STATISTICS

APPLIED STATISTICAL TECHNIQUES AND STOCHASTIC MODELLING

Proceedings of National Seminar in Applied Statistics and Symposium on Stochastic Modelling, NSASSSM-2020

Edited by

NICY SEBASTIAN, SAJESH T. A.

Sponsored By

DST, UGC & Mathematical and Statistical Sciences Trust.

Published by Publications Division, St.Thomas' College Thrissur

APPLIED STATISTICAL TECHNIQUES AND STOCHASTIC MODELLING

Proceedings of National Seminar in Applied Statistics and Symposium on Stochastic Modelling, NSASSSM-2020

(4-6 February 2020)

Edited by

Nicy Sebastian, Sajesh T A

Sponsored by

DST, UGC & Mathematical and Statistical Sciences Trust

Organized by

Research and Post Graduate Department of Statistics & B.Voc. Data Science, St. Thomas' College, Thrissur

As part of centenary celebrations of the college

&

85^{th} ***Birth Anniversary of Professor A. M. Mathai***

A Publication of

Publications Division

St.Thomas' College (Autonomous)

Aided by Govt. of Kerala & Affiliated to University of Calicut

Thrissur, Kerala-680001, INDIA

E-mail : stcthrissur@gmail.com

Phone : +91 487 2420435

EDITORIAL BOARD

Language: English, Year: 2020, ISBN 978-81-944730-1-5
Publisher: Publications Division, St. Thomas' College (Autonomous), Thrissur.

PREFACE

This book consists of 17 papers which were presented at the National Seminar in Applied Statistics and Symposium on Stochastic Modelling held during February 4-6, 2020 at St. Thomas College (Autonomous), Thrissur, Kerala as part of centenary celebrations of the college & 85th birth anniversary of Professor A. M. Mathai. Topics to be covered in this book are Distribution Theory, Special functions, Multivariate Analysis, Reliability and Survival Analysis, Classical and Bayesian Inference, Biostatistics and Time Series and Stochastic Modelling. This conference, comprised of around 44 talks regarding various areas of applied statistics and stochastic modelling, was attended by about 150 participants.

St. Thomas College, Thrissur is one of the leading academic institutions in the higher education sector of Kerala since 1919. It has a long and proud tradition of excellence in training, teaching and research in many academic disciplines of Science, Arts, Commerce and Humanities.

Professor A. M. Mathai is an Emeritus Professor (Full Professor) of Mathematics and Statistics, McGill University, Montreal, Canada, Director, Centre for Mathematical and Statistical Sciences, Kerala, India and Former President of the Indian Mathematical Society and Chairman of the Kerala State Statistical Commission. Dr A. M. Mathai was honored by several institutions and national societies in India and thrice by the United Nations. By 2020 he has finished nearly 60 years of his teaching career and nearly 54 years of research career. He has published over 300 research papers, most of them in SCI journals, and in 37 books he is the sole author or principal co-author. He has edited another 35 books also, including two for the United Nations. He has over 28,000 citations by 2019 as per Google counts, possibly the most cited mathematician/statistician in India and one of the top cited in the world.

The Organizers of NSASSSM-2020 are very much grateful to University Grants Commission (UGC), Science and Engineering Research Board, Department of Science and Technology (DST) and Mathematical and Statistical Sciences Trust for giving financial assistance. We gratefully acknowledge the support extended by these agencies for the successful conduct of the conference.

All the papers submitted for publication in this book were refereed rigorously. The help offered in this regard by numerous referees is gratefully acknowledged. We honestly hope that the readers of this book will find the papers to be useful and of interest. We also thank all the authors for submitting their papers for publication in this book. Special thanks goes to our Advisory Committee, National Organizing Committee and Local Organizing Committee for their support and involvement in organising this seminar and for helping with the final production of this book. Special thanks to research scholars, teaching and

non-teaching staff, graduate and post graduate students of Department of Statistics and B. Voc. Data Science, for the excellent cooperation rendered in bringing out this conference.

The editors are not responsible for the correctness and originality of the results in various papers. The responsibility lies solely with the authors of the respective papers.

Thrissur
25^{th} April 2020

Nicy Sebastian
Sajesh T A

CONTENTS

	Preface	iii
	Contents	v
1	Lectures On Factor Analysis: A New Look, A. M. Mathai	1 - 22
2	Some Unbiased Estimations Under Random Censoring For Gompertz Inliers Model, K. Muralidharan and Bavagosai Pratima	23 - 34
3	q-Gompertz Distribution And Its Applications, A. Mohammed Shabeer and K. Jayakumar	35 - 48
4	Generalized Polya-Aeppli Process and Applications, K. K. Jose and Shalitha Jacob	49 - 64
5	Estimation of Stress-Strength Reliability Function Using Weibull Distribution, Krishnendu K., Drisya M. and Joby K. Jose	65 - 78
6	Product Development: A Case Study, E. V. Gijo and U. H. Acharya	79 - 90
7	Correlation Coefficients and its Misinterpretation, Sudheesh K. Kattumannil	91 - 96
8	Some Properties and Applications of Burr III-Weibull Distribution, Deepthy G. S., Nicy Sebastian and Reshma Rison	97 - 112
9	HIV Replication- INAR(1) Model- Comparison Between MLE And CLS Estimates, Mary George K., K. Kalyanaraman and G. Meenakshi	113 - 124

10 A Study on Some Generalizations of Gompertz Distribution and Gompertz-Makeham Distribution, Rani Sebastian and Aleesha Sebastian 125 - 154

11 Prevalence and Risk Factors of Non-Alcoholic Steatohepatitis and Cardiovascular disease in Type 2 Diabetes Mellitus Childbearing Women, Ramya M. S. and Prasanth C. B. 155 - 164

12 Robust Estimation For Multivariate Zero-Inflated Poisson Regression, Jeena Joseph and Pinky Peter 165 - 182

13 Compound Extended Gamma Density and Connection to Inverse Gaussian, Dhannya P. Joseph 183 - 192

14 Robust Multivariate Analysis of Variance, Sajesh T. A. and Arya Damodharan 193 - 206

15 Generalized Gamma Model and Pathway Fractional Operator, Seema S. Nair 207 - 216

16 Three Parameter Quasi Lindley Distribution, Ashlin Mathew P. M. and Sneha Robinson 217 - 232

17 Quasi-Bayesian Estimation of Time to Test Transform in Lomax Model, Sowbhagya S. Prabhu and E. S. Jeevanand 233 - 240

Author Index 241

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 01-22.

Lectures On Factor Analysis: A New Look

A.M. Mathai
Emeritus Professor of Mathematics and Statistics, McGill University, Canada
Email:directorcms458@gmail.com

Abstract

The topic of Factor Analysis in Multivariate Statistical Analysis is usually a difficult topic for students to understand, to interpret and to see the significance of the various procedures and approximations. This article is based on the lectures of the author. In this article this area is examined through Jacobians of matrix transformations and some properties of special functions of matrix argument. This approach makes the derivations easy and to some extend interpretable also. All the factor loadings and error variances are estimated explicitly and testing of hypothesis is also done under one set of identification conditions.

AMS Subject Classification: 62H25, 62H12, 62H15.

Key words: Factor analysis, principal components analysis, linear models, estimation and test of hypotheses.

1. Introduction

The following notations will be used in this article. Small letters x, y, z etc will denote real scalar variables whether mathematical variables or random variables. Capital letters X, Y etc will be used for real vector/matrix-variate variables whether square or rectangular matrix is involved. A tilde will be used for the variables in the complex domain such as $\tilde{x}, \tilde{y}, \tilde{X}, \tilde{Y}$ etc. Constant vectors/matrices will be denoted by A, B, C etc. A tilde will not be used for constant matrices unless the point is to be stressed that the constant matrix is in the complex domain. In the real case, determinant of a square matrix A will be denoted by $|A|$ or $\det(A)$ and in the complex case the absolute value of the determinant of A will be denoted as $|\det(A)|$. When matrices are square then their order will be taken as

$p \times p$ unless specified otherwise. When A is of full rank in the complex domain then AA^* is Hermitian positive definite where A^* indicates complex conjugate transpose of A. Also, $\mathrm{d}X$ will indicate the wedge product of all distinct differentials in the matrix X. If $X = (x_{ij})$, $p \times q$ and of distinct real scalar variables x_{ij}'s then $\mathrm{d}X = \wedge_{i=1}^{p} \wedge_{j=1}^{q} \mathrm{d}x_{ij}$. For the complex matrix $\tilde{X} = X_1 + iX_2, i = \sqrt{(-1)}$, where X_1 and X_2 are real, then $\mathrm{d}\tilde{X} = \mathrm{d}X_1 \wedge \mathrm{d}X_2$.

Factor Analysis was originally developed in connection with mental tests. A score available in a mental test is contributed by many mental faculties or mental factors. It is assumed that a certain linear function of these contributions from the various mental factors is producing the final score. Hence there is a parallel to linear regression models and analysis of variance or design of experiment models. In order to introduce the topic of Factor Analysis we will examine a linear regression model and a design of experiment model first.

Since the primary audience is expected to be students, the material will be written in the form of a lecture note. Those who are already exposed to the preliminaries may delete Sections 1 and 2 and go directly to Section 3.

1.1. Linear Models from Different Disciplines

(a) A linear regression model

Let x be a real scalar random variable and let $t_1, ..., t_r$ be either r fixed numbers or given values of r random variables. Let the conditional expectation of x, given $t_1, ..., t_r$, be of the form

$$E[x|t_1, ..., t_r] = a_o + a_1 t_1 + ... + a_r t_r$$

or the corresponding model be

$$x = a_o + a_1 t_1 + ... + a_r t_r + e$$

where $a_o, a_1, ..., a_r$ are unknown constants, $t_1, ..., t_r$ are given values and e is the error part or the sum total of contributions coming from unknown or uncontrolled factors plus experimental error. Here x may be inflation index taking a particular year, say 2010, as the base year. Here t_1 may be the change or deviation in the average price per kilogram of staple vegetable items from the base year 2010, t_2 may be the change or deviation in the average price of the staple item rice per kilogram compared to the base year 2010, t_3 may be the change or deviation in the average price of lentils per kilogram compared to the base year 2010, and so on, and t_r may be the change or deviation in the average price per kilogram of beans compared to the base year 2010. Here the notation $t_j, j = 1, ..., r$ is used to denote the given values as well as the corresponding random variables. Since we are taking deviations from the base value we may assume, without loss of generality, that the expected value of t_j is zero or $E[t_j] = 0, j = 1, ..., r$. We may also take the expected

value of the error term e to be zero or $E[e] = 0$. Let x_1 be the inflation index, x_2 be the calorie intake index per person, x_3 be the general health index and so on. In all these cases the same $t_1, ..., t_r$ can act as the independent variables in a regression set up. Then in such a case a multivariate linear regression model will have the following format:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ \vdots & \vdots & \dots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pr} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_r \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix}. \tag{1}$$

We may write this model in matrix notation as

$$X = \mu + \Lambda f + e$$

where $\Lambda = (\lambda_{ij})$ is $p \times r$, $r \leq p$ and of full rank r, e is $p \times 1$ and f is $r \times 1$. In (1), $\lambda_{ij} = a_{ij}, f_j = t_j$. Then $E[X] = \mu + \Lambda E[f] + E[e] = \mu$ since we have assumed that $E[f] = O$ (null), $E[e] = O$. Then when f and e are uncorrelated then the covariance matrix in X, denoted by $\mathrm{Cov}(X) = \Sigma$ is the following:

$$\begin{aligned} \Sigma = \mathrm{Cov}(X) &= E\{(X-\mu)(X-\mu)'\} = E\{(\Lambda f + e)(\Lambda f + e)'\} \\ &= \Lambda \mathrm{Cov}(f)\Lambda' + \mathrm{Cov}(e) + O = \Lambda\Phi\Lambda' + \Psi \\ \Sigma &= \Lambda\Phi\Lambda' + \Psi \end{aligned} \tag{2}$$

where we have denoted the covariance matrices of f and e by $\Phi > O$ (positive definite) and $\Psi > O$ respectively. In the above formulation, f is taken as a real vector random variable. In a simple linear model usually the covariance matrix of e, namely Ψ, is taken as $\sigma^2 I$ where $\sigma^2 > 0$ is a real scalar quantity and I is the identity matrix. In a more general model we take Ψ as a diagonal matrix with positive diagonal elements or we assume that the e_j's are non-correlated and the variances of the e_j's need not be equal. In (2) we will take Ψ to be a diagonal matrix with positive diagonal elements.

(b) A Design of Experiment model

Consider a completely randomized experiment where one set of treatment is tried. Here the experimental plots are assumed to be fully homogeneous with respect to all known factors of variation which may contribute towards the final observation. The observation may be the yield of a particular variety of corn from an experimental plot. Let the set of treatments be a set of r different fertilizers $F_1, ..., F_r$ where the effects of these fertilizers are denoted by $\alpha_1, ..., \alpha_r$. If no fertilizer is applied the yield from a test plot need not be zero. Let μ_1 be a general effect when F_1 is applied so that we may take α_1 as deviation from the general effect μ_1 due to F_1. Let e_1 be the sum total contributions coming from all unknown or uncontrolled factors plus experimental error, if any, when F_1 is applied. Then a simple linear one-way classification model for F_1 is

$$x_1 = \mu_1 + \alpha_1 + e_1$$

where x_1 is the yield from the test plot where F_1 is applied. Then corresponding to $F_1, ..., F_r$ we have the following:

$$\begin{aligned} x_1 &= \mu_1 + \alpha_1 + e_1 \\ \vdots \quad &= \vdots \qquad \text{or } X = \mu + \Lambda f + e \\ x_p &= \mu_p + \alpha_p + e_p \end{aligned} \tag{3}$$

where

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, e = \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix}, f = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}, \Lambda = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Here the elements of Λ are decided by the design itself. If the vector f is fixed then we call the model in (3) as the fixed effect model and if f is assumed to be random then it is called the random effect model. With one observation per cell, as stated in (3), we will not be able to estimate the parameters and test hypotheses. We will have to replicate the experiment. Then in the j-th replicate the observation vector will be

$$X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{pj} \end{bmatrix}, j = 1, ..., n; \Sigma, \Phi, \Psi$$

remaining the same for each replicate for the random effect model. In the regression model in (1) also the j-th replication or repetition vector will be X_j, same format as above, with Σ, Φ, Ψ there remaining the same for each sample.

We will consider a general linear model of the type in (1) and (3) and do a complete analysis in the sense of checking for the existence and uniqueness of such a model, estimation of the parameters and testing of various types of hypotheses. This will be called Factor Analysis.

2. A General Linear Model for Factor Analysis

Consider the following general linear model:

$$X = \mu + \Lambda f + e \tag{4}$$

where

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}, e = \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix}, f = \begin{bmatrix} f_1 \\ \vdots \\ f_r \end{bmatrix}, r \leq p$$

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2r} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pr} \end{bmatrix}$$

with μ_j's, λ_{ij}'s, f_j's being real scalar parameters, $x_j, j = 1, ..., p$ as real scalar quantities, Λ is $p \times r, r \leq p$ and of full rank r. When we talk about expected values, variances, covariance etc then X, f, e are assumed to be random quantities and when we talk about estimates then X represents a vector of observations. This convention will be used throughout the discussion in this lecture so that multiplicity of symbols for the variables and the corresponding observations can be avoided.

Geometrically speaking, the r columns of Λ are linearly independent and these r columns can determine a r-dimensional subspace in the p-dimensional Euclidean space. In this case the $r \times 1$ vector f is a point in this r-subspace and this subspace is usually called the *factor space*. Then if the $p \times r$ matrix Λ is multiplied on the right by a matrix then this will correspond to taking a new set of coordinate axes for the factor space.

Factor Analysis is a subject dealing with identification or unique determination of a model of the type in (4), estimation of parameters and testing of various types of hypotheses in (4). The subject matter was originally developed in connection with mental tests. Suppose that one test or a battery of tests is administered to an individual to evaluate the individual's reading ability or computational ability or mathematical ability or language ability etc, then the test will result in a test score. There will be a component here representing the expected score. If the test is administered among students in the 10th grade of a school then the grand average of such test scores among all 10th graders across the nation may be taken as the expected score. Then there will be contributions coming from various mental factors, or certain mental faculties or a combination of several mental faculties. All such factors may be contributing towards the observed test score. If $f_1, ..., f_r$ are the contributions coming from r such mental factors then certain linear functions of these contributions will be the final quantity entering into the observed test score when we assume a linear model. If x_1 is the test score then it may be taken as a linear model in the following form:

$$x_1 = \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + ... + \lambda_{1r} f_r + e_1$$

where $\lambda_{11}, ..., \lambda_{1r}$ are the coefficients for $f_1, ..., f_r$, where $f_1, ..., f_r$ are contributions from r factors toward x_1 and these factors may be called the *main mental factors* and the

coefficients $\lambda_{11}, ..., \lambda_{1r}$ may be called the *factor loadings* for the main factors. Here μ_1 is the general expected values and e_1 is the error part or the sum total contributions coming from all unknown factors plus experimental error, if any. Note that the contributions $f_1, ..., f_r$ coming from main mental factors can vary from individual to individual and hence it is appropriate to treat $f_1, ..., f_r$ as random variables rather than as fixed unknown quantities. These $f_1, ..., f_r$ are not observable as in the case of design model in (3) whereas in the regression type model in (1) they may be given values of observable variables called the independent variables. Analysis of the above model may be done treating $f_1, ..., f_r$ as fixed quantities or as random variables. If they are treated as random variables then we can assume some joint distribution over $f_1, ..., f_r$. Usually a joint normality is assumed for $f_1, ..., f_r$. Since $f_1, ..., f_r$ are deviations from the general effect μ_1 due to various main mental faculties we may take, without loss of generality, that the expected value as null or $E[f] = O$ (null), and we will denote the covariance matrix in f as Φ or $\mathrm{Cov}(f) = \Phi > O$ (real positive definite). Note that the error part e_j is always a random variable. Let $x_1, ..., x_p$ be the test scores on p individuals then we have the error vector $e' = (e_1, ..., e_p)$. Without loss of generality we may take the expected value here as null or $E[e] = O$ (null). For a very simple situation we may assume the covariance matrix in e as $\mathrm{Cov}(e) = \sigma^2 I$ where $\sigma^2 > 0$ is a real positive scalar quantity and I is the identity matrix. For a slightly more general situation we may assume $\mathrm{Cov}(e) = \Psi$ as a real positive definite diagonal matrix or a diagonal matrix with real positive diagonal elements. In the most general case we may take Ψ as a real positive definite matrix. In our model in (4), we will assume that Ψ is diagonal with positive diagonal elements. We will assume that f and e are uncorrelated. If Σ is the covariance matrix of X then we have the following:

$$\begin{aligned}\Sigma &= E[(X-\mu)(X-\mu)'] = E[(\Lambda f + e)(\Lambda f + e)'] \\ &= \Lambda E(ff')\Lambda' + E(ee') + 0 = \Lambda\Phi\Lambda' + \Psi \\ \Sigma &= \Lambda\Phi\Lambda' + \Psi. \end{aligned} \tag{5}$$

We will assume $\Sigma > O$ (real positive definite) and hence $\Lambda\Phi\Lambda' + \Psi > O$.

2.1. Identification problem

Is the model in (4) a unique model or can it represent different situations or can it be quite arbitrary? Does the model (4), as stated there, make sense as a model? Let A be any $r \times r$ nonsingular matrix. Let $Af = f^*$ and $\Lambda A^{-1} = \Lambda^*$. Then $\Lambda^* f^* = \Lambda A^{-1} A f = \Lambda f$. In other words,

$$X = \mu + \Lambda f + e = \mu + \Lambda^* f^* + e. \tag{6}$$

Therefore the model in (4) is not identified or not uniquely determined.

Identification problem can also be stated as follows: Given the real positive definite $p \times p$ matrix $\Sigma > O$ with $p(p+1)/2$ distinct elements, can it be uniquely represented as

$\Lambda\Phi\Lambda'+\Psi$ where Λ has pr distinct elements, $\Phi > O$ has $r(r+1)/2$ distinct elements, $\Psi =$ a diagonal matrix has p distinct elements? The answer is no as seen in (6). Note that a $r \times r$ arbitrary matrix A represents r^2 distinct elements. From (6) we have seen that we can impose r^2 conditions on the parameters in Λ, Φ, Ψ. Also we can modify the question. Can the $p(p+1)/2$ distinct elements in Σ plus the r^2 elements in A (r^2 conditions) uniquely determine all elements in Λ, Ψ, Φ? Let us see how many elements are there in total. Λ, Ψ, Φ, have a total of $pr + p + r(r+1)/2$ elements and A and Σ have a total of $r^2 + p(p+1)/2$ elements. Hence the difference, denoted by δ, is the following:

$$\delta = \frac{p(p+1)}{2} + r^2 - [pr + \frac{r(r+1)}{2} + p] = \frac{1}{2}[(p-r)^2 - (p+r)]. \tag{7}$$

Note that the right side in (5) is not a linear function of Λ, Φ, and Ψ. Hence if $\delta > 0$ then we cannot guarantee but we can anticipate existence and uniqueness, if $\delta = 0$ we can hope for existence and uniqueness and if $\delta < 0$ then we can expect existence but possibly not uniqueness. From (5) note that

$$\Sigma = \Psi + \Lambda\Phi\Lambda' \Rightarrow \Sigma - \Psi = \Lambda\Phi\Lambda'$$

where $\Lambda\Phi\Lambda'$ is positive semi-definite of rank r since Λ is $p \times r, r \le p$ and of full rank r and $\Phi > O$ (positive definite). Then the existence question can also be stated as follows: Given a $p \times p$ real positive definite matrix $\Sigma > O$, can we find a $\Psi =$ diagonal with positive diagonal elements such that $\Sigma - \Psi$ is real positive semi-definite of rank r? *If a set of parameters exist and if the model is uniquely determined then we say that the model is identified.*

If we assume $\Phi = I$ then this will impose $r(r+1)/2$ conditions. But $r^2 = \frac{r(r+1)}{2} + \frac{r(r-1)}{2}$. Hence we can impose $r(r-1)/2$ conditions more after imposing the condition $\Phi = I$. Note that for $\Phi = I$, $\Lambda^*\Phi\Lambda'^* = \Lambda^*\lambda'^* = \Lambda A^{-1}A'^{-1}\Lambda'$ and if this is equal to $\Lambda\Lambda'$ under $\Phi = I$ then this means $(A'A)^{-1} = I$ or $A'A = I$ or A is an orthonormal matrix. Under the condition $\Phi = I$ the arbitrary $r \times r$ matrix A becomes an orthonormal matrix. In this case the transformation $Y = \Lambda A$ is an orthonormal transformation or a rotation of the coordinate axes. The following $r \times r$ symmetric matrix of $r(r+1)/2$ distinct elements

$$\Delta = \Lambda'\Psi^{-1}\Lambda \tag{8}$$

is needed when we do estimation and tests of hypotheses and hence we can impose $r(r-1)/2$ conditions by requiring Δ to be diagonal with distinct diagonal elements, $\Delta = \text{diag}(\delta_1, ..., \delta_r), \delta_j > 0, j = 1, ..., r$. This imposes $\frac{r(r+1)}{2} - r = \frac{r(r-1)}{2}$ conditions. Hence for identification of the model or for the unique determination of all the parameters in Λ, Φ, Ψ we can impose the conditions $\Phi = I$ and that $\Delta = \Lambda'\Psi^{-1}\Lambda$ is diagonal with positive diagonal elements. These two conditions will provide $\frac{r(r+1)}{2} + \frac{r(r-1)}{2} = r^2$ restrictions on the model and the model will be identified.

When $\Phi = I$ the main factors are orthogonal. If $\Phi =$ a diagonal matrix (including identity matrix) the covariances are zeros and it is an orthogonal situation and in this case we say that the *main factors are orthogonal*. If Φ is not diagonal (including identity matrix) then we say that the *main factors are oblique*.

One can also impose $r(r-1)/2$ conditions on the $p \times r$ matrix Λ. Consider the first $r \times r$ block or the leading $r \times r$ sub-matrix or the upper $r \times r$ block in the $p \times r$ matrix, call it B. Impose the condition that this $r \times r$ block B is lower triangular. Then this will result in $r^2 - \frac{r(r+1)}{2} = \frac{r(r-1)}{2}$ conditions. Hence $\Phi = I$ and the condition that this leading $r \times r$ block B is lower triangular will guarantee r^2 restrictions and then the model will be identified. One can also take a preselected $r \times r$ matrix B_1 and then impose the condition that $B_1 B$ is lower triangular. This will also produce $\frac{r(r-1)}{2}$ conditions. Hence $\Phi = I$ and $B_1 B$ is lower triangular will complete the identification of the model.

When we put conditions on Φ and Ψ we are requiring the unknown covariance matrices to be of certain formats. Such conditions are justified. But can we put conditions on Λ, the factor loadings? If we say that the fist $r \times r$ block B in the $p \times r$ matrix Λ as lower triangular then we are saying that $\lambda_{12} = 0 = \lambda_{13} = ... = \lambda_{1r}$ or that $f_2, ..., f_r$ do not enter into the model for x_1, $X' = (x_1, x_2, ..., x_p)$. Such restrictions are justified if we can design the experiment in such a way that x_1 depends on f_1 alone and not on $f_2, ..., f_r$. In psychological tests it is possible to design the test in such a way that there are contributions from certain main factors only. Hence a triangular format that in x_1 there are no contributions from $f_2, ..., f_r$ or the corresponding factor loadings $\lambda_{12}, ..., \lambda_{1r}$ are zeros, or, in general, in x_i there are no contributions from $f_{i+1}, ..., f_r$, or the factor loading $\lambda_{i\ i+1}, ..., \lambda_{ir}$ are zeros for $i = 1, ..., r-1$ is a feasible proposition and hence such a condition is justified. Suppose that the first r tests are designed in such a way that x_i has only contributions from $f_1, ..., f_i$ and from no other factor or $x_i = \mu_i + \lambda_{i1} f_1 + ... + \lambda_{ii} f_i + e_i, i = 1, ..., r$, this is also a feasible proposition. We can also measure the contribution from f_i in λ_{ii} units or we can take $\lambda_{ii} = 1$. Then by taking $B = I_r$ we can impose r^2 conditions without requiring $\Phi = I$. This means that the first r tests are specifically designed so that x_1 has one unit contribution from f_1 only, x_2 has one unit contribution from f_2 only etc and x_r has one unit contribution from f_r only. If B is taken as diagonal then there are factor loadings $\lambda_{11}, \lambda_{22}, ..., \lambda_{rr}$ respectively but x_i has contribution from f_i only for $i = 1, ..., r$. Hence the following are some model identification conditions:

(1): $\Phi = I$ and $\Lambda' \Psi^{-1} \Lambda$ is diagonal with distinct diagonal elements;

(2): $\Phi = I$ and the leading $r \times r$ sub-matrix B in the $p \times r$ matrix Λ is triangular;

(3): $\Phi = I$ and $B_1 B$ is lower triangular where B_1 is a preselected matrix;

(4): The leading $r \times r$ sub-matrix B in the $p \times r$ matrix Λ is an identity matrix.

Observe that when $r = p$ then the condition (4) above will correspond to the design model

considered in (3).

2.2. Scaling or units of measurements

Any analysis based on a covariance matrix Σ has the basic defect that the covariances depend on the units of measurements of the individual variables. If the units are changed then the covariances change. If y_i and y_j are two real scalar random variables with variances σ_{ii} and σ_{jj} and covariance between them σ_{ij} then a way of getting rid of the effect of scaling or changes in units of measurements is to consider the variables $z_i = y_i/\sqrt{\sigma_{ii}}$ and $z_j = y_j/\sqrt{\sigma_{jj}}$ so that $\text{Cov}(z_i, z_j) = r_{ij}$ = correlation between y_i and y_j and the correlation is free of the units of measurements. Let $Y' = (y_1, ..., y_p)$ and let $D = \text{diag}(\frac{1}{\sqrt{\sigma_{11}}}, ..., \frac{1}{\sqrt{\sigma_{pp}}})$ then consider $Z = DY$. We note that $\text{Cov}(Y) = \Sigma \Rightarrow \text{Cov}(Z) = D\Sigma D = R$ = the correlation matrix in Y.

In psychological testing situation or in the model in (4), when a test score x_j is multiplied by a scalar quantity c_j then the factor loadings $\lambda_{j1}, ..., \lambda_{jr}$ are multiplied by c_j, the error e_j is multiplied by c_j and the general effect μ_j is multiplied by c_j or $c_j x_j = c_j\mu_j + c_j(\lambda_{j1}f_1 + ... + \lambda_{jr}f_r) + c_j e_j$. Let $\text{Cov}(x_i, x_j) = \sigma_{ij}$, $\text{Cov}(X) = \Sigma, X' = (x_1, ..., x_p)$ and let $D = \text{diag}(\frac{1}{\sqrt{\sigma_{11}}}, ..., \frac{1}{\sqrt{\sigma_{pp}}})$. Consider the model

$$DX = D\mu + D\Lambda f + De \Rightarrow D\Sigma D = D\Lambda\Phi\Lambda' D + D\Psi D. \tag{9}$$

If $X^* = DX, \mu^* = D\mu, \Lambda^* = D\Lambda, e^* = De$ then we have the following model and the resulting covariance matrix:

$$\begin{aligned} X^* = \mu^* + \Lambda^* f + e^* &\Rightarrow \Sigma^* = \text{Cov}(X^*) = \Lambda^* \text{Cov}(f)\Lambda'^* + \Psi^* \\ &\Rightarrow D\Sigma D = D\Lambda\Phi\Lambda' D + D\Psi D \\ &\Rightarrow R = \Lambda^*\Phi\Lambda'^* + \Psi^* \end{aligned} \tag{10}$$

where $R = (r_{ij})$ is the correlation matrix in X. One interesting point to be noted is that the identification condition $\Phi = I$ and $\Lambda'^*\Psi^{*-1}\Lambda^*$ = diagonal becomes the following: $\Phi = I$ and $\Lambda'^*\Psi^{*-1}\Lambda^* = \Lambda' D D^{-1}\Psi^{-1}D^{-1}D\Lambda = \Lambda'\Psi^{-1}\Lambda$ = diagonal or $\Lambda'\Psi^{-1}\Lambda$ is invariant under scaling transformation on the model or under $X^* = DX, \Psi^* = D\Psi D$.

