shows the error variances have been the equality by Levene’s Test. (p-value > 0.05). Levene’s test [2] is used to test if k samples have equal variances. Equal variances across samples are called homogeneity of variance. Thus, the analysis of variance is followed on the assumption, the used of ANOVA sufficient in the analysis.

Table 5: Test of two-way ANOVA

Parameters	Type III Sum of Squares	df	Mean Square	F	Sig
Model	10544.5	5	2108.9	77.5	<0.01
Intercept	56815	1	56815	2088.9	<0.01
Missing values	10388.7	3	3462.9	127.3	<0.01
Sample size	155.7	2	77.8	2.8	0.134
Error	163.1	6	27.1		
Total	67522.7	12			

R Squared = 0.985(Adjusted Squared = 0.972)

Table 5 shows the model of testing has been the significant (p-values < 0.05). The missing values have been the significant (Sig values < 0.05). But the sample size has not been a significant (Sig values > 0.05). The R^2 (Coefficient of Determination) equals 0.972.

Table 6: Test of one-way ANOVA

Parameters	Type III Sum of Squares	df	Mean Square	F	Sig
Model	10388.7	3	3462.9	86.8	<0.01
Intercept	56815.0	1	56815	1424.9	<0.01
Missing values	10388.7	3	3462.9	86.8	<0.01
Error	318.9	8	39.8		
Total	67522.7	12			

R Squared = 0.97(Adjusted Squared = 0.955)

Table 6 shows the model of testing have been the significant (p-values < 0.05). The missing values have been a significant (p-values < 0.05). It means the Binary Logistic Regression is created by the missing values data. There is the effect when the data was the loss. The Coefficient of Determination equals 0.955 is told to know that the model has the confident of 95.5%.

4. Conclusion

This study shows the experiment is analyzed by Binary Logistic Regression, when the data loss is assumed to be 5%, 10%, 15% and 25%. The samples are the three sizes of missing values datasets with sizes of 60, 90 and 120 respectively. The datasets have been generated by the R statistical software. The Monte Carlo simulation method is primarily used in the experiment. Each dataset is analyzed to predict the binary dependent value by using the Logistic Regression technique. The results can be concluded that the lost datasets can be replaced using the mean of the remaining datasets. The missing values above 5% an effect on the predictions. But the forecast values of dependent variables do not depended on the sample size at significance level of 0.05.

Acknowledgements

The author would like to thank the two anonymous readers and Jeffrey for their insightful comments which improved this paper. Also, I would like to acknowledge the Faculty of Science, King Mongkut's University of Technology of Thonburi for the financial support.

References

[1] Kalton, G, Kasprzyk, D. The treatment of missing survey data. Survey methodology Journal. 1986; 12: 1-16.

[2] Levene, H, Contributions to probability and statistics. California: Stanford University Press; 1960.

[3] Hilbe, J.M., Logistic regression models. Chapman & Hall/CRC Press, Boca Raton; 2009.

[4] Chan, P. and Dunn, O.J. The treatment of missing values in discriminant analysis. Journal of the American Statistical Association, 1972; 6: 473-477.

[5] Fisher, R.A., The design of Experiments. Eight Edition. New York; 1969.

[6] Little, R.J., Rubin, D.B. Statistical analysis with missing data. 2*nd* ed. John Wiley and Sons, New York; 2002.

[7] Senol, S, Ulutagay, G. Logistic regression analysis to determine the factors that affect “Green Card” usage for health services. Journal of the Faculty of Science 2006; 29: 18-26..

[8] Sanogo, S., Yang, XB. Overview of selected multivariate statistical methods and their use in phytopathological research. Phytopathology 2004; 94; 1004-1006.

[9] Shapiro, S. S. and Wilk, M. B. An analysis of variance test for normality. Biometrika. 1965: 52 (3–4): 591–611.

Appendix A. R Code for Simulated Data

```
function (ROUND,N,MISS)
{
SHOWMISSING=rep(0,ROUND)
ytestold=rep(0,N)
ytestnew=rep(0,N)
for (I in1:ROUND)
{x1=rnorm(N)
x2=rnorm(N)
yHAT1=rep(0,N)
yHAT2=rep(0,N)
yMISSNG=rep(,N)
for (u in 1:N)
{yHAT1[u]=0.2+0.3*x1[u]+0.4*x2[u]}
P=1/(1+exp(-yHAT1))
for(k in 1:N)
{
if(P[k]>=0.5)
}ytestold[k]=1
}
Else
{ytestold[k]=0
}
```

On the exact and the approximate mean integrated square error for the kernel distribution function estimator

Abdel-Razzaq Mugdadi*[1] and Rawan Bani-Melhem[2]
[1]Department of Mathematics and Statistics, Jordan University of Science and Technology, Irbid, Jordan, aamugdadi@just.edu.jo
[2]Department of Mathematics and Statistics, Jordan University of Science and Technology, Irbid, Jordan, rawanmelhem552@yahoo.com

Abstract

The asymptotic mean integrated square error (AMISE) is used as an approximate measure of error for the mean integrated square error (MISE). The exact MISE for kernel density estimator is discussed by Marron and Wand (1992). In this investigator we will discuss the exact MISE for the cumulative distribution estimate and compare it with the AMISE. Also, we compare between the optimal bandwidth that minimize the AMISE and that minimize MISE. In addition, through simulation these optimal bandwidths are compared with the bandwidth selectors using the least square cross -validation (LSCV), biased cross - validation (BCV), and direct plug -in (DPI) techniques.

Keywords: Density estimation, cumulative distribution function estimation, bandwidth, mean square error

*Corresponding Author
E-mail Address: aamugdadi@just.edu.jo

Stochastic orders of negative binomial-generalized exponential distribution

P. Thongchan[1], W. Wongrin[2] and W. Bodhisuwan[3*]

[1]*Department of Statistics, Kasetsart University, P.O.Box 1086, Chatuchak, Bangkok, 10903, Thailand, e-mail :panu.tho@hotmail.com*

[2]*Department of Statistics, Kasetsart University, P.O.Box 1086, Chatuchak, Bangkok, 10903, Thailand, e-mail : weerinradaj@gmail.com*

[3]*Department of Statistics, Kasetsart University, P.O.Box 1086, Chatuchak, Bangkok, 10903, Thailand, e-mail : fsciwnb@ku.ac.th*

Abstract

The objective of this work was to compare two random variables; the negative binomial distribution and negative binomial - generalized exponential distribution. The study was done by stochastic orders. We characterized the comparison such as usual stochastic order, likelihood ratio order, convex order and expectation order based on theorem .We also provided some numerical examples.

Keywords: Negative Binomial Distribution, negative binomial - generalized exponential distribution, stochastic orders

*Corresponding Author
E-mail Address: fsciwnb@ku.ac.th

1. Introduction

Poisson distribution is a discrete distribution and widely use to analyze count phenomena, but sometime the count data occurred over-dispersion problem (the mean is less than variance). Since Poisson distribution is not appropriate to analyze, the Negative-Binomial (NB) distribution is an alternative distribution to solve the over-dispersion problem. The NB distribution is a mixed poisson distribution with gamma distribution that is a flexible functional form to describe the count data. However, the recent distribution was proposed by Aruyeuan and Bodhisuwan,. [1] called Negative Binomial – Generalized Exponential (NB-GE) distribution which is a mixture distribution based on NB distribution when the prior distribution of parameter (p) is Generalized Exponential distribution. NB-GE is better than fitted to the count data of insurance claims. If Y is a NB-GE random variable with parameters r,α and β then its probability mass function is

$$g(y)=\binom{r+y-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)},y=0,1,2,....$$

where r,α and $\beta>0$ with

$$E(Y)=r(\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}-1)$$

and

$$Var(Y)=(r^2+r)\frac{\Gamma(\alpha+1)\Gamma(1-\frac{2}{\beta})}{\Gamma(1-\frac{2}{\beta}+\alpha)}-(2r^2+r)\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}+r^2-\left[r(\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}-1)\right]^2$$

Some probability mass functions are plotted as shown in figure (Fig. 1).

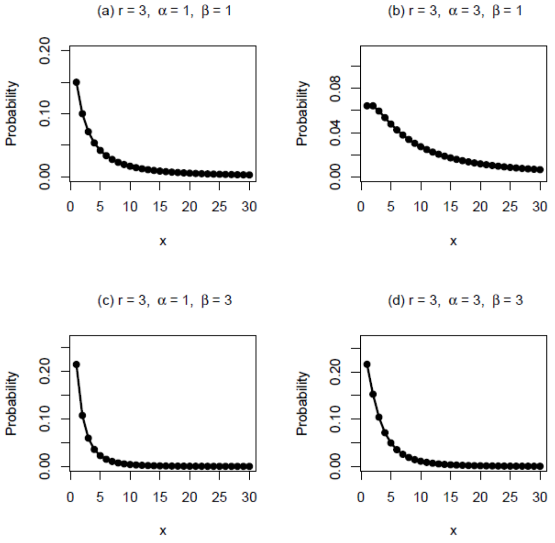


Figure 1. The Examples of Probability Mass Function for difference values of r , α and β

The aim of this work was to compare the NB distribution with the NB-GE distribution based on stochastic orders such as usual stochastic order, likelihood ratio order, convex order and expectation order.

2. Stochastic Orders

Stochastic orders is a method that have been used to compare two probability mass functions, in the areas of probability and statistics.

Definition 1

Let X and Y be random variables with densities function f and g, respectively, if $g(y)/f(x)$ is an increasing function in k over the union of the support of X and Y $(k\in x,y\mid f(\mathrm{k})\neq 0, g(\mathrm{k})\neq 0)$, or equivalently. Then X is said to be smaller than Y in sense of the likelihood ratio order which is denoted by $X\leq_{lr}Y$. The likelihood ratio order is a stronger ordering of random variables.

Definition 2

Let X and Y be two random variables such that $P(X\leq k)\geq P(Y\leq k)$ for all $k\in\mathbb{R}$. Then X is said to be smaller than Y in sense of the usual stochastic order which is denoted by $X\leq_{st}Y$.

Definition 3

Let X and Y be two random variables such that $E(\varphi(X))\leq E(\varphi(Y))$, for every real valued convex function φ where expectations are assumed to be existed. Then X is said to be smaller than Y in sense of the convex order which is denoted by $X\leq_{cv}Y$.

Definition 4

Let X and Y be two random variables such that $E(X)\leq E(Y)$, where expectations are assumed to be existed. Then X is said to be smaller than Y in sense of the expectation order which is denoted by $X\leq_{E}Y$.

3. Random Variable Comparisons

We make comparisons between the negative binomial random variable and negative binomial – generalized exponential with respect to the likelihood ratio order, stochastic order, convex order, expectation order and uniform more variable order. The following lemma will be useful in proving the main result.

Lemma 1

Defined

$$a(k)=1-\frac{\displaystyle\sum_{j=0}^{k+1}\binom{k+1}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}}{\displaystyle\sum_{j=0}^{k}\binom{k}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}}$$

and

$$\varphi_k(m)=\sum_{j=0}^{k}\binom{j+r-1}{j}m(1-m^{\frac{1}{r}})^{j},\quad \forall m, 0<m<1, k=0,1,2,\ldots.$$

Then

I. $a(k)$ is a decreasing function of k $k\in\{0,1,2,\ldots\}$

II. for each fixed $k\in\{0,1,2,\ldots\},\varphi_k(m)$ is concave function of $m\in(0,1).$

Proof

(i) We may write $a(k)$ in form

$$a(k)=1-\frac{\int_0^{\infty}e^{-\lambda r}(1-e^{-\lambda})^{k+1}g(\lambda\mid\alpha,\beta)d\lambda}{\int_0^{\infty}e^{-\lambda r}(1-e^{-\lambda})^{k}g(\lambda\mid\alpha,\beta)d\lambda}=E(W_k),\qquad k=0,1,2,\ldots$$

where W_k is a random variable with the pdf

$$\psi_k(\mathrm{x})=\frac{1}{\displaystyle\sum_{j=0}^{k}\binom{k}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}}e^{-rx}(1-\mathrm{e}^{-x})^{k}g(\mathrm{x}\mid\alpha,\beta)$$

For fixed $k\in\{0,1,2,\ldots\}$, the ratio $\psi_{k+1}(x)/\psi_k(x)$ is obviously a decreasing function of $x>0$. Then, by Definition 1 and 2 , we have $W_k\geq_{lr}W_{k+1}\Rightarrow W_k\geq_{st}W_{k+1}\Rightarrow E(W_k)\geq E(W_{k+1})$ or, equivalently, $a(k)\geq a(k+1)$ [2-3].

This proves $a(k)$ is a decreasing function of $k\in\{0,1,2,\ldots\}$.

(ii) For $k=0,$ note that $\varphi_0(m)=m$ is both convex and concave function. For $k=1,2,\ldots$ we can write

$$\varphi_k(m)=1-\sum_{j=k+1}^{\infty}\binom{j+r-1}{j}m(1-m^{\frac{1}{r}})^{j},\forall m,0<m<1.$$

The relationship between negative binomial and beta probabilities is of the form

$$\sum_{j=k}^{\infty}\binom{j+r-1}{j}p^{r}(1-p)^{j}=\frac{(k+r-1)!}{(k-1)!(r-1)!}\int_{0}^{1-p}t^{k-1}(1-t)^{r-1}dt,\quad k=0,1,2,\ldots$$

Therefore, $\varphi_k(m)$ in Equation 2 can be written as

$$\varphi_k(m)=1-\frac{(k+r)!}{k!(r-1)!}\int_{0}^{1-m^{\frac{1}{r}}}t^{k}(1-t)^{r-1}dt,\quad \forall m, 0<m<1.$$

Thus,

$$\frac{\partial^2\varphi_k(m)}{\partial m^2}=-\frac{1}{r}\frac{(k+r)!}{(k-1)!r!}m^{\frac{1}{r}-1}(1-m^{\frac{1}{r}})^{k-1}<0,\quad \forall m, 0<m<1$$

which proves concavity

Theorem 1

Let $X\sim NB(r,p), Y\sim\ NB-GE(r,\alpha,\beta)$ and

$$p_0=\frac{\Gamma(1+\frac{r+1}{\beta})\Gamma(1+\frac{r}{\beta}+\alpha)}{\Gamma(1+\frac{r}{\beta})\Gamma(1+\frac{r+1}{\beta}+\alpha)},$$

$$p_1=\sqrt[r]{\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r}{\beta})}{\Gamma(1+\frac{r}{\beta}+\alpha)}},$$

$$p_2=\frac{\Gamma(1-\frac{1}{\beta}+\alpha)}{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}$$

and

Then $0<p_2<p_1<p_0<1.$ Furthermore,

(I). $X\le_{lr}Y$ if and only if $p\ge p_0$,

(II). $X\le_{st}Y$ if and only if $p\ge p_1$,

(III). $E(X)\le E(Y)$ if and only if $p\ge p_2$,

Proof

(i) The likelihood ratio order between X and Y is given by

$$l(k)=\frac{\mathrm{P}(Y=k)}{\mathrm{P}(X=k)}=\frac{\sum_{j=0}^{k}\binom{k}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}}{p^{r}(1-p)^{k}},\quad k=0,1,2,\ldots.$$

By Definition 1 , we have

$$X\le_{lr}Y\Leftrightarrow l(k)\le l(k+1),\quad \forall k=0,1,2,\ldots$$

$$\Leftrightarrow p\ge 1-\frac{\sum_{j=0}^{k+1}\binom{k+1}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}}{\sum_{j=0}^{k}\binom{k}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}},\forall k=0,1,2,\ldots$$

$$\Leftrightarrow\ p\ge p_0=a(0)$$

Since $a(k)$ is decreasing in k (Lemma 1), then $p\ge p_0$ which provides a necessary and sufficient condition for the $l(k)$ in Equation 3 to be non-decreasing. This completes the proof of result.

(ii) Let $X\le_{st}Y$ by Definition 2, then we have

$$P(X\le 0)\ge P(Y\le 0)\Rightarrow\binom{r+x-1}{x}p^{r}(1-p)^{x}$$

$$\ge\binom{r+y-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}$$

$$\Rightarrow p\ge\sqrt[r]{\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r}{\beta})}{\Gamma(1+\frac{r}{\beta}+\alpha)}}\qquad\Rightarrow p\ge p_1.$$

Conversely, suppose that $0<p_1<p<1.$ For $k=0,1,2,\ldots$ consider

$$\Delta_p(k)=\mathrm{P}(X\le k)-\mathrm{P}(Y\le k)$$

$$=\sum_{i=0}^{k}\binom{i+r-1}{i}p^{r}(1-p)^{i}$$

$$-\sum_{i=0}^{k}\left[\binom{i+r-1}{i}\sum_{j=0}^{i}\binom{i}{j}(-1)^{j}\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}\right],$$

and If $X\sim\ NB(r,p)$ and $X_1\sim\ NB(r,p_1),$ then $X\le_{lr}X_1$. Thus, we get $X\le_{st}X_1$.

Hence,

$$\sum_{i=0}^{k}\binom{i+r-1}{i}p^r(1-p)^i \ge \sum_{i=0}^{k}\binom{i+r-1}{i}p_1^{\,r}(1-p_1)^i,$$

and therefore $\Delta_p(k)\ge\Delta_{p_1}(k)$.

For fixed

$$k\in\{0,1,2,\ldots\},\ p_1=\sqrt[r]{\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r}{\beta})}{\Gamma(1+\frac{r}{\beta}+\alpha)}},\ p=exp(-\lambda)\quad \text{and}\quad \lambda\sim GE(\alpha,\beta).$$

We get

$$\Delta_{p_1}(k) = \sum_{i=0}^{k}\binom{i+r-1}{i}p^r(1-p)^i-\sum_{i=0}^{k}\left[\binom{i+r-1}{i}\sum_{j=0}^{i}\binom{i}{j}(-1)^j\frac{\Gamma(\alpha+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(1+\frac{r+j}{\beta}+\alpha)}\right]$$

$$=\sum_{i=0}^{k}\binom{i+r-1}{i}E(P^r)\left[1-\sqrt[r]{E(P^r)}\right]^i-\sum_{i=0}^{k}\binom{i+r-1}{i}E(P^r(1-P)^i)$$

Using concave function from part(ii) in Lemma 1, $\Delta_{p_1}(k)$ can be written as

$$\Delta_{p_1}(k)\ =\ \varphi_k(E(P^r))-E(\varphi_k(P^r))$$

Applying Jensen's inequality to concave function, we have $\varphi_k(E(P^r))\ge E(\varphi_k(P^r))$ and $\Delta_{p_1}(k)\ge 0$ for $\forall k\in\{0,1,2,\ldots\}$.Conversely, $Y\le_{st}Z$ implies that $\mathrm{P}(Z\le 0)\le\ \mathrm{P}(Y\le 0)$. This proves $p\ge p_1$.

(iii)The proofs of the results are obvious.

By Definition 4, we have

$$X\le_E Y\Leftrightarrow E(X)\le E(Y)\Leftrightarrow\frac{r(1-p)}{p}\le r(\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}-1)$$

$$\Leftrightarrow p\ge\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}\Leftrightarrow p\ge p2$$

Theorem 2

Suppose that $p=p_2$,then $X\le_{cx}Y$.

Proof

By Definition 3 and the result of Shake [2-3] we have

$$X\le_{cx}Y\Leftrightarrow E(X)=E(Y)\Leftrightarrow\frac{r(1-p)}{p}=r(\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}-1)$$

$$\Leftrightarrow p=\frac{\Gamma(\alpha+1)\Gamma(1-\frac{1}{\beta})}{\Gamma(1-\frac{1}{\beta}+\alpha)}$$

$$\Leftrightarrow p=p_2$$

However, in Figure 2 has shown to clarify that the NB-GE random variable is higher than NB random variable.

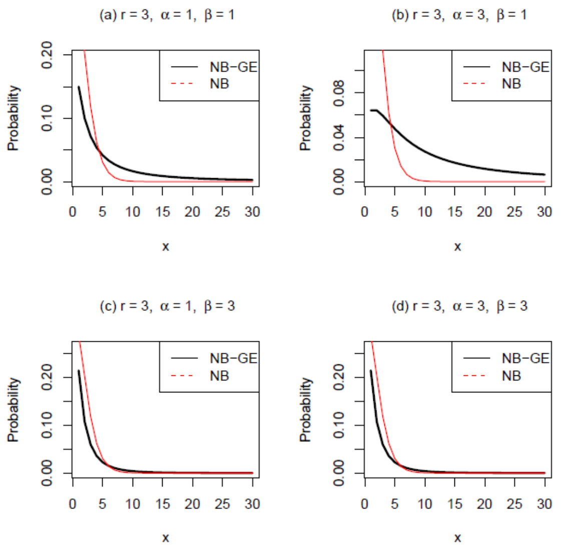


Figure 2. Probability of NB(r = 5; p = 0:65) and NB-GE(r;α;β)

4. Conclusion

In this paper presents comparisons of negative binomial random variable with negative binomial-generalized exponential random variable in the sense of stochastic orders included usual stochastic order, likelihood ratio order, convex order and expectation order. It is known that if Y is bigger than X in usual stochastic order, then it gives a condition for X is smaller than Y in the expectation order. In addition, if mean of X and Y are the same it implies that X is smaller than Y in the sense of convex order.

Acknowledgements

This research was made possible by Department of statistic, Kasetsart University, we thanks reviewers anonymous for comments made on an earlier version of this manuscript and for many inspiring discussions.

References

[1] Aryuyuen, S. and Bodhisuwan, W. The negative binomial-generalized exponential distribution. Applied Mathematical Sciences. 2013, 7(22): 1093.

[2] Shaked M. and Shanthikumar J. Stochastic Orders. New York: Academic Press, 2006.

[3] Kochar, S. Stochastic comparisons of order statistics and spacings: A review. International Scholarly Research Network. 2012: 47.

[4] Pudprommarat C. and Bodhisuwan, W. Stochastic orders comparisons of negative binomial distribution with negative binomial

- lindley distribution. Open Journal of Statistics.2012, . 2(2) : 208–212.

[5] Pudprommarat, C. Negative binomial - beta exponential distribution [Dissertation]. Bangkok: Kasetsart University, 2012.

[6] Li H. and Li, X. Eds., Stochastic Orders in Reliability and Risk in Honor of Professor Moshe Shaked. Lecture Notes in

Statistics 208. Springer, 2013.

[7] Kijima M. and Ohnishi, M. Stochastic orders and their anpplications in financial optimization. Mathematical Methods of

Operations Research.1999, 50(2): 351–572.

[8] Zakerzadeh H. and DolatiA. Generalized lindley distribution. Journal of Mathematical Extension.2009, 3(2): 13–25.

Some methods for addressing publication bias in meta-analysis

April Albertine

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA, april9@umbc.edu

Abstract

In meta-analysis, the science of combining the results of many different primary studies in order to arrive at an overall conclusion about the question of interest, the researcher must make a serious effort to collect all relevant available studies (ranging from peer-reviewed articles to master's theses) for inclusion. However, the difficulty may still remain that some studies may be unpublished and unavailable to the researcher. This is the well-known "file drawer problem," in which studies with nonsignificant results languish in the file drawers of authors and journal editors, who are presumably less likely to submit or accept for publication studies that fail to reject the null hypothesis. This publication bias may lead to an overall conclusion of significance when in fact, including the unavailable studies might have yielded an overall nonsignificant result. In order to account for this bias, we build on two approaches outlined and developed by Iyengar and Greenhouse (1988). The first approach determines the number of nonsignificant studies that must exist in order for the meta-analysis to yield a just barely nonsignificant result; we provide a correction. The second approach incorporates a parameter for publication bias into the likelihood of the effect size, and then estimates the publication bias parameter and the effect size using maximum likelihood estimation. We simplify the procedure using a normal approximation. Finally, we present an application of both approaches.

Keywords: Fail safe sample size; file drawer problem; meta-analysis; publication bias; selection models

Corresponding Author
E-mail Address: april9@umbc.edu

Confidence intervals for a coefficient of quartile variation with bootstrap method

Anurak Tongkaw[1*] and Vanida Pongsakchat[2]

[1]*Department of Mathematics, Burapha University, Mueang, Chonburi 20131, Thailand, airstatistic@hotmail.com*
[2]*Department of Mathematics, Burapha University, Mueang, Chonburi 20131, Thailand, vanida@buu.ac.th*

Abstract

The new confidence intervals for a coefficient of quartile variations are proposed and compared with the approximate confidence interval of Bonett [1]. The proposed methods are modified from the Bonett method and bootstrap method. Monte Carlo simulation results for selected normal distribution and non- normal distributions show that the new confidence intervals perform better than the Bonett confidence interval in terms of coverage probability and average length in most cases.

Keywords: Confidence interval, coefficient of quartile variation, bootstrap, bootstrap-*t*, coverage probability, average length

*Corresponding Author
E-mail Address: airstatistic@hotmail.com

1. Introduction

The population coefficient of quartile variation (*CQV*) is a descriptive measure of dispersion which may be preferred when sampling from a non-normal distribution (Bonett [1]). The *CQV* can be obtained from

$$CQV = (Q_3 - Q_1)/(Q_3 + Q_1) \qquad (1)$$

where Q_1 is the population 25th percentile and Q_3 is the population 75th percentile.

Bonett [1] proposed a confidence interval for the coefficient of quartile variation and found in simulation study that the proposed method has a coverage probability very close to confidence levels when sampling from non-normal distributions. However, Bonett's confidence interval performs well only when a sample size is large and needs to estimate the probability density function of sample 25th and 75th percentiles.

Efron and Tibshirani [2] introduced the bootstrap method as a tool to estimate the standard error of a statistic and use this method to find confidence interval for unknown parameters.

This paper aims to find the confidence intervals for a coefficient of quartile variation by using the Bonett and bootstrap methods. The coverage probability and average length of these confidence intervals are compared.

2. The Bonett confidence interval

Let Y be a continuous random variable with positive support and $Y_1, Y_2, ..., Y_n$ be a random sample of size *n*. Let $\hat{Q}_1$ and $\hat{Q}_3$ denote the sample 25th and 75th percentiles respectively.

The Bonett $(1-\alpha)100\%$ confidence interval for a coefficient of quartile variation is

$$\exp\left\{\log(D/S)c \pm z_{1-\alpha/2}v^{1/2}\right\} \qquad (2)$$

where $c = n/(n-1)$, $D = \hat{Q}_3 - \hat{Q}_1$ and $S = \hat{Q}_3 + \hat{Q}_1$. The estimate variance of $\log(D/S)$ is

$$v = (1/16n)\left\{\left(\frac{3}{\hat{f}_1^{\,2}} + \frac{3}{\hat{f}_3^{\,2}} - \frac{2}{\hat{f}_1^{\,2}\hat{f}_3^{\,2}}\right)\Big/D^2 + \left(\frac{3}{\hat{f}_1^{\,2}} + \frac{3}{\hat{f}_3^{\,2}} + \frac{2}{\hat{f}_1^{\,2}\hat{f}_3^{\,2}}\right)\Big/S^2 - 2\left(\frac{3}{\hat{f}_3^{\,2}} - \frac{3}{\hat{f}_1^{\,2}}\right)\Big/DS\right\}$$

$$\hat{f}_1^{\,2} = 3(z_{1-\alpha^*/2})^2 \Big/ \left\{4n(Y_{(b)} - Y_{(a)})^2\right\}$$

$$\hat{f}_3^{\,2} = 3(z_{1-\alpha^*/2})^2 \Big/ \left\{4n(Y_{(d)} - Y_{(c)})^2\right\}$$

where $\hat{f}_1$ and $\hat{f}_3$ are the estimates of the probability density function of $\hat{Q}_1$ and $\hat{Q}_3$ respectively and $Y_{(j)}$ is the *j*th order statistic,

$a = n/4 - 1.96(3n/16)^{1/2}$
$b = n/4 + 1.96(3n/16)^{1/2}$

$c = n+1-b$, $d = n+1-a$

Round *a* and *b* up to the nearest integer with $a \ge 1$. The $z_{1-\alpha^*/2}$ is the $1-\alpha^*/2$ quantile of the standard normal distribution and $\alpha^* = 1 - \sum_{i=a}^{b-1} \binom{n}{i} (1/4)^i (3/4)^{n-i}$. For large *n* set $\alpha^* = 0.05$.

3. Proposed confidence intervals

In this section, two new confidence intervals for a coefficient of quartile variation are proposed.

The first proposed confidence interval, the Bonett-bootstrap confidence interval, is created by combining the Bonett method with bootstrap. The point estimate of variance of $\log(D/S)$ in (2) is obtained using bootstrap. Hence, the Bonett-bootstrap $(1-\alpha)100\%$ confidence interval for a coefficient of quartile variation is

$$\exp\left\{\log(D/S)c \pm t^*_{1-\alpha/2} v_B^{1/2}\right\} \tag{3}$$

where $c = n/(n-1)$. Let $\hat{Q}_1^*$ and $\hat{Q}_3^*$ denote the sample 25th and 75th percentiles obtained from bootstrap method respectively and $D_B = \hat{Q}_3^* - \hat{Q}_1^*$, $S_B = \hat{Q}_3^* + \hat{Q}_1^*$. The bootstrap estimate variance of $\log(D/S)$ is given by

$$v_B = (1/16n)\left\{ \left(\frac{3}{\hat{f}_{1B}^2} + \frac{3}{\hat{f}_{3B}^2} - \frac{2}{\hat{f}_{1B}^2 \hat{f}_{3B}^2} \right) \Big/ D_B^2 \right.$$
$$+ \left(\frac{3}{\hat{f}_1^2} + \frac{3}{\hat{f}_3^2} + \frac{2}{\hat{f}_{1B}^2 \hat{f}_{3B}^2} \right) \Big/ S_B^2$$
$$\left. -2\left(\frac{3}{\hat{f}_{3B}^2} - \frac{3}{\hat{f}_{1B}^2} \right) \Big/ D_B S_B \right\}$$

$$\hat{f}_{1B}^2 = 3(z_{1-\alpha^*/2})^2 \Big/ \left\{ 4n(\hat{Y}_{(b)}^* - \hat{Y}_{(a)}^*)^2 \right\}$$

$$\hat{f}_{3B}^2 = 3(z_{1-\alpha^*/2})^2 \Big/ \left\{ 4n(\hat{Y}_{(d)}^* - \hat{Y}_{(c)}^*)^2 \right\}$$

where $\hat{f}_{1B}$ and $\hat{f}_{3B}$ are the estimate of the probability density function of $\hat{Q}_1$ and $\hat{Q}_3$ from bootstrap method. $\hat{Y}_{(j)}^*$ is the *j*th order statistic from bootstrap method and for large *n*, $t^*_{1-\alpha/2}$ from bootstrap method is used.

The second proposed confidence interval for a coefficient of quartile variation is created using the bootstrap percentile method. Hence, the bootstrap $(1-\alpha)100\%$ confidence interval for a coefficient of quartile variation is

$$[\widehat{CQV}_{BL}, \widehat{CQV}_{BU}] \tag{4}$$

where $\widehat{CQV}_{BL}$ and $\widehat{CQV}_{BU}$ are the $(\alpha/2)100$th and $(1-\alpha/2)100$th percentile of the bootstrap sample coefficient of quartile variation.

4. Simulation results

This section provides simulation studies for the coverage probabilities and the average lengths of confidence intervals as proposed in section 3 and the Bonett confidence interval. The nominal level is 95%. 10,000 samples of size $n = 10, 15, 25, 50$ and 100 were generated from Normal(4,2), Lognormal(0,1) and Gamma(1.5,1). For the bootstrap method, 1,000 samples were drawn from the original sample. The simulation results are reported in Table 1, 2 and 3.

The results suggest that when the distribution is normal the bootstrap confidence interval (4) has coverage probabilities closer to the nominal level compare to the other two confidence intervals. A part from a coverage probability, the bootstrap confidence interval (4) also has the shortest average length.

For the non-normal distributions, the bootstrap confidence interval has coverage probabilities closer to the nominal level when *n* = 10, 15, 25 and 50. However, when *n*=100, the Bonett (2) and Bonett-bootstrap confidence interval (3) has better coverage probability.

In the case of small sample sizes and the distribution is non-normal, the Bonett confidence interval (2) has problem with the average length and gives the longest average length.

5. Conclusion

The Bonett-bootstrap confidence interval (3), the bootstrap confidence interval (4) and the Bonett confidence interval (2) for a coefficient of quartile variation were studied and compared. Considering coverage probabilities, the bootstrap confidence interval (4) was the best method compare to the other two methods especially when the sample size is small for both normal and non-normal distributions. The advantages of the bootstrap confidence interval are that this method uses the basic bootstrap percentile method to obtain the confidence interval of a coefficient of quartile variation does not need any assumptions about the distribution of the statistic and the calculation is simple.

6. Acknowledgements

The author would like to thank the reviewers for their constructive advice which have improved this paper.

Table 1: The estimated coverage probability and average length of a 95% confidence intervals for Normal (4,2).

Method		Sample sizes (n)				
		10	15	25	50	100
Bonett	Coverage probability	0.9679	0.9826	0.9831	0.9746	0.9350
	Average length	0.8665	0.6430	0.4775	0.2944	0.1690
Bonett-bootstrap	Coverage probability	0.9260	0.9618	0.9754	0.9752	0.9350
	Average length	0.7118	0.5708	0.4996	0.2912	0.1673
Bootstrap	Coverage probability	0.9442	0.9464	0.9570	0.9564	0.9560
	Average length	0.6222	0.4571	0.3696	0.2506	0.1756

Table 2: The estimated coverage probability and average length of a 95% confidence intervals for Lognormal (0,1).

Method		Sample sizes (n)				
		10	15	25	50	100
Bonett	Coverage probability	0.9760	0.9867	0.9800	0.9800	0.9499
	Average length	6.36e+40	2.9e+15	5.7281	0.3854	0.2076
Bonett-bootstrap	Coverage probability	0.9266	0.9676	0.9757	0.9842	0.9497
	Average length	0.7556	1.1337	0.7912	0.3801	0.2042
Bootstrap	Coverage probability	0.9535	0.9589	0.9686	0.9704	0.9689
	Average length	0.6413	0.5254	0.4348	0.2979	0.2095

Table 3: The estimated coverage probability and average length of a 95% confidence intervals for Gamma (1.5,1).

Method		Sample sizes (n)				
		10	15	25	50	100
Bonett	Coverage probability	0.9749	0.9814	0.9793	0.9781	0.9414
	Average length	4.8706	1.6080	0.5549	0.3458	0.2027
Bonett-bootstrap	Coverage probability	0.9254	0.9576	0.9642	0.9802	0.9460
	Average length	0.6221	0.5800	0.4900	0.3403	0.2007
Bootstrap	Coverage probability	0.9488	0.9512	0.9674	0.9666	0.9652
	Average length	0.6218	0.5154	0.4332	0.3002	0.2131

References

[1] Bonett DG, Confidence interval for a coefficient of quartile variation. Comput Stat Data Anal. 2006; 50: 2953 - 2957.

[2] Efron B, Tibshirani RJ. An Introduction to the bootstrap. Boca Raton: Chapman & Hall; 1994.

Parameter estimation of the length-biased exponentiated inverted Weibull distribution

Palakorn Seenoi[1*] and Winai Bodhisuwan[2]
[1]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok 10903, Thailand, seenoi_p@hotmail.com*
[2]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok 10903, Thailand, fsciwnb@ku.ac.th*

Abstract

We obtained the maximum likelihood and Bayes estimators of parameters of the length-biased exponentiated inverted Weibull distribution. Bayesian estimation procedure has been discussed under the consideration of the square error loss function while the model parameters follow the gamma prior distributions. The performances of the maximum likelihood and Bayes estimators are compared in term of the simulation study. Bayes estimates are, generally, better than the maximum likelihood against the proposed prior in sense of having smaller mean square error.

Keywords: Length-biased exponentiated inverted Weibull distribution, maximum likelihood estimation, Bayesian estimation, square error loss function, Markov chain Monte Carlo method

*Corresponding Author
E-mail Address: seenoi_p@hotmail.com

1. Introduction

Recently the length-biased exponentiated inverted Weibulll (LBEIW) distribution has been proposed by Seenoi, Supapakorn, and Bodhisuwan [1] in 2014.

The LBEIW distribution has the following density function

$$g(x)=\frac{\beta\theta^{1-\frac{1}{\beta}}}{\Gamma(1-\frac{1}{\beta})}x^{-\beta}\{\exp(-x^{-\beta})\}^{\theta}; \qquad (1)$$

$$x>0,\beta>1,\theta>0,$$

and the distribution function

$$G(x)=\frac{\Gamma(1-\frac{1}{\beta},\frac{\theta}{x^{\beta}})}{\Gamma(1-\frac{1}{\beta})}, \qquad (2)$$

where $\Gamma(s,x)=\int_x^{\infty}t^{s-1}e^{-t}\,\mathrm{d}t$ is an upper incomplete gamma function.

Also, the survival and hazard functions of the LBEIW distribution with two shape parameters β and θ are given by

$$S(x)=\frac{\gamma(1-\frac{1}{\beta},\frac{\theta}{x^{\beta}})}{\Gamma(1-\frac{1}{\beta})}, \qquad (3)$$

and

$$h(x)=\frac{\beta\theta^{1-\frac{1}{\beta}}x^{-\beta}\{\exp(-x^{-\beta})\}^{\theta}}{\gamma(1-\frac{1}{\beta},\frac{\theta}{t^{\beta}})}, \qquad (4)$$

where $\gamma(s,x)=\int_0^{x}t^{s-1}\exp(-t)\,\mathrm{d}t$ is a lower incomplete gamma function.

Some plots of the LBEIW distribution function with specific parameter values are shown in Figure 1.

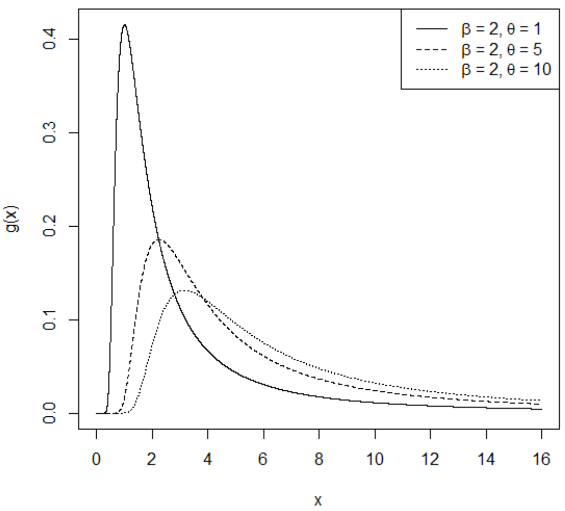


Figure 1: The probability density function of the LBEIW distribution for selected values of β and θ.

In Bayesian approach, we need to integrate over the posterior distribution and the problem is that the integrals are usually impossible to evaluate analytically. Markov chain Monte Carlo (MCMC) technique is a Monte Carlo integration method which draws samples from the target posterior distribution. MCMC procedure provided a convenient and efficient way to sample from complex, high-dimensional statistical distributions. Recently, application of the MCMC method to the estimation of parameters or some other vital properties about statistical models is very common. Pang et al. [2] claimed that MCMC is quite versatile and flexible for use in parameter estimation of the three-parameter Weibull distribution. Soliman et al. [3] considered the maximum likelihood and Bayesian inferences of the unknown parameters of modified Weibull distribution and they used MCMC technique to obtain the Bayes estimates.

The main objective of this article is to estimate the two unknown parameters of the LBEIW distribution. We compute the approximate Bayes estimators under the assumption of independent gamma priors of the unknown parameters and compare them with the maximum likelihood estimators (MLE) by Monte Carlo simulations.

The rest of the paper is organized as follows. In the next section, the MLE of the unknown parameters is presented. In Section 3, we propose the Bayes estimators of the unknown parameters under squared error loss function (SELF). Numerical results of the approximate MLE and Bayes estimator based on MCMC technique are provided in Section 4. Finally conclusions appear in Section 5.

2. The maximum likelihood estimation

Maximum likelihood estimation is one of the most popular methods for parameter estimation of continuous distributions because of its attractive properties, such as consistency, asymptotic unbiased, asymptotic efficiency, and asymptotic normality. In this section we discuss the MLE of the parameters of the LBEIW distribution. Suppose that $X_1, X_2, ..., X_n$ is a sample of size n obtained from the LBEIW distribution with parameters β and θ. It is assumed that both parameters β and θ are unknown, the likelihood function for the parameters β and θ is given by

$$L(\beta,\theta;x)=\frac{\beta^n\theta^{n(1-\frac{1}{\beta})}}{\left[\Gamma(1-\frac{1}{\beta})\right]^n}\prod_{i=1}^{n}x_i^{-\beta}\exp(-\theta\sum_{i=1}^{n}x_i^{-\beta}). \quad (5)$$

By taking logarithm of Eq. (5), the log-likelihood function of X can be written as

$$\begin{aligned} l(\beta,\theta;x) &= n\log\beta+n\log\theta-\frac{n}{\beta}\log\theta \\ &\quad -\beta\sum_{i=1}^{n}\log x_i-\theta\sum_{i=1}^{n}x_i^{-\beta} \\ &\quad -n\log\Gamma\left(1-\frac{1}{\beta}\right). \end{aligned} \quad (6)$$

The MLE solutions of parameters β and θ are obtained by setting the first partial derivatives of Eq. (6) to zero with respective to β and θ, respectively. These simultaneous equations are

$$\begin{aligned} \frac{\partial l}{\partial\beta} &= \frac{n}{\beta}+\frac{n}{\beta^2}\log\theta-\sum_{i=1}^{n}\log x_i \\ &\quad +\theta\left[\sum_{i=1}^{n}x_i^{-\beta}\right]\left[\sum_{i=1}^{n}\log x_i\right]-n\psi\left(1-\frac{1}{\beta}\right), \end{aligned} \quad (7)$$

$$\frac{\partial l}{\partial\theta} = \frac{n}{\theta}-\frac{n}{\beta\theta}-\sum_{i=1}^{n}x_i^{-\beta}, \quad (8)$$

where $\psi(z)=\frac{d}{dz}\ln\Gamma(z)=\frac{\Gamma'(z)}{\Gamma(z)}$ be a digamma function, which is the logarithmic derivative of the gamma function. It may be noted that this is an implicit equations in $\hat{\beta}_{MLE}$ and $\hat{\theta}_{MLE}$, so it cannot be solved analytically. We propose to solve it by using a numerical procedure with the Newton-Raphson method.

3. Bayes estimation

Now, we deal with the problem of estimating the parameters β and θ of LBEIW distribution under SELF. Since the parameters β and θ are assumed to be unknown, the prior distribution for β and θ are taken to be $Gamma(a,b)$ and $Gamma(c,d)$ respectively. So, the joint prior distribution of β and θ is of the form

$$\begin{aligned} &\pi(\beta,\theta)\propto\beta^{a-1}\theta^{c-1}e^{-b\beta-d\theta}, \\ &\qquad \beta>0,\theta>0,a>0,b>0,c>0,d>0. \end{aligned} \quad (9)$$

Combining the prior given by Eq. (9) with likelihood given by Eq. (5), we can easily obtain the posterior distribution of β and θ as

$$\pi(\beta,\theta\,|\,x)=\frac{J_1}{J_0}, \quad (10)$$

where

$$J_1 = \left(\frac{\prod_{i=1}^{n} x_i^{-\beta}}{\left[\Gamma(1-\frac{1}{\beta}) \right]^n} \right) \beta^{n+a-1} \theta^{n+c-\frac{n}{\beta}-1}$$

$$\times \exp(-\theta \sum_{i=1}^{n} x_i^{-\beta} - b\beta - d\theta),$$

and

$$J_0 = \int_0^\infty \int_0^\infty J_1 d\beta d\theta .$$

Usually the Bayes estimators are obtained under SELF

$$l(\phi,\hat{\phi}) = E[(\phi-\hat{\phi})^2],$$

where $\hat{\phi}$ is the estimate of the parameter ϕ and the Bayes estimator $\hat{\phi}$ of ϕ come out to be $E_\phi[\phi]$, where E_ϕ denotes the posterior expectation.

Then, the Bayes estimator $\hat{\phi}$ of ϕ under SELF is given by

$$\hat{\phi} = E_\phi[(\phi-\hat{\phi})^2].$$

It may be noted here that the integrals involved in the expressions for the Bayes estimators $\hat{\beta}_{BE}$ and $\hat{\theta}_{BE}$ cannot be obtained analytically and one needs numerical techniques for computations. Therefore, we have proposed to use MCMC methods. In MCMC techniques, we considered the Metropolis-Hastings [4] algorithms to generate samples from posterior distributions and these samples are used to compute Bayes estimates. The Gibbs is an algorithm for simulating from the full conditional posterior distributions while the random-walk Metropolis algorithm generates samples from an arbitrary proposal distribution.

The Bayes estimates of β and θ against the SELF are respectively obtained as

$$\hat{\beta}_{BE} = E[\beta \mid x] = E_\beta[\beta,\theta \mid x],$$

and

$$\hat{\theta}_{BE} = E[\theta \mid x] = E_\theta[\beta,\theta \mid x].$$

4. Numerical comparison study

In this section a Monte Carlo simulation study is presented to illustrate all the estimation methods described in the preceding sections. All the computations are performed using R code [5].

We compare the performances of the MLE and Bayes estimates with respect to the SELF in terms of mean squares errors (MSE).

In order to assess the statistical performances of these estimates, a simulation study is conducted. The random samples are generated as follows:

1). For given value of the prior parameters (a,b,c,d), generate random values for β and θ from the gamma distributions.

2). Using β and θ obtained in step 1), and generate random samples of different sizes: $n = 15,30,50,100,150,200,300,500,1000$ from the LBEIW distribution.

3). The MLE of the parameters β and θ are obtained by iteratively solving the Eq. (7) and (8).

4). We run the Gibbs sampler with Metropolis-Hastings to generate a Markov chain with 10,000 observations. Based on MCMC samples, the Bayes estimates relative to SELF.

5). The above steps are repeated 1,000 times and the MSE is computed for different sample sizes n .

In all above cases the generated value of (β , θ) are (2,1), (2,5) and (2,10) as the true values. The average Bayes estimates (BE), variance and MSE are displayed in Tables 1-3.

From Tables 1-3, the BE relative to SELF are better than their corresponding MLE, for most cases of n . When the effective sample sizes n are increase the MSE of the all estimates are decrease.

5. Conclusion

In this work, estimation of the parameters is obtained when data are drawn from the LBEIW distribution. The maximum likelihood and Bayes methods are used in estimation. In the Bayes approach, the estimators are obtained under squared error loss function. The MCMC method provides an alternative method for parameter estimation of the LBEIW distribution. It is more flexible as compared to the MLE method. Indeed, the MCMC sample may be used to completely summarize posterior distribution about the parameters, through a kernel estimate. The MCMC procedure can easily be applied to complex Bayesian modeling relating to LBEIW distribution. The estimation methods are compared based on the MSE. From Tables 1-3, it may be noticed that the Bayes estimates are, generally, better than the MLE against the proposed prior in sense of having smaller MSE. Even for sample size as small as $n = 1000$, good Bayes estimates with smaller MSE.

Table 1: Simulation study of length-biased exponentiated inverted Weibull distribution (True parameters: $\beta = 2$, $\theta = 1$)

Sample size	Parameter	MLE			BE		
		Estimate	VAR	MSE	Estimate	VAR	MSE
15	β	2.1253	0.0984	0.1140	2.0383	0.0829	0.0843
	θ	1.2506	0.4964	0.5587	0.9916	0.2745	0.2743
30	β	2.0585	0.0375	0.0408	2.0217	0.0346	0.0351
	θ	1.1091	0.1222	0.1339	1.0002	0.0991	0.0990
50	β	2.0293	0.0223	0.0232	2.0073	0.0212	0.0212
	θ	1.0577	0.0613	0.0645	0.9938	0.0535	0.0535
100	β	2.0187	0.0095	0.0098	2.0076	0.0092	0.0093
	θ	1.0258	0.0240	0.0247	0.9946	0.0226	0.0226
150	β	2.0114	0.0066	0.0067	2.0038	0.0065	0.0065
	θ	1.0254	0.0173	0.0179	1.0035	0.0166	0.0166
200	β	2.0101	0.0046	0.0047	2.0044	0.0045	0.0045
	θ	1.0158	0.0110	0.0112	0.9996	0.0106	0.0106
300	β	2.0065	0.0034	0.0035	2.0026	0.0034	0.0034
	θ	1.0109	0.0075	0.0076	0.9999	0.0073	0.0073
500	β	2.0042	0.0019	0.0019	2.0018	0.0019	0.0019
	θ	1.0061	0.0050	0.0050	0.9992	0.0049	0.0049
1000	β	2.0011	0.0010	0.0010	1.9999	0.0009	0.0009
	θ	1.0038	0.0024	0.0024	1.0002	0.0024	0.0024

Table 2: Simulation study of length-biased exponentiated inverted Weibull distribution (True parameters: $\beta = 2$, $\theta = 5$)

Sample size	Parameter	MLE			BE		
		Estimate	VAR	MSE	Estimate	VAR	MSE
15	β	2.1480	0.0979	0.1196	1.9894	0.0636	0.0637
	θ	7.5796	40.6067	47.2199	5.1589	11.4148	11.4286
30	β	2.0709	0.0370	0.0419	2.0006	0.0305	0.0304
	θ	6.0175	6.9347	7.9631	5.1291	4.6152	4.6272
50	β	2.0374	0.0217	0.0231	1.9961	0.0192	0.0192
	θ	5.5770	2.7325	3.0628	5.0800	2.2107	2.2149
100	β	2.0186	0.0106	0.0109	1.9986	0.0099	0.0099
	θ	5.2540	1.0663	1.1297	5.0223	0.9606	0.9602
150	β	2.0140	0.0070	0.0072	2.0006	0.0068	0.0068
	θ	5.1626	0.6289	0.6547	5.0112	0.5950	0.5945
200	β	2.0096	0.0051	0.0052	2.0000	0.0049	0.0049
	θ	5.1338	0.4994	0.5168	5.0252	0.4755	0.4757
300	β	2.0051	0.0032	0.0032	1.9982	0.0031	0.0031
	θ	5.0561	0.2989	0.3018	4.9785	0.2892	0.2894
500	β	2.0047	0.0020	0.0021	2.0005	0.0020	0.0020
	θ	5.0679	0.1944	0.1989	5.0214	0.1909	0.1912
1000	β	2.0039	0.0009	0.0010	2.0018	0.0009	0.0009
	θ	5.0331	0.0886	0.0896	5.0093	0.0876	0.0876

Table 3: Simulation study of length-biased exponentiated inverted Weibull distribution (True parameters: $\beta = 2$, $\theta = 10$)

Sample size	Parameter	MLE			BE		
		Estimate	VAR	MSE	Estimate	VAR	MSE
15	β	2.1252	0.0988	0.1144	1.9342	0.0567	0.0610
	θ	16.7403	377.0209	422.0599	9.7780	51.9299	51.9272
30	β	2.0494	0.0384	0.0408	1.9645	0.0301	0.0313
	θ	12.2678	41.6895	46.7906	9.8828	21.2611	21.2535
50	β	2.0321	0.0214	0.0224	1.9854	0.0196	0.0198
	θ	11.2998	16.2086	17.8817	10.1493	13.2233	13.2324
100	β	2.0143	0.0095	0.0097	1.9906	0.0090	0.0091
	θ	10.4830	5.0584	5.2867	9.9380	4.6262	4.6254
150	β	2.0129	0.0066	0.0068	1.9968	0.0063	0.0063
	θ	10.4417	3.3393	3.5310	10.0710	3.0670	3.0690
200	β	2.0074	0.0049	0.0049	1.9952	0.0047	0.0047
	θ	10.3549	2.6372	2.7604	10.0734	2.4666	2.4696
300	β	2.0049	0.0034	0.0035	1.9967	0.0034	0.0034
	θ	10.1530	1.7352	1.7568	9.9651	1.6594	1.6590
500	β	2.0050	0.0018	0.0018	2.0000	0.0018	0.0018
	θ	10.1337	0.9048	0.9218	10.0201	0.8906	0.8901
1000	β	2.0015	0.0010	0.0010	1.9989	0.0010	0.0010
	θ	10.0483	0.4562	0.4581	9.9906	0.4533	0.4529

Acknowledgements

We are grateful to Department of Statistics and the Commission on Faculty of Science Kasetsart University Postgraduate Studentship (ScKUPGS) for supporting the Scholarships.

References

[1] Seenoi P, Supapakorn T, Bodhisuwan W. The length-biased exponentiated inverted weibull distribution. International Journal of Pure and Applied Mathematics. 2014; 92(2): 191-206.

[2] Pang K, Hou H, Yu T. On a proper way to select population failure distribution and a stochastic optimization method in parameter estimation. European Journal of Operational Research. 2007; 177: 604-611.

[3] Soliman A, Abd-Ellah H, Abou-Elheggag A, Ahmed A. Modified weibull model: A bayes study using mcmc approach basedon progressive censoring data. Reliability Engineering & System Safety. 2012; 100: 48-57.

[4] Metropolis N, Rosenbluth W, Rosenbluth N, Teller H, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics. 1953; 21: 1087-1092.

[5] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. [Online]. Available: http://www.R-project.org/

Modified Anderson-Darling test based on the likelihood ratio for skewed distributions

Phannaphat Saethow[1*] and Jutaporn Neamvonk [2]
[1]Department of Mathematics, Burapha University, Chonburi, 20000, Thailand, panmod99@gmail.com
[2] Department of Mathematics, Burapha University, Chonburi, 20000, Thailand, jutaporn@buu.ac.th

Abtract

Before analyzing the data in statistical studies, we need to know the form of distribution of the data. A goodness of fit test is used to determine how well the data fits. The Anderson-Darling test is a well known goodness of fit test based on the empirical distribution function. Many years ago, researchers suggested modifications of Anderson-Darling test for several distributions. In this study, the author proposes a modified Anderson-darling Z_A test which is a combination of the two modified Anderson-Darling test (Ahmad et al., 1988; Zhang, 2002). The critical values of the modified Anderson-darling Z_A test were obtained with a variety of sample sizes. The probability of type I error and the performance of the proposed test among Kolmogorov-Smirnov, Anderson-Darling, and modified Anderson-Darling (Ahmad et al., 1988) tests are investigated by using Monte Carlo simulation. The results show that the modified Anderson-darling Z_A test is more powerful than the other tests in most cases.

Keywords: Goodness of fit test, Kolmogorov-Smirnov test, Anderson-Darling test, Modified Anderson-Darling test, Modified Anderson-darling Z_A test, type I error probability, power of the test

1. Introduction

Natural data have different types of distributions. To do statistical analysis or forecasting, the researcher need to know about the form of distribution that fits to the data in order to choose correct statistical tools for analyzing data. Therefore, the results obtained are accurate and reliable. Goodness of fit test is a method to examine whether the data fits any specified distribution. The two well-known statistical goodness of fit tests based on the empirical distribution functions are Kolmogorov-Smirnov test (KS) and Anderson-Darling test (AD).

The KS test was first suggested by Kolmogorov [1]. This test is widely used in solving goodness of fit problems. However, this test has lower power than the AD test that was proposed by Anderson and Darling [2]. The AD test is the most powerful when the given data comes from symmetric distribution (see also Stephens [3]). Furthermore, the AD test is more powerful than other tests for several skewed distributions: özmen [4] for Gamma, Sriamporn [5] for the Lognormal and Weibull, Pajjayakarn [6] for the Generalized Exponential, and Abd-Elfattah [7] for the Generalized Frechet models.

However, the traditional AD test gives equal weights to both tails of distribution. It seems the AD test is not appropriate for skewed distributions. To overcome this disadvantage, Ahmad et al. [8] presents an improvement of the AD test for skewed distributions, Modified Anderson-Darling test (MAD), which gives more weight to the lower tail or upper tail of the distributions by defining a weight function consistent with the tail that focuses on. For instance, in hydrology the frequency analysis of floods focuses on the upper tail of the distribution. Power studies of the MAD test have been proceeded with various distributions: Arshad et al. [9] for Generalized Pareto and Shin et al. [10] for Generalized Extreme Value and Generalized Logistic distributions.

In addition, Zhang[11] proposed Anderson-darling Z_A test (ZAD), a modification of the AD test for symmetric distributions, by using the likelihood ratio statistic together with adjusting weight function. Some studies focused on the power study of the proposed test such as Zhang and Wu [12] for normal and Abidi et al. [13] for Gumbel models.

In this study, modified Anderson-darling Z_A test (ZMAD), integrating concepts of the MAD and ZAD tests altogether, is proposed. The critical values of ZMAD test are created only for positive skewed distributions including Gamma and Lognormal distributions. Then the efficiency of the ZMAD test is investigated and compared with the KS, AD, and MAD tests in a significant level 0.05.

2. Research Methodology

2.1 The Gamma and Lognormal distributions

Let X be a random variable. Probability density function and cumulative distribution function of Gamma distribution are

$$f(x)=\frac{1}{\Gamma(k)\theta^{k}}x^{k-1}e^{-\frac{x}{\theta}},\qquad x>0,\ k>0,\ \theta>0$$

and $F(x)=\dfrac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)}$,

where $\gamma(k,x)=\int_{0}^{x}t^{k-1}e^{-t}dt$ and $\Gamma(k)=\int_{0}^{\infty}x^{k-1}e^{-x}dx$.

The k and θ are the shape and scale parameters respectively.

Probability density function and cumulative distribution function of Lognormal distribution are

$$f(x)=\frac{1}{x\sqrt{2\pi\sigma^{2}}}e^{-\frac{(\log(x)-\mu)^{2}}{2\sigma^{2}}},\ x>0,\ \mu>0,\ \sigma^{2}>0$$

$$F(x)=\frac{1}{2}\left[1+erf\left(\frac{\log(x)-\mu}{\sigma\sqrt{2}}\right)\right]$$

where $erf(x)=\frac{2}{\sqrt{\pi}}\int_{0}^{x}e^{-t^{2}}dt$.

The μ and σ^{2} are the scale and shape parameters respectively.

2.2 Goodness of fit tests

Let $X_1, X_2, X_3, ..., X_n$ be continuous random samples of n independent observations with order statistics $X_{(1)}, X_{(2)}, X_{(3)}, ..., X_{(n)}$. Suppose the null hypothesis is

$H_0 : F(x)=F_0(x)$ for all $x\in(-\infty,\infty)$

against the alternative

$H_1 : F(x)\neq F_0(x)$ for some $x\in(-\infty,\infty)$

where $F(x)$ is a distribution function, and $F_0(x)$ is a hypothesized distribution function.

Zhang [11] showed that the statistic for testing H_0 *versus* H_1 can be defined by

$$Z=\int_{-\infty}^{\infty}Z_t\,dw(t) \tag{1}$$

or

$$Z_{\max}=\sup_{t\in(-\infty,\infty)}\left\{Z_t w(t)\right\} \tag{2}$$

for all $t\in(-\infty,\infty)$, where $w(t)$ is a weight function and Z_t is replaced with the Pearson chi square test statistic

$$\chi_t^2=\frac{n\left\{F_n(t)-F_0(t)\right\}^2}{F_0(t)\left\{1-F_0(t)\right\}} \tag{3}$$

or the likelihood ratio test statistic

$$G_t^2=2n\left[F_n(t)\log\left\{\frac{F_n(t)}{F_0(t)}\right\}\right.$$
$$\left.+\left\{1-F_n(t)\right\}\log\left\{\frac{1-F_n(t)}{1-F_0(t)}\right\}\right]. \tag{4}$$

where $F_n(t)$ is the empirical distribution function of X, and $F_0(x)$ is the hypothesized distribution function.

2.2.1 Kolmogorov-Smirnov test (KS)

The KS test is a well known test which is suggested by Kolmogorov [1]. The KS statistic is created by replacing Z_t in equation (2) with χ_t^2 in equation (3) and defining $w(t)=\frac{1}{n}F_0(t)\left\{1-F_0(t)\right\}$ as the weight function. Therefore,

$$KS=\sup\left|F_n(t)-F_0(t)\right|.$$

2.2.2 Anderson-Darling test (AD)

The AD test is a powerful test which is suggested by Anderson and Darling [2]. The AD statistic is created by replacing Z_t in equation (1) with χ_t^2 in equation (3) and defining the weight function as $dw(t)=dF_0(t)$. Therefore,

$$AD=\int_{-\infty}^{\infty}\frac{n\left\{F_n(t)-F_0(t)\right\}^2}{F_0(t)\left\{1-F_0(t)\right\}}dF_0(t)$$
$$=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log F\left(x_{(i)}\right)\right.$$
$$\left.+\log\left\{1-F\left(x_{(n+1-i)}\right)\right\}\right].$$

2.2.3 Modified Anderson-Darling test (MAD)

The MAD test which is suggested by Ahmad et al. [8] is a modification of AD test that concentrate on a tail of the distribution. MAD statistics is created by replacing Z_t in equation (1) with χ_t^2 in equation (3) and the weight function is defined by the tail that is focused on.

For the lower tail, the weight function is $dw(t)=\left\{1-F_0(t)\right\}dF_0(t)$. Therefore,

$$AL=n\int_{-\infty}^{\infty}\frac{\left\{F_n(t)-F_0(t)\right\}^2}{F_0(t)}dF_0(t)$$
$$=-\frac{3n}{2}+2\sum_{i=1}^{n}F_0\left(x_{(i)}\right)$$

$$-\sum_{i=1}^{n}\left[\left(\frac{2i-1}{n}\right)\log\left\{F_0\left(x_{(i)}\right)\right\}\right].$$

For the upper tail, the weight function is $dw(t)=F_0(t)dF_0(t)$. Hence,

$$AU=n\int_{-\infty}^{\infty}\frac{\left\{F_n(t)-F_0(t)\right\}^2}{\left\{1-F_0(t)\right\}}dF_0(t)$$

$$=\frac{n}{2}-2\sum_{i=1}^{n}F_0\left(x_{(i)}\right)$$

$$-\sum_{i=1}^{n}\left[\left\{2-\left(\frac{2i-1}{n}\right)\right\}\log\left\{1-F_0\left(x_{(i)}\right)\right\}\right].$$

Note that the sum of the AL and the AU statistic is equal to traditional AD statistic, i.e. $AL+AU=AD$.

2.2.4 Modified Anderson-Darling Z_A test (ZMAD)

Zhang [11] proposes Anderson-Darling Z_A test (ZAD). This proposed statistics is a modification of the AD test based on the likelihood ratio statistics. The Z_A statistic was created by replacing Z_t in equation (1) with G_t^2 in equation (4) and defining the weight function as $dw(t)=\dfrac{1}{F_n(t)\left\{1-F_n(t)\right\}}dF_n(t)$. Therefore,

$$Z_A=2n\int_{-\infty}^{\infty}\left[\frac{1}{\left\{1-F_n(t)\right\}}\log\left\{\frac{F_n(t)}{F_0(t)}\right\}\right.$$

$$\left.+\frac{1}{F_n(t)}\log\left\{\frac{1-F_n(t)}{1-F_0(t)}\right\}\right]dF_n(t)$$

$$=2\sum_{i=1}^{n}\left[\frac{n}{n-i+\frac{1}{2}}\log\left\{\frac{i-\frac{1}{2}}{F_0\left(x_{(i)}\right)}\right\}\right.$$

$$\left.+\frac{n}{i-\frac{1}{2}}\log\left\{\frac{n-i+\frac{1}{2}}{n\left\{1-F_0\left(x_{(i)}\right)\right\}}\right\}\right].$$

However, the ZAD test is more powerful than traditional AD test only for symmetric distributions.

Here, the new statistics which is created by integrating the concepts of ZAD and MAD tests, Modified Anderson-Darling Z_A test (ZMAD), is proposed for skewed distributions. The ZMAD statistics is generated by replacing Z_t in equation (1) with G_t^2 in equation (4) and the weight function defined by the tail that is focused on. For the lower tail, the weight function is $dw(t)=\dfrac{1}{F_n(t)}dF_n(t)$. Therefore,

$$Z_{AL}=2n\int_{-\infty}^{\infty}\left[\log\left\{\frac{F_n(t)}{F_0(t)}\right\}\right.$$

$$\left.+\frac{\left\{1-F_n(t)\right\}}{F_n(t)}\log\left\{\frac{1-F_n(t)}{1-F_0(t)}\right\}dF_n(t)\right]$$

$$=2\sum_{i=1}^{n}\left[\log\left\{\frac{i-\frac{1}{2}}{nF_0\left(x_{(i)}\right)}\right\}\right.$$

$$\left.+\frac{n-i+\frac{1}{2}}{i-\frac{1}{2}}\log\left\{\frac{n-i+\frac{1}{2}}{n\left\{1-F_0\left(x_{(i)}\right)\right\}}\right\}\right].$$

For the upper tail, the weight function is $dw(t)=\dfrac{1}{\left\{1-F_n(t)\right\}}dF_n(t)$. Then,

$$Z_{AU}=2n\int_{-\infty}^{\infty}\left[\frac{F_n(t)}{\left\{1-F_n(t)\right\}}\log\left\{\frac{F_n(t)}{F_0(t)}\right\}\right.$$

$$\left.+\log\left\{\frac{1-F_n(t)}{1-F_0(t)}\right\}\right]dF_n(t)$$

$$=2\sum_{i=1}^{n}\left[\frac{i-\frac{1}{2}}{n-i+\frac{1}{2}}\log\left\{\frac{i-\frac{1}{2}}{nF_0\left(x_{(i)}\right)}\right\}\right.$$

$$\left.+\log\left\{\frac{n-i+\frac{1}{2}}{n\left\{1-F_0\left(x_{(i)}\right)\right\}}\right\}\right].$$

Note that the sum of the Z_{AL} and the Z_{AU} statistic is equal to Z_A statistic, i.e. $Z_{AL}+Z_{AU}=Z_A$.

2.3 Approximate critical values of the test statistics

This research considers only gamma and lognormal distributions which have positive skewness. Coefficient of skewness of the distributions are varied from heavy right skewed to nearly symmetric shape and the parameter of the distributions are evaluated for the specified coefficient of skewness. Monte Carlo simulation is used to generate critical values by the following steps.

1. The coefficient of skewness and the parameters of the distributions are defined in Table 1.

Table 1: The parameters of distributions

Coefficient of skewness	Parameter of Gamma distribution		Parameters of Lognormal distribution	
	θ	k	μ	σ^2
2.000	1	1	0	0.305
1.414	1	2	0	0.178
1.155	1	3	0	0.127
1.000	1	4	0	0.1
0.895	1	5	0	0.08
0.500	1	16	0	0.027
0.360	1	30	0	0.014
0.025	1	6400	0	0.00007

2. A random sample $X_1, X_2, X_3, ..., X_n$ from gamma and lognormal distributions are generated at each situation with sample sizes of n = 10, 20, 30, 50, 100 and 200.

3. The distribution parameters are estimated using maximum likelihood method and substituted in cumulative distribution function.

4. KS, AD, MAD, and ZMAD statistics were calculated

5. Repeat step 1-4 for 100,000 times.

6. For each statistic, 100,000 values of test statistic are ranked in ascending order. The 95th percentile is the critical value for significance level of 0.05.

7. Repeat step 1-6 for various sets of distribution parameters.

8. At each sample size, the critical values for various sets of parameter are very similar with the difference at the second decimal places. This shows that the critical values do not depend on the parameters. Therefore, the critical values of all sets of parameter are averaged to represent critical values at that sample size.

2.4 Type I error probability

The following steps are the investigation on probability of type I error for Gamma and Lognormal distributions by Monte Carlo simulation method.

1. A random sample $X_1, X_2, X_3, ..., X_n$ from null distributions: Gamma and Lognormal distributions are generated with sample sizes of n = 10, 20, 30, 50, 100 and 200.

2. The parameters are estimated by maximum likelihood method and substituted in cumulative distribution function.

3. The KS, AD, MAD, and ZMAD statistics are evaluated and compared with critical values corresponding to the null distribution and sample size of the tests.

4. Repeat step 1-3 for 10,000 times.

5. For each statistics, the type I error probabilities are calculated by dividing numbers of rejecting null hypothesis by 10,000.

2.5 Power study

For power study, the alternative positive skewness cases; Gamma, Lognormal, Weibull, and Beta distributions, are considered. Monte Carlo simulation is used to compute power of the test by the following steps.

1. A random sample $X_1, X_2, X_3, ..., X_n$ from alternative distributions is generated with sample sizes of n = 10, 20, 30, 50, 100 and 200.

2. The parameters are estimated by using maximum likelihood method and substituted in cumulative distribution function.

3. The KS, AD, MAD, and ZMAD statistic are calculated and compare with critical values corresponding to the null distribution and sample size for the tests.

4. Repeat step 1-3 for 10,000 times.

5. For each statistic, the powers of the tests are calculated by dividing numbers of rejecting the null hypothesis by 10,000.

3. Research Results and Discussion

The results of the critical values for the KS AD MAD ZMAD tests at the upper tail areas of the Gamma and Lognormal distributions with the sample sizes n = 10, 20, 30, 50, 100, and 200 at significant level 0.05 are shown in Table 2 and 3, respectively.

Table 2: Critical values of KS AD MAD ZMAD tests for Gamma distribution at α =0.05

n	10	20	30	50	100	200
ZMAD	2.9917	4.2879	5.1076	6.2174	7.7779	9.3764
MAD	0.3778	0.3873	0.3900	0.3929	0.3946	0.3953
AD	0.7315	0.7461	0.7511	0.7561	0.7596	0.7604
KS	0.2681	0.1948	0.1609	0.1260	0.0900	0.0640

Table 3: Critical values of KS AD MAD ZMAD tests for Lognormal distribution at α =0.05

n	10	20	30	50	100	200
ZMAD	3.0270	4.3420	5.1721	6.2719	7.8398	9.4437
MAD	0.3753	0.3854	0.3878	0.3893	0.3897	0.3907
AD	0.7247	0.7407	0.7438	0.7467	0.7485	0.7509
KS	0.2660	0.1933	0.1595	0.1248	0.0890	0.0634

Table 4: The probability of type I error

n	Gamma (3,1)				Lognormal (0,0.127)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0487	0.0512	0.0484	0.0503	0.0475	0.0470	0.0472	0.0524
20	0.0509	0.0508	0.0477	0.0455	0.0496	0.0527	0.0521	0.0553
30	0.0502	0.0498	0.0452	0.0464	0.0479	0.0505	0.0461	0.0522
50	0.0480	0.0501	0.0485	0.0480	0.0470	0.0469	0.0488	0.0472
100	0.0493	0.0499	0.0503	0.0486	0.0526	0.0489	0.0472	0.0459
200	0.0537	0.0531	0.0539	0.0505	0.0460	0.0483	0.0471	0.0485
n	Gamma (4,1)				Lognormal (0,0.1)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0513	0.0500	0.0498	0.0505	0.0511	0.0490	0.0486	0.0488
20	0.0512	0.0504	0.0490	0.0463	0.0516	0.0500	0.0540	0.0558
30	0.0527	0.0532	0.0524	0.0493	0.0489	0.0514	0.0490	0.0520
50	0.0496	0.0504	0.0539	0.0511	0.0515	0.0511	0.0515	0.0456
100	0.0507	0.0506	0.0507	0.0520	0.0484	0.0506	0.0512	0.0509
200	0.0482	0.0518	0.0503	0.0554	0.0526	0.0482	0.0485	0.0516
n	Gamma (5,1)				Lognormal (0,0.08)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0527	0.0536	0.0528	0.0520	0.0513	0.0483	0.0490	0.0482
20	0.0502	0.0509	0.0481	0.0454	0.0490	0.0524	0.0530	0.0538
30	0.0495	0.0475	0.0486	0.0468	0.0481	0.0488	0.0495	0.0532
50	0.0475	0.0488	0.0488	0.0458	0.0523	0.0478	0.0481	0.0508
100	0.0498	0.0477	0.0495	0.0499	0.0503	0.0498	0.0500	0.0480
200	0.0499	0.0492	0.0476	0.0520	0.0498	0.0449	0.0463	0.0487
n	Gamma (16,1)				Lognormal (0,0.027)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0497	0.0484	0.0489	0.0486	0.0500	0.0486	0.0500	0.0484
20	0.0512	0.0482	0.0484	0.0439	0.0531	0.0532	0.0524	0.0520
30	0.0471	0.0477	0.0470	0.0481	0.0487	0.0503	0.0524	0.0516
50	0.0481	0.0499	0.0530	0.0486	0.0530	0.0455	0.0475	0.0471
100	0.0512	0.0542	0.0504	0.0506	0.0497	0.0521	0.0516	0.0453
200	0.0476	0.0493	0.0504	0.0459	0.0447	0.0486	0.0446	0.0430

Table 4 presents the results of type I error probabilities of the KS, AD, MAD, and ZMAD tests for testing the Gamma and Lognormal distributions with various sample sizes and distribution parameters at significant level 0.05. It can be seen that the type I error probabilities of all tests are in the range of the criteria of Cochran [14] which ranges from 0.04 to 0.06 Therefore, all tests can control type I error probability very well.

Then, Table 5 and 6 present the powers of the KS, AD, MAD and ZMAD tests for testing Gamma and Lognormal distributions of sample sizes *n* at a significant level 0.05

Table 5: Power of the KS, AD, MAD, and ZMAD for Gamma distribution

n	Lognormal (0,0.127)				Weibull (0.5,1.433)				Beta (2,21)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0733*	0.0677	0.0604	0.0553	0.0439	0.0518	0.0566*	0.0558	0.0423	0.0458	0.0486	0.0494*
20	0.0834*	0.0800	0.0725	0.0667	0.0534	0.0613	0.0688*	0.0653	0.0443	0.0544	0.0576	0.0584*
30	0.0938	0.0950*	0.0881	0.0778	0.0615	0.0632	0.0714*	0.0705	0.0506	0.0537	0.0584	0.0613*
50	0.1132	0.1167*	0.1100	0.0857	0.0892*	0.0778	0.0870	0.0811	0.0549*	0.0510	0.0524	0.0514
100	0.1574	0.1792*	0.1779	0.1365	0.1272*	0.1054	0.1192	0.1043	0.0728*	0.0606	0.0650	0.0645
200	0.2480	0.2995	0.3174*	0.2289	0.2240*	0.1839	0.1977	0.1595	0.1096*	0.0747	0.0751	0.0708
n	Lognormal (0,0.1)				Weibull (0.5,1.563)				Beta (2,12.5)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0735*	0.0623	0.0575	0.0557	0.0486	0.0560	0.0576*	0.0567	0.0424	0.0484	0.0514	0.0516*
20	0.0827*	0.0784	0.0691	0.0632	0.0577	0.0634	0.0742*	0.0737	0.0455	0.0514	0.0574	0.0622*
30	0.0871	0.0888*	0.0825	0.0735	0.0744	0.0694	0.0815*	0.0787	0.0559	0.0588	0.0626	0.0650*
50	0.0962	0.1018*	0.0928	0.0775	0.1004*	0.0851	0.0964	0.0893	0.0734*	0.0585	0.0597	0.0631
100	0.1305	0.1514*	0.1489	0.1184	0.1799*	0.1437	0.1537	0.1282	0.1182*	0.0817	0.0881	0.0794
200	0.2109	0.2446	0.2520*	0.1867	0.3147*	0.2569	0.2698	0.2182	0.1916*	0.1276	0.1261	0.1071
n	Lognormal (0,0.08)				Weibull (0.5,1.668)				Beta (2,9.27)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0658*	0.0566	0.0541	0.0527	0.0487	0.0555	0.0633*	0.0585	0.0424	0.0494	0.0511	0.0542*
20	0.0745*	0.0705	0.0651	0.0621	0.0599	0.0666	0.0749*	0.0779	0.0554	0.0553	0.0641	0.0672*
30	0.0812	0.0833*	0.0784	0.0664	0.0836	0.0795	0.0916*	0.0845	0.0640	0.0628	0.0678	0.0708*
50	0.0878	0.0934*	0.0897	0.0710	0.1262*	0.1012	0.1138	0.1032	0.0921*	0.0716	0.0779	0.0724
100	0.1144	0.1324*	0.1246	0.1063	0.2047*	0.1598	0.1792	0.1466	0.1665*	0.1112	0.1122	0.0952
200	0.1588	0.2006	0.2097*	0.1599	0.3724*	0.3161	0.3426	0.2584	0.3135*	0.1954	0.1898	0.1508
n	Lognormal (0,0.027)				Weibull (0.5,2.2)				Beta (2,4.25)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0580*	0.0555	0.0529	0.0507	0.0537	0.0646	0.0732*	0.0684	0.0500	0.0597	0.0675*	0.0634
20	0.0622*	0.0609	0.0578	0.0574	0.0887	0.0804	0.0967*	0.0937	0.0906	0.0848	0.0954*	0.0852
30	0.0619	0.0662*	0.0599	0.0582	0.1239	0.1058	0.1293*	0.1123	0.1532*	0.1234	0.1323	0.1108
50	0.0662	0.0644*	0.0595	0.0545	0.2104*	0.1608	0.1799	0.1413	0.2829*	0.1852	0.1877	0.1480
100	0.0689	0.0799*	0.0713	0.0641	0.3897*	0.2931	0.3234	0.2392	0.5910*	0.3778	0.3648	0.2614
200	0.0733	0.0990	0.0994*	0.0776	0.6576*	0.5706	0.5945	0.4367	0.9251*	0.7091	0.6820	0.4850

Table 5 shows power of the tests when the null distribution is Gamma distribution and the alternative distributions are Lognormal, Weibull and Beta distributions with various sets of parameters. When the alternative distributions is Lognormal distribution, the ZMAD is more powerful than the other tests only for $n \le 20$, and when the sample sizes increased the MAD tests is better than all of the tests. For the alternative Weibull distribution, the AD test has higher power than other tests $n \le 30$. However, when $n \ge 50$, the ZMAD is the most powerful overall comparing to the KS, AD, and MAD. For the Beta distribution, the KS test is the best when $n \le 30$. For $n \ge 50$, the ZMAD is the most powerful. Except for Beta (2,4.25), when $n \le 20$, the AD test is better than the other tests. And the ZMAD is the best when $n \ge$ 30.

Table 6 shows powers of the test when the null distribution is Lognormal distribution and the alternative distribution are Gamma, Weibull and Beta distribution. The results show that the ZMAD test is the most powerful for every alternative distributions and sample sizes except for n =10. The AD test has higher power than the other tests when n =10 and also is a second powerful test. Moreover the power of ZMAD test increased with increasing of the sample sizes.

Table 6: Power comparison of KS AD MAD ZMAD tests for Lognormal distribution

n	Gamma(3,1)				Weibull (0.5,1.433)				Beta (2,21)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0644	0.0675	0.0827*	0.0715	0.1166	0.1129	0.1417*	0.1132	0.0836	0.0814	0.1010*	0.0868
20	0.1172*	0.0917	0.1127	0.0949	0.2908*	0.2318	0.2666	0.1910	0.1921*	0.1473	0.1770	0.1353
30	0.1706*	0.1298	0.1554	0.1139	0.4494*	0.3526	0.3916	0.2767	0.3051*	0.2230	0.2543	0.1798
50	0.2766*	0.2016	0.2396	0.1770	0.6878*	0.5647	0.5967	0.4449	0.4830*	0.3665	0.4027	0.2868
100	0.5081*	0.4077	0.4429	0.3121	0.9421*	0.8787	0.8880	0.7291	0.8077*	0.6812	0.7016	0.5142
200	0.7941*	0.7031	0.7318	0.5399	0.9995*	0.9968	0.9967	0.9589	0.9819*	0.9462	0.9527	0.8193
n	Gamma(4,1)				Weibull (0.5,1.563)				Beta (2,12.5)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0587	0.0587	0.0714*	0.0628	0.1154	0.1118	0.1368*	0.1146	0.0901	0.0896	0.1127*	0.0943
20	0.1010*	0.0824	0.0960	0.0765	0.2938*	0.2267	0.2626	0.1866	0.2163*	0.1671	0.1940	0.1428
30	0.1381*	0.0979	0.1221	0.0956	0.4500*	0.3458	0.3872	0.2698	0.3451*	0.2510	0.2781	0.2045
50	0.2174*	0.1609	0.1873	0.1390	0.6850*	0.5674	0.5997	0.4360	0.5620*	0.4269	0.4578	0.3178
100	0.3848*	0.2939	0.3251	0.2367	0.9426*	0.8790	0.8893	0.7293	0.8708*	0.7454	0.7595	0.5647
200	0.6576*	0.5616	0.6021	0.4225	0.9987*	0.9940	0.9950	0.9583	0.9955*	0.9723	0.9731	0.8672
n	Gamma(5,1)				Weibull (0.5,1.663)				Beta (2,9.27)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0554	0.0563	0.0676*	0.0603	0.1137	0.1152	0.1393*	0.1119	0.0989	0.0978	0.1184*	0.0986
20	0.0878*	0.0709	0.0854	0.0736	0.2811*	0.2301	0.2614	0.1870	0.2392*	0.1834	0.2071	0.1543
30	0.1251*	0.0964	0.1102	0.0861	0.4487*	0.3489	0.3836	0.2743	0.3825*	0.2810	0.3129	0.2185
50	0.1890*	0.1382	0.1641	0.1268	0.6837*	0.5645	0.5880	0.4284	0.6171*	0.4740	0.4951	0.3513
100	0.3264*	0.2431	0.2754	0.2028	0.9387*	0.8789	0.8904	0.7299	0.9137*	0.8035	0.8119	0.6280
200	0.5468*	0.4578	0.4942	0.3468	0.9992*	0.9948	0.9960	0.9553	0.9979*	0.9856	0.9852	0.9040
n	Gamma(16,1)				Weibull (0.5,2.2)				Beta (2,4.25)			
	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS	ZMAD	MAD	AD	KS
10	0.0433	0.0459	0.0537*	0.0524	0.1147	0.1122	0.1415*	0.1190	0.1311	0.1280	0.1494*	0.1223
20	0.0619*	0.0545	0.0604	0.0520	0.2928*	0.2288	0.2672	0.1993	0.3483*	0.2653	0.2938	0.2040
30	0.0669*	0.0557	0.0649	0.0599	0.4378*	0.3410	0.3789	0.2659	0.5583*	0.4109	0.4309	0.2965
50	0.0897*	0.0689	0.0801	0.0715	0.6861*	0.5726	0.5989	0.4380	0.8389*	0.6705	0.6714	0.4794
100	0.1361*	0.0982	0.1118	0.0961	0.9393*	0.8747	0.8863	0.7205	0.9930*	0.9528	0.9484	0.8024
200	0.2163*	0.1604	0.1790	0.1335	0.9988*	0.9965	0.9965	0.9583	1.0000*	0.9996	0.9994	0.9786

4. Conclusion

In this article, the modified Anderson-darling Z_A test (ZMAD) for the skewed distributions is established from two concepts; the first uses the likelihood ratio statistics by Zhang [11] and the second uses the weight function to the tail that focuses on by Ahmad et al. [8]. The critical values of the ZMAD, MAD, AD, and KS were created only for the upper tails of the Gamma and Lognormal distributions by using Monte Carlo simulation. The power study show that the ZMAD is the most powerful for testing Lognormal distribution with $n \geq 20$. For testing Gamma distributions, the ZMAD test is better than all tests at large sample sizes. Except the case that alternative is Lognormal, the power of ZMAD is the best for small sample sizes.

Acknowledgement

I would like to acknowledge Dr. Jutaporn Neamvonk my adviser and Dr. Vanida Pongsakchat for all helpful comments, suggestions, and support. I also would like to thank faculty of science for financial support of this conference.

References

[1] Kolmogorov AN. Sulla Determinazione Empirica Di una Legge Di Distribuzione. Giornale Dell' Istituto Italiano Degli Attuari. 1933; 4: 83-91.

[2] Anderson TW, Darling DA. A Test of Goodness of fit. Journal of the American statistical association. 1954; 49: 765-769.

[3] Stephens MA. EDF Statistics for goodness of fit and some comparisons. Journal of the American statistical associatio. 1974; 69: 730-737.

[4] özmen T. A modified Anderson Darling goodness-of-fit test for the gamma distribution with unknown scale and location parameters [Dissertation]. Ohio: Air Univ; 1993.

[5] Sriamporn R. Comparison of Power of Some Goodness of Fit Tests for Testing Lognormal and Weibull Distribution [Dissertation]. Bankok: King Mongkut's institute of technology north Bangkok; 2006.

[6] Pajjayakarn Promdan 2009: Some Goodness of Fit Tests for Generalized Exponential Distribution [Dissertation]. Bankok: Kasetsart Univ; 2009.

[7] Abd-Elfattah AM, Hala AF, Omima AM. Goodness of fit tests for generalized frechet distribution. Australian journal of basic and applied sciences. 2010; 4(2): 286-301.

[8] Ahmad MI, Sinclair CD, Spurr BD. Assessment of flood frequency models using empirical distribution function statistics. Water resources research. 1988; 24(8): 1323-1328.

[9] Arshad M, Rasool MT, Ahmad MI. Anderson Darling and modified Anderson Darling tests for generalized pareto distribution. Pakistan journal of applied sciences. 2003; 3(2): 85-88.

[10] Shin HJ, Jung YG, Jeong CS, Heo JH. Assessment of modified Anderson-Darling test statistics for the generalized extreme value and generalized logistic distributions. stochastic environmental research and risk assessment. 2011; 26: 105-114.

[11] Zhang J. Powerful goodness-of-fit tests based on the likelihood ratio. Royal statistical society. 2002; 64(2): 281-294.

[12] Zhang J, Wu Y. Likelihood-ratio tests for normality. Computational statistics & data Analysis. 2005; 49: 709-721.

[13] Abidin NZ, Adam MB, Midi H. The goodness of fit test for gumbel distribution: A comparative study. Matematika. 2012; 28(1): 35-48.

[14] Cochran WG. Some methods of strengthening the common chi-square tests. Biometrics. 1954; 10: 417-451.

Confidence interval for the two-parameter exponential distribution based on the double type II censored data

Nassamon Bootwisas[1]* and Wuttichai Srisodaphol[2]
[1]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, nassamon_ae@hotmail.com*
[2]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, wuttsr@kku.ac.th*

Abstract

This study aims to propose the confidence intervals of the location parameter and the scale parameter for the two-parameter exponential distribution based on double Type II censored data which is the confidence intervals that use Jackknife method to adjust the confidence intervals of Fauzy (2004). Coverage probability and interval width are criteria to compare the efficiency of the confidence intervals. The results of simulation study show that proposed confidence interval is better than Faucy's confidence interval for the scale parameter. For the location parameter, these two confidence intervals perform poorly in term of the coverage probability which is lower to nominal level. In the results of real data sets show that the proposed confidence intervals have shorter interval width than Faucy's confidence intervals for the location parameter and the scale parameter.

Keywords: Confidence interval, two-parameter exponential distribution, double type II censored data

* Corresponding Author
E-mail Address: nassamon_ae@hotmail.com

1. Introduction

The exponential distribution is the probability distribution that describes the period of time waiting and events. The probability density function of two-parameter exponential distribution is

$$f\left(x;\mu,\theta\right)=\frac{1}{\theta}\exp\left(-\frac{x-\mu}{\theta}\right)\quad ;x\geq\mu,\mu>0,\theta>0$$

where μ and θ are the location parameter and the scale parameter, respectively. The exponential distribution is used immensely in the analysis of lifetime. Historically, the exponential distribution was the first lifetime model of statistical methods in survival analysis that has been developed extensively. What distinguishes survival analysis from other fields of statistics is censoring. The three types of censoring are type I censoring, type II censoring and random censoring [4]. In reliability studies, due to time limitation or other restrictions on data collection, several lifetime of units put on test may not be observed. In addition, sometimes few lowest and highest observations in a sample could be due to some negligence or some other extraordinary reasons. In such cases, it is convenient to remove those outlying observations. Suppose some initial observations are censored in addition to some final observations being censored. Out of the n components put to test, suppose the experimenter fails to observe the first r and s the last, observations x are then said to be double Type II censoring [3]: $x_{r+1;n}\leq x_{r+2;n}\leq\ldots\leq x_{n-s;n}$.

In 1990 Balakrishnan [2] used the maximum likelihood estimation to find the point estimators of the location parameter and the scale parameter for the two-parameter exponential distribution based on multiply Type II censored sample. In 2004 Fauzy [3] used point estimators by the maximum likelihood estimation and bootstrap percentile method to construct the confidence intervals for the two-parameter exponential distribution based on double Type II censoring. In 2007 and 2010 Wu [6,7] used the pivotal quantity to construct the confidence interval of the scale parameter and the joint confidence region of the two parameters for the two-parameter exponential distribution based on double Type II censored sample and a progressively Type II censored sample, respectively. In 2011 Asgharzadeh [1] used the pivotal quantity to construct the confidence intervals of the location parameter and the scale parameter, and the joint confidence region of the two parameters for the two-parameter exponential distribution based on records.

The maximum likelihood estimation is frequently used for parameter estimation but it can be highly biased for small sample. Jackknife method which is resampling method that assumes a sample of repeated samples from the same set, to become a random repeat sample of the actual population can reduce this bias. Therefore, in this study, we use Jackknife method for estimation of parameter.

In this paper we propose the confidence intervals of the location parameter and the scale parameter for the two-parameter exponential distribution based on double Type II censored data which is adjusted the confidence interval that use the point estimator of Jackknife method in the confidence interval of Fauzy. Coverage probability and interval width are used to compare efficiency between the proposed confidence intervals and the confidence intervals of Fauzy.

2. Research Methodology

2.1 Proposed confidence intervals

We find the point estimators of the location parameter and the scale parameter by using Jackknife method, and then replace these point estimators instead of the point estimators in Fauzy's confidence intervals. The following steps are our proposed confidence intervals.

Step 1: we choose at random for $n-s-r$ sample from double Type II censored data $x_{r+1;n}$, $x_{r+2;n}$,…, $x_{n-s;n}$ which is distributed as exponential distribution based on double Type II censored data. Let $\hat{\theta}_i$ and $\hat{\mu}_i$ be the point estimators of θ and μ for $i=1,2,...,n-s-r$, respectively. We obtain $\hat{\theta}_i$ and $\hat{\mu}_i$ as follows.

First, we remove $x_{r+1;n}$ out from $x_{r+1;n}$, $x_{r+2;n}$,…, $x_{n-s;n}$ and then use the maximum likelihood estimation based on double Type II censored data to find the point estimators from $x_{r+2;n}, x_{r+3;n}, ..., x_{n-s;n}$. Let $\hat{\theta}_1$ and $\hat{\mu}_1$ be the point estimators. The maximum likelihood estimators based on double Type II censored data of Fauzy [3] are

$$\hat{\mu}_1 = x_{r+2;n} + \hat{\theta}_1 \ln\left(\frac{n-(r+1)}{n}\right)$$

with

$$\hat{\theta}_1 = \frac{\sum_{i=r+2}^{n-s} x_{i;n} + s x_{n-s;n} - (n-(r+1)) x_{r+2;n}}{(n-s-(r+1))}$$

where n is the total data

r is the number of the first observation being censored

s is the number of the last observation being censored

$x_{i;n}$ is value of data order i ,

$i = r+1,...,r+k,r+k+t+1,...,n-s$.

Second, we remove $x_{r+2;n}$ out from $x_{r+1;n}$, $x_{r+2;n}$,…, $x_{n-s;n}$ and then use the maximum likelihood estimation based on double Type II censored data to find the point estimators from $x_{r+1;n}, x_{r+3;n}, ..., x_{n-s;n}$. Let $\hat{\theta}_2$ and $\hat{\mu}_2$ be the point estimators. The maximum likelihood estimators based on multiply Type II censored data of Balakrishnan [2] are

$$\hat{\mu}_2 = x_{r+1;n} + \hat{\theta}_2 \ln\left(\frac{n-r}{n}\right)$$

with

$$\hat{\theta}_2 = \frac{1}{(A-t\alpha)}\left[\sum_{i=r+1}^{r+k} x_{i;n} + \sum_{i=r+k+t+1}^{n-s} x_{i;n} + s x_{n-s;n} + t\beta x_{r+k;n} + t(1-\beta) x_{r+k+t+1;n} - (n-r) x_{r+1;n}\right]$$

where $A = n-s-r-t$

$$p_r = \frac{r}{n}$$

$$q_r = 1 - p_r$$

$$\beta = \frac{q_{r+k}}{q_{r+k} - q_{r+k+t+1}} - \frac{q_{r+k} q_{r+k+t+1}}{\left(q_{r+k} - q_{r+k+t+1}\right)^2} \times \ln\left(\frac{q_{r+k}}{q_{r+k+t+1}}\right)$$

$$\alpha^* = \frac{\left(q_{r+k} \ln q_{r+k+t+1} - q_{r+k} \ln q_{r+k}\right)}{\left(q_{r+k} - q_{r+k+t+1}\right)}$$

$$\alpha = \alpha^* + \beta \ln q_{r+k} + (1-\beta) \ln q_{r+k+t+1}$$

t is the number of the middle observation being censored

k is the number of the observation between r and t .

.

.

.

Then, we remove $x_{n-s-1;n}$ out from $x_{r+1;n}$, $x_{r+2;n}$,…, $x_{n-s;n}$ and then use the maximum likelihood estimation based on double Type II censored data to find the point estimators from $x_{r+1;n}, x_{r+2;n}, ..., x_{n-s-2;n}, x_{n-s;n}$. Let $\hat{\theta}_{n-s-r-1}$ and $\hat{\mu}_{n-s-r-1}$ be the point estimators. The maximum likelihood estimators based on multiply Type II censored data of Balakrishnan [2] are

$$\hat{\mu}_{n-s-r-1} = x_{r+1;n} + \hat{\theta}_{n-s-r-1} \ln\left(\frac{n-r}{n}\right)$$

with

$$\hat{\theta}_{n-s-r-1} = \frac{1}{(A-t\alpha)}\left[\sum_{i=r+1}^{r+k} x_{i;n} + \sum_{i=r+k+t+1}^{n-s} x_{i;n} + s x_{n-s;n} + t\beta x_{r+k;n} + t(1-\beta) x_{r+k+t+1;n} - (n-r) x_{r+1;n}\right].$$

Finally, we remove $x_{n-s;n}$ out from $x_{r+1;n}$, $x_{r+2;n}$,…, $x_{n-s;n}$ and then use the maximum likelihood estimation based on double Type II censored data to find the point estimators from $x_{r+1;n}, x_{r+2;n}, ..., x_{n-s-1;n}$. Let $\hat{\theta}_{n-s-r}$ and $\hat{\mu}_{n-s-r}$ be the point estimators. The maximum likelihood estimators based on double Type II censored data of Fauzy [3] are

$$\hat{\mu}_{n-s-r} = x_{r+1;n} + \hat{\theta}_{n-s-r} \ln\left(\frac{n-r}{n}\right)$$

with

$$\hat{\theta}_{n-s-r} = \frac{\sum_{i=r+1}^{n-s-1} x_{i;n} + (s+1) x_{n-s-1;n} - (n-r) x_{r+1;n}}{(n-(s+1)-r)}.$$

Step 2: we find the mean of the point estimators from Step 1,

$$\hat{\mu}_J = \frac{1}{n-s-r}\sum_{i=1}^{n-s-r}\hat{\mu}_i$$

and

$$\hat{\theta}_J = \frac{1}{n-s-r}\sum_{i=1}^{n-s-r}\hat{\theta}_i \ .$$

We replace the point estimators $\hat{\mu}_J$ and $\hat{\theta}_J$ instead of point estimators in Fauzy's confidence intervals. Our proposed confidence intervals are

$$\mu_L < \mu < \mu_U$$

where $\mu_L = \hat{\mu}_J - \dfrac{(n-s-r)\hat{\theta}_J F_{1-\frac{\alpha}{2},(2,2(n-s-r)-2)}}{n(n-s-r-1)}$

$$\mu_U = \hat{\mu}_J - \frac{(n-s-r)\hat{\theta}_J F_{\frac{\alpha}{2},(2,2(n-s-r)-2)}}{n(n-s-r-1)}$$

and

$$\frac{2(n-s-r)\hat{\theta}_J}{\chi^2_{1-\frac{\alpha}{2},(2(n-s-r)-2)}} < \theta < \frac{2(n-s-r)\hat{\theta}_J}{\chi^2_{\frac{\alpha}{2},(2(n-s-r)-2)}}$$

where

$F_{\frac{\alpha}{2},(2,2(n-s-r)-2)}$ is the right-tailed $(\alpha/2)$ quantile for F distribution with 2 and $2(n-s-r)-2$ degrees of freedom.

$F_{1-\frac{\alpha}{2},(2,2(n-s-r)-2)}$ is the left-tailed $(1-\alpha/2)$ quantile for F distribution with 2 and $2(n-s-r)-2$ degrees of freedom.

$\chi^2_{\frac{\alpha}{2},(2(n-s-r)-2)}$ is the right-tailed $(\alpha/2)$ quantile for chi-square distribution with $2(n-s-r)-2$ degrees of freedom.

$\chi^2_{1-\frac{\alpha}{2},(2(n-s-r)-2)}$ is the left-tailed $(1-\alpha/2)$ quantile for chi-square distribution with $2(n-s-r)-2$ degrees of freedom.

2.2 Comparison of confidence intervals

To compare the confidence intervals for two-parameter exponential distribution, we generate random sample from the exponential distribution with $\mu = 1.5$, $\theta = 20, 40, 60$ and $\mu = 18$, $\theta = 45, 65, 85$ for sample size $n = 10, 20, 30$ with 1,000 times for the simulation data. We specify the percentiles of censoring are 10% and 20% for the first observation is censored (r) and the last observation is censored (s), respectively, as shown in Table 1.

Table 1: Number of censored data

Censored data		$n=10$		$n = 20$		$n = 30$	
r	s	r	s	r	s	r	s
10%	20%	1	2	2	4	3	6

There are two real data sets from Lawless [5] for comparing the confidence intervals as follows.

Data set 1: The following data represent failure times (minutes) for a specific type of electrical insulation in an experiment in which the insulation was subjected to continuously increasing voltage stress.

18.5	21.7	35.1	40.5	42.3	48.7
79.4	86.0	121.9	147.1	150.2	219.3

Data set 2: The following data represent failure times (minutes) for electronic components in an accelerated life test.

1.4	5.1	5.3	10.8	12.1	18.5	19.7
22.2	23.0	30.6	37.3	46.3	53.9	59.8
66.2	73.5	80.3	87.2	95.7	102.4	

3. Result

3.1 The simulation data

The simulated result at 95% and 99% confidence interval are tabulated in Table 2 and Table 3, respectively. The two criteria that use to compare confidence intervals are coverage probability and interval width. The best performance occurs when coverage probability is close to nominal level and provides shorter interval width.

Table 2 and Table 3 show the coverage probability and interval width at 95% and 99% confidence interval of confidence intervals the location parameter and the scale parameter. For scale parameter, proposed confidence interval and Fauzy's confidence interval provide the same coverage probability and are close to the nominal level but proposed confidence interval gives the shorter interval width than Fauzy's confidence interval. For location parameter, proposed confidence interval and Fauzy's confidence interval perform poorly in term of the coverage probability which is lower to nominal level.

3.2 The real data

The interval width of the confidence intervals for two real data sets at 95% and 99% confidence intervals are tabulated in Table 4 and Table 5, respectively.

Table 4 and Table 5 show that the interval widths of our proposed confidence intervals are shorter than Fauzy's confidence intervals.

4. Conclusion

In this paper, we proposed the confidence intervals of the location parameter and the scale parameter for the two-parameter exponential distribution based on double Type II censored data.

The simulation results show that the proposed confidence intervals had the better performance in all situations for the scale parameter but performed poorly in term of the coverage probability for the location parameter.

Moreover, our proposed confidence intervals of the location parameter and the scale parameter had shorter interval width than Fauzy's confidence intervals for the two real data sets.

Table 2: Coverage probability and interval width for μ and θ at the level of 95%

n	(μ,θ)	CP				IW			
		μ		θ		μ		θ	
		F	FJ	F	FJ	F	FJ	F	FJ
10	$(1.5, 20)$	0.715	0.727	**0.956**	**0.957**	10.151	10.078	44.261	**43.942**
	$(1.5, 40)$	0.689	0.707	**0.954**	**0.956**	20.258	20.091	88.327	**87.598**
	$(1.5, 60)$	0.720	0.725	**0.951**	**0.950**	30.705	30.435	133.875	**132.699**
	$(18, 45)$	0.714	0.728	**0.951**	**0.954**	23.211	23.037	101.244	**100.443**
	$(18, 65)$	0.729	0.732	**0.953**	**0.954**	32.632	32.395	142.277	**141.245**
	$(18, 85)$	0.722	0.730	**0.955**	**0.955**	43.293	42.990	188.759	**187.437**
20	$(1.5, 20)$	0.638	0.627	**0.962**	**0.963**	8.395	8.388	49.812	**49.767**
	$(1.5, 40)$	0.613	0.607	**0.954**	**0.953**	8.489	8.483	50.369	**50.334**
	$(1.5, 60)$	0.609	0.614	**0.959**	**0.958**	12.541	12.529	74.406	**74.335**
	$(18, 45)$	0.588	0.597	**0.961**	**0.959**	9.491	9.481	56.312	**56.255**
	$(18, 65)$	0.596	0.595	**0.950**	**0.951**	13.686	13.679	81.202	**81.161**
	$(18, 85)$	0.584	0.588	**0.956**	**0.958**	18.101	18.086	107.398	**107.310**
30	$(1.5, 20)$	0.524	0.522	**0.959**	**0.958**	2.672	2.671	19.174	**19.168**
	$(1.5, 40)$	0.521	0.524	**0.954**	**0.954**	5.434	5.433	39.001	**38.990**
	$(1.5, 60)$	0.552	0.551	**0.960**	**0.961**	7.970	7.968	57.198	**57.187**
	$(18, 45)$	0.553	0.550	**0.958**	**0.958**	6.077	6.075	43.612	**43.596**
	$(18, 65)$	0.515	0.512	**0.955**	**0.955**	8.708	8.707	62.493	**62.487**
	$(18, 85)$	0.522	0.525	**0.951**	**0.950**	11.477	11.473	82.370	**82.343**

F is the confidence interval of Fauzy, FJ is adjusted confidence interval using the point estimation of Jackknife method in the confidence interval of Fauzy.

Table 3: Coverage probability (CP) and interval width (IW) for μ and θ at the level of 99%

n	(μ,θ)	CP				IW			
		μ		θ		μ		θ	
		F	FJ	F	FJ	F	FJ	F	FJ
10	$(1.5, 20)$	0.777	0.796	**0.990**	**0.991**	17.027	16.904	69.672	**69.170**
	$(1.5, 40)$	0.753	0.769	**0.994**	**0.993**	33.979	33.699	139.038	**137.891**
	$(1.5, 60)$	0.765	0.776	**0.991**	**0.991**	51.502	51.049	210.737	**208.884**
	$(18, 45)$	0.773	0.790	**0.992**	**0.991**	38.949	38.640	159.371	**158.109**
	$(18, 65)$	0.780	0.789	**0.991**	**0.992**	54.734	54.337	223.962	**222.337**
	$(18, 85)$	0.780	0.791	**0.991**	**0.992**	72.616	72.107	297.130	**295.049**

F is the confidence interval of Fauzy, FJ is adjusted confidence interval using the point estimation of Jackknife method in the confidence interval of Fauzy.

Table 3: Coverage probability (CP) and interval width (IW) for μ and θ at the level of 99% (cont.)

n	(μ,θ)	CP				IW			
		μ		θ		μ		θ	
		F	FJ	F	FJ	F	FJ	F	FJ
20	$(1.5, 20)$	0.710	0.701	**0.993**	**0.993**	12.941	12.929	70.934	**70.869**
	$(1.5, 40)$	0.686	0.690	**0.990**	**0.990**	13.086	13.077	71.727	**71.678**
	$(1.5, 60)$	0.691	0.703	**0.992**	**0.993**	19.330	19.312	105.956	**105.856**
	$(18, 45)$	0.681	0.686	**0.990**	**0.990**	14.630	14.615	80.191	**80.110**
	$(18, 65)$	0.669	0.679	**0.995**	**0.995**	21.096	21.085	115.634	**115.575**
	$(18, 85)$	0.647	0.653	**0.993**	**0.993**	27.901	27.879	152.938	**152.813**
30	$(1.5, 20)$	0.602	0.604	**0.993**	**0.994**	4.023	4.021	26.532	**26.524**
	$(1.5, 40)$	0.629	0.631	**0.992**	**0.992**	8.182	8.180	53.967	**53.953**
	$(1.5, 60)$	0.638	0.637	**0.990**	**0.991**	12.000	11.998	79.148	**79.133**
	$(18, 45)$	0.634	0.629	**0.993**	**0.993**	9.150	9.146	60.349	**60.327**
	$(18, 65)$	0.609	0.605	**0.991**	**0.992**	13.111	13.109	86.476	**86.467**
	$(18, 85)$	0.612	0.617	**0.989**	**0.990**	17.281	17.275	113.980	**113.943**

F is the confidence interval of Fauzy, FJ is confidence interval was adjusted using the point estimation of Jackknife method in the confidence interval of Fauzy.

Table 4: Lower bound (LB), upper bound (UB) and interval widths (IW) for μ and θ at the level of 95% and 99% for data set 1

CI	95% Confidence interval					
	μ			θ		
	LB	UB	IW	LB	UB	IW
F	-17.966	14.964	32.930	47.023	196.361	149.338
FJ	-16.480	16.036	**32.516**	46.432	193.892	**147.460**
	99% Confidence interval					
F	-37.938	15.108	53.046	39.583	263.778	224.195
FJ	-36.201	16.179	**52.380**	39.085	260.461	**221.376**

F is the confidence interval of Fauzy, FJ is confidence interval was adjusted using the point estimation of Jackknife method in the confidence interval of Fauzy.

Table 5: Lower bound (LB), upper bound (UB) and interval widths (IW) for μ and θ at the level of 95% and 99% for data set 2

CI	95% Confidence interval					
	μ			θ		
	LB	UB	IW	LB	UB	IW
F	-10.921	0.133	11.054	32.335	97.920	65.585
FJ	-10.648	0.363	**11.011**	32.209	97.537	**65.328**
	99% Confidence interval					
F	-16.853	0.186	17.039	28.072	121.467	93.395
FJ	-16.556	0.415	**16.971**	27.962	120.992	**93.030**

F is the confidence interval of Fauzy, FJ is confidence interval was adjusted using the point estimation of Jackknife method in the confidence interval of Fauzy.

Acknowledgements

This paper was partially supported by Department of Statistics, Faculty of Science, Khon Kaen University.

References

[1] Asgharzadeh, A. Confidence Intervals and Joint Confidence Regions for the Two-Parameter Exponential Distribution based on Records. Communication of the Korean Statistical Society. 2011; 18: 103–110.

[2] Balakrishnan, N. On the maximum likelihood estimation of the location and scale parameters of exponential distribution based on Type II censored samples. Journal of Applied Statistics. 1990; 17: 55–61.

[3] Fauzy, A. Interval Estimation for Parameters Exponential Distribution Under Double Type II Censoring. Journal of Applied Mathematrics Islamic Azad University Lahijan. 2004; 71–79.

[4] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated data. New York: Springer; 2003.

[5] Lawless, J.F. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons; 1982.

[6] Wu, S.F. Interval Estimation for the Two-Parameters Exponential Distribution Based on Doubly Type II Censored sample. Quality & Quantity. 2007; 41: 489–496.

[7] Wu, S.F. Interval Estimation for the Two-Parameters Exponential Distribution under progressive censoring. Quality & Quantity. 2010; 44: 181–189

The beta length-biased Pareto distribution with an application for Norwegian fire claims

Winai Bodhisuwan* and Nareerat Nanuwong
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand, 10900,
e-mail: fsciwnb@ku.ac.th

Abstract

In this article, a new four-parameter beta length-biased Pareto distribution is developed and studied. Various properties of this distribution are discussed. The new distribution has either a unimodal or a decreasing hazard rate. The expressions for the mean, mean deviation, variance, skewness, kurtosis and Rényi and Shannon entropies are obtained. The relationship between these moments and the parameters are presented. The method of maximum likelihood is proposed to estimate the parameters of the distribution. Finally, the new distribution is applied to Norwegian fire claims data.

Keywords: Beta length-biased Pareto distribution, hazard rate, Rényi and Shannon entropies, maximum likelihood estimation

*Corresponding Author
E-mail Address: fsciwnb@ku.ac.th

1. Introduction

The family of the Pareto distribution is a versatile probability model which was first presented in 1909 by Vilfredo Pareto, the Italian-born Swiss economist, sociologist and philosopher. The Pareto distribution usually describes the distribution of personal income and wealth. The various forms of the Pareto distribution are very versatile, and a variety of uncertainties can be usefully modelled by them. They arise as tractable ‘life time’ models in actuarial sciences, economics, finance, life testing and climatology, where it usually describes the occurrence of extreme weather. Censoring is common in many lifetime studies due to time and money limitations and other data collection restrictions. In failure-censored tests, a known number of observations in an ordered sample may be missing at either end or at both ends [1-6].

Many transformations and generalization of the Pareto distribution have been proposed in order to get more flexible models [7-10]. Different methods may be used to introduce a shape parameter to the Pareto model. Patil and Rao [11] presented the length-biased Pareto (LBP) distribution by concept of a weighted distribution with probability density function (pdf):

$$g(x) = \frac{(\gamma-1)}{\theta}\left(\frac{x}{\theta}\right)^{-\gamma}, \qquad x \geq \theta;\ \gamma > 1 \qquad (1)$$

and the cumulative distribution function (cdf) of LBP distribution is given by:

$$G(x) = 1 - \left(\frac{x}{\theta}\right)^{-(\gamma-1)}. \qquad (2)$$

More recently, Probability weighted moment inequalities and variability orderings for weighted and unweighted reliability measures and related functions were presented by Oluyede [12]. Also, stochastic comparisons and moment inequalities were given. More recently, Das and Roy [13] developed the length-biased form of the weighted Weibull distribution and discussed various properties of it. The result of this distribution suggested a good fitted to consecutive years data. Further, Nanuwong and Bodhisuwan [14] introduced the length-biased beta-Pareto distribution, it has several sub-models include in the length-biased Pareto, arcsine, log-beta and exponential distributions.

This motivates us to propose in this article another generalization of the Pareto distribution, referred to the beta length-biased Pareto (BLBP) distribution. In Section 2, we define the BLBP distribution and outline some special sub-models of this distribution. We investigate some properties of the distribution in Section 3. Section 4 is devoted to the discussion on the moments of the BLBP distribution. The mean deviation from the mean and the median are provided in Section 5. Rényi and Shannon entropies are discussed in Section 6. Section 7 consists of the maximum likelihood estimates of the parameters and provide application of the BLBP distribution to Norwegian fire claims data in Section 8. We conclude in Section 9 with some remarks on the main results and their significance.

2. The beta length-biased Pareto distribution

One such class of distributions generated from the logit of a beta random variable (r.v.) which extends the original beta family of distributions with the incorporation of two additional parameters that control the skewness and the tail weight. This class of generalized distributions has been receiving considerable attention over the last years, in particular after the works of Eugene *et al.* [15]. Various authors of this beta generalized distributions have been developed in recent years [16-24].

Let F(x) denote the cdf of a r.v. X. The cdf for a generalized class of distribution for the r.v. X, as defined by Eugene *et al.* [15], is generated by applying the inverse cdf to a beta distributed r.v. to obtain

$$F(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_{0}^{G(x)} t^{\alpha-1}(1-t)^{\beta-1}\,dt, \qquad \alpha,\beta>0.$$

The corresponding pdf for G(x) in Equation (2) as given below

$$f(x)=\frac{1}{B(\alpha,\beta)}\left[G(x)\right]^{\alpha-1}\left[1-G(x)\right]^{\beta-1}G'(x), \qquad (3)$$

where $B(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$. From Equation (3), the pdf for the BLBP r.v. as follows

$$f(x)=\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)\beta-1},$$

$$x \geq \theta,\ \alpha,\ \beta,\ \theta > 0,\ \gamma > 1. \qquad (4)$$

By setting $(x/\theta)^{-(\gamma-1)} = y$ in Equation (4), it is easy to show that $\int_{\theta}^{\infty} f(x)dx=1$. The cdf of the BLBP r.v. denoted as F(x), can rewrite as: let $F^{*}(x) = 1 - F(x)$, then, by setting $y = (t/\theta)^{-(\gamma-1)}$, $F^{*}(x)$ for the BLBP distribution with density function given in Equation (4) is

$$F^{*}(x)=\frac{B(z;\beta,\alpha)}{B(\alpha,\beta)}$$

where,

$$B(z;\beta,\alpha)=z^{\beta}\left\{\frac{1}{\beta}+\frac{1-\alpha}{\beta+1}z+\cdots+\frac{(1-\alpha)(2-\alpha)\cdots(n-\alpha)}{n!(\beta+n)}z^{n}+\cdots\right\}$$

is an incomplete beta function with $z=(x/\theta)^{-(\gamma-1)}$. Hence,

$$F(x)=1-\frac{B(z;\beta,\alpha)}{B(\alpha,\beta)}. \qquad (5)$$

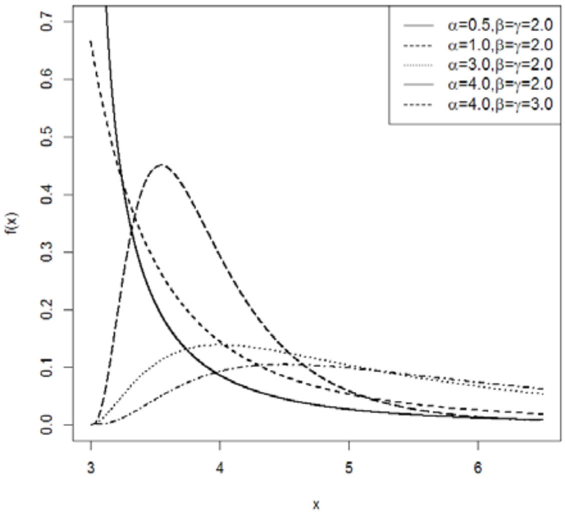


Figure 1: The pdf of a BLBP r.v. X for θ = 3 of some values of α and β = γ.

2.1 Some well-known sub-models of BLBP distribution

We may consider some special case of BLBP distribution as the corollaries.

Corollary 1: Let suppose X~BLBP(α=β=1,θ,γ), the BLBP distribution in Equation (4) reduces to the LBP distribution in Equation (1) with parameters θ and γ.

Corollary 2: Let X~BLBP(α=1,β,θ,γ), the BLBP distribution with parameters α, β, θ and γ reduces to the Pareto distribution with parameters (γ-1)β = k and θ, with density function

$$f(x)=\frac{k}{\theta}\left(\frac{x}{\theta}\right)^{-(k+1)}, \quad x \geq \theta,\ k > 0,\ \theta > 0.$$

Corollary 3: If X~BLBP(α=β=1/2,θ,γ), then the r.v. $Y=(X/\theta)^{-(\gamma-1)}$ has the arcsine distribution as follows:

$$f(y)=\frac{1}{\pi\sqrt{y(1-y)}}, \qquad 0 < y < 1.$$

Corollary 4: If X~BLBP(α,β,θ,γ), then the r.v. Y=βln(X/θ) has the log-beta distribution, with parameters a = α, b = β and c = γ-1.

Proof Using the transformation method, we can show that the r.v. Y has the density function given by

$$f(y)=\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(e^{\frac{y}{\beta}}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(e^{\frac{y}{\beta}}\right)^{-(\gamma-1)\beta-1}\left|\frac{\theta}{\beta}e^{\frac{y}{\beta}}\right|$$

$$=\frac{(\gamma-1)}{\beta B(\alpha,\beta)}\left(1-e^{\frac{-(\gamma-1)y}{\beta}}\right)^{\alpha-1}e^{-(\gamma-1)y}$$

$$=\frac{c}{bB(\alpha,\beta)}\left(1-e^{\frac{-cy}{b}}\right)^{a-1}e^{-cy},\ \ 0< \text{ y } <\infty.$$

Corollary 5: If X~BLBP(α=β=1,θ,γ), then the r.v. Y=βln(X/θ) follows the exponential distribution with mean 1/(γ-1) with pdf:

$$f(y)=(\gamma-1)e^{-(\gamma-1)y}\text{ , } 0 < \text{y} < \infty.$$

3. Some properties of the BLBP distribution

We investigate some properties of the BLBP distribution in this section.

3.1 Transformation

If Y is a beta r.v. with parameters α and β, then the r.v. $X=\theta(1-Y)^{-1/(\gamma-1)}$ has a BLBP distribution with parameters α, β, θ and γ-1.

Proof: Using the transformation method, we can show that the r.v. X has pdf in Equation (4) as follow:

$$f(x)=\frac{1}{B(\alpha,\beta)}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}$$

$$\left\{1-\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]\right\}^{\beta-1}\left|\frac{(\gamma-1)}{\theta}\left(\frac{x}{\theta}\right)^{-\gamma}\right|$$

$$=\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)\beta-1}.$$

3.2 Hazard rate

By definition, the hazard rate of a r.v. X with density f(x) and cdf F(x) is defined as

$$h(x)=\frac{f(x)}{1-F(x)}$$

Using Equation (4) and Equation (5), the hazard rate of the BLBP distribution as given below:

$$h(x)=\frac{\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)\beta-1}}{1-\left[1-\frac{B(z;\beta,\alpha)}{B(\alpha,\beta)}\right]}$$

$$=\frac{(\gamma-1)\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)\beta-1}}{\theta z^{\beta}\left\{\frac{1}{\beta}+\frac{1-\alpha}{\beta+1}z+\cdots+\frac{(1-\alpha)(2-\alpha)\cdots(n-\alpha)}{n!(\beta+n)}z^{n}+\cdots\right\}}$$

$$=\frac{(\gamma-1)/x\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}}{\left\{\frac{1}{\beta}+\frac{1-\alpha}{\beta+1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)}+\cdots+\frac{(1-\alpha)(2-\alpha)\cdots(n-\alpha)}{n!(\beta+n)}\left(\frac{x}{\theta}\right)^{-(\gamma-1)n}+\cdots\right\}}.$$

We display some graphs of hazard rate for the BLBP distribution in Figure 2.

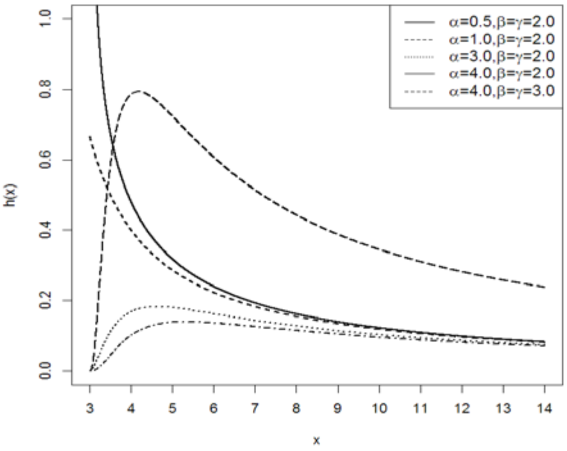


Figure 2: The hazard rate of a BLBP distribution r.v. X for θ = 3 of some values of α and β = γ.

4. The *r* th Moments of the BLBP distribution

It is easier to find the moment of the quantity $(X/\theta)^r$ first, and with appropriate manipulation, obtain the moment of X^r. By definition,

$$E\left(\frac{X}{\theta}\right)^{r}=\int_{\theta}^{\infty}\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x}{\theta}\right)^{-(\gamma-1)\beta+r-1}dx$$

Let, $y=(x/\theta)^{-(\gamma-1)}$, then

$$E\left(X^{r}\right)=\frac{\theta^{r}B\left(\alpha,\beta-\frac{r}{\gamma-1}\right)}{B(\alpha,\beta)}\tag{6}$$

When r = 1 in Equation (6), the mean of BLBP distribution may be expressed as

$$E(X)=\frac{\theta\Gamma(\alpha+\beta)\Gamma\left(\beta-\frac{1}{\gamma-1}\right)}{\Gamma(\beta)\Gamma\left(\alpha+\beta-\frac{1}{\gamma-1}\right)}.$$

By using appropriate moment expressions, the variance, skewness and kurtosis of the BLBP distribution may be written, respectively, as

$$\mathrm{Var}(\mathrm{X})=\theta^2\left\{\frac{\Gamma(\alpha+\beta)\Gamma\left(\beta-\frac{2}{\gamma-1}\right)}{\Gamma(\beta)\Gamma\left(\alpha+\beta-\frac{2}{\gamma-1}\right)}-\left[\frac{\Gamma(\alpha+\beta)\Gamma\left(\beta-\frac{1}{\gamma-1}\right)}{\Gamma(\beta)\Gamma\left(\alpha+\beta-\frac{1}{\gamma-1}\right)}\right]^2\right\}$$

Skewness(X) = $[\omega(\alpha,\beta,\gamma,3)-3\omega(\alpha,\beta,\gamma,2)\omega(\alpha,\beta,\gamma,1)$
$+2\omega^3(\alpha,\beta,\gamma,1)]/T^3$

and,

Kurtosis(X) = $[\omega(\alpha,\beta,\gamma,4)-4\omega(\alpha,\beta,\gamma,3)\omega(\alpha,\beta,\gamma,1)-$
$3\omega^4(\alpha,\beta,\gamma,1)+6\omega(\alpha,\beta,\gamma,2)\omega^2(\alpha,\beta,\gamma,1)]/T^4$

where, $\omega(\alpha,\beta,\gamma,i)=\dfrac{\Gamma(\alpha+\beta)\Gamma\left(\beta-\frac{i}{\gamma-1}\right)}{\Gamma(\beta)\Gamma\left(\alpha+\beta-\frac{i}{\gamma-1}\right)}$, $\beta > i/(\gamma-1)$, $i \in I^+$ and $T=\sqrt{\omega(\alpha,\beta,\gamma,2)-\omega^2(\alpha,\beta,\gamma,1)}$. Notice that both the skewness and the kurtosis of the BLBP distribution are independent from parameter θ.

5. The mean deviation

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from $\delta_\mu(X)=\int_\theta^\infty|x-\mu|\,f(x)\,dx$ and, $\delta_M(X)=\int_\theta^\infty|x-M|\,f(x)\,dx$ respectively, where $\mu=E(X)$ and M denotes the median [15, 25-27]. These measures can be calculated using the relationships

$$\delta_\mu(X)=\int_\theta^\mu|\mu-x|\,f(x)\,dx+\int_\mu^\infty|x-\mu|\,f(x)\,dx$$
$$=2\mu\int_\theta^\mu f(x)\,dx-2\int_\theta^\mu xf(x)\,dx$$

$$\delta_\mu(X)=2\mu F(\mu)-2\int_\theta^\mu xf(x)\,dx \qquad (7)$$

and,

$$\delta_M(X)=\int_\theta^M|M-x|\,f(x)\,dx+\int_M^\infty|x-M|\,f(x)\,dx$$
$$=2\int_\theta^M(M-x)\,f(x)\,dx+\int_\theta^\infty(x-M)\,f(x)\,dx$$
$$=2MF(M)-2\int_\theta^M xf(x)\,dx+E(X-M)$$

$$\delta_M(X)=\mu+2MF(M)-M-2\int_\theta^M xf(x)\,dx \qquad (8)$$

Using the series representation

$$(1+s)^\alpha=\sum_{j=0}^\infty\binom{\alpha}{j}s^j=\sum_{j=0}^\infty\frac{\Gamma(\alpha+1)}{\Gamma(\alpha-j+1)}\frac{s^j}{j!}$$

one can evaluate the integral term in Equation (7) and Equation (8) as

$$\int_\theta^c xf(x)\,dx=\frac{(\gamma-1)\theta}{B(\alpha,\beta)}\sum_{j=0}^\infty\frac{(-1)^j\,\Gamma(\alpha)}{\Gamma(\alpha-j)\,j!}\frac{1-\left(\frac{c}{\theta}\right)^{1-(\gamma-1)(\beta+j)}}{(\gamma-1)(\beta+i)-1} \qquad (9)$$

Substituting Equation (9) into Equation (7) and Equation (8), one obtains the following expressions for the mean deviations:

$$\delta_\mu(X)=2\mu F(\mu)-\frac{2(\gamma-1)\theta}{B(\alpha,\beta)}\sum_{j=0}^\infty\frac{(-1)^j\,\Gamma(\alpha)}{\Gamma(\alpha-j)\,j!}\frac{1-\left(\frac{\mu}{\theta}\right)^{1-(\gamma-1)(\beta+j)}}{(\gamma-1)(\beta+i)-1}$$

and,

$$\delta_M(X)=\mu+2MF(M)-M$$
$$-\frac{2(\gamma-1)\theta}{B(\alpha,\beta)}\sum_{j=0}^\infty\frac{(-1)^j\,\Gamma(\alpha)}{\Gamma(\alpha-j)\,j!}\frac{1-\left(\frac{M}{\theta}\right)^{1-(\gamma-1)(\beta+j)}}{(\gamma-1)(\beta+i)-1}.$$

6. The Rényi and Shannon entropies

Statistical entropy is a probabilistic measure of ignorance about the outcome of a random experiment and is a measure of a reduction in that uncertainty. Entropy of r.v. X with pdf f(x) is a measure of variation of the uncertainty [28-30]. For a BLBP r.v. X with density f(x), the Rényi entropy is defined by,

$$I_R(\xi)=\frac{1}{1-\xi}\log\left[\int f^\xi dx\right] \qquad (10)$$

where ξ > 0 and ξ ≠ 1. By using the BLBP density, we have

$$\int_{\theta}^{\infty} f^{\xi}(x)\,dx = \frac{(\gamma-1)^{\xi}}{\left[\theta B(\alpha,\beta)\right]^{\xi}} \int_{\theta}^{\infty}\left[1-\left(\frac{x}{\theta}\right)^{-(\gamma-1)}\right]^{\xi(\alpha-1)}\left(\frac{x}{\theta}\right)^{-\xi[(\gamma-1)\beta-1]} dx \tag{11}$$

By using the substitution y=(x/θ)$^{-(\gamma-1)}$, Equation (11) may be written as

$$\int_{\theta}^{\infty} f^{\xi}(x)dx = \left(\frac{\gamma-1}{\theta}\right)^{\xi-1}\frac{B\left(\xi(\alpha-1)+1,\xi\left[\beta+\frac{1}{(\gamma-1)}\right]-\frac{1}{(\gamma-1)}\right)}{B^{\xi}(\alpha,\beta)}$$

The Rényi entropy can now be written as

$$I_R(\xi) = -\log\left(\frac{\gamma-1}{\theta}\right)+\frac{1}{1-\xi}\log\left\{\frac{B\left(\xi(\alpha-1)+1,\xi\left[\beta+\frac{1}{(\gamma-1)}\right]-\frac{1}{(\gamma-1)}\right)}{B^{\xi}(\alpha,\beta)}\right\} \tag{12}$$

A special case of Equation (10) is defined by Shannon as E{-log[f(X)]}. This is obtained by taking the limit of the Rényi entropy when ξ → 1.

$$E\left\{-\log\left[f(X)\right]\right\} = \lim_{\xi\to 1}\left\{\frac{1}{1-\xi}\log\left[\int f^{\xi}dx\right]\right\} \tag{13}$$

By taking the limit of Equation (13) as ξ → 1, using the L'Hospital's rule and simplifying the result, we obtain

$$E\left\{-\log\left[f(X)\right]\right\} = -\log\left(\frac{\gamma-1}{\theta}\right)-(\alpha-1)\Psi(\alpha) -\left(\beta+\frac{1}{\gamma-1}\right)\Psi(\beta)+\log B(\alpha,\beta) +\left(\alpha-1+\beta+\frac{1}{\gamma-1}\right)\Psi(\alpha+\beta)$$

where Ψ(z)=Γ′(z)/Γ(z) is a digamma function.

7. Maximum likelihood estimates of the parameters

The log-likelihood function of BLBP distribution can be algorithmically described as follows: let $X_1,..,X_n$ be a random sample from X~BLBP(α,β,θ,γ) and let Θ=(α,β,θ,γ)^T be the vector of the model parameters. The likelihood function for Θ can be expressed as:

$$L(x;\Theta) = \prod_{i=1}^{n}\left\{\frac{(\gamma-1)}{\theta B(\alpha,\beta)}\left[1-\left(\frac{x_i}{\theta}\right)^{-(\gamma-1)}\right]^{\alpha-1}\left(\frac{x_i}{\theta}\right)^{-(\gamma-1)\beta-1}\right\}$$

The log-likelihood function for Θ reduces to:

$$\ln L(\Theta) = n\ln(\gamma-1)-n\ln\theta-n\ln\Gamma(\alpha)-n\ln\Gamma(\beta) +n\ln\Gamma(\alpha+\beta)+(\alpha-1)\sum_{i=1}^{n}\ln\left[1-\left(\frac{x_i}{\theta}\right)^{-(\gamma-1)}\right] -\left[(\gamma-1)\beta+1\right]\sum_{i=1}^{n}\ln\left(\frac{x_i}{\theta}\right) \tag{14}$$

Differentiating Equation (14) with respect to α, β and γ, respectively, and setting the results equal to zero, we have

$$\frac{\partial \ln L(\Theta)}{\partial\alpha} = n\Psi(\alpha+\beta)-n\Psi(\alpha)+\sum_{i=1}^{n}\ln\left[1-\left(\frac{x_i}{\theta}\right)^{-(\gamma-1)}\right] \tag{15}$$

$$\frac{\partial \ln L(\Theta)}{\partial\beta} = n\Psi(\alpha+\beta)-n\Psi(\beta)-(\gamma-1)\sum_{i=1}^{n}\ln\left(\frac{x_i}{\theta}\right) \tag{16}$$

and,

$$\frac{\partial \ln L(\Theta)}{\partial\gamma} = \frac{n}{(\gamma-1)}-\sum_{i=1}^{n}\left\{\beta+(\alpha-1)\left[1-\left(\frac{x_i}{\theta}\right)^{\gamma-1}\right]^{-1}\right\}\ln\left(\frac{x_i}{\theta}\right) \tag{17}$$

Because of x ≥ θ, hence, $\hat{\theta}$ = $x_{(1)}$, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ for the parameters α, β and γ respectively, are obtained by solving iteratively Equation (15)-(17). The expectations of the partial derivatives are not in simple form. A Newton-Raphson method can be employed to obtain the expectations.

8. Application of the BLBP distribution

In this section, the BLBP distribution is fitted to real data set, we consider the data set in the field of insurance which has received extensive attention in the actuarial literature. This data set is one among the twenty sets of Norwegian fire claims (in millions of Norwegian krones) is presented in Fernández [6]. The resulting parameter estimates for Norwegian fire claims data show that in Table 1. We can see they based on K-S statistic the p-value of BLBP distribution is provided the best fitted when compare to the Weibull and beta-Pareto distributions.

Table 1: Parameter estimates and K-S statistics for Norwegian fire claims data.

Distribution	Weibull	beta-Pareto	BLBP

Parameter
estimates
	Threshold=0.5
	Shape=0.5792
	Scale=0.8276
		$\hat{\theta}$ =0.5
		$\hat{\gamma}$ =169.1104
		$\hat{\alpha}$ =0.3183
		$\hat{\beta}$ =0.0069
			$\hat{\theta}$ =0.5
			$\hat{\gamma}$ =135.4664
			$\hat{\alpha}$ =0.3468
			$\hat{\beta}$ =0.0088

K-S statistic	0.0983	0.0444	0.0413
p-value	0.128	0.942	0.969

9. Conclusion

We proposed the BLBP distribution, it has some sub-models, such as LBP, arcsine, log-beta, exponential distributions. Rényi and Shannon entropies and hazard rate are provided. We derive the r th moments and apply maximum likelihood estimation to estimate parameters of the distribution. An application to real data set shows that the fit of the BLBP distribution is best fit to the data with high p-value. We hope BLBP distribution may attract extensive applications in lifetime data analysis and other fields. The future research may consider in parameter estimation using Bayesian or other approaches.

Acknowledgements

This research was made possible by Department of Statistics, Faculty of Science, Kasetsart University and National Statistics Office, Thailand. We thank anonymous reviewers for comments made on an earlier version of this manuscript and for many inspiring suggestions.

References

[1] Abdel-All N, Mahmoud M, Abd-Ellah H. Geometrical properties of Pareto distribution, Appl. Math. Comput. 2003; 145(2): 321-339.

[2] Wu J, Lee W, Chen S. Computational comparison for weighted moments estimators and BLUE of the scale parameter of a Pareto distribution with known shape parameter under type II multiply censored sample, Appl. Math. Comput. 2006; 181(2): 1462-1470.

[3] Hong C, Wu J, Cheng C. Computational procedure of performance assessment of lifetime index of businesses for the Pareto lifetime model with the right type II censored sample, Appl. Math. Comput. 2007; 184(2): 336-350.