3. Maximum Likelihood Estimators for the Parameters

From the model $X = \mu + \Lambda f + e$ in (4) when we say we have a simple random sample of size n the meaning is that we are considering independently and identically distributed

(iid) $X_j, j = 1, ..., n$ where

$$X_j = \mu + \Lambda f + e_j, j = 1, ..., n, X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}, e = \begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{pj} \end{bmatrix} \tag{11}$$

and iid. Let X_j and e_j be independently normally distributed, that is $e_j \sim N_p(O, \Psi)$ and $X_j \sim N_p(\mu, \Sigma), \Sigma = \Lambda\Phi\Lambda' + \Psi$ where $\Phi > O, \Psi > O, \Sigma > O$ where Ψ is diagonal with positive diagonal elements. Then the likelihood function is the following:

$$\begin{aligned} L &= \prod_{j=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(X_j-\mu)'\Sigma^{-1}(X_j-\mu)} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^{\frac{n}{2}}} \mathrm{e}^{-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu)'\Sigma^{-1}(X_j-\mu)}. \end{aligned} \tag{12}$$

The sample matrix is denoted by the bold-faced $\mathbf{X} = (X_1, ..., X_n)$. Let J be the $n \times 1$ vector of unities, $J' = (1, 1, ..., 1)$. Then

$$\mathbf{X} = (X_1, ..., X_n) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix}$$

$$\Rightarrow \frac{1}{n}\mathbf{X}J = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{1j}) \\ \vdots \\ \frac{1}{n}(\sum_{j=1}^{n} x_{pj}) \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \bar{X}$$

where $\bar{X}$ is the sample average vector or the sample mean vector. Let the bold-faced $\bar{\mathbf{X}}$ be the $p \times n$ matrix $\bar{\mathbf{X}} = (\bar{X}, \bar{X}, ..., \bar{X})$. Then

$$(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})' = S = (s_{ij}), s_{ij} = \sum_{k=1}^{n}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j}) \tag{13}$$

where S is the sample sum of products matrix or the "corrected" sample sum of products matrix. Note that

$$\begin{aligned} \bar{X} &= \frac{1}{n}\mathbf{X}J \\ &\Rightarrow \bar{\mathbf{X}} = (\bar{X}, ..., \bar{X}) = \mathbf{X}(\frac{1}{n}JJ') \\ &\Rightarrow \mathbf{X} - \bar{\mathbf{X}} = \mathbf{X}(I - \frac{1}{n}JJ'). \end{aligned}$$

Then

$$S = \mathbf{X}(I - \frac{1}{n}JJ')(I - \frac{1}{n}JJ')'\mathbf{X}' = \mathbf{X}(I - \frac{1}{n}JJ')\mathbf{X}'. \tag{14}$$

Since $(X_j - \mu)'\Sigma^{-1}(X_j - \mu)$ is a real scalar quantity we have the following:

$$\begin{aligned}
\sum_{j=1}^{n}(X_j - \mu)'\Sigma^{-1}(X_j - \mu) &= \sum_{j=1}^{n}\mathrm{tr}(X_j - \mu)'\Sigma^{-1}(X_j - \mu) \\
&= \sum_{j=1}^{n}\mathrm{tr}[\Sigma^{-1}(X_j - \mu)(X_j - \mu)'] \\
&= \mathrm{tr}[\Sigma^{-1}\sum_{j=1}^{n}(X_j - \bar{X} + \bar{X} - \mu)(X_j - \bar{X} + \bar{X} - \mu)'] \\
&= \mathrm{tr}[\Sigma^{-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})'] \\
&+ n\mathrm{tr}[\Sigma^{-1}(\bar{X} - \mu)(\bar{X} - \mu)'] \\
&= \mathrm{tr}(\Sigma^{-1}S) + n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu).
\end{aligned} \tag{15}$$

Hence

$$L = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}}\mathrm{e}^{-\frac{1}{2}\{\mathrm{tr}(\Sigma^{-1}S) + n(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu)\}}. \tag{16}$$

Differentiating (16) with respect to μ and equating to a null vector and then solving we have the estimator for μ, denoted by $\hat{\mu}$, as $\hat{\mu} = \bar{X}$. Then $\ln L$ at $\mu = \bar{X}$ is the following:

$$\begin{aligned}
\ln L &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}S) \\
&= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Lambda\Phi\Lambda' + \Psi| - \frac{1}{2}\mathrm{tr}[(\Lambda\Phi\Lambda' + \Psi)^{-1}S].
\end{aligned} \tag{17}$$

3.1. Maximum likelihood estimators under an identification condition

One of the conditions for identification of the model is $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda =$ a diagonal matrix with positive diagonal elements. We will examine the maximum likelihood estimators (MLE)/maximum likelihood estimates (MLE) under this identification condition. Then from (17)

$$\ln L = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Lambda\Lambda' + \Psi| - \frac{1}{2}\mathrm{tr}[(\Lambda\Lambda' + \Psi)^{-1}S]. \tag{18}$$

By opening up the following determinant in two different ways we have the following relationship:

$$\begin{vmatrix} \Psi & -\Lambda \\ \Lambda' & I_r \end{vmatrix} = |\Psi|\ |I + \Lambda'\Psi^{-1}\Lambda| = |\Psi + \Lambda\Lambda'|. \tag{19}$$

Hence

$$\begin{aligned}\ln|\Lambda\Lambda'+\Psi| &= \ln|\Psi|+\ln|I+\Lambda'\Psi^{-1}\Lambda| \\ &= \sum_{j=1}^{p}\ln\psi_{jj}+\sum_{j=1}^{r}\ln(1+\delta_j), \\ \Delta &= \Lambda'\Psi^{-1}\Lambda = \mathrm{diag}(\delta_1,\delta_2,...,\delta_r) \qquad (20)\end{aligned}$$

where $\psi_{jj}, j=1,...,p$ are the diagonal elements in the diagonal matrix Ψ, and the identification condition is that $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda = \Delta = \mathrm{diag}(\delta_1,...,\delta_r)$. Hence if we can write $\mathrm{tr}(\Sigma^{-1}S) = \mathrm{tr}[(\Lambda\Lambda'+\Psi)^{-1}S]$ in terms of $\psi_{jj}, j=1,...,p$ and $\delta_j, j=1,...,r$ then the likelihood equation can be evaluated directly from (18) and (20) and the estimators can be obtained. In this connection we can obtain the following result:

Theorem 3.1. *For $\Lambda\Lambda'+\Psi$ nonsingular, which here means real positive definite, the inverse is given by*

$$(\Lambda\Lambda'+\Psi)^{-1} = \Psi^{-1}-\Psi^{-1}\Lambda(\Delta+I)^{-1}\Lambda\Psi^{-1} \qquad (21)$$

where the Δ is given in (20).

It is easily verified that pre and post multiplications of $\Psi^{-1}-\Psi^{-1}\Lambda(\Delta+I)^{-1}\Lambda'\Psi^{-1}$ by $\Lambda\Lambda'+\Psi$ give the identity matrix I_p.

3.2. Simplifications of $|\Sigma|$ and $\mathrm{tr}(\Sigma^{-1}S)$

From (20)

$$\begin{aligned}|\Sigma| &= |\Lambda\Lambda'+\Psi| = |\Psi|\ |\Lambda'\Psi^{-1}\Lambda+I| \\ &= |\Psi|\ |I+\Delta| = \{\prod_{j=1}^{p}\psi_{jj}\}\{\prod_{j=1}^{r}(1+\delta_j)\}.\end{aligned}$$

Then observe the following: In $\Lambda(\Delta+I)^{-1} = \Lambda\mathrm{diag}(\frac{1}{1+\delta_1},...,\frac{1}{1+\delta_r})$ the j-th column of Λ is multiplied by $\frac{1}{1+\delta_j}$, $j=1,...,r$ and

$$\Lambda(\Delta+I)^{-1}\Lambda' = \sum_{j=1}^{r}\frac{1}{1+\delta_j}\Lambda_j\Lambda_j'$$

where Λ_j is the j-th column of Λ and the δ_j's are given in (20). Then

$$\ln|\Sigma| = \sum_{j=1}^{p} \ln \psi_{jj} + \sum_{j=1}^{r} \ln(1+\delta_j) \tag{22}$$

$$\begin{aligned}
\operatorname{tr}(\Sigma^{-1}S) &= \operatorname{tr}[(\Lambda\Lambda' + \Psi)^{-1}S] = \operatorname{tr}[(\Psi^{-1}S)] - \operatorname{tr}[\Psi^{-1}\Lambda(\Delta + I)^{-1}\Lambda'\Psi^{-1}S] \\
&= \operatorname{tr}(\Psi^{-1}S) - \sum_{j=1}^{r} \frac{1}{1+\delta_j} \operatorname{tr}(\Lambda_j\Lambda_j'(\Psi^{-1}S\Psi^{-1}) \\
&= \operatorname{tr}(\Psi^{-1}S) - \sum_{j=1}^{r} \frac{1}{1+\delta_j} \operatorname{tr}(\Lambda_j'(\Psi^{-1}S\Psi^{-1})\Lambda_j) \\
&= \operatorname{tr}(\Psi^{-1}S) - \sum_{j=1}^{r} \frac{1}{1+\delta_j} \Lambda_j'(\Psi^{-1}S\Psi^{-1})\Lambda_j
\end{aligned} \tag{23}$$

where Λ_j is the j-th column of Λ and by using the property $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and then observing that $\Lambda_j'(\Psi^{-1}S\Psi^{-1})\Lambda_j$ is a quadratic form.

3.3. Special case $\Psi = \sigma^2 I_p$

Let $\Psi = \sigma^2 I$ where σ^2 is a real scalar. Then $\Psi^{-1} = \sigma^{-2} I_p = \theta I_p$ where $\theta = \sigma^{-2}$. Then the log likelihood function reduces to the following:

$$\begin{aligned}
\ln L = &-\frac{np}{2}\ln(2\pi) + \frac{np}{2}\ln\theta - \frac{n}{2}\sum_{j=1}^{r}\ln(1+\delta_j) \\
&- \frac{\theta}{2}\operatorname{tr}(S) + \frac{\theta^2}{2}\sum_{j=1}^{r}\frac{1}{1+\delta_j}\Lambda_j' S \Lambda_j
\end{aligned}$$

where $1 + \delta_j = 1 + \theta\Lambda_j'\Lambda_j$ with Λ_j being the j-th column of Λ. Consider the equation

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\ln L = 0 \Rightarrow \\
&\quad \frac{np}{\theta} - n\sum_{j=1}^{r}\frac{\Lambda_j'\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} - \operatorname{tr}(S) \\
&\qquad + 2\theta\sum_{j=1}^{r}\frac{\Lambda_j' S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} - \theta^2\sum_{j=1}^{r}\frac{\Lambda_j'\Lambda_j}{(1+\theta\Lambda_j'\Lambda_j)^2}\Lambda_j' S\Lambda_j = 0.
\end{aligned} \tag{24}$$

For a specific j, consider

$$\begin{aligned}
&\frac{\partial}{\partial\Lambda_j}\ln L = O \Rightarrow \\
&-\frac{n}{2}\frac{2\theta\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} + \frac{\theta^2}{2}\frac{2S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} - \frac{\theta^2}{2}\frac{\Lambda_j' S\Lambda_j}{[1+\theta\Lambda_j'\Lambda_j]^2}(2\theta)\Lambda_j = O.
\end{aligned} \tag{25}$$

Cancel one θ and multiply on the left by $[1+\theta\Lambda_j'\Lambda_j]^2$. Then

$$-n(1+\theta\Lambda_j'\Lambda_j)\Lambda_j+\theta[(1+\theta\Lambda_j'\Lambda_j)S\Lambda_j]-\theta^2(\Lambda_j'S\Lambda_j)\Lambda_j=O. \tag{26}$$

Pre-multiply (26) by Λ_j' to obtain the following:

$$-n(1+\theta\Lambda_j'\Lambda_j)\Lambda_j'\Lambda_j+\theta[(1+\theta\Lambda_j'\Lambda_j)\Lambda_j'S\Lambda_j-\theta^2(\Lambda_j'S\Lambda_j)\Lambda_j'\Lambda_j=0 \tag{27}$$

This simplifies to the following:

$$\theta[\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2]=n\Lambda_j'\Lambda_j \tag{28}$$

which gives, for $\Lambda_j'S\Lambda_j\neq n(\Lambda_j'\Lambda_j)^2$,

$$\theta=\frac{n\Lambda_j'\Lambda_j}{\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2}, j=1,...,r. \tag{29}$$

Substituting this θ value from (29) to (26) we have the following, observing that

$$1+\theta\Lambda_j'\Lambda_j=\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2}.$$

$$-n\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2}\Lambda_j+\frac{n(\Lambda_j'\Lambda_j)}{\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2}[\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2}]S\Lambda_j$$
$$-\frac{(n\Lambda_j'\Lambda_j)^2}{[\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2]^2}(\Lambda_j'S\Lambda_j)\Lambda_j=O.$$

This simplifies to the following:

$$-n[\Lambda_j'S\Lambda_j-n(\Lambda_j'\Lambda_j)^2]\Lambda_j+n(\Lambda_j'\Lambda_j)S\Lambda_j-n^2(\Lambda_j'\Lambda_j)^2\Lambda_j=O$$

which gives

$$\left[S-\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'\Lambda_j}I\right]\Lambda_j=O\Rightarrow$$
$$\left|S-\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'\Lambda_j}I\right|=0. \tag{30}$$

This means that $\frac{\Lambda_j'S\Lambda_j}{\Lambda_j'\Lambda_j}$ is an eigenvalue of S and Λ_j is an eigenvector of S. There are p eigenvalues and the corresponding p eigenvectors for S. Note that for $\hat{\Lambda}_j$ as a column of $\hat{\Lambda}$ there are only r of them but $\hat{\Lambda}_j$ as an eigenvector of S there are $p\geq r$ of them. If η_j is an eigenvalue of S then

$$S\hat{\Lambda}_j=\eta_j\hat{\Lambda}_j\Rightarrow\eta_j=\frac{\hat{\Lambda}_j'S\hat{\Lambda}_j}{\hat{\Lambda}_j'\hat{\Lambda}_j}.$$

Take the largest r eigenvalues of S. Compute the corresponding eigenvectors $\hat{\Lambda}_j, j = 1, ..., r$. Then these give the estimates of the factor loadings. Then the sum of the eigenvalues is the trace of S and hence we have the following result:

Theorem 3.2. *The sum of all eigenvalues of S from equation (30), including the estimates of the r factor loadings $\hat{\Lambda}_j$'s is given by*

$$\sum_{j=1}^{p} \frac{\hat{\Lambda}_j' S \hat{\Lambda}_j}{\hat{\Lambda}_j' \hat{\Lambda}_j} = \text{tr}(S). \tag{31}$$

Hence, compute the eigenvalues and the corresponding eigenvectors of the sample sum of products matrix S. The estimates for the factor loadings, denoted by $\hat{\Lambda}_j$, are available from the eigenvectors $\hat{\Lambda}_j$'s of S. Take the first r largest eigenvalues of S and then compute the corresponding eigenvectors to obtain estimates for all the factor loadings. Then the procedure is also connected to the Principal Components Analysis. The estimates of the variances of the principal components will be $\hat{\Lambda}_j' S \hat{\Lambda}_j / \hat{\Lambda}_j' \hat{\Lambda}_j$ for $j = 1, ..., r$.

Verification

Does the θ value in (29) satisfy the likelihood equation (24)? Since θ is estimated through Λ_j for each $j = 1, ..., r$ we may replace θ in (24) as θ_j and insert inside the summation symbol. Then equation (24) will be the following:

$$\begin{aligned} & n \sum_j \frac{1}{\theta_j} - n \sum_j \frac{\Lambda_j' \Lambda_j}{1 + \theta_j \Lambda_j' \Lambda_j} - \text{tr}(S) \\ & \qquad + 2 \sum_j \theta_j \frac{\Lambda_j' S \Lambda_j}{1 + \theta_j \Lambda_j' \Lambda_j} - \sum_j \theta_j^2 \frac{\Lambda_j' \Lambda_j (\Lambda_j' S \Lambda_j)}{(1 + \theta_j \Lambda_j' \Lambda_j)^2} = 0. \end{aligned} \tag{32}$$

Now, substituting the value of θ_j from (29) into (24), the left side of (24) reduces to the following:

$$\begin{aligned} & n \sum_j \frac{[\Lambda_j' S \Lambda_j - n(\Lambda_j' \Lambda_j)^2]}{n \Lambda_j' \Lambda_j} - n \sum_j \frac{\Lambda_j' \Lambda_j}{\Lambda_j' S \Lambda_j} [\Lambda_j' S \Lambda_j - n(\Lambda_j' \Lambda_j)^2] - \text{tr}(S) \\ & + 2 \sum_j n \Lambda_j' \Lambda_j - \sum_j \frac{\Lambda_j' \Lambda_j}{\Lambda_j S \Lambda_j} (n \Lambda_j' \Lambda_j)^2 \\ & = \sum_j \frac{\Lambda_j' S \Lambda_j}{\Lambda_j' \Lambda_j} - \text{tr}(S) = 0 \end{aligned}$$

because of Theorem 3.2. Hence equation (32) holds for the θ value from (29) and Λ_j value from (26).

Since the basic estimating equation for $\hat{\theta}$ is coming from (28) as

$$\theta[\Lambda_j' S \Lambda_j - n(\Lambda_j'\Lambda_j)^2] = n\Lambda_j'\Lambda_j$$

we may sum up over j on both sides and then we have the combined estimate for θ as follows:

$$\hat{\theta} = \frac{n\sum_{j=1}^r \hat{\Lambda}_j'\hat{\Lambda}_j}{\sum_{j=1}^r [\hat{\Lambda}_j' S \hat{\Lambda}_j - n(\hat{\Lambda}_j'\hat{\Lambda}_j)^2]} = \frac{1}{\hat{\sigma}^2} \tag{33}$$

or the estimate of σ^2 is the reciprocal of the estimate of θ.

Maximum value of the exponent

We have the estimate $\hat{\theta}$ of θ coming from (29) at the estimated value $\hat{\Lambda}_j$ of Λ_j for each j. This $\hat{\Lambda}_j$ is an eigenvector of S coming from (30). The exponent of the likelihood function is $-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S)$ and our $\Sigma = \Lambda\Phi\Lambda' + \Psi$ and our identification conditions are $\Phi = I_p$ and $\Lambda'\Psi^{-1}\Lambda =$ a diagonal matrix with positive diagonal elements. Under these conditions and for the special case $\Psi = \sigma^2 I_p, \sigma^{-2} = \theta$ we have seen that the exponent in the log-likelihood function reduces to $-\frac{1}{2}\theta\mathrm{tr}(S) + \frac{1}{2}\theta^2\sum_{j=1}^r \frac{\Lambda_j' S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j}$. Now, consider $\theta\mathrm{tr}(S) - \sum_{j=1}^r \theta^2 \frac{\Lambda_j' S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} = \delta$ say . Since an estimate of θ is available from each j we may replace θ by θ_j whenever it is convenient. Then

$$\delta = \theta\mathrm{tr}(S) - \theta\sum_{j=1}^r \theta_j \frac{\Lambda_j' S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j}.$$

Substituting an expression for $1+\theta\Lambda_j'\Lambda_j$ from (29) we have the following:

$$\sum_j \theta_j \frac{\Lambda_j' S\Lambda_j}{1+\theta\Lambda_j'\Lambda_j} = \sum_j n\Lambda_j'\Lambda_j \text{ and } \delta = \theta[\mathrm{tr}(S) - n\sum_j \Lambda_j'\Lambda_j].$$

From Theorem 3.2, sum of the eigenvalues of S is $\sum_j \frac{\Lambda_j' S\Lambda_j}{\Lambda_j'\Lambda_j}$ and substituting this we have

$$\begin{aligned}\delta &= \theta[\sum_j \frac{\Lambda_j' S\Lambda_j}{\Lambda_j'\Lambda_j} - n\sum_j \Lambda_j'\Lambda_j] = \theta[\sum_j (\frac{\Lambda_j' S\Lambda_j - n(\Lambda_j'\Lambda_j)^2}{\Lambda_j'\Lambda_j})] \\ &= \theta\sum_j \frac{n}{\theta} = n\sum_j 1 = np\end{aligned}$$

which establishes the result that the exponential part simplifies to $\mathrm{e}^{-\frac{np}{2}}$. Thus, the estimates coming from (29) and (30) for θ and $\Lambda_j, j = 1, ..., r$ maximize the likelihood function.

4. General case

Let us consider the original parameters Λ_j and Θ. Then $1+\delta_j = 1+\Lambda_j'\Theta^2\Lambda_j, \Psi^{-1} = \Theta^2, \Psi^{-\frac{1}{2}} = \Theta, \Theta = \text{diag}(\theta_1, ..., \theta_p)$. Then

$$\ln L = -\frac{np}{2}\ln(2\pi) + n\sum_{j=1}^{p}\ln\theta_j - \frac{n}{2}\sum_{j=1}^{r}\ln(1+\delta_j)$$
$$-\frac{1}{2}\text{tr}(\Theta^2 S) + \frac{1}{2}\sum_{j=1}^{r}\frac{1}{1+\delta_j}\Lambda_j'\Theta^2 S\Theta^2\Lambda_j.$$

Differentiating $\ln L$ with respect to Λ_j, for a specific j, and equating to a null vector we have the following:

$$-\frac{n}{2}\frac{2\Theta^2}{1+\delta_j}\Lambda_j + \frac{1}{2}\frac{2\Theta^2 S\Theta^2}{1+\delta_j}\Lambda_j - \frac{1}{2}\frac{2\Theta^2(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)}{(1+\delta_j)^2}\Lambda_j = O. \tag{34}$$

Premultiply (34) by Λ_j' and then sum up over j to obtain

$$-n\sum_{j=1}^{r}\frac{\Lambda_j'\Theta^2\Lambda_j}{1+\delta_j} + \sum_{j=1}^{r}\frac{\Lambda_j'\Theta^2 S\Theta^2\Lambda_j}{1+\delta_j} - \sum_{j=1}^{r}\frac{\Lambda_j'\Theta^2\Lambda_j}{(1+\delta_j)^2}(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j) = 0. \tag{35}$$

Now, consider the derivative of $\ln L$ with respect to a specific θ_j equated to zero. For constructing this equation, observe the following results:

$$\theta_j\frac{\partial}{\partial\theta_j}(\Lambda_j'\Theta^2\Lambda_j) = 2\theta_j\Lambda_j'\text{diag}(0, ..., 0, \theta_j, 0, ..., 0)\Lambda_j$$
$$= 2\Lambda_j'\text{diag}(0, ..., 0, \theta_j^2, 0, ..., 0)\Lambda_j; \tag{36}$$
$$\theta_j\frac{\partial}{\partial\theta_j}(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j) = 2\Lambda_j\text{diag}(0, ..., 0, \theta_j^2, 0, ..., 0)S\Theta^2\Lambda_j$$
$$+ 2\Lambda_j'\Theta^2 S\ \text{diag}(0, ..., 0, \theta_j^2, 0, .., 0)\Lambda_j. \tag{37}$$

Now, consider the operator $\beta = [\theta_1\frac{\partial}{\partial\theta_1} + ... + \theta_p\frac{\partial}{\partial\theta_p}]$ operating on $\Lambda_j'\Theta^2\Lambda_j$ and $\Lambda_j'\Theta^2 S\Theta^2\Lambda_j$. Then we have the following:

$$\beta(\Lambda_j'\Theta^2\Lambda_j) = 2\Lambda_j'\Theta^2\Lambda_j, \beta(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j) = 2[2\Lambda_j'\Theta^2 S\Theta^2\Lambda_j]. \tag{38}$$

Since $\frac{\partial}{\partial\theta_j}\ln L = 0, j = 1, ..., p$ the operator β operating on $\ln L$ is also zero. That is,

$$np - n\sum_{j=1}^{r}\frac{\Lambda_j'\Theta^2\Lambda_j}{1+\delta_j} + \sum_{j=1}^{r}\frac{2\Lambda_j'\Theta^2 S\Theta^2\Lambda_j}{1+\delta_j}$$
$$-\sum_{j=1}^{r}\frac{\Lambda_j'\Theta^2\Lambda_j}{(1+\delta_j)^2}(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j) - \sum_{i=1}^{p}\theta_i^2 s_{ii} = 0. \tag{39}$$

Comparing (35) and (39) we have

$$np+\sum_j \frac{\Lambda_j'\Theta^2 S\Theta^2\Lambda_j}{1+\Lambda_j'\Theta^2\Lambda_j}-\sum_{i=1}^p \theta_i^2 s_{ii}=0. \tag{40}$$

Since (34) holds for each j, multiply by $(1+\delta_j)^2$ then we have the following observing that $1+\delta_j=1+\Lambda_j'\Theta^2\Lambda_j$.

$$-n\Theta^2\Lambda_j+\Theta^2 S\Theta^2\Lambda_j-\Theta^2\frac{(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)}{1+\delta_j}\Lambda_j=O\Rightarrow$$
$$-n(1+\Lambda_j'\Theta^2\Lambda_j)\Theta^2\Lambda_j+(1+\Lambda_j'\Theta^2\Lambda_j)\Theta^2 S\Theta^2\Lambda_j-(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)\Theta^2\Lambda_j=O.$$

Pre-multiplying by Λ_j' and simplifying we have

$$-n(1+\Lambda_j'\Theta^2\Lambda_j)(\Lambda_j'\Theta^2\Lambda_j)+(1+\Lambda_j'\Theta^2\Lambda_j)(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)-(\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)(\Lambda_j'\Theta^2\Lambda_j)=0$$
$$\Rightarrow (\Lambda_j'\Theta^2 S\Theta^2\Lambda_j)=n(1+\Lambda_j'\Theta^2\Lambda_j)(\Lambda_j'\Theta^2\Lambda_j). \tag{41}$$

After substituting the value of $\Lambda_j'\Theta^2 S\Theta^2\Lambda_j$ from (41) into (34) and simplifying we have

$$-n(\Theta\Lambda_j)-(\Lambda_j'\Theta^2\Lambda_j)(\Theta\Lambda_j)+(\Theta S\Theta)(\Theta\Lambda_j)=O.$$

Let $U_j=\Theta\Lambda_j$. Then

$$[\Theta S\Theta-n(1+U_j'U_j)I]U_j=O. \tag{42}$$

Now, after substituting the value of $\Lambda_j\Theta^2 S\Theta^2\Lambda_j$ from (41) into (40) we have

$$np+\sum_j n\Lambda_j'\Theta^2\Lambda_j-\sum_{j=1}^p \theta_i^2 s_{ii}=0.$$

Let

$$c=\frac{1}{p}\sum_{j=1}^r \Lambda_j'\Theta^2\Lambda_j=\frac{1}{p}\sum_{j=1}^r U_j'U_j. \tag{43}$$

Then from (4.10)

$$\sum_{i=1}^p[(1+c)-\theta_i^2\frac{s_{ii}}{n}]=0.$$

One solution here is

$$\theta_i^2=(1+c)\frac{n}{s_{ii}}, i=1,...,p. \tag{44}$$

But from, (4.11), $\hat{\Theta}S\hat{\Theta}=(1+c)nR, R=$ sample correlation matrix. Then from (4.10) an eigenvalue of nR is of the form $\frac{n(1+\hat{U}_j'\hat{U}_j)}{1+c}$. Hence the procedure is the following: Compute

all eigenvalues of the sample correlation matrix R. Let $U_j, j = 1, ..., r$ be the eigenvectors corresponding to the largest r eigenvalues of R. Then $c = \frac{1}{p}\sum_{j=1}^{r} U_j'U_j$ and

$$\hat{\Lambda}_j = \hat{\Theta}^{-1}U_j = \frac{1}{\sqrt{n(1+c)}}\text{diag}(\sqrt{s_{11}}, ..., \sqrt{s_{pp}})U_j \text{ and}$$
$$\hat{\Theta} = \sqrt{n(1+c)}\text{diag}(\frac{1}{\sqrt{s_{11}}}, ..., \frac{1}{\sqrt{s_{pp}}}). \tag{45}$$

This completes the estimation of $\Lambda_j, j = 1, ..., r$ and Θ.

Note 4.1. If the parameters in (5), under the identification conditions $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda =$ diagonal, are taken as $\Psi^{-\frac{1}{2}} = \Theta = \text{diag}(\theta_1, ..., \theta_p)$ and $U_j = \Theta\Lambda_j, j = 1, ..., r$ then it is easily seen that each θ_i is estimated by $\hat{\theta}_i^2 = \frac{n}{s_{ii}}, i = 1, ..., p$ so that $\hat{\Theta}S\hat{\Theta} = nR$ where R is the sample correlation matrix.

Maximum value of the likelihood function

The exponential part in $\ln L$ is $-\frac{1}{2}\text{tr}(\hat{\Theta}S\hat{\Theta}) + \frac{1}{2}n\sum_{j=1}^{r} U_j'U_j$. Observe that from the definition of c, we have $n\sum_{j=1}^{r} U_j'U_j = ncp$ and $\text{tr}(\hat{\Theta}S\hat{\Theta}) = (1+c)n\text{tr}(R) = (1+c)np$. Hence the exponent simplifies to $-\frac{1}{2}(1+c)np + \frac{1}{2}npc = -\frac{np}{2}$. Hence the estimates given (4.12) will maximize the likelihood function.

5. Testing of Hypotheses

The usual test in this topic is to test for identifiability or to test the hypothesis H_o that the population covariance matrix $\Sigma > O$ can be represented as $\Sigma = \Lambda\Phi\Lambda' + \Psi$ when $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda =$ a diagonal matrix with positive diagonal elements and $\Psi > O$ is a diagonal matrix and $\Lambda = (\lambda_{ij})$ is a $p \times r, r \leq p$ matrix of full rank r and of the factor loadings λ_{ij}'s. That is,

$$H_o : \Sigma = \Lambda\Lambda' + \Psi. \tag{46}$$

When the model is of the form $X = \mu + \Lambda f + e$ where X, with $X' = (x_1, ..., x_p)$, is the $p \times 1$ vector of observed scores on p tests or p batteries of tests, μ is $p \times 1$ vector of general effect, f is $r \times 1$ vector of unknown factors, $\Lambda = (\lambda_{ij})$ is the unknown $p \times r$ matrix of factor loadings and e is the $p \times 1$ error vector. When e and f are uncorrelated then the covariance matrix of X is given by

$$\Sigma = \Lambda\Phi\Lambda' + \Psi$$

where $\text{Cov}(f) = \Phi > O, \text{Cov}(e) = \Psi > O$ with Φ being $r \times r$ and Ψ being $p \times p$ and diagonal. A simple random sample from this X means iid (independently and identically distributed) $p \times 1$ vectors $X_j, j = 1, ..., n$ where n is the sample size, $X_j' = (x_{1j}, x_{2j}, ..., x_{pj})$. The sample sum of products matrix or "corrected" sample sum of products matrix is

$S=(s_{ij}), s_{ij}=\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$, where, for example, the average of x_i, namely $\bar{x}_i$, is $\bar{x}_i=\sum_{k=1}^{n}x_{ik}/n$. If e and f are independently normally distributed then the likelihood ratio criterion or λ-criterion is given by

$$\lambda=\frac{\max_{H_o}L}{\max L}=\frac{|\hat{\Sigma}|^{\frac{n}{2}}}{|\hat{\Lambda}\hat{\Lambda}'+\hat{\Psi}|^{\frac{n}{2}}} \tag{47}$$

where $\hat{\Sigma}=\frac{1}{n}S$ and under H_o the covariance matrix is $\Sigma=\Lambda\Lambda'+\Psi$ and it is assumed that $\Phi=I$, an identity matrix, and the $r\times r$ matrix $\Lambda'\Psi^{-1}\Lambda=\text{diag}(\delta_1,...,\delta_r)$, a diagonal matrix with diagonal elements $\delta_j>0, j=1,...,r$. From Section 4.2 it is seen that

$$|\Lambda\Lambda'+\Psi|=|\Psi|\ |\Lambda'\Psi^{-1}\Lambda+I|. \tag{48}$$

and $\delta_j=1+\Lambda_j'\Psi^{-1}\Lambda_j=1+U_j'U_j$ where $U_j=\Psi^{-\frac{1}{2}}\Lambda_j$ and Λ_j is the j-th column of Λ for $j=1,...,r$. It is shown in (42) that U_j is an eigenvector of the sample correlation matrix R and

$$\prod_{j=1}^{r}(1+\delta_j)=\prod_{j=1}^{r}(1+U_j'U_j)=|\Lambda'\Psi^{-1}\Lambda+I|.$$

But from (44) an eigenvalue of R is of the form $\frac{1+U_j'U_j}{(1+c)}, j=1,...,r$. That is, $1+U_j'U_j=(1+c)\nu_j$ where let $\nu_j, j=1,...,p$ be the eigenvalues of R and let the largest r eigenvalues be $\nu_1,...,\nu_r$. From (43), $\Psi^{-\frac{1}{2}}=\Theta, \hat{\Theta}^2=n(1+c)\text{diag}(\frac{1}{s_{11}},...,\frac{1}{s_{pp}})$.

$$\begin{aligned}\frac{|\hat{\Sigma}|}{|\hat{\Psi}|}&=|\hat{\Theta}\frac{S}{n}\hat{\Theta}|=|(1+c)R|=(1+c)^p\nu_1...\nu_p\\ &\Rightarrow\frac{|\frac{1}{n}S|}{|\hat{\Psi}|\ |\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda}+I|}=\frac{(1+c)^p\nu_1...\nu_r...\nu_p}{(1+c)^r\nu_1...\nu_r}\\ &=(1+c)^{p-r}\nu_{r+1}...\nu_p.\end{aligned} \tag{49}$$

Hence we reject the null hypothesis for small values of the product $(1+c)^{p-r}\nu_{r+1}...\nu_p$ the product of the smallest $p-r$ eigenvalues of the sample correlation matrix R, multiplied by $(1+c)^{p-r}$. For evaluating the critical point one needs the null distribution of the product of the eigenvalues $(1+c)^{p-r}\nu_{r+1}...\nu_p$, which is difficult to evaluate. For large samples or for n large we have a chisquare approximation giving the result $-2\ln\lambda$ is approximately a chisquare with k degrees of freedom where k is the number of parameters restricted by our hypothesis H_o, where λ is the likelihood ratio criterion. In our case the total number of parameters available is $\frac{p(p+1)}{2}$ in $\Sigma>O$ plus $\frac{r(r-1)}{2}$ the number of conditions imposed by assuming $\lambda'\Psi^{-1}\Lambda$ to be a diagonal matrix. The number of parameters estimated is pr in Λ plus p parameters in Ψ. Hence the number of parameters restricted by the hypothesis H_o is

$$k = \frac{p(p+1)}{2} + \frac{r(r-1)}{2} - pr - p = \frac{1}{2}[(p-r)^2 - (p+r)]. \tag{50}$$

Therefore

$$-2\ln\lambda \approx \chi_k^2 \tag{51}$$

when n is large, where k is given in (50). We reject H_o for large values of this χ_k^2 or reject H_o if the observed value of this $\chi_k^2 \geq a$ with the probability $Pr\{\chi_k^2 \geq a\} = \alpha$ where α is the probability for the type-1 error or the probability of rejecting H_o when in fact H_o is true.

References

[1] Anderson, T. W. An Introduction to Multivariate Statistical Analysis (3rd ed.). New York: John Wiley, 2003.

[2] Bartlett, M. S. The Statistical Conception of Mental Factors. British Journal of Psychology, 28 (1937), 97-104.

[3] Bartlett, M. S. A Note on Multiplying Factors for Various Chi-Squared Approximations. Journal of the Royal Statistical Society (B) 16 (1954),296-298.

[4] Harmon, H. H. Modern Factor Analysis (3rd ed.). Chicago: The University of Chicago Press, 1976.

[5] Härdle, Wolfgang Karl and Simar, Léopold Applied Multivariate Statistical Analysis, New York, Springer, 2011.

[6] Johnson, R.A. and Wichern, D.W. Applied multivariate statistical analysis, 3rd Edn., New Delhi, Prentice Hall of India, 2001.

[7] Joreskog, K. G. Factor Analysis by Least Squares and Maximum Likelihood. In Statistical Methods for Digital Computers, edited by K. Enslem, A. Ralston, and H. S. WIlf. New York: John Wiley, 1975.

[8] Kaiser, H.F. The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika,23 (1958), 187-200.

[9] Lawley, D. N., and A. E. Maxwell. Factor Analysis as a Statistical Method (2nd ed.). New York: American Elsevier Publishing Co., 1971.

[10] Mardia, K.V., Kent, J.T., and Bibby, J.M. Multivariate Analysis, New York, Academic Press. 1979

[11] Mathai, A. M. Jacobians of Matrix Transformations and Functions of Matrix Argument, New York, World Scientific Publishing. 1997.

[12] Mathai, A. M. On non-central generalized Laplacianness of quadratic forms in normal variables, Journal of Multivariate Analysis, 45 (1993), 239246.

[13] Mathai, A. M., Provost, S.B., Quadratic Forms in random Variables: Theory and Applications, New York, Marcel Dekker. 1992.

[14] Maxwell, A. E. Multivariate Analysis in Behavioral Research. London: Chapman and Hall, 1977.

[15] Morrison, D. F. Multivariate Statistical Methods (4th ed.). Belmont, CA: Brooks/Cole Thompson Learning, 2005.

[16] Wright, S. The Interpretation of Multivariate Systems. In Statistics and Mathematics in Biology, edited by O. Kempthorne and others. Ames, lA: Iowa State UmvefSlty Press, 1954, 11-33.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 23-34.

Some Unbiased Estimations Under Random Censoring For Gompertz Inliers Model

K. Muralidharan and Bavagosai Pratima
Department of Statistics, Faculty of Science, The Maharaja Sayajirao University of Baroda, Vadodara 390002, India.
Email:lmv_murali@yahoo.com, pratimabava@outlook.com

Abstract

It is known that inliers are natural occurrences of a life test, where some of the items fail immediately (instantaneously) or within a short time (early) of the life test due to mechanical failure, inferior quality or faulty construction of items and components etc. We consider random censored inliers to investigate the estimation of parameters of Gompertz distribution. Gompertz distribution is used in many situations in medical studies, where time-to-event occurrence of life times is observed. Because of the nature of the data, observing complete sample is not always viable, and hence random censoring concept is used to study the data. We propose the classical estimation of parameters including the Uniformly Minimum Variance Unbiased Estimation for density function, reliability function and some parametric functions of the model. As an illustration, we also discuss a randomly censored real data set and study its characteristics.

Key words: Early failures, Failure time distribution, Inliers, Instantaneous failures, Randomly censored sample.

1. Introduction

In failure time studies, *inliers* in a data set are subsets of observations sufficiently small, relative to the rest of the observations, which appear to be inconsistent with the remaining data set. They are either the results of instantaneous failures, or early failures, experienced in life testing experiments as well as in clinical trials. The test items that fail at time are called *instantaneous failures*, and the test items that fail prematurely are called *early failures*. These occurrences may be due to inferior quality or faulty construction or due to no

response to the treatments. Such failures usually discard the assumption of a single mode distribution and hence the usual method of modeling and inference procedures may not be accurate in practice. There are series of papers developed and studied on this concept since year 2000. One may refer to the early paper by Kale and Muralidharan (2000), who have introduced the term *inliers* in connection with the estimation of parameters of early failure models with modified failure time distribution being an exponential distribution with mean θ and the number of inliers known. See also Kale (2003), Muralidharan and Lathika (2006), Kale and Muralidharan (2008), and Muralidharan (2010), Muralidharan and Arti (2008, 2013), Bavagosai and Muralidharan (2016), Muralidharan and Bavagosai (2016a, b) etc for more details and for complete sample studies.

In life testing and reliability, most of the experiments often take a long time to terminate, so the experiment does not observe all failure times due to cost and time considerations. The concept of censoring is introduced to account for these considerations. The censored observations contain only partial information (incomplete observations) about the random variable of interest. Among various censoring concepts, the random censoring is commonly used in time-to-event studies of medical applications including clinical trials. For development of models and inference, one may refer to Meeker and Escobar (1998), Lawless (2003), Klein and Moeschberger, 2003) and so on.

The Gompertz distribution was first introduced by the British actuary Benjamin Gompertzin 1825 to describe human mortality and establish actuarial tables Gompertz (1825). The Gompertz distribution with shape parameter α and scale parameter θ , have the following CDF

$$F(x;\alpha,\theta) = \begin{cases} 1 - e^{-\frac{\theta}{\alpha}(e^{\alpha x} - 1)}, & x > 0; \alpha > 0; \theta > 0 \\ 0, & o.w., \end{cases} \tag{1}$$

This distribution has been widely used, especially in actuarial science, epidemiological, biomedical studies and in demography and is now applied in fields including reliability and life testing studies. This and a particular inliers prone example motivated us to try Gompertz distribution to model the inliers situations using randomly censored data.

The organization of the paper is as follows: The inliers prone Gompertz model under random censoring is presented in Section 2. We propose the maximum likelihood estimator (MLE) and its asymptotic distribution followed by the unbiased estimation for various parametric functions in Section 3. The Section 4 deals with a numerical example and its inference.

2. The Gompertz inliers model under random censoring

Consider the model $\mathcal{F} = \{F(x;\theta), x \geq 0, \theta \in \Theta\}$ where $F(x;\theta)$ is a continuous failure time CDF with $F(0) = 0$. To accommodate a real life situation, where instantaneous failures are observed at the origin, the model F is modified to $G = \{G(x;p,\theta), x \geq \delta, \theta \in \Theta, 0 \leq p \leq 1\}$ by using a mixture in the proportions $(1-p)$ and p respectively of a singular random variable at δ and a random variable X with CDF $F \in \mathcal{F}$. Here δ is known and is sufficiently small. Thus, the modified failure time distribution is given by the CDF:

$$G(x;p,\theta,\delta) = \begin{cases} 0, & x < \delta \\ (1-p) + pF(x;\theta), & x \geq \delta. \end{cases} \tag{2}$$

For $\delta = 0$ the model reduces to the instantaneous failures case and if $\delta > 0$, it reduces to the case of early failures. If we use $F(x;\theta)$ as in (1) in the model (2), we get the pdf of Gompertz inliers model as

$$g_X(x;p,\alpha,\theta) = \begin{cases} 0, & x < \delta \\ 1 - pe^{-\frac{\theta}{\alpha}\left(e^{\alpha\delta}-1\right)}, & x = \delta \\ p\theta e^{\alpha x} e^{-\frac{\theta}{\alpha}(e^{\alpha x}-1)}, & x > \delta. \end{cases} \tag{3}$$

Here δ is sufficiently small and assumed known. Consider X be the i.i.d. random variable denoting life time having CDF defined by (3) and Y be the censoring variable having CDF $G_Y(x) = 1 - e^{-\frac{\varrho}{\alpha}(e^{\alpha x} - 1)}$, where $\varrho > 0$ (known). Let Y is independent of X . Let $T = \min(X,Y), X > \delta$. Then under random censoring setting, the distribution function of T after some simplification is obtained as

$$g_T(t;p,\propto,\theta^*) = \begin{cases} 0, & t < \delta \\ 1 - pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta} - 1\right)}, & t = \delta \\ p\theta^* e^{\alpha t} e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha t} - 1\right)}, & t > \delta, \end{cases} \tag{4}$$

where $\theta^* = \theta + \zeta$.

3. Parameter estimation under random censoring

If we define

$$I(t) = \begin{cases} 1, & \text{if } t = \delta \\ 0, & \text{otherwise,} \end{cases}$$

then the pdf in (4) may be written as

$$g_T\left(t;p,\propto,\theta^*\right)=\left(1-\ pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}\ -\ 1\right)}\right)^{I(t)}\left(p\ \theta^*\ e^{\alpha t}e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha t}\ -\ 1\right)}\right)^{1-I(t)},t>0$$

Let $r\left(\leq n\right)$ denote the number of life times greater than δ . Then the joint density function of T is expressed as

$$g_T\left(\underline{t};p,\alpha,\theta^*\right)=\left(1-\ pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}\ -\ 1\right)}\right)^{n-r}p^r\theta^{*r}e^{\alpha\sum_{t_i>\delta}t_i}e^{\frac{\theta^*}{\alpha}\sum_{t_i>\delta}\left(e^{\alpha t_i}-1\right)} \tag{5}$$

3.1 The Maximum Likelihood Estimation and its asymptotic distribution

Assuming α known, the log-likelihood function is

$$\begin{aligned}\log L\left(p,\alpha,\theta^*|\underline{x}\right)=(n-r)\log\left(1-\ pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}\ \ 1\right)}\right)+r\log p+r\log\theta^*\\+\alpha\sum_{t_i>\delta}t_i-\frac{\theta^*}{\alpha}\sum_{t_i>\delta}\left(e^{\alpha t_i}-1\right).\end{aligned} \tag{6}$$

Taking partial derivative of log-likelihood function with respect to θ^* and p and equating to zero we get MLE of θ^* and p as $\hat{\theta}^*_{MLE}$ and $\hat{p}_{MLE}$, say, respectively as follows:

$$\hat{\theta}^*_{MLE}=\frac{r\alpha}{e^{\alpha\delta}\sum_{t_i>\delta}\left(e^{\alpha(t_i-\delta)}\ -\ 1\right)} \tag{7}$$

and

$$\hat{p}_{MLE}=\frac{re^{\frac{\widehat{\theta*}}{\alpha}\left(e^{\alpha\delta}-1\right)}}{n} \tag{8}$$

The Fisher information matrix $I_g\left(p,\theta^*\right)$ is given by

$$I_g\left(p,\theta^*\right)=\begin{bmatrix}I_{pp} & I_{p\theta^*}\\ I_{\theta^*p} & I_{\theta^*\theta^*}\end{bmatrix}$$

where,

$$I_{pp}=E\left(-\frac{\partial^2\log g\left(p,\alpha,\theta^*|t\right)}{\partial p^2}\right)=\frac{e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}}{p\ \left(1-pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}\right)}=\frac{e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}}{p\ p^*},$$

$$I_{\theta^*\theta^*}=E\left(-\frac{\partial^2\log g\left(p,\alpha,\theta^*|t\right)}{\partial{\theta^*}^2}\right)=\frac{(1-p^*)\left\{\theta^{*2}\left[\frac{e^{\alpha\delta}-1}{\alpha}\right]^2+p^*\right\}}{{\theta^*}^2p^*}$$

and

$$I_{p\theta^*} = E\left(-\frac{\partial^2 logg\,(p,\alpha,\theta^*|x)}{\partial p \partial \theta^*}\right) = -\frac{e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}\ \frac{\left(e^{\alpha\delta}-1\right)}{\alpha}}{p^*}$$

where, $p^* = 1 - pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}$. The determinant of $I_g\,(p,\theta^*)$ is $\Delta = \frac{e^{-\frac{2}{\alpha}\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}}{\theta^{*2}p^*}$ and hence the inverse matrix $I_g^{-1}\,(p,\theta^*)$ is

$$I_g^{-1}\,(p,\theta^*) = \begin{bmatrix} \frac{p\left(\theta^{*2}\frac{\left(e^{\alpha\delta}-1\right)^2}{\alpha^2}+p^*\right)}{e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}} & -\frac{\theta^{*2}\left(e^{\alpha\delta}-1\right)}{\alpha\ e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}} \\ -\frac{\theta^{*2}\left(e^{\alpha\delta}-1\right)}{\alpha\ e^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}} & \frac{\theta^{*2}}{1-p^*} \end{bmatrix}.$$

Using the standard result of MLE, we have $\left(\hat{p}_{MLE}, \widehat{\theta^*}_{MLE}\right)' \sim AN^{(2)}\left[(p,\theta^*)',\ \frac{1}{n}I_g^{-1}\,(p,\theta^*)\right]$. The approximate confidence interval for p and θ^* with confidence coefficient $(1-\alpha)$ are respectively given by:

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\left(\hat{\theta}^{*2}\frac{\left(e^{\alpha\delta}-1\right)^2}{\alpha^2}+\hat{p}^*\right)}{n\ \ e^{-\frac{\hat{\theta}^*}{\alpha}\left(e^{\alpha\delta}-1\right)}}}$$

and

$$\hat{\theta}^* \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}^{*2}}{n\left(1-\hat{p}^*\right)}}$$

where, $\hat{p}^* = 1 - \hat{p}\ e^{-\frac{\widehat{\theta_*}}{\alpha}\left(e^{\alpha\delta}-1\right)}$.

3.2 UMVUE of parametric functions

The pdf $g\,(t;p,\alpha,\theta^*)$ given by (4) with α known, can be expressed as:

$$g_T\,(t;p,\propto,\theta^*) = \left(1 - pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}\ \ 1\right)}\right)^{I(t)}\left(pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}\ \ 1\right)}\right)^{(1-I(t))} \times \left[\frac{e^{\alpha t}\ \left(e^{-\theta^*}\right)^{\left[\frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}\ \ 1\right)}{\alpha}\right]}}{\left(\frac{1}{\theta^*}\right)}\right]^{(1-I(t))}. \tag{9}$$

Also,

$$g_T\,(t;p,\propto,\theta^*) = a\,(t)^{(1-I(t))}\,\frac{h\,[\theta^*]^{d(t)(1-I(t))}}{\left[\frac{g(\theta^*)}{1-p^*}\right]}\left[g\,(\theta^*)\left(\frac{p^*}{1-p^*}\right)\right]^{I(t)}, \tag{10}$$

where $a(t)=e^{at}$; $\mathrm{h}(\theta^*)=e^{-\theta^*}$; $\mathrm{d}(t)=\frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}\ -\ 1\right)}{\alpha}$; $\mathrm{g}(\theta^*)=\frac{1}{\theta^*}$; $p^*=1-pe^{-\frac{\theta^*}{\alpha}\left(e^{\alpha\delta}-1\right)}$; $a(t)>0$ and $g(\theta^*)=\int_{t>\delta}a(t)\,h[\theta^*]^{d(t)}\,dt$.

The density in (10) so obtained is defined with respect to a measure $\mu(t)$ which is the sum of Lebesgue measure over (δ,∞) and a singular measure at δ . Also, it is a well know form of a two-parameter exponential family with natural parameters $(\eta_1,\eta_2)=\left(\log\left[g(\theta^*)\left(\frac{p^*}{1-p^*}\right)\right],\log\left[h(\theta^*)\right]\right)$ generated by underlying indexing parameters (p^*,θ^*).

Hence $(I(x),d(x)(1-I(x)))$ is jointly minimal sufficient for (p^*,θ^*), as $I(x)$ and $\frac{\left(e^{\alpha\delta}-1\right)}{\alpha}(1-I(x))$ do not satisfy any linear restriction. The η s do not satisfy any linear constraint too and hence natural parameter space is convex set E_2 containing (10) a full-rank family. Thus, $(I(t),d(t)(1-I(t)))$ is complete. Also, the joint density function of T as in (3.1) can be expressed as

$$\begin{aligned} g(\underline{t};p,\alpha,\theta^*) &= (p^*)^{n-r}(1-p^*)^r\,\theta^{*r}e^{\alpha\sum_{t_i>\delta}t_i}\ e^{-\theta^*\frac{e^{\alpha\delta}}{\alpha}\sum_{t_i>\delta}\left(e^{\alpha(t_i-\delta)}\ -\ 1\right)} \\ &= (p^*)^{n-r}(1-p^*)^r\theta^{*r}\ e^{-\theta^*z}, \\ &= P(n-R=n-r)\quad \mathrm{G}(z;\theta^*|(n-r)), \end{aligned} \tag{11}$$

where $z=\frac{e^{\alpha\delta}}{\alpha}\sum_{t_i>\delta}\left(e^{\alpha(t_i-\delta)}\ -\ 1\right)$. Therefore, by Neymans factorization theorem $(n-R,Z)$ are jointly sufficient for (p^*,θ^*). Also, $n-R$ is binomial which is the same as that of R with parameter (n, p*), and is a complete family and the variable $(Z|R=r,r>0)$ is distributed as a gamma random variable having density function:

$$g(z;\ \theta^*|r)=\frac{\theta^{*r}}{\Gamma r}z^{r-1}e^{-\theta^*z},\ z>\delta;\theta>0. \tag{12}$$

Hence, $Z|R$ is completely sufficient for $g(\theta^*)=\frac{1}{\theta^*}$. This preserves the exponential structure for (12). Using the result of Jayade (1993), $(n-R,Z)$ is complete sufficient for (p^*,θ^*) . The joint pdf of $(n-R,Z)$ can be written as:

$$g(z;n-r,p,\alpha,\theta^*,d)=\begin{cases}(p^*)^n, & z=\delta;r=0\\ B(z,r,n)\,\dfrac{e^{-\theta^*z}\left(\frac{p^*}{\theta^*(1-p^*)}\right)^{n-r}}{\left(\frac{1}{\theta^*\ (1-p^*)}\right)^n}, & z>\delta;r=1,2,\ldots,n.\end{cases} \tag{13}$$

where,

$$B(z,r,n)=\begin{cases}1, & z=\delta;r=0\\ \binom{n}{r}\frac{z^{r-1}}{\Gamma r}, & z>\delta;r=1,2,\ldots,n.\end{cases} \tag{14}$$

is such that

$$(p^*)^n+\sum_{r=1}^{n}\int_{z>\delta}B(z,r,n)\,e^{-\theta z}\left(\frac{p^*}{\theta^*(1-p^*)}\right)^{n-r}\left(\frac{1}{\theta^*\ (1-p^*)}\right)^{-n}dz=1.$$

Following Roy and Mitra (1957) and Jani and Singh (1995), the UMVUEs of parametric function $\phi(p,\theta^*)$ exists if and only if $\phi(p,\theta^*)$ can be expressed in the form:

$$\phi(p,\theta^*) = \alpha(\delta,\delta,n)(p^*)^n + \sum_{r=1}^{n}\int_{z>\delta}\alpha(z,r,n)\frac{e^{-\theta z}\left(\frac{p^*}{\theta^*(1-p^*)}\right)^{n-r}}{\left(\frac{1}{\theta^*\ (1-p^*)}\right)^{n}}dz$$

Thus, the UMVUE of function $\phi(p,\theta^*)$ in $g(t;p,\alpha,\theta^*)$ is given by $\psi(Z,R,n) = \frac{\alpha(Z,R,n)}{B(Z,R,n)}$, $\mathrm{B}(Z,R,n) \neq 0$. The UMVUE for some of the parametric functions are given as results below:

Result 1: *Let $T_1, T_2, \ldots, T_n$ be n censored sample from (10), then for fixed t , the UMVUE of the density function $g(t;p,\alpha,\theta^*)$ is obtained as:*

$$\phi_t(z,r,n) = \begin{cases} a(x)\frac{B(z-d(t),r-1,n-1)}{B(z,r,n)}, & t>\delta; r=1,2,\ldots,n; z>d(t), \\ \frac{B(z,r,n-1)}{B(z,r,n)}, & t=\delta; r=0,1,\ldots,n-1; z>d(t), \\ 0, & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \frac{r(r-1)}{\mathrm{n\ z}}e^{\alpha t}\left(1-\frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-2}, & t>\delta; z>\frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha}; r=1,2,\ldots,n \\ \frac{n-r}{n}, & t=\delta; r=0,1,\ldots,n-1, \\ 0, & \text{otherwise.} \end{cases}$$

Result 2: *The UMVUE of the variance of $\phi_x(Z,R,n)$ is obtained as:*

$$\widehat{var}\left[\phi_t(z,r,n)\right] = \begin{cases} \phi_t^2(z,r,n) - \phi_t(z,r,n)\,\phi_t(z-d(x),r-1,n-1), & t>\delta; z>2d(t)\text{ and} \\ & ; r=2,3,\ldots,n, \\ \phi_t^2(z,r,n), & t>\delta; d(t)<z<2d(t), \\ \phi_t^2(z,r,n) - \phi_t(z,r,n)\,\phi_t(,r,n-1), & t=d; r=0,1,\ldots,n-1, \\ 0, & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \left[\frac{r(r-1)}{\mathrm{n\ z}}e^{\alpha t}\left(1-\frac{\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-2}\right]^2 & \\ -\frac{r(r-1)^2(r-2)e^{2\alpha t}\left(1-\frac{\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-2}\left(1-\frac{\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha\left(z-\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)\right)}\right)^{r-3}}{n(n-1)\ z\left(z-\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)\right)}, & t>\delta; r=2,3,\ldots,n\text{and} \\ & ; z>\frac{2e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right) \\ \left[\frac{r(r-1)}{\mathrm{n\ z}}e^{\alpha t}\left(1-\frac{\frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-2}\right]^2, & t>\delta\text{and} \\ & ; \left(e^{\alpha(t-\delta)}-1\right) < \frac{\alpha z}{e^{\alpha\delta}} < \left(e^{\alpha(t-\delta)}-1\right), \\ \frac{r(n-r)}{n^2(n-1)}, & t=\delta; r=0,1,\ldots,n-1 \\ 0, & \text{otherwise.} \end{cases}$$

Result 3: *For fixed z and r , the UMVUE of the reliability function $R(t) = p(T > t)$, $t \geq d$ is obtained as: $\overline{G}_T(t; p, \propto, \theta^*) = e^{-\frac{\zeta}{\alpha}\left(e^{\alpha t} - 1\right)} R(t)$, where, the UMVUE of the reliability function for (10) is*

$$\begin{aligned}\widehat{\overline{G}_T}(t; p, \propto, \theta^*) &= \int_{x>t} \phi_x(z, r, n)\, dx \\ &= \begin{cases} \frac{r}{n}\left(1 - \frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-1}, z > \frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)} - 1\right) \\ 0, \qquad \text{otherwise.} \end{cases}\end{aligned}$$

Hence, the UMVUE of the reliability function $\hat{R}(t)$ of (4) is

$$\hat{R}(t) = \begin{cases} \frac{r}{n} e^{\frac{\zeta}{\alpha}\left(e^{\alpha t} - 1\right)}\left(1 - \frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-1}, z > \frac{e^{\alpha\delta}}{\alpha}\left(e^{\alpha(t-\delta)} - 1\right) \\ 0, \qquad \text{otherwise} \end{cases}$$

Result 4: *The UMVUE of the variance of $\widehat{R}(t)$ is obtained as:*

$$\widehat{var}\left[\hat{R}(t)\right] = e^{2\frac{\zeta}{\alpha}\left(e^{\alpha t} - 1\right)}\ \widehat{var}\left[\widehat{\overline{G}_T}(t; p, \propto, \theta^*)\right]$$

where,

$$\widehat{var}\left[\widehat{\overline{G}_T}(t; p, \propto, \theta^*)\right]$$

$$= \begin{cases} \left[\frac{r}{n}\left(1 - \frac{e^{\alpha\delta}}{\alpha}\frac{\left(e^{\alpha(t-\delta)}-1\right)}{z}\right)^{r-1}\right]^2 - \frac{r(r-1)}{n(n-1)}\left(1 - \frac{2e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-1}, z > \frac{2e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha} \\ \left[\frac{r}{n}\left(1 - \frac{e^{\alpha\delta}\left(e^{\alpha(t-\delta)}-1\right)}{\alpha z}\right)^{r-1}\right]^2, \qquad \frac{\left(e^{\alpha(t-\delta)}-1\right)}{\alpha} < \frac{\alpha z}{e^{\alpha\delta}} < \frac{2\left(e^{\alpha(t-\delta)}-1\right)}{\alpha} \\ 0, \qquad \text{otherwise.} \end{cases}$$

Remark 1: The UMVUE of the density function (3) is $\phi_x(z, r, n)$ we may get as

$$\phi_x(z, r, n) = -\left[\frac{d\hat{R}(t)}{dt}\right]_{t=x}$$

$$= \begin{cases} \frac{r}{n} e^{\alpha x} e^{\frac{\zeta}{\alpha}\left(e^{\alpha x} \ 1\right)}\left(1 - \frac{e^{\alpha\delta}\left(e^{\alpha(x-\delta)} \ 1\right)}{\alpha z}\right)^{r-2}\left[\frac{(r-1)}{z} - \zeta\left(1 - \frac{e^{\alpha\delta}\left(e^{\alpha(x-\delta)} \ 1\right)}{\alpha z}\right)\right], \\ \qquad\qquad x > \delta; z > \frac{e^{\alpha\delta}\left(e^{\alpha(x-\delta)}-1\right)}{\alpha}; r = 1, 2, \ldots, n \\ \frac{n-r}{n}, \qquad\qquad x = \delta; r = 0, 1, \ldots, n-1 \\ 0, \qquad\qquad \text{otherwise.} \end{cases}$$

4. Application

In this section, we consider a randomly censored real data set on patients of acute myelogenous leukemia reported to the International Bone Marrow Transplant Registry from

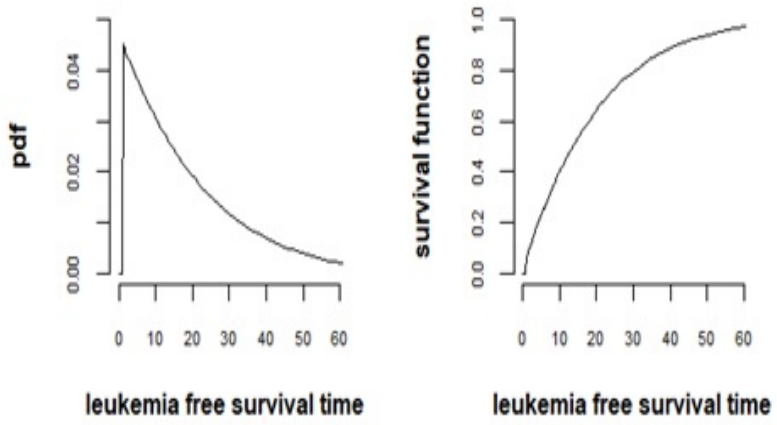


Figure 1: Plots of density function and survival function for leukemia free survival time data..