[4] Wu S. Interval estimation for a Pareto distribution based on a doubly type II censored sample, Comput. Statist. Data Anal. 2008; 52: 3779-3788.

[5] Soliman A. Estimations for Pareto model using general progressive censored data and asymmetric loss, Commun. Statist. - Theory Meth. 2008; 37(9): 1353-1370.

[6] Fernández A. Smallest Pareto confidence regions and applications, Comput. Statist. Data Anal. 2013; 62: 11-25.

[7] Newman MEJ. Power laws, Pareto distributions and Zipf’s law, Contemp. Phys. 2005; 46(5): 323–351.

[8] Ali M, Nadarajah S. A truncated Pareto distribution, Comput. Commun. 2006; 30: 1-4.

[9] Manas A. The paretian ratio distribution - an application to the volatility of GDP, Economics Letters. 2011; 111: 180-183.

[10] Nassar M, Nada N. A new generalization of the Pareto geometric distribution, J. Egy. Math. Soc. 2013; 21: 148-155.

[11] Patil G, Rao C. Weighted distributions and size-biased sampling with applications to wildlife populations and human families, Biometrics. 1978; 34(2): 179-189.

[12] Oluyede B. A note on probability weighted moment inequalities for reliability measures, J. Inequal. Pure Applied Math. 2006; 7(1): Art. 28.

[13] Das K, Roy T. On some length-biased weighted Weibull distribution, Adv. Applied Sci. Res. 2011; 2(5): 465-475.

[14] Nanuwong N, Bodhisuwan W. Length biased beta-Pareto distribution and its structural properties with application, J. Math. Stat. 2014; 10(1): 49-57.

[15] Eugene N, Lee C, Famoye F. The beta-normal distribution and its applications, Commun. Statist. - Theory Meth. 2002; 31(4): 497-512.

[16] Cintra R, Rêgo L, Cordeiro G, Nascimento A. Beta generalized normal distribution with an application for SAR image processing, Statistics. 2014; 48(2): 279-294.

[17] Singla N, Jain K, Sharma S. The beta generalized Weibull distribution: Properties and applications, Reliab. Eng. Syst. Saf. 2012; 102: 5-15.

[18] Paranaiba P, Ortega E, Cordeiro G, Pescim R. The beta Burr XII distribution with application to lifetime data, Comput. Statist. Data Anal. 2011; 55: 1118-1136.

[19] Barreto-Souza W, Santos A, Cordeiro G. The beta generalized exponential distributions, J. Statist. Comput. Simulat. 2010; 80(2): 159-172.

[20] Fischer M, Vaughan D. The beta-hyperbolic secant distribution, Austrian Journal of Statistics. 2010; 39(3): 245-258.

[21] Pescim R, Demétrio C, Cordeiro G, Ortega E, Urbano M. The beta generalized half-normal distribution, Comput. Statist. Data Anal. 2010; 54: 945-957.

[22] Shuaib M. The beta inverse Weibull distribution, Int. Trans. Math. Sci. Comput. 2010; 3(1): 113-119, ISBN-0975-3753.

[23] Silva G, Ortega E, Cordeiro G. The beta modified Weibull distribution, Lifetime Data Anal. 2010; 16: 409-430.

[24] Famoye F, Lee C, Olumolade O. The beta-Weibull distribution, J. Statist. Theory Appl. 2005; 4(2): 121-136.

[25] Mahmoudi E. The beta generalized Pareto distribution with application to lifetime data, Math. Comput. Simulat. 2011; 81: 2414-2430.

[26] Cordeiro G, Gomes A, Silva C, Ortega E. The beta exponentiated Weibull distribution, J. Statist. Comput. Simulat. 2013; 83(1): 114-138.

[27] Lemonte AJ, Cordeiro GM. An extended Lomax distribution, Statistics. 2013; 47(4): 800-816.

[28] Akinsete A, Famoye F, Lee C. The beta-Pareto distribution, Statistics. 2008; 42(6): 547-563.

[29] Nadarajah S, Kotz S. The beta Gumbel distribution, Math. Probab. Eng. 2004; 10: 323-332.

[30] Sunoj S, Linu M. Dynamic cumulative residual Rényi’s entropy, Statistics. 2013; 46(1): 41-56.

The new Hodges-Lehmann estimator control charting technique for the known process distributions

Adisak Pongpullponsak[1*] and Vadhana Jayathavaj[2]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, adisak.pon@kmutt.ac.th*
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, vadhana.j@rsu.ac.th*

Abstract

This research studies the performances (average run length) of the Hodges-Lehmann estimator control chart for the symmetric standard normal distribution and 11 shapes (skewness = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9) of the standard Weibull distribution, both when the process is in control and when the shift occurs (shift in δ times of standard deviation, δ = 0.25, 0.50, 0.75, 1, 1.5, 2, 2.5, and 3). The Hodges-Lehmann estimator (the median of Walsh average) is the nonparametric (distribution free) statistical inference method. The Wilcoxon signed rank statistics are assumed to be the location of the Walsh average to determine the Hodges-Lehmann estimator control limits in the original Hodges-Lehmann quality control charting technique. For the aforementioned process distributions, the average run length of the Hodges-Lehmann estimator control chart is not performed corresponding to the probability distribution of the Wilcoxon signed rank test statistics. The new charting approach by constructing the control limits from the Hodges-Lehmann estimator distribution through simulation is reflected the expected real process average run length. For the process that needs the robust to the outliers test statistic of the Hodges-Lehmann estimator, this new charting method is also an alternative.

Keywords: Control chart, Hodges-Lehmann estimator, nonparametric statistic, statistical process control, Walsh average, Wilcoxon signed rank

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

1. Introduction

The traditional parametric Shewhart type $\overline{X}$ control chart tests hypothesis for the process variable that has or assumed to be the normal distribution by using the probability distribution of $\overline{X}$ within ±3σ control limits [1]. The tested statistic in the Shewhart style nonparametric control charts, the binomial distribution with the probability = ½ (the distribution of the number of sign +/- from the process median) is used in the Sign test, the Mann-Whitney U statistic (the distribution of the sum of the rank of the treatment data after combined with the control data) is used in the Mann-Whitney control chart, and etc., the control chart names after their nonparametric statistic used in hypothesis testing and assuming that the underlying process distribution is followed that nonparametric statistic [2-5]. But in the Hodges-Lehmann estimator control chart (HLCC), this control chart tests hypothesis differ from another nonparametric control charts. The median of Walsh average transforms process data to the Hodges-Lehmann estimator (HLE), the HLE is tested by assuming that the HLE control limits at the given Wilcoxon signed rank statistic locations have the Wilcoxon signed rank probability mass function (WSR) at that locations, WSR is the distribution of sum of the rank by the plus (+) or minus (-) sign from the process median (not the distribution of Hodges-Lehmann estimator) [4]. Through the process data transformation by the Walsh average, there are no relationships between the statistic from the underlying process distribution (the Hodges-Lehmann estimator) and the Wilcoxon signed rank distribution. Many articles persuaded to use nonparametric control charts (the distribution free control charts) for the process which has less to no information about the process distribution (possibly in the initial stage – Phase I in quality control system), Chakraborti and Van de Weil also showed many advantages of using the nonparametric control charts [3]. Neuhäuser [6] cited from many authors that a normality assumption was often not justified in the statistical practice, especially in the field of health sciences. A branch of statistics known as nonparametric statistics or distribution free statistics are used when the population from which selected samples are not normally distributed or normality cannot be met. Several nonparametric tests can be applied in case of nominal or ordinal data. The Hodges-Lehmann estimator is the nonparametric statistic that has very robust to the outliers [7]. In the quality control operations, after implement HLCC for a period of time, the process data probability distribution have been gathered and may follow the Central Limit Theorem (fit to the normal distribution) or perhaps the distribution with some degree of skewness. For the known process distribution, the type I error (α), the in control average run length (ARL_0) will not follow the control limits determined by using the Wilcoxon signed rank statistic. Once the process distribution is known, it is possible to use the Hodges-Lehmann estimator probability distribution in hypothesis testing vis-à-vis using the distribution of $\overline{X}$ in hypothesis testing in the $\overline{X}$ control chart. In the parametric control chart, the standard normal distribution is the standard choice for studying the symmetrical process distribution, and for the asymmetrical shapes, the 11 shapes of standard Weibull distribution represent the skewed process distribution from symmetry to asymmetry (elongated tail at the right, more data in the right tail than would be expected in the normal distribution) with the positive skewnesses ranging from 0.1, 0.5,1, 2, 3, 4, 5, 6, 7, 8, and 9 had been studied in many parametric control charts [8-10]. The chart performance in terms of the average run length (ARL) of the HLCC fixed parameters control charts are investigated for the known process variable of standard normal distribution and the standard Weibull distribution with above 11 skewnesses and the process mean (μ) has shift in δ times of standard deviation (σ) from μ to $\mu+\delta\sigma$ when δ = 0, 0.25, 0.50, 0.75, 1, 1.5, 2, 2.5, and 3.

2. Research Methodology

To verify that the theoretical average run length (ARL) from the selected Wilcoxon signed rank statistic control limits of the HLCC has the same values with ARL from the HLE distribution (The Shewhart type $\overline{X}$ chart tests the distribution of the sample mean ($\overline{X}$) from the sample, if the random variable X has the normal distribution, $\overline{X}$ from the sample size *n* also has the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$) [1], the ARL from the selected WSR control limits is to be computed.

For the known process distribution, the distribution of the order statistic can be derived from the location of that order [11] (For example, the median is always at the middle of data), but the distribution of the HLE cannot be derived from the order statistics, because the median of Walsh average is not always forming from the observations at the same ranks, the HLE depends on the magnitude of observations instead [4]. The simulation is used to identify the control limits at the determined control region probabilities from the process distribution and to count the run lengths, both the in control state and when the shift occurs. This is the new approach to the Hodges-Lehmann estimator control chart for the known process distribution.

The related terms and theory are as follows.

2.1 Performance and Shift Detection

A popular measure of chart performance is the expected value of the run length (the number of samples or subgroups that need to be collected before the first out of control signal is given, by a chart is a random variable called the run length) distribution called the average run length (ARL). It is desirable to have a large value of ARL when the process is in control and small value when the process is out of control.

The in control average run length (ARL_0) is the average number of samples before the control chart signals an out of control condition. The ARL_0 can be calculated from

$$ARL_0 = 1/\alpha \qquad (1)$$

where α is the probability of the type-I error or the probability that any point exceeds the control limits.

The out of control average run length (ARL_1) is the number of samples to detect the process shift, on the average. The type-II error (β) is the error that occurs when the chart does not give out of control signal when actually the process is out of control.

$$ARL_1 = 1/(1-\beta) \qquad (2)$$

2.2 Distributions of Process Variable

The process variable in the real operation has many forms of distribution. The shape of distribution vary from symmetry to asymmetry explains by skewness of the distribution. The normal distribution is selected to represent the symmetry, and the Weibull distribution with various skewnesses are selected to represent the asymmetry.

2.2.1 The Standard Normal Distribution

Assume that X to be the process variable, $X \sim N(\mu,\sigma)$ with mean μ and standard deviation σ, the change variable $z=(x-\mu)/\sigma$ converts X from an $N(\mu,\sigma)$ random variable into an $N(0,1)$ random variable, the standard normal distribution. The standard normal distribution is the test statistic in the Shewhart $\overline{X}$ chart, and also help us follow the sense Type I error (α) of $\pm 3\sigma$ control limits in this nonparametric control chart.

2.2.2 The Standard Weibull Distribution

For the process X that assumed to be a Weibull distribution with the scale parameter (θ) and the shape parameter (β), when θ =1, a Weibull distribution is called the standard Weibull distribution. From Nelson [9] as cited in [10] , 11 skewnesses ranging from 0.1, 0.5, 1, 2, …, 9 are selected to evaluate the performance of the control system. The shape parameters of these 11 skewnesses are shown in Table 1.

Table 1: Skewness coefficient and shape parameter of Weibull distribution with the scale parameter $\theta = 1$

Skewness coefficient	Shape parameter (β)
0.1	3.2219
0.5	2.2110
1	1.5630
2	1.0000
3	0.7686
4	0.6478
5	0.5737
6	0.5237
7	0.4873
8	0.4596
9	0.4376

2.3 Process Shift

The statistical process control in Shewhart style control chart, the Type I error probability (α) and the Type II error probabilty (β) are exhibited by plotting the process data on the control limits in an in control state with mean μ and when the process mean is shifted to $\mu + \delta\sigma$ respectively, as shown in Figure 1. In Figure 1 the +k and –k are the upper and lower action limits, and the +w and -w are the upper and lower warning limits in an in control state. The Type I error probability (α) is the probability that $x < -k$ or $x > +k$ for x in an in control state and the Type II error probability (β) is the probability that $x \geq -k$ and $x \leq +k$ for x from the process that the mean is shifted from μ to $\mu + \delta\sigma$ (out of control).

The in control state (the process mean = μ)

α_1 = the probability that $x > +k$

α_2 = the probability that $x < -k$

$\alpha = \alpha_1 + \alpha_2$

The out of control state (the process mean is shifted from μ to $\mu + \delta\sigma$)

γ_1 = the probability that $x > +k$

γ_2 = the probability that $x < -k$

$1 - \beta = \gamma_1 + \gamma_2$

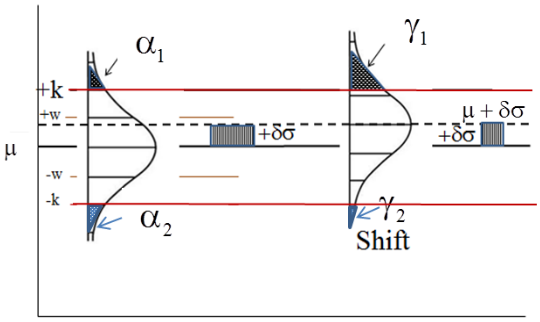


Figure 1: Process shift

2.4 The Hodges-Lehmann Estimator Control chart

The implementation of the nonparametric control chart is recommended for the quality control process that the normality cannot be met or less information or not enough detail information about the probability distribution of the process variable. In a distribution free inference, whether for testing or estimation, the methods are based on functions of the observation which does not depend on the specific distribution function of the population from which the sample was drawn [11]. The test statistic in the Hodges-Lehmann estimator control chart is the median of Walsh averages, and this statistic assumes the symmetric discrete Wilcoxon signed rank distribution in hypothesis testing [4].

2.4.1 The Original Hodges-Lehmann estimator Control Charting Technique

Alloway and Raghavachari [4] as cited in Das [5] considered a Shewhart type control chart for the point of symmetry θ of a continuous, symmetric population based on a distribution free confidence interval for θ, and calculated the Hodges–Lehmann estimator as follows.

Let $X_1, X_2, ..., X_n$ be a random sample of n observations.

Define the $M = \dfrac{n(n+1)}{2}$

The Walsh averages WA_r ,

$$WA_r = \frac{(X_i + X_j)}{2} \qquad \begin{matrix} r = 1, 2, ..., M \\ i \leq j \\ i = 1, 2, ..., n \\ j = 1, 2, ..., n \end{matrix} \qquad (3)$$

The Hodges-Lehmann estimator of $\hat{\theta}$ or $\widehat{hl}$ is defined as the median of the Walsh averages for the sample. In other words, if $WA_{(r)}$ for $r = 1, 2, ..., M$ are the ordered Walsh averages then

$$\hat{\theta} = \begin{cases} WA_{(k+1)} & \text{if } M \text{ is odd} \\ \left(WA_{(k)} + WA_{(k+1)}\right)/2 & \text{if } M \text{ is even} \end{cases}$$

where

$$k = \begin{cases} M/2 & \text{if } M \text{ is odd} \\ (M-1)/2 & \text{if } M \text{ is even} \end{cases} .$$

If WSR^{+} is the Wilcoxon signed rank statistic for testing the null hypothesis that the underlying distribution of the individual observations is symmetric about θ then WSR^{+} is the number of Walsh averages greater than θ.

The control chart is based on the Hodges–Lehmann estimator and the associated Wilcoxon signed rank confidence interval. The approach differs from a Shewhart control chart that the control limits are based on the order statistics instead of traditional measures of mean and dispersion. Let $WSR^{+}(\alpha/2,n)$ be the upper $100(1-\alpha)^{\text{th}}$ percentile point of the null distribution of the Wilcoxon signed rank statistic for a sample of size n. The 100(1–α)% confidence interval for θ is given by the following order statistics of the Walsh averages [12]

$$\left(WA_{\left(M+1-WSR^{+}(\alpha/2,n)\right)}, WA_{\left(WSR^{+}(\alpha/2,n)\right)}\right).$$

The positions of the two ordered Walsh averages determine the subgroup control limit values by using the connection with Wilcoxon signed-rank statistic.

The steps for calculating the Hodges-Lehmann estimator control chart are as follows.

For each subgroup

Step 1. Collect a subgroup of size $n, X_1, X_2, ..., X_n$. A minimum size of 10 is required to achieve a significance level closes to ±3σ control limits in traditional used.

Step 2. Using the table of positions of the two ordered Walsh averages to determine the subgroup control limit values for the upper and the lower control values of Walsh averages.

Step 3. Compute the Hodges–Lehmann estimator.

To determine the center line for the control chart

Determine the average of the Hodges–Lehmann estimators from each subgroup form step 3.

To determine the upper and lower control limits

Determine the median of the upper (lower) values from step 2. above for all subgroups. This is the upper (lower) control limit.

2.4.2 Performance evaluation - ARL_{0WSR} from given Wilcoxon signed rank statistic control limits

The symmetrical discrete Wilcoxon signed rank statistic is assumed for the Hodges-Lehmann estimator control chart. To evaluate the performance of the Hodges-Lehmann estimator control chart, the ARL_0 derives from the Wilcoxon signed rank probability mass function (ARL_{0WSR}) can be computed in the following steps,

- Determine the sample size (n)
- Use the Wilcoxon signed rank statistic table [13] to determine the Wilcoxon signed rank statistics for lower and upper control limits (LCL_{WSR}, UCL_{WSR}) at the desired Type I error probability level (α_{WSR}).
- Compute the theoretical ARL_0 by using the desired probability value at the Wilcoxon signed rank statistic (α_{WSR}),

$$ARL_{0WSR} = \frac{1}{\alpha_{WSR}}. \tag{4}$$

2.4.3 The average run length from the probability of the process variable in the range of the Hodges-Lehmann estimator control limits (ARL_{0X})

In the Shewhart $\overline{X}$ chart, the hypothesis is tested against the probability distribution of $\overline{X}$. If the process variable X has the normal distribution with mean (μ) and standard deviation (σ) $(X \sim N(\mu,\sigma))$, then the sampling distribution of the sample mean $(\overline{X})$ also has the normal distribution with $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. To verify that the Hodges-Lehmann estimator will belong to the distribution of the process variable, the probability of x in the range of the upper control limit (UCL_{HL}) and the lower control limit (LCL_{HL}) of the Hodges-Lehmann estimator control chart can be computed directly from the distribution of process data,

α_X = Probability($(x < LCL_{HL})$ and $(x > UCL_{HL})$),

and the in control average run length from the process variable data distribution is,

$$ARL_{0X} = \frac{1}{\alpha_X}. \tag{5}$$

2.4.4 Performance of Hodges-Lehmann estimator control limits for the known distribution

The simulation will perform to verify the Hodges-Lehmann estimator control chart in the case of known process distribution by using the following computation steps,

Create the Hodges-Lehmann estimator control chart as stated in 2.4.1 above

- Generate 100,000 subgroups of Walsh averages of sample size n using the given process distribution.
- Using the table of positions of the two ordered Walsh averages [13] to determine the subgroup control limit values to obtain the upper and lower control values of Walsh averages.
- Determine the median of the upper values from 100,000 subgroups above. This is the upper control limit (UCL_{HL}). The lower control limit (LCL_{HL}) is determined in the same fashion.

2.4.5 The in control average run length (ARL_{0HL})

Using the control limits from 2.4.4

- let NRL_i for $(i = 1, 2, ..., 10000)$ be the 10,000 run lengths of Hodges-Lehmann estimators of the sample size n from the given distribution

- Each run length computation

    - count the consecutive number of Hodges-Lehmann estimators $\left(\widehat{hl}\right)$ which greater than LCL_{HL} and lesser than UCL_{HL}

    - If the first out of control ($\widehat{hl}$ is lesser than LCL_{HL} or greater than UCL_{HL}) Hodges-Lehmann estimator $\left(\widehat{hl}\right)$ is found, then the run length equals to the number of the previous Hodges-Lehmann estimator before out of control.

    - the in control average run length

$$ARL_{0HL} = \sum_{i=1}^{10000} NRL_i \bigg/ 10000 \qquad (6)$$

2.4.6 Shift detection – the out of control average run length (ARL_{1HL}^{δ})

Using the control limits from 2.4.4

- given NRL_i^{δ} for $(i = 1, 2, ..., 10000)$ = the 10,000 run lengths of Hodges-Lehmann estimators of the sample size n using the given process distribution with shift in δ times standard deviation (σ)

- Each run length computation (i)

    - continuously generate random numbers of sample size n from $N(0,1)$ with shift in δ times of standard deviation (σ)

    - compute the Hodges-Lehmann estimator $\left(\widehat{hl}\right)$ for each sample

    - count the consecutive number of Hodges-Lehmann estimators $\left(\widehat{hl}\right)$ which greater than LCL_{HL} and lesser than UCL_{HL} (in control)

    - if the Hodges-Lehmann estimator $\left(\widehat{hl}\right)$ lesser than LCL_{HL} or greater than UCL_{HL} (out of control), then i^{th} run length equals to the number of this first out of control sample.

- the out of control average run length

$$ARL_{1HL}^{\delta} = \sum_{i=1}^{10000} NRL_i^{\delta} \bigg/ 10000 \qquad (7)$$

2.5 The new charting approach - The Hodges-Lehmann estimator control chart using probability distribution of the Hodges-Lehmann estimator itself

In the case of unknown process variable data distribution, the control limits from Wilcoxon signed rank are used Hodges-Lehmann estimator control chart as shown in 2.4.1 are the typical implementation. But if the process variable has their own distribution, the Hodges-Lehmann estimator should test hypothesis against the Hodges-Lehmann estimator distribution as same as the Shewhart $\overline{X}$ chart tests the sample $\overline{X}$ statistic against the probability distribution of $\overline{X}$.

2.5.1 The Hodges-Lehmann estimator control limits from the Hodges-Lehmann estimator distribution by simulation.

As earlier mentioned that the Hodges-Lehmann estimator distribution cannot be obtained from the order statistics that form the median of Walsh average, the simulation is brought to explore the performance in the case of known process distribution.

The simulation steps are as follows,

- determine the Type I error probability and the probabilities for the action region

- determine the sample size

- compute 100,000 Hodges-Lehmann estimators from the random number generate under the given distribution and sort them in ascending order.

- identify the new Hodges-Lehmann estimator control limits, the two new action limits (the new lower control limit ($LCL_{\widehat{HL}}$) and the new upper control limit $\left(UCL_{\widehat{HL}}\right)$ from the locations correspond to the given probabilities of the 100,000 Hodges-Lehmann estimators. For example, given the probability of 0.0027 (as at +/-3 σ in the standard normal distribution), the locations at $\frac{\alpha}{2}$ for the upper control limits and the lower control limits are 135 and 99865 (100000 - 135).

2.5.2 The New Control limits

To find the ARL_0 and ARL_1 from this new charting technique (by using the Hodges-Lehmann estimator distribution to create the control limits), the control limits must be constructed from the given sample size using the given process distribution.

Simulated the Hodges-Lehmann estimator control limits

- determine the sample size n

- determine the type I probability for upper control limit ($\alpha / 2$) and the lower control limit ($\alpha / 2$)

- compute 100,000 set of Walsh averages of the sample size n using the given process distribution

- get the upper and lower control limits for the the Hodges-Lehmann estimator control chart from the $\widehat{hl}$ at the locations correspond to the type I probability for the new upper limit and the new lower control limit ($UCL_{\widehat{HL}}$ and $LCL_{\widehat{HL}}$)

2.5.3 The in control average run length ($ARL_{0\widehat{HL}}$)

- let NRL_i for $(i = 1, 2, ..., 10000)$ be the 10,000 run lengths of Hodges-Lehmann estimators of the sample size n from the standard normal distribution

- Each run length computation
  - count the consecutive number of Hodges-Lehmann estimators $\left(\widehat{hl}\right)$ which greater than $LCL_{\widehat{HL}}$ and lesser than $UCL_{\widehat{HL}}$
  - If the first out of control Hodges-Lehmann estimator $\left(\left(\widehat{hl} < LCL_{HL}\right)\right.$ or $\left.\left(\widehat{hl} > UCL_{\widehat{HL}}\right)\right)$ is found, then the run length equals to the number of the previous Hodges-Lehmann estimator before out of control.
  - the in control average run length

$$ARL_{0\widehat{HL}} = \sum_{i=1}^{10000} NRL_i \Big/ 10000 \qquad (8)$$

2.5.4 *Shift detection – the out of control average run length ($ARL^{\delta}_{1\widehat{HL}}$)*

- given NRL^{δ}_{i} for $(i = 1, 2, ..., 10000)$ = the 10,000 run lengths of Hodges-Lehmann estimators of the sample size n using the given process distribution with shift in δ times of standard deviation (σ)

- Each run length computation (i)
  - continuously generate random numbers of sample size n from $N(0,1)$ with shift in δ times of standard deviation (σ)
  - compute Hodges-Lehmann estimator $\left(\widehat{hl}\right)$ for each sample
  - count the consecutive number of Hodges-Lehmann estimators $\left(\widehat{hl}\right)$ which greater than $LCL_{\widehat{HL}}$ and lesser than $UCL_{\widehat{HL}}$
  - If Hodges-Lehmann estimator $\left(\widehat{hl}\right)$ is lesser than $LCL_{\widehat{HL}}$ or greater than $UCL_{\widehat{HL}}$, then the i^{th} run length equals to the number of that out of control sample.

- the out of control average run length for the new Hodges-Lehmann estimator control limits is

$$ARL^{\delta}_{1\widehat{HL}} = \sum_{i=1}^{10000} NRL^{\delta}_{i} \Big/ 10000 . \qquad (9)$$

3. Research Results and Discussion

The performance of the Hodges-Lehmann estimator control chart is evaluated, the standard normal distribution is used in the original method, and both the symmetric standard normal distribution and the 11 skewnesses of standard Weibull distribution are used in the new purposed charting technique. MATLAB [14] and MINITAB [15] are the computing program in this research.

3.1 Performance of the original Hodges-Lehmann estimator control charting technique.

The sample size of 10 is the smallest starting sample size to obtain the small enough α probability to get the acceptable ARL (at least greater than 370 of +/-3 σ control limits when compares to the traditional $\overline{X}$ chart), if the sample size less than 10, the starting α probability is high, the ARL is quite low. The Wilcoxon signed rank statistic at the sample size n =10 with 11 control limits (the lower control limit varies from 1, 2, 3,…, 11, and the upper control limit varies from 54, 53, 52, …, 44 respectively) or 11 schemes bring to study the control limits behavior and performance of the Hodges-Lehmann estimator control charts.

3.1.1 The average run length from the assumed probability associated to the Wilcoxon signed rank control limits $\left(ARL_{0WSR}\right)$.

The α probability comes from the Wilcoxon signed rank table [13] . Using the procedure steps as shown in subsection 2.4.2, the results are shown in Table 2, the ARL_{0WSR} are 510.20, 256.41, …, 11.91 for the lower limit at 1, 2, 3, …, 11 (the scheme 1 to 11) respectively. From scheme 1 to 11, when the Wilcoxon signed rank control limits are narrower (Type I error is larger), the ARL_{0WSR} are smaller too.

The Hodges-Lehmann control limits derive from the given Wilcoxon signed rank control limits in Table 2 compute by using the original charting technique as shown in 2.4.1 above. The 100,000 samples of 10 observations are simulated and the Hodges-Lehmann control limits that shown in Table 2 also narrower corresponding to the Wilcoxon signed rank control limits. The 100,000 simulated samples comes from the test run on the Shewhart type $\overline{X}$ chat, when the number of simulated run length increase, the simulated ARL approaches the theoretical ARL (ARL_0 = 258.80, 395.44, 380.00, 362.43, and 368.49 at 10, 100,1000, 10000 and 100000 runs, respectively) 368.49 is closed to 370.37 for +/-3 σ control limits.

3.1.2 The average run length from the process distribution in the Hodges-Lehmann estimator control limits derive from the Wilcoxon signed rank control limits $\left(ARL_{0X}\right)$

To verify that the Type I error (α) probability of the standard normal distribution process data using the Hodges-Lehmann control limits (α_X) in 3.1.1 is conformed to the Type I error probability from Wilcoxon sign rank control limits (α_W). From Table 2, the α_X (compute by the procedure in subsection 2.4.3) are very large due to the robust to the outliers of the median of Walsh average which transform process samples close to their process mean (μ =0), the ARL_{0X} are very small, and ARL_{0X} and ARL_{0W} are in different magnitude. The Type I error probabilities (α) from the process data and the test statistic are in different magnitude in the original Hodges-Lehmann estimator control chart implementation, while they are the same in the parametric Shewhart type $\overline{X}$ chart.

Table 2 Hodges-Lehmann estimator control chart for fixed sample size n=10
for 11 action limits of the standard normal distribution

Scheme	1	2	3	4	5
n(n+1)/2	55	55	55	55	55
Wilcoxon Signed Rank Statistics control limit					
- UCL_{WSR}	54	53	52	51	50
- LCL_{WSR}	1	2	3	4	5
Right tailed probability					
at LCL_{WSR}	0.00098	0.00195	0.00293	0.00488	0.00684
α_{WSR}	0.00196	0.0039	0.00586	0.00976	0.01368
ARL_{0WSR}	**510.20**	256.41	170.65	102.46	73.10
Hodges-Lehmann estimator control limits from 100,000 runs simulation					
- UCL_{HL}	1.2544	1.1517	1.0299	0.9607	0.8863
- LCL_{HL}	-1.4993	-1.2538	-1.1518	-1.0339	-0.9591
Probability of process data					
in control region	0.8283	0.7703	0.7238	0.6811	0.6435
α_X	0.1717	0.2297	0.2762	0.3189	0.3565
ARL_{0X}	5.82	4.35	3.62	3.14	2.81
10,000 run lengths simulation					
ARL_{0HL}	**13281.35**	3374.91	926.33	385.53	188.54
ARL_{1HL}^{δ} with shift in times of standard deviation					
Shift $(\delta\sigma)$					
0.25	896.34	335.75	113.16	66.86	37.90
0.50	94.12	41.66	18.93	12.32	8.51
0.75	16.04	8.17	5.17	3.83	2.94
1.00	4.63	3.13	2.14	1.84	1.57
1.50	1.29	1.16	1.08	1.05	1.03
2.00, 2.50, 3.00	1.00	1.00	1.00	1.00	1.00

3.1.3 The in control average run length from the given process distribution Hodges-Lehmann estimator simulation $\left(ARL_{0HL}\right)$

To find the in control average run length from the given process distribution Hodges-Lehmann estimator simulation $\left(ARL_{0HL}\right)$, the simulation results using the procedural steps in subsection 2.4.4 and 2.4.5 show in Table 2 that ARL_{0HL} are very high and sharply downward from 13821.35 to 12.20 for the lower control limits from 1 to 11 (the scheme 1 to 11).

Table 2 Hodges-Lehmann estimator control chart for fixed sample size n=10
for 11 action limits of the standard normal distribution (continue)

Scheme	6	7	8	9	10	11
n(n+1)/2	55	55	55	55	55	55
Wilcoxon Signed Rank Statistics control limit						
- UCL_{WSR}	49	48	47	46	45	44
- LCL_{WSR}	6	7	8	9	10	11
Right tailed probability						
at LCL_{WSR}	0.00977	0.01367	0.01855	0.02441	0.03223	0.04199
α_{WSR}	0.01954	0.02734	0.0371	0.04882	0.06446	0.08398
ARL_{0WSR}	51.18	36.58	26.95	20.48	15.51	11.91
Hodges-Lehmann estimator control limits from 100,000 runs simulation						
- UCL_{HL}	0.8223	0.7628	0.7111	0.6570	0.6090	0.5646
- LCL_{HL}	-0.8851	-0.8200	-0.7600	-0.7092	-0.6557	-0.6079
Probability of process data						
in control region	0.6065	0.5711	0.5379	0.5053	0.4727	0.4422
α_X	0.3935	0.4289	0.4621	0.4947	0.5273	0.5578
ARL_{0X}	2.54	2.33	2.16	2.02	1.90	1.79
10,000 run lengths simulation						
ARL_{0HL}	102.88	60.28	38.26	24.92	17.54	12.20
Shift $(\delta\sigma)$	ARL_{1HL}^{δ} with shift in times of standard deviation					
0.25	24.54	17.15	12.41	9.23	7.25	5.75
0.50	6.13	4.64	3.83	3.13	2.70	2.38
0.75	2.42	2.05	1.81	1.64	1.49	1.40
1.00	1.41	1.32	1.23	1.18	1.13	1.10
1.50	1.02	1.01	1.01	1.00	1.00	1.00
2.00, 2.50, 3.00	1.00	1.00	1.00	1.00	1.00	1.00

3.1.4 The out of control average run length from the given process distribution Hodges-Lehmann estimator simulation $\left(ARL_{1HL}^{\delta}\right)$.

Using the procedure in subsection 2.4.4 and 2.4.6, the results (as shown in Table 2), when the shift occurs, the ARL_{1Hl}^{δ} in each scheme also responses to detect the process shift effectively, and for the narrower control limits, the ARL_{1Hl}^{δ} also detect shift faster too.

3.1.5 The original Hodges-Lehmann estimator control charting technique is impractical for the known process distribution.

The in control average run length from subsection 3.1.1, 3.1.2, and 3.1.3 are in different magnitudes, this confirms Gibbons's statement that "in a distribution free inference, whether for testing or estimation, the methods are based on functions of the observation which does not depend on the specific distribution function of the population from which the sample was drawn" [11]. The comparative of ARL_{0W} and ARL_{0HL} for 11 schemes from Table 2 are shown in Figure 2.

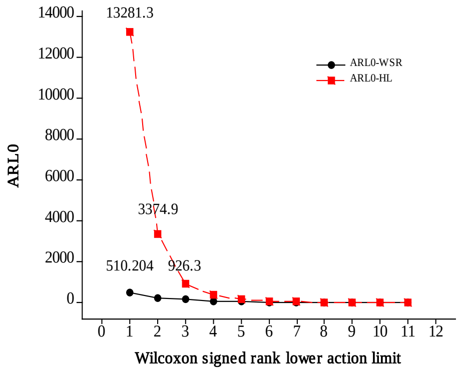


Figure 2: The in control average run length of the Hodges-Lehmann estimator control chart n=10 for 11 control limits

3.2 Performance of the new Hodges-Lehmann estimator control chart for the standard normal distribution.

In the traditional $\overline{X}$ chart, the $\pm 3\sigma$ are the action limits, $\pm 2\sigma$ are the warning limits, the Hodges-Lehmann estimator control limits derive from 100,000 simulated Hodges-Lehmann estimator (from 100,000 sets of Walsh average of sample size n from the standard normal distribution) will be at the location as shown in Table 3.

Table 3. Hodges- Estimator Control Limits to match +/-3 σ and +/-2 σ of the standard normal distribution

	control limits in times of σ			
	-3	-2	2	3
Cumulative probability distribution				
	0.00135	0.01138	0.98863	0.99865
Location for 100,000 values				
	135	1137.5	98862.5	99865
adjusted location				
	135	1138	98862	99865
adjusted cumulative probability				
	0.00135	0.01138	0.98862	0.99865

The Hodges-Lehmann estimator control limits derived from the locations in Table 3 from 100000 runs simulation of the Hodges-Lehmann estimator are shown in Table 4 for n=5 to 20 and Figure 3. From Figure 3, at the closed to +/-3σ and +/-2σ of the standard normal distribution for n=5, 6, … , 20, the control limits also narrower when the sample size is increasing. This narrower control limits show that this implementation has the same effect as the narrower control limits from +/-$\sigma/\sqrt{n}$ when the sample size is increasing in the Shewhart type $\overline{X}$ chart. The new control limits approach will be implemented, if the process distribution is known.

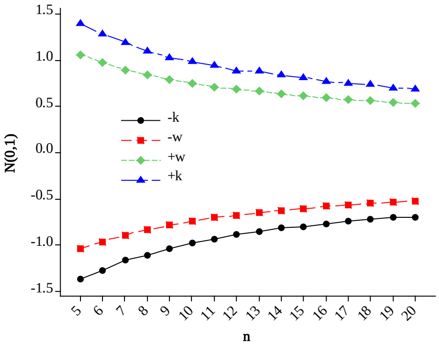


Figure 3 The new control limits of Hodges-Lehmann estimator control charts for the standard normal distribution for n=5,6,7,…,20

Table 4 The new Hodges-Lehmann estimator control limits from 100,000 runs simulation of the standard normal distribution data

n	-k	-w	+w	+k
5	-1.3736	-1.0459	1.0580	1.3906
6	-1.2766	-0.9662	0.9691	1.2742
7	-1.1664	-0.8931	0.8924	1.1915
8	-1.1135	-0.8359	0.8373	1.0911
9	-1.0386	-0.7873	0.7902	1.0245
10	-0.9840	-0.7456	0.7462	0.9831
11	-0.9436	-0.7062	0.7070	0.9426
12	-0.8853	-0.6812	0.6793	0.8779
13	-0.8574	-0.6506	0.6613	0.8754
14	-0.8162	-0.6283	0.6288	0.8342
15	-0.8086	-0.6096	0.6078	0.8114
16	-0.7722	-0.5829	0.5891	0.7663
17	-0.7446	-0.5687	0.5679	0.7476
18	-0.7189	-0.5522	0.5595	0.7326
19	-0.7066	-0.5360	0.5362	0.6979
20	-0.7008	-0.5258	0.5274	0.6878

3.3 The New Hodges-Lehmann control limits with the standard normal distribution

Using the simulation procedures from section 2.5 for evaluating the performance of the new Hodges-Lehmann control limits (at +/-3σ) for the standard normal distribution, the performances of the process shifts in δ times of standard deviation (σ) from 0, 0.25, 0.50, 0.75, 1, 1.5, 2, 2.5, and 3 for the sample size n =10,11,12,13,14,15 are shown in Table 5. The in control average run length ($ARL_{0\widehat{HL}}$), and the out of control average run length ($ARL^{\delta}_{1\widehat{HL}}$) also correspond to the determined control probability ($\alpha = 0.0027$, $ARL_0 = 370$).

3.4 The New Hodges-Lehmann control limits with the standard Weibull distribution

Using the simulation procedures from section 2.5 for evaluating the performance of the new Hodges-Lehmann control limits (at +/-3σ) for the 11 shapes (skewness = 0.1, 0.5, 1, 2, 3, …, 9) of the standard Weibull distribution, the performances of the process shifts from 0, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, and 3, and the sample size n=10,11,12,13,14,15 are shown in Table 6. This new approach shows that the average run length is in the range of the +/-3σ control limits of the standard normal distribution ($\alpha = 0.0027$, $ARL_0 = 370$) due to the robust to the outliers of the Hodges-Lehmann estimator.

Table 5 The Hodges-Lehmann control chart for the standard normal distribution using the locations from 100000 simulated Hodges-Lehmann estimators

n	10	11	12	13	14	15
Hl Location at : +3 σ =99865, -3 σ = 135						
Type I error (α) = 0.0027						
$UCL_{\widehat{HL}}$ at +3 σ	-0.99573	-0.93255	-0.90214	-0.87512	-0.82383	-0.79610
$LCL_{\widehat{HL}}$ at -3 σ	0.99179	0.92749	0.89708	0.86389	0.82901	0.78779
	$ARL_{0\widehat{HL}}$ and $ARL^{\delta}_{1\widehat{HL}}$ when the shift occurs					
shift						
0.00	401.03	345.64	376.66	402.13	362.69	336.38
0.25	84.20	66.80	67.60	61.94	55.49	46.29
0.50	15.03	11.72	10.93	9.65	8.49	7.08
0.75	4.42	3.55	3.21	2.93	2.63	2.22
1.00	1.96	1.70	1.57	1.48	1.36	1.26
1.50	1.06	1.03	1.02	1.01	1.01	1.00
2.00	1.00	1.00	1.00	1.00	1.00	1.00
2.50	1.00	1.00	1.00	1.00	1.00	1.00
3.00	1.00	1.00	1.00	1.00	1.00	1.00

4. Conclusion

The performance of the Hodges-Lehmann estimator control chart is not performed corresponding to the probability distribution of the Wilcoxon signed rank test statistics in control chart hypothesis testing for the known process distribution.

For the quality control process that needs the robust to the outliers test statistic of the Hodges-Lehmann estimator, the new approach to Hodges-Lehmann estimator control charting technique by using the Hodges-Lehmann estimator distribution to construct the control limits through simulation is also an alternative.

The variable parameters and economic design should be carried out for further studies.

Acknowledgements

This research is supported by Faculty of Science, King Mongkut’s University of Technology Thonburi, and Faculty of Oriental Medicine, Rangsit university for work hours allowance to the co-author.

The corresponding author, Assoc. Prof. Adisak Pongpullponsak, as a statistician, he is continually performing the academic research in his specialized fields – Statistical methodology, Nonparametric statistics, Statistical process control, Fuzzy logic, and Stochastic, etc.

Table 6 Average Run length of Hodges-Lehmann estimator control chart of Weibull distribution

	Sample Size (n)						Sample Size (n)					
Shift	10	11	12	13	14	15	10	11	12	13	14	15
	Skewness = 0.1						Skewness = 5.0					
0.00	339.92	325.47	350.62	361.94	355.22	400.66	373.44	391.75	405.67	387.64	370.90	330.25
0.25	76.06	71.13	65.46	63.84	59.96	54.13	245.82	214.58	212.98	183.53	151.23	119.47
0.50	14.26	12.62	11.20	10.28	9.48	8.41	76.91	55.82	51.92	42.17	28.81	20.46
0.75	4.31	3.84	3.36	3.03	2.79	2.52	22.03	13.87	12.26	9.71	5.75	3.83
1.00	1.98	1.78	1.63	1.54	1.43	1.33	6.25	3.65	3.06	2.45	1.46	1.15
1.50	1.07	1.05	1.03	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness = 0.5						Skewness = 6.0					
0.00	417.87	341.24	370.41	370.94	376.66	331.17	349.41	390.56	341.34	317.14	383.20	384.31
0.25	111.94	81.08	80.17	78.89	72.64	61.27	219.96	211.14	193.24	177.03	141.67	135.87
0.50	21.69	15.51	14.77	13.40	11.83	9.82	66.97	54.36	45.74	39.88	24.71	22.03
0.75	6.12	4.56	4.13	3.71	3.32	2.86	18.58	12.71	9.84	8.59	4.30	3.42
1.00	2.49	2.00	1.85	1.68	1.56	1.42	4.82	2.88	2.17	1.90	1.10	1.02
1.50	1.11	1.05	1.04	1.02	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness = 1.0						Skewness = 7.0					
0.00	377.73	370.12	380.32	381.75	343.57	365.87	362.48	397.86	367.71	359.36	366.76	342.21
0.25	122.45	115.83	105.07	94.60	80.28	79.36	260.82	215.47	193.89	199.57	147.55	105.35
0.50	26.94	23.63	19.58	17.17	13.93	13.02	77.39	52.49	43.14	45.55	23.66	14.68
0.75	7.48	6.52	5.41	4.64	3.85	3.54	20.33	10.78	8.37	8.96	3.54	2.06
1.00	2.93	2.53	2.16	1.94	1.69	1.55	4.77	2.15	1.61	1.74	1.00	1.00
1.50	1.14	1.09	1.04	1.03	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness = 2.0						Skewness = 8.0					
0.00	392.51	369.79	391.49	372.65	357.96	396.69	378.36	398.93	418.91	378.54	354.41	388.07
0.25	163.52	138.98	140.11	121.04	113.31	114.31	250.38	215.22	205.26	171.35	115.22	132.41
0.50	42.19	32.88	30.00	24.42	21.64	20.88	73.05	48.37	44.54	36.57	16.51	16.98
0.75	12.22	9.16	8.06	6.54	5.49	5.04	17.50	9.14	7.76	6.31	2.07	1.94
1.00	4.37	3.33	2.86	2.30	2.01	1.84	3.60	1.53	1.29	1.15	1.00	1.00
1.50	1.20	1.08	1.05	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness = 3.0						Skewness = 9.0					
0.00	406.69	373.30	358.82	372.83	376.24	352.02	366.92	360.60	390.12	338.04	358.40	352.90
0.25	243.94	172.30	173.64	151.60	143.08	108.95	234.55	177.35	184.84	161.59	126.85	93.55
0.50	68.80	43.99	40.76	34.11	28.65	20.38	63.33	36.12	36.99	32.50	16.32	10.33
0.75	20.32	12.23	10.43	8.51	6.77	4.67	13.86	5.85	5.65	5.08	1.70	1.11
1.00	6.55	3.91	3.26	2.68	2.12	1.61	2.35	1.00	1.00	1.00	1.00	1.00
1.50	1.26	1.04	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness = 4.0											
0.00	346.17	342.40	396.72	373.40	368.02	357.89						
0.25	201.46	184.23	208.07	174.56	129.00	131.22						
0.50	59.58	48.42	49.63	40.54	25.99	24.25						
0.75	17.85	13.22	12.32	9.81	5.85	4.95						
1.00	5.57	3.84	3.39	2.71	1.70	1.47						
1.50	1.05	1.00	1.00	1.00	1.00	1.00						

Note : For the shift 2.00, 2.50, and 3.00 the average run length of every sample size = 1.00.

References

[1] Montgomery DC. Statistical quality control. 7th ed. Asia: John Wiley & Sons Singapore; 2013.

[2] Amin R, Reynolds Jr MR, Bakir ST. Nonparametric quality control charts based on the sign statistic. Communication in Statistics Theory and Methods. 1995; 24:1579–1623. DOI 10.1080/03610929508831574.

[3] Chakraborti S, van de Wiel MA. A nonparametric control chart based on the Mann-Whitney statistic. IMS Collections. Institute of Mathematical Statistics. 2008; 1:156–172.

[4] Alloway JA, Raghavachari M. (1991) Control chart based on Hodges–Lehmann estimator. Journal of Quality Technology. 1991; 23:336–347. DOI: 10.1214/193940307000000112

[5] Das N. A comparison study of three non-parametric control charts to detect shift in location parameters. The International Journal of Advanced Manufacturing Technology. 2009; 41:799-807.

[6] Neuhäuser M. Nonparametric Statistical tests : A computational Approach. Florida : CRC Press; 2012.

[7] Duchnowski R. Hodges–Lehmann estimates in deformation analyses. Journal of Geodynamics. 2013; 87:873–884. DOI 10.1007/s00190-013-0651-2

[8] Pongpullponsak A, Suracherdkiati W, Panthong C. The economic model of $\overline{X}$ control chart using Shewhart method for skewed distributions. Thailand Statistician Journal of the Thai Statistician Association. 2009; 7(1): 81-89.

[9] Nelson PR. Control Charts for Weibull Processes with Standards Given. IEEE Transactions On Reliability. 1979; 28: 283-288.

[10] Pongpullponsak A, Suracherdkiati W, Kriweradechachai P. The comparison of efficiency of control chart by weighted variance method, Nelson Method, Shewhart method for skewed population, Proceeding of the 5th Applied Statistics Conference of Northern Thailand; 2004 May 27-29; Chiang Mai, Thailand. 2004.

[11] Gibbons JD. Nonparametric Statistical Inference. Tokyo: McGraw-Hill Kogakusha; 1971.

[12] Lehmann EL. Non-parametric confidence intervals for a shift parameter. The Annals of Mathematical Statistics. 1963; 34: 1507–1512. DOI 10.1214/aoms/1177703882

[13] Beaumont GP, Knowles JD. Statistical tests An introduction with MINITAB commentary. Hertfordshire: Prentice Hall international (UK) Limited; 1996.

[14] The Math Works™, MATLAB 7.6.0(R2009a), License Number 350306, February 12, 2009.

[15] MINITAB Thailand. MINITAB 16 Order Number 100004968850, Single License, February 02, 2010.

An empirical likelihood ratio based goodness-of-fit test for some lifetime distributions

Ratchadaporn Meksena[1*] and Supunnee Ungpansattawong[2]
[1]*Department of Statistics, Khon Kaen University, Khon Kaen 40002, Thailand, ratchadaporn.m@kkumail.com*
[2]*Department of Statistics, Khon Kaen University, Khon Kaen 40002, Thailand, supunnee@kku.ac.th*

Abstract

In this paper, an empirical likelihood ratio based goodness-of-fit test for three important lifetime distributions is proposed. The three distributions are a generalized exponential distribution, a Weibull distribution, and a log-normal distribution. For each distribution, the proposed test statistic and three other test statistics are examined, namely the Kolmogorov-Smirnov test statistic, the Cramer-von Mises test statistic and the Anderson-Darling test statistic. Method for parameter estimation in this study is a maximum likelihood. Some investigations on type I error control and power of the test of these statistics are studied by simulation. The selected sample sizes are 10, 25, 50 and 100. The simulation studies show that the proposed test statistics with the optimal values of δ which locate below or around 0.5 can control type I error well. Power studies using several different distributional forms show that the proposed test statistics are competitive when compared with other available test statistics.

Keywords: Empirical likelihood, Likelihood ratio, goodness-of-fit test, generalized exponential distribution, Weibull distribution, log-normal distribution

*Corresponding Author
E-mail Address: ratchadaporn.m@kkumail.com

1. Introduction

Statistical reliability analysis applied in industrial field is concerned with failure probability of the interested (i.e. machines, equipments or products), inference for any lifetime distribution, lifetime testing methods, system reliability, and maintenance and replacement. The concerned data called lifetime data which be the time to occurrence of some event is of interest such as failure or death for individuals in some population. Measured values of lifetime data may be in term of numbers or cycles of failure, or of time length. Various parametric families of models are used in lifetime data analysis. Some important models are including of the exponential, Weibull, log-normal, log-logistic, and gamma distributions. In this paper, we consider three models: generalized exponential, Weibull and log-normal distributions. Due to their usefulness in wide range of situations, it is meaningful to develop a corresponding goodness-of-fit test, which has satisfactory statistical properties. Several goodness-of-fit tests are available in the literature such as Chi-square test, Kolmogorov-Smirnov test, Cramer-von Mises test, Anderson-Darling test etc. Hegacy and Green [4] classified tests of goodness-of-fit into four categories: the likelihood ratio and Pearson tests, tests based on the empirical distribution function (i.e. the Kolmogorov-Smirnov tests, the Kuiper V tests, Pyke's C tests, Brunk's B test, Durbin's D test, the Cramer-von Mises W^2 test, Durbin's M^2 test, Watson's U^2 test, the Anderson-Darling A test, Fisher π and π' tests, the Hartley-Pfaffenberger s^2 test), tests based on sample moments (i.e. the standard third moment $\sqrt{b_1}$ test, the standard fourth moment b_2 test, the Gurland and Dahiya Q test), and tests based upon sample ordered statistics (i.e. Shapiro-Wilk W test, the D'Agostino test, the David el al. u test, and Hegacy and Green's test).

Recently a new goodness-of-fit test based on empirical likelihood ratio (*ELR*) was introduced by Vexler and Gurevich [12] and be available in the literature [5, 13]. The empirical likelihood method has been proposed and developed by Owen [6, 7]. An outline of the empirical likelihood methodology can be presented as follows. Let $X_1, X_2, ..., X_n$ be independently and identically distributed observations, which follow an unknown population distribution F. The empirical likelihood function of F be defined as $L(F)=\sum_{i=1}^{n} p_i$, where the component $p_i, i=1,2,...,n$, maximize the likelihood $L(F)$ and satisfy empirical constraints corresponding to hypotheses of interest. For example, when a population parameter θ identified by $E(X)=\theta$ is of interest, and the true value of θ is θ_0. The null hypothesis is $H_0 : E(X)=\theta_0$. To maximize $L(F)$, the values of component $p_i\, p_i$in $L(F)$ should be chosen given the constraints $p_i \geq 0, \sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i X = \theta_0$, where the constraint $\sum_{i=1}^{n} p_i X = \theta_0$ is an empirical version of $E(X)=\theta_0$. In this case, the likelihood function has the form of $\prod_{i=1}^{n}\{F(X_i)-F(X_i-)\}$ [8].

In this paper, we will follow the similar idea by Vexler and Gurevich [12] to construct an empirical likelihood ratio based goodness-of-fit test for generalized exponential, Weibull and log-normal distributions. We provide the procedure of critical values calculation to obtain the tables of critical values for the proposed test statistics and three other test statistics which mostly used, namely the Kolmogorov-Smirnov test statistic, the Cramer-von Mises test statistic and the Anderson-Darling test statistic. Moreover, we conduct the simulations to investigate the performance of these test statistics in controlling type I error and compare the power of the test among these test statistics.

2. Research Methodology

This section is organized as follows. In Section 2.1, we derive the empirical likelihood ratio based goodness-of-fit test for generalized exponential, Weibull and log-normal distributions. We provide the procedure of calculating the critical values for the empirical likelihood ratio, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling test statistics in Section 2.2. In Section 2.3 and Section 2.4, we investigate type I error control and power of the test of these test statistics, respectively. To obtain the critical values and investigate type I error control and power of the test, we conduct the simulations in R (R version 3.0.2).

We consider the problem of testing the composite hypothesis that $X_1, X_2, ..., X_n$ are distributed as the distribution being tested, hypothesized distribution, with unknown parameters and assume that the density function, under the null hypothesis, say f_{H_0}, is known up to parameters or completely defined, whereas the alternative density function, say f_{H_1}, is completely unknown.

2.1 The Empirical Likelihood Ratio Test for composite hypotheses

Let $X_1, X_2, ..., X_n$ be a random sample from a population with a probability density function *f* and a finite variance. We consider the following hypothesis:

$$H_0 : f = f_{H_0}$$

$$H_1 : f = f_{H_1},$$

where f_{H_1} is completely unknown, whereas f_{H_0} is known up to the vector of parameters $\boldsymbol{\theta}=\left(\theta_1,\theta_2,...,\theta_p\right), p\ge 1$ is a dimension of the vector $\boldsymbol{\theta}$.

In case of f_{H_0} and f_{H_1} are completely known, Neyman-Pearson lemma shows that the likelihood ratio

$$LR=\frac{\prod_{i=1}^{n} f_{H_1}\left(X_i\right)}{\prod_{i=1}^{n} f_{H_0}\left(X_i\right)} \tag{1}$$

is the most powerful test statistic. To approximate the optimal parametric likelihood ratio test, under the null hypothesis, which f_{H_0} is known up to the vector of parameters $\boldsymbol{\theta}=\left(\theta_1,\theta_2,...,\theta_p\right)$, we apply the maximum likelihood method to estimate the unknown vector $\boldsymbol{\theta}$ approximating the likelihood function of the null hypothesis. However, under the alternative hypothesis, the likelihood function can be estimated nonparametrically by applying a maximum empirical likelihood technique [12, 13]. We rewrite the likelihood function of the alternative hypothesis

$$L_{f_{H_1}}=\prod_{i=1}^{n} f_{H_1}\left(X_i\right)=\prod_{i=1}^{n} f_{H_1}\left(X_{(i)}\right)=\prod_{i=1}^{n} f_i,$$

$$f_i = f_{H_1}\left(X_{(i)}\right),$$

where $X_{(1)}\le X_{(2)}\le ...\le X_{(n)}$ are the order statistics based on the observations $X_1, X_2, ..., X_n$. Following the maximum empirical likelihood methodology, we must derive values of $f_i, i=1,2,...,n$ that maximize $L_{f_{H_1}}$ and satisfy the empirical constraints $\int_{-\infty}^{\infty} f\left(u\right)du=1$ corresponding to the alternative hypothesis. The empirical form of this constraint can be obtained by the following lemma by Vexler, Gurevich [12].

Lemma 1. Let *f(x)* be a density function. Then

$$\sum_{i=1}^{n}\int_{X_{(i-m)}}^{X_{(i+m)}} f(x)dx = 2m\int_{X_{(1)}}^{X_{(n)}} f(x)dx-\sum_{k=1}^{m-1}\left(m-k\right)\int_{X_{(n-k)}}^{X_{(n-k+1)}} f(x)dx$$

$$-\sum_{k=1}^{m-1}\left(m-k\right)\int_{X_{(k)}}^{X_{(k+1)}} f(x)dx$$

$$\cong 2m\int_{X_{(1)}}^{X_{(n)}} f(x)dx-\frac{m\left(m-1\right)}{n},$$

where $X_{(i-m)}=X_{(1)}$, if $i-m\le 1$, and $X_{(i+m)}=X_{(n)}$, if $i+m\ge n$.

Proof. See the proof of Proposition 2.1 by Vexler, Gurevich [12] and of Lemma 2.1 by Vexler et al. [13].

Since $\int_{X_{(1)}}^{X_{(n)}} f(x)dx\le\int_{-\infty}^{\infty} f(x)dx=1$ and we denote $\mathrm{T}_m=\frac{1}{2m}\sum_{i=1}^{n}\int_{X_{(i-m)}}^{X_{(i+m)}} f(x)dx\le 1$, using the empirical approximation to the remainder term in Lemma 1, we have

$$\mathrm{T}_m\cong\int_{X_{(1)}}^{X(n)} f(x)dx-\frac{\left(m-1\right)}{2n}\le 1-\frac{\left(m-1\right)}{2n}.$$

From Lemma 1, we can empirically estimate T_m via

$$\hat{\mathrm{T}}_m = \int\limits_{F_n(X_{(1)})}^{F_n(X_{(n)})} dx - \frac{(m-1)}{2n} = 1 - \frac{1}{n} - \frac{(m-1)}{2n}.$$

We observe that $\mathrm{T}_m \to 1$ when $m/n \to 0$ as $m, n \to \infty$. By applying the mean value integration theorem to the term of $\sum_{i=1}^{n}\int_{X_{(i-m)}}^{X_{(i+m)}} f(x)dx$, we have

$$\sum_{i=1}^{n}\int\limits_{X_{(i-m)}}^{X_{(i+m)}} f(x)dx \cong \sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f(X_{(i)})$$

$$= \sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f_i.$$

Thus, the empirical constraint under the alternative hypothesis is given by

$$\mathrm{T}_m = \frac{1}{2m}\sum_{i=1}^{n}\int\limits_{X_{(i-m)}}^{X_{(i+m)}} f(x)dx \cong \frac{1}{2m}\sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f_i$$

$$\triangleq \hat{\mathrm{T}}_m \le 1. \tag{1}$$

Consequently, we apply the Lagrange multiplier method to maximize $\sum_{i=1}^{n}\log f_i$ and satisfy $\hat{\mathrm{T}}_m \le 1$. The Lagrange function defined by

$$\Lambda(f_1, f_2, \ldots, f_n, \lambda_1) = \sum_{i=1}^{n}\log f_i$$

$$+\lambda_1\left(\frac{1}{2m}\sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f_i - 1\right),$$

where λ_1 is a Lagrange multiplier. By taking the derivative of the above equation with respect to each $f_i, i = 1, 2, \ldots, n$, and λ_1 , we obtain

$$\frac{1}{f_i} + \frac{\lambda_1}{2m}\left(X_{(i+m)} - X_{(i-m)}\right) = 0 \tag{2}$$

and

$$\frac{1}{2m}\sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f_i - 1 = 0\,, \tag{3}$$

respectively. From the equation (3), we have

$$f_i = -\frac{2m}{\lambda_1\left(X_{(i+m)} - X_{(i-m)}\right)}.$$

Then multiply equation (2) by f_i and taking summation, we have

$$n + \lambda_1 \frac{1}{2m}\sum_{i=1}^{n}\left(X_{(i+m)} - X_{(i-m)}\right) f_i = 0\,.$$

Since equation (1), we have $\lambda_1 = -n$. Finally, we obtain the estimate value of f_j to maximize $\sum_{i=1}^{n}\log f_i$, which also maximizes $\prod_{i=1}^{n} f_i$ as

$$f_i = \frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)}$$

where $X_{(i-m)} = X_{(1)}$, if $i - m \le 1$, and $X_{(i+m)} = X_{(n)}$, if $i + m \ge n$.

Thus, using the maximum empirical likelihood method, we can construct the empirical likelihood ratio test statistic as

$$ELR_{mn} = \frac{\prod_{i=1}^{n}\frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)}}{\prod_{i=1}^{n} f_{H_0}\left(X_i; \hat{\boldsymbol{\theta}}\right)}\,, \tag{4}$$

where $0 < \delta < 1$ and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of vector of parameters $\boldsymbol{\theta}$.

We notice that the distribution of the test statistic ELR_{mn} strongly depends on the integer *m*. Thus, the optimal values of *m* should be evaluated to make the test more efficient. We follow the same argument by Vexler and Gurevich [12] to reconstruct the test statistic according to the properties of the empirical likelihood method. We adopt their idea here to reconstruct the test statistic in (4) as

$$ELR_n = \frac{\min\limits_{1\le m<n^{\delta}}\prod_{i=1}^{n}\frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)}}{\prod_{i=1}^{n} f_{H_0}\left(X_i; \hat{\boldsymbol{\theta}}\right)}\,. \tag{5}$$

To examine asymptotic properties of the test statistic ELR_n , we denote

$h_i\left(x; \boldsymbol{\theta}\right) = \dfrac{\partial \ln f_{H_0}\left(x; \boldsymbol{\theta}\right)}{\partial \theta_i}, i = 1, 2, \ldots, p$, and assume the following conditions hold:

(1) $E\left(\ln f\left(X_1\right)\right)^2 < \infty$

(2) Under the null hypothesis, we define $\left|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right| = \max\limits_{1\le i\le p}\left|\theta_i - \hat{\theta}_i\right|$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. By a consistency property of the maximum likelihood estimator, we have $\left|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right| \xrightarrow{P} 0$ as $n \to \infty$.

(3) Under the alternative hypothesis, $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$ as $n \to \infty$ where $\boldsymbol{\theta}_0$ is a constant vector with finite components.

(4) There are open intervals $\Theta_0 \subseteq \mathbb{R}^p$ and $\Theta_1 \subseteq \mathbb{R}^p$ containing $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, respectively. There also exists a function $s\left(x\right)$ such that $E\left(s\left(X_1\right)\right) < \infty$ and $\left|h_i\left(x; \eta\right)\right| \le s\left(x\right)$ for all $x \in \mathbb{R}$ and $\eta \in \Theta_0 \cup \Theta_1$.

Proposition 1. Assume the conditions (1) – (4) hold. Then, under the null hypothesis,

$$\frac{1}{n}\ln\left(ELR_n\right) \xrightarrow{P} 0 \text{ as } n \to \infty\,.$$

Proposition 2. Assume the conditions (1) – (4) hold. Then, under the alternative hypothesis,

$$\frac{1}{n}\ln\left(ELR_n\right)\xrightarrow{P} E\ln\left(\frac{f_{H_1}\left(X_1\right)}{f_{H_0}\left(X_1;\boldsymbol{\theta_0}\right)}\right)>0 \text{ as } n\to\infty .$$

Proof. See the proof of Proposition 2.2 by Vexler, Gurevich [12].

Given conditions (1) – (4), Proposition 2 shows that $P_{H_1}\left(\ln\left(ELR_n\right)>C_\alpha\right)\xrightarrow[n\to\infty]{}1$, where C_α is a critical value related to the probability of type I error α. That is, the proposed test statistic is consistent.

2.1.1 The Empirical Likelihood Ratio Test for Generalized Exponential Distributions

We consider the generalized exponential distribution which introduced by Gupta and Kundu [3] in 1999. The generalized exponential distribution has the probability density function defined as:

$$f\left(x;k,\lambda\right)=k\lambda\left(1-e^{-\lambda x}\right)^{k-1}e^{-\lambda x}, \tag{6}$$

where $x>0$, $k>0$ is the shape parameter, and $\lambda>0$ is the scale parameter. We say that the random variable X~ $GE\left(k,\lambda\right)$ if it has the probability density function (6).

Suppose that the data consist of independent observations $X_1,X_2,...,X_n$. Consider the following hypothesis:

$$H_0: f = f_{H_0} \sim GE\left(k,\lambda\right)$$
$$H_1: f = f_{H_1} \nsim GE\left(k,\lambda\right),$$

In accordance with the technique mentioned in Section 2.1, we can construct the empirical likelihood ratio based goodness-of-fit test for generalized exponential distribution as

$$GE_n=\frac{\min\limits_{1\le m<n^\delta}\prod\limits_{i=1}^{n}\frac{2m}{n\left(X_{(i+m)}-X_{(i-m)}\right)}}{\hat{k}^n\,\hat{\lambda}^n\left[\prod\limits_{i=1}^{n}\left(1-e^{-\hat{\lambda}x_i}\right)\right]^{\hat{k}-1}e^{-\hat{\lambda}\sum\limits_{i=1}^{n}x_i}}, \tag{7}$$

where $0<\delta<1$, $X_{(i-m)}=X_{(1)}$, if $i-m\le 1$, $X_{(i+m)}=X_{(n)}$, if $i+m\ge n$, and $\hat{k},\hat{\lambda}$ are the maximum likelihood estimators of the parameters k,λ, respectively.

2.1.2 The Empirical Likelihood Ratio Test for Weibull Distributions

We consider the Weibull distribution which has the probability density function defined as:

$$f\left(x;k,\lambda\right)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^k}, \tag{8}$$

where $x>0$, $k>0$ is the shape parameter and, $\lambda>0$ is the scale parameter. We say that the random variable X~ $WB\left(k,\lambda\right)$ if it has the probability density function (8).

Suppose that the data consist of independent observations $X_1,X_2,...,X_n$. Consider the following hypothesis:

$$H_0: f = f_{H_0} \sim WB\left(k,\lambda\right)$$
$$H_1: f = f_{H_1} \nsim WB\left(k,\lambda\right),$$

In accordance with the technique mentioned in Section 2.1, we can construct the empirical likelihood ratio based goodness-of-fit test for Weibull distribution as

$$WB_n=\frac{\min\limits_{1\le m<n^\delta}\prod\limits_{i=1}^{n}\frac{2m}{n\left(X_{(i+m)}-X_{(i-m)}\right)}}{\frac{\hat{k}^n}{\hat{\lambda}^n}\left[\prod\limits_{i=1}^{n}\left(\frac{x_i}{\hat{\lambda}}\right)\right]^{\hat{k}-1}e^{-\sum\limits_{i=1}^{n}\left(\frac{x_i}{\hat{\lambda}}\right)^{\hat{k}}}}, \tag{9}$$

where $0<\delta<1$, $X_{(i-m)}=X_{(1)}$, if $i-m\le 1$, $X_{(i+m)}=X_{(n)}$, if $i+m\ge n$, and $\hat{k},\hat{\lambda}$ are the maximum likelihood estimators of the parameters k,λ, respectively.

2.1.3 The Empirical Likelihood Ratio Test for Log-Normal Distributions

We consider the log-normal distribution which has the probability density function defined as:

$$f\left(x;\mu,\sigma\right)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}}, \tag{10}$$

where $x>0$, $\mu\in\mathbb{R}$ is the mean of the data set after transformation by taking the logarithm, and $\sigma>0$ is the standard deviation of the data set after transformation. We say that the random variable X~ $LN\left(\mu,\sigma\right)$ if it has the probability density function (10).

Suppose that the data consist of independent observations $X_1,X_2,...,X_n$. Consider the following hypothesis:

$$H_0: f = f_{H_0} \sim LN\left(\mu,\sigma\right)$$
$$H_1: f = f_{H_1} \nsim LN\left(\mu,\sigma\right),$$

In accordance with the technique mentioned in Section 2.1, we can construct the empirical likelihood ratio based goodness-of-fit test for log-normal distribution as

$$LN_n=\frac{\min\limits_{1\le m<n^\delta}\prod\limits_{i=1}^{n}\frac{2m}{n\left(X_{(i+m)}-X_{(i-m)}\right)}}{\prod\limits_{i=1}^{n}\left[\frac{1}{x_i\sqrt{2\pi\hat{\sigma}^2}}e^{-\frac{\left(\ln x_i-\hat{\mu}\right)^2}{2\hat{\sigma}^2}}\right]}, \tag{11}$$

where $0<\delta<1$, $X_{(i-m)}=X_{(1)}$, if $i-m\le 1$, $X_{(i+m)}=X_{(n)}$, if $i+m\ge n$, and $\hat{\mu},\hat{\sigma}$ are the maximum likelihood estimators of the parameters μ,σ, respectively.

2.2 Calculation of Critical Values

The aim of this section is to obtain tables of goodness-of-fit critical values for the empirical likelihood ratio (*ELR*) test statistic, the Kolmogorov-Smirnov (*KS*) test statistic, the Cramer-von Mises (*CM*) test statistic, and the Anderson-Darling (*AD*) test statistic. We calculate the critical values for fixed sample sizes n = 10, 25, 50, 100 related to the probability of type I error = 0.1, 0.05 and 0.01 using a simulation. The following steps are used in calculating the critical values:

Step 1: Generate a random sample $X_1, X_2, ..., X_n$ from the hypothesized distribution with specified parameters as Table 1.

Step 2: Estimate the unknown parameters based on the generated random sample by method of maximum likelihood.

Step 3: The resulting maximum likelihood estimators of the unknown parameters under each case are used to calculate the test statistic for the given values of sample size *n*, as follows:

1. The *ELR* test statistics are ln(GE_{n}), ln(WB_{n}), and ln(LN_{n}) for a generalized exponential, a log-normal and a Weibull distribution tests, respectively.
2. The *KS* test statistic [2] is

$$KS = \max_{1\le i\le n}\left\{D^+, D^-\right\},$$

where $D^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F_0\left(x_{(i)}\right)\right\}$,

$$D^- = \max_{1\le i\le n}\left\{F_0\left(x_{(i)}\right) - \frac{(i-1)}{n}\right\}$$

and $F_0\left(x_{(i)}\right)$ is a cumulative distribution function for the distribution being tested.

3. The *CM* test statistic [1] is

$$CM = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F_0\left(x_{(i)}\right)\right]^2.$$

4. The *AD* test statistic is

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)\ln\left(F_0\left(x_{(i)}\right)\right)$$
$$+\left(2(n-i)+1\right)\ln\left(1-F_0\left(x_{(i)}\right)\right)].$$

This study uses the following modified *AD* test statistics given by D’ Agostino and Stephens which used in the research by Promdan [9] and Raibankoh [10],

$AD^* = AD\left(1+\frac{0.75}{n}+\frac{2.25}{n^2}\right)$ for testing generalized exponential or log-normal distribution and

$AD^* = AD\left(1+\frac{0.2}{\sqrt{n}}\right)$ for testing Weibull distribution.

Step 4: This procedure is repeat 5,000 times, thus we obtain 5,000 independent values of the test statistic.

Then, rank them and choose 90th, 95th and 99th percentiles to be the critical values corresponding to the significance levels α = 0.1, 0.05 and 0.01, respectively.

Table 1: Parameters of the hypothesized distribution

Hypothesized distribution	Case	Parameters
Generalized exponential distribution (*GE*)	I	k = 1 , λ , = 1
	II	k = 2 , λ , = 1
Weibull distribution (*WB*)	I	k = 1 , λ , = 1
	II	k = 2 , λ , = 1
Log-Normal distribution (*LN*)	I	μ = 0 , σ , = 1
	II	μ = 0 , σ , = 1.5

2.3 Investigation of Type I Error Control

In this section, we investigate the performance of all test statistics in controlling the type I error with the significance levels α = 0.1, 0.05 and 0.01. For *ELR* test statistic, we try the test with different values of δ = 0.2, 0.4, 0.5, 0.6, 0.8 in order to select the appropriated test statistics before investigate the power property. We conduct simulations 5,000 times under the specified hypothesized distribution with sample sizes n = 10, 25, 50, 100. For each sample, we calculate the test statistic and compare to its respective critical values and count the number of rejecting null hypothesis. The percentage of the rejections will be the probability of a type I error.

2.4 Power Comparison

To investigate the power of the proposed test with the given nominal level α = 0.05, a power comparison is made among the selected *ELR* test statistics, the *KS* test statistic, the *CM* test statistic, and the *AD* test statistic for testing generalized exponential, Weibull and log-normal distributions. For each hypothesized distribution, the power of the test is determined by generating 5,000 random samples of sizes n = 10, 25, 50, 100 from the six alternatives as shown in Table 2 for each test statistic. Then, we calculate the test statistic and compare to its respective critical values and count the number of rejecting null hypothesis. The percentage of the rejections will be the power of the test.

Table 2: The alternatives for each hypothesized distribution

No.	Hypothesized distributions		
	GE	*WB*	*LN*
1	*WB*(1, 1)	*GE*(1, 1)	*GE*(1, 1)
2	*WB*(2, 1)	*GE*(2, 1)	*GE*(2, 1)
3	*LN*(0, 1)	*LN*(0, 1)	*WB*(1, 1)
4	*LN*(0, 1.5)	*LN*(0, 1.5)	*WB*(2, 1)
5	*Gamma*(1, 2)	*Gamma*(1, 2)	*Gamma*(1, 2)
6	*Gamma*(5, 1)	*Gamma*(5, 1)	*Gamma*(5, 1)

3. Research Results and Discussion

3.1 Critical Values

Some results of the simulation to create the tables of critical values for the test statistics are listed in Table 3.

3.2 Type I Error Control

The simulation study demonstrate the type I error of the proposed tests statistics with δ = 0.2, 0.4, some situations of δ = 0.5, the *KS* test statistic, the *CM* test statistic, and the *AD* test statistic are well controlled for the given significance levels. A selection of the simulated results of investigating type I error of the test statistics are listed in Table 4. Therefore, *ELR* tests with δ = 0.2, 0.4, 0.5 can be used for further power investigation.

3.3 Power Comparison

Based on the simulated results, the power of the test statistics increase as the sample size increase. We observe that in case of a hypothesized distribution is a generalized exponential distribution or a Weibull distribution, in most cases, the *ELR* test statistics with δ = 0.5 perform better than the other test statistics, except in some cases such as in case of the alternative is log-normal distribution. However, in case of a hypothesized distribution is a log-normal distribution, we observe that the proposed test statistics are the most powerful goodness-of-fit test among the competitors in most cases, except a little worse in some cases. Some power comparison results with generalized exponential distribution, Weibull distribution and log-normal distribution are displayed in Table 5, Table 6 and Table 7, respectively**.**

4. Conclusion

In this paper, we propose an empirical likelihood ratio based goodness-of-fit test which is a nonparametric approximation to the traditional likelihood ratio test for a generalized exponential, a Weibull and a log-normal distribution. The tables of goodness-of-fit critical values for all test statistics are created. Simulations indicate that the proposed test statistics with the optimal value of δ which locate below or around 0.5 can control type I error well. Power studies using several different distributional forms show that the proposed test statistic is competitive when compared with other available test statistics.

Table 3: Critical values of the tests at the significance level *α* for testing *WB*(1, 1)

α	Sample	Test statistics							
	Size *n*	ln($WB_{n,\ \delta=0.2}$)	ln($WB_{n,\ \delta=0.4}$)	ln($WB_{n,\ \delta=0.5}$)	ln($WB_{n,\ \delta=0.6}$)	ln($WB_{n,\ \delta=0.8}$)	*KS*	*CM*	*AD*
0.1	10	7.0948	5.0371	4.5025	4.5225	4.1742	0.2417	0.0982	0.6488
	25	12.0922	6.7986	6.3653	6.0905	5.8206	0.1584	0.1014	0.6527
	50	11.9222	8.6591	7.7837	7.5596	7.2982	0.1127	0.1011	0.6498
	100	19.5724	10.6342	9.8000	9.5441	9.1244	0.0800	0.1024	0.6477
0.05	10	8.2084	5.8442	5.1181	5.0734	4.5899	0.2628	0.1180	0.7656
	25	13.5274	7.7299	7.2119	6.8576	6.5139	0.1719	0.1231	0.7725
	50	13.2022	9.6266	8.7097	8.4405	8.2618	0.1222	0.1225	0.7659
	100	21.0363	11.7146	10.8839	10.6440	10.2616	0.0870	0.1229	0.7761
0.01	10	10.5462	7.4251	6.2870	6.2244	5.5049	0.3017	0.1671	1.0452
	25	16.5120	9.4933	8.7996	8.7077	8.0434	0.2000	0.1736	1.0373
	50	15.6031	11.6574	10.6337	10.4259	10.0895	0.1408	0.1764	1.0197
	100	24.3823	14.0801	13.0574	12.8146	12.6640	0.1003	0.1777	1.0618

Table 4: Type I error of the tests at the significance level *α* with *WB*(1, 1)

α	Sample	Test statistics							
	size *n*	ln($WB_{n,\ \delta=0.2}$)	ln($WB_{n,\ \delta=0.4}$)	ln($WB_{n,\ \delta=0.5}$)	ln($WB_{n,\ \delta=0.6}$)	ln($WB_{n,\ \delta=0.8}$)	*KS*	*CM*	*AD*
0.1	10	0.0960	0.0968	**0.1256**	**0.1220**	**0.4508**	0.0954	0.1068	0.1022
	25	0.1010	0.0926	0.1074	**0.1646**	**0.4394**	0.0958	0.1018	0.1042
	50	0.1044	0.0974	**0.1272**	**0.2062**	**0.3830**	0.0952	0.1008	0.0998
	100	0.0996	0.1068	0.1198	**0.1762**	**0.2958**	0.1106	0.1026	0.1018
0.05	10	0.0486	0.0508	0.0536	0.0584	**0.3434**	0.0474	0.0556	0.0492
	25	0.0502	0.0472	0.0538	**0.0804**	**0.3570**	0.0474	0.0506	0.0500
	50	0.0488	0.0506	**0.0664**	**0.1140**	**0.3072**	0.0476	0.0512	0.0502
	100	0.0548	0.0544	**0.0640**	**0.1042**	**0.2470**	0.0584	0.0542	0.0518
0.01	10	0.0102	0.0102	0.0132	0.0144	**0.1298**	0.0088	0.0100	0.0100
	25	0.0102	0.0096	0.0114	0.0086	**0.1832**	0.0104	0.0102	0.0138
	50	0.0106	0.0100	0.0108	**0.0228**	**0.1834**	0.0100	0.0088	0.0124
	100	0.0096	0.0114	0.0144	**0.0312**	**0.1494**	0.0144	0.0124	0.0122

Table 5: Power comparison with *GE*(1, 1), α = 0.05

	Alternative: *WB*(1, 1)					Alternative: *WB*(2, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($GE_{n,\ \delta=0.2}$)	0.0522	0.0530	0.0472	0.0460	ln($GE_{n,\ \delta=0.2}$)	0.0688	0.0916	0.1748	0.2346
ln($GE_{n,\ \delta=0.4}$)	0.0470	0.0510	0.0470	0.0444	ln($GE_{n,\ \delta=0.4}$)	0.1094	0.1816	0.2778	0.4536
ln($GE_{n,\ \delta=0.5}$)	**0.0684**	**0.0662**	**0.0704**	**0.0570**	ln($GE_{n,\ \delta=0.5}$)	**0.2370**	**0.2858**	**0.4804**	**0.5830**
KS	0.0460	0.0456	0.0504	0.0484	*KS*	0.0542	0.0838	0.1504	0.2976
CM	0.0412	0.0474	0.0518	0.0522	*CM*	0.0578	0.0948	0.1856	0.3984
AD	0.0450	0.0508	0.0512	0.0470	*AD*	0.0610	0.1102	0.2216	0.4470
	Alternative: *LN*(0, 1)					Alternative: *LN*(0, 1.5)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($GE_{n,\ \delta=0.2}$)	0.0794	0.1412	0.3306	0.5078	ln($GE_{n,\ \delta=0.2}$)	0.0856	0.2540	0.5962	0.8530
ln($GE_{n,\ \delta=0.4}$)	0.0884	0.1982	0.3670	0.6094	ln($GE_{n,\ \delta=0.4}$)	0.0672	0.2656	0.5956	0.8902
ln($GE_{n,\ \delta=0.5}$)	0.0986	0.2036	0.3124	0.5418	ln($GE_{n,\ \delta=0.5}$)	0.0382	0.2018	0.3552	0.7920
KS	0.1116	0.2142	0.4160	0.6776	*KS*	0.1812	0.4450	0.7510	0.9600
CM	0.1238	0.2512	0.4874	0.7952	*CM*	**0.2076**	0.5250	0.8246	0.9854
AD	**0.1304**	**0.2712**	**0.5224**	**0.8126**	*AD*	0.2052	**0.5346**	**0.8376**	**0.9876**
	Alternative: *Gamma*(1, 2)					Alternative: *Gamma*(5, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($GE_{n,\ \delta=0.2}$)	0.0558	0.0506	0.0484	0.0452	ln($GE_{n,\ \delta=0.2}$)	0.0590	0.0616	0.0880	0.0790
ln($GE_{n,\ \delta=0.4}$)	0.0500	0.0546	0.0492	0.0416	ln($GE_{n,\ \delta=0.4}$)	0.0920	0.1124	0.1268	0.1502
ln($GE_{n,\ \delta=0.5}$)	**0.0666**	**0.0676**	**0.0678**	**0.0528**	ln($GE_{n,\ \delta=0.5}$)	**0.2034**	**0.1812**	**0.2466**	**0.2308**
KS	0.0486	0.0454	0.0476	0.0456	*KS*	0.0328	0.0378	0.0510	0.0650
CM	0.0490	0.0452	0.0430	0.0480	*CM*	0.0342	0.0384	0.0576	0.0860
AD	0.0500	0.0478	0.0442	0.0438	*AD*	0.0398	0.0502	0.0696	0.1000

Table 6: Power comparison with *WB*(1, 1), α = 0.05

	Alternative: *GE*(1, 1)					Alternative: *GE*(2, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($WB_{n,\ \delta=0.2}$)	0.0492	0.0524	0.0500	0.0492	ln($WB_{n,\ \delta=0.2}$)	0.0574	0.0654	0.0800	0.1056
ln($WB_{n,\ \delta=0.4}$)	0.0476	0.0472	0.0456	0.0506	ln($WB_{n,\ \delta=0.4}$)	0.0750	0.0876	0.1230	0.1700
ln($WB_{n,\ \delta=0.5}$)	**0.0548**	**0.0526**	**0.0578**	**0.0588**	ln($WB_{n,\ \delta=0.5}$)	**0.1126**	**0.1120**	**0.1962**	**0.2068**
KS	0.0452	0.0514	0.0566	0.0566	*KS*	0.0446	0.0610	0.0906	0.1346
CM	0.0506	0.0526	0.0478	0.0462	*CM*	0.0566	0.0742	0.1050	0.1602
AD	0.0472	0.0512	0.0496	0.0442	*AD*	0.0494	0.0740	0.1096	0.1738
	Alternative: *LN*(0, 1)					Alternative: *LN*(0, 1.5)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($WB_{n,\ \delta=0.2}$)	0.0730	0.1514	0.3528	0.6084	ln($WB_{n,\ \delta=0.2}$)	0.0650	0.1310	0.3096	0.5792
ln($WB_{n,\ \delta=0.4}$)	0.0912	0.1984	0.4157	0.7372	ln($WB_{n,\ \delta=0.4}$)	0.0434	0.0882	0.2542	0.5308
ln($WB_{n,\ \delta=0.5}$)	0.0997	0.1972	0.3478	0.6468	ln($WB_{n,\ \delta=0.5}$)	0.0268	0.0612	0.0861	0.2742
KS	0.0897	0.1930	0.3780	0.6762	*KS*	0.0835	0.1848	0.3767	0.6734
CM	**0.1155**	0.2630	0.4951	0.8182	*CM*	**0.1117**	0.2566	0.4929	0.8112
AD	0.1019	**0.2856**	**0.5614**	**0.8764**	*AD*	0.0983	**0.2816**	**0.5567**	**0.8730**
	Alternative: *Gamma*(1, 2)					Alternative: *Gamma*(5, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($WB_{n,\ \delta=0.2}$)	0.0464	**0.0546**	0.0454	0.0518	ln($WB_{n,\ \delta=0.2}$)	0.0696	0.0790	0.1336	0.2028
ln($WB_{n,\ \delta=0.4}$)	0.0408	0.0462	0.0480	0.0568	ln($WB_{n,\ \delta=0.4}$)	0.1038	0.1378	0.2026	0.3550
ln($WB_{n,\ \delta=0.5}$)	0.0454	0.0508	**0.0638**	**0.0626**	ln($WB_{n,\ \delta=0.5}$)	**0.2078**	**0.1988**	**0.3478**	**0.4254**
KS	0.0464	0.0442	0.0490	0.0544	*KS*	0.0576	0.0770	0.1400	0.2552
CM	**0.0544**	0.0512	0.0494	0.0516	*CM*	0.0760	0.0988	0.1744	0.3264
AD	0.0530	0.0534	0.0510	0.0498	*AD*	0.0678	0.1026	0.1976	0.3764

Table 7: Power comparison with *LN*(0, 1.5), α = 0.05

	Alternative: *GE*(1, 1)					Alternative: *GE*(2, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($LN_{n,\ \delta=0.2}$)	0.0872	0.1546	0.4208	0.6784	ln($LN_{n,\ \delta=0.2}$)	0.0834	0.0910	0.2172	0.3344
ln($LN_{n,\ \delta=0.4}$)	0.1412	0.3546	0.6108	0.8834	ln($LN_{n,\ \delta=0.4}$)	0.1530	0.2608	0.3980	0.6342
ln($LN_{n,\ \delta=0.5}$)	**0.2460**	**0.4350**	**0.7426**	**0.9124**	ln($LN_{n,\ \delta=0.5}$)	**0.3132**	**0.3850**	**0.6422**	**0.7618**
KS	0.1120	0.2390	0.4310	0.7272	*KS*	0.0700	0.1126	0.2036	0.3774
CM	0.1304	0.3018	0.5388	0.8472	*CM*	0.0784	0.1440	0.2556	0.4940
AD	0.1542	0.3412	0.5944	0.8910	*AD*	0.0936	0.1664	0.2900	0.5536
	Alternative: *WB*(1, 1)					Alternative: *WB*(2, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($LN_{n,\ \delta=0.2}$)	0.0878	0.1560	0.4208	0.6948	ln($LN_{n,\ \delta=0.2}$)	0.1090	0.1692	0.4830	0.7122
ln($LN_{n,\ \delta=0.4}$)	0.1426	0.3508	0.6060	0.8812	ln($LN_{n,\ \delta=0.4}$)	0.2270	0.4798	0.7348	0.9378
ln($LN_{n,\ \delta=0.5}$)	**0.2562**	**0.4406**	**0.7348**	**0.9164**	ln($LN_{n,\ \delta=0.5}$)	**0.4798**	**0.6432**	**0.9170**	**0.9776**
KS	0.1102	0.2366	0.4254	0.7238	*KS*	0.1084	0.2230	0.4354	0.7330
CM	0.1348	0.3026	0.5492	0.8502	*CM*	0.1318	0.2888	0.5446	0.8574
AD	0.1574	0.3412	0.5980	0.8958	*AD*	0.1554	0.3304	0.5998	0.8948
	Alternative: *Gamma*(1, 2)					Alternative: *Gamma*(5, 1)			
Sample size *n*	10	25	50	100	Sample size *n*	10	25	50	100
ln($LN_{n,\ \delta=0.2}$)	0.0908	0.1556	0.4332	0.6946	ln($LN_{n,\ \delta=0.2}$)	0.0820	0.0842	0.1440	0.1836
ln($LN_{n,\ \delta=0.4}$)	0.1388	0.3568	0.6278	0.8910	ln($LN_{n,\ \delta=0.4}$)	0.1522	0.2358	0.3044	0.4442
ln($LN_{n,\ \delta=0.5}$)	**0.2518**	**0.4430**	**0.7496**	**0.9202**	ln($LN_{n,\ \delta=0.5}$)	**0.3752**	**0.3884**	**0.6330**	**0.6676**
KS	0.1082	0.2334	0.4382	0.7320	*KS*	0.0556	0.0834	0.1220	0.1986
CM	0.1300	0.3060	0.5572	0.8518	*CM*	0.0614	0.0882	0.1446	0.2522
AD	0.1532	0.3480	0.6128	0.8958	*AD*	0.0764	0.1044	0.1562	0.2828

Acknowledgements

This work was partially supported by Department of Statistics, Faculty of Science, Khon Kaen University.

References

[1] Anderson TW. On the distribution of the two-sample Cramer-von Mises criterion. The Annals of Mathematical Statistics. 1962; 33(3): 1148-1159.

[2] Chakravarti IM, Laha RG, Roy J. Handbook of Methods of Applied Statistics. 1st ed. New York: John Wiley & Sons; 1967.

[3] Gupta RD, Kundu D. Generalized exponential distributions. Australian and New Zealand Journal of Statistics. 1999; 41(2): 173–188.

[4] Hegazy YAS, Green JR. Some new goodness-of-fit tests using order. Applied Statistics. 1975; 24(3): 299–308.

[5] Ning W, Ngunkeng G. An empirical likelihood ratio based goodness-of-fit test for skew normality. Statistical Methods & Applications. 2013; 22(2): 209–226.

[6] Owen AB. Empirical likelihood ratio confidence intervals for a single functional. Biometrika. 1988; 75(2): 237–249.

[7] Owen AB. Empirical likelihood ratio confidence regions. The Annals of Statistics. 1990; 18(1): 90–120.

[8] Owen AB. Empirical likelihood. 1st ed. New York: Chapman and Hall/CRC; 2001.

[9] Promdan P. Some goodness of fit tests for generalized exponential distribution [Dissertation]. Bangkok: Kasetsart Univ; 2009.

[10] Raibankoh S. Comparison of power of some goodness of fit test for testing lognormal and Weibull distribution [Dissertation]. Bangkok: King Mongkut’t Institue of technology North Bangkok; 2006.

[11] Stephens MA. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. 1974; 69(347): 730–737.

[12] Vexler A, Gurevich G. Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. Computational Statistics and Data Analysis. 2010; 54(2): 531–545.

[13] Vexler A, Kim S, Tsai W, Tian L, Hutson AD. An empirical likelihood ratio based goodness-of-fit test for inverse Gaussian distributions. Journal of Statistical Planning and Inference. 2011; 141(6): 2128–2140.

Modeling for extreme temperature of central northeast region of Thailand

Benjawan Charin[1*], Wuttichai Srisodaphol[2] and Piyapatr Busababodhin[3]
[1] Department of Statistics, Faculty of Science, Khon Kaen University, Thailand, cbenjawan@kkumail.com.
[2]Department of Statistics, Faculty of Science, Khon Kaen University, Thailand, wuttsr@kku.ac.th.
[3]Department of Mathematics, Faculty of Science, Mahasarakham University, Thailand. piyapatr99@gmail.com

Abstract

The aim of this study is to model monthly extreme temperature in central northeast of Thailand by using the Generalized Extreme Value distribution (GEV) and Generalized Pareto distribution (GPD). Time series data that used is the maximum monthly temperature during 1977 to 2013 from 6 meteorological stations which are set in the central northeast of Thailand. These data were obtained from the Meteorological Department of Thailand. An "extreme" package in R program is provided to directly model by using the GEV and GPD with stationary and non-stationary process. The parameter estimation of the extreme value theory is based on Maximum Likelihood Estimation (MLE). The criterion of model selection is Akaikes Information Criterion (AIC). Results of the study found that the estimates of parameters in GEV and GPD of Khon Kean station are stationary. Otherwise, 5 of meteorological stations, the estimates of parameters in GEV and GPD are non-stationary. Furthermore, the Weibull and Gamma distributions are the best model for GEV and GPD, respectively.

Keywords: Maximum temperature, generalized extreme value distribution, generalized pareto distribution, modeling of extreme temperature

*Corresponding Author
E-mail Address: cbenjawan@kkumail.com

1. Introduction

Since 1861, the global temperature was increasing as 0.6 ± 0.2 degrees Celsius which effect to the world. The losses of increasing economic and also human life were due to this situation [3]. So, the research on extreme temperatures and their variation have increased.

In Thailand, the temperature has changed since recorded data by period of the years 1951 to 2007. The temperature is increasing trend and also the average temperature, average maximum temperature, and average minimum temperature [6]. In 2012, the Meteorology Department of Thailand studied about the overall climate and classified it as a region. They found that the average annual temperature in Thailand was higher than normal value (normal value of temperature is 27 degrees Celsius). Hence, for classification of a region, they found that the climate in Northeast was higher than the normal value [10]. Especially in the central area of northeast which was consisted four provinces; Kalasin, Khon Kean, Roi-Et and Mahasarakham had temperatures above normal value because the most drought and the affected from the southwest monsoon cannot reach. However, the drought is mainly related to the temperature because the drought had been caused by climate warming in summer than usual temperature or the impact of the greenhouse phenomenon, etc [8].

Furthermore, if we want to know the probability of the incident that the extreme value is on the tail such as the highest-lowest daily temperatures, the highest-lowest monthly rainfall, the maximum monthly wind speed and so on. In order to find the ways to prevent and resolve situations such as drought, floods, storm, and earthquake, Statistical tool that will play a big role on this favor is the extreme value theory.

In 2007, Nadarajah and Choi [9] proposed the model for annual maxima of daily rainfall for five locations of South Korea that is Seoul, Gangneung, Busan, Gwangju and Chupungryong, during the years 1961 to 2001. They used the GEV distribution to fit the data from each location and then they described the extremes of rainfall and predicted its future behavior. The study found that Gumbel distribution provided the most reasonable model for four of five locations. The estimation of return levels for 10, 50, 100, 1000, 5000, 10,000, 50,000 and 100,000 year were described depending on locations.

In 2009, Unkasevic and Tosic [11] studied the fluctuation in extremes daily winter and summer temperature in Belgrade. They used GEV distribution and GPD distribution to fit the absolute minimum winter (AMINW) temperatures, the maximum summer (AMAXS) temperatures, the daily minimum winter (DMINW) temperatures and the maximum summer (DMAXS) temperatures. They found that both of distributions were reasonable. Furthermore, they also fitted the trend and the North Atlantic Oscillations index as covariates in the location parameter when applying the GEV to fit the AMINW temperatures. The results showed that it was significantly improvement over the model without covariates. Finally, they estimated the return levels for 100-year and 10-year of return periods by using the GEV distribution and GPD distribution for summer and winter seasons.

In 2011, Xu *et. al* [12] presented extreme value theory for fitting disaster area-based drought losses. They found that GPD model was built based on the extreme value distribution of these data. The results showed that the extreme value distribution fitted with the actual data and it could be significantly improve forecast accuracy.

In 2013, Khongthip *et. al* [5] studied the model of extreme rainfall data in upper northern region of Thailand by using the GEV distribution and estimated the return levels for various return periods during 1957 to 2009 from twenty-six stations which covered this area. They also provided an R program which was able to directly model a data for each station. The results showed that only the 17th station at Chiengkong district of Chiengrai province had GEV distribution. The location parameter was changed depending on linear trend. Only two stations had GEV distribution when the location parameter was changed depending on quadratic trend, that is, the 14th station at Lee district of Lampoon province and the 20th station at Muang district of Chiengrai province, respectively. The remainder of stations had GEV distribution with stationary of location parameter. Since the 23rd station at Masai district of Chiengrai had the highest return level for various return periods, so it should be the first consideration station in preventing or reducing the severity of floods. By the way, the 13rd station at Matha district of Lampoon province which had the smallest return level for various return periods, it should be the last consideration.

In 2014, Bootchamruei *et. al* [1] studied the model of extreme rainfall data in central northeast region of Thailand by using the GEV distribution. They estimated the return level for various return periods during the years 1957 to 2009 from six stations which covered this area. They found that extreme rainfall data of Roi-et station at Muang district of Roi-et province is the form of parameters μ, σ and ξ are constant and the rest are the form of parameters μ changed depending on linear trend, parameter σ changed depending on exponential trend and ξ is constant. Since the Kamalasai station at Kamalasai district of Kalasin province has the highest return level from various return periods, so it should be the first consideration station in preventing or reducing the severity of floods. By the way, the Khon Kean station at Muang district of Khon Kean province which has the smallest return level for various return periods should be the last consideration.

In this study, we propose the models of maximum temperature of central northeast of Thailand by using the GEV distribution and GPD distribution. In section 2, the background of extreme value theory is reviewed. The data and steps of the analysis are described in Section 3. In section 4, the results of GEV distribution and GPD distribution are shown. Finally, the conclusions and discussion are shown.

2. Background

2.1 The generalized extreme value distribution (GEV)

Let $X_1, X_2, ..., X_n$ be a sequence of independent variables with common distribution function F, the maximum value of random variable $X_1, X_2, ..., X_n$ is $X_{(n)} = \max\left(X_1, X_2, ..., X_n\right)$. The cumulative distribution function (cdf.) of the GEV distribution is [2]

$$F\left(x\right) = \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right\},$$

and its probability density function (pdf.) is as follow.

$$f\left(x\right) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{(-1/\xi)-1}\exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}$$

where $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

The GEV distribution which has three parameters, i.e. [4],

(1) Location parameter, denote by μ, specifies the center of the GEV distribution,

(2) Scale parameter, denote by σ, determines the size of deviations of μ, and

(3) Shape parameter which denoted by ξ, shows how rapidly the upper tail decays.

The representation of $F\left(x\right)$ is combined single model which can lead to three types of non-degenerate distribution function families, i.e., [2]

Type I, Gumbel family which corresponds to case $\xi = 0$ i.e., GEV family with limits as $\xi \to 0$:

$$F\left(x\right) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < x < \infty,$$

Type II, Fréchet family which corresponds to case $\xi > 0$ of GEV family:

$$F\left(x\right) = \begin{cases} 0, & x \le \mu \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}, & x > \mu, \end{cases}$$

Type III, Weibull family which corresponds to case $\xi < 0$ of GEV family:

$$F\left(x\right) = \begin{cases} \exp\left\{-\left(-\frac{x-\mu}{\sigma}\right)^{\alpha}\right\}, & x < \mu \\ 1, & x \ge \mu \end{cases}$$

2.2 The generalized Pareto distribution (GPD)

It is natural to regard as extreme events those of the x_i that exceed some high threshold, u. For large enough u, the distribution function of $x_i - u$, conditional on $x_i > u$, is approximately [2],

$$F(x) = 1 - \left(1 + \frac{\xi x}{\tilde{\sigma}}\right)^{-1/\xi} \quad (1)$$

defined on $\{x : x > 0, 1 + \xi x / \tilde{\sigma} > 0\}$, $\tilde{\sigma} = \sigma + (u - \mu)$ and with scale parameter σ $(\sigma > 0)$ and shape parameter ξ $(-\infty < \xi < \infty)$. If ξ > 0 (ξ < 0) then the GPD is simplified into the Pareto (Gamma) distribution. For $\xi \to 0$, GPD is simplified into the Exponential distribution. The family of distribution defined by (1) is called the generalized Pareto family. Denoting by σ_u the value of the GPD scale parameter for a threshold, $u > u_0$, where $\sigma_u = \sigma_{u0} + \xi(u - u_0)$, so that the scale parameter changes with u unless $\xi = 0$. A modified scale is obtained by parameterizing the GPD scale parameter as $\sigma^* = \tilde{\sigma} - \xi u$, which is constant with respect to u. A threshold μ_0 is selected as the lowest value of μ for which the estimates of σ^* and ξ remain near constant. The probability density function (cdf) of GPD is in the form of:

$$f(x) = 1 + \left[\xi\left(\frac{x - u}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

where $\sigma > 0$ and $-\infty < \xi < \infty$.

2.3 Maximum Likelihood Estimation

Let $X_1, X_2, ..., X_n$ be independent random variables with probability density function $f(x;\theta)$ where θ denotes d-dimensional parameter. The aim to estimate, then the likelihood function is [4],

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta),$$

and the log-likelihood function is

$$l(\theta) = \log L(\theta).$$

Finding the maximum likelihood estimator $\hat{\theta}$ of θ, differentiating $L(\theta)$ or $l(\theta)$ with respect to θ and setting its derivative equal to zero, the maximum likelihood estimator $\hat{\theta}$ can be obtained from the equations,

$$\frac{\partial L(\theta)}{\partial \theta} = 0, \text{ or } \frac{\partial l(\theta)}{\partial \theta} = 0.$$

2.4 Akaike's Information Criterion (AIC)

The Akaike's Information Criterion (AIC) is constructed from variance estimate of Kullback Leibler Information between the model and the truth. It is unbiased theory when n is large. The best model is lowest value of AIC [7].

In the general case, the form of AIC is

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximized value of the likelihood function for the estimated model.

3. Research Methodology

3.1 Data

This study, the maximum temperature data of central northeast of Thailand are used to analyze. They are obtained from the Meteorological Department of Thailand during January 1, 1977 to September 30, 2013 for six meteorological stations which are located in Khon Kean, Mahasarakham, Kalasin and Roi-Et province. The stations and summary statistics of the corresponding data sets are showed in Table 1.

Table 1: Stations and some summary statistics of data

Stations	Years of data	Latitude	Longitude	Min	Max	Skewness
1) Khon Kaen	1977-2013	16.4625	102.7857	15.50	41.90	-0.087
2) Thrapha Agromet.	1977-2013	16.3377	102.8235	15.90	42.60	-0.161
3) Kosum Phisai	1977-2013	16.2472	103.0681	17.20	42.00	-0.222
4) Kamalasai	1999-2013	16.3325	103.0681	17.10	42.30	-0.209
5) Roi Et	1977-2013	16.0500	103.6833	18.60	41.20	-0.108
6) Roi Et Agromet.	1982-2013	16.0666	103.6167	16.60	41.20	-0.121

3.2 Analysis of GEV

The analysis of GEV is formed of three steps as follows.

Step 1: To find the monthly maximum temperature data from original data. These are used to be as the block-maxima method to define the extreme temperature as the maximum of monthly temperature within each year.

Step 2: To find the estimates of parameters in the GEV distribution, we use maximum likelihood estimation with the R program. The parameters that we would to estimate are showed in Table 2. Table 2 gives the details of the forms of parameter and parameters in the GEV distribution.

Table 2: The forms of parameter and the parameters in the GEV distribution

Form of parameter	Parameters
Stationary process	
Form 1: μ is constant, σ is constant and ξ is constant	μ,σ and ξ
Non-Stationary process	
Form 2: $\mu(t)=\beta_0+\beta_1 t$, σ is constant and ξ is constant	$\beta_0,\ \beta_1,\sigma$ and ξ
Form 3: μ is constant, $\sigma(t)=\alpha_0+\alpha_1 t$ and ξ is constant	$\mu,\alpha_0,\ \alpha_1$ and ξ
Form 4: μ is constant, $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ and ξ is constant	$\mu,\alpha_0,\ \alpha_1$ and ξ
Form 5: $\mu(t)=\beta_0+\beta_1 t,\quad \sigma(t)=\alpha_0+\alpha_1 t$ and ξ is constant	$\beta_0,\ \beta_1,\alpha_0,\alpha_1$ and ξ
Form 6: $\mu(t)=\beta_0+\beta_1 t,\quad \sigma(t)=\exp(\alpha_0+\alpha_1 t)$ and ξ is constant	$\beta_0,\ \beta_1,\alpha_0,\alpha_1$ and ξ

Form Table 2, if $\xi=0$, the distribution of data is Gumbel, $\xi>0$, the distribution of data is Fréchet, and $\xi<0$, the distribution of data is Weibull.

Step 3: The adequacy distribution with various the forms of parameters of each station is used AIC.

3.3 Analysis of GPD

The analysis of GPD is formed of three steps as follows.

Step 1: To find the extreme temperature as the maximum of excess over threshold is used. The values of threshold and the number of excesses for each station are presented in Table 3. The GPD distribution is fitted to the tails of daily maxima maximum temperature data using threshold around 38.6 to 40.1 degree Celsius for each stations.

Step 2: The maximum likelihood estimation is used with the R program. The parameters that we would to estimate are showed in Table 4. Table 4 gives the details of the forms of parameter and parameters in the GPD distribution.

Table 3: The threshold and number of exceedances for each station.

Stations	Threshold	Number of excesses
1) Khon Kaen	40.1	127
2) Thrapha Agromet.	39.9	96
3) Kosum Phisai	40.1	128
4) Kamalasai	38.6	98
5) Roi Et	39.2	122
6) Roi Et Agromet.	39.2	102

Table 4: The forms of parameter and the parameters in the GPD distribution

Form of parameter	Parameters
Stationary process	
Form 1: σ is constant and ξ is constant	σ and ξ
Non-Stationary process	
Form 2: $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ and ξ is constant	$\alpha_0,\ \alpha_1$ and ξ

Form Table 4, if $\xi=0$, the distribution of data is Exponential, if $\xi>0$, the distribution of data is Pareto, and if $\xi<0$, the distribution of data is Gamma.

Step 3: The adequacy distribution with various the forms of parameters of each station is used AIC.

4. Results of extreme value theory

Table 5 shows the best model and the best form of parameters of the best model for each station for GEV distribution.

The results of GEV distribution, the values estimates of shape parameter ξ can be indicated that data are best fitted by the Weibull distribution for all stations. In the other hand, the form of parameters fitted to the maximum temperature data is stationary only Khon Kean station since the values of estimates of three parameters are constant and the rest fitted to non-stationary. In additions, parameter μ change depending on linear trend and, parameters σ and ξ are constant for Thrapha Agromet. station, Kosum Phisai station, Roi Et station and Roi Et. Agromet station. The parameter σ change depending on exponential trend and, parameters μ and ξ are constant for Kamalasai station.

Table 5: The fitted model and parameters estimates for GEV distribution

Stations	Model	Form	Parameter estimates						
			$\hat{\mu}$ (se)	$\hat{\beta}_0$ (se)	$\hat{\beta}_1$ (se)	$\hat{\sigma}$ (se)	$\hat{\alpha}_0$ (se)	$\hat{\alpha}_1$ (se)	$\hat{\xi}$ (se)
Khon Kaen	Weibull	From 1	35.13(0.11)	-	-	2.03(0.08)	-	-	-0.07(0.04)
Thrapha Agromet.	Weibull	From 2	-	34.98(0.11)	0.26(0.08)	2.01(0.08)	-	-	-0.06(0.04)
Kosum Phisai	Weibull	From 2	-	35.54(0.11)	0.42(0.08)	2.06(0.08)	-	-	-0.14(0.04)
Kamalasai	Weibull	From 4	34.52(0.18)	-	-	-	0.57(0.12)	0.19(0.09)	-0.14(0.05)
Roi-Et	Weibull	From 2	-	34.37(0.11)	0.21(0.09)	2.04(0.08)	-	-	-0.10(0.04)
Roi Et Agromet.	Weibull	From 2	-	34.33(0.13)	0.33(0.23)	2.05(0.09)	-	-	-0.09(0.05)

Table 6: The fitted model and parameters estimates for GPD distribution

Stations	Model	Form	Parameter estimates			
			$\hat{\sigma}$ (se)	$\hat{\alpha}_0$ (se)	$\hat{\alpha}_1$ (se)	$\hat{\xi}$ (se)
Khon Kaen	Gamma	Form 1	0.91(0.09)	-	-	-0.49(0.06)
Thrapha Agromet.	Gamma	Form 2	-	0.05(0.12)	0.31(0.08)	-0.52(0.08)
Kosum Phisai	Gamma	Form 2	-	0.06(0.11)	0.31(0.05)	-0.45(0.07)
Kamalasai	Gamma	Form 2	-	0.21(0.15)	0.31(0.09)	-0.50(0.08)
Roi Et	Gamma	Form 2	-	-0.12(0.10)	0.24(0.05)	-0.48(0.06)
Roi Et Agromet.	Gamma	Form 2	-	-0.24(0.11)	0.18(0.06)	-0.41(0.06)

Table 6 shows the best model and the best form of parameters of the best model for each station for GPD distribution. The values estimates of shape parameter ξ can be indicated that data are best fitted by the Gamma distribution for all stations. In the other hand, the form of parameters fitted to the maximum temperature data is stationary only Khon Kean station since the values of estimates of three parameters are constant and the rest fitted to non-stationary. In additions, parameter σ change depending on exponential trend and parameters ξ are constant for Thrapha Agromet. station, Kosum Phisai station, Kamalasai station, Roi Et station and Roi Et. Agromet station.

5. Conclusions

The study aims to model monthly maximum temperature in central northeast of Thailand by using the generalized extreme value distribution and the generalized pareto distribution. We found that maximum temperature data of the Khon Kean station at Muang district of Khon Kean province is stationary. Maximum temperature data of the Thrapha Agromet station, Kosum Phisai station, Kamalasai station, Roi Et station and Roi Et Agromet station are non-stationary. The Weibull distribution and Gamma distribution is the best model for GEV and GPD, respectively.

Acknowledgements

The authors would like to thank the Department of Statistics, Khon Kean University, Khon Kean, Thailand, for financial support and Meteorological Department of Thailand for temperature data.

Reference

[1] Bootchamruei P, Busababodin P, Kaewman A. Modeling Monthly Extreme Precipitation in Central Northeast of Thailand[Thesis]. Mahasarakham: Mahasarakham Univ; 2014.

[2] Coles S, Nadaraja S. An Introduction to Statistical Modeling of Extremes Values. Great Britain: Springer-Varlag London Limited; 2001.

[3] Easterling DR, Horton B. Maximum and minimum temperature trends for the globe. Science 1997; 227: 364–367

[4] Gong S. Estimation of hot and cold spells with extreme value theory[Dissertation]. Uppsala: Uppsala Univ; 2012.

[5] Khongthip P, Khamkong M, Bookamana P. Modeling Annual Extreme Precipitation in upper Northern Region of Thailand. Burapha Science Journal. 2013; 18 (1), 95-104.

[6] Khonrawee. Variability and climate change [Internet]. 2007 [updated 2008 Mar 28; cited 2013 Dec 22]. Available from:
www.tmd.go.th/ncct/article/2550.pdf

[7] McQuarrie A, Robert S, Tsai C.L. The Model Selection Criterion AICu. Statistics and Probability Letters. 1997; 34(3) : 285-292.

[8] Meteorological station in Narathiwat. Drought and sugar cane industries.[Internet]. [Cited 2014

Jan 15]. Available from: www.metnara.tmd.go.th/Patchai%20Kan%

[9] Nadarajah S, Dongseok C. Maximum daily rainfall in South Korea. J. Earth Syst. Sci. 2007; 116(4): 311-320.

[10] Thai Meteorological department. Climate of Thailand in year 2012[Internet]. [Cited 2013 Dec 22]. Available from: www.tmd.go.th/programs%5CyearlySum

[11] Unkasevic M, Tosic I. Changes in extreme daily winter and summer temperatures in Belgrade. Theoretical and Applied Climatol. 2009; 2009(95) : 27-38.

[12] Xu L, Wang H, Chen J. Application of Extreme Value Analysis to Extreme Drought Disaster Area in China. Springer-Verlag Berlin Heidelberg. 2011; 349-357.

Four team sport tournament with minimum number of traveling

Tinnaluk Rutjanisarakul[1] and Thiradet Jiarasuksakun[2*]

[1]*Department of Mathematics , King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
r.tinnaluk@gmail.com

[2]*Department of Mathematics , King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
thiradet.jia@mail.kmutt.ac.th

Abstract

In a sport tournament, each team normally plays with each other team twice: one home game and one opponent's home game. The sport traveling tournament condition could be considered a type of Traveling Tournament Problem (TTP) or double round robin. If the TTP of *n* teams requires each team plays every other once in the first $n - 1$ games, then it is called Half Traveling Tournament Problem (HTTP). If the HTTP also requires that the last $n - 1$ games are ordered exactly like the first $n - 1$ games with reversed venues, then it is called Mirror Traveling Tournament Problem (MTTP). This research aims to study the four team sport tournament in order to schedule the tournament with minimum total number of traveling of all teams. We also counts a number of all possible scheduling, presents proofs and gives an example of schedule with minimum total number of traveling. The result shows that the minimum total number of all team traveling is 17 and it is an MTTP.

Keywords: Sport tournament, Traveling Tournament Problem, Mirrored Traveling Tournament Problem, Half Traveling Tournament Problem, minimum total number of traveling

*Corresponding Author
E-mail Address: thiradet.jia@mail.kmutt.ac.th

1. Introduction

Nowadays football leagues are getting more and more popular in each nation all over the world. Each football league is also considered a tournament. Each playing team has to find sponsors to support many costs for a tournament such as advertisement, traveling and accommodation. So it would be beneficial to all if the tournament organizer could help to reduce some costs of each team. Thus the tournament scheduling problems affecting the traveling cost has become an important class of optimization problems for recent years. In football league, the traveling of each team is a class of traveling tournament problem. It can be represented by some notions in graph theory.

2. Research Methodology

In mathematics and computer science, graph theory is the study of graphs: mathematical structures used to model pairwise relations between objects from a certain collection. A "graph" in this context refers to a collection of vertices or nodes and a collection of edges that connect pairs of vertices. Let $G = (V, E)$ be a graph on *n* vertices with a vertex set *V* and an edge set *E*. There are some vocabularies and classification of graphs needed in this paper as shown in the following.

K_n denotes a complete graph on *n* vertices if every two distinct vertices of *G* are adjacent.

P_n denotes a path on n vertices if one can label vertices so that $E = \{\{v_0, v_1\}, \{v_1, v_2\}, ..., \{v_{n-2}, v_{n-1}\}\}$, where $V = \{v_0, v_1, ..., v_{n-1}\}$.

C_n denotes a cycle on *n* vertices if one can label vertices so that $E = \{\{v_0, v_1\},\{v_1, v_2\}, ...,\{v_{n-2}, v_{n-1}\}, \{v_{n-1},v_0\}\}$, where $V = \{v_0, v_1, ..., v_{n-1}\}$.

Definition 1.1. [7] A direct graph or digraph *E* consists of a set *V* of vertices and a set *E* of edges such that $e \in E$ is associated with an ordered pair of vertices. In other words, if each edge of the graph *G* has a direction then the graph is called directed graph. In the diagram of directed graph, each edge $e = (u,v)$ is represented by an arrow of directed curve from initial point *u* to the terminal point *v*.

Definition 1.2. [7] A cycle in a graph *G* that contains every vertices of *G* is called a Hamiltonian cycle of *G*. Thus a Hamiltonian cycle of *G* is a spanning cycle of *G*. A Hamiltonian graph is a graph that contains a Hamiltonian cycle. Certainly the graph C_n $(n \geq 3)$ is Hamiltonian. Also, for $n \geq 3$, the complete graph K_n is a Hamiltonian graph.

Definition 1.3. [7] A path in a graph *G* that contains every vertices of *G* is called a Hamiltonian path in *G*. If a graph contains a Hamiltonian cycle, then it contains a

Hamiltonian path. In fact, removing any edge from a Hamiltonian cycle produces a Hamiltonian path. If a graph contains a Hamiltonian path, however, it does not need contain a Hamiltonian cycle.

Definition 1.4. [3] A digraph is called Round robin tournament if a digraph is an orientation of a complete graph such that for every pair *u*, *v* of distinct vertices, exactly one of *(u, v)* and *(v, u)* is an arc.

Definition 1.5. [4] Traveling Tournament Problem (TTP) is a considered a double round robin (DRR). A scheduling to a double round robin (DRR) tournament, played by *n* teams, where *n* is a even number, consisting in a schedule where each team plays with each other twice, one game in its home and other in its opponent's home.

Definition 1.6. Half Traveling Tournament Problem (HTTP) is a generalization of TTP. An HTTP is a tournament where each team plays every other once in the first *n* − *1* rounds.

Definition 1.7. [6] Mirrored Traveling Tournament Problem (MTTP) is a generalization of TTP that represents the common structure in Latin America tournaments. An MTTP is a tournament where each team plays every other once in the *n* − *1*, followed by the same games with reversed venues in the last *n* − *1* rounds. It is also a type of HTTP.

In this paper, we start by studying the traveling sequences of four teams. Then we count all possible numbers of scheduling without any condition and with some conditions. HTTP and MTTP are also discussed in each scheduling. Finally, we schedule the tournaments with minimum number of traveling.

2.1 Notations and Observations

The figure below shows a complete graph on four vertices. Four teams can be represented by letters *A*, *B*, *C and D*. Each line represents two games between two corresponding teams.

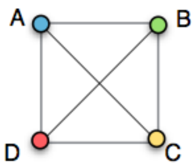


Figure 1: A complete graph on *4* vertices

Then, we describe the TTP by constructing a new graph *G′* as shown in Figure 2. For each team *X*, it will be represented by two points: *X* and *X ′*.

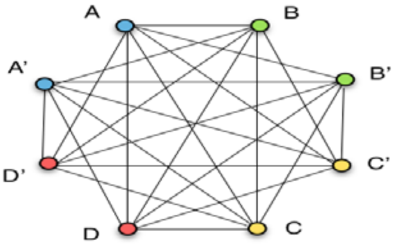


Figure 2: A graph *G′*

The graph G'_A represents all possible traveling of team *A*. Each line represents the traveling of team *A* from start to finish the tournament.

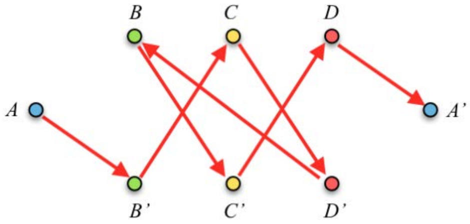


Figure 3: One possible traveling for team *A*

The expression *A* → *B′* of graph G'_A represents the game between *A* and *B* at home *A*. The expression *A* → *B′*→ *C* represents the second week; team *A* has a game with team *C* at home *C*. Then we assume that point *A* is a source and point *A′* is a sink. We find a traveling sequence starting at *A* and ending at *A′* in graph G'_A. For example, one possible traveling sequence shown in Figure 3 is

$$A \to B' \to C \to D' \to B \to C' \to D \to A'. \qquad (1)$$

After we obtain the traveling sequence (1), we can find the traveling sequence of team *B* by filling in the four empty spots in the sequence: *B*→*A*→...→...→*A′* →...→...→*B′* of team *B*.

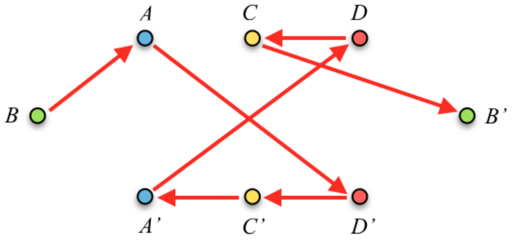


Figure 4: One possible traveling for team *B*

Then the traveling sequence of team *B* shown in Figure 4 could be

$$B \to A \to D' \to C' \to A' \to D \to C \to B'. \qquad (2)$$

After we get the traveling sequences (1) and (2), we can find the traveling sequence of team *C* by filling in the only two empty spots in the sequence: $C \to \ldots \to A' \to B \to \ldots \to A \to B' \to C'$ of team *C*.

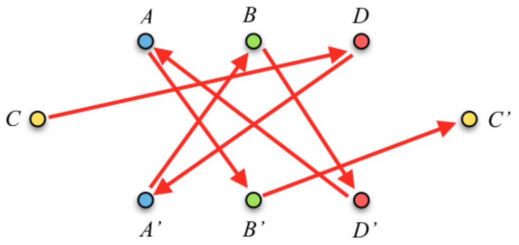


Figure 5: One possible traveling for team *C*

Then the traveling sequence of team *C* shown in Figure 5 could be

$$C \to D \to A' \to B \to D' \to A \to B' \to C'. \qquad (3)$$

After we get the traveling sequences (*1*), (*2*) and (*3*), there is no empty spot left. So the traveling sequence of team D is fixed as shown in sequence (*4*) and Figure 6;

$$D \to C' \to B \to A \to C \to B' \to A' \to D'. \qquad (4)$$

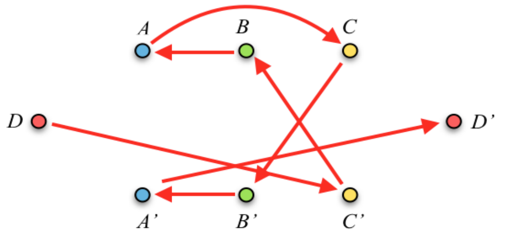


Figure 6: One possible traveling for team *D*

From all traveling sequences, we can also describe them by constructing a complete graph with two parallel edges for each pair of vertices to represent a TTP as shown in Figure 7.

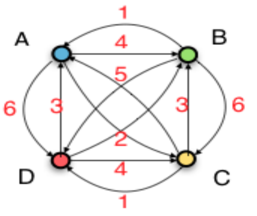


Figure 7: A multi-complete graph K_4 representing four team tournament

The number on each arrow shows the week number of each pair's game and the arrowhead represents the host of the game. After having all traveling sequences, one can make a scheduling of all four teams in seven weeks (last week for traveling back home) as shown in Table *1*.

Table 1: An Example of a Scheduling

Week	Tournament		Traveling to			
	Game *1*	Game *2*	Team *A*	Team *B*	Team *C*	Team *D*
1	*(A, B)*	*(D, C)*	*A*	*A*	*D*	*D*
2	*(C, A)*	*(B, D)*	*C*	*B*	*C*	*B*
3	*(A, D)*	*(B, C)*	*A*	*B*	*B*	*A*
4	*(B, A)*	*(C, D)*	*B*	*B*	*C*	*C*
5	*(A, C)*	*(D, B)*	*A*	*D*	*A*	*D*
6	*(D, A)*	*(C, B)*	*D*	*C*	*C*	*D*
7			*A*	*B*	*C*	*D*
	Number of traveling		*6*	*5*	*6*	*4*

Note *(X, Y)* means a game between *X* and *Y* at home *X*. This is also an MTTP. Then, the total number of traveling of all teams is *21*.

3. Research Results and Discussion

Theorem *3.1*. There are *5,760* possibilities of scheduling for four team tournament.

Proof. (Without loss of generality) Starting the traveling sequence of team *A*:
$A \to \ldots \to \ldots \to \ldots \to \ldots \to \ldots \to \ldots \to A'$, where each spot is filled by $x \in \{B, B', C, C', D, D'\}$. There are *6!* possibilities of traveling. For example, one possible traveling sequence is

$$A \to B \to D' \to B' \to D \to C' \to C \to A'.$$

Next, considering the traveling sequence of team *B*, there are only 4 empty spots left as indicated below

$$B \to A' \to \ldots \to A \to \ldots \to \ldots \to \ldots \to B'.$$

There are *2 x 2* possibilities because in week *2* and *4*, *C* or *C'* could be chosen which are *2* possibilities and in week *3* and *6*, *D* or *D'* could be chosen which are *2* possibilities. For example,

$$B \to A' \to C \to A \to C' \to D' \to D \to B'.$$

Next, considering the traveling sequence of team *C*, there are only *2* empty spots left which as indicated below

$$C \to \ldots \to B' \to \ldots \to B \to A \to A' \to C'.$$

There are *2* possibilities because in week *1* and *3*, *D* or *D'* could be chosen. For example,

$$C \to D \to B' \to D' \to B \to A \to A' \to C'.$$

Finally, there is no empty spot left for the traveling sequence of team *D*. So the traveling sequence of team *D* is $D \rightarrow C' \rightarrow A \rightarrow C \rightarrow A' \rightarrow B \rightarrow B' \rightarrow D'$.

Thus, there are $6!\ x\ 2\ x\ 2\ x\ 2 = 5{,}760$ possibilities of scheduling for four team tournament.

Theorem 3.2. There are *1,920* possibilities of scheduling for four team tournament with additional condition: a game between *X* and *Y* at *Y*'s home cannot be followed by the game between *X* and *Y* at X's home.

Proof. (Without loss of generality) Starting the traveling sequence of team *A*:
$A \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow A'$, where each spot is filled by $x \in \{B, B', C, C', D, D'\}$
Let T^a be a set of all traveling sequences of team A with no condition. Let $|T^a|$ be a cardinal number of set T^a, then $|T^a|$ equal $6!$.

Let *T* be a set of possible traveling sequences of team *A* and satisfying a given condition: a game between *X* and *Y* at *Y*'s home cannot be followed by the game between *X* and *Y* at *X*'s home.

Let *T'* be a complement set of *T*. It is a set of traveling sequences of team *A*, where at least one team plays with team *A* at home and away games consecutively. Then, T'_i is defined to be a set of traveling sequences of team *A* such that there are *i* teams which are in $\{B, C, D\}$, playing with team *A* at home and away games, consecutively.

So, $|T'|=|T'_1|-|T'_2|+|T'_3|$ by inclusion and exclusion principle. Then, $|T'_1| = \binom{3}{1}\ x\ 5!\ x\ 2$, $|T'_2| = \binom{3}{2}\ x\ 4!\ x\ 2^2$, $|T'_3| = \binom{3}{3}\ x\ 3!\ x\ 2^3$.

Thus, $|T'|=|T'_1|-|T'_2|+|T'_3|= 720-288+48 = 480$. So, we find $|T|$ by $|T| = |T^a| - |T'| = 720 - 480 = 240$. Thus, there are *240* possibilities traveling sequence of team *A*. For example, $A \rightarrow B \rightarrow C' \rightarrow D \rightarrow C \rightarrow B' \rightarrow D' \rightarrow A'$.

Next, the traveling sequences of team *B*, *C* and *D* are considered by the same idea of proof of Theorem *3.1*. So, there are *2x2x2* possibilities of traveling sequences of these teams.
Therefore, there are *1,920* possibilities of scheduling for four team tournament.

Theorem *3.3*. There are *1,536* possibilities of scheduling for four team Half Traveling Tournament Problem with additional condition: a game between *X* and *Y* at *Y*'s home cannot be followed by the game between *X* and *Y* at *X*'s home.

Proof. (Without loss of generality) Starting the traveling sequence of team *A*:
$A \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow A'$, where the first *3* weeks, each team must play with other team once.
Let H^a be a set of all half traveling sequences of team *A* with no condition, then $|H^a|$ equal $3!x3!x2^3 = 288$.

Let *H* be a set of possible half traveling sequences of team *A* and satisfying a given condition: a game between *X* and *Y* at *Y*'s home cannot be followed by the game between *X* and *Y* at *X*'s home.

Let *H'* be a complement set of *H*. It is a set of half traveling sequences of team *A*, where there is a team plays *2* games with team *A* in week *3* and *4* consecutively. Then, H'_t is a set of team $t \in \{B, C, D\}$ plays with team *A* consecutively in week *3* and *4*. So, $|H'|=|H'_B|+|H'_C|+|H'_D|$.

Then, $|H'_B|= 2x4x2^2$, $|H'_C|= 2x4x2^2$, $|H'_D|= 2x4x2^2$. Thus, $|H'|= |H'_B|+|H'_C|+|H'_D|= 32+32+32= 96$. So, we find $|H|$ by $|H|= |H^a|-|H'|= 288-96 = 192$. For example,

$$A \rightarrow B \rightarrow C' \rightarrow D \rightarrow C \rightarrow B' \rightarrow D \rightarrow A'.$$

Next, the traveling sequences of team *B*, *C* and *D* are considered by the same idea of proof of Theorem *3.1*. So, there are *2x2x2* possibilities of traveling sequences of these teams. For example,

$B \rightarrow A' \rightarrow D \rightarrow C \rightarrow D' \rightarrow A \rightarrow C' \rightarrow B'$,
$C \rightarrow D \rightarrow A \rightarrow B' \rightarrow A' \rightarrow D' \rightarrow B \rightarrow C'$and
$D \rightarrow C' \rightarrow B' \rightarrow A' \rightarrow B \rightarrow C \rightarrow A \rightarrow D'$.

Therefore, there are *1,536* possibilities of scheduling for four team Half Traveling Tournament Problem.

Theorem *3.4*. There are *384* possibilities of scheduling for four team Mirrored Traveling Tournament Problem with additional condition: a game between *X* and *Y* at *Y*'s home cannot be followed by the game between *X* and *Y* at *X*'s home.

Proof. (Without loss of generality) Starting the traveling sequence of team *A*:
$A \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow \dots \rightarrow A'$, where the first *3* weeks, each team must play with other team once and the last *3* games are ordered exactly like the first *3* games with reversed venues. With this condition, it is impossible to have a game between *X* and *Y* at *Y*'s home followed by a game between *X* and *Y* at *X*'s home.

Let *M* be a set of all mirrored traveling sequences of team *A*, then $|M| = 3!\ x2^3$. For example,

$A \rightarrow B \rightarrow C' \rightarrow D \rightarrow B' \rightarrow C \rightarrow D' \rightarrow A'$.

Next, the traveling sequences of team *B*, *C* and *D* are considered by the same idea of proof of Theorem *3.1*. So, there are *2x2x2* possibilities of traveling sequences of these teams. For example,

$B \rightarrow A' \rightarrow D \rightarrow C \rightarrow A \rightarrow D' \rightarrow C' \rightarrow B'$,
$C \rightarrow D \rightarrow A \rightarrow B' \rightarrow D' \rightarrow A' \rightarrow B \rightarrow C'$ and
$D \rightarrow C' \rightarrow B' \rightarrow A' \rightarrow C \rightarrow B \rightarrow A \rightarrow D'$.

Therefore, there are *384* possibilities of scheduling for four team Mirrored Tournament Problem.

Theorem *3.5*. For four team tournament, the minimum number of traveling of each team is *4*.

Proof. Each team must have *3* away games. There would be at least *3* traveling for *3* away games, and plus *1* for going back home. Therefore, the minimum number of traveling is *4*.

Lemma *3.6*. The minimum number of traveling occurs when a team has all away games consecutively.

Proof. (by contradiction) Since each team has to play *6* games, assume one team has one home game between *2* away games. For example, Home-Away-Home- Away-Away-Home. We can find the number of traveling is at least *5*. That contradicts with minimum number of traveling is *4*.

Theorem *3.7*. For four team tournament, not all team could attain a minimum number of traveling.

Proof. There are only *4* possibilities of minimum number of traveling which are shown in Table *2*.

Table 2: All possible minimum traveling for each team

Week	1	2	3	4	5	6
Case 1	Home	Home	Home	Away	Away	Away
Case 2	Home	Home	Away	Away	Away	Home
Case 3	Home	Away	Away	Away	Home	Home
Case 4	Away	Away	Away	Home	Home	Home

From Table *2*, it is not possible to schedule this tournament because there must be *2* home teams and *2* away teams each week.

However, there exists a possible scheduling for four teams which attains a minimum total number of traveling. One example is shown in Table *3*.

Table 3: Schedule of tournament with minimum number of traveling

Week	Tournament		Traveling to			
	Game *1*	Game *2*	Team *A*	Team *B*	Team *C*	Team *D*
1	*(A, B)*	*(C, D)*	*A*	*A*	*C*	*C*
2	*(A, C)*	*(B, D)*	*A*	*B*	*A*	*B*
3	*(A, D)*	*(B, C)*	*A*	*B*	*B*	*A*
4	*(B, A)*	*(D, C)*	*B*	*B*	*D*	*D*
5	*(C, A)*	*(D, B)*	*C*	*D*	*C*	*D*
6	*(D, A)*	*(C, B)*	*D*	*C*	*C*	*D*
7			*A*	*B*	*C*	*D*
	Number of traveling		*4*	*5*	*4*	*4*

Then, the number of traveling of all team is *17*. It is a Mirrored Traveling Tournament Problem (MTTP). For this solution, it is a minimum number of traveling but there is a team having more traveling than other teams. If we want to make the tournament fair, we can reschedule so that each team has the same minimum number of traveling. In this case, each team has to travel *5* times as shown in Table *4*.

Table 4: Schedule of tournament with each team having the same number of traveling

Week	Tournament		Traveling to			
	Game *1*	Game *2*	Team *A*	Team *B*	Team *C*	Team *D*
1	*(B, A)*	*(C, D)*	*B*	*B*	*C*	*C*
2	*(D, A)*	*(C, B)*	*D*	*C*	*C*	*D*
3	*(A, C)*	*(D, B)*	*A*	*D*	*A*	*D*
4	*(A, D)*	*(B, C)*	*A*	*B*	*B*	*A*
5	*(C, A)*	*(B, D)*	*C*	*B*	*C*	*B*
6	*(A, B)*	*(D, C)*	*A*	*A*	*D*	*D*
7			*A*	*B*	*C*	*D*
	Number of traveling		*5*	*5*	*5*	*5*

The total number of traveling of all teams is *20*, where each team travels *5* times. It is a Half Traveling

Tournament Problem (HTTP), but not a Mirrored Traveling Tournament Problem (MTTP).

Remark 3.1. For four team tournament, Mirrored Traveling Tournament Problem cannot be made with each team traveling *5* times.

Proof. Consider the first three weeks which could be three home games, or two home and one away games, or one home and two away games, or three away games. In case of two home and one away games, there are *3!/2! = 3* options. In case of one home and two away games, there are also *3!/2! = 3* options. So, there are *8* possible Mirrored Traveling Tournaments, which are shown in Table *5*.