Klein and Moeschberger (2003). A sample of fifty patients had an allogeneic (allo) bone marrow transplant where marrow from an HLA (Histocompatibility Leukocyte Antigen) matched sibling was used to replenish their immune systems. The leukemia free survival times (in months) for 50 transplant patients were: 0.030, 0.493, 0.855, 1.184, 1.283, 1.480, 1.776, 2.138, 2.500, 2.763, 2.993, 3.224, 3.421, 4.178, 4.441^+, 5.691, 5.855^+, 6.941^+, 6.941, 7.993^+, 8.882, 8.882, 9.145^+, 11.480, 11.513, 12.105^+, 12.796, 12.993^+, 13.849^+, 16.612^+, 17.138^+, 20.066, 20.329^+, 22.368^+, 26.776^+, 28.717^+, 28.717^+, 32.928^+, 33.783^+, 34.211^+, 34.770^+, 39.539^+, 41.118^+, 45.033^+, 46.053^+, 46.941^+, 48.289^+, 57.401^+, 58.322^+ and 60.625^+. The + denotes a censored observation.

The distribution fit of the data was carried out for all those uncensored observations, and Gompertz distribution found to be a good fit. Many other authors have also concluded this, and hence the details are not discussed further here. Figure 4.1 show the pdf and survival functions curve for the leukemia free survival time data. Table 4.1 gives a summary of estimates of models for some selected values of parameters of the leukemia free survival time data. The entry in square brackets is the standard error (SE) of the estimate.

A visual look at the data supports the fact that there are few early failure observations. Since the inliers detection is not part of this paper, we have not discussed the same here. Note that the estimate of p and θ is comparable in instantaneous and early failure model and the standard error is very small for estimates. The UMVU estimates of the pdf and

Table 1: Summary of estimates of models for the leukemia free survival times: Estimates for $\alpha = 0.0057$ & $\zeta = 0.01271$

Estimator (SE)	Instantaneous Failure Model d = 0	Early Failure Model d = 1.0
MLE of p	1.00000 (-)	0.98630 (0.03707)
MLE of θ^*	0.04859 (0.00687)	0.04794 (0.01184)
MLE of θ	0.03589 (0.00687)	0.03531 (0.01184)
95% CI for p	-	(0.91360, 1.00000)
95% CI for θ^*	(0.03512, 0.06206)	(0.02469, 0.07111)
95% CI for θ	(0.02242, 0.04935)	(0.01210, 0.05851)
UMVUE of pdf in (10):		
$\phi_2(z, r, n)$	0.04817 (0.00688)	0.04461 (0.00677)
$\phi_{20}(z, r, n)$	0.05337 (0.00762)	0.04943 (0.00750)
$\phi_{50}(z, r, n)$	0.06332 (0.00904)	0.05864 (0.00890)
UMVUE of pdf in (3):		
at t =2	0.03306 (0.00320)	0.03187 (0.005430)
at t =20	0.01901 (0.00021)	0.01864 (0.004353)
at t =50	0.00613 (0.000435)	0.00620 (0.005603)
UMVUE of reliability function $\overline{G}_T(t; p, \propto, \theta^*)$:		
at t =2	0.90845 (0.01247)	0.89653 (0.03296)
at t =20	0.36031 (0.05281)	0.36866 (0.05282)
at t =50	0.05845 (0.02340)	0.06393 (0.02419)
UMVUE of reliability function $R(t)$:		
at t =2	0.93197 (0.02401)	0.91974 (0.02481)
at t =20	0.47160 (0.03063)	0.48253 (0.03166)
at t =50	0.12191 (0.04881)	0.13333 (0.05045)

reliability function decreases as point of time increases. It is seen that as number of inliers increases, the estimate of parameters decreases.

References

[1] Bavagosai, P. and Muralidharan, K. (2016). Some inferential study on inliers in Lindley distribution. *International Journal of Statistics and Reliability Engineering*, 3(2): 108 129.

[2] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality and on the new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society A*, 115: 513 580.

[3] Jani, P. N. and Singh, A. K. (1995). Minimum variance unbiased estimation in multiparameter exponential family of distributions. *Metron*, 53: 93 106.

[4] Jayade, V. P. and Prasad, M. S. (1990). Estimation of parameters of mixed failure time distribution. *Communication in Statistics Theory Methods*, 19(12): 4667 4677

[5] Kale, B. K. (2003). Modified failure time distributions to accommodate instantaneous and early failures. *Industrial Mathematics and Statistics*, Ed. J. C. Misra, Narosa Publishing House, New Delhi: 623 648.

[6] Kale, B. K. and Muralidharan, K. (2000). Optimal estimating equations in mixture distributions accommodating instantaneous or early failures. *Journal of the Indian Statistical Association*, 38: 317 329.

[7] Kale, B. K. and Muralidharan, K. (2008). Maximum Likelihood estimation in presence of inliers. *Journal of Indian Society for Probability and Statistics*, 10: 65 80.

[8] Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*, Second Edition, Springer.

[9] Kleyle, R. M. and Dahiya, R. L. (1975). Estimation of parameters of mixed failure time distribution from censored data. *Communication in Statistics Theory and Methods*, 4(9): 873 882.

[10] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, Second Edition, Wiley: Hoboken, NJ, USA.

[11] Meeker, W. Q., and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.

[12] Muralidharan, K. (2010). Inlier prone models: A review. *ProbStat Forum*, 3: 38 51.

[13] Muralidharan, K. and Arti, M. (2008). Analysis of instantaneous and early failures in Pareto distribution. *Journal of statistical theory and Applications*, 7: 187 204.

[14] Muralidharan, K. and Arti, M. (2013). Inlier proness in Normal distribution. *Reliability: Theory and Applications*, 8(1): 86 99.

[15] Muralidharan, K. and Bavagosai, P. (2016a). Some inferential studies on inliers in Gompertz distribution. *Journal of Indian Society for Probability and Statistics*, 17: 35 55.

[16] Muralidharan, K. and Bavagosai, P. (2016b). A revisit to early failure analysis in life testing. *Journal of the Indian Statistical Association*, 54(1 & 2): 43 69.

[17] Muralidharan, K. and Lathika, P. (2006). Analysis of instantaneous and early failures in Weibull distribution. *Metrika*, 64(3): 305 316.

[18] Roy, J. and Mitra, S. K. (1957). Unbiased minimum variance estimation in a class of discrete distributions. *Sankhya*, 18: 371 378.

[19] Shinde, R. L. and Shanubhogue, A. (2000). Estimation of parameters and the mean life of a mixed failure time distribution. *Communications in Statistics Theory and Methods*, 29(1): 2621 2642.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 35-48.

q-Gompertz Distribution and its Applications

A. Mohammed Shabeer and K. Jayakumar
Department of Statistics, University of Calicut, Kerala, India
Email:jkumar19@rediffmail.com

Abstract

In this paper, we introduce a new generalization of Gompertz distribution, namely q-Gompertz distribution and study its properties. We show that for large values of the parameter, distribution belong to class L. Expressions for moments, mode and quantile function are derived. Characterizations of the new distribution are obtained. The unknown parameters of the distribution are estimated using maximum likelihood method. The new distribution is fitted to a real data set to show the flexibility/competency of the model.

Key words: Gompertz distribution, Mortality Rate, Maximum Likelihood, Order Statistics, Pathway models.

1. Introduction

Gompertz distribution was proposed by Gompertz in 1825 and he showed that age specific mortality rates increase exponentially with age, over much of adult life span. Gompertz distribution is applied in various contexts of mortality studies and lifetime analysis. Gompertz curve has been utilized as a growth curve (see, Winsor (1932)). A random variable X on $(0, \infty)$ is said to have a Gompertz distribution and write X $\sim$ Gompertz(η, b), if its probability density function (pdf) is

$$f(x) = \eta b e^{-\eta(e^{bx}-1)} \quad \eta > 0, b > 0. \tag{1}$$

Many authors have studied the q-analog of distributions, that is, introducing pathway parameter to the existing distribution. The addition of this parameter stretches the distribution and improve the flexibility of the distribution for modelling data. The word

q-distribution is used, because the parameter added to the base distribution is q through this construction. For details of study on q-analogs such as q-Weibull, q-Logistic, q-Mittag-Leffler and q-Esscher transformed Laplace distribution see, Jose and Naik(2009), Mathai and Provost(2006), Sebastian and Gorenflo(2016) and Rimsha and George(2018).
A wide variety of socio economic variables have distributions that are heavy-tailed. The motivation of this study is to introduce a new statistical distribution which has thicker tail compared to Gompertz distribution and can be used to provide good fit to some real data sets.

1.1. Pathway models

In this section, we present some basic aspects of pathway models introduced by Mathai(2005). Let us consider a type-2 beta density given by,

$$f(x) = \frac{\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta-\alpha)} x^{\alpha-1}(1+x)^{-\beta}, \quad 0 < x < \infty, \alpha > 0, \beta > 0. \tag{2}$$

Now q-analog of (2) is given by,

$$f_1(x) = c_1 x^{\alpha-1}(1+\eta(q-1)x^{\delta})^{\frac{-\beta}{q-1}}, \quad 0 < x < \infty, \alpha > 0, \beta > 0, q > 1, \eta > 0, \delta > 0 \tag{3}$$

where c_1 is normalizing constant given by,

$$c_1 = \frac{\delta \eta^{\frac{\alpha}{\delta}}(q-1)^{\frac{\alpha}{\delta}}\Gamma(\frac{\beta}{q-1})}{\Gamma(\frac{\alpha}{\delta})\Gamma(\frac{\beta}{q-1}-\frac{\alpha}{\delta})}.$$

Here q is the pathway parameter. For $q < 1$, we can replace $(q-1)$ in (3) by $-(1-q)$ and we get,

$$f_2(x) = c_2 x^{\alpha-1}(1-\eta(1-q)x^{\delta})^{\frac{\beta}{1-q}}, \quad 1-\eta(1-q)x > 0, \alpha > 0, \beta > 0, q < 1, \delta > 0, \tag{4}$$

where c_2 is,

$$c_2 = \frac{\eta^{\frac{\alpha}{\delta}}(1-q)^{\frac{\alpha}{\delta}}\Gamma(\frac{\beta}{q-1}+\frac{\alpha}{\delta}+1)}{\Gamma(\frac{\alpha}{\delta})\Gamma(\frac{\beta}{1-q}-\frac{\alpha}{\delta})}.$$

Note that putting $q = 0, \eta = 1, \beta$ replaced by $\beta - 1$, we get type-I beta distribution. For various sub-models of (3) and (4), see Mathai and Haubold (2008).
Rest of the paper is organized as follows: In Section 2, we propose a new generalization of Gompertz distribution, namely q-Gompertz distribution based on path way models described above and discuss the nature of density function, distribution function, hazard rate function and showed the membership in class L of the q-Gompertz distribution. Various structural properties of q-Gompertz distribution such as moments, cumulants, quantile

function and mode are studied in section 3. Characterizations of q-Gompertz distribution are obtained in Section 4. In Section 5, we study the estimation of parameters of the q-Gompertz distribution, using the method of Maximum Likelihood. In Section 6, we have fitted the model to a real life data to show the flexibility of the new distribution. Concluding remarks are presented in Section 7.

2. The q-Analoque of the Gompertz distribution

Here we propose a q-analogue of the Gompertz distribution using the transformation $y = ln(x)$ in (3) and (4) and putting $\alpha = \delta = b, \beta = 1$. Then for $x > 0$ and $q > 1$, (3) reduces to,

$$f_3(y) = \frac{(1+\eta(q-1))^{\frac{-q+2}{q-1}} b\eta(q-1)\Gamma(\frac{1}{q-1})}{\Gamma(\frac{1}{q-1}-1)} e^{by}(1+\eta(q-1)e^{by})^{\frac{-1}{q-1}},$$
$$y > 0, q > 1, \eta > 0, b > 0. \quad (5)$$

If we restrict q between 1 and 2, equation (5) reduce to

$$f_4(y) = (1+\eta(q-1))^{\frac{-q+2}{q-1}} b\eta(2-q)e^{by}(1+\eta(q-1)e^{by})^{\frac{-1}{q-1}} \quad , y > 0, 1 < q < 2, \eta > 0, b > 0. \quad (6)$$

Now for $q < 1$, we have,

$$f_5(y) = \frac{(1-\eta(1-q))^{\frac{q-2}{1-q}} b\eta(1-q)\Gamma(\frac{1}{1-q}+2)}{\Gamma(\frac{1}{1-q}+1)} e^{by}(1-\eta(1-q)e^{by})^{\frac{1}{1-q}}$$
$$, 1-\eta(1-q)e^{by} > 0, q < 1, \eta > 0, b > 0. \quad (7)$$

which further reduces to,

$$f_5(y) = (1-\eta(1-q))^{\frac{q-2}{1-q}}(2-q)b\eta e^{by}(1-\eta(1-q)e^{by})^{\frac{1}{1-q}},$$
$$0 < y < \frac{1}{b} ln(\frac{1}{\eta(1-q)}), q < 1, \eta > 0, b > 0. \quad (8)$$

Both equation (6) and (8) reduce to classical Gompertz distribution, that is,

$$f(y) = \eta b e^{-\eta(e^{by}-1)}$$

as $q \to 1$.

In Figure 1, we have presented the density function of q-Gompertz for different values of q between 1 and 2 for $\eta = .5$ and $b = 1$. From graph we can observe that the curves get thicker tails as q increases from 1 to 2. Also note that as the values of q decreases from 2

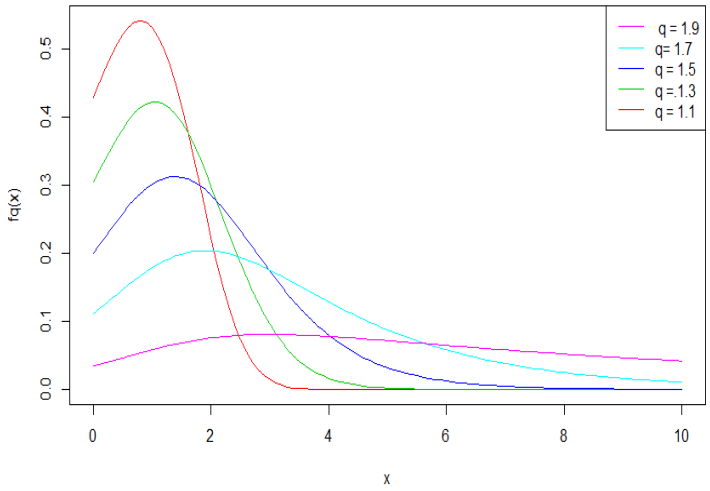


Figure 1: Pdf of q-Gompertz distribution for q between 1 and 2.

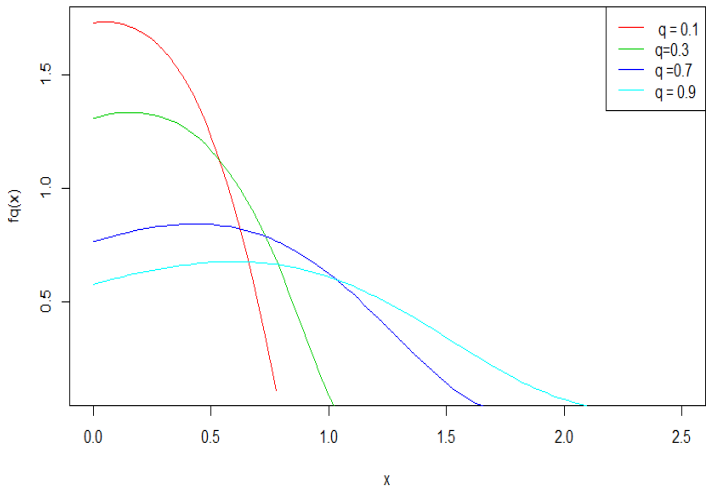


Figure 2: Pdf for different values of $q < 1$.

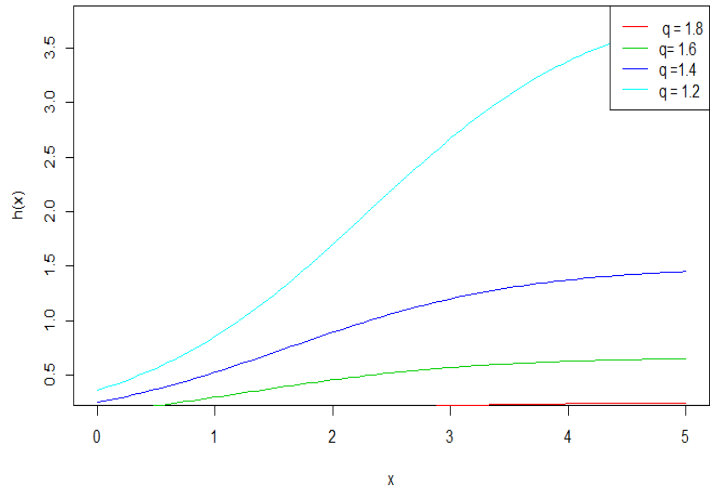


Figure 3: Hazard rate function for different values of q.

to 1, the peakedness increases and slowly converge to Gompertz density. In Figure 2, we have presented the density function of q-Gompertz for different values of q between 0 and 1 for $\eta = .5$ and $b = 1$. It can be seen that as q moves from 0 to 1, the peakedness reduces and slowly converge to Gompertz density.
The cumulative distribution function of q-Gompertz for $q > 1$ is given by,

$$F(y) = \left(1 - (1 + \eta(q-1))^{\frac{-q+2}{q-1}}(1 + \eta(q-1)e^{by})^{\frac{q-2}{q-1}}\right), \tag{9}$$

and for $q < 1$ is

$$F(y) = \left(1 - (1 - \eta(1-q))^{\frac{q-2}{1-q}}(1 - \eta(1-q)e^{by})^{\frac{2-q}{1-q}}\right). \tag{10}$$

Thus, hazard rate function is given by (for $q > 1$)

$$\mu(y) = b\eta(2-q)e^{by}(1 + \eta(q-1)e^{by})^{-1}, \tag{11}$$

and for $q < 1$,

$$\mu(y) = (2-q)b\eta e^{by}(1 - \eta(1-q)e^{by})^{-1} \quad 0 < y < \frac{1}{b}ln(\frac{1}{b(1-q)}). \tag{12}$$

The graph of hazard rate function for various values of q is given in Figure 3. We can observe that the hazard rate function is non-decreasing.

An important class of distributions used in risk theory and queuing theory is the class L distributions.

Definition 2.1. *A distribution F belongs to the class L if*

$$\lim_{x\to\infty} \frac{1-F(x-y)}{1-F(x)} = 1, \forall y \in R.$$

Lemma 2.1. *A distribution $F \in L$ if and only if $\lim_{x\to\infty} \mu(x) = 0$, where $\mu(x)$ is the hazard rate function .* For details of Lemma 2.1, see Klüppelberg (1988).

Theorem 2.1. *The q-Gompertz distribution belong to class L as $q \to 2$.*
Proof. From the hazard rate function (11), we have

$$\lim_{y\to\infty} \mu(y) = \frac{(2-q)b}{(q-1)}. \tag{13}$$

Now

$$\lim_{q\to 2} \lim_{y\to\infty} \mu(y) = 0.$$

□

3. Structural properties

In this Section, we derive expressions for moments and cumulants of q-Gompertz distribution. Expressions for quantile function and mode are also derived.

3.1. Moments

Let random variable Y follows q-Gompertz distribution. Then we can obtain the r^{th} moment of Y from its moment generating function. Let $\mu_r^{(y)}$ denote r^{th} moment of Y. Then,

$$\mu_r^{(y)} = \frac{\partial^r}{\partial t^r}[E(e^{tY})] \quad at \quad t = 0. \tag{14}$$

Also note that we get q-anlog of Gompertz distribution using the transformation $Y = ln(X)$ where X has the pdf (3) and (4). Then,

$$E(e^{tln(X)}) = E(X^t). \tag{15}$$

Thus,

$$\mu_r^{(y)} = \frac{\partial^r}{\partial t^r}[E(e^{tY})] = \frac{\partial^r}{\partial t^r}[E(X^t)] \quad at \quad t = 0. \tag{16}$$

But from (3), letting $\beta = 1, \alpha = \delta = b$ we have,

$$E(X^t) = \frac{\Gamma(\frac{t}{b}+1)\Gamma(\frac{1}{q-1}-\frac{t}{b}-1)}{\eta^{\frac{t}{b}}(q-1)^{\frac{t}{b}}\Gamma(\frac{1}{q-1}-1)}. \tag{17}$$

Then,

$$\mu_r^{(y)} = \frac{(1+\eta(q-1))^{\frac{2-q}{q-1}}}{\Gamma(\frac{1}{q-1}-1)}\frac{\partial^{r-1}}{\partial t^{r-1}}\frac{\Gamma(\frac{t}{b}+1)\Gamma(\frac{1}{q-1}-\frac{t}{b}-1)}{\eta^{\frac{t}{b}}(q-1)^{\frac{t}{b}}}$$
$$\times\left(\frac{1}{b}\psi(\frac{t}{b}+1)-\frac{1}{b}\psi(\frac{1}{q-1}-\frac{t}{b}-1)-\frac{1}{b}ln(\eta)-\frac{1}{b}ln(q-1)\right) \quad at \quad t=0, \tag{18}$$

where the $\psi(.)$ is the logarithmic derivative of the gamma function. Thus,

$$\mu_r^{(y)} = \sum_{n=0}^{r-1}\binom{r-1}{n}\mu_n^{(y)}K_{r-n}^{y)}, \tag{19}$$

where $K_{r-n}^{y)}$ is the $(r-n)^{th}$ cumulant of Y. Here

$$K_{r-n}^{(y)} = \frac{\partial^{r-n}}{\partial t^{r-n}}log[E(X^t)] \quad at \quad t=0$$
$$= \frac{\partial^{r-n-1}}{\partial t^{r-n-1}}\left(\frac{1}{b}\psi(\frac{t}{b}+1)-\frac{1}{b}\psi(\frac{1}{q-1}-\frac{t}{b}-1)-\frac{1}{b}ln(\eta)-\frac{1}{b}ln(q-1)\right), \quad at \quad t=0, \tag{20}$$

where,

$$K_m^{(y)} = (m-1)!\left((-1)^m\zeta(m,1)+\zeta(m,\frac{1}{q-1}-1)\right), \tag{21}$$

where $\zeta(.,.)$ is the generalized Riemann zeta function (Mathai (1993)). Moments of Y can be obtained recursively using equation (19) and (21). Similarly for $q < 1$, from (8) we have,

$$E(X^t) = \frac{\Gamma(\frac{t}{b}+1)\Gamma(\frac{1}{1-q}+2)}{\eta^{\frac{t}{b}}(q-1)^{\frac{t}{b}}\Gamma(\frac{1}{1-q}+\frac{t}{b}+2)}, \tag{22}$$

for $\eta > 0, b > 0, q < 1, \frac{t}{b}+1 > 0, \frac{1}{1-q}+2 > 0, \frac{1}{1-q}+\frac{t}{b}+2 > 0$. Thus the form of (19) will be same but (20) changes, that is

$$K_{r-n}^{y} == \frac{\partial^{r-n-1}}{\partial t^{r-n-1}}\left(\psi(\frac{t}{b}+1)-\frac{1}{b}\psi(\frac{1}{q-1}+\frac{t}{b}+2)-\frac{1}{b}ln(\eta)-\frac{1}{b}ln(q-1)\right),$$
$$at \quad t=0, \tag{23}$$

where,

$$K_m^{y} = (m-1)!(-1)^m\left(\zeta(m,1)-\zeta(m,\frac{1}{q-1}+2)\right). \tag{24}$$

3.2. Quantiles

The quantile function of the random variable Y for $1 < q < 2$ is given by,

$$y_p = \frac{1}{b} ln \left(\frac{(1-p)^{\frac{q-1}{q-2}}(1+\eta(q-1))}{\eta(q-1)} - \frac{1}{\eta(q-1)} \right), \tag{25}$$

for $q < 1$,

$$y_p = \frac{1}{b} ln \left(\frac{1}{\eta(1-q)} - \frac{(1-p)^{\frac{1-q}{2-q}}(1-\eta(1-q))}{\eta(1-q)} \right). \tag{26}$$

3.3. Mode

For $0 < q < 2$ mode of the random variable Y is given by,

$$mode = \frac{1}{b} ln \left(\frac{1}{\eta(2-q)} \right).$$

3.4. Simulation

A random variable Y having q-Gompertz distribution can be simulated, for $1 < q < 2$ as,

$$Y = \frac{1}{b} ln \left(\frac{(1-U)^{\frac{q-1}{q-2}}(1+\eta(q-1))}{\eta(q-1)} - \frac{1}{\eta(q-1)} \right), \tag{27}$$

for $q < 1$,

$$Y = \frac{1}{b} ln \left(\frac{1}{\eta(1-q)} - \frac{(1-p)^{\frac{1-q}{2-q}}(1-\eta(1-q))}{\eta(1-q)} \right), \tag{28}$$

where $U \sim U(0,1)$.

4. Characterization

The following Propositions establishes a characterization of q-Gompertz in terms of the hazard function.

Proposition 4.1. *Let $X : \Omega \to (0, \infty)$ be a continuous random variable. The pdf of X, is (6) if and only if its hazard function $\mu(x)$ satisfies the differential equation*

$$\mu'(x) - b\mu(x) = -\frac{b^2\eta^2(q-1)(2-q)e^{2bx}}{(1+\eta(q-1)e^{bx})^2}.$$

Proof. If X has pdf (6), then clearly the above differential equation holds. Now, if the differential equation holds, then

$$\frac{d}{dx}\left(e^{-bx}\mu(x)\right) = (2-q)b\eta\frac{d}{dx}\left(\frac{1}{1-\eta(q-1)e^{by}}\right)$$

or, equivalently,

$$e^{-bx}\mu(x) = (2-q)b\eta\frac{1}{1-\eta(q-1)e^{by}}$$

which gives hazard function of q-Gompertz distribution for $q > 1$. □

Proposition 4.2. *Let $X : \Omega \to (0, \frac{1}{b}ln\left(\frac{1}{b(1-q)}\right))$ be a continuous random variable. The pdf of X, is (8) if and only if its hazard function $\mu(x)$ satisfies the differential equation*

$$\mu'(x) - b\mu(x) = \frac{b^2\eta^2(1-q)(2-q)e^{2bx}}{(1-\eta(1-q)e^{bx})^2}, 0 < x < \frac{1}{b}ln\left(\frac{1}{b(1-q)}\right)$$

with the boundary condition $lim_{x\to\frac{1}{b}ln\left(\frac{1}{b}(1-q)\right)}\mu(x) = (2-q)b\eta/(b-\eta)(1-q)$.