Table 5: All possible Mirrored Traveling Tournament for each team

Week	1	2	3	4	5	6
Case 1	Home	Home	Home	Away	Away	Away
Case 2	Home	Home	Away	Away	Away	Home
Case 3	Home	Away	Home	Away	Home	Away
Case 4	Away	Home	Home	Home	Away	Away
Case 5	Home	Away	Away	Away	Home	Home
Case 6	Away	Home	Away	Home	Away	Home
Case 7	Away	Away	Home	Home	Home	Away
Case 8	Away	Away	Away	Home	Home	Home

There are only *2* cases (case *4* and *7*) having number of traveling equal to *5*. Therefore, the four team tournament cannot be Mirrored Traveling Tournament Problem with number of traveling equal to *5* for each team.

4. Conclusion

There are *1,536* possibilities of scheduling for four team HTTP and *384* possibilities of scheduling for four team MTTP. The minimum number of traveling of each team is four for four team tournament, but not all teams could attain a minimum number of traveling. The minimum total number of traveling is *17* and it is an MTTP. A scheduling that is made fair to every team is an HTTP, but not an MTTP.

Acknowledgements

The authors would like to thank Human Resource Development in Science Project (Science Achievement Scholarship of Thailand, SAST) for financial support. This research is a part of Ph.D. thesis at department of mathematics, faculty of science at KMUTT.

References

[1] Ahmed AH. Genetic Algorithm for the Traveling Salesman Problem using Sequential Constructive Crossover Operator. International Journal of Biometrics & Bioinformatics. 2010; 3(6): 96-105.

[2] Biajoli FL, Lorena LAN. Mirrored Traveling Tournament Problem: An Evolutionary Approach. Advances in Artificial Intelligence - IBERAMIA-SBIA 2006 Lecture Notes in Computer Science. 2006; 4140: 208−217.

[3] Chartrand G, Zhang P. Introduction to Graph Theory. McGraw-Hill Companies; 2005.

[4] Easton K, Nemhauser G, Trick M. The Traveling Tournament Problem Description and Benchmarks. Seventh International Conference on the Principles and Practice of Constraint Programming. 2001; 580−589.

[5] Falkenauer E. Genetic Algorithms and Grouping Problems. New York: John Wiley & Sons; 1998.

[6] Ribeiro CC, Urrutia S. Heuristics for the Mirrored Traveling Tournament Problem, Fifth International Conference on the Practice and Theory of Automated Timetabling; 2004 August 18-20; Pittsburgh, USA. 2004. p. 323-342.

[7] Vasudev C. Graph Theory with Application. Delhi: New Age International (P) Ltd.; 2006.

[8] West D. Introduction to Graph Theory. 2nd edition, Prentice Hall: 2001.

Quality control chart based on the Kolmogorov-Smirnov structure

Preecha Khrueasom[1*] and Adisak Pongpullponsak[2]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
*55090800006@*st.*kmutt.ac.th*
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140 Thailand,*
adisak.pon@kmutt.ac.th

Abstract

This research develops the distribution-free or unknown distribution quality control chart based on the Kolmogorov-Smirnov statistic (KS). The performance is Average Run Length (ARL) which derives from the probability that a single point falls outside the control limits when the process is in-control. The unknown process distribution is generated by simulating the uniform (0, 1). The results show that the average run length (ARL) of KS is equal to 469.11 more than the general model of control chart is value equal to 370.

Keywords: Distribution-free, Kolmogorov-Smirnov, distribution-free, nonparametric, quality control chart, average run length

E-mail Address: 55090800006@st.kmutt.ac.th

1. Introduction

According to Dr. Shewhart who provided the first control chart in 1924, currently the quality control is used widely in the industry and the control charts was developed from past to present in multiple models.

However, we need to know the sampling properties of monitoring statistic, in order to develop the proper chart study the characteristics of the chart and evaluate its performance, including comparing with other chart. In most case, a normal distribution or a distribution with a know form was assumed for variable data. Hence the question is: What will we do if we had no knowledge of the underlying population distribution or the known distribution does not help us derive the necessary sampling properties? Using a nonparametric approach seems to be a reasonable alternative, from "A new nonparametric EWMA Sing Control Chart" by Su-Fen Yang, Jheng-Sian Lin and Smiley W Cheng [8].

The Kolmogorov-Smirnov statistic (KS) is also distribution-free, knowledge of their sampling distribution would make them useful in nonparametric statistics inference as well. Their exact sampling distributions are considerably easier to calculate then that for D_n statistic, from "Nonparametric statistical Inference" by Gibbons, J.D. [2]. Saad T. Bakir. [7] was made to the charting statistics is a modified version of the two - sample Kolmogorov-Smirnov (KS) test statistic where the difference of the reference and test empirical distribution function is maximized only over the training sample values.

From "Statistical quality control" by Pongpullponsak, A. [4], considered that the performance evaluation of the control chart was measured by average run length (ARL) and the probability that a single point falls outside the limited when the process is in-control.

This research is to study and made to develop quality control chart for distribution-free or unknown distribution based on the KS structure and application of the general theory control chart.

2. Research Methodology

2.1 The Kolmogorov-Smirnov

A random sample $X_1, X_2, ..., X_n$ is drawn from a population with unknown cumulative distribution function $F_X(x)$. For any value of *x*, the empirical distribution function of the sample $S_n(x)$ provides a consistent point estimator for $F_X(x)$. The step function $S_n(x)$, with jumps at the values of the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ for the sample, approaches the true distribution function for all *x*. Therefore, the deviations between the true distribution function and its statistical. This result suggests that the statistic (see Empirical processes, Kolmogorov-Smirnov Statistic Math [1], Nonparametric statistical Inference by Gibbons, J.D. [2] and Section 13 Kolmogorov-Smirnov test [6]).

$$D_n = \sup_x \left| S_n(x) - F_X(x) \right|. \qquad (1)$$

This D_n statistic, called the Kolmogorov-Smirnov one-sample statistic is particularly useful in nonparametric statistical inference because the probability distribution of D_n does not depend upon $F_X(x)$ as long as F_X is continuous and D_n may be called a distribution-free statistic from “Nonparametric statistical Inference” by Gibbons, J.D. [2].

The direction deviations defined as

$$D_n^+ = \sup_x \left| S_n(x) - F_X(x) \right|$$
$$D_n^- = \sup_x \left| F_X(x) - S_n(x) \right|$$

the D_n^+ and D_n^- are called the one-sided Kolmogorov-Smirnov one-sample statistics.

There measures are also distribution-free, as is proved in the following by theorem:

Theorem 1. The statistics of D_n, D_n^+ and D_n^- are completely distribution-free for any continuous F_X [2].

Proof $D_n = \sup_x \left| S_n(x) - F_X(x) \right| = \max_x \left(D_n^+, D_n^- \right)$

We can write $S_n(x) = \frac{i}{n}, i = 0,1,...,n,$

for $X_{(i)} \le x \le X_{(i-1)}$

therefore

$$\begin{aligned} D_n^+ &= \sup_x \left[S_n(x) - F_X(x) \right] \\ &= \max_{0\le i\le n} \sup_{X_{(i)} \le x \le X_{(i+1)}} \left[S_n(x) - F_X(x) \right] \\ &= \max_{0\le i\le n} \sup_{X_{(i)} \le x \le X_{(i+1)}} \left[\frac{i}{n} - F_X(x) \right] \\ &= \max_{0\le i\le n} \left[\frac{i}{n} - \inf_{X_{(i)} \le x \le X_{(i+1)}} F_X(x) \right] \\ &= \max_{0\le i\le n} \left[\frac{i}{n} - F_X\left(x_{(i)}\right) \right] \\ &= \max\left\{ \max_{0\le i\le n} \left[\frac{i}{n} - F_X\left(x_{(i)}\right) \right], 0 \right\} \end{aligned}$$

similarly

$$D_n^- = \max\left\{ \max_{0\le i\le n} \left[F_X\left(x_{(i)}\right) - \frac{i-1}{n} \right], 0 \right\}$$

$$D_n = \max\left\{ 0, \max_{0\le i\le n} \left[\frac{i}{n} - F_X\left(x_{(i)}\right) \right], \max_{0\le i\le n} \left[F_X\left(x_{(i)}\right) - \frac{i-1}{n} \right] \right\}.$$

The probability distribution of D_n, D_n^+ and D_n^- therefore are seen to depend only on the random variables $F_X\left(x_{(i)}\right),\ i = 0,1,...,n.$ There are the order statistics from the uniform distribution (0, 1).

The empirical (sample) distribution function of a random sample of size *n*, denoted by $S_n(x)$, is the proportion of sample values which do not exceed the number *x*. Thus $S_n(x)$ is the step function which increases by the amount 1/*n* at its jump points, which are the order statistics of the sample. Letting $X_{(1)}, X_{(2)}, ..., X_{(n)}$ denote the order statistics of a random sample, its empirical distribution function is denote symbolically as from “Nonparametric statistical Inference” by Gibbons, J.D. [2]

$$S_n(x) = \begin{cases} 0 & if\ x \le X_{(1)} \\ \frac{k}{n} & if\ X_{(k)} \le x \le X_{(k+1)}\ for\ k = 1,2,...,n-1. \\ 1 & if\ x \ge X_{(n)} \end{cases}$$

Theorem 2. For the random variable $S_n(x)$, which is the empirical distribution function of a random sample $X_{(1)}, X_{(2)}, ..., X_{(n)}$ from a distribution F_X from “Nonparametric statistical Inference” by Gibbons, J.D. [2], we have

$$P\left[S_n(x) = \frac{i}{n} \right] = \binom{n}{i} \left[F_X \right]^i \left[1 - F_X \right]^{n-i},\ i = 0,1,...,n.$$

Proof Define the indicator random variables

$$\delta_i(t) = \begin{cases} 1 & if\ X_i \le t \\ 0 & otherwise. \end{cases}$$

The $\delta_1(t), \delta_2(t), ..., \delta_n(t)$ constitute a set of *n* independent random variables from the Bernoulli distribution with parameter θ [2], where

$$\theta = P\left[\delta_i(t) = 1 \right] = P\left(X_i \le t \right) = F_X(t).$$

Since we can write

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_i(x)$$

the random variables $nS_n(x)$ is the sum of *n* independent Bernoulli random distribution, which follows the binomial distribution with parameter $\theta = F_X(x)$.

Computation of $F_X(x)$

The sample for the test is made of $F_X(x)$ scores, each of the denoted X_i. The sample mean is denoted M and the sample variance is denoted S^2, it computed from *z* scores which are obtained the following formula, from “New Table and Numerical Approximations for Kolmogorov-Smirnov/Lilliefors/Van Soest Normality Test” by Paul Molin and $Herve'$ Abdi [5]:

$F_X(x)$ is equal to $z = \frac{X_i - M}{S}$

let S is the square root of

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - M)^2}{N - 1}$$

and

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Computation of V_{D_n} and proof

This research used exact moments of the order statistics from “Nonparametric statistical Inference” by Gibbons, J.D. [2]

$$E\left(X_{(r)}^{k}\right) = \frac{n!}{(r-1)!(n-r)!}\int_{0}^{1} x^{r+k+1}(1-x)^{n-r}\,dx$$

$$= \frac{n!}{(r-1)!(n-r)!}B(r+k, n-r+1)$$

$$= \frac{n!(r+k-1)!}{(r-1)!(n-r)!}$$

$$= \frac{(r+k-1)(r+k-2)\ldots(r+1)r}{(n+1)(n+k-1)\ldots(n+2)(n+1)}.$$

For any $1 \le r \le n$ and integer *k*. In particular, the mean is

$$E\left(X_{(r)}\right) = \frac{r}{n+1}$$

$$Var\left(X_{(r)}\right) = E\left(X_{(r)}^{2}\right) - \left[E\left(X_{(r)}\right)\right]^{2}$$

and

$$= \frac{r(r+1)}{(n+2)(n+1)} - \frac{r^2}{(n+1)^2}$$

$$= \frac{r(n-r+1)}{(n+2)(n+1)^2}.$$

The distribution-free techniques is the order statistics for the uniform distribution over the interval (0, 1) the case where the set $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. Then the marginal distribution as follows from “Nonparametric statistical Inference” by Gibbons, J.D. [2]:

$$f_{X_{(r)},X_{(s)}}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}x^{r-1}$$

$$(y-x)^{s-r-1}(1-y)^{n-s} \quad ,0 < x < y < 1.$$

In order to determine the covariance of the two order statistics $X_{(r)}$ and $X_{(s)}$ from the marginal distribution as follows from “Nonparametric statistical Inference” by Gibbons, J.D. [2]:

$$E\left(X_{(r)}X_{(s)}\right) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

$$\int_{0}^{1}\int_{0}^{y} x^{r}y(y-x)^{s-r-1}(1-y)^{n-s}\,dx\,dy$$

$$= \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

$$\int_{0}^{1} y(1-y)^{n-s}\left[y^{s}B(r+1,s-r)\right]dy$$

$$= \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

$$B(s+2,n-s+1)B(r+1,s-r)$$

$$= \frac{n!(s+1)!(n-s)!r!(s-r-1)!}{(r-1)!(s-r-1)!(n-s)!(n+2)!s!}$$

$$= \frac{r(s+1)}{(n+1)(n+2)}.$$

Now the covariance is found using from “Nonparametric statistical Inference” by Gibbons, J.D. [2]

$$\operatorname{cov}\left(X_{(r)}X_{(s)}\right) = E\left(X_{(r)}X_{(s)}\right) - E\left(X_{(r)}\right)E\left(X_{(s)}\right)$$

$$= \frac{r(s+1)}{(n+1)(n+2)} - \frac{rs}{(n+1)^2}$$

$$= \frac{r(n-s+1)}{(n+1)^2(n+2)}.$$

From the above data, the author can provide the following equation by applying from the computation of V_{D_n} and variance properties.

Let $S_n(x)$ and $F_X(x)$ are marginal distribution, given a and b are constant so that

$$V_{D_n}\left(a^2S_n(x) \pm b^2F_X(x)\right) = \left|a^2V\left[S_n(x)\right] + b^2V\left[F_X(x)\right] \pm 2ab\operatorname{cov}\left(S_n(x), F_X(x)\right)\right|$$

note $E\left(S_n(x)\right) = F_X(x)$, $V\left[S_n(x)\right] = \dfrac{F_X(x)\left[1-F_X(x)\right]}{n}$

and

$E\left[F_X(x)\right] = E\left(X_{(r)}\right) = \dfrac{r}{n+1}$, $V\left[F_X(x)\right] = E\left[F_X^2(x)\right] - \left[E\left(F_X(x)\right)\right]^2$.

2.2 Quality control chart models

In order to evaluate the performance of the various control charts, their entire relevant characteristics were collected from the literature used in this study.

2.2.1 Statistical basis of the control chart

We may give the general model of the control chart. If *w* is a sample statistic to measures some quality characteristics of interest and supposes that the mean of *w* is μ_w, the standard deviation of w is σ_w. Then the center line (CL), the upper control limit (UCL) and lower control limit (LCL), *L* is the distance of the control limits from the center line (from Montgomery, D.C [3] and Pongpullponsak, A [4]). The general model of the control chart as follows:

$$\begin{aligned} UCL &= \mu_w + L\sigma_w \\ CL &= \mu_w \\ LCL &= \mu_w - L\sigma_w \end{aligned} \qquad (2)$$

if $\sigma_w = \dfrac{\sigma}{\sqrt{n}}$, *n* is sample size and *L* equal to 3.

2.1.2 KS control chart

This research develops of the control chart based on the Kolmogorov-Smirnov statistic (KS) is applied of the general model of the control chart from (1), (2).

The KS control chart as follows:

$$\begin{aligned} UCL &= D_n + L\frac{\sqrt{V_{D_n}}}{\sqrt{n}} \\ CL &= D_n \\ LCL &= D_n - L\frac{\sqrt{V_{D_n}}}{\sqrt{n}} \end{aligned} \qquad (3)$$

D_n is the value of $\left|S_n(x) - F_X(x)\right|$ at the rank of sample in the median

L is the distance of the control limits from the center line equal to 3

n is the sample size equal to 5

$\sqrt{V_{D_n}}$ is the standard deviation of max values of $D_n = \left|S_n(x) - F_X(x)\right|$

3. Example KS control chart

This research used simulations generate random numbers the uniform (0, 1) and the model for the KS control chart by (3), we will use the following example from table 1.

Table 1: Example of KS

Number of Subgroup	Order of Sample Size	z	$F_X(x)$	$S_n(x)$	D_n	V_{D_n}
1	6	-2.1306	0.0166	0.0500	0.0334	0.0012
2	8	-1.7149	0.0432	0.1000	0.0568	0.0019
3	9	-1.507	0.0659	0.1500	0.0841	0.0025
4	11	-1.0913	0.1376	0.2000	0.0624	0.0011
5	13	-0.6755	0.2497	0.2500	0.0003	0.0011
6	15	-0.2598	0.3975	0.3000	0.0975	0.0027
7	16	-0.052	0.4793	0.3500	0.1293	0.0024
8	16	-0.052	0.4793	0.4000	0.0793	0.0018
9	16	-0.052	0.4793	0.4500	0.0293	0.0013
10	17	0.1559	0.5619	0.5000	0.0619	0.001
11	17	0.1559	0.5619	0.5500	0.0119	0.001
12	17	0.1559	0.5619	0.6000	0.0381	0.0012
13	18	0.3638	0.642	0.6500	0.0080	0.0008
14	19	0.5716	0.7162	0.7000	0.0162	0.0001
15	19	0.5716	0.7162	0.7500	0.0338	0.0009
16	19	0.5716	0.7162	0.8000	0.0838	0.0019
17	21	0.9873	0.8383	0.8500	0.0117	0.0002
18	22	1.1952	0.884	0.9000	0.0160	0.0004
19	23	1.4031	0.9197	0.9500	0.0303	0.0002
20	23	1.4031	0.9197	1.0000	0.0803	0.0016

The KS control chart from (3) as follows:

$$\begin{aligned} UCL &= 0.06 + 3\frac{\sqrt{0.0024}}{\sqrt{5}} \approx 0.13 \\ CL &= 0.0619 \approx 0.06 \\ LCL &= 0.06 - 3\frac{\sqrt{0.0024}}{\sqrt{5}} \approx 0.00 \end{aligned}$$

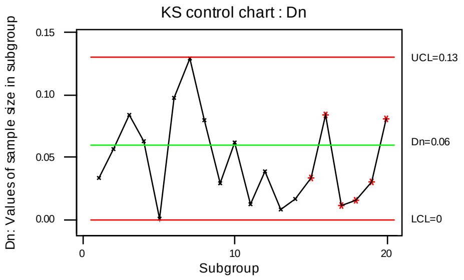


Figure 1: KS control chart.

4. Computation the ARL of the KS control chart

The performance evaluation of the control chart was measured by average run length (ARL) and the probability that a single point falls outside the limited when the process is in-control, the ARL can be computed from Montgomery, D.C [3] and Pongpullponsak, A [4]

$$\text{ARL} = \frac{1}{p}.$$

However, for determining the ARL of control charts KS by counting the number of random sample before, the process is out of control (or the run) from the graph it can be difficult, due to the large number of samples to process out of control, this research used method of Siegmund from Montgomery, D.C [3] and Yindee.P [9].

Recall that the proposed one-side KS chart signals at the first sampling instance *i*

$$S_n(x) = \frac{i}{n}, i = 0,1,...,n,$$

for which

$$D_n \le \left|S_n(x) - F_X(x)\right| = \max_x\left(D_n^+, D_n^-\right).$$

Thus a signal occurs if

$$D_n^+ \le S_n(x) - F_X(x)$$

or $D_n^- \ge F_X(x) - S_n(x)$

Using *n*= *h* = 5 and *k*= $S_n(x)$ will generally provide a KS that has good ARL properties against a shift of about 1σ in the process, for a one-side KS with parameters *h* and *k*, by Siegmund's approximation is

$$\text{ARL} = \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2}$$

and ARL is symmetry $\frac{1}{\text{ARL}} = \frac{1}{\text{ARL}^+} + \frac{1}{\text{ARL}^-}$

Example* 1. *ARL of KS

Define $\delta^* = 0, \Delta = \delta^* - k = 0 - k = (-k)$ and $b = h + 1.166$

$$\text{ARL}_0^+ = \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2}$$
$$= \frac{\exp(2k(h+1.166)) + 2k(h+1.166) - 1}{2k^2}$$

by symmetry

$$\frac{1}{\text{ARL}_0} = \frac{1}{\text{ARL}_0^+} + \frac{1}{\text{ARL}_0^-}$$

the ARL of KS at $S_n(x)$ or $CL = D_n$ = 0.50 is not a shift and equal to 469.11, This table 2 to show the ARL are shift 0.5, 1.0, 1.5, 2.0, 2.5 and 3.00.

Table 2: ARL Performance of the with $k = S_n(x)$ and h = 5

Shift of k ((σ)	*h* = 5
0	469.11
0.5	0.00
1.0	5.17
1.5	2.83
2.0	1.94
2.5	1.48
3.0	1.19

This table 3 to show the ARL of KS at $S_n(x)$ or $D_n \le \left|S_n(x) - F_X(x)\right| = \max_x\left(D_n^+, D_n^-\right)$ = 0.35 is not a shift and equal to 142.01 and are shift 0.5, 1.0, 1.5, 2.0, 2.5 and 3.00.

Table 3: ARL Performance of $D_n \le \max\left|S_n(x) - F_X(x)\right|$

Shift of k ((σ)	*h* = 5
0	142.01
0.5	11.19
1.0	4.15
1.5	2.49
2.0	1.78
2.5	1.38
3.0	1.13

5. Research Results and Discussion

The objective of this research is made to develop quality control chart based on the Kolmogorov-Smirnov statistic (KS) and with computation the ARL of the KS control chart by Siegmund's approximation.

The results show that the performance for ARL of D_n in the table 2 is value equal to 469.11 comparison with the ARL for the general model of control chart is value equal to 370, we see that the KS is more than, when k $\left(\text{or}\, S_n\left(x\right)\right)$ is increasing and h is constant the trend of ARL is decreasing.

This research, the value of ARL is equal to 469.11 and consistent with Montgomery, D.C. [3] and Yindee.P. [9]. However, the model obtained in this research may be defective. The author will improve to make it better and this problem could be an interesting issue for further study.

Acknowledgements

The authors would like to thank the referees for their constructive comments and suggestions. In this research, was supported by Department of mathematics, King Mongkut's University of Technology Thonburi.

References

[1] Empirical processes, Kolmogorov-Smirnov Statistic Math 6070, Spring 2006 [Internet]. [updated 2006 ;cited 2014 Feb 27]. Available from: http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf

[2] Gibbons, J.D. Nonparametric statistical Inference. McGRAW-HILL Inc; 1971.

[3] Montgomery, D.C. Introduction to statistical quality control 6th ed. New York: John Wiley & Sons; 2009.

[4] Pongpullponsak, A. Statistical quality control. 3rd ed. Bangkok: 2011.

[5] Paul Molin and $\text{Herve}'\text{Abdi}$. New Table and Numerical Approximations for Kolmogorov-Smirnov/Lilliefors/Van Soest Normality Test. (1998).

[6] Section 13 Kolmogorov-Smirnov test [Internet]. [updated 2014 Feb 27; cited 2014 Feb 27]. Available from: http://ocw.mit.edu/courses/mathematics/18-443 statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf

[7] Saad T. Bakir. A Nonparametric Shewhart-Type Quality Control Chart for Monitoring Broad Changes in a Process Distribution. International Journal of Quality, Statistics, and Reliability. Volume 2012, Article 147520, 10 pages.

[8] Su-Fen Yang, Jheng-Sian Lin and Smiley W Cheng. A new nonparametric EWMA Sing Control Chart. Expert Systems with Applications. 38 (2011): 6239–6243.

[9] Yindee.P. Determination of ARL of CUSUM control chart under the change of control chart parameters with SIEGMUND'S method [Special Problems]. Chon Buri:Burapha Univ; 2012.

Missing imputation using combining Lagrange interpolation and principal component multiple linear regression methods

Wirawan Puttamat[1]*, Wuttichai Srisodaphol [2] and Prem Junsawang [3]
[1]*Department of Statistics, Faculty of Science Khon Kaen University, Khon Kaen, Thailand, pwirawan@kkumail.com*
[2]*Department of Statistics, Faculty of Science Khon Kaen University, Khon Kaen, Thailand, wuttsr@kku.ac.th*
[3]*Department of Statistics, Faculty of Science Khon Kaen University, Khon Kaen, Thailand, prem@kku.ac.th*

Abstract

This paper focuses on missing-data imputation by combining Lagrange Interpolation (LI) and Principal Component Multiple Linear Regression (PCMR) methods. In this work, the proposed method called Weighted Average (WA) which is two combining approximation methods (LI and PCMR) is applied in different aspects. LI is applied in aspect of the known data within the same variable and PCMR is applied in aspect of the known data from relevant variables. Then, a missing value is handled by weighed average criteria from two estimated values. The weighted values are derived by minimum variance method. The proposed method is empirically evaluated on sales of frozen fruit data set with four proportions (5%, 10%, 15% and 20%) of randomly missing data. The results show that the Root Mean Square Error (RMSE) values of the proposed method are less than five traditional imputation methods including Mean, Mode, Multiple Linear Regression (MR), LI and PCMR for all proportions of randomly missing data.

Keywords: Missing data, Lagrange interpolation, multiple linear regression, principal component analysis

*Corresponding Author
E-mail Address: pwirawan@kkumail.com

1. Introduction

Survey or research need to collect the huge of data for analyzing and processing in order to obtain the conclusion which lead to solve the research problem. In the collection of data from a survey may prone to have missing data problems [6]. Currently, there are three major problems that may arise when dealing with missing data. First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation to the irregularity in data structure and the impossibility of using standard software. Third, the most important, there may be bias due to systematic differences between observed and unobserved data. It is important to prevent or reduce the missing data, but in fact, the missing data is difficult to control. Therefore, the estimation of missing data is very important before applying, and it helps to reduce the bias [5].

Generally, three types of missing data were made by Rubin [9] including 1) Missing At Random (MAR); if the observed data are missing independently of unobserved data, 2) Missing Completely At Random (MCAR) if observed data are missing independently of both observed and unobserved data and 3) Missing Not At Random (MNAR) if missing observations related to values of unobserved data.

A straightforward approach for handling the missing problem is to replace a missing value by a statistical value such as mean or mode value. However, the replacement method affects the original distribution of data which is distorted. Currently, many missing imputation researches have been developed and studied. In 2007, Dankyu [1] proposed a robust least square estimation with principal components method based on the local least square imputation (LLSimpute) [4]. The basic idea of this method was to employ quantile regression for missing value estimation by using the estimated principal components of a selected set of similar genes. In 2007, Norazian [7] compared the result of missing imputation of linear interpolation with the result of mean replacement methods on environmental data set. In 2009, Nunlaong [2] proposed the combination of three missing value estimations by using weighted average. These three methods were weighted by using Equally Weight (EW), Least Absolute Value (LAV) and Minimum Variance (MV). In 2012, Sompomgnawakij [10] proposed two composite imputation methods. The composite imputation methods are KRMI1 and KRMI2. They combined three single imputation methods; k-Nearest Neighbor Imputation method (KNNI), Regression Imputation method (RI) and Multiple Imputation method (MI). KRMI1 was weighted by using LAV and KRMI2 was weighted by using EW.

In this paper, we propose a method of Weighted Average (WA) between LI and PCMR by minimum variance method for missing data imputation. The efficiency of our propose method is compared with Mean, Mode, MR, LI and PCMR methods by using RMSE.

2. Background

2.1 Lagrange Interpolation

If there are n data values, a polynomial of degree $n+1$ can be found that will pass through all the points. The Lagrange interpolation provides a convenient alternative to solve the simultaneous equations that result from requiring the polynomials to pass through the data values. The Lagrange interpolation formula [6] is summarized as follows.

Let $\left\{(x_k, y_k) \in \mathbb{R}^2 \middle| \text{for } k = 1,2,\ldots,n\right\}$ be the set of n data points

$$P(y) = \sum_{i=1}^{n} L_i(y) \qquad (1)$$

where

$$L_i(y) = y_i \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

$P(y)$ is the estimated value from data n order.

2.2 Principal Component Analysis

Let Σ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, ..., X_p]$. The covariance matrix Σ has the eigenvalue-eigenvector pairs

$$(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), ..., (\lambda_p, \mathbf{e}_p)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the k^{th} principal component is given by [8]

$$\begin{aligned} PC_k &= \mathbf{e}_k'\mathbf{X} \\ &= e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kp}X_p,\ k = 1,2,\ldots,p \end{aligned} \qquad (2)$$

2.3 Minimum Variance Method

Let $\hat{Y}_1$ and $\hat{Y}_2$ be estimated values from estimation method 1 and method 2, respectively that provide the minimum error.

Let $e_{\hat{Y}_1}$ and $e_{\hat{Y}_2}$ be errors from estimation method 1 and method 2, respectively [11].

The weighted average method formula is

$$CF = w\hat{Y}_1 + (1-w)\hat{Y}_2 \qquad (3)$$

where

CF is the estimated value from weighted average method

$\hat{Y}_1$ is the estimated value from method 1

$\hat{Y}_2$ is the estimated value from method 2

w is the weighted value by minimum variance method.

The weighted value by minimum variance method are estimated by using

$$\hat{w} = \frac{\hat{\sigma}^2_{e_{\hat{Y}_2}} - \hat{\sigma}_{e_{\hat{Y}_1} e_{\hat{Y}_2}}}{\hat{\sigma}^2_{e_{\hat{Y}_1}} + \hat{\sigma}^2_{e_{\hat{Y}_2}} - 2\hat{\sigma}_{e_{\hat{Y}_1} e_{\hat{Y}_2}}}.$$

where

$\hat{\sigma}^2_{e_{\hat{Y}_1}}$ is the estimated value of variance of error for method 1.

$\hat{\sigma}^2_{e_{\hat{Y}_2}}$ is the estimated value of variance of error for method 2.

$\hat{\sigma}_{e_{\hat{Y}_1} e_{\hat{Y}_2}}$ is the estimated value of covariance of error between method 1 and 2.

2.4 Multiple Linear Regressions Analysis

Multiple Linear Regressions is a statistical method used to examine the relationship between one dependent variable Y and two or more independent variables X. The multiple linear regression formula is summarized as follow.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (4)$$

The regression coefficients β_0, $\beta_1, \ldots,$ β_k are estimated by using the method of Ordinary Least Squares method (OLS).

3. Research Methodology

The proposed method is imputing missing value in two directions. Firstly, we impute the missing value by using the data within the same variable with LI method. Secondly, we impute the missing value by using the data in the other variables with PCMR method.

Our proposed method deal with the real data set that is sales of frozen fruit (Thousand unit sales) between March 31st, 2000 and December 31st, 2007 which are obtained from UCI machine learning repository [3]. There are 6 variables, Unit Sales (Y), Price (X_1), Advertising expenditures (X_2), Competitors' Price (X_3) Income (X_4) and Population (X_5), with sample sizes is 144.

The proposed method consists of four steps as follows.

3.1 Determine the number of missing data

We determine the numbers of missing data in four proportions of randomly missing data as 5%, 10%, 15% and 20%. The numbers of the missing values are 7, 14, 22, and 29, respectively.

3.2 Random the position of missing data

Let Y be a variable with incomplete cases and X be a variable with complete cases. We random the positions of Y by denoted Y^* as a variable that specify the position of Y. The variable Y^* is distributed as binomial distribution, with p is the probability of the data in Y that will be missed ($p = 0.05, 0.1, 0.15, 0.2$).

There are two possible outcomes for Y^* as zero or one. If values in Y^* is zero, the position of Y will be missed.

3.3 Individual methods for estimating of missing values

3.3.1 Multiple Linear Regressions

Multiple linear regressions is the estimation of missing values using complete cases by OLS method. Let response variable be a variable with incomplete cases and independent variables be a variable with complete cases. The process of the imputation the missing values by MR is divided into two cases.

3.3.1.1 The case of independent variables $X_1, X_2, \dots, X_5$ are independent

This case, variance inflation factors (VIF) value of $X_1, X_2, \dots, X_5$ less than or equal 10 and correlation coefficient (r) value less than or equal 0.8. The steps of this case as follows.

1) *Calculating regression coefficients*

We calculate regression coefficients of Y on X using complete cases by OLS. We have that

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* \tag{6}$$

where

$\mathbf{X}^*$ is matrix of independent variables with complete case

$\mathbf{Y}^*$ is vector of response variable with complete case.

2) Computing the fitted value

The formula for impute the missing values by MR is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_5 x_5 \tag{7}$$

where

$\hat{Y}$ is the estimated value

$x_1, x_2, \dots, x_5$ are the observed values of $X_1, X_2, \dots, X_5$

$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_5$ are estimated regression coefficients.

3.3.1.2 The case of independent variables $X_1, X_2, \dots, X_5$ are dependent

This case, variance inflation factors (VIF) value of $X_1, X_2, \dots, X_5$ more than 10 and correlation coefficient (r) value more than 0.8. The steps of this case as follows.

1) Principal Component Analysis

Since the independence variables are dependent, we have to group the variables by using PCA. The k^{th} principal component is given by (2).

2) Calculating regression coefficients

We calculate regression coefficients of Y on X using complete cases by OLS. We have that

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{PC}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{PC}^{*\prime}\boldsymbol{Y}^* \tag{8}$$

where

$\mathbf{PC}^*$ is matrix of independent variables from PCA with complete case

$\boldsymbol{Y}^*$ is vector of response variable with complete case.

3) Computing the fitted value

The formula for impute the missing values by MR is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 PC_1 + \hat{\beta}_2 PC_2 + \dots + \hat{\beta}_k PC_k \tag{9}$$

where

$\hat{Y}$ is the estimated value

PC_k is the k^{th} principal component

$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are estimated regression coefficients

k is number of variable from PCA.

3.3.2 Lagrange Interpolation

This step, we impute the missing value by Lagrange Interpolation method with three data values and three position $(n = 3)$ of the complete cases in Y. The Lagrange interpolation formula is (1).

3.4. Weighted Average

Let $\hat{Y}$ be an estimated value from MR or PCMR depending on the correlated independence variables and $P(y)$ be an estimated value from data n order. The weighted between $\hat{Y}$ and $P(y)$ by using minimum variance method is defined as

$$WA = \hat{w}\hat{Y} + (1-\hat{w})P(y). \tag{10}$$

4. Results and Discussions

In this paper, we use the Root Mean Square Error (RMSE) to evaluate the performances among the proposed (WA), Mean, Mode, MR, LI and PCMR methods for missing data estimation. The empirical results are classified by the proportions of randomly missing data (5% 10% 15% and 20%), for iterative 500 times. The results are showed in Tables 1-4, respectively.

From Tables 1-4, we observe that our proposed method is less RMSE than Mean, Mode, MR, LI and PCMR methods when proportion of randomly missing data are 5%, 10%, 15% and 20%. It turns out that our proposed method has preference over others methods.

Our proposed method is combined in two different aspects of the known data within the same variable and relevant variables. It is more accuracy than the traditional imputation methods using one aspect of variable. Moreover, if we replace the missing data by mode of the same variable, the performance of this method is very poor.

5. Conclusion

This paper, we present weighted average by minimum variance method between Lagrange Interpolation (LI) and Principal Component Multiple Linear Regression (PCMR) methods. The empirical results of this paper are showed by using data of sales of frozen fruit with the proportions of randomly missing data 5%, 10%, 15% and 20%, respectively. The RMSE is the criterion to compare the efficiency of the methods. It turns out that our proposed method has preference over Mean, Mode, MR, LI and PCMR methods.

Table 1: RMSE with mean $\pm sd$ for 5% randomly missing value on each method

Method	RMSE
Mean	4.420 ± 4.282
Mode	19.730 ± 1.212
MR	4.248 ± 4.106
LI	3.360 ± 1.898
PCMR	1.743 ± 1.205
WA	$\mathbf{1.356 \pm 1.098}$

Table 2: RMSE with mean $\pm sd$ for 10% randomly missing value on each method

Method	RMSE
Mean	4.182 ± 3.485
Mode	20.082 ± 1.428
MR	4.189 ± 3.213
LI	3.423 ± 11.106
PCMR	2.087 ± 1.089
WA	$\mathbf{1.592 \pm 0.945}$

Table 3: RMSE with mean $\pm sd$ for 15% randomly missing value on each method

Method	RMSE
Mean	4.919 ± 1.362
Mode	19.850 ± 1.875
MR	4.905 ± 1.338
LI	3.506 ± 1.763
PCMR	1.861 ± 0.541
WA	$\mathbf{1.087 \pm 0.208}$

Table 4: RMSE with mean $\pm sd$ for 20% randomly missing value on each method

Method	RMSE
Mean	2.004 ± 6.060
Mode	8.916 ± 10.416
MR	1.982 ± 5.983
LI	2.105 ± 8.664
PCMR	0.880 ± 1.972
WA	$\mathbf{0.788 \pm 0.870}$

Acknowledgements

This paper is partially supported by the Department of Statistics, Faculty of Science, Khon Kaen University, Thailand.

References

[1] Dankyu Y. Robust imputation method for missing values in microarray data. BMC Bioinformatics Journal. 2007;8(2):1-7.

[2] Nunlaong J. A Comparison of Missing Value Estimation Methods for Forecasting Models [Dissertation]. Bangkok : King Mongkut's University of Technology North Bangkok; 2009.

[3] The UCI machine learning repository [Internet]. 2013 [updated 2013 Jun 01; cited 2014 Jan 05]. Available from: https://www.archive.ics.uci.edu/ml/datasets.html.

[4] Kim H, Golub GH, Park H. Missing Value Estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics Journal. 2005; 21(2)**:**187-198.

[5] Kwang KJ. Variance Estimation after Imputation [Dissertation]. United States: Iowa State Univ; 2000.

[6] Sunitha L, BallRaju M, SasiKiran J. Data Mining : Estimation of Missing Values Using Lagrange Interpolation Technigue. International Journal of Advanced Research in Computer Engineering & Technology. 2013; 4(2): 1579-1582.

[7] Norazian MN. Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set, Proceeding of the 2nd Malaysian Technical Universites Conference on Engineering and Technology; 2008 March 8-10; Perlis, Malaysia. 2007. p. 1-4.

[8] Johnson A, Wichern W. Applied Multivariate Statistical Analysis. 5th ed. New Jersey: Prentice-Hall; 2002.

[9] Rubin DB. Inference with missing data. Biometrika Journal. 1976; 63(1): 581-592.

[10] Sompomgnawakij S. A Comparison of Composite Imputation Methods [Dissertation]. Bangkok: Kasetsart Univ; 2012.

[11] Thompolos NT. Applied Forecasting Methods. New Jersey: Prentice-Hall; 2002.

Ratio estimators of a population mean in simple random sampling using two auxiliary variables

Jidapa Nittayanon[1*] and Supunnee Ungpansattawong[2]

[1]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, nam_yaroki@hotmail.com*

[2]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, supunnee@kku.ac.th*

Abstract

In this study we proposed the ratio estimators for population mean by using two auxiliary variable in simple random sampling and compared with other four estimators; ratio estimators in Sisodia and Dewide (1981), Singh and Kakran (1993), Upadhyaya and Singh (1999), Kadilar and Cingi (2005). A mean of squared error (MSE) is using for MSE is using for decide that is better the other. The result shows that the proposed estimator has less MSE than other estimators. This result is also supported by a numerical illustration.

Keywords: Ratio estimators, auxiliary variable, mean square error, simple random sampling, regression

*Corresponding Author
E-mail Address: nam_yaroki@hotmail.com

1. Introduction

The classical ratio estimator for the population mean $\bar{Y}$ is defined by Kadilar and Cingi (2004) [4]

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}}\bar{X} \tag{1}$$

Where it is assumed that the population mean $\bar{X}$ of auxiliary variable x is know.

MSE of classical ratio estimator was given by :

$$MSE\left(\bar{y}_r\right) = \frac{(1-f)}{n}\left[R^2S_x^2 - 2R\rho_{xy}S_yS_x + S_y^2\right] \tag{2}$$

In 1981 Sisodia and Dwivedi [6] used the population coefficient of variation of auxiliary variate C_x is know. Sisodia and Dwivedi suggested a modified ratio estimator for $\bar{Y}$ as

$$\bar{y}_{SD} = \bar{y}\left[\frac{\bar{X}+C_x}{\bar{x}+C_x}\right] \tag{3}$$

MSE of this estimator was given by :

$$MSE(\bar{y}_{SD}) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + C_x^2\alpha\left(\alpha - 2K\right)\right] \tag{4}$$

Where C_y is the population coefficient of variation of variate of interest.

$$\alpha = \frac{\bar{X}}{\bar{X}+C_x} \quad , \quad K = \rho_{xy}\frac{C_y}{C_x}$$

Later, In 1993 Singh and Kakran [7] developed ratio estimator for $\bar{Y}$ as

$$\bar{y}_{SK} = \bar{y}\left[\frac{\bar{X}+\beta_2\left(x\right)}{\bar{x}+\beta_2(x)}\right] \tag{5}$$

MSE of this estimator was given by :

$$MSE(\bar{y}_{SK}) = \frac{1-f}{n}\bar{Y}^2\left(C_y^2 + C_x^2\delta(\delta - 2\rho\frac{C_y}{C_x})\right) \tag{6}$$

Where $\beta_2(x)$ is the population coefficient of kurtosis of auxiliary variate.

$$\delta = \frac{\bar{X}}{\bar{X}+\beta_2(x)}$$

Upadhyaya and Singh (1999) [5] consider both coefficient of variation and kurtosis in their estimator. Suggested the following ratio estimator for $\bar{Y}$ as

$$\bar{y}_{US_1} = \bar{y}\left[\frac{\bar{X}\beta_2\left(x\right)+C_x}{\bar{x}\beta_2\left(x\right)+C_x}\right] \tag{7}$$

$$\bar{y}_{US_2} = \bar{y}\left[\frac{\bar{X}C_x+\beta_2\left(x\right)}{\bar{x}C_x+\beta_2\left(x\right)}\right] \tag{8}$$

Where $\beta_2(x)$ is the population coefficient of kurtosis of auxiliary variate.

C_x is the population coefficient of variation of variate of interest.

And the MSE of two estimators was given by :

$$MSE(\bar{y}_{US_1}) = \frac{1-f}{n}\bar{Y}^2\left[C_y^2 + \omega C_x^2(\omega - 2K)\right] \tag{9}$$

$$MSE(\bar{y}_{US_2}) = \frac{1-f}{n}\bar{Y}^2\left[C_y^2 + \theta(\theta - 2K)C_x^2\right] \tag{10}$$

Where $\omega = \frac{\bar{X}\beta_{2(x)}}{\bar{X}\beta_{2(x)}+C_x}$, $\theta = \frac{\bar{X}C_x}{\bar{X}C_x+\beta_{2(x)}}$, $K = \rho_{xy}\frac{C_y}{C_x}$

Supposed that an auxiliary variable x_i , correlated with variate of interest y_i , is obtained for each unit in the sample which is drawn by simple random sampling and that the population mean $\bar{X}$ of the x_i , is know. The regression estimate of $\bar{Y}$, The population mean of the y_i , is

$$\bar{y}_{reg_1} = \bar{y} + b(\bar{X} - \bar{x}) \tag{11}$$

When these are two auxiliary variates as x_1 and x_2 , the regression estimate of $\bar{Y}$ will be [3]

$$\bar{y}_{reg_2} = \bar{y} + b_1(\bar{X}_1 - \bar{x}_1) + b_2(\bar{X}_2 - \bar{x}_2) \tag{12}$$

Where $b_1 = \dfrac{S_{x_1 y}}{S_{x_1}^2}$, $b_2 = \dfrac{S_{x_2 y}}{S_{x_2}^2}$

Abu-Dayyeh et al. (2003) [1] using information of two auxiliary variable x_1 and x_2 . We propose the following of estimator of the population mean :

$$\bar{y}_a = \bar{y}\left(\frac{\bar{x}_1}{\bar{X}_1}\right)^{a_1}\left(\frac{\bar{x}_2}{\bar{X}_2}\right)^{a_2} \tag{13}$$

$$\bar{y}_w = w_1\bar{y}\left(\frac{\bar{x}_1}{\bar{X}_1}\right)^{a_1} + w_2\bar{y}\left(\frac{\bar{x}_2}{\bar{X}_2}\right)^{a_2} \tag{14}$$

Where a_1, a_2 are real number and w_1 , w_2 denote the weight that satisfy the condition $w_1 + w_2 = 1$
And MSE of this estimator was given by :

$$MSE_{\min}(\bar{y}_a) = \frac{(1-f)}{n}S_y^2\left[1 - \frac{\rho_{x_1y}^2 + \rho_{x_2y}^2 - 2\rho_{x_1y}\rho_{x_2y}\rho_{x_1x_2}}{1-\rho_{x_1x_2}^2}\right] \tag{15}$$

$$MSE_{\min}(\bar{y}_w) = \bar{Y}^2\frac{f}{n}\begin{bmatrix} C_y^2 + w_1^{*2}\begin{pmatrix} a_1^2C_{x_1}^2 - a_2^2C_{x_2}^2 \\ -2a_1a_2\rho_{x_1x_2}C_{x_1}C_{x_2} \end{pmatrix} \\ +a_2^2C_{x_2}^2 + 2a_2\rho_{x_2y}C_yC_{x_2} \\ -2w_1^*\begin{pmatrix} a_2^2C_{x_2}^2 - a_2^2C_{x_2}^2 - a_1\rho_{x_1y}C_yC_{x_1} \\ +a_2\rho_{x_2y}C_yC_{x_2} - a_1a_2\rho_{x_1x_2}C_{x_1}C_{x_2} \end{pmatrix} \end{bmatrix} \tag{16}$$

Where $w_1^* = \dfrac{a_2^2C_{x_2}^2 + a_2\rho_{x_2y}C_{x_2}C_y - a_1\rho_{x_1y}C_{x_1}C_y - a_1a_2\rho_{x_1x_2}C_{x_1}C_{x_2}}{a_1^2C_{x_1}^2 + a_2^2C_{x_2}^2 - 2a_1a_2\rho_{x_1x_2}C_{x_1}C_{x_2}}$

Later, Kadilar and Cingi (2005) [3] using two auxiliary variates given in (13) instead of $\bar{y}$ in regression (12) . suggested a modified ratio estimators for $\bar{y}$ as

$$\bar{y}_{pr} = y\left(\frac{\bar{X}_1}{\bar{x}_1}\right)^{\alpha_1}\left(\frac{\bar{X}_2}{\bar{x}_2}\right)^{\alpha_2} + b_1\left(\bar{X}_1 - \bar{x}_1\right) + b_2\left(\bar{X}_2 - \bar{x}_2\right) \tag{17}$$

Where α_1 and α_2 were real number.

The study found , They can obtain minimum MSE of the suggested estimate using the optimal equation of α_1^* and α_2^* as

$$MES_{\min}(\bar{y}_{pr}) \cong \frac{(1-f)}{n}S_y^2\begin{bmatrix} 1 + C_1^2 + C_2^2 - 2C_1\rho_{x_1y} \\ -2C_2\rho_{x_2y} + 2C_1C_2\rho_{x_1x_2} \end{bmatrix} \tag{18}$$

Where $\alpha_1^* = \dfrac{S_y}{R_1S_{x_1}}\rho_1^*$, $\alpha_2^* = \dfrac{S_y}{R_2S_{x_2}}\rho_2^*$

$$C_1 = \rho_1^* + \rho_{x_1y} \text{ , } C_2 = \rho_2^* + \rho_{x_2y}$$

$$\rho_1^* = \frac{\rho_{x_1x_2}\left(\rho_{x_1y}\rho_{x_1x_2} - \rho_{x_2y}\right)}{1-\rho_{x_1x_2}^2}, \rho_2^* = \frac{\rho_{x_1x_2}\left(\rho_{x_2y}\rho_{x_1x_2} - \rho_{x_1y}\right)}{1-\rho_{x_1x_2}^2}$$

2. Research Methodology

In this paper, we proposed ratio estimators for population mean using two auxiliary variable in simple random sampling. Adaping the estimators given in (7)–(8) to the given in (17) . we proposed ratio estimators as

$$\bar{y}_{KU} = \bar{y}\left[\frac{\bar{X}_1\beta_2(x_1) + C_{x1}}{\bar{x}_1\beta_2(x_1) + C_{x1}}\right]^{\alpha_1}\left[\frac{\bar{X}_2\beta_2(x_2) + C_{x2}}{\bar{x}_2\beta_2(x_2) + C_{x2}}\right]^{\alpha_2} + b_1\left(\bar{X}_1 - \bar{x}_1\right) + b_2\left(\bar{X}_2 - \bar{x}_2\right) \tag{19}$$

Therefore, MSE or this estimators can be found using Taylor Series method defined is

$$h\left(\bar{x}_1, \bar{x}_2, \bar{y}\right) \cong \begin{bmatrix} h\left(\bar{X}_1, \bar{X}_2, \bar{Y}\right) + \dfrac{\partial h\left(\bar{x}_1, \bar{x}_2, \bar{y}\right)}{\partial \bar{x}_1}|_{\bar{X}_1,\bar{X}_2,\bar{Y}}\left(\bar{x}_1 - \bar{X}_1\right) \\ + \dfrac{\partial h\left(\bar{x}_1, \bar{x}_2, \bar{y}\right)}{\partial \bar{x}_2}|_{\bar{X}_1,\bar{X}_2,\bar{Y}}\left(\bar{x}_2 - \bar{X}_2\right) \\ + \dfrac{\partial h\left(\bar{x}_1, \bar{x}_2, \bar{y}\right)}{\partial \bar{y}}|_{\bar{X}_1,\bar{X}_2,\bar{Y}}\left(\bar{y} - \bar{Y}\right) \end{bmatrix} \tag{20}$$

Therefore, MSE or this estimators is

$$\bar{y}_{KU} \cong \begin{bmatrix} \bar{Y} + \bar{y}\left(-\dfrac{\bar{X}_1\beta_2(x_1) + C_{x_1}}{(\bar{x}_1\beta_2(x_1) + C_{x_1})^2} - b_1\right)\left(\dfrac{\bar{X}_2\beta_2(x_2) + C_{x_2}}{\bar{x}_2\beta_2(x_1) + C_{x_2}}\right)|_{\bar{X}_1,\bar{X}_2,\bar{Y}}(\bar{x}_1 - \bar{X}_1) \\ +\bar{y}\left(\dfrac{\bar{X}_1\beta_2(x_1) + C_{x1}}{\bar{x}_1\beta_2(x_1) + C_{x1}}\right)\left(-\dfrac{\bar{X}_2\beta_2(x_2) + C_{x2}}{(\bar{x}_2\beta_2(x_2) + C_{x2})^2} - b_2\right)|_{\bar{X}_1,\bar{X}_2,\bar{Y}}(\bar{x}_2 - \bar{X}_2) \\ +\left(\dfrac{\bar{X}_1\beta_2(x_1) + C_{x1}}{\bar{x}_1\beta_2(x_1) + C_{x1}}\right)^{\alpha_1}\left(\dfrac{\bar{X}_2\beta_2(x_2) + C_{x2}}{\bar{x}_2\beta_2(x_2) + C_{x2}}\right)^{\alpha_2}|_{\bar{X}_1,\bar{X}_2,\bar{Y}}(\bar{y} - \bar{Y}) \end{bmatrix}$$

Where $\bar{y}_{KU} = h\left(\bar{x}_1, \bar{x}_2, \bar{y}\right)$ and $\bar{Y} = h\left(\bar{X}_1, \bar{X}_2, \bar{Y}\right)$

$$(\bar{y}_{KU} - \bar{Y}) \cong \begin{bmatrix} -\left(\dfrac{\bar{Y}}{\bar{x}_1\beta_2(x_1) + C_{x_1}} - b_1\right)(\bar{x}_1 - \bar{X}_1) \\ -\left(\dfrac{\bar{Y}}{\bar{x}_2\beta_2(x_2) + C_{x2}} - b_2\right)(\bar{x}_2 - \bar{X}_2) + (\bar{y} - \bar{Y}) \end{bmatrix}$$

$$E(\bar{y}_{KU} - \bar{Y}) \cong E\left[(\bar{y} - \bar{Y}) - \left(A_{KU_1} - B_1\right)(\bar{x}_1 - \bar{X}_1) - \left(A_{KU_2} - B_2\right)(\bar{x}_2 - \bar{X}_2)\right]^2$$

The MSE of this estimator is as

$$MSE\left(\bar{y}_{KU}\right) = \frac{(1-f)}{n}S_y^2\begin{bmatrix} 1 + (A_{KU_1}W_1 + \rho_{x_1y})^2 + (A_{KU_2}W_2 + \rho_{x_2y})^2 \\ -2(A_{KU_1}W_1 + \rho_{x_1y})\rho_{x_1y} - 2(A_{KU2}W_2 + \rho_{x_2y})\rho_{x_2y} \\ +2(A_{KU_1}W_1 + \rho_{x_1y})(A_{KU2}W_2 + \rho_{x_2y})\rho_{x_2x_2} \end{bmatrix} \tag{21}$$

Where $A_{KU} = \dfrac{\bar{Y}}{\bar{x}_1\beta_2\left(x_1\right) + C_{x_1}}, A_{KU} = \dfrac{\bar{Y}}{\bar{x}_2\beta_2\left(x_2\right) + C_{x_2}}$

$$W_1 = \frac{S_{x_1}}{S_y} \text{ , } W_2 = \frac{S_{x_2}}{S_y}$$

3. Result

Numerical Illustration

We have used the data of Daroga Singh [2] in this section. The following values were obtained using the whole data set :

Table 1: Data Statistics

$N = 170$	$S_{x_1} = 733.1407$	$\rho_{x_1 y} = 0.4453$
$n = 34$	$S_{x_2} = 150.5059$	$\rho_{x_2 y} = 0.9801$
$\bar{x}_1 = 856.4118$	$S_y = 150.215$	$W_1 = 4.8806$
$\bar{x}_2 = 208.8824$	$S_{x_1 y} = 47599.99$	$W_2 = 1.0021$
$\bar{y} = 199.4412$	$S_{x_2 y} = 21506.35$	$\beta_2(x_1) = 12.2697$
$\alpha_1 = -0.4857$	$\alpha_2 = -0.2133$	$\beta_2(x_2) = 0.0975$

Table 2**:** MSE value of ratio estimators

Estimators		MSE
Proposed	$\bar{y}_{KU}$	60.0618
Sisodia and Dwidei [1]	$\bar{y}_{SD}$	678.5226
Singh and Kakran [2]	$\bar{y}_{SK}$	679.2601
Upadhyaya and Singh [4]	$\bar{y}_{US_1}$	679.2871
Kadilar and Cingi [3]	$\bar{y}_{pr}$	60.0518

From table 2, we observe that the proposed estimator $\bar{y}_{KU}$ have a smaller MSE value among all ratio estimators given in section 1.

4. Conclusion

We develop some ratio estimators from the ratio estimators in Upadhyaya and Singh (1999) and Kadilar and Cingi (2005) using two auxiliary variable. We show the proposed estimators $\bar{y}_{KU}$ have a smaller MSE than the ratio estimators in Sisodia and Dwidei (1981), Singh and Kakran (1993), Upadhyaya and Singh (1999) and Kadilar and Cingi (2005). In addition, we support this theoretical result by a numerical example illustration. In future, we hope to expend the information presented here to product estimators in simple random sampling by using two auxiliary variables.

Acknowledgements

The authors are thankful to the editor Asso.Dr. Supunnee Ungpansattawong for their valuable suggestions for the improvement of the paper.

References

[1] Abu-Dayyeh et al, Some estimators of a finite population mean using auxiliary information. Journal of Applied Mathematics and Computation. 2003; 139 : 287-298

[2] Daroga Singh F.S. Chaudhary, Theory and Analysis simple Survey desins. Indian Agricultural Statistics Research Tnstitute, New Delhi India. 1986: 177

[3] Kadilar C. and Cingi H.,A new estimator using two auxiliary variable. Journal of Applied Mathematics and Computation. 2005; 162: 901-908

[4] Kadilar C. and Cingi H, Ratio estimator in simple random sampling. Journal of Applied Mathematics and Computation. 2004; 151: 893-902

[5] Lakshmi N. Upadhaya and Housila P. Singh, Use of transformed auxiliary variable in estimating the finite population mean. Biometrical Journal., 1999; 41(5): 627-636

[6] Sisodia, B.V.S and Dwividi, V.K., A Modified ratio estimator using coefficient of variation of auxiliary variable. Journal of the Indian society of Agricultural Statistics. 1981; 33(1): 13-18

[7] Singh H.P., Tailor R., and Kakran, M.S., An Improved estimator of Population mean using power transformed. Journal of the Indian society of Agricultural Statistics. 2004; 58(2): 223-230

Extending Zelterman's estimator

Krisana Lanumteang[1*]and Dankmar Böhning [2]
[1]*Section of Statistics, Maejo University, SanSai, ChiangMai 50290, THAILAND, k.lanumteang@mju.ac.th*
[2]*School of Mathematics, University of Southampton Highfield, Southampton, SO17 1BJ,*
UK, D.A.Bohning@soton.ac.uk

Abstract

Two new estimators of population size are introduced in this paper. The new estimators are developed as a modification of Zelterman's estimator. Simulation technique is applied to study the performance of the proposed estimators. The simulation results show that the modified Zelterman's estimators can improve the efficiency of the original Zelterman's estimator. Overall, for a particularly small population size the proposed estimators give a smaller relative bias, relative variance and relative mean square error. This is in comparison to Zelterman's estimator for both homogeneity and heterogeneity Poisson capture probabilities.

Keywords: Zelterman's estimator, capture-recapture model, population size, Poisson capture probability

*Corresponding Author
E-mail Address: k.lanumteang@mju.ac.th

1. Introduction

Based on capture-recapture models, the identifying system generally provides a count $Y_i > 0$ of how many times the individual i^{th} has been captured, for $i = 1, 2, \ldots, n$ and $Y_i = 0$ denotes unobserved cases in the system for $i = n+1, n+2, \ldots, N$. Hence, it can be written that the total number of a target population (N) consists of an observed part (zero-truncated) of size n and unobserved part of unknown size $f_0 = N- n$ as well as $N = n+f_0$. In order to investigate an estimate of N based on available sample $Y_1,Y_2,\ldots,Y_n$, it is usually required to assume a model for the capture probability of Y, p_{j} = Prob(Y=j). A typical example of such a model is the Poisson or the binomial distribution.

In 1988, Zelterman proposed an estimator of population size based on the zero-truncated Poisson approach, which is widely used nowadays, particularly in the social sciences, see [1]. Suppose that the capture probability of individuals $p_j = (e^{-\lambda}\lambda^j)/(j!(1-e^{-\lambda}))$; j = 1,2,3,... where λ is the location parameter of the zero-truncated Poisson, then we have that $\lambda = (j+1)p_{j+1}/p_j$. Replacing the probability functions p_j and p_{j+1} with their association observed frequency f_j/N and f_{j+1}/N, respectively, an estimation of λ can be simply obtained as $\hat{\lambda} = (j+1)f_{j+1}/f_j$. Finally, this can lead to the estimator of population size by taking $\hat{p}_0 = \exp\{-(j+1)f_{j+1}/f_j\}$ into Horvitz-Thompson equation $\hat{N}_{HT} = n/(1- \hat{p}_0)$, see [2] for review. Therefore, the estimator of population size in terms of Zelterman approach is $\hat{N}_{Zel} = n/(1-\exp\{-(j+1)f_{j+1}/f_j\})$. In practice, Zelterman suggested that j is usually chosen as one or two, due to the fact that in many capture-recapture studies the majority of counts are contributing to f_1 and f_2. These counts might be more similar to those individual that were not observed, f_0 . For j = 1, this achieves the Zelterman's estimator of the form $\hat{N}_{Zel} = n/(1-\exp\{-2f_2/f_1\})$.

It is clearly seen that Zelterman's formula is very simple to understand and to use. This is perhaps one of the reasons why it has been widely used in many fields in particular social section [3]. In addition, using only f_1 and f_2 makes the estimator more robust in the sense that there are no effects from the fluctuation of higher frequency counts. It can also be thought of as being less sensitive to model violations since it might be applied for other count data behaving like the Poisson distribution. However, there are also some critical remarks on using Zelterman's estimator. Firstly, it uses limited information from observed counts to achieve an estimation of population size. Using only f_1 and f_2 to estimate the location parameter, $\hat{\lambda}_{Zel} = 2f_2/f_1$, might not be suitable, especially for long tail data. Another issue is that Zelterman's estimator seems to provide overestimation. This issue has been discussed in many simulation studies such as [3] and [4]. In addition, the Zelterman's estimator also typically gives a lager variance compared with other estimators. Hence, we should be concerned about how to overcome with these critical remarks. This is the motivation for the study of an extension of Zelterman's estimator. Two modified Zelterman's estimators will be examined in the next section.

2. Proposed Estimators

The first modified form of Zelterman's estimator is provided by keeping the original idea of using only f_1 and f_2 to estimate the location parameter λ of the Poisson probability. We consider $\hat{N}_{Zel} = n/(1-\exp\{-2f_2/f_1\})$ where the denominator term is fixed. The overestimation bias yielded in the Zelterman's estimator therefore generates from only the numerator term, *n*. Hence, in order to reduce the overestimation of Zelterman's estimator, it is necessary to decrease the size of the numerator term. In this modification form, $n = f_1 + f_2 + \ldots + f_m$ (*m* is the maximum times of identifying individuals) is decreased to have only the term $f_1 + f_2$, and then the other counts will be added up again after computing the main part $(f_1+f_2)/(1-\exp\{-2f_2/f_1\})$. Finally, the first modified Zelterman's estimator is given as:

$$\hat{N}_{ZelM1} = \frac{f_1 + f_2}{1-\exp\{-2f_2/f_1\}} + f_3 + f_4 + \ldots + f_m \qquad (1)$$

$$= \frac{f_1 + f_2}{1-\exp\{-2f_2/f_1\}} + (n - f_1 + f_2). \qquad (2)$$

Furthermore, consider the drawback of limited use of available information of frequency counts of the Zelterman's estimator. Assuming the Poisson model, we do not only have $\lambda = (j+1)p_{j+1}/p_j$, but also that $\lambda = \sum_{j=1}^{k}(j+1)p_{j+1} / \sum_{j=1}^{k} p_j$; $k = 1, 2, \ldots, m-1$. Consequently, substituting associated observed frequency count f_j for p_j leads to a family of estimation λ as follows:

$$\hat{\lambda}_1 = \frac{2f_2}{f_1}, \qquad (3)$$

$$\hat{\lambda}_2 = \frac{2f_2 + 3f_3}{f_1 + f_2}, \qquad (4)$$

$$\hat{\lambda}_3 = \frac{2f_2 + 3f_3 + 4f_4}{f_1 + f_2 + f_3}, \qquad (5)$$

$$\vdots$$

$$\hat{\lambda}_{m-1} = \frac{2f_2 + 3f_3 + 4f_4 + \ldots + mf_m}{f_1 + f_2 + \ldots + f_{m-1}}. \qquad (6)$$

Recall again the Zelterman's estimator $\hat{N}_{Zel} = n/(1-\exp\{-2f_2/f_1\})$, it can be also written as

$$\hat{N}_{Zel} = \frac{n}{1-\exp(\frac{2f_2}{f_1})}$$

$$= \frac{f_1 + f_2 + \ldots + f_m}{1-\exp(\frac{2f_2}{f_1})}$$

$$= \frac{f_1 + f_2}{1-\exp(\frac{2f_2}{f_1})} + \frac{f_3}{1-\exp(\frac{2f_2}{f_1})} + \ldots$$

$$+ \frac{f_m}{1-\exp(\frac{2f_2}{f_1})}. \qquad (7)$$

Then, if we use a series of estimators $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \ldots, \hat{\lambda}_{m-1}$ for each linear combination in (7) in stage of using only $\hat{\lambda}_{Zel} = \frac{2f_2}{f_1}$, the second modified Zelterman's form is now obtained as follows:

$$\hat{\lambda}_{ZelM2} = \frac{f_1 + f_2}{1-\exp\{-\frac{2f_2}{f_1}\}} + \frac{f_3}{1-\exp\{-\frac{2f_2 + 3f_3}{f_1 + f_2}\}}$$

$$+ \ldots + \frac{f_m}{1-\exp\{-\frac{2f_2 + 3f_3 + \ldots + mf_m}{f_1 + f_2 + \ldots + f_{m-1}}\}}. \qquad (8)$$

Furthermore, in order to evaluate an efficiency of the two modification estimators in terms of accuracy and precision, the numerical study will be investigated in the next section.

3. Simulation Study

3.1 Aim of study

The main aim of this simulation is to study the performance of proposed estimators and to compare their behaviours with the other well-known estimators, which will also be examined include: maximum likelihood, Zelterman's estimator, Chao's lower bound estimator, and Turing methods.

3.2 Scope of Study

3.2.1 The data was generated by the Monte Carlo technique running on program MINITAB, each condition was repeated 1,000 times.

3.2.2 There were two type of target populations, homogeneity and heterogeneity cases, which were generated from Poisson models and mixture Poisson models respectively.

3.2.3 The population size of each model was 20, 50, 100 and 500.

3.2.4 To achieve a better judgment of the proposed estimator we include the following estimators in the comparison:

1)Maximum likelihood estimator

$$\hat{N}_{MLE} = \frac{n}{1-\exp(-\hat{\lambda}_{MLE})}; \qquad (9)$$

where $\hat{\lambda}_{MLE}$ is the maximum likelihood estimator of the unknown parameter under the zero truncated Poisson distribution, see [5].

2) Chao's estimator

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}, \text{ see [6].} \qquad (10)$$

3) Zelterman's estimator

$$\hat{N}_{Zel} = \frac{n}{1-\exp(-\frac{2f_2}{f_1})}, \text{ see [1].} \qquad (11)$$

4) Turing's estimator

$$\hat{N}_{Tu} = \frac{n}{1-\frac{f_1}{\sum_{j=1}^{m} jf_j}}, \text{ see [5].} \qquad (12)$$

3.2.5 The criteria of comparing the performance of each estimator were relative bias (*RBias*), relative variance (*RVar*) and relative mean squared error (*RMSE*).

Bias is commonly defined as the difference between the expected value of estimator and the population parameter. In other words, the small bias shows a high accuracy of estimation. Consequently, in order to determine which estimator provides a more appropriate value of estimation, it is traditionally considered which estimator gives a small value of bias or provides the relative bias approaching to zero. In addition, the positive value of *RBias* shows an overestimation whereas the negative *RBias* presents an underestimation.

Variance of each estimator commonly shows the variation of estimation in terms of the difference in average between an individual value of estimator and the expected value of estimator. Therefore, a small variance of estimator can imply that most individual values of estimators are closer to their mean. To be more precise, the relative variances here is calculated (*RVar*) as the relative ratio of the variance and the expected value of estimator squared.

Relative Mean square error (*RMSE*) shows the relative ratio of the difference between individual value of estimator and the true parameter of interest over the true parameter. The estimator, which provides a small value of *RMSE* normally indicates that this estimator shows the highest efficient estimation, on average is closest to the true parameter of interest.

3.3 Study Designs/Simulation Scenarios

This simulation study was carried out to investigate the performance of the two proposed estimators and to compare with other estimators by means of the Monte Carlo method. The total number of the target population for each condition was assumed at 20, 50, 100, and 500. Two types of populations, homogeneity and heterogeneity populations, were generated arising from Poisson distribution and two components mixture of Poisson distribution, respectively. The simulation procedure of each scenario is as follows:

3.3.1 Homogeneity Poisson Models

There were fifteen cases of homogeneity Poisson model, which were generated under a Poisson distribution with parameter $\lambda \in \{0.5, 0.6, 0.7,\ldots,1.9, 2.0\}$.

3.3.2 Heterogeneity Poisson Model

There were fifty-four cases of a Poisson mixture model, which were generated arising from 50 percent: 50 percent mixture of Poisson distribution as well as the two components of contaminated Poisson distribution, $0.5Poi(\lambda)+0.5Poi(\mu)$. The Poisson parameter of each component was varied in this ways:

$\lambda \in \{0.5, 0.6, 0.7,\ldots, 1.0\}$

$\mu \in \{1.5, 1.6, 1.7,\ldots, 2.0, 3.0, 4.0, 5.0\}$.

An amount of contamination of these mixture Poisson is given as the difference between λ and μ as well as $\Delta = |\lambda-\mu|$. The highest amount of contamination is the case where $\Delta = 4.5$ ($\lambda = 0.5$ and $\mu = 5.0$), whereas $\Delta = 0.5$ ($\lambda = 1.0$ and $\mu = 1.5$) is the smallest cases of contamination.

The simulation procedures were the following: 1) Generate the population of size N for each scenario 2) Count the frequencies of identifying individuals exactly j times; f_j, $j = 0,1,2,\ldots m,$ where $N = f_0 + f_1 + f_2 + \ldots + f_m$ and $n = f_1 + f_2 + \ldots + f_m$ 3) Truncate f_0 and set as unknown 4) Estimate N by each method and then compute *RBias*, *RVar* and *RMSE* from 1,000 repeated times.

3.4 Results

3.4.1 Homogeneity Poisson Models

Overall, it is clearly seen from Table 1 that almost all estimators provide an overestimation for all conditions. There is only the Modified Zelterman's estimator 1 that gives an underestimation. Turing and MLE estimators, respectively, seem to show the highest performance of accuracy, which generally give the smallest *RBias* among the other methods for all population sizes. Interestingly, the Modified Zelterman's estimator 1 and 2 give less bias than the original Zelterman's estimator and the Modified Zelterman's estimator 1 tends to show an underestimation for a large population size. Moreover, it is found that both Poisson parameter and population sizes have an effect on the biasness. An increase in values of Poisson parameter and sizes of population leads to a slight decrease in biasness of all estimators.

The *RVar* of each estimator from simulation study are shown in Table 2. MLE and Turing's estimator tends to provide the minimum *RVar* for both small and large population sizes. Although the proposed estimators, Modified Zelterman's estimator 1 and 2, do not give the smallest value of *RVar*, these methods seem to provide a smaller *RVar* among the original Zelterman's estimator. For a small population size (N = 20 and 50), an increase in Poisson parameter has been responsible for a dramatic decline of *RVar* for all estimators. On the other hand, the *RVar* of estimation dropped slowly in particular cases of larger population size (N = 100 and 500).

The *RMSE* of each estimator for all scenarios are presented in Table 3. Similar to what was just stated in the investigation of *RVar*, MLE estimator also gives the least value of *RMSE* for all studied cases. According to a comparison between the two modification forms among the Zelterman's estimator, these modified forms seem to provide a smaller *RMSE* than the original one. Undoubtedly, a consequence of the increase in both Poisson parameter and population size is the significant decline in *RMSE* for all estimators.

3.4.2 Heterogeneity Poisson Model

The cases of heterogeneity population were generated in the difference among the two Poisson parameters of the mixture model as defined in subsection 3.3.2 . The results are presented as an overall of the studied conditions. However, only cases with 0.5Poi(0.5) + 0.5Poi(μ) is shown in this paper. Overall, under contaminated Poisson models almost all estimators yield values below the true population size. In contrast, only Zelterman's estimator remarkably overestimates in cases where there is a large amount of heterogeneity. Modified Zelterman's estimator 2 and Chao's estimator tends to give a smallest *RBias*, particularly

Table1: *RBias of population size estimators for counts drawn from Poi(λ)*

Estimator	λ				λ			
	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
	N = 20				*N* = 50			
Chao	0.0943	0.2006	0.1167	0.0651	0.2303	0.0724	0.0381	❸0.0229
MLE	0.1473	❷0.1286	❷0.0415	❷0.0146	❷0.1700	❸0.0396	❸ 0.0147	❷0.0095
Modified Zel I	❶0.0687	❸0.1588	❸0.0689	❸0.0222	❸0.2084	❶0.0274	❷-0.0137	-0.0244
Modified Zel II	❸0.0921	0.1946	0.1150	0.0664	0.2268	0.0708	0.0380	0.0236
Turing	0.1180	❶0.1164	❶0.0351	❶0.0118	❶0.1616	❷0.0361	❶ 0.0123	❶0.0072
Zelterman	❷0.0739	0.2346	0.1686	0.1197	0.2394	0.0849	0.0553	0.0417
	N = 100				*N* = 500			
Chao	0.1535	0.0318	0.0186	❸0.0098	0.0222	❸0.0065	0.0033	❸0.0022
MLE	❷0.1025	❸ 0.0146	❷0.0091	❷0.0036	❸0.0153	❶0.0006	❷0.0025	❷0.0007
Modified Zel I	❸0.1319	❷-0.0140	-0.0349	-0.0390	❶0.0002	-0.0398	-0.0514	-0.0480
Modified Zel II	0.1531	0.0310	❸0.0183	0.0101	0.0220	0.0068	❸0.0030	0.0023
Turing	❶0.1016	❶ 0.0134	❶0.0075	❶0.0025	❷0.0152	❷0.0011	❶0.0018	❶0.0006
Zelterman	0.1606	0.0384	0.0264	0.0184	0.0232	0.0086	0.0044	0.0041

❶,❷ and ❸ : giving the smallest value of *RBias*, respectively

Table 2: *RVar of population size estimators for counts drawn from Poi(λ)*

Estimator	λ				λ			
	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
	N = 20				*N* = 50			
Chao	❸0.4383	❸0.4173	❸0.1643	0.0690	❸0.5599	❸0.0827	❸0.0298	0.0127
MLE	❷0.4333	❷0.2452	❶0.0351	❶0.0144	❷0.3669	❶0.0344	❶0.0112	❶0.0053
Modified Zel I	0.4432	0.4254	0.1665	❸0.0681	0.5682	0.0849	0.0299	❸0.0120
Modified Zel II	0.4467	0.4442	0.1817	0.0762	0.5810	0.0922	0.0344	0.0142
Turing	❶0.4126	❶0.2388	❷0.0385	❷0.0160	❶0.3634	❷0.0357	❷0.0123	❷0.0057
Zelterman	0.4597	0.5427	0.2979	0.1601	0.6174	0.1071	0.0496	0.0275
	N = 100				*N* = 500			
Chao	❸0.2647	❸0.0333	❸0.0124	0.0055	❸0.0228	❸0.0054	❸0.0021	❸0.0009
MLE	❶0.1688	❶0.0150	❶0.0054	❶0.0023	❶0.0145	❶0.0027	❶0.0010	❶0.0004
Modified Zel I	0.2692	0.0343	❸0.0124	❸0.0051	0.0234	0.0055	❸0.0021	❸0.0009
Modified Zel II	0.2754	0.0379	0.0144	0.0062	0.0244	0.0062	0.0024	0.0011
Turing	❷0.1691	❷0.0164	❷0.0060	❷0.0027	❷0.0152	❷0.0030	❷0.0011	❷0.0005
Zelterman	0.2858	0.0444	0.0201	0.0108	0.0250	0.0070	0.0033	0.0018

❶,❷ and ❸ : giving the smallest value of *RVar*, respectively

Table 3: *RMSE of population size estimators for counts drawn from Poi(λ)*

Estimator	λ				λ			
	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
	N = 20				*N* = 50			
Chao	❷0.4472	0.4575	0.1780	0.0732	0.6130	0.0879	0.0313	0.0132
MLE	0.4550	❷0.2617	❶0.0368	❶0.0146	❷0.3959	❶0.0360	❶0.0115	❶0.0054
Modified Zel I	❸0.4479	❸0.4506	❸0.1712	❸0.0686	❸0.6117	❸0.0856	❸0.0301	❸0.0126
Modified Zel II	0.4552	0.4820	0.1949	0.0806	0.6324	0.0972	0.0359	0.0148
Turing	❶0.4265	❶0.2523	❷0.0397	❷0.0161	❶0.3895	❷0.0370	❷0.0124	❷0.0058
Zelterman	0.4651	0.5977	0.3263	0.1744	0.6747	0.1143	0.0527	0.0292
	N = 100				*N* = 500			
Chao	0.2883	❸0.0343	❸0.0128	❸0.0056	❸0.0233	❸0.0054	❸0.0021	❸0.0010
MLE	❶0.1793	❶0.0152	❶0.0055	❶0.0023	❶0.0147	❶0.0027	❶0.0010	❶0.0004
Modified Zel I	❸0.2866	0.0345	0.0136	0.0066	0.0234	0.0071	0.0047	0.0032
Modified Zel II	0.2989	0.0389	0.0148	0.0063	0.0249	0.0062	0.0024	0.0011
Turing	❷0.1794	❷0.0166	❷0.0061	❷0.0027	❷0.0154	❷0.0030	❷0.0011	❷0.0005
Zelterman	0.3116	0.0459	0.0208	0.0111	0.0255	0.0071	0.0033	0.0019

❶,❷ and ❸ : giving the smallest value of *RMSE*, respectively

Table 4**:** *RBias of population size estimators for counts drawn from 0.5Poi(0.5)+0.5Poi(μ)*

Estimator	μ					μ				
	1.5	2.0	3.0	4.0	5.0	1.5	2.0	3.0	4.0	5.0
	N = 20					N = 50				
Chao	❸ 0.0335	❷ 0.0197	❷-0.0589	❷-0.0685	❷-0.0519	❸-0.0210	❸-0.0889	❸-0.1175	❷-0.0920	❷-0.0489
MLE	-0.0925	-0.1669	-0.2465	-0.2696	-0.2881	-0.1222	-0.1944	-0.2525	-0.2787	-0.2918
Modified Zel I	❶-0.0038	❶-0.0163	❸-0.0905	❸-0.0970	❸-0.0765	-0.0605	-0.1279	-0.1499	❸-0.1164	❸-0.0678
Modified Zel II	❷ 0.0331	❸ 0.0282	❶-0.0470	❶-0.0554	❶-0.0380	❶-0.0147	❶-0.0790	❷-0.1033	❶-0.0749	❶-0.0289
Turing	-0.0857	-0.1450	-0.2195	-0.2396	-0.2564	-0.1087	-0.1738	-0.2241	-0.2479	-0.2612
Zelterman	0.0853	0.1142	0.0970	0.1531	0.2590	❷ 0.0155	❷-0.0389	❶-0.0052	0.1545	0.3824
	N = 100					N = 500				
Chao	❸-0.0784	❸-0.1114	❸-0.1359	❸-0.1200	❷-0.0925	❸-0.1002	❸-0.1285	❸-0.1548	❸-0.1498	❷-0.1228
MLE	-0.1507	-0.1995	-0.2539	-0.2781	-0.2923	-0.1563	-0.2050	-0.2580	-0.2808	-0.2921
Modified Zel I	-0.1185	-0.1516	-0.1693	-0.1455	❸-0.1123	-0.1413	-0.1698	-0.1891	0.1764	❸-0.1432
Modified Zel II	❷-0.0711	❷-0.1006	❷-0.1211	❷-0.1018	❶-0.0721	❷-0.0922	❷-0.1166	❷-0.1395	❷-0.1323	❶-0.1016
Turing	-0.1354	-0.1776	-0.2249	-0.2464	-0.2620	-0.1395	-0.1815	-0.2285	-0.2496	-0.2615
Zelterman	❶-0.0533	❶-0.0700	❶-0.0428	❶0.0722	0.2488	❶-0.0831	❶-0.0963	❶-0.0791	❶0.0004	0.1526
	1.0	1.5	2.5	3.5	4.5	1.0	1.5	2.5	3.5	4.5
	$\Delta = \lvert 0.5 - \mu \rvert$					$\Delta = \lvert 0.5 - \mu \rvert$				

❶,❷ and ❸ : giving the smallest value of *RBias*, respectively

Table 5: *RVar of population size estimators for counts drawn from 0.5Poi(0.5)+0.5Poi(μ)*

Estimator	μ					μ				
	1.5	2.0	3.0	4.0	5.0	1.5	2.0	3.0	4.0	5.0
	N = 20					N = 50				
Chao	❸0.2849	❸0.2203	❸0.1282	❸0.0925	❸0.0904	❸0.1146	❸0.0497	❸0.0329	❸0.0495	❸0.0807
MLE	❶0.0845	❶0.0397	❶0.0145	❶0.0090	❶0.0075	❶0.0290	❶0.0117	❶0.0053	❶0.0033	❶0.0030
Modified Zel I	0.2889	0.2230	0.1290	0.0929	0.0909	0.1172	0.0504	0.0332	0.0504	0.0822
Modified Zel II	0.3101	0.2421	0.1438	0.1037	0.1010	0.1265	0.0566	0.0378	0.0556	0.0892
Turing	❷0.0922	❷0.0460	❷0.0180	❷0.0122	❷0.0102	❷0.0314	❷0.0139	❷0.0067	❷0.0045	❷0.0041
Zelterman	0.4036	0.3792	0.3410	0.3325	0.3740	0.1741	0.0867	0.1012	0.2728	0.5927
	N = 100					N = 500				
Chao	❸0.0287	❸0.0173	❸0.0155	❸0.0175	❸0.0340	❸0.0046	❸0.0031	❸0.0019	❸0.0022	❸0.0030
MLE	❶0.0099	❶0.0055	❶0.0027	❶0.0019	❶0.0015	❶0.0020	❶0.0011	❶0.0005	❶0.0003	❶0.0003
Modified Zel I	0.0295	0.0175	0.0156	0.0178	0.0348	0.0047	❸0.0031	❸0.0019	❸0.0022	0.0031
Modified Zel II	0.0333	0.0202	0.0179	0.0204	0.0379	0.0053	0.0036	0.0023	0.0026	0.0036
Turing	❷0.0110	❷0.0065	❷0.0035	❷0.0026	❷0.0020	❷0.0022	❷0.0012	❷0.0006	❷0.0004	❷0.0004
Zelterman	0.0427	0.0302	0.0504	0.0824	0.2768	0.0064	0.0052	0.0049	0.0090	0.0210
	1.0	1.5	2.5	3.5	4.5	1.0	1.5	2.5	3.5	4.5
	$\Delta = \lvert 0.5 - \mu \rvert$					$\Delta = \lvert 0.5 - \mu \rvert$				

❶,❷ and ❸ : giving the smallest value of *RVar*, respectively

Table 6**:** *RMSE of population size estimators for counts drawn from 0.5Poi(0.5)+0.5Poi(μ)*

Estimator	μ					μ				
	1.5	2.0	3.0	4.0	5.0	1.5	2.0	3.0	4.0	5.0
	N = 20					N = 50				
Chao	❸0.2861	❸0.2207	❸0.1316	❸0.0972	❸0.0931	❸0.1150	❸0.0576	❶0.0467	❶0.0579	❷0.0831
MLE	❶0.0931	❷0.0676	❷0.0752	❷0.0816	❷0.0905	❷0.0440	❷0.0495	0.0690	0.0810	❸0.0882
Modified Zel I	0.2889	0.2233	0.1372	0.1023	0.0967	0.1209	0.0668	❸0.0557	❸0.0640	0.0868
Modified Zel II	0.3112	0.2428	0.1460	0.1067	0.1024	0.1267	0.0629	❷0.0485	❷0.0612	0.0900
Turing	❷0.0995	❶0.0670	❶0.0662	❶0.0697	❶0.0760	❶0.0432	❶0.0441	0.0569	0.0659	❶0.0723
Zelterman	0.4108	0.3923	0.3504	0.3559	0.4411	0.1743	0.0882	0.1012	0.2967	0.7389
	N = 100					N = 500				
Chao	❸0.0348	❶0.0297	❷0.0340	❷0.0319	❷0.0426	❸0.0146	❸0.0196	❸0.0258	❸0.0246	❶0.0181
MLE	❷0.0326	0.0453	0.0672	0.0792	0.0869	0.0265	0.0431	0.0670	0.0791	0.0856
Modified Zel I	0.0436	0.0405	❸0.0443	❸0.0390	❸0.0474	0.0246	0.0319	0.0377	0.0333	❸0.0236
Modified Zel II	0.0383	❷0.0303	❶0.0326	❶0.0307	❶0.0431	❷0.0138	❷0.0172	❷0.0217	❷0.0201	❷0.0139
Turing	❶0.0293	0.0380	0.0541	0.0633	0.0707	0.0217	0.0342	0.0528	0.0627	0.0688
Zelterman	0.0455	❸0.0351	0.0522	0.0876	0.3387	❶0.0133	❶0.0144	❶0.0111	❶0.0090	0.0443
	1.0	1.5	2.5	3.5	4.5	1.0	1.5	2.5	3.5	4.5
	$\Delta = \lvert 0.5 - \mu \rvert$					$\Delta = \lvert 0.5 - \mu \rvert$				

❶,❷ and ❸ : giving the smallest value of *RMSE*, respectively in cases with a high contamination of the mixture model, ($\Delta = |\lambda - \mu|$). As is expected, the *RBias* of Zelterman's estimator not only runs out from zero but also goes largely above zero, especially in cases that have a high heterogeneity and large population sizes. Interestingly, the *RBias* of Modified Zelterman's estimator 1 and 2, and Chao's estimator yield the same pattern and there are lines more close to zero in cases of larger heterogeneity. In short, it can be stated that both modified Zelterman's estimators show a good performance of estimation in terms of accuracy, similar to Chao's estimator for a particular highly contaminated Poisson model and a large size of population, see Table 4.

According to *RVar*, MLE estimator gives the lowest *RVar* of estimation for all conditions, while Zelterman's estimator remarkably provides the highest amount of *RVar*. Significantly, an increase in the amount of heterogeneity results in a decrease in *RVar* of the two modified Zelterman estimators. In comparison, both modified forms significantly give a smaller *RVar* among the original Zelterman's estimator, see Table 5 .

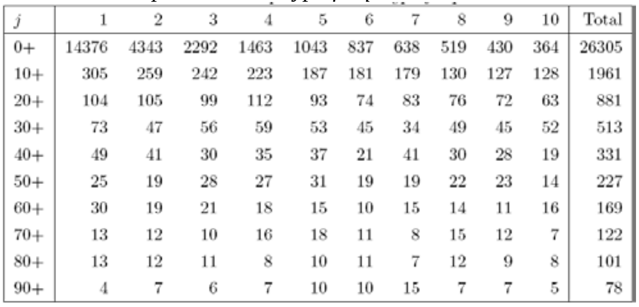
As can be seen from Table 6, MLE and Turing estimator generally give the lowest *RMSE* for the condition of having a small population size and less heterogeneity. On the other hand, Chao's estimator provides the smallest *RMSE* particularly in cases with a high amount of heterogeneity. In addition, the modified Zelterman estimator 2 also yields a good performance of *RMSE* in the cases with a substantial amount of contamination and large population size. Remarkably, both modified forms are superior to the Zelterman's estimator in terms of providing less *RMSE*, but only for particular small population size. Overall, there is a causal link between an increase in the size of population and a decrease in the amount of *RMSE* for all estimators.

4. Empirical Application

To illustrate, consider the study of Efron and Thisted [7], which aims to estimate the number of words that Shakespeare actually knew but do not appear in any of his publicized works. The number of word types used exactly *j* times is shown in Table 7. Shakespeare's works have a total of 884,647 written words which contain 31,534 different words. Of these word types used, each appears on an average of $\bar{y}$ = 884,647/31,534 = 28.05 times. As can been seen from Table 7, there are 14,376 word types used only once and 4,343 word types appearing just twice, these yield an estimate of average appeared $\hat{\lambda}_{Zel}$ = (2*4,343)/14,376 = 0.6042 times. It can be clearly seen that the difference between the sample mean ($\bar{y}$) and estimated mean $\hat{\lambda}_{Zel}$ is very distinct. This might be due to the fact that the tail of this data is very long so that the Poisson model might not fit well. Therefore, Zelterman's estimator of mean might not be appropriate and misleading.

According to an application of estimating the number of words that Shakespeare knew but do not use in any of his recognized works, the results of estimation from a variety of methods are shown in Table 8. Obviously, each estimator gives a larger difference in the amount of estimation. As a consequence of misleading homogeneity Poisson model, MLE, and Turing estimator yields a small number of estimating words that Shakespeare knew, which are 31,589 and 34,039 words, respectively. Nonetheless, the modified Zelterman's estimator 1 and 2 give a smaller estimation (54,093 and 57,373) than the original estimator (69,537), as is expected. Interestingly, the two modified Zelterman forms yield a result similar to Chao's estimator (55,328), which is known as the lower bounds of estimation population size.

Table 7: *Shakespeare's word type frequencies*



j	1	2	3	4	5	6	7	8	9	10	Total
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78

* There are 846 word types which appear more than 100 times, for a total of 31,534 word types.

Table 8: *Estimated total number of words that Shakespeare knew*

Estimator	Number of unseen words	Total number of words knew
Chao	23,794	55,328
MLE	55	31,589
Modified Zel I	22,559	54,093
Modified Zel II	25,839	57,373
Turing	2,505	34,039
Zelterman	38,003	69,537

* The total numbers of word types appeared is 31,534.

5. Conclusion/Discussion

Zelterman's estimator [1] is one of the most popular estimators used to estimate the total number of a target population. This method is developed under zero-truncated Poisson assumption. Nowadays, it has been widely used in many fields, particularly in social science. This might be due to the fact that it is easier to understand and to apply. In addition, Zelterman's estimator is a robust estimator in the sense that there are no effects from the large frequency counts. It can also be applied for the data set which behaves like the Poisson distribution. However, there are some drawbacks of using this estimator. Some critical points limit the use of available information, giving the overestimation and proving large variance of estimation. To overcome these problems two new modification forms of Zelterman's estimator are proposed in this paper, see section 2.

The simulation technique was carried out to study the performance of proposed estimators and to compare their behaviour with other well-known estimators. The scenarios of the study considered both homogeneity and heterogeneity population generated from the Poisson distribution and 50 percent:50 percent Poisson mixture distribution. The sizes of a target population were 20, 50, 100 and 500. For the homogenous Poisson model, it is clear that despite the fact that the MLE method does mostly not provide the smallest *RBias*, this method tends to give the minimum *RVar* for all conditions. In addition, *RMSE* for this method gives the lowest results out of all the cases studied. Therefore, it can be stated that the MLE method is the most suitable for a particular homogeneous Poisson model, as can be expected from efficiency results available. According to the two modified forms, Modified Zelterman's estimator 1 and 2 tend to provide a smaller *RBias*, *RVar* and *RMSE* among the ordinary Zelterman's estimator. Noticeably, the Modified Zelterman's estimator 2 shows the underestimation for a large population size, whereas the other estimators provide an overestimation. An increase in both Poisson parameter and population size can cause a decrease in *RBias*, *RVar* and *RMSE* for all estimators.

For the heterogeneous Poisson model, MLE and Turing estimator seem to provide the best performance of estimators in the cases of small heterogeneity. On the other hand, the modified Zelterman estimator 2 and Chao's estimator perform well, especially for high heterogeneity and large population size. In comparison, the two modified forms show good performance of estimation against the original Zelterman's estimator only in cases of small population size. In short, although the two modified Zelterman's estimators are not the most appropriate methods for all conditions of the study, these estimators seem to provide a better performance of estimation than the original Zelterman estimator as well as generally giving less amount of *RBias*, *RVar* and *RMSE*. Consequently, it can be stated that both modified Zelterman's estimators can be an alternative form of the Zelterman's estimator.

Lastly, there are some significant aspects of this study which will need further study and improvement for the further work. Estimation of the size of an elusive target population is of considerable interest in several fields. A variety of applications of estimating population size might still be initiative and viable research topics for statisticians and other researchers. A variance approximation and interval estimation should be also undertaken in consideration for further works. The confidence interval could be constructed by using both approximate normal and a bootstrap method. Due to computational burden of bootstrapped approach, profile-likelihood confidence interval might be an alternative. Further insights could be discussed along the lines of previous publications such as [6], [8] and [9].

References

[1] Zelterman D. Robust Estimation in truncated discrete distributions with application to capture-recapture experiments. Journal of Statistical Planning and Inference. 1988; 18:225-237.

[2] Horvitz G, Thompson J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47(260):663-685.

[3] Böhning D, van der Heijden P. Recent developments in life and social science applications of capture-recapture methods. AStA Advances in Statistical Analysis. 2009; 93:1-3.

[4] Böhning D, van der Heijden P. A covariate adjustments for zero-trucated approaches to estimating the size of hidden and elusive population. Annals of Applied Statistics. 2009; 3(2):595-610.

[5] Chao A, Lee S. Estimating the Number of Classes via Sample Coverage. Journal of the American Statistical Association. 1992; 87(417):210-217.

[6] Chao A, Estimating the population size for capture-recapture data with unequal catchability. Biometrics. 1987; 43(4):783-791.

[7] Efron B, Thisted R. Estimating the number of useen species: How many words did Shakespeare know?, Biometrika. 1976; 63(3):435-447.

[8] Böhning D. A simple variance formula for population size estimators by conditioning. Statistical Methodology. 2008; 5:410-423.

[9] Evans M A, Kim H M and O'Brien T E, An Application of Profile-Likelihood Based Confidence Interval to Capture-Recapture Estimators. Journal of Agricultural, Biological, and Environmental Statistics. 1996; 1(1):131-140.

Appendix

Let j be the number of times of identifying individuals which follows the probability density function of zero-truncated Poisson $p_j = (e^{-\lambda}\lambda^j)/(j!(1-e^{-\lambda}))$; $j = 1,2,3,...$. It can be written that $\lambda = (j+1)p_{j+1}/p_j$, and also $\lambda = \sum_{j=1}^{k}(j+1)p_{j+1} / \sum_{j=1}^{k} p_j$; $k = 1, 2, \ldots, m-1$ where m is the maximum of observed j.

Proof I.

$$\frac{(j+1)p_{j+1}}{p_j} = \frac{\dfrac{(j+1)e^{-\lambda}\lambda^{j+1}}{(j+1)!(1-e^{-\lambda})}}{\dfrac{e^{-\lambda}\lambda^j}{j!(1-e^{-\lambda})}}$$

$$= \frac{(j+1)e^{-\lambda}\lambda^{j+1}}{(j+1)j!(1-e^{-\lambda})}\frac{j!(1-e^{-\lambda})}{e^{-\lambda}\lambda^j}$$

$$= \frac{\lambda^{j+1}}{\lambda^j}$$

$$= \lambda$$

$$\therefore \quad \lambda = \frac{(j+1)p_{j+1}}{p_j}$$

Proof II.

If λ is the parameter of a Poisson probability truncated at zero and above *m* we have that:

$$2\frac{e^{-\lambda}\lambda^2}{2!}+3\frac{e^{-\lambda}\lambda^3}{3!}+...+m\frac{e^{-\lambda}\lambda^m}{m!}=\lambda\{2\frac{e^{-\lambda}\lambda}{(2)1!}+3\frac{e^{-\lambda}\lambda^2}{(3)2!}+...+m\frac{e^{-\lambda}\lambda^{m-1}}{m(m-1)!}\}$$

$$2\frac{e^{-\lambda}\lambda^2}{2!}+3\frac{e^{-\lambda}\lambda^3}{3!}+...+m\frac{e^{-\lambda}\lambda^m}{m!}=\lambda\{\frac{e^{-\lambda}\lambda}{1!}+\frac{e^{-\lambda}\lambda^2}{2!}+...+\frac{e^{-\lambda}\lambda^{m-1}}{(m-1)!}\}$$

$$\sum_{j=2}^{m} j\frac{e^{-\lambda}\lambda^j}{j!}=\lambda\sum_{j=1}^{m-1}\frac{e^{-\lambda}\lambda^j}{j!}$$

$$\lambda = \frac{\sum_{j=2}^{m} j\frac{e^{-\lambda}\lambda^j}{j!}}{\sum_{j=1}^{m-1}\frac{e^{-\lambda}\lambda^j}{j!}}$$

$$= \frac{\sum_{j=2}^{m} jp_j}{\sum_{j=1}^{m-1} p_j}$$

$$= \frac{\sum_{j=1}^{m-1} (j+1)p_{j+1}}{\sum_{j=1}^{m-1} p_j}.$$

Hence, if *j* is truncated at only zero we have that $\lambda=\sum_{j=1}^{m-1}(j+1)p_{j+1}/\sum_{j=1}^{m-1}p_j$ .This can be also satisfied $\lambda=\sum_{j=1}^{k}(j+1)p_{j+1}/\sum_{j=1}^{k}p_j$; *k* = 1, 2, 3, …, *m*–1 for a Poisson probability truncated at zero and above *k*.

Application of length biased generalized gamma distribution

Satsayamon Suksaengrakcharoen [1] and Winai Bodhisuwan [2*]

[1] Department of Statistics, Kasetsart University, Bangkok, 10900, Thailand, g5517400132@ku.ac.th

[2] Department of Statistics, Kasetsart University, Bangkok, 10900, Thailand, fsciwnb@ku.ac.th

Abstract

When the weight function depends on the lengths of the units of interest, the resulting distribution is called length biased. In this paper, the length biased generalized gamma (LBGG) distribution is considered; a particular case of weighted generalized gamma distribution, taking the weights as the variate values has been defined. The LBGG distribution is more flexible and has some interesting properties such as coefficient of kurtosis, coefficient of skewness, hazard rate and the rth moments. There are several sub-models include in the length biased exponential, length biased gamma and exponential distributions. We apply maximum likelihood estimation to estimate parameters of the distribution. We illustrate the superiority of the LBGG distribution to the flood rates data. The LBGG distribution seems to be the most appropriate model for this data set, since it provides a significantly better fit than the generalized gamma and three parameters Weibull distributions. We are also deriving the survival functions of the LBGG distribution is an alternative distribution can be used in lifetime data analysis and other fields.

Keywords: Length biased generalized gamma distribution, maximum likelihood estimation, lifetime data

*Corresponding Author
E-mail Address: fsciwnb@ku.ac.th

1. Introduction

The weighted distributions arise when the observations generated from a stochastic process is not given equal chance of being recorded; instead they are recorded according to some weight function. When the weight function depends on the lengths of the units of interest, the resulting distribution is called length biased. First introduced by Fisher [5] to model ascertainment bias, these were later formalized in a unifying theory by Rao [10]. These distributions arise in practice when observations from a sample are recorded with unequal probability. The concept of length biased distribution found in various applications in lifetime area such as family history disease and survival events. The study of human families and wildlife populations were the subjects of the article developed by Patill and Rao [8]. Patill *et al*. [9] presented a list of the most common forms of the weight function useful in scientific and statistical literature as well as some basic theorems for weighted distributions and length biased as special case they arrived at the conclusion. For example, Nanuwang and Bodhisuwan [6] presented the length biased Beta Pareto distribution. However, length biased distribution simultaneously provides great flexibility in modeling data in practice. which extends the the length biased distributions, provide powerful and popular tools for generating flexible distributions with attractive statistical and probabilistic properties.

The Length biased generalized gamma (LBGG) distribution presents a flexible family in the varieties of shapes and hazard functions for modeling duration. It was introduced by Ahmed *et al*. [1]. The LBGG family, which encompasses exponential and length biased exponential as a subfamilies and length biased gamma distribution as a particular case is introduced. The pdf of the LBGG distribution is given by:

$$g(x)=\frac{\lambda\beta}{\Gamma\left(\alpha+\frac{1}{\beta}\right)}(\lambda x)^{\alpha\beta}\,e^{-(\lambda x)^{\beta}},x>0;\alpha,\beta,\lambda>0 \quad (1)$$

where α and β are shape parameters and λ is a scale parameter. $\Gamma(a)$ is the gamma function, defined by $\Gamma(a)=\int_0^{\infty}y^{a-1}e^{-y}dy$. As well as the cumulative distribution function (cdf) of LBGG distribution, denoted as *G*(*x*), can be expressed as follows:

$$G(x)=1-\frac{\Gamma\left(\alpha+\frac{1}{\beta},(\lambda x)^{\beta}\right)}{\Gamma\left(\alpha+\frac{1}{\beta}\right)} \quad (2)$$

where $\Gamma(a,b)=\int_b^\infty y^{a-1}e^{-y}dy$ is the upper incomplete gamma function.

In Fig. 1-3, we show graphs of LBGG distribution, for different values of α,β and λ respectively.

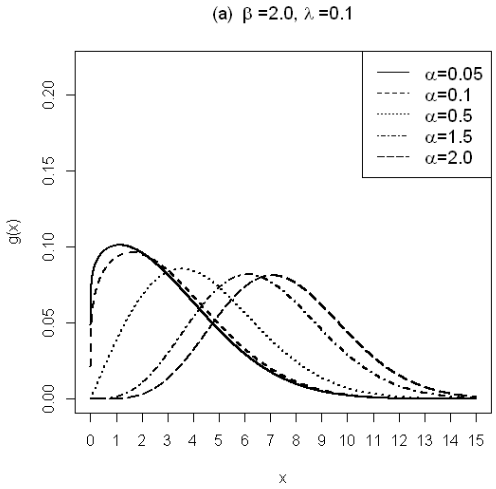


Figure1: The probability density function of LBGG distribution for different values of α

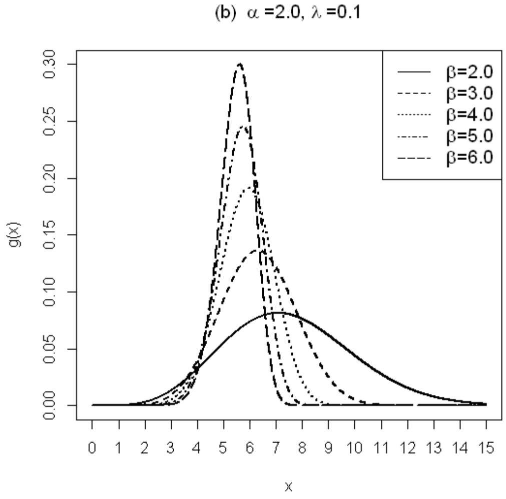


Figure2: The probability density function of LBGG distribution for different values of β

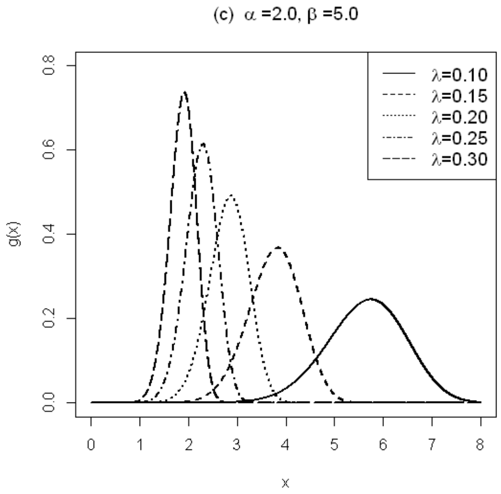


Figure 3: The probability density function of LBGG distribution for different values of λ

The rth moments of the LBGG distribution is given by:

$$E\left(X^r\right)=\frac{\Gamma\left(\alpha+\frac{r+1}{\beta}\right)}{\lambda^r\Gamma\left(\alpha+\frac{1}{\beta}\right)}\quad ; r = 1, 2, 3, ... \qquad (3)$$

From (3), it is simple to deduce mean and variance

of X which are given in (4) and (5) respectively.

$$E\left(X\right)=\frac{\Gamma\left(\alpha+\frac{2}{\beta}\right)}{\lambda\Gamma\left(\alpha+\frac{1}{\beta}\right)}. \qquad (4)$$

$$\operatorname{Var}\left(X\right)=\frac{\Gamma\left(\alpha+\frac{3}{\beta}\right)}{\lambda^2\Gamma\left(\alpha+\frac{1}{\beta}\right)}-\left[\frac{\Gamma\left(\alpha+\frac{2}{\beta}\right)}{\lambda\Gamma\left(\alpha+\frac{1}{\beta}\right)}\right]^2. \qquad (5)$$

There are some sub-model of the LBGG distribution for example

- If α = β = 1 then the LBGG distribution deduces to length biased exponential distribution Ahmed *et al.* [2] and its pdf is given by:

$$g\left(x\right)=\lambda^2xe^{-\lambda x}$$

- If β = 1 then the LBGG distribution reduces to length biased gamma distribution which presented by Ahmed *et al.* [2] as follows:

$$g(x)=\frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)}x^{\alpha}e^{-\lambda x}$$

- If $\alpha = 0$ and $\beta = 1$, then the LBGG distribution reduces to exponential distribution and its pdf can be written as:

$$g(x)=\lambda e^{-\lambda x}$$

In this work, we investigate some mathematical properties are coefficient of kurtosis, coefficient of skewness, the hazard rate of distribution and devoted to the discussion on the rth moments. Maximum Likelihood Estimation (MLE) is addressed in Section 2 and we provide application for the LBGG distribution to real data set and discussion in Section 3. Finally, we offer some concluding remarks on the main results and their significance.

2. Material and Method

2.1 Some mathematical properties

In this section, we will consider the rth moment of r.v. X~LBGG (α,β,λ). The LBGG distribution presents various properties including: the rth moment, coefficient of kurtosis, coefficient of skewness and harzard rate are provided as follows:

From (3), it is straightforward to the second four moments, coefficient of kurtosis and coefficient of skewness respectively as:

$$E(X^2)=\frac{\Gamma\left(\alpha+\frac{3}{\beta}\right)}{\lambda^2\Gamma\left(\alpha+\frac{1}{\beta}\right)}$$

$$E(X^3)=\frac{\Gamma\left(\alpha+\frac{4}{\beta}\right)}{\lambda^3\Gamma\left(\alpha+\frac{1}{\beta}\right)}$$

$$E(X^4)=\frac{\Gamma\left(\alpha+\frac{5}{\beta}\right)}{\lambda^4\Gamma\left(\alpha+\frac{1}{\beta}\right)}$$

We set

$$\omega(\alpha,\beta,i)=\frac{\Gamma(\alpha)\Gamma\left(\alpha+\frac{i+1}{\beta}\right)}{\Gamma(\alpha)\Gamma\left(\alpha+\frac{1}{\beta}\right)}$$

note that, $\omega(\alpha,\beta,i)$ is defined when $i \in I^+$ and let,

$$W=\sqrt{\omega(\alpha,\beta,2)-\omega^2(\alpha,\beta,1)}$$

consequently, the coefficient of skewness (α_3) in (6) and the coefficient of kurtosis (α_4) in (7) can be written as:

$$\alpha_3=\frac{\left[\omega(\alpha,\beta,3)-3\omega(\alpha,\beta,2)\omega(\alpha,\beta,1)+2\omega^3(\alpha,\beta,1)\right]}{W^3} \quad (6)$$

$$\alpha_4=\left[\omega(\alpha,\beta,4)-4\omega(\alpha,\beta,3)\omega(\alpha,\beta,1) +6\omega(\alpha,\beta,2)\omega^2(\alpha,\beta,1)-3\omega^4(\alpha,\beta,1)\right]/W^4 \quad (7)$$

Hazard rate (or failure rate) are expansively apply in several fields. By definition, the hazard rate of a r.v. X with pdf f(x) and cdf F(x) can be written by:

$$h(x)=\frac{g(x)}{1-G(x)}.$$

Using (9) and (10), the hazard rate of the MGG distribution may be expressed as:

$$h(x)=\frac{\lambda\beta(\lambda x)^{\alpha\beta}e^{-(\lambda x)^{\beta}}}{\Gamma\left(\alpha+\frac{1}{\beta}\right)-\Gamma\left(\alpha+\frac{1}{\beta},(\lambda x)^{\beta}\right)}$$

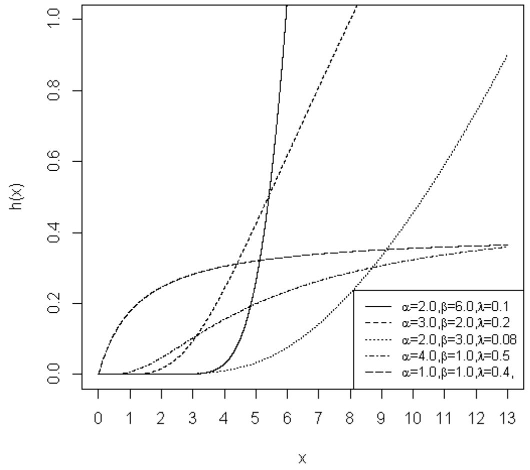


Figure 4: Plot of the hazard rates of the LBGG distribution for different values of parameters.

2.2 Maximum likelihood Estimators

The estimation of parameters for the LBGG distribution via the MLE will be discussed. Let $X_1,...,X_n$ be a random sample from X~LBGG(α,β,λ) the likelihood function is given by:

$$\mathrm{L}\left(\mathrm{X};\theta\right)=\prod_{\mathrm{i=1}}^{\mathrm{n}}\frac{\lambda\beta}{\Gamma\left(\alpha+\frac{1}{\beta}\right)}\left(\lambda\beta\right)^{\alpha\beta}e^{-\left(\lambda x\right)^{\beta}}$$

$$\mathrm{L}\left(\mathrm{X};\theta\right)=\frac{\lambda^{n\left(1+\alpha\beta\right)}\beta^{n}}{\Gamma^{n}\left(\alpha+\frac{1}{\beta}\right)}\prod_{\mathrm{i=1}}^{\mathrm{n}}x_i^{\alpha\beta}e^{-\lambda^{\beta}\sum_{i=1}^{n}x_i^{\beta}}$$

from which we calculate approximately the log-likelihood function:

$$\log L\left(\theta\right) = n\left(1+\alpha\beta\right)\log\left(\lambda\right)+n\log\beta - n\log\Gamma\left(\alpha+\frac{1}{\beta}\right) + \alpha\beta\sum_{i=1}^{n}\log\left(x_i\right) - \lambda^{\beta}\sum_{i=1}^{n}x_i^{\beta}$$

The first order conditions for finding the optimal values of the parameters were obtained by differentiating with respect to α,β and λ we get the following differential equations:

$$\frac{\partial}{\partial\alpha}\mathrm{logL}\left(\theta\right)=\mathrm{n}\beta\log\left(\lambda\right)\text{-}\mathrm{nlog}\Gamma\left(\alpha+\frac{1}{\beta}\right)+\beta\sum_{\mathrm{i=1}}^{\mathrm{n}}\log\left(\mathrm{x_i}\right) \qquad (8)$$

$$\frac{\partial}{\partial\beta}\mathrm{logL}\left(\theta\right)=\mathrm{n}\alpha\log\left(\lambda\right)+\frac{\mathrm{n}}{\beta}\text{-}\frac{\mathrm{nlog}\left(\lambda^{\beta}\right)}{\beta^{2}}+\mathrm{nlog}\Gamma\left(\alpha+\frac{1}{\beta}\right) +\alpha\sum_{\mathrm{i=1}}^{\mathrm{n}}\log\left(\mathrm{x_i}\right)\text{-}\lambda^{\beta}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{x_i^{\beta}logx_i}\text{-}\lambda^{\beta}\log\left(\lambda\right)\sum_{\mathrm{i=1}}^{\mathrm{n}}\log\left(\mathrm{x_i}\right)$$

(9)
and

$$\frac{\partial}{\partial\lambda}\mathrm{logL}\left(\theta\right)=\frac{\mathrm{n}\left(1+\alpha\beta\right)}{\lambda}\text{-}\beta\lambda^{\beta-1}\sum_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{x_i^{\beta}} \qquad (10)$$

These three derivative equations cannot be solved analytically, as they need to rely on Newton-Raphson: The Newton-Raphson method is a powerful technique for solving equations numerically. In practice $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$ are the solution of the estimating equations obtained by differentiating the likelihood in terms of α,β and λ solving in (8) - (10) to zero. Therefore, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$ can be obtained by solving the resulting equations simultaneously using a numerical procedure with the Newton-Raphson method.

3. Results and Discussion

3.1 Results

The results for the LBGG distribution are fitted to a real data set. This was the flood rates data from the Floyd River located in James, Iowa, USA for the years 1935-1973 from Akinsete *et al.* [3]. The maximum likelihood method provides parameters estimation. By comparing these fitting distribution in **Table 1** based on the p-value of this comparison.

The results have shown that the LBGG distribution provided a better fit than the GG and the three parameters Weibull distributions. Since, Mahdi and Gupta [6] presented the three parameters Weibull distribution obtained the pdf as:

$$f_W\left(x\right)=\frac{\beta}{\lambda}\left(\frac{x-\alpha}{\lambda}\right)^{\beta-1}e^{-\left(\frac{x-\alpha}{\lambda}\right)^{\beta}} \quad \text{for } x > 0;\ \alpha,\beta,\lambda > 0.$$

Table 1: Maximum likelihood estimates and K-S distances with their associated p-values for the three distributions fitted to depressive condition data

Distribution	Maximum Likelihood Estimates	K-S statistic	p-value
LBGG	$\hat{\alpha}$ =0.5908, $\hat{\beta}$ =0.5758, $\hat{\lambda}$ =0.0008	0.1041	0.7535
GG	$\hat{\alpha}$ =0.9999, $\hat{\beta}$ =1.0025, $\hat{\lambda}$ =0.0001	0.1505	0.3083
Weibull	$\hat{\alpha}$ =318, $\hat{\beta}$ =0.7032, $\hat{\lambda}$ =5127.8	0.1660	0.2081

3.2 Discussion

The LBGG distribution is a consequence of length biased distribution method which is a new generalized of gamma distribution. In this study, the LBGG distribution found that it provides significantly better fit than the GG and three parameters weibull distributions. As well as the research of Das and Roy [4], the length biased Weibull distribution provided fit to data of June rainfall in Tezpur Assam, India better than its sub-models. Furthermore, the result of this study consistent with the findings of Nanuwong, and Bodhisuwan [7], the length biased beta Pareto distribution using the exceedances of Norwegian fire claims data provides a better fit than the length biased Pareto and the beta-Pareto distributions

4. Conclusion

This work presents the LBGG distribution which is obtained by weight GG distribution. We showed that the length biased exponential, length biased Gamma and exponential distributions are sub-models of this distribution. We have derived several properties of the LBGG distribution which includes skewness, kurtosis and hazard rate. Additionally, parameters estimation are also implemented using MLE method and the usefulness of this distribution is illustrated by real data set. Based on p-values of goodness of fit test, we found that the LBGG distribution provide highest p-values when we compared with GG and three parameters Weibull distributions as shown in Table 1. According to the classical statistics, the LBGG distribution is the best fit for these data. In conclusion, it is believed that the LBGG distribution may attract wider application in lifetime data from diverse disciplines.

Acknowledgements

The authors wish to thank School of Science University of Phayao and Department of Statistics, Kasetsart University. Also, the authors thank the editor and referees for their comments that aided in improving this article

References

[1] Ahmed, A., Mir, K.A. and Reshi, J.A. On new method of estimation of parameters of size-biased generalized gamma distribution and its structural properties. IOSR Journal of Mathematics. 2013; 5 (2): 34-40.

[2] Ahmed, A., Mir, K.A. and Reshi, J.A. Structural properties of size-biased gamma distribution. IOSR Journal of Mathematics.2013; 5 (2): 55-61.

[3] Akinsete, A., Famoye, F. and Lee, C. The beta-Pareto distribution. Statistics. 2008; 42 (6): 547-563. DOI: 10.1080/02331880801983876

[4] Das, K.K. and T.D. Roy. On some length-biased weighted Weibull distribution. Adv. Applied Sci.Res. 2011; 2: 465-475.

[5] Fisher, R.A. The effects of methods of scertainment upon the estimation of frequencies. Ann. Eugenics.1934; 6: 13-25.

[6] Mahdi, T. and Gupta, A.K. A generalization of the gamma distribution. Journal of Data Science. 2013; 11(2013): 403–414.

[7] Nanuwong, N. and W. Bodhisuwan. Length biased beta-pareto distribution and its structural properties with application. J. Math. Stat. 2014; 10: 49-57.

[8] Patil, G.P. and Rao, C.R. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics. 1978; 34: 179-189.

[9] Patill, G.P. Rao, C.R. and Ratnaparkhi, M.V. On discrete weighted distribution and their use in model choice for observed data. Commun. Statisit-Theory Math. 1986; 15(3): 907–918.

[10] Rao, C.R. On discrete distributions arising out of method of ascertainment in classical and Contagious Discret. Pergamum Press and Statistical publishing Society, Calcutta. 1965; 320-332

Autocorrelation of southern Thailand (east coast) rainfall data using wavelet transform

Budsaraphorn Luangmalawat [1*], Usa Humphries[2] and Suwon Tangmanee [3]

[1]*Deptment. of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Thailand, budsaraphorn@tni.ac.th,*

[2]*Deptment. of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Thailand, usa.wan@kmutt.ac.th*

[3]*Centre of Excellence in Mathematics, CHE, Thailand,kruanchun@gmail.com*

Abstract

The wavelet transform is a powerful tool to analyse a non-stationary time series since it allows analysing different scales of temporal variability. This study, the total monthly rainfall data of 62 years, 1951-2012, in the east coast (Gulf of Thailand) of southern Thailand were analysed using Morlet wavelet. The wavelet power spectrum was defined as the wavelet transform of the autocorrelation function which used to investigate the periodic oscillation of rainfall variability in the rainfall time series. The analysing revealed two events of oscillations, the period of 4-8 month and 8-16 month, with 95% confidence interval using a white noise process.

Keywords: Autocorrelation, rainfall data, rainfall variability, east coast of southern Thailand, wavelet transform

*Corresponding Author
E-mail Address: budsaraphorn@tni.ac.th

1. Introduction

Southern Thailand is located on the Malay Peninsula, as shown in Figure 1. with rain 8 months a year. The region is divided into two parts for meteorological purposes, the east coast (Gulf of Thailand) and the west coast (Andaman Sea). The region is considered as an important economic crop region of Thailand and also including tourism industry that generates revenue to the country. Rainfall is a significant impact factor on those productivities. The time series of rainfall data is actually a non-stationary time series. In the past, the Short Time Fourier Transform (STFT) was considered as a powerful tool to analyse such the series. Until the application of wavelet transform has been developed for signal processing by Grossman and Morlet in 1984. The wavelet transform is more appropriate than the STFT since it allows analysing different scales of temporal variability. The wavelet transform has been applied to analyse rainfall time series data by many researchers. Santos et al. [1] in 2001 analysed the variability of the total monthly rainfall data of Matsuyama city Japan. Yueqing [2] in 2004 used wavelet to analyse rainfall variation in Hebei China. Kumar, Manchanda, and Sontakke [3] in 2006 used wavelet method for Indian rainfall data.

The purpose of this study is using wavelet transform to investigate the periodic oscillation of rainfall variability in the past 62 years in the east coast (Gulf of Thailand) of southern Thailand.

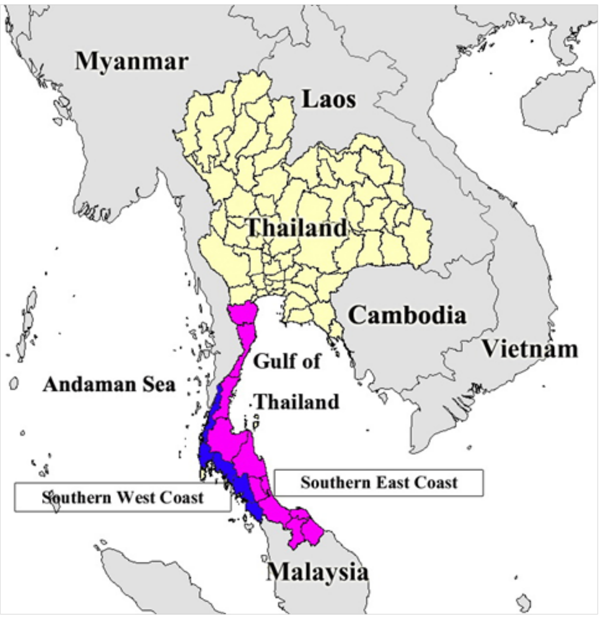


Figure 1: The studied area

2. Rainfall data

Southern Thailand receives on average 2400 mm.[4] of rainfall annually. There are only two seasons, wet and dry, but the seasons do not run at the same time on both the east and the west side of the peninsular. The east coast, the northeast monsoon brings heavy rain falls between September and December. While the west coast, the southwest monsoon brings rain and often heavy storms from April through to October. The monthly rainfall in the region was defined as the mean of total monthly rainfall over the stations of the region. The data since 1951 to 2012 were obtained from the Thai Meteorological Department. Figure 2 shows the monthly mean rainfall of the east coasts.

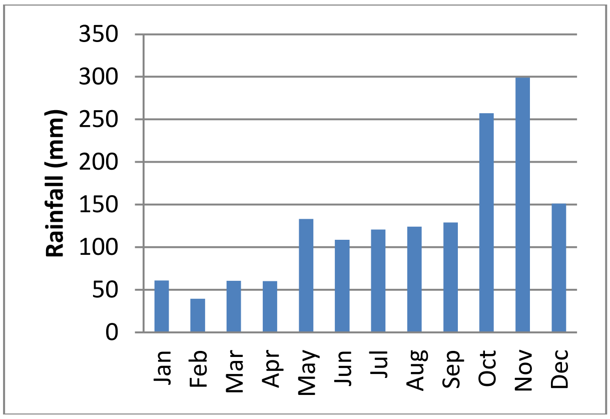


Figure 2: The monthly mean rainfall

3. Wavelet transform

The wavelet transform has been developed to analyse a nonstationary time series. The continuous wavelet transform of a discrete-time series $x_n = x(t_n)$ at time $t_j = j \cdot \delta t$ on scale s of length N and a time interval δt with respect to the wavelet function or the mother wavelet $\psi_0(\theta)$ is defined as the inner product of the time series x_n with a scaled and translated mother wavelet (called the daughter wavelet)

$$W_j(s) = \left(\frac{\delta t}{s}\right)^{1/2} \sum_{n=0}^{N-1} x_n \psi_0^* \left[\frac{(n-j)\delta t}{s}\right], \qquad (1)$$

where ψ^* denotes the complex conjugate of ψ.

The wavelet function $\psi_0(\theta)$, real or complex, must be square integrable and satisfy the admissibility conditions

$$\int_{-\infty}^{\infty} |\psi_0(\theta)|^2 d\theta < \infty, \qquad (2)$$

$$2\pi \int_{-\infty}^{\infty} \frac{|\hat{\psi}_0(\omega)|^2}{|\omega|} d\omega < \infty, \qquad (3)$$

where $\hat{\psi}_0(\omega) = \int_{-\infty}^{\infty} \psi_0(\theta) e^{-i\omega\theta} d\theta$ is the Fourier transform of $\psi_0(\theta)$.

The scale s $(s > 0)$, can be interpreted as a dilation when $s > 1$ or a compression when $s < 1$ of $\psi_0(\theta)$. The factor of $(\delta t / s)^{1/2}$ is a normalization to keep the energy of the scaled wavelet function equal to the energy of the original wavelet function. In this study, the Morlet wavelet is used as the wavelet function. The Morlet wavelet, see Figure 3, is given by

$$\psi_0(\theta) = \frac{1}{\pi^{1/4}} e^{i\omega_0\theta} e^{-\theta^2/2}, \qquad (4)$$

where ω_0 is a unit-less frequency. In this study, $\omega_0 = 6$ is chosen in order to guarantee the admissibility condition.

The Fourier transform of $\psi_0(\theta)$ is given by

$$\hat{\psi}_0(\omega) = \frac{1}{\pi^{1/4}} e^{-(\omega-\omega_0)^2/2}. \qquad (5)$$

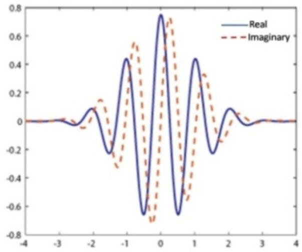


Figure 3: The Morlet wavelet with $\omega_0 = 6$

The calculation of wavelet transform (1) can be done in Fourier space [5]. The convolution theorem indicates that the wavelet transform of the time series x_n with the mother wavelet $\psi_0(\theta)$ can be obtained as the inverse Fourier transform of the product of the individual Fourier transforms:

$$W_j(s) = \left(\frac{2\pi s}{\delta t}\right)^{1/2} \sum_{k=0}^{N-1} \hat{x}_k \hat{\psi}_0^*(s\omega_k) e^{i\omega_k j\delta t}, \qquad (6)$$

where the factor of $(2\pi s / \delta t)^{1/2}$ is a normalization to keep unit energy, $\|\hat{\psi}_0(s\omega_k)\| = 1$, $\hat{x}_k$ is the discrete Fourier transform of x_n:

$$\hat{x}_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{-2\pi ikn/N}, \qquad (7)$$

and $\hat{\psi}_0(s\omega_k)$ is the Fourier transform of $\psi_0(t/s)$.

The angular frequency ω_k is defined as:

$$\omega_k = \begin{cases} \frac{2\pi k}{N\delta t}: & k \le \frac{N}{2} \\ -\frac{2\pi k}{N\delta t}: & k > \frac{N}{2} \end{cases} . \qquad (8)$$

3.1 Autocorrelation

This study, autocorrelation method on the wavelet transform was described in detecting the periodic oscillation of the rainfall time series. The wavelet power spectrum, $WPS_j(s)$, was defined by the autocorrelation of the wavelet transform:

$$WPS_j(s) = \left\langle W_j(s)W_j^*(s) \right\rangle = |W_j(s)|^2, \qquad (9)$$

where $\langle \cdot \rangle$ denotes expectation value.

The values of the wavelet power spectrum $WPS_j(s)$ were visualized by using contour plots. The $WPS_j(s)$ could be normalized by $1/\sigma^2$,where σ^2 is the variance. The normalized wavelet power spectrum gave a measure of the power relative to white noise. For a white noise time series, the expectation value for the wavelet transform is $|W_j(s)|^2 = \sigma^2$.

3.2 Edge effect

The finite-length time series was used in this study, but the wavelet transform in (6) assumes the time series is cyclic. The wavelet transform at a point in time t_j always contains information of neighbouring data points. The number of these points depends on the wavelet function and the respected scale. The edge effect will occur when the wavelet function is centred close to the beginning and the end of each scale of the computations. Thus, padding with sufficient zeros, 1×10^{-5} in this study, at the beginning and the end is necessary to decrease the amplitude near the edges as close enough to zero. The region of the wavelet transform where edge effects become important is called the cone of influence. The cone of influence is defined as the *e*-folding time, τ_s, for the autocorrelation of wavelet power at each scale [5], which for the Morlet wavelet $\tau_s = \sqrt{2}s$.

3.3 Significance testing

In this study, a white-noise process was assumed to be a background spectrum. If the time series x_n is a normally distributed, then the real and imaginary parts of $\hat{x}_k$ are both normally distributed. The square of a normally distributed number is chi-square distributed with one degree of freedom, therefore $|\hat{x}_k|^2$ is chi-square distributed with two degrees of freedom, denoted by χ_2^2. Since the local wavelet power spectrum follows the mean Fourier spectrum, the distribution of $|W_j(s)|^2$ is the same as $|\hat{x}_k|^2$. Therefore, every value of the wavelet power spectrum is distributed as:

$$\frac{\left|W_j(s)\right|^2}{\sigma^2} \sim \frac{1}{2}\chi_2^2 \qquad (10)$$

After choosing a specific confidence level for χ_2^2, 95% in this study, (10) could be calculated at each scale and the 95% confidence contour lines were constructed.

4. Results and interpretation

The calculation and plot of the wavelet power spectrum in this study were done by modifying from the wavelet software written by C. Torrence and G. Compo, available at the URL http://paos.colorado.edu/reseaech/wavelets/. Since the rainfall data distributed monthly, the parameters for the analysis were set as the time interval $\delta t = 1$ month, the starting scale $s_0 = 2$ months, the scale width $dj = 1/12$ and a white noise time series was used as the background.

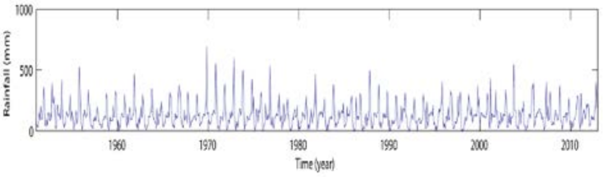


Figure 4: Monthly Rainfall of Southern Thailand East Coast

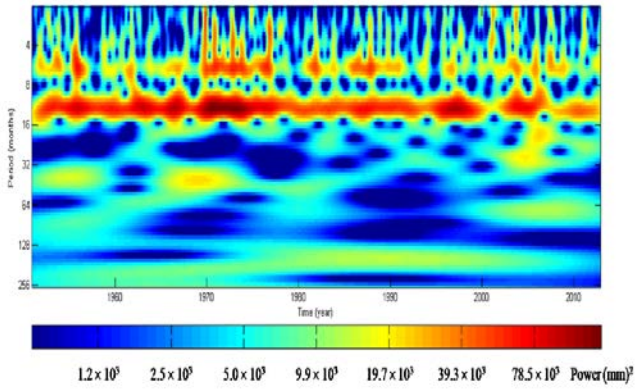


Figure 5: Southern Thailand East Coast Monthly Rainfall Wavelet Power Spectrum

Figure 5 shows the wavelet power spectrum of the monthly rainfall in the east coast of southern Thailand, Figure 4, using the Morlet wavelet. The left axis is the Fourier period (in month) corresponding to the wavelet scale on the right axis. Since the wavelet transform $W_j(s)$ are complex, the wavelet power spectrum ($|W_j(s)|^2$) is a way to visualize them. The higher value of $|W_j(s)|^2$ means the more similarity between the time series and the wavelet function at that particular time shift and scale. A high value will result in a dark red point and a low value in a dark blue point as illustrate in the colour bar.

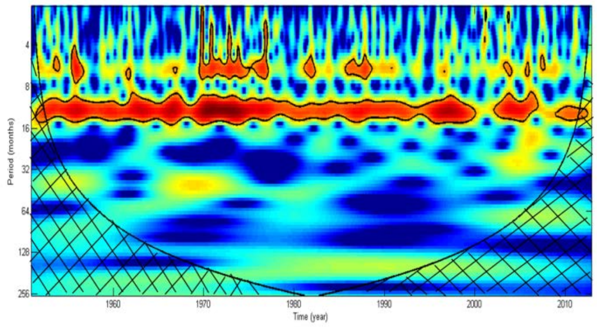


Figure 6: Southern Thailand East Coast Monthly Rainfall Wavelet Power Spectrum with 95% confidence contour

The thick contour in Figure 6 encloses regions of greater than 95% confidence interval for a white noise process. The cross-hatched region is called the cone of influence, where edge effects become important. Since the calculation deals with finite-length time series, errors will occur at the beginning and the end of the wavelet power spectrum [5]. It is observed that, see Figure 6, there are two events of oscillations, the period of 4-8 month and 8-16 month bands. The variance of power in 8-16 month band also shows the dry and the wet year i.e., when the power decreases substantially in the band, it means a dry year and when the power is maximum means a wet year [1].

5. Conclusion

Wavelet analysis was applied to analyse the variability of the monthly rainfall time series on the east coast (Gulf of Thailand) of southern Thailand. The analysis was done with the mean of total monthly rainfall over the stations of the region during 1951 to 2012 which were obtained from the Thai Meteorological Department. The periodic oscillation of rainfall variability could be investigated by the plot of the wavelet power spectrum, which defined as the autocorrelation of the wavelet transform. There were two events of oscillations occur, the period of 4-8 month and 8-16 month, with 95% confidence interval using a white noise process.

Acknowledgement

The authors wish to acknowledge Dr. Christopher Torrence and Dr. Gilbert P. Compo for providing the wavelet software.

References

[1] Santos C, Galvao C, Suzuki K, Trigo R. Matsuyama City Rainfall Data Analysis Using Wavelet Transform. Annual Journal of Hydraulic Engineering. 2001; 45: 211-216.

[2] Xu Y, Li S, Cai Y. Wavelet analysis of rainfall variation in the Hebei Plain. Science in China Ser. D Earth Sciences. 2005; 48(1): 2241-2250.

[3] Kumar J, Manchanda P, Sontakke NA. Mathematical Models and Methods for Real World Systems. CRC Press; 2005.

[4] Andrew J. The report Thailand 2009**.** Oxford Business Group; 2009.

[5] Torrence C, Compo GP. A practical guide to wavelet analysis. Bulletin of the American Meteorological Society. 1998; 79: 61-78.

[6] Daubechies I. Ten Lectures on Wavelets; 1992.

[7] Emery W, Thomson R. Data Analysis Methods in Physical Oceanography. 2nd ed. Amsterdam: Elsevier Science B.V; 1999.

[8] Kim S. Wavelet analysis of precipitation variability in Northern California, U.S.A. Journal of Civil Engineering. 2004; 8(4): 471-477.

[9] Labat D. Wavelet analysis of the annual discharge records of the world's largest rivers. Advances in Water Resources. 2008; 31: 109-117.

[10] Percival DB. On estimation of the wavelet variance. Biometrika. 1995; 82: 619-631.

[11] Santos C, Galvao C, Trigo R. Rainfall data analysis using wavelet transform, Proceedings of the International Conference on Hydrology of the Mediterranean and Semiarid Regions; 2003 April 1-4; Montpellier, France. 2003. p. 195-201.

Thailand river basin flood prediction using fuzzy rules

Chalermchai Puripat[1] and Adisak Pongpullponsak[2*]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand,*
poolsup973@gmail.com,chalermchai.pur@kbu.ac.th
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand,*
adisak.pon@kmutt.ac.th

Abstract

This paper gives an approach to predict flood risk using fuzzy logic, fuzzy set and mandani inference. These methods provide a logical and systematic analysis of uncertainties, which can deal with many factor affect flooding in a relative simple and concrete way. However, A way of preventing flooding that is base a natural idea of Thai people because Thailand is in the tropical moisture. These causes were very flooding of the basin in Thailand. The aim of this paper is to perform demonstrate a simple risk system approach for this crisis. The disaster warning Thailand flood will be useful for people. This condition simulation is improved on qualitative planning issues like future flood prevention, flood early warning systems, preparedness, emergency response protocol during crisis, flooding mitigation systems arrangement and lessons learned. Rainfalls in the year have a different amount.

Keywords: flood prediction, fuzzy logic, flood risk,

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

1. Introduction

The rainy Global warming affects climate change to each country around the globe, the significantly change in season like La Nina and El Nino[1] are followed with weather-related disasters and more degree of severity and also inevitably effecting to Thailand. For Thailand, The natural disasters from the report of the Department of Disaster Prevention and Mitigation, Ministry of Interior during 2009-2010 were floods, landslide, storms, droughts, colds, wildfires, earthquakes, tsunamis[2], diseases and pests. The most damages to agricultural sectors were floods and droughts. The consequence from flooding is household garbage, waste water, unemployment, and etc. The causes of flooding are amount and duration of rainfalls deposit on the ground and flow into canals and rivers, if volume of water is over the capacity of the rivers, water surpasses the river banks and deposits in the lower land for a long period of time. To predict flooding accurately, the rain fall, the depth of river, the contour, and the dam level are the input data to the fuzzy logic model. The result will reduce the risk from flood in the future.

2. Research Methodology

2.1 Fuzzy logic

The basic concept in fuzzy logic

Fuzzy logic is a tool to assist in decision making under uncertainty by allowing flexible. The main reason for using a simulation approach, the complexity of the human mind. Fuzzy logic is a special logical false (Boolean logic), a concept that has been expanded in terms of the actual (partial true) by the fact that in the period between the (completely true) to false. (completely false) the same logic is true and false only. Shown in Figure 2-1.

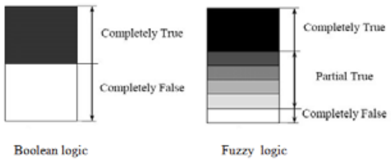


Fig 2-1 Boolean logic and fuzzy logic

A fuzzy(fuzziness) is called multi freelancers (multivalance) whose value is more than two values and differences with Britain that millions of France (bivalance) is a member of only two values, fuzzy. the sets (Fuzzy set) is a mathematical tool of the media. "The uncertainty (uncertainty)" is not only the second case by the theory of fuzzy sets to use meaningful variable (linguistic) rather than quantity. (quantitative) variables. Fuzzy math shows how to create an ambiguity. Uncertainties associated with the idea of human feelings. Considering the uncertainties in various components to set the terms of the decision (Decision making) by the set of non-members (Set membership).

2.2 Fuzzy set

Fuzzy Set is a set with a smooth boundary. Theory of fuzzy sets theory to cover typical. By fuzzy sets allow a subset of values between 0 and 1 in the world of reality, not a set, only a set pattern only Are fuzzy sets as well. Fuzzy sets have fuzzy boundaries, not abrupt change from white to black.

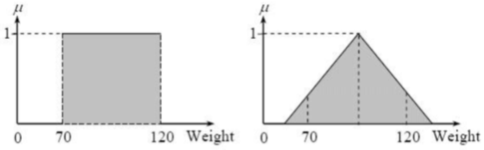


Fig 2-2 The configuration a membership.

Definition of fuzzy sets assigned to X is a subset of the fuzzy set unknown characterize the functions of a member. $\mu_{\underset{\sim}{A}}(x): X \rightarrow [0,1]$

Membership Function

The membership function to define a function with variable degree requirements. Start by replacing the representative with uncertainty and ambiguity, so it is not essential to the property or operations of the fuzzy number. The shape of the membership function is critical thinking and problem solving. The function is to be asymmetric or symmetric all the time.

The type of the member function.

Type of function is a common species. But here are just some of the six types mentioned.

Triangular membership function A triangle with all three parameters, namely {a, b, c}.

$$triangular(x:a,b,c)=\begin{cases} 0 & x<a \\ (x-a)/(b-a) & a\le x<b \\ (c-x)/(c-b) & b\le x<c \\ 0 & x>c \end{cases}$$

Trapezoidal membership function

That trapezoidal function with 4 parameters {*a*, *b*, *c*, *d*}

$$trapezoidal(x:a,b,c,d)=\begin{cases} 0 & x<a \\ (x-a)/(b-a) & a\le x<b \\ 1 & b\le x<c \\ (d-x)/(d-c) & c\le x<d \\ 0 & x>d \end{cases}$$

Linguistic variable

The concept is very important in fuzzy logic. Language variables that determine the value of what is described in both quality Using literal language (linguistic term) and in volume. Function as a member (membership function) that displays a set of fuzzy. Language term used for the representation of concepts and knowledge in human communication. The function is useful in dealing with numerical input data.

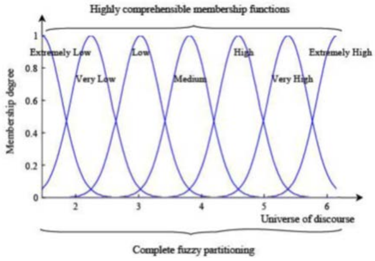


Fig 2-3 Linguistic Variable

2.3 fuzzy rules

Science on fuzzy logic has a number. But the most popular and most applications will include comments. Fuzzy rules are if - then (fuzzy if-then rule).

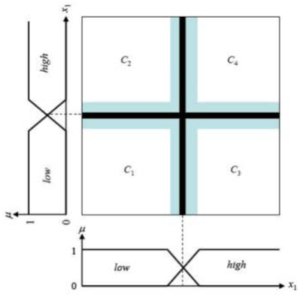


Fig 2-4 The fuzzy Grouping

we can be written as a rule in the following sentence.

Rule 1: If x_1 is low and x_2 is low, then the data (x_1, x_2) is C_1

Rule 2: If x_1 is low and x_2 is high, then the data (x_1, x_2) is C_2

Rule 3: If x_1 and x_2 are the high and low values for (x_1, x_2) is a C_3

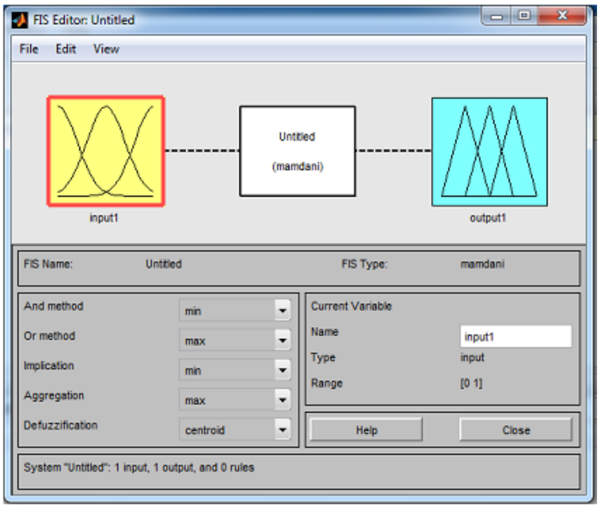
Rule 4: If x_1 and x_2 is high, then the data is high (x_1, x_2) is a C_4

Fuzzy cognitive system can be expressed in a sentence.

IF premise (antecedent), THEN conclusion (consequent)

The above is a well known name. "If the rule base - so" (IF-THEN rule-based form) or a priori (deductive form) in the form of the inference. If we know the truth (and the assumptions or field), then we can infer. Fact or conclusions, also called settlement or by.

2.4 Mamdani min/max of inference



FIS Name: Untitled

FIS Type: mamdani

And method	min
Or method	max
Implication	min
Aggregation	max
Defuzzification	centroid

Current Variable

Name	input1
Type	input
Range	[0 1]

System "Untitled": 1 input, 1 output, and 0 rules

Fig 2-5 Mamdani fuzzy system

Assigned a Mamdani fuzzy system with two inputs x1 and x2 (antecedent) and one output y (consequent) fuzzy rules which are

$$IF \quad x_1 \quad is \quad \underset{\sim}{A}_{k1} \quad and \quad x_2 \quad is \quad \underset{\sim}{A}_{k2} \quad THEN \quad y \quad is \quad \underset{\sim}{B}_k \quad for \quad k=1,2,...,r$$

Total output is obtained. By using the elements Maximum – minimum (max-min composition) and the highest value component - a multiple (max-product composition).

Method compose a maximum value - minimum.

$$\mu_{\underset{\sim}{B}_k}(y)=\max\left[\min\left(\mu_{\underset{\sim}{A}_{k_1}}\left(Input(i)\right),\mu_{\underset{\sim}{A}_{k_2}}\left(Input(j)\right)\right)\right] \quad for \quad k=1,2,...,r$$

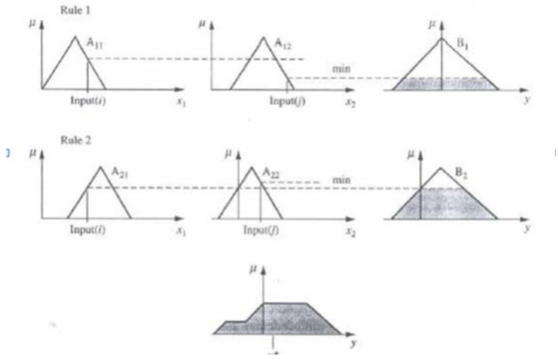


Fig 2-6 Graphic inference mamdani method

Finally, if you want the output to a common value. Can find a way to convert the fuzzy values are. (defuzzification method).

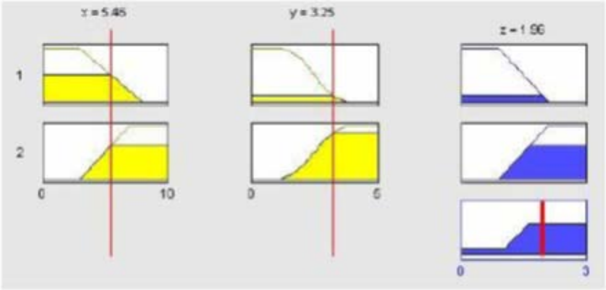


Fig 2-7 The fuzzy inference system mamdani method

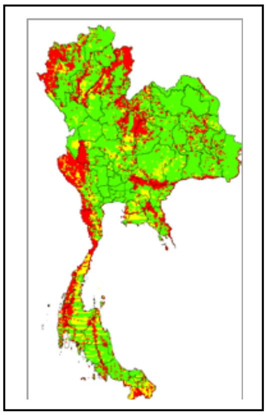


Fig 2-8 map shows flood risk areas[14]

Recall that the volume of the water in the dam Volume is related to the height of fluid long term.

3. Result

Main reason for the effectiveness of a flood prediction system is a quite complex task. The Importance, a correct measure would be a cost-benefit analysis, an amount of the damages from an incorrect forecast. Although, one has always to resort to fuzzy logic evaluations. We study the factors associated with flood i.e. water level, water value, rain, water in the dam and height of area. Which various factors, We study and split level of the factor for find the flood risk using fuzzy rule and to create a map showing the risk of flood for to use as a method to prevent flood

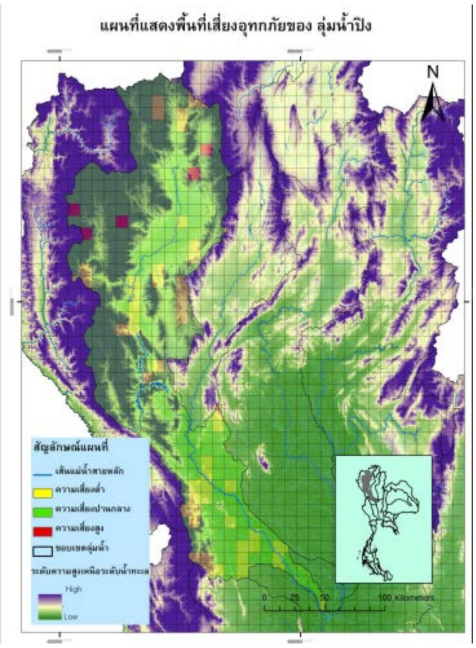


Fig 3-1 Map of Ping basin

Table 3-1 The variables are defined as three levels.

Water level	(m)	Water level	(ml)	Rainfall	Average per day. (ml)	Water in dam	%	Height above sea level	(m)
Normal	< 3.00	Normal	<300	Normal	< 50	Little	<30	normal	<700
Much	2.50-4.00	Much	250-450	Much	30-90	Normal	25-100	medium	500-1500
Overmuch	> 3.70	Overmuch	>350	Overmuch	> 75	Overmuch	>95	highest	>1000

In this graph it can be seen high water level and lot volume of water

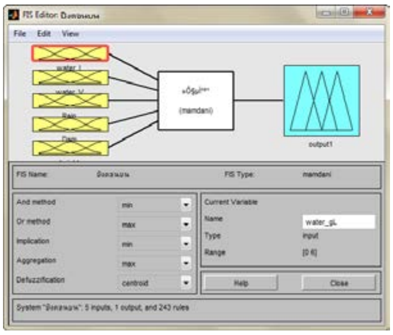
Table 3-2 The risk of water level in each area.

Flood risk areas.	Level of risk	Probability risks
Very low severe flood risk areas.	1	0.0- 0.2
Low severe flood risk areas.	2	0.1 - 0.4
Moderate severe (normally) flood risk areas.	3	0.3 - 0.7
Very severe flood risk areas.	4	0.6 - 0.9
Most severe flood risk areas.	5	0.8 - 1.0

We will use program math lab to help create membership function of the 25 basin. Variables used to create the membership function, Thus

water level	= Water_gL
volume of water	= Water_gV
rainfall	= Rain
percentage of water in the dam	= Dam
height of the area	= height

We show example to create membership function in Ping basin follows



FIS Name: FIS Type: mamdani

And method min
Or method max
Implication min
Aggregation max
Defuzzification centroid

Current Variable
Name water_gL
Type input
Range [0 6]

System "[illegible]": 5 inputs, 1 output, and 243 rules

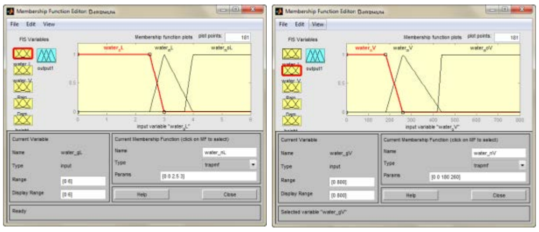
Fig 3-2 variable sets the language of fuzzy sets and Extremely Low, Very Low, Low, Medium, High, Very High and Extremely High.

Membership function of Variableฯ water level

Water_nL restrictive water level < 3

Water_mL restrictive 2.5 < water level < 4

Water_moL restrictive water level >3.7



Current Variable
Name water_gL
Type input
Range [0 6]
Display Range [0 6]

Current Membership Function (click on MF to select)
Name water_nL
Type trapmf
Params [0 0 2.5 3]

Ready

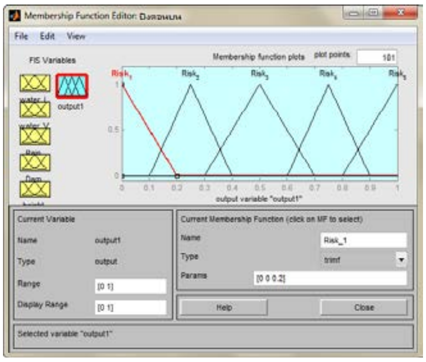
Current Variable
Name water_gV
Type input
Range [0 800]
Display Range [0 800]

Current Membership Function (click on MF to select)
Name water_nV
Type trapmf
Params [0 0 180 260]

Selected variable "water_gV"

Fig 3-3 Determine level out put variable is risk.

Create output variable and create membership function of output which divided 5 level are Risk_1, Risk_2, Risk_3, Risk_4 and Risk_5. We use program math lab to create membership function such as determine level factor.



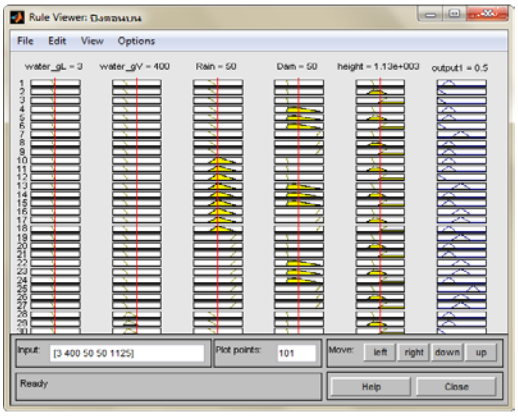
Current Variable
Name output1
Type output
Range [0 1]
Display Range [0 1]

Current Membership Function (click on MF to select)
Name Risk_1
Type trimf
Params [0 0 0.2]

Selected variable "output1"

Fig 3-4 Membership function Editor of Output

The study of the factors and principles of Fuzzy Rule of Flood Risk in the statistical models used to analyze the risk in different areas.



water_gL = 3 water_gV = 400 Rain = 50 Dam = 50 height = 1.13e+003 output1 = 0.5

Input: [3 400 50 50 1125]

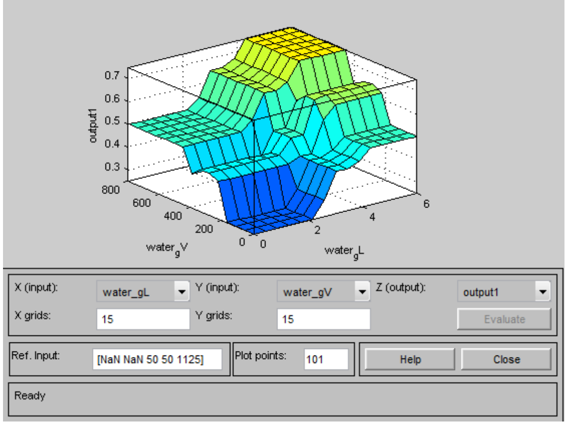
Plot points: 101

Ready

Fig 3-5 Rule Viewer to calculate

Effects of the different factors using the Fuzzy Rule with risk

For example the rules. IF Water_gL and Water_gV and Rain and Dam and Height THEN Risk



X (input): water_gL Y (input): water_gV Z (output): output1

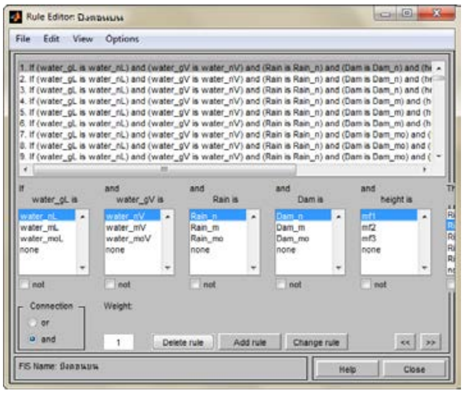
X grids: 15 Y grids: 15

Ref. Input: [NaN NaN 50 50 1125]

Plot points: 101

Ready

Fig 3-6The condition in simulate the risk of water level .



1. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_n) and (he
2. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_n) and (he
3. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_n) and (he
4. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_m) and (h
5. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_m) and (h
6. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_m) and (h
7. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_mo) and (
8. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_mo) and (
9. If (water_gL is water_nL) and (water_gV is water_nV) and (Rain is Rain_n) and (Dam is Dam_mo) and (

If water_gL is	and water_gV is	and Rain is	and Dam is	and height is
water_nL	water_nV	Rain_n	Dam_n	mf1
water_mL	water_mV	Rain_m	Dam_m	mf2
water_moL	water_moV	Rain_mo	Dam_mo	mf3
none	none	none	none	none
□ not	□ not	□ not	□ not	□ not

Connection ○ or ● and

Weight: 1

FIS Name:

Figure 3-7 a rule in the following variable is risk

1. If (water_gL is water_nL) and (water_gV is water_nV) and (RAIN is Rain_n) and (Dam is Dam_n) and (height is height_n) then (output is Risk_2)

2. If (water_gL is water_nL) and (water_gV is water_nV) and (RAIN is Rain_n) and (Dam is Dam_n) and (height is height_m) then (output is Risk_1)

3. If (water_gL is water_nL) and (water_gV is water_nV) and (RAIN is Rain_n) and (Dam is Dam_n) and (height is height_mo) then (output is Risk_1)

4. If (water_gL is water_nL) and (water_gV is water_nV) and (RAIN is Rain_n) and (Dam is Dam_m) and (height is height_n) then (output is Risk_2)

3. Conclusion

The fuzzy-set approach described above allows for a logical and systematic analysis of uncertainties. The models of fuzzy reliability [1] could be used for fuzzy reliability determining of a basin structure cross section or of a basin number. Uncertain parameters can be expressed as fuzzy sets. It is possible to process the fuzzy uncertainties in reliability analysis of a basin member. Fuzzy uncertainty could be incorporated in the estimated probability of failure. Then the optimization models designed for reliable design problems [12] can be modified by included fuzzy parameters and the a posteriori verification of results can also be generalized in the similar way [13]. The approach allows an assessment of the derivative that a particular concrete cross-section (or all basin member studied) will have a higher failure probability than the failure probability of the deterministic designed cross section or member [11]

Acknowledgements

The authors wish to thanks advisor, Royal Irrigation Department and Department of Disaster Prevention and Mitigation for the valuable co-operation and all the quantity water in each basin whose data were studied.

Reference

[1] McCarthy. et al , (2001) “The power of collaborative governance: Managing the risks associated with flooding and sea-level rise in Cape Town” Cambridge University Press, Cambridge. (Book)

[2] Centre for Disaster Mitigation Department of Disaster Prevention and Mitigation (2553)

[3] Geist, E.L., 1998. Local Tsunamis and Earthquake Source Parameters, Advances in Geophysics, 39, 117-209

[4] Adisak Pongpulponsak, Sukuman Sarikavanij and Thiradet Jiarasuksakun. (2009)

[5] Zadeh L.A.: Probability Measures and Fuzzy Events, J. of Math. Analysis and Applications, 23(2), p. 421–427

[6] Klir G.J., Yuan B.: Fuzzy Sets and Fuzzy Logic, 1st ed., Prentice Hall, New Jersey, 1995,ISBN 0-13-101171-5

[7] Mareˇ s M.: Computation over Fuzzy Quantities, CRC Press, Boca , Florida, 1994. ISBN 0849376351

[8] Karp´ ıˇ sek Z.: Fuzzy Probability and its Properties, In MENDEL ’00, 6th International Conference on Soft Computing, Brno 2000, p. 262–266, ISBN 80-214-1609-2

[9] Karp´ ıˇ sek Z., Posp´ ıˇ sek M., Slav´ ıˇ cek K.: Properties of a Certain Class of Fuzzy Numbers, In Proceedings East West Fuzzy Colloquium 2000, 8th Zittau Fuzzy Colloquium, Zittau, 2000, p. 42–51, ISBN 3-00-006723-X

[10] Karp´ ı sek Z.: Fuzzy Probability Distribution – Characteristics and Models, In Proceedings East West Fuzzy Colloquium 2001, 9th Zittau Fuzzy Colloquium. Zittau, 2001, p. 36–45, ISBN 3-9808089-0-4

[11] Step´ anek P.: New Methods and Trends for Strengthening of Concrete and Masonry Structures,In WTA Almanach 2008 Restauration and Building-Physics, Munchen, 2008, p. 83–109, ISBN 978-3-937066-08-0

[12] Plsek J., Step´ anek P., Popela P.: Deterministic and Reliability Based Structural Optimization of Concrete Cross-section, Journal of Advanced Concrete Technology, Vol. 5(1), 63–74, 2007

[13] Zampachov´ a E., Popela P., Mr´ azek M.: Optimum Beam Design via Stochastic Programming,Kybernetika, Vol. 46(3), pp. 571–582, 2010

[14] Malczewski J., 1999, GIS and Multicriteria Desision Analysis, New York: John Willey and Sons, Inc.

Combined ratio estimator of the population total in stratified random sampling

Radarut Mungta[1*] and Supannee Ungpansattawong[2]
[1]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, pug_puo_radarut@hotmail.com*
[2]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, supannee@kku.ac.th*

Abstract

This study aims to propose a new ratio estimator in stratified random sampling based on the Phanida estimator (2008). Theoretically, mean square error (MSE) has been used to indicate the quality of ratio estimators and to compare the efficiency with a former ratio estimator. This study shows that the proposed estimator has more efficiency than combined ratio estimator. Furthermore, a numerical example also has been provided to support the better theoretical result.

Keywords: Stratified random sampling, combined ratio estimator, mean square error, population total

*Corresponding Author
E-mail Address: pug_puo_radarut@hotmail.com

1. Introduction

There are two ways to make a ratio estimate of the population total Y can be made in two ways. One is to make a separate ratio estimate of the total of each stratum and add these totals. An alternative estimate is derived from a single combined ratio. From the sample data, we compute sample total of the variates in stratified random sampling method as

$$\hat{Y}_{st} = N\sum_{h=1}^{L} W_h \bar{y}_h \quad \text{and} \quad \hat{X}_{st} = N\sum_{h=1}^{L} W_h \bar{x}_h$$

L is the number of stratum , $W_h = \frac{N_h}{N}$ is the h stratum weight, N is the number of units in population, N_h is the number of units in stratum h, $\bar{y}_h$ is the sample total of the study variable in stratum h and $\bar{x}_h$ is the sample total of auxiliary variable in stratum h.

These are the standard estimates of the population totals Y and X, respectively, made from a stratified sample. The combined ratio estimate, $\hat{Y}_{RC}$ (c for combined) is

$$\hat{Y}_{RC} = \hat{R}_C X = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X$$

where $\bar{y}_{st} = \frac{\hat{Y}_{st}}{N}, \bar{x}_{st} = \frac{\hat{X}_{st}}{N}$, are the estimated population means from a stratified sample.
X is the total population of auxiliary variable.

The variance of combined ratio estimator is

$$V\left(\hat{Y}_{RC}\right) \approx \sum_{h=1}^{L} N_h^2 \theta_h \left[S_{yh}^2 - 2RS_{yxh} + R^2 S_{xh}^2 \right]$$

where $\theta_h = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$ and $R = \frac{\bar{Y}}{\bar{X}}$ is the population ratio, n_h is the number of units in sample stratum h, S_{yh}^2 is the population's variance of the study variable in stratum h, S_{xh}^2 is the population variance of auxiliary variable in stratum h and S_{yxh} is the population covariance between auxiliary variable and interested variable in stratum h (Cochran,1977).

2. Ratio estimator and its mean square errors

When first degree approximation is used in obtaining the mean square error (MSE) of a ratio estimate, it is know that MSE is equal to the variance, so MSE of combined ratio estimator can be written as follows:

$$MSE\left(\hat{Y}_{RC}\right) = V\left(\hat{Y}_{RC}\right) \approx \sum_{h=1}^{L} N_h^2 \theta_h \left[S_{yh}^2 - 2RS_{yxh} + R^2 S_{xh}^2 \right]. \qquad (1)$$

2.1 Phanida estimator

It is developed by Ray and Singh (1981), Sisodia and Dwivedi (1981), and Kadilar and Cingi (2004). Ratio estimator type of $\hat{Y}$ is written as

$$\hat{Y}_{PR3} = \hat{R}_{PR3}(X + C_x)$$

where,

$$\hat{R}_{PR3} = \frac{y}{x + C_x}.$$

In stratified random sampling, we propose this estimator as

$$\hat{Y}_{st1} = \hat{R}_{st1}(X + C_x)$$

where,

$$\hat{R}_{st1} = \frac{y_{st}}{x_{st} + C_x}$$

and $C_x = \sum_{h=1}^{L} W_h C_{xh}$.

Therefore, the MSE of this estimator is

$$MSE(\hat{Y}_{st1}) = E\left(\hat{Y}_{st1} - Y_{st1}\right)^2$$
$$= E\left(\hat{R}_{st1}\left(X + C_{xh}\right) - R_{st1}\left(X + C_{xh}\right)\right)^2$$
$$= \left(X + C_{xh}\right)^2 E\left(\hat{R}_{st1} - R_{st1}\right)^2$$

where

$$R_{st1} = \frac{Y}{X + C_{xh}}.$$

MSE of this estimator will be defined by using Taylor’s series method as

$$h\left(x_{st}, y_{st}\right) \cong h\left(X, Y\right) + \frac{\partial h\left(x_{sy}, y_{st}\right)}{\partial x_{st}}\bigg|_{X,Y}\left(x_{st} - X\right) + \frac{\partial h\left(x_{st}, y_{st}\right)}{\partial y_{st}}\bigg|_{X,Y}\left(y_{st} - Y\right)$$

where,

$$h\left(x_{st}, y_{st}\right) = \hat{R}_{st1} \quad \text{and} \quad h\left(X, Y\right) = R_{st1}.$$

When we replace the above variable into Taylor’s series, MSE can be derived as follows:

$$\hat{R}_{st1} - R_{st1} \cong \frac{\partial}{\partial x_{st}}\left(\frac{y_{st}}{x_{st} + C_{xh}}\right)\bigg|_{X,Y}\left(x_{st} - X\right) - \frac{\partial}{\partial y_{st}}\left(\frac{y_{st}}{x_{st} + C_{xh}}\right)\bigg|_{X,Y}\left(y_{st} - Y\right)$$
$$\cong -\left(\frac{Y}{(X + C_{xh})^2}\right)\left(x_{st} - X\right) + \frac{1}{(X + C_{xh})}\left(y_{st} - Y\right)$$
$$(\hat{R}_{st1} - R_{st1})^2 \cong \left(\frac{Y}{(X + C_{xh})^2}\right)^2\left(x_{st} - X\right)^2 + \frac{1}{(X + C_{xh})^2}\left(y_{st} - Y\right)^2 - 2\left(\frac{Y}{\left(X + C_{xh}\right)^3}\right)\left(x_{st} - X\right)\left(y_{st} - Y\right).$$
$$E\left(\hat{R}_{st1} - R_{st1}\right)^2 \cong \left(\frac{Y}{\left(X + C_{xh}\right)^2}\right)^2\left(V\left(x_{st}\right) + (1 - f)^2 X^2\right) + \frac{1}{(X + C_{xh})^2}\left(V\left(y_{st}\right) + (1 - f)^2 Y^2\right)$$
$$-2\left(\frac{Y}{(X + C_{xh})^3}\right)\left(\operatorname{cov}\left(x_{st}, y_{st}\right) + (1 - f)^2 XY\right),$$
$$\cong \frac{1}{(X + C_{xh})^2}\left[\frac{Y^2}{\left(X + C_{xh}\right)^2}\left(V\left(x_{st}\right) + (1 - f)^2 X^2\right) + \left(V\left(y_{st}\right) + (1 - f)^2 Y^2\right)\right.$$
$$\left. -2\frac{Y}{(X + C_{xh})}\left(Cov\left(x_{st}, y_{st}\right) + (1 - f)^2 XY\right)\right].$$

Therefore, the MSE of this estimator is

$$MSE\left(\hat{Y}_{st1}\right) = (X + C_{xh})^2 E\left(\hat{R}_{st1} - R_{st1}\right)^2$$
$$\cong \frac{Y^2}{\left(X + C_{xh}\right)^2}V\left(x_{st}\right) + V\left(y_{st}\right) - 2\frac{Y}{(X + C_{xh})}\left(Cov\left(x_{st}, y_{st}\right)\right)$$
$$+\left[\frac{Y^2}{\left(X + C_{xh}\right)^2}\left((1 - f)^2 X^2\right) + \left((1 - f)^2 Y^2\right) - 2\frac{Y}{(X + C_{xh})}\left((1 - f)^2 XY\right)\right].$$

Finally, MSE of this ratio estimator can be written as

$$MSE(\hat{Y}_{st1}) \cong \sum_{h=1}^{L}\left[N_h^2\theta_h\left({R_{st1}}^2 S_{xh}^2 + S_{yh}^2 - 2R_{st1}\rho_h S_{yh} S_{xh}\right)\right] + \left[(1 - f)^2\left(Y - R_{st1}X\right)^2\right]. \qquad (2)$$

where, $R_{st1} = \frac{Y}{X + C_{xh}}$, $V\left(x_{st}\right) = N^2\sum_{h=1}^{L} W_h^2\theta_h S_{xh}^2$,

$V\left(y_{st}\right) = N^2\sum_{h=1}^{L} W_h^2\theta_h S_{yh}^2$ and

$Cov\left(x_{st}, y_{st}\right) = N^2\sum_{h=1}^{L} W_h^2\theta_h\rho_h S_{yh} S_{xh}$.

3. Comparison of efficiency

We compare the combined ratio estimator with former combined ratio estimator by using the conditions as follows:

$$MSE\left(\hat{Y}_{st1}\right) < MSE\left(\hat{Y}_{RC}\right)$$

$$MSE(\hat{Y}_{st1}) \cong \sum_{h=1}^{L}\left[N_h^2\theta_h\left({R_{st1}}^2 S_{xh}^2 + S_{yh}^2 - 2R_{st1}\rho_h S_{yh} S_{xh}\right)\right] + \left[(1 - f)^2\left(Y - R_{st1}X\right)^2\right] < \sum_{h=1}^{L} N_h^2\theta_h\left(R^2 S_{xh}^2 - 2R\rho_h S_{xh} S_{yh} + S_{yh}^2\right)$$

Let

$A = \sum_{h=1}^{L} N_h^2\theta_h S_{xh}^2.$, $B = \sum_{h=1}^{L} N_h^2\theta_h S_{yh}^2.$, $C = \sum_{h=1}^{L} N_h^2\theta_h\rho_h S_{xh} S_{yh}.$

$$R_{st1}^2 A + B - 2R_{st1}C + (1 - f)^2(Y - R_{st1}X)^2 < R^2 A + B - 2RC$$
$$2C(R - R_{st1}) + (1 - f)^2(Y - R_{st1}X)^2 < (R^2 - R_{st1}^2)A$$
$$2C(R - R_{st1}) + (1 - f)^2(Y - R_{st1}X)^2 < (R - R_{st1})(R + R_{st1})A$$
$$2C + \frac{(1 - f)^2(Y - R_{st1}X)^2}{(R - R_{st1})} < (R + R_{st1})A$$
$$\frac{(1 - f)^2(Y - R_{st1}X)^2}{(R - R_{st1})} < \frac{(R + R_{st1})A}{2C}. \qquad (3)$$

Finally, when condition (3) is become as the expectation, it can be stated that the proposed estimator gives more efficiency than combined ratio estimator. When comparing, the value of proposed estimator is smaller than MSE of combined ratio estimator.

4. Numerical example

In this section, we used the data of Kadilar and Cingi (2005). We have applied our proposed and combined ratio estimators on the data of apple production amount (as interest of variate) and number of apple trees (as auxiliary variate) in 854 villages of Turkey in 1999 (Source: Institute of Statistics, Republic of Turkey). First, we have stratified the data by regions of Turkey and from each stratum (region); we have randomly selected the samples (villages). By using the Neyman allocation (Cochran, 1977),

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{k} N_h S_h}.$$

Table 1 : Data Statistics

Total	Stratum	1	2	3	4	5	6
$N = 854$	N_h	106	106	94	171	204	173
$n = 140$	n_h	9	17	38	67	7	2
$X = 32110400$	X_h	20816250	23417534	61837286	63507710	22580614	8406776
$Y = 2502220$	Y_h	1311744	1889048	8013936	4772152	825818	345016
$C_x = 3.85$	C_{xh}	2.02	2.1	2.22	3.84	1.72	1.91
$C_y = 5.83$	C_{yh}	0.0049	0.0061	0.0037	0.0060	0.0029	0.0027
$S_x = 144794$	S_{xh}	49189	57461	160757	285603	45403	18794
$S_y = 17106$	S_{yh}	6425	11552	29907	28643	2390	946
$\rho = 0.92$	ρ_h	0.82	0.86	0.9	0.99	0.71	0.89
$R = 0.077926$	θ_h	0.102	0.049	0.016	0.009	0.138	0.006
$R_{st1} = 0.077926$	W_h^2	0.015	0.015	0.012	0.04	0.057	0.041

In Table 1, we observed the statistics about the population, strata, and sample's size.

Table 2: MSE value of ratio estimator

Estimators	Estimated value	MSE values
$\hat{Y}_{RC}$	2502220	8.863×10^{11}
$\hat{Y}_{st1}$	2502220	7.035×10^{11}

Table 2 is the values of MSE of proposed ratio estimator and combined ratio estimator.

It seems that the MSE value of the proposed ratio estimator is smaller than the MSE value of combined ratio estimator**.**

5. Conclusion

We have arised a new type of ratio estimator in stratified random sampling from the estimator of Phanida (2008) and obtained its MSE equation. In theory, the proposed estimator has a smaller MSE than the MSE of combined ratio estimator and it has a numerical example to support the better theoretical result.

References

[1] Cochran, W. G. Sampling Techniques. New York: John Wiley and Sons; 1977.

[2] Kadilar, C., Cingi, H. Ratio estimators in stratified random sampling. Biometrical J. 2003; 45(2):218–225.

[3] Kadilar,C.,and Cingi, H. Ratio estimatir in simple random sampling. Applied Mathematics and Computation. 2004; 151: 893-902.

[4] Kadilar, C., Cingi, H. A new ratio estimators in stratified random sampling. Communications in Statistics-Theory and Methods. 2005; 34: 1-6.

[5] Phanida Krailamao. Ratio estimators of the population tatal. Department of Statistics, Faculty of Science, Khon Kaen University; 2008.

[6] Pluttipat Sopradit. Combined ratio estimators of population mean in stratified random sampling. Department of Statistics, Faculty of Science, Khon Kaen University; 2012.

[7] Ray, S. K., Singh, R. K. Difference-cum-ratio type estimators. Journal of Indian Statistical Association. 1981; 19:147-151.

[8] Sisodia, B.V.S and Dwividi, V.K. A Modified ratio estimator using coefficient of variation of auxiliary variable. Journal of the Indian society of Agricultural Statistics. 1981; 33(1):13-18.

Development of fuzzy variable parameters $\overline{X}$ control charts by weighted variance method using α –cut under non-normality

Chaowalit Panthong [1] and Adisak Pongpullponsak[2*]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, chao-chao-kmutt@hotmail.com*
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140 Thailand, adisak.pon@kmutt.ac.th*

Abstract

Shewhart $\overline{X}$ Control Charts was developed continuously. Recently, Fuzzy theory was applied in uncertain data. This research came about the idea to develop a change of variable parameter $\overline{X}$ control charts by weighted variance method (VP-WV) to Fuzzy variable parameter $\overline{X}$ control charts by weighted variance method (FVP-WV) using $\alpha-$cut. In case of non-normality distribution data, lognormal distribution and Weibull distribution data study would be applied. The average number of observations to signal (ANOS), the adjusted average time to signal (AATS) and the average time to signal (ATS) were also used for assessing the efficiency of the control charts.

Keywords: VP control charts, fuzzy $\overline{X}$ control charts, $\alpha-$cut, non-normality distribution

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

.

1. Introduction

Many studied about adaptive $\overline{X}$ control charts these are developed by variable sampling intervals (VSI), variable sample sizes (VSS), variable sample sizes and sampling intervals (VSSI) and variable parameter(VP) $\overline{X}$ chart [1]. These studies are under the assumption that the data come from a normal distribution. The two topics in non-normal distribution data was presented by Yan-Kwang Chen [2] in the variable sampling interval (VSI) $\overline{X}$ control charts using Burr's distribution and Yu-Chang Lin and Chao-Yu Chou [3] proposed "On the design of variable sample size and sampling intervals $\overline{X}$ charts under non-normality" using Burr's distribution and "Non-normality and the variable parameters $\overline{X}$ control charts" using Gamma distribution and t distribution. The previous papers used the standard normal distribution to transform non-normality data to normal before constructing the control charts. The performance indicators of the control charts are calculated using the Markov chain approach.A. Pongpullponsak, W.Suracherkeiti and C.Panthong [4], studied the variable parameter $\overline{X}$ control charts using shewhart method(SH), the weighted variance method (WV) and the scaled weighted variance method (SWV) for skewed distribution. The weighted variance method (WV) and the scaled weighted variance method (SWV) create control chart from skewed data (which generate from Weibull, lognormal, and Burr's distribution) directly without any transformation. The economic model was the tools in investigating the performance of these control charts.

The problem in control charts making causes from uncertain data e.g. human errors, measuring devices, and environmental conditions. Many studied were done to combine statistical methods and fuzzy set theory. Fuzzy sets theory was first introduced by Zadeh [5]. In 2005 Zadeh outlined generalized theory of uncertainty (GTU) which presented a change of perspective and direction in thinking about the system and uncertainties [5]. Gullbay [6] suggested the $\alpha-$cut fuzzy control charts for linguistic data. Kahraman, C [9] proposed an alternative approach to fuzzy control charts: Direct fuzzy approach (DFA) was developed fuzzy approaches to control charts based on fuzzy transformation methods, used the trapezoidal membership function. Zarandi [8] presented a new hybrid method based on a combination of fuzzified sensitivity criteria and fuzzy adaptive sampling rules to determine the sample size and sample interval of the control charts in order to determine the sample size and sample interval of the control charts. in order to control and improve process efficiency at its best. It was discovered by Senturk and Erginel [5] that control charts could be used to solve the problem of uncertain data by using fuzzy theory. The topic of the research studied was fuzzy $\tilde{\bar{X}} - \tilde{R}$ and $\tilde{\bar{X}} - \tilde{S}$ control charts using $\alpha-$cut. The methods used in the transformation of fuzzy sets into scalars are fuzzy mode, fuzzy median and $\alpha-$level fuzzy midrange. Which one you choose to use depends on the difficulty of the computation or preference as in Wang [9]. The $\alpha-$level fuzzy midrange was selected.

approach. Thirdly $\alpha-$level fuzzy midranges for FVP-WV are calculated by using $\alpha-$level fuzzy midrange transformation techniques. Finally, we can use ANOS, AATS and ATS to determine the efficiency of the charts.

2. Research Methodology

In this study, FVP-WV will be considered which are developed from WV $\bar{X}$ control charts studied by Pongpullponsak, Suracherkiati and Panthong [4].These control charts have non–normal distribution data which are Weibull and lognormal

2.1 Control charts

2.1.1 Fuzzy Variable parameters $\bar{X}$ control charts by weighted variance method: FVP - WV

The studies of control charts of process of mean, which is often used by control chart for average, control charts for the range or control chart for the standard deviation. But the study of Pongpullponsak, Suracherkiati and Panthong [4] and Senturk, Erginel [8] modified VP-WV to FVP-WV, that uses membership represented by a triangular fuzzy number (a,b,c) .Therefore, the control limits are

$$U\tilde{C}L_i = \overline{\overline{X}}_i + \frac{k_i W_{U_i} \overline{R}_i}{3} = \overline{\overline{X}}_{a,i} + \frac{k_i W_{U_{a,i}} \overline{R}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{k_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}_{c,i} + \frac{k_i W_{U_{c,i}} \overline{R}_{c,i}}{3}, \quad (1)$$

$$U\tilde{C}L_i = \overline{\overline{X}}_i + \frac{w_i W_{U_i} \overline{R}_i}{3} = \overline{\overline{X}}_{a,i} + \frac{w_i W_{U_{a,i}} \overline{R}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{w_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}_{c,i} + \frac{w_i W_{U_{c,i}} \overline{R}_{c,i}}{3}, \quad (2)$$

$$C\tilde{L}_i = (\overline{\overline{X}}_{a,i}, \overline{\overline{X}}_{b,i}, \overline{\overline{X}}_{c,i}), \quad (3)$$

$$L\tilde{W}L_i = \overline{\overline{X}}_i - \frac{w_i W_{U_i} \overline{R}_i}{3} = \overline{\overline{X}}_{a,i} - \frac{w_i W_{U_{a,i}} \overline{R}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{w_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}_{c,i} - \frac{w_i W_{U_{c,i}} \overline{R}_{c,i}}{3}, \quad (4)$$

$$L\tilde{C}L_i = \overline{\overline{X}}_i - \frac{k_i W_{U_i} \overline{R}_i}{3} = \overline{\overline{X}}_{a,i} - \frac{k_i W_{U_{a,i}} \overline{R}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{k_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}_{c,i} - \frac{k_i W_{U_{c,i}} \overline{R}_{c,i}}{3}, \quad (5)$$

where $i = 1, 2$

2.1.2 $\alpha-$cut fuzzy weighted variance method control chart

An $\alpha-$cut consists of any elements whose membership is greater than or equal to α . Applying $\alpha-$cut of fuzzy sets, the values of $\overline{\overline{X}}_{a,i}$, $\overline{\overline{X}}_{c,i}$, $\overline{R}_{a,i}$, $\overline{R}_{c,i}$ are determined as follows. First of all, we

The aim of this study is to introduce the framework of FVP-WV which are lognormal and Weibull distributions, using $\alpha-$cut with the transform VP-WV to FVP-WV. To obtain FVP-WV charts, triangular fuzzy numbers (a,b,c) are used. Secondly $\alpha-$cut FVP-WV charts are developed by using $\alpha-$cut as follows:

$$\overline{\overline{X}}^{\alpha}_{a,i} = \overline{\overline{X}}_{a,i} + \alpha(\overline{\overline{X}}_{b,i} - \overline{\overline{X}}_{a,i}), \quad (6)$$

$$\overline{\overline{X}}^{\alpha}_{c,i} = \overline{\overline{X}}_{c,i} - \alpha(\overline{\overline{X}}_{c,i} - \overline{\overline{X}}_{b,i}), \quad (7)$$

$$\overline{R}^{\alpha}_{a,i} = \overline{R}_{a,i} + \alpha(\overline{R}_{b,i} - \overline{R}_{a,i}), \quad (8)$$

$$\overline{R}^{\alpha}_{c,i} = \overline{R}_{c,i} - \alpha(\overline{R}_{c,i} - \overline{R}_{b,i}), \quad (9)$$

Therefore, the $\alpha-$cut fuzzy mean control limits by weighted variance method are

$$U\tilde{C}L_i^{\alpha} = \overline{\overline{X}}^{\alpha}_{a,i} + \frac{k_i W_{U_{a,i}} \overline{R}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{k_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}^{\alpha}_{c,i} + \frac{k_i W_{U_{c,i}} \overline{R}^{\alpha}_{c,i}}{3}, \quad (10)$$

$$U\tilde{W}L_i^{\alpha} = \overline{\overline{X}}^{\alpha}_{a,i} + \frac{w_i W_{U_{a,i}} \overline{R}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{w_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}^{\alpha}_{c,i} + \frac{w_i W_{U_{c,i}} \overline{R}^{\alpha}_{c,i}}{3}, \quad (11)$$

$$C\tilde{L}_i^{\alpha} = \left(\overline{\overline{x}}^{\alpha}_{a,i}, \overline{\overline{x}}_{b,i}, \overline{\overline{x}}^{\alpha}_{c,i} \right), \quad (12)$$

$$L\tilde{W}L_i^{\alpha} = \overline{\overline{X}}^{\alpha}_{a,i} - \frac{w_i W_{U_{a,i}} \overline{R}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{w_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}^{\alpha}_{c,i} - \frac{w_i W_{U_{c,i}} \overline{R}^{\alpha}_{c,i}}{3}, \quad (13)$$

$$L\tilde{C}L_i^{\alpha} = \overline{\overline{X}}^{\alpha}_{a,i} - \frac{k_i W_{U_{a,i}} \overline{R}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{k_i W_{U_{b,i}} \overline{R}_{b,i}}{3}, \overline{\overline{X}}^{\alpha}_{c,i} - \frac{k_i W_{U_{c,i}} \overline{R}^{\alpha}_{c,i}}{3}, \quad (14)$$

where $i = 1, 2$

2.1.3 $\alpha-$level fuzzy midrange for $\alpha-$cut fuzzy mean by weighted variance method control chart

A $\alpha-$level fuzzy midrange for $\alpha-$cut fuzzy mean control limits by weighted variance method control charts is developed to suggest construction of control charts, which is more efficient, suitable, and statistical tool. Therefore, the $\alpha-$level fuzzy midrange for $\alpha-$cut fuzzy mean control limits by weighted variance method control limits are

$$U\tilde{C}L^{\alpha}_{mr-i} = \left(\frac{\overline{\overline{X}}^{\alpha}_{a,i} + \overline{\overline{X}}^{\alpha}_{c,i}}{2} \right) + \frac{k_i W_{U,i} \overline{R}^{\alpha}_{a,i} + \overline{R}^{\alpha}_{c,i}}{6}, \quad (15)$$

$$U\tilde{W}L^{\alpha}_{mr-i} = \left(\frac{\overline{\overline{X}}^{\alpha}_{a,i} + \overline{\overline{X}}^{\alpha}_{c,i}}{2} \right) + \frac{w_i W_{U,i} \overline{R}^{\alpha}_{a,i} + \overline{R}^{\alpha}_{c,i}}{6}, \quad (16)$$

$$\tilde{C}L^{\alpha}_{mr-i} = \left(\frac{\overline{\overline{X}}^{\alpha}_{a,i} + \overline{\overline{X}}^{\alpha}_{c,i}}{2} \right), \quad (17)$$

$$L\tilde{W}L^{\alpha}_{mr-i} = \left(\frac{\overline{\overline{X}}^{\alpha}_{a,i} + \overline{\overline{X}}^{\alpha}_{c,i}}{2} \right) - \frac{w_i W_{U,i} \overline{R}^{\alpha}_{a,i} + \overline{R}^{\alpha}_{c,i}}{6}, \quad (18)$$

$$U\tilde{C}L^{\alpha}_{mr-i} = \left(\frac{\overline{\overline{X}}^{\alpha}_{a,i} + \overline{\overline{X}}^{\alpha}_{c,i}}{2} \right) - \frac{k_i W_{U,i} \overline{R}^{\alpha}_{a,i} + \overline{R}^{\alpha}_{c,i}}{6}, \quad (19)$$

where $i = 1, 2$

2.2 Fuzzy transformation techniques

Fuzzy transformation techniques have four types: fuzzy mode, fuzzy median, fuzzy average and $\alpha-$level fuzzy midrange, in this study, the $\alpha-$level fuzzy midrange transformation technique is used for FVP-WV charts.

2.2.1 $\alpha-$level Fuzzy midrange

The $\alpha-$level fuzzy midrange f^{α}_{mr} be defined as the midpoint of the $\alpha-$level cuts. Let A^{α} is $\alpha-$level cuts, nonfuzzy sets that consist of any elements whose membership is greater than or equal to α. If a^{α} and b^{α} are end points of A^{α} then

$$f^{\alpha}_{mr} = \frac{1}{2}(a^{\alpha} + c^{\alpha}) \qquad (20)$$

In fact the fuzzy mode is a special case of $\alpha-$level fuzzy midrange when $\alpha=1$. The definition of $\alpha-$level fuzzy midrange of sample j for fuzzy $\tilde{\overline{\overline{X}}}$ control charts is

$$S^{\alpha}_{mr-\overline{X},j} = \frac{(\overline{X}_{a_j} + \overline{X}_{c_j}) + \alpha\left[(\overline{X}_{b_j} - \overline{X}_{a_j}) - (\overline{X}_{c_j} - \overline{X}_{b_j})\right]}{2} \qquad (21)$$

Then, the condition of process control for each sample can be defined as
Process control =

$$\begin{cases} warning\ control & ; UCL^{\alpha}_{mr-\overline{X}} \geq S^{\alpha}_{mr-\overline{X},j} > UWL^{\alpha}_{mr-\overline{X}} \\ & and\ LWL^{\alpha}_{mr-\overline{X}} < S^{\alpha}_{mr-\overline{X},j} \leq LWL^{\alpha}_{mr-\overline{X}} \\ in\ \ control & ; LWL^{\alpha}_{mr-\overline{X}} \leq S^{\alpha}_{mr-\overline{X},j} \leq UWL^{\alpha}_{mr-\overline{X}} \\ out\ of\ control & ; S^{\alpha}_{mr-X,j} < LCL^{\alpha}_{mr-\overline{X}} \\ & and\ \ S^{\alpha}_{mr-\overline{X},j} > UCL^{\alpha}_{mr-\overline{X}} \end{cases} \qquad (22)$$

2.3 Distributed Data

2.3.1 Weibull distribution

Weibull distribution is continuous distribution that is used widely. Let X be continuous random variables that are Weibull distribution with $\beta > 0$ and $\theta > 0$.

Density function

$$f(x;\theta,\beta) = \frac{\beta}{\theta^{\beta}} x^{\beta-1} e^{-(x/\theta)^{\beta}}\ ;\ x > 0, \qquad (23)$$

Cumulative distribution function

$$F(x;\theta,\beta) = 1 - e^{-(x/\theta)^{\beta}}\ ;\ \ x > 0, \qquad (24)$$

Mean

$$\mu = E(X) = \frac{\theta}{\beta}\Gamma(\frac{1}{\beta}), \qquad (25)$$

Variance

$$\sigma^2 = V(X) = \frac{\theta^2}{\beta}\left\{2\Gamma(\frac{2}{\beta}) - \frac{1}{\beta}\left[\Gamma(\frac{1}{\beta})\right]^2\right\} \qquad (26)$$

where θ is scale parameter and
β is shape parameter.

In this study $\theta = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9$ and β are relevant with a coefficient of skewness at $\tau \in \{0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ shown in table 1.

Table 1: A coefficient of skewness and shape parameter of Weibull distribution

Coefficient of skewness (τ)	Shape parameters (β)
0.1	3.2219
0.5	2.2110
1	1.5630
2	1.0000
3	0.7686
4	0.6478
5	0.5737
6	0.5237
7	0.4873
8	0.4596
9	0.4376

2.3.2 Lognormal distribution

Lognormal distribution is correlated with normal distribution but random variables have positive values. Let X be continuous random variables that are lognormal distribution.

Density function

$$f(x;\mu,\sigma) = \frac{1}{x\hat{\sigma}\sqrt{2\pi}} \exp\left(\frac{\left(\ln x - \hat{\mu}\right)^2}{2\hat{\sigma}^2}\right), \qquad (27)$$

Cumulative distribution function

$$F_X(x,\mu,\sigma) = \frac{1}{2}\left[1 + erf\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right] = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \qquad (28)$$

Mean

$$\mu = E(X) = e^{\mu+\frac{\sigma^2}{2}} , \qquad (29)$$

Variance

$$\sigma^2 = V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) , \qquad (30)$$

where μ is scale parameter and

σ is shape parameter

In this study $\mu = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9$ and σ are relevant with a coefficient of skewness at $\tau \in \{0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ shown in table 2.

Table 2: A coefficient of skewness and shape parameter of lognormal distribution

Coefficient of skewness (τ)	Shape parameters (β)
0.1	0.0334
0.5	0.1641
1	0.3142
2	0.5513
3	0.7156
4	0.8326
5	0.9202
6	0.9889
7	1.0446
8	1.0911
9	1.1307

2.4 Calculations of ANOS, ATS and AATS

The operation of VP-WV, the design parameter should be chosen such that $n_1 > n_2$, $h_1 < h_2$, $w_1 < w_2$ and $k_1 < k_2$ as shown in Costa [11]. If the sample point falls in the central region $(LWL_i < \overline{X}_i < UWL_i)$, the control is relaxed by using the small size(n_1), the long interval (h_1) and the wide action limit coefficient (k_1).If the sample point falls in the warning region

$$(LCL_i < \overline{X}_i < LWL_i \ and \ UWL_i < \overline{X}_i < UCL_i) ,$$

the control is tighten the large size(n_2), the short interval (h_2) and the narrow action limit coefficient (k_2). If the sample point fall in the action region $(\overline{X}_i < LCL_i \ and \ \overline{X}_i > UCL_i)$, then the process is considered out of control.

The parameters calculation based the probability p_0 as follows,

$$p_0 = P(|M| < w_i \mid |M| < k_i) , i = 1, 2 , \qquad (31)$$

where M is non-normal random variable .

Calculations of *ANOS*, *ATS* and *AATS* The calculated of this value are applied. The Markov chain approach according to the condition as following.

State1. The process is in-control and the sample point fall in the central regions.

State2. The process is in-control and the sample point fall in the warning regions.

State3. The process is out-of-control and the sample point fall in the action regions (absorbing state).

The widely used performance indicators for adaptive control charts include [2]:

1. *ANOS*—the average number of observations to signal, which is defined as the expected number of individual observations from the start of the process to the time when the chart indicates an out-of-control signal.
2. *ATS*—the average time to signal, which is defined as the expected value of the time from the start of the process to the time when the chart indicates an out-of-control signal.
3. *AATS*—the adjusted average time to signal, which is defined as the expected value of the time from the occurrence of an assignable cause to the time when the chart indicates an out-of control signal.

Let Q be the state transition probability matrix:

$$Q = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} , \qquad (32)$$

where p_{ij} are probability which sample in stage *i* and sample point fall in state *j*. Let P be the transition probability matrix when the process is in – control:

$$P = \begin{bmatrix} P_1 & 1 - P_1 \\ P_2 & 1 - P_2 \end{bmatrix} , \qquad (33)$$

where P_i as the conditional probability that the sample size is n_i and the sample point fall in the central regions, i = 1, 2 the *ANOS*, *ATS* and *AATS* can be calculated by

$$ATS = r^{'}(I-Q)^{-1}h', \qquad (34)$$

$$ANOS = r^{'}(I-Q)^{-1}n', \qquad (35)$$

$$AATS = \pi^{'}\left[(I-Q)^{-1}-\frac{1}{2}\right]h', \qquad (36)$$

where $n^{'} = [n_1, n_2]$ is the vector of the sample size, $h^{'} = [h_1, h_2]$ is the vector of the sampling interval, I is the I dentity matrix with order 2, 1 is a 2×1 unit column vector and $r^{'} = [r_1, r_2]$ as the vector of steady – state probability when the process is in-control.

3. Results and Discussion

The purpose of this study is to compare the efficiency of FVP-WV control charts using $\alpha-$cut under non-normality are Weibull and lognormal distributions which have various values of the coefficient of skewness which are 0.1, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, number of class , number of sample size are randomly generated from Weibull and lognormal distributions with relevant with a coefficient of skewness shown in table 1 and 2. The procedure is repeated 10,000 times for shift sizes of δ = 0.5, 1.0, 1.5, 2.0, 2.5 and 3.0. From this study, results are shown as follows:

3.1 The process in control

Table 3: The *ATS* of lognormal and Weibull distribution when $\delta = 0\sigma$

Δ	n_1	n_2	k_{11}	k_{12}	k_{21}	k_{22}	w_{11}	w_{12}
0.1	3	4	3.14	3.15	2.28	2.29	2.07	2.5
0.5	3	5	3.41	3.14	2.50	2.50	2.00	1.99
1	5	7	3.09	3.22	3.00	3.00	2.22	2.03
2	6	9	3.33	2.90	2.42	2.53	2.13	2.43
3	3	8	3.12	2.58	2.17	2.56	2.10	2.00
4	4	8	3.28	3.70	2.14	2.44	1.10	2.10
5	6	7	3.09	3.41	2.02	1.79	1.05	1.45
6	4	9	3.32	3.09	2.55	2.92	1.98	2.01
7	3	7	3.10	3.30	2.77	2.65	2.06	2.22
8	5	8	3.22	3.10	2.01	2.19	2.00	2.74
9	6	12	3.09	3.12	2.82	2.60	1.04	2.10

Table 3: (continuous) The *ATS* of lognormal and Weibull distribution when $\delta = 0\sigma$.

Δ	w_{21}	w_{22}	h_1	h_2	$ATS-\log$	$ATS-$Weibull
0.1	2.00	2.10	1.50	0.10	255.50	235.32
0.5	1.95	1.99	1.32	0.50	240.87	220.42
1	1.11	2.03	1.30	1.00	240.22	220.40
2	1.92	1.24	1.20	1.05	228.12	220.02
3	1.75	2.00	2.30	1.07	214.48	218.23
4	2.02	0.50	2.15	0.09	208.76	210.50
5	2.00	1.45	3.50	0.44	200.11	205.00
6	1.92	2.01	1.30	1.00	192,37	200.65
7	0.95	1.33	1.20	1.10	182.09	193.55
8	0.92	0.74	2.80	1.67	175.07	188.87
9	1.19	2.12	1.50	0.10	160.46	170.00

From Table 3 and Figure 1 show *ATS* in the process in control both distributions will be reduced *ATS* when coefficient of skewness increasing at coefficient of skewness at 0.1 , 0.5 , 1.0 and 2.0. The lognormal distribution will provide *ATS* more than the Weibull distribution. On the other hand the coefficient of skewness 3.0 , 4.0 , 5.0 , 6.0 , 7.0 , 8.0 and 9.0 The Weibull distribution will provide *ATS* more than lognormal distribution.

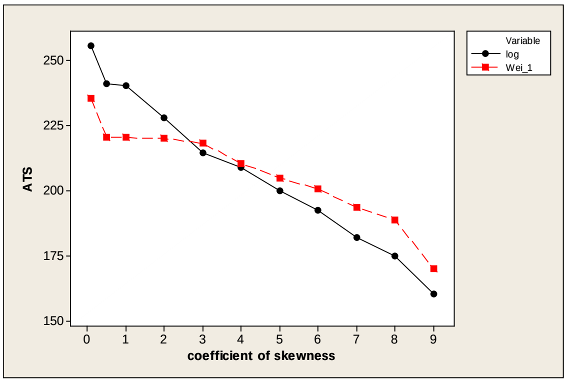


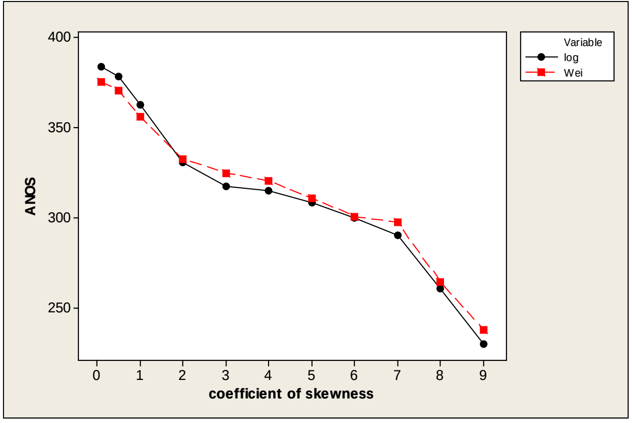
Figure 1: The *ATS* of lognormal and Weibull distribution when $\delta = 0\sigma$

3.2 The process out of control

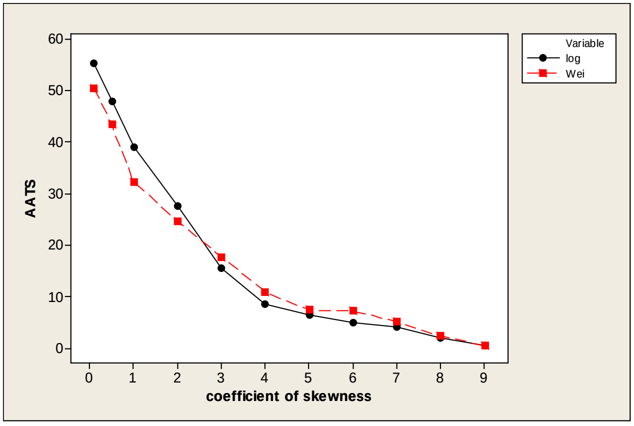
The process is shifted, we will present experimental results with graph only as shown in Figure 2(a-l).

$\delta = 0.5\sigma$

(a)

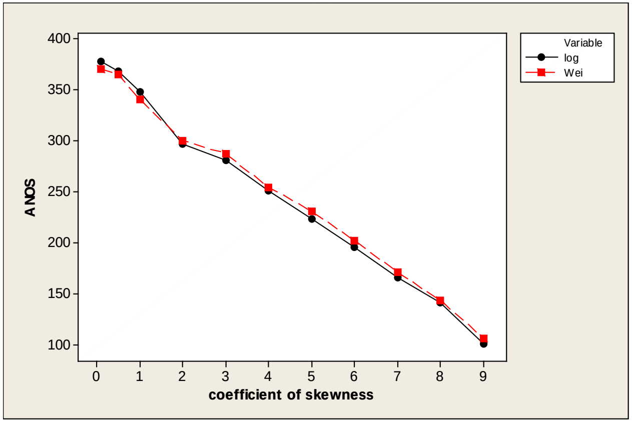


(b)

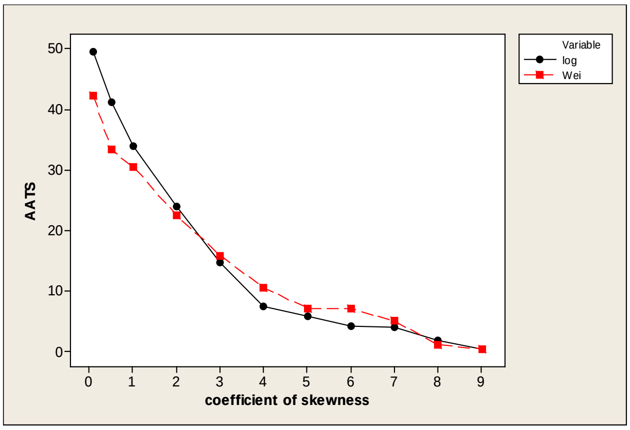


$\delta = 1.0\sigma$

(c)

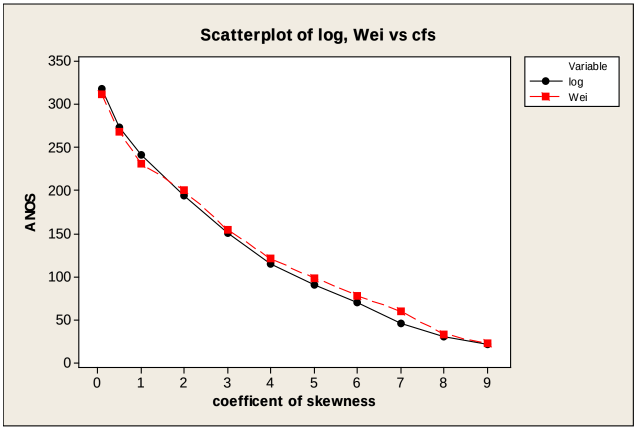


(d)

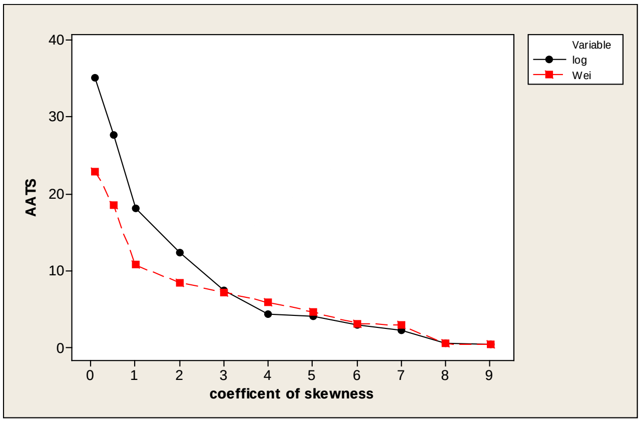


$\delta = 1.5\sigma$

(e)

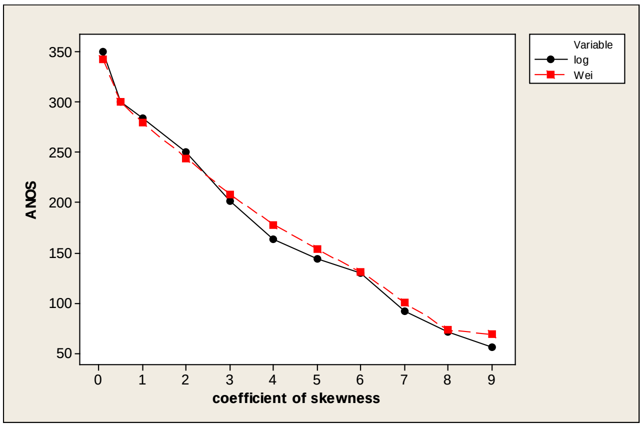


(f)

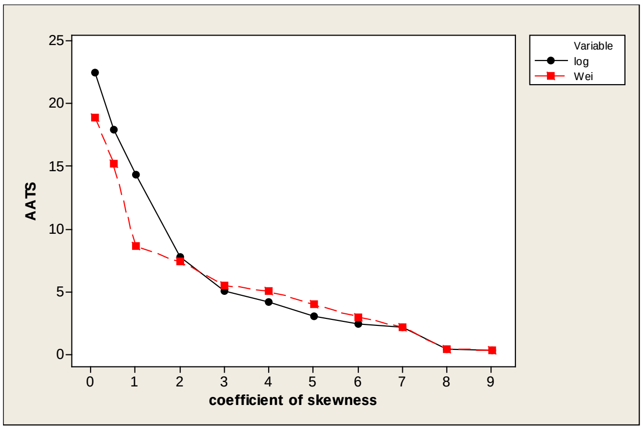


$\delta = 2.0\sigma$

(g)



(h)



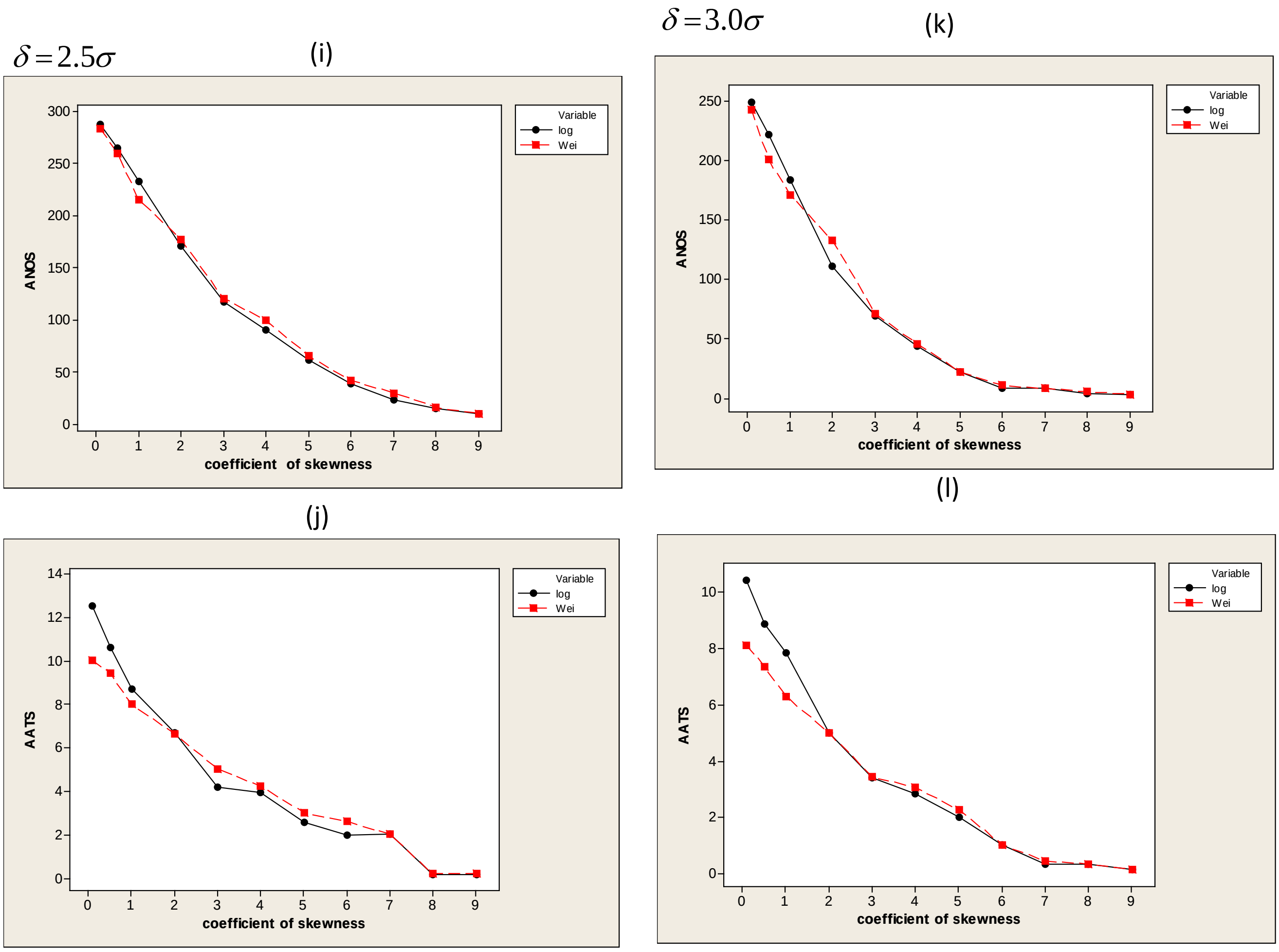


Figure 2: The ANOS and AATS of lognormal and Weibull distribution when process is shifted.

In Figure 2, to monitor the performance of the control charts by *ANOS* and *AATS* calculated by topic 2.4, the result is the level of process control shifts and the coefficient of skewness increase in *ANOS* and decline in *AATS*. When *ANOS* is compared to both distributions in coefficient of skewness 0.1, 0.5 and 1.0 from Weibull distribution and the result of *ANOS* is lower than lognormal distribution in every level of change control process. *AATS* at coefficient of skewness is 0.1 , 0.5 , 1.0 and 2.0. The result of *ANOS* in Weibull distribution will lower than lognormal distribution in every level of process control charts.

4. Conclusions

The purpose of this research is to propose the development of control charts from the variable parameters $\overline{X}$ control charts by weighted variancemethod (VP-WV) to fuzzy variable parameter control charts by weighted variance method (FVP-WV) by fuzzy theory starts from a triangular fuzzy number. To generates random data to control the resolution and coverage.Then cut unnecessary information by using $\alpha-$cut and random sampling technique with $\alpha-$level fuzzy midrange. The process of control charts to consider with two cases. First one the process is in control take ATS to check coefficient of skewness start at less value (0.1, 0.5, 1.0 and 2.0) from lognormal distribution,which butter than Weibull distribution. Another part of coefficient of skewness from Weibull distribution have more effienced than lognormal distribution. Second one the process in control is change and determined from *AATS* and *ANOS* at the lowest when coefficient of skewness is less (0.1, 0.5 and 1.0) from lognormal distribution have more efficocied of Weibull distribution. The coefficient of skewness increase the Weibull distribution have more efficacied than lognormal distribution both of distribution from the level of process control shifts. Further research,we will use the fuzzy thory to developd the chart as weighted variance method

(WV), scaled weighted variance method (SWV) or we may use the examine the cost of effectiveness of control charts or we may study data under other distributions such as student's t , gramma or burr's distributions etc.

Acknowledgements
The authors wish to thank Advisor for the valuable co-operation and Faculty of Mathematics**.**

References

[1] De Magalhaes MS., Costa A.F.B. and MouraNeto, F.D. A hierarchy of adaptive $\overline{X}$ control charts. Int. J. Production Economics. 2009; 11: 271–283.

[2] Yan-Kwang C. Economic design of control chart for non-normal data using variable sampling policy. International Journal of Production Economics. 2004; 92: 61-74.

[3] Lin Y.C. and Chou C.Y. Non-normality and the variable parameters $\overline{X}$ control charts. European Journal of Operational Research. 2007; 176: 361–373.

[4] Pongpullponsak A., Suracherkiati W. and Panthong C. The economic model of $\overline{X}$ control chart using shewhart method for skewed distribution. Thailand Statistician. 2009; 7(1): 81-99.

[5] Zadeh L.A. Fuzzy Sets. Information and Control . 1965; 8: 338-353.

[6] Gullbay M. and Kahraman C. An Alternative Approach to Fuzzy Control Charts:Direct Fuzzy Approach. Information Science. 2006; 77: 1463-1480.

[7] Zarandi M.H., Alaeddini A. and Turksen I.B. A Hybrid Fuzzy Adaptive Sampling Run Rules for Shewhart Control Charts. Information Sciences. 2008; 178: 1152-1170.

[8] Senturk S. and Erginel N. Development of Fuzzy $\overline{X}-R$ and $\overline{X}-S$ Control Charts Using α-cuts. Information Science. 2009; 179: 1542-1551.

[9] Wang J.H. and Raz T. On the Construction of Control Charts Using Linguistic Variables. Intelligent Journal of Production Research. 1990; 28: 477-487.

[10] Pongpullponsak A., Suracherkiati W. and Intaramo K. Development of fuzzy extreme value theory control using $\alpha-$cut for skewed populations. Applied Mathematical Sciences. 2012; 6(117): 5811 – 5834.

[11] Costa A.F.B. Joint $\overline{X}$ chart with variable parameters.Journal of Quality Technology 31(1999): 408–416.

Symptoms and anthropometric data related to normals and severity of Obstructive Sleep Apnea patients by using factor analysis

Thaya Maranate[1], Adisak Pongpullponsak[2*] and Pimon Ruttanaumpawan[3]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
thaya_mar@hotmail.co.th
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140 Thailand,*
adisak.pon@kmutt.ac.th
[3]*Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, 10700, Thailand,*
pimon.rut@mahidol.ac.th

Abstract

Obstructive sleep apnea (OSA) is common and potentially causes adverse consequences if left untreated. Recently, the awareness of OSA amongst Thai physicians has been increasing. Meanwhile, the gold standard to diagnose OSA, polysomnography (PSG), requires sophisticated equipment, well trained sleep technicians and physicians. Therefore, there has been a problem of shortage of sleep laboratories that can accommodate the patients in a timely manner. If there are clinical factors that can predict the severity of OSA, the physicians will be able to determine the priority of patients' waiting for PSGs. Factor analysis was applied to analyze sleep questionnaires and anthropometric data of 1,042 patients suspected to have OSA at Siriraj Sleep Center in 2010-2011. The objective was to determine the clinical factors that can predict the severity of OSA, represented by Respiratory Disturbance Index(RDI). All 113 variables obtained from sleep questionnaires and anthropometric measurement were categorized into 13 domains as the followings: underlying diseases; body built; alcohol and tobacco consumption; decreased work-hour per day; decreased mental and physical performance; excessive daytime sleepiness; decreased sleeping hour per night; insomnia; surreal dream; causes and frequency of nighttime awakening; witnessed snoring; witnessed apnea and most comfortable sleep posture. It is suggested that sleep questionnaires in specific domains and anthropometric data may be useful to determine the priority of patients' waiting for PSGs.

Keywords: Factor analysis, severity of OSA, RDI, sleep questionnaire, anthropometric data

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

1. Introduction

Obstructive sleep apnea (OSA) is characterized by repetitive partial or complete collapse of the upper airway during sleep, resulting in sleep fragmentation and cyclic oxygen desaturations. Epidemiological studies revealed the prevalence of OSA in general population to be approximately 2-5%. Although OSA can occur at any age, it typically affects people aged between 40 and 60 years. The prevalence of OSA increases with age and is twice higher among men than women (4% vs. 2%). Additionally, obesity and weight gain have been shown to be important risk factors in the development and progression of OSA in middle-age adults [1].

Polysomnography (PSG) is the gold standard to diagnose and classify the severity of OSA. Respiratory Disturbance Index (RDI) is defined as the total number of apneas, hypopneas and respiratory effort-related arousals (RERAs) per hour of sleep (events/hour). It is derived from PSG and is widely used to define the presence as well as the severity of OSA, i.e. normal: RDI < 5, mild OSA: RDI 5-14.9, moderate OSA:RDI 15-30, and severe OSA: RDI> 30. Unfortunately, the PSG has not been widely implemented in Thailand due to the shortage of sleep laboratories. This results in a long waiting list which ranges from several months to years in some sleep laboratories. The priority for the appointment should be made on the basis of severity of OSA instead of "First come, first serve". Thus, physicians need a simple and non-expensive tool for predicting the severity of OSA. Sleep questionnaires and anthropometric measurement may be useful. To date, several studies have failed to demonstrate any clinical factors that can predict the severity of OSA [2]. Therefore, we aimed to find such clinical factors by using different statistical method. The objective of our study was to define the groups of clinical factors obtained from sleep questionnaires and anthropometric measurement that would be significantly related to each level of OSA severity and normals in patients suspected to have OSA by using factor analysis. The study was approved by the Committee of Siriraj Institutional Review Board, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand.

2. Research Methodology

This research is a quantitative study with analyzed materials included Siriraj sleep questionnaires and anthropometric data of 1,042 patients suspected to have OSA from medical sleep clinic, Siriraj Hospital during 2010-2011. The sleep questionnaire consisted of patient characteristics, underlying diseases, chief complaints, troubles at night or during sleep, frequency and causes of nighttime awakening, Epworth sleepiness scale (excessive daytime sleepiness), sleep related symptoms and history of related-accidents etc. The anthropometric measurement included body weight, height, neck, waist and hip circumferences, and thyromental distance. The databases were checked for accuracy by looking for overlapped information, missing items and incorrect coding. All 1,042 patients suspected to have OSA were classified into 4 groups by using RDI derived from PSGs: 62 normals; 178 mild OSA, 262 moderate OSA and 540 severe OSA. There were altogether 113 variables which were independent variables, whereas RDI was dependent variable.

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The number of variables can be reduced by the combining multiple variables in the same factor. The factors can be considered as a new variable and can be called factor score for further statistical analysis [3].

The hypothesis about the structure of that factor can be tested. Each factor consists of some variables and each variable should be correlated with the rate of weight gain and concluded that in order to test whether this factor models meet or match existing theories or not. This model is called confirmatory factor analysis model (CFA) which is a technique of factor analysis. Study of the structure of relationships among variables is processed by making the correlation structure of the variables. Factor analysis technique is to find the correlation coefficient for each pair of variables, then to include variables related to the same factor. It can analyze the structure showing the relationship of the variables that are in the same component and is applied to explain the meaning of each of the factors that is the meaning of the variables which are in the factor to be used in the planning. To understand the meaning of the factors in order to bring factors to variable for analysis is a very useful practice which can explain the meaning or comparable factors. The used theories are as followed.

2.1 Pearson correlation

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (1)$$

Where, r = Pearson correlation coefficient

x = Values in first set of data

y = Values in second set of data

n = Total number of values.

However, we need to perform a significance test to decide whether based upon this sample, there is any or no evidence to suggest that linear correlation is present in the population. To do this we test the null hypothesis, H_0, that there is no correlation in the population against the alternative hypothesis, H_1, that there is correlation; our data will indicate which of these opposing hypotheses is most likely to be true. We can thus express this test as: H_0: $\rho = 0$, H_1: $\rho \neq 0$ [3]

2.2 Kaiser-Mayer-Olkin (KMO) and Bartlett's test of sphericity

$$KMO = \frac{\sum r_i^2}{\sum r_i^2 + \sum (partial\ \ correlation)^2} \quad (2)$$

Where r = correlation coefficient, which is 0 < KMO <1

The KMO value shows whether the technique is appropriate for factor analysis. Generally, if the KMO < 0.5 is assumed that it is not reasonable to use factor analysis.

Bartlett's test of sphericity is used to test statistical hypotheses. H_0: Correlation matrix is the identity matrix, or all variables were correlated , H_1: the variables are not related. Therefore, the H_0 indicates that the variables are related and factor analysis should be used (significant < 0.05) [4].

2.3 Principal Component Analysis (PCA)

PCA is a mathematical technique that aims to gather the details of the more variables into fewer factors, by considering on the details of each variable.PCA analysis will create a linear combination of variables. Linear combination of factor 1 has the details of all the variables or the maximum variance. Linear combination of factor 2 is able to bring a detail remaining as possible of factor 1 into factor 2. Factor 1 will be orthogonal with factor 2 and factor 1 not correlated with factor 2 (likewise Factor 4… 5.). Factor loading is the value used to determine the relation among variables, should be in the same factor variables which are great value converge +1 or -1 are in the same factor. [4].

$$\begin{array}{l} \mathrm{x}_1 = \ell_{11}\mathrm{F}_1 + \ell_{12}\mathrm{F}_2 + \ldots + \ell_{1\mathrm{m}}\mathrm{F}_\mathrm{m} + \mathrm{e}_1) \\ \quad \cdot \quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \quad \cdot \quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \quad \cdot \quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \mathrm{x}_\mathrm{p} = \ell_{\mathrm{p}1}\mathrm{F}_1 + \ell_{\mathrm{p}2}\mathrm{F}_2 + \ldots + \ell_{\mathrm{pm}}\mathrm{F}_\mathrm{m} + \mathrm{e}_\mathrm{p}) \end{array}$$

Where X_1 , X_2 , . . ., X_p are p variables and F_1, F_2 ,…F_m are m factor orders, ℓ_{ij} is coefficient /weight of X_i which are called "factor loading", Communality = h^2 $(h^2) = \sum_{m=1}^{m} \ell_{ij}^2$, m = No. of common factor , p = No. of variable ,E_i = unique factor of i variable , F_i = w_1x_1+ w_2x_2+…+ w_ix_i ,F_i = score of factor i , w_i = weighted of variable i, and x_i= observed variable of i. It can be written in matrix form [5].

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1m} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{p1} & \ell_{p1} & \cdots & \ell_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix}$$

Eigen value is all variance in the original variables which can be explained by the factor loading [6].

$$\lambda_i = \sum_{i}^{n} \ell_{ik}^2 \qquad (3)$$

λ_i = Eigen value, i = number of variables

ℓ_{ik} = factor loading of variables i in factor order k

- Factor; a number of extracted factor must be or less than a number of variable.
- The two power of summation factor loading from all variables.
- Eigen value of Factor >1 can explain variance of variable set.

2.4 Factor Rotation (Varimax rotation)

The purpose of the axis rotation is to change more or less of the factor loading until the variables ought to be in some suitable factors and an orthogonal method of rotation that minimizes the number of variables with high loadings on a factor.

$$\upsilon = \sum_{j=1}^{r} \frac{n\sum_{i=1}^{n}(a_{ij}^2 / h_i^2)^2 - \left(\sum_{i=1}^{n} a_{ij}^2 / h_i^2\right)^2}{n^2} \qquad (4)$$

where h_i^2 is the communality of the *i* th test. Thus, the matrix acted upon by the normalized varimax method of rotation is the matrix $H^{-1}A$, where H^2 is an n×n diagonal matrix of communalities, and A is the initial n×r unrotated-factor-structure matrix. The result matrix, $B^* = H^{-1}A T_\upsilon$, where T_υ is an r×r orthogonal-transformation matrix determined by the normalized varimax criterion, is then modified into the “original metric” of the variables by premultiplying it by the diagonal matrix of the square roots of the communalities, ie. $B = HB^*$, with B the resulting rotated-factor-loading matrix to be interpreted [3].Therefore, the procedure of the research is summarized in Fig. 1 [7].

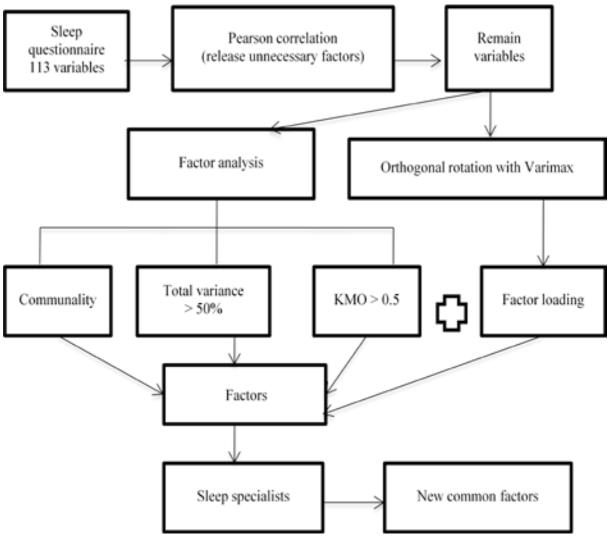


Figure 1: Procedure of the research

3. Research Results and Discussion

The results of the factor analysis of 4 groups which were normals (non OSA), mild, moderate and severe OSA respectively were demonstrated as followed.

3.1. Normal group (non OSA)

Data of the 62 normals to find the Pearson correlation with other variables significant at 0.05 and 0.01 had remaining 11 from 113 variables. KMO = 0.506 was assumed that it was reasonable to use factor analysis. Bartlett's test of sphericity had the significance = 0.000 and the result of 11 variables revealed correlation matrix was the identity matrix. The factor extraction from PCA had values > 0.5 with lowest value in the variable “you sweat profoundly during sleep” which was still not too low that it could be clearly grouped in some factors (Table 1).

Table 1: Communalities (normals)

Variables	Initial	Extraction
The most comfortable sleep posture	1.000	0.857
Periodical grasping	1.000	0.925
You sweat profoundly during sleep	1.000	0.626
Kidney diseases	1.000	0.775
Chest discomfort	1.000	0.823
Age	1.000	0.792
Stomach/intestinal diseases	1.000	0.824
Hypertension	1.000	0.809
Routine medicine use	1.000	0.650
Sitting inactive in a public place (sleepiness)	1.000	0.874
As a passenger in a car for an hour without break (sleepiness)	1.000	0.894

Each of 3 factors had Eigen value > 1. The total variances explained 3 factors was 80.455 % by extraction method: PCA (Table 2).The result of factor rotation (Varimax rotation) revealed 3 factors (11 variables) with new naming by sleep specialists that corresponded with variables (Table 3) and the scree plot showed new naming factors (Fig. 2).

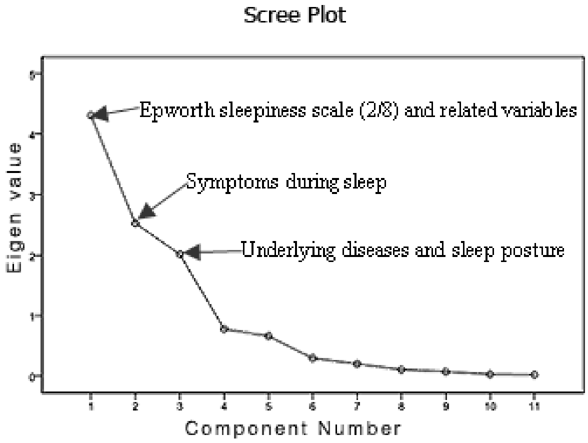


Figure 2: The 3 new naming factors in normal group on the slope of scree line that had Eigen value >1

Table 2: The initial Eigen values, extraction and rotation sums of squared loadings in normal group

Components (factor)	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.308	39.165	39.165	4.308	39.165	39.165	3.634	33.037	**33.037**
2	2.527	22.969	62.134	2.527	22.969	62.134	2.781	25.279	**58.316**
3	2.015	18.321	80.455	2.015	18.321	80.455	2.435	22.139	**80.455**
⋮	⋮	⋮	⋮						
11	.020	.185	100.000						

Table 3: The result of factor loading of 3 factors in normal group & new naming

New naming by sleep specialists	Variables	Components (factors)		
		1	2	3
1. Epworth sleepiness scale (2/8) and related variables	As a passenger in a car for an hour without break	**-.934**	.145	.003
	Sitting inactive in a public place	**-.929**	-.070	.071
	Age	**.880**	-.049	.120
	Routine medicine use	**.754**	-.158	.236
	Hypertension	**.666**	-.297	.527
2.Symptoms during sleep	Periodical grasping	.002	**.953**	-.129
	Chest discomfort	-.254	**.864**	.110
	You sweat profoundly during sleep	-.077	**.780**	-.110
3.Underlying diseases and sleep posture	The most comfortable sleep posture	.196	.010	**.905**
	Kidney diseases	.009	-.279	**.835**
	Stomach/intestinal diseases	.037	.545	**.725**

3.2 Mild OSA group

Data of the 178 mild OSA patients to find the Pearson correlation with other variables significant at 0.05 and 0.01 had remaining 17 from 113 variables. KMO = 0.667 was assumed that it was reasonable to use factor analysis. Bartlett's test of sphericity showed the significance = 0.000 and the result of 17 variables revealed correlation matrix was the Identity matrix. The factor extraction from PCA had values > 0.5 with lowest value in the variable “asthmatic attack” which was still not too low that it could be clearly grouped in some factors. (Table 4)

Each of 5 factors had Eigen value > 1. The total variances explained 5 factors was 78.656% by extraction method: PCA (Table 5). The result of factor rotation (Varimax rotation), showed 5 factors (17 variables) with new naming by sleep specialists that corresponded with variables (Table 6) and the scree plot showed new naming factors (Fig. 3).

Table 4: Communalities (mild OSA group)

Variables	Initial	Extraction
Work-hours/day	1.000	.863
Age	1.000	.726
Hypertension	1.000	.671
Average sleeping hours in each night	1.000	.866
Feeling like choking	1.000	.814
Asthmatic attack	1.000	.536
Chest discomfort	1.000	.853
Hot, cold, noisy, light	1.000	.739
Thirsty & hungry	1.000	.734
Loudness of your snoring	1.000	.693
Hearing your own snoring	1.000	.833
Lung diseases	1.000	.782
Headache & stomachache	1.000	.822
You wake up during late night: times/night	1.000	.7 66
Sitting and talking to someone(sleepiness)	1.000	.884
Sitting quietly after lunch(sleepiness)	1.000	.874
Sitting inactive in a public place (sleepiness)	1.000	.913

Table 5: The initial Eigen values, extraction and rotation sums of squared loadings in mild OSA group

Components (factor)	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.802	28.248	28.248	4.802	28.248	28.248	3.768	22.166	**22.166**
2	3.236	19.034	47.282	3.236	19.034	47.282	2.854	16.786	**38.951**
3	2.124	12.491	59.773	2.124	12.491	59.773	2.677	15.746	**54.697**
4	1.833	10.784	70.557	1.833	10.784	70.557	2.079	12.231	**66.928**
5	1.377	8.098	78.656	1.377	8.098	78.656	1.994	11.728	**78.656**
6	.737	4.337	82.992						
⋮	⋮	⋮	⋮						
17	.070	.412	100.000						

Table 6: The result of factor loading of 5 factors in mild OSA group & new naming

New naming by sleep specialists	Variables	Components (factors)				
		1	2	3	4	5
1.Causes of nighttime awakening	Chest discomfort	**.871**	-.120	.259	.016	.112
	Headache & stomachache	**.869**	.033	.223	.016	-.127
	Hot ,cold, noisy, light	**.843**	-.089	.084	.116	.005
	Thirsty & hungry	**.717**	.385	.267	-.004	.025
2.Epworth sleepiness scale(3/8)	Sitting inactive in a public place	-.102	**.936**	-.045	-.117	.103
	Sitting quietly after lunch	-.010	**.934**	.032	.014	.008
	Sitting and talking to someone	.113	**.916**	-.130	.019	.124
3.Lung diseases and sleep-wake pattern	Average sleeping hours in each night	-.125	.000	**-.917**	-.032	-.099
	Lung diseases	.299	-.144	**.695**	.235	-.366
	You wake up during late night: times /night	.347	.064	**.619**	.450	-.237
	Asthmatic attack	.306	-.097	**.601**	.175	.200
4.Risk variables and work	Age	.007	.042	.269	**.806**	-.043
	Hypertension	-.128	-.141	.260	**.742**	-.132
	Work-hours/day	-.439	-.006	.399	**-.715**	-.002
5.Choking and witnessed snoring	Hearing your own snoring	.017	.083	-.169	-.094	**.888**
	Feeling like choking	.492	.159	.167	-.200	**.692**
	Loudness of your snoring	-.486	.078	.196	.038	**.642**

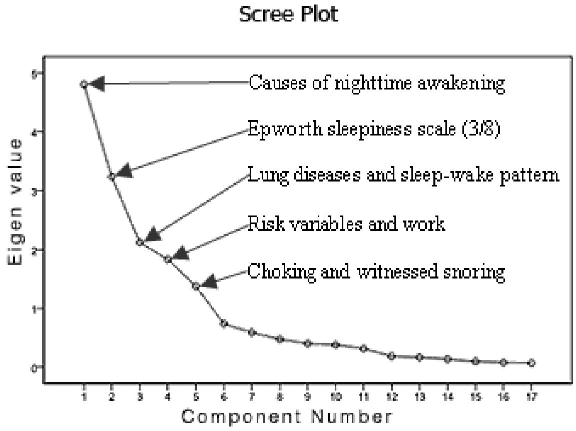


Figure 3: The 5 new naming factors in mild OSA group on the slope of scree line that had Eigen value >1

3.3 Moderate OSA group

Data of the 262 moderate OSA patients to find the Pearson correlation with other variables significant at 0.05 and 0.01 had remaining 19 from 113 variables. KMO = 0.780 was assumed that it was reasonable to use factor analysis. Bartlett's test of sphericity showed the significance = 0.000 and the result of 19 variables revealed correlation matrix was the identity matrix. The factor extraction from PCA had values > 0.5 which had lowest value in the variable “Epilepsy” which was still not too low that it could be clearly grouped in some factors (Table 7).

Table 7: Communalities (moderate OSA group)

Variables	Initial	Extraction
Epilepsy	1.000	.507
Surreal dream	1.000	.707
Loudness of your snoring	1.000	.692
Tobacco smoking	1.000	.667
Drink alcohol or beer: quantity	1.000	.832
Exhaust/inactive	1.000	.637
Decreased concentration	1.000	.707
Periodically stop breathing	1.000	.754
Depressing	1.000	.708
Sitting and reading (sleepiness)	1.000	.814
Sitting inactive in a public place (sleepiness)	1.000	.886
Sitting and talking to someone (sleepiness)	1.000	.752
Decreased your work or study ability	1.000	.782
As a passenger in a car for an hour without break (sleepiness)	1.000	.832
Stroke	1.000	.720
Watching TV (sleepiness)	1.000	.777
Lung diseases	1.000	.626
Chest discomfort	1.000	.692
Sitting quietly after lunch with no alcohol (sleepiness)	1.000	.835

Each of 5 factors had Eigen value > 1. The total variances explained 5 factors was 73.303% by extraction method: PCA (Table 8). The result of factor rotation (Varimax rotation), revealed 5 factors (19 variables) with new naming by sleep specialists that corresponded with variables (Table 9) and the scree plot showed new naming factors (Fig. 4).

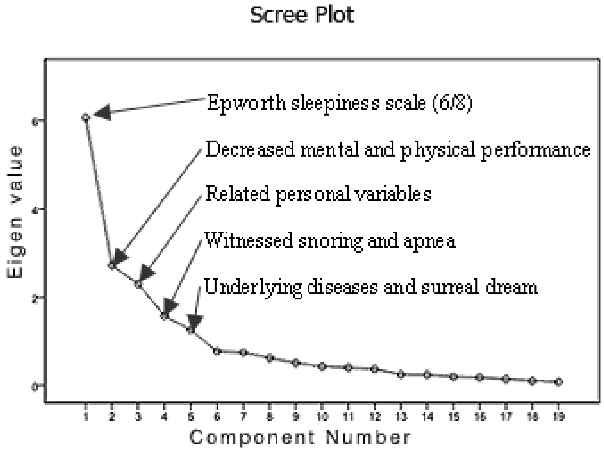


Figure 4: The 5 new naming factors in moderate OSA group on the slope of scree line that had Eigen value >1

Table 8: The initial Eigen values, extraction and rotation sums of squared loadings in moderate OSA group

Components (factor)	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.066	31.927	31.927	6.066	31.927	31.927	5.344	28.128	**28.128**
2	2.724	14.336	46.263	2.724	14.336	46.263	2.823	14.859	**42.987**
3	2.301	12.113	58.376	2.301	12.113	58.376	2.200	11.580	**54.567**
4	1.580	8.318	66.694	1.580	8.318	66.694	2.191	11.533	**66.100**
5	1.256	6.608	73.303	1.256	6.608	73.303	1.369	7.203	**73.303**
⋮	⋮	⋮	⋮						
19	.080	.423	100.000						

Table 9: The result of factor loading of 5 factors in moderate OSA group & new naming

New naming by sleep specialists	Variables	Components (factors)				
		1	2	3	4	5
1.Epworth sleepiness scale(6/8)	Sitting inactive in a public place	**.920**	.167	-.004	.000	.108
	Sitting quietly after lunch	**.894**	.075	.004	-.014	-.173
	Sitting and reading	**.893**	.053	-.085	.033	.069
	Watching TV	**.858**	.026	-.141	.126	.063
	Sitting and talking to someone	**.843**	.159	.006	.102	-.078
	As a passenger in a car for an hour without break	**.839**	.349	.023	.011	.083

Table 9: The result of factor loading of 5 factors in moderate OSA group & new naming (cont.)

2.Decreased mental and physical performance	Decreased concentration	.225	**.805**	-.040	.037	.076
	Depressing	-.139	**.770**	.046	.306	-.023
	Decreased your work or study ability	.377	**.747**	.130	-.245	-.070
	Exhaust/inactive	.224	**.727**	.066	.229	.040
3.Related personal variables	Drink alcohol or beer: quantity	-.001	.064	**.901**	-.124	.027
	Tobacco smoking	.053	.141	**.795**	.040	-.105
	Epilepsy	-.256	-.092	**.658**	.011	.004
4.Witnessed snoring and apnea	Periodically stop breathing	.158	.155	.048	**.836**	.064
	Chest discomfort	-.212	.141	-.228	**.756**	-.064
	Loudness of your snoring	.470	-.005	.138	**.668**	.071
5.Underlying diseases and surreal dream	Stroke	-.166	-.041	-.246	-.026	**.793**
	Lung diseases	-.315	-.064	-.369	-.058	**-.620**
	Surreal dream	.043	.480	-.044	.467	**.504**

3.4 Severe OSA group

Data of the 540 severe OSA patients to find the Pearson correlation with other variables significant at 0.05 and 0.01 had remaining 24 from 113 variables. KMO = 0.823 was assumed that it was reasonable to use factor analysis. Bartlett's test of sphericity showed the significance = 0.000 and the result of 24 variables revealed correlation matrix was the Identity matrix. The Factor extraction from PCA had values > 0.5 with lowest value in the variable "Thyromental distance" which was still not too low that it could be clearly grouped in some factors (Tables 10).

Each of 6 factors had Eigen value > 1. The total variances explained 6 factors was 80.239% by extraction method: PCA (Table 11). The result of factor rotation (Varimax rotation) revealed 6 factors (24 variables) with new naming by sleep specialists that corresponded with variables (Table 12) and the scree plot showed new naming factors (Fig. 5).

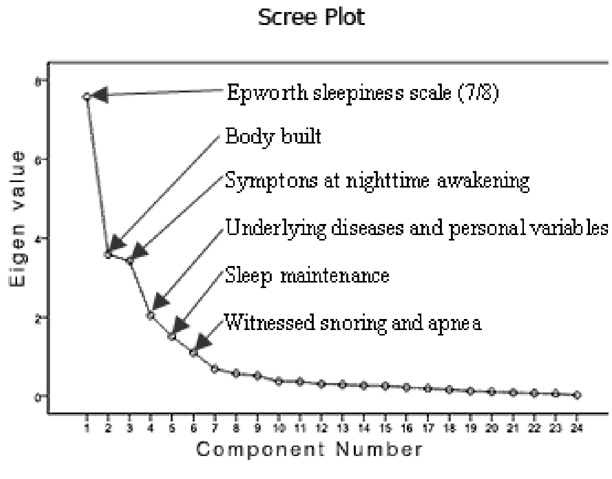


Figure 5: The 6 new naming factors in severe OSA group on the slope of scree line that had Eigen value >1

Table 10: Communalities (severe OSA group)

Variables	Initial	Extraction
Body mass index	1.000	.793
Waist circumference	1.000	.915
Hip circumference	1.000	.929
Sitting inactive in a public place (sleepiness)	1.000	.919
Lying down to rest in the afternoon (sleepiness)	1.000	.817
Sitting quietly after lunch with no alcohol (sleepiness)	1.000	.858
Sitting and reading (sleepiness)	1.000	.783
In a car for an hour without break (sleepiness)	1.000	.850
Snoring	1.000	.870
Loudness of your snoring	1.000	.849
Periodically stop breathing	1.000	.837
Difficult breathing like something obstruct in the throat	1.000	.898
Leg or arm jerking	1.000	.800
Age	1.000	.869
Diabetes mellitus	1.000	.603
Thyromental distance	1.000	.512
Routine Medicine use	1.000	.788
Get up earlier than your expectation	1.000	.769
Urinate very often	1.000	.810
Watching TV (sleepiness)	1.000	.748
Feeling like choking	1.000	.834
Hypertension	1.000	.760
You wake up during late night: times / night	1.000	.695
Sitting and talking to someone (sleepiness)	1.000	.751

Table 11: The initial Eigen values, extraction and rotation sums of squared loadings in severe OSA group

Components (factor)	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.572	31.552	31.552	7.572	31.552	31.552	5.567	23.194	**23.194**
2	3.593	14.970	46.522	3.593	14.970	46.522	3.483	14.511	**37.705**
3	3.428	14.284	60.806	3.428	14.284	60.806	2.931	12.214	**49.919**
4	2.047	8.530	69.336	2.047	8.530	69.336	2.624	10.933	**60.852**
5	1.515	6.310	75.646	1.515	6.310	75.646	2.348	9.782	**70.634**
6	1.102	4.593	80.239	1.102	4.593	80.239	2.305	9.605	**80.239**
⋮	⋮	⋮	⋮						
24	.029	.122	100.000						

Table 12: The result of factor loading of 6 factors in severe OSA group

New naming by sleep specialists	Variables	Components (factors)					
		1	2	3	4	5	6
1.Epworth sleepiness scale(7/8)	Sitting inactive in a public place	**.923**	.081	.168	-.147	.000	.100
	Sitting quietly after lunch with no alcohol	**.899**	-.030	.096	-.122	.099	.121
	In a car for an hour without break	**.853**	-.068	.172	-.240	.021	.174
	Lying down to rest in the afternoon	**.836**	-.035	-.023	-.094	.297	.138
	Sitting and reading	**.836**	-.199	.076	-.133	-.083	.119
	Watching TV	**.804**	-.023	.239	-.019	.162	.130
	Sitting and talking to someone	**.784**	.043	.362	-.019	-.040	-.039
2.Body built	Waist circumference	.000	**.954**	-.044	.006	.004	-.059
	Hip circumference	.044	**.947**	.110	.103	.060	-.064
	Body mass index	-.139	**.863**	.054	.140	-.080	-.011
	Thyromental distance	-.058	**.683**	-.012	-.138	-.114	-.102
3.Symptoms at night time awakening	Difficult breathing like something obstruct in the throat	.212	.099	**.909**	-.082	.083	.062
	Feeling like choking	.191	.075	**.882**	-.035	-.048	.105
	Leg or arm jerking	.294	-.048	**.799**	-.119	.243	.000
4.Underlying diseases and personal variables	Hypertension	-.216	.113	-.045	**.831**	.051	-.064
	Routine Medicine use	-.169	-.146	-.163	**.817**	.114	-.178
	Age	-.151	-.415	-.136	**.653**	.294	-.377
	Diabetes mellitus	-.096	.320	.013	**.645**	.165	-.220
5.Sleep maintenance	Urinate very often	.102	.030	-.018	.266	**.849**	.078
	Get up earlier than your expectation	.145	-.176	.004	.030	**.846**	.000
	You wake up during late night (times/night)	.028	.017	.360	.109	**.727**	-.151
6.Witnessed snoring and apnea	Snoring	.100	-.121	.043	-.153	.059	**.904**
	Loudness of your snoring	.315	-.048	-.028	-.264	-.147	**.809**
	Periodically stop breathing	.276	-.185	.481	-.230	.055	**.663**

It is noteworthy that most variables were positively and significantly related to all RDI,whereas some variables were negatively and significantly related to some RDI eg. Excessive daytime sleepiness: As a passenger in a car for an hour without break (factor loading = -0.934); Sitting inactive in a public place (factor loading = -0.929) in normal group which imply that normal group has tendency to be alert even with an inactive activities surrounded by sleepiness induced atmosphere. Likewise, work-hour/day (factor loading = -0.715) and average sleeping hours in each night (factor loading = -0.917) in mild OSA group may explain the high RDI with less sleeping hour/night and thus less work-hour/day. Furthermore, Lung diseases should be common in mild OSA group (factor loading =0.695) and uncommon in moderate OSA group (factor loading = -0.620). However, by taking all variables of each factor in non OSA and each severity level of OSA patients into consideration, some observations which might be of clinical value, revealed significant variables found correlated with only non OSA and only each severity level of OSA including witnessed snoring in all severity of OSA and witnessed apnea in moderate and severe OSA group (Table 13).

Table 13: Significantly related variables in only each severity level of OSA and only normals by factor analysis.

Variables	Non OSA	OSA Severity		
		Mild	Moderate	Severe
1. Underlying diseases: Kidney, GI diseases	✓	-	-	-
Diabetes mellitus	-	-	-	✓
Lung diseases	-	✓	-	-
Epilepsy	-	-	✓	-
Stroke	-	-	✓	-
2. Body Built – obesity, thyromental distance	-	-	-	✓
3. Alcohol & tobacco smoking	-	-	✓	-
4. Decreased work -hour per day	-	✓	-	-
5. Decreased mental and physical performance	-	-	✓	-
6. Epworth sleepiness scale (Daytime sleepiness)	2/8	3/8	6/8	7/8
Maximal Expected Score	≤ 6	≤ 9	≤ 18	≤ 21
7. Decreased sleeping hour/night	-	✓	-	-
8. Most comfortable sleep posture	✓	-	-	-
9. Insomnia: Get up earlier than expected	-	-	-	✓
Wake up during late night	-	-	-	✓
10. Surreal dream	-	-	✓	-
11. Causes of nighttime awakening				
· Witnessed, periodical grasping	✓	-	-	-
· Sweat profoundly	✓	-	-	-
· Physical pain: stomachache, headache	-	✓	-	-
· Environment: hot, cold, noise, light	-	✓	-	-
· Difficult breathing like something obstruct in the throat	-	-	-	✓
· Witnessed leg & arm jerking	-	-	-	✓
· Frequent urination	-	-	-	✓
12. Witnessed snoring	-	✓	✓	✓
13. Witnessed apnea	-	-	✓	✓

4. Conclusion

Sleep questionnaires and anthropometric data from 1,042 suspected OSA patients were grouped by factor analysis and found to be significantly related to normal and each level of severity of OSA as followed: Firstly, in normal (non-OSA) group, there were three factors consisted of Epworth sleepiness scale (2/8) and related variables; symptoms during sleep; and underlying diseases and sleep posture. Secondly, in mild OSA group, there were five factors consisted of causes of nighttime awakening; Epworth sleepiness scale (3/8); lung diseases and sleep-wake pattern; risk variables and work; and choking and witnessed snoring. Thirdly, in moderate OSA group, there were five factors consisted of Epworth sleepiness scale (6/8); decreased mental and physical performance; related personal variables; witnessed snoring and apnea; and underlying diseases & surreal dream. Fourthly, in severe OSA group, there were six factors consisted of Epworth sleepiness scale (7/8); body built; symptoms at nighttime awakening; underlying diseases and personal variables; sleep maintenance; and witnessed snoring and apnea. Finally, particular variables of symptoms and anthropometric data related to only normals and only each severity level of OSA might be of clinical value in consideration of appropriate queuing for PSGs.

Acknowledgements

The authors wish to thank Faculty of Medicine Siriraj Hospital for the valuable co-operation and all the patients whose data were studied.

References

[1] Friedman, M. Sleep apnea and snoring surgical and non-surgical therapy. 1st ed.Elsevier Inc. 2008.

[2] Sleep Society of Thailand, The 4thPolysomnography (PSG) training: basic level, March 4-8, Bangkok: Siriraj Hospital, 2013.

[3] Mulaik, S A. Foundation of factor analysis. 2nd ed. New York: McGraw-Hill Book Company, 2010.

[4] Vanichbuncha, K . Advance statistics analysis by SPSS. 9thed.Bangkok: Chulalongkorn University, 2009.

[5] Vanichbuncha, K. Multivariate analysis. 2nd ed. Bangkok: Chulalongkorn University, 2009.

[6] Kaiyawan, Y. Multivariate statistical analysis for research. 1st ed. Bangkok: Chulalongkorn University, 2013.

[7]. Pongpullponsak, A. Final report of project on KPI, connection, output, outcome and strategy for development of economic structure. Bangkok: King Mongkut’s University of Technology Thonburi, 2008.

Bullying behavior in schools, Mueang district, Chiang Mai province

W. Sriwattanapongse*[1] and P. Taninpong[2]

Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai ,Thailand

Abstract

Bullying behavior in schools is well known among students, parents, teachers and educational personnel. The objectives of this research were to estimate the prevalence of bullying in schools and to analyze the relationship between risk factors and bullying in Chiang Mai, Thailand. Risk factors analyzed included gender, school, age group, school type, parental discipline, parental marital status, and domestic abuse in the home, both verbal and physical. Pearson's chi-squared test and odds ratio was used to assess the associations between the outcome and the various categorical determinants. The prevalence of physical bullying was to found to be 15.51%, with a victim rate of 22.99%. We found that gender, school, age group, school type, parental discipline, and domestic abuse were associated with student bullying. In addition, exposure to domestic physical abuse in the home was clearly the most strongly associated determinant, much more strongly linked to bullying than not having been exposed to domestic physical abuse (OR 2.10, 95% CI 1.11-3.97)

Keywords: Bullying behavior, prevalence of bullies, risk factor

*Corresponding Author
E-mail Address: wattanavadee.s@cmu.ac.th

1. Introduction

Bullying behavior in schools is well-known among students, parents, teachers and educational personnel. School bullying is a serious problem which affects students' quality of life, inflicting psychological, emotional and physical damage, and occurs throughout the world. Bullying might be classified into a variety of ways including physical assaults and psychological or emotional or verbal harassment [1], and Reference [2] explained that physical bullying is action oriented and intended to intimidate or physically hurt the victim through pinching, pushing, kicking, and hitting, while verbal bullying is using words to humiliate or hurt someone's feelings through teasing, name-calling, insulting, or threatening behavior. The major reasons that children bully others are to enjoy exercising power and status over their victims, boredom, jealousy, attention seeking, showing off, anger, revenge, and self-protection [3]. The targets or victims of school bullying are at risk of a variety of negative outcomes. There are many causes of bullying, such as domestic violence [4], preferring cartoon violence [5], older students [6], male gender [7], and religion [8]. Studies have indicated that 56% of students in South Africa bully [9], 38% in Netherlands [10], 34% in Australia [11], 30% in Nigeria [12], 22% in Italy [13], 21% in Canada [14], 20% in Malaysia [15] and 16% in Portugal [16]. In Thailand, the prevalence of students' bullying is 42%, with a regional breakdown as follows: 44% in the North, 40% in Central, 39% in the East, 35% in the Northeast, and 27% in the South [17]. Reference [18] studied the relationship between individual characteristics, the family context and cyber-bullying. The results showed a considerable degree of cyber bullying among Thai youth. In addition, more than half had ever seen or heard of cyber-bullying occurring to their friends.

The objective of this research was to estimate the prevalence of student bullying and to analyze the relationship between the risk factors and student bullying in Chiang Mai, Thailand.

2. Research Methodology

This study used a cross-sectional study design involving interviews and surveys of school students in a sample selected from the target population studied. The target population comprised all students at Chiang Mai schools during January1, 2014 and February 15, 2014. The participants were selected by using a stratified two-stage cluster sample Design for primary school and a two-stage stratified sampling method for secondary school as showed students sample in Table1.

Table 1 Cross-tabulation between school and school type

	School type		
School	Public	Private	
Primary	112	261	373
Secondary	249	113	362
	361	374	735

Table 1 shows a cross-tabulation between school and school type. Of the 735 students in this study, 364 students (49.5%) come from publics' school.

The interest variables of the study comprised nine determinants and a binary outcome. The determinants in the study provided information on (a) characteristics of students (gender, age group, punishment and school type), (b) family environment (father's occupation, mother's occupation, marital status of parent and parental violence).The physical bullying was binary outcome.

In this study, Descriptive analysis was conducted for measuring the prevalence of bullying. Pearson's chi-squared test was used to assess the associations between the outcome and the various categorical determinants.

Pearson's Chi-square

Pearson's chi-squared statistics for independence were used to assess the association between the determinant variables and the outcome of this study .In this study the chi-squared statistics takes the form

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{1}$$

Where O_{ij} is the observe count in category i of the determinant and category j of the outcome, and E_{ij} is the corresponding expected count, defined as before by dividing the product of the marginal totals by the overall total sample size, that is

$$E_{ij} = \frac{[\sum_{j=1}^{c} O_{ij}][\sum_{i=1}^{r} O_{ij}]}{n} \tag{2}$$

When the null hypothesis of independent is true, the right-hand side of equation has a chi-square distribution with $(r-1)(c-1)$ degree of freedom

Odds

The odds of an event or outcome are defined as the ratio of the probability of the event to the probability of corresponding non-event. In this study, we are using the term "probability "to mean the proportion of outcomes in the population of interest. Thus if an event has probability π and the corresponding non-event has probability $(1-\pi)$ the odds of the event is $\pi(1-\pi)$.

The estimated odds of an outcome are obtained by replacing the population proportion by the sample proportion *p*.

Graphs of estimated odds and 95% confidence intervals of odds can be used to represent the proportions of physical bullying in the two groups. The graph of odds includes a 95% confidence interval. The confidence interval is graphed as a horizontal line containing a dot denoting the estimated odds.

If *p* is the proportion of outcomes in a sample of size n, the estimated of odds is

$$odds = \frac{p}{1-p} \tag{3}$$

An asymptotically valid (for large n) formula for the standard error of the log-odds, defined as $ln(odds) = ln(p) - ln(1-p)$, is

$$SE(ln(odds)) = \sqrt{\frac{1}{np} + \frac{1}{n-np}} \tag{4}$$

A 95% confidence interval (95%CI) for the population odds is thus given by

$$odds \times exp(-1.96 \times SE), odds \times exp(1.96 \times SE)$$

The estimated odds of an outcome are thus the ratio of the number of outcomes of interest to the number of non-outcome. The odd is always greater than the corresponding proportion or probability. But unlike a proportion, which is restricted to being between 0 and 1, the odds can be greater than 1[19].

3. Research Results and Discussion

Descriptive method used to describe percentage of sample characteristics in Table 2 and Table 3.

Table 2 shows that 69.80% of participants were girls, approximately half from primary school (50.75%) and half from secondary school (49.25%).The students were divided into age groups as follows: 9-11 years (37.01%),age group 12-14 years(48.16%) and 15-17 years(14.83%).The type of school attended was either public (49.12%) or private (50.88%). The most common form of parental discipline was described as "punish occasionally" (59.46%).Most parents were married (73.74%) and only 11.84% were divorced. Although 14.29% had witnessed parental verbal abuse once or twice a year, most had not (73.74%). Most had not seen parental physical abuse (87.76), but a few had witnessed physical abuse at home (4.63%).

Most students reported that they had never been involved in physical bullying at school (93.47%) both as a bully and victim. However, 8.98% admitted that they had physically bullied other students and16.46% had been the victims of student bullying. While 6.53% reported that they had been both a victims of bullying and bullied other students, 18.91% reported that they had been one or the other at some time.

Boys were more likely than girls to be merely the victim of bullying (22.52% vs 13.84%, respectively), or report ever being bullied, 33.33 vs 18.52 respectively. Boy were more likely to report being bully (13.06% vs 7.21%, respectively). While those who admitted to being both a bully and a victim of bullying was also more common for boys (10.81% vs4.68%, respectively). Overall 23.87% of boys and 11.89% of girls had physically bullied other students. Therefore, there appears to be a large difference in the percentage of boys versus girls involved in bullying, both as the victim and as the perpetrators.There were approximately 15.51% of students reported having bullied others.

Table 2 Sample characteristics among students in Chiang Mai Thailand, 2013

Demographics	All students	Boys	Girls
	N=735(100 %)	N=222(30.20%)	N=513(69.80 %)
School			
Primary School	373(50.75%)	120(54.05%)	253(49.32%)
Secondary School	362(49.25%)	102(45.95%)	260(50.68%)
Age group(years)			
ageGrp1:9-11	272(37.01%)	89(40.09%)	183(35.67%)
ageGrp2:12-14	354(48.16%)	93(41.89%)	261(50.88%)
ageGrp3:15-17	109(14.83%)	40(18.02%)	69(13.45%)
School type			
Public	361(49.12%)	107(48.20%)	254(49.51%)
Private	374(50.88%)	115(51.8%)	259(50.49%)
Parental discipline			
I never do anything wrong	95(12.93%)	28(12.61%)	67(13.06%)
I don’t get punished	169(22.99%)	50(22.52%)	119(23.20%)
I get punished occasionally	437(59.46%)	127(57.21%)	310(60.43%)
I get punished frequently	31(4.22%)	15(6.76%)	16(3.12%)
missing	3(0.41%)	2(0.90%)	1(0.19%)
Marital status of parents			
Married	542(73.74%)	162(72.97%)	380(74.07%)
Separated	74(10.07%)	31(13.96%)	43(8.38%)
Divorced	87(11.84%)	18(8.11%)	69(13.45%)
Father deceased	27(3.67%)	8(3.60%)	19(3.70%)
Mother deceased	4(0.54%)	3(1.35%)	1(0.19%)
Both parent deceased	1(0.14%)	0(0.00%)	1(0.19%)
parents verbally abuse			
Never	428(58.23%)	134(60.36%)	294(57.31%)
Once or twice a week	53(7.21%)	11(4.95%)	42(8.19%)
Once or twice a month	60(8.16%)	15(6.76%)	45(8.77%)
Once or twice a year	105(14.29%)	33(14.86%)	72(14.04%)
Other	84(11.43%)	27(12.16%)	57(11.11%)
missing	5(0.68%)	2(0.90%)	3(0.58%)
parents physically abuse			
Never	645(87.76%)	194(87.39%)	451(87.91%)
Once or twice a week	8(1.09%)	3(1.35%)	5(0.97%)
Once or twice a month	2(0.27%)	0(0.00%)	2(0.39%)
Once or twice a year	34(4.63%)	11(4.95%)	23(4.48%)
Other	34(4.63%)	8(3.60%)	26(5.07%)
missing	12(1.63%)	6(2.70%)	6(1.17%)
Nature of bullying			
Uninvolved	669(91.02%)	193(86.94%)	476(92.79%)
Bully	66(8.98%)	29(13.06%)	37(7.21%)
Uninvolved	614(83.54%)	172(77.48%)	442(86.16%)
Victim	121(16.46%)	50(22.52%)	71(13.84%)
Uninvolved	596(93.47%)	167(89.19%)	429(95.32%)
Bully and Victim	48(6.53%)	24(10.81%)	24(4.68%)

Table 3 Physical bullying others during the last semester (May to 31 October 2013)

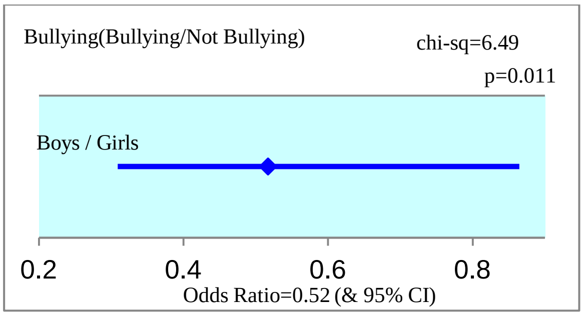
Physical bully	All student 66 (100%)	Boys 29(43.94%)	Girls 37(56.06 %)
Bullied			
1.Slapped	3(4.55%)	1(3.45%)	2(5.41%)
2.Hit	42(63.00%)	13(44.83%)	29(78.38%)
3.Kicked	2(3.03%)	2(6.9%)	0(0.00%)
4.Punched	3(4.55%)	3(10.34%)	0(0.00%)
5.Weapons	0(0.00%)	0(0.00%)	0(0.00%)
6.Other	6(9.09%)	3(10.34%)	3(8.11%)
7.Slapped,Hit,Kicked	3(4.55%)	3(10.34%)	0(0.00%)
8.Hit,Punched	1(1.52%)	0(0.00%)	1(2.70%)
9.Slapped,Hit	3(4.55%)	2(6.9%)	1(2.70%)
10.Hit,Kicked,Punched	2(3.03%)	2(6.9%)	0(0.00%)
11.Slapped,Hit,Kicked	1(1.52%)	0(0.00%)	1(2.70%)
Time of Day			
1.Before school	1(1.52%)	1(3.45%)	0(0.00%)
2.During lessons	14(21.21%)	7(24.14%)	7(18.92%)
3.At lunch time	35(53.03%)	14(48.28%)	21(56.76%)
4.After school	13(19.70%)	5(17.24%)	8(21.62%)
5.During holiday times	3(4.55%)	2(6.90%)	1(2.70%)
Location			
1.Classroom	42(63.64%)	18(62.07%)	24(64.86%)
2.Activities	0(0.00%)	0(0.00%)	0(0.00%)
3.Sport Stadium	9(13.64%)	6(20.69%)	3(8.11%)
4.Toilets	4(6.06%)	1(3.45%)	3(8.11%)
5.Canteen	2(3.03%)	0(0.00%)	2(5.41%)
6.Behind building	0(0.00%)	0(0.00%)	0(0.00%)
7.Outside school	4(6.06%)	2(6.90%)	2(5.41%)
8.Other	5(7.58%)	2(6.90%)	3(8.11%)
Reason			
1.Physically weaker	1(1.52%)	1(3.45%)	0(0.00%)
2 To create an enemy with other student(s)	6(9.09%)	3(10.34%)	3(8.11%)
3.Persecuted	4(6.06%)	2(6.90%)	2(5.41%)
4.Pay back or revenge	24(36.36%)	13(44.83%)	11(29.73%)
5.Provoked	7(10.61%)	3(10.34%)	4(10.81%)
6.Physical appearance or personality not liked	4(6.06%)	0(0.00%)	4(10.81%)
7.Boyfriend or girlfriend	1(1.52%)	1(3.45%)	0(0.00%)
8.Want to be accepted by peers	9(13.64%)	2(6.90%)	7(18.92%)
9.Other	10(15.15%)	4(13.79%)	6(16.22%)
Frequency			
1.Never	52(78.79%)	22(75.86%)	30(81.08%)
2.Once of twice	1(1.52%)	1(3.45%)	0(0.00%)
3.Once or twice a month	0(0.00%)	0(0.00%)	0(0.00%)
4.Once or twice a year	6(9.09%)	1(3.45%)	5(13.51%)
5.Other	4(6.06%)	3(10.34%)	1(2.70%)
missing	3(4.55%)	2(6.90%)	1(2.70%)

Table 4 Associations between bullying and study determinants

Demographics		Bullied behavior			
		Not Bullied	Bullied	Chi-square	p-value
		669	66		
Gender				5.49	0.011*
	Boys	193(86.94%)	29(13.06%)		
	Girls	476(92.79%)	37(7.21%)		
School				20.41	0.000*
	Primary	322(86.33%)	51(13.67%)		
	Secondary	347(95.86%)	15(4.14%)		
Age group				24.01	0.000*
	ageGrp1:9-11	230(84.6%)	42(15.4%)		
	ageGrp2:12-14	332(93.8%)	22(6.2%)		
	ageGrp3:15-17	107(98.2%)	2(1.8%)		
School type				5.91	0.015*
Public		338(45.99%)	23(3.13%)		
Private		331(45.03%)	43(5.85%)		
Parental discipline				10.09	0.006*
I never do anything wrong		94(98.95%)	1(1.05%)		
I don't get punished		156(98.95%)	13(7.69%)		
I get punished.		419(88.96%)	52(11.04%)		
Marital status of parents				6.88	0.076
Married		494(91.14%)	48(8.86%)		
Separated		63(85.14%)	11(14.86%)		
Divorced		84(96.55%)	3(3.45%)		
Father or Mother deceased		28(87.5%)	4(12.5%)		
parents verbally abusive				2.64	0.104
Not witnessed		396(92.52%)	32(7.48%)		
Witnessed		273(88.93%)	34(11.07%)		
parents physically abusive				5.43	0.020*
Not witnessed		593(91.94%)	52(8.06%)		
Witnessed		76(84.44%)	14(15.56%)		

* p-value < 0.05

Figure 1-6:95% confidence intervals of odds ratio plot by six different risk factors.



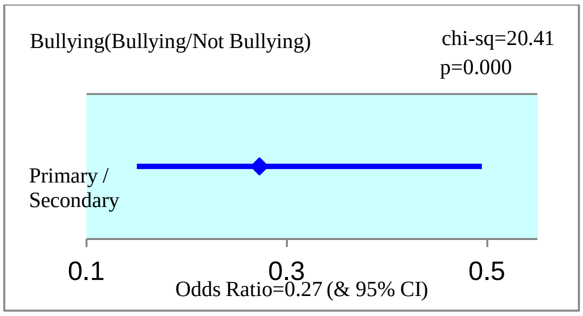


Figure 1: Odds ratio of bullying by gender

Figure 2: Odds ratio of bullying by School

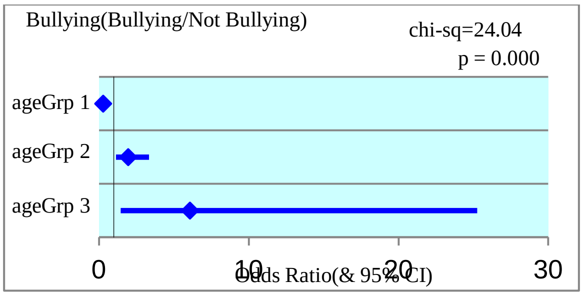


Figure 3: Odds ratio of bullying by age Group

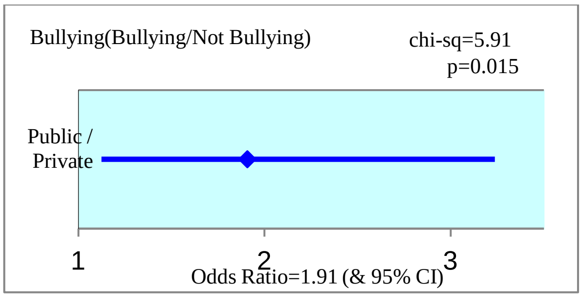


Figure 4: Odds ratio of bullying by type

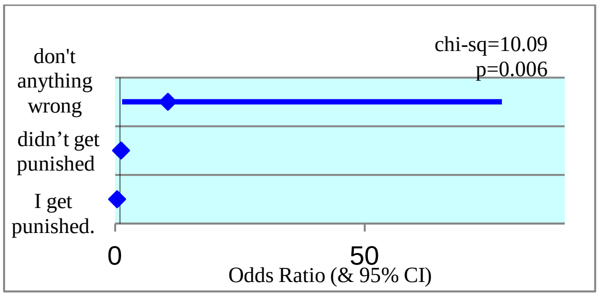


Figure 5: Odds ratio of bullying by parental discipline

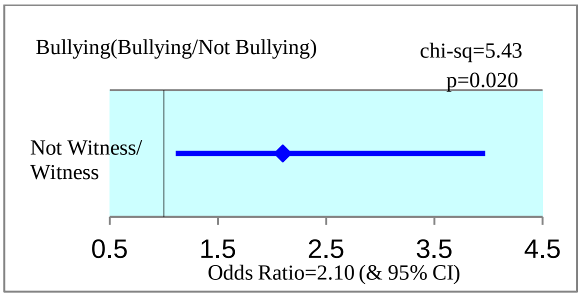


Figure 6: Odds ratio of bullying by domestic violence

Table 3 shows that “hitting” is the most common form of physical bullying (63%) and that the classroom is the most common location (53%) and that lunch time is the most common time for bullying (63.64%).The main reasons given for bullying include “pay back or revenge” (36.36%),and “wanting to be accepted by their peers”(13.64%). In the last semester, the girls were more likely to report bullying “once or twice a year” than the boys (13.51% vs 3.45%, respectively).

Table4 shows the associations between bullying and the eight study determinants. Since all of the variables are categorical, Pearson’s chi-squared test is used to assess the statistical significance of the association in each case. Results show gender, school, age group, school type, parental discipline, and domestic abuse were strongly associated with bullying.

Figure 1 shows the odds ratio plot of bullying and student gender. Boys are more often involved in the bullying than girls (OR 0.52, 95% CI 0.31-0.86).

Figure 2 shows the odds ratio plot of bullying and school, Primary school students more often reported having than secondary school students (OR 0.27, 95% CI 0.15-0.49).

Figure 3 shows the odds ratio plot of bullying and student age group. Younger students (15-17 years) were more likely to report bullying than older age group.(9-11 years and 12-14 years), (OR 0.30, 95% CI 0.18-0.51; OR 1.97, 95% CI 1.16-3.36; OR 6.09, 95% CI 1.47-25.27).