Proof. If X has pdf (8), then clearly the above differential equation holds. Now, if the differential equation holds, then

$$\frac{d}{dx}\left(e^{-bx}\mu(x)\right) = (2-q)b\eta\frac{d}{dx}\left(\frac{1}{1-\eta(1-q)e^{by}}\right)$$

or, equivalently,

$$e^{-bx}\mu(x) = (2-q)b\eta\frac{1}{1-\eta(1-q)e^{by}}$$

which gives hazard function of q-Gompertz distribution for $q < 1$. □

5. Estimation

Let $y_1, y_2, ..., y_n$ be an observed random sample from q-Gompertz distribution with unknown parameters η, b, q. The log-likelihood function for $1 < q < 2$ is given by,

$$ln(L(y,\eta,b,q)) = \frac{-n(q-2)}{(q-1)}ln(1+\eta(q-1)) + nln(2-q) + nln(\eta) + nln(b) + b\sum_{i=1}^{n} y_i - \frac{1}{q-1}\sum_{i=1}^{n} ln(1+\eta(q-1)e^{by_i}), \quad (29)$$

for $q < 1$,

$$ln(L(y,\eta,b,q)) = \frac{n(q-2)}{(1-q)} ln(1-\eta(1-q)) + nln(2-q) + nln(\eta) + nln(b) + b\sum_{i=1}^{n} y_i + \frac{1}{1-q}\sum_{i=1}^{n} ln(1-\eta(1-q)e^{by_i}). \quad (30)$$

The partial derivative of the log likelihood with respect to the parameters for $1 < q < 2$ are

$$\frac{\partial ln(L)}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n} y_i - \frac{1}{q-1}\sum_{i=1}^{n} \frac{\eta(q-1)y_i e^{by_i}}{1+\eta(q-1)e^{by_i}} \quad (31)$$

$$\frac{\partial ln(L)}{\partial \eta} = \frac{-n(q-2)}{\eta(q-1)+1} + \frac{n}{\eta} - \frac{1}{q-1}\sum_{i=1}^{n} \frac{(q-1)e^{by_i}}{1+\eta(q-1)e^{by_i}}, \quad (32)$$

$$\frac{\partial ln(L)}{\partial q} = \frac{-n}{2-q} - \frac{n}{(q-1)^2} ln(1+\eta(q-1)) - \frac{n\eta(q-2)}{q-1(1+\eta(q-1))} + \frac{1}{(q-1)^2}\sum_{i=1}^{n} ln(1+\eta(q-1)e^{by_i}) - \frac{1}{q-1}\sum_{i=1}^{n} \frac{\eta e^{by_i}}{1+\eta(q-1)e^{by_i}}. \quad (33)$$

The partial derivative of the log likelihood with respect to the parameters for $q < 1$ are

$$\frac{\partial ln(L)}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n} y_i - \frac{1}{q-1}\sum_{i=1}^{n} \frac{\eta(1-q)y_i e^{by_i}}{1-\eta(1-q)e^{by_i}} \quad (34)$$

$$\frac{\partial ln(L)}{\partial \eta} = \frac{-n(q-2)}{1-\eta(1-q)} + \frac{n}{\eta} - \frac{1}{1-q}\sum_{i=1}^{n} \frac{(1-q)e^{by_i}}{1-\eta(1-q)e^{by_i}}, \quad (35)$$

$$\frac{\partial ln(L)}{\partial q} = \frac{-n}{2-q} - \frac{n}{(1-q)^2} ln(1-\eta(1-q)) - \frac{n\eta(q-2)}{(1-q)(1-\eta(1-q))} + \frac{1}{(1-q)^2}\sum_{i=1}^{n} ln(1-\eta(1-q)e^{by_i}) - \frac{1}{1-q}\sum_{i=1}^{n} \frac{\eta e^{by_i}}{1-\eta(1-q)e^{by_i}}. \quad (36)$$

The Maximum Likelihood estimates(MLE) of (b,η,q) are the solution of the simultaneous equations $\frac{\partial LogL}{\partial b} = 0, \frac{\partial LogL}{\partial \eta} = 0, \frac{\partial LogL}{\partial q} = 0$. These equations cannot be solved analytically. So numerical technique such as Newton-Raphson method can be employed to get the MLEs.

6. Data analysis

In this section we consider a real life data and illustrate the flexibility of the q-Gompertz distribution. We fit the model to the data set and compare it with Gompertz distribution.

We consider a data consists of observations on breaking stress of carbon fibres(in Gba). This data had been studied previously by Nichols and Padgett(2006), Fatima and Roohi (2015) and Jayakumar et al. (2018) to compare generalizations of Pareto distributions. The data is given in Table 1.

Table 1: Breaking stress of carbon fibres data.

3.70	2.74	2.73	2.50	3.60	3.11	3.27	2.87	1.47	3.11
4.42	2.41	3.19	3.22	1.69	3.28	3.09	1.87	3.15	4.90
3.75	2.43	2.95	2.97	3.39	2.96	2.53	2.67	2.93	3.22
3.39	2.81	4.2	3.3	2.55	3.31	3.31	2.85	2.56	3.56
3.15	2.35	2.55	2.59	2.38	2.81	2.77	2.17	2.83	1.92
1.41	3.68	2.97	1.36	.98	2.76	4.91	3.68	1.84	1.59
3.19	1.57	.81	5.56	1.73	1.59	2	1.22	1.12	1.71
2.17	1.17	5.08	2.48	1.18	3.51	2.17	1.69	1.25	4.38
1.84	.39	3.68	2.48	.85	1.61	2.79	4.7	2.03	1.8
1.57	1.08	2.03	1.61	2.12	1.89	2.88	2.82	2.05	3.65

We now consider Total Time on Test (TTT) plot, a graphical method to identify the shape of the hazard rate function of the data. In Figure 4 we have presented the TTT plot the data. It is clear from the figure that the hazard rate function is increasing. We have fitted the q-Gompertz distribution for the data and compared it with Gompertz distribution having pdf (1).

For both the distributions we have computed log-Likelihood (-logL), Akaike Information Criterion(AIC), Bayesian Information Criterion(BIC) and Hannan-Quinn information criterion(HQIC) to check the goodness of fit of the model to the data. The computed values are presented in Table 2. Clearly the q-Gompertz distribution gives a better fit to this data. The fitted densities are presented in Figure 5.

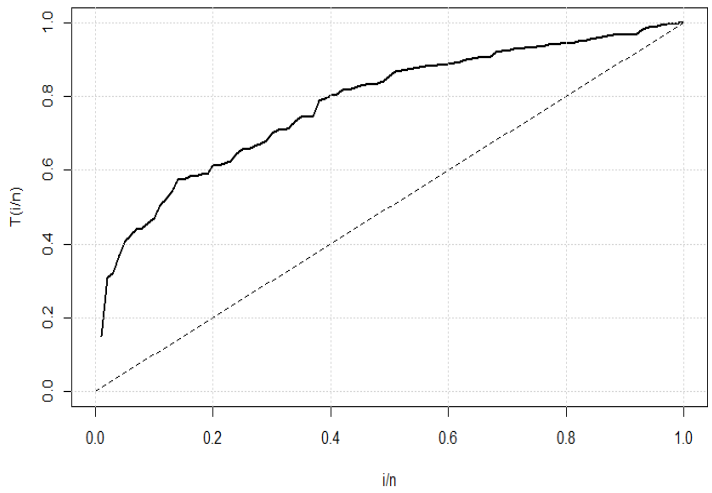


Figure 4: TTT plot

Figure 5: Estimated pdf.

Table 2: Parameter estimates and goodness of fit statistics for models fitted to the data.

Model	Estimates	-logL	AIC	BIC	HQIC
Gompertz	$\hat{\eta}$=.123	149.117	302.234	307.443	304.342
	$\hat{b}$=.7909				
q-Gompertz	$\hat{\eta}$=.026	142.0921	290.1842	297.9997	293.3472
	$\hat{b}$=1.6995				
	$\hat{q}$=1.5085				

7. Conclusion

In this paper, as a generalization of Gompertz distribution, q-Gompertz distribution is introduced. The new distribution belongs to class L. Characterizations of q-Gompertz distribution are obtained. It is shown that the new model is a competitor of Gompertz distribution, for modelling certain types of data sets. Also generation of random variates from the new model is simple. We expect that the new model may attract the attention of researchers, as a viable competitor of the Gompertz distribution.

Acknowledgement

The first author wishes to thank the University Grants Commission, New Delhi for the financial assistance in the form of Junior Research Fellowship.

References

[1] Fatima, A. and Roohi, A. (2015). Transmuted exponentiated Pareto-I distribution. Pakistan Journal of Statistics 32:6380.

[2] Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new method of determining the value of life contingencies. Philosophical Transactions of the Royal Society 115: 513-585.

[3] Jayakumar, K., Krishnan, B. and Hamedani, G. G. (2018). On a new generalization of Pareto distribution and its applications. Communications in Statistics - Simulation and Computation
DOI: 10.1080/03610918.2018.1494281.

[4] Jose, K. K. and Naik, S. R. (2009). On the q-Weibull distribution and its applications. Communications in StatisticsTheory and Methods 38: 912-926.

[5] Klüppelberg, C. (1988). Subexponential distributions and integrated tails. Journal of Applied Probability 25:132-141.

[6] Mathai, A. M. (1993). A Handbook of Generalized Special Functions for Statistical and Physical Sciences. Oxford University Press, Oxford.

[7] Mathai, A. M. (2005). A pathway to matrix-variate gamma and normal densities. Linear Algebra and its Applications 396:317328.

[8] Mathai, A.M and Provost, S.B.(2006). On q-logistic and related models. IEEE transactions in Reliability 55: 237-244.

[9] Mathai, A.M. and Haubold, H. J.(2008). On generalized distributions and pathways. Physics Letters A 372 : 21092133.

[10] Nichols, M. D. and Padgett, W. J. (2006). A bootstrap control chart for Weibull percentiles. Quality and Reliability Engineering International 22 :14151.

[11] Rimsha, H. and George, D.(2018). q-Esscher transformed Laplace distribution. Communications in StatisticsTheory and Methods 48: 1563-1578.

[12] Sebastian, N. and Gorenflo, R.(2016). Time series models associated with Mittag-Leffler type distributions and its properties. Communications in StatisticsTheory and Methods 45: 7210-7225.

[13] Winsor, C. P. (1932) The Gompertz curve as a growth curve. PNAS, USA, 18: 1-8.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 49-64.

Generalized Polya-Aeppli Process and Applications

K.K.Jose[1] and Shalitha Jacob[2]
[1]Department of Statistics, St.Thomas College Pala, Arunapuram,
Mahatma Gandhi University, Kottayam, Kerala, India-686574
Email:kkjstc@gmail.com
[2]Department of Mathematics
Mar Thoma College for Women Perumbavoor, Kerala, India
Email:shalithajacob@yahoo.in

Abstract

In this paper, we introduce a compound Poisson process with truncated negative binomial compounding distribution, called a Generalized Polya-Aeppli Process. We derive expressions for its p.m.f and discuss several properties. We propose a risk model with Generalized Polya-Aeppli process as the counting process. The joint distribution of the time to ruin and deficit at the time of ruin is derived. The differential equation of the ruin probability is given. As an example, we consider the case in which the claim size has an exponential distribution.

Key words: Generalized Polya-Aeppli distribution, Ruin probability, aggregate claims distribution, stop loss moment.

1 Introduction

The Poisson process is a stochastic counting process that appears in a large variety of daily life situations. But it is a good fit only when the count data at hand is equidispersed, that is, when the variance of the data is equal to the mean. It is observed that for many available count data, the sample variance is smaller or greater than the sample mean, which is referred to as under and over dispersion, respectively. This motivated the researchers to search for alternative models. As a result, there are two directions in which the Poisson process be generalized: by compounding and by mixing distributions. Mixing

is just a mechanism for constructing new distributions.

Starting from the parameterized distribution $g(x/\theta)$ of a random variable X with unknown parameter θ we may obtain a new family of distributions, if we allow parameter θ to be itself a random variable with distribution function $H(\theta)$. Then the unconditional distribution of X is said to be a mixture distribution and is given by

$$g(x) = \int g(x/\theta) dH(\theta).$$

Usually, $g(x/\theta)$ is called the mixed distribution and the parameters distribution $H(\theta)$ is called the mixing distribution. Mixtures are usually considered as alternative models that offer more flexibility. For discrete distributions "compounding" is commonly used in place of "mixing". The process of compounding creates a large class of distributions.

The compound distributions can be constructed as follows. Let M be a counting random variable and $X_1, X_2, X_3...$ be i.i.d random variables independent of M. Then the distribution of $S = X_1 + X_2 + ... + X_M$ is called a compound distribution and is given by

$$P_S(z) = \sum_{k=0}^{\infty} P(M = k) g^{*k}(z) \quad ,$$

where g^{*k} is the k-fold convolution of distribution of X. In this regard, the distribution of X is called the compounding distribution, while that of M is the compounded distribution.

The compound Poisson process is a generalization of the Poisson process obtained by compounding with a suitable distribution. It has wide applications in various fields such as transport, ecology, radiology, quality control, telecommunications etc. The compound Poisson process assures a better description for clustering of events.

The compound Poisson process $\{M(t), t \geq 0\}$ is given by the sum

$$M(t) = \sum_{i=1}^{N(t)} X_i \quad ,$$

where $N(t)$ is a homogenous Poisson process and $X_1, X_2, X_3...$ is a sequence of i.i.d random variables independent of $N(t)$. The distribution of X is called compounding distribution. Assume that the compounding random variable X has truncated geometric distribution. Then we get Polya-Aeppli process $\{M(t), t \geq 0\}$. Minkova (2004) introduced Polya-Aeppli process as a compound Poisson process with the truncated geometric compounding distribution. They showed that Polya-Aeppli process provides greater flexibility in modeling count data when it possesses overdispersion property.

The negative binomial model is one of the most popular models to count data. Among specific fields where negative binomial distribution have been found to provide useful representations may be mentioned in accident statistics, Econometrics, quality control and biometrics. In many cases, however, the entire distribution of counts is not observed. In particular, more often zeros are not observed. The negative binomial distribution often arises in practice where the zero group is truncated. It is a known fact that the truncated geometric distribution is a special case of the truncated negative binomial distribution. Therefore, in the Polya-Aeppli process instead of the geometric distribution we consider the truncated negative binomial as compounding distribution. Consequently, Generalized Polya-Aeppli process will be obtained.

2 Generalized Polya-Aeppli Distribution

Consider a random variable

$$M = X_1 + X_2 + ... + X_N,$$

where N has a Poisson distribution with parameter λ, independent of the i.i.d random variales $X_1, X_2, X_3....$ Suppose that $X_1, X_2, X_3...$ are truncated negative binomial with parameters r and $1-\rho,\ \rho \in [0,1]$ and $r > 0$.
The PMF and the PGF of the compounding random variable X are given by

$$P(X = x) = \frac{\binom{r+x-1}{x}\rho^x(1-\rho)^r}{1-(1-\rho)^r},\ x = 1, 2, ...,\ 0 < \rho < 1,\ r > 0.$$

and

$$P(s) = Es^X = \frac{(1-\rho s)^{-r} - 1}{(1-\rho)^{-r} - 1}.$$

Then, we can obtain the PGF and PMF of the random variable N as follows:

$$\Psi_M(s) = e^{-\lambda(1-P(s))} = e^{-\lambda(1-\frac{(1-\rho s)^{-r}-1}{(1-\rho)^{-r}-1})},$$

$$\begin{aligned} P(M = m) &= e^{-\lambda}, \quad m = 0 \\ &= e^{-\lambda}\rho^m \sum_{i=1}^{m}\sum_{k=1}^{i} \frac{(-1)^{i+k}\binom{i}{k}(\frac{\lambda}{(1-\rho)^{-r}-1})^i\binom{rk+m-1}{m}}{i!},\ m = 1, 2, \dots \end{aligned} \tag{1}$$

The notation that we used to refer generalized Polya-Aeppli distribution with parameters λ, ρ and r is $GPA(\lambda, \rho, r)$.

The cumulative distribution function of N is given by

$$\begin{aligned} F(0) &= e^{-\lambda}, \\ F(x) &= e^{-\lambda}\sum_{j=0}^{x}\sum_{i=1}^{x-j}\sum_{k=1}^{i} \frac{(-1)^{i+k}(\frac{\lambda}{(1-\rho)^{-r}-1})^i\binom{rk+n-j-1}{n-j}\rho^{n-j}}{i!},\ x \geq 1. \end{aligned} \tag{2}$$

3 Generalized Polya-Aeppli Process

Let $M(t)$ denotes the no of occurrence in the interval $(0,t]$. For the Generalized Polya-Aeppli process, $M(t)$ has a Generalized Polya-Aeppli distribution, $GPA(\lambda t,\rho,r)$ and is given by

$$P(M(t)=m)=\begin{cases} e^{-\lambda t}, & m=0 \\ e^{-\lambda t}\rho^m \sum_{i=1}^{m}\sum_{k=1}^{i} \frac{(-1)^{i+k}\binom{i}{k}(\frac{\lambda t}{(1-\rho)^{-r}-1})^i\binom{rk+m-1}{m}}{i!}, & m=1,2,\dots \end{cases} \tag{3}$$

To express $\{M(t), t\geq 0\}$ is a Generalized Polya-Aeppli process with parameters λ,ρ and r, we use the notation $M(t)\sim \text{GPAP}(\lambda,\rho,r)$.

Remark 3.1. *Taking into account the equality* $\sum_{k=1}^{i}(-1)^k\binom{i}{k}\binom{k+m-1}{m}=(-1)^i\binom{m-1}{i-1}$, *when* $r=1$, *the Generalized Polya-Aeppli process* $GPAP(\lambda,\rho,1)$ *reduces to the Polya-Aeppli process. If* $r=1$ *and* $\rho=0$, *then it is a homogeneous Poisson process with intensity* λ. *Thus the Poisson process and the Polya-Aeppli process are the special cases of Generalized Polya-Aeppli process.*

Definition 3.1. *A counting process* $\{M(t), t\geq 0\}$ *is called a Generalized Polya-Aeppli process with parameters* λ,ρ *and* r *if it satisfies*

1 *M(0)= 0, i.e, it starts at zero;*

2 *M(t) has independent increments;*

3 *For each* $t>0$, *the number of occurrence* $M(t)$ *in any interval of length* t *has Generalized Polya-Aeppli distribution with parameters* λt, ρ *and* r.

We have

$$EM(t)=\frac{r\rho\lambda t}{(1-\rho)(1-(1-\rho)^r)} \tag{4}$$

and

$$Var(M(t))=\frac{r\rho\lambda t(1+r\rho)}{(1-(1-\rho)^r)(1-\rho)^2}. \tag{5}$$

using (4) and (5) it can be shown that autocovariance between $N(s)$ and $N(t), s<t$ is

$$c(s,t))=\frac{r\rho\lambda s(1+r\rho)}{(1-(1-\rho)^r)(1-\rho)^2}.$$

Hence the autocorrelation function

$$\begin{aligned}\rho(s,t)&=\frac{c(s,t)}{\sqrt{\text{var }(M(s)).\ \text{var}(M(t))}}\\ &=\left(\frac{s}{t}\right)^{\frac{1}{2}}.\end{aligned}$$

It can be shown that

$$\rho(t,t') = \frac{\min(t,t')}{(t,t')^{\frac{1}{2}}}.$$

This is the autocorrelation function of the process.
The Fisher index of dispersion is given by

$$\begin{aligned} FI(M(t)) &= \frac{\text{var}(M(t))}{EM(t)} \\ &= \frac{1+r\rho}{1-\rho} \\ &= 1 + \frac{(1+r)\rho}{1-\rho} > 1. \end{aligned}$$

If r=0 the Generalized Polya-Aeppli process is over dispersed, which offer more flexibility in modeling count data compared to the standard Poisson process.

4 Alternate Definition of Generalized Polya-Aeppli process

In this section we define Generalized Polya-Aeppli process as a pure birth process.

Definition 4.1.

A counting process $\{M(t), t \geq 0\}$ is called a Generalized Polya-Aeppli process with parameters λ, ρ and r if

1. $M(0) = 0$;
2. $M(t)$ has stationary independent increments;
3. the state transition probabilities are defined as follows:

$$P\left(M(t+h) = m / M(t) = n\right) = \begin{cases} 1 - \lambda h + o(h), & m = n \\ \frac{\binom{r+i-1}{i}\rho^i(1-\rho)^r}{1-(1-\rho)^r}\lambda h + o(h), & m = n+i, i = 1,2,\ldots \end{cases} \tag{6}$$

for every $n = 0, 1, \ldots$ where $o(h) \to 0$ as $h \to 0$.

Let $P_m(t) = P(M(t) = m), \quad m = 0, 1, 2, \ldots.$
From the above postulates we get the following Kolmogorov forward equations:

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t), \\ P_m'(t) &= -\lambda P_m(t) + \frac{\lambda}{(1-\rho)^{-r} - 1} \sum_{i=1}^{m} \binom{r+i-1}{i} \rho^i P_{m-i}(t), \ m \geq 1, \end{aligned} \tag{7}$$

with initial conditions.

$$P_0(0) = 1 \text{ and } P_m(0) = 0, \quad m = 1, 2, \ldots$$

Using the equations of (7) we shall have the following differential equation for $\psi_{M(t)}$

$$\frac{\partial}{\partial t}\psi_{M(t)}(s) = -\lambda(1 - P(s))\psi_{M(t)}(s).$$

With $\psi_{M(t)}(1) = 1$, the above differential equation admit of the solution

$$\psi_{M(t)}(s) = \mathrm{e}^{-\lambda t(1-P(s))}$$

But this is the PGF of the GPAP(λ, ρ, r), which leads to (3).
Therefore two definitions of the generalized Polya-Aeppli Process are equivalent.

5 Properties of GPAP(λ, ρ, r)

In this section, we discuss some properties of GPAP(λ, ρ, r).

5.1 Interarrival Times Distributions

Theorem 5.1. *For the Generalized Polya-Aeppli process GPAP(λ, ρ, r) interval of time Z_1 to the first occurrence is exponential random variable with parameter λ and Z_2, the time between 1st and 2nd occurrence of the process takes value zero with probability $1 - \frac{r\rho}{(1-\rho)^{-r}-1}$ and with probability $\frac{r\rho}{(1-\rho)^{-r}-1}$, exponentially distributed with parameter λ.*

Proof. Let Z_k be the time of the k^{th} arrival, for $k = 1, 2, \ldots$. Let $W_n = \sum_{i=1}^{n} Z_i$ be the waiting time up to the n^{th} occurrence and $M(t)$ denote the number of occurrence up to the instant t. For any $t \geq 0$ and $n \geq 0$, we have the following relation.

$$P(M(t) = n) = P(W_n \leq t) - P(W_{n+1} \leq t), n = 0, 1, \ldots \tag{8}$$

For $n = 0$, equation (8) yields

$$P(M(t) = 0) = 1 - P(Z_1 \leq t) = 1 - F_{Z_1}(t), \tag{9}$$

where $F_{Z_1}(t)$ is the distribution function of Z_1.
According to (3),

$$P(M(t) = 0) = \mathrm{e}^{-\lambda t}. \tag{10}$$

From (9) and (10), we get

$$F_{Z_1}(t) = 1 - \mathrm{e}^{-\lambda t}.$$

Hence the density function of Z_1 is

$$f_{Z_1}(t) = \lambda \mathrm{e}^{-\lambda t}, t \geq 0.$$

i.e, Z_1 is exponential random variable with parameter λ.
Now from (8), for $n = 1$, we have,

$$P(M(t) = 1) = P(W_1 \leq t) - P(W_2 \leq t).$$

Then taking Laplace transform on both sides of above equation, we get

$$\frac{r\rho}{(1-\rho)^{-r}-1}\frac{\lambda}{\lambda+s}\frac{s}{s+\lambda} = LS_{W_1}(s) - LS_{W_2}(s).$$

On simplification, we get

$$LS_{Z_1+Z_2}(s) = \frac{\lambda}{\lambda+s}\left[1 - \frac{r\rho}{(1-\rho)^{-r}-1} + \frac{r\rho}{(1-\rho)^{-r}-1}\frac{\lambda}{s+\lambda}\right],$$

It follows that Z_1 and Z_2 are independent. Furthermore, Z_2 has an exponential distribution with parameter λ and takes value zero with probability $1 - \frac{r\rho}{(1-\rho)^{-r}-1}$.

Thus, the P.D.F of of Z_2 is given by

$$f_{Z_2}(t) = \left(1 - \frac{r\rho}{(1-\rho)^{-r}-1}\right)\delta_0(t) + \left(\frac{r\rho}{(1-\rho)^{-r}-1}\right)\lambda \mathrm{e}^{-\lambda t}, t \geq 0,$$

where $\delta_0(t)$ is the dirac delta function. □

Remark 5.1. *However the Laplace transform of the random variable $W_n, n = 3, 4, \ldots$ is too complicated. Therefore, distributions of remaining interarrival times cannot be expressed in an explicit form.*

5.2 The Waiting Time Distribution

Theorem 5.2. *The distribution function of the waiting time up to the n^{th} occurrence is given by*

$$F_{W_n}(t) = 1 - \mathrm{e}^{-\lambda t}\left(1 + \sum_{m=1}^{n-1} a_{m,t}\rho^m\right),$$

where

$$a_{m,t} = \sum_{i=1}^{m}\sum_{k=1}^{i}\frac{(-1)^{i+k}\binom{i}{k}\left(\frac{\lambda t}{(1-\rho)^{-r}-1}\right)^i}{i!}\binom{rk+m-1}{m}.$$

Proof. Let Z_n denotes the time between $(n-1)^{\text{th}}$ and n^{th} occurrence of the process, $n = 2, 3, \ldots$.
For any given integer $n \geq 1$ and time $t > 0$, the relation between waiting time up to the n^{th} occurrence W_n and counting random variable M(t), is given by

$$\{W_n \leq t\} = \{M(t) \geq n\}$$

Hence

$$P(M(t) \geq n) = P(W_n \leq t).$$

The cumulative distribution function of W_n is given by

$$\begin{aligned} F_{W_n}(t) &= P(W_n \leq t) \\ &= P(M(t) \geq n) \\ &= 1 - P(M(t) \leq n-1) \\ &= 1 - e^{-\lambda t}\left(1 + \sum_{m=1}^{n-1} \rho^m \sum_{i=1}^{m} \sum_{k=1}^{i} \frac{(-i)^{i+k}\binom{i}{k}(\frac{\lambda t}{(1-\rho)^{-r}-1})^i \binom{rk+m-1}{m}}{i!}\right) \\ &= 1 - e^{-\lambda t}\left(1 + \sum_{m=1}^{n-1} a_{m,t}\rho^m\right). \end{aligned}$$

□

5.3 Martingale Property

Theorem 5.3. *For $M(t) \sim GPAP(\lambda, \rho, r)$, the process $N(t) = M(t) - \frac{r\rho\lambda t}{(1-\rho)(1-(1-\rho)^r)}$ is a martingale with respect to $(\Omega, \mathcal{F}_t, \rho)$, where*

$$\mathcal{F}_t = \sigma\{M(s), 0 \leq s \leq t\}.$$

Proof. Since $E(M(t) - \frac{r\rho\lambda t}{(1-\rho)(1-(1-\rho)^r)}) = 0$ and $M(t)$ has independent increments, for any $t \geq s$ we have

$$\begin{aligned} E(N(t)/\mathcal{F}_s) &= E\left(M(t) - \frac{r\rho\lambda t}{(1-\rho)(1-(1-\rho)^r)}/\mathcal{F}_s\right) \\ &= E\left(M(t) - M(s) - \frac{r\rho\lambda(t-s)}{(1-\rho)(1-(1-\rho)^r)}/\mathcal{F}_s\right) \\ &\quad + M(s) - \frac{r\rho\lambda s}{(1-\rho)(1-(1-\rho)^r)} \\ &= E\left(M(t-s) - \frac{r\rho\lambda(t-s)}{(1-\rho)(1-(1-\rho)^r)}\right) \\ &\quad + M(s) - \frac{r\rho\lambda s}{(1-\rho)(1-(1-\rho)^r)} \\ &= N(s). \end{aligned}$$

Therefore, $\{N(t), t \geq 0\}$is a martingale. □

5.4 Relation between GPAP(λ, ρ, r) and Uniform distribution

Theorem 5.4. *Given that only one occurrence of the generalized Polya−Aeppli process, $\{M(t), t \geq 0\}$ has occurred in $[0, t]$, then the distribution of time interval of that occurrence is uniform in $[0, t]$.*

Proof. Let Z_1 denotes the time of first occurrence.

$$\begin{aligned} P(Z_1 \leq x | M(t) = 1 &= \frac{P(\text{ only one occurrence in } (0,x], \text{there was no occurrence in } (x,t])}{P(M(t)=1)} \\ &= \frac{P(M(x)=1, M(t-x)=0)}{P(M(t)=1)} \\ &= \frac{\frac{e^{-\lambda x} r\rho\lambda x \cdot e^{-\lambda(t-x)}}{((1-\rho)^{-r}-1)}}{\frac{e^{-\lambda t} r\rho\lambda t}{(1-\rho)^{-r}-1}} \\ &= \frac{x}{t}, 0 \leq x \leq t. \end{aligned}$$

□

6 Application to Risk Theory

Consider the standard risk model $\{X(t), t \geq 0\}$ of an insurance company given by

$$X(t) = ct - \sum_{k=1}^{M(t)} Y_k, \left(\sum_{1}^{0} = 0\right). \tag{11}$$

Here, c is a premium income per unit time and the claim size sequence $\{Y_i\}_{i=1}^{\infty}$ are i.i.d random variables having common distribution function F such that $F(0) = 0$ and with mean value μ, independent of the counting process $\{M(t), t \geq 0\}$.
We assume that the counting process $\{M(t), t \geq 0\}$ in the risk model given in (11) is a Generalized Polya-Aeppli process and the resulting risk model obtained is called the generalized Polya-Aeppli risk model.
The relative safety loading θ is given by

$$\theta = \frac{EX(t)}{E\sum_{k=1}^{M(t)} Y_k} = \left(\frac{c(1-\rho)(1-(1-\rho)^r)}{r\lambda\rho\mu} - 1\right)$$

and in connection with positive safety loading $\theta > 0, c > \frac{r\rho\mu\lambda}{(1-\rho)(1-(1-\rho)^r)}$.
Suppose that the time of ruin of the company with initial capital u is denoted by τ and is defined as $\tau = \inf\{t : X(t) + u < 0\}$. Here we take the convention that $\inf \phi = \infty$

The ruin probability of a company having initial capital u is defined as

$$\Psi(u) = P(\tau < \infty). \tag{12}$$

Then the non ruin probability is given by $\Phi(u) = 1 - \Psi(u)$.