Figure 4 shows the odds ratio plot of bullying for the student’s school type. Private school students reported bullying behavior more than public schools students. (OR 1.91, 95% CI 1.13-3.27).

Figure 5 shows the odds ratio plot of bullying for type of parental discipline. Students reporting that they “got punished” were more likely to report bullying than both “don’t get punished” and “never did anything wrong” categories. (OR 10.63, 95% CI 1.46-77.5; OR 1.24, 95% CI 0.66-1.23; OR .45, 95% CI 0.25-0.83).

Figure 6 shows the odds ratio plot of bullying for student’s exposure to domestic violence. Those who had witness domestic violence at home were more likely to report bullying at school (OR 2.1, 95% CI 1.11-3.39).

Exposure to domestic physical abuse at home was clearly the strongest associated determinant to reporting bullying at school.

Exposure to domestic violence is notably to negative student behaviors at school, such as bullying other students. This is in accordance with the studies of Rossman et al [20], Herrera et al [21], and Reference [22] who all reported that parental modelling of aggression and violence promotes the development of a child’s negative behavior. The child learns from the parents to use violence to achieve social dominance in their own social setting. Reference [23] found that the problem of student violence in Thailand is affected by social conditions based on the presence of conflict in the family leading to verbal abuse. Reference [24] studied the meanings and configurations of school violence and the effect on children in order to understand the whole social structure of school-based abuse and youth violence via the experiences and ideas of the student or others involved. In addition, this research found that the nature of and motivations behind school violence were related to conduct of teachers, friends, and people working at the school, who were actors that harmed students physically, mentally and sexually in addition to exacting latent violence by violating their rights. The configurations of school violence were diverse and depended on individual factors and local surroundings. Violence was related to the relationship structure between teachers and students, students and students, teachers and teachers, teachers and administrators, students and school staff. Violence in the case of abused youths was associated with conflict over shared social spaces among actors with different statuses, including over the legitimacy of ecological boundaries, school rules and individual status.

4. Conclusion

The prevalence of physical bullying was 15.15 % (including 8.98% bullying only and 6.53% both bully and victim).Also, 22.99% were victims of bullying (16.46% in the victim only and 6.53% both). Boys were more likely to be the victim of physical bullying than girls (22.52% vs 13.84%) as well as more likely to be bullies (23.87%vs 11.89%). Approximately 44.83% of boys who bullied admitted to hitting others, while girls bullied hit others in 78.38% of the bullying episodes. As follow-up to Reference [25], this study assessed bullying and its associated factors in school-going adolescents in Thailand. The study found that the predominant forms of bullying were hitting, kicking, pushing, shoving and locking indoors among boys and, among girls, lewd sexual remarks, comments and gestures were more common. Risk factors among the boys for being victims of bullying included younger age (adjusted odds ratio) having been in a physical fight, physically inactive, truancy, and psychosocial distress. Among girls bullied, risk factors included having been in a physical fight, lack of parental bonding, and psychosocial distress. These results may help inform school health programs on the prevalence and risk factors associated with bullying at school.

Most of the bullying took place in the classroom during lunchtime with the main reason being revenge. Reference [26] examined the relationship between perceptions of bullying behaviors and actual bullying behaviors among groups of Chiang Mai students. The result of this study showed that the students *often* perceived bullying behavior at school. 48.1% of the bullying behaviors were conducted by a group of 2-3 students. When perceiving bullying acts, the students preferred sharing their experience with friends. 62.7% of the bully victims were seldom bullied. Mockery was identified as the most performed bullying act. 56.7% of the students seldom bullied other students. Locations of bullying occurred in classrooms with no presence of teachers, hallways or stairways, fields, toilets and canteens.

In this study, we also found that risk factor for bullying at school in Chiang Mai, Thailand included gender, school, age group, school type, parental discipline, and exposure to domestic abuse.

The study also indicates a relationship between the perception of bullying behaviors and the actual bullying behaviors. The research findings suggest campaigns against bullying behaviors should be carried out to raise awareness among teachers, parents and students to create a social norm which considers bullying behaviors as non-acceptable and encourages effective solutions to behavioral problems. Bullying behavior prevention courses should be implemented at schools to follow up the problem in the long term.

Acknowledgements

We would like to thank the Faculty of Science, Chiang Mai University for funding this research. We extend appreciation to the students who participated in this research, and would like to thank the school principals and teachers for their facilitation of this study.

References

[1] Beale AV, Bully busters': Using drama to empower students to take a stand against bullying behavior. Professional School Counseling. 2001; 4: 300-305.

[2] Woods S, and Wolke D, Direct and relational bullying among primary school children and academic achievement. Journal of School Psychology. 2004; 42: 135-155.

[3] Besag VE, Bullying among girls: Friends or foes? School Psychology International. 2006; 27(5): 535–551.

[4] Baldry AC, Bullying in schools and exposure to domestic violence. Child Abuse & Neglect. 2003; 27(7): 713–732.

[5] Theppipidh P, Relationships between preferences of comics, television programs, computer games, and aggressive behavior of prathom suksa six students, Bangkok Metropolis [Thesis]. Chulalongkorn University; 1990.

[6] Wolke D, Woods S, Stanford K, Schulz H, Bullying and victimization of primary school children in England and Germany: Prevalence and school factors. British Journal of Psychology. 2001; 92: 673–696.

[7] Mouttapa M, Valente T, Gallaher P, Rohrbach LA, Unger JB, Social network predictors of bullying and victimization. Adolescence. 2004; 39: 315-336.

[8] Ian, J., Wendy, MC., William, FB.,William, P. Associations between overweight and obesity with bullying behaviors in school-aged children. [Internet]. 2004. [Updated August 7, 2007] Available from: http://pediatrics.aappublications.org/cgi/content/full/113/5/1187.

[9] Greeff P, The nature and prevalence of bullying during the intermediate school phase [Thesis] University of the Free State; 2004.

[10] Veenstra R., Lindenberg S, Oldehinkel AJ, De Winter AF, Verhulst FC, Ormel J, Bullying and victimization in elementary schools: A comparison of bullies, victims, bully/victims, and uninvolved preadolescents. Developmental Psychology.2005; 41: 672-682.

[11] Ahmed E, Braithwaite V, "What, Me Ashamed?" Shame management and school bullying, Journal

of Research in Crime and Delinquency. 2004; 41(3): 269-294.

[12] Egbochuku EO, Bullying in Nigerian schools: Prevalence study and implications for counselling, J. Soc. Sci., 2007; 14(1): 65-71.

[13] Gini G, Associations between bullying behavior, psychosomatic complaints, emotional and behavioral problems. Journal of Pediatric and Child Health. 2008; 44: 492–497.

[14] Hawkins DL, Pepler DJ, Craig WM, Naturalistic observations of peer interventions in bullying. Social Development. 2001; 10 (4): 512-527.

[15] Wan Salwina WI, Susan MK, Nik Ruzyanei NJ, Tuti Iryani MD, Syamsul S, Aniza A, Zasmani S, School bullying amongst standard students attending primary national schools in the federal territory of Kuala Lumpur: The prevalence and associated socio demographic factors. Malaysian Journal of Psychiatry. 2009; 18(1):5-12.

[16] Pereira B, Mendonca D, Neto C, Valente L,Smith, P, 2004. Bullying in Portuguese schools. School Psychology International, 25: 241-254.

[17] Tapanya S, A survey of bullying problem of students in Thailand. Chiang Mai, Faculty of Medicine, Chiang Mai University, 2006.

[18] Songsiri N, Musikaphan W, Cyber-bullying among Secondary and Vocational Students in Bangkok. Journal of Population and Social Studies, 2011; 19 (2). 235-242.

[19] McNeil D, Modern statistics: A graphical introduction. Sydney, Australia: Macquarie University, 1998.

[20] Rossman BR, Hughes HM, Rosenburg MS, Children and inter parental violence: The impact of exposure. Philadelphia, PA: Brunner/Mazel; 2000.

[21] Herrera VM, McCloskey LA, Gender differences in the risk for delinquency among youth exposed to family violence. Child Abuse & Neglect. 2001; 25: 1037–1051.

[22] Baldry AC, *Bullying in schools and exposure to domestic violence.* Child Abuse & Neglect. 2003; 27(7): 713–732.

[23] Taosound N, The relationship between child Rearing Practices and Aggression of Secondary Students. Chiang Mai Province [Thesis] Chiang Mai: Chiang Mai University; 2004.

[24] Junprasert T, Soadmanee O, Bhookong S, Mukpradit M, Social Structures Related to Child and Youth Violence at School: A Case Study of School in Central Thailand. Journal of Behavioral Science 2011; 17 (1): 93-107.

[25] Pengpid S, Peltzer K, Bullying and Its Associated Factors among School-Aged Adolescents in Thailand [Internat].2013
[update 2013 Jan; cited 2014 Feb18]
Available from:
http://dx.doi.org/10.1155/2013/254083

[26] Junlapiya P, The relationship between the perception of bullying behaviors and the actual bullying behaviors among groups of students in Chiangmai.[Thesis].Bangkok:Mahidol University;2007.

A review on parametric and nonparametric empirical likelihood method based on interval-censored data

Nor Azelah Zainduin [2], F. A. M. Elfaki [1*] and M. Yeakub Ali [2]

[1] *Department of Science in Engineering, Faculty of Engineering, International Islamic University Malaysia, 50728, Kuala Lumpur, Malaysia*

[2] *Department of Manufacturing and Materials Engineering, Faculty of Engineering, International Islamic University Malaysia, 50728, Kuala Lumpur, Malaysia*

Abstract

This paper review on the estimation of distribution functions for both parametric and nonparametric empirical likelihood method. The focus of this review is on empirical likelihood regression method behavior on censored data and how those methods affect the result of the simulation studies. Moreover, we look into the case which is involve the partly interval censored data.

Keywords: Censored data; empirical likelihood; parametric estimation; nonparametric estimation; approximation of variance

*Corresponding Author
E-mail Address: faizelfaki@yahoo.com

Forecasting model of malaria incidence with climatic variables: A case study in Ratchaburi, Thailand

Ngamphol Soonthornworasiri*, Supalarp Puangsa-art, Patiwat Sa-angchai,
Chotipa Kulrat and Jaranit Kaewkungwal
Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University,
Bangkok, 10200, Thailand, ngamphol.soo@mahidol.ac.th

Abstract

Malaria is an infectious disease in the tropics and is a major global health problem with a significantly increasing due to climate change. Malaria still remains a public health problem in Thailand, especially, in some provinces along the Thai-border area. This study aimed to develop a forecasting model of malaria incidence based on the time series analysis incorporating climatic variables in the endemic area near the Thai-Myanmar border in Ratchaburi province, Thailand. The retrospective study was carried out using monthly malaria cases and the climatic variables, including maximum and minimum temperature, rainfall, humidity, between January 1999 and December 2011 from The Rajanagarindra Tropical Disease International Centre (RTIC), located in a rural community in Ratchaburi province. A Seasonal Autoregressive Moving Average (SARIMA) model was developed and validated by dividing the time series data into two parts; the data from January 1999 to December 2010 were used to construct a model and those from January to December 2011 were used to validate it. Cross-correlation analyses were performed to reveal a significant correlation between monthly malaria cases and the climatic variables. The SARIMA model can produce a reliable forecast of malaria cases. In additional, the model with the climatic variables fit better than the model without the climate variables. Therefore, the SARIMA forecasting model can be use for planning and managing malaria prevention and control program which can strengthen an early warning system in this endemic area.

Keywords: Malaria, climate, time series, SARIMA, forecasting

*Corresponding Author
E-mail Address: ngamphol.soo@mahidol.ac.th

Use of factor analysis to compare agronomic trait interrelationships in determinate and indeterminate types of sesame

M. İlhan Çağırgan[1], Selçuk Özerden[2], Mehmet Tekin[1*] and M. Onur Özbaş[1,3]

[1]*Antalya Mutation Project, Dept. of Field Crops, Faculty of Agriculture, Akdeniz University, TR07059 Antalya, Turkey.*

[2]*Directorate of Agricultural Protection and Quarantine, Ministry of Agriculture, Mersin, Turkey*

[3]*Present address: Enza Zaden Tarım Ar-Ge, Hisaraltı Mevkii, P.O. Box 87, Antalya, Turkey*

Abstract

Sesame (*Sesamum indicum* L.) is one of the most important oil crops and is grown throughout the tropics and sub-tropics. Sesame seeds are used for confectionary over the globe, as high quality of edible oil and making tahini. Although it is well suited to different farming systems, its indeterminate growth habit causes non-uniform capsule maturation and thus difficulty at deciding when to harvest. Despite the progress, improved determinate lines of sesame are still low yielding. To make further genetic progress in yield capacity, agronomic traits affecting yield in the novel determinate (dtdt) types should be compared with the indeterminate (DtDt) types. The common obstacle regarding to integration of a new trait in to the elite breeding populations is that the novel trait, dtdt in our case, generates many unwanted side effects in other characters, restricting its direct use. To determine the selection criteria for yield potential, all characters affecting yield should be considered simultaneously. Factor analysis is designed to do this and it is a multivariate statistical method that can be successfully utilized in understanding the patterned variation in a set of variables based on the structural relationships among them. We applied factor analysis on 11 agronomic traits measured in the determinate and indeterminate types in segregating 45 F2 populations derived from the dtdt x DtDt crosses. The factor analysis reduced 11 correlated variables to 4 factors, smaller number of unrelated grouped variables called \`factors\` or \`patterns\` both in the determinate and indeterminate types. Together these four factors accounted for 85.6 % of the variance for the eleven correlated traits in the determinate type. The communalities accounted for by all factors taken together were between 0.684 (no of seeds per capsule) and 0.978 (Seed weight per capsule). Each factor accounted for the total variance 25.1%, 21.0%, 20.5%, and 19.0%, respectively, and explained 85.6% in total. Regarding indeterminate type, the communalities ranged between 0.225 (no of seed per capsule) and 0.973 (days to stopping flowering). Each factor accounted for the total variance 22.3%, 20.6%, 17.8%, and 14.1%, respectively, and explained 74.9% in total. The lesser variance accounted for in the indeterminate group attributed to the lesser variability in this group than the determinate group, as expected. Factor analysis is a useful technique in the study of the patterned variability resulting simultaneous changes in a novel trait for sesame such as determinacy in comparison to its wild type, the indeterminacy.

Keywords: Eigenvalues, interrelationships, patterned variability, factor analysis, varimax

*Corresponding Author
E-mail Address: mehmettekin@akdeniz.edu.tr

Comparison between proportional hazard model (PHM) approach and hierarchical summary receiver operating characteristic (HSROC) for SROC in meta-analysis of diagnostic test studies

Supada Charoensawat[1*] and Dankmar Böhning[2]
[1]*Health Sciences Program, Udon Thani Rajabhat University, Udon Thani, Thailand, suphadac@hotmail.com*
[2]*School of Mathematics and Southampton Statistical Sciences Research Institute, University of Southampton, UK, D.A.Bohning@soton.ac.uk*

Abstract

The number of meta-analysis of diagnostic studies is increasing and the models which deal with the summary receiver operating characteristic (SROC) have become popular, for example, the bivariate random effects model (BREM) and the hierarchical summary receiver operating characteristics (HSROC). The two existing models have reached considerable statistical complexity, required expertise and knowledge. However, these two models are mathematically equivalent when no covariates are included. A model named the proportional hazard model (PHM) is developed. The PHM has a simple form, namely it has only one parameter of interest (θ), called the diagnostic accuracy. Also, the PHM is easy to interpret that the smaller theta, the higher the diagnostic accuracy. The improved PHM structure is largely enriched by allowing a random effects component. The adjusted profile maximum likelihood estimator (APMLE) was chosen as the best, among several estimators to estimate theta. A comparison between the existing model, the HSROC, and the PHM is the central issue of this study. An application of meta-analysis of diagnostic studies is applied in this study.

Keywords: The PHM, SROC, HSROC, meta-analysis, diagnostic study

*Corresponding Author
E-mail Address: suphadac@hotmail.com

Weibull regression model for testing factors affecting survival of cancer patients with *Electro-Capacitive Cancer Therapy* (ECCT)

Hamid Dimyati[1] and Sri Haryatmi[2]
[1]*Mathematics Department, Universitas Gadjah Mada, Yogyakarta, 3404, Indonesia, hamid.dimyati@mail.ugm.ac.id*
[2]*Mathematics Department, Universitas Gadjah Mada, Yogyakarta, 3404, Indonesia, s_kartiko@yahoo.com*

Abstract

According to data from the World Health Organization (WHO), as many as 8.2 million people worldwide deaths caused by cancer in 2012. Approximately 70% of deaths caused by cancer occured in under-developed and developing countries, including Indonesia. Various methods have been found to treat cancer, but currently there is no cure. Therapeutic cancer treatment developed by Dr. Warsito Purwo Taruno named *Electro-Capacitive Cancer Therapy* (ECCT) have invited a lot of medical attention. He used his new technologies finding –*Electrical Capacitance Volume Tomography* (ECVT) to be applied to the cancer therapy apparel. Through his Cancer Research Clinic –C-Care , Dr. Warsito have applied the therapeutic treatment to more than 11,000 cancer patients since 2012. Although the therapy does not guarantee in total cure, but it saved the patients from early death due to the spread of cancer cells that are not controlled. By using Weibull regression model in survival analysis, it was found that the factors that cause ECCT therapy patients survived longer include the frequency of monitoring, alternative treatments, chemotherapy, radiotherapy and type of the cancer. In addition, this research also sought the patient's chances of survival comparison based on type of cancer and by sex.

Keywords: Cancer, ECCT, survival analysis, Weibull regression

1. Introduction

Cancer is a disease caused by an abnormal growth of body tissue cells of which transformed into cancer cells. During its development, these cancer cells can spread to other body parts that can cause death. Various methods have been found to treat cancer, but currently there is no cure. According to data from the World Health Organization (WHO), as many as 8.2 million people worldwide deaths caused by cancer in 2012. Approximately 70% of deaths caused by cancer occured in under-developed and developing countries, including Indonesia.

As a developing country, Indonesia can not escape from the threat of cancer. Based on Indonesia Hospital Information System (SIRS) in 2010, cancer became the third cause of death with an incidence of 7.7% of all causes of death from non-communicable diseases, after a stroke and heart disease. Meanwhile, breast cancer and cervical cancer are two highest on the type of cancer inpatients and outpatients in all hospitals in Indonesia with a proportion of 28.7% for breast cancers and cervical cancers are 12.8%, leukemia 10.4%, lymphoma 8.3%, and 7.8% of lung cancers.

At the end of 2011, an Indonesia physicist named Dr. Warsito Purwo Taruno found cancer treatment technology by utilizing static electricity. Starting from concern over the fourth stage breast cancer of his older sister, Suwarni, which then drives the instinct of a researcher in the field of tomographic to create a tool of cancer treatment by utilizing the concept of Electrical Capacitance Volume-Tomography (ECVT). This tool can scan 3D moving objects or scan the volume based on the movement of time and able to determine the contents of the object without unload it. ECVT is the only method that help scientists get an accurate three-dimensional picture corresponding to the object of scan.

From these findings ECVT technology, Warsito create an apparel that can be used for cancer treatment therapies. Method of cancer healing which he created named Electro-Capacitive Cancer Therapy (ECCT). ECVT technology allows preventing cancer cells to grow and spread, thus help the patient to survive longer. Although the therapy does not guarantee in total cure, but it saved the patients from early death due to the spread of cancer cells that are not controlled. ECCT has several steps including a patient's medical record document submission, scan, use of cancer clothes or apparel, alternative treatment, herbs, chemotherapy, radiotherapy and periodic monitoring.

The goal of this research is to determine the effect of this therapy on the survival of a cancer patient who is registered as a patient C-Care -Cancer Research Clinic. In addition, to know the differences in survival of patients based on the type of cancer suffered by the patient and based on sex of patients. Weibull regression model for survival function will answer these questions, as well as identifying factors that affect the survival of the patient's cancer treatment.

2. Survival Analysis

Survival data is a long time until an event occurs or often referred to as time-to-event data. In survival analysis, there are several components that must be fulfilled, namely,

- event ; something that concerns the research,
- origin ; starting point that is used to measure the length of time until an event occurs,
- the unit of measurement used.

The problems that often appear in the analysis of survival data is the presence of incomplete observations, which can generally be categorized into censored data and truncated data.

Survival function is the probability of an individual living or staying in a state for longer than t, denoted by;

$$S(t) = P(T > t). \tag{2.1}$$

The variable t is obtained by calculating the length of time from the origin until the occurrence of the event. If there is right censoring data, t is calculated from the origin to the end of research.

2.1 Identification Distribution

Several parametric distributions commonly used in the analysis of survival data, among others, Exponential distribution, Weibull distribution, Gamma distribution, Log-normal distribution, Gompertz-makeham distribution, Log logistic distribution and many more. In identifying survival data can use analysis of probability-plot or hazard plot and through correlation coefficient.

Here are some guidelines for suspected distribution of survival data;

Table 1: Distribution Characteristics of Exponential, Weibull and Log-normal

Distribution	Probability-plot	Hazard plot
Exponential	Plot a straight line between t and $log\left[\frac{1}{1-\hat{F}(t)}\right]$	Plot a straight line between t and $\hat{H}(t)$
Weibull	Plot a straight line between $\log t$ and $log\frac{1}{\lambda}+\frac{1}{\alpha}log\left[log\frac{1}{1-\hat{F}(t)}\right]$	Plot a straight line between $\log t$ and $log\left[\hat{H}(t)\right]$
Log-normal	Plot a straight line between $\log t$ and $\phi^{-1}\left(\hat{F}(t)\right)$	Plot a straight line between $\log t$ and $\phi^{-1}\left(1-e^{-\hat{H}(t)}\right)$

To test the linearity probability-plot can use the correlation coefficient (ρ):

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{2.2}$$

where σ_{XY} is the value of the covariance between x (time) and y (median rank), σ_X is the standard deviation of x and σ_Y is the standard deviation of y. Estimator ρ denoted by $\hat{\rho}$ gives

$$\hat{\rho} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} \tag{2.3}$$

with $-1 \le \hat{\rho} \le 1$.

2.2 Linear Regression Model

In linear regression models involve the independent variables and the dependent variable. Independent variables commonly is denoted by the letter $\mathbf{X} = (x_1, x_2, \dots, x_p)$, while the regression parameters is denoted by $\theta = (\beta_1, \beta_2, \dots, \beta_p)$.

The independent variables and the regression parameters, stated by $\psi(X;\theta)$, can be a linear combination

$$\psi(\mathbf{X};\theta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{2.4}$$

In general, the regression model (2.4) containing error variables are denoted with ε_i. Mean value $E(\varepsilon_i) = 0$ and Variance $V(\varepsilon_i) = \sigma^2$. By other terms, ε_i and ε_j are not correlated so $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all values of i and $i \neq j$; $i = 1,2,3,\dots,n$.

The linear regression equation is obtained as follows

$$\hat{Y} = \psi(\mathbf{X};\hat{\theta}) + \varepsilon_i \tag{2.5}$$

Estimator $b_0, b_1, \dots, b_k$ searched using Least Squared Error (LSE) which is a method for finding the estimator by minimizing the sum of squared errors.

$$\begin{aligned} L &= \sum_{i=1}^{N} {\varepsilon_i}^2 \\ &= \sum_{i=1}^{N}\left(\hat{Y}_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_p X_{pi}\right)^2 \end{aligned} \tag{2.6}$$

The L model (2.6) will be worth a maximum if the partial derivative of the $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ equal to zero. Obtained system of linear equations with p variables estimator. Such equations will be easier if calculated in the form of a matrix;

$$X'Xb = X'Y, \tag{2.7}$$

with

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{bmatrix}, \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}.$$

From equation (2.7), we obtain estimator vector b :

$$b = (X'X)^{-1}X'Y.$$

Finally, the estimated regression equation is given by

$$Y = b_0 + b_1X_1 + \cdots + b_pX_p \qquad (2.8)$$

2.3 Weibull Regression Model

The influence of the independent variable X in the Weibull regression model is expressed through the scale parameter $\lambda = f_\lambda(\mathbf{X};\beta)$, and the shape parameter $\alpha = f_\alpha(\mathbf{X};\beta)$ in the form of an exponential function $exp(\mathbf{X}\beta)$. Models with $\lambda = f_\lambda(\mathbf{X};\beta) = \exp(\mathbf{X}\beta)$ is the most frequently used so that the survival function for the Weibull regression is

$$\begin{aligned} S(t|\mathbf{X}) &= exp(-(f_\lambda(\mathbf{X};\beta)t)^\alpha) \\ &= exp(-(exp(\mathbf{X}\beta)t)^\alpha) \qquad (2.9) \end{aligned}$$

If the survival random variable T is transformed into $Y = \log T$, then the survival function of Y is

$$\begin{aligned} S(y|\mathbf{X}) &= exp(-(exp(\mathbf{X}\beta)e^y)^\alpha) \\ &= exp\left(-\left(exp(y+\mathbf{X}\beta)\right)^\alpha\right) \\ &= exp\left(-(exp(y+\mathbf{X}\beta)\alpha)\right), \qquad (2.10) \end{aligned}$$

or can be written as

$$S(y|\mathbf{X}) = exp\left[-exp\left(\frac{y-\mathbf{X}\beta}{\sigma}\right)\right], \qquad (2.11)$$

which is known as the survival function of the extreme value distribution, with location parameter $\mu = -\mathbf{X}\beta$ and scale parameter $\sigma = \frac{1}{\alpha}$.

Likelihood function models prepared by the general form likelihood function is;

$$L(\beta,\alpha) = \prod_{i=1}^{n} f(t_i,\theta|\mathrm{X}_i)^{\delta_i}.S(t_i,\theta|\mathrm{X}_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \begin{array}{c} \left\{\frac{1}{\sigma} exp\left[\frac{y-\mu}{\sigma} - exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\delta_i} \\ \left\{exp\left[-exp\left(\frac{y-\beta\mathrm{X}}{\sigma}\right)\right]\right\}^{1-\delta_i} \end{array} \qquad (2.12)$$

2.4 Coefficient Regression Test of Significance

In testing the regression coefficient of the survival data will use hypothesis testing rule, the same as the linear regression coefficient test in general. The test has a purpose of determining significance of each regression coefficients whether influence on dependent variable. To test whether there is a relationship between the dependent variable Y with the independent variables $X_1, X_2, \ldots, X_p$ use test statistic $F = \frac{MSR}{MSE}$, while partial testing the regression coefficients β_i use test statistic $z = \frac{b_i - \beta_i}{\hat{\sigma}_{b_i}}$.

Hypothesis Testing for β_i (coefficients);

- H_0 : $\beta_i = 0$ (coefficient of X is not feasible included into the model).
- H_1 : $\beta_i \neq 0$ (coefficient of X is feasible included into the model).
- Test statistic used is
$$z = \frac{b_i - \beta_i}{\hat{\sigma}_{b_i}}. \qquad (2.13)$$
- Area criticism ;

H_0 is rejected if $\left|z_{hitung}\right| \geq \left|z_{\frac{\alpha}{2}}\right|$ or $p - value < \alpha$

3. Research Methodology

3.1 Data Sources

This research used data treatment of cancer patients from Edwar Health Care (EHC) from February 1, 2013 to January 31, 2014. Samples were taken as 212 people.

3.2 Research Variables

Dependent variable is the survival time of patients and the variable status (0 for censored data, 1 for uncensored data). While the independent variables that will be analyzed are the frequency of monitoring, gender (1 for female, 2 for male), age, metastatic (0 for no metastatic, 1 for metastatic), herbs (0 for not use herbs, 1 for use herbs), alternatives (0 for not use alternative treatments, 1 for use other alternative treatments), chemotherapy (0 for no use of chemotherapy, 1 for use of chemotherapy), radiotherapy (0 for no use of radiotherapy, 1 for use radiotherapy) and types of cancer (Bone/Spine cancer as the reference category)

3.3 Research Steps

The steps used in this research are as follows;

- Determine the survival data of patients initiating therapy time until an event occurs (died).
- Describe the characteristics of cancer patients who researched.
- Predict survival data distribution.
- Identify patterns of association between survival data with independent variables using parametric regression model for survival function.
- Eliminate the factors that are not significant.
- Get the most suitable model for survival data samples.

4. Research Results and Discussion

4.1 Characteristics of Cancer Patients

Of the 212 cancer patients were used as samples of research, the following description;

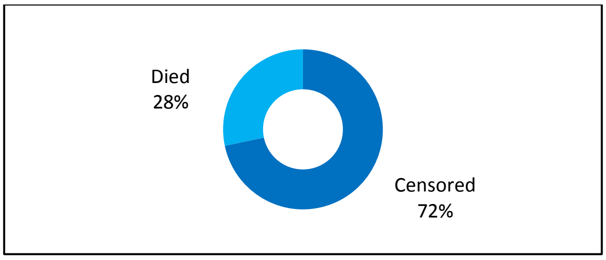


Figure 1 Proportion of patient status

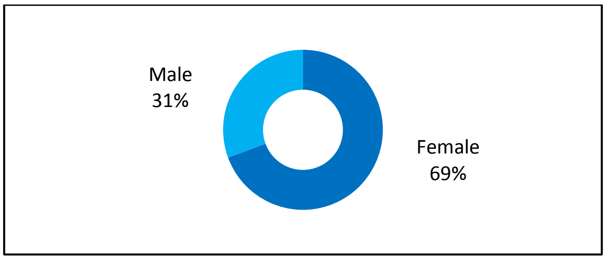


Figure 2 Proportion of patient gender

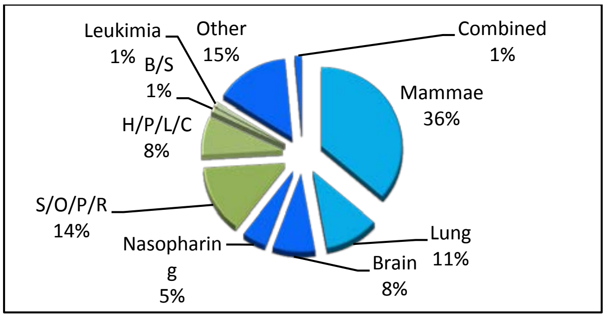


Figure 3 Proportion of cancer patients

Figure 1 shows that censored data has a larger number of samples is 72% while the samples affected by events as much as 28%. Figure 2 shows the proportion of patients in whom the sex of the sample is still dominated by women by 69% and men 31%. In Figure 3 describes the percentage of cancers suffered by the samples.

Table 2: Other Variables Proportion of Patients

Factors	Number of Patients	
	(+)	(-)
Metastatic	42	170
Herbs	21	191
Alternatives	6	206
Chemotherapy	5	207
Radiotherapy	1	211

Table 2 shows the proportion of the other independent variables. For metastatic variable, the number of patients who were positive metastatic cancer totaled 42 people, while 170 patients were negative status. The number of patients taking herbal treatment as many as 21 people, while 191 patients did not. Patients who received alternative treatment amounted to 6 people and the rest did not. There were 5 people who receiving chemotherapy and the rest did not. Finally, the number of patients who acquired a radiotherapy only one person from the samples, while the other 211 patients did not receive radiotherapy.

4.2 Distribution Estimation

In survival analysis, the primary variables required are time variable and status variable. Time variable, as already noted above is the survival value of the sample, while the status indicates right-censored state. In terms of if a patient died before the end of research then the patient is considered uncensored and weighted with a value of 1. As for patients who survive past the end of research or they are still alive until January 31, 2014, the patients is said to censored and rated 0.

Before continuing to the next analysis, we need to know first the time data distribution. Identification is done by looking at the plot and correlation. The suitable data distribution is indicated by the correlation coefficients value close to 1.

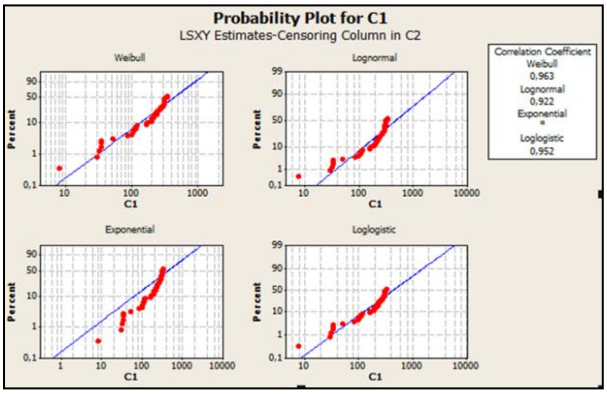


Figure 4 Comparison of Probability Plots in Identifying Distribution

From Figure 4, we can see from the values of Correlation Coefficient; Data has a value of 0.963 to the Weibull distribution, 0.922 of the Log-normal distribution and 0.952 of the Log-logistic. So it can be concluded that the data is to follow the Weibull distribution with the value of Correlation Coefficient of 0.963.

4.3 Parametric Regression Model

The variables that significantly affect the length of survival of cancer patients are the Frequency of Monitoring, Types of Cancer, Alternatives Treatment, Chemotherapy and Radiotherapy. Weibull regression model for survival function is as follows:

$$S(t|X) = exp(-(exp(-4{,}7784 - 0{,}1177 * Frek - 1{,}1072 * JKmammae - 0{,}7325 * JKlung - 0{,}9192 * JKbrain - 0{,}9692 * JKnasopharing - 0{,}8215 * JKsopr - 0{,}8645 * JKhplc - 0{,}8768 * JKleukimia - 0{,}8691 * JKother + 0{,}0187 * JKcombined - 7{,}6460 * Alter - 7{,}3986 * kaliKemo - 1{,}1348 * kaliRadiasi)\, t)^{2{,}3364}). \qquad (4.1)$$

Model (4.1) shows that the opportunity to live a cancer patient is supported by Frequency of Monitoring, Alternative Treatment, Chemotherapy, Radiotherapy and Type of Cancer.

Survival function that have been obtained can be used to make comparison of survival based on the type of cancer and gender. Suppose patients are not getting alternative treatments, chemotherapy and radiotherapy as well as monitor 3 times, within 1 year or 366 days, the chances of survival of patients suffering from different cancers by using the model (4.1) are listed in Table 3;

Table 3: Patients Probability to Survive by Type of Cancer

Type of Cancer	S(t|X)
Mammae	0.634
Nasopharing	0.533
Brain	0.493
Leukemia	0.458
Other	0.451
Hepar/Pancreas/Lymp/ Colon	0.448
Servics/Ovarium/Prostate/Rectum	0.411
Lung	0.335
Bone/Spine	0.002
Combined	0.002

From the result that most major types of cancer that causes death are Bone/Spine cancer and Combined (patients who have more than one type of cancer). Two types have a chance to live for only 0.002 with conditions as noted above. The type of cancer that most likely to survive are breast cancer with odds of 0.634 with prescribed conditions.

Build Weibull regression model for survival function with a variable gender re-enter into model (4.1) for comparing survival function based on gender of the patients. The factors that go into the model are the Frequency of Monitoring, Gender, Type of Cancer, Alternative Treatment, Chemotherapy, and Radiotherapy

Weibull regression model for survival functions obtained is as follows:

$$S(t|X) = exp(-(exp(-4{,}7988 - 0{,}1176 * Frek + 0{,}0127 * Gender - 1{,}1004 * JKmammae - 0{,}733 * JKlung - 0{,}9150 * JKbrain - 0{,}9728 * JKnasopharing - 0{,}8152 * JKsopr - 0{,}8656 * JKhplc - 0{,}8693 * JKleukimia - 0{,}8690 * JKother + 0{,}0235 * JKcombined - 7{,}6210 * Alter - 7{,}3998 * kaliKemo - 1{,}1134 * kaliRadiasi)\, t)^{2{,}3364}). \qquad (4.2)$$

With the condition that the patient is suffering from brain cancer, did not receive: alternative treatment, chemotherapy, and radiotherapy as well as monitor 3 times, within 1 year or 366 days ahead, the patients probability to survive by sex are as follows;

Female patients = 0.495

Male patients = 0.485

It can be concluded that female patients had probability to survive greater than male patients.

5. Conclusion

Based on the discussion in the previous section, it can be concluded that parametric regression model for the survival function can be used to examine the factors that affect the long life of patients with cancer. The factors that significantly affect the chances of survival of patiens after ECCT are Frequency of Monitoring, Alternative Treatment, Chemotherapy, Radiotherapy and Type of Cancer.

Using model (4.1), the most major types of cancer that causes death are Bone/Spine cancer and Combined (patients who have more than one type of cancer). Two types have a chance to live for only 0.002 with conditions as noted above. The type of cancer that most likely to survive are breast cancer with odds of 0.634 with prescribed conditions. Using model (4.2), it can be concluded that female patients have better survival rate than male patients with the condition that the patient is suffering from brain cancer, did not receive: alternative treatment, chemotherapy, and radiotherapy as well as monitor 3 times, within 1 year or 366 days ahead.

Acknowledgements

This work was supported by Mathematics Department, Universitas Gadjah Mada (UGM) and Edwar Health Care (EHC).

References

[1] D. R. Cox. Regression Models and Life-Tables. Journal of the Royal Statistical Society. 1972; Series B (Methodological), Vol. 34, No. 2, pp.187-220

[2] Edwin MMO, Gauss MC, Michael WK. The log-beta Weibull regression model with application to predict recurrence of prostate cancer. Springer-Verlag. 2013; 54: 113-132

[3] Skinner and Humphreys. Weibull Regression for Lifetime Measured with Error. Lifetime Data Analysis, 5. 1999; 23-37

[4] Fa'rifah, Riska Y., Purhadi. Analisis Survival Faktor-Faktor yang Mempengaruhi Laju Kesembuhan Pasien Penderita Demam Berdarah Dengue (DBD) di RSU Haji Surabaya dengan Regresi Cox. Jurnal Sains dan Seni Institut Teknologi Sepuluh Nopember Vol.1. 2012; 2301-928X

[5] Lawless, Jerald F. Statistical Models and Methods for Lifetime Data (second edition). New Jersey: John Wiley & Sons; 2003.

[6] Danardono. Analisis Data Survival. Yogyakarta: Statistics-Universitas Gadjah Mada; 2012

[7] Chen, Cheng-Mei & Friends. Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees, Proceeding of the 7th International Conference on Natural Computation; 2011 July 26-28; Shanghai, China. 2011.

[8] Anonymous. C-Tech Labs Profile. Tangerang Selatan: C-Tech Labs; 2013

[9] Parkway Cancer Center. About Cancer [Internet]. 2013 [cited 2014 February 5]. Available from: http://www.parkwaycancercentre.com/id/

[10] World Health Organization. Cancer Facts in the World [Internet]. 2014 [cited 2014 February 5]. Available from: http://www.who.int/cancer/en/

[11] Indonesia's Health Ministry. Cancer Facts in Indonesia [Internet]. 2013 [cited 2014 February 20]. Available from: http://www.depkes.go.id/index.php?vw=2&id=2233

Economic model for fuzzy Weibull distribution

P. Charongrattanasakul[1] and A. Pongpullponsak[2*]

[1]*Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand, motecooler@hotmail.com*

[2]*Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand, adisak.pon@kmutt.ac.th*

Abstract

Investigation of economic model under fuzzy quality control environments is proposed in this paper. The fuzzy concepts are applied for analysis of the Weibull Distribution. The goal of this article is to construct the economic model was combining Fuzzy Weibull Distribution with Integrated model (Statistical process control and Maintenance management). The genetic algorithm is a method to find optimal values of six variables $(n, h, k, \kappa_{\min}, \kappa_{\max}, \alpha - \text{cut})$ that minimize the hourly cost $(E[H]_{\text{Lower}}, E[H]_{\text{Upper}})$. Finally, multiple regressions are employed to demonstrate the effect of cost parameters.

Keywords: Integrated model, maintenance management, fuzzy Weibull distribution, genetic algorithm

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

1. Introduction

Nowadays, control charts are widely used to maintain and established statistical control of a process. The control charting technique plays an important role in production process monitoring. The major function of control chart detecting the occurrence of assignable causes, so that, the necessary corrective action may be taken before a large quantity of nonconforming product is manufactured. The control charting technique may be considered as the graphical expression and operation of statistical hypothesis testing. When a control chart is applied to monitor process, some test parameters should be determined, the sample size, the sampling interval between successive samples, and the control limits or critical region of the chart.

Statistical process control (SPC) is an efficient method for improving a firm's quality and productivity. The main objective of SPC is to quickly detect the occurrence of assignable causes or process shifts, so that investigation of the process and corrective action may be undertaken before large numbers of non-conforming units are manufactured. SPC has two main tools to control the process. One is ‘‘acceptance sampling” and the other one is ‘‘control charts”. The control charts are on-line process control techniques, widely used to monitor the process. With the use of control charts and collecting few but frequent samples, the SPC can effectively detect changes in the process that may affect its quality.

Most of our traditional tools for formal modeling, reasoning, and computing are crisp, deterministic, and precise in character. Crisp means dichotomous, that is, yes-or-no type rather than more-or-less type. In traditional dual logic, for instance, a statement can be true or false and nothing in between. In set theory, an element can either belong to a set or not; in optimization a solution can be feasible or not. Precision assumes that the parameters of a model represent exactly the real system that has been modeled. This, generally, also implies that the model is unequivocal, that is, contains no ambiguities. Certainty eventually indicates that structures and parameters of the model to be definitely known. There are no doubts about their values or their occurrence. Unluckily these assumptions and beliefs are not justified if it is important, that the model describes well reality (which is neither crisp nor certain). In addition, the complete description of a real system would often require far more detailed data than a human being could ever recognize simultaneously, processing, and understanding. This situation has already been recognized by thinkers in the past. In 1923, the philosopher Russell [1] referred to the first point by written that “All traditional logic habitually assumes that precise symbols are being employed”. It is, therefore, not applicable to this terrestrial life but only to an imagined celestial existence. Zadeh [2] referred to the second point, by written that “As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics”. For a long time, probability theory and statistics have been the predominant theories and tools to model uncertainties of reality (Lodwick [3], Jamison and Zadeh [4]). Zimmermann [5-6] referred to base as all formal theories on certain axiomatic assumptions, which are hardly ever tested, when these theories are applied to reality. In the meantime more than 20 other "uncertainty theories" have been developed, which partly contradict each other and partly complement each other. Fuzzy set theory formally speaking is one of these theories, which was initially intended to be an extension of dual logic (and/or) classical set theory. During the last decades, it has been developed in the direction of a powerful "fuzzy mathematics". When it is used, however, as a tool to model reality better than traditional theories, then an empirical validation is very desirable. Stojakovi [7] referred to deals with a new concept introducing notion of fuzzy set to one mathematical model describing an economic system. A finite pure exchange economy is considered. Hussien Aliwi and Fawzi Hamza [8] referred to find the type of relation and to decrease the variations in values through finding a combination between several probability distributions and membership function in fuzzy logic to be either continuous or discrete. Weibull distribution has been exploited with a continuous membership function, Gaussian membership and applied failure rate function. Different methods of estimation are discussed for the parameters of Weibull distribution when the available data are in the form of fuzzy numbers (Pak, Parham and Saraj [9]). They include the maximum likelihood estimation, Bayesian estimation and method of moments. The estimation procedures are discussed in details and compared via Monte Carlo simulations in terms of their average biases and mean squared errors. Finally, a real data set taken from a light emitting diodes manufacturing process is investigated to illustrate the applicability of the proposed methods.

Investigation of reliability characteristics under fuzzy environments is proposed. Fuzzy Weibull distribution and lifetimes of components are selected to describe the process distribution. Formulas of a fuzzy reliability function, fuzzy hazard function and their α-cut set are presented. Furthermore, the fuzzy functions of series systems and parallel systems are discussed, respectively. Finally, some numerical examples are presented to illustrate how to calculate the fuzzy reliability characteristics and their α-cut set (Jamkhaneh [10]). After investigating the advantages and disadvantages of current methods of statistical process control, it becomes important to overcome the disadvantages and then use the advantages to improve a method for monitoring a process with categorical observations. An approach which considers uncertainty and vagueness is tried for this study; and for this purpose, fuzzy set theory is inevitable to use. So, a new approach based on fuzzy set theory is introduced in this research for monitoring attribute quality characteristics. This approach is then compared with the current related approach to see the difference in performance (Sorooshian [11]).

The aim of this work is to develop the integrated economic design for Fuzzy Weibull Distribution economic model by EWMA Control Chart. For determining the values of four test variables of the chart (the sampling interval (h), the sample size (n), the number of samples taken before Planned Maintenance (k), the Alpha-cut $(\alpha\text{-cut})$, the minimum vagueness coefficient $(\kappa_{\min})$ and the maximum vagueness coefficient $(\kappa_{\max})$). By using this developed genetic algorithm to optimize these parameters (four test parameters), the total cost per hour $(E[H]_{\text{Lower}}, E[H]_{\text{Upper}})$ is expectedly minimized.

2. Research Methodology

In this work, an integrated model of control chart with reference to the three-scenario integrated model firstly purposed by Linderman and Anderson [12] is developed. Then a generalized analytic model is employed to determine the optimal policy for coordinating Statistical Process Control and Planned Maintenance in minimizing the total expected cost. Recently, Zhou and Zhu [13] modified the Linderman model from three to four policies under determination of the optimal policy for minimization of the total expected cost with coordination of Statistical Process Control and Planned Maintenance. For this research, an integrated model between Statistical Process Control and Planned Maintenance of the EWMA control chart is conducted. In developing, integrated model for Fuzzy Weibull Distribution economic model by EWMA Control Chart, as shown in Figure 1, the framework of the integrated model illustrates four different scenarios, in which each scenario is further elaborated as following. In Scenario 1, the process begins with a "in-control" state and inspections occur after h hours of monitoring as to whether the process has shifted from an "in-control" to an "out-of-control" state. There is an alert signal in the control chart before the scheduled time when maintenance should be performed. But the signal is false, that is to say, the process is still in-control. Since searching and determining false signal take time and incur cost, Compensatory Maintenance is performed. In Scenario 2, similar to Scenario 1, there is also a signal. While the signal is valid and the process shifts to an "out-of-control" state, it results in Reactive Maintenance. In Scenarios 3 and 4, no signal occurs in the control chart before the scheduled time. Then at the $(k+1)^{\text{th}}$ sampling interval, appropriate maintenance should be arranged. In Scenario 3, the process is always "in-control", Planned Maintenance is performed. When the process shifts to an "out-of-control" state in Scenario 4, Reactive Maintenance takes place because the "out-of-control" condition occurred before the scheduled time, and additional time and expense will be incurred to identify and solve the equipment problem.

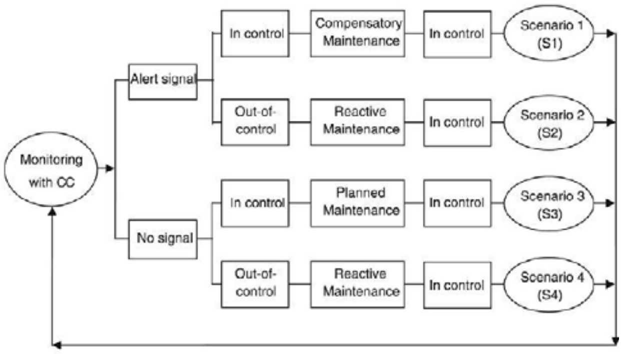


Figure 1: Integrated model of Scenarios

Fuzzy Weibull distribution (Karpisek [14])

This model results the first model of fuzzy probability distribution. Suppose that the values of a fuzzy random variable $\underset{\sim}{T}$ are the fuzzy number $\underset{\sim}{t} = \left([0,\infty), \mu_t\right)$ and $\underset{\sim}{t} = \underset{\sim}{\kappa} t$ where t is the observed value of a crisp random variable is T and $\underset{\sim}{\kappa}$ is a so-call vagueness coefficient. The vagueness coefficient is real triangular fuzzy number $\underset{\sim}{\kappa} = \left([0,\infty), \mu_{\underset{\sim}{\kappa}}\right)$ with the main value $\kappa = 1$ and membership function

$$\mu_{\underset{\sim}{\kappa}}(\kappa) = \begin{cases} \dfrac{\kappa - \kappa_{\min}}{1-\kappa_{\min}}, & \kappa \in [\kappa_{\min}, 1]\ , \\ \dfrac{\kappa - \kappa_{\max}}{1-\kappa_{\max}}, & \kappa \in [1, \kappa_{\max}]\ , \\ 0 & \text{otherwise} \end{cases}$$

(1)

where $0 < \kappa_{\min} \le 1 \le \kappa_{\max}$, and boundary values $\kappa_{\min}, \kappa_{\max}$ they are given by an expert's estimate. Figure 2 shows graph of $\mu_{\underset{\sim}{\kappa}}$.

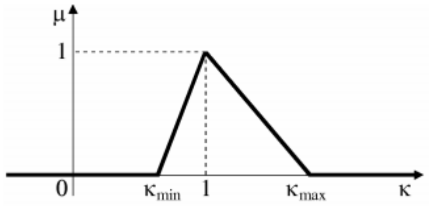


Figure 2: Triangular fuzzy number of $\mu_{\underset{\sim}{\kappa}}(\kappa)$.

If a random variable T has a crisp Weibull probability distribution $W(b,\delta)$ then the corresponding fuzzy random variable $\underset{\sim}{T}$ with the Weibull Fuzzy probability distribution $\underset{\sim}{W}(b,\delta)$ has the following fuzzy characteristics. For $\forall t \in [0,\infty)$ the fuzzy distribution cumulative function

$\underset{\sim}{F}(t) = 1 - e^{-[\lambda t/(\underset{\sim}{\kappa})]^{\nu}}$, so that for $\forall \alpha \in [0,1]$ (2)

The $\alpha-$cut of fuzzy distribution cumulative function

$$F_{\alpha}(t) = [F_{1\alpha}(t), F_{2\alpha}(t)]$$
$$= 1 - e^{-[\lambda t/[(1-\kappa_{\max})\alpha + \kappa_{\max}]]^{\nu}} \ , 1 - e^{-[\lambda t/[(1-\kappa_{\min})\alpha + \kappa_{\min}]]^{\nu}}$$

(3)

Production cycle

The production cycle, Zhou and Zhu [13] is defined in production duration. The production process will be assumed to be constantly under control at the beginning. Production cycle composes of two time periods which are under control period and not under control period as the following details:

Cycle Time $E[T]$

T_0 : The expected time searching for a false alarm,

T_P : The expected time to identify maintenance requirement and to perform a Planned Maintenance,

T_A : The expected time to determine occurrence of assignable causes,

T_R : The expected time to identify maintenance requirement and to perform a Reactive Maintenance,

T_C : The expected time to perform a Compensatory Maintenance,

τ : The mean elapse time from the last sample before the assignable cause to the occurrence of the assignable cause,

ARL_I : The average runs length during in-control period,

ARL_O : The average runs length during out-of-control period,

E : The expected time to sample and chart one item,

$\gamma_P(\gamma_R, \gamma_C, \gamma_A)$: The indicator variable (If it equals 1 production continuous during Planned Maintenance (Reactive Maintenance, Compensatory Maintenance, validate assignable cause) or 0 otherwise),

P_i^I : The probability that run length of control chart equals i during in-control period $P_i^I = \alpha(1-\alpha)^{i-1}$,

P_i^O : The probability that run length of control chart equals i during out-of-control period $P_i^O = (1-\beta)\beta^{i-1}$,

L : The width of control limit in units of standard deviation,

The cost model

The cost model used for determining the optimal values of chart parameters is built upon the general cost function of Lorenzen and Vance [15].

The expected Hourly cost per time unit $(E[\text{Hourly Cost}])$ is the Ratio of the expected cost per cycle $E[\text{Cycle Cost}]$ to the expected cycle time $E[\text{Cycle Time}]$, that is

$$E[\text{Hourly Cost}] = \frac{E[\text{Cycle Time}]}{E[\text{Cycle Cost}]}$$

Cycle Cost $E[C]$

C_I : The cost of quality loss per unit time (the process is in an in-control state) often estimated by a Taguchi Loss function,

C_O : The cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function,

C_P : The cost of performing Planned Maintenance,

C_R : The cost of performing Reactive Maintenance,

C_C : The cost of performing Compensatory Maintenance,

C_F : The fixed cost of sampling,

C_V : The variable cost of sampling,

C_f : The cost to investigate a false alarm,

Optimal Variable

h : The interval between sampling. (h^* for optimal),

n : The sampling size. (n^* for optimal),

k : The number of sample taken before Planned Maintenance.(k^* for optimal).

$\alpha-\text{cut}$: Alpha-cut. ($\alpha-\text{cut}^*$ for optimal)

$\kappa_{\max}$: vagueness coefficient (max). ($\kappa_{\max}{}^*$ for optimal)

$\kappa_{\min}$: vagueness coefficient (min) ($\kappa_{\min}{}^*$ for optimal)

Cost analysis of integrated model

The process begins with an in-control state with a Process Failure Mechanism that follows a Fuzzy Weibull Distribution. Denote

$$\underset{\sim}{f}(t) = \lambda^{\nu}\nu\left(\frac{t}{\underset{\sim}{\kappa}}\right)^{\nu-1} e^{-\left(\frac{\lambda t}{\underset{\sim}{\kappa}}\right)^{\nu}} \quad \text{where } \lambda,\nu,t \ge 0 \text{ , } \kappa > 0 \quad (4)$$

$\alpha-$cut of fuzzy distribution function

$$f_{\alpha}(t) = \quad [\, f_{L\alpha}(t), f_{U\alpha}(t)]$$

$$= \begin{bmatrix} \lambda^{\nu}\nu\left(\dfrac{t}{(1-\kappa_{\min})\alpha+\kappa_{\min}}\right)^{\nu-1} e^{-\left(\frac{\lambda t}{(1-\kappa_{\min})\alpha+\kappa_{\min}}\right)^{\nu}} \\ ,\lambda^{\nu}\nu\left(\dfrac{t}{(1-\kappa_{\max})\alpha+\kappa_{\max}}\right)^{\nu-1} e^{-\left(\frac{\lambda t}{(1-\kappa_{\max})\alpha+\kappa_{\max}}\right)^{\nu}} \end{bmatrix} \quad (5)$$

The Fuzzy Weibull Distribution cumulative function has the form

$$\underset{\sim}{F}(t) = 1-e^{-[\lambda t/(\underset{\sim}{\kappa})]^{\nu}} \qquad \text{where} \quad \lambda,\nu,t \ge 0 \text{ , } \kappa > 0 \quad (6)$$

$\alpha-$cut of fuzzy distribution cumulative function

$$F_{\alpha}(t) = [F_{L\alpha}(t), F_{U\alpha}(t)]$$

$$= \begin{bmatrix} 1-e^{-[\lambda t/[(1-\kappa_{\max})\alpha+\kappa_{\max}]]^{\nu}} \\ , 1-e^{-[\lambda t/[(1-\kappa_{\min})\alpha+\kappa_{\min}]]^{\nu}} \end{bmatrix} \quad (7)$$

Scenario 1 (S1) the process in-control alert signal

The expected cycle time of EWMA control chart for scenario 1.

In S1, the process begins with a "in-control" state and inspections occur after h hours of monitoring as to whether the process has shifted from an "in-control" to an "out-of-control" state. And there is an alert signal in the control chart before the scheduled time when maintenance should be performed. But the signal is false, that is to say, the process is still "in-control". Since searching and determining false signal take time and incur cost, Compensatory Maintenance is performed. (See Fig. 3)

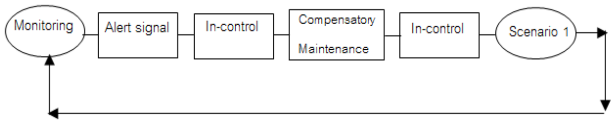


Figure 3: Integrated model of Scenario 1

$$E[T|S_1]_{\alpha} = \left[\, E[T|S_1]_{L\alpha}, E[T|S_1]_{U\alpha} \right]$$

$$= \begin{bmatrix} h\sum_{i=1}^{k} ip_i^I(1-F_{L\alpha}(ih)) + T_0 + T_C \\ , h\sum_{i=1}^{k} ip_i^I(1-F_{U\alpha}(ih)) + T_0 + T_C \end{bmatrix} \quad (8)$$

$$E[C|S_1]_{\alpha} = \left[\, E[C|S_1]_{L\alpha} \text{ , } E[C|S_1]_{U\alpha} \right]$$

$$= \begin{bmatrix} C_I[h\sum_{i=0}^{k} ip_i^I(1-F_{L\alpha}(ih))+\gamma_C T_C] \\ + (C_F+nC_V)\sum_{i=0}^{k} ip_i^I(1-F_{L\alpha}(ih))+C_f+C_C \\ ,C_I[h\sum_{i=0}^{k} ip_i^I(1-F_{U\alpha}(ih))+\gamma_C T_C] \\ +(C_F+nC_V)\sum_{i=0}^{k} ip_i^I(1-F_{U\alpha}(ih))+C_f+C_C \end{bmatrix} \quad (9)$$

Scenario 2 (S2) consider the process out-of-control alert signal

The expected cycle time of EWMA control chart for scenario 2.

In S2, It assumes that the process shifts to an "out-of-control" state prior to the Planned Maintenance and process failure mechanism follows a Fuzzy Weibull distribution, the in-control time follows a truncated Fuzzy Weibull distribution. (See Fig. 4)

$$f_\alpha(t|(k+1)h) = \frac{f_\alpha(t)}{F_\alpha((k+1)h)}$$

$$= \frac{\lambda^v v\left(\frac{t}{\underset{\sim}{\kappa}}\right)^{v-1} e^{-\left(\frac{\lambda t}{\underset{\sim}{\kappa}}\right)^v}}{1-e^{-(\lambda\frac{(k+1)h}{\underset{\sim}{\kappa}})^v}}; \; 0 \le t \le (k+1)h \tag{10}$$

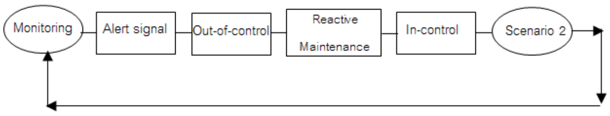


Figure 4: Integrated model of Scenario 2

$$E[T|S_2]_\alpha = \left[E[T|S_2]_{L\alpha}\ ,\ E[T|S_2]_{U\alpha}\right]$$

$$= \begin{bmatrix} \int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{L\alpha}(t|(k+1)h)dt + hARL_O - \tau + nE + T_A + T_R \\ ,\ \int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{U\alpha}(t|(k+1)h)dt + hARL_O - \tau + nE + T_A + T_R \end{bmatrix} \tag{11}$$

$$E[C|S_2]_\alpha = \left[E[C|S_2]_{L\alpha}\ ,\ E[C|S_2]_{U\alpha}\right]$$

$$= \begin{bmatrix} C_I \left[\int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{L\alpha}(t|(k+1)h)dt\right] \\ +C_O\left[hARL_O - \tau + nE + \gamma_A T_A + \gamma_R T_R\right] \\ +\frac{1}{h}E[T|S_2](C_F + nC_V) + C_R \\ ,\ C_I \left[\int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{U\alpha}(t|(k+1)h)dt\right] \\ +C_O\left[hARL_O - \tau + nE + \gamma_A T_A + \gamma_R T_R\right] \\ +\frac{1}{h}E[T|S_2](C_F + nC_V) + C_R \end{bmatrix} \tag{12}$$

Scenario 3 (S3) consider the process in-control chart no signal

The expected cycle time of EWMA control chart for scenario 3. In S3, No signal occurs in the control chart before the scheduled time. Then, at the $(k+1)^{\text{th}}$ sampling interval, appropriate maintenance should be arranged. In S3, the process is always "in-control", Planned Maintenance is performed. (See Fig. 5)

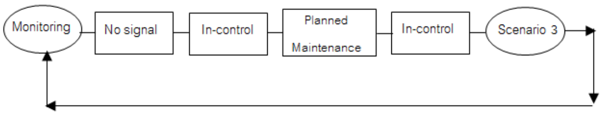


Figure 5: Integrated model of Scenario 3

$$E[T|S_3]_\alpha = \left[E[T|S_3]_{L\alpha}\ ,\ E[T|S_3]_{U\alpha}\right]$$
$$= \left[(k+1)h + T_P\ ,\ (k+1)h + T_P\right] \tag{13}$$

$$E[C|S_3]_\alpha = \left[E[C|S_3]_{L\alpha}\ ,\ E[C|S_3]_{U\alpha}\right]$$

$$= \begin{bmatrix} C_I\left[(k+1)h + \gamma_P T_P\right] + k(C_F + nC_V) + C_P \\ ,\ C_I\left[(k+1)h + \gamma_P T_P\right] + k(C_F + nC_V) + C_P \end{bmatrix} \tag{14}$$

Scenario 4 (S4) consider the process out-of-control chart no signal

The expected cycle time of EWMA control chart for scenario 4.

In S4, The process begins in control. When the process shifts to an "out-of-control" state, Reactive Maintenance takes place because the "out-of-control" condition occurred before the scheduled time, and additional time and expense will be incurred to identify and solve the equipment problem. (See Fig.6)

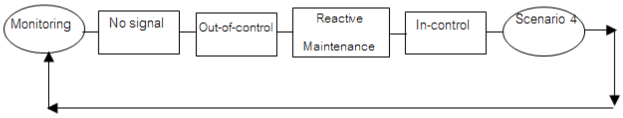


Figure 6: Integrated model of Scenario 4

$$E[T|S_4]_\alpha = \left[E[T|S_4]_{L\alpha}\ ,E[T|S_4]_{U\alpha}\right]$$
$$= \left[(k+1)h + T_R\ ,\ (k+1)h + T_R\right] \tag{15}$$

$$E[C|S_4]_\alpha = \left[E[C|S_4]_{L\alpha}\ ,\ E[C|S_4]_{U\alpha}\right]$$

$$= \begin{bmatrix} C_I\left[\int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{L\alpha}(t|(k+1)h)dt\right] \\ +C_O\left[(k+1)h - \int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{L\alpha}(t|(k+1)h)dt + \gamma_R T_R\right] + k(C_F + nC_V) + C_R \\ ,\ C_I\left[\int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{U\alpha}(t|(k+1)h)dt\right] \\ +C_O\left[(k+1)h - \int_0^{kh} \frac{t}{\underset{\sim}{\kappa}} f_{U\alpha}(t|(k+1)h)dt + \gamma_R T_R\right] + k(C_F + nC_V) + C_R \end{bmatrix}$$

(16)

Expected Hourly Cost $E[H]$

The model can be considered as a renewal-reward process; hence, the expected cost per hour $E[H]$ can be expressed as

$$E[H] = \frac{E[C]}{E[T]} \tag{17}$$

where

$$E[T] = E[T|S_1]P(S_1) + E[T|S_2]P(S_2) + E[T|S_3]P(S_3) + E[T|S_4]P(S_4) \tag{18}$$

$$E[C] = E[C|S_1]P(S_1) + E[C|S_2]P(S_2) + E[C|S_3]P(S_3) + E[C|S_4]P(S_4) \tag{19}$$

and

Probability of scenario 1

$$P_\alpha(S_1) = \sum_{i=1}^{k} P(\text{In-control} \cap \text{Alert Signal})$$

$$= \sum_{i=1}^{k} P(\text{In-control}|\text{Alert Signal})P(\text{Alert Signal}) \tag{20}$$

$$= \sum_{i=1}^{k} P_i^I(1 - F_{L\alpha}(ih))\ ,\ \sum_{i=1}^{k} P_i^I(1 - F_{U\alpha}(ih))$$

Probability of scenario 2

$$P_\alpha(S_2) = \sum_{i=1}^{k} P(\text{Out-of-control} \cap \text{Alert Signal})$$

$$= \sum_{i=1}^{k} P(\text{Out-of-control} \mid \text{Alert Signal})\, P(\text{Alert Signal}) \quad (21)$$

$$= \begin{bmatrix} \sum_{i=1}^{k} [F_{L\alpha}(ih) - F_{L\alpha}(i-1)h](1 - \sum_{j=1}^{i-1} P_j^I) \sum_{l=1}^{k-i+1} P_l^O \\ , \sum_{i=1}^{k} [F_{U\alpha}(ih) - F_{U\alpha}(i-1)h](1 - \sum_{j=1}^{i-1} P_j^I) \sum_{l=1}^{k-i+1} P_l^O \end{bmatrix}$$

Probability of scenario 3

$$P_\alpha(S_3) = \sum_{i=1}^{k} P(\text{In-control} \cap \text{No Signal})$$

$$= \sum_{i=1}^{k} P(\text{In-control} \mid \text{No Signal})\, P(\text{No Signal}) \quad (22)$$

$$= \begin{bmatrix} (1 - F_\alpha(kh)) - \sum_{i=1}^{k} P_i^I (1 - F_\alpha(ih)) \\ , (1 - F_\alpha(kh)) - \sum_{i=1}^{k} P_i^I (1 - F_\alpha(ih)) \end{bmatrix}$$

Probability of scenario 4

$$P_\alpha(S_4) = \sum_{i=1}^{k} P(\text{Out-of-control} \cap \text{No Signal})$$

$$= \sum_{i=1}^{k} P(\text{Out-of-control} \mid \text{No Signal})\, P(\text{No Signal})$$

$$= \begin{bmatrix} F_{L\alpha}(kh) - \sum_{i=1}^{k} [F_{L\alpha}(ih) - F_{L\alpha}(i-1)h] \\ (1 - \sum_{j=1}^{i-1} P_j^I) \sum_{l=1}^{k-i+1} P_l^O \\ , F_{U\alpha}(kh) - \sum_{i=1}^{k} [F_{U\alpha}(ih) - F_{U\alpha}(i-1)h] \\ (1 - \sum_{j=1}^{i-1} P_j^I) \sum_{l=1}^{k-i+1} P_l^O \end{bmatrix} \quad (23)$$

The economic design of integrated model of EWMA chart is to determine the optimal values of the six test variables $(n, h, k, \kappa_{\min}, \kappa_{\max}, \alpha - \text{cut})$ such that the expected total cost per hour in Equation (17) is minimized.

From examination of the components in Equation (18) and (19), it can be seen that determining the economically optimal values of the six test variables for the EWMA chart is not straightforward. To illustrate the nature of the solutions obtained from economic design of EWMA chart, a particular numerical example is provided.

3. Research Results and Discussion

The solution procedure is carried out using genetic algorithms (GA) with MATLAB 7.6.0(R2008a) software to obtain the optimal values of $n, h, k, \kappa_{\min}, \kappa_{\max}$ and $\alpha - \text{cut}$ that minimize $(E[H])$.

The GA, based on the concept of natural genetics, is directed toward a random optimization search technique. The GA solves problems using the approach inspired by the process of Darwinian evolution. The current GA in science and engineering refers to the models introduced and investigated by (Holland [16]). In the GA, the solution of a problem is called a ‘‘chromosome’’. A chromosome is composed of genes (i.e., features or characters). Although there are several kinds of numerical optimization methods, such as neural network, gradient-based search, GA, etc., the GA has advantages in the following aspects:

1. The operation of GA uses the fitness function values and the stochastic way (not deterministic rule) to guide the search direction of finding the optimal solution. Therefore the GA can be applied for many kinds of optimization problems.
2. The GA can lead to a global optimum by mutation and crossover technique to avoid trapping in the local optimum.
3. The GA is able to search for many possible solutions (or chromosomes) at the same time. Hence, it can obtain the global optimal solution efficiently. Based on these points, GA is considered as an appropriate technique for solving the problems of combinatorial optimization and has been successfully applied in many areas to solve optimization problems ([17-19]) The solution procedure for our example using the GA by MATLAB is briefly described as follows:

Step1. Initialization

One hundred initial solutions that satisfy the constraint condition of each test variable are randomly produced. Meanwhile, the constraint condition for each test variable is set as follows:

$0.1 \le h \le 5$

$5 \le n \le 15$

$20 \le k \le 40$

$0 \le \kappa_{\min} \le 1$

$1 \le \kappa_{\max} \le 15$

$0 \le \alpha - \text{cut} \le 1$

Step2. Evaluation

The fitness of each solution is evaluated by calculating the value of fitness function. The fitness function for our example is the cost function shown in Equation (14).

Step3. Selection

The survivors (i.e., 30 solutions) are selected for the next generation according to the better fitness of chromosomes. (In the first generation, the chromosome with the highest cost is replaced by the chromosome with the lowest cost.)

Step4. Crossover

A pairs of survivors (from the 30 solutions) are selected randomly as the parents used for crossover operations to produce new chromosomes (or children) for the next generation. In this example, the arithmetical crossover method with crossover rate 0.8 is applied as follows:

$$D_1 = 0.8R + 0.2M$$
$$D_2 = 0.2R + 0.8M$$

where D_1 is the first new chromosome, D_2 is the second new chromosome, R and M are the parents chromosomes. If 30 parents are randomly selected, then there are 60 children that will be produced. Thus, the population size increases to 90 (i.e., 30 parents + 60 children) in this step.

Step5. Mutation

Suppose that the mutation rate is 0.1. In this example, we use non-uniform method to carry out the mutation operation. Since we have 90 solutions, we can randomly select nine chromosomes (i.e., 90*0.1=9) to mutate some parameters (or genes)

Step6. Repeat Step 2 to Step 5 until the stopping criteria is found. In this example, we use until cannot be changed value as our stopping criteria.

In this section, numerical example and sensitivity analysis are conducted to study the effect of model parameters on the solution of economic design of the EWMA chart. The sensitivity analysis is carried out using orthogonal-array experimental design and multiple regression, in which the model parameters are considered as the independent variables and the six test parameters (i.e., $n, h, k, \kappa_{\min}, \kappa_{\max}, \alpha-\text{cut}$), as well as the average total hourly cost $E[H]$, are treated as the dependent variables.

Eight independent parameter (i.e., the model parameters) considered in the sensitivity analysis and their corresponding level planning are shown in Table 1

Table 1: Eight model parameters and their level planning

Model Parameters	Level 1	Level 2
C_I	10	20
C_O	200	400
C_P	3000	6000
C_R	2000	4000
C_C	1000	2000
C_F	10	20
C_V	0.1	0.2
C_f	100	200

The experimental design by L16 orthogonal array finds the eight independent parameters costs per hour for the Lower $E[H]$ and for the Upper $E[H]$ as shown in Table 3 and Table 6 respectively. In the L16 orthogonal array experiment design, there are 16 trials (i.e., 16 different level combinations of the independent variables). For each trial, the Genetic Algorithm is applied to produce the optimal solution of the economic design, with fixed parameters as shown in Table 2:

Table 2: Fix Parameter

$\gamma_1 = \gamma_2 = \gamma_3 = 1$	$\lambda = 0.05$	$L = 3$	$v = 2$
$T_0 = 0.2$	$r = 0.2$	$T_A = 0.3$	$T_C = 0.6$
$T_R = 1$	$T_P = 0.8$		

Table 3: Model parameter assignment in the L16 orthogonal array and the corresponding solution (Lower)

Trial	Model Parameter							
	C_I	C_O	C_P	C_R	C_C	C_F	C_V	C_f
1	10	200	3000	2000	1000	10	0.1	100
2	20	200	3000	2000	1000	20	0.2	200
3	10	400	3000	2000	2000	10	0.2	200
4	20	400	3000	2000	2000	20	0.1	100
5	20	200	6000	2000	2000	20	0.2	100
6	10	200	6000	2000	2000	10	0.1	200
7	20	400	6000	2000	1000	20	0.1	200
8	10	400	6000	2000	1000	10	0.2	100
9	20	200	3000	4000	2000	20	0.1	200
10	10	200	3000	4000	2000	10	0.2	100
11	20	400	3000	4000	1000	20	0.2	100
12	10	400	3000	4000	1000	10	0.1	200
13	10	200	6000	4000	1000	10	0.2	200
14	20	200	6000	4000	1000	20	0.1	100
15	10	400	6000	4000	2000	10	0.1	100
16	20	400	6000	4000	2000	20	0.2	200

Trial	Result						
	h	n	k	$\alpha-\text{cut}$	$\kappa_{\max}$	$\kappa_{\min}$	$E[H]_L$
1	3.04	11.58	23.89	0.015	20.12	0.41	206.395
2	3.957	14.34	22.89	0.003	23.07	0.759	208.849
3	3.196	5.357	20.12	0.021	19.28	0.549	406.005
4	3.042	5.795	20.06	0.205	22.63	0.1	409.186
5	3.653	6.14	24.94	0.03	21.94	0.663	209.861
6	3.137	13.45	23.05	0.007	20.26	0.536	206.908
7	2.839	6.429	23.43	0.014	22.21	0.585	408.986
8	2.507	6.404	20.00	0.003	15.08	0.369	406.350
9	3.586	14.23	25.50	0.034	20.50	0.364	212.363
10	3.217	14.19	25.23	0.019	19.20	0.042	201.796
11	2.912	5.89	27.09	0.015	22.17	0.632	412.095
12	2.405	5.895	26.28	0.028	22.93	0.619	408.396
13	3.222	5.618	25.27	0.003	22.29	0.504	208.085
14	3.586	5.232	25.02	0.003	22.18	0.349	210.588
15	2.23	5.75	32.14	0.002	20.75	0.195	411.501

16	3.04	11.59	23.89	0.015	20.11	0.413	412.791

Table 4: Optimal value for model parameter ($E[H]_L$)

Parameter	Value	Parameter	Value
C_I	10	C_C	2000
C_O	200	C_F	10
C_P	3000	C_V	0.2
C_R	4000	C_f	100

Table 5: Optimal value for six variable and optimal value total hourly costs (Lower)

Parameter	Value	Parameter	Value
h^*	3.217	$\kappa^*_{\max}$	19.20
n^*	14.19	$\kappa^*_{\min}$	0.042
k^*	25.23	$E[H]^*_{\text{Lower}}$	201.796
$\alpha-\text{cut}^*$	0.019		

Table 6: Model parameter assignment in the L16 orthogonal array and the corresponding solution (Upper)

Trial	Model Parameter						
	C_I	C_O	C_P	C_R	C_C	C_F	C_V	C_f
1	10	200	3000	2000	1000	10	0.1	100
2	20	200	3000	2000	1000	20	0.2	200
3	10	400	3000	2000	2000	10	0.2	200
4	20	400	3000	2000	2000	20	0.1	100
5	20	200	6000	2000	2000	20	0.2	100
6	10	200	6000	2000	2000	10	0.1	200
7	20	400	6000	2000	1000	20	0.1	200
8	10	400	6000	2000	1000	10	0.2	100
9	20	200	3000	4000	2000	20	0.1	200
10	10	200	3000	4000	2000	10	0.2	100
11	20	400	3000	4000	1000	20	0.2	100
12	10	400	3000	4000	1000	10	0.1	200
13	10	200	6000	4000	1000	10	0.2	200
14	20	200	6000	4000	1000	20	0.1	100
15	10	400	6000	4000	2000	10	0.1	100
16	20	400	6000	4000	2000	20	0.2	200

Trial	Result						
	h	n	k	$\alpha-\text{cut}$	$\kappa_{\max}$	$\kappa_{\min}$	$E[H]_U$
1	0.537	5	20	0.862	2.5	0.8	350.033
2	4.088	6.062	20	0.136	22.65	0.25	211.736
3	0.416	5	20	0.862	1	1	545.546
4	0.424	5	20	0.862	1	1	568.596
5	4.438	6.19	38.18	0.002	2.094	0.04	229.992
6	0.7	5.143	21	0.998	1.008	0.941	426.011
7	0.537	5	20	0.862	1	1	658.482
8	0.536	5	20	0.863	1	1	644.319
9	0.588	5.037	20	0.999	1.016	0.549	403.022
10	0.546	5	20	0.862	1	1	390.992
11	0.412	5	20	0.862	1	1	553.218
12	0.412	5	20	0.862	1	1	538.633
13	0.665	5.196	21	0.9	1.014	0.66	430.891
14	0.699	5.142	21	0.998	1.007	0.940	446.565
15	0.537	5	20	0.862	1	1	675.821
16	0.537	5	20	0.862	1	1	700.067

Table 7: Optimal value for model parameter ($E[H]$ Upper)

Parameter	Value	Parameter	Value
C_I	20	C_C	1000
C_O	200	C_F	20
C_P	3000	C_V	0.2
C_R	2000	C_f	200

Table 8: Optimal value for six variable and optimal value total hourly costs (Upper)

Parameter	Value	Parameter	Value
h^*	4.088	$\kappa^*_{\max}$	22.65
n^*	6.062	$\kappa^*_{\min}$	0.25
k^*	20.00	$E[H]^*_{\text{Lower}}$	211.736
$\alpha-\text{cut}^*$	0.136		

Data Analysis

Consider Lower α-cut value

The output of the GA for each trial is also recorded in Table 3.

To study the effect of model parameters on the solution of economic design of EWMA chart, based on the data in Table 3, the statistical software SPSS 15.0 is used to run the regression analysis for each dependent variable. For each dependent variable, the output of SPSS includes an ANOVA table for regression and a table of regression coefficients, showing the corresponding information about statistical hypothesis testing.

Demonstrating in Table 9 is the SPSS output for the interval between sampling $(\hat{h})$. Considered the ANOVA in Table 9(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of interval between sampling $(\hat{h})$.

By examining Table 9(b), It is show that the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) , the fixed cost of sampling (C_F) and the variable cost of sampling (C_V) significantly affect the value of interval between sampling $(\hat{h})$. It is noticed that the sign of the coefficients of the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the variable cost of sampling (C_V) are negative, indicating that the higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the variable cost of sampling (C_V) generally reduces interval between sampling $(\hat{h})$. And the sign of the fixed cost of sampling (C_F) is positive, indicating that the fixed cost of sampling (C_F) generally increases interval between sampling $(\hat{h})$, which is consistent with the principle of statistical hypothesis testing.

Table 9: SPSS output for the interval between sampling $(\hat{h})$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	2.757	3	0.919	22.022	0.00(c)
Residual	0.501	12	0.42		
Total	3.257	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	3.047	0.270	11.276^{**}	0.00
C_O	-0.003	0.001	-6.397^{**}	0.00
C_F	0.046	0.010	4.481^{**}	0.001
C_V	-0.005	1.021	2.251^{*}	0.044

Table 10 is the SPSS output for the sampling size $(\hat{n})$. Considered the ANOVA in Table 10(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of the sampling size $(\hat{n})$.

By examining Table 10(b), It is show that the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) significantly affect the value of the sampling size $(\hat{n})$. It is noticed that the sign of the coefficients of the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) is negative, indicating that the higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) generally reduces the sampling size $(\hat{n})$, which is consistent with the principle of statistical hypothesis testing.

Table 10: SPSS output for the sampling size $(\hat{n})$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	62.687	1	62.687	5.791	0.030(a)
Residual	151.535	14	10.824		
Total	214.222	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	14.556	2.601	5.597^{**}	0.00
C_O	-0.02	0.008	-2.407^{**}	0.030

Table 11 is the SPSS output for the sampling size $(\hat{n})$. Considered the ANOVA in Table 11(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of the number of samples taken before Planned Maintenance $(\hat{k})$.

By examining Table 11(b), It is show that the cost of performing Reactive Maintenance (C_R) significantly affect the value of the value of the number of samples taken before Planned Maintenance $(\hat{k})$. It is noticed that the sign of the coefficients of the cost of performing Reactive Maintenance (C_R) is positive, indicating that the higher cost of performing Reactive Maintenance (C_R) generally increases the number of samples taken before Planned Maintenance $(\hat{k})$, which is consistent with the principle of statistical hypothesis testing.

Table 11: SPSS output for the number of samples taken before Planned Maintenance $(\hat{k})$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	64.610	1	64.610	12.516	0.003(a)
Residual	71.768	14	5.126		
Total	135.928	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	18.292	1.790	10.220^{**}	0.00
C_R	0.002	0.001	3.538^{**}	0.003

Table 12 is the SPSS output for the Hourly Cost $(E[\hat{H}])$. Considered the ANOVA in Table 12(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of Hourly Cost $(E[\hat{H}])$.

By examining Table 12(b), It is show that the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the fixed cost of sampling (C_F) significantly affect the value of Hourly Cost $(E[\hat{H}])$. It is noticed that the sign of the coefficients of the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the fixed cost of sampling (C_F) are positive, indicating that the higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the fixed cost of sampling (C_F) generally increases Hourly Cost $(E[\hat{H}])$, which is consistent with the principle of statistical hypothesis testing.

Table 12: SPSS output for the Hourly Cost $(E[\hat{H}])$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	162153.4	2	81076.7	16538.828	0.00(b)
Residual	63.729	13	4.902		
Total	162217.1	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	1.307	2.413	0.542	0.597
C_O	1.007	0.006	181.843^{**}	0.00
C_F	0.366	0.111	3.306^{**}	0.06

Consider Upper α-cut value

The output of the GA for each trial is also recorded in Table 6.

To study the effect of model parameters on the solution of economic design of EWMA chart, based on the data in Table 6, the statistical software SPSS 15.0 is used to run the regression analysis for each dependent variable. For each dependent variable, the output of SPSS includes an ANOVA table for regression and a table of regression coefficients, showing the corresponding information about statistical hypothesis testing.

Demonstrating in Table 13 is the SPSS output for the vagueness coefficient (min) $(\hat{\kappa}_{\min})$. Considered the ANOVA in Table 13(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of vagueness coefficient (min) $(\hat{\kappa}_{\min})$.

By examining Table 13(b), It is show that the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) significantly affect the value of vagueness coefficient (min) $(\hat{\kappa}_{\min})$. It is noticed that the sign of the coefficients of the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) is positive, indicating that the higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) generally increases vagueness coefficient (min) $(\hat{\kappa}_{\min})$, which is consistent with the principle of statistical hypothesis testing.

Table 13: SPSS output for the number of samples taken before Planned Maintenance $(\hat{k})$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	0.497	1	0.497	8.128	0.013(a)
Residual	0.856	14	0.061		
Total	1.353	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	0.295	0.195	1.509^{**}	0.154
C_R	0.002	0.001	2.851^{**}	0.013

Table 14 is the SPSS output for the Hourly Cost $(E[\hat{H}])$. Considered the ANOVA in Table 14(a), if the significance level is set to be 0.05, we observe that at least one model parameters significantly affect the value of Hourly Cost $(E[\hat{H}])$.

By examining Table 14(b), It is show that the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the cost of performing Planned Maintenance (C_P) significantly affect the value of Hourly Cost $(E[\hat{H}])$. It is noticed that the sign of the coefficients of the cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the cost of performing Planned Maintenance (C_P) are positive, indicating that the higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O) and the cost of performing Planned Maintenance (C_P) generally increases Hourly Cost $(E[\hat{H}])$, which is consistent with the principle of statistical hypothesis testing.

Table 14: SPSS output for the Hourly Cost $(E[\hat{H}])$.

(a) ANOVA Table

Model	Sum of square	df	Mean Square	F	P-Value
Regression	275297.7	2	137648.89	28.686	0.00(b)
Residual	62380.4	13	4798.495		
Total	337678.2	15			

(b) Table of regression coefficients

Independent Variable	Coefficient	Std. error	t	P-Value
(Constant)	-10.220	75.487	-0.135	0.894

C_O	1.247	0.173	7.202^{**}	0.00
C_P	0.027	0.012	2.347^{**}	0.035

4. Conclusion

An integrated model which takes advantage of two traditions but separately uses manufacturing process control tools (statistical process control and maintenance management) was proposed. In the present paper, we developed the integrated economic for Fuzzy Weibull distribution design of EWMA control chart to determine the values of six tested variables of the chart (i.e., the sampling interval (h), the sample size (n), the number of samples taken before Planned Maintenance (k), the Alpha-cut $(\alpha-\text{cut})$, the minimum vagueness coefficient $(\kappa_{\min})$ and the maximum vagueness coefficient $(\kappa_{\max})$ such that the expected total cost per hour $(E[H]_{\text{Lower}}$, $E[H]_{\text{Upper}})$ would be minimized. The cost function is based on the cost model described in Linderman and Kathleen [12]. An illustrative example was provided and the Genetic Algorithm was employed to search for the solution of the economic design. A sensitivity analysis was then carried out to study the effect of model parameters on the solution of the economic design. Based on the sensitivity analysis, the following results are observed:

Consider Lower α-cut value

1. A higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O). The result show that the interval between sampling (h) is decreasing.
2. A higher fixed cost of sampling (C_F). The result show that the interval between sampling (h) is increasing.
3. A higher variable cost of sampling (C_V). The result show that the interval between sampling (h) is decreasing.
4. A higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O). The result show that sampling size (n) is decreasing.
5. A higher cost of performing Reactive Maintenance (C_R). The result show that the number of samples taken before Planned Maintenance (k) is increasing.
6. A higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O). The result show that the Hourly Cost $(E[H])$ is increasing.
7. A higher fixed cost of sampling (C_F). The result show that the Hourly Cost $(E[H])$ is increasing.

Consider Upper α-cut value

1. A higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O). The result show that the vagueness coefficient (min) $(\kappa_{\min})$ is increasing.
2. A higher cost of quality loss per unit time (the process is in an out-of-control state) often estimated by a Taguchi Loss function (C_O). The result show that the Hourly Cost $(E[H])$ is increasing.
3. A higher cost of performing Planned Maintenance (C_P). The result show that the Hourly Cost $(E[H])$ is increasing.

Acknowledgements

The authors would like to thank the reviewers for their useful suggestions and remarks, which helped in improving the manuscript. This work is supported by Department Mathematics KMUTT

References

[1] Russell B. Vagueness. Australasian J Psychol Philos. 1923; 1: 84–92.

[2] Zadeh LA. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybernet. 1973; 3: 28–44.

[3] Lodwick WA. Jamison KD. Interval-valued probability in the ananlysis of problems containing a mixture of possibilistic, probabilistic and interval uncertainty. Fuzzy Set Syst. 2008; 159: 2845–2858.

[4] Zadeh LA. From imprecise to granular probabilities. Fuzzy Set Syst. 2005; 154: 370–374.

[5] Zimmermann H-J. An application-oriented view of modelling uncertainty. Eur J Oper Res. 2000; 122: 190–199.

[6] Zimmermann H-J. Testability and meaning of mathematical models in social sciences. Math Modelling . 1980; 1: 123–139.

[7] Stojakovic M. Fuzzy Random Variable in Mathematical Economics. Novi Sad J. Math. 2005; 35(1): 103-112.

[8] Bushra Hussien Aliwi, Kawther Fawzi Hamza. Numerical Comparison Function for Weibull Distribution Probability Values with Possibility Values for Fuzzy Logic. Education Collage (Ibn Hayan). Department of Mathematics University of Babylon, Iraq 2010.

[9] Abbas Pak, Gholam Ali Parham, Mansour Saraj. Inference for the Weibull Distribution Based on Fuzzy Data. Revista Colombiana de Estadística. 2013; 36(2): 339-358.

[10] Baloui Jamkhaneh E. Analyzing System Reliability Using Fuzzy Weibull Lifetime Distribution. International Journal of Applied Operational Research. 2014; 4(1): 93-102.

[11] Shahryar Sorooshian. Fuzzy Approach to Statistical Control Charts. Hindawi Publishing Corporation Journal of Applied Mathematics. 2013; 2013

[12] Linderman, K, McKone-Sweet J.C. An integrated systems approach to process control and maintenance. European Journal of Operational Research. 2005 ; 164: 324–340.

[13] Wen-Hui Zhou, Gui-Long Zhu. Economic design of integrated model of control chart and maintenance. Mathematical and Computer Modeling. 2008; 47: 1389–1395.

[14] Karpisek Z, Stepanek P, and Jurak P. Weibull Fuzzy Probability Distribution for Reliability of concrete structure. Engineering MECHANICS. 2010; 17: 363–372.

[15] Lorenzen T.J, Vance L.C. The economic design of control charts: a unified approach. Technometrics. 1986; 28: 3-10.

[16] Holland J.H. Adaptation in Natural and Artificial Systems. Univ. Michigan Press: Ann Arbor; 1975.

[17] Jensen M.T. Generating robust and flexible job shop schedules using genetic algorithm. IEEE Transactions on Evolutionary Computation. 2003; 7: 275–288.

[18] Chou C.Y., Chen C.H. and Chen C.H. Economic design of variable sampling intervals T^2 control charts using genetic algorithms. Expert Systems with Applications. 2006; 30: 233–242.

[19] Chou C.Y., Wu C. and Chen C.H. Joint economic design of variable sampling intervals X-bar and R charts using genetic algorithms, Communications in Statistics. Simulation and Computation. 2006; 35(4): in press.

Pricing multiname credit derivatives by multicorrelated market factor model

Supalak Phetcharat
Department of Banking and Finance, Chulalongkorn University, Bangkok, 10330, Thailand,
Supalak.P@student.chula.ac.th

Abstract

Intensity-based Models with dependent market factors are used to produce correlation implied by multi-name credit products such as Collateralized Debt Obligations (CDOs). The market factors are modeled as multivariate jump-diffusion distributions that have continuous or drift-diffusion parts and jump parts driven by Gamma-Poisson mixture processes. The Gamma-Poisson mixture processes have capacity to capture tail dependence and rare events. As a result, the model produces strong default dependency enough to be calibrated to the CDO tranches in financial crisis situation. For single-name asset of a credit portfolio, market factor loadings are used to fit its Credit Default Swap (CDS) curves. The challenge thing about pricing CDOs is to find the solutions of the portfolio loss distribution. The problem becomes more complicated when factors of the model are not independent. We suggest optional methods, a recursive method and a Mimicking Markov Chain method, which can be efficiently used with both homogeneous and heterogeneous CDOs without time discretization. Numerical results show that a drift-diffusion part of market factors are important for generating correlated defaults under normal situation. The correlation parameters also help to fit the equity and the junior mezzanine without affecting other tranches. However, during financial crisis the jump part plays a prominent role. The estimated parameters hint that the event of jump are rare but significant.

Keywords: Credit derivatives, correlated defaults, market risk, tail dependence, intensity-based model

Corresponding Author
E-mail Address: Supalak.P@student.chula.ac.th

1. Introduction

Credit derivatives are financial instruments used by market participants such as banks and hedge funds for risk management and trading of credit risk. Derivatives can be distinguished by the number of underlying credits being referenced. In single-name credit derivatives, the product relates to only one underlying entity. In multi-name credit derivatives, there are multiple credit references. The well-known single-name products and multi-name products are CDSs and CDOs respectively. CDOs have underlying assets such as bonds and CDSs. Each CDO tranche is classified with the different level of the portfolio loss. However the risk of a multi-name credit portfolio literally is not diversified, there are unobservable dependencies between defaults in a portfolio.

For pricing credit derivatives, it requires to model a loss distribution. One example of credit risk models is the standard Gaussian copula. The model default times are constructed by calibration of marginal default distributions and their default correlations. Although there are fast algorithms provided to compute loss distributions, the Gaussian copula is static and unable to fit different tranches with one correlation parameter. The model has no capacity to measure dependence of defaults.

Intensity-based models have been successfully used to pricing single-name credit derivatives. Nevertheless, the standard intensity models has an issue with weak default correlations for multi-name credit derivatives. There are numerous models that have been invented to defeat these problems. For example, Mortensen (2006) uses the affine-jump diffusion model. His model is accomplished enough to capture the default dependency for both homogenous and heterogeneous portfolios. Peng and Kou (2009) proposed the Conditional Survival (CS) model. They use cumulative intensities instead to generate joint default events. Their market factors are also allowed to have jumps in cumulative terms. However, the context of correlated market factors is not supported in here.

In our research, we develop the intensity-based multi-correlated-market-factor model that is an extension of Duffie and Garleanu (2001)'s notable multi-issuer default model to capture correlated defaults and rare events. We demonstrate our work to be able to calibrate the model to CDS spreads and CDO tranche spreads. In the Duffie and Garleanu (2001) model the firm defaults are determined by the first jumps of point processes with jump-diffusion intensities. The individual firm intensity consists of market factors and an idiosyncratic factor. In particular, each market factor independently responds to different types of default risk such regional risk and global risk, causing the dependence between default events. There are other literatures motivated by the model of Duffie and Garleanu (2001), for example, Mortensen (2006) and Peng and Kou (2009). Unlike any of them, our market factors are allowed to be correlated.

With positive correlations between market factors, it helps to increase the default dependencies in a portfolio. If market factors are negatively dependent, credits in a portfolio are more diversified. More importantly, the structure of the dependence between defaults in a multi-credit portfolio such as a CDO is described by correlations among market factors and the market factors' model. The sensitivity of a firm's default risk to each market factor can be measured by the magnitude of its market factor loadings. Hence, we use the individual firm's market factor loadings and idiosyncratic parameters to describe its term structure of the marginal default probabilities. The process of a market factor is formed of a continuous part and jump parts. For simplicity, the continuous part is assumed to follow the Ornstein Ublenbeck (OU) process that can be decomposed into a drift term and a Brownian-Motion-driven diffusion part. We also control model parameters to avoid the possibility of intensity rates being negative.

We also specify market factors to incorporate jumps. There are jump processes shared among market intensities with different jump sizes, which technically makes market factors correlated. We use a Gamma-Poisson mixture process to determine jump times, since this process can represent the contagious defaults or fat-tailed dependence. Specifically, any two events modeled by a Gamma-Poisson mixture process are time interdependent, which means that occurrence of an event triggers an increase of the probability of other events occurring.

Due to the complexity of the model, the distribution function of the portfolio loss cannot be written in explicit form. As another objective of this work, we show how to implement our model using two alternative methods: a recursive method and a Mimicking Markov Chain simulation method. Before that, we must find the explicit formula of our model particularly for the marginal distribution of single firm's default time. Owing to the fact that market factors are dependent, the individual firm's state cannot be described by the marginal default probability of each market factor separately. Instead, it must be done by given the path of a vector of market-factor processes in the system. We derive the Laplace transform of a vector-valued market-factor process associated with their integrals by applying multivariable Ito's lemma. The Laplace transform is used to solve for the survival probability of each firm and employ in the implementation mechanisms we have mentioned.

To calculate the portfolio loss, we first define the conditional loss distribution given the path of market processes and then estimate the unconditional ones. For this issue, we use the recursive method that is applied from Andersen et al. (2003). The conditional loss distribution given the path of a vector-valued market-factor process can be written in a recursive form. The portfolio (unconditional) loss distribution is found by taking the expectation of the conditional loss distribution. How difficult it is to solve for the unconditional loss distribution from the recursive form is based on the number of firms in the portfolio. For a large heterogeneous portfolio, it is hard to find a crystal-clear closed-form solution. Moreover, it requires high-precision computation. The computation of the portfolio loss has to involve with numerous mathematical operations on numbers in the large range that is from very small to very large number. Therefore, the number of firms in a portfolio should be limited. However, if a portfolio is homogeneous, the loss of a portfolio is distributed as binomial. We can derive the conditional mass function of the loss given the path of a vector of market-factor processes in a simple way. By taking the expectation, the unconditional one is done in a close form. That closed form is applied from the Laplace transform we derive.

The Mimicking Markov Chain simulation method is adopted from Giesecke et al. (2010). Giesecke et al. (2010) don't present how to use their method pricing correlated credit products such CDOs. But we choose to apply their method because it provides the time and the firm that defaults without any time discretization. For that reason, we apply their simulation scheme with our model to price the CDO tranches. To implement the model, we need to mimic our own Markov chain that represents the dependent structure of firms' state in a portfolio. Instead of using usual firms' intensities to determine default times themselves, the arrival times for the next default is determined by the current portfolio transition rate of the mimicking Markov chain. The portfolio mimicking-Markov transition rate is the sum of transition rates of all firms. If a firm defaults, its transition rate becomes zero. The individual firm's transition rate has the meaning of the expectation of its intensity conditioning on a vector of states of the firms in the portfolio. It is challenging to find the explicit solution for the conditional expectation of each market factor's intensity given the state vector of firms. Especially when market factors are not independent, For instance, one step is to solve for the probability that the vector of firms' states happens. We have to define that conditional probability given the path of a vector of correlated-market-factor processes and then take the expected value of that result. The solutions of the firm transition rates are done in really sophisticated nested form. If the number of firms exceeds 100, using this method is very time-consuming.

We solve exhausting computation of transition rates in closed forms by generating sample paths of market-factor intensities, then computing the conditional expectations given the market-factor paths, and averaging conditional expectations to get their expectation terms. In our research, the path of market-factor intensities and their integrals are simulated without any discretization. First, we generate a Gamma-random arrival rate for each jump process and use it to determine arrival times, where the jump sizes are drawn from the exponential distribution. Once we have a set of arrival times and sizes of jumps, we plug them into market-factor processes, treating them as constants. Then we generate samples from the distributions of market-factor intensities and their integrals exactly as if their distributions were normal.

It is difficult to model the default correlation in a multi-name portfolio since market data on default dependencies is unobservable. Fortunately, there is no much difference between the performance of the model fitting the CDO tranche spreads under the assumption of a homogenous and a heterogeneous portfolio. It means that the tranches of a CDO are sensitive to systematic risk. As mentioned before, the market parameters are estimated from the spreads on CDO tranches, whereas each firm's specific parameters such as market factor loadings are calibrated to its CDS curves. In addition, we present a calibration algorithm for the model parameter estimation.

2. Review of Existing Models

There are many existing models that have been proposed for credit risk pricing. Their main contribution is usually to produce fat-tailed, long-tailed loss distributions for correlated defaults, which cannot be explained by the standard Gaussian Copula model.

The processes that are generally used to drive default intensities are an affine-jump diffusions which is the combination between a continuous part and a jump part. The basic affine-jump process X with parameters $(k, \theta, \sigma, \mu, \ell)$ solves

$$dX(t) = k\big(\theta - X(t)\big)dt + \sigma\sqrt{X(t)}dW(t) + dJ(t),$$

where k is the speed of adjustment, θ is the long-term mean, σ is volatility, W is a Brownian motion, and J is the jump Poisson process that has μ as the mean of exponential-distributed jump sizes and ℓ as the jump arrival rate.

2.1 Duffie and Garleanu (2001)'s Multi-Issuer Default Model

They assume that a portfolio is homogeneous, that is all firms' intensity processes in the portfolio have the same model parameters. Consider a n-firm portfolio. There are S sectors which each firm particularly belongs to. The ith firm's intensity process λ_i is adapted to the filtration F generated by the firm default processes, idiosyncratic risk factors X_i, $1 \le i \le n$, sectorial risk factors $Y_{c(i)}$, $c(i) \in \{1, \dots, S\}$ and a global risk factor Z, where X_i, $Y_{c(i)}$ and Z are supposed to be independent affine-jump processes sharing the same parameters k, σ and μ, having different long-term mean $\theta_i, \theta_{c(i)}, \theta_z$ and jump arrival rate $\ell_i, \ell_{c(i)}, \ell_z$ respectively. Then λ^i is a basic affine-jump process with parameters $(k, \theta, \sigma, \mu, \ell)$, defined as

$$\lambda_i(t) = X_i(t) + \ Y_{c(i)}(t) + Z(t),$$

where $\theta = \ \theta_i + \theta_{c(i)} + \ \theta_z$ and $\ell = \ell_i + \ \ell_{c(i)} + \ell_z$.

2.2 Mortensen (2006)'s Multi-Name Intensity Model

He modifies Duffie and Garleanu (2001)'s work to handle heterogeneous portfolios. Let F denote the filtration generated by the firm default processes, idiosyncratic risk factors $X_i, 1 \le i \le n$, and a market risk factor $Y.$ X_i and a_iY are supposed to be independent affine-jump processes with parameters $\big(k, \theta_i, \sqrt{a_i}\sigma, a_i\mu, \ell\big)$ and $\big(k, a_i\theta_Y, \sqrt{a_i}\sigma, a_i\mu, \ell\big)$ respectively, where $\ell = \ell_i + \ell_y$.

The ith firm's intensity process λ_i takes the form

$$\lambda_i(t) = a_iY(t) + X_i(t),$$

where the parameter a^i refers to the sensitivity of firm i to the market factor Y. It is implied that λ_i is an affine-jump process with parameters $\big(k, a_i\theta_Y + \theta_i, \sqrt{a_i}\sigma, a_i\mu, \ell\big)$.

2.3 Peng and Kou (2009)'s Conditional Survival Model

Peng and Kou (2009) propose new Conditional Survival (CS) Model to capture default clustering. They found that the default clustering occurs across time and section and consider that Duffie and Garleanu (2001)'s the multi-issuer default model cannot produce strong default correlation. They illustrate that once the jump of market factor occurs, it just increases the probabilities that firms might default, but doesn't make several firms defaulted simultaneously. To solve that problem, their CS Model has dynamics in form of constituent cumulative intensities, an idiosyncratic factor and multi market factors. Market factors are allowed to be jump processes themselves, not being part of intensity processes.

Consider a portfolio of n firms, Λ_i is the cumulative intensity of firm i, whose default time in this case is defined as

$$\tau^i = inf\{t \ge 0 \colon \Lambda_i(t) \ge \epsilon_\mathrm{i}\}, 1 \le i \le n,$$

where ϵ_i are independent exponential random variables with mean 1. The ith firm's cumulative intensity process Λ_i is adapted to the filtration F generated by the firm default processes, cumulative market factors M_j, $1 \le j \le J$, cumulative idiosyncratic factors $X_i, 1 \le i \le n$, specified as

$$\Lambda_i(t) = \sum_{j=1}^{J} a_{i,j}M_j(t) + X_i(t), 1 \le i \le n, t \ge 0,$$

where the factor loading $a_{i,j}$ represents the sensitivity of firm i to market factor j.

Peng and Kou (2009) used Polya processes and integral of CIR processes to model market-factor cumulative intensities. For a Polya process M_j, it can be viewed as the Poisson process that has a Gamma random variable as an arrival rate of default. Peng and Kou (2009) state that the Polya process can generate strong cross-sectional correlation for financial crisis due to the property of positively correlated increments. When an event occurs, it triggers jumps in overall firms' cumulative intensities. In addition, they use integral of CIR processes $\int_0^t \lambda(s)ds$ to provide dynamic and describe dependency structure of defaults under normal situation.

3. Our Proposed Model

Suppose that there are n underlying firms in a portfolio and m market factors in a system. We represent τ^i as the time that ith firm defaults in the portfolio, which is determined by its intensity process λ_i. Denote $(\Omega, \mathcal{F}, P)$ as a complete probability space with a filtration $\mathrm{F} = (\mathcal{F}_t)_{t\geq 0}$ of its σ-algebra $\mathcal{F}$. P is a risk neutral measure. Additionally, the filtration F is generated by the firm default processes, market factors $X_j, 1 \leq j \leq m$, and idiosyncratic factors Y_i , $1 \leq i \leq n$. The ith firm's intensity process is specified as

$$\lambda_i(t) = \sum_{j=1}^{m} \beta_{i,j} X_j(t) + Y_i(t), 1 \leq i \leq n, \tag{1}$$

where $\beta_{i,j}$ is the market factor loading representing the sensitivity of ith firm to market factor $j, 1 \leq j \leq m$.
Let Z^j be jump processes, $1 \leq j \leq J$ and X_i be an Ornstein–Uhlenbeck (OU)-process of the ith market factor that has dynamics

$$\begin{aligned} dX_i(t) = k_i\big(\theta_i - X_i(t)\big)dt + \sigma_i dW_i(t) \\ + \sum_{j=1}^{J} \zeta^{i,j} dZ^j(t), \end{aligned} \tag{2}$$

where k_i is the rate of mean-reversion, θ_i is the long-term mean, σ_i is the volatility, and $\zeta^{i,j}$ controls the jth jump loading for firm $i, 1 \leq j \leq J$. For the Brownian motions $W_1(t), W_2(t), \ldots, W_m(t)$, they are assumed to be correlated such that

$$dW_i(t)dW_j(t) = \rho_{ij}dt. \tag{3}$$

Brownian motions are allowed to be correlated among market factors because it provides dynamic dependence among market-factor processes. Moreover, these correlation parameters can be represented behavior and economic information of markets. If correlations between market factors are negative, a portfolio becomes more diversified. Conversely, the model that has positive correlation between market factors produces stronger default dependency.

As can be seen in the expression (1), the factor loadings $\beta_{i,j}$ of market factor $j, 1 \leq j \leq m$ are varied across firm $i, 1 \leq i \leq n$. It implies that we cannot rely on factor loadings only to create default correlations. Thus, the factor loadings are used to fit CDS curves of individual firm and the hidden parameters such Brownian motion correlations and other parameters of market factors are used to fit the CDO tranches.

In addition, the compound processes $Z^i, 1 \leq i \leq J$ are independent. In particular, Z is defined by

$$Z^i(t) = \sum_{n=1}^{\Pi_i(t)} \Upsilon_n^{\mathrm{i}}, \tag{4}$$

where $\Pi_i(t)$ is a counting Poisson process with Gamma random intensity Λ^i with a shape parameter α^i and an inversed scaled parameter B^i, and the jump sizes of $\Upsilon_1^{\mathrm{i}}, \Upsilon_2^{\mathrm{i}}$ ... are exponentially distributed random variables with respective means μ^i. Each compound process Z^i triggers events that cause state jumps in market intensities simultaneously with different sizes of jump.

We choose Gamma-Poisson mixture processes to model jumps because default times are interdependent and have positive correlations among them. A normal Poisson distribution has no capability to produce defaults that are correlated across time. The normal Poisson process constricts that the variance are equal to the mean, but the variance of the Gamma-Poisson mixture process can be selected arbitrarily. As a result, this process has ability to capture the fat-tailed dependence that are needed for pricing correlated products such as CDOs.

4. Suggested Methods in Computing Loss Distribution

To compute loss distribution, we must define the default times or the number of defaulted firms in a portfolio. It is known that the default times are determined by its intensity processes. There are several ways to estimate default times such as numerical analysis, Monte Carlo simulation and Semi-analytical transform techniques. Euler time-discretization method is easily used to approximate simulation of any stochastic equations. However, the bias is unacceptable and it is time consuming. In this paper, we suggest two alternative methods, a recursive method and a Mimicking Markov Chain method. Both methods are efficient. To use a recursive method, we need to assume that the defaults occur between coupon payment dates. Conversely for a Mimicking Markov Chain method, the default times and the defaulted firms are acknowledged. In addition, it is useful for the model that has unknown distributions of cumulative or integral intensities, for example, CIR processes. The Mimicking Markov Chain method uses transition rates instead to determine default times of firms in a portfolio. There is no need to discretize time to calculate cumulative intensities.

To use those methods, we have more works to do. We provide the Laplace transform function that is implemented in the recursive method and the

Mimicking Markov Chain method. Usually, the Laplace transform of each process X_i for $1 \le i \le n$ is given by

$$E\left[\exp\left(-u_i\int_t^T X_i(s)ds - z_iX_i(T)\right)\right]. \qquad (5)$$

Nevertheless, the distribution of market factor $X_1(t), X_2(t), .., X_m(t)$ cannot be transformed into their own characteristic functions because these intensity processes are not assumingly independent. Alternatively, we try to study the distribution of $X_1(t), X_2(t), .., X_m(t)$ through the sum $\sum_{i=1}^m X_i(t)$. Therefore, we define the Laplace transform of a vector-valued market-factor process associated with their integrals as

$$\phi(t,u,z,X) = E\left[e^{-\sum_{i=1}^m u_i\int_0^t X_i(s)ds - z_iX_i(t)}\right]. \qquad (6)$$

Before solving the above formula, let us introduce the set $u = \{u_i : 1 \le i \le m\}$, the set $z = \{z_i : 1 \le i \le m\}$, the jump-diffusion set $X = \{X_i : 1 \le i \le m\}$, the set of arrival rates of jumps $\Lambda = \{\Lambda_j : 1 \le j \le J\}$, and the exponentially-affine characteristic function f. For all $(t,u,z,\Lambda,X) \in [0,T] \times \mathrm{R}^\mathrm{m} \times \mathrm{R}^\mathrm{m} \times [0,\infty]^J \times [0,\infty]^m$, the characteristic function $f(t,u,z,\Lambda,X)$ has the stochastic representation

$$\begin{aligned} &f(t,u,z,\Lambda,X) \\ &= E\left[e^{-\sum_{i=1}^m u_i\int_t^T X_i(s)ds - z_iX_i(T)} \mid \mathcal{F}_t\right]. \end{aligned} \qquad (7)$$

We want to write the function f in the affine form

$$\begin{aligned} &f(t,u,z,\Lambda,X_t) \\ &= e^{a(T-t,u,z)+\sum_{i=1}^m b^i(T-t,u_i,z_i)X_t^i + \sum_{j=1}^J c^j(T-t,u,z)\Lambda^j}, \end{aligned} \qquad (8)$$

where $a: [0,T] \times \mathrm{R}^\mathrm{m} \times \mathrm{R}^\mathrm{m} \to \mathrm{R}$, $b^i: [0,T] \times \mathrm{R} \times \mathrm{R} \to \mathrm{R}$, and $c: [0,T] \times \mathrm{R}^\mathrm{m} \times \mathrm{R}^\mathrm{m} \to \mathrm{R}$ with $a(0,u,z) = 0$, $b^i(0,u_i,z_i) = -z_i, c^j(0,u,z) = 0, 1 \le i \le m, 1 \le j \le J$.

However, our proposed model has jump arrival intensities following Gamma distributions, not constant variables. Hence we find the solution of the function f that has jumps as Poisson processes and then find the expected value of the function f conditioning on the arrival rates Λ that are distributed as Gamma.

The Laplace transform of a vector-valued market-factor process associated with their integrals can be written as

$$\begin{aligned} &\phi(T-t,u,z,X_t) \\ &= E\left[E\left[e^{\sum_{i=1}^m\left(-u_i\int_t^T X_i(s)ds - z_iX_i(T)\right)} \mid \Lambda\right]\right] \\ &= E\left[E[f(t,u,z,\Lambda,X_t) \mid \Lambda]\right] \\ &= e^{a(T-t,u,z)+\sum_{i=1}^m b^i(T-t,u_i,z_i)X_t^i}\prod_{j=1}^J E\left[e^{c^j(T-t,u,z)\Lambda^\mathrm{j}}\right], \end{aligned} \qquad (9)$$

where the moment generating function of the gamma distribution $g(\Lambda;\alpha,\mathrm{B}) = \Lambda^{\alpha-1}\exp(-\Lambda/\mathrm{B})/(\Gamma(\alpha)\mathrm{B}^\alpha)$ is

$$E\left[e^{c\Lambda^\mathrm{j}}\right] = \int_0^\infty e^{c\Lambda^\mathrm{j}}\ g(\Lambda^\mathrm{j};\alpha,\mathrm{B})d\Lambda^\mathrm{j} = (1-c\mathrm{B})^{-\alpha}, \qquad (10)$$

4.1 Recursive Method

Let the integral market process be $v_t^i = \int_0^t X_s^i ds, 1 \le i \le m$. The conditional survival probability given the path of a vector of correlated-market-factor processes $(X_s)_{s\le t}$ is defined as

$$P(\tau^i > t|(X_s)_{s\le t}) = \mathrm{e}^{-\sum_{j=1}^m \beta_{i,j}v^j}E\left[\mathrm{e}^{-\int_0^t Y_s^i ds}\right]. \qquad (11)$$

For $0 \le l \le n$, the conditional mass function of the loss given the path of a vector of correlated-market-factor processes $(X_s)_{s\le t}$ is suggested by Andersen et al. (2003), specified as

$$\begin{aligned} &P^i(\mathrm{N}_t = i|(X_s)_{s\le t}) \\ &= P^{i-1}(\mathrm{N}_t = i-1|(X_s)_{s\le t})[1 - P(\tau^i > t|(X_s)_{s\le t})], \\ &P^i(\mathrm{N}_t = 0|(X_s)_{s\le t}) \\ &= P^{i-1}(\mathrm{N}_t = 0|(X_s)_{s\le t})P(\tau^i > t|(X_s)_{s\le t}), \\ &P^i(\mathrm{N}_t = l|(X_s)_{s\le t}) \\ &= P^{i-1}(\mathrm{N}_t = l|(X_s)_{s\le t})P(\tau^i > t|(X_s)_{s\le t}) \\ &\ + P^{i-1}(\mathrm{N}_t = l-1|(X_s)_{s\le t})[1 - P(\tau^i > t|(X_s)_{s\le t})] \end{aligned}$$

where $i = 1, .., l-1$. (12)

The unconditional loss distribution is the expectation of conditional loss distribution

$$P(\mathrm{N}_t = l) = E[P(\mathrm{N}_t = l|(X_s)_{s\le t})]. \qquad (13)$$

However, the computation of the unconditional mass function of the loss becomes intensive when the number of assets in an underlying portfolio is large. For example, CDX IG NA and Itraxx Europe have 125 firms in their portfolios. Assuming that the underlying portfolio is homogeneous and then $Y^1, \dots, Y^n$ are randomly independent, we obtain

$$\begin{aligned} &P(\mathrm{N}_t = l|(X_s)_{s\le t}) \\ &= \binom{n}{l}\left(1 - P(\tau > t|(X_s)_{s\le t})\right)^l\left(P(\tau > t|(X_s)_{s\le t})\right)^{n-l}. \end{aligned} \qquad (14)$$

The conditional loss distribution above can be rewritten as

$$\begin{aligned} &P(\mathrm{N}_t = l|(X_s)_{s\le t}) \\ &= \binom{n}{l}\sum_{i=1}^l\binom{l}{i}(-1)^{l-i}\,\mathrm{e}^{-(n-i)\sum_{j=1}^m \beta_j v^j}\left(E\left[\mathrm{e}^{-\int_0^t Y_s ds}\right]\right)^{n-i}. \end{aligned} \qquad (15)$$

Note that entire firms use the same factor loadings $\beta = \{\beta_1, \dots, \beta_m\}$. By taking expectation, its unconditional loss distribution becomes

$$P(\mathrm{N}_t = l) \\ = \binom{n}{l} \sum_{i=1}^{l} \binom{l}{i} (-1)^{l-i} E\left[\mathrm{e}^{-(n-i)\sum_{j=1}^{m} \beta_j v^j}\right] \left(E\left[\mathrm{e}^{-\int_0^t Y_s ds}\right]\right)^{n-i}, \tag{16}$$

where $E\left[\exp\left(-(n-i)\sum_{j=1}^{m} \beta_j\, v^j\right)\right] = \phi(t, u, 0_m, X_0)$ from (9), and $u_j = (n-i)\beta_j, 1 \le j \le m$.

4.2 Mimicking Markov Chain Method

Giesecke et al. (2010) develop the simulation approach that is exact and efficient for a vector process. In order to avoid discretizing the vector process $\lambda_t = (\lambda_t^1, \dots, \lambda_t^n)$, they construct the mimicking Markov chain $M_t = (M_t^1, \dots, M_t^n) \in \{0,1\}^n$, which has the same properties as the portfolio default process N_t, in its own filtration $\mathbb{G} = (G_t)_{t\ge 0}$ generated by M. The mimicking Markov chain M_t is determined by the transition rates $h^i(t, M)$ instead of intensity processes $\lambda_t^i, 1 \le i \le n$. The transition rate $h^i(t, B)$ has the meaning of the expectation of $\lambda_t^i I(\tau^i > t)$ conditioning on the vector-valued default counting process $N_t = B$, defined as

$$h^i(t,B) = E\left(\lambda_t^i I(\tau^i > t)\middle| N_t = B\right), B \in \{0,1\}^n.$$

For our proposed model, the transition rate $h^i(t, B)$ can be rewritten as

$$h^i(t,B) = (1 - B^i) E\left(Y_t^i \middle| \tau^i > t\right) \\ + (1 - B^i) \sum_{j=1}^{m} \beta_{i,j} E\left(X_t^j \middle| N_t = B\right). \tag{17}$$

Because the idiosyncratic factor Y^i is independent, the expectation of the ith idiosyncratic-factor intensity given that the firm is survival $E\left(Y_t^i \middle| \tau^i > t\right) = -\frac{\partial_z \phi\left(t,1,z,Y_0^i\right)|_{z=0}}{\phi\left(t,1,0,Y_0^i\right)}$ is simply solved. However, the solution becomes more complicated when market factors are correlated. By Bayes' theorem and the law of iterated expectations, we obtain

$$E\left(X_t^i | N_t = B\right) = \frac{E\left(X_t^i I(N_t = B)\right)}{P(N_t = B)} \\ = \frac{E\left(X_t^i P(N_t = B | (X_s)_{s\le t})\right)}{E\left(P(N_t = B | (X_s)_{s\le t})\right)}.$$

The conditional probability at time t that the portfolio default process N_t is equivalent to B given the path of a vector of correlated-market-factor processes $(X_s)_{s\le t}$ is given by

$$P(N_t = B | (X_s)_{s\le t}) \\ = \prod_{j=1}^{n} P\left(N_t^j = B^j | (X_s)_{s\le t}\right) \\ = \prod_{j=1}^{n} \left[B^j - (2B^j - 1) P(\tau^j > t | (X_s)_{s\le t})\right].$$

By iterated expectations and conditional independence, the conditional survival probability of firm i given the path of a vector of correlated-market-factor processes $(X_s)_{s\le t}$ is

$$P(\tau^i > t | (X_s)_{s\le t}) \\ = \exp\left(-\sum_{j=1}^{m} \beta_{i,j} \int_0^t X_s^j ds\right) E\left[\exp\left(-\int_0^t Y_s^i ds\right)\right].$$

Thus, there are 2^n terms in the expansion of

$$P(N_t = B | (X_s)_{s\le t}) \\ = \prod_{j=1}^{n} \left[B^j - (2B^j - 1)\mathrm{e}^{-\sum_{j=1}^{m} \beta_{i,j} \int_0^t X_s^j ds} E\left[\mathrm{e}^{-\int_0^t Y_s^i ds}\right]\right] \\ = \sum_{k=0}^{2^n - 1} c_k(t) \exp\left(-\sum_{j=1}^{m} b_{k,j} \int_0^t X_s^j ds\right),$$

where $c_k(t)$ is the coefficient of the k-th term and each constant $b_{k,j}$ is a sum of values $\beta_{i,j}$ for particular i. Denote $b_k = \left(b_{k,j}\right)_{k;j=1,\dots,m}$ as a vector of non-negative constant and take expectation on both sides of the equation above

$$E[P(N_t = B | (X_s)_{s\le t})] \\ = \sum_{k=0}^{2^n - 1} c_k(t)\, \mathrm{E}\left[\exp\left(-\sum_{j=1}^{m} b_{k,j} v^j\right)\right] \\ = \sum_{k=0}^{2^n - 1} c_k(t)\, \phi(t, b_k, 0_m, X_0).$$

For the $E\left(X_t^i P(N_t = B) | (X_s)_{s\le t}\right)$ term, it can be solved by taking the derivative of $E[P(N_t = B | (X_s)_{s\le t})]$ with respect to z_j, and substituting $z_1 = z_2 = \cdots. = 0$. Now, we have

$$h^i(t,B) \\ = -\frac{\partial_z \phi\left(t,1,z,Y_0^i\right)|_{z=0}}{\phi\left(t,1,0,Y_0^i\right)} \\ - \sum_{j=1}^{m} \beta_{i,j} \frac{\sum_{k=0}^{2^n-1} c_k(t)\, \phi_{z_j}(t, b_k, z, X_0)|_{z_1 = z_2 = \cdots. = 0}}{\sum_{.k=0}^{2^n-1} c_k(t)\, \phi(t, b_k, 0_m, X_0)}. \tag{18}$$

In the filtration $\mathbb{G}$, $(T_k)_{k=1,2,..}$ is the sequence of default times, the inter-arrival intensity function is

$$H(t,k) = \sum_{i=1}^{n} h\left(t, M_{T_k}\right), \quad T_k \le t, \quad k = 0,1,2, \dots$$

As will be seen, the inter-arrival intensity function $H(t, k)$ plays an important role in the thinning scheme.

4.2.1 Simulation of market factors

The computation of transition rates $h^i(t, B), 1 \le i \le n$ in the expression (17) has to operate at least 2^n steps. It is hard to estimate transition rates for the portfolio has the number of firms $n \ge 100$. The most cumbersome part is to compute $E(X_i(t) | N_t = B)$. Therefore, we generate the path of correlated market-factor intensities $(X_s)_{s\le t}$ to calculate the expectations that are conditional on the path of market-factor processes such as $P(N_t = B|(X_s)_{s\le t})$ and $X_t^i P(N_t = B|(X_s)_{s\le t})$, and then average them to get unconditional expectations.

To generating paths for the intensities $X_i(t)$ and their integral $\int_0^t X_i(s)ds$ of market factors, we use the procedure described as follows:

1. Simulate each Gamma-mixture of Poisson process Z^j with arrival rate $\Lambda^j \sim \text{Gamma}(\alpha^j, \mathrm{B}^j)$.
2. Within the time interval $[0,T]$, generate the set of jump times and their jump sizes $\{\Gamma_h^j, \varepsilon_h^j\}$ which is defined by $\Gamma_h^j = \inf\{t : N^j(t) = h\}$ and ε_h^j generated by sampling from an exponential distribution with mean μ^j.
3. Given the set of jump times and their jump sizes $\{\Gamma_h^j, \varepsilon_h^j\}$, generate samples from the distributions of X_i and $\int_0^t X_i(s)ds$ for the i-th market factor as if they were normal distributions.

4.2.2 Thinning Scheme

In Giesecke et al. (2010), the thinning algorithm is applied to simulate the mimicking chain M by generating the firm defaults' identities and their default times. The advantage of this scheme is that there is no need for time scaling and discrete-time integration.

First, we determine the appropriate value of the number of intervals $\mathcal{M}$ for the intensity $H(t,k)$, next create a partition of the given interval $[0,T]$ such that $0 \le L_0 < L_1 <\ldots< L_{\mathcal{M}} = T$ to obtain a subinterval $[L_i, L_{i-1}]$ and then find the majorizing function $H^*(i,k)$ such that

$$H^*(i,k) = \sup\{H(s,k) \colon L_{i-1} \le s < L_i\}, \qquad (19)$$

where $i = 1,\ldots,\mathcal{M}$. After an exponential random arrival time $x, x \in (L_{i-1}, L_i)$ having intensity rate $H^*(i,k)$ is generated, we need to accept the event occurring at time x with probability $H(x,k)/H^*(i,k)$.

5. Numerical Results

In this chapter, we show how to calibrate our proposed model to market data to fit both CDO tranche spreads and single name CDS spreads.

Similar to Peng and Kou (2009)'s work, we assume that the idiosyncratic part Y^{i} of the firm i's specific intensity is small as error. Hence $E\left[\mathrm{e}^{-\mathrm{Y}^{\mathrm{i}}}\right] = 1$. Given initial values of market factor parameters and the firm i's CDS spreads $S_{mkt}^i(t,T)$ that is observed from the market at time t and matures at time T. Denote the survival probability $q_i(t) = P(\tau^i > t)$ which is calibrated to the firm i's CDS spreads for each maturity. The firm i's CDS spreads $S_{model}^i(t,T)$ is calculated by using estimated parameters of the model. The factor-loading coefficient $\beta_{i,1},\ldots,\beta_{i,j}$ for firm i can be estimated as follows:

$$\text{Min} \qquad \sum_T \left(\frac{S_{model}^i(t,T) - S_{mkt}^i(t,T)}{S_{mkt}^i(t,T)}\right)^2$$

$$\text{s.t.} \quad \frac{q_i(t_k)}{E\left[\mathrm{e}^{-\int_t^{t_k}\sum_{j=1}^m \beta_{i,j}X_s^j ds}\right]} \le \frac{q_i(t_{k-1})}{E\left[\mathrm{e}^{-\int_t^{t_{k-1}}\sum_{j=1}^m \beta_{i,j}X_s^j ds}\right]}$$

$$t \le t_k \le T \text{ and } 0 \le \beta_{i,j}, j = 1,\ldots,m.$$

5.1 Calibration Algorithm

Let S_k be the kth tranche spread form the model and S_k^a, S_k^b be the ask price and bid price of CDO tranche spread from the market. To fit spreads of CDO tranche and CDS for our model, we use the following processes:

1. Calibrate survival probabilities $q_i(t)$ which is bootstrapped from firm i's CDS spreads.
2. Initialize values of market factor parameters.
3. Estimate the factor-loading coefficient $\beta_{i,1},\ldots,\beta_{i,j}$ according to (18).
4. Use $\frac{q_i(t_k)}{\left[\mathrm{e}^{-\int_t^{t_k}\sum_{j=1}^m \beta_{i,j}X_s^j ds}\right]}$ to estimate values of the idiosyncratic factor parameter for coupon payment dates $t \le t_1, t_2, \ldots, t_M = T$.
5. Calculate spreads for the CDO tranches by using the methods that we have mentioned.
6. Update the parameters and the path of the market factors and repeat until the root mean square error (RMSE) is small enough which is given by

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\frac{s_k - \frac{(s_k^a + s_k^b)}{2}}{s_k^a - s_k^b}\right)^2}$$

5.2 Results

We fit the model to market data such CDOs and CDSs and then compare our results with those that are from Mortensen (2006) and Peng and Kou (2009).

CDX NA IG and Itraxx Europe are the most liquid CDS indices and have therefore no arbitrage. Both of them have 125 firm assets in their portfolio. The tranches of a CDO are classified by the level of the portfolio loss. The equity tranche has different mechanism of pricing from other tranches, paying an upfront fee with a running spread instead.

For CDX NA IG, our model is calibrated to CDX NA IG S2 5Y and CDX NA IG S5 5Y on August 23, 2004 and December 5, 2005 respectively. CDX NA IG S2 5Y was launched on March 23, 2004 and matured on June 22, 2009 with quarterly coupons. Similarly, CDX NA IG S4 5Y was launched on September 21, 2005 and matured on Dec. 20, 2010. Interest rates are extracted from USD swap rates. The recovery rate is assumed to be 40%. The equity tranche is quoted with a running spread of 500 bps. Table 1 displays that our model can solve the issue of overpricing equity tranche (0-3%), the mezzanine 1 tranche (3-7%), and the mezzanine 3 tranche (10-15%).

Table 1: Comparison of calibration results from our model and Mortensen (2006) model for the CDX NA IG S2 5Y on August 23, 2004

Tranches %	Source			
	Market	B/A	Mortensen	Model
0-3	40.0%	2.0%	46.9%	39.4%
3-7	312.5	15.0	340.2	318.3
7-10	22.5	7.0	19.7	18.4
10-15	42.5	7.0	61.9	45.0
15-30	12.5	3.0	14.3	12.7
RMSE			2.1	0.37

Likewise, Table 2 shows that our model fits the mezzanine 1 (3-7%) better that the jump-diffusion model of Mortensen (2006). RMSE is lesser.

Table 2: Comparison of calibration results from our model and Mortensen (2006) model for the CDX NA IG S5 5Y on December 5, 2005

Tranches %	Source			
	Market	B/A	Mortensen	Model
0-3	41.1%	0.8%	43.2%	40.9%
3-7	117.5	6.8	125.9	122.0
7-10	32.9	5.3	30.6	33.27
10-15	15.8	3.0	21.3	15.8
15-30	7.9	1.0	8.8	7.3
RMSE			1.58	0.40

ITraxx Europe S8 5Y was launched on September 20, 2007 and matured on December 20, 2012. We fit the model to market data on March 14, 2008. Table 3 shows that our result is not much different from Peng and Kou (2009)'s.

Table 3: Comparison of calibration results from our model and Peng and Kou (2008) model for the ITraxx Europe S8 5Y on March 14, 2008

Tranches %	Source			
	Market	B/A	Peng&Kou	Model
0-3	51.4%	1.6%	50.5%	51.9%
3-6	649.0	24.3	691.14	658.9
6-9	401.1	24.5	395.47	374.5
9-12	255.3	19.8	261.23	249.5
12-22	143.4	11.8	168.62	166.6
22-100	69.9	2.9	66.96	68.0
RMSE			1.27	1.07

Two market factors are efficient to fit CDO tranches for our sample series. The obvious difference between those two market factors are jump parts. Each market factor has its unique jump process with the jump loading $\zeta = 1$. Table 4 represents that either market factor has the jump size u significantly greater than market factor 2, whereas the shape parameter α is remarkably lesser. Interestingly, we found that the correlation parameter helps fitting the equity tranche and the mezzanine 1 tranche more efficient without affecting other tranches. When increasing correlation, the equity tranche spread is decreasing and the mezzanine 1 tranche spread is increasing. Conversely, if we want to reduce the spread of mezzanine 1 tranche and increase the spread of equity tranche, we decrease the value of correlation parameter.

Table 4: Estimated Parameters

Date (Series)	Part		Market factors	
			1	2
23/4/2004 (CDX S2)	Continuous part	k	0.1	1.3
		θ	0.004	0.0042
		x_0	0.0004	0.0006
		σ	0.0008	0.0006
	Jump part	u	8	0.072
		α	0.001	10
		B	25	0.01
	Correlation	ρ		-1
5/12/2005 (CDX S5)	Continuous part	k	0.6	2
		θ	0.003	0.004
		x_0	0.002	0.0008
		σ	0.001	0.0048
	Jump part	u	1	0.0168
		α	0.0016	0.07
		B	2	2.5
	Correlation	ρ		-1
23/4/2004 (Itraxx S8)	Continuous part	k	0.1	1.6
		θ	0.0008	0.0102
		x_0	0.0004	0.0012
		σ	0.002	0.0006
	Jump part	u	1.2	0.18
		α	0.01	25
		B	15	0.004
	Correlation	ρ		1

6. Conclusion

For Mortensen (2006), they use Poisson processes to drive jump parts. The result shows that Gamma-Poisson mixture process can fit tranche better than normal Poisson process. For Peng and Kou (2009), they claim that cumulative intensity can produce correlated defaults better. However, our model can emulate their achievement in capturing clustering defaults. It shows that the Gamma-Poisson process is the real factor that generates tail dependence.

References

[1] Duffie D, Garleanu N. Risk and Valuation of Collateralized Debt Obligations. Financial Analysts Journal. 2001; 57: 41-59.

[2] Giesecke K, Kakavand H, Mousavi M, Takada H. Exact and efficient simulation of correlated defaults. SIAM J. Financial Math. 2010; 1: 868–896.

[3] Lewis P, Shedler G. Simulation of nonhomogeneous Poisson processes by thinning. Naval Logistics Quart. 1979; 26: 403–413.

[4] Mortensen A. Semi-analytical valuation of basket credit derivatives in intensity-based models. Journal of Derivatives. 2006; 13: 8–26.

[5] Peng S, Kou X. Default clustering and valuation of collateralized debt obligations. 2009.

[6] Andersen L, Sidenius J, and Basu S. All your hedges in one basket. Risk, November, 2003: 67-72.

Appendix A. The Exponentially-Affine Characteristic Function

The exponentially-affine characteristic function f has jumps distributed as Poisson with arrival rates $\Lambda = \{\Lambda_i,\ 1 \leq i \leq J\}$, for all $(t, u, z, \Lambda, X) \in [0, T] \times \mathrm{R}^{\mathrm{m}} \times \mathrm{R}^{\mathrm{m}} \times [0, \infty]^J \times [0, \infty]^m$ is written as

$$f(t, u, z, \Lambda, X_t)$$
$$= e^{a(T-t,u,z)+\sum_{i=1}^{m} b^i(T-t,u_i,z_i)X_t^i+\sum_{j=1}^{J} c^j(T-t,u,z)\Lambda^j},$$

where

$$b^i(t, u_i, z_i) = \left(-z_i + \frac{u_i}{k_i}\right) e^{-k_i t} - \frac{u_i}{k_i},$$

$$c^j(t, u, z) = \int_0^t \frac{1}{1 - \mu^j \sum_{i=1}^{m} \zeta^{i,j} b^i(s, u_i, z_i)} ds,$$

$$a(t, u, z)$$
$$= -\sum_{i=1}^{m} \theta_i \left((e^{-k_i t} - 1)\left(-z_i + \frac{u_i}{k_i}\right) + t u_i \right)$$
$$+ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=i}^{m} \left[\rho_{ij}\sigma_i\sigma_j \left(\frac{t u_i u_j}{k_i k_j} \right.\right.$$
$$+ \frac{1}{k_i}(e^{-k_i t} - 1)\frac{u_j}{k_j}\left(-z_i + \frac{u_i}{k_i}\right)$$
$$+ \frac{1}{k_j}\left(e^{-k_j t} - 1\right)\frac{u_i}{k_i}\left(-z_j + \frac{u_j}{k_j}\right)$$
$$\left.\left. - \frac{\left(e^{-(k_i+k_j)t} - 1\right)\left(-z_i + \frac{u_i}{k_i}\right)\left(-z_j + \frac{u_j}{k_j}\right)}{k^i + k^j} \right) \right]$$

Appendix B. Thinning Scheme Algorithm

The algorithm of Thinning scheme is used to generate sequences of $(T_k, I_k)_{k=1,2,..}$where the kth default time $T_k \leq T$ and the kth default $I_k \leq n$. Denote the number of intervals $\mathcal{M}$, the transition rate function $h(t, B)$ where $B \in \{0,1\}^n$, the inter-arrival intensity function $H(t, k)$, and the majoring intensity function $H^*(i, k)$. Inputs are the current interval i such that $i = \{i^*: L_{i^*-1} \leq t \leq L_{i^*}\}$, the firms' status vector M, the current time t, and the number of the firms that have defaulted k. First, we initialize t=0, k=0, T=0$_\mathrm{n}$, M=0$_\mathrm{n}$ and i=1. Then we proceed as follows:

1. Generate x~exponential random variable with the mean $H^*(i, k)$.
2. If $t + x < L_i$ set $t \leftarrow t + x$ and if $t > T$ or $i > M$ stop else go to step 4. Else if $t + x \geq L_i$ go to step 3.
3. Set $x \leftarrow H^*(i, k)(t + x - L_i)/H^*(i + 1, k), t \leftarrow t_i, i \leftarrow i + 1$. Go to step 2.
4. Generate $\omega_2 \sim random[0,1]$.
   If $\omega_2 \leq H(t, k)/H^*(i, k)$, set $k \leftarrow k + 1$and $T_k \leftarrow t$. Otherwise, go to 1.
5. Draw the defaulted firm J from the pool of the firms that have survived with probabilities $h^j(t, M_{T_{k-1}})/H(t, k - 1)$. Then set $I_k = J$ and update $M_{T_k}[J] = 1$. Return to step 1.

An analytical of ARL for seasonal MA(1)$_s$ on CUSUM chart

Piyapatr Busababodhin

Department of Mathematics, Faculty of Science, Mahasarakham University, Thailand. Piyapatr99@gmail.com

Abstract

This paper studies on the observations are from a seasonal first order moving average (seasonal MA(1)$_s$) model with exponential white noise on the Cumulative Sum (CUSUM) chart. In this paper, the explicit formulas of the Average Run Length (ARL) and Average Delay Time (ADT) are derived, and a numerical integration method for evaluating ARL is also presented. The numerical results from explicit formulas and the numerical integration method are presented. The results illustrate that the explicit formulas can reduce computational times to evaluate the ARL when compared with the results obtained from the numerical integration method. According to the proposed explicit formulas for the ARL, it is very useful in practical applications in order to design an optimal CUSUM chart.

Keywords: Cumulative sum, seasonal first order moving average, average run length, exponential distribution, integral equation

Corresponding Author
E-mail Address: Piyapatr99@gmail.com

1. Introduction

Since the Cumulative Sum (CUSUM) control chart was proposed by Page in 1954, lots of CUSUMs have been developed and then improved to use for different process data. The CUSUM is popular procedure in statistical quality control charts as theirs sensitivities of small shifts of changed parameters. Traditional, the CUSUM procedure is based on an assumption that the observations are normally and identically independent random variables. This assumption, however, is not appropriated in practice such as in continuous manufacturing which most of observations are autocorrelated, refinery operations, smelting operations, wood product manufacturing, waste-water processing and the operation of nuclear reactors. Alwan and Robert [1] showed that 85% of a sample of 235 control chart applications displayed incorrect control limits and more than half of these displacements were restricted to violation of the independence assumption. Many researchers investigated new methods to measure the CUSUM chart when observations are autocorrelation whether the process is stationary or not (see [2], [3], [6], [7], [9], [17], [23], [24]).

There are many characteristics to show the performance of chart; such as Average Run Length (ARL) and the Average Delay Time (ADT), they are two conflicting criteria that must be balanced in chart. Both of them are frequently method used in chart for evaluating the detection performance of various control charts (see [12], [15], [15], [21]). In recently, there are many methodologies to measure the ARL and ADT of chart such as Monte Carlo simulation (MC), Markov Chain Approach (MCA) (see [12]), martingale approaches (see [19, 20]) and Integral Equations (IE) (see [5], [10], [18]). The first three methods can approximate the ARL and ADT as closed-form formulas, while the IE can derive the ARL and ADT as explicit formulas.

In this article, the ARL and ADT of the CUSUM chart when observations are from a seasonal first order moving average (seasonal MA(1)$_s$) model with exponential white noise are studied. The derivation of integral equations for the ARL and ADT are solved by using the Gauss-Legendre numerical integration rule. Furthermore, the results obtained from the numerical integration with the results obtained from explicit formulae are compared.

2. The characteristics of CUSUM chart

The discussion of CUSUM chart's characteristic, which was proposed by Pallak and Siegmund (1985). Especially, it is an effective approach for detecting small changes. Its properties have been investigated by many authors (see [7], [22]). Generally, this chart is based on the assumption that $\xi_1, \xi_2, ..., \xi_n$ are sequentially observed identically independent distributed (i.i.d) random variables with exponential distribution function $F(x, \lambda)$. To assume that the parameter λ has the value λ_0 in the in-control state, the value $\lambda \neq \lambda_0$ in an out-of-control state, and the change occurs at a change-point time $\theta \leq \infty$. The parameters λ and λ_0 are assumed to be known.

A typical method of detecting change-point in CUSUM chart is to define some statistic X_n and a control boundary limit h of X_n such that an alarm signal is given when X_n exceeds h. Typically, a first exit time (τ) over a boundary is defined as

$$\tau_h = \inf\{n \geq 0; X_n \geq h\},$$

is used for the alarm signal.

To define $\mathbb{E}_\theta(\cdot)$ as the expectation under distribution $F(x,\lambda_0)$ that the change-point occurs at time θ from the in-control value λ_0 to an out-of-control value λ. Typically, measures for alarm times τ are

$$ARL \approx \mathbb{E}_\infty \tau_h \geq T,$$

where T is given (usually large) and

$$ADT \approx \mathbb{E}_1 \tau_h \leq (\tau | \tau \geq 1).$$

The ARL is a measure of the average time before a process that is still in-control is signaled as being out-of-control and ADT is a measure of the average time before a process that has gone out-of-control is signaled as being out-of-control.

To define X_n as the statistics which satisfies the following recursive equation

$$X_n = (X_{n-1} + \xi_n - a)^+, \quad n = 1,2,..., \quad X_0 = x,$$

where X_n is the CUSUM value of a statistic after n observations, x is an initial value for X_n, $y^+ = \max(0, y)$ and a is a constant.

In this paper, to consider CUSUM charts for the case where observations are from a seasonal MA(1)$_s$ model with exponential white noise and define

$$X_n = X_{n-1} + Z_n - a, \qquad n = 1,2,...,X_0 = x, \qquad (1)$$

with

$$Z_n = \xi_n - \phi_1 \xi_{1-s} \text{ , } -1 < \phi_1 < 1 \text{ and } \xi_n \sim \exp(\lambda).$$

where n is the time of sampling, Z_n is the sample value at time n, a is reference value, ϕ is the moving average coefficient $(-1 < \phi < 1)$, s is periodicity and ξ_n is the autoregressive white noise at time n following $\xi_n \sim \exp(\lambda)$.

3. The Evaluation of Average Run Length

In this section, first present is the explicit formulae discovered for the ARL and ADT, then propose a numerical integral equation approach based on the Gauss-Legendre rule.

3.1 Explicit Formulae

According to Banach's Fixed Point Theorem, we present the existence and uniqueness of our solutions. To evaluate the ARL of CUSUM chart defined as a function $j(x) = \mathbb{E}_X \tau_h$. Let $\mathbb{P}_X$ and $\mathbb{E}_X$ be the probability measure and the induced expectation corresponding to the initial value $X_0 = x$. Varderman and Ray (1985) and Venkateshwara et al.(2001) showed that the ARL for CUSUM at a given level, defined as $j(x) = ARL = \mathbb{E}_X \tau_h < \infty$, is a solution of the following integral equation

$$j(x) = 1 + \mathbb{E}_X\left[I\{0 < X_1 < h\}\, j(X_1)\right] + \mathbb{P}_X\{X_1 = 0\}\, j(0). \qquad (2)$$

For this case, ξ_n are exponential distributed observations which have been shown by Busaba et al. (2011) and Mititelu et al. (2010) and ξ_n are exponential distribution white noise in the MA(1) model by Petcharat (2013). This paper also define ξ_n as exponential distribution white noise in the seasonal MA(1)$_s$ model as in (1) so (2) can be written as

$$j(x) = 1 + \lambda e^{-\lambda(a - x + \phi_1 \xi_{1-s})} \int_0^h j(y) e^{-\lambda y} dy + \left(1 - e^{-\lambda(a - x + \phi_1 \xi_{1-s})}\right) j(0), \qquad (3)$$

where $x \in [0, a)$.

It is shown that solutions of the integral equation (3) are continuous functions because the right hand side of (3) contains only continuous functions.

As on the metric space of all continuous functions $(\mathbb{C}(\mathbb{I}), \| \|_1)$, where $\mathbb{I}$ is a compact interval, and the norm is defined as $\|j\| = \sup_{x \in \mathbb{I}} |j(x)|$, the operator T is named a contraction if there exists a number $0 \leq q < 1$ such that

$\|T(j_1) - T(j_2)\| \leq q \|j_1 - j_2\|$ for all $j_1, j_2 \in X$. Now, define the operators T as

$$T(j(x)) = 1 + \lambda e^{-\lambda(a - x + \phi_1 \xi_{1-s})} \int_0^h j(y) e^{-\lambda y} dy + \left(1 - e^{-\lambda(a - x + \phi_1 \xi_{1-s})}\right) j(0), \qquad (4)$$

where $x \in [0, a)$.

Then the integral equations in (3) can be written as $T(j(x)) = j(x)$. Recalling Banach's Fixed Point Theorem if the operator T is contraction, then the fixed-point equation $T(j(x)) = j(x)$ has a unique solution. To show the uniqueness of the solution of (4), it is shown as the prove in Theorem 3.1 that T is a contraction. Define the norms $\|j\|_1 = \sup_{x \in \mathbb{I}_1} |j(x)|$,

Theorem 3.1 On the metric spaces $(\mathbb{C}(\mathbb{I}_1), \| \|_1)$ the operator T is a contraction.

Proof. First, to prove T is contraction we may check that for any $x \in \mathbb{I}_1$, and $j_1, j_2 \in \mathbb{C}(\mathbb{I}_1)$, the inequality $\|T(j_1) - T(j_2)\|_1 \leq q \|j_1 - j_2\|_1$, where q is a positive constant, $0 \leq q < 1$. According to (4) it shows that:

$$\left\|T(j_1)-T(j_2)\right\| = Sup\left|j(x)\right|$$

$$= \underset{x\in[0,a)}{Sup}\Big|\left(j_1(0)-j_2(0)\right)$$

$$\left(1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\right)+\lambda e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}$$

$$\int_0^h\left(j_1(y)-j_2(y)\right)e^{-\lambda y}dy\Big|$$

$$\le \underset{x\in[0,a)}{Sup}\Big|\left\|j_1(0)-j_2(0)\right\|_1$$

$$\left(1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\right)$$

$$+\left\|j_1-j_2\right\|_1\lambda e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\int_0^h e^{-\lambda y}dy\Big|$$

$$=\left\|j_1-j_2\right\|_1\underset{x\in[0,a)}{Sup}\left[1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}-\lambda h\right]$$

$$=\left[1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}-\lambda h\right]\left\|j_1-j_2\right\|_1$$

$$=q_1\left\|j_1-j_2\right\|,$$

where $q_1=\left[1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}-\lambda h\right]<1.$

The triangular inequality and the fact is shown that,

$$\left|j_1(0)-j_2(0)\right|\le \underset{x\in[0,a)}{Sup}\left|j_1(x)-j_2(x)\right|=\left\|j_1-j_2\right\|.$$

To consider the explicit formulae and the numerical integral equation to solve the solutions for the seasonal MA(1)$_s$ model. The explicit formulas are based on an integral equation approach, "Fredholm integral equation of the second type". In Theorem 3.2, to derive and propose explicit solutions which are guaranteed existence and uniqueness by Theorem 3.1.

Theorem 3.2 The solution of (3) is

$$j(x)=\left(1+e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}-\lambda h\right)e^{\lambda h}-e^{\lambda x},\ x\ge 0.$$

Proof.

$$j(x)=1+\lambda e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\int_0^h j(y)e^{-\lambda y}dy+$$

$$\left(1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\right)j(0),\quad x\in[0,a).$$

Set $d=\int_0^h j(y)e^{-\lambda y}dy.$ Now, we have

$$j(x)=1+\lambda e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}d\ +\left(1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\right)j(0).\ (5)$$

If $x=0$ then

$$j(0)=1+\lambda e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}d+\left(1-e^{-\lambda\left(a-x+\phi_1\xi_{1-s}\right)}\right)j(0),$$

$$=1+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}+\lambda d.$$

Substituting $j(0)$ into (5), we found that

$$j(x)=1+\lambda d+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}-e^{\lambda x}\qquad(6)$$

Now the constant d can be found as

$$d=\int_0^h j\left(1+\lambda d+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}-e^{\lambda y}\right)e^{-\lambda y}dy$$

$$=\frac{e^{\lambda h}}{\lambda}\left(1-e^{-\lambda h}\right)\left(1+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}\right)-he^{\lambda h}.$$

Substituting the constant d into (6), we have

$$j(x)=\left(1+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}-\lambda h\right)e^{\lambda h}-e^{\lambda x},\quad x\ge 0.$$

The explicit formulas for the ARL and ADT are presented as follows,

$$ARL=\left(1+e^{\left(a+\phi_1\xi_{1-s}\right)}-h\right)e^{h}-e^{x},\qquad(7)$$

and

$$ADT=\left(1+e^{\lambda\left(a+\phi_1\xi_{1-s}\right)}-\lambda h\right)e^{\lambda h}-e^{\lambda x},\qquad(8)$$

where λ is a parameter of the exponential distribution, ϕ_1 is an first order of moving average observations model, h is boundary value, and a is reference value.

3.2. Numerical Integral Equation Approach

This approach was first studied by Crowder (1978) for approximating the ARL of a Gaussian distribution. He derived and used a Fredholm Integral Equation of the second type. Later, Champ and Rigdon (1991) applied this approach to evaluate the ARL for both the CUSUM and EWMA charts and compared the results obtained with Monte Carlo simulation.

To apply the approach to the CUSUM chart for an seasonal MA(1)$_s$ process, it has to be assumed that the system is in-control at time n if the CUSUM statistic X_n is in the range $H_L\le X_n\le H_U$ and out-of-control if $X_n>H_U$ or $X_n>H_L$, where H_L is a constant lower bound $\left(H_L=0\right)$, and H_U is a constant upper bound $\left(H_U=h\right)$. Also assume that the system is initially in an in-control state x, i.e., $X_0=x$ and $0\le x\le h.$ Then, a function $j^{IE}(x)$ can be defined as follows:

$$j^{IE}(x)=E_x\tau_h<\infty$$

$$=\ 1+\int_0^h j(y)\,f\left(y+a_1-x\right)dy$$

$$+F(a-x)\,j(0)\qquad(9)$$

where τ_h is the first exit time defined in (1). Then $j^{IE}(x)$ is the ARL for initial value $x.$

Next, the present of a numerical scheme for evaluating solutions of the integral equations (9) for the CUSUM chart which can be written as follows:

$$j^{IE}(x)=1+j(0)F\left(a-x+\phi_1\xi_{1-s}\right)$$

$$+\int_0^h j(y)\,f\left(a-x+\phi_1\xi_{1-s}+y\right)dy,\qquad(10)$$

where $F(x)=1-e^{-\lambda x}$ and $f(x)=\dfrac{dF(x)}{dx}=\lambda e^{-\lambda x}$.

For a given quadrature rule for integrals on $[0,h]$, the integral equation can be approximated by

$$j(a_i)\approx 1+j(a_1)F(a-a_i+\phi_1\xi_{1-s}).$$
$$+\sum_{k=1}^{m} w_k\, j(a_k)\, f(a_k+a-a_i+\phi_1\xi_{1-s}),$$
$$i=1,2,...,m \qquad .. \quad (11)$$

Without loss of generality, the integral by a sum of areas of rectangles can be approximated with bases h/m and heights chosen as the values of $f(a_k)$ at the midpoints of intervals of length h/m beginning at zero, i.e. on the interval $[0,h]$ with the division points $0\le a_1\le a_2\le ...\le a_m\le h$ and weights w_k. Then, to obtain $\int_0^h j(y)dy\approx\sum_{k=1}^{m} w_k\, f(a_k)$, where $a_k=\dfrac{h}{m}\left(k-\dfrac{1}{2}\right)$, $k=1,2,...,m.$

Equation (11) is a system of m linear equations in the m unknowns $j(a_1), j(a_2),..., j(a_m)$, and it can be written in matrix form as

$$J_{m\times1} = 1_{m\times1}+R_{m\times m}J_{m\times1}$$
$$(I_m-R_{m\times m})J_{m\times1} = 1_{m\times1} \qquad (12)$$

where

$$J_{m\times1}=\begin{pmatrix} j(a_1)\\ j(a_2)\\ \vdots\\ j(a_m)\end{pmatrix},\quad 1_{m\times1}=\begin{pmatrix}1\\1\\ \vdots\\ 1\end{pmatrix},$$

$$R_{m\times m}=\begin{pmatrix} F(a-a_1+\phi_1\xi_{1-s})+w_1 f(a) & w_2 f(a_2+a-a_1+\phi_1\xi_{1-s}) & \cdots & w_m f(a_m+a-a_1+\phi_1\xi_{1-s})\\ F(a-a_1+\phi_1\xi_{1-s})+w_1 f(a_1+a-a_2+\phi_1\xi_{1-s}) & w_2 f(a) & \cdots & w_m f(a_m+a-a_2+\phi_1\xi_{1-s})\\ \vdots & \vdots & \ddots & \vdots\\ F(a-a_m+\phi_1\xi_{1-s})+w_1 f(a_1+a-a_m+\phi_1\xi_{1-s}) & w_2 f(a_2+a-a_m+\phi_1\xi_{1-s}) & \cdots & w_m f(a)\end{pmatrix}$$

and $I_m=diag(1,1,...,1)$ is the unit matrix of order m. If there exists $(I_m-R_{m\times m})^{-1}$, then the solution of the matrix in (12) is as follows:

$$J_{m\times1}=(I_m-R_{m\times m})^{-1}1_{m\times1}. \qquad (13)$$

To solve this set of equations for the approximate values of $j(a_1), j(a_2),..., j(a_m)$, the function $j^{IE}(x)$ can be approximated with $w_k=\dfrac{h}{m}$ and $a_k=\dfrac{h}{m}\left(k-\dfrac{1}{2}\right)$ as

$$j^{IE}(x)\approx 1+j(a_1)F(a-x+\phi_1\xi_{1-s})$$
$$+\sum_{k=1}^{m} w_k\, j(a_k)\, f(a_k+a-a_i+\phi_1\xi_{1-s}) \qquad (14)$$

The numerical scheme displays to evaluate solutions of the integral equations in (7) and (8) from section 3.1, which are compared with the approximate function $j(x)$ as in (14), by Gauss-Legendre quadrature rule. All results give a comparison of the approximated solutions $j^{IE}(x)$, the exact solutions $j(x)$, the absolute percentage difference

$$Diff\,(\%)=\frac{\left|j(x)-j^{IE}(x)\right|}{j(x)}\times100\%$$

for several values of a, h and the number of divisions m.

4. Comparison results

Tables 1 and 4 show a comparison of the approximate values $j^{IE}(x)$ of the ARL obtained from the integral equations with the exact solutions $j(x)$ obtained from the explicit formulae for both negative and positive values of the seasonal MA(1)$_s$.

On Table 1, the solution of ARL of CUSUM chart obtained from explicit formulas against numerical Integration Equation (IE) approximation for Seasonal MA(1)$_1$ model with exponential distribution white noise and s=1 are shown, where $\lambda_0=1$. The ARL is given to 100 and 370 when the process is in control for CUSUM chart which they are in good agreement. Notice that the absolute percentage difference is less than 0.1%. The computational time based on our technique takes less than one second while the numerical integral equation takes approximately 10-15 minutes.

On Table 2, the numerical solutions obtained from explicit formula and IE are compared Table 1. It is given the ARL =100 and 370, when the process is in-control and the parameter of Seasonal MA(1)$_s$ model with exponential distribution white noise and s=1 for CUSUM chart and the shift of parameters $\lambda=\lambda_0+\delta$, $\delta=0.1$, 0.3 and 0.5. Notice that $\lambda=\lambda_0=1$ is the value assumed for the in-control parameters, or the ARL and for $\lambda>\lambda_0>1$ corresponds to values of out-of-control parameters, or the ADT. It found that the analytical explicit solutions are in good agreement with results obtained from numerical integral equation approximation.

On table 3, the solution of ARL of CUSUM chart obtained from explicit formulas against numerical Integration Equation (IE) approximation for Seasonal MA(1)$_4$ model with exponential distribution white noise and s=4 are shown, where $\lambda_0=1$ at which they are in good agreement. Notice that the absolute percentage difference is less than 0.1%. The computational time based on our technique takes less than one second while the numerical integral equation takes approximately 10-15 minutes.

Table 1: ARL of CUSUM chart for Seasonal MA(1)$_1$ model with exponential distribution white noise

ϕ	h	Methods	$a=3.5$		$a=4$	
			$x=0$	$x=2$	$x=0$	$x=2$
0.23	0.38	$j(x)$	60.853	54.464	100.391	94.002
		$j^{IE}(x)$	60.381 (11.09[1]) 0.776[2]	54.400 (11.08) 0.118	100.353 (1.09) 0.038	93.967 (11.07) 0.037
	2.0	$j(x)$	299.580	293.191	499.366	492.977
		$j^{IE}(x)$	298.995 (11.16) 0.195	292.619 (11.20) 0.195	498.381 (11.09) 0.197	492.005 (11.09) 0.197
0.53	0.38	$j(x)$	110.959	104.570	183.001	176.612
		$j^{IE}(x)$	110.917 (11.12) 0.038	104.531 (11.35) 0.037	182.932 (10.88) 0.038	176.545 (11.12) 0.038
	2.0	$j(x)$	552.768	546.278	916.802	908.605
		$j^{IE}(x)$	551.676 (11.37) 0.198	545.299 (11.36) 0.179	914.981 (11.23) 0.199	908.605 (11.11) 0.199

[1] CPU time used (minutes) [2] *Diff* (%)

Table 2: ARL of CUSUM chart for Seasonal MA(1)$_1$ model with exponential distribution white noise when $\phi = 0.23$

δ	Methods	CUSUM(a,h)	
		(4, 0.38)	(4,2.0)
0.1	$j(x)$	66.008	281.652
	$j^{IE}(x)$	65.985 0.034[1]	282.154 0.178
0.3	$j(x)$	34.628	116.898
	$j^{IE}(x)$	34.618 0.028	117.071 0.148
0.5	$j(x)$	21.576	61.305
	$j^{IE}(x)$	21.571 0.024	61.381 0.125

[1] *Diff* (%)

Table 3: *ARL* of CUSUM chart for Seasonal MA(1)$_4$ model with exponential distribution white noise

ϕ	h	Methods	$a = 3.5$		$a = 4$	
			$x = 0$	$x = 2$	$x = 0$	$x = 2$
0.23	1.36	$j(x)$	100.111	93.722	166.614	160.225
		$j^{IE}(x)$	100.263 (11.29[1]) 0.152[2]	93.872 (11.38) 0.160	166.934 (10.13) 0.192	160.525 (11.17) 0.187
	2.72	$j(x)$	370.518	364.129	628.335	621.946
		$j^{IE}(x)$	371.181 (11.08) 0.179	364.839 (11.25) 0.195	629.579 (11.15) 0.198	623.115 (11.09) 0.188
0.53	1.69	$j(x)$	100.897	94.508	169.425	163.036
		$j^{IE}(x)$	101.036 (11.12) 0.138	94.637 (11.35) 0.137	169.659 (10.88) 0.138	163.261 (11.12) 0.138
	3.06	$j(x)$	370.78	364.391	640.463	634.074
		$j^{IE}(x)$	371.440 (11.27) 0.178	365.032 (11.26) 0.176	641.673 (11.34) 0.189	635.241 (11.27) 0.184

[1] CPU time used (minutes) [2] *Diff* (%)

Table 4: *ADT* of CUSUM chart for Seasonal MA(1)$_4$ model with exponential distribution white noise

δ	Methods	CUSUM$\left(a,h;\phi\right)$	
		(3.5, 2.72;0.23)	(3.5,3.06;0.53)
0.1	$j(x)$	212.326	210.51
	$j^{IE}(x)$	212.698 0.175[1]	210.876 0.174
0.3	$j(x)$	90.081	88.135
	$j^{IE}(x)$	90.223 0.158	88.266 0.149
0.5	$j(x)$	48.105	46.703
	$j^{IE}(x)$	48.165 0.124	46.761 0.125

[1] *Diff* (%)

On table 4, the solution of *ADT* of CUSUM chart obtained from explicit formulas against numerical Integration Equation (IE) approximation for Seasonal MA(1)$_s$ model with exponential distribution white noise and s=4 are shown, where $\lambda = \lambda_0 + \delta, \delta = 0.1$, 0.3 and 0.5. The results are also in good agreement.

It can be seen that the analytical explicit solutions are in good agreement with the results obtained from the numerical integral equation approach with 500 nodes in the integration rule. The computational times of the numerical integral equation approach take approximately 10-15 minutes while the results obtained from the explicit formula take less than 1 second which is much less than the former. To compare the values of $j(x)$ and $j^{IE}(x)$ obtained from the explicit formulae and the numerical approximations for varying values of the parameters, as in table 1 and 4, respectively. It found that the numerical results obtained from the IE approach have similar accuracy to the results obtained from the explicit formulae.

5. CONCLUSIONS

The *ARL* and *ADT* for the CUSUM chart when observations are seasonal MA(1)$_s$ with exponential white noise have been evaluated by two methods based on the integral equation approach. The integral equations have been solved by numerical methods, while the explicit formulas have been obtained the solutions. The results obtained from the two methods are in excellent agreement. The amount of time required of the numerical computations were approximately 10-15 minutes compared with less than one second for the explicit formulas. In addition, the results can easily be implemented in any computer program which is very useful for design of optimal CUSUM charts.

References

[1] Alwan, LC , Roberts HW. Time series modeling for statistical process control. J. Busi. Statis. 1988; 6: 87-95.

[2] Busaba J, Sukparungsee S, Areepong Y. An analysis of average run length for first order of autoregressive observations on CUSUM procedure. International Journal of Applied Mathematics and Statistics.. 2013; 34(4): 20-35.

[3] Busaba J, Sukparungsee S, Areepong Y. Numerical approximations of average run length for AR(1) on Exponential CUSUM, Proceeding of the International Multiconference of Engineers and Computer Scientists 2012 Vol II (IMECS 2012); 2012 March 14-16; Hong Kong. 2012. ISSN:2078-0966(online).

[4] Champ CW, Rigdon SE. A comparison of the markov chain and the integral equation approaches for evaluating the run length distribution of quality control charts. Communication Statistics: Part B Simulation and Computation. 1991; 20 : 191-204.

[5] Crowder SV. A simple method for studying run length distributions of exponentially weighted moving average charts. Technometrics. 1978; 29: 401-407.

[6] Harris TJ, Ross WH. Statistical process control procedure for correlated observations. Canadian Journal of Chemical Engineering. 1991; 69 ; 48-57.

[7] Hawkins DG, Olwell DH. Cumulative sum charts and charting for quality improvement. New York: Springer; 1998.

[8] Johnson RA, Bagshaw M. The effect of serial correlation on the performance of CUSUM tests. Technometrics. 1974; 16 : 103-112.

[9] Karaoglan AD, Bayhan GM. Performance comparison of residual control charts for trend stationary first order autoregressive processes. Gazi university journal of science, 2011; 24(2): 329-339.

[10] Knoth S, Schmid W. Control charts for time series. A review. In Frontiers in Statistical Quality Control (Edited by H.J.Lenze abd P'T' Wilrich). 2002; 7: 210-236.

[11] Lorden G. Procedures for reacting to a change in distribution. Annual Mathematics Statistics. 1971; 42: 1897-1908.

[12] Lucas JM, Saccucci MS. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics. 1990; 32: 1-29.

[13] Page ES. Continuous Inspection Schemes. Biometrika. 1954; 41: 100-114.

[14] Pallak M, Siegmund D. A diffusion process and its application to detecting a change in the drift of Brownian motion. Biometrika. 1985; 72: 267-280.

[15] Pallak M. Average run lengths of an optimal method of detecting a change in distribution. Annual Statistics. 1987; 15: 749-779.

[16] Petcharat K. An analysis of average run length for trend stationary first order of autoregressive observations on CUSUM procedure [Dissertation]. Bangkok: King's Mongkut University of North Bangkok; 2013.

[17] Rao BV, Disney RL, Pignatiello JJ. Uniqueness and converges of solutions to average run length integral equations for cumulative sum and other control charts. IEEE Transactions. 2001; 33(6): 463-469.

[18] Srivastava MS, Wu Y. Evaluation of optimum weights and average run lengths in EWMA control schemes. Communications in Statistics: Theory and Methods. 1997; 26: 1253-1267.

[19] Sukparungsee S, Novikov AA. On EWMA procedure for detection of a change in observations via martingale approach. KMITL Science Journal: An International Journal of Science and Applied Science. 2006; 6 : 373-380.

[20] Sukparungsee S, Novikov AA. Analytical approximations for detection of a change-point in case of light-tailed distributions. Journal of Quality Measurement and Analysis. 2008; 4(2): 49-56.

[21] Taylor HM. A stopped Brownian motion formula. Annual Probability. 1975; 3 : 234-246.

[22] Woodall WH, Adams BM. The statistical design of CUSUM charts. Quality Engineering. 1993; 4: 559-570.

[23] Woodal WH, Faltin F. Autocorrelated data and SPC. ASQC Statistics Division Newsletter. 1993; 13 : 18-21.

[24] Yashchin E. Performance of CUSUM control schemes for serially correlated observations. Technometrics. 1993; 35 : 37-52.

Fuzzy model for academicians

P.G. Khot
Department of Statistics, RTM Nagpur University, Nagpu, INDIA, pgkhot@gmail.com

Abstract

Worldwide National policies on higher education are giving increasing importance to improve the quality of education. Consequently, the evaluation of teachers' performance in teaching activity is especially relevant for the academic institutions. It helps to define efficient plans to guarantee quality of teachers and the teaching learning process. This paper contributes towards the estimates of teachers' performance in one of the Engineering Institutes in Nagpur (INDIA). Here we make use of fuzzy concept. This Fuzzy Evaluation Model considers the various aspects of the performance of teachers, like personal skills, academic development, teaching learning process, other performance & abilities, Students' attendance, Students Feedback and Results. Considering all indicator values we introduce fuzzy values and define membership function (Trapezoidal function) the performance are calculated for all the involving factors. Using all these values as input values and considering the output membership functions we turn towards the inference process. Ultimately at last we need the defuzzification (Performances) and we find the output values which are direct and calculated by Fuzzy technique.

We observed the difference in the direct value and the values determined by using fuzzy model. This is due to the weightage given on some important parameters related to teaching learning process and overall development of the institute while framing the rules. So the overall performance of a teacher determined by fuzzy model is more realistic than the direct values.

Teachers' regular assessment is suggested to maintain quality in higher education. There is a vast potential of the applications of fuzzy expert system (FES) in teachers' assessment. Expert system technology using Fuzzy Logic is very interesting for quantitative and qualitative facts evaluation. A model fuzzy expert system (FES) is proposed to evaluate teachers overall performance on the basis of various activities tending towards the perfection. The qualitative variables are mapped into numeric results by implementing the fuzzy expert system (FES) model through various input examples and provided a basis to use the system for further decision making. In this way the teaching staff is encouraged to reflect on quality, adequacy, satisfaction, efficiency and innovation in teaching.

Corresponding Author
E-mail Address: *pgkhot@gmail.com*

Variant of constants in subgradient optimization method over planar 3-index assignment problems

Sutitar Maneechai

Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, 90112, Thailand, sutitar.m@psu.ac.th

Abstract

Planar 3-index assignment problem is an NP-complete problem. Its global optimal solution can be found by a branch and bound algorithm. The efficiency of the algorithm depends on the best lower and upper bounds. Subgradient optimization method is one of the methods which can provide a good lower bound for the problem. The method can be applied to the root node or leave of the branch and bound tree. Some conditions used may lead the method converts to optimal solution. Computational experiments are used for finding a suitable exact value for some constants in the method. This paper shows computational results of applying a modified subgradient optimization method in order to find a good lower bound for planar 3-index assignment problem. The method consists of a few steps with some constants whose effect the lower bound value. A variant of initial step length constants is focused in these experiments. The best initial step length constant for small problem size (n<20) is rather to be chosen between 0.1 and 1. For the bigger problem size, 0.075 is the best initial step length constant has been found by the experiments.

Keywords: Planar 3-index assignment problem, NP-complete, lower bound, upper bound, subgradient optimization method, step length constant

Corresponding Author
E-mail Address: sutitar.m@psu.ac.th

Effects of forced response and question display on web survey's completion rate

Chatpong Tangmanee[1]* and Phattharaphong Niruttinanon[2]

[1]*Chulalongkorn Business School, Chulalongkorn University, Bangkok 10330, Thailand, chatpong@cbs.chula.ac.th*

[2]*Chulalongkorn Business School, Chulalongkorn University, Bangkok 10330, Thailand, phattharaphong@gmail.com*

Abstract

A web survey has gained remarkable acceptance, especially among social science researchers. Previous studies have examined factors contributing to a completion rate. However, virtually no empirical work has examined the effects of forced responses and question display together on a web survey's response rate. The current study attempted to fill this gap. Using a quasi experiment approach, we obtained 778 unique responses to six (i.e., 3 levels of forced responses x 2 styles of question display) comparable web questionnaires of identical contents. The analysis confirmed that (1) the effect of forced responses on the completion rate was statistically significant at a 0.05 level but (2) the effect of question display on it was not significant. In addition to extending the theoretical insight into factors contributing to a web survey's completion rate, the findings have offered recommendations to enhance the completion rate in a web survey project.

Keywords: Forced response; question display; scrolling; paging; completion rate; web survey

*Corresponding author
E-mail address: chatpong@cbs.chula.ac.th

1. Problem statement and research objectives

Online questionnaires are tools social science researchers have adopted to gather data from samples through major web browsers. The increasing number of publications have addressed issues on how to implement a survey using online questionnaires because they have certain advantages and limitations [14, 15].

Compared to the offline counterpart, online questionnaires offer three major advantages. They include (1) a small amount of error in recording the collected data into a file since the data were saved as soon as a sample responded to questionnaire items, (2) quick data analysis and data collection processes because of the Internet's worldwide accessibility, and (3) a cost-justified survey on a general topic since researchers could reach a large group of targeted sample. However, an online survey project do have two limitations that researchers must have a proper plan to minimize prior to starting the data collection. First, an online survey always reaches only the Internet users. If the project's target population taps those whose profiles are not largely shared with the Internet user profiles, researchers may have to give up the online version. Second, given the Internet nature, the samples' responses may not be the same as when the paper-based questionnaires are used. Such responses include those from the same subject or from unqualified samples. They could immensely distort the finding's validity and reliability. As a result, researchers may have a set of screening questions to eliminate the unqualified samples or check the samples' IP address of their online responses. If two responses of the same IP address are given within a short period of time interval, researchers may have to pay close attention to all responses from that IP address [1].

One quality check of a survey project is through a completion rate. It is the number of completed responses divided by the total number of responses. It also indicates the extent to which samples are determined to respond to the entire questionnaires. Using an online survey, it is fairly difficult for a researcher to design the questionnaire that could retain a sample's focus so he or she could respond to the entire questionnaire [8]. Dillman [6] remarked that there is no single design solution to gain the sample's attention throughout the answering session. The researcher must be attentive to the holistic detail of the design in order to increase the completion rate. Polonsky and Vocino [13] suggested based on their experiment that old and employed subjects were more likely to complete web-based questionnaires than the young or unemployed samples.

Among many attempts to examine factors affecting an online survey's completion rate, forced response and question display are of our interest since no previous attempt has addressed them in the same study. Forced responses refer to an online survey execution through which a sample is reminded to answer to a questionnaire item, if he or she has missed it. This feature is impossible in a traditional paper-based survey. With certain programmability, it is easy to detect any missing questionnaire items and forcibly remind the sample to answer. The sample could not proceed to the next step unless he or she must respond to the missing item. The “forced choice” style is similar to the forced response design but they are notably different. The forced choice refers to the survey design that suggests possible choices of answer to which a subject could respond. For instance, a researcher may adopt a four-level scale (e.g., least, less, more, or most), instead of a typical five-level scale (e.g., least, less, neutral (or average), more, or most). This is how a researcher forces a subject to agree to certain choices. Yet, the forced response is the design that requires a sample to respond to a questionnaire item, it is left unanswered. The item could be a Likert scale or an open-ended question. The sample can not proceed to the next step unless he or she respond to it.

Based on the experiment approach, Derouvray and Couper [5] discovered that the forced-response condition had lower performance than did the no-forced condition. Similarly, Stieger and colleagues [18] conducted a survey on students’ well-being issues in Europe and confirmed that the forced-response increased the number of survey dropout. In addition, male samples dropped out faster than female subjects. The poor performance empirically supports Dillman’s [6] statement in which the forced-response may be so annoying that samples would want to give up the survey participation or even turn off a web browser. However, Albaum and colleagues [1] failed to offer empirical evidence of which the forced answering could have lowered a completion rate. Its effect of the forced responses is still inconclusive.

How to display questionnaire items to attract samples’ attention and to retain it until they submit the completely-filled-in questionnaires to a researcher has gain remarkable attention [3, 8]. Presenting a too-wide table on a web-based survey led to more dropouts than a simple one [11]. Yan and colleagues’ [20] experiment verified that the presence of a progress indicator (i.e., visual feedback information to tell samples how far they had responded to the survey questionnaires) led to fewer dropouts only when the questionnaire length was perceived short. Recently, a survey of radiologists validated that the long questionnaire was not a problem as long as incentives were justified [21].

One of the design guidelines for question display is choosing between scrolling and paging layouts. The scrolling style displays the entire questionnaire in one single webpage. It thus requires a sample to scroll down while completing it. The paging style displays the questionnaire in many webpages requiring the samples to “flip” to the next page or the next section. The flip could be through a click on the next or the continue buttons. According to Dillman [6], the scrolling design demands less computer resources because it requires one single submission of the questionnaire. Moreover, the samples are able to scroll back to review their responses since it appears solely on one webpage. However, the paging design allows different structures of the same questionnaire. In other words, the paging design could have outperformed the scrolling style if the response to each questionnaire item is non-linear [10]. For instance, the samples whose responses to the gender item are male would have to answer the different section of the questionnaire, as compared to those who replied female to the gender item. Nevertheless, previous research could not verify if the scrolling is better or worse than the paging design [7, 12]. This is perhaps why Elliott and colleagues [7] suggested the hybrid version combining the scrolling and paging styles. The hybrid design requires a researcher’s great effort to balance the number of questionnaire webpages and the amount of up-down scrolling. Das and coworkers [4] remarked that a display of one questionnaire item per page could draw a sample’s attention to the survey; yet, they failed to verify their remark. In addition, a few projects were unsuccessful to substantiate similar statements [2].

A review of previous literature indicated two gaps for possible research. First, while a large amount of research has examined a response rate, a relatively small portion has investigated a completion rate. The response rate could show the percentage of those submitting their survey responses compared to the total number of those who happened to visit the first page of the online questionnaire. Given the context of web survey, researchers should have extended their effort to cover the completion rate. Second, there is virtually no experiment investigating the effect of varying degrees of forced responses (e.g., 100%-, 50%- or 0%-forcing) or different styles of question display (e.g., scrolling and paging) on a web survey’s completion rate. As a result, the current study’s objectives were to (1) compare the completion rate of responses to web surveys using 100%-, 50%- and 0%-forced responding conditions, and (2) compare the completion rate to web surveys with scrolling and paging design of the question items.

2. Research methodology

To achieve the study’s objectives, this section describes five methodology issues. They are the research approach; experimental units; questionnaire content and experimental execution; reliability and validity issues; and data analysis framework and hypothesis statements.

2.1 Research approach

Given the study’s casual investigative style, we strived to adopt the quasi-experimental approach. The two independent variables are (1) forced responses and (2) question display. The forced-response variable has three possible values. They are (a) 100%-forced responses (i.e., subjects are forced to respond to every questionnaire item), (b) 50%-forced response (i.e., subjects are forced to respond to half of all items) and (c) 0%-forced response (i.e., subjects are free to leave any questionnaire items unanswered). The selection of the 100% and 0% forced categories was challenged by previous studies [1, 6]. The 50% choice of forced response was added to see if the forced response should not be dichotomy and researchers may want to force the responses only to a few items.

The question-display variable has two possible values. They are (a) scrolling and (b) paging styles of display. Using the scrolling design, the entire questionnaire appears in one webpage. On the contrary, the paging style would display a few questionnaire items per page (the detail of questionnaire development will be in the next section).

The dependent variable is the completion rate measured by dividing the number of questionnaires that all items are answered by the number of returned questionnaires.

2.2 Experimental units

Given the quasi experiment approach, the participants must not only represent the target population, but also share large compatibility such that the difference of completion rate, if any, is due to the two independent variables, not to the subjects' incompatibility.

We were fortunate to receive assistance from the Stock2morrow website. They allowed us to invite their subscribers to participate in our experiment. Based on the six conditions (3 levels of forced responses x 2 styles of question display), the number of subjects per condition should be at least 30-40 samples [16]. The Stock2morrow website administration agreed that we posted a call for research participation on the website. Within the two-month data collection, 778 unique visitors to the website took part in the experiment, of which each condition had about 128-132 experimental subjects.

2.3 Questionnaire content and experimental execution

The six experimental conditions require similar questionnaires of identical content. The difference among these questionnaires must be from the manipulation of the two independent variables. The Stock2morrow website requested that the questionnaire content should help them to improve the website and perhaps their business. We therefore included in the questionnaire the total of 36 question items asking subjects about their demographics, web usage, their life style, and their reaction when they were asked to participate in an online survey.

Given the three levels of forced responses, those in the 100%-forced-response group were forced to answer all 36 items. We forced those in the 50% group to answer every other items. The subjects in the 0% group were not forced to respond to the items they may have missed.

Regarding the question display, the entire questionnaire appears in one webpage for the scrolling style. For the paging style, the questionnaire was divided into four sections, each of which appeared in one page. The total number of pages for the paging style is thus four pages.

Once the six versions of questionnaires with identical content were crafted, we pretested them on twelve graduate students in Chulalongkorn Business School and made a few adjustments. When all questionnaires are ready, we posted during June – July 2012 messages inviting subscribers to the Stock2morrow website to respond to the questionnaires. We randomized a subject to one of the experiment's six conditions. When the first six samples had placed in all six conditions, the next subject was again randomized to one of the six conditions. We repeated the process until the number of samples in each condition exceeded the minimum threshold of 50 subjects.

At the end of the end of July 2012, all questionnaires were replied 912 times but when duplications were removed, we had the total of 778 usable records for further analyses.

2.4 Reliability and validity issues

We strived to respond to the two objectives validly and reliably. Such efforts include the followings.

- The questionnaires were carefully crafted and thoroughly tested to ensure their acceptable quality. To minimize a chance of duplicating participation, we developed a session to control the number of responses. We did not keep track of any IP address as suggested in previous online research. Such tracking may have reduced the number of responses from those who might have shared the same computer stations and the IP address.
- To motivate the Stock2morrow website subscribers to take part in our experiment, we explained in a call for research participation that their participation is critically important to the research community. Also, to show our appreciation toward their participation, we offer a lucky draw at the end of the project to win an iPad mini.
- To conform with an approach in using an experiment in social science research, we selected the condition of the scrolling style with no forced-response as the control group. The selection was because it is typically a default design available in many online questionnaire services.
- The pretest was deemed useful. It helped us to learn about technical incompatibility and prepare for it. During the pretest, we discovered few flaws in the data we collected. They occurred when pretest samples used different browsers or worked on

diverse platforms to do our questionnaires. To minimize the chance of such differences, we popped up a window suggesting the samples a set of acceptable choices.

2.5 Data analysis framework and hypothesis statements

We used descriptive statistics to describe the sample's demographics. The two hypotheses are (1) the difference of completion rates across three levels of forced responses, or (2) the difference of completion rates between the two styles of display each is statistically significant. The hypothesis testing was through the z test of proportion and it is two-tailed.

3. Analysis results

The demographic profile of the experimental subjects are in Table 1. The followings are the highlight.

- Most of experimental subjects are male, 36-55 years of age with at least college degrees. 26% of those who submitted questionnaires (the largest portion) earn monthly income of 10,000-24,999 baht.
- We expected to have the Stock2morrow website subscribers as the participants since we posted an invitation to take part in our project on the website. However, only 39% of the participants claimed they are the subscribers. The remaining who submitted the questionnaires admit they were not.

According to Tables 2 and 3, 24.3% of 778 who had accessed to the questionnaires provided complete responses. Given three levels of the forced-response condition, the samples who responded to the 100%-, 50%- and 0%-forced responding categories are 44.7%, 17.7% and 10.7%, respectively. Regarding the scrolling and paging styles of question display, the percentages of those who responded to the questionnaires are 23.4% and 25.2%, respectively.

Results of the hypothesis testing are in Table 4. They confirm statistically significant differences of the completion rates among the three levels of forced responses. Based on Table 2, it seems that the more forced the sample received, the higher the completion rates were. However, the difference of completion rates between the two styles of question display is not significant (see Table 4).

Table 1: Profiles of experimental subjects

Profiles	N(%)
Gender (N=470)	
Male	323(69)
Female	147(31)
Age (N=471)	
Less than 23 yrs	42(9)
24-35	184(39)
36-55	210(45)
At least 56 yrs	35(7)
Highest education (N=468)	
Less than college	32(7)
College degrees	287(61)
Graduate level	149(32)
Monthly salary in Thai baht (N=463)	
Less than 10,000	38(8)
10,000-24,999	118(26)
25,000-39,999	99(21)
40,000-54,999	81(18)
55,000-69,999	46(10)
70,000-100,000	29(6)
Higher than 100,000	52(11)
Whether a subject subscribes to the Stock2morrow website (N=474)	
Already a subscriber	184(39)
Not yet a subscriber	290(61)

Table 2: Complete responses categorized by three levels of forced responses

Issues	Levels of forced responses			Total
	100%-forced	50%-forced	0%-forced	
Number of complete responses	115	46	28	189
Number of questionnaire distributed	257	260	261	778
Percentages of complete responses over distributed questionnaires	44.7	17.7	10.7	24.3

Table 3: Complete responses categorized by two styles of question display

Issues	Question display styles		Total
	Scrolling	Paging	
Number of complete responses	91	98	189
Number of questionnaire distributed	389	389	778
Percentages of complete responses over distributed questionnaires	23.4	25.2	24.3

Table 4: Results of hypothesis testing

Comparing pair	Z-Statistics,	Significant level
100% - vs. 50%-forced responses	6.642	.000
100% - vs. 0% -forced responses	8.659	.000
50% - vs. 0% -forced responses	2.277	.000
Scrolling vs. paging styles	-.585	.278

4. Conclusion and discussion

The study attempted to examine the effect of three levels of forced responses and two styles of questionnaire display on a completion rate in one online survey project. Via a call for research participation posted on the Stock2morrow website, the subjects were asked to respond to one of the six comparable questionnaires of identical contents. Most of the subjects are men, in between 24-35 years of age and largely holding at least college degrees. Yet, 6 in 10 have not as yet subscribed to the Stock2morrow website. Although the demographics appear to confirm the representative samples of Stock2morrow subscribers or those of the Internet users in Thailand, it could be premature to have such a claim since only 4 in 10 of the participants are current website subscribers. Readers must therefore be cautious when using the study's findings.

The differences of completion rates among the samples who submitted complete questionnaires with varying degrees of forced responses were statistically significant (see Table 2). Based on the findings, the more the samples were forced to answer, the higher proportion of complete questionnaires they returned. This finding contradicts to those in previous work [1, 5, 19] in which the forced responding category was found as effective as the no-forced group. In addition, Stieger and colleagues [18] found that the dropout in the forced category was higher than that in the no-forced counterpart. Nevertheless, Polonsky and Vocino [13] discovered that the old and employed subjects responded to online questionnaires better than the young or unemployed samples. Given that our samples are mostly over 30 years old and having a career, it is thus reasonable that the 100%-forced group had the highest completion rate as compared to the other two groups that has less degree of the forced response. Such speculation is however highly uncertain and urging more empirical research to verify it.

An attempt to validate the effect of question display styles on completion rate has failed since the rates on the scrolling and the paging styles are about the same. Such trivial finding replicates results in previous studies [10, 12, 17]. The possible explanation could be that our questionnaires with the paging style consisted of only four pages, each of which did not fit in one screen. As such, the samples may have not perceived much different from the ones with the scrolling design. The second possible explanation of the nonsignificance is based on Peytchev and colleagues' [12] experiment. The trivial finding in theirs was due to the long survey. In other words, the paging style should have led samples to (1) perceive that the questionnaire is relatively short and (2) subsequently make an effort to respond to it completely. However, if the paging design does not lead to the perceived short survey (e.g., in the current study, each of the four pages in the paging design was so long that the scroll bar popped up on the right edge of the screen), we would assume that the samples might not have perceived the short questionnaire; thereby, were not motivated to give complete answer. Such explanation still needs more empirical verification.

Our findings offer theoretical and practical contribution. Theoretically, our findings have extended insight into the design features of online questionnaire in the Thai context. Practically, we could offer two recommendations. First, researchers who attempt to use an online channel to gather data using a questionnaire may increase the number of forced-responding items since it may lead to the high completion rate. Second, researchers may not need to choose between the scrolling or the paging style of question display for no significance between the two styles. However, in a project where a certain group of samples must respond to certain groups of questions, the paging style should still be a researcher's choice [6]. This is because the responses to such project are non-linear. Only if all samples must answer the same set of survey items (i.e., the responses are on a linear pattern) should the scrolling style be considered.

Similar to other research projects, the current study have two limitations. First, we adopted the quasi experiment approach, Although the external validity is acceptable, the internal validity must be inevitably compromised. Second, our conclusion is substantial only among the subscribers to the

Stock2morrow website. As a result, a generalization across other contexts may be made with high caution.

References

[1] Albaum, G, Roster, CA, Wiley, J, Rossiter, J, Smith SM. Designing web surveys in marketing research: Does use of forced answering affect completion rates? Journal of marketing theory and practice. 2010; 18(3): 285-293.

[2] Batagelj, Z, Manfreda, KL, Vehovar, V. Design of Web Survey Questionnaires: Three Basic Experiments. Journal of computer-mediated communication. 2002.

[3] Crowford, SD, McCabe, SE, Pope, D. Applying web-based survey design standards. Journal of prevention and intervention in the community. 2005; 29: 43-66.

[4] Das, M, Soest, AV, Toepoel, V. Design of web questionnaires: The effects of the number of items per screen. Field Methods. 2009, 21(2).

[5] Derouvray, C, Couper, MP. Designing a strategy for reducing "no opinion" responses in web-based surveys. Social science computer review. 2002; 20(3): 3-9.

[6] Dillman, DA. Mail and Internet surveys: The tailored design method. Imprint New York : J. Wiley: 2007.

[7] Elliott, MN, Fricker, RD, Schonlau, M. Conducting research survey via e-mail and the web. CA:RAND: 2002.

[8] Fan W, Yan, Z. Factors affecting response rates of the web survey: A systematic review. Computers in human behavior. 2009; 26: 132-139.

[9] Galesic, M, Bosnlak, M. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public opinion quarterly. 2009: 73(2): 349-360.

[10] Norman, KL, Friedman, Z, Norman, K, Stevenson, R. Navigational issues in the design of on-line self-administered questionnaires. Behavior & information technology. 2001; 20(1): 1-14.

[11] O'Neil, KM, Penrod, SD, Bornstein, BH. Web-based research: Methodological variables' effect on dropout and sample characteristics. Behavioral research methods, instruments, & computers. 2003; 35(2): 217-226.

[12] Peytchev, A, Couper, MP, McCabe, SE, Crawford, SD. Web survey design: Paging versus scrolling. Public opinion quarterly. 2006: 70(4): 596-607.

[13] Polonsky, M, Vocino, A. Survey completion speed of online panelists: The role of demographics and experience. Proceeding of the 2010 Australian and New Zealand Marketing Academy Conferences, NZ; 2010.

[14] Reips, U-D. Standards for Internet-based experimenting. Experimental psychology. 2002; 49(4): 243-256.

[15] Reips, U-D, Dtieger, S, Voracek, M. Forced-response in online surveys: Bias from reactance an increase in sex-specific dropout. Journal of the American society for information science and technology. 2007; 58(11).

[16] Roscoe, JT. Fundamental research statistics for the behavioral sciences. 2nd edition. New York: Holt, Rinehart and Winston. 1975.

[17] Sedley, A, Callegaro, M. Effects of pagination on short online surveys. Proceedings of The American association for public opinion research (AAPOR), 67th Annual Conference, 2012.

[18] Stieger, S, Reips, U-D, Varocek, M. Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. Journal of the American society for information science and technology. 2007; 58(11): 1653-1660.

[19] Swan, J E, Epley, DE. Completion and response rates for different forms of income questions in a mail survey. Perceptual and motor skills. 1981: 52: 219-222.

[20] Yan, T. Conrad, FG, Tourangeau, R, Couper, MP. Should I stay or should I go? The effects of progress indicators, promised duration, and questionnaire length on completing web surveys. Proceedings of The American association for public opinion research (AAPOR), 62nd Annual Conference, 2007.

[21] Ziegenfuss, JY, Niedergauser, BD, Kallmes, D, Beebe, TJ, An assessment of incentives versus survey length trade-offs in a web survey of radiologists. Journal of medical Internet research. 2013; 15(3): 3 pages.

Effect of some cover crops on soil moisture content under water deficit condition at Mahidol University Kanchanaburi Campus

Nuengruithai Tharawatcharasart*, Suravoot Yooyongwech and Chai Rukkachat

School of Interdisciplinary Studies Mahidol University Kanchanaburi Campus, Kanchanaburi 71150, Thailand, nuengruithai.tha@mahidol.ac.th

Abstract

Currently rainfall is not seasonally and is causing drought in many areas that leads to water scarcity in agriculture. One of methods to save conserving soil and water is to grow the cover crops such that Leguminosae is very popular for this purpose. Karopo (*Calopogonium mucunoides*), Kudzu (*Pueraria phaseoloides*) and Pinto (*Arachis pintoi*) were experimented and compared to study the morphology of the bean roots by applying the manga farming (Koshi Kawachi, 2008) in the Completely randomized design (CRD). The analysis of variance for mean differences of the morphology of the bean roots is significantly different ($p < 0.05$). It is found that Pinto beans have the highest growth rate followed by Karopo and Kudzu, respectively.

Keywords: Completely randomized design, cover crops, manga farming, morphology, water scarcity

*Corresponding Author
E-mail Address: nuengruithai.tha@mahidol.ac.th

Solar 48 quality control analysis in Pusdiklat Cepu refining

Purwo Nur Hidayat
Mathematics Department, Universitas Gadjah Mada ,3404, Yogyakarta, Indonesia,
purwo.nur.h@mail.ugm.ac.id

Abstract

Pusdiklat Migas Cepu is a corporation which has petroleum refining which can produce over 5.000 barrels per day. This corporates also produce crude oil processed called solar. General Directory of Oil and Gasoline's Decisions released specification about solar characteristic in 2006. Pusdiklat Migas Cepu has to obey the specification of production before the products distribute to each parts of Indonesia, the products must be under control and fulfill the specification. Quality control analysis has a big contribution to solve this specification problem, such as out-of-control data. Because it gives an absolute conclusion about product's specification. This analysis give control chart which reflects how that product is in control, however the capability processes also can be shown in this method. And if the process capable, it means that the process can work to the next production. If the process isn't capable, the production must be evaluated because it gives the worst quality of product.

Keywords : Quality control, specification, control chart, capability processes

*Corresponding Author
E-mail Address: purwo.nur.h@mail.ugm.ac.id

1.Introduction

a. Company Profile

Pusat Pendidikan dan Pelatihan Minyak dan Gas Bumi Cepu (Education and Training Centre of Oil and Gasoline Cepu) or shortened to Pusdiklat Migas Cepu, is an institute which under authority of Indonesia's Ministry of Energy and Mineral Resources. This institute refines petroleum and crude gases. This corporation has petroleum refining which can produce over 5.000 barrels per day.

Based on Keputusan Direktur Jenderal Minyak dan Gas Bumi No. 3674 K/24/DJM/2006 (General Directory of Oil and Gasoline's Decisions) about standard and specification of petroleum, production of this corporate on under control or not. Pusdiklat Migas Cepu has two oil refining, Kawengan (for crude oil paraffin specialist) and Ledok (for crude oil asphalt specialist). All refining and central office are located in Central Java, Indonesia.[8]

This place corporates product called solar, pertasol CA, pertasol CB, pertasol CA and residue. But in this research only focusing in solar production. Solar or gas oil is fuel for diesel machine for vehicles like bus, truck, train and tractor.

Before resulting solar, crude oil must pass some processes,

1. Atmospheric Distillation Process

This process separate crude oil into each fractions base on boiling point difference in 1 atm. pressure.

2. Treating Process

This process decrease or eliminate impurities in the product.

Solar characteristics which determined by General Directory of Oil and Gasoline's Decisions No. 3674 K/24/DJM/2006, as follow in Table 1.

Table 1. General Directory of Oil and Gasoline's Decisions about solar characteristics

No	Characteristic	Unit	Limit	
			Min	Max
1	Cetana rate - Cetana digit - Cetana index	 - -	 48 45	 - -
2	Density (in 15°C)	kg/m^3	815	870
3	Viscosity (in 40°C)	mm^2/s	2	5
4	Sulphur content	%m/m	-	0,35
5	Distillation : T 90	°C	-	370
6	Flash Point	°C	60	-
7	Pour Point	°C	-	18
8	Carbon residue	%m/m	-	0,1
9	Water content	mg/kg	-	500
10	*Biological growth*	-	None	
11	FAME content	%v/v	-	10
12	Metanol & etanol content	%v/v	Not detected	
13	Copper corrosion	Merit	-	1st class
14	Ash content	%m/m	-	0,01
15	Sediment content	%m/m	-	0,01
16	Strong acid rate	mg KOH/g	-	0
17	Total acid rate	mg KOH/g	-	0,6

18	Particulate	mg/l	-	-
19	Visual presentation	-	Purify & bright	
20	Color saybolt	No. ASTM		3

This work focused on controlling density, distillation, flash point, pour point and color saybolt.

2 Stastistical Quality Control

In industrial sector, there are so much product variations. And these products have qualities. From quality, we can consider about good or not that product is.

Many expertise explain about qualities, there are :

- Juran (1980), quality means fitness for use[11]
- Crosby (1983), Quality is conformance to requirement[2]
- Feigenbaum (1985), Quality is the total composite product and service characteristics of a marketing, engineering, manufacture, and maintenance through which the product and service in use will meet expectation by customer.[3]
- Taguchi (1986), Quality is the loss imparted to society from the time product is shipped.[10]

By several opinions about quality, the conclusion is quality have goal to fulfill costumer's need.

Quality can be determined by seven tools, there are checksheet, pareto diagram, ishikawa diagram, box-plot diagram, histogram-steam and leaf, defect concentration diagram and control chart.

2.1 Control Chart

In this section, only explain quality process by control chart specially $\bar{X} - R$ chart, because characteristic of data and capability process need. Control chart describe stability of a process work, and monitoring operational process and production. This method also will quickly anticipate uncontrolled process.[6]

2.2 $\bar{X} - R$ *Control Chart*

$\bar{X} - R$ chart is quantity control chart which use in small data (n < 10). Given quality characteristic with normal distributed with mean (μ) and deviation standard (σ). If x_1, x_2, …,x_n is sample with n range, thus the mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

The upper and lower control limit are :

$$\mu + Z_{\alpha/2}\sigma_{\bar{x}} = \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and}$$

$$\mu - Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = \mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If μ and σ, were not given, mean estimated by total mean, $\bar{\bar{x}} = \frac{\sum_{i=1}^{m} \bar{x_i}}{m}$. so the central line for the control chart.

The alternative calculation for UCL (Upper Control Limit) and LCL (Lower Control Limit) can be :

UCL $= \bar{\bar{x}} + \frac{3}{d_2\sqrt{n}}\bar{R}$

Central line $= \bar{\bar{x}}$

LCL $= \bar{\bar{x}} - \frac{3}{d_2\sqrt{n}}\bar{R}$

Control limits for R :

UCL $= \left(1 + \frac{3\,d_3}{d_2}\right)\bar{R}$

Central line $= \bar{R}$

LCL $= \left(1 - \frac{3\,d_3}{d_2}\right)\bar{R}$

With :

$\bar{R} = \frac{\sum_{i=1}^{m} R_i}{m}$

R = $x_{maks} - x_{min}$

$d_2 = \frac{\bar{R}}{\sigma}$

If there are data out from the control limit or uncontrolled, it must be eliminated and to be analyze again until all data under control.

2.3Capability Process

A process can call capable if

1. Process under control
2. Require all limits
3. High precision and accuracy

Capability process index can estimated by :

$$Cp = \frac{USL - LSL}{6\sigma}$$

Where :

USL : Upper Specification Limit
LSL : Lower Specification Limit
σ : deviation standard

There are 3 probability of Cp score :

1. Cp < 1, means specification limit smaller than observes. Indicate that process in bad condition.
2. Cp = 1, means specification limit equal with observes. But, need to increasing quality
3. Cp > 1, means specification higher than observes. Indicate that the process in good condition[9]

Performance Process Index can be estimated by :

Cpk = minimum {Cp_A, Cp_B }

where

$Cp_A = \frac{USL-\mu}{3\sigma}$ $\quad Cp_B = \frac{\mu-LSL}{3\sigma}$

There are 7 conditions for score of Cpk :

1. Cpk < 0, means process mean out of specification
2. Cpk = 0, means process mean equal with one of specification
3. Cpk < 0,9, means process uncontrolled. It must be re-analyzed
4. 0,9 ≤ Cpk < 1 , means diverge product not always exist. But must be eliminated
5. 1 ≤ Cpk < 1,1, means a little movement should be diverge product
6. 1,1 ≤ Cpk < 1,3, means variance in the limit
7. Cpk ≥ 1,3, means small probability to diverge from specification

3. Research methodology

All data about solar characteristic were collected daily by Pusdiklat Migas Cepu, Indonesia since December 2013 to January 2014.

Research methodology show on Fig. 1, consist of these following steps :

1. Data entry
2. Apply to control chart
3. Check the chart, if there are out-of-control data, eliminate until all data in control.
4. Normality assumption, if the assumption not met, fix it by transform data
5. Do capability process
6. Analyze which might really important, such as the ‘capable’ point of each characteristic

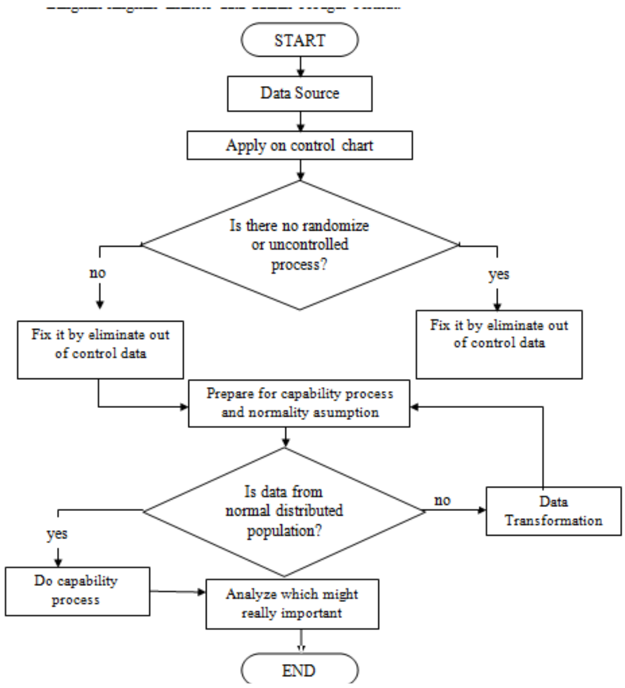

START
Data Source
Apply on control chart
Is there no randomize or uncontrolled process?
no
yes
Fix it by eliminate out of control data
Fix it by eliminate out of control data
Prepare for capability process and normality assumption
Is data from normal distributed population?
no
Data Transformation
yes
Do capability process
Analyze which might really important
END

Fig. 1. Flowchart of Research Step

4. Research results

By MINITAB 14, the result of quality control analyze of characteristic Solar 48 are :

4.1 Density 15˚C kg/m^3

Is owned by the specific weight of a substance in a given volume, the characteristics related to the calorific value and the power generated by the diesel engine fuel per unit volume. Diesel fuel density was measured at a temperature of 15 ° C using ASTM method D-1298.

Fig. 2 and Fig. 3, describe control chart for Density 15˚C, before and after uncontrolled data handling.

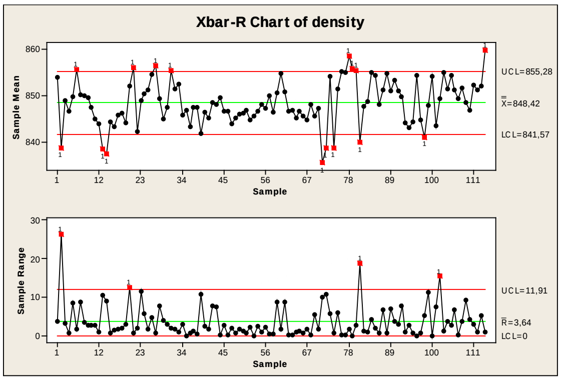


Fig. 2. Control Chart for Density 15˚C

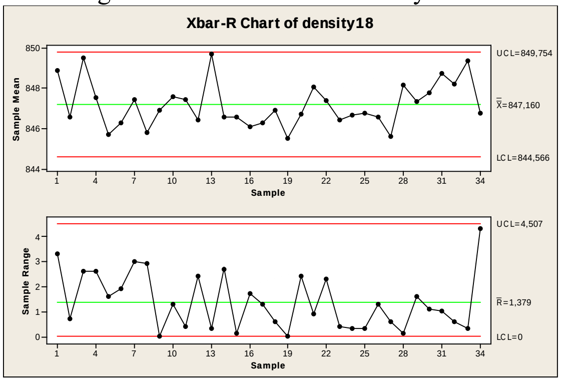


Fig. 3. Control Chart for Density 15˚C (after uncontrolled data handling)

After all data under control, we can proceed to capability process. According to Table 1, LSL = 815 kg/m^3 and USL = 870 kg/m^3 .

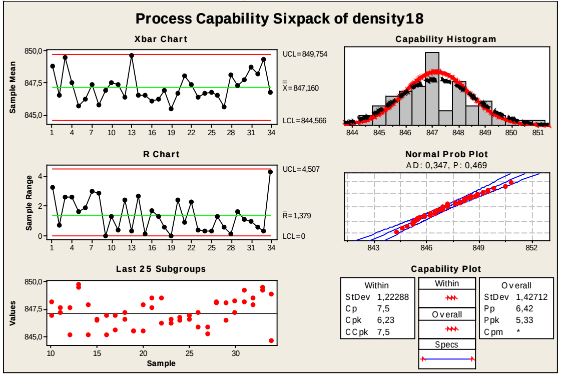


Fig. 4. Capability Process for Density 15˚C

By Figure 4, the conclusions are :

- $Cp = \frac{870-815}{6(1.22288)} = 7.5$, means specification higher than observes. Indicate that the process in good condition
- $Cpk = \min\left(\frac{870-847.16}{3(1.22288)}, \frac{847.16-815}{3(1.22288)}\right)$
  $= min(6.23\,, 8.766) = 6.23$
  means small probability to diverge from specification

4.2 Distillation T 90

is a chemical separation methods based on differences in the speed or ease evaporate (volatility) material. The mixture was boiled substance that evaporates, and the vapor is then cooled back into a liquid form. Substances which have a lower boiling point will evaporate first. This method is included as an operating unit of chemical mass transfer types.

Fig. 5 and Fig. 6 describe control chart for Distillation T 90, before and after uncontrolled data handling.

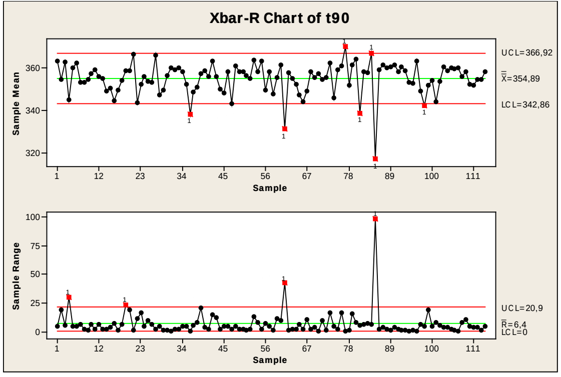


Fig. 5. Control Chart for Distillation T 90

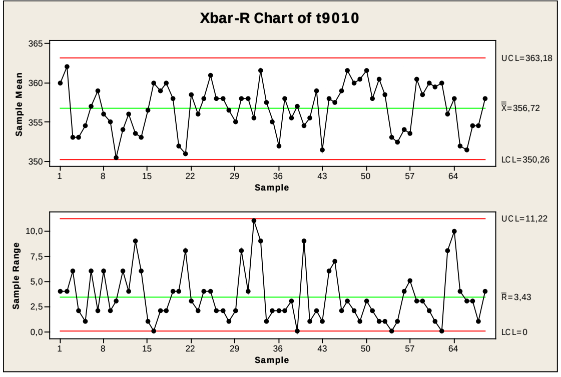


Fig. 6. Control Chart for Distillation T 90 (after uncontrolled data handling)

After all data under control, we can proceed to capability process. According to Table 1, USL for Distillation T 90 is 370˚C.

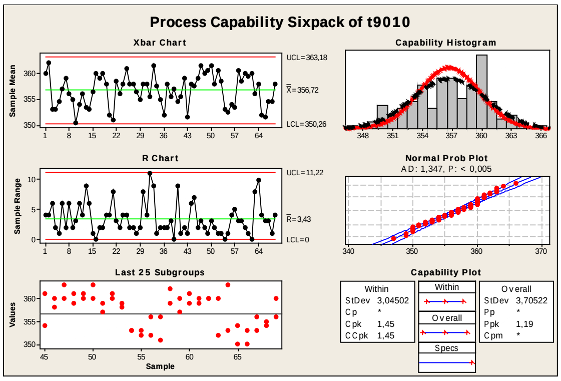


Fig. 7. Capability Process for Distillation T 90

By Figure 7, the conclusion is :

- Cp can't be calculate because there is no LSL
- Cpk = $\min\left(\frac{370-356.72}{3(3.04502)}, \frac{356.72-(-\infty)}{3(3.05402)}\right)$

$= min(1.45\ ,\infty) = 1.45$

, means small probability to diverge from specification. Indicate that the process in good condition.

4.3 Flash point ˚C

is the lowest temperature to which a fuel will undergo ignition when brought near a flame.

Fig. 8 and Fig. 9 describe control chart for Flash point, before and after uncontrolled data handling.

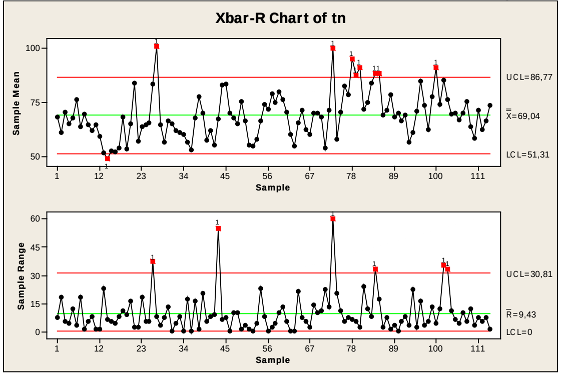


Fig. 8. Control Chart for Flash Point

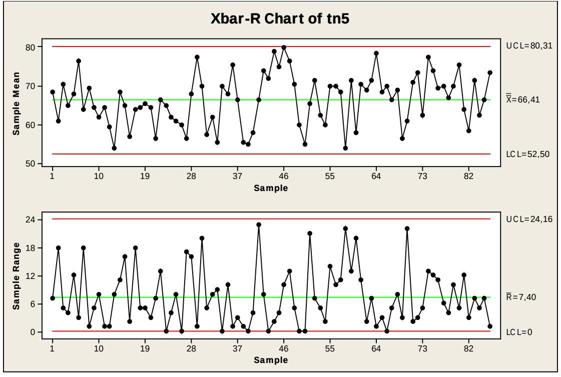


Fig. 9. Control Chart for Flash Point (after uncontrolled data handling)

After all data under control, we can proceed to capability process According to Table 1, LSL for Flash Point is 60˚C.

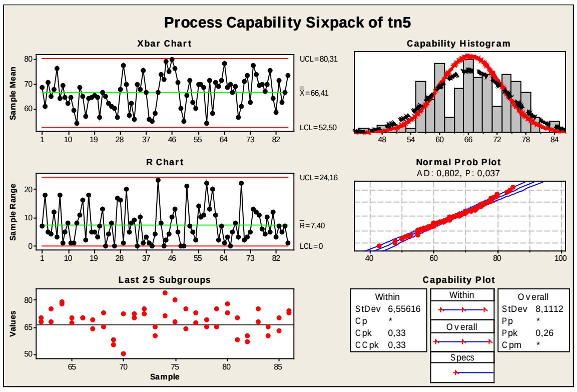


Fig. 10. Capability Process for Flash Point

By Figure 10, the conclusion is :

- Cp can't be calculate because there is no USL
- Cpk = $\min\left(\frac{\infty-66.41}{3(6.55616)}, \frac{66.41-60}{3(6.55616)}\right)$

$= min(\infty, 0.33) = 0.33$

, means process uncontrolled. It must be re-analyzed

4.4Pour point ˚C

is a number that states the lowest temperature of the fuel oil so that the oil can still flow due to gravity. Value pour point is required in connection with the practical requirements of the procedure stockpiling and use of fuel oil. This is because the fuel oil is pumped fatherly often difficult when the temperature has been below its pour point.

Fig. 11 and Fig. 12 describe control chart for Pour point, before and after uncontrolled data handling.

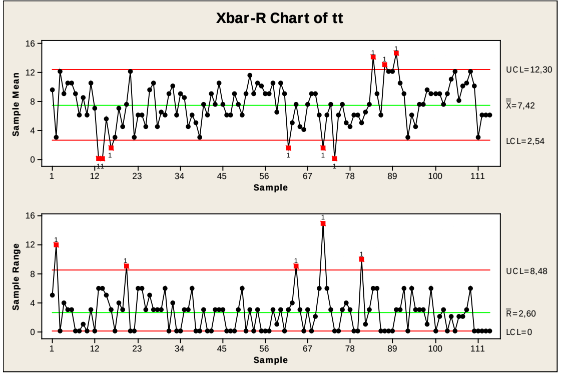


Fig. 11. Control Chart for Pour Point

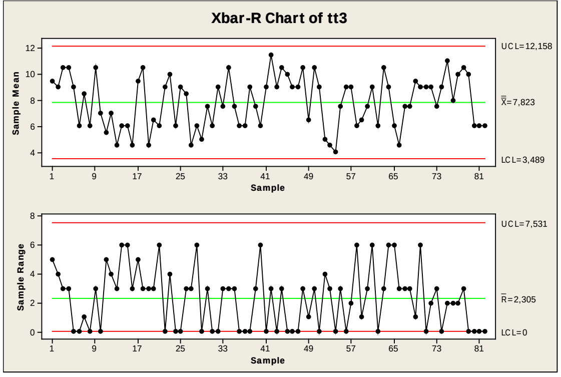


Fig. 12. Control Chart for Pour Point (after uncontrolled data handling)

After all data under control, we can proceed to capability process. According to Table 1, USL for Pour Point is 18˚C.

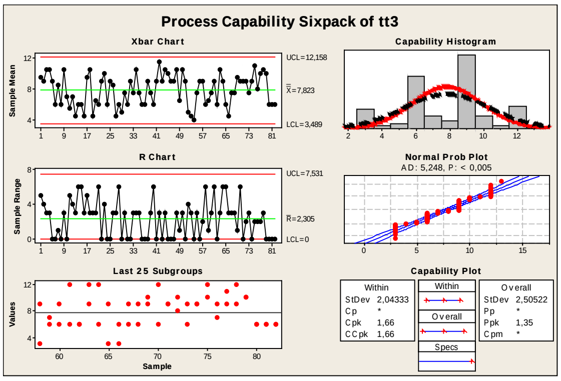


Fig. 13. Capability Process for Pour Point

By Figure 13, the conclusion is :

- Cp can’t be calculate because there is no LSL
- Cpk = $\min\left(\frac{18-7.823}{3(2.04333)}, \frac{7.823-(-\infty)}{3(2.04333)}\right)$

  $= min(1.6, \infty) = 1.6$

  means small probability to diverge from specification

4.5 Color Saybolt

Is an observation using the standard color comparator which the fuel compared to the existing standard color, when the color of the fuel beyond the standard specifications then there is the possibility of Solar mixed with other materials.

Fig. 14 and Fig. 15 describe control chart for Color saybolt, before and after uncontrolled data handling.

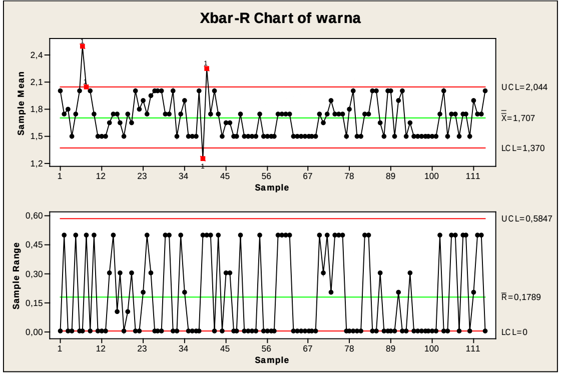


Fig. 14. Control Chart for Color Saybolt

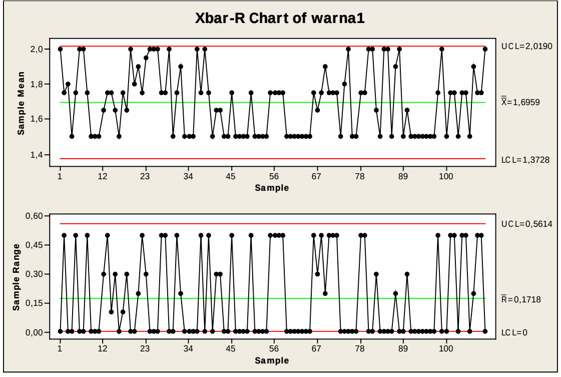


Fig. 15. Control Chart for Color Saybolt (after uncontrolled data handling)

After all data under control, we can proceed to capability process. According to Table 1, USL for Color Saybolt is 3 No. ASTM.

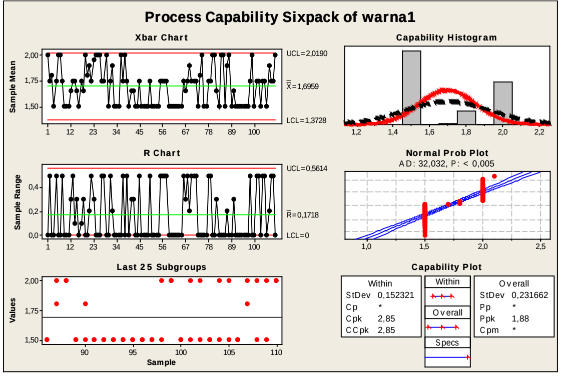


Fig. 16. Capability Process for Color Saybolt

By Figure 16, the conclusion is :

- Cp can’t be calculate because there is no LSL
- Cpk = $\min\left(\frac{3-1.6959}{3(0.152321)}, \frac{1.6959-(-\infty)}{3(0.152321)}\right)$

  $= min(2.85, \infty) = 2.85$

  means small probability to diverge from specification

5. Conclusion

Based on discussion that has been done in the previous section, these are the conclusions

Characteristic	Cpk	Keterangan
Density 15°C	6,23	Capable
Distillation T 90	1,45	Capable
Flash Point	0,33	Capable , but need to increase quality
Pour Point	1,6	Capable
Color Saybolt	2,85	Capable

Table
* Based on General Directory of Oil and Gasoline's Decisions in 2006

According to the result, to make the characteristic become capable, reduce error or variance in production. Make the production straight in mean, can minimize variance and minimize out-of-control data.

Acknowledgements

This work was supported by Mathematics Department, Universitas Gadjah Mada (UGM) and Pusat Pendidikan dan Pelatihan Minyak dan Gas Bumi, Cepu , Indonesia (PUSDIKLAT MIGAS CEPU)

References

[1]. Anonymous. Modul Pengendalian Kualitas Statistik, Laboratorium Komputasi Statistika Jurusan Matematika FMIPA UGM, Yogyakarta. 2011

[2]. Crosby, Phillip . Quality Without Tears . New York : McGraw-Hill. 1983

[3]. Feigenbaum, Armand V .Total Quality Control , New York : McGraw-Hill. 1983

[4]. Goetsch, David L. & Stanley B. Davis . Quality Management for Organizational Excellence . Pearson Higher Ed USA

[5]. Hidayat, Purwo Nur. Analisis Pengendalian Kualitas Statistik Solar 48 Kilang Pusdiklat Migas Cepu, Jawa Tengah. Program Studi Statistika, FMIPA UGM. 2014

[6]. Montgomery, Douglas C.. Introduction into Statistical Quality Control (6th edition), Wiley. 2008

[7]. Noviani, Laras Nur,. Analisis Kualitas Ethanol Prima I Produksi Pabrik Spiritus PT. Madubaru, Yogyakarta dengan Statistical Quality Control. Program Studi Statistika FMIPA Universitas Gadjah Mada, Yogyakarta. 2013

[8]. Rahmawati , Febrian D. . Analisis Kualitas Air Minum Produksi Pusdiklat Migas Cepu dengan Statistical Quality Control . Program Studi Statistika FMIPA Universitas Gadjah Mada, Yogyakarta. 2012

[9]. Sari, Trya . N. Implemetasi Statistical Quality Cotrol terhadap Produksi Pertasol CA di Unit Kilang PMC, Jateng , Program Studi Statistika FMIPA Universitas Gadjah Mada, Yogyakarta. 2012

[10]. Taguchi, Genichi . Introduction to Quality Engineering . Asian Productivity Organization. 1986

[11]. Utami, Susi. Statistical Quality Control Solar 48 Hasil Olahan Kilang Minyak Pusdiklat Migas Cepu Jawa Tengah. Program Studi Statistika FMIPA Universitas Gadjah Mada, Yogyakarta. 2012

The fuzzy FMEA method to improve the defects in sanitary ware manufacturing process

Pinnarat Nuchpho[1], Santirat Nansaarng[2] and Adisak Pongpullponsak[3]*

[1]*Department of Learning Innovation and Technology, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, pinnacomsc@gmail.com*

[2]*Department of Production Technology Education, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, santirat.nan@kmutt.ac.th*

[3]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, iadinsak@kmutt.ac.th*

Abstract

The purpose of this study was to find the affecting factors of the detective products in the process of plated bath products type page settle (K-160) of sanitary ware. The problem was discovered scratches on the surface of the plated bath products. The causes of the scratches had been analyzed via Total Quality Management (TQM) by using pareto chart, cause and effect diagram, and Failure mode and effects analysis (FMEA) techniques for continuous improvements in product and determining the Risk Priority Numbers (RPN), which indicate the levels of risks associated with potential problems. These numbers are generally attained from past experience and engineering judgments, and this way of risk assessment sometimes leads to inaccuracies and inconsistencies during priority numbering. Fuzzy logic approach is preferable in order to remove these deficiencies in assigning the risk priority numbers. A fuzzy FMEA is applied to reduce the defects on sanitary ware products. The results indicate that the application of fuzzy FMEA method has arisen from traditional FMEA and they can solve the problems and efficiently discover the potential failure modes and effects.

Keywords: Total quality management (TQM), failure mode and effects analysis (FMEA), risk priority number (RPN), fuzzy logic, fuzzy FMEA

*Corresponding Author
E-mail Address: iadinsak@kmutt.ac.th

1. Introduction

At present, Thailand industrial business is regarded as one of the key factors in country development and the manufacturing process, both domestically and internationally in the industry sector has higher competition. The organization which will succeed and survive in a strong competitive situation needs the capability of decreasing or controlling the production cost which can help them in this situation. From the above reasons, the discovery and production process development are created in order to make the least losses and errors with tools such as samplings for examination and final inspection [1]. When the production system goes into mass production, the checking of product's quality cannot solve the quality problems as soon as possible. Therefore, the organization should manage the system which protect the errors and create the quality from every stages of production [1], [2], [3]. The most popular strategy, which also is regarded as the core strategy, lies in the quality of their products. It must be constantly acceptable by the markets and consumers. One of the most important principles is the quality control (QC) principle [4], [5], [6], [7], [8].

In the production process, which is typically lacks in continuously monitoring the quality of the products. Sometimes a lack of sufficient quality control occurs because of outsourcing a part of production to other companies, thus resulting in a large number of defective parts of the products [1], [2], [3]. For example, for ordering sanitary ware factories to plate some parts of the sanitary ware, they have to hire outside companies to do the plating. When this occurs, the factory has no control over the quality of the products. Even with an initial attempt to resolve this issue, the control standard system remains insufficiently.

Such problems lead to our research in identifying causes of problems and priority number of the problems associated with defective products. Total quality management (TQM), Risk Priority Number (RPN) and fuzzy FMEA will be utilized and analyzed for what influencing an occurrence of the product loss and identifying the root of the problem. Also, it will be used to seek ways to reduce a number of defective product and to increase the effectiveness of the production process.

2. Research Methodology

In this study, there are steps in studying the problems as follows:

2.1 Study the operation's condition of production process and current problems

On this step, the study starts from the entire production procedure and collect the tendency of errors from the production process to make the Pareto Diagram [3] and choose the main problem to improve it.

2.2 Review Literature

In this research, it aims at studying defects reduction on production process so it needs basic studying and surveying to find out the instrument which can be used for applying the organization's condition and situation appropriately. After that, review the related research to find out the research instrument whether it can be used for reducing the defects.

2.3 Analysis the cause by using Cause and Effect Diagram

When there is the production flow of each problem, they will be used as the components for writing the Cause and Effect Diagram [3]. On this step, the team is selected from related organization to discuss together on risk assessment and cause of problem by using brainstorming. These errors can occur in every stage of production by showing the data of the errors' characteristics occurred.

2.4 Failure mode and effect analysis (FMEA)

2.4.1 Analysis the errors by using FMEA technique

FMEA is a tool for solving problems in the production process, to analyze the problem, process failure and perdition effects of factors [9]. From collecting all occurred errors and possible causes in the form of Cause and Effect Diagram, it will be analyzed to find out the Risk Priority Number (RPN). Then, analyze the solution starting from the cause which has the most RPN to the least RPN. The guideline for risk assessment is gained from these 3 components; Severity (S), Occurrence (O) and Detection (D). So the RPN is received from RPN= S x O x D [10], [11], [12], [13].

Traditional FMEA uses five scales and 10 scores of 1-10, to measure severity, the probability of occurrence and the probability of not detection by asking an analyst or an expert to assign scores ranging from 1 to 10 for the different factors. After the errors are analyzed with FMEA technique, they are collected and considered RPN. Then, RPN is categorized from the most to the least.

2.5 Fuzzy approach to FMEA

Fuzzy Logic is the method to manage the uncertainty; it is the possibility theory which emerged from the Fuzzy set [14]. The fuzzy logic variables may have a membership value of not only 0 or 1, but a value inclusively between 0 and 1 [15]. Thus, the fuzzy logic provides a basis for approximate reasoning, that is, a mode of reasoning which is not exact or very inexact. Fuzzy logic is described for the following steps

2.5.1 Determining linguistic variable for input and output

Linguistic variables are the input or output variables of the system whose values are words or sentences from a natural language, instead of numerical values. To determine variable can be helped for explaining for human communication and it will be translated to numerical value for processing data by using membership function.

2.5.2 *Procedure of converting data to Fuzzy logic relationship of input and output*

This procedure is to convert input to be Fuzzy logic and build up the membership function by finding the model to cover the receiving data. The number of set terms of each variable should be determined in order to have the deduction of output value be alike the real data the most. This relationship between variables of data is related with the membership function using in this study. There are different forms of membership functions such as triangular, trapezoidal, piecewise linear, Gaussian, or singleton.

Fuzzy logic concept can be expressed as follows. Let X be a nonempty set. A fuzzy set A in X is characterized by its membership function $\mu_A : X \to [0,1]$ and $\mu_A(x)$ is interpreted as the degree of membership of element x in fuzzy set A for each $x \in X$.

It is clear that A is completely determined by the set of tuples $A = ((u, \mu_A(u)) / u \in X)$. Frequently $A(x)$ is used instead of $\mu_A(x)$. The family of all fuzzy sets in X is denoted by $F(X)$. If $X = (x_1, ..., x_n)$ is a finite set and A is a fuzzy set in X then the following notation is often used.

$$A = \frac{\mu_1}{x_1} + ... + \frac{\mu_n}{x_n} \tag{1}$$

Where the term μ_1 / x_1 , i = 1,..., n signifies that μ_1 is the grade of membership of x_1 in A and the plus sign represents the union. From among various membership functions the triangular one is to be exemplified below.

A fuzzy set A is called triangular fuzzy number with peak a, left width $\alpha > 0$ and right width $\beta > 0$ if its membership function has the following:

$$A(t)=\begin{cases}1-(a-t)/\alpha & \text{if } a-\alpha\le t\le a\\ 1-(t-a)/\beta & \text{if } a\le t\le a+\beta\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

and we use the notation $A=(a,\alpha,\beta)$. A triangular fuzzy membership function is depicted in Figure 2:

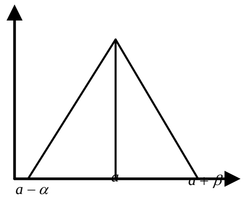


Figure 1: Triangular fuzzy membership function

A triangular fuzzy number with center a may be seen as a fuzzy quantity “ x is approximately equal to a ”

2.5.3 *Deduction of the Fuzzy logic relationship from Mamdani-style Inference*

Deduction of the fuzzy logic relationship from Mamdani-style Inference by using fuzzy rule composed of IF and THEN. The conditional value IF will have some of Fuzzy. THEN value is assessed by the relationship. The conditional value of IF and THEN can have various values and every condition of IF will be assessed together in the same time by AND. Generally, in the system the number of conditional values is limited by choosing necessary rules. The number of rules condition depends on the variables in the system. In this study, the deduction of Mamdani-style Inference [16] is used and it composes of 4 steps as follows:

2.5.3.1 Fuzzification

To do fuzzification is to calculate the value of member of relationship level of linguistic variables set. In this step, it needs to find out the value of member of relationship level of input and the input is in the form of numerical value. After that, the value of member of relationship level can be obtained from the membership function.

2.5.3.2 Fuzzy rule evaluation

After calculating the value of member of relationship level for the input variable, then assess the obtained variables by Fuzzy rule in order to assess the conditional value of input and what rules will be used in THEN because there might have more conditional rules in the same time according to more input and condition of each input assessed by fuzzy set, for example, AND to obtain the final result which is numerical value and THEN is assessed for seeking out the value of member of relationship level for the output variable.

2.5.3.3 Aggregation

After the other rules are assessed and the rule which is not equal zero, it will be assembled together by membership function. As for output, it will be assembled by union and use the result of aggregation to convert into single number for data processing.

2.5.3.4 Defuzzification

Defuzzification is the step of converting of aggregation in the form of crisp value by using Center of gravity, COG) [17], [18], [19]. Centroid defuzzification method finds a point representing the center of gravity of the fuzzy set, A, on the interval, ab

$$COG=\frac{\int_a^b \mu_A(\mathrm{X})\,xdx}{\int_a^b \mu_A(\mathrm{X})\,xdx} \qquad (3)$$

2.5.3.5 Consideration of improving error

After finishing Fuzzy Assessment, it will be considered by crisp value categorized from the most to the least.

2.5.4 Comparison

A comparison is made between the ranking orders of the traditional FMEA and fuzzy FMEA methods.

3. Research Results and Discussion

3.1 Identifying the main problem

Looking at various parts of defective product in the production process, the most defective parts were the plated bath products type page model K-160. This problem then was chosen to be the focus of this research study. We then considered types and numbers of defective parts, as shown in Table 1.

Table 1: The percentage of defect [20]

Type	Feature of defective product	Number of defective product	Percentage of total detective product [%]
1	Scratched	17,654	49.95
2	Blistered	11,826	33.46
3	Spotted	5,443	15.40
4	Stuck to the mold	255	0.72
5	Not fully injected	166	0.47
	Total	35,344	100.00

From Table 1 the most defective parts were scratches found on the surface of plated bath products type K-160, with a percentage of 49.95 of a total number of defective parts. It was thus the highest priority for the manufacturer to resolve the issue and Pareto Chart show in Figure 2.

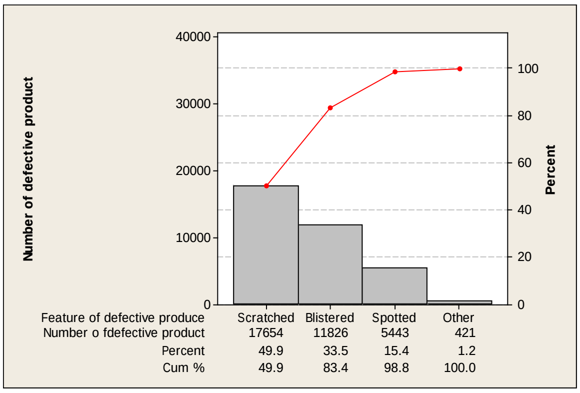


Figure 2: Pareto Chart of Feature of defective product

3.2 Finding what caused the problem

The brainstorming technique was used to find what caused the scratch problem. The chiefs of the mold-making and the quality-control departments, as well as operating workers participated in a brainstorming session. These groups of people were specialized in operating the machines, and had direct experiences with the problem of defective parts. The brainstorm was not limited to quantitative or qualitative thoughts, and everyone involved were allowed to share their ideas freely. All of the ideas shared during the brainstorm were summarized in a fishbone diagram shown in Figure 3.

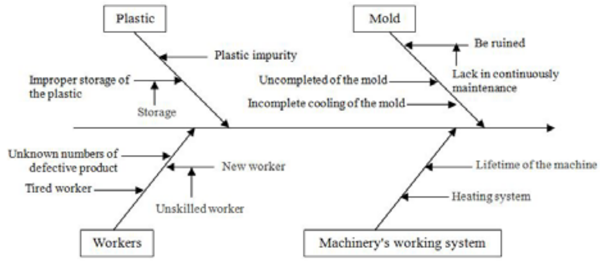

Plastic
Mold
Plastic impurity
Be ruined
Improper storage of the plastic
Uncompleted of the mold
Lack in continuously maintenance
Storage
Incomplete cooling of the mold
Unknown numbers of defective product
New worker
Lifetime of the machine
Tired worker
Heating system
Unskilled worker
Workers
Machinery's working system

Figure 3: The cause-effect fishbone diagram for plated bath products type page settle (K-160) of scratched sanitary ware

According to the fishbone diagram in Figure 3, there were too many preliminary items which might not be the actual cause of the problem. Thus the process of relating brainstormed items must be executed in order to quantify how relevant each item was. Prioritizing causes of a problem was done by using Risk Priority Number (RPN). The RPN takes into account the most relevant components which include the severity of the cause (S), the frequency of the occurrence (O), and the probability of detecting the problem (D). The resulting RPN which led to scratched parts from various causes is shown in Table 2.

Table 2: The Risk Priority Number (RPN), for each cause

Cause	S	O	D	S × O × D	Rank
1. Dented mold	10	10	5	500	1
2. Incomplete cooling of the mold	10	8	5	400	2
3. Unknown numbers of defective product	10	10	3	300	3
4. Tired worker	8	5	5	200	4
5. Uncompleted of the mold	10	5	3	150	5
6. Improper storage of the plastic	10	5	3	150	5
7. Lifetime of the machine	5	5	5	125	7
8. Unskilled worker	8	5	3	120	8
9. Uncompleted heating system	8	5	3	120	8
10. Plastic impurity	8	5	3	120	8

The number RPN enabled us to prioritize leading causes of the problem, which are due to 1) dented mold, 2) incomplete cooling of the mold, and 3) unknown number of defective product. This finding gave clues for the brainstorming session and thus was used for planning out an approach to solve the problem systematically. The issue with the highest RPN was firstly taken care of, followed by those with the next highest RPN and Pareto Chart show in Figure 4.

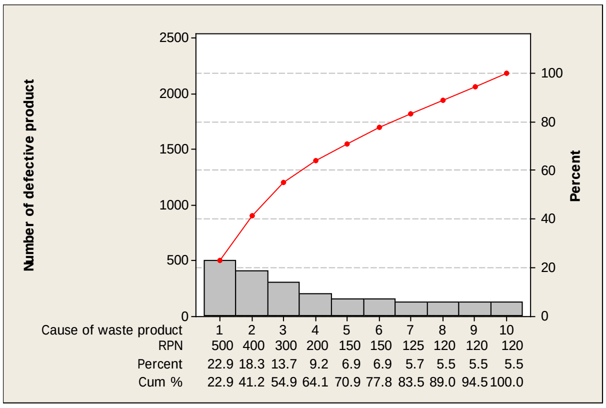


Figure 4: Pareto Chart cause of defective product

3.3 Fuzzy FMEA method

A model was established for the FMEA technique having 3 inputs and 1 output variable. The RPN values were calculated by combining the associated 3 inputs. For the input (severity, occurrence and not detection) and output variables triangular membership functions were used as shown in figure 5-6, membership functions for input values have the 10-level scale that divided into 5 different regions. Being represented by triangular membership function, these sub-regions respectively are almost none, low, medium, high and very high. For the output variable RPN the 10-level scale ranging from 0,1,2,…,10 mean none, very low, low high low, low medium, medium, high medium and high respectively. The severity, occurrence and not detection values of the failures were identified with the help of expert opinions and by using a database of 125 decision rules (see Appendix A) determined specifically.

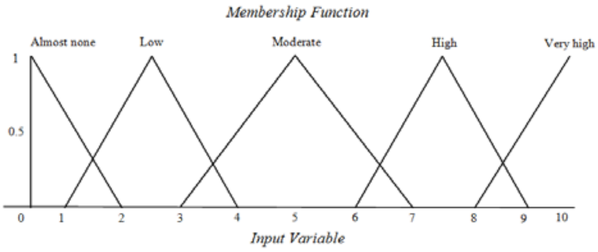


Figure 5: Input variables membership function.

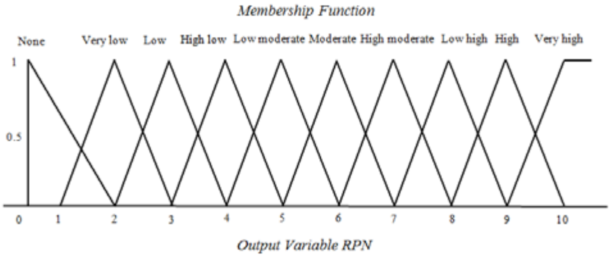


Figure 6: Output variables membership function.

Mamdani min/max method of inference mechanism (input method: min; aggregate method: max) was used and the results were defuzzified by center of gravity method.

As to the types of failure, the fuzzy RPN values provided in the model are given in a descending order in Table 3 in comparison with the RPN values of traditional FMEA. The failure types containing the same RPN values were arranged according to the values of severity, occurrence and not detection. The average number of Fuzzy RPN was found to be 5.60 (moderate–High moderate).

Table 3: Ranking of failure modes

Failure mode	RPN	Ranking	Fuzzy RPN	Ranking
1. Dented Mold	500	1	9.0	1
2. Incomplete cooling of the mold	400	2	7.0	3
3. Unknown numbers of defective product	300	3	8.0	2
4. Tired worker	200	4	5.0	5
5. Uncompleted of the mold	150	5	4.0	6
6. Improper storage of the plastic	150	5	4.0	6
7. Lifetime of the machine	125	7	7.0	3
8. Unskilled worker	120	8	4.0	6
9. Uncompleted heating system	120	8	4.0	6
10. Plastic impurity	120	8	4.0	6

3.4 Comparison

In this section, a comparison is made between the ranking orders of the traditional FMEA and fuzzy FMEA methods. In Table 4, the results can be seen the main problem in the traditional FMEA methodology is that it puts two critical sub-assemblies of the uncompleted of the mold and improper storage of the plastic as having the same priority. The unskilled worker, uncompleted heating system and plastic impurity are also placed at the same ranking level. But, applying the proposed methodology reveals that there is a noticeable difference between their ranking orders. On the other side, there is some noticeable difference between the ranking orders of some sub-assemblies (such as Unknown numbers of defective product and Tired worker) using the traditional FMEA and the Fuzzy FMEA methods. From the Table 4, create scatterplot for comparison between traditional FMEA and the fuzzy FMEA method.

Figure 7 shows the scatterplot of comparison between the ranking traditional FMEA and fuzzy FMEA method which the x-Axis is a failure mode of sanitary ware products ranging from 1, 2, 3, ..., 10 mean dented mold, incomplete cooling of the mold, unknown numbers of defective product, tired worker, uncompleted of the mold, improper storage of the plastic, lifetime of the machine, unskilled worker, uncompleted heating system and plastic impurity respectively. For the y-axis, there are 8 ranking, which show a comparison of the traditional FMEA (solid line) and fuzzy FMEA (dotted lines). When using fuzzy logic, it was found that the ranking would be changed by changing the most obvious that was the failure mode at 6, 7, 8 (improper storage of the plastic, lifetime of the machine, unskilled worker). When the ranking of failure mode is compared between traditional FMEA and fuzzy FMEA, it is found that the incomplete cooling of the mold will change from ranking 2 into ranking 3, the unknown numbers of defective product changes from ranking 3 into ranking 2, the tired worker changes from ranking 4 into ranking , the uncompleted of the mold and the improper storage of the plastic changes from ranking 5 into ranking 6, the lifetime of the machine changes from ranking 7 into ranking 3, the unskilled worker, the uncompleted heating system and plastic impurity changes from ranking 8 into ranking 6. However, the root mean square error of fuzzy FMEA (0.618) is less than the traditional FMEA (0.823) and failure mode of fuzzy FMEA which obviously changes; that is improper storage of the plastic, lifetime of the machine, unskilled worker, uncompleted heating system, and plastic impurity which have lower ranking traditional FMEA. Therefore it doesn’t show the certainty of the traditional FMEA characteristics enough. These results show that a more accurate ranking can be achieved by the application of the fuzzy FMEA.

Table 4: Ranking comparisons between traditional FMEA and the fuzzy FMEA method

Rank	Traditional FMEA	Fuzzy FMEA
1	Dented Mold	Dented Mold
2	Incomplete cooling of the mold	Unknown numbers of defective product
3	Unknown numbers of defective product	Incomplete cooling of the mold/ Lifetime of the machine
4	Tired worker	
5	Uncompleted of the mold/ Improper storage of the plastic	Tired worker
6		Uncompleted of the mold/ Improper storage of the plastic/ Unskilled worker/ Uncompleted heating system/ Plastic impurity
7	Lifetime of the machine	
8	Unskilled worker/ Uncompleted heating system/ Plastic impurity	
9		
10		

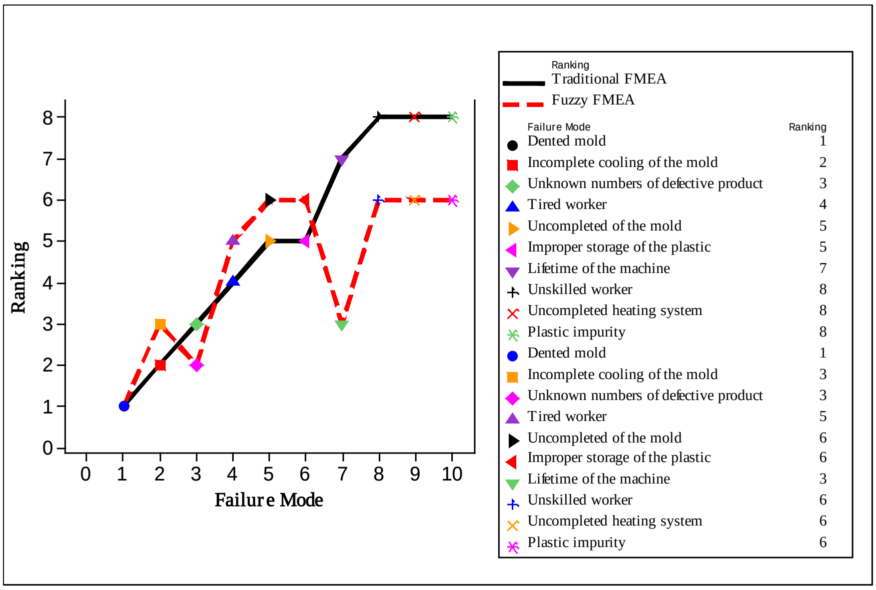


Figure 7: Scatterplot of comparison between the ranking traditional FMEA and fuzzy FMEA method

4. Conclusion

FMEA has been extensively used for examining potential failures in products. In the failure modes are determined and can be evaluated by risk factors like severity, occurrence and detection but the FMEA method has been criticized to have several deficiencies so this study applied fuzzy theory as an appropriate tool to eliminate the conversion debate by directly evaluating the linguistic assessment of factors to obtain RPN and to improve the defects in sanitary ware can be summarized as follows:

- The proposed fuzzy FMEA approach can be useful when the failure data is unavailable or unreliable.
- The use of linguistic terms in the analysis enables the experts to express their judgments more realistically and hence improving the applicability of the FMEA technique.
- Fuzzy FMEA method has arisen from conventional FMEA, and can solve the problems and can efficiently discover the potential failure modes and effects. It can also provide the stability of product and process assurance.
- Fuzzy FMEA approach might be helpful to the management processes in manufacturing and service sectors it is quite possible to use this technique successfully.

Acknowledgements

The authors express sincere appreciation to the editor and reviewers for their constructive comments and suggestions which are very helpful in improving the quality of the paper. This work was partially supported by Learning Industry Project of Assoc. Prof. Adisak Pongpullponsak, Savitree Thongrattanatai and Siraprapa Taechamaneerat.

References

[1] Montgomery DC. Introduction to Statistical Quality Control, 6th ed. United States of America: John Wiley & Sons; 2009.

[2] Rawlins RA. Total quality management. Central Milton Keynes: Author House UK, Ltd; 2008.

[3] Pongpullponsak A. Statistical Quality Control. Bangkok: Jaransanitwong kanpim Co.,Ltd; 2013.

[4] Chung YC, Tien SW, Hsieh CH, Tsai CH. A study of the business value of Total Quality Management. TQM. 2008; 19: 367-379.

[5] Klefsjö B., Bergquist B., Garvare R. Quality management and business excellence, customers and stakeholders: Do we agree on what we are talking about, and does it matter? The TQM Journal. 2008; 20(2): 120 - 129.

[6] Salaheldin SI. Critical success factors for TQM implementation and their impact on performance of SMEs. International Journal of Productivity and Performance Management. 2009; 58(3): 215-237.

[7] Juneja D, Ahmad S, Kumar S. Adaptability of Total Quality Management to Service Sector. International Journal of Computer Science & Management Studies. 2011; 11(2): 93-98.

[8] Talib F, Rahman Z, Qureshi MN. Total Quality Management in Service Sector: A literature Review. International Journal of Business Innovation and Research. 2012; 6(3): 259-301.

[9] Stamatis DH. Failure mode and effect analysis: FMEA from theory to execution. New York: ASQC Press; 1995.

[10] Palady P. Failure Modes and Effects Analysis: Predicting & Preventing Problems Before They Occur. West Palm Beach, FL: PT Publications; 1995.

[11] Pillay A., Wang J. Modified failure mode and effects analysis using approximate reasoning. Reliability and System Safety. 2003; 79: 69-85.

[12] Chang KH. Evaluate the orderings of risk for failure problems using a more general RPN methodology. Microelectronics Reliability. 2009; 49: 1586-1596.

[13] Chang KH., Cheng CH. A risk assessment methodology using intuitionistic fuzzy set in FMEA. International Journal of Systems Science. 2010; 41: 1457-1471.

[14] Zadeh LA. Fuzzy sets. Information and Control. 1965; 8(3): 338–353.

[15] Wikipedia. Fuzzy logic [Internet]. 2011 [cited 2014 February 5]. Available from: http://en.wikipedia.org/wiki/Fuzzy logic

[16] Mamdani EH. Application of fuzzy logic to approximate reasoning using linguistic Systems. IEEE Transactions on Computers. 1977; 26 (12): 1182-1191.

[17] Klir GJ., Yuan B. Fuzzy Sets and Fuzzy Logic: Theory and Applications. NY: Prentice Hall, Inc.; 1995.

[18] Mendel JM. Fuzzy logic systems for engineering: a tutorial, Proceedings of the IEEE. 1995; 83 (3): 345-377.

[19] Zimmermann HJ. Fuzzy Set Theory and its Applications, 4th ed. Norwell, Massachusetts, USA: Kluwer Academic Publishers; 2001.

[20] Thongrattanatai S, Taechamaneerat S, Pongpullponsak A. Quality control by TQM in order to reduce the defects in plated bath product type (K-160) of sanitary ware [Project]. Bangkok: King King Mongkut's University of Technology Thonburi; 2010.

Appendix A. Rule base for fuzzy output

No	Severity	Occurrence	Not detection	Fuzzy output
1	Almost none	Almost none	Almost none	None
2	Almost none	Almost none	Low	None
3	Almost none	Almost none	Moderate	Very low
4	Almost none	Almost none	High	Low
5	Almost none	Almost none	Very high	Low
6	Almost none	Low	Almost none	Very low
7	Almost none	Low	Low	Low
8	Almost none	Low	Moderate	Low
9	Almost none	Low	High	High low
10	Almost none	Low	Very high	Low moderate
11	Almost none	Moderate	Almost none	Very low
12	Almost none	Moderate	Low	Low
13	Almost none	Moderate	Moderate	Low
14	Almost none	Moderate	High	High low
15	Almost none	Moderate	Very high	High low
16	Almost none	High	Almost none	Low
17	Almost none	High	Low	High low
18	Almost none	High	Moderate	Low moderate
19	Almost none	High	High	Moderate
20	Almost none	High	Very high	High moderate
21	Almost none	Very high	Almost none	High low
22	Almost none	Very high	Low	Low moderate
23	Almost none	Very high	Moderate	Moderate
24	Almost none	Very high	High	High moderate
25	Almost none	Very high	Very high	High
26	Low	Almost none	Almost none	None
27	Low	Almost none	Low	None
28	Low	Almost none	Moderate	Very low
29	Low	Almost none	High	Low

No	Severity	Occurrence	Not detection	Fuzzy output
30	Low	Almost none	Very high	Low
31	Low	Low	Almost none	Very low
32	Low	Low	Low	Low
33	Low	Low	Moderate	High low
34	Low	Low	High	Low moderate
35	Low	Low	Very high	Moderate
36	Low	Moderate	Almost none	High low
37	Low	Moderate	Low	Low moderate
38	Low	Moderate	Moderate	Moderate
39	Low	Moderate	High	High moderate
40	Low	Moderate	Very high	Low high
41	Low	High	Almost none	Low moderate
42	Low	High	Low	Moderate
43	Low	High	Moderate	High moderate
44	Low	High	High	Low high
45	Low	High	Very high	High
46	Low	Very high	Almost none	Low moderate
47	Low	Very high	Low	Low moderate
48	Low	Very high	Moderate	Moderate
49	Low	Very high	High	High
50	Low	Very high	Very high	High
51	Moderate	Almost none	Almost none	Very low
52	Moderate	Almost none	Low	Very low
53	Moderate	Almost none	Moderate	Low
54	Moderate	Almost none	High	High low
55	Moderate	Almost none	Very high	Low moderate
56	Moderate	Low	Almost none	Low
57	Moderate	Low	Low	High low
58	Moderate	Low	Moderate	Low moderate
59	Moderate	Low	High	Moderate
60	Moderate	Low	Very high	High moderate
61	Moderate	Moderate	Almost none	Low moderate
62	Moderate	Moderate	Low	Moderate
63	Moderate	Moderate	Moderate	High moderate
64	Moderate	Moderate	High	Low high
65	Moderate	Moderate	Very high	High
66	Moderate	High	Almost none	Low
67	Moderate	High	Low	High low
68	Moderate	High	Moderate	Low moderate
69	Moderate	High	High	High moderate
70	Moderate	High	Very high	Low high
71	Moderate	Very high	Almost none	Low high
72	Moderate	Very high	Low	Moderate
73	Moderate	Very high	Moderate	High moderate
74	Moderate	Very high	High	Low high
75	Moderate	Very high	Very high	High
76	High	Almost none	Almost none	None
77	High	Almost none	Low	Very low
78	High	Almost none	Moderate	Low
79	High	Almost none	High	Low
80	High	Almost none	Very high	High low
81	High	Low	Almost none	Very low
82	High	Low	Low	Low
83	High	Low	Moderate	High low
84	High	Low	High	Very low
85	High	Low	Very high	Moderate
86	High	Moderate	Almost none	Low
87	High	Moderate	Low	High low
88	High	Moderate	Moderate	Low moderate
89	High	Moderate	High	Moderate
90	High	Moderate	Very high	High moderate
91	High	High	Almost none	Low moderate
92	High	High	Low	Moderate
93	High	High	Moderate	High moderate
94	High	High	High	Low high
95	High	High	Very high	High
96	High	Very high	Almost none	Moderate
97	High	Very high	Low	High moderate
98	High	Very high	Moderate	Low high
99	High	Very high	High	High
100	High	Very high	Very high	Very high
101	Very high	Almost none	Almost none	Very low
102	Very high	Almost none	Low	Very low
103	Very high	Almost none	Moderate	Low
104	Very high	Almost none	High	Low
105	Very high	Almost none	Very high	High low
106	Very high	Low	Almost none	Low
107	Very high	Low	Low	High low
108	Very high	Low	Moderate	High low
109	Very high	Low	High	Low moderate
110	Very high	Low	Very high	Moderate
111	Very high	Moderate	Almost none	High low
112	Very high	Moderate	Low	High low
113	Very high	Moderate	Moderate	Low moderate
114	Very high	Moderate	High	Moderate
115	Very high	Moderate	Very high	High moderate
116	Very high	High	Almost none	Low moderate
117	Very high	High	Low	Moderate
118	Very high	High	Moderate	High moderate
119	Very high	High	High	Low high
120	Very high	High	Very high	High
121	Very high	Very high	Almost none	High moderate
122	Very high	Very high	Low	Low high
123	Very high	Very high	Moderate	High
124	Very high	Very high	High	Very high
125	Very high	Very high	Very high	Very high

Multinomial logistic regression of benefits of renewable energy-examples from biogas consumers of Nepal

Jyoti U. Devkota[1*], Chanda Prajapati[2], Swechhya Singh[2] and Binu Hada[2]
[1]*Department of Natural Sciences - Mathematical Sciences, Kathmandu University, Dhulikhel, Kavre, Nepal, drjdevkota@ku.edu.np*
[2]*Department of Environmental Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal*

Abstract

Benefits of a source of renewable energy namely biogas are predicted using multinomial logistic models. This analysis is based on consumer profile database constructed with the data of 400 households. An extensive survey of 400 households was conducted between September to November 2010. It helped understand the change in the socio-economic parameters of biogas users after they used this source of renewable energy. Exercises in the generation of error free data during this survey ensured its quality. This plays a crucial role in the developing world where even data from governmental sources cannot be relied upon for the quality and accuracy. Through these models the probability of change in time saved is predicted for a unit change in several independent variables such as distance travelled in the collection of firewood before and after the construction of plant, reduced pollution, reduced fuel expenses, amount of firewood saved after the switch over to this renewable energy source. The interrelationships between these variables including the benefits to women in terms of reduced pollution and reduced fuel expenses are analyzed. This study helps understand the dynamics of improvement in the lifestyle of people with the help of multinomial regression. Multivariate analysis has been used to minutely analyze and predict the interrelationships. Further benefits to the consumers in general and women in particular after they started using biogas are minutely analyzed with the help of statistics.

Keywords: Multinomial regression, applied statistics, renewable energy, statistical analysis, interdisciplinary application

*Corresponding Author
E-mail Address: drjdevkota@ku.edu.np

1. Introduction

Twelve hours of power outages daily from the national grid in the current season and up to sixteen hours in the lean period is the state of commercial energy in Nepal. Major section of the population depends on firewood for cooking. A lot of time of women is spent in the collection of firewood. With ever increasing population in the developing world and depleting fossil fuel reserve, this dependence shows a rising trend. According to 2011 census 94 % of the total households in mountain depend on wood and firewood for fuel and cooking [1]. The rural urban differential in the use of wood/ firewood for cooking is stark. The use of firewood is in 73 % household in the rural areas whereas it is 25% in the urban areas. This dependence also implies that the women spend their times in the collection of firewood. This time could have been utilized for working in the farm, helping children in studies and other income generating activities. Cost of firewood per kilogram ranges from Rs 5- Rs 8. A switch over to a source of renewable energy is not only a suitable adaptation strategy to climate change but it also provides solutions to other problems typical to the developing world.

Biogas plants were promoted by the government of Nepal in the agriculture year 1974/75 as a part of special program. This program installed 250 plants in different parts of Nepal under the supervision of governmental and non governmental agencies. As a part of alternate energy year 2009/2010 government of Nepal planned to install 100,000 biogas plants in 70 districts [2]. Problems of energy supply, shortage of cheap and efficient fuel, shortage of other many usable commodities and growing environmental pollution a switch over to a source of renewable energy seems to be a plausible solution.

In modern age of information and communication technology, data based research is important. It can be used not only for exploration but also for confirmation. Data is power as many important problems can be isolated through the correct analysis and prognosis of this data. Its generation, verification and rectification play a crucial role in interdisciplinary research. Especially in the developing world data are erroneous and faulty. In such cases results are unreliable. It is mainly due to the lack of awareness of the various organization and individuals involved in the collection and supply of data. Conduction of well planned surveys and data collections processes result in minimum errors and lack of ambiguity [3]. Good quality data with maximum information requires proper planning. Steps from the stage of design of experiments, collection of data to the final analysis and interpretation has to be properly worked out. It results in minimum error data ensuring wider interpretability and application. It also minimizes statistical bias. The latent information stored in the data can extracted with the help of appropriate statistical methodologies. The interrelationships between various variables of minimum error data can be studied using advanced statistical techniques like principle components analysis, factor analysis, logistic regression, multiple regressions. Various methods very popular in practice such as pie charts, bar chart, histogram, mean, mode, and median are effective but they cannot exploit full potential of the data. Thus most of the information stored in the data goes underutilized if only very simple statistical tools such as mean, median or mode are used. These aspects have played a prominent role here.

In this paper various benefits related to the use of renewable energy are minutely analyzed. Time spent in the collection of fire wood before and after the construction of plant, distance travelled in the collection of firewood before and after the plant, time saved and firewood saved are some of the variables which try to assess the benefits of biogas plant. These results are based on data collected from the socio-economic survey of 400 households. It was designed with an objective of keeping biogas use in the core and getting all the possible information about a typical middleclass Nepalese family inhabiting in rural areas, its economic and social background and change after biogas was used in their household. It was conducted during September – December 2010 in three different rural settings of Nepal. The questionnaire and the survey were carefully designed. All the possible sources of error were carefully eliminated. These results give a better understanding of the benefits of biogas. This will help policy makers and planners in making better strategy in popularizing biogas in particular and renewable energy in general.

2. Research Methodology

Generation of error free data is of crucial importance especially in countries like Nepal. The data collected by different governmental and nongovernmental bodies cannot be relied on for their accuracy. Lack of awareness of all the stake holders of this process including public is one of the main reasons [3]. Further lack of coordination between several governmental and nongovernmental bodies has also resulted in the duplication of the collected data. The generation of data for this study was thoroughly exercised for identification and elimination of possible sources of error.

2.1 The Questionnaire and its Pretesting

Interdisciplinary research is data based. The data should be generated from a good experimental design and well tested questionnaire. The survey and the questionnaire were designed keeping these principles under consideration. The questionnaire was thoroughly exercised on 30 households. Then it was finalised to 59 questions. A research assistant specially trained for it required 40-45 minutes. The draft questionnaire comprised of 62 questions before it was tested on 30 households in Sudal VDC, Bhaktapur [4]. Not only the response of the consumers was noted and the answer options were accordingly refined and updated to remove errors, but also the ambiguity of answers were traced and corrected. In these 59 questions information was collected on very relevant topics such as the age distribution of 400 households comprising of 2272 individuals of different age groups, amount of landholdings, livestock, their fuel wood expenses before and after the installation of plants.

2.2 The Survey and Database

The entire data collection process took 15 days. It was conducted during September to December 2010. Information on 467 variables related to the socio-economic consumption pattern of biogas was collected [4]. The details of the survey are given in the Table 1.

Table1: Details of the Survey of Biogas consumers

Sr. No.	Area of Survey	No of households	Survey
1	Sudal VDC	30	Pretest
2	Simara	300	Main Survey
3	Sarlahi	70	Main Survey

2.3 Statistical Methodology

Most statistical analyses distinguish between response variable and explanatory variable. Categorical data analysis of the dependent and independent variables where these are various responses of biogas consumers highlights the interdisciplinary application of statistical methodology. Although categorical data are often referred as qualitative in nature, but here they are treated as ordinal data of quantitative nature. They have been assigned ordered scores according to their categories [5]. Here two key distributions have been fitted to the categorical data binomial and multinomial. Response variables with yes/no options are binomial and those with more than two options are multinomial.

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership.

Let J denote the number of categories of Y. Here Y is a multinomial response variable. Let $\{\pi_1, \pi_2,\ldots,\pi_j\}$ denote the response probabilities, satisfying the condition that their sum is equal to 1. Logit models for multinomial response pair each category with a baseline category [5]. When the last category (J) is the baseline, the baseline-category logits are

$\log(\pi_i/\pi_j)$, j = 1,….J-1

Given that the response falls in the category j or J, this is the log odds that the response is j.

The baseline category logit model with predictor x is

$\log(\pi_j/\pi_J)=\alpha_j+\beta_j x$, j = 1, …,.J-1

The model has J-1 equations with separate parameters for each. The effects vary according to the category paired with the baseline. When J = 2, this model simplifies to single linear equation for $\log(\pi_1/\pi_2)=\text{logit}(\pi_1)$, resulting in ordinary logistic regression for binary responses.

There is an odds ratio associated with each predictor. It is denoted by Exp(B). It is more than 1 in cases where predictors increase the logit, Exp(B) is equal to 1 in cases where predictor don't have any influence on the logit and Exp(B) is less than 1 in cases where predictors decrease the logit.

Box plots help visually summarise the data in terms of quartiles and outliers. The box plot uses the median, the approximate quartiles, and the lowest and highest data points to convey the level, spread, and symmetry of a distribution of data values. Here box plot have been used instead of tabular data for different independent variables. This has helped us understand distribution and spread of different variables.

The interdependence between two variables is also analysed with the help of chi square test of independence of attributes. Here H_0 represents that the two attributes are independent and $H_1$is the opposite of the null hypothesis representing that the two attributes are dependent. Large χ^2values are more contradictory to $H_{0,}$ the P-value is the null probability that $\chi^2$is at least as large as the observed value [5].

3. Research Results and Discussion

Different variables obtained from the survey are classified in Table 2. A consumer profile database was constructed. It comprised of 23 Tables classified under various names as mentioned in Column of Table 2. Family background and family description were two tables that gave details of the families in these 400 households. There are 72 variables giving various details and have data that are either binary nominal or discrete ratio in nature. Here the benefits of biogas to the consumers were indirectly assessed with the questions tabulated later in the database under following names:

1. Distance travelled for firewood
2. Comparison of types of fuel used
3. Time spent on firewood collection
4. Source of firewood
5. Positive impacts after installation of plant

Different questions assessed the benefits of biogas to its consumers. For example for the question how much time is saved after the construction of biogas plant? Options of no time saved, up to 60 min, 1-3 hours, 3 – 5 hours, more than 5 hours were given. These options are categorised as 1, 2, 3, 4 and 5. So saves in time data is ordinal. Similarly questions asking distance travelled for the collection of firewood before and after the construction of plant had 5 options namely none, up to 100 m, 100 – 200 m, 200-500 m and more than 500 m. These were categorised as 1, 2, 3, 4 and 5 and are ordinal in nature. Similarly questions seeking response to reduced pollution as a benefit to the use of biogas are also of yes/no nature. It is also binary with codes 1 and 2 respectively. The questions seeking responses to reduced fuel expenses from male or female are of yes/no type. They are binary in nature. The response to the question on amount of firewood saved after the construction of the plant was categorised as nothing saved, up to 30 kg, 30-50 kg and above 50 kg. Ordinal data 1, 2, 3 and 4 was used to quantify these responses.

Chi-square test of independence of attributes is applied for testing the dependence between various attributes.

A relative profile of the attributes across 400 households is portrayed in seven different box plots. These attributes indirectly assessed the benefits of biogas to its consumers. The dependence between distance travelled for the collection of firewood before and time saved response is highly significant. Comparative profile of the distribution these responses across 400 households of biogas consumers are given in the box plot of Figure 1. The chi-square value at 16 degree of freedom was 44.883. Similarly the dependence between the responses of distance travelled after for the collection of firewood and time saved is also highly significant. The chi-square value was 78.789 at 16 degree of freedom. Box plot of Figure 2 portrays this dependence. The concentration of most of the values is in the category 200 m – 500 m as portrayed in Figure 1. But after the construction of plant most values are concentrated in the category not travelled as shown in Figure 2. This reflects a significant reduction in distance travelled for the collection of fuel wood. Women were more sensitive to the benefits of biogas in reducing the pollution effects. The dependence of this response to response of time saved is highly significant in case of female respondents. The chi-square value was 125.10 at 4 degree of freedom. But in males this dependence was not significant with a chi-square value of 1.387 at 4 degree of freedom and a p value of 0.846.

As in rural settings of Nepal most of the time of the women is spent in the traditional kitchen using firewood, these results statistically validate this fact. Time spent by female is substantial hence this test shows that men are not sensitive to the benefits of biogas in terms of reduced pollution effects. The gender differential to this response can be obtained from Figure 3 and Figure 4. A comparison between different categories of responses between time saved and firewood saved is given by box plot of Figure 5. There is a highly significant dependence between these two responses with a chi-square of 39.168 at 8 degrees of freedom.

Table 2: Overview of Consumer Profile Database

Table No.	Heading of the Table	Variables	Types of Data
1	Family Background	10	Yes/No & Nominal
2	Family Description	60	Discrete
3	Occupation	6	Nominal
4	Livestock	12	Discrete & Continuous
5	Business	3	Nominal and Continuous
6	Other Occupation	8	Discrete
7 & 8	Land	6	Yes/No
9	Socio-economic condition	54	Yes/No, Nominal
10	Water Needs	17	Yes/No
11	Biogas Questions	75	Nominal, Ordinal
12	Size of Plant	18	Yes/No
13	Woman Empowerment	44	Nominal
14	Source of Firewood	14	Yes/No
15	Distance covered for firewood	8	Yes/No
16	Comparison of Types of fuel used	40	Yes/No, Continuous
17	Time spent on Firewood collection	10	Yes/No
18	Frequency of Feeding	15	Yes/No
19	Cultivation of Crops	32	Yes/No
20	Application of Fertilizers	20	Yes/No
21	Health Related Issues	8	Nominal
22 & 23	Positive Impacts of Biogas	9	Yes/No
	Total	467	

Table 3: Descriptive Statistics

Parameters	Mean	Mode	Median	Standard Deviation
A	4.08	5.0	5.0	1.3248
B	2.37	2.0	1.0	1.633
C	2.85	3.0	3.0	0.931
D	1.94	2.0	2.0	0.233
E	1.44	1.0	1.0	0.496
F	1.99	2.0	2.0	0.100
G	1.72	2.0	2.0	0.448
H	3.51	4	4	0.67

The yes/no response of reduced fuel expenses to the categorical-ordinal data response of time saved in portrayed in Figure 6 and Figure 7. A comparison between these two figures also gives the gender differential. Chi-square test of dependence of response of reduced fuel expenses to time saved showed significant results for male respondents with a p value of 0.037 at 4 degrees of freedom. The chi-square value is 10.193. But the response of female is not significant at 5 percent level of significance with a p value of 0.069. The chi-square value is 8.707 at 4 degrees of freedom. In the following table (Table 3), we represent distance travelled for the collection of firewood before the construction of plant by A, distance travelled after by B, saves time by C, reduced fuel expenses by female by D, reduced fuel expenses by male by E, reduced pollution by male by F , reduced pollution by female by G and firewood saving by H. Although these are ordinal and nominal data, but the measures of central tendencies such as mean, mode and median gives us an idea of the average response and most popular response. The standard deviation gives an idea of the consistency of the response.

From Table 3 it can be seen that for question A on average option 4 was chosen. But since the modal value is 5, it is the most popular choice. This implies that out of 400 households of biogas consumers people covered on average 200-500 m per day for the collection of firewood. After the construction of plant half of the population covered up to 100m as reflected by the median value of question B. So the distance travelled is reduced substantially. In response to the query on how much time is saved per day represented by C, the category 3 was a popular choice. This implies that 1 – 3 hours per day is saved. This time is utilized for income generating activities. Female work in the farm thus money spent on the payment of extra help is saved. Similarly in the response to firewood saved represented by H, category 4 is the mode. This represents saving of more than 50 kg. The benefits of reduced fuel expenses have been ranked higher by male respondents than female.