The joint probability distribution $W(u,z)$ of the time to ruin τ and deficit at the time of ruin $D = |u + X(\tau)|$ is given by

$$W(u,z) = P(\tau < \infty, D \leq z), \quad z \geq 0, \tag{13}$$

It is obvious that

$$\lim_{z\to\infty} W(u,z) = \Psi(u). \tag{14}$$

We can obtain the following equation by using the postulates in (6).

$$W(u,z) = (1-\lambda h)W(u+ch,z) + \frac{\lambda}{(1-\rho)^{-r}-1}\sum_{k=1}^{\infty}\binom{r+k-1}{k}\rho^k h \times\left[\int_0^{u+ch} W(u+ch-x,z)dF^{*k}(x) + (F^{*k}(u+ch+z) - F^{*k}(u+ch))\right] + o(h),$$

where $F^{*k)}(x), k = 1,2,\ldots$ is the distribution function of $Y_1 + Y_2 + \ldots + Y_k$.

or, equivalently

$$\frac{W(u+ch,z)-W(u,z)}{ch} = \frac{\lambda}{c}W(u+ch,z) - \frac{\lambda\sum_{k=1}^{\infty}\binom{r+k-1}{k}\rho^k}{c((1-\rho)^{-r}-1)} \times\left[\int_0^{u+ch} W(u+ch-x,z)dF^{*k}(x) + (F^{*k}(u+ch+z) - F^{*}(u+ch))\right] + o(h),$$

In the limit, as $h \to 0$,

$$\frac{\partial}{\partial u}W(u,y) = \frac{\lambda}{c}\left[W(u,y) - \int_0^u W(u-x,z)dG(x) - (G(u+z)-G(u))\right] \tag{15}$$

where

$$G(x) = \frac{1}{(1-\rho)^{-r}-1}\sum_{k=1}^{\infty}\binom{r+k-1}{k}\rho^k F^{*k}(x),$$

is the nondefective distribution function of the claims with

$$G(0) = 0, G(\infty) = 1.$$

Related to safety loading above equation can be written as

$$\frac{\partial}{\partial u}W(u,z) = \frac{(1-\rho)(1-(1-\rho)^r)}{rp\mu(1+\theta)} \times\left[W(u,z) - \int_0^u W(u-x,z)d\,G(x) - (G(u+z)-G(u))\right] \tag{16}$$

From (14) and (15) we get the following integro differential equation for ruin probability.

$$\frac{d}{du}\Psi(u) = \frac{\lambda}{c}\left[\Psi(u) - \int_0^u \Psi(u-x)dG(x) - (1-G(u))\right],\ u \geq 0.$$

Theorem 6.1. *The function $W(0,z)$ is given by*

$$W(0,y) = \frac{\lambda}{c}\int_0^y (1-G(u))du. \tag{17}$$

Proof. Integrating (15) from 0 to ∞ and then using $W(\infty, z) = 0$, we have

$$-W(0,z) = \frac{\lambda}{c}\left[\int_0^\infty W(u,z)du - \int_0^\infty \int_0^u W(u-x,z)dG(x)du - \int_0^\infty (G(u+z) - G(u))du\right].$$

Substitution in the double integral and after simplification, we get

$$W(0,z) = \frac{\lambda}{c}\int_0^\infty (G(u+z) - G(u))du.$$

Hence

$$W(0,z) = \frac{\lambda}{c}\int_0^y (1 - G(u))du.$$

□

Theorem 6.2. *The ruin probability with no initial capital satisfies*

$$\Psi(0) = \frac{\lambda r \rho \mu}{c(1-\rho)(1-(1-\rho)^r)} \tag{18}$$

Proof. From (14) and (17) we obtain,

$$\Psi(0) = \lim_{z\to\infty} W(0,z) = \frac{\lambda}{c}\int_0^\infty (1 - G(u))du. \tag{19}$$

Suppose that $G(x)$ be the distribution function of a random variable X. Using the result $E(X) = \int_0^\infty (1 - G(x))dx$, (19) becomes.

$$\Psi(0) = \frac{\lambda}{c}E(X) \tag{20}$$

Considering the definitions of $G(x)$ and $EY = \mu$, we get

$$E(X) = \frac{\mu}{(1-\rho)^{-r} - 1}\sum_{k=1}^{\infty} k\binom{r+k-1}{k}\rho^k = \frac{\mu r \rho}{(1-\rho)(1-(1-\rho)^r)}. \tag{21}$$

From the Equations (20) and (21) we get the result.

□

7 Exponentially Distributed claims

Suppose that the claim sizes have exponential distribution with mean μ. i.e., $\overline{F}(x) = 1 - e^{-\frac{x}{\mu}}$, $x \geq 0$, $\mu > 0$. Now the density function $g(x) = \frac{d}{dx}G(x)$ obtained is an Erlang mixture and is given by

$$g(x) = \sum_{k=1}^{\infty} \frac{q_k (\frac{x}{\mu})^{k-1} e^{-\frac{x}{\mu}}}{\mu(k-1)!}, \qquad x > 0,$$

where $q_k = P(Y = k) = \frac{\binom{r+k-1}{k}\rho^k}{(1-\rho)^{-r}-1}, k = 1, 2, \ldots$ is the mixing distribution, see Willmot and Lin(2001).
The survival function corresponding to the above density is given by

$$\overline{G}(x) = \sum_{k=1}^{\infty} \frac{\overline{Q}_{k-1} (\frac{x}{\mu})^{k-1} e^{-x/\mu}}{(k-1)!}, x > 0,$$

where

$$\begin{aligned} \overline{Q}_{k-1} &= P(Y > k-1) \\ &= \sum_{i=k}^{\infty} \frac{\binom{r+i-1}{i}\rho^i}{(1-\rho)^{-r}-1} \quad k = 1, 2, 3... \end{aligned}$$

Applying the theorem 6.1, for the case when claim size has exponential distribution, $W(0, y)$ is given by

$$W(0, Z) = \frac{\lambda\mu}{c} \sum_{k=1}^{\infty} \frac{\overline{Q}_{k-1}\gamma(k, Z/\mu)}{(k-1)!},$$

where $\gamma(j, y) = \int_0^y u^{j-1}e^{-u}\, du$ is the incomplete gamma function.

8 Derivation of Total Claims Distribution, Total Loss(gain) Distribution and Stop Loss Moment

A central problem in risk theory is the modeling of the probability distribution of the aggregate claims. The aggregate claims distribution and its components, the claim count and claim amount distributions are used to compute ruin probabilities and to provide other information of interest to the decision makers. Panjer(1981) found that a compound Poisson process approximately modeled the aggregate claims distribution, based on the collective risk assumption.

Here we focus on the case where the aggregate claims distribution is modelled by a compound generalized Polya-Aeppli process. This is equivalent to assuming that the

counting process is Generalized Polya-Aeppli process. We assume that the claim sizes have a continuous distribution with distribution function F such that $F(0) = 0$, and the mean value μ.
Denote by $N(t)$ the no of claims, by Z_i the i^{th} claim amount and by $S(t)$ the aggregate claim amount in a time period of length t

$$S(t) = \sum_{i=1}^{M(t)} Z_i.$$

In this case $ES(t)$ corresponds to the pure premium and is given by

$$\begin{aligned} ES(t) &= EM(t)E(Z) \\ &= \frac{r\rho\lambda t}{(1-\rho)(1-(1-\rho)^r)}\mu \end{aligned}$$

Let $H(x,t)$ denotes the cumulative distribution function of the aggregate claims and $F^{*k}(x)$ is the k-fold convolution of claim amount distribution function which can be calculated recursively as

$$F^{*k}(x) = \int_0^x F^{*k-1}(x-y)f(y)dy.$$

with

$$\begin{aligned} F^{*0}(x) &= 1, \qquad x \geq 0 \\ &= 0, \qquad x < 0. \end{aligned}$$

Recalling that the number of claims has Generalized Polya-Aeppli distribution, We have

$$\begin{aligned} H(x,t) &= \sum_{k=0}^{\infty} P(M(t)=k)F^{*k}(x) \\ &= e^{-\lambda t}\left[I_{[0,\infty)}(x) + \sum_{k=1}^{\infty} a_{k,t}\rho^k F^{*k}(x)\right], \qquad x \geq 0, \end{aligned} \tag{22}$$

where $a_{k,t} = \sum_{i=1}^{k}\sum_{j=1}^{i}(-1)^{i+j}\binom{i}{j}\binom{rj+k-1}{k}(\frac{\lambda t}{(1-\rho)^{-r}-1})^i$ and $I_A(x)$ is the indicator function of the set A.

Finding the distribution function of the aggregate gain(loss) is one of the principal problems in the collective risk theory. Here we derive the distribution function of aggregate loss (gain) from generalized Polya-Aeppli risk model using the distribution function of aggregate claims.

Consider the generalized Polya-Aeppli risk model

$$X(t) = ct - S(t),$$

mentioned in (11) , where $X(t)$ denotes the aggregate gain (loss) and $S(t)$ is the aggregate claims in a time period of length t. In this model the number of claims occurring in a period of length t has the Generalized Polya-Aeppli distribution.
Then the cumulative distribution function $G(x,t)$ of $X(t)$ is given by

$$\begin{aligned} G(x,t) &= P(X(t) \leq x) \\ &= P(S(t) \geq ct - x) \\ &= \overline{H}(ct - x, t) \\ &= \sum_{k=1}^{\infty} a_{k,t} \rho^k \overline{F}^{*k}(ct - x), \end{aligned} \tag{23}$$

Where $\overline{F}^{*k}(x)$ is the survival function of $F^{*k}(x)$.
Stop-loss moment of any positive order can be obtained using (22) and is given by

$$\int_y^{\infty} (x-y)^m dH(x,t) = e^{-\lambda t} \sum_{k=1}^{\infty} a_{k,t} \rho^k \int_y^{\infty} (x-y)^m f^{*k}(x). \tag{24}$$

where f^{*k} is the k-fold convolutions of pdf of claims.
Note that the case $m = 0$ we get $\overline{F}(a)$, the tail function of the aggregate claim amount. When $m = 1$, the stop-loss premium results and is given by

$$\int_y^{\infty} \overline{H}(x,t) dx = \mu e^{-\lambda t} \sum_{k=1}^{\infty} k a_{k,t} \rho^n \overline{F}_I^{*k}(y), \tag{25}$$

where $F_I^{*k}(x) = \frac{1}{EX} \int_0^x \overline{F}_X^{*k}(u) du$, is the integrated tail distribution of F^{*k}.
Now we discuss a particular case in which the claim sizes have an exponential distribution with mean μ. i.e, $F(x) = 1 - e^{-x/\mu}, x \geq 0, \mu > 0$.
In this case the k fold convolution of claim sizes is given by

$$\begin{aligned} F^{*k}(x) &= \frac{\gamma(k, x/\mu)}{\Gamma k} \\ &= 1 - e^{-x/\mu} \sum_{j=0}^{k-1} \frac{(\frac{x}{\mu})^j}{j!} \\ &= 1 - e^{-x/\mu} e_k(x/\mu), \end{aligned}$$

where Γk is the gamma function and $\gamma(a,x) = \int_0^x t^{a-1} e^{-t} dt$ is the incomplete gamma function.
Hence the distribution function of aggregate claims with respect to Generalized Polya-Aeppli risk model with exponential claims is

$$H(x,t) = e^{-\lambda t} \left(I_{[0,\infty)}(x) + \sum_{k=1}^{\infty} a_{k,t} \rho^k [1 - e^{-x/\mu} e_k(x/\mu)] \right)$$

and
the distribution function of aggregate loss(gain) is given by

$$G(x,t) = e^{-\lambda t}\left(I_{[0,\infty)}(x) + e^{-\frac{(ct-x)}{\mu}}\sum_{k=1}^{\infty} a_{k,t}\rho^k e_k(\frac{ct-x}{\mu})\right).$$

Based on (24), for exponential claims, the stop loss moment is

$$\int_y^{\infty}(x-y)^m dH(x,t) = e^{-(\lambda t+\frac{y}{\mu})}\sum_{k=1}^{\infty}\sum_{i=0}^{m}\frac{(-1)^{m-i}\binom{m}{i}a_{k,t}\rho^k y^{m-i}\mu^i\Gamma(k+i)\ e_{k+i}(y/\mu)}{\Gamma(k)}$$

From (25) we can obtain stop-loss premium as

$$\int_y^{\infty}\overline{H}(x,t)dx = e^{-(\lambda t+y/\mu)}\sum_{k=1}^{\infty} a_{k,t}\rho^k\left[k\mu e_{k+1}(y/\mu) - y e_k(y/\mu)\right]$$

9 Discussion

In this paper, we introduced a new compound Poisson process, called the Generalized Polya-Aeppli process, which is obtained by compounding with truncated negative binomial distribution. We have shown that it is a generalization of Polya-Aeppli process. We found that this model is capable of handling over-dispersed count data. We have defined the risk model with Generalized Polya-Aeppli counting process and is thereby called Generalized Polya-Aeppli risk model. This model can be used as a model for insurance business. We have studied the key focus of ruin theory, the probability of ruin for this model and as a special case, we have obtained an expression for the ruin probability with zero initial capital.

References

[1] Johnson N.L, Kemp A.W. and Kotz S. (2005): Univariate Discrete Distributions. *Wiley Series in Probability and Mathematical Statistics*, third edition.

[2] Minkova, L.D. (2004). The Polya - Aeppli process and ruin problems. *J. Appi. Math. Stoch. Analysis*,3, 221 - 234.1,2.1,3.

[3] Chukova, S. and Minkova, L.D. (2013). Characterization of Polya - Aeppli process. *Stochastic Analysis and Applications*, 31, 590-599.2.

[4] Willmot, G.E. and Lin, X.S. (2010). Risk modeling with mixed Erlang distribution. *Applied Stochastic Models in Business and Industry* 27, 2-16.

[5] Dufresne, F.and Gerber, H.U.(1989). Three methods to calculate the probability of ruin. *ASTIN Bulletin*, 19(1), 71-90.

[6] Willmot, G.E. and Lin, X.S.(2001). Lundberg Approximation for Compound distribution with insurance applications. Berlin:*Springer.*

[7] Dickson, D.C.M.(2007). Some finite time ruin problems. *Annals of Acturial Science* 2, 217-232.

[8] Geber, H.U.(1982). On the numerical evaluation of the distribution of aggregate claims and its stop-loss premium. *insurance Mathematics and Economics*, 1(1), 13-18.

[9] Rufresne, F., Gerber, H.U. and Shiu, E.W.(1991). Risk theory with Gamma process *Astin Bulletin*. 21, 177-192.

[10] Gerber, H.U.(1979). An introduction to Mathematical Risk theory. *S.S Huebner Foundation, Philadelphia, P.A.*

[11] Mao, Z.C. and Lia, J.E.(2005). A risk model and ruin probability with compound Poisson-geometric process(in Chinese). *Acta Mathematicae Applicatae Sinica* 28(3), 419-428.

[12] Panjer, H.(1981). Recursive Evaluation of a Family of Compound Distributions. *ASTIN Bulletin*, 12(1), 22.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 65-78.

Estimation of Stress-Strength Reliability Function Using Weibull Distribution

Krishnendu K. [1], **Drisya M.** [2] **and Joby K. Jose** [3]
Department of Statistical Sciences, Kannur University, Kerala 670567, India.
[1]Email:krishnendu61195@gmail.com
[2]Email:drisyam.m@gmail.com
[3]Email:jobydsskannur@rediffmail.com

Abstract

Time-dependent stress-strength reliability deals with the chance of survival for systems with dynamic strength and/or dynamic stress. When a system is allowed to run continuously, each run of the system will cause a change in the strength of the system. And the repeated occurrence of stress on the system over each run will affect the survival capacity of the system. In this paper we consider the distribution of time taken for the completion of a run by the system as gamma and the strength of the system as Weibull or Weibull mixture. Moreover, the strength of the system is assumed to decrease by a constant value by the end of each run. The stress acting on the system is assumed either as fixed throughout the observation period, or as increasing by a constant over each run of the system. We obtain expressions for stress-strength reliability and also discuss numerical illustrations of the result.

Key words: Stress-strength reliability, Gamma Renewal process. Weibull distribution, Finite mixture distribution, EM algorithm.

1. Introduction

In reliability theory, stress-strength reliability measures the chance of strength of a system to overcome the stress acting on it. Every object or individual has its own strength for survival. When they are subject to any kind of stress, they will survive only if their strength exceed the stress. Stress-strength reliability model can be used for comparing the

effectiveness of two treatments, to compare the life length of two equipments etc. Let Y denote the random strength of the system under consideration and X be the stress acting on that system. Then the stress-strength reliability of the system is denoted by R and is defined as $R = P[X < Y]$.

The concept of stress-strength reliability theory was originated by [2]. [7] collected several point and interval estimation of stress-strength models using different approaches. [1] proposed an estimator of R based on kernel estimators of the densities of X and Y. [12] illustrated estimation of R using bootstrap method. Recently [6] and [10] studied the stress-strength reliability estimation of single system using various generalizations of half logistic distribution.

Nowadays, many of the studies in stress-strength reliability estimation concentrates on the case where the stress or strength or both of them changes with respect to time, and hence the term time dependent stress-strength reliability. Let $Y(t)$ represent the strength of a system at time t and $X(t)$ be the stress on the system at t. Under time-dependent stress-strength reliability model, we are interested in the estimation of the stress-strength reliability function

$$R(t) = P[X(t) < Y(t)], \tag{1}$$

which gives the chance of survival of the system at time t. For example, many often we have to download files to mobile phones. The downloaded files consumes the memory space of the phone corresponding to the size of that file. It will cause a reduction in the speed of functioning of the phone. So each time we download a new file, there is an increase in the number of files piled up in the phone memory and a reduction in the functioning speed of the phone.Time dependent stress-strength reliability models were studied in [11], [4, 5], [3] and [8, 9].

When a system is allowed to run continuously, each run of the system will cause a change in the strength of the system. And the repeated occurrence of stress on the system over each run will affect the survival capacity of the system. In this paper we consider the distribution of time taken for the completion of a run by the system as gamma and the strength of the system as Weibull or Weibull mixture. Moreover, the strength of the system is assumed to decrease by a constant value by the end of each run. The stress acting on the system is assumed either as fixed throughout the observation period, or as increasing by a constant over each run of the system.

This paper is organized as follows: Estimation of stress-strength reliability with gamma cycle times under fixed and random fixed stress is discussed in Section 2. A brief description of EM algorithm for estimating parameters of Weibull mixture distribution is given in Section 3. A numerical illustration of the models based on simulated data are given in Section 4.

2. Estimation of $R(t)$ with gamma cycle times

Consider a system which is allowed to work continuously. The system executes several runs during the time period of observation say $(0, t)$. The time taken for completion of a run by the system is a random variable and we call it as cycle time. In this paper we assume that the cycle times are gamma distributed. Hence the total number of runs within the entire time period will have a renewal process. Let the cycle time Z follows gamma distribution with p.d.f.,

$$f(z) = \frac{a^k z^{k-1} e^{-az}}{(k-1)!}; z \geq 0. \tag{2}$$

Then the number of runs during the time interval (0,t) has the following distribution.

$$P_n(t) = e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!}; n = 0, 1, 2, ... \tag{3}$$

We consider the estimation of $R(t)$ under two different situations. In the first case, we consider systems which are subject to fixed stress throughout the observation period and in the second case systems subject to random stress are considered .

2.1. Estimation of $R(t)$ when stress is fixed

Here, we assume that the strength of the system decreases by a constant say, a_0. It is also assumed that the stress acting on the system do not vary throughout the observation period. Let X_j be the stress imposed on the system during j^{th} cycle time and the corresponding strength of the system be Y_j. Then probability of functioning of the system after n runs is given by

$$\begin{aligned} R_n &= P[(X_1 < Y_1) \cap (X_2 < Y_2) \cap \cdots \cap (X_n < Y_n)] \\ &= P[(x_0 < Y_0 - a_0) \cap (x_0 < Y_0 - 2a_0) \cap \cdots \cap (x_0 < Y_0 - na_0)] \\ &= P[(x_0 + na_0 < Y_0)] \\ &= \int_{x_0+na_0}^{\infty} f(y_0) dy_0 \end{aligned} \tag{4}$$

Therefore, the value of $R(t)$ can be obtained as

$$\begin{aligned} R(t) &= \sum_{n=0}^{\infty} P_n(t) R_n \\ &= \sum_{n=0}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} \int_{x_0+na_0}^{\infty} f(y_0) dy_0. \end{aligned}$$

2.1.1. Estimation of $R(t)$ with Weibull initial strength

In many practical situations, it is observed that the distribution of life time random variable have Weibull distribution. Let the strength of the system have Weibull distribution with shape parameter α and scale parameter β having p.d.f

$$f(y_0) = \frac{\alpha}{\beta} y_0^{\alpha-1} e^{-y_0^{\alpha}/\beta}; y_0 \geq 0 \tag{5}$$

Then R_n is given by

$$R_n = e^{-(x_0+na_0)^{\alpha}/\beta} \tag{6}$$

and the corresponding $R(t)$ is obtained as

$$R(t) = \sum_{n=0}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} e^{-(x_0+na_0)^{\alpha}/\beta}. \tag{7}$$

Change in $R(t)$ corresponding to change in parameters are given in Figure 1. From Figure 1, we can see that the value of $R(t)$ increases with an increase in shape parameter values and decreases with an increase in scale parameter value of strength when initial strength of the distribution is Weibull distributed. And $R(t)$ increases with an increase in shape parameter value and decreases with an increase in rate parameter value of cycle time distribution.

2.1.2. Estimation of $R(t)$ with Weibull mixture initial strength

In the previous subsection we have considered the distribution of initial strength as Weibull. In this subsection, we consider mixture of Weibull distributions, which is a more flexible choice for initial strength distribution. The p.d.f. of initial strength distribution be

$$f(y_0) = \sum_{i=1}^{m} \pi_i \frac{\alpha}{\beta_i} y_0^{\alpha-1} e^{-y_0^{\alpha}/\beta_i}.; y_0 \geq 0 \tag{8}$$

The chance of the system works after the completion of n runs is,

$$R_n = \sum_{i=1}^{m} \pi_i e^{-(x_0+na_0)^{\alpha}/\beta_i} \tag{9}$$

and the corresponding stress-strength reliability is obtained as

$$R(t) = \sum_{n=0}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} \sum_{i=1}^{m} \pi_i e^{-(x_0+na_0)^{\alpha}/\beta_i} \tag{10}$$

$$= \sum_{i=1}^{m} \pi_i \sum_{n=0}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} e^{-(x_0+na_0)^{\alpha}/\beta_i} \tag{11}$$

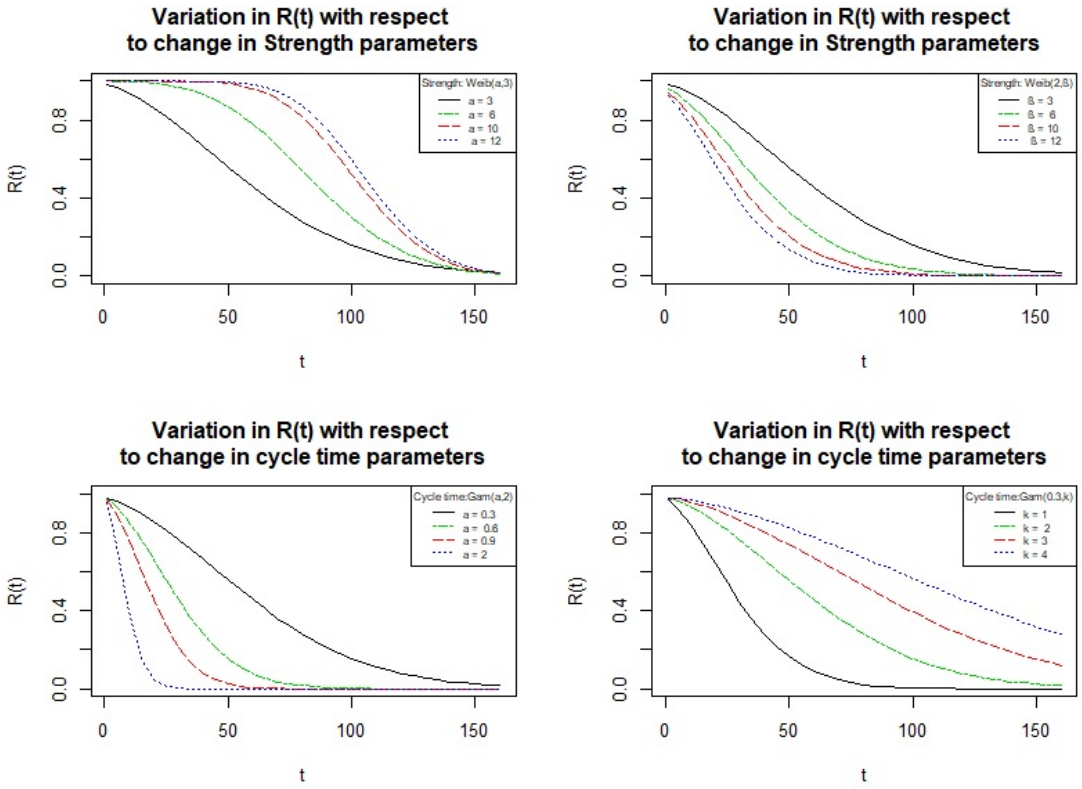


Figure 1: Variation in $R(t)$ corresponding to change in parameters

From equation (11) we can observe that, the reliability of a system with gamma cycle times and Weibull mixture initial strength is same as the mixture of reliability with gamma cycle times and Weibull initial strength with the same mixing probabilities. Change in $R(t)$ corresponding to change in parameters are given in Figure 2. From this figure we can see that the value of $R(t)$ increases with an increase in shape parameter value and decreases with increase in rate parameter value of cycle time distribution.

2.2. Estimation of $R(t)$ with gamma cycle times and random fixed stress

Let the initial strength of the system, say Y_0 be a continuous random variable with density function $f(y_0)$ and the initial stress on the system X_0 be also a continuous random variable with p.d.f $g(x_0)$. The system is allowed to run continuously and when the system runs, its strength decreases by a_0 and the stress increases by b_0 on completion of each run.

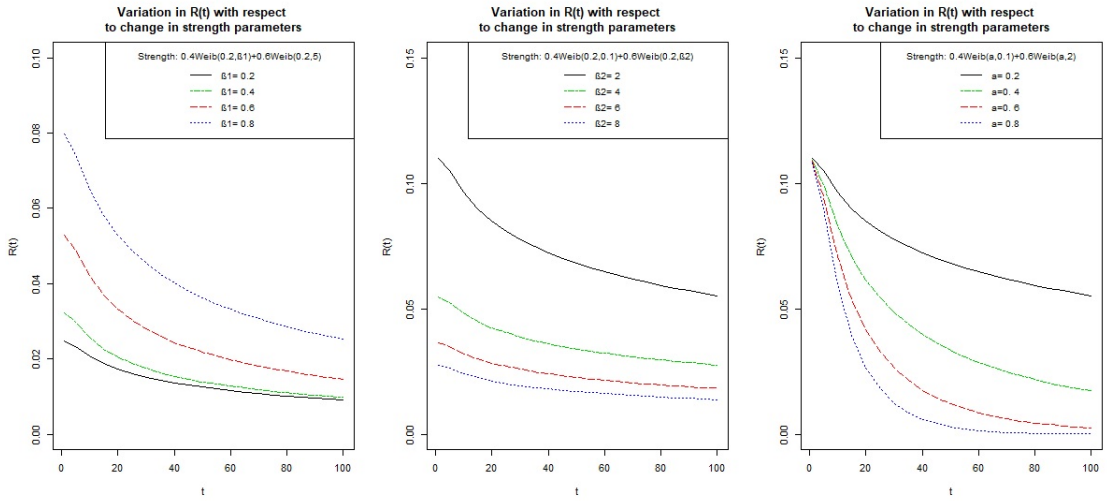


Figure 2: Variation in $R(t)$ corresponding to change in parameters

Hence, the probability that the system works after n runs is given by

$$\begin{aligned} R_n &= P((X_1 < Y_1) \cap (X_2 < Y_2) \cap \cdots \cap (X_n < Y_n)) \\ &= P((x_0 + b_0 < y_0 - a_0) \cap (x_0 + 2b_0 < y_0 - 2a_0) \cap \cdots \cap (x_0 + nb_0 < y_0 - na_0)) \\ &= P(x_0 + n(a_0 + b_0) < y_0) \\ &= \int_0^\infty \int_{x_0+n(a_0+b_0)}^\infty f(y_0)g(x_0)dy_0dx_0 \end{aligned} \tag{12}$$

Therfore the reliability of the system at time t is

$$R(t) = \sum_{n=0}^{\infty} P_n(t) \int_0^\infty \int_{x_0+n(a_0+b_0)}^\infty f(y_0)g(x_0)dy_0dx_0 \tag{13}$$

$$= \sum_{n=0}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} \int_0^\infty \int_{x_0+n(a_0+b_0)}^\infty f(y_0)g(x_0)dy_0dx_0 \tag{14}$$

2.2.1. Estimation of $R(t)$ with Weibull initial stress and strength

Let the initial strength of the system as well as the initial stress on the system are Weibull distributed with parameters (α, β_1) and (α, β_2) respectively. The time taken for completion of a run is assumed to be a gamma variate. Then the chance for survival of the system after n runs is

$$R_n = e^{-(n(a_0+b_0))^\alpha/\beta_1}; n = 1, 2, ... \tag{15}$$

with

$$R_0 = \frac{\beta_1}{\beta_1 + \beta_2} \tag{16}$$

Then the corresponding stress-strength reliability is obtained as

$$R(t) = e^{-at} \sum_{r=0}^{k-1} \frac{(at)^r}{r!} \frac{\beta_1}{\beta_1 + \beta_2} + \sum_{n=1}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} e^{-(n(a_0+b_0))^{\alpha}/\beta_1} \tag{17}$$

Change in $R(t)$ corresponding to change in different parameters of stress and strength

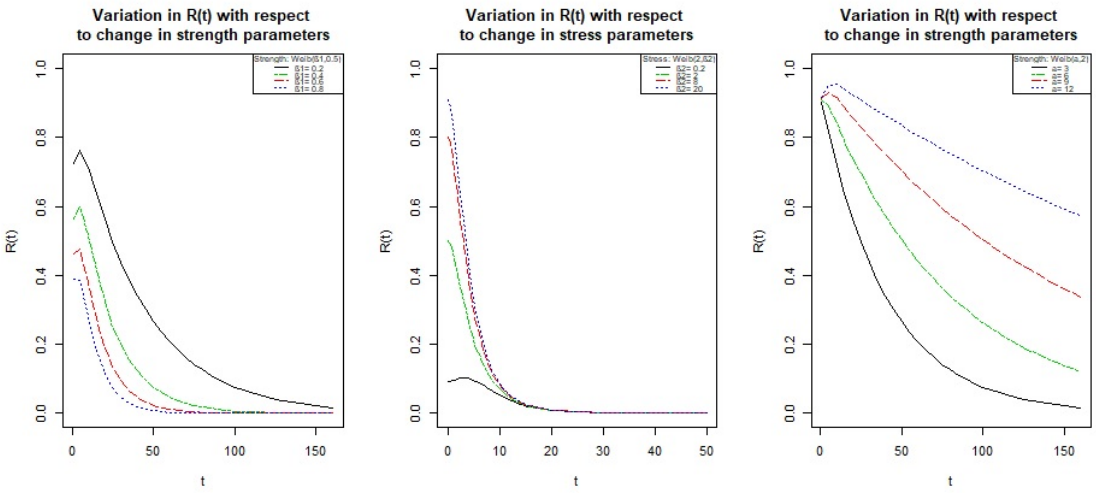


Figure 3: Variation in $R(t)$ corresponding to change in parameters

distributions are given in Figure 3. From the figure we can see that the value of $R(t)$ increases with an increase in shape parameter values and decreases with and increase in scale parameter value of strength when initial strength of the distribution is Weibull distributed. Also $R(t)$ increases with an increase in shape parameter value of stress distribution.