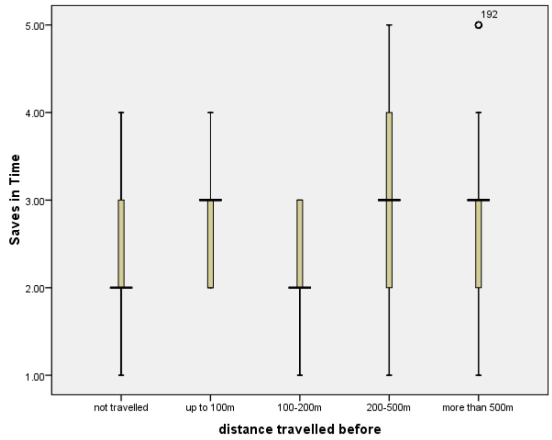


Figure 1: Relation between distances travelled before and time saved

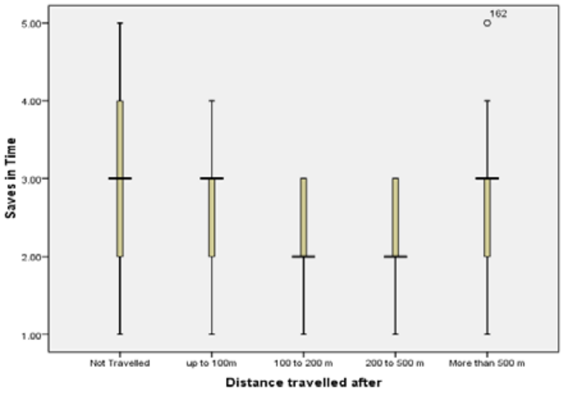


Figure 2: Relation between distances travelled after and time saved

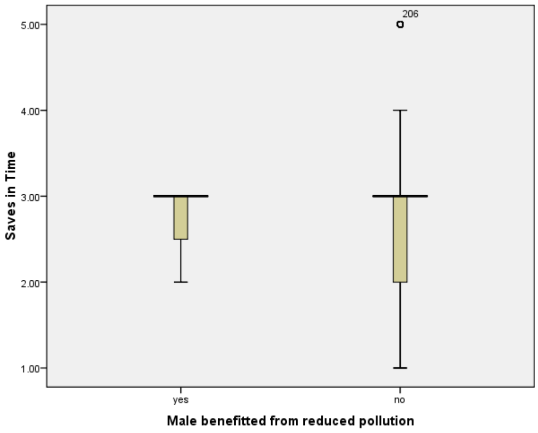


Figure 3: Relation between reduced pollution and time saved (male)

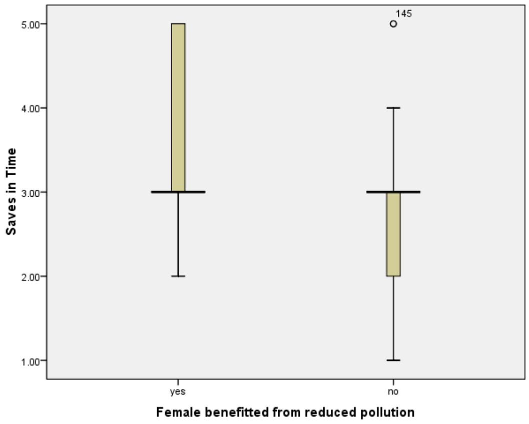


Figure 4: Relation between reduced pollution and time saved (female)

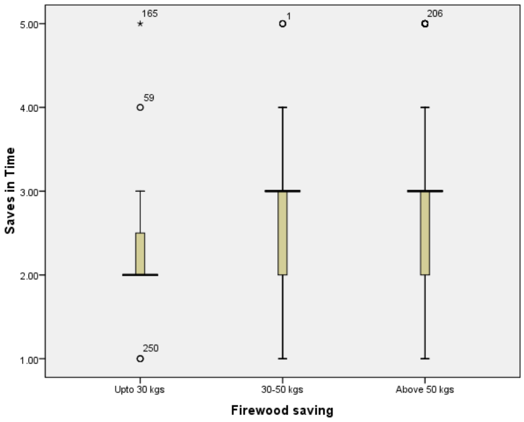


Figure 5: Relation between firewood saved and time saved

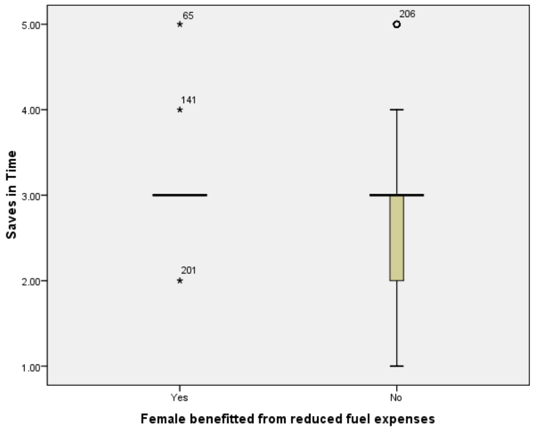


Figure 6: Relation between response of reduced fuel expenses and time saved (female)

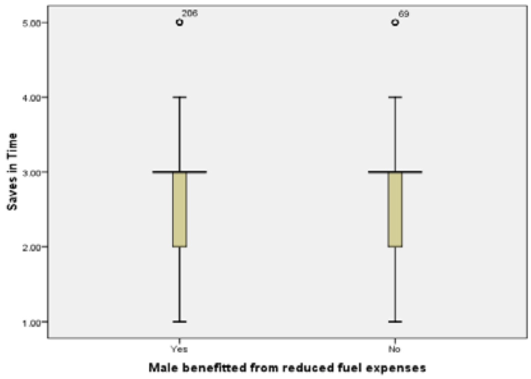


Figure 7: Relation between response of reduced fuel expenses and time saved (male)

Further multinomial logistic regression is fitted to these variables. Here the variable saves time is regressed on distance travelled before, distance travelled after, female benefitting from reduced fuel expenses, male benefitting from reduced fuel expenses, male benefitting from reduced pollution, female benefitting from reduced pollution and amount of firewood saved per month. Here saves in time is an ordinal data with categories 1, 2, 3, 4, 5 representing no time saved, up to 60 minutes saved, 1- 3 hours saved, 3-5 saved and more than 5 hours saved respectively. Similarly distance travelled before is an ordinal data with categories 1, 2, 3, 4 and 5 representing not travelled, up to 100m, 100-200m, 200-500m and more than 500m respectively. Similarly distance travelled after also has same categories. Firewood saving is also an ordinal data with three categories 1, 2 and 3 represented by up to 30 kg saved, 30 – 50 kg and more than 50 kg saved. The rest of four variables are binary yes/no variables. Here 1 is represented by yes and 2 by no. This regression is highly significant. The likelihood ratio test of regression of saves time (dependent variables) on ordinal data distance travelled after, distance travelled before and firewood saved shows highly significant results. Similarly the response of reduced fuel expenses by men and reduced pollution by women has also shown significant results. Through multinomial logistic models the benefits have been quantified. The results can be summarized in the following manner. Those who travelled up to 100 for the collection of firewood before the construction of plant their likelihood of claiming the biogas saved their time increased by 19730000, 23230000 and 21670000 for up to 60 min, 1- 3 hours and 3 – 5 hours respectively. Similarly those who covered 100 m – 200 m for the collection of firewood their likelihood of responding that it saved their time increased by 1.945 and 1.582 times for up to 60 min and 1 - 3 hours respectively. This is in comparison to those households who had to cover more than 500 m for the collection of firewood before the construction of plant. Similarly those who don’t have to travel at all for the collection of firewood after the construction of the plant are 3.478 and 4.706 times more likely to respond that biogas saved 3-5 hours and more than 5 hours than those who still cover more than 500m. These are all the Exp(B) values. They are also called the odds ratio. There were 229 households which covered more than 500m before and now after the construction of the plant it is 88. This shows that there is a considerable reduction in the distance travelled. Less distance travelled implies more time saved which in turn implies more money saved. As this time saved is used in income generating activities by these household. So a switch over to renewable energy is not only an adaptation strategy to climate change but it can be also increase the socio-economic status of areas hit with scarcity of fossil fuel. Similarly the odd of those women favoring biogas as source of energy that has reduced effects of pollution also supporting its time saving benefits is very highly significant. This feature us reflected more prominently in women as they have to suffer from the pollution of traditional cooking stoves.

Table 4: Model fitting information

Model	-2 Logliklihood	Chi-Square	df	Sig
Intercept Only	642.490			
Final	364.621	277.869	56	.000

The results of fitting a multinomial regression model are given in Table 4 and Table 5. We see that this regression is highly significant with a chi square value of 56. The likelihood ratio tests in Table 5 also show significant results at 5 % level of significance for most of the independent variables. The accuracy prediction with the help of multinomial regression is given in Table 6. Observed values are compared with predicted values in this table. The accuracy of this model is 62.8 %. This is quiet high for a data based on response of human beings. These results are not based on closely controlled laboratory conditions. Although all the possible sources of error have been traced out and eliminated in this survey experiments based on humans are more likely influenced by other psychological factors. Hence an accuracy of 62.8 % is a good result.

Table 5: Likelihood ratio tests

Effect	-2 logliklihood	Chi - square	df	Sig.
Distance travelled after	426.16	61.54	16	.000
Male benefit from reduced fuel	374.286	9.666	4	.046
Female benefit from reduced fuel	373.394	8.774	4	.067
Male benefit from reduced pollution	365.784	1.164	4	.884
Female benefit from reduced pollution	436.983	72.363	4	.000
Firewood Saved	390.633	26.012	8	.001
Distance travelled before	403.043	38.422	16	.001

Table 6: Observed versus predicted in multinomial regression

	Predicted					
Observed	60 mins	1-3 hr	3-5 hr	>5 hr	No time saved	% Correct
60 mins	57	71	1	3	0	43.2
1-3 hr	25	160	3	7	0	82.1
3-5 hr	5	10	6	2	0	26.1
>5 hr	1	4	4	28	0	75.7
No time saved	8	5	0	0	0	.0
Overall Percentage	24	62.5	3.5	10.0	0	62.8

4. Conclusion

This paper promotes the interdisciplinary application of statistical methodology in a problem from renewable energy. The results of this paper are based on a survey of 400 households of biogas consumers conducted from September – December 2010 in three different rural settings of Nepal. This survey was thoroughly planned before its actual implementation and all the possible sources of error were identified and eliminated. This increased the reliability of data where in countries like Nepal even the official sources don't have accurate and error free data. This extensive data collection tried to get information about change in the socio-economic conditions of a typical biogas consumer household. All 59 questions of the questionnaire of this survey required 40-45 minutes for completion per household. The survey was done in collaboration with a partner industry that has installed these plants in these households. General and multinomial logistic regression in particular is applied in quantifying the benefits of biogas to its consumers. Here the distribution of time saved is studied with respect to several variables such as distance travelled before for the collection of firewood, distance travelled after for the collection of firewood, amount of firewood saved, response of reduced fuel expenses by male, response of reduced fuel expenses by female, amount of reduced pollution by male and amount of reduced pollution by female. Box plots give a comparative profile of the distribution of the response variables with respect to each of these independent variables across 400 households. The interdependence of these attributes is analyzed using chi-square test of independence of attributes. The time saved by the household after biogas is installed is regressed on distance travelled before, distance travelled after, female benefitting from reduced fuel expenses, male benefitting from reduced fuel expenses, male benefitting from reduced pollution, female benefitting from reduced pollution and amount of firewood saved per month. Saves in time is an ordinal data with categories 1, 2, 3, 4, 5 representing no time saved, up to 60 minutes saved, 1- 3 hours saved, 3-5 saved and more than 5 hours saved respectively. Similarly distance travelled before is an ordinal data with categories 1, 2, 3, 4 and 5 representing not travelled, up to 100m, 100-200m, 200-500m and more than 500m respectively. Similarly distance travelled after also has same categories. Firewood saving is also an ordinal data with three categories 1, 2 and 3 represented by up to 30 kg saved, 30 – 50 kg and more than 50 kg saved. The rest of four variables are binary yes/no variables.

Here 1 is represented by yes and 2 by no. This regression is highly significant. Large Exp(B) values show that there are greater benefits of biogas plants. Here these benefits have been quantified with the help of odds ratio by fitting multinomial logistic models.

Acknowledgements

The authors gratefully acknowledge the contributions of Rapti Renewable and Energy Services and other members of the data collection team. We are also very thankful to Renewable Nepal for funding this work. This work is funded by NORAD, SINTEF under Renewable Nepal Project grant number RENP-10-06-PID-379.

References

[1] Central Bureau of Statistics. National report on national population and housing census 2011. Central Bureau of Statistics, Kathmandu, Nepal 2012.

[2] WECS. Energy Sector Synopsis Report 2010. Water and Energy Commission Secretariat, Kathmandu, Nepal, 2010.

[3] Devkota JU. Mortality and Fertility Models for Countries with Limited Data - Results Based on Demographic Data of Nepal, India and Germany. Lambert Academic Publishing, Saarbruecken, Germany, 2012.

[4] Devkota J. U., Hada B., Prajapati C., Singh S. The importance of research data digitization and its statistical analysis-with examples of biogas consumers of Nepal. International Journal for Environmental Science and Development 2012, 3 (2), 103-108.

[5] Agresti A. An introduction to categorical data analysis. 2 nd ed, New Jersey, John Wiley & Sons, 2007.

A new estimator of population size based upon the conditional independent Poisson Mixture Model

Rattana Lerdsuwansri[1*] and Dankmar Böhning[2]
[1]*Department of Mathematics and Statistics, Thammasat University, Pathumthani, 12121, Thailand,*
rattana@mathstat.sci.tu.ac.th
[2]*School of Mathematics & Southampton Statistical Sciences Research Institute,*
University of Southampton, Southampton, SO17 1BJ, UK, d.a.bohning@soton.ac.uk

Abstract

To estimate the size of a closed population via capture-recapture experiments, the Lincoln-Petersen approach is a classical model relying on a two-source situation and binary source/listing variables. A new estimator is proposed by extending the Lincoln-Petersen estimator to non-binary source/listing variables. We consider a bivariate count variable where counts are used to summarize how often a unit was identified from source/list 1 and source/list 2. The Poisson mixture model is adopted to model unobserved population heterogeneity by assuming independence for a homogeneous component leading to discrete mixtures of bivariate, conditional independent Poisson model. The EM algorithm is discussed for maximum likelihood estimation. The model is selected on the basis of Bayesian Information Criterion (BIC). A simulation study was conducted to compare several estimators, including the MLE, the Turing, the Chao, the Zelterman and the proposed estimator. The results of the simulation experiments showed that the new estimator performs the best with smallest mean square error for all population sizes under heterogeneity and even under homogeneous Poisson model. As an application, estimating the number of heroin users in Bangkok in the year 2001 is examined using the proposed method.

Keywords: Capture-recapture, population size estimation, zero-truncated bivariate Poisson mixtures, EM algorithm

*Corresponding Author
E-mail Address: rattana@mathstat.sci.tu.ac.th

1. Introduction

Capture-recapture methods have been widely used in enumerating a population of size N that is difficult to approach. Ordinarily, the methods are well established to estimate wildlife abundance [1–2]. A diversity of application areas has adopted capture-recapture methods to estimate missing units as well as the total number in the population. For instance, in social sciences, the interest is in determining the amount of illegal behavior such as driving without license or immigrating without permission [3–4]. In medical sciences/public health, there is concern about finding the number of illicit drug users as well as estimating the number of outbreaks of particular disease [5–6].

We assume that the unknown population size N remains constant during the period of the study (no birth, death or migration) which is referred to a closed population. To formulate an estimate of population size N, a capture mechanism, e.g. trapping, register, diagnostic device, is used to identify units having a characteristic of interest. The Lincoln-Petersen estimator is one of the most popular approaches in capture-recapture. The estimator is used in a closed population and based on independence of two sources. Each source is treated as a binary variable taking values 0 and 1 for an unidentified and identified unit, respectively and a 2×2 contingency table is formed (see Table 1). A population size N is partitioned into 4 groups by n_{11}, n_{10}, n_{01}, n_{00}, the number of units identified by both sources, by source 1 but not source 2, by source 2 but not source 1, and not both of source 1 and source 2. n_{00} is unknown because units who were never identified did not appear. Consequently, n_{00} is required to be estimated leading to an estimate of N.

Table 1: A 2×2 table of a two-source situation

		Source 2		
		not identified (0)	identified (1)	
Source 1	not identified (0)	n_{00}	n_{01}	n_{0+}
	identified (1)	n_{10}	n_{11}	n_{1+}
		n_{+0}	n_{+1}	N

If two sources were independent, the odds ratio $\frac{n_{11}n_{00}}{n_{10}n_{01}} \approx 1$. Under independence, we have $\hat{n}_{00} = \frac{n_{10}n_{01}}{n_{11}}$ and the Lincoln-Petersen estimate is then given by

$$\hat{N}_{LP} = n_{11} + n_{10} + n_{01} + \hat{n}_{00} = \frac{n_{1+}n_{+1}}{n_{11}}. \quad (1)$$

See [7] for more details.

A new estimator proposed in this paper is built upon the extension of the Lincoln-Petersen approach. We restrict ourselves to repeated identifications in which the same unit is identified repeatedly during the period of the study. Instead of 0 (not identified) and 1 (identified), we focus on the number of identifications, 0, 1, 2, … , m where m is the largest observed count. As a consequence, we consider bivariate counts (i,j) where counts are used to summarize how often a unit was identified by source 1 and source 2, respectively. The number of units identified exactly i times by source 1 and j times by source 2 denoted by f_{ij} is obtained and can be expressed as in Table 2. It is interesting to note that the observed data do not include (0,0) counts since the units who never apprehend do not appear in both of the two sources.

Table 2: Observed frequencies in terms of contingency table

	Source 2				
	not identified	identified			
	0	1	2	…	m
not identified 0	-	f_{01}	f_{02}	…	f_{0m}
1	f_{10}	f_{11}	f_{12}	…	f_{1m}
Source 1 identified 2	f_{20}	f_{21}	f_{22}	…	f_{2m}
⋮	⋮	⋮	⋮	…	⋮
m	f_{m0}	f_{m1}	f_{m2}	…	f_{mm}

The methods are presented in section 2. Performance of the proposed estimator is evaluated by a simulation study in section 3. In section 4, the new estimator is applied to real data describing heroin user contacts in Bangkok, Thailand. Some points of work are remarked in section 5.

2. Discrete mixtures of bivariate, conditional independent Poisson distributions

2.1 Count probability model

Suppose that a population is closed with size N. Let $\mathbf{Y}_d = (Y_{d1},Y_{d2})'$ denote the number of times that unit d is identified by source 1 and source 2 in the observational period, d = 1, 2,…, N. $\mathbf{Y}_d$ is a vector of two dimensions and a bivariate count variable having values in {0, 1, 2, 3, …}. If $\mathbf{Y}_d = \mathbf{0}$ then a unit is unidentified from both of two sources. Since a unit which has never been identified does not appear in the data, the observed data set is $\{\mathbf{Y}_d = (Y_{d1},Y_{d2})'|Y_{d1}+Y_{d2} \geq 1, d = 1,2, \dots, N\}$. The associated, observed distribution of counts is therefore referred to zero-truncated distribution. Let f_{ij} denote the number of units identified i times by source 1 and j times by source 2. Hence, f_{00} is the frequency of units identified zero times by both sources. f_{00} is unknown. We have the number of observed units $n = f_{01}+ f_{10}+ f_{11}+\dots+ f_{mm}$ and $N = n+ f_{00}$. Consequently, f_{00} requires determination in order to obtain an estimate for the population size N.

Let $p_{ij} = \Pr(\mathbf{Y}_d = (i,j)')$ denote probability for identifying a unit i times by source 1 and j times by source 2. Accordingly, p_{00} is the probability of not identifying a unit. The unobserved p_{00} might be replaced by the expected value Np_{00}. If p_{00} is known then solving for $N = n+ f_{00}$ leads to the well-known Horvitz-Thompson estimator

$$\hat{N} = \frac{n}{1-\hat{p}_{00}}. \qquad (2)$$

See [3] for more details.

As p_{00} is unknown, modelling for count probability p_{ij} has to be assumed. However, count data modelled by identical parameter $\boldsymbol{\theta}$ are rare in practice. An alternative model incorporating heterogeneity of the population might be more appropriate [8–11]. Based upon repeated identifications in observational period, we postulate that $\mathbf{Y}$ arises from Poisson distribution having parameter $\boldsymbol{\theta} = (\lambda,\mu)'$ where λ and μ are parameters of count distribution in source 1 and source 2, respectively. Independence of Y_1 and Y_2 is assumed by conditioning on a homogeneous component. The distribution of count $\mathbf{Y}$ is provided as

$$f(\boldsymbol{y};\boldsymbol{Q}) = \sum_{k=1}^{s} q_k \,\mathrm{Po}(y_1;\lambda_k)\,\mathrm{Po}(y_2;\mu_k) \qquad (3)$$

with respect to the unobserved variable Z having distribution *Q*. A discrete mixing distribution $\boldsymbol{Q} = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots\lambda_S \\ \mu_1 & \mu_2 & \cdots\mu_S \\ q_1 & q_2 & \cdots q_S \end{pmatrix}$ gives weight q_k to parameters λ_k and μ_k for *k* = 1, 2, …, *s* where *s* is the number of unobserved components. The component weight q_k satisfy two constraints, $q_k \geq 0$ and $\sum_{k=1}^{s} q_k = 1$. Eq.(3) is referred to as discrete mixtures of bivariate, conditional independent Poisson distributions. For an introduction to mixture model, see [12].

2.2 Maximum Likelihood Estimation for bivariate zero-truncated Poisson mixtures

Assume that $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ are observed and drawn from mixture density. The observed, incomplete data log-likelihood is of the form

$$l(Q) = \sum_{i=0}^{m}\sum_{j=0}^{m} f_{ij}\log\left(\frac{\sum_{k=1}^{s} q_k\,\mathrm{Po}(i;\lambda_k)\mathrm{Po}(j;\mu_k)}{1-\sum_{k=1}^{s} q_k e^{-\lambda_k} e^{-\mu_k}}\right). \qquad (4)$$

An estimate of *Q* can be achieved by maximizing zero-truncated Poisson likelihood (4) leading to the nonparametric maximum likelihood estimate (NPMLE). The EM algorithm has become popular for maximum likelihood estimation particularly in connection with mixture models. To carry on the EM algorithm, the complete data log-likelihood is required [13]. At the E-step, the unobserved frequency f_{00} is replaced by its expected value given observed frequencies and current values of *Q*. Let the expected value of f_{00} is denoted by $\hat{f}_{00}$ which can be shown to be

$$\hat{f}_{00} = \mathrm{E}(f_{00} \,|\, \text{observed data}\,; \boldsymbol{Q}) = \frac{n \sum_{k=1}^{s} q_k e^{-\lambda_k} e^{-\mu_k}}{1 - \sum_{k=1}^{s} q_k e^{-\lambda_k} e^{-\mu_k}}.$$

The associated complete data log-likelihood is given by

$$l(Q) = \hat{f}_{00} \log\left(\sum_{k=1}^{s} q_k e^{-\lambda k} e^{-\mu k} \right) + \sum_{\substack{i=0 \\ i+j\geq 1}}^{m} \sum_{j=0}^{m} f_{ij} \log\left(\sum_{k=1}^{s} q_k Po(i;\lambda k) Po(j;\mu_k) \right). \quad (5)$$

To manipulate the maximum likelihood estimate $\hat{Q}$, the log-likelihood is maximized by applying the EM algorithm as well. In this case, a variable indicating component to which the count (i,j) belongs is ignored. Let z_{ijk} denote indicator variable defined as 1 if count (i,j) was drawn from component k; 0 otherwise.

If z_{ijk} were observed, the log-likelihood for the complete data would be given by

$$l(Q) = \sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} \sum_{k=1}^{s} z_{ijk} \, log\left(q_k\right) + \sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} \sum_{k=1}^{s} z_{ijk} \, log\left(\mathrm{Po}(\mathrm{i}; \lambda_k)\mathrm{Po}(\mathrm{j}; \mu_k)\right). \quad (6)$$

At the E-step, the unobserved indicator z_{ijk} is replaced by $e_{ij,k}$, its expected value conditional upon the observed data and current values of Q leading to

$$e_{ij,k} = E\left(z_{ijk} \,\middle|\, f_{ij}, \boldsymbol{Q} \right) = \frac{\mathrm{q_k} Po(i;\lambda k) Po(j;\mu_k)}{\sum_{\mathrm{k}=1}^{\mathrm{s}} \mathrm{q_k}\, Po(i;\lambda k) Po(j;\mu_k)}. \quad (7)$$

Substituting $e_{ij,k}$ into (6) yields the expected log-likelihood which is of the form

$$\sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} \sum_{k=1}^{s} e_{ij,k} \, log\left(q_k\right) + \sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} \sum_{k=1}^{s} e_{ij,k} \, log\left(\mathrm{Po}(\mathrm{i}; \lambda_k)\mathrm{Po}(\mathrm{j}; \mu_k)\right). \quad (8)$$

At the M-step, the new values of $\hat{Q}$ are updated by maximizing (8). The estimates of component weights $\mathrm{q_k}$ are achieved as

$$\hat{q}_k = \frac{\sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} e_{ij,k}}{\hat{N}}, \; k = 1, 2, \ldots, \mathrm{s}. \quad (9)$$

The estimates of component parameters are accomplished as

$$\hat{\lambda}_k = \frac{\sum_{i=0}^{m} \sum_{j=0}^{m} i f_{ij} e_{ij,k}}{\sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} e_{ij,k}}, \; k = 1, 2, \ldots, \mathrm{s}. \quad (10)$$

$$\hat{\mu}_k = \frac{\sum_{i=0}^{m} \sum_{j=0}^{m} j f_{ij} e_{ij,k}}{\sum_{i=0}^{m} \sum_{j=0}^{m} f_{ij} e_{ij,k}}, \; k = 1, 2, \ldots, \mathrm{s}. \quad (11)$$

Consequently, the population size estimator based upon discrete mixtures of bivariate, conditional independent Poisson model through the Horvitz-Thompson approach is

$$\hat{N}_{DBP} = \frac{n}{1 - \hat{p}_{00}} = \frac{n}{1 - \sum_{\mathrm{k}=1}^{\mathrm{s}} \hat{q}_k e^{-\hat{\lambda}_k} e^{-\hat{\mu}_k}}. \quad (12)$$

We attach the estimator an index DBP and call $\hat{N}_{DBP}$ the new estimator. Maximum likelihood estimation discussed above is along the line of Böhning and Schön [11].

2.3 Model selection criteria

Choosing an appropriate number of components is an important concern of finite mixture estimation. Several criteria are possible. Once the parameters have been estimated by maximum likelihood for each value of s, the information criteria are commonly recommended to select the number of components as well as an appropriate model. It is widely accepted that Bayesian information criterion (BIC) performs well for model selection in the context of mixture models. The BIC is defined as

$$\mathrm{BIC} = -2\log L(\hat{Q}_S) + (3s-1)log(n) \quad (13)$$

where $\log L(\hat{Q}_S)$ is the maximum log-likelihood of the model with *s* components and (3*s*–1) is the number of parameters of the model. According to such a criterion, the model with the smallest BIC is preferred.

2.4 Population size estimation

By means of capture-recapture methods, a given registration system identifies n cases that are observed and leaves a number of cases unobserved. According to an identification system, each unit is identified repeatedly by the same mechanism during the period of study. The number of identifications leads to count data x and frequency f_x of a unique unit identified exactly x times. It is clear from the sample that $f_1, f_2, \dots, f_m$ except f_0 can be obtained where m is the largest observed count. Many works contributed to zero-truncated count models have been proposed to estimate unobserved f_0 and the size N of a target population [3–4].

There are a large number of estimators which are derived from homogeneity and heterogeneity of population. Examples of estimators based on homogeneity are $\hat{N}_{MLE} = \dfrac{n}{1-e^{-\hat{\lambda}}}$ where $\hat{\lambda}$ is the maximum likelihood estimate for the parameter λ of the zero-truncated Poisson distribution, $\hat{N}_{Turing} = \dfrac{n}{1-\left(f_1/S\right)}$ where S = $0f_0 + 1f_1 + \dots + mf_m$ [14]. Furthermore, two popular estimators of Chao [15] and Zelterman [16] allow population heterogeneity. Chao's estimator and Zelterman's estimator are given by $\hat{N}_{Chao} = n + \dfrac{f_1^2}{2f_2}$ and $\hat{N}_Z = \dfrac{n}{1\text{-}e^{\left(-2f_2/f_1\right)}}$, respectively. As presented in the preceding section, the new estimator $\hat{N}_{DBP} = \dfrac{n}{1-\sum_{k=1}^{S}\hat{q}_k e^{-\hat{\lambda}_k} e^{-\hat{\mu}_k}}$ allows homogeneity or heterogeneity relying on one or more component sizes.

With a single list in which the same individual is identified repeatedly during the period of the study. Not only count x and frequency f_x but also two sources can be generated. An informative source is exploited and split that up to time components such as 1st half and 2nd half of the year. Then 1st half and 2nd half are treated as source 1 and source 2, respectively. As a result of two-source situation, the new estimator $\hat{N}_{DBP}$ is obtained. This regarding induces an advantage of the proposed model.

3. A simulation study

3.1 A simulation plan

To compare the suggested estimator with existing estimators, a simulation study was carried out. Data were generated from discrete mixture of bivariate Poisson. One was one-component Poisson model with $p_{ij} = Po(i;1)Po(j;\mu)$, $\mu \in \{1,2,3,4\}$. The other was two-component Poisson model having equal weights with $p_{ij} = 0.5Po(i;1)Po(j;1) + 0.5Po(i;\lambda_2)Po(j;\mu_2)$, $\lambda_2, \mu_2 \in \{1,2,3,4\}$ indicating weak, moderate and strong heterogeneity, respectively. Each simulated data set was arranged in a form of frequencies f_{ij}. Then, f_{00} was left out and f_x was in use where x = i+j. The population size to be estimated was N = 100, 1000 and 10000. Estimates of population sizes are computed by means of MLE, Turing, Chao, DBP and Zelterman where $\hat{N}_{DBP}$ was based on the basis of BIC-selected modelling approach. A large number of replications would be not feasible since the proposed approach is computationally intensive. As a result, the experiment was based on 100 replicates.

Performance of population size estimators is evaluated in terms of bias and mean square error. Due to the fact that with increasing N the expected values and variance increase, we take Relative bias (Rbias), and Relative mean square error (RMSE) to be the following:

$$\text{Relative bias} = \frac{E(\hat{N}) - N}{N} \qquad (14)$$

$$\text{Relative mean square error} = \frac{E\left(\hat{N} - N\right)^2}{N^2} \qquad (15)$$

3.2 Simulation results

Shown in Table 3 and Table 4 are results of the simulation study. Estimating the size N of populations is considered under homogeneity and heterogeneity. We summarize a few major results as following

3.2.1. Results for homogeneity

Since $\hat{N}_{MLE}$ and $\hat{N}_{Turing}$ are derived on the basis of homogeneity Poisson model, they might be expected to be an appropriate choice. It is found from Table 3 that $\hat{N}_{MLE}$ and $\hat{N}_{Turing}$ produce the smallest relative bias among the other. Furthermore, it should be noted that the MLE results are identical to $\hat{N}_{DBP}$ containing one component. The numerical difference is sometimes caused by computations. Similar to $\hat{N}_{MLE}$, $\hat{N}_{Turing}$ and $\hat{N}_{Chao}$, the proposed estimator $\hat{N}_{DBP}$ is asymptotically unbiased with respect to N. It is also found that $\hat{N}_Z$ has the worst performance of accuracy with the largest relative bias for N = 100, 1000 and 10000. It is interesting look at the case of small population (N=100). With increasing values of parameter μ, $\hat{N}_{MLE}$ and $\hat{N}_{DBP}$ have the better bias whereas Zelterman's bias becomes larger.

Based on simulation results, it is clearly seen from Table 4 that under homogeneity $\hat{N}_{MLE}$ as well as $\hat{N}_{DBP}$ perform the best with the smallest relative mean square error (RMSE) for N = 100, 1000 and 10000. Performance of $\hat{N}_{Turing}$ is fairly close to both of $\hat{N}_{MLE}$ and $\hat{N}_{DBP}$. For all population sizes, $\hat{N}_Z$ performs the worst with the largest RMSE.

Table 3: Relative bias of five population size estimators

	Capture probability p_{ij}	MLE	Turing	Chao	DBP	Zelterman
N = 100	Po(1)Po(1)	0.0042	0.0003	0.0016	0.0043	0.0057
	Po(1)Po(2)	0.0034	0.0004	0.0010	0.0034	0.0071
	Po(1)Po(3)	0.0004	-0.0014	-0.0004	0.0004	0.0101
	Po(1)Po(4)	0.0001	0.0003	0.0040	0.0002	0.0359
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	-0.0072	-0.0057	0.0002	-0.0069	0.0086
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	-0.0357	-0.0277	-0.0175	-0.0352	-0.0039
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	-0.0289	-0.0207	-0.0084	-0.0168	0.0106
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	-0.0427	-0.0275	-0.0042	-0.0126	0.0378
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	-0.0541	-0.0362	-0.0102	-0.0034	0.0350
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	-0.0486	-0.0344	-0.0141	-0.0064	0.0217
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	-0.0500	-0.0300	0.0034	0.0186	0.0641
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	-0.0589	-0.0378	-0.0021	0.0103	0.0593
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	-0.0610	-0.0406	-0.0018	0.0052	0.0698
N = 1000	Po(1)Po(1)	0.0000	-0.0014	-0.0024	0.0001	-0.0030
	Po(1)Po(2)	0.0002	-0.0003	-0.0004	0.0002	0.0001
	Po(1)Po(3)	0.0007	0.0003	0.0003	0.0007	0.0014
	Po(1)Po(4)	0.0000	0.0000	0.0003	0.0000	0.0030
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	-0.0164	-0.0122	-0.0073	-0.0137	-0.0023
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	-0.0330	-0.0211	-0.0082	-0.0032	0.0074
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	-0.0328	-0.0198	-0.0040	0.0027	0.0167
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	-0.0444	-0.0269	-0.0050	0.0037	0.0277
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	-0.0562	-0.0361	-0.0087	-0.0011	0.0350
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	-0.0482	-0.0307	-0.0093	-0.0008	0.0216
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	-0.0546	-0.0343	-0.0061	0.0014	0.0396
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	-0.0612	-0.0400	-0.0060	-0.0009	0.0526
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	-0.0629	-0.0418	-0.0040	-0.0004	0.0608
N = 10000	Po(1)Po(1)	0.0005	0.0005	0.0004	0.0005	0.0002
	Po(1)Po(2)	0.0000	0.0002	0.0005	0.0000	0.0009
	Po(1)Po(3)	-0.0001	0.0000	0.0000	-0.0001	0.0003
	Po(1)Po(4)	0.0000	0.0000	0.0000	0.0000	0.0002
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	-0.0153	-0.0100	-0.0039	-0.0001	0.0023
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	-0.0334	-0.0206	-0.0058	-0.0001	0.0131
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	-0.0334	-0.0206	-0.0058	0.0004	0.0130
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	-0.0469	-0.0288	-0.0056	0.0004	0.0294
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	-0.0551	-0.0343	-0.0042	0.0005	0.0455
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	-0.0467	-0.0287	-0.0056	0.0000	0.0294
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	-0.0551	-0.0343	-0.0038	0.0000	0.0471
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	-0.0605	-0.0387	-0.0030	0.0000	0.0588
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	-0.0636	-0.0419	-0.0016	-0.0005	0.0693

Table 4: Relative mean square error of five population size estimators

	Capture probability p_{ij}	MLE	Turing	Chao	DBP	Zelterman
N = 100	Po(1)Po(1)	0.0022	0.0023	0.0038	0.0022	0.0072
	Po(1)Po(2)	0.0007	0.0007	0.0012	0.0007	0.0042
	Po(1)Po(3)	0.0002	0.0002	0.0003	0.0002	0.0022
	Po(1)Po(4)	0.0001	0.0001	0.0002	0.0002	0.0092
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	0.0011	0.0013	0.0025	0.0011	0.0059
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	0.0020	0.0016	0.0019	0.0020	0.0044
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	0.0015	0.0012	0.0017	0.0016	0.0055
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	0.0025	0.0015	0.0018	0.0038	0.0098
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	0.0035	0.0020	0.0017	0.0028	0.0075
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	0.0031	0.0019	0.0016	0.0014	0.0067
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	0.0032	0.0016	0.0021	0.0026	0.0166
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	0.0040	0.0020	0.0016	0.0015	0.0119
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	0.0042	0.0022	0.0019	0.0012	0.0178
N = 1000	Po(1)Po(1)	0.0003	0.0003	0.0005	0.0003	0.0009
	Po(1)Po(2)	0.0001	0.0001	0.0001	0.0001	0.0003
	Po(1)Po(3)	0.0001	0.0001	0.0001	0.0001	0.0001
	Po(1)Po(4)	0.0000	0.0000	0.0000	0.0000	0.0001
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	0.0004	0.0003	0.0003	0.0004	0.0005
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	0.0012	0.0005	0.0002	0.0003	0.0005
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	0.0012	0.0005	0.0002	0.0002	0.0009
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	0.0021	0.0008	0.0002	0.0002	0.0015
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	0.0032	0.0014	0.0003	0.0002	0.0022
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	0.0024	0.0010	0.0002	0.0001	0.0011
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	0.0031	0.0013	0.0002	0.0002	0.0023
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	0.0038	0.0017	0.0002	0.0001	0.0036
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	0.0040	0.0018	0.0002	0.0002	0.0050
N = 10000	Po(1)Po(1)	0.0000	0.0000	0.0000	0.0000	0.0001
	Po(1)Po(2)	0.0000	0.0000	0.0000	0.0000	0.0000
	Po(1)Po(3)	0.0000	0.0000	0.0000	0.0000	0.0000
	Po(1)Po(4)	0.0000	0.0000	0.0000	0.0000	0.0000
	0.5Po(1)Po(1) + 0.5Po(1)Po(2)	0.0002	0.0001	0.0000	0.0000	0.0001
	0.5Po(1)Po(1) + 0.5Po(2)Po(2)	0.0011	0.0004	0.0001	0.0000	0.0003
	0.5Po(1)Po(1) + 0.5Po(1)Po(3)	0.0011	0.0004	0.0001	0.0000	0.0002
	0.5Po(1)Po(1) + 0.5Po(2)Po(3)	0.0022	0.0008	0.0001	0.0000	0.0010
	0.5Po(1)Po(1) + 0.5Po(3)Po(3)	0.0030	0.0012	0.0000	0.0000	0.0022
	0.5Po(1)Po(1) + 0.5Po(1)Po(4)	0.0022	0.0008	0.0000	0.0000	0.0009
	0.5Po(1)Po(1) + 0.5Po(2)Po(4)	0.0030	0.0012	0.0000	0.0000	0.0023
	0.5Po(1)Po(1) + 0.5Po(3)Po(4)	0.0037	0.0015	0.0000	0.0000	0.0036
	0.5Po(1)Po(1) + 0.5Po(4)Po(4)	0.0040	0.0018	0.0000	0.0000	0.0049

3.2.2. Results for heterogeneity

With respect to relative bias, $\hat{N}_{MLE}$, $\hat{N}_{Turing}$ and $\hat{N}_{Chao}$ are underestimating whereas $\hat{N}_{Z}$ is overestimating. In the limited simulation studies it is shown that $\hat{N}_{DBP}$ performs the best with the smallest relative bias under any degree of heterogeneity for N = 1000 and N = 10000 but $\hat{N}_{DBP}$ is not doing good for small population size (N = 100) with weak heterogeneity. It is found that $\hat{N}_{Chao}$ provides the minimum of bias for N = 100 under weak and strong heterogeneity. The underestimation bias of $\hat{N}_{MLE}$ is similar and relatively constant over the different sizes of population and any degree of heterogeneity. It becomes the worst with the largest relative bias under heterogeneity for all population sizes. $\hat{N}_{Z}$ is also the worst for N = 100 under strong heterogeneity.

Achieving the smallest relative mean square error, $\hat{N}_{DBP}$ performs best under moderate and strong heterogeneity for medium size of population (N = 1000) and becomes an excellent estimator under any degree of heterogeneity for large population size (N = 10000). Additionally, it is found that RMSE of $\hat{N}_{DBP}$ is smaller with increasing degree of heterogeneity (results presented with more decimal places are not reported here). $\hat{N}_{Chao}$ performs good as well, particularly for N = 1000 and N = 10000. Both of $\hat{N}_{MLE}$ and $\hat{N}_{Z}$ are not doing good when compared to the other estimators. As can be seen, $\hat{N}_{Z}$ provides the largest RMSE for N = 100 and $\hat{N}_{MLE}$ has the highest RMSE for N = 1000 and N = 10000.

4. Real data example

Estimating the size of a drug use population is of great interest. The data discussed relate to heroin users in the year 2001. The list of the surveillance system is from 61 private and public treatment centers in the Bangkok metropolitan area. The information is constructed on the basis of frequencies of the treatment episodes permitted to treat drug addicts and arise from the surveillance system of the Office of the Narcotics Control Board (ONCB) of the Ministry of Public Health (Thailand). More details of the data source are provided in Lanumteang [17].

Presented in Table 5 is the number of heroin users that contacted the treatment centers in 2001. The count variable of interest is the number of occasions that a specific drug user contacted the treatment centers i times in 1st half year and j times in 2nd half year. It is found that the observed number of heroin users n = 5515.

Analysis of nonparametric maximum likelihood estimation is presented in Table 6. The number of components starts from s = 1 up to s = 3 in which the log-likelihood is not increasing. The NPMLE is provided for three components as well as BIC. Maximum likelihood estimator of the mixing distribution is given by

$$\hat{Q} = \begin{pmatrix} 0.1038 & 0.5944 & 1.8984 \\ 0.5123 & 0.0297 & 1.5903 \\ 0.4419 & 0.4158 & 0.1423 \end{pmatrix}$$

leading to $\hat{N}_{DBP}$ = 10323 for the total number of heroin users and $\hat{f}_{00}$ = 4808 for the hidden.

Table 5**:** Count distribution of heroin user contacts in Bangkok, Thailand in the year 2001

		2nd half year							
		0	1	2	3	4	5	6	
	0	-	1401	369	98	23	1	1	1893
	1	1736	315	129	50	26	1	0	2257
1st	2	445	137	105	53	20	4	0	764
half	3	164	89	75	49	30	1	2	410
year	4	47	25	48	34	8	0	0	162
	5	5	7	8	2	3	0	0	25
	6	1	0	1	1	0	0	0	3
	8	0	0	0	1	0	0	0	1
		2398	1974	735	288	110	7	3	5515

The NPMLE indeed carries the risk for a spurious estimate since small changes in the log-likelihood can cause large changes in population size estimates. For further details see Kuhnert et al. [18]. Therefore, the results based on the NPMLE are cautioned and the estimates of alternative estimators are calculated for comparison (See the lower part of Table 6). As can be seen, MLE and Turing which rely upon the assumption of homogeneity as well as one-component mixture model are close together. The NPMLE here provides a reasonable value lying between Chao's estimate and Zelterman's estimate.

5. Conclusion

Although capture-recapture contributions have experienced theoretical developments, there is not much work available for bivariate count variables. This study focuses on estimating the population size in the two-source situation. We consider a bivariate count variable where counts are used to summarize how often a unit was identified from source/list 1 and source/list 2. Independence for a homogeneous component is assumed and the mixture model is presented to model unobserved population heterogeneity. We propose discrete mixtures of bivariate, conditional independent Poisson distribution to fit the arising two-dimensional frequency table. Parameters of discrete mixing distribution are estimated by means of maximum likelihood estimation via the EM algorithm. The number of components is unknown and is a controversial issue. We have suggested utilizing BIC for model selection. To evaluate the performance of the suggested estimator, comparisons are done among existing estimators derived for homogeneity and heterogeneity. The simulation study provides the positive evidence that the suggested estimator shows good performance and become a candidate for use.

Table 6: Analysis of heroin drug user data in 2001

s	$\hat{\lambda}_k$	$\hat{\mu}_k$	$\hat{q}_k$	$l(\hat{Q}_k)$	BIC	$\hat{f}_{00}$	$\hat{N}$
1	0.8174	0.6750	1	-13242.69	26502.61	1599	7114
2	0.3023 1.8196	0.2481 1.5113	0.8480 0.1520	-12617.17	25277.42	5395	10910
3	0.1038 0.5944 1.8984	0.5123 0.0297 1.5903	0.4419 0.4158 0.1423	-12436.76	24942.44	4808	10323
	MLE					1599	7114
	Turing					2314	7829
	Chao					4358	9873
	Zelterman					5232	10747

It is not easy to deal with computation of maximum likelihood estimation for the bivariate mixture model. The suggested approach is computationally intensive. With regard to long run time, bootstrap resampling technique is not investigated to construct confidence interval for population size N. However, CI associated with the profile mixture likelihood approach could be done based upon the unconditional maximum likelihood suggested by Lerdsuwansri [19].

Acknowledgements

We are grateful to the Ministry of Science & Technology, the Royal Thai Government for providing Ph.D.-funding for Rattana Lerdsuwansri.

References

[1] Borchers DL, Buckland ST and Zucchini W. Estimating Animal Abundance : Closed Populations. Berlin: Springer; 2002.

[2] Seber G. Estimation of Animal Abundance and Related Parameters. Caldwell: Blackburn; 2002.

[3] van der Heijden P, Bustami R, Cruyff M, Engbersan G, and Houwelingen H. Point and interval estimation of population size using the truncated Poisson regression model. Statistical Modelling**.** 2003**;** 3: 305–322.

[4] van der Heijden P, Cruyff M, and Houwelingen H. Estimating the size of a criminal population from police records using the trunated poisson regression model. Statistica Neerlandica. 2003**;** 57: 289–304.

[5] Gallay A, Vaillant V, Bouvet P, Grimont P. and Desenclos J. How many foodborne outbreaks of salmonella infection occurred in France in 1995? Application of the capture-recapture method to three surveillance systems. American Journal of Epidemiology. 2000; 152: 171–177.

[6] Hook E. and Regal R. Capture-Recapture Methods in Epidemiology: Methods and Limitations. Epidemiologic Reviews. 1995; 17: 243–264.

[7] Brittain S. and Böhning D. Estimators in capture-recapture studies with two sources. AStA Adv Stat Anal. 2009; 93: 23–47.

[8] Norris J. and Pollock K. Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics. 1996; 52: 639–649.

[9] Pledger S. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics. 2000; 56: 434–442

[10] Dorazio R. and Royle J. Mixture Models for Estimating the Size of a Closed Population When Capture Rates Vary among Individuals. Biometrics. 2003; 59: 351–364.

[11] Böhning D. and Schön D. Nonparametric maximum likelihood estimation of population size based on the counting distribution. Journal of the Royal Statistical Society C. 2005; 54, 721-737.

[12] McLachlan G. and Peel D. Finite Mixture Models. New York: Wiley; 2000.

[13] McLachlan G. and Krishnan, T. The EM Algorithm and Extensions. New York: Wiley; 1997.

[14] Good I. On the Population Frequencies of species and the Estimation of Population Parameters. Biometrika. 1953; 40: 237–264.

[15] Chao, A. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics. 1987; 43, 783-791.

[16] Zelterman D. Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments. Journal of Statistical Planning and Inference**.** 1998**;** 18: 225–237.

[17] Lanumteang K. Estimation of the Size of a Target Population Using Capture-Recapture Methods Based upon Multiple Sources and Continuous Time Experiments [Dissertation]. University of Reading; 2010.

[18] Kuhnert R., Del Rio Vilas VJ., Gallagher J. and D. Böhning D. A Bagging-Based Correction for the Mixture Model Estimator of Population Size. Biometrical Jurnal. 2008**;** 50: 993–1005.

[19] Lerdsuwansri R. Generalisation of the Lincoln-Petersen Approach to Non-binary Source Variables [Dissertation]. University of Reading; 2012.

Multivariate statistical analysis in three morphologically-related fish species of the genus *Cyclocheilichthys* (family Cyprinidae, order Cypriniformes)

Anan Kentho[1] and Pornpimol Jearranaiprepame[2*]
[1]*Department of Biology, Faculty of Science, Khon Kaen University, Muang, Khon Kaen, 40002, Thailand,*
anan.k@kkumail.com
[2]*Department of Biology, Faculty of Science, Khon Kaen University, Muang, Khon Kaen, 40002, Thailand,*
porjea@kku.ac.th

Abstract

Multivariate morphometric is a method involving mathematics and statistics to determine shape or pattern of interested objects and also various organisms. The present study was to characterise body shape of three morphologically-related fish of the genus *Cyclocheilichthys* which caught in Northeast of Thailand by using statistical methods. Multivariate analysis was carried out to examine the variations and the differences within and among the three fish species including *C. apogon*, *C. armatus* and *C. repasson*. Morphometric variables retrieved from traditional morphometric measurement were performed and then subjected to remove size-dependent variations from morphometric data by Allometric transformation. The multivariate analysis of variance revealed the value of Wilk's Lambda = 0.1870, $F_{(48,544)}$ = 14.8732, and *p-value* < 0.001, which indicated significant differences among the three species with the least of a pair of population differences. The first three principal components of the principal component analysis accounted 57.41% of total variance, suggesting morphological variations among three species. The dynamics of the fish characters of each species was apparently displayed on various body parts of head, body and fin in correlation to their ecological niches. Discriminant function analysis additionally distinguished fish sample to their actual species in original and cross-validated classification tests with 87.58 and 85.2%, respectively with the confidential value of Wilk's Lambda = 0.2170, approx. $F_{(24,568\ )}$= 27.1878, and *p-value* < 0.001. Therefore, it was concluded that statistical analysis could be of use to elucidate the variations and the differences of morphological structure within and among the three species cyprinid fish.

Keywords: Statistical analysis, multivariate morphmetrics, traditional measurement, cyprinid fish

*Corresponding Author
E-mail Address: porjea@kku.ac.th

Analysis of fatal accidents using association rules

Desmond Lobo
Faculty of Science, Naresuan University, Phitsanulok, Thailand, DesmondLobo@yahoo.com

Abstract

As in several other developed countries, the number of senior citizens in the UK is increasing at a dramatic rate. A concern is that many of these elderly individuals continue to drive because there is no upper age limit for operating a motor vehicle in the country. Aging can affect individuals in various ways: reaction time slows, hearing and vision deteriorates, the neck and joints stiffen, muscles weaken, etc.

The objective of this research was to determine which factors could be associated with fatal accidents, as opposed to accidents that caused just severe injuries, when seniors over the age of 75 were behind the wheel. Road safety data was obtained from the UK’s Department of Transport for the years 2011 and 2012. For each accident, the following information was analyzed: accident severity, number of vehicles involved in the accident, number of casualties, day of the week, speed limit, light conditions, weather conditions, road surface conditions, urban or rural area, purpose of the journey, and age/sex of the driver. An advanced data mining technique known as association rules was then used to uncover relationships in the data.

For senior gentlemen, the results of this research showed that fatal accidents could be associated with driving at a high speed in rural parts of the country. Fatal accidents were also linked with senior ladies driving in rural areas, although speed was not a factor. Lack of public transportation in the countryside might account for some seniors having no other alternative but to drive.

Keywords: Association rules, senior drivers, fatal accidents, road safety, United Kingdom

Corresponding Author
E-mail Address: DesmondLobo@yahoo.com

Fitting linear mixed models to longitudinal linked data

Klairung Samart[1*] and Ray Chambers[2]

[1]*Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, 90110, Thailand, klairung.s@psu.ac.th*

[2]*National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, NSW 2522, Australia, ray@uow.edu.au*

Abstract

Linked data sets are particularly useful in many areas such as epidemiology, health, demography and sociology. One problem arising from linking process is the occurrence of linkage errors. Although statistical methods for linking data sets are now well established and the research on impact of linkage errors on analysis of linked data has been carried out, it was mainly focused on linked records from two distinct data sets. In this research we develop unbiased regression parameter estimates when fitting a linear mixed model to longitudinal linked data where registers are linked over time. Furthermore, we develop appropriate modifications to standard methods of variance components estimation in order to account for linkage error. In particular, we focus on three widely used methods of variance components estimation: analysis of variance, maximum likelihood and restricted maximum likelihood. Simulation results and application on real life data show that our estimators perform reasonably better than standard estimation methods that ignore linkage errors.

Keywords: Analysis of variance, linkage error, longitudinal analysis, maximum likelihood, restricted maximum likelihood, weighted least squares

*Corresponding Author
E-mail Address: klairung.s@psu.ac.th

Comparing results between fuzzy set interpretation and raw data interpretation

A. Pongpullponsak[1] and R. Chonchaiya[2]*
[1]*Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, Thailand, adisak.pon@kmutt.ac.th*
[2]*Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand, hengmath@hotmail.com*

Abstract

In this paper, we focus on how to use fuzzy logic in order to interpret the level of happiness among Thai teachers in Ratchathewi, Bangkok. In order to interpret the survey results, we used 2 different methods, i.e. the direct interpretation method and the fuzzy set method for the mathematically data analysis; and then compared the responses from each method. The result shows that there is some significantly statistical difference between two methods. In addition, the fuzzy set method is better than the raw data method since the result after defuzzification has a smaller variance.

Keywords: Component, fuzzy set, happiness at work, questionaire interpretation

*Corresponding Author
E-mail Address: hengmath@hotmail.com

1. Introduction

One of the popular techniques among researchers in social sciences is using questionaires. Cohen, Manion and Morrison [9] provided a list of questionnaire designings; for example, closed and open questions comparing, dichotomous questions, multiple choice questions, rating scale, open- ended questions and matrix questions.

In order to gather the data from evaluators, using Likert scale is one of the most common and helpful methods. When we looked thoroughly at collecting the data using questionaire, we had some concerns. When a respondent needs to make a decision and fill out a questionnaire, there might be some questions that they probably agree partially or they might not sure if they should choose "agree" or "strongly agree". Lazim, Salihin and Osman [10] said that the subject preference, such as "Strongly Agree" is fuzzy. It is naturally quality rather than the lack of information or knowledge to judge the available rating of the respondent.

Therefore, the fuzzy set model is needed in order to reduce the vague judgement caused by the discreteness of the rating scale.

2. Research work on happiness at work

Nowadays, in the era of information and technology, everyone needs to be well educated and open to advance technologies now and in the future. That means Thai educational institutions also need to prepare their teachers to be up-to-date and focus on students. Moreover, in the national education act B.E. 2542 (see [8]), in the section 57 : "Educational agencies shall mobilize human resources in the community to participate in educational provision by contributing their experience, knowledge, expertise, and local wisdom for educational benefits. Contributions from those who promote and support educational provision shall be duly recognized". That means teachers have to work harder than before which might affect their happiness at work as well.

There are several research papers on happiness at work and its influencing factors. In 2009, Geounuppakul et. al. [4] explored the level of happiness and factors influencing happiness of people working at Boromarajonani College of Nursing Bangkok by using questionaires modified from Rodtiang, 2007 [5] . The original questionaires included those asking about personal factors, family relation, organizational factors and happiness in the workplace. The authors modified the questionaires in order to satisfy the behavioral, environmental and cultural factors of Boromarajonani College of Nursing, Bangkok.

In 2013, Bryson and MacKerron [7] explored the link between individuals' well being measured at random points in time and their experiences of paid work. The questionaires measured on 3 factors of momentary happiness: how happy, how relaxed, and how awake they feel.

3. Fuzzy set model

Fuzzy concepts have distracted people's mind for many decades and became a popular and very interesting subject among computer engineers, mathematicians and statisticians as well as philosophers and psychologists. One of the reasons why fuzzy or vague concept cannot be formulated by ordinary mathematics easily is this concept does not include definitive results. So, we need to know the different mathematical concepts to explain the mathematical modeling of the fuzzy idea.

The idea of fuzzy concept is related to Frege's (1904) boundary-line view. A concept is fuzzy if there are some objects which cannot be classified either to the concept or to its complements but are members of the concept's boundary. The first successful approach to fuzziness was the notion of a fuzzy set proposed by Lotfi Zadeh [6]. In this approach, sets are defined by partial membership in contrast to crisp membership used in classical definition of a set.

Fuzziness can be occurred in so many areas where human judgment, conclusion, and decision are involved, for example, engineering, medicine, artificial intelligence, pattern recognition, meteorology, computer science education, psychology, sociology etc.

There are some benefits of studying about fuzzy concept. One of them is to give the interpretation of "non-random uncertainty" since the vague concept provides a fantastic tool to measure the level of uncertainty or imprecision. This concept also helps when information is lacking regarding a particular respond of the subject and a conclusion is needed.

There are several research work using fuzzy set method. For example, in 1998, Lafuente et al. [11] conducted a research by interviewing wheelchair users concerning the adaptation of his/her wheelchair to the office workplace. The authors used fuzzy model in order to interpret the users' preference mathematically.

In 1999, Hassall [12] studied how to analyse the ordinal or interval questionnaire data using fuzzy set method. The author proposed 2 formulations as treatments of the result from the questionnaire which recognizes the inherently imprecise nature of the judgement beingmade; triangular fuzzy number and best hypothesis determination.

In 2012, Gomez et. al. [13] applied fuzzy logic techniques to the response given by the professors from Universidad Complutense de Madrid (UCM) and other Spanish faculties. The questionnaire asked about the professional competences that are basic for social work in Spain. The results showed that the processing of the data obtained from the research questionaires through fuzzy logic technique makes decision taking easier.

In 2013, Chonchaiya and Pongpullponsak [14] used fuzzy logic for grade evaluation and compared the results with another 2 system of grade evaluation, namely, criteria reference system and group reference system. It has been shown that the fuzzy method gave the most reasonable grade evaluation.

In this paper, we are studying on using fuzzy set method in order to interpret the factors influencing happiness at work among teachers who teach in Primary school level in Ratchathewi, Bangkok. We applied iOpener Performance happiness model [3] in 5 categories : Contribution, Conviction, Culture, Commitment and Confidence. This questionaires, asked about teachers' attitude or happiness at work and contained 29 questions of Likert-type from a scale 1 to 5. For example, "I am proud to be a teacher" is follows by a 5 point rating scale of strongly agree (5), agree(4), neither agree nor disagree (3), disagree (2) and strongly disagree (1).

4. Terminology and definition

Definition 1 [1] (membership function) For a set A, we define a membership function μ_A such as

$$\mu_A(x) = \begin{cases} 1, & \text{if and only if} \quad x \in A \\ 0, & \text{if and only if} \quad x \notin A \end{cases}$$

We can say that the function μ_A maps each element in the universal set X to the set $\{0,1\}$, i.e.

$$\mu_A : X \to \{0,1\}.$$

More precisely, the membership function μ_A in crisp set maps whole members in the universal set to $\{0,1\}$.

Definition 2 [1] (membership function of fuzzy set) In a fuzzy sets A, each element is mapped to $[0,1]$ by a membership function

$$\mu_A : X \to [0,1],$$

where $[0,1]$ means the set of real numbers between 0 and 1 (including 0 and 1).

Definition 3 [1] (Fuzzy number) If a Fuzzy set is convex and normalized, and its membership function is defined in R and piecewise continuous, it is called as "Fuzzy number". So fuzzy number (fuzzy set) represents a real number interval whose boundary is fuzzy.

Definition 4 [1] (Triangular Fuzzy number) It is a fuzzy number represented with three points as follows: $\tilde{A} = (a_1, a_2, a_3)$. This representation is interpreted as membership functions and holds the following conditions (Figure 1).

(i) it is an increasing function from a_1 to a_2

(ii) it is a decreasing function from a_2 to a_3

(iii) $a_1 \le a_2 \le a_3$.

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & \text{for} \quad x < a_1 \\ \dfrac{x-a_1}{a_2-a_1} & \text{for} \quad a_1 \le x \le a_2 \\ \dfrac{a_3-x}{a_3-a_2} & \text{for} \quad a_2 \le x \le a_3 \\ 0 & \text{for} \quad x > a_3 \end{cases}$$

Definition 5 [1] (Trapezoidal Fuzzy number) We can define a trapezoidal fuzzy number $\widetilde{A} = (a_1, a_2, a_3, a_4)$. The membership function of this fuzzy number will be interpreted as follows (Figure 2).

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & \text{for} \quad x < a_1 \\ \dfrac{x - a_1}{a_2 - a_1} & \text{for} \quad a_1 \le x \le a_2 \\ 1 & \text{for} \quad a_2 \le x \le a_3 \\ \dfrac{a_4 - x}{a_4 - a_3} & \text{for} \quad a_3 \le x \le a_4 \\ 0 & \text{for} \quad x > a_4 \end{cases}$$

Definition 6 [6] (Centroid Method) This procedure (also called center of area or center of gravity) is the most prevalent and physically appealing of all the defuzzification methods. It is given by the algebraic expression

$$x^* = \frac{\int \mu_{\tilde{A}}(x) \cdot x dx}{\int \mu_{\tilde{A}}(x) dx},$$

Where $\int$ denotes an algebraically integration.

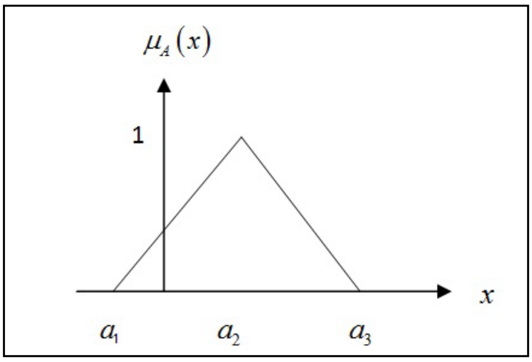


Figure 1 Triangular Fuzzy number

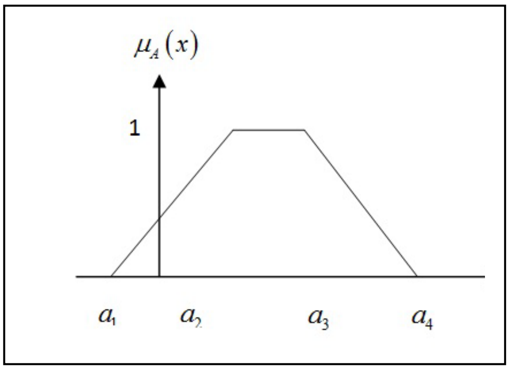


Figure 2 Trapezoidal Fuzzy number

5. The numerical Results

This research used a quantitative approach to collect data using questionaires from 100 primary teachers in Ratchathewi, Bangkok. We measured 5 factors of happiness at work by using a questionnaire applying the iOpener Performance Happiness Model. In order to interpret the survey results, we used 2 different methods, i.e. the raw data interpretation method and the fuzzy set method for the mathematically data analysis.

First of all, what we mean by "raw data interpretation" is the interpretation by adding all the rating scores from each question and then considering from the following criteria.

Total score	Translation
29-43	Very Unhappy
44-72	Unhappy
73-101	Neither happy nor unhappy
102-130	Happy
131-145	Very Happy

Table 1: Results from the raw data Interpretation Method

Group	Percent of respondents
Very Unhappy	0
Unhappy	0
Neither happy nor unhappy	5
Happy	67
Very Happy	28

In Table 1, We analysed the survey responses and counted the number of respondents in each happiness group. We found that more than 25% of the teachers are very happy to work in their school even there are so many stressful environments and difficulties, for example, their degree is nothing about what they are teaching, too much workload from the school or the time spending on traveling from home to the school.

As we have mentioned before that some responses might be biased or unclear because of the respondents. We are now applying the Fuzzy Set Method in order to interpret the result and compare whether this method gives the difference significantly.

Next, we are going to compare the results between the raw data interpretation method and the fuzzy set method. In this study, "happiness" is a variables whose values are "strongly agree (5)" , "agree (4)", "neither agree nor disagree (3)", "disagree (2)" and "strongly disagree (1)".

Since a fuzzy set is characterized by its membership function, we have to choose the shape for those five values of "happiness". For the rating scale 1 and 5, we choose the trapezoidal shape membership functions and the rest we choose the triangular membership functions. Figure 3 shows the membership functions of the input and output of the fuzzy set method.

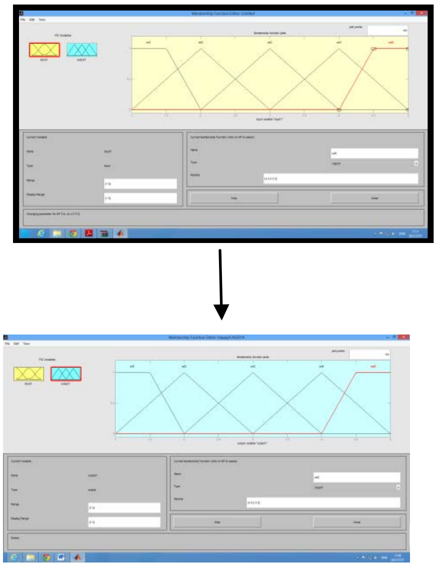


Figure 3: The Membership Functions of the Input and Output for fuzzy set analyzing.

Table 2: Results from the fuzzy set Interpretation Method

Group	Percent of respondents
Very Unhappy	0
Unhappy	0
Neither happy nor unhappy	5
Happy	91
Very Happy	4

Considering from Table 1 and Table 2, we have seen the differences between the numbers of each group of happiness which are given by using those 2 methods. One of the reasons for the differences is the fuzziness of the data collection. In Figure 4, the blue graph showed that the traditional method gave a higher or equal result compare to the red graph of the Fuzzy Set method.

The next step is to test whether the traditional method is statistical different significantly from the Fuzzy Set method.

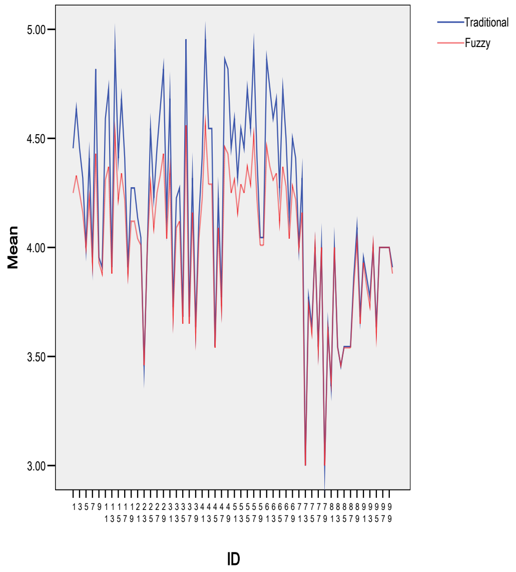
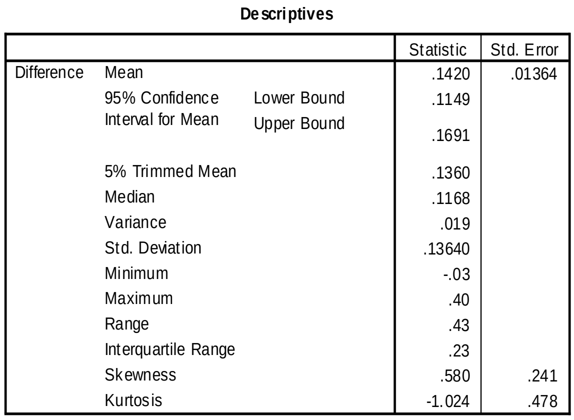


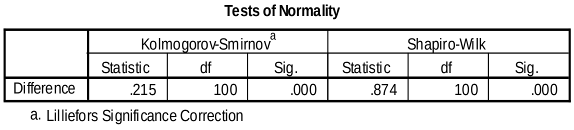
Figure 4: The Comparison of the Results Between the Direct Method (blue) and the Fuzzy Set Method (red)

Table 3: The output table on the normality of the difference between the results between the raw data interpretation method and the fuzzy set interpretation method.



Descriptives

			Statistic	Std. Error
Difference	Mean		.1420	.01364
	95% Confidence Interval for Mean	Lower Bound	.1149	
		Upper Bound	.1691	
	5% Trimmed Mean		.1360	
	Median		.1168	
	Variance		.019	
	Std. Deviation		.13640	
	Minimum		-.03	
	Maximum		.40	
	Range		.43	
	Interquartile Range		.23	
	Skewness		.580	.241
	Kurtosis		-1.024	.478

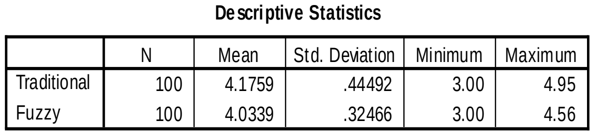


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Difference	.215	100	.000	.874	100	.000

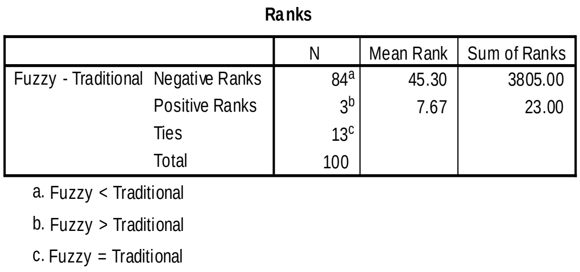
a. Lilliefors Significance Correction

Table 4: The output table on Statistical Analysis on the differences between the results from the raw data interpretation method and the fuzzy set interpretation method using wilcoxon test.



Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Traditional	100	4.1759	.44492	3.00	4.95
Fuzzy	100	4.0339	.32466	3.00	4.56



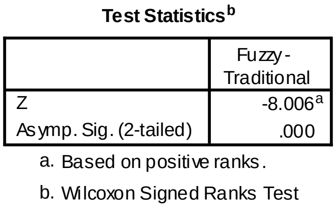
Ranks

		N	Mean Rank	Sum of Ranks
Fuzzy - Traditional	Negative Ranks	84[a]	45.30	3805.00
	Positive Ranks	3[b]	7.67	23.00
	Ties	13[c]		
	Total	100		

a. Fuzzy < Traditional

b. Fuzzy > Traditional

c. Fuzzy = Traditional



Test Statistics[b]

	Fuzzy - Traditional
Z	-8.006[a]
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

6. Conclusion

The aim in order to confirm whether this new method of interpretation is really statistical different significantly from the previous method. The suggested null and alternative hypothesis are:

H_0: There is no difference between the result from the raw data interpretation method and the result from the fuzzy set method

H_1: There is a difference between the result from the raw data interpretation method and the result from the fuzzy set method

First of all, the normality test is needed. From the output in Table 3, using Kolmogorov-Smirnov Test, it can be concluded that the normality of the difference between the results from those 2 interpretation methods cannot be assumed. Therefore, we decided to use Wilcoxon Test.

From the output in Table 4, we compared the data from both methods. We can see that sig.=0.00 < 0.05 which can be concluded that these two methods are different significantly. One of the reasons why a new method makes a difference is the fuzzy set model can extract the subjective criteria that played a role in the respondent perception of the happiness at work which depends on the emotions and experiences.

Moreover, we can notice from the Table 4 that the variance of the responses of the method using the raw data is more than the variance of the response from the Fuzzy Set Method. This shows that the interpretation using the Fuzzy Set Method is better than the original method.

References

[1] Lee KH, First Course on Fuzzy Theory and Applications. Springer Berlin Heidelberg NewYork; 2009.

[2]Jessica P. Happiness at Work Maximizing Your Psychological Capital For Success. 1st ed. UK: Wiley-Blackwell; 2010

[3] iOpener Institute for People and Performance, The Science of Happiness At Work [Update 2011; Cite 2014 Feb 20] Available from: http://www.iopenerinstitute.com/the-science-of-happiness-at-work%E2%84%A2.aspx

[4] Geounuppakul M. Tounprommarat A. and Prinpijarn A., Factors Influencing Happiness in Personal Working at Boromarajonani College of Nursing, Bangkok, Thailand. Paper of the 57th Session of the International Statistical Institute; 2009.

[5]Rodtiang N. Factor Influence on Happiness at Work Among Personel in North-Eastern Region Health Promotion Center, Department of Health Ministry of Public Health. [Dissertation] Bangkok: Mahidol University; 2007.

[6] Zadeh LA. Fuzzy sets. Information Control, 8, 338-353; 1965.

[7] Bryson A and MavKerron G. Are You Happy While You Work?, CEP Discussion Paper No 1187; February 2013

[8] Office of the National Educational Commission , National Education Act B.E. 2542 (1999) [Update 2011; Cite 2014 Feb 20] Available from: http://planipolis.iiep.unesco.org/upload/Thailand/Thailand_Education_Act_1999.pdf

[9] Cohen L. Manion L and Morrison K. Research Methods in Education. London: Routledge Falmer; 2007

[10] Lazim M Salihin W and Osman A. Fuzzy Sets In The Social Sciences: An Overview Of Related Researches, Jurnal Teknologi. 2004; 41(E): 43-54.

[11] Lafuente R Page A Sanchez-Lacuesta J. Application of Fuzzy Logic Techniques for the Qualitative Interpretation of Preferences in a Collective Questionaire for Users of Wheelchairs, Journal of Rehabilitation Research and Development. 1998; Vol.35, No.1: 91-107.

[12] Hassall J. Method of Analysing Ordinal/ Interval Questinaire Data using Fuzzy Mathematical Principle, Wolverhamton Bussiness School, Management Research Centre, University of Wolverhampton. 1999. ISSN 1363-6839

[13] Gomez FG Gomez MPM and Gans AG. Fuzzy Logic Techniques (FLT) in the interpretation of the responses Given to a Questionaire Filled out by Professor in Spain, International Journal of Humanities and Social Science. 2012; Vol. 2 No. 20 [Special Issue October].

[14] Susan J. Likert scales: how to (ab)use them, Medical education. December 2004; Vol. 38, Issue 12 : 1217-1218.

[15] Chonchaiya R and Pongpullponsak A. Fuzzy-Set Method for Grade Evaluation. Paper of the International Conference on Applied Statistics; 2013

[16] The Math WorksTM, MATLAB, 7.6.0 (R2009a), License Number 350306, February 12, 2009

Forecasting models for Thai ethanol prices using multiple regression and artificial neural networks

Rojanee Homchalee[1*] and Weerapat Sessomboon[2]
[1]*Department of Mathematics, Faculty of Science, Mahasarakham University, Mahasarakham 44150, Thailand,*
e-mail address: rojaneeh@hotmail.com
[2]*Department of Industrial Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand,*
e-mail address: weerapatkku@hotmail.com

Abstract

In this paper, forecasts of domestic and export prices of Thai ethanol were modeled by considering factors that are expected to affect the forecasting. The data used to develop the forecasting models were collected from several agencies. Two methods, multiple regression and artificial neural networks were used to determine the forecasting models. The results showed that although the stepwise multiple regression analysis provided the predicted models, so called MR models, and contained statistically significant factors affecting the dependent variables, they were unsatisfactory when compared with artificial neural networks models, so called ANN models. In addition, integration of stepwise multiple regression analysis and artificial neural networks by taking affected factors from the stepwise multiple regression analysis as inputs in artificial neural networks, were used to formulate the forecasting models, so called MR-ANN models. The MR-ANN models were superior because they provided the lowest MAPE and highest R^2, particularly the prediction model for domestic ethanol price. Therefore, the integration of multiple regression and artificial neural networks was proposed to model the forecasting of Thai ethanol price for both domestic and export market.

Keywords: Artificial neural networks, ethanol price, forecasting model, multiple regression analysis

*Corresponding Author
E-mail Address: rojaneeh@hotmail.com

1. Introduction

Thailand set ethanol as renewable energy in the alternative energy plan in order to reduce dependence on petroleum. Ethanol consumption in Thailand is associated with gasohol consumption. Fuel grade ethanol (99.5% purity ethanol) has been blended with gasoline as gasohol including gasohol 95 E10, gasohol 91 E10, gasohol E20 and gasohol E85 for domestic consumption. Under the Liquor Act, sale of ethanol is only permitted to oil traders or oil refineries. Ethanol has been denatured with 0.05% gasoline and shipped from ethanol plants to gasohol blending depots, refineries and oil terminals with blending facilities. Of such government policies and the relatively high price of gasoline, gasohol consumption in Thailand increased continuously. Total consumption of gasohol in 2007 was 1,762.76 million liters and increased to 4,454.73 million liters in 2012 with a major consumption of gasohol 95 E10 and gasohol 91 E10. Consequently, ethanol consumption is raised. In 2007 the use of 99.5% purity ethanol was 176.28 million liters and increased to 508.94 million liters in 2012 [1].

Domestic price of ethanol depends on the agreement between sellers (ethanol manufacturers) and buyers (fuel traders). Thai government in particular, the Ministry of Energy set the price as reference for domestic market [2]. The ethanol reference price is used for decision making of ethanol manufacturers and fuel traders, especially price negotiation. In the last few years, ethanol reference prices of Thailand were calculated based on ethanol price of Brazil and freight rate from Brazil to Thailand, since Brazil is the main ethanol producer and exporter [3]. Starting from the year 2009, the ethanol reference prices were calculated from the production cost of molasses-based and cassava-based ethanol, and the production of molasses and cassava, since they are major feedstocks for ethanol production in Thailand. This formula was called “cost-plus” and it was used until end of the year 2011. From January 2012 to present, the ethanol reference price was calculated from average sales price of all ethanol manufacturers weighted by their sales volume. This calculation principle is expected to be a reference price that closes to actual price of domestic market [4].

However, ethanol price of Thailand is fluctuating monthly. In 2007, average ethanol price was 17.52 Baht/liter and increased continuously until the year 2011 which was 24.27 Baht/liter, subsequently it fell to 20.80 Baht/liter in 2012 [5]. This may be due to a change in the rules to calculate the ethanol reference price. Nonetheless, the rules for calculating the price of ethanol may change in the future. It may also relate to the quantity and price of feedstocks used in the production since the feedstocks are significant cost of ethanol production in Thailand.

Besides domestic consumption, Thailand also has been exported ethanol to many countries, particularly industrial grade ethanol (95% purity ethanol). Exported ethanol is not denatured by gasoline. The export volume of ethanol increased continuously. In 2007, the total export volume of 80% purity or higher undenatured ethanol was 46.47 million liters which mainly export to Japan and Singapore. For 2012, the export volume highly increased to 321.07 million liters but mainly export to Philippines, Japan and Korea [6]. The export price depended on the export volume and FOB value, and the price was fluctuating. During 2009 to 2010, the export price was high then decreased in the later two years. However, the price movement was not explicitly. Consequently, prediction making was difficult.

Forecasting of domestic and export ethanol price is necessary for both ethanol manufacturers and fuel traders, and also useful for exporters and overseas customers. The forecast is a key decision making in their business. To find appropriate forecasting models, this study aimed to formulate the forecasting models using multiple regression analysis and artificial neural networks.

2. Literature review

Forecasting methods can be divided into two groups. The first group is called qualitative forecasting examples include good guesses, a Jury of Executive Opinion, the Delphi method. Another group is quantitative forecasting which could be time series analysis, regression analysis and econometric analysis [7, 8]. Another quantitative forecasting method is artificial neural networks, which is recognized and widely used today [9]. However, each method has different advantages and disadvantages depending on the nature of the available data and the limitations of predicting or implementation.

2.1 Multiple regression analysis

Regression analysis is a statistical tool used to predict a dependent variable that depends on one or more independent variables. The purpose of regression analysis is to construct forecasting model [10]. For multiple regression analysis, dependent variable depends on more than one independent variable. Although there are several methods in multiple regression analysis for selecting the independent variables to determine a forecasting model, a stepwise method is popular and widely used because it is effective [11].

However, there are important assumptions that should be examined in the multiple regression analysis, such as autocorrelation and multicollinearity problems. The model of multiple regression analysis reasonably examined both problems. Autocorrelation problems occur if the errors are serially correlated. The forecasting model will not exhibit autocorrelation problems if the Durbin-Watson statistic has a value between 1.5 and 2.5 [12]. Multicollinearity occurs if there is a high degree of correlation between independent variables. In this case, estimation is not possible, or, at a minimum, the accuracy of the estimates is reduced. The degree of multicollinearity can be found by calculating the variation inflation factor (VIF). If the value of the VIF is less than 5, the multicollinearity is not significant, whereas if the VIF is greater than 5, the multicollinearity is significant. Multicollinearity is a more serious issue if the VIF is greater than 10 [13]. Additionally, multicollinearity can be determined by using tolerance statistics. If the tolerance statistics of all the independent variables are approximately equal to 1, then multicollinearity is not an issue [14].

Additionally, the use of regression analysis to determine a forecasting model can provide information regarding the performance of the model by using the coefficient of determination (R^2) and the standard error of estimation (SE).The coefficient of determination is used to interpret the percent of variation in the dependent variable, which could explain all of the independent variables in the forecasting model [11].

2.2 Artificial neural networks

Artificial neural networks are computational model that was inspired from human nervous system [9]. These neural networks widely used in many fields because they are capable of solving many complex or real-world problems including nonlinearity, a high degree of parallelism, faults and noise. They are additionally capable of learning and generalization [9, 15].

The development of an artificial neural network requires dividing a dataset into three subsets: training, test and validation. The training set is used to update the weights of the network. The test set is used during the learning process to verify the network response to data that were not used for training. The validation set is used after selecting the best network for further examination of the network or to confirm its accuracy before it is implemented in the neural system. A large test set can emphasize the generalisation capability better, whereas a smaller training set might not be adequate to train the network. Currently, there are no mathematical rules for determining the required sizes of the various sets [15]. In general, the test set should be 20% of the entire dataset [16]. Another recommendation is to use 65% for the training set, 10% for the validation set and 25% for the test set [17].

There are a number of different learning algorithms that can be used to train a neural network; however, the backpropagation paradigm has become the most popular algorithm for prediction and classification problems [15, 18, 19]. A multilayer feed-forward network typically has three layers (the input, hidden, and output layers). One hidden layer is the most popular because it is sufficient for most applications. A typical approach to determine the optimum number of hidden nodes is by using trial and error [15]. Various training algorithms have been used for network training including gradient descent with momentum, scaled conjugate gradient, Levenberg-Marquardt, quasi-Newton, conjugate gradient and variable learning rate [20]. A high learning rate will accelerate the training process by changing the weights considerably from one cycle to the next. In contrast, a low learning rate drives the search steadily (but slowly) in the direction of the global minimum [15]. Therefore, it is recommended that the learning rate be in the range of 0.0-1.0 [21]. Additionally, a momentum coefficient is typically used to update weights and allow the search to avoid local minima, which reduces the likelihood of search unsteadiness. A high momentum reduces the risk of the network becoming stuck at a local minimum, but it increases the risk of overshooting the solution, as does a high learning rate. Depending on the problem, the completion of training varies with the selected momentum through a trial and error procedure that is typically used [15]. It is recommended that the momentum value is set in the range of 0.0-1.0 [21]. Another method relates momentum to the adaptive learning rate such that momentum decreases as learning accelerates, for example, a learning rate of 0.1 and a momentum of 0.9. This determination is consistent with suggestions that the learning rate plus the momentum is approximately one [22].

Many studies have used multiple regression and artificial neural networks for forecasting because they have different advantages and disadvantages [9, 23-25]. For example, both multiple regression and artificial neural networks were used to estimate the energy demand in South Korea for decision making of energy policy. Artificial neural networks model with a feed-forward multilayer network and a backpropagation algorithm were used. The model provided a more accurate forecast than a linear regression model or an exponential model in terms of RMSE without any over-fitting problem [26]. In other example, models for predicting fuel consumption of tractor were developed using backpropagation artificial neural network models with six training algorithms. The highest performance was obtained from the network with two hidden layers each having 10 neurons, and which employed the Levenberg-Marquardt training algorithm. The results indicated that the artificial neural networks and stepwise regression models produced similar determination coefficients, but the artificial neural networks provided relatively better prediction accuracy compared to stepwise regression [20].

From literatures, the predictive models obtained with multiple regression and artificial neural networks were applied in several case studies including the comparison of multiple regression and artificial neural networks with actual and simulated data. Kumar [25] showed that regression was much better than neural networks for skewed data. Kim [27] compared the performance among linear regression, artificial neural networks and decision trees. A set of data generated by varying a numbers, types and classes of independent variables, and sample sizes, was used in the comparison. In a case of continuous independent variables, results showed that linear regression was superior to artificial neural networks and decision trees regardless of the number of variables and the sample size. In a case of categorical independent variables, linear regression was the best when the number of categorical variables was one, whereas the artificial neural network was superior when the number of categorical variables was two or more. Moreover, performance of artificial neural network improved faster than the other methods as the number of classes of categorical variables increased.

The choice between artificial neural networks and statistical techniques depends on the problem to be solved. However, for modeling the data which is low dimensionality or for approximating simple functions, classical techniques based on statistics should be tried first, and artificial neural networks may then be employed if higher accuracy is needed [15]. Therefore, we proposed the use of multiple regression and artificial neural networks to forecast ethanol price for both domestic and export market that consider factors affecting them. Furthermore, the integration of multiple regression and artificial neural networks was used in this study to design the predicted model in order to increase the forecasting accuracy.

3. Research Methodology

In this paper, the forecasting of domestic ethanol price and export price of Thailand was modeled by considering factors that expected affecting them. The ethanol reference price (Erp) as domestic price of 99.5% ethanol, base on cassava production (Cpr), sugarcane production (Spr), cassava chips price (CCp) and molasses price (Mp) for domestic use. There is also an important factor, the ethanol price in Brazil ($AEBp$) which is the main producer of the world. Then, these factors are expected affect the domestic price of ethanol as in (1).

$$Erp = f(CCp, Mp, Cpr, Spr, AEBp) \tag{1}$$

The factors expected to influence export prices of 80% or higher undenatured ethanol ($UEep$) including ethanol export volume ($UEeq$), ethanol reference price (Erp), ethanol price in Brazil ($AEBp$), the exchange rate (ER) and world crude oil prices ($COilp$) as seen in (2).

$$UEep = f(UEeq, Erp, AEBp, ER, COilp) \tag{2}$$

These variables are monthly historical data for the period January 2007 to December 2011 and were collected from several agencies such as the Department of Alternative Energy Development and Efficiency, the

Energy Policy and Planning Office, the Department of Energy Business, the Office of the Cane and Sugar Board, the Excise Department, the Thai Sugar Millers Corporation Limited, the Thai Tapioca Trade Association, and the Thai Tapioca Starch Association.

Two methods, multiple regression and artificial neural networks were used to determine the forecasting models. The artificial neural networks models are called ANN models, and the stepwise multiple regression models are called MR models. Furthermore, we integrated two methods by taking the independent variables from the stepwise multiple regression analysis as inputs in artificial neural networks. These models from integrated approach are called MR-ANN models as shown in Fig. 1.

For the artificial neural networks, the data was divided into three sets including training, validation and test set with the ratio of 65%, 10%, and 25%, respectively. Dividing data into three parts is to prevent over-fitting problem. Data was divided randomly by repeating 100 times; and then select the closest pattern in order to be representative data for constructing the ANN models. Therefore, this dividing data by randomly did not affect to forecast the time series data with seasonal variation. ANN models were constructed by feedforward multilayer and backpropagation algorithm with Levenberg–Marquardt training algorithm (LM). Input nodes of each model are independent variables and the output node is dependent variable. We used only one hidden layer with a varying a number of hidden nodes ranging from 5 to 20. The training mode was designed with 500 training cycles with a learning rate of 0.0001. Log-sigmoid and tan-sigmoid functions were used as transfer function. Then, the performance of neural networks was measured by a coefficient of determination (R^2) and mean absolute percentage error (MAPE). These key performances are unaffected by outliers, independent of scale, and are widely used. Additionally, these criteria are easy to implementation and intuitive for practitioner [19]. Consequently, ANN, MR, and MR-ANN were again compared by R^2 and MAPE for all data. Then, appropriate models with lowest MAPE and highest R^2, were proposed to forecast domestic and export price of Thai ethanol.

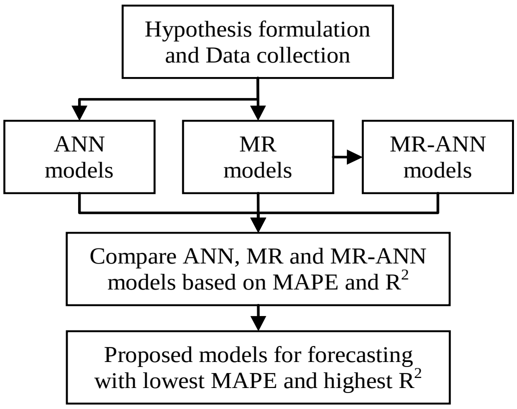

Hypothesis formulation and Data collection
ANN models
MR models
MR-ANN models
Compare ANN, MR and MR-ANN models based on MAPE and R2
Proposed models for forecasting with lowest MAPE and highest R2

Figure 1: Framework of methodology

4. Research Results

As our hypothesis in (1), stepwise multiple regression analysis was used to obtain a forecasting model of ethanol reference price, domestic price. The results showed that there were many factors significantly affecting ethanol reference price, including cassava production, sugarcane production, cassava chip price and molasses price. The forecasting model, MR model as in (3) indicated that these factors explain the variation of domestic ethanol price with R^2 of 69.4%. The most influential factor considered from the standardized coefficient was cassava chip price, and following influential factor was cassava production as shown in Table 1. This indicated that cassava is a factor significantly impacting the price of ethanol in Thailand.

$$\begin{aligned}\hat{E}rp &= 15.463 + 1.366CCp - 11.941(Mp)^{-1} \\ &\quad + 1.218Cpr - 0.168Spr\end{aligned} \tag{3}$$

Furthermore, all expected independent variables affecting domestic ethanol price were used as input nodes in artificial neural networks. Appropriate neural network for this ANN model was obtained with 20 hidden nodes and tan-sigmoid transfer function, as listed in Table 2. The performance of this model from the test set obtained an acceptable MAPE and R^2 value of 6.287 and 0.808, respectively as shown in Table 3.

In addition, four factors affecting the domestic ethanol price from stepwise multiple regression analysis were used as inputs in artificial neural networks. This integrated method provided an appropriate neural network with 20 hidden nodes and tan-sigmoid transfer function, as listed in Table 2. The performance of this MR-ANN model from the test set, as shown in Table 3, produced MAPE and R^2 value of 6.211 and 0.727, respectively.

In order to forecast export price of Thai ethanol as hypothesis in (2), the factors including export volume, domestic ethanol price, ethanol price in Brazil, exchange rate and world crude oil prices, were used as inputs in artificial neural networks. Appropriate neural network of this ANN model was obtained with the training of 5 hidden nodes and tan-sigmoid transfer function, as listed in Table 2. From the test set, the performance model provided a MAPE of 8.187 and R^2 of 0.722, as shown in Table 3.

Some of the independent variables, export volume, domestic ethanol price and world crude oil prices were statistically significant by stepwise multiple regression analysis. They could explain the variation of ethanol export price of 62% as R^2 and acquired the predicted model, MR model as in (4). When considering the standardized coefficients, the most influential factor was export volume, and the following influential factors were domestic ethanol price and world crude oil prices, respectively.

Table 1: Summary of Statistics for Two Forecasting models from Stepwise Multiple Regression Analysis (MR models)

Dependent variable	Independent variables	Coefficient	Standardized Coefficient	t-value	t-sig.	F-value	F-sig.	R^2	SE
Erp	*Constant*	15.463	-	9.157	<0.0005	24.392	<0.0005	0.694	1.860
	CCp	1.366	0.613	6.988	<0.0005				
	$(Mp)^{-1}$	-11.947	-0.401	-4.642	<0.0005				
	Cpr	1.218	0.601	4.377	<0.0005				
	Spr	-0.168	-0.433	-3.190	0.003				
$UEep$	*Constant*	-26.911	-	-2.442	0.018	30.519	<0.0005	0.620	4.464
	$\ln(UEeq)$	-3.000	-0.513	-5.302	<0.0005				
	$\ln(Erp)$	19.760	0.455	5.520	<0.0005				
	$COilp$	-0.078	-0.232	-2.396	0.020				

$$U\hat{E}ep = -26.911 - 3.000\ln(UEeq) + 19.760\ln(Erp) - 0.078COilp \quad (4)$$

Additional, independent variables selected from the stepwise multiple regression analysis were used as input nodes in artificial neural networks. This MR-ANN model was appropriately trained with 15 hidden nodes, and tan-sigmoid transfer function, as listed in Table 2. From the test set, the performance of the model obtained a MAPE of 8.468 and R^2 of 0.812, as shown in Table 3.

Table 2: Artificial Neural Networks Structures of ANN and MR-ANN Models

Output	Model	Number of input nodes	Number of hidden layers	Number of hidden nodes	Activation function
Erp	ANN	5	1	20	tan-sigmoid
	MR-ANN	4	1	20	tan-sigmoid
$UEep$	ANN	5	1	5	tan-sigmoid
	MR-ANN	3	1	15	tan-sigmoid

Table 3: Performance of ANN and MR-ANN Models

Output	Model	Training set		Validation set		Test set	
		MAPE	R^2	MAPE	R^2	MAPE	R^2
Erp	ANN	5.266	0.802	11.994	0.492	6.287	0.808
	MR-ANN	0.469	0.993	11.661	0.262	6.211	0.727
$UEep$	ANN	4.917	0.942	9.968	0.650	8.187	0.722
	MR-ANN	4.716	0.946	15.650	0.810	8.468	0.812

5. Discussion

For domestic price, factors affecting ethanol reference price were cassava production, sugarcane production, cassava chip price and molasses price. It had been found that the feedstocks of ethanol production influence ethanol price. Major operating cost of ethanol plant in Thailand is a feedstocks cost, which account for 60-70%. Feedstocks price varies with crop production, sugarcane and cassava production. Moreover, feedstocks of ethanol production, molasses and cassava, are also raw material used in others industries such as animal feed, monosodium glutamate, and liquor industries. In addition, molasses and cassava chips have been exported in massive volume continuously. This means that quantity and price of feedstocks supplied to ethanol industry is uncontrollable because they depend on the mechanism of domestic and export market.

For export price, factors affecting ethanol export price were export volume, ethanol reference price, and world crude oil price. Export volume and FOB value influence the export price. FOB value is a direct proportion, meanwhile export volume is an inverse proportion to export price. Moreover, quantity and price of feedstocks cause variation to ethanol reference price, consequently, influence export price. In addition, crude oil price influence ethanol export price due to a complex relationship between gasohol and gasoline. If crude oil price is high, it increases gasoline price. Conseqently, the domestic consumption of gasohol is increased and the export of ethanol is decreased. Therefore, it causes variation to export price.

Factors affecting both domestic and export price of Thai ethanol were derived from stepwise multiple regression analysis, which provided forecasting equations. However, performance of MR models was lower than ANN models. In addition, there are limitations of multiple regression since it has assumptions of analysis, especially autocorrelation and multicollinearity problems [11]. Moreover, regression analysis model also needs to be carefully set in the case of nonlinearity of relationship between independent variables and dependent variable. In contrast, artificial neural networks are able to solve nonlinear problems straightforwardly, and also to accomplish problem with data containing a high degree of parallelism, faults and noise. Nonetheless, artificial neural networks do not show a prediction equation and lack of testing for statistical significance [9].

Furthermore, when affected factors from stepwise multiple regression analysis were used as inputs in artificial neural networks, the MR-ANN models were more precise. This can be clearly seen from the comparison of MAPE and R^2 for all data. MR-ANN models were superior because their lowest MAPE and highest R^2, particularly the prediction model for domestic ethanol price, as shown in Table 4 and more clearly seen in Figs 2-3.

Table 4: Comparison of Forecasting Accuracy between ANN, MR and MR-ANN Models

Dependent variables	MAPE			R^2		
	ANN	MR	MR-ANN	ANN	MR	MR-ANN
Erp	6.503	6.539	3.537	0.708	0.694	0.799
UEep	6.492	13.753	7.294	0.868	0.620	0.890

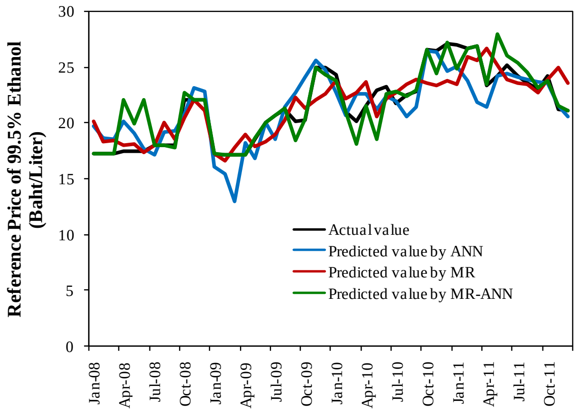


Figure 2: Actual and Predicted Value of Reference Price of 99.5% Ethanol in Thailand

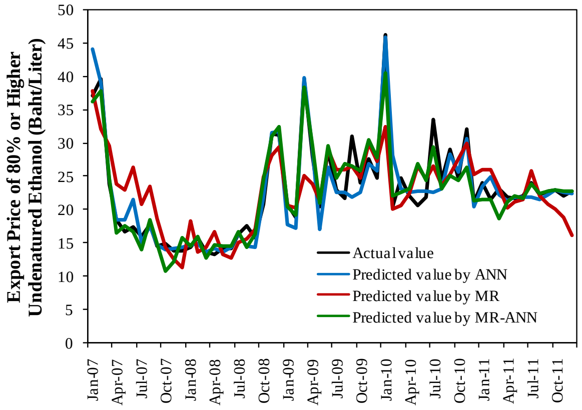


Figure 3: Actual and Predicted Value of Export Price of 80% or Higher Undenatured Ethanol in Thailand

In addition, we have also re-examined the MR-ANN models to confirm its precision by using the actual value of the year 2012 as inputs. Then, the inputs were used to forecast both domestic and export price of ethanol. The results showed that predicted value closed to the actual value with MAPE of 10.06% and 5.85%, respectively. It indicated that MR-ANN models which is an integration of multiple regression and artificial neural networks, were appropriate models to forecast Thai ethanol prices for both domestic and export.

6. Conclusion

In order to construct a forecasting model of domestic and export prices for Thai ethanol, a number of factors were considered in this study. Three types of analytical methods were used, multiple regression, artificial neural networks and an integration of multiple regression and artificial neural networks. A set of data related to the considered factors during the year 2007-2011 were collected and used as inputs in the three methods. It has been found from the formulation that the integrated models, MR-ANN, outperformed the other two in term of MAPE and R^2. For domestic price forecasting, influential factors were cassava chips price, molasses price, cassava production and sugarcane production. Meanwhile, export volume, ethanol reference price and world crude oil prices affected ethanol export price.

In addition, the proposed models were applied and verified the predicted prices and actual prices of the year 2012. The results showed that the proposed models were accurate with satisfied MAPE. Therefore, the proposed models can be useful in decision making for ethanol manufacturers, fuel traders, ethanol exporters and overseas customers.

Acknowledgements

The authors would like to thank Khon Kaen University for the financial support of this study under the National Research University of Thailand program, Contract Numbers NRU543045 and BiF-2553-Ph.D-07.

References

[1] Department of Energy Business. Fuel consumption in Thailand. [4 July 2013]. Available from: http://www.doeb.go.th/info/info_procure.php.

[2] Petroleum and Petrochemical Policy Bureau. Ethanol reference price. [25 May 2013]. Available from: http://www.eppo.go.th/petro/price/index.html.

[3] Department of Alternative Energy Development and Efficiency. Ethanol pricing formulation. [25 September 2012]. Available from: www.dede.go.th.

[4] Bank of Thailand, Ethanol: Opportunities and Challenges of Thai Energy Policy. 2012, Regional Economy Division, Bank of Thailand, Northeastern Region Office: Khon Kaen.

[5] Energy Policy and Planning Office. Ethanol reference price of Thailand. [26 May 2013]. Available from: http://www.eppo.go.th.

[6] The Customs Department, Export Statistics. 2013, The Customs Department, Ministry of Finance, Thailand.

[7] Spyros M, Stevec CW. The Handbook of Forecasting. New York: John Wiley & Son; 1982.

[8] Spyros M, Stevec CW, Victor EM. Forecasting methods and applications. Singapore: John Wiley & Son; 1983.

[9] Mukta P, Usha AK. Neural networks and statistical techniques: A review of applications. Expert Systems with Applications. 2009; 36(1): 2-17.

[10] Jay LD. Probability and statistics for engineering and the science. California: Wadsworth; 1995.

[11] Norman RD, Harry S. Applied Regression Analysis. 3rd ed. New York: John Wiley and Sons; 1998.

[12] Barbara GT, Linda. SF. Using multivariate statistics. 4th ed. New York: Harper & Row; 2000.

[13] Intan MMG, Sabri A. Stepwise Multiple Regression Method to Forecast Fish Landing. Procedia - Social and Behavioral Sciences. 2010; 8(0): 549-554.

[14] Jeremy M, Mark S. Applying Regression & Correlation. London: SAGE Publications; 2001.

[15] Basheer, I.A. and M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods. 2000; 43(1): 3-31.

[16] Swingler, K., Applying Neural Network: A Practical Guide. New York: Academic Press; 1996.

[17] Carl GL. Advances in Feedforward Neural Networks: Demystifying Knowledge Acquiring Black Boxes. IEEE Transactions on knowledge and data engineering. 1996; 8(2).

[18] Bo KW, Vincent SL, Jolie L. A bibliography of neural network business applications research: 1994–1998. Computers & Operations Research. 2000; 27(11–12): 1045-1076.

[19] Jeffrey ES, Venkatachalam, AR. Venkatachalam, A neural network approach to forecasting model selection. Information & Management. 1995; 29(6): 297-303.

[20] Fatemeh RA, Yousef AG. Artificial Neural Network and stepwise multiple range regression methods for prediction of tractor fuel consumption. Measurement. 2011; 44(10): 2104-2111.

[21] Fu, L., Neural Networks in Computer Intelligence. New York: McGraw-Hill; 1995.

[22] Zipan, J. and J. Gasteiger, Neural Networks for Chemists: An Introduction. New York: VCH; 1993.

[23] Azadeh, A. and Behshtipour. The effect of neural network parameters on the performance of neural network forecasting. in 6th IEEE International Conference on Industrial Informatics, 2008. INDIN 2008.

[24] Chia KS, Rahim H.A., Rahim R.A.. A comparison of Principal Component Regression and Artificial Neural Network in fruits quality prediction. in Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on. 2011.

[25] Usha AK. U.A., Comparison of neural networks and regression analysis: A new insight. Expert Systems with Applications. 2005; 29(2): 424-430.

[26] Zong Woo G, William ER. Energy demand estimation of South Korea using artificial neural network. Energy Policy. 2009; 37(10): 4049-4054.

[27] Yong Soo K. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. Expert Systems with Applications. 2008; 34(2): 1227-1234.

A note on the combined use of statistical methods to study trait interrelationships in plant breeding

M. İlhan Çağırgan[1*], M. Onur Özbaş[1,2], Hasan Topuz[1,2], Alper Adak[1] and Mehmet Tekin[1]

[1]*Antalya Mutation Project, Dept. of Field Crops, Faculty of Agriculture, Akdeniz University, TR07059 Antalya, Turkey.*

[2]*Present address: Enza Zaden Tarım Ar-Ge, Hisaraltı Mevkii, P.O. Box 87, Antalya, Turkey*

Abstract

Plant breeding is the science and art of developing better crops in yield, quality, and adaptation to environmental stresses and pests to the benefit of the mankind. To develop better crops, a plant breeder makes his or her selections from a population of plants using the data recorded on the traits which may be correlated negatively or positively. To determine any relationship between traits, correlation analysis has been traditionally applied to the data in breeding programs. However, a correlation coefficient between two traits does not show any cause-effect relationship. Alternately, in the studies in which a number of traits are recorded and their contribution to yield are examined, a multiple regression analysis is used. In this case, yield is logically chosen as the dependent variable while other yield components or plant traits are the independent variables. One of the problems related with this approach is the problem of multicollinearity. Path analysis is used to reveal the causes in the associations between traits and to partition the correlation coefficients into direct and indirect effects through a response variable. Path analysis is basically the form of a multiple regression and it is useful when cause-and effect relationship is efficiently set with a reasonable number of indicators on the response characters. When one deals with high number of traits with no priori information about their association and avoiding multicollinearity problem, factor analysis is useful to understand structural interrelationships and can be used to select a set of fewer traits based on the usefulness and easiness of the measurements for that characters. While factor analysis deals with grouping observed characters of genotypes, cluster analysis groups genetically similar genotypes based on a phenotypic or molecular basis (e.g., using DNA markers). With the advent of computers, multivariate statistical methods are used widely in plant breeding particularly by combining different methods. However, improper combination of the methods produces a mess of information. In this communication, we exemplify an efficient execution of a combined use of statistical procedures in the study of interrelationships among agronomic traits recorded in our barley breeding program.

Keywords: Correlation, path analysis, stepwise regression procedure, factor analysis

*Corresponding Author
E-mail Address: cagirgan@akdeniz.edu.tr

Spatiotemporal pattern of extreme rainfall events in Indochina Peninsula

Muhammad Yazid[1], Usa Humphries[2*] and Triyono Sudarmadji[3]
[1]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
zy_d26@yahoo.com
[2]*Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,*
usa.wan@kmutt.ac.th
[3]*Soil and Water Conservation Laboratory, Mulawarman University, Samarinda, 75119, Indonesia,*
anni_tri@yahoo.com

Abstract

Rainfall is the most influential weather parameter affecting human lives, besides being a natural resource that is needed by humans it can also be a source of disasters if it gets to its extreme. This extreme rainfall is a big problem for the society, thus the analysis of extreme rainfall is needed to design mitigation strategies. This study describes the extreme rainfall phenomena based on statistics that focused on the existence of trends in Consecutive Dry Days (CDD) and Consecutive Wet Days (CWD) and characteristic of its distribution. The trends were obtained from high-quality grid precipitation data compiled by Asian Precipitation-Highly-Resolved Observational Data Integration towards Evaluation of Water Resources (APHRODITE) over Indochina Peninsula (4°-25°N and 90°-112°E). This analysis were selected from the list of climate change indices recommended by World Meteorological Organization-Commission for Climatology (WMO-CCI) and the research program on climate variability and predictability (CLIVAR). Linear trends were calculated by least squares fitting and significant or non-significant trends were identified using Mann-Kendall test. The result revealed contrasting trends of each index in the eastern and western Indochina Peninsula. In the eastern Indochina Peninsula mostly indicated positive trends in CWD and negative trends in CDD with some grids showing significant trends, contrary to western Indochina Peninsula. The percentages of positive and negative significant trends of CDD are 10.88% and 2.41% respectively, while for CWD index are 6.85% and 14.51% respectively from a total of 248 grids.

Keywords: Indochina Peninsula, Extreme Rainfall, CDD, CWD, Trends

*Corresponding Author
E-mail Address: usa.wan@kmutt.ac.th

1. Introduction

Weather is the most important geographical parameter affecting human lives, while rainfall is the most influential weather parameter. Rainfall is a natural resource that is needed by humans, but it can be a source of disasters if it gets to its extreme. Extreme rainfall events often damage the environment, because they are often followed by floods, lightening, and strong winds [1]. As a result of extreme weather and climate, some areas are vulnerable to disaster such as floods, landslides, and severe drought.

The impacts of climate change and extreme weather events are the most serious problems for society [2]. Extreme events will be more frequent, more widespread and will increase in intensity in the 21st century [3]. Various problems that arise due to extreme weather and climate change range from disease outbreaks, health problems, fishermen dare not go to sea due to high waves, farmers can not do harvest and also social vulnerability.

Lately, extreme rainfall events are getting a serious attention in the community, because of their great impact on nature and a great loss for the community. In rural areas, these extreme rainfall events often cause damages to agricultural crops and livestock, while in urban areas they can cause floods because of inability of drainage systems to accommodate high rainfall [4].

Indochina Peninsula is a region in Southeast Asia, including Thailand, Cambodia, Laos, Vietnam, Myanmar, and some parts of Malaysia. Some disaster events that caused extreme rainfall in the region include in Thailand; long floods in the year 2011 that started in June and ended in mid-January 2012. The flooding affected the provinces of northern, northeastern and central Thailand along the Mekong and Chao Phraya river basins, as well as parts of the capital city of Bangkok [5].

In Myanmar, rainfall caused flooding in several places at the end of July 2013. The flash initially displaced over 38.300 people, leaving six dead and one person missing, and damaged residential buildings, roads and bridges [6]. On the other hand, extreme rainfall also occurs annually in the Upper Mekong countries of China, Laos, Myanmar and Thailand, where these events greatly impact on the environment and society [7]. While in Malaysia floods triggered by extreme rainfall in northeastern, central and southern parts of Malaysia killed 33, and affected 158,000 in December 2007[5].

The statement of World Meteorological Organization (WMO) based on the analysis of extreme events [2], says that the development of the economy depends on our ability to deal with some risks that are caused by extreme events. The knowledge of extreme events is necessary, because our lives depend on food, water, energy, and transportation which are very sensitive to extreme events. It is a priority to use advancement in scientific knowledge of extreme rainfall to develop solutions to the water-related challenges faced by the society [8]. It is therefore important to study the information and knowledge of extreme weather and climate change. By knowing the pattern of extreme weather and climate change, the impacts of extreme events can be anticipated as early as possible.

2. Research Methodology

The steps of this study include determining the data, QC and homogeneity data, data processing, and data analysis as follows:

2.1 APHRODITE Datasets

In this study, the indices of extreme rainfall events were generated using series data. Period of data is crucial in determining extreme rainfall events [9]. Long period of time is needed to estimate the frequency and intensity of extreme rainfall events [2] because it would eliminate the effect of bias [10]. On the other hand, the used data should have a high resolution at least daily data [2].

The data in this study was obtained from Asian Precipitation-Highly-Resolved Observational Data Integration towards Evaluation of Water Resources (APHRODITE) datasets for 48 year period from 1960 to 2007, this grid data were extracted from the station data that had been collected by the APHRODITE team. The station data from various sources such as National Hydro-Meteorological Service (NHMS) in each country such as Thai Meteorological Agency and the Royal Irrigation Department in Thailand, Department of Hydrology and Meteorology in Myanmar, Ministry of Water Resources and Meteorology in Cambodia, and National Hydro-Meteorological Service in Vietnam. This data was also collected from Precompiled datasets by other projects such as the Global Energy and Water Cycle Experiment (GEWEX) Asian Monsoon Experiment and-Tropics (GAME-T) and from Global Telecommunication Systems (GTS) report based global datasets such as the Global Surface Summary of the Day (GSOD) [11].

2.2 Quality Control and Homogeneity Test of APHRODITE Datasets

This study uses APHRODITE dataset, which the selection of this data is based on APHRODITE daily rainfall data in the grid form that is extracted from the station data. The data is ready to be processed because it has been through a long quality control (QC) process. Automated QC system developed in this APHRODITE datasets basically designed to detect errors in daily rainfall station data such as recording errors, clerical errors and so on.

There were 14 steps that had been developed to process APHRODITE QC data such as; QC for errors in station metadata, errors identified in single station records such as; erroneous values inherent to particular data, values exceeding national/regional records, contamination with different weather elements, repetition of constant values, duplication of monthly or sub-monthly records, outliers, homogeneity test, and also errors identified in multiple station records such as; spatiotemporally isolated values, errors in units of measurement, and ambiguity in recorded date [12].

2.3 Methods for Identification of Extreme Rainfall

Adapted from WMO [2], the analysis of extreme rainfall events were determined by several indices that are widely used in determining extreme weather events. The indices were selected from the list of climate change indices recommended by World Meteorological Organization-Commission for Climatology (WMO-CCI) and the research program on Climate Variability and Predictability (CLIVAR). Table 1 explains the definition of Consecutive Dry Day (CDD) and Consecutive Wet Day (CWD) that were used in this study.

Table 1: Definition of CDD and CWD

Indices	Indices Name	Indices Calculation	Definition	Unit
CDD	Consecutive Dry days	$RR_{ij} < 1_{mm}$	Maximum number of consecutive days with rainfall (RR) <1 mm	Day
CWD	Consecutive Wet days	$RR_{ij} \geq 1_{mm}$	Maximum number of consecutive days with RR >1 mm	Day

RR=Rainfall on consecutive days

In this study, the rainfall data was processed using the RClimDex software package to calculate indices and trends, a R-language software package that was developed by CCI-CLIVAR Expert Team for Climate Change Detection, Monitoring and Indices (ETCCDMI) with the main focus for detection and monitoring the extreme climate events using daily data. This RClimDEX software packages are available at http://cccma.seos.uvic.ca.

2.4 Methods for Temporal Trend Analysis

Man-Kendall test is a statistical test widely uutilized for the analysis of trend in climatology [13-16]. The benefit of Man-Kendall test; it is non parametric statistics and does not require a normally distributed data, and has a low sensitivity to empty data due to non homogeneity data. Null hypothesis (H_0) states that data is no trend (the data is independent and randomly ordered), and this is tested against the alternative hypothesis (H_1) which assumes that there are trends. The Kendall's statistics (S) is computed as follows:

2.4.1 Man-Kendall Statistical Calculation

The computational procedure of Mann-Kendall test considers the time series of n data points and x_j and x_k as two subsets of data which j =1,2,3,... n and k = $j+1$, $j+2$, $j+3 \ldots, n$. The data values are evaluated as an ordered time series. Each data value is compared with all subsequent data value. The initial value is assumed with zero (0) or no trend, if the value of the data from the later time period is greater than the earlier time period, therefore S is added by 1. On the other hand, if the value of the data in later time period is less than the earlier time period, therefore S is reduced by 1. The end result of the addition and subtraction is final value of S .

The formula to calculate the S is:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign\,(x_j - x_k)\,. \qquad (3.1)$$

Here,

$$sign(x_j - x_k) = \begin{cases} 1, & \text{if } x_j - x_k > 0 \\ 0, & \text{if } x_j - x_k = 0 \\ -1, & \text{if } x_j - x_k < 0 \end{cases} , \qquad (3.2)$$

where x_j and x_k are the annual values in years j and k , $j > k$, respectively. If $n < 10$, the value of $|\,S\,|$ is compared directly to the theoretical distribution of S derived by Mann and Kendall. The two tailed test is used. At certain probability (H_0) is rejected in due to (H_1) if the absolute value of S equals or exceeds a specified value of $S_{\alpha/2}$, where $S_{\alpha/2}$ is the smallest S which has the probability less than $\alpha / 2$ to appear in case of no trend. A positive (negative) value of S indicates an upward (downward) trend. For $n > 10$, the statistics S is approximately normally distributed with mean and variance as follows:

The variance of S is given by

$$\sigma^2 = \frac{n(n-1)(2n+5) - \sum t_i (i)(i-1)(2i+5)}{18} \qquad (3.3)$$

which t_i denotes the number of ties to extent i . The summation term in the numerator is used only if the data series contains tied values.

2.4.2 The Standard Test Statistics Z_c Calculation

The standard test statistics Z_c is calculated as follows:

$$Z_C = \begin{cases} \dfrac{s-1}{\sqrt{\sigma}}, & if \;\; S > 0 \\ 0 \;, & if \;\; S = 0 \\ \dfrac{s-1}{\sqrt{\sigma}}, & if \;\; S < 0 \end{cases} . \qquad (3.4)$$

Positive values of Z_c indicates an upward trend while negative Z_c shows downward trend. When testing either upward or downward trends at a significance level α , the null hypothesis was rejected for absolute value of Z_c greater than $Z_{\alpha/2}$. In this study, significance level α of 0.05 was applied.

2.4.3 To Compute the Probability Associated with This Normalized Test Statistics Expressed in P-Value.

The probability density function for a normal distribution is given by the following equation:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} , \qquad (3.5)$$

$$p - value = 1 - f(z) ,$$

$$\text{when} \quad \begin{matrix} p - value < \alpha = significant \\ p - value > \alpha = non\,significant \end{matrix} \qquad (3.6)$$

2.4.4 Decide level of Significant (α=5% *typically*)

Conclude the trend using this criterion, the trend is said to be downward if Z is negative and is said to be upward if the Z is positive. The trend is statistically significant if *p-value* is less then α and non-significant if *p-value* is greater than α.

2.5 Methods for Spatial Analysis

Since rainfall is not distributed proportionally, therefore, the spatial patterns analysis is needed because topography affects the rainfall in some regions. In this study, mean climatology of annual indices is executed using interpolation methods of Radial Basis Function. Spatial analysis of temporal trends is done with thematic maps using dots (.) and plus (+) signs, where the plus (+) sign indicates a positive trend and a dot (.) sign indicates a negative trend. The color indicates the significant of changes, the blue color indicates significant trends, and the red color indicates non-significant trends. And the gradient of plus (+) or dot (.) sign indicates how significant the trend is.

3. Research Result and Discussion

The results of this study include mean climatology of annual indices, temporal trends analysis, and spatial pattern analysis of CDD and CWD index, as follows:

3.1 Mean Climatology of Annual Indices

The annual indices in this study were calculated based on annual block that means only one value per year. Thus, in this study with period from 1960 to 2007 there would be 48 values for each grid. To analyze the spatial patterns of each extreme rainfall indices, each index was averaged to obtain the mean value of each grid of climatology.

The mean climatology of consecutive dry day or consecutive days without rainfall (CDD) in Indochina Peninsula ranged from 8 to 110 days each year (Fig. 1). The lowest number of consecutive days without rainfall is on grid Lon 101.25 and Lat 4.25 with total 8 days, while the greatest number of consecutive days without rainfall is on grid Lon 95.25 and Lat 18.25 with total 110 days. The area with the greatest number of consecutive days without rainfall is mostly situated on the West coast of Indochina Peninsula; in most of Myanmar, central and northern Thailand, most of Cambodia, some part of Laos, and southern Vietnam. While the East coast of the Indochina peninsula, extending from southern China to Vietnam, southern Thailand, eastern and northern Laos and a part of Malaysia include small part of Indonesia experienced a lower number of consecutive days without rainfall.

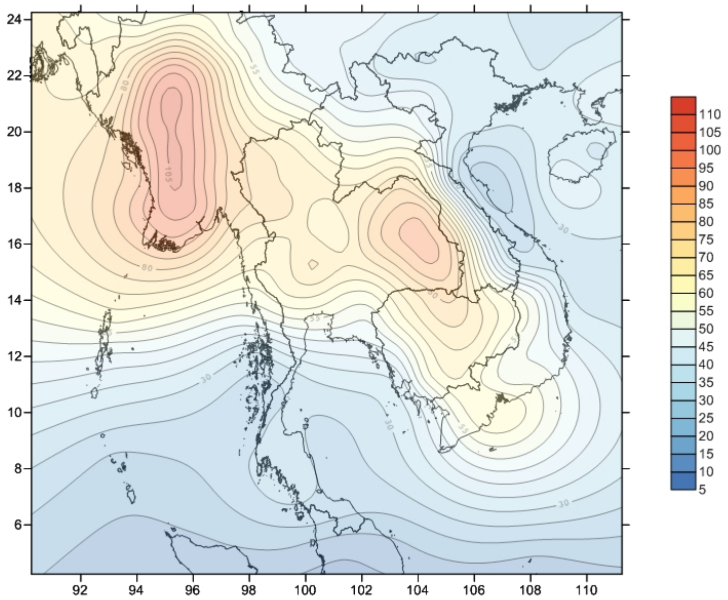


Figure 1: Mean climatology of CDD index

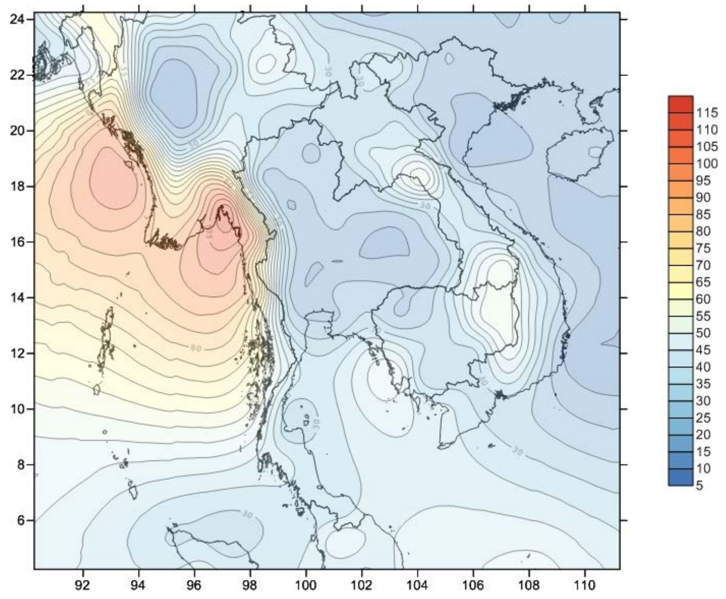


Figure 2: Mean climatology of CWD index

Figure 2 shows the spatial pattern of mean number of consecutive wet days or consecutive rainy days (CWD) in Indochina Peninsula that ranged from 11 to 111 days each year. The lowest number of consecutive rainy days is on grid Lon 106.25 and Lat 20.35 with a total 11 days, while the greatest number of consecutive rainy days is on grid Lon 97.25 and Lat 17.35 with total 111 days. This figure shows that most of the Indochina Peninsula covering the whole territory of Vietnam, Thailand, Laos, and Cambodia has a consecutive rainy days on average from 11 to 40 days, and a bit different in the coastal areas of Myanmar which has a consecutive rainy day on average more than 55 days.

3.2 Temporal Trends of Extreme Rainfall

Trends in extreme rainfall indices are calculated using statistical software RClimDex R-based language program. The positive (negative) values indicate upward (downward) trends. This study uses a significance level (α) of 5%. The slope represents the magnitude of the variable changes each year.

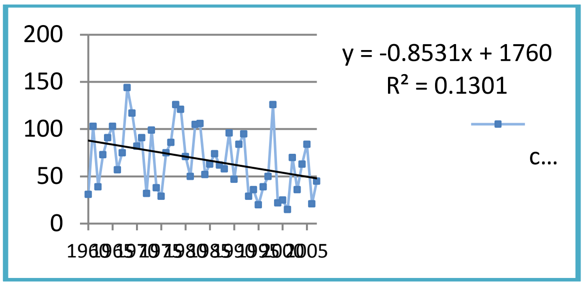


Figure 3: Significant negative trend of CDD index on grid Lon 107.25 and Lat 10.25

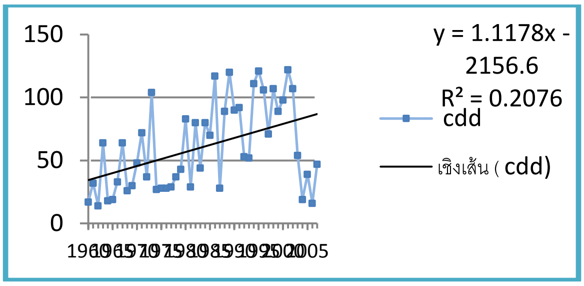


Figure 4: Significant positive trend of CDD index on grid Lon 105.25 and Lat 17.25

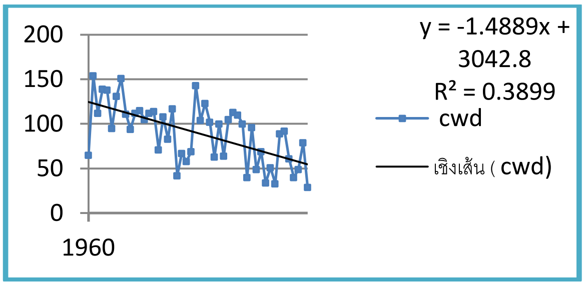


Figure 5: Significant negative trend of CWD index on grid Lon 95.25 and Lat 16.25

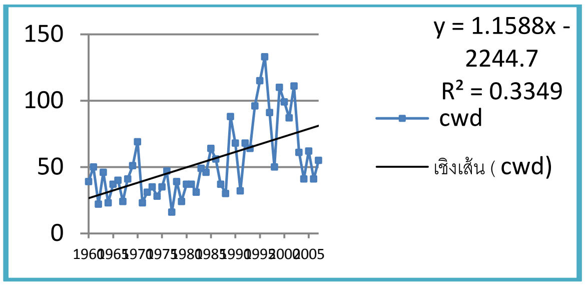


Figure 6: Significant positive trend of CWD on grid Lon 104.25 and Lat 18.25

Consecutive days without rainfall (CDD) index was the first to be processed, the results show that non-significant trend changes dominate the study area. From a total of 248 grids, there are 215 grids showing non-significant trends which 65.72% shows positive trends and 20.96% shows negative trends of the overall number of grids. While 32 of other grids demonstrate significant trends which 10.88% shows positive trends and 2.41% shows negative trends of the overall number of grids. Figures 3 and 4 show the most significant of positive and negative trends of CDD index. The most significant of positive trends is on the grid Lon 105.25 and Lat 17.25 with the slope 1.117, while the most significant of negative trends is on grid Lon 107.25 and Lat 10.25 with the slope -0.853.

The next calculated index is consecutive rainy days (CWD) index, the results indicate 194 grids showing non-significant trends which 36.29% shows positive trends and 41.93% shows negative trends of the total grids. While 53 of other grids demonstrate significant trends which 6.85% shows positive trends and 14.51% shows negative trends of the total grids. Figures 6 and 7 indicate the most significant of negative and positive trends of CWD index. The most significant of positive trend is on grid Lon 104.25 and Lat 18.25 with the slope 1.1588, while the most significant of negative trend found on grid Lon 95.25 and Lat 16.25 with a slope -1.488. Here Table 2 shows the percentages of significant and non-significant trends.

Table 2: Percentage of Significant and non-significant trends

Index	Positive Significan t Trend (%)	Positive Non Significan t Trend (%)	Negative Significan t Trend (%)	Negative Non Significan t Trend (%)
CDD	10.88	65.72	2.41	20.96
CWD	6.85	36.29	14.51	41.93

3.3 Spatial Pattern of Detected Trends

The purpose of the spatial patterns analysis of extreme precipitation trends is to identify which areas are experiencing a positive (negative) significant (non-significant) trend, whether the resulting pattern is a random pattern, spread, or to form a particular pattern. This study used a point pattern analysis by plotting each grid into Indochina Peninsula Map.

The spatial pattern of consecutive days without rainfall (CDD) index and consecutive rainy days (CWD) index are shown in Figure 7 and Figure 8. In the CDD index, the pattern shows that most of the Eastern coast of the Indochina Peninsula covering the East coast of Vietnam and the East coast of southern Thailand experienced negative trends with some grids show significant trends. Otherwise, a unique pattern occurred in the West coast of the Indochina Peninsula, most grids experienced positive trends with some grids show a significant trend. While for CWD index, the positive trends occurred in the eastern Indochina peninsula along the East coast of Vietnam and the East coast of southern Thailand with some grids show significant trends. While the West coast of the Indochina Peninsula shows negative trends with some grids show significant trends. This became a unique pattern because it shows the opposite pattern between the eastern and western Indochina Peninsula.

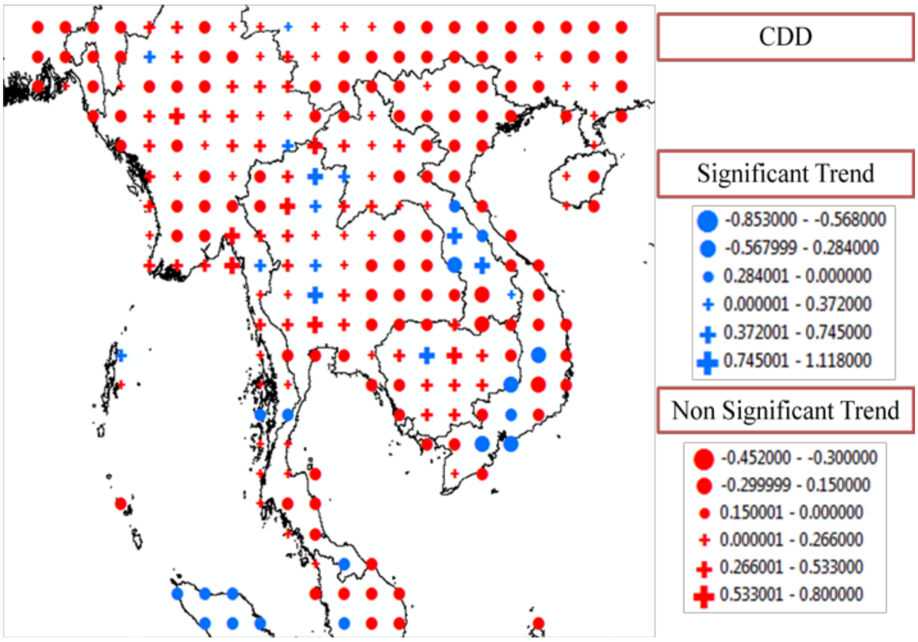


Figure 7: Spatial analysis of CDD index over Indochina

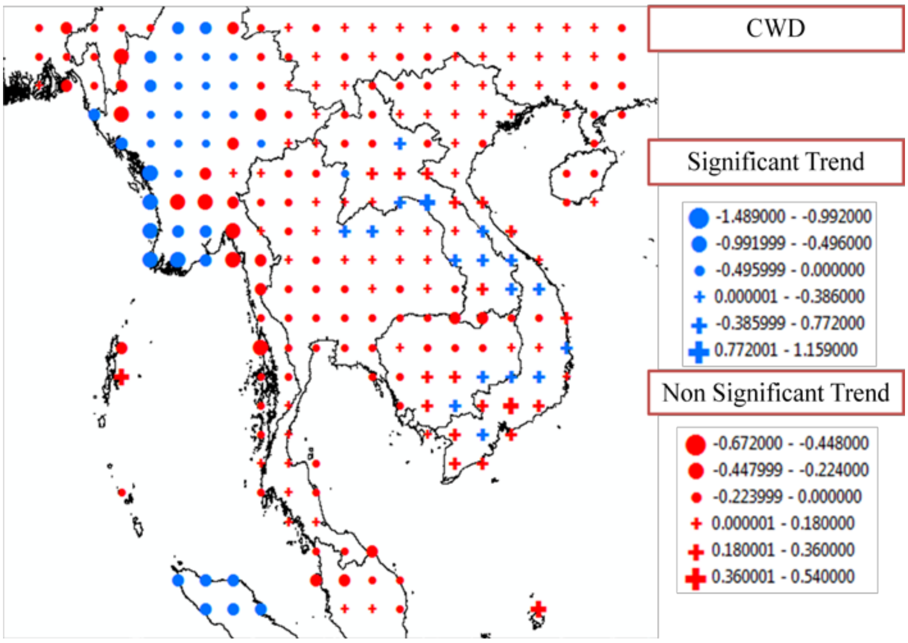


Figure 8: Spatial analysis of CWD index over Indochina Peninsula

4. Conclusion

Some remarkable findings can be concluded from this study. Data analysis and homogeneity data become very important because they can affect the resulting trends and will cause a bias interpretation, hence the selection of high quality data should be considered very important before doing research on trend changes of extreme rainfall indices. This study used APHRODITE datasets with high quality data that has been through long QC process.

Overall, the pattern of the mean climatology of each index can be well described. These two indices have a pattern and specificity to the geographic location of this area. The highest mean climatology was noticed in two indices situates in the coast of North Myanmar, North Vietnam, and southern Thailand (border with Malaysia). While other regions have mean climatology from medium to lower that dominated in the central Indochina Peninsula.

This study also found that the Indochina peninsula is dominated by non-significant trends that spread evenly across it with the percentage of positive and negative significant trends respectively, 10.88% and 2.41% for CDD index, 6.85% and 14.51% for CWD index.

For CDD and CWD indices show a unique pattern that contrasted between West and East Indochina Peninsula. In the CDD index, the pattern shows that most of the eastern coast of Indochina Peninsula covering the East coast of Vietnam and the East coast of southern Thailand experienced negative trends with some grids show significant trends. Otherwise, for CWD index, which the negative trends occurred in the West coast of the Indochina Peninsula with some grids show significant trends.

Acknowledgements

This master project detailed in this research was supervised by Dr. Atsamon Limsakul (Ministry of Natural Resources Thailand), his advices and suggestions were very much acknowledged. Special thanks are dedicated to Mathematics Department for travel expenses and accommodation during this conference. The authors would also like to thank to East Kalimantan Government of Indonesia for providing the scholarship through Kaltim Cemerlang Scholarship during studying at KMUTT.

References

[1] Jones C, Waliser DE, Lau KM, Stern W. Global Occurrences of Extreme Precipitation and the Madden–Julian Oscillation: Observations and Predictability. American Meteorological Society. 2004; 17: 4575-4589.

[2] Tank, AMGK, Zwiers FW, Zhang X. Guidelines on Analysis of Extremes in Changing Climate in Support of Informed Decisions for Adaptation. Climate Data Monitoring, WCDMP-No. 72. Switzerland: WMO; 2009.

[3] Solomon S, Qin D, Manning M, Marquis M, Averyt K., Tignor MMB, Jr HLRM, Chen Z. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), Climate Change 2007: The Physical Science Basis. Cambridge, United Kingdom and New York: Cambridge University Press; 2007

[4] Carvalho LMV, Jones C, Liebmann B. Extreme Precipitation Events in Southeastern South America and Large-Scale Convective Patterns in the South Atlantic Convergence Zone. American Meteorological Society. 2002: 15; 2377-2394.

[5] The official web of an Asian Disaster Reduction Centre (ADRC) [Internet].2008[Cited 2013 November 6]
Available from:
http://www.adrc.asia/nationinformation.php?NationCode=764&Lang=en&Mode=country

[6] The official web of United Nations Office for the Coordination of Humanitarian Affairs, OCHA [Internet][Cited 2013 November 6]
Available from: www.unocha.org.

[7] The official web of Disaster Reduction Program for Cambodia, Lao PDR, and Vietnam (DRV-CLV) [Internet].2003[Cited 2013 November 6]
Available from:
http://www.adpc.net/drp/clv/cambodia/index.php

[8] Testik FY, Gebrenichael M. Rainfall: State of the Science. American Geophysical Union as a Part of the Geophysical Monograph Series. 2013;191.

[9] Frei C, Schar C. Detection Probability of Trends in Rare Event: Theory and Application to Heavy Precipitation in the Alpine Region. Journal of Climate. 2000; 14: 1568-1584.

[10] Manton MJ, Della-Marta PM, Haylock MR, Hennessy K.J, Nicholls N, Chambers LE, Collins DA, Daw G, Finet A, Gunawan D, Inape K, Isobe H, Kestin TS, Lefale P, Leyu CH, Lwin T, Maitrepierre L, Ouprasitwong N, Page CM, Pahalad C, Plummer N, Salinger MJ, Suppiah R, Tran VL, Trewin B, Tibig I, Yee D. Trends in Extreme Daily Rainfall and Temperature in Southeast Asia and The South Pacific: 1961–1998. International Journal of Climatology. 2001; 21: 269-284.

[11] Yatagai A, Kamighuci K, Arakawa O, Hamada A, Yasutomi N, Kitoh A., 15. APHRODITE Constructing a Long-Term Daily Gridded Precipitation Dataset for Asia Based on a Dense Network of Rain Gauges. American Meteorological Society. 2012: 1401-1415.

[12] Hamada A, Arakawa O, Yatagai A. An Automated Quality Control Method for Daily Rain-Gauge Data. Global Environmental Research. 2011; 15: 183-192.

[13] Zhang X, Hogg WD, Mekis E. Spatial and Temporal Characteristics of Heavy Precipitation

Events over Canada. Journal of Climate. 2000; 14: 1923-1936.

[14] Santos CAC, Brito JIB, Junior CHFS, Dantas LG. Trends in Precipitation Extremes over the Northern Part of Brazil from ERA40 Dataset. Revista Brasileira de Geografia Fisica**.** 2012; 4: 836-851.

[15] Fu G, Viney NR, Charles SP, Liu J. Long-Term Temporal Variation of Extreme Rainfall Events in Australia: 1910–2006. Journal of Hydrometeorology. 2010; 11: 950-965.

[16] Atsamon L, Sangchan L, Thavivongse S. Assessment of Extreme Weather Evetns along the Coastal Areas of Thailand. Proceeding of 21th Conference on Climate Variability and Change**.** Washington. 2009.