2.2.2. Estimation of $R(t)$ with Weibull mixture initial stress and strength

Let the initial strength of the system follows a mixture of Weibull distributions with p.d.f.

$$f(y_0) = \sum_{i=1}^{m_1} \pi_i \frac{\alpha}{\beta_i} y_0^{\alpha-1} e^{-y_0^{\alpha}/\beta_i}; y_0 \geq 0 \tag{18}$$

and initial stress on the system follows mixture of Weibull distribution with p.d.f.

$$g(x_0) = \sum_{j=1}^{m_2} p_j \frac{\alpha}{\theta_j} x_0^{\alpha-1} e^{-x_0^{\alpha}/\theta_j}; x_0 \geq 0. \tag{19}$$

When the system runs, its strength decreases by a_0 and the stress increases by b_0 on completion of each run. The time taken for completion of a run is assumed to be a gamma variate. Then the chance for survival of the system after n runs is

$$R_n = \sum_{i=1}^{m_1} \pi_i \sum_{j=1}^{m_2} p_j e^{-(n(a_0+b_0))^{\alpha}/\beta_i}; n = 1, 2, ... \tag{20}$$

with

$$R_0 = \sum_{i=1}^{m_1} \pi_i \sum_{j=1}^{m_2} p_j \frac{\beta_i}{\beta_i + \theta_j} \tag{21}$$

Then the corresponding stress-strength reliability is obtained as

$$\begin{aligned} R(t) &= e^{-at} \sum_{r=0}^{k-1} \frac{(at)^r}{r!} \sum_{i=1}^{m_1} \pi_i \sum_{j=1}^{m_2} p_j \frac{\beta_i}{\beta_i + \theta_j} \\ &+ \sum_{n=1}^{\infty} e^{-at} \sum_{r=nk}^{(n+1)k-1} \frac{(at)^r}{r!} \sum_{i=1}^{m_1} \pi_i \sum_{j=1}^{m_2} p_j e^{-(n(a_0+b_0))^{\alpha}/\beta_i} \end{aligned} \tag{22}$$

Change in $R(t)$ corresponding to change in different parameters stress and strength distributions are given in Figure 4.

3. EM algorithm for estimating parameters of Weibull mixture distribution

Here we give a brief description of the EM algorithm for estimating the parameters of a finite mixture of weibull distribution. Consider the data consist of n independent and identically distributed observations $y_0 = (y_1, y_2, ..., y_n)$ from a finite weibull mixture with pdf

$$f(y_0), \psi) = \sum_{i=1}^{k} \pi_i f(y_0, \psi).$$

Where

$$f(y_0, \psi) = \frac{\alpha}{\beta_i} y_0^{\alpha-1} e^{-\frac{y_0^{\alpha}}{\beta_i}}; y_0 > 0$$

and

$$\psi = (\pi_1, \pi_2, ..., \pi_k, \alpha, \beta_1, \beta_2, ...\beta_k)$$

The associated loglikelihood function is

$$L(y, \psi) = \sum_{i=1}^{n} log f(y_0, \psi). \tag{23}$$

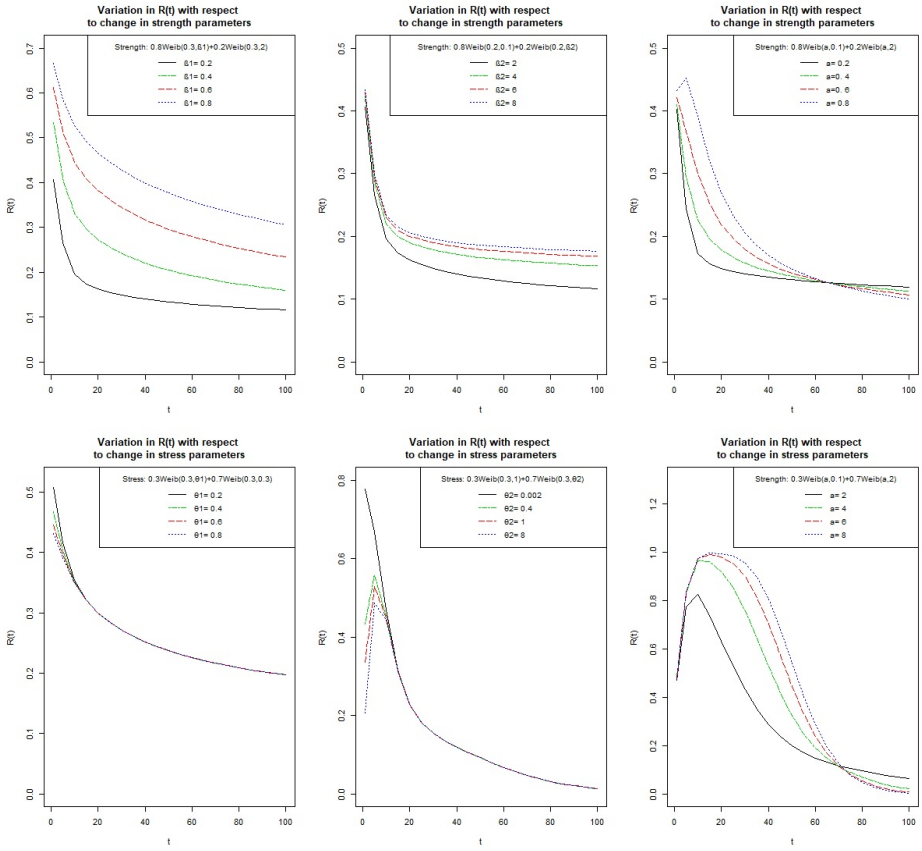


Figure 4: Variation in $R(t)$ corresponding to change in parameters

The MLE of ψ, say $\widehat{\psi}$, is determined such that

$$L(y, \widehat{\psi}) = sup_{\psi} L(y, \psi). \tag{24}$$

Define a variable z_{ij} such that $z_{ij} = 1$ if j^{th} unit of the sample comes from the i^{th} component and $z_{ij} = 0$ otherwise. Since each component comes from exactly one component, we have $\sum_{i=1}^{k} z_{ij} = 1, \quad \pi_i = P[z_{ij} = 1]$.

$$Y_i / z_{ij=1} \sim \texttt{Weibull}(\alpha, \beta_i), i = 1, 2, ..., k.$$

In missing data setup y can be considered as a incomplete data and $x = (x_1, x_2, ..., x_n)$ where $x_j = (y_j, z_j)$ and $z_j = (z_{ij}, i = 1, 2, ..., k)$ as a complete data set. The density function corresponding to one observation in complete data set is

$$f_c(x_j, \psi) = f_c(y_j, z_j, \psi) = \sum_{i=1}^{k} \pi_i I_{z_{ij}} f_i(y_j, \psi). \tag{25}$$

and the loglikelihood function is

$$L_c(x,\psi) = \sum_{j=1}^{n} log f_c(x_j,\psi). \tag{26}$$

The EM algorithm iteratively maximizes $Q(\psi/\psi^{(t)}) = E(L_c(x,\psi/y,\psi^{(t)}))$ instead of maximizing $L(y,\psi)$, where $\psi^{(t)}$ is the current value at t and then compute the expectation

$$E_{\psi^{(t)}}(L_c(x,\psi)/y) = \sum_{i=1}^{n}\sum_{j=1}^{k} E_{\psi^{(t)}}(z_{ij}/y)(log\pi_j + log f_j(y_i,\theta_j)) \tag{27}$$

$$\begin{aligned} E_{\psi^{(t)}}(z_{ij}/y) &= P_{\psi^{(t)}}(z_{ij}=1/y) \\ &= P_{ij}(\psi^{(t)}) \\ &= \frac{\pi_i^{(t)} f(y_j,\theta_i)}{\sum_{i=1}^{k}\pi_i^{(t)} f(y_j,\theta_i)}, j=1,2,...,n; i=1,2,...,k \end{aligned} \tag{28}$$

It is the posterior probability that j^{th} observation belongs to the i^{th} component in the t^{th} iteration. Thus we have

$$Q(\psi/\psi^{(t)}) = \sum_{j=1}^{n}\sum_{i=1}^{k} P_{ij}(\psi^{(t)})(log\pi_i + log f_i(y_i,\theta_i)). \tag{29}$$

Where $\theta = (\alpha,\beta)$ Hence the EM algorithm consist of the following two steps.

Step1.E-step: *Compute* $Q(\psi/\psi^{(t)})$
Step2.M-step:*Compute the value of* $\psi^{(t+1)}$ *that maximizes* $Q(\psi/\psi^{(t)})$.

As in the case of MLE from a multinomial likelihood, we have

$$P_{ij}^{(t+1)} = \frac{\sum_{j=1}^{n} P_{ij}(\psi^{(t)})}{n}, i=1,2,...,k. \tag{30}$$

For the $(t+1)^{th}$ update other parameters $(\theta_1,\theta_2,...,\theta_k)$, we have to obtain the solution of

$$\sum_{i=1}^{n}\sum_{j=1}^{k} P_{ij}(\pi^{(t)})\frac{\partial}{\partial\theta} log f_j(y_i,\theta_j)) = 0 \tag{31}$$

We repeat the procedure until the desired accuracy is obtained. We get the estimates as:

$$\hat{\alpha_i} = \left[\frac{\sum_{j=1}^{n} z_{ij}(y_j)^{\beta_i}}{\sum_{j=1}^{n} z_{ij}}\right]^{\frac{1}{\beta_i}} \tag{32}$$

$$\hat{\beta}_i = \left[\frac{\sum_{j=1}^{n} z_{ij} y_j^{\hat{\beta}_i} log(y_j)}{\sum_{j=1}^{n} z_{ij} y_j^{\hat{\beta}_i}} - \frac{\sum_{j=1}^{n} z_{ij} log(y_j)}{\sum_{j=1}^{n} z_{ij}} \right]^{-1} \tag{33}$$

4. Numerical Analysis

In this section we carry out Monte Carlo simulation to illustrate the estimation of the value of $R(t)$ of the time dependent system with gamma cycle times. Here we consider two examples. In the first example consider the system with finite mixture Weibull distribution for initial strength and in the second example we consider the system with finite mixture Weibull distribution for initial stress and initial strength.

To simulating samples for $N(t)$,the number of cycles during $(0, t)$, generate $(c_1, c_2, ..., c_{n_i})$ of size n_i from Gamma distribution, so that $c_1 + c_2 + ... + c_{n_i} = t$, $c_i's$ represent cycle times. The procedure is repeated n times to generate sample of size n for $N(t)$, (Size of the simulated data set is fixed as 10000.)

Example 1:Assume that the cycle time of the system is distributed as Gam(1,2) and the initial strength of the system follow, a mixture of Weibull distribution, 0.2 Weibull(1,2)+ 0.8 Weibull(1,6). The strength of the system is assumed to decrease by a fixed value say, 0.001 and the initial stress on the system is fixed as 0.5 throughout the period of observation. Suppose that we are interested in estimating the value of $R(t)$ at t=50.

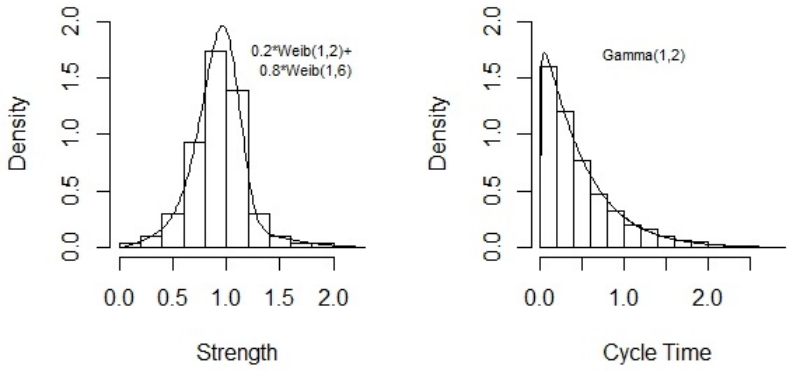


Figure 5: Histogram of simulated data sets for initial strength and cycle time

We have simulated observations on cycle time, initial strength, the number of cycles during $(0, t)$. The parameters are estimated using EM algorithm. Using the estimates of the parameters we have obtained the value of $R(50)$ as 0.7841. The estimated values of

Table 1: Estimated values of the parameters

	α	β_1	β_2	p_1	p_2	a	k
True Value	1	2	6	0.2	0.8	1	2
MLE	0.9932	2.2781	6.1552	0.2401	0.7599	1.0961	2.2650

parameters are given in Table 1. The histogram of initial strength and cycle time data along with fitted density curve is given in Figure 5.

Table 2 gives values of $R(t)$ for different values of time points and different values of parameters of cycle time and strength distributions.

Table 2: Estimated values of the parameters

Cycle time	Strength	a_0	x_0	t	$R(t)$
G(0.1,5)	0.2W(0.2,0.6)+0.8W(0.2,2)	0.02	0.05	10	0.2974
				25	0.2963
				50	0.2914
				75	0.2865
				100	0.2822
G(0.3,2)	0.7W(0.1,0.6)+0.3W(0.1,0.3)	0.1	0.03	10	0.5855
				25	0.5463
				50	0.5109
				75	0.4891
				100	0.4733

Example 2.

Assume that cycle time is distributed as Gam(1.5,4), the initial strength of the system follow a finite mixture of Weibull distribution say, $0.9 * W(1,5) + 0.1 * W(1,3)$ and the initial stress on the system follow a finite mixture of Weibull distribution say, $0.2W(1,3) + 0.8W(1,7)$. The strength of the system is assumed to decrease by a fixed value 0.002. and the stress is increase by a fixed value 0.004, after each run of the system. Suppose that we are interested in estimating the value of $R(t)$ at t=15. We have simulated observations on cycle time, initial strength, initial stress and the number of cycles during $(0,t)$. The parameters are estimated using EM algorithm. and estimated the values of the parameters of stress, strength and cycle time distributions based on simulated data set using EM algorithm. The estimated values of parameters are given in Table 3. The histogram of initial strength, initial stress and cycle time corresponding to the simulated data along with fitted density curve is given in Figure 6. Using the estimates of the parameters we have obtained the value of $R(15)$ as 0.9703.

Table 4 gives values of $R(t)$ for different values of time points and different values of parameters of cycle time, initial stress and initial strength distributions.

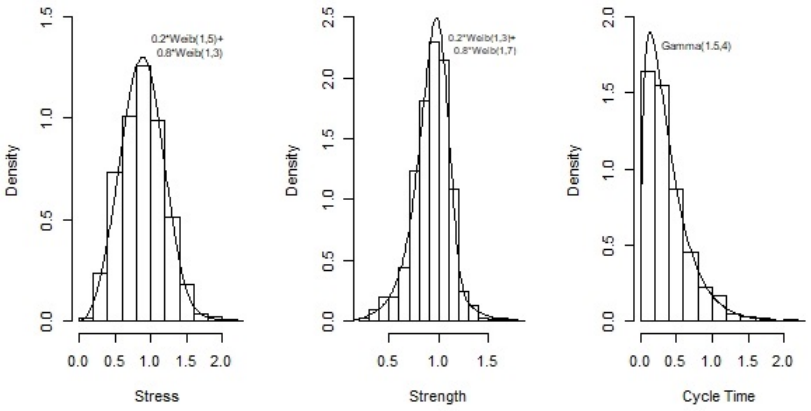


Figure 6: Histogram of simulated data with fitted density curve

Table 3: Estimated values of the parameters

	α	θ_1	θ_2	p_1	p_2	β_1	β_2	π_1	π_2	a	k
Actual Value	1	5	3	0.2	0.8	3	7	0.2	0.8	1.5	4
MLE	1.0318	3.4359	2.1848	0.1005	0.8995	3.3172	7.7769	0.2411	0.7589	1.5539	4.0789

Table 4: Estimated values of the parameters

Cycle time	Stress and Strength	a_0	b_0	t	$R(t)$
Gam(0.5,2)	Strength:	1	0.02	10	0.1964
	0.8W(0.3,0.6)+0.2W(0.3,2)			25	0.1548
				50	0.1336
	Stress:			75	0.1231
	0.3W(0.3,1)+0.7W(0.3,0.3)			100	0.1161
Gam(0.5,1)	Strength:	0.001	0.008	10	0.9984
	0.6W(5,0.3)+0.4W(5,2)			25	0.9980
				50	0.9978
	Stress:			75	0.9863
	0.3W(5,0.1)+0.7W(5,0.2)			100	0.9491

References

[1] A. Baklizi and O. M. Eidous, Nonparametric estimation of $P(X < Y)$ using kernel methods, Metron **LXIV** (2006), 47–60.

[2] Z. M. Birnbaum, On a use of the Mann-Whitney Statistic, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, **1**,(1956), 13–17.

[3] S. Eryilmaz, On stress-strength reliability with a time-dependent strength, Journal of Quality and Reliability Engineering, **2013** (2013).

[4] M. N. Gopalan and P. Venkateswarlu, Reliability analysis of time-dependent cascade system with deterministic cycle times, **22**, (1982), 841–872.

[5] M. N. Gopalan and P. Venkateswarlu, Reliability analysis of time-dependent cascade system with random cycle times, Microelectronics Reliability, **23(2)**,(1983), 355–366.

[6] J. K. Jose, T. Xavier and M. Drisya, Estimation of stress-strength reliability Using Kumaraswamy half-logistic distribution, Journal of Probability and Statistical Science **17(2)**,(2019), 141–154.

[7] S. Kotz, Y. Lumelskii and M. Pensky, The stress-strength model and its generalizations: theory and applications, World Scientific Publishing Co. Pvt. Ltd., Singapore, 2003.

[8] K. C. Siju and M. Kumar, Reliability analysis of time dependent stressstrength model with random cycle times, Perspectives in Science, **8**, (2016), 654–657.

[9] K. C. Siju and M. Kumar, Reliability computation of a dynamic stressstrength model with random cycle times, International journal of pure and applied Mathematics , **117**, (2017), 309–316.

[10] T. Xavier and J. K. Jose, A study of stress-strength reliability using a generalization of power transformed half-logistic distribution, Communications in Statistics- Theory and Methods, https : //doi.org/10.1080/03610926.2020.1716250, (2020).

[11] R. P. S. Yadav, A reliability model for stress strength problem, Microelectronics Reliability, **12(2)**, (1973), 119–123.

[12] W. Zhou, Statistical inference for $P(X < Y)$, Statistics in Medicine, **27**, (2008), 257–279.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 79-90.

Product Development: A Case Study

E V Gijo and U H Acharya
SQC & OR Unit, Indian Statistical institute
8th Mile, Mysore Road, Bangalore 560059
Email:gijo@isibang.ac.in

Abstract

Gas analyzers are used by regulating authorities for checking and measuring emission levels of carbon monoxide and other such polluting gases in vehicles. The company was in the process of overhauling its existing model. It had become necessary Government of India imposing strict regulations on the emission level by vehicles. With this objective, the study was framed to design exhaust gas analyzer of Fit and Forget Quality, qualifying all customer and regulatory requirements. The existing product was subjected to a Failure Mode and Effect Analysis (FMEA) for identifying the weaknesses if any in the present design. At the same time, the team thought that it would be appropriate to understand and include the customer perceptions in the new design. Accordingly an exhaustive customer survey was initiated and conducted. The evaluation of the current design in the light of voice of customer (VOC) revealed that improvement was necessary in several areas. Based on the analyses, a House of Quality Matrix was developed and deployed. The prototype was assembled and tested for performance. The test reports showed favorable results validating product expectations. Also, other tests like vibration testing, endurance test etc. were performed on the product and the performance was found satisfactory. The new product was subsequently launched countrywide.

Key words: Product Development, FMEA, Voice of customer, Quality function deployment.

1. Introduction

Exhaust gas analyzers are indispensable service instruments in the workshop, necessary for optimal mixture adjustment and effective trouble shooting on the engine upon analyzing

the exhaust emission gas from the vehicle. The Unit consists of a Gas Bench, which is for measuring the amount of carbon monoxide (CO), carbon dioxide (CO2) and hydrocarbons present in exhaust emission by NDIR (Non-Dispersive Infra Red method) and an Oxygen Sensor and Nox sensor to measure the amount of Oxygen and Nitrogen. The Unit consists of three filters, two Pumps, One Mini-Electro Valve and a Pressure switch. The Pumps used to transfer the filtered exhaust gas to gas bench and to pump out the condensed water particle from the moisture filter. The Electro-valve is used to switch between the atmospheric gas and sample gas. The Pressure switch is to identify the leak in the tubings. The Gas bench has the powerful Microprocessor unit, which processes the pollutants, measures in the bench and converts to digital values. A graphical LCD is present in the front panel for clear viewing with backlight.

Operating principle is so easy: simply insert a probe into the exhaust pipe and select measurement option and press enter key to initiate measurements. Readings appear on a graphical LCD.The measuring process begins with automatic zeroing and sensitivity calibrations. During prolonged testing periods, the analyzer automatically runs intermittent zero calibrations to ensure accuracy. The attachable sampling unit has a probe with a pre-filter that fits into all types of exhaust pipes.

The company was in the process of overhauling its existing model. It had become necessary because of sudden opening in the market with Government of India imposing strict regulations on the emission level by vehicles. The demand for accuracy and reliability on the measurement is also increasing. The Ministry of Road Transport and State enforcement authorities have put forth stringent requirement on the supplier of these emission testing instruments. This company had a product, which was basically meeting the non-euro requirements. So it was necessary to develop a Euro norms compliant product. With this objective, the study was framed to design exhaust gas analyzer of Fit and Forget Quality, qualifying all customer and regulatory requirements.

The remaining part of this article is arranged as follows. A literature review on the methodology adopted is presented in section 2. In section 3, the details steps involved in the product design are explained. Concluding remarks including the limitations and future research directions are presented in section 4.

2. Literature Review

To design a product that meets all the requirements, the design team needs to know what they are designing and what the end-users expect from it. Quality Function Deployment (QFD) is a structured technique for solving problems associated with the development or improvement of any product or service (Akao, 1990). QFD was developed in Japan in

1966. It consists in translating customer requirements into design characteristics for each stage of product development (Wasserman, 1993; Glen et al, 1996). By 1972 the power of the approach had been well demonstrated at the Mitsubishi Heavy Industries Kobe Shipyard (Sullivan, 1986). The goal of QFD is to translate often subjective quality criteria into objective ones that can be quantified and measured and which can then be used to design and manufacture the product (Mallon and Mulligan, 1993; Shigeru and Akao, 1994). QFD emphasizes active participation from the customer and helps integrate the engineering efforts of teams with skills from multiple disciplines (Douglas et al, 1995). It uses a series of product and process matrices to relate customer needs to technical requirements, analyze their relative importance, and evaluate their technical capabilities (Zeithml et al, 1990; Cohen, 1995).

Failure mode and effect analysis is a tool that examines potential product or process failures, evaluates risk priorities, and helps determine remedial actions to avoid identified problems (AIAG, 2001). The spreadsheet format allows easy review of the analysis. Failure mode and effect analysis is primarily a quality planning tool. It is useful in developing features and goals for both products and processes, in identifying critical product/process factors and designing counter measures to potential problems, in establishing controls to prevent process errors, and in prioritizing process subunits to ensure reliability (Kmenta and Ishii, 2004). In FMEA, failures are prioritized according to how serious their consequences are, how frequently they occur and how easily they can be detected (Stamatis, 2003). An FMEA also documents current knowledge and actions about the risks of failures for use in continuous improvement. FMEA is used during the design stage with an aim to avoid future failures (Mcdermott et al, 1996).

3. Product Design

The approach adopted for the study was as follows. First, to understand the weak areas in the design, an FMEA was performed and the weak areas in the design were identified. This was followed by a VOC data collection and translating it to company requirements. Finally a prototype was developed and tested.

3.1. Failure Mode Effects and Analysis

The existing product was subjected to a Failure Mode and Effect Analysis in order to identify the weaknesses in the product. All the functions performed by the gas analyzer were considered with its potential failures while preparing the FMEA. The FMEA thus performed is presented in Table1. From the FMEA given in table 1, the following parts/components emerged as weak areas in the existing design.

1. Bench
2. Sensor
3. Digital IC relays and crystals
4. Filter

These parts were selected for in depth design analysis, and improvement.

Table 1: Failure Mode and Effect Analysis

Process Function / Requirements	Potential Failure Mode	Potential Effects of Failure	SEV	Potential causes/ Mechanism of failure	OCC	Current Process Controls	DET	RPN
To measure and display correct values of gas concentration under all environmental conditions.	No proper display	a. No display b. Improper display	7	a. Cable failure b. Liquid leaks c. Back light failure	5 5 4	Visual Detection	2 2 2	70 70 56
	Improper values	a. Incorrect output	6	a. Improper calibration b. Faulty gas bench c. Failure of infrared sensor	4 3 3	No current controls	10 10 10	240 180 180
	Equipment Malfunctioning	a. Incorrect output b. Garbage output	8	a. Failure of digital ICs and relays b. Crystal variation	4 6	Visual Error Alerts	3 3	96 144
	Equipment Malfunctioning	a. Incorrect outputs	5	a. Failure of cooling fan	5	Audible hum	3	75
	Equipment not responding	a. Unable to feed the data	3	a. Sticky keys b. Keys stick with panel sticker	3 2	Visual Detection	3 2	27 12
	Improper values	Incorrect output	5	Filter leaks	6	No current controls	10	300

3.2. Voice of Customer (VOC) Analysis

The voice of the customer is the cornerstone of QFD process for product development (Day, 1996). Once you have identified key customers to be interviewed / surveyed, its time to gather the customer wants (Hayes, 1992). There are many useful techniques in gathering customers wants. Which techniques you choose, are based on considerations such as timing, cost and resources available (Fink and Kosecoff, 1985). The following approaches are the most popular methods in obtaining customer wants.

Focus groups: Additional insight into customers wants and perceptions can be gained through use of small groups of eight to twelve people who engage in free flowing discussion.

Table 2: Summary of Voice of Customer

Sl. No	Design attributes	Option-1	Option-2	Option-3
1	No. of Gases Checked	Two gas (25)	Four gas (25)	Govt. Regulations (50)
2	Printer Facility inbuilt	Yes (75)	No (0)	Govt. Regulations (25)
3	Exhaust Gas Hose 6mts with tip	Yes (70)	No (0)	Govt. Regulations (30)
4	Pre-filters in Gas Inlet	Yes (100)	No (0)	Govt. Regulations (0)
5	PC interface	Yes (70)	No (25)	Govt. Regulations (5)
6	Max Ambient temp 45 deg C	OK (65)	More (30)	Not specific (5)
7	Min Ambient Temp 5 deg C	OK (65)	Lower (30)	Not specific (5)
8	Battery Supply Provision – 12 V DC	Not Required (15)	Required (70)	Not Specific (15)
9	Power Supply thru' Cigarette Lighter	Not Required (60)	Required (10)	Not specific (30)
10	Display	Backlit LCD (75)	6 segment LED (20)	Not specific (5)
11	Control Keys	Membrane (90)	Selector Switch (10)	Not Specific (0)
12	Water Drain	Auto Drain (100)	Manual (0)	Not Specific (0)
13	Condensate Sensing	Not Required (20)	Required (80)	Not Specific (0)
14	Calibration	Gas (25)	Gas/Electronic (60)	Not specific (15)
15	With stand Power Fluctuations +/- 10%	OK (60)	More (40)	Not specific (0)
16	Interfacing with Web Camera, SIAM s/ware	Not Required (10)	Required (90)	Not Specific (0)
17	Portability – Handle	Not Required (10)	Required (90)	Not Specific (0)
18	Dimensions 270 x 300 x 120	OK (60)	Less (30)	Not Specific (10)
19	Weight < 12 kg	OK (60)	Less (30)	Not Specific (10)
20	All gases display	OK (90)	Toggle (10)	Not specific (0)
21	Manual (Operations Sheet)	English (70)	Regional (30)	Not specific (0)
22	Oil Temperature Probe	Optional (70)	Standard (20)	Not specific (10)
23	Up gradation to 5/6 gas	Not required (30)	Required (60)	Not specific (10)
24	Warm up Time < 15 min	OK (35)	Less (55)	Not Specific (10)
25	Auto Calibration/ Auto zeroing	Not Required (10)	Required (80)	Not specific (10)
26	Range CO : 0 – 9.99 %	OK (90)	More (10)	Not Specific (0)
27	Range HC : 0 – 1995 PPM	OK (70)	More (30)	Not Specific (0)
28	Range CO2 : 0 – 19.9 %	OK (90)	More (10)	Not Specific (0)
29	Range O2 : 0 – 21.4 %	OK (85)	More (0)	Not Specific (15)
30	Response Time < 10 sec	OK (60)	Less (40)	Not Specific (0)
31	Power consumption < 250W	OK (60)	Less (40)	Not Specific (0)
32	Two Wheeler Extension	Required (100)	Not Required(0)	Not Specific (0)
33	Lambda Value printout	Not Required (25)	Required (50)	Not Specific (25)
34	External Serial Printer	Optional (60)	Standard (30)	Not Specific (10)
35	Audio (Buzzer) Error	Not Required (10)	Required (85)	Not Specific (5)
36	Shop details in printout	Not Required (10)	Required (90)	Not Specific (0)

*Percentage in brackets representing the % of respondents.