Comparing the type I error and power of one population variance test between the standard and the asymptotic chi-square distribution of -2 log likelihood ratio test statistic

Jularat Chumnaul

Department of Mathematics and Statistics, Prince of Songkla University, Hatyai, Songkhla 90112, Thailand, jularat.c@psu.ac.th

Abstract

This research aimed to propose a test statistic for one-sample variance testing based on an asymptotic chi-square distribution of -2 log likelihood ratio and determine an effectiveness of the proposed test statistic with the standard test statistic. The probability of type I error and powers of the test are also investigated analytically. The comparative simulation was conducted using MINITAB 14.0 program to generate the normal random variables and compute the probability of type I error and power of the test and an experiment repeated 1,000 times for each situation. From the study of the simulation, we find that the probability of type I error of the test statistic based on an asymptotic chi-square distribution of -2 log likelihood ratio and the standard test statistic are not different and close to a significance level. The powers of the proposed test statistic are more powerful than the standard test statistic. In both test statistics, the powers of the test increase with sample sizes. However the powers decrease when the difference between the true variance of population and variance for testing is smaller.

Keywords: Likelihood ratio, asymptotic chi-square distribution, type I error, power of the test

Corresponding Author
E-mail Address: Jularat.c@psu.ac.th

A new practical approach to goodness-of-fit test for logistic regression models

Kanyaphorn Hankla[1*] and Veeranun Pongsapukdee[2]
[1]Statistics, Silpakorn University, Nakorn Pathom, 73000, Thailand, e-mail: fon0243@gmail.com
[2]Statistics, Silpakorn University, Nakorn Pathom, 73000, Thailand

Abstract

Logistic regression model is a wildly model used in many research fields. To describe the association between the response variable and the explanatory variables, at least one continuous variable, classical Pearson and Deviance tests to assess logistic regression model are invalid. While the Hosmer - Lemeshow test can be used in this situations, simple to perform and wildly used but it does not have desirable power. The objective of this research was to propose a new practical approach for goodness-of-fit test that used Principal Component Analysis (PCA) with continuous variables. Principal Component score obtained from PCA and response variable fitted logistic regression model and used the known tests for goodness-of-test. The results of the study were as follows: the proposed test (Using PCA) performs as well as or better than many of the known tests.

Keywords: Logistic regression model, goodness-of-fit, component score, type I error

*Corresponding Author
E-mail Address: fon0243@gmail.com

1. Introduction

Logistic regression model is a branch of the generalized linear models and is widely used in many areas of scientific research. The logit link function and the binary dependent variable of interest make the logistic regression model distinct from the linear regression model. After fitting a model to the observed data, one of the next essential steps is to investigate how well the posed model fits the observed data. A model is said to fit poorly if either the model's residual variation is large, systematic, or does not follow the variability postulated by the model [1]. There are many ways that cause the logistic regression model to fit the data inadequately, the most important of which involves the problem with the linear component [2] such as omission of higher order terms of covariates or important covariates related to the response variables from the model. Influential observations and outliers can also lead to a poor fit.

Goodness-of-fit tests are designed to determine formally the adequacy or inadequacy of the fitted logistic regression model. A poorly fitted model can give biased or invalid conclusions on the statistical inference based on the fitted model. Therefore, we must test the lack-of-fit of a model before we can use it to make statistic inferences.

Pearson's classical chi-square test and Deviance test are well known goodness-of-fit tests of logistic regression model, and they work very well when the covariates are categorical. When one or more covariates are continuous, the disadvantages of Pearson chi-square test and Deviance test provide incorrect p-values[3]. The Hosmer-Lemeshow test can be used in this situations but it does not have desirable power in many cases and providea no furtherinformation on the source of any detectable lack of fit[4].

In this research, proposed a new practical approach for goodness-of-fit test that used Principal Component Analysis (PCA) with continuous variables. Component score obtained from PCA and response variable fitted logistic regression model and used goodness-of-test. Our research focused on the methodologies proposed primarily for the logistic regression model and the generalized linear model. i.e. Hosmer-Lemeshow test, Pearson's chi-square test and Deviance test.

2. Research Methodology

2.1 Generalized Linear Models: GLMs

Generalized linear model (GLMs) is proposed by McCullagh and Nelder [5] the dependent or outcome variable $\mathbf{y}$ is assumed to be generated from a particular probabiliry distribution function from exponential family. Generalized linear models are defined by 3 components:

2.1.1 Random Components: the probability distribution function f for the dependent variable $\mathbf{Y} = (y_1,...,y_N)'$ is from an exponential family of distributions.

2.1.2 Systematic components: a linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, where the matrix $\mathbf{X}$ contains columns of explanatory variables which is called the design matrix, $\boldsymbol{\beta}$ are the unknown parameters and $\boldsymbol{\eta} = (\eta_1,...,\eta_N)'$.

2.1.3 A link function g such that $E(\mathbf{y}\,|\,\mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$ which provides the link between the predictor and the mean of the probability distribution function f of y. Let $E(\mathbf{y}_i\,|\,\mathbf{x_i}) = \boldsymbol{\mu}_i, i = 1,2,...,N$ then $\mathbf{x_i}$ is the ith row of $\mathbf{X}$; the explanatory variables association with the ith response y$_i$ vector of independent variables.

The goal of logistic regression models is used to model the probability of the occurrence of an event depending on the value of covariates x. The model is of the following form:

$$\Pr(y = 1\,|\,\mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{(\mathbf{x}\boldsymbol{\beta})}}{1 + e^{(\mathbf{x}\boldsymbol{\beta})}}$$

In logistic regression, there are several goodness-of-fit tests obtained by comparing the overall difference between the observed and fitted values.

2.2 Hosmer-Lemeshow

Hosmer and Lemeshow proposed a goodness-of-fit, now universally refered to as the Hosmer-Lemeshow

In this method, the subjects are grouped into g groups with each group containing $\frac{n}{10}$ subjects. The number of groups g is about 10 and can be less than 10, due to fewer subjects Ideally, the first group contains $n_1' = \frac{n}{10}$ subjects having the second smallest estimated success probabilities, and so on. Let $\overline{\pi}_k$ be the average estimated success probability based on the fitted model corresponding to the subjects in the subjects in the kth group with y=1 and let o$_k$ be the number of subjects with y=1 in the kth group.

The test proposed

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n_k'\overline{\pi}_k\right)^2}{n_k'\overline{\pi}_k(1 - \overline{\pi}_k)}$$

Where $\sum_{k=1}^{g} n_k' = n$ and n_k' k=1,2,…,g; g is usually chosen to be 10. is the total subjects in the kth group.

Under the null hypothesis, the test statistic is approximately distributed as a chi-square distribution with g-2 degrees of freedom.

2.3 Pearson Chi-Square test

Pearson chi-square test as a measure of goodness-of fit was proposed by Karl Pearson

Let $\hat{\pi}_j$ be the maximum likelihood estimate of π_j associated with the jth covariate pattern, then the experted number of success observed by subject in the jth group with jth covariate pattern is

$$\hat{y}_{j1} = m_j\hat{\pi}(x_j)$$

The likelihood function and the log-likelihood function can be written respectively as below for type two covariate pattern

$$L(\beta) = \prod_{j=1}^{J}\binom{m_j}{y_{j1}}(\pi(x_j))^{y_{j1}}(1 - \pi(x_j))^{m_j - y_{j1}}$$

$$LogL(\beta) = \prod_{j=1}^{J}\binom{m_j}{y_{j1}}(\pi(x_j))^{y_{j1}}(1 - \pi(x_j))^{m_j - y_{j1}} \,..$$

The Pearson residual is defined as

$$r(y_{j1}, \hat{\pi}(x_j)) = \frac{(y_{j1} - m_j\hat{\pi}(x_j))}{\sqrt{m_j\hat{\pi}(x_j)(1 - \hat{\pi}(x_j))}}$$

The Pearson Chi-Square test statistic is

$$\chi^2 = \sum_{j=1}^{J} r(y_{j1}, \hat{\pi}(x_j))^2$$

If the fitted model is correct, the sampling distribution of the Pearson chi-square test statistic can be approximated by a chi-square distribution with degrees of freedom J-(p+1), where p is the number of the parameters in the model. For the type one pattern case with J=n i.e. the number of unique covariate patterns is equal or almost equal to the number of subjects. Pearson’s chi-square[1] test statistic is not an applicable goodness-of-fit, because the test statistic will not have a Chi-square distribution. This due to the fact that when m$_j$=1, one subject for each covariate pattern, the Pearson residual does not asymptotically have a normal distribution for large n.

2.4 Deviance Test

The deviance as a measure of goodness-of-fit was first proposed by Nelder wedderburn[1]. The deviance residual for the jth covariate pattern is defined as .

$$d(y_{j1}, \hat{\pi}(x_j)) = \pm 2\left[y_{j1}\ln\left(\frac{y_{j1}}{m_j\hat{\pi}(x_j)}\right) + (m_j - y_{j1})\ln(\frac{m_j - y_{j1}}{m_j(1 - \hat{\pi}(x))})\right]\Big]\Big\}$$

Where the sign of the deviance residual is the same as that of $(y_{i1} - m_j\hat{\pi}(x_j))$

The deviance test statistic is

$$D = \sum_{j=1}^{J} d(y_{j1}, \hat{\pi}(x_j))^2$$

If the model is correct, the test statistic of Deviance has approximately a Chi-Square distribution with degree of freedom J-(p+1).

Under the null model and Type one covariate pattern, the deviance residual is reduced to

$$d(y_{i1}, \hat{\pi}(x_i)) = \pm\{2\left[y_i\ln(\frac{y_{i1}}{\hat{\pi}(x_i)}) + (1 - y_{i1})\ln\frac{(1 - y_{i1})}{(1 - \hat{\pi}(x_i))}\right]\}^{1/2}$$

Deviance residual measures the difference of the log likelihood between the assumed model and the saturated model. y_{i1} has two values 0 and 1, which indicates that $y_{i1}\log y_{i1}$ and $(1- y_{i1})\log(1- y_{i1})$ will be both 0. Then under Type one covariate pattern the test statistics of deviance can be expressed as

$$D = -2\sum\{\hat{\pi}(x_i)\log\hat{\pi}(x_i) + (1-\hat{\pi}(x_i))\log(1-\hat{\pi}(x_i))\}$$

The test statistic of Deviance does not involves in the comparison of the observed and fitted frequency of success but involves a comparison of the log of the maximum likelihood of the saturated model and the assumed model.

2.5 Principal component analysis

PCA decomposes a correlation matrix with ones on the diagonals. The amount of variance is equal to the trace of the matrix, the sum of the diagonals, or the number of observed variables in the analysis. PCA minimizes the sum of the squared perpendicular distance to the component axis. Components are uninterpretable, e.g., no underlying constructs. Principal components retained account for a maximal amount of variance[6].

The component score is a linear combination of observed variables weighted by eigenvectors. Component scores are a transformation of observed variables (C1 = b11x1 + b12x2 + b13x3 + . . .). With this done, these component scores could be used either as predictor variables or as criterion variables in subsequent analyses.

The PCA Model is Y = XB

Where Y is a matrix of observed variables

X is a matrix of scores on components

B is a matrix of eigenvectors (weights)

3. Research Results and Discussion

One option in investigating the power is to generate data under an alternative model, perform logistic regression on the generated data, and determine how often each goodness-of –fit test rejected the null hypothesis of an adequate logistic regression model

Part1: No Using PCA

Suppose the assumed model is of the form

$$H_{0:}\ \log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad 1)$$

Where $\beta_0 = -1.3$, $\beta_1 = 0.26$, $\beta_2 = 0.26$, $\beta_3 = 0.23$

In this simulation, 1000 random samples of size n=300, 500 and 1000 were respectively generated from model with covariate x_1 independently generated from an uniform distribution over the interval (-3, 3), covariate x_2 generated from beta distribution over the interval (4, 2) and covariate x_3 generated from normal distribution over the interval (0, 1). The response variable y associated with the covariate x_1, x_2 and x_3 was generated from a uniform distribution over the interval (0,1). The response y was assigned a value 1, if $\pi(x) \geq u$ and 0 otherwise at level of significance $\alpha = 0.05$. The observed rejection rates are presented in table1, where the known test $\hat{C}$ is the Hosmer - Lemeshow's $\hat{C}$ test; D is the Deviance test and P is the Pearson chi-square test in the table1.

Part2: Using PCA

In this simulation, 1000 random samples of size n=300, 500 and 1000 were respectively generated from model with covariate x_1 independently generated from an uniform distribution over the interval (-3, 3), covariate x_2 generated from beta distribution over the interval (4, 2) and covariate x_3 generated from normal distribution over the interval (0, 1). The response variable y associated with the covariate x_1, x_2 and x_3 was generated from a uniform distribution over the interval (0,1). The response y was assigned a value 1, if $\pi(x) \geq u$ and 0 otherwise. To reduce number of the variables by Using PCA with X_1, X_2 and X_3 and use component score from PCA as predictor variable (X) and fit logistic regression model.at level of significance $\alpha = 0.05$. The obsevered rejection rates are presented in table1, where the known test $\hat{C}$ is the Hosmer and Lemeshow's $\hat{C}$ test; D is the Deviance test and P is the Pearson chi-square test in the table1.

Table 1: Observed Type I for Simulation Study

Sample	Methods	No PCA	PCA
300	Pearson	0.006	0.005
	Deviance	0.430	0.560
	$\hat{C}$	0.040	0.040
500	Pearson	0.001	0.001
	Deviance	1	1
	$\hat{C}$	0.040	0.045
1000	Pearson	0.001	0.001
	Deviance	1	1
	$\hat{C}$	0.040	0.050

The results from the simulations for the assumed models when the sample sizes are 300,500 and 1000 are summarized in table1. The Tabled value is the percent of times the p-value from the goodness-of-fit test was less than indicate that under hypothesis, Part using PCA: the rejection rate of Deviance are larger than the upper bound of 95% confidence interval. the rejection rate of Pearson Chi-square test are less than the lower bound of 95% confidence with different sample. Type I error rate of Hosmer-Lemeshow' $\hat{C}$ close to or fall within 95%confiddence interval under different sample sizes. The results of using PCA showed that the ability of controlling type I error rate of the known tests as well as the results of no using PCA.

Next, researcher used five different interaction models to study the power with omission of interaction terms from the model. Researcher generated the outcome variable the random samples were generated by the following model:

$$\log((\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 x_3 \quad 2)$$

where β_4 takes the values 0.0, 0.2, 0.4, 0.6, 0.8 respectively. The known test were applied to the random sample generated from 2) to test the goodness-of-fit of the assumed model 1) at 5% level of significance. The simulated rejection rates of this test were presented in Table2-Table3.

Part1: No using PCA

Table 2: Rejection Rate for the known test for Simulation Study : No using PCA

Sample Size	β_4	$\hat{C}$	D	P
300	0.0	0.040	0.890	0.001
	0.2	0.070	0.960	0.002
	0.4	0.150	0.980	0.004
	0.6	0.200	0.990	0.002
	0.8	0.250	0.990	0.002
500	0.0	0.040	1	0.001
	0.2	0.080	1	0.001
	0.4	0.180	1	0.000
	0.6	0.270	1	0.000
	0.8	0.310	1	0.000
1000	0.0	0.04	1	0.001
	0.2	0.07	1	0.001
	0.4	0.15	1	0.000
	0.6	0.28	1	0.000
	0.8	0.40	1	0.000

Part2: To reduce number of the variables by Using PCA with continuous variables (X_1, X_2 and X_3) and use component score from PCA as predictor variable (X) and fit logistic regression model.

Table 3: Rejection Rate for the known test for Simulation Study : Using PCA

Sample Size	β_4	$\hat{C}$	D	P
300	0.0	0.050	0.890	0.005
	0.2	0.040	0.960	0.001
	0.4	0.090	0.980	0.003
	0.6	0.220	0.990	0.001
	0.8	0.250	1	0.001
500	0.0	0.040	1	0.003
	0.2	0.080	1	0.002
	0.4	0.200	1	0.001
	0.6	0.270	1	0.001
	0.8	0.320	1	0
1000	0.0	0.050	1	0.001
	0.2	0.080	1	0.001
	0.4	0.100	1	0.000
	0.6	0.280	1	0.000
	0.8	0.350	1	0.000

Table2-Table3 show the power, the percent of time each of this tests rejects the hypothesis of fit at the $\alpha = 0.05$ level. The simulation study, results suggest that the power of detecting lack of fit for Pearson Chi-square test is poor for different sample sizes departure from the true model. Deviance test has very high type I error rate. This simulation study gives a strong evidence that Pearson Chi-square test and Deviance test but the rejection rate of Hosmer-Lemeshow' $\hat{C}$ which using PCA is the best test among the known tests because it controls the type I error rate and has strong power to reject the assumed model when it is not true one, the power increases when the sample size increases.

The simulation study that, the power of Deviance test is getting poorer and poorer when the assumed model departures from the true model. For Pearson Chi-square test, bigger sample size and further departure from true model cannot enhance its power to detect the lack of fit. This simulation study showed that the Deviance test and Pearson chi-square test are not applicable to the type covariate pattern (J=N).

4. Conclusion

When continuous variable Pearson Chi-square test and Deviance test perform poorly for which they either cannot control type I error or have weak power in detecting the goodness-of-fit of the assumed model. Hosmer-Lemeshow tests are recommended to solve this problem. The proposed method which using PCA has good control of the type I error rate, and it has higher power in detecting omission of interaction terms from the model. It seems to perform well with overall steady rejection rate, with the type I rate lying within 95% confidence interval in all situations considered.

References

[1] Hosmer, D.W., hosmer, T., le Cessie, S., and Lemeshow, S. A comparison of gooness-of-fit for logistic regression model. Statistic in medicine. 1997: 16:9, 965-80.

[2] Collet, D. Modelling Binary data. London: Chapman and Hall; 1991.

[3] Ying Liu. On the goodness-of-fit logistic regression model [Dissertation]. Kansas: Kansas state University; 2007.

[4] Xian-Jin Xie, Jane Pendergast, William Clarke. Increasing the power: A practical approach to goodness-of –fit test for logistic regression models with continuous predictors. ScienceDirect. 2008; 52: 2703-2713.

[5] McCaullagh, P., J.A. Nelder. Generalized Linear Models. Biometrical Journal. 2nd ed. London: Chapman and Hall; 1989.

[6] Diana D., Suhr. Behaviour Principal Component Analysis vs Exploratory Factor Analysis [Internet]. Available from: http://www2.sas.com/proceedings/sugi30/203-30.pdf

Factors affecting English learning skills of the third year King Mongkut's University of Technology Thonburi Science students in 2012 academic year

Chunchom Pongchavalit[1*], Prapaphan Poolek[2] and Tananun Ninju
[1]*Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi*, *pongchavalit@yahoo.com*

Abstract

The purpose of this research is to study some factors which affecting the English learning skills of the third year King Mongkut's University of Technology Thonburi (KMUTT) Science students in 2012 academic year. The raw data that used to analyze in this research are English test scores and these answer from the questionnaires. The sample group in this research consists of 179 third year KMUTT science students, whose questionnaires are completely answered and English scores are variable. Two statistical methods used to analyze those data are Factor Analysis and Pearson Correlation analysis. The result from factor analysis showed that are all 26 variables were grouped into 5 main groups. These include the factor about the personal skills, the factor about benefits of English, the factor about English for every use, the factor about English for self-development, and the factor about the positive attitude in English. In addition, the result from correlation analysis also provides that there are only 3 factors from 5 factors that significantly correlated to the English test scores. These three factors are the factor about benefits of English, the factor about English for self development, and the factor about the positive attitude in English.

Keywords: English learning skills, factor analysis, Pearson correlation coefficients

*Corresponding Author
E-mail Address: pongchavalit@yahoo.com

1. Introduction

With the effective Asean Economic Community (AEC), economic activities among member countries, including Thailand, could rapidly expand. Foreign investment may flow into Thailand to use it as a production base for the Southeast Asian market by enjoying the zero tariffs under AEC. Apart from the expansion of market and production, leading to the creation of more employment opportunities for Thai workers, AEC, on the other hand, could allow the free movement of labor among member countries. This could present Thai workers with a competitive challenge from other ASEAN's workers. In order to either reap benefits from liberalization or survive the strong competition among themselves and from ASEAN workers, Thai skilled workers need to improve their necessary working skills, for example, English language, computer usage, and so on. Therefore, it is very crucial for Thai skilled workers to improve their working skills so that they can catch up with the globalizing trend and increase their job performance. One of the reasons for this lies in Thailand's historical background. The country had never been under colonial rules so its education system is mainly monolingual. English language has always been one of the weakest features of Thai skilled labor and it can be a decisive factor for any employment opportunities of workers. The interest on the determinants of English language capacity belongs to a broader field of the economics of language, which has developed since the late 1970s by Posel and Casale, [1]. Language capacity can be analyzed using the human capital framework. According to human capital theory, language skills can be considered as human capital because they can satisfy the three requirements for human capital, that it is productive, costly to produce, and embodied in the person by Chiswick, 2008 [2]. The purpose of this research is to study some factors which affect the English learning skills of the third year KMUTT Science students in 2012 academic year. The raw data that used to analyze in this research are English test scores and those answer from the questionnaires. Two statistical methods used to analyze those data are Factor Analysis and Pearson Correlation analysis.

2. Methodology

2.1 Population and sample group

2.1.1 Population

Populations used in this research are third year student in faculty of science, King Mongkut's University of Technology Thonburi in the academic year of 2012. The population consisted of students from four different department, Mathematics, Chemistry, Physics and Microbiology.

2.1.2 Sample Group

Sample group used in this research and reference are third year students in Faculty of Science, King Mongkut's University of Technology. The group was selected by the method of simple random sampling. The size of the sample groups are calculated by the formula of Taro Yamane at confident 95 % and error 5 %. Thus the calculation of sample size can use the following formula $n = \frac{N}{1+Ne^2}$ (1)

In this research the size of sample group is equal to 182 people which 179 of these the data was collected entirely.

2.2 Implementation in the research

2.2.1 Characteristics of the implementation

The implementation used to gather the information for this survey research is questionnaire. In order to assemble information about factors of English learning which is an importance of entering the AEC. In addition, the English proficiency test result of the students, organized by the university is also used.

Questionnaires are divided in 2 parts as following

Part 1 Asking about general information of the respondents: sex, grade in the subjects of LNG 101 General English, LNG102 English Skills and Strategies, LNG103 Academic English, their willingness of English learning, they do to enhance their English proficiency and skills in each area of the language. The questionnaire consists of check-lists and ranking questions.

Part 2 Questions are about factors effecting English learning process. The questions are designed to be answer by rating scale of 1-5. The appointed rating scales of the questions that are wording are given by Likert scale.

2.2.2. English proficiency test results

The test results were kindly given by the School of Liberal Arts. The results are divided in to 3 categories: Grammar, Reading and Writing skills. Te total score is 75 points. The numbers of third year, faculty of science student are 333. However, there were only 269 who did the test. Which 179 of these the data was completely collected. The results acquired are used in the analysis of relationship between the results itself and the factor effecting English learning process.

The steps of structuring the implementation are the following

1. Study from the existing research to be as a guideline for the questionnaire.

2. Specify the objective of the questionnaire to be as a guideline to create the questions which covers all the objectives of the research.

3. Make the questionnaire to cover all aspects of the objectives as recommended by the advisor.

4. Present the questionnaire to the advisor for corrections and improvement.

5. Try out the questionnaire with 30 people from sample group.

6. Questionnaires were attained to be test the reliability by using the Cronbach's Alpha coefficient determination formula. The value of Cronbach's Alpha (α) must not be less than 0.7. This is to indicate that the questionnaire is reliable and can be put to use with the chosen sample group.

The questionnaires consist of the 30 people experimental group. The value of Cronbach's Alpha coefficient was attained by using the SPSS program equal to 0.933. Thus the questionnaires use in this research is confirmed reliable.

2.3 Data collecting

1. The researcher requested aid from the third year student in Faculty of Science, King Mongkut's University of Technology. There were 182 students who complete the questionnaire. The group was selected by the method of simple random sampling.

2. The collected questionnaires were examined and the incomplete questionnaires were separated. The number of complete questionnaire were 179 which a 98.35 percent of the specific sample group.

2.4 Data Management

The collected questionnaire were examine thoroughly then enter these following process are Coding the data by the IBM SPSS Statistics 20 program and data processing through the data analysis procedure.

2.5 Data Analysis

The researcher had a progress through the process to fulfill the following objectives:

1. Gained the general information of respondents from part one on the questionnaire. Data are analyzed by frequency distribution, percentage and illustration by graph.

2. Learned the factors effecting English learning process from part 2 on the questionnaire. Data are analyzed by factor analysis. The components can explain the English learning process by the method of principle component analysis. This method can ensure the components are independence yet relational. The chose method for orthogonal rotation is varimax method.

3. English proficiency test results were analyze by frequency distribution, percentage averaging and finding relationship between factors of learning and test results by the method of Pearson Product Moment Correlation Coefficient.

2.6 Statistics

The analysis of the data obtained from the sample group was process by the IBM SPSS Statistics 20 program. The calculation are used consist of

1. Percentage $$P = \frac{f \times 100}{n} \quad (2)$$

2. Mean $$\bar{x} = \frac{\sum x}{N} \quad (3)$$

3. Cronbach's Alpha Coefficient

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum {s_i}^2}{{s_t}^2}\right] \quad (4)$$

4. Pearson Product Moment Correlation

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]\}}}$$

(5).

Table 1: Illustrating percentage of respondents' skills.

Rank of Skill	Grammar	Vocabulary	Reading	Listening	Writing	Conversation
1	16.2	14.0	**46.9**	6.7	4.5	12.3
2	10.1	**24.6**	26.8	10.1	13.4	15.1
3	10.1	20.7	11.2	21.2	**21.8**	15.1
4	10.6	18.4	9.5	**29.1**	15.6	17.3
5	17.9	14.5	3.9	17.9	19.6	**25.1**
6	**35.2**	7.8	1.7	15.1	25.1	15.1

Table 1 shows percent of respondents' skills. Rank 1, the students have skill reading 46.9%. Rank 2, the students have skill reading 26.8%. Rank 3, the students have skill writing 21.8%. Rank 4, the students have skill listening 29.1%. Rank 5, the students have skill conversation 25.1%. The last rank 6, the students have skill grammar 35.2%.

By factor analysis from 26 factors can analyze remain only 5 Factors show that.

Factor 1 Personal skills consist of: Satisfaction of oneself speaking skills, listening skills, reading skills, writing skills, vocabulary skills, grammar skills and factors relating to satisfaction of oneself English language skills.

Factor 2 Benefits of English consist of: Benefiting the chance of pursue to higher education locally, the chance to study abroad, future professional endeavors. English is considered to be the most common for international communication. When possess of English language skills, more opportunities to grow in the career. The factors relate to the benefits of having skills in English, either as a student or as a professional.

Factor 3 Daily usages of English consist of: English benefits the daily life, movie viewing, music listening, book reading, internet browsing. English encourage curiosity. English is ever learning process. English can help in socializing. The understanding of English also increases the understanding of other academic subjects. Realize of the importance of English. Show proud to use English to conversation. The factors are relating the benefit of usage of English in daily life, either as a hobby or building up confidence.

Factor 4 Personal development consist of: The desire to communicate with a foreigner, achieve goals of English learning, to be an advanced learner of the language. The aspiration to work can abroad with multinational organization. The factors are relating to greater opportunities to be successful. The understanding in English is important for being more successful than others.

Factor 5 English Learning consist of: Effortless understanding of English, confident in learning English, interests about other nation culture by using English as a median. The factors are relating to the attitude of the ability to understand English, confident level. Lastly, the ability is to apply English into the understanding of other cultures.

Table2: Show mean and percent of skill in English score.

Skill	scores	mean	percent
Grammar	30	14.81	49.37
Reading	25	12.38	49.52
Writing	20	3.04	15.2
total	75	30.23	40.31

Table 2 shows mean and percent of English score. Mean and percent of English for student are the same.

Table 3: Analyze correlation in each factor

	Personal skills	Benefits of English	Daily usages of English	Personal development	English Learning
Pearson Correlation	0.104	0.191*	0.016	0.179*	0.204**
p-value	0.166	0.01	0.829	0.017	0.006
N	179	179	179	179	179

1. The correlation of score and Personal skills is 0.104. The alpha is 0.166 greater than 0.05 that means no correlation between score and Personal skills.
2. The correlation of score and Benefits of English is 0.191. The alpha is 0.01 less than 0.05 that means correlation between score and Benefits of English.
3. The correlation of score and Daily usages of English is 0.016. The alpha is 0.829 greater than 0.05 that means no correlation between score and Daily usages of English.
4. The correlation of score and Personal development is 0.179. The alpha is 0.008 less than 0.05 that means correlation between score and Personal development.
5. The correlation of score and English Learning is 0.204. The alpha is 0.003 less than 0.05 that means correlation between score and English Learning.

3. Conclusion and Suggestions

From the research, the factor in learning English of the third year students in Faculty of Science, King Mongkut's University of Technology Thonburi. The researcher had concluded the following suggestions.

1. The number of sample group in the factor analysis technique should be 10 times greater than the factor or the question in the questionnaire. In which there were 26 factors, therefore the sample group should be at least 260 people, but due to the number of 179 students were present for the test so only 179 questionnaires were obtained.
2. In this research, the population chosen was only the third year student. This population is only a tally of 1 in 4 of the whole population of the faculty. Analyze of the relation between factors in English learning and the English proficiency test results. The researcher had requested the 2011 academic year test result from the School of Liberals Arts. The results given were only of the second year student (now third year). Consequently, if the research was to be more accurate and reliable then it should study on every year of student better than selecting only one year of student.
3. The implementation should be more varied as well as the method of data collecting. Adding to using of questionnaire, perhaps use the method of interviews or observation. This can result in additional data collected.
4. In the questionnaire where the respondents were asked about their grade in the passing 3 LNG subjects. The sample group was uncertain about their grade as they fail to recall it or in some instance the student start with the subjects LNG 102 or LNG 103. Hence the data collected were incomplete.
5. On account of finding relation between the relation between factors in English learning and the English proficiency test results. The necessity to use the student identity number to match with each student attitude have raised suggestion from several respondents that the personal information should not be shown. However due time restraint, another English proficiency test cannot be organized.
6. The test results given from the School of Liberals Arts were considered personal information. Therefore the research can only demonstrate in the average value

and percentage value as the raw score cannot be shown individually.

7. From the English proficiency test results, it indicates that the most problematic skill is the writing skills as the result was only 3.04 out of 20 points. The development is needed for this skill. The approach can be added to the normal academic plan or by laying the foundation during the orientation before entering the university. Likewise for grammatical skills and reading skills the gradual development is needed. This is to strive to be at the same level with the neighboring country and the ease of entering the ASEAN Economic Community.

References

[1] Posel, D. and Casale, D. (2010). English Language Proficiency and Earnings in a Developing Country: the Case of South Africa. Working Paper Number 181, University of KwaZulu-Natal, KwaZulu-Natal

[2] Chiswick, B.R. (2008). The Economics of Language: An Introduction and Overview. Discussion Paper No. 3568, Institute for the Study of Labor, Bonn.

Factors affecting employment selection of Mathematics graduate students at King Mongkut's University of Technology Thonburi

Chunchom Pongchavalit* and Pakavat Srisakdanuwat
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi,
pongchavalit@yahoo.com

Abstract

The objective of this research was to study about needs, attitudes, factors affecting employment selection of Mathematics which are used in the future occupation of graduated students of Mathematics Department, Faculty of Science at King Mongkut's University of Technology Thonburi. The factors affecting employment selection of Mathematics of this research were 71 employed graduated students by Simple Random Sampling. The research methodology is exploratory research. The research instruments used for gathering information were questionnaires with rating scale. The data was analyzed by percentage, frequency; mean standard deviation, Chi – square test and factor analysis. The findings of the study can be summarized as follows: from the studying relationship of seeking occupation and work place of graduates shows that choosing occupation does not depend on parents' occupation but depends on cousins' occupation. The place for work does not depend on hometown. Moreover, achievement of graduated students does not have only their work experience but also other factors. The studying relationship between occupation and subjects shows that most of graduated students work in computer, finance, accountancy, marketing, management and network, respectively.

Keywords: Factor analysis, occupation, gratification, attitude

*Corresponding Author
E-mail Address: pongchavalit@yahoo.com

The efficiency comparison of tests for equality of two variances

Treerit Chotisathiensup[1*], Kamon Budsaba[2] and Supranee Lisawadi[3]

[1]*Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani, Thailand, treerit@mathstat.sci.tu.ac.th*

[2]*Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani, Thailand, kamon@mathstat.sci.tu.ac.th*

[3]*Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani, Thailand, supranee@mathstat.sci.tu.ac.th*

Abstract

Monte Carlo studies were conducted to compare on robustness and power of four statistical tests for equality of two variances. The equality of variance tests were (1) a nonparametric Wald test (called R test), (2) the Brown-Forsythe test, (3) the Levene test and (4) the F test where each test was analyzed under five distributions. The distributions were (1) normal, (2) chi-square, (3) exponential, (4) gamma and (5) Weibull distributions. The results from the Monte Carlo simulation studies demonstrated the Brown-Forsythe test was very good in terms of robustness, but it was poorly in terms of power when compared to the other tests. The F test performed well when the distribution is normal at level of significance 0.05. The R test outperformed the Levene test in terms of robustness and power at level of significance 0.01. The R test was nearly as robust as the Brown-Forsythe test at significance level 0.01 and nearly as power as the F test at significance level 0.05.

Keywords: R test, Brown-Forsythe test, Levene test, F test

*Corresponding Author
E-mail Address: jackkrup@hotmail.com

1. Introduction

The analysis of variance (ANOVA) is the most powerful technique for testing hypotheses about this phenomenon when the assumptions of normality, homogeneity of variance, and independence of errors are met. Failure of any assumption would impair the utility of the test, leading to the wrong and invalid conclusions (Cochran,1947; Bodhisuwan, 1991; Srisunsanee, 1998). Therefore, it is necessary to test the assumptions before using analysis of variance (Vorapongsathorn, Taejaroenkul and Viwatwongkasem 2004). There are several literatures search about the equality test of variances.

Bodhisuwan and Kuvattana (2002) compare the powers of three tests for homogeneity of variance of four populations using Monte Carlo simulation technique. Three homogeneity of variance tests were Bartlett's test, Layard chi square test, and squared ranks test under different ratios of variances and different sample sizes and three distributions.

Vorapongsathorn, Taejaroenkul, and Viwatwongkasem (2004) study compared the probability of type I error and the power of three statistical tests (Bartlett, Levene and Cochran) by varying the sampling distribution, variances and sample sizes. Monte Carlo methods were used to generate responses based on sample sizes and distributions with 1,000 repetitions. The sample sizes were both equal and unequal: 15, 30 and 45. The data distributions were normal, gamma and chi-square. It was found that Bartlett's test was sensitive to the normality assumption whereas Cochran's test and Levene's test were robust when the normal assumption was violated. Moreover, Levene's test was quite good for both equal and small sample sizes. In the case of power, Bartlett's test had the highest power in all cases. When one variance was large, Cochran's test was the best test.

Lee, Katz, and Restori (2010) compared seven homogeneity of variance tests on robustness and power using Monte Carlo studies. The homogeneity of variance tests were Levene's test, modified Levene test, Z-variance test, Overall-Woodward modified Z-variance test, O'Brien test, Samiuddin cube-root test, and F-max test. Each test was subjected to Monte Carlo analysis through different five shaped distributions is normal, platykurtic, leptokurtic, moderate skewed, and highly skewed.

Allingham and Rayner (2011) gave a new test for equality of variances based on what might be called a nonparametric version of a very natural Wald test which is introduced and called it the R test. In an indicative empirical study they showed that, in moderately-sized samples, the R test is nearly as powerful as the F test when normality may be assumed, and is nearly as robust as Levene's test when normality is in doubt.

In this study, four tests for equality of variances will be compared on robustness and power using Monte Carlo studies. The four equality of variance tests are (1) F test, (2) Levene test, (3) Brown-Forsythe test and (4) R test. The main point of study is examining whether R test is good when compared to each test determined by test the robustness and power of tests.

In Section 2, four tests for the equality of variance are presented (1) F test, (2) Levene test, (3) Brown-Forsythe test and (4) R test. In Section 3 it is shown type I errors rate and powers of tests and the summary of result for four sample sizes (n=20, 30, 40 and 50) and five distributions are shown. The distributions were (1) normal distribution, (2) chi-square distribution, (3) exponential distribution, (4) gamma distribution and (5) Weibull distribution.

2. Tests for Equality of Two variances

Let $X_{11},\ldots,X_{1n1}$ and $X_{21},\ldots,X_{2n2}$ are two independent random samples, and independent and identically distributed (iid). The two sample sizes are equal ($n_1=n_2=n$). We wish to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

2.1 F-test

F-test [5,7] is used to test the equality of variances of two populations.

The F-test is defined as the following: H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

$$F = \frac{S_1^2}{S_2^2}\,;\; S_1^2 > S_2^2$$

where: S_1^2 and S_2^2 are sample variances.

Decision rule: Reject H_0 if F > F(α/2,m-1,n-1) or F < F(1-α/2,m-1,n-1) otherwise do not reject H_0, where: F(α/2,m-1,n-1) is the upper critical value and F(1-α/2,m-1,n-1) is the lower critical value of the F-distribution with m-1 and n-1 degrees of freedom at a significance level of α.

2.2 Levene's test

Levene [4] is a popular test which is used to test for equality of variances for variable of two or more groups. Levene's test is just the pooled t-test applied to the sample residual.

The Levene test is defined as the following: H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\overline{Z}_i - \overline{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_i)^2}$$

where:
k is the number of different groups
N is the total number of sample size
n_i is the number of sample size for group
$Z_{ij} = \left|X_{ij} - \overline{X}_i\right|$ $\overline{X}_i$ is the mean of the ith subgroup

$\overline{Z}_i = \dfrac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$ is the group mean of the Z_{ij}

$\overline{Z} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}}{N}$ is the overall mean of the Z_{ij}

Decision rule: Reject H_0 if W > F(α,k-1,N-k) otherwise do not reject $H_{0,}$ where: F(α/2,m-1,n-1) is the upper critical value of the F-distribution with k-1 and N-k degrees of freedom at a significance level of α.

2.3 Brown Forsythe test

Brown Forsythe [2] is an application of Levene test by using median instead of the mean in computing Z_{ij}. The Brown Forsythe test statistic is:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\tilde{Z}_i - \tilde{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \tilde{Z}_i)^2}$$

where:
k is the number of different groups
N is the total number of sample size
n_i is the number of sample size for group
$Z_{ij} = \left|X_{ij} - \tilde{X}_i\right|$ $\tilde{X}_i$ is the median of the ith subgroup

$\tilde{Z}_i = \dfrac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$ is the group mean of the Z_{ij}

$\tilde{Z} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}}{N}$ is the overall mean of the Z_{ij}

Decision rule: Reject H_0 if W > F(α,k-1,N-k) otherwise do not reject H_0, where: F(α/2,m-1,n-1) is the upper critical value of the F-distribution with k-1 and N-k degrees of freedom at a significance level of α.

2.4 R-test

The R test [1] is a nonparametric Wald test. It can be used when normality is in doubt but appropriate without assumption of normality.

The R test is defined as the following: H_0: $\sigma_1^2 - \sigma_2^2 = 0$
H_1: $\sigma_1^2 - \sigma_2^2 \neq 0$

$$R = \frac{\left(S_1^2 - S_2^2\right)^2}{\left(m_{14} - S_1^4\right)/n_1 + \left(m_{24} - S_2^4\right)/n_2}$$

where:
S_i^2 is sample variance for ith group
m_{i4} is the fourth central sample moments for the ith group
n_i is the number of sample size for ith group

Decision rule: Reject H_0 if R > χ_1 = 3.841 otherwise do not reject H_0, where: χ_1 = 3.841 is the critical value of the chi-square distribution with 1 degrees of freedom at a significance level of α.

3. Comparison Type I Error Rate and Empirical Power of Tests under Different Distributions

Monte Carlo study is used to evaluate in terms of robustness and power for a comparison of all test statistics in this study. For robustness, we will compare type I errors rate of each test based on Cochran.

The computer programing (R, 2014) was written to analyse type I errors. The random numbers follow five distributions: normal, chi-square, exponential, gamma, and Weibull distribution. 1000 simulated two group experiments were analyzed for four different sample sizes of n=20, 30, 40, and 50 of two significance level 0.01 and 0.05 (α = 0.01 and 0.05). The result of these tests are shown in Table 1 and Table 2.

3.1 Nominal significance level 0.01 (α = 0.01)

Table 1 gives type I errors rate and empirical power of tests at level α = 0.01. In terms of robustness, the results from these Monte Carlo studies found in the normal distribution, the F and the Levene tests could control type I error for all sample sizes. The R test could control type I error for all sample sizes (except small sample size n=20) when the distribution was normal. The Brown-Forsythe test could control type I error when sample sizes were 40 and 50. The Levene and the R tests could control type I error for all sample sizes (except sample size of 20) when the distribution was chi-square. The Brown-Forsythe test could control type I error for all sample sizes when the distribution was exponential. For gamma distribution, the Levene and R tests could control type I error for the sample size of 30 and the type I error rates were 0.010 for the Levene test and 0.012 for the R test. The Brown-Forsythe test was the only test that could control type I error for sample sizes of 20, 30 and 40 when the distribution was Weibull distribution. For F, Levene and R tests could not control type I error for all sample sizes.

In terms of power, the empirical power of tests tend to get higher as the sample size increases in every distributions and test statistics. In normal distribution, the F test had the highest empirical powers for all sample sizes (except small sample size n=20) compared to the other test statistics. The F test outperformed the other test statistics in chi-square distribution for sample size of 30. The R test had empirical powers higher than the Levene test for gamma distribution.

Table 1: Type I error rates and empirical powers of tests at significance level 0.01

Sample size	F test		Levene test		Brown-Forsythe test		R test	
	Type I errors	Power	Type I errors	Power	Type I errors	Power	Type I errors	Power
Normal Distribution								
20	0.014*	0.394	0.010*	0.248	0.004	0.192	0.034	0.422
30	0.014*	0.622	0.012*	0.452	0.004	0.420	0.008*	0.586
40	0.012*	0.798	0.007*	0.624	0.010*	0.578	0.010*	0.756
50	0.012*	0.902	0.008*	0.806	0.010*	0.784	0.012*	0.862
Chi-Square Distribution								
20	0.012*	0.394	0.018	0.216	0.006	0.182	0.034	0.386
30	0.010*	0.628	0.008*	0.484	0.008*	0.414	0.010*	0.580
40	0.020	0.808	0.010*	0.660	0.004	0.632	0.014*	0.722
50	0.016	0.884	0.008*	0.798	0.004	0.752	0.012*	0.842
Exponential Distribution								
20	0.146	0.896	0.028	0.696	0.012*	0.340	0.020	0.410
30	0.150	0.958	0.044	0.878	0.014*	0.670	0.016	0.528
40	0.152	0.992	0.044	0.974	0.010*	0.874	0.014*	0.630
50	0.190	0.998	0.048	0.994	0.008*	0.964	0.012*	0.702
Gamma Distribution								
20	0.018	0.438	0.016	0.262	0.006	0.196	0.026	0.368
30	0.022	0.626	0.010*	0.458	0.002	0.424	0.010*	0.506
40	0.036	0.782	0.016	0.664	0.006	0.580	0.016	0.674
50	0.052	0.858	0.020	0.784	0.006	0.728	0.020	0.766
Weibull Distribution								
20	0.132	0.908	0.020	0.638	0.012*	0.320	0.018	0.428
30	0.160	0.970	0.046	0.876	0.010*	0.632	0.016	0.542
40	0.172	0.998	0.060	0.960	0.008*	0.846	0.016	0.642
50	0.184	1.000	0.060	0.996	0.002	0.958	0.022	0.792

*estimated type I error rate is inside (0.007,0.014)

Table 2: type I errors rate and empirical power of tests at significance level 0.05

Sample size	F test		Levene test		Brown-Forsythe test		R test	
	Type I errors	Power	Type I errors	Power	Type I errors	Power	Type I errors	Power
				Normal Distribution				
20	0.053*	0.640	0.044*	0.568	0.040*	0.480	0.092	0.692
30	0.044*	0..824	0.042*	0.752	0.042*	0.722	0.070	0.834
40	0.048*	0.932	0.054*	0.888	0.042*	0.864	0.063	0.930
50	0.050*	0.962	0.056*	0.920	0.044*	0.914	0.067	0.954
				Chi-Square Distribution				
20	0.073	0.634	0.060*	0.544	0.038	0.528	0.098	0.656
30	0.050*	0.840	0.052*	0.748	0.048*	0.674	0.070	0.822
40	0.057*	0.928	0.050*	0.862	0.040*	0.832	0.062	0.920
50	0.060*	0.968	0.054*	0.944	0.044*	0.920	0.060*	0.968
				Exponential Distribution				
20	0.264	0.942	0.130	0.892	0.042*	0.744	0.088	0.718
30	0.290	0.984	0.152	0.972	0.048*	0.934	0.075	0.806
40	0.280	0.994	0.130	0.990	0.052*	0.984	0.068	0.894
50	0.272	1.000	0.144	0.996	0.038	0.994	0.066	0.888
				Gamma Distribution				
20	0.080	0.644	0.058*	0.516	0.030	0.494	0.100	0.606
30	0.096	0.806	0.054*	0.756	0.048*	0.682	0.074	0.780
40	0.106	0.908	0.070	0.868	0.054*	0.812	0.072	0.878
50	0.096	0.960	0.072	0.930	0.054*	0.896	0.062	0.924
				Weibull Distribution				
20	0.256	0.964	0.126	0.890	0.036	0.702	0.086	0.708
30	0.284	0.992	0.124	0.976	0.052*	0.890	0.068	0.820
40	0.298	1.000	0.138	0.996	0.032	0.976	0.072	0.940
50	0.302	1.000	0.150	1.000	0.052*	0.992	0.074	0.940

* estimated type I error rate is inside (0.04,0.06)

3.2 Nominal significance level 0.05 (α = 0.05)

Table 2 shows type I errors rate and empirical power of tests in different distributions with sample sizes of n = 20, 30, 40 and 50 at significance level of α = 0.05. The results demonstrate that the significance level was increased 5% (α = 0.05) as well as the type I errors rate and the empirical power of tests.

In terms of robustness, all test statistics underlying normal distribution could control type I error for all sample sizes (except R test). The R test could control type I error for sample size of 50 when the distribution was chi-square while the remaining case could not control. The Brown-Forsythe test was the only test that could control type I error for sample sizes of 20, 30 and 40 when the distribution was Weibull distribution. For F, Levene and R tests could not control type I error for all sample sizes.

In terms of power, the F test had empirical powers higher than the Levene and the Brown-Forsythe test for normal distribution and all sample sizes. The empirical power of the R and F tests were 0.968 for sample size of 50 in chi-square distribution. For gamma distribution, the Levene test had empirical powers higher than the Brown-Forsythe test for sample sizes of 30.

4. Conclusion

Monte Carlo studies were used to simulate the data in calculating type I errors and powers of tests for the comparison tests for equality of two variance underlying five distributions and four sample sizes at significance level 0.01 and 0.05 for 1000 experiments that demonstrate the result in section 3.

In conclusion, a comparison of test statistics showed that the Brown-Forsythe test exhibited the greatest robust for both level of significance and all distributions and all sample sizes. For power of tests, the F test is the best power for both levels of significance, all distributions, and moderate sample sizes (n=30, 40, and 50). The R test was nearly as robust as the Brown-Forsythe test at significance level of 0.01 and nearly as power as the F test at significance level of 0.05.

References

[1] Allingham D, and Rayner J. Two-Sample Testing for Equality of Variances, Proceeding of the 4th Annual ASEARC; 2011 February 17-18; Paramatta, Australia.

[2] Brown M.B., and A.B.Forsythe. Robust tests for the equality of variances, Journal of the American Statistical Association 1974; 69:364-367

[3] Howard B.L, Katz G.S, and Restori A.F. A Monte Carlo Study of Seven Homogeneity of Variance Tests, Journal of Mathematics and Statistics 2010; 6(3): 359-366

[4] Levene, H. Robust Tests for Equality of variances.In I. Olkin (ed.), Contributions to Probability and Statistics, Stanford University Press, Stanford, CA, 1960; pp. 278–292

[5] Snedecor and Cochran. Bartlett's test [Internet].1983 [updated 2012 April 10; cited 2013 October 30]. Available from: http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm

[6] Vorapongsathorn, T., Taejaroenkul, S., and Viwatwongkasem, C. (2004). A Comparison of Type I Error and Power of Bartlett's Test, Levene's Test, and Cochran's Test under Violation of Assumptions. Songklanakarin Journal Science and Technology 26, 537-547.

[7] Zieliński W. Robust test for comparison of two variances, Acta Universitatis Lodziensis, Folia Oeconomica, 235, 109-118

[8] Cochran, W.G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometics, 3: 22-38

[9] Bodhisuwan, W. 1991. A comparison of the test statistics for homogeneity of variances. [M.S.Thesis in Statistics]. Bangkok: Faculty of Graduate Studies, Chulalongkorn University.

[10] Srisunsanee, J. 1998. Effects of failure to meet assumptions of homogeneity of variance and normal on Type I error in one-way analysis of variance. [M.S. Thesis in statistics]. Bangkok: Faculty of Graduate Studies, Kasetsart University.

[11] R Core Team (2014). R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

An empirical comparison of homogeneity of variance tests

Yada Pornpakdee[1*], Kamon Budsaba[2] and Wararit Panichkitkosolkul[3]

[1]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, yada@mathstat.sci.tu.ac.th*

[2]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, kamon@mathstat.sci.tu.ac.th*

[3]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, wararit@mathstat.sci.tu.ac.th*

Abstract

The objective of this research was to compare homogeneity tests of variances when sample sizes are equal. Analysis of Means for Variances (ANOMV), Samiuddin Cube Root test, and Bartlett's test were compared under only normal distribution. Analysis of Means for Variances Version of Levene's test (ANOMV-LEV), Analysis of Means for Variances Version of Transformed Ranks (ANOMV-TR), Levene's test, Modified Levene's test**,** and Trimmed Mean Levene's test were compared under normal distribution, t distribution, lognormal distribution, double exponential distribution, gamma distribution, and logistic distribution. The comparison criteria capability to control type I error rate and its empirical power at significance level 0.05 were focused. The data were simulated by the Monte Carlo technique with 10,000 time replications. In the case of the normal population, Samiuddin Cube Root test and ANOMV test can control the type I error rate. However, Bartlett's test showed highest empirical powers. In the case of the non-normal population, Modified Levene's test can control of type I error rate for all distributions. ANOMV-LEV test, ANOMV-TR test, Levene's test, Modified Levene's test and Trimmed Mean Levene's test performed empirical power all distributions.

Keywords: Type I error rate, empirical power, analysis of means for variances, Levene's test

*Corresponding Author
E-mail Address: yada@mathstat.sci.tu.ac.th

1. Introduction

Statistics is the most widely used branch of mathematics in the quantitative research. Statistical methods are used extensively within fields such as economics, social sciences and biology. Quantitative research using statistical methods start with the collection of data, based on the hypothesis or theory.

Tests for homogeneity of variances are often of interest as a preliminary to other analyses such as analysis of variance or a pooling of data from different sources to yield an improved estimated variance.

The classical approach to hypothesis testing usually begins with the likelihood ratio test under the assumption of normal distributions. However, the distribution of the statistic in the likelihood ratio test for equality of variances in normal populations depends on the kurtosis of the distribution ((George, E.P. (1953) [12]), which helps to explain why that test is so sensitive to departures from normality. This nonrobust (sometimes called "puny") property of the likelihood ratio test has prompted the invention of many alternative tests for variances. Some of these are modifications of the likelihood ratio test. Others are adaptations of the F test to test variances rather than means. Many hypothesis testing are based on nonparametric methods, although their modification for the case in which the means are unknown often makes these tests distributionally dependent.

The assumption is the homogeneity of variance (HOV). That is, in an ANOVA, we assume that treatment variances are equal:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Moderate deviations from the assumption of equal variances do not seriously affect the results in the ANOVA (Keselman,HJ. (2007) [13]). Therefore, the ANOVA is robust to small deviations from the HOV assumption. We only need to be concerned about large deviations from the HOV assumption.

Evidence of a large heterogeneity of variance problem is easy to detect in residual plots. Residual plots also provide pattern informations among the variance. Some researchers like to perform a hypothesis test to validate the HOV assumption. We will consider three common HOV tests; Bartlett's test, Levene's test, and the Brown-Forsythe test.

These tests are not powerful for detecting small or moderate dierences in variances (Keselman,HJ.(2007)

[13]). This is satisfied because we are only concerned about large deviations from the HOV assumption.

When the assumption is not the normal variability in the test, the performance is needed. Levene (1960) [4] performed and transformed each observations. The observations were applied by the ANOVA F test to the resulting absolute deviations from the median (ADM), which have to convert the observations y_{ij} into $y_{ij} =| x_{ij} - \tilde{x}_i |$ for i =1,…,i and j =1,…, n_i . where $\tilde{x}_i$ is the median of sample i. The analysis of means (ANOM) version of Levene’s test (ANOMV-LEV) is move efficient than a non-normal distribution population.

Bartlett's test was used with chi-squared distribution and the degrees of freedom of the test with the test population of k groups. Gartside PS (1972) [3] the population distribution is normal. Bartlett test has high test power. It can be seen that the results of research will be of the good quality only depends on the technique or how to test by selecting the most suitable test statistic for the situation of the data. That is use the empirical power, so it should be up to the test of a statistical test for equality of variances.

Fligner and Killeen (1981) [5] applied the ANOVA-F test to the transformed ranks of the absolute deviations from the median, which have to convert the observed E_{ij} into $E_{ij} = \Phi^{-1}\left(0.5+\frac{r_{ij}}{2N+1}\right)$; $r_{ij} = \text{rank}(y_{ij})$; $y_{ij} =| x_{ij} - \tilde{x}_i |$ and $\tilde{x}_i$ is the median of sample i and Φ^{-1} is the inverse normal score and N is the total number of observations. Which ANOM version of the Fligner–Killeen (ANOMV-TR) is applied to the transformed ranks of the absolute deviations from the median.

Since 1997, Wludyka and Nelson [1] had proposed a statistics test used to the HOV hypothesis test. This test is called the analysis of means for variances (ANOMV) and is accomplished by transforming the observations y_{ij} into $t_{ij} = (y_{ij} - \bar{y}_{i.})^2$ which when averaged become sample variances (with a multiplier). Power comparisons given by (Wludyka and Nelson (1997) [1]) indicate that ANOMV is a good choice for testing hypothesis when the assumption of normality is reasonable.

In case that the population is non-normal distribution for the homogeneity of variance tests of population more than two groups have the statistics for multiple testing, such as Levene, Modified Levene, Z-variance, Overall–Woodward Modified Z-variance, O'Brien, Samiuddin Cube Root and F-Max. Lee (2010) [2] said that the modified Levene test showed very good robustness when compared to the other tests but lower power than other tests. The Samiuddin test is the best in terms of robustness and power when the distribution is normal.

The researchers investigate the statistical tests both under normal and non-normal distributions. ANOMV, Samiuddin Cube Root test and Bartlett's test were compared under normal distribution. ANOMV-LEV, ANOMV-TR, Levene's test, Modified Levene's test and Trimmed Mean Levene’s test were compared under normal distribution, t distribution, lognormal distribution, double exponential distribution, gamma distribution, and logistic distribution. These tests were compared using a Monte Carlo approach at significance level 0.05 and replicated 10,000 times.

2. Background

2.1 ANOMV for Balanced Studies

Letting $\sigma^2 = \sigma.^2$ and $\tau_i = \sigma_i^2 - \sigma^2$.

They can be used to from the decision lines

$$UDL = U(\alpha; k, n-1)kMS_e$$

$$CL = MS_e$$

$$LDL = L(\alpha; k, n-1)kMS_e$$

where: $MS_e =$

$$\frac{(n_1 -1)(s_1^2) + (n_2 -1)(s_2^2) + ... + (n_i -1)(s_i^2)}{N}$$

n_i is the sample size of the i th group

k is the number of groups

$U_{(\alpha,k,n-1)}$ is upper critical value

$L_{(\alpha,k,n-1)}$ is lower critical value

$U_{(\alpha,k,n-1)}$ and $L_{(\alpha,k,n-1)}$ is table ANOMV critical values (Wludyka and Nelson (1997) [1])

2.2 Samiuddin Cube-Root Test

Samiuddin and Atiqullah (1976) [6] developed a homogeneity of variance test they refers to as the Bayesian test of homogeneity. Samiuddin and Atiqullah (1976) showed that the “cube-root” test is superior over some other tests such as the Bartlett test when the sample distributions are not homogeneous. However, Levy (1978) [7] had shown Samiuddin and Atiqullah’s findings to be flawed or misleading. This study will re-examine the Samiuddin test in terms of robustness and power and in comparison to some other homogeneity of variance tests not tested by Samiuddin and Atiqullah (1976) [6] and Levy (1978).[7]

If $X_{ij}(i = 1, 2, ..., k; j = 1, 2, ..., n_i)$ are normally distributed with mean μ_i and variance σ_i^2 , Samiuddinand Atiqullah (1976) define

$$s_i^2 = \sum_{j=1}^{k}(X_{ij} - M_i)^2 \text{ and } s_i^2 = S_i^2 / v_i ,$$

where $M_i = \sum_{j=1}^{k} X_{ij} / n_i$ and $v_i = n_i - 1$. When $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, Samiuddin and Atiqullah (1976) [6] shows $\sum (m_i - m)^2 / a_i^2$ is approximately distributed as a χ^2 with k-1 degrees of freedom where:

$$m_i = (\frac{v_i}{S_i^2})^{1/3}(1 - \frac{2}{9v_i})$$

$$a_i^2 = \frac{2}{9}(S_i^2)^{2/3} v_i^{1/3}$$

and: $$m = \frac{(\sum \frac{m_i}{a_i^2})}{(\sum \frac{1}{a_i^2})}$$

2.3 Bartlett's Test

Bartlett's test (Snedecor and Cochran, 1983) [8] is used to test the homogeneity of variances. Equal variances across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples.

$$B = \frac{1}{c}[\sum_{i=1}^{k}(n_i - 1)\ln s^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2]$$

where:

$$c = 1 + \frac{1}{3(k-1)}[\sum_{i=1}^{k}\frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{k}(n_i - 1)}]$$

$$s_i^2 = \frac{\sum_{j=1}^{n}(X_{ij} - \overline{x}_i)^2}{n_i - 1}$$

$$s^2 = \frac{\sum_{j=1}^{k}(n_i - 1)s_i^2}{\sum_{j=1}^{k}(n_i - 1)}$$

n_i is the sample size of the i th group

k is the number of groups

s_i^2 is the variance of the i th group

s^2 is the variance of the k th group.

2.4 ANOMV-LEV: The ANOMVersion of Levene's Test

The homogeneity of variance test (HOV) by Levene (1960) is performed by transforming each observation using (1) and then applying the ANOVA *F* test to the resulting ADMs. In the ANOM version of Levene's test (ANOMV-LEV), the ANOM test is applied to the ADM; that is

$$y_{ij} = | x_{ij} - \tilde{x}_i | \qquad (1)$$

for

i = 1,2,…,*I*

j = 1,2,…, n_i

$\tilde{x}_i$ = is the median of sample *i*

The overall mean $\overline{Y}.. = \frac{\sum_{i=1}^{k}\overline{Y}_i}{k}$

when $\overline{Y}_i = \frac{\sum_{i=1}^{n_i} y_{ij}}{n_i}$

They can be used to from the decision lines

$$UDL = \overline{Y}.. + h(\alpha; k, N-k)\sqrt{MSe}\sqrt{\frac{k-1}{N}}$$

$$LDL = \overline{Y}.. - h(\alpha; k, N-k)\sqrt{MSe}\sqrt{\frac{k-1}{N}}$$

where:

$U_{h(\alpha,k,N-k)}$ is upper critical value

$L_{h(\alpha,k,N-k)}$ is lower critical value

$U_{h(\alpha,k,N-k)}$ and $L_{h(\alpha,k,N-k)}$ is table balanced ANOM critical values $h(\alpha, k, N-k)$ (Wludyka and Nelson (1997) [1])

2.5 ANOMV-TR: The ANOMVersion of the Fligner–Killeen Test

In the HOV test proposed by Fligner and Killeen (as presented in Conover et al. (1981)) [5], the ANOVA *F* test is applied to the transformed ranks of the absolute deviations from the median (2).ANOMV-TR, in which ANOM is applied to the transformed ranks of theADMs, is described by Wludyka and Nelson (1997), who used a Monte Carlo study to demonstrate ANOMVTR's robustness.

$$y_{ij} = | x_{ij} - \tilde{x}_i |$$

$$r_{ij} = \mathrm{Rank}(y_{ij})$$

$$E_{ij} = \Phi^{-1}\left(0.5 + \frac{r_{ij}}{(2N+1)}\right) \qquad (2)$$

where

$\tilde{x}_i$ = is the median of sample *i*

Φ^{-1} = is the inverse normal score

N = is the total number of observations

Grand Means $$\bar{\bar{E}} = \frac{\sum_{i=1}^{k} \bar{E}_i}{k}$$

when; $$\bar{E}_i = \frac{\sum_{i=1}^{n_i} E_{ij}}{n_i}$$

Pooled Variance $$S^2 = \frac{\sum_{i=1}^{k} S_i^2}{k}$$

when; $$S_i^2 = \frac{\sum_{j=1}^{n_i} (E_{ij} - \bar{E}_i)^2}{n-1}$$

They can be used to from the decision lines

$$UDL = \bar{\bar{E}} + h_{(\alpha,k,N-k)}\, S\sqrt{\frac{k-1}{nk}}$$

$$LDL = \bar{\bar{E}} - h_{(\alpha,k,N-k)}\, S\sqrt{\frac{k-1}{nk}}$$

where:

$U_{h(\alpha,k,N-k)}$ is upper critical value

$L_{h(\alpha,k,N-k)}$ is lower critical value

$U_{h(\alpha,k,N-k)}$ and $L_{h(\alpha,k,N-k)}$ is table balanced ANOM critical values $h(\alpha,k,N-k)$ (Wludyka and Nelson (1997) [1])

2.6 Levene's test

In 1960, Levene proposed an alternative method to the Bartlett Test (Klotz and Johnson, 1993) [11] for testing the assumption of homogeneity of variance for independent sample t-test and ANOVA designs. The Bartlett test works well for data that are normally or approximate normally distributed. The Bartlett test does not fare well for data that follow a leptokurtic or skewed distribution (Overall and Woodward, 1974) [10]. According to Levene (Gastwirth *et al*., 2009) [9], the test he proposed was less sensitive to departures from normality. This says that the Levene test had fewer Type 1 errors than the Bartlett Test for distributions that were aberrant from normality.

If Y_{ij} 's represent the raw scores on the dependent

measure: $$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_i)^2}$$

where:

N = Total sample size

n_i = Sample size for group

$$Z_{ij} = | Y_{ij} - \bar{Y}_i |$$

$\bar{Y}_i$ = The mean of the ith subgroup

$\bar{Z}_i$ = The group means of the Z_{ij}

$\bar{Z}$ = The overall mean of the Z_{ij}

2.7 Modified Levene's test

The modified Levene test is nearly identical to the original Levene test. The difference is that the median is used instead of the mean in computing Z_{ij}. That is $Z_{ij} = | Y_{ij} - \widetilde{Y}_i |$, where $\widetilde{Y}_i$ is the median of the ith subgroup. This is the modification studied earlier by Brown and Forsyth is referenced in Carroll and Schneider (1985) [14].

2.8 Trimmed Mean Levene's test

The trimmed mean Levene test is nearly identical to the original Levene test. The difference is that the trimmed mean is used instead of the mean in computing Z_{ij}. That is $Z_{ij} = | Y_{ij} - \widetilde{Y}'_i |$, where $\widetilde{Y}'_i$ is the 10% trimmed mean of the ith subgroup. This is the modification studied earlier by Brown and Forsyth is referenced in Carroll and Schneider (1985) [14].

3. Research Methodology

This research study to compare of tests for homogeneity of variance, include ANOMV, Samiuddin Cube Root test and Bartlett's test under normal ANOMV-LEV, ANOMV-TR, Modified Levene's test and Trimmed Mean Levene's test under normal distribution, t distribution, lognormal distribution, double exponential distribution, gamma distribution, and logistic distribution.

The criteria used for comparison is the study are of type I error rate and empirical power were criteria. The Monte Carlo study follow the step bellow.

3.1 Generate data with various distributions

3.2 Write computed property values of the test statistic using program R version 3.0.3 to be used to test the type I error rate.

3.3 Once created, the data is done through sampling distribution of the population according to the specified size, to test for equality of population variance in each situation with 8 statistics test.

3.4 Take the test statistic calculated relative to the crisis to conclude that reject or accept the null hypothesis at significance level 0.05 for repeated 10,000 times.

3.5 In simulation the type I error rate by recording the number of times that reject the null hypothesis when the null hypothesis is true to determine the type 1 error rate is estimated as follows.

$$\text{The probability of type I error} = \frac{\text{The number of reject the null hypothesis}}{10,000}$$

3.6 Consider the power of the statistic test of the 8 characters, the case that can control the type I error rate, by recording the number of rejected null hypothesis when the null hypothesis is not true, and determine the empirical power.

4. Results

4.1 Normal distribution

Table 1: Observed relative frequencies of type 1 error rate for analysis of samples with equal variance (Robustness)

Normal Distribution (k=3)			
	n=15	n=30	n=45
ANOMV	0.038	0.047	0.041
Samiuddin Cube Root	0.031	0.048	0.040
Bartlett	0.047	0.050	0.049

Monte Carlo study of robustness: In case of the normal population, Samiuddin Cube Root test and ANOMV test have the observed relative frequencies of type I error rate less than those of the Bartlett test.

Table 2: Observed relative frequencies of power for analysis of samples with unequal variance (empirical power)

Normal Distribution (k=3)			
	n=15	n=30	n=45
ANOMV	0.872	0.951	0.984
Samiuddin Cube Root	0.896	0.969	0.996
Bartlett	0.958	0.999	1

Monte Carlo study of power: In case of the normal population is typically that when the sample size (n = 15, 30 and 45) Bartlett's test is the highest, followed by the Samiuddin. Cube Root test and ANOMV test respectively for the normal distribution.

4.2 Non-Normal distribution

Monte Carlo study of robustness: In case of the non-normal population to be that the normal distribution ANOMV-TR test, Modified Levene's test and ANOMV-LEV test have the observed relative frequencies of type I error rate less than those of the Levene's test and Trimmed Mean Levene's test for n=15. ANOMV-TR test and Modified Levene's test can control the type I error rate for n=30 and n=45

t distribution, Modified Levene's test and ANOMV-TR test can control the type I error rate.

Lognormal distribution, Modified Levene's test and ANOMV-TR test have the observed relative frequencies of type I error rate less than those of the Trimmed Mean Levene's test, ANOMV-LEV test and Levene's test for n=15. ANOMV-TR test and Modified Levene's test can control the type I error rate for n=30 and n=45

Double exponential distribution, Modified Levene's test can control the type I error rate for n=15 and n=30. Modified Levene's test, ANOMV-TR test and Trimmed Mean Levene's test have the observed relative frequencies of type I error rate less than those of the Levene's test and ANOMV-LEV for n=45.

Gamma distribution, ANOMV-TR test, ANOMV-LEV test and Modified Levene's test can control the type I error rate.

Logistic distribution, Modified Levene's test, Trimmed Mean Levene's test and ANOMV-TR test have the observed relative frequencies of type I error rate less than those of the Levene's test and ANOMV-LEV test for n=30. Modified Levene's test and ANOMV-TR test can control the type I error rate for n=15 and n=45. The best values in Table 3 are in bold print

Monte Carlo study of power: In case of the non-normal population to be that the normal distribution of sample size n=15, n=30 and n=45 were tested, including ANOMV-LEV test, ANOMV-TR test, Levence's test, Modified Levene's test and Trimmed Mean Levene's test. (Efficiency did not differ.)

t distribution, ANOMV-LEV test, ANOMV-TR test, Levence's test, Modified Levene's test and Trimmed Mean Levene's test. Efficiency was almost the same for n=15, n=30 and n=45.

Lognormal distribution, the sample size case n = 15, n = 30and n = 45 were tested, including ANOMV-LEV test, ANOMV-TR test and Levence's test. Effectively is no different except the case of Modified Levene's test and Trimmed Mean Levene's test. The sample size n = 15, n = 30 and n = 45 was less effective than the other test statistics.

Double exponential distribution, ANOMV-LEV test, ANOMV-TR test, Levence's test, Modified Levene's test and Trimmed Mean Levene's test. All sample sizes had the some in efficiency. Efficiency did not differ for n=15, n=30 and n=45.

Gamma distribution, ANOMV-LEV test, ANOMV-TR test, Levence's test, Modified Levene's test and Trimmed Mean Levene's test. Efficiency did not differ for n=15, n=30 and n=45.

Logistic distribution, ANOMV-LEV test, ANOMV-TR test, Levence's test, Modified Levene's test and Trimmed Mean Levene's test. Efficiency did not differ for n=15, n=30 and n=45. The best values are printed in bold in Table 4.

Table 3: Observed relative frequencies of type 1 error rate for analysis of samples with equal variance (Robustness)

	normal	t	lognormal	double exponential	gamma	logistic
n =15						
ANOMV-LEV	0.050	0.077	0.131	0.071	0.034	0.074
ANOMV-TR	0.022	0.042	0.047	0.064	0.029	0.042
Levene	0.058	0.088	0.264	0.068	0.192	0.062
Modified Levene	0.030	0.031	0.037	0.035	0.042	0.033
Trimmed Mean Levene	0.055	0.066	0.135	0.060	0.130	0.056
n=30						
ANOMV-LEV	0.063	0.062	0.017	0.064	0.031	0.063
ANOMV-TR	0.031	0.037	0.038	0.057	0.025	0.050
Levene	0.055	0.079	0.251	0.058	0.183	0.053
Modified Levene	0.039	0.039	0.038	0.044	0.049	0.040
Trimmed Mean Levene	0.053	0.053	0.087	0.053	0.102	0.049
n =45						
ANOMV-LEV	0.065	0.079	0.003	0.060	0.030	0.060
ANOMV-TR	0.043	0.045	0.019	0.050	0.022	0.055
Levene	0.052	0.084	0.254	0.056	0.179	0.050
Modified Levene	0.043	0.043	0.042	0.046	0.049	0.044
Trimmed Mean Levene	0.051	0.056	0.088	0.050	0.102	0.051

Table 4: Observed relative frequencies of power for analysis of samples with unequal variance (empirical power)

	normal	t	lognormal	double exponential	gamma	logistic
n =15						
ANOMV-LEV	0.826	0.759	0.997	0.956	0.735	0.815
ANOMV-TR	0.890	0.805	0.852	0.872	0.896	0.897
Levene	0.877	0.891	0.948	0.965	0.962	0.985
Modified Levene	0.793	0.663	0.043	0.932	0.852	0.963
Trimmed Mean Levene	0.867	0.790	0.253	0.953	0.923	0.978
n =30						
ANOMV-LEV	0.982	0.949	0.998	0.971	0.814	0.836
ANOMV-TR	0.959	0.973	0.898	0.965	0.918	0.956
Levene	0.999	0.999	0.976	0.998	0.997	1
Modified Levene	0.997	0.986	0.095	0.997	0.997	1
Trimmed Mean Levene	0.998	0.988	0.148	0.997	0.998	1
n =45						
ANOMV-LEV	1	0.912	0.995	0.996	0.865	0.871
ANOMV-TR	0.990	0.869	0.973	0.980	0.971	0.983
Levene	1	0.959	0.985	1	1	1
Modified Levene	1	0.911	0.142	1	1	1
Trimmed Mean Levene	1	0.930	0.186	1	1	1

5. Conclusions

When considering robustness, ANOMV test, Samiuddin Cube Root test, and Bartlett’s test under a normal distribution have the ability to control the type I error rate better is not different. ANOMV-LEV test is able to control the type I error rate, better in the normal distribution and the gamma distribution for n = 15. However, ANOMV-LEV test is able to control the type I error rate in the lognormal distribution and gamma distribution for n=30 and n=45. ANOMV-TR test has the ability to control the type I error rate for all distributions, except the double exponential distributed for n = 15, 30 and logistic distribution in cases n = 45. Besides, Levene’s test had the ability to control the type I error rate in logistic distribution for n = 45, Modified Levene’s test is to control the type I error rate all distributions and Trimmed Mean Levene’s test is to control the type I error rate in the case of logistic distribution n = 30 and double exponential distribution in case n = 45.

When considering power, ANOMV test, Samiuddin Cube Root test and Bartlett's test under a normal distribution, there are high efficiency test. ANOMV-LEV test, ANOMV-TR test, Levene's test, Modified Levene's test and Trimmed Mean Levene’s test have high efficiency test all distributions. Although, in the lognormal distribution, Modified Levene's test and Trimmed Mean Levene’s test had lower efficiency than other methods.

References

[1] Wludyka and Nelson. The Analysis of Mean: A Graphical Method for Comparing Means, Rates, and Proportions. Virginia: American Statistical Association, 1997.

[2] Lee Howard B. A Monte Carlo Study of Seven Homogeneity of Variance Tests. Mathematics and Statistics Journal. 2010;6(3): 359-366.

[3] Gartside, P.S. A Study of Methods for Comparing Several Variances. American Statistics Association Jourmal. 1972; 67: 342-346.

[4] Levene, H. Robust Tests for Equality of variances.In I. Olkin (ed.), Contributions to Probability and Statistics, Stanford University Press, Stanford, CA, 1960; pp. 278–292.

[5] Fligner, M. A., and Killeen, T. J. Distribution-Free Two-Sample Tests for Scale. American Statistics Association Jourmal. 1981; 71, pp. 210–233.

[6] Samiuddin, M, M. Atiqullah. A Test of Equality of Variance. Biometrika,1976; 63: 206-208.

[7] Levy, K.J. An Empirical Study of The Cube-Root Test for Homogeneity of Variance with Respect to The Effects of Non-Normality and Power. J. Stat. Comput. Simul Journal.1978; 7: 71-78.

[8] Snedecor and Cochran. Bartlett's test [Internet].1983 [updated 2012 April 10; cited 2013 October 30]. Available from: http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm

[9] Gastwirth, J.L., Y.R. Gel ,W. Miao. The Impact of Levene' s Test of Equality of Variances on Statistical Theory And practice. Stat. Sci., 24: 343-360.

[10] Overall, J.E., J.A. Woodward. A Simple Test for Homogeneity of Variance in Complex Factorial Design. Psychometrika.1974; 39: 311-218.

[11] Klotz, S. , N.L. Johnson. Breakthroughs in Statistics: Volume1: Foundations and Basic Theory. 1st Edn., Springer, New York,1993; pp: 680.

[12] George, E.P. Box. Non-normality and Tests on Variances. Biometrika,1953; 40: 318-335.

[13] Keselman, HJ. Test for homogeneity. University of Manitoba Winnipeg.2007;p113-129.

[14] Carroll, R.J. and H. Schneider. A note on Levenc's tests for equality of variances. Stat. Probab. Lett. 1985; 3: 191-194.

The efficiency comparison of test for differences among several population means under heterogeneity of variances

Uparittha Intarasat[1*], Kamon Budsaba[2] and Saengla Chaimongkol[3]
[1]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, uparittha@mathstat.sci.tu.ac.th*
[2]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, kamon@mathstat.sci.tu.ac.th*
[3]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, saengla@mathstat.sci.tu.ac.th*

Abstract

The purpose of this study is to compare the efficiency of the statistical tests for testing differences among several population means under heterogeneity of variances. Heteroscedastic analysis of means test (HANOM), analysis of mean test (ANOM), Welch test, Brown-Forsythe test and Analysis of variance F-test (ANOVA-F) under 3 and 5 group means are investigated. The distributions of considered population are normal, beta, t and chi-square. The methods are compared by considering the ability to control the type I error rate and empirical power. The test is based on 0.05 levels of significance. In case of the under 3 group and 5 group means, HANOM test can control type I error rate. HANOM, ANOM, Welch test, Brown-Forsythe test and ANOVA-F exhibit highest empirical powers. However, Welch test has lower empirical power test of normal distribution and beta distribution for all sample sizes in case the number of treatment groups is 5 (k=5).

Keywords: Type I error rate, empirical power

*Corresponding Author
E-mail Address: uparittha@mathstat.sci.tu.ac.th

1. Introduction

Most of the studies in practice concern comparison of the difference of group means. The one-way fixed effects analysis of variance F-test (ANOVA-F) is a commonly used technique for comparing the effects of k independent group means. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. Additionally, in its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and generalizes the t-test to more than two groups. Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. Therefore, ANOVAs are useful in comparing three or more means for statistical significance. Besides, analysis of means is used for the same purpose. ANOM test is developed under the assumptions of normality and homogeneity of variance as in ANOVA. ANOM test is used for comparing the group means, proportions or rates, and testing the homogeneity of the variances. Since it is a graphical method, understanding and interpreting the results are quite easy. At the same time, ANOM graphics provide information about practical significance of the differences in question, as well as their statistical significance. Despite having important advantages over variance analysis, ANOM test has some incompetence such as limited use as a method, and ambiguity of its performance when the normality and homogeneity of variance assumptions are not met. If the populations from which data to be analyzed by ANOVA were sampled violate one or more of its assumptions, the results of the analysis may be incorrect or misleading. For example, if the assumption of independence is violated, then the ANOVA is simply not appropriate, although another test may be appropriate. If the assumption of normality is violated, or outliers are present, then the ANOVA may not be the most powerful test available, and this could mean the difference between detecting a true difference among the population means or not. A nonparametric test is a transformation may result in a more powerful test. A potentially more damaging assumption violation occurs when the population variances are unequal, especially if the sample sizes are not approximately equal (unbalanced). Often, the effect of an assumption violation on the one-way ANOVA result depends on the extent of the violation (such as how unequal the population variances are, or how heavy-tailed one or another population distribution is). Some small violations may have little practical effect on the analysis, while other violations may render the one-way ANOVA result uselessly incorrect or uninterpretable. In particular, small or unbalanced sample sizes can increase vulnerability to assumption violations. In 1947, Welch B.L. [6] has described a method for comparing the mean of 2 normal populations, when the ratio of their variances is unknown. The same of method can be used to compare 2 regression coefficients. In the present paper we derive by similar method a test for the equality of k = (r+1) mean, or regression coefficients. In 1997, Keselman HJ, Wilcox RR. [7] presented an 'improved' Brown and Forsythe (1974) statistic which is designed to provide a valid test of mean equality in independent groups designs when variances are heterogeneous. In particular, the usual Brown and Fosythe procedure is modified by using a Satterthwaite approximation for numerator degrees of freedom instead of the usual value of number of groups minus one. In 2011, Oyeyemi GM. and Adeleke LB [8] The analysis of means (ANOM) is a graphical and statistical method used in illustrating important variations among group of means (populations) and it is commonly employed in statistical quality control. ANOM is a technique originally developed by Ott (1967) for comparing a group of treatment means in order to ascertain if any of them differs significantly from the overall mean at a specified significance level. ANOM methodology compares the mean of each group to overall mean to detect statistically significant differences among means. It is not limited to comparison of group means only but also used to compare rates, proportions and variances. While analysis of variance (ANOVA) is the most commonly used method to compare the means of several groups, analysis of means (ANOM) serves as alternative procedure for comparing means. In 2013, Mendes M. and Yigit S. [1] At this time, There is no study in the literature that ANOM test is compared with ANOVA-F or other tests in terms of type I error rate and test power. Many studies are available in the literature that ANOVA-F was compared with some alternative tests in terms of type I error rate and test power. The main purpose of this study is to compare performances of ANOM test with the ANOVA-F under various conditions.

Since, there is a not clear result in the selection of test methods all 5 ways. Therefore, research in this study is the comparison of the average population in 5 ways to test methods, include an Analysis of variance (ANOVA) Analysis of mean test (ANOM) , Brown-Forsythe test ,Welch test and Heteroscedastic analysis of mean (HANOM) under heterogeneity of variances, purpose of the study to compare the ability to control the Type I error and empirical power statistic of the population that shall be used in any statistical test comparing the average population under heterogeneity of variances. This study requires a comparison by simulation in Monte Carlo; they are compared to the type I error rate and empirical power.

2. Statistical tests

2.1 One-way ANOVA-F

ANOVA-F (Mendes M. and Yigit S., 2013) [1] ANOVA-F is used to test the hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ versus the alternative that H_1 : at least one of the μ_i is different. The F-ratio is computed as the ratio of the mean square between treatment group means (MST) to treatments group within means (MSE).

$$F = \frac{MST}{MSE}$$

where

$$MST = \frac{SSTr}{k-1}$$

$$MSE = \frac{SSE}{N-k}$$

$$SSE = SST - SSTr$$

$$SSTr = \sum_{i=1}^{k} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{\sum_{i=1}^{k} n_i}$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{\sum_{i=1}^{k} n_i}$$

k = the number of treatment groups

n_i = the sample sizes of the i th group

The critical test statistic is obtained from the F-distribution with k − 1 and N − k degrees of freedom. If F-ratio is equal or greater than critical F-value, then H_0 will reject otherwise H_0 will not reject.

2.2 ANOM test

ANOM test, (Mendes M. and Yigit S., 2013) [1] as reported by many authors, can be taken into account as an alternative to the ANOVA. (Balamurali and Kalyanasundaram, 2007) [2] report that the ANOM is sometimes referred to as an alternative to the ANOVA. Therefore, ANOM test can be used to test the hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ versus the alternative that H_1 : at least one of the μ_i is different as well ANOVA-F. ANOM is a graphical analogue to ANOVA, and tests the equality of population means. One important difference is that ANOVA tests whether the treatment means differ from each other, while ANOM tests whether the treatment means differ from the grand mean. The ANOM is performed by computing UDL and LDL and checking to see whether any of the means fall outside decision lines or not.

The steps of ANOM test are:

Treatment mean are calculated as

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$

The overall mean is calculated as

$$\bar{Y}_{..} = \frac{\bar{Y}_{1.} + \bar{Y}_{2.} + ... + \bar{Y}_{k.}}{k}$$

Sample variances are calculated as

$$S_i^2 = \frac{\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}{n-1}$$

MSE used as an estimate of the true population variance is computed as

$$MSE = \frac{S_1^2 + S_2^2 + ... + S_k^2}{k}$$

for equal sample size and

$$MSE = \frac{\sum_{i=1}^{k} \left(n_i - 1\right) S_i^2}{N-k}$$

for unequal sample size, respectively.

UDL and LDL are computed as

$$UDL = \bar{Y}_{..} - h\left(\alpha, k, N-k\right)\sqrt{MSE}\sqrt{\frac{k-1}{N}}$$

and $$LDL = \bar{Y}_{..} - h\left(\alpha, k, N-k\right)\sqrt{MSE}\sqrt{\frac{k-1}{N}}$$

for equal sample size and

$$UDL = \bar{Y}_{..} + h\left(\alpha, k, N-k\right)\sqrt{MSE}\sqrt{\frac{N-n_i}{Nn_i}}$$

and $$LDL = \bar{Y}_{..} - h\left(\alpha, k, N-k\right)\sqrt{MSE}\sqrt{\frac{N-n_i}{Nn_i}}$$

for equal sample size, respectively. where k is the number of treatment groups, N is the total number of observation, n_i is the sample size for the *i*th group and $h\left(\alpha, k, N-k\right)$ is found in Table Balanced ANOM Critical Values $h\left(\alpha, k, N-k\right)$ on significance level (α), number of mean being compared (k) and degrees of freedom for means square error (N-k).

2.3 Welch test

Welch, B.L. (1951) [3] The Welch test for general k compares the statistic

$$F_{Welch} = \frac{\sum_{j=1}^{k} w_j (\bar{y}_j - \tilde{y}_{..})^2 / \left(k-1\right)}{1 + \left[2\left(k-2\right) / \left(k^2 - 1\right)\right] \sum_{j=1}^{k} h_j}$$

to the F(k-1,v) distribution, where

$$w_j = n_j / s_j^2 \quad , \quad \hat{\mu} = \sum_{j=1}^{k} w_j y_j / W$$

$$W = \sum_{j=1}^{k} w_j$$

$$h_j = \left(1 - w_j / W\right)^2 / \left(n_j - 1\right)$$

$$v = \left(k^2 - 1\right) / \left(3 \sum_{j=1}^{k} h_j\right)$$

2.4 Brown-Forsythe test

The procedure presented by Brown and Forsythe (1974) [5] is based on a test statistic, in which the numerator and denominator have the same expected value under null hypothesis,

$$F_{BF} = \frac{\sum_{i=1}^{k} n_{.i}\left(\bar{y}_{.i} - \bar{y}_{..}\right)^2}{\sum_{i=1}^{k} \left(1 - n_i / N\right) S_i^2}$$

The F^* statistic is approximately distributed as an F variable with $df_1 = k-1$ and $df_2 = f$ degrees of freedom, where f is obtained with the Satterthwaite (1941) approximation as

$$\frac{1}{f} = \sum_{i=1}^{k} c_i^2 / \left(n_i - 1\right)$$

with

$$c_i = \left(1 - n_i / N\right) S_i^2 / \left[\sum_{i=1}^{k} \left(1 - n_i / N\right) S_i^2\right]$$

2.5 HANOM test

Nelson PR, Wludyka PS, and Copeland KAF. (2005) [4] let k be the number of treatments being compared, and let y_{ij} be the *j*th observation from the *i*th population, where

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

and all the observations are independent. Now, however, we assume that $\varepsilon_{ij} \sim N\left(0, \sigma_i^2\right)$.

Compute

$$n_i = \max\left\{n_0 + 1, \left\lfloor \left(w / \delta\right)^2 s_i^2 \right\rfloor + 1\right\}$$

for each i, where $\lfloor y \rfloor$ denotes the greatest integer in y, and take $n_i - n_0$ additional observations $y_{i,n_0+1}, ..., y_{i,n_i}$ from population i.

For each i calculate the sample mean

$$\bar{y}_i = \frac{y_{i,n_0+1} + ... + y_{i,n_i}}{n_i - n_0}$$

of the second set of observations from population i.

For each i compute

$$b_i = \frac{n_i - n_0}{n_i}\left[1 + \sqrt{\left(\frac{n_0}{n_i - n_0}\right)\left(\left[\frac{\delta}{w}\right]^2 \frac{n_i}{s_i^2} - 1\right)}\right]$$

and

$$\tilde{\bar{y}}_i = \left(1 - b_i\right)\bar{y}_{0i} + b_i \bar{y}_i$$

and computed as

$$\tilde{\bar{y}}_\bullet = \left(\tilde{\bar{y}}_1 + \ldots + \tilde{\bar{y}}_k\right) / k$$

Compute decision lines

$$\tilde{\bar{y}}_\bullet \pm H\left(\alpha; k, n_0 - 1\right)\frac{\delta}{w}$$

where $H\left(\alpha; k, n_0 - 1\right)$ is found in Table HANOM Critical Values $H\left(\alpha; k, n_0 - 1\right)$, and reject the hypothesis $H_0 : \mu_1 = \ldots = \mu_k$ if any of the $\tilde{\bar{y}}_i$ fall outside these decision lines.

3. Research Methodology

This research is experimental research to study the statistical tests of population means under heterogeneity of variances. The statistical tests, Analysis of variance F-test (ANOVA-F), Analysis of mean test (ANOM), Brown-Forsythe test, Welch test and Heteroscedastic analysis of mean (HANOM) when a random sample from a population with a normal distribution. beta distribution, t distribution and chi-square distribution, the criteria used for comparison study of the probability of I error type and the test population means in this research, the simulation data with the Monte Carlo Simulation Technique with a computer program written in R version 3.0.3 Repeat 10,000 times with the research process as follow.

3.1 Generate data with normal distribution beta distribution, t distribution and chi-square distribution.

3.1.1. Create normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

3.1.2. Create t distribution with $\text{d.f.} = 10$.

3.1.3. Create chi-square distribution with $\text{d.f.} = 3$.

3.1.5. Create a beta distribution with parameters α = 10 and β = 10

3.2 Random from the sample size

When creating the data with different distribution, the next step is sampled from a population to test a population mean in each scenario, the scope of research with all 5 of the test statistic.

3.3 Comparison of the statistics on the crisis

Take the test statistic calculated relative to the crisis, to conclude that the rejection or acceptance of the null hypothesis at 0.05 levels of significance do this 10,000 times.

3.4 Estimating the probability of Type 1 error

Approximately the probability of Type I error by recording the number of times that reject the null hypothesis when the null hypothesis is true, to determine the relative frequency of rejecting the null hypothesis when the null hypothesis is true.

$$\text{type I error} = P(\text{reject null hypothesis} \mid \text{null hypothesis is true})$$

3.5 Estimated Empirical power

Estimated the statistical empirical power by recording the number of 5 methods of the denial null hypothesis, when the null hypothesis is false to determine the relative frequency of rejecting the null hypothesis when the null hypothesis is false.

$$\text{power} = P(\text{reject null hypothesis} \mid \text{null hypothesis is false})$$

3.6 The Criterion of estimating the probability of Type 1 error

If the probability value is below 0.05 ,so we can control the type I error rate.

3.7 The Criterion of estimated Empirical power

If empirical power value is nearby 1 show that a good efficiency.

4. Results

This research is studied to compare the statistics test of the population under unequal variances which comparison with the average many ways. There is 5 ways in this example, include Analysis of variance F-test (ANOVA-F), Analysis of mean test (ANOM), Brown-Forsythe test , Welch test, and Heteroscedastic analysis of means test (HANOM), when random the sample from a population with a normal distribution, beta distribution, t distribution and chi-square distribution. In this case study has considered 2 important issues are type I error rate and empirical power, that can be summarized the results as follows.

4.1 Empirical type I error rate estimates when number of treatment groups (k = 3) under heterogeneity of variances.

Monte Carlo study of robustness: The results in case of the under 3 group means demonstrate that the ANOM test perform the best in producing overall the fewest type I errors rate across beta distribution, t distribution and chi-square distribution for all sample sizes.

ANOVA-F test perform the best in producing overall the fewest type I errors rate across beta distribution, t distribution and chi-square distribution for n=10 and n=20. ANOVA-F test can control the type I error rate across t distribution and chi-square for n=30.

Brown-Forsythe test perform the best in producing overall the fewest type I errors rate across beta distribution, t distribution and chi-square distribution for n=10 and n=20. Brown-Forsythe test can control the type I error rate across t distribution and chi-square for n=30.

Welch test perform the best in producing overall the fewest type I errors rate across normal distribution and t distribution for all sample sizes.

HANOM test perform the best in producing overall the fewest type I errors across all distributions for all sample sizes.

4.2 Empirical type I error rate estimates when number of treatment groups (k = 5) under heterogeneity of variances.

Monte Carlo study of robustness: The results in case of the under 5 group means demonstrate that the ANOM test perform the best in producing overall the fewest type I errors across all distributions for all sample sizes.

ANOVA-F test perform the best in producing overall the fewest type I errors rate across beta distribution, t distribution and chi-square distribution for all sample sizes.

Brown-Forsythe test perform the best in producing overall the fewest type I errors across all distributions for n=10. Brown-Forsythe test can control the type I error rate across beta distribution, t distribution and chi-square distribution for n=20 and n=30.

Welch test perform the best in producing overall the fewest type I errors rate across t distribution and chi-square distribution for n=10.Welch test no good in producing overall the fewest type I errors rate across all distributions for n=20 and n=30.

HANOM test perform the best in producing overall the fewest type I errors across all distributions for all sample sizes.

4.3 Empirical power estimates when number of treatment groups (k=3) under heterogeneity of variances.

Monte Carlo study of power: The results in case of the under 3 group means demonstrate that ANOM test has empirical power does not differ all distributions for all sample sizes.

ANOVA-F test has empirical power does not differ all distributions for all sample sizes.

Brown-Forsythe test has empirical power does not differ all distributions for all sample sizes.

Welch test has empirical power does not differ normal distribution and t distribution for all sample sizes.

HANOM has high empirical power for does not differ all distributions for n=20 and n=30. But, in the chi-square distribution has lower empirical power for n=10.

4.4 Empirical power estimates when number of treatment groups (k=5) under heterogeneity of variances.

Monte Carlo study of power: The results in case of the under 5 group means demonstrate that ANOM test has high empirical power does not differ all distributions for all sample sizes.

ANOVA-F test has high empirical power does not differ all distributions for all sample sizes.

Brown-Forsythe test has high empirical power does not differ all distributions for all sample sizes.

Welch test has empirical power does not differ t distribution and chi-square distribution for n=10.

HANOM has high empirical power in the three distributions is normal, t and chi-square for all sample sizes.

Table 1: Type I error rates for k = 3 and k = 5 under heterogeneity of variance

	Test	k = 3				k = 5			
		normal	beta	t	chi-square	normal	beta	t	chi-square
n = 10	ANOM	0.0574	0.0456*	0.0379*	0.0394*	0.0401*	0.0380*	0.0385*	0.0394*
	ANOVA-F	0.0602	0.0471*	0.0430*	0.0430*	0.0518	0.0462*	0.0428*	0.0434*
	BF	0.0565	0.0414*	0.0395*	0.0377*	0.0499*	0.0420*	0.0372*	0.0372*
	WELCH	0.0470*	0.5239	0.0410*	0.2478	0.0718	0.0552	0.0061*	0.0074*
	HANOM	0.0460*	0.0030*	0.0018*	0.0285*	0.0079*	0.0004*	0.0010*	0.0273*
n = 20	ANOM	0.0536	0.0468*	0.0422*	0.0411*	0.0388*	0.0388*	0.0394*	0.0390*
	ANOVA-F	0.0610	0.0495*	0.0479*	0.0464*	0.0511	0.0490*	0.0455*	0.0475*
	BF	0.0585	0.0473*	0.0464*	0.0440*	0.0503	0.0481*	0.0431*	0.0450*
	WELCH	0.0450*	0.9596	0.0420*	0.7680	0.6196	0.5135	0.0971	0.1014
	HANOM	0.0442*	0.0011*	0.0002*	0.0339*	0.0072*	0.0012*	0.0003*	0.0155*
n = 30	ANOM	0.0522	0.0477*	0.0448*	0.0424*	0.0446*	0.0405*	0.0354*	0.0443*
	ANOVA-F	0.0601	0.0528	0.0480*	0.0484*	0.0506	0.0487*	0.0485*	0.0499*
	BF	0.0573	0.0517	0.0478*	0.0472*	0.0506	0.0484*	0.0331*	0.0487*
	WELCH	0.0390*	0.9991	0.0500*	0.9620	0.9433	0.9002	0.0852	0.3739
	HANOM	0.0323*	0.0008*	0.0010*	0.0093*	0.0013*	0.0014*	0.0013*	0.0062*

* Type I errors rate could control

Table 2: Empirical power for k = 3 and k = 5 under heterogeneity of variances

	Test	k = 3				k = 5			
		normal	beta	t	chi-square	normal	beta	t	chi-square
n = 10	ANOM	1	0.9983	0.9882	0.9452	1	1	0.9996	0.9998
	ANOVA-F	1	0.9984	0.9897	0.9560	1	1	0.9997	1
	BF	1	0.9975	0.9883	0.9510	1	1	0.9997	1
	WELCH	0.9870	0.0126	0.9930	0.1368	0.0001	0.0257	0.9980	0.9990
	HANOM	0.9238	0.8351	0.8046	0.3903	0.9471	0.0956	0.9699	0.9109
n = 20	ANOM	1	1	0.9984	0.9997	1	1	1	1
	ANOVA-F	1	1	0.9984	0.9998	1	1	1	1
	BF	1	1	0.9984	0.9998	1	1	1	1
	WELCH	0.9960	0.0358	0.9870	0.3630	0.0001	0.2001	0.0699	0.0033
	HANOM	0.9998	0.8401	0.9936	0.9372	1	0.0741	0.9999	0.9622
n = 30	ANOM	1	1	0.9998	1	1	1	1	1
	ANOVA-F	1	1	0.9998	1	1	1	1	1
	BF	1	1	0.9998	1	1	1	1	1
	WELCH	0.9990	0.075	0.9850	0.6085	0.0011	0.4978	0.2138	0.0061
	HANOM	1	0.951	0.9994	0.9302	1	0.0132	1	0.9977

5. Conclusions

When considering robustness in case of the under 3 group means demonstrate that: HAOM test has the ability to control the type I error rate for all distributions for all sample sizes. ANOM test has the ability to control the type I error rate for all distributions, except the normal distribution for all sample sizes.

ANOVA-F test and Brown-Forsythe test have the ability to control the type I error rate in the beta distribution, t distribution and chi-square distribution for n=10, n=20. However, in case sample size n=30 has the ability to control the type I error rate for t distribution and chi-square distribution. Welch test has the ability to control the type I error rate for normal distribution and t distribution for all sample size.

When considering robustness in case of the under 5 group means demonstrate that: ANOM test and HAOM test have the ability to control the type I error rate for all distributions for all sample sizes.

ANOVA-F test has the ability to control the type I error rate for all distributions except the normal distribution for all sample sizes.

Brown-Forsythe test has the ability to control the type I error rate for all distributions for n=10. However, in case sample size n=20 and n=30 have the ability to control the type I error rate for beta distribution, t distribution and chi-square distribution.

Welch test no has the ability to control the type I error rate for all distributions for n=20 and n=30. But, in case n=10 has the ability to control the type I error rate for t distribution and chi-square distribution.

When considering power in case of the under 3 group means demonstrate that: ANOM test, ANOVA-F test and Brown-Forsythe test have high empirical power test all distributions for all sample sizes. HANOM has high empirical power test all distributions for all sample sizes, except chi-square distribution for n=10 lower empirical power. Welch test has high empirical power test normal distribution and t distribution for all sample sizes.

When considering power in case of the under 5 group means demonstrate that: ANOM test, ANOVA-F test and Brown-Forsythe test have high empirical power test all distributions for all sample sizes. HANOM test has high empirical power test normal distribution, t distribution and chi-square distribution for all sample sizes. Welch test has high empirical power test t distribution and chi-square distribution for n=10.

References

[1] Mendes M., Yigit S. Comparison of ANOVA-F and ANOM tests with regard to type I error rate and test power. Statistical Computation and Simulation Journal. 2013; 83(11): 2093-2104.

[2] Balamurali, Kalyanasundaram. The improved brown and forsythe test for mean equality: some things can't be fixed. 3rd ed. Canada: Taylor & Francis; 2007.

[3] Welch BL. On the comparison of several means: An alternative approach. Biometrika Journal. 1951; 34(1-2): 28-35.

[4] Nelson PR, Wludyka PS, Copeland KAF. Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions. Pennsylvania: ASA-SIAM; 2005.

[5] Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of means. Biomathematics Journal. 1974; 16(1): 129-132.

[6] Welch BL. The generalization of Student’s problem when several different population variances are involved. Biometrika Journal. 1947; 34: 28–35.

[7] Keselman HJ, Wilcox RR.The 'improved' brown and forsythe test for mean equality: some things can't be fixed. Communications in Statistics Journal. 1999; 28(3): 627-636.

[8] Oyeyemi GM., Adeleke LB. Sequential analysis of means for testing equality of several means. Statistics Journal. 2011; 99(1): 90-99.

n4Studies: Sample size calculation on a smart device

Chetta Ngamjarus[1, 2*], Virasakdi Chongsuvivatwong[3], and Edward McNeil[4]

[1]Ph.D. Student at Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110, Thailand, nchett@kku.ac.th

[2]Department of Biostatistics and Demography, Khon Kaen University, Khon Kaen, 40002, Thailand, nchett@kku.ac.th

[3]Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110, Thailand, cvirasak@gmail.com

[4]Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110, Thailand, edward.m@psu.ac.th

Abstract

Sample size should be determined before conducting a health science research study. Manual calculation needs the availability of formula. Using statistical software on a personal computer is more convenient. A smart device (phone or tablet) with appropriate application can be even more handy. In this study, we developed such an application (called “n4Studies”) for free use on iPhone and Android Operating Systems. The application can calculate the sample size and power for various epidemiological study designs including comparison of two population means and two population proportions, matched and unmatched case-control studies, cohort studies, randomized controlled trials, non-inferiority trials, equivalence trials and so on. It can be downloaded from the Apple App store and Google play store. Comparing n4Studies with other applications, n4Studies covers several more types of epidemiological study designs, gives the same result for estimation of infinite population mean from the N application, and for comparison of two independent means from WhatStat. Moreover, results are highly comparable with established PC software/packages such as STATA, R/epicalc, PS, G*Power, and OpenEpi, using the same input parameters. It is therefore useful to researchers for providing the required sample size outside the office.

Keywords: Sample size, calculation, application, smart device, smart phone, tablet

*Corresponding Author
E-mail Address: nchett@kku.ac.th

Statistical methods for modeling socioeconomic indicators affecting all-cause mortality in Thailand

Jurairat Ardkaew[1*], Kanitta Bundhamcharoen[2] and Pattapan Odton[3]

[1] *Mathematics and Statistics Program, Department of Sciences, Faculty of Science and Technology, Loei Rajabhat University, Loei, Thailand, Jurairat_p@hotmail.com*

[2] *International Health Policy Program, Ministry of public health, Nonthaburi, Thailand, kanitta@ihpp.thaigov.net*

[3] *International Health Policy Program, Ministry of public health, Nonthaburi, Thailand, podton@yahoo.com*

Abstract

This study aims to model the socioeconomic indicators affecting all-cause mortality pattern with different 18 age groups. Based on the data from Thailand's 2000 Population and Housing Census and the Vital Registration database from 1999 to 2001, this study incorporates the factor analysis into the multivariate multiple regression analysis to determine five factors: four composite factors (farming, urban, comfort and goods) derived from twenty socioeconomic indicators and one unexplained mortality factor. The result shows that the mortality rates of three predictors: farming, goods and the unexplained, are higher in all age groups except for the group with less than one year of age. The mortality due to the unexplained factor of both genders is high in northern part of the country, particularly the super-districts in Chiang Mai, Chiang Rai and Phayao. The finding provides useful guidelines for the policy makers in healthcare authorities to set up health promotion and education programs, especially for 30 to 39 years old males in urban area and 25 to 29 years old females in farming group.

Keywords: All-cause mortality, socioeconomic indicators, multivariate multiple regression model, factor analysis

*Corresponding Author
E-mail Address: Jurairat_p@hotmail.com

1. Introduction

Death rate is one of population health indices based on mortality that is an important for monitoring on human health. In Thailand, the rate remained high and fluctuated in range 6.86 to 7.72 deaths per 1,000 populations in last decade year [1]. The leading causes of death among males are stroke (9.4%); transport accidents (8.1%); HIV/AIDS (7.9%); ischemic heart diseases (6.4%); and chronic obstructive lung diseases (5.7%). Among females, the leading causes are stroke (11.3%); diabetes (8%); ischemic heart disease (7.5%); HIV/AIDS (5.7%); and renal diseases (4%) [2]. Therefore, it needs to know all-cause mortality that is useful in monitoring. The studies in this issue found that most of studies have focused on cause-specific mortality. For instance, Faramnuayphol et al [3] studied age-specific, cause-specific and all-cause mortality ratios at the district level. Next, Pattaraarchachai et al [4] studied the cause-specific mortality patterns among hospital deaths in Thailand.

In fact, there are great variations in mortality between different age groups. Also, it is likely to be affected by socioeconomic factors such as level of education, consumer goods, employment etc. Few studies have been done on all-cause mortality by considering socio-economic factors therefore we emphasize on these.

In this study, we describe the appropriate statistical method to feature the all-cause mortality pattern affecting socioeconomic indicators with different 18 age groups both male and female. The study used statistical methods that combined factor analysis and multivariate multiple regression analysis (MMR). It could be provided fundamental state information of mortality in Thailand to improve the inherent strategies for planning the effective policies of public health officer.

2. Research Methodology

2.1 Data

The vital registration of information of mortality cases in period 1999 to 2001 of Thailand was used as source of data and the number of population including socioeconomic indicators were obtained from the 2000 Population and Housing Census. Twenty socioeconomic indicators (Table 1) were selected and were calculated in the proportion of populations in region r (r was represented as a 'super-district', a district or group of contiguous districts within the same province having population approximately 200,000 persons; i.e. 235 super-districts in Thailand).

2.2 Analyses

In this study we used two main statistical methods, which mentioned as follows.

First, factor analysis is performed on the socioeconomic indicators with the aim of substantially reducing correlations between them that could mask their associations with the outcome variables. Each factor identifies correlated groups of variables.

The number of factors selected was based on obtaining an acceptable statistical fit using the chi-squared test, and these factors were fitted using maximum likelihood with promax rotation in preference to varimax, which requires the rotation to be orthogonal [5-6]. It revealed four factors; farming factor, urban factor, comfort factor, and goods factor (see Table 1) and the resulting weights are called "loadings". We calculate the new weights from loadings with the criteria that were adapted from a Likert rating scale as follows,

a) If loading < 0.4 weight = 0;

b) If loading in (0.4, 0.75) weight = 1;

c) If loading > 0.75 weight = 2;

d) If a variable has loading > 0.4 for 2 factors put it in the factor with the larger loading and increase the weight = 2 for it in that factor.

The new weights of four composite factors of socioeconomic variables were transformed into Z-scores (standardized to a mean of 0 and a standard deviation of 1).

Second, we do multivariate multiple regression (MMR) using predictors obtained by the four new weights and one unexplained mortality variable. This method is used to evaluate the effects of multiple predictor variables on multiple response variables (mortality rate in 18 age groups).

From the Multivariate Linear Regression Model:

$$Y = X\,B + \mathcal{E}, \qquad (1)$$

Where Y is the matrix of outcome variables, Y = $f(m_{rx})$ = $\log(m_{rx})$, and columns represents n_a = 18 age groups (0,1-4,5-9,…, and 80-84) which rows represents n_r = 235 super-districts (region),

B is the $(p+2)\times n_a$ matrix of parameters (a_x, b_{jx}).

X is the matrix with n_r rows.

$\mathcal{E}$ is an error.

There are p+2 columns in X matrix, the first column contains 1's and the next p+1 columns are $\sum_{j=1}^{p+1} g_{rj}$ matrix.

The $\sum_{j=1}^{p+1} g_{rj}$ matrix contains p+1 columns.

The first p columns contains the composite factors of socio-economic predictors, and the last column contains the unexplained explanatory variable, $\overline{h}_r$, where $\overline{h}_r$ obtains from the least-squares fit. The product of B and $\overline{h}_r$ is the error ($\mathcal{E}$) in equation (1).

Since the last column of X matrix is the error, equation (1) can be re-written as follows:

$$f(m_{rx}) = a_x + \sum_{j=1}^{p+1} b_{jx} g_{rj}\ , \qquad (2)$$

where $f(m_{rx})$ is a function of the mortality incidence rate and the subscripts x and r denote age-group and region, respectively and $g_{r,p+1}$ = $\overline{h}_r$ (an explanatory variable encapsulating the unexplained information on how mortality varies with region).

The model fit assessed by using the set of r-squared values [7] for the response variables to see how much of the variation in each is accounted for the model. The method also provides standard errors for each of the $j \times x$ regression coefficients thus providing p-values for testing their statistical significance after appropriate allowance for multiple hypothesis testing. The multivariate analysis of variance (MANOVA) decomposition is also used to assess the overall association between each predictor and the set of outcomes [8-9]. All statistical analyses and graphs were carried out using the R program [10].

3. Research Results and Discussion

3.1 Factor analysis

The result of factor analysis was illustrated in Table 1, the loadings greater in magnitude than 0.4 are considered; the four factors contain any overlapping of socioeconomic variables. The composite factor 1: *Farming group* includes positive loadings, such as the proportion of households with bicycle(s), the proportion of population who are employed (aged 13+), the proportion of households that have agricultural machine(s), the proportion of households that have tractor(s), the proportion of population in agricultural field and the proportion of population (aged 15+) graduated from secondary school or lower. The composite factor 2: *Urban group* includes five variables: the proportion of households without motorcycle, the population density, the proportion of population living in urban area, the proportion of households with tap water and the proportion of households with air-conditioner(s). The composite factor 3: *Comfort group,* comprises positive loadings, such as the proportion of households with television(s), the proportion of households with fan(s), proportion of households with durable construction materials, and the proportion of households with toilet(s). The composite factor 4: *Goods group,* contains positive loadings, such as the proportion of households with car(s), the proportion of households with washing machine(s), the proportion of households radio(s), the proportion of households with refrigerator(s) and the proportion of telephone(s).

The four composite socioeconomic factors accounted for the variances of 17.6%, 17.2%, 16.5% and 13.4% respectively, and the cumulative variance of 64.8%.

3.2 Multivariate multiple regression analysis

The coefficients and standard errors from fitting multivariate multiple regression models of male and female are shown in Table 2. More understanding, the 95 % confidence interval (CIs) of regression coefficient of MMR as appears in Figure 1. In addition, a range map of unexplained mortality of 235 super-districts also appears in Figure 2.

Table 2 shows the corresponding individual regression coefficients, standard errors and r-squared values for each age group after fitting the MMR model with all four composite socioeconomic factors and the unexplained mortality predictors included. The corresponding result for the models containing all predictors were statistically significant at 5% in MANOVA (Table 3). Figure 1 shows 95% confidence interval of regression coefficient of male on the left panel and female on the right panel. Both models show higher mortality rate in 3 characteristic groups (farming, goods and unexplained) of 17 age groups except children aged less than one year. Other two groups indicated lower mortality rate except children aged less than one year of urban group. For male, the highest mortality of four composite socioeconomic factors found in age group 35 to 39 year old of goods group whereas age group 25 to 29 of farming group of female indicated the highest mortality. The results, both models clearly indicate that farming and goods group in age group between 20-24 and 45-49 year old have higher mortality rates. In Thailand, framing occupations provide a livelihood for more than 70 per cent of the population. Hence, it is to be expected that the high rates will show in the agriculture group, especially in countries that rely heavily on manpower for production. Particularly occur in farming group that may take more risk from the ill-effects of pesticides that may cause of death [11].

Figure 2 shows range map of unexplained mortality in 235 super-district of male on the left panel and female on the right panel. The highest range groups to the lowest range groups of mortality represent by dark shade to right shade (brown to yellow). The highest range groups mainly occur in super-district in the northern of Thailand especially in Chiang Mai, Chiang Rai and Phayao provinces which are be also in the top 10 list of super-district of both gender as shown in table 4. The highest group of unexplained mortality in super-district of Chiang Mai, Chiang Rai and Phayao provinces in this period may be affect from HIV/AIDS according to the report of the ministry of public health [12].

Table 1: Factor analysis of categorize of socioeconomic variables (with loadings below 0.4 are not shown)

Socioeconomic Variables	Factor loadings			
	Farming	Urban	Comfort	Goods
Proportion of households with bicycle(s)	0.89			
Proportion of population who are employed (aged 13+)	0.77			
Proportion of households that have agricultural machine(s)	0.77			
Proportion of households that have tractor(s)	0.71			
Proportion of population in agricultural field	0.56			
Proportion of population (aged 15+) graduating secondary school or lower.	0.54			
Proportion of households without motorcycle		0.99		
Population density (1000 persons/sq.km.)		0.85		
Proportion of population living in urban area		0.59		
Proportion of households with tap water		0.72		
Proportion of households with air-conditioner(s)		0.65		0.40
Proportion of households with television(s)			1.09	
Proportion of households with fan(s)			0.90	
Proportion of households with durable construction materials			0.83	
Proportion of households with toilet(s)			0.69	
Proportion of households with car(s)				0.97
Proportion of households with washing machine(s)				0.65
Proportion of households radio(s)				0.53
Proportion of households with refrigerator(s)				0.47
Proportion of households that have telephone(s)		0.41		0.44
% Total variance	17.6	17.2	16.5	13.4
% Cumulative variance	17.6	34.8	51.4	64.8

Table 2: Coefficients and standard errors from fitting multivariate multiple regression models of male and female (omitted coefficient not significant at 0.05 of significant level)

Age	(Intercept)	Farming	Urban	Comfort	Goods	Unexplained	R^2
Male							
0	1.80(0.03)	-	0.08(0.04)	-0.11(0.04)	-	0.11(0.03)	0.17
1-4	0.25(0.02)	-	-0.16(0.03)	-0.06(0.03)	-	0.10(0.02)	0.24
5-9	-0.40(0.01)	0.13(0.02)	-0.13(0.02)	-0.07(0.02)	0.08(0.03)	0.23(0.01)	0.63
10-14	-0.73(0.02)	-	-0.13(0.03)	-0.06(0.03)	0.12(0.04)	0.05(0.02)	0.14
15-19	0.50(0.01)	0.06(0.02)	-0.15(0.02)	-0.04(0.02)	-	0.10(0.01)	0.54
20-24	1.05(0.01)	0.16(0.02)	-0.12(0.01)	-0.04(0.02)	-	0.16(0.01)	0.74
25-29	1.66(0.01)	0.19(0.02)	-0.17(0.02)	-0.07(0.02)	0.10(0.02)	0.29(0.01)	0.81
30-34	1.83(0.01)	0.18(0.02)	-0.16(0.01)	-0.08(0.01)	0.18(0.02)	0.30(0.01)	0.85
35-39	1.73(0.01)	0.13(0.01)	-0.10(0.01)	-0.08(0.01)	0.21(0.02)	0.25(0.01)	0.85
40-44	1.71(0.01)	0.12(0.01)	-0.07(0.01)	-0.06(0.01)	0.20(0.01)	0.20(0.01)	0.83
45-49	1.89(0.01)	0.13(0.01)	-0.07(0.01)	-0.06(0.01)	0.18(0.02)	0.16(0.01)	0.72
50-54	2.16(0.01)	0.11(0.02)	-0.06(0.01)	-	0.13(0.02)	0.14(0.01)	0.55
55-59	2.44(0.01)	0.07(0.02)	-0.04(0.01)	-	0.11(0.02)	0.12(0.01)	0.37
60-64	2.83(0.01)	0.05(0.02)	-0.04(0.01)	-	0.09(0.02)	0.10(0.01)	0.33
65-69	3.22(0.01)	0.05(0.02)	-	-	0.13(0.02)	0.07(0.01)	0.30
70-74	3.60(0.01)	0.05(0.02)	-0.05(0.01)	-	0.10(0.02)	0.08(0.01)	0.27
75-79	4.06(0.01)	-	-0.06(0.01)	-	0.12(0.02)	0.06(0.01)	0.28
80-84	4.49(0.01)	-	-0.04(0.01)	-	0.08(0.02)	0.07(0.01)	0.26
Female							
0	1.62(0.03)	-	0.13(0.04)	-0.12(0.04)	0.14(0.06)	0.07(0.03)	0.19
1-4	0.10(0.02)	0.11(0.04)	-0.15(0.03)	-0.08(0.03)	0.19(0.05)	0.09(0.02)	0.18
5-9	-0.73(0.02)	0.19(0.03)	-0.17(0.02)	-0.13(0.02)	0.17(0.04)	0.24(0.02)	0.57
10-14	-1.15(0.03)	-	-0.25(0.03)	-	0.12(0.06)	0.10(0.03)	0.27
15-19	-0.59(0.02)	0.16(0.03)	-0.17(0.02)	-0.11(0.02)	0.09(0.04)	0.13(0.02)	0.48
20-24	0.17(0.01)	0.30(0.03)	-0.17(0.02)	-0.15(0.02)	0.11(0.03)	0.33(0.01)	0.81
25-29	0.69(0.01)	0.35(0.02)	-0.21(0.02)	-0.15(0.02)	0.20(0.03)	0.43(0.01)	0.89
30-34	0.67(0.01)	0.26(0.02)	-0.18(0.01)	-0.15(0.01)	0.19(0.02)	0.35(0.01)	0.88
35-39	0.67(0.01)	0.18(0.01)	-0.16(0.01)	-0.11(0.01)	0.18(0.02)	0.26(0.01)	0.88
40-44	0.85(0.01)	0.17(0.02)	-0.11(0.01)	-0.09(0.01)	0.17(0.02)	0.19(0.01)	0.75
45-49	1.21(0.01)	0.16(0.02)	-0.10(0.01)	-0.08(0.01)	0.16(0.02)	0.16(0.01)	0.67
50-54	1.57(0.01)	0.17(0.02)	-0.09(0.01)	-	0.12(0.02)	0.15(0.01)	0.57
55-59	1.93(0.01)	0.16(0.02)	-0.06(0.02)	-0.05(0.02)	0.14(0.03)	0.15(0.01)	0.47
60-64	2.36(0.01)	0.15(0.02)	-0.06(0.02)	-0.05(0.02)	0.14(0.03)	0.13(0.01)	0.39
65-69	2.82(0.01)	0.13(0.02)	-0.06(0.02)	-0.06(0.02)	0.17(0.03)	0.12(0.02)	0.33
70-74	3.25(0.01)	0.13(0.02)	-0.06(0.02)	-0.07(0.02)	0.15(0.03)	0.11(0.01)	0.32
75-79	3.79(0.01)	0.10(0.02)	-0.04(0.02)	-0.06(0.02)	0.15(0.03)	0.09(0.01)	0.29
80-84	4.31(0.01)	0.08(0.02)	-0.06(0.01)	-0.03(0.01)	0.11(0.02)	0.09(0.01)	0.35

Table 3: MANOVA decomposition for multivariate multiple regression model with five predictors of male and female

Source of Variation	Df	Pillai	Approx F	Df (num)	Df (denom)	p-value
Male						
Intercept	1	1	425098*	18	212	0.000
F1	1	1	680*	18	212	0.000
F2	1	1	573*	18	212	0.000
F3	1	0.27	4*	18	212	0.000
F4	1	1	514*	18	212	0.000
Unexplained	1	1	4879*	18	212	0.000
Female						
Intercept	1	1	62811*	18	212	0.000
F1	1	1	1162*	18	212	0.000
F2	1	1	496*	18	212	0.000
F3	1	1	69*	18	212	0.000
F4	1	1	310*	18	212	0.000
Unexplained	1	1	2843*	18	212	0.000

* p-value < 0.05

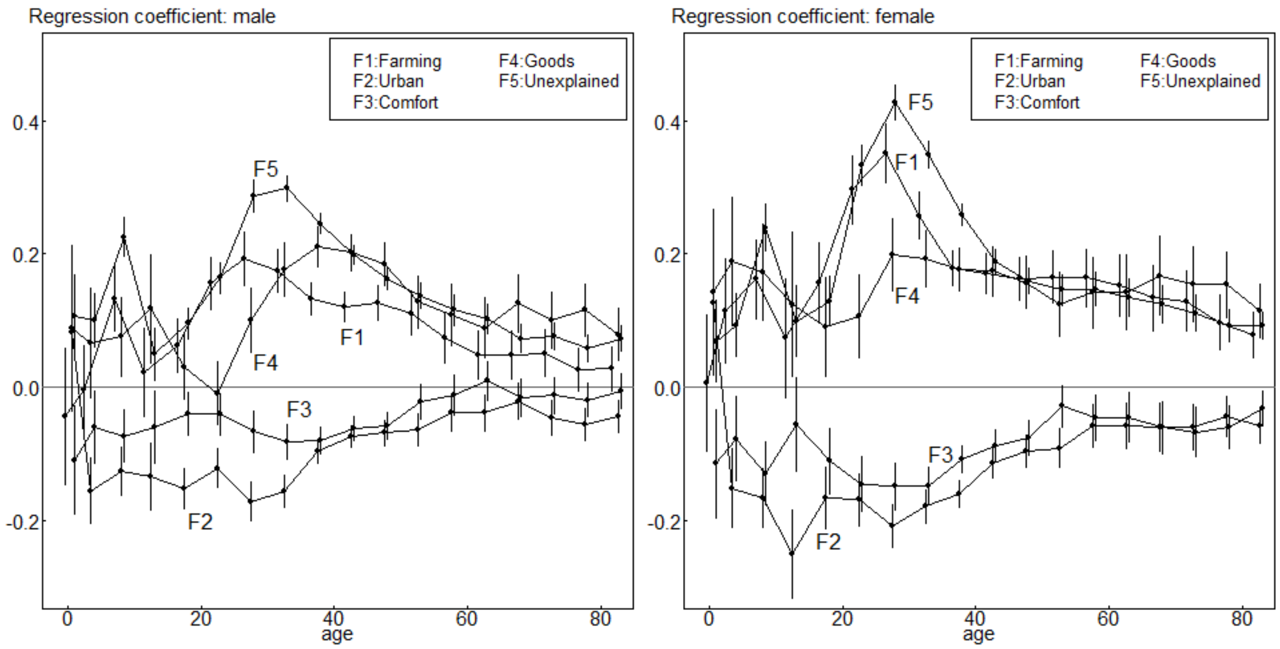


Figure 1: 95% confidence interval of regression coefficient of male and female models

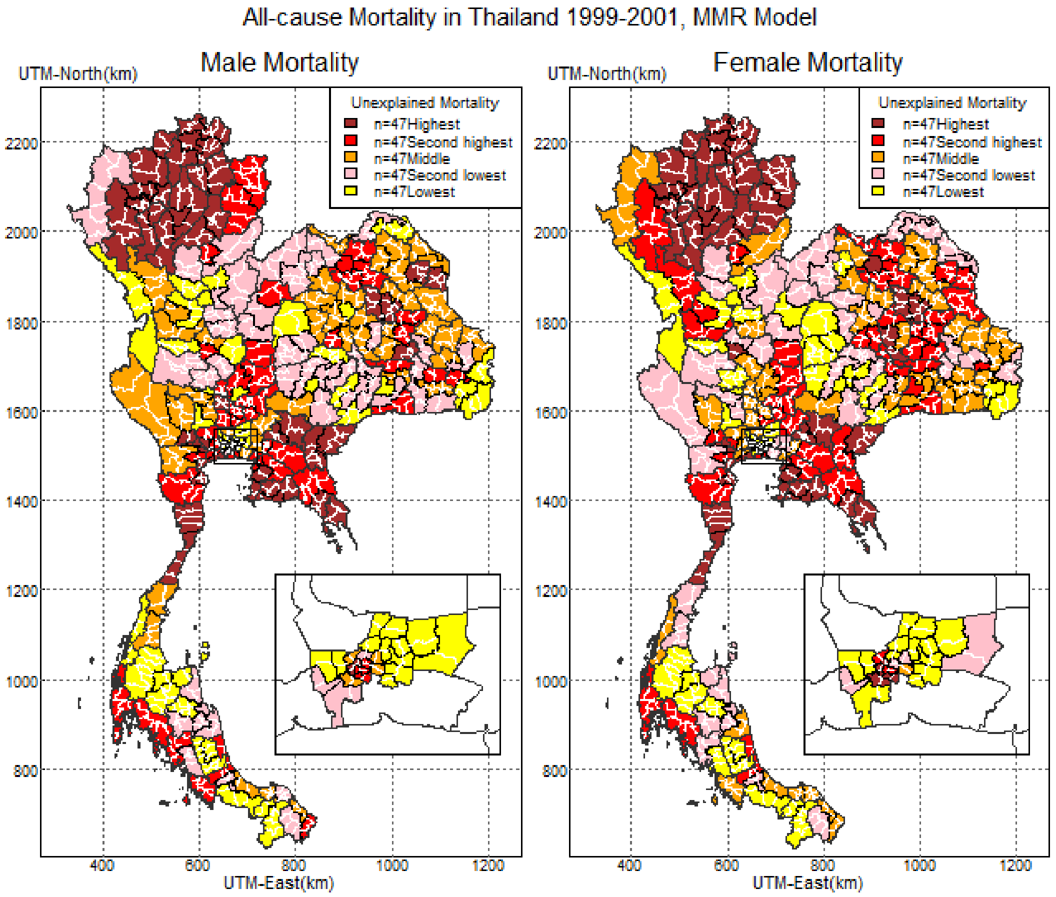


Figure 2: Range maps of unexplained mortality in super-district of male and female

Table 4: Top 10 lists of unexplained mortality in 235 super-districts in Thailand of male and female

Rank	Male		Female	
	Super-district	Province	Super-district	Province
1	Phan, MaeSa-ruai, WiangPaPao	Chiang Rai	Phan, MaeSa-ruai, WiangPaPao	Chiang Rai
2	SanPaTong, HangDong, MaeWang, DoiLo	Chiang Mai	WiangChai, Thoeng, PaDaet, PhayaMengrai, KhunTan	Chiang Rai
3	MaeRim, Samoeng, SanSai	Chiang Mai	Chun, ChiangKham, ChiangMuan, Pong, PhuSang	Phayao
4	ChiangDao, MaeTaeng, Phrao, WiangHaeng	Chiang Mai	MuangPhayao, DokKhamTai, MaeChai, PhuKamYao	Phayao
5	Fang, MaeAi, ChaiPrakarn	Chiang Mai	MuangChiangRai, MaeLao	Chiang Rai
6	MuangLamphun, MaeTha, BanThi	Lamphun	ChiangDao, MaeTaeng, Phrao, WiangHaeng	Chiang Mai
7	WiangChai, Thoeng, PaDaet, PhayaMengrai, KhunTan	Chiang Rai	SanPaTong, HangDong, MaeWang, DoiLo	Chiang Mai
8	MuangPhayao, DokKhamTai, MaeChai, PhuKamYao	Phayao	Fang, MaeAi, ChaiPrakarn	Chiang Mai
9	DoiSaket, SanKamphaeng, Saraphi, MaeOn	Chiang Mai	MaeRim, Samoeng, SanSai	Chiang Mai
10	MuangChiangRai, MaeLao	Chiang Rai	ChiangKhong, ChiangSaen, WiengKaen, WiengChiangRung, DoiLuang	Chiang Rai

4. Conclusions

The aim of this study was to model the all-cause mortality patterns from socioeconomic indicators with different 18 age groups by gender. To handle groups of correlated predictors by single variables, factor analysis was used to remove correlations between socioeconomic parameters that effect on the mortality rate.

Multivariate multiple regression analysis was used to examine the relations between the all causes mortality rate and the composite of socioeconomic predictors including unexplained mortality. Each of these five predictors was found to be affected by the all-cause mortality, the corresponding MANOVA decomposition found all of them to be statistically significant.

The both MMR models containing all five predictors give six associations between the composite socioeconomic factors and mortality rates that are highly statistically significant (p-value < 0.05) and high r-square in age group between 20 to 24 and 45 to 49 year old.

There are some limitations in our study. It is based on secondary data and we could not incorporate recent data of socioeconomic indicators data that obtained from the 2010 Population and Housing Census of Thailand, which is under development.

In conclusion, the finding provides the preliminary information for the Ministry of Public Health for monitoring and planning the policy by putting preventive measures in place for the demographic group at risk. Addressing these issues needs to be prioritized including further research into in person's morbidity and also into interventions to reduce death rate. In addition, the influence of existing and future policy needs to be evaluated carefully for potential effects on the determinants of people's health and mortality. Moreover, the health care units need to be set up in order to provide more health promotion and education programs in Thailand.

Acknowledgements

We would like to express our sincere gratitude to Prof. Dr. Don McNeil. We also would like to thank the Ministry of Public Health and the National Statistical Office, Ministry of Information and Communication Technology, for allowing us to use the data. This study was funded by the International Health Policy Program, Ministry of Public Health.

References

[1] Central Intelligence agency [Internet]. 2014 [update 2014 Jan 15; cited 2014 Mar 10] Available from: https://www.cia.gov/library/publications/the-world-factbook/fields/2066.html

[2] Rao C, Porapakkham Y, Pattaraarchachai J, Polprasert W, Swampunyalert N, Lopez AD. Verifying causes of death in Thailand: rationale and methods for empirical investigation. Population Health Metrics. 2010; 8-11.

[3] Faramnuayphol P, Chongsuvivatwong V, Pannarunothai S. Geographical variation of mortality in Thailand. Journal of the Medical Association of Thailand. 2008; 91: 1455-1460.

[4] Pattaraarchachai J, Rao C, Polprasert W, Porapakkham Y, Pao-in W, Singwerathum N, Lopez AD. Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. Population Health Metrics. 2010; 8-12.

[5] Browne MW. An overview analytic rotation in exploratory factor analysis. Multivariate Behavioral Research; 2001; 36: 111-150.

[6] Abdi H. Factor rotations in factor analysis. In B. Lewis-Beck and Futing (Eds.), Encyclopedia of Social Sciences Research Methods; 2003; 978–982.

[7] Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer Verlag; 2002.

[8] Olson CL. On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin. 1976; 83, 579-586.

[9] Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, 4th ed. Taxus: Prentice-Hall; 1998.

[10] R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.

[11]Global healing center. Effect from pesticides [internet] 2012 [cited 2013 Jan 15]. Available from: http://www.globalhealingcenter.com/effects-of-pesticides/effects-of-pesticides.

[12]Wibulpolprasert S, Gajeena A, Wattanamano S, Taverat R, Molee J, Ukachoke K. Thailand health profile 1999-2000. Bangkok: Printing Press Express Transportation Organization; 2002.

Comparing robust properties of D-optimal exact designs and weighted D-optimal exact designs in mixture experiments

Wanida Limmun[1*], John J. Borkowski[2] and Boonorm Chomteecirst [3]

[1]*Department of Mathematics and Statistics, Walailak University, Thasala, Nakhon Si Thammarat, Thailand*

[2]*Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA*

[3]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok, Thailand*

Abstract

We examine robust properties of D-optimal exact designs and weighted D-optimal exact designs in mixture experiments when the experimental region is an irregularly shaped polyhedral region. A genetic algorithm (GA) is used for generating designs when the model under consideration is the Scheffé quadratic model. The GA does not require selection of points from a user-defined candidate set of mixtures and allows movement through a continuous region. The D-optimal exact designs are based on optimizing the D-criterion while the weighted D-optimal exact designs are based on the optimization of the weighted D-criterion over the continuous region. The use of graphical approaches has been explored extensively for comparing the quality of prediction over the design space. We adopt fraction of design space (FDS) plots for comparing designs across a set of all possible reduced models when the full model is the Scheffé quadratic model. The FDS plots show the distributions of the scaled prediction variance (SPV) over the region of interest for different models on the same graph. Illustrating examples will be provided. The results indicate that the weighted D-optimal exact designs appear to be more robust than the D-optimal exact designs.

Keywords: Mixture experiments, genetic algorithm, D-efficiency, exact design, fraction of design space plots

*Corresponding Author
E-mail Address: wanida.waa@gmail.com

Statistical analysis of RAPD results in understanding genetic relationship of four fish species of the genus *Cyclocheilichthys* (family Cyprinidae)

Theerachat Kampaengsri[1], Anupong Sukjai[2], Preeya Puangsomlee Wangsomnuk[3] and Pornpimol Jearranaiprepame[4]*

[1]*Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand, june.thk39@gmail.com*

[2]*Department of Biology, Khon Kaen University, Khon Kaen, 40002, Thailand, sanupongia@gmail.com*

[3]*Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand, preeyakku@gmail.com*

[4]*Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand, pimolsingnoy@gmail.com*

Abstract

Statistical methodologies have been used in analyzing molecular information to estimate the diversity of organism including fish. In the present study Random amplified polymorphic DNA (RAPD), a technique with single primer of arbitrary nucleotide sequence, is applied to identify and determine genetic diversity of fish species. A total of thirty-eight fish from nine populations of four *Cyclocheilichthys* fish in northeast of Thailand was examined, including one population of *Cyclocheilichthys enoplos,* two populations of *C. armatus,* four populations of *C. apogon* and two populations of *C. repasson*. Shannon's information index, polymorphism percentage and Fisher's Exact tests were employed to examine genetic diversity and discriminate the fish species and population levels. The results showed that there were high value in both of genetic materials and percentage of polymorphism. The higher number of percentages of polymorphism indicated the more genetic diversity. The Shannon's information index within species ranged from 0.0538 ± 0.1663 to 0.0893 ± 0.2161 in *C. apogon* and *C. enoplos*, respectively, whereas the range within populations were from 0.0254 ± 0.1218 to 0.0893 ± 0.2161 in *C. armatus* from Ubolratana Dam, Khon Kaen (Car01) and *C. enoplos* from Chi River, Khon Kaen (Cen01), respectively. Percentage of polymorphism was congruent with the index, ranged from 12.61 to 15.13 within species and 4.20 to 15.13 within populations. The results from Fisher's Exact tests confirmed that the genetic difference between species was significant with the value of P=0.000, whereas among populations was not significant with the value of P=1.000.

Keywords: Statistical analysis, Genus *Cyclocheilichthys***,** RAPD, genetic relationship

*Corresponding Author
E-mail Address: pimolsingnoy@gmail.com

AUTHOR INDEX

Sponsors

Minitab™

QUALITY. ANALYSIS. RESULTS.

Booknet