Individual Interviews: Focus groups can tend to represent a reinforcement of the ideas offered by the most vocal members of the group. The wants and perceptions of more quiet

members may be lost in the process. Individual interviews overcome this weakness and are an effective method for obtaining insight into the customers wants. This process can be very effective in getting at the emotional side of the customers purchasing decision.

Listening and watching: A very effective technique used by Japanese is the simple art of listening / watching. Engineers, managers and others can be stationed at trade shows and at retail outlets to casually listen to the comments and conversations of the people.

Existing information: Most companies have a variety of information representing the voice of their customers. Much of these may be negative in the form of complaints, service call records and letters. However, these are valuable sources of information about the perception and expectations of the purchasers and should be included in the development of the voice of the customer (Burchill and Brodie, 1997).

Product designers were involved in surveys and interviews to provide them with the opportunity to develop a real understanding of the voice of the customer(s).The goal was to develop a meaningful list of unconstrained root wants. Lot of effort was put to get the exact words of the customer, develop a list of quotable statements as opposed to one word service attribute like cost, delivery, or service quality.

The summary finding of the VOC exercise is given in Table 2. The design option with maximum preference scores were chosen for consideration in the product design. The evaluation of the current design in the light of VOC revealed that improvement was necessary in eight areas.

The design options preferred by respondents were compared with the existing product as well as competitors product. The summary of comparison is given Table 3.

The above findings, highlighted in bold letters, were decided to be incorporated into the product. This was accomplished by the Quality Function Deployment (QFD) approach.

3.3. The QFD Process

The most critical QFD process is the House of Quality matrix (Hauser and Clausing, 1988). It was accomplished through the following steps.

The Objective: Describes the goal, problem, or objective of the team effort. The customer plays an important role in defining this objective since it gets to the heart of the problem to be solved and directly drives the product development. In this case it was to develop a Gas Analyzer of Fit and Forget quality.

Customer requirements: Determine the prioritized customer requirements for the product

Table 3: Existing product vis-a–vis Competitor product analysis

Sl. No	Description	Requirement Score	Current Product	Comp. 1	Comp. 2	Comp. 3
1	No. of Gases Checked	Govt. Regulations (50)	Y	Y	Y	Y
2	Printer Facility inbuilt	Yes (75)	Y	Y	N	N
3	Exhaust Gas Hose 6mts with tip	Yes (70)	Y	Y	Y	Y
4	Pre-filters in Gas Inlet	Yes (100)	Y	Y	Y	Y
5	PC interface	Yes (70)	Y	Y	Y	Optional
6	Max Ambient temp 45 deg C	OK (65)	Y	Y	Y	Y
7	Min Ambient Temp 5 deg C	OK (65)	Y	Y	Y	Y
8	**Battery Supply Provision - 12 V DC**	**Required (70)**	**N**	**N**	**Y**	**N**
9	Power Supply thru' Cigarette Lighter	Not Required (60)	N	N	Y	N
10	**Display**	**Backlit LCD (75)**	**N**	**N**	**Y**	**N**
11	Control Keys	Membrane (90)	Y	Y	Y	N
12	Water Drain	Auto Drain (100)	Y	Y	Y	Y
13	Condensate Sensing	Required (80)	Y	Y	Y	Y
14	Calibration	Gas/Electronic (60)	Gas/ Elec.	Gas	Gas	Gas
15	With stand Power Fluctuations +/- 10%	OK (60)	Y	Y	Y	Y
16	Interfacing with Web Camera, SIAM s/ware	Required (90)	Y	N	Y	Y
17	Portability - Handle	Required (90)	Y	Y	Y	Y
18	**Dimensions 270 x 300 x 120**	**OK (60)**	**N**	**N**	**Y**	**Y**
19	**Weight < 12 kg**	**OK (60)**	**N**	**N**	**Y**	**N**
20	All gases display	OK (90)	Y	N	N	N
21	Manual (Operations Sheet)	English (70)	Y	Y	Y	Y
22	Oil Temperature Probe	Optional (70)	Optional	N	N	N
23	**Up gradation to 5/6 gas**	**Required (60)**	**N**	**N**	**N**	**N**
24	Warm up Time < 15 min	Less (55)	15 Min	15 Min	3 Min	15 Min
25	Auto Calibration/ Auto zeroing	Required (80)	Y	Y	Y	N
26	Range CO : 0 - 9.99 %	OK (90)	Y	Y	Y	Y
27	Range HC : 0 - 1995 PPM	OK (70)	Y	Y	Y	Y
28	Range CO2 : 0 - 19.9 %	OK (90)	Y	Y	Y	N
29	Range O2 : 0 - 21.4 %	OK (85)	Y	Y	Y	N
30	**Response Time < 10 sec**	**OK (60)**	**N**	**N**	**Y**	**N**
31	Power consumption < 250W	OK (60)	Y	Y	Y	Y
32	**Two Wheeler Extension**	**Required (100)**	**N**	**Y**	**N**	**Y**
33	Lambda Value Printout	Required (50)	Y	N	N	N
34	**External Serial Printer**	**Optional (60)**	**N**	**N**	**N**	**N**
35	Audio (Buzzer) Error	Required (85)	Y	Y	Y	N
36	Shop details in printout	Required (90)	Y	Y	N	N

or service to be delivered. Since trade off decisions always exist, this step ensures that they favour customer needs and desires to the maximum extent and are not based only on

whats convenient for the developer (refer Table 2).

Competitive evaluations: Determine how well an organization and its competitors products and services meet the customers requirements (Table 3).

Technical requirements: Correspond to translation of operational requirements into technical specification language.

Relationship matrix: Determines the strength of the relationship between the customer requirements and the technical requirements.

Importance weighting: Compare the strength of the customer requirements, the technical requirements, and the customer importance information to identify technical requirements that are most important.

Correlation matrix: Explores the strength of the relationship between pairs of technical requirements. Weak correlations are traded off to find the best compromise, and strong correlations are studied to prevent duplication of effort.

Based on the above steps a House of Quality Matrix was developed and deployed. The Voice of Customer (VOC) data shortlisted as above from table 3 was deployed through QFD method to company requirements (Refer Table 4).

3.4. Prototype Development

The company engineers further worked on the company requirements as identified in the QFD matrix and developed a prototype. For example in case of LCD display, the existing product firmware was modified to incorporate LCD display. This new development not only delighted the customer by way of easy viewing with graphics. This exercise involved extensive design review and modification. Similarly the remaining seven VOCs were worked upon and the product meeting all the government regulations and better than competitors was developed. The components/ parts requirements were worked out; specifications were derived and subsequently ordered on the suppliers. The prototype was assembled and tested for performance. The test reports given in Table 5 show favorable results, validating product expectations. Also, other tests like vibration testing, endurance test etc. were performed on the product and the performance was found satisfactory. Product was successfully launched in the market subsequently.

4. Concluding Remarks

The company was primarily using reaction based approach for up gradation/ modification of the existing products. By this methodology they could realize the benefits of VOC

Table 4: First level QFD – VOC mapping

Company Requirement / Customer Requirement	Importance	Firmware	Optical sensing	Future up gradation facility	Alternate power supply	Meeting Govt. Standard	LCD Graphics	Low operating cost	Membrane particulars	Suitable accessory	Easy user interface	Ease of Handling	Quick viewing	Effective utilization
Battery Supply - 12 V DC	9				H			L						
LCD Display	9	H					H				H		H	
Dimensions 270x300x120	3											H		
Weight < 12 kg	3											H		
Up gradation to 5/6 gases	3	M	H	H		H			L					M
Response Time <10 Seconds	3	H									M			H
Two wheeler extension	9									H				
External Serial Printout	1	H										H		

and QFD in product design and development with maximum speed. The company could launch the new product precisely made as per VOC at much reduced cycle time. However the new requirements keep surfacing every now and then. This methodology therefore must become the basic knowledge for every designer.

The potential benefits of this approach include:

- *Customer focused product development.* System and product design requirements and objectives can be traced from customer requirements. This facilitates inclusion of the Voice of the Customer into the early system design process.

- *Shorter system development cycles.* Application of QFD helps avoid the need for fire-fighting during detail system design. Emphasis is placed on a before-the-fact approach, rather than after-the-fact.

Table 5: Test Results

	TEST REPORT	Sl.No.	VIB 002
		Date	
Test	VIBRATION TEST		
Product Reference	Model/Part No. : ECOMATE Sample Size : 01 No. Description : EXHAUST GAS ANALYZER		
Customer Reference	Customer : DC No. 1067 Date :		
Equipment Reference	Vibration Chamber RI 70006. Make : TIRA - Germany.		
Calibration Status	Calibration Valid Up to : MAR 2005		
Test condition	Frequency: 5-150 Hz Amplitude: 3 mm (peak to peak)-5 to 9Hz Acceleration : 1G –9- 150Hz Duration: 1 Hour Direction: vertical- Device under On condition.		
Test results **No visible distortion, damages, slackness failure in the product.** Equipment run satisfactorily.			
Remarks : –			
Test Status :	☑ Accepted	☐ Rejected	
Tested by / Date	Approved by / Date		Sheet 01 of 01
Tested at :			

- *Fewer production start-up problems.* The proactive approach inherent within the QFD method involves early consideration of downstream issues pertaining to manufacturing, distribution, installation, operation, and sustaining support. This reduces start-up problems.

The limitation of this study is that as the regulatory requirements for emission in vehicle changes, the current product may become obsolete after a specified period. Hence a further research can be conducted for streamlining the method of collection of VOC from the customer and regulatory requirements and establishing a system for product up gradation to suit any future requirements.

References

[1] Akao, Y. (1990), Quality Function Deployment, Productivity Press, Cambridge MA.

[2] Automotive Industry Action Group (2001), Potential Failure Mode and Effects Analysis, Reference Manual, 3rd edn., AIAG, South Filed, MI.

[3] Burchill, G. and Brodie, C.H. (1997), Voices into choices: Acting on the voice of the customer, Joiner Associates, Madison WI.

[4] Cohen, L. (1995), Quality Function Deployment: How to make QFD work for you, Addison Wesley, MA.

[5] Day, R.G. (1996), Quality function deployment: linking a company with its customers, Tata McGraw-Hill, New Delhi.

[6] Douglas, D., Barnard, W. and Norman, R. (1995), Customer integration: The quality function deployment (QFD) leaders guide for decision making, John Wiley and Sons, New York.

[7] Fink, A. and Kosecoff, J. (1985), How to conduct surveys: A step-by-step guide, Sage Publications, Thousand Oaks, CA.

[8] Glen, P., Motwani, J., Kumar, A. and Cheng, C. H. (1996), QFD application in an educational setting: A pilot field study, International Journal of Quality and Reliability Management, Vol. 13 No. 4, pp. 99-108.

[9] Hauser, J. R. and Clausing, D. (1988), "The House of Quality," The Harvard Business Review, Vol. 66, No. 3, pp. 63-73.

[10] Hayes, B.E. (1992), Measuring customer satisfaction: development and use of questionnaires, ASQC Quality Press, Milwaukee, WI.

[11] Kmenta, S. and Ishii, K. (2004), "Scenario-Based Failure Modes and Effects Analysis Using Expected Cost". Journal of Mechanical Design, Vol. 126, No. 6, pp. 1027-1035.

[12] Mallon, J. C. and Mulligan, D. E. (1993), Quality function deployment - a system for meeting customers needs, Journal of Construction Engineering and Management, Vol. 119. No. 3, pp. 516-531.

[13] Mcdermott, R.E., Mikulak, R.J. and Beauregard, M.R. (1996), Basics of FMEA, Productivity Press, Portland, OR.

[14] Shigeru, M. and Akao, Y. (1994), QFD: The customer driven approach to quality planning and Deployment, Asian Productivity Organisation, Tokyo.

[15] Stamatis, D.H. (2003), Failure Mode and Effect Analysis: FMEA from Theory to Execution, 2nd edn., ASQ Quality Press, Milwaukee, WI.

[16] Sullivan, L.P. (1986), "Quality Function Deployment", Quality Progress, June, pp 39-50.

[17] Wasserman, G. S. (1993), On how to prioritize design requirements during the QFD planning process, IIE Transactions, Vol. 25, No. 3, pp. 59-63.

[18] Zeithml, V.A., Parasuraman, A. and Berry, L.L. (1990), Delivering quality service: balancing customer perceptions and expectations. The Free Press, New York.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 91-96.

Correlation Coefficients and Its Misinterpretation

Sudheesh K. Kattumannil
Indian Statistical Institute, Chennai, India.
Email:skkattu@isichennai.res.in

Abstract

The correlation coefficient is often used or misinterpreted as the measure of linear association. We review the concepts correlation coefficients and studied its properties. We provide some examples to clear several misconceptions about the correlation coefficients. The examples are reproduced here with proper references.

Key words: Correlation coefficient, zero covariance, Independence.

1. Introduction

Despite the increasing use of statistical methods in different fields it is noted that several statistical concepts are still more or less unclear to scientists in other fields. For examples, in regression analysis, the coefficient of determination, R^2 some time misinterpreted as strength of the relationship. However, it (squared correlation coefficient) is simply a measure of how much of the variation measured as the sum of squares of the Y variable that is accounted for by a mathematical model involving independent variables. Another misconception is that the correlation between two variables is zero when the variable are independent. Hence we made an attempt to discuss the properties of the correlation coefficient in details. We refer to interest readers to Mathai (1998), Zhang (2007) and Mukhopadhyay (2010) and the references therein.

Before defining correlation coefficients, we starts with some questions that we will be answered later in this manuscript.

- Correlation zero implies whether the events are independent?
- Correlation tends to one implies whether the relation becomes linear?
- Does covariance equals zero implies correlation equals zero?

- Does independence implies correlation zero?

Answer to all the questions above is wrong and we discuss this in detail.

2. Correlation coefficients

Let X and Y be two random variables with distribution function F and G, respectively. We define the covariance between X and Y, as

$$Cov(X,Y) = E\big((X - E(X))(Y - E(Y))\big).$$

Next we give the definition of correlation.

Definition 2.1.

The correlation between X and Y, denoted by ρ, is defined as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}}, \tag{1}$$

for all non-degenerate random variables for which the covariance, $Cov(X,Y)$, and the variances, $Var(X)$ and $Var(Y)$, exist.

What does ρ really measure? At some extent, this question was answered by Mathai (1998). Mathai (1998) used the following example to illustrate some of the misconcepts about the correlation coefficient.

Let X be a real random variable having symmetric distribution. Clearly the odd order moments are zero. That is, $E(X^{2r+1}) = 0, r = 0, 1, ...$ Let $\sigma^2 = Var(X) = E(X^2)$ and that σ^2 is finite. This assumption is taken to ease the calculation of ρ. Consider a quadratic relation given by

$$Y = a + bX + cX^2, \tag{2}$$

where a, b and c are some real constants. Clearly $E(Y) = a + c\sigma^2$. Substituting these quantities in equation (1) we obtain

$$\rho = \frac{b}{\sqrt{b^2 + c^2\sigma^2[\frac{E(X^4)}{\sigma^4} - 1]}}, \tag{3}$$

provided $E(X^4) < \infty$. We can easily see that ρ equal to zero for $b = 0$. In this case we have quadratic relationship between X and Y.

We can see that ρ gives a small large positive and negative values for the relation specified in equations (2). Next we illustrate it. See Mathai (1998) for more details. Let X be a standard normal variate. Clearly $E(X) = 0$. Hence the coefficient of kurtosis defined by $\frac{E(X^4)}{(E(X^2))^2}$ has value 3. Then using equation (3) we obtain

$$\rho = \pm\frac{1}{\sqrt{1 + 2\frac{c^2}{b^2}}}, \quad b \neq 0.$$

Then we obtain small and large positive and negative values of ρ when the relationship between X and Y is specified in the equation (2) and for specific values of a, b and c.

(i) $Y = a + X + X^2$, $\rho = 0.58$.
(ii) $Y = a - X + X^2$, $\rho = -0.58$
(iii) $Y = a + X + 10X^2$, $\rho = 0.82$
(iv) $Y = a + 2X + X^2$, $\rho = 0.07$
(v) $Y = a - 3X + X^2$, $\rho = -0.90$.

Next, we answer to the question, does correlation near one in magnitude implies that the all points will tends to fall in a straight line? Suppose (X,Y) is a discrete random vector having joint uniform distribution on the 4 points $(-k, -a-\epsilon), (-k, -a+\epsilon), (k, a-\epsilon), (k, a+\epsilon)$. For these variables $E(X) = E(Y) = 0$. $Var(X) = k^2$, $Var(Y) = a^2 + \epsilon^2$ and $Cov(X, Y) = ak$. Hence, for a positive k, the correlation coefficient is given by

$$\rho = \frac{a}{\sqrt{(a^2 + \epsilon^2)}}.$$

Note that as $|a| \to \infty$, $\rho \to 1$ for any fixed $\epsilon \neq 0$. However the points $(k, a+\epsilon)$ and $(k, a-\epsilon)$ remain $2|\epsilon|$ units apart. This implies that ρ does not measure the linear relationship between two random variables. What can we say about the linearity. Using Cauchy-Schwartz inequality we can say that the relationship between X and Y are linear if and only if ρ equal to plus or minus one. As soon as ρ is away from plus or minus one, there could be perfect non-linear relationship between X and Y. All other values of ρ can not have meaningful interpretation as measure of relationship.

All these discussions highlighted the fact that ρ can not be used as a measure of association. What ρ really measure. We know that variance can be considered as a measure of spread of X_i's from its mean $\bar{X}$, center of gravity. Can we give a similar interpretation for ρ as the case of variance? Covariance can be considered as a measure of spread of the points (X_i, Y_i) from the centre $(\bar{X}, \bar{Y})$. Or it can be views as measure of angular dispersion between the points $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ with the origin shifted to $(\bar{X}, \bar{Y}$. And the correlation coefficient, ρ, can be considered as scale free covariance. See Mathai (1998) for more discussion on the interpretation of ρ. Next we some examples to which may help us to remove the misconception related to ρ.

Next we give an example of zero covariance without independence. See Zhang (2007) for more details. Let X_1 and X_2 be a sample of two independent observations drawn from a Bernoulli distribution with parameter $p(0 < p < 1)$. The sample mean and sample variance in this case is given by

$$\bar{X} = \frac{X_1 + X_2}{2} \quad \text{and} \quad S^2 = \frac{(X_1 - X_2)^2}{2}.$$

Using the joint probability distribution of $\bar{X}$ and S^2 (derive!), we can find that the third central moment of X_1 is equal to $p(1-p)(1-2p)$. Using an identity connecting the sample

mean and sample variance to the third central moment (see Remark 1 below) we obtain

$$Cov(\bar{X}, S^2) = \frac{p(1-p)(1-2p)}{2}.$$

From the above identity we observe that $Cov(\bar{X}, S^2) = 0$ for $p = 0.5$. However, $\bar{X}$ and S^2 are not independence. This can be verified using the conditional probability of $S^2 = 0$ given $\bar{X}$ which is given below

$$Pr(S^2 = 0|\bar{X} = 1) = 1 \neq 1/2 = Pr(S^2 = 0).$$

Remark 2.1.

In the above example, we use an identity connecting the sample mean and sample variance to the third central moment given by

$$n.Cov(\bar{X}, S^2) = E(X - \mu)^3,$$

where $\bar{X}$ and S^2 are the sample mean and sample variance based on the observations $X_1, \ldots, X_n$. The proof this identity is given in Zhang (2007). One can use this identity to study the skewness of random variables.

We end our discussion by reproducing some examples from Mukhopadhyay (2010). This will help us to understand the concepts of correlation more meaningfully. We next give examples to answer the question does independent implies zero covariance or correlation? Consider independent standard normal random variables U_1, U_2 and define

$$X = U_1 \quad \text{and} \quad Y = U_2^{-1}.$$

We know that ratio of two independent standard normal random variable has the standard Cauchy distribution. Hence $E(XY)$ is not finite, accordingly covariance can not be defined. Hence X and Y are independent does not implies covariance is zero.

Again, consider independent standard normal random variables $U_1, \ldots, U_5$ and define

$$X = U_1^2 \quad \text{and} \quad Y = \left(U_2^2 + U_3^2 + U_4^2 + U_5^2\right)^{-1}.$$

Clearly X and Y are independent and X and $1/Y$ follows chi-square distribution with one and 4 degrees of freedom, respectively. Hence Y follows inverse chi square distribution and $E(Y) = 1/2$. Also $E(X) = 1$. The distribution of $4XY$ is $F(1, 4)$ distribution. Hence $E(XY) = 2/4 = 1/2$ giving $Cov(X, Y) = 0$. Since $Var(Y)$ is not finite ρ is not finite in this case. This also reiterate the fact that the existence of second moment is important while defining ρ. What happens when the second moment of X and Y do not exist? Next section we address this question.

3. Gini covariance

There are several other correlation coefficients are discussed in literature. For example, Spearman rank correlation and Kendals τ, etc. Recent times correlation based on Gini mean difference has received much attention due to its applications in regression and time series analysis. Especially in time series analysis, it is desirable to allow for heavy tailed situations by developing concepts and methods that impose only first order moment assumptions. The Gini autocorrelation has great significance due to its potential application in infinite variance time series analysis. Preliminary work in this direction can be found Carcea and Serfling (2015).

Let X_1 and X_2 be two independent and identically distributed random variable having distribution function F. Gini mean difference(GMD) is defined as

$$GMD = E|X_1 - X_2|. \tag{4}$$

And Gini index is given by

$$GMD = \frac{E|X_1 - X_2|}{2\mu}.$$

Using the identity $|X_1 - X_2| = 2\max(X_1, X_2) - X_1 - X_2$ and noting that $F^2(X)$ is the distribution function of $\max(X_1, X_2)$, the GMD given in (4) can be rewritten as

$$GMD = 4.Cov(X, F(X)).$$

It is observed that GMD is four times the covariance between X variate and its rank F. This paves the way to define Gini covariance between X and Y. Next we define Gini covariance and correlation.

Definition 3.1. *Let (X, Y) be a bivariate random vector with joint distribution function F_{XY}. Also let F_X and F_Y be the respective marginal distribution functions. The Gini covariance between Y and X is defined as*

$$C(Y, X) = 4Cov(Y, F_X(X)). \tag{5}$$

Definition 3.2.

The Gini correlation between Y and X is defined as

$$\rho_g(Y, X) = \frac{Cov(Y, F_X(X))}{Cov(Y, F_Y(Y))}. \tag{6}$$

Definition 3.3. *The Gini regression parameter of Y on X is defined as*

$$\beta_g(Y, X) = \frac{Cov(Y, F_X(X))}{Cov(X, F_X(X))}. \tag{7}$$

Similarly, one can define the Gini covariance, Gini correlation between X and Y and Gini regression parameter of X on Y and we denote it as $C(X,Y), \rho_g(X,Y)$ and $\beta_g(X,Y)$ respectively. Unlike the other correlation coefficients Gini correlation is based on the variate value of one variable and the ranks of the other variable. And it can be defined in case of first moments of the random variables are finite.

Note that $\rho_g(X,Y) \neq \rho_g(Y,X)$ in general. If the distribution of (X,Y) is exchangeable up to a linear transformation then $\rho_g(X,Y) = \rho_g(Y,X)$. Moreover, if the distribution of (X,Y) is bivariate normal, then $\rho_g(X,Y) = \rho_g(Y,X) = \rho$, where ρ is the Pearson's correlation coefficient. Under bivariate normal assumption of the random vector (X,Y), $\beta_g(Y,X)$ reduces to the ordinary least square regression coefficient of Y on X. This parallelism also motivate researchers to develop alternate theory based on Gini covariance.

4. Concluding remarks

We made an attempt to avoid the confusion about the Pearson correlation coefficient. The discussion given in this paper reiterate the fact that one must give more care while using or interpreting ρ. We also discuss an alternate of correlation coefficient, Gini correlation which has numerous application in time series analysis. Some recent work based on Gini correlation and Gini regression coefficient we refer to Charpentier et al. (2019) and Shelef and Schechtman (2019).

References

[1] Carcea, M. and Serfling, R. (2015). A gini autocovariance function for time series modelling. *Journal of Time Series Analysis,* 36, 817-838.

[2] Charpentier, A., Ka, N., Mussard, S., and Ndiaye, O. H. (2019). Gini Regressions and Heteroskedasticity. *Econometrics,* 7, 4.

[3] Mathai, A. M. (1998). The concept of correlation and misinterpretations. *International Journal of Mathematical and Statistical Sciences*, 7, 157-168.

[4] Mukhopadhyay, N. (2010). When finiteness matters: counterexamples to notions of covariance, correlation, and independence. *The American Statistician*, 64, 231-233.

[5] Shelef, A., and Schechtman, E. (2019). A Gini-based time series analysis and test for reversibility. *Statistical Papers*, 60, 337-366.

[6] Zhang, L. (2007). Sample mean and sample variance: Their covariance and their (in) dependence. *The American Statistician*, 61, 159-160.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 97-112.

Some Properties and Applications of Burr III-Weibull Distribution

Deepthy G S, Nicy Sebastian and Reshma Rison
Department of Statistics, St.Thomas College, Thrissur, Kerala, India
Email:deepthygs@gmail.com, nicycms@gmail.com

Abstract

In this paper, we introduce a new distribution called Burr III-Weibull(BW) distribution using the concept of competing risk. We derive moments, conditional moments, mean deviation and quantiles of the proposed distribution. Also the Renyi's entropy and order statistics of the distribution are obtained. Estimation of parameters of the distribution is performed via maximum likelihood method. A simulation study is performed to validate the maximum likelihood estimator (MLE). A real practical data set is analyzed for illustration.

Key words:Burr III distribution, Weibull distribution, Maximum Likelihood Estimation.

1. Introduction

Burr type III distribution with two parameters was first introduced in the literature of Burr [1] for modelling lifetime data or survival data. It is more flexible and includes a variety of distributions with varying degrees of skewness and kurtosis. This distribution has a wide application in areas of statistical modelling such as forestry Gove et al.[3], meteorology Mielke [2], and reliability Mokhlis [7]. Burr type XII distribution can be derived from Burr type III distribution by replacing X with $\frac{1}{X}$. The usefulness and properties of Burr distribution are discussed by Burr and Cislak [8] and Johnson et al. [9].

The hazard function of distributions may include one or more of the following behavioural patterns; increasing, decreasing or constant shapes. Thus, they cannot be used to model lifetime data with a bathtub shaped hazard function, such as human mortality and machine life cycles. For last few decades, statisticians have been developing various

extensions and modified forms of the Weibull distribution and other related models. The two-parameter, flexible Weibull extension of Bebbington et al. [4] has a hazard function that can be increasing, decreasing or bathtub shaped. Zhang and Xie [5] studied the characteristics and application of the truncated Weibull distribution, which has a bathtub shaped hazard function. A new modified Weibull distribution by Saad and Jingsong [6] considered an increasing and a bathtub shaped hazard function.

The cumulative distribution function(cdf) and probability density function(pdf) of the weibull distribution are given by,

$$F_W(x;\lambda,\beta) = 1-e^{-(x/\lambda)^\beta}, x\geq 0, \lambda>0, \beta>0 \quad (1)$$

$$f_W(x;\lambda,\beta) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} e^{-(x/\lambda)^\beta} \quad (2)$$

where λ and β are the scale and shape parameters. The cumulative distribution(cdf) and probability density function(pdf) of the BurrIII distribution is given by,

$$G_B(x;c,k) = (1+x^{-c})^{-k}, x\geq 0, k>0, c>0 \quad (3)$$

$$g_B(x;c,k) = ckx^{-c-1}\left(1+x^{-c}\right)^{-k-1} \quad (4)$$

where c and k are shape parameters.

2. Burr III-Weibull Distribution

The reliability function of the new distribution, say Burr III Weibull(BW) distribution, can be constructed by combining the corresponding reliability functions of Burr III and Weibull distributions. The resulting reliability function, the cumulative distribution function and the probability density function are given by,

$$\bar{F}_{BW}(x;c,k,\lambda,\beta) = \left(1-\left(1+x^{-c}\right)^{-k}\right)\left(e^{-(x/\lambda)^\beta}\right) \quad ;c,k,\lambda,\beta>0 \quad (5)$$

$$F_{BW}(x;c,k,\lambda,\beta) = 1-\left(1-(1+x^{-c})^{-k}\right)\left(e^{-(x/\lambda)^\beta}\right) \quad (6)$$

for c,k,λ,$\beta>0$.

$$f_{BW}(x;c,k,\lambda,\beta) = e^{-(x/\lambda)^\beta}\left[ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{\beta-1}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right] \quad (7)$$

for $c,k,\lambda,\beta>0$. The hazard rate h(x) and reverse hazard rate r(x) are given respectively as,

$$h(x) = \frac{ck(1+x^{-c})^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{\beta-1}(1-(1+x^{-c})^{-k})}{1-(1+x^{-c})^{-k}} \quad (8)$$

$$T(x) = \frac{e^{-(x/\lambda)^\beta}[ck(1+x^{-c})^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{\beta-1}(1-(1+x^{-c})^{-k})]}{1-[1-(1+x^{-c})^{-k}e^{-(x/\lambda)^\beta}} \quad (9)$$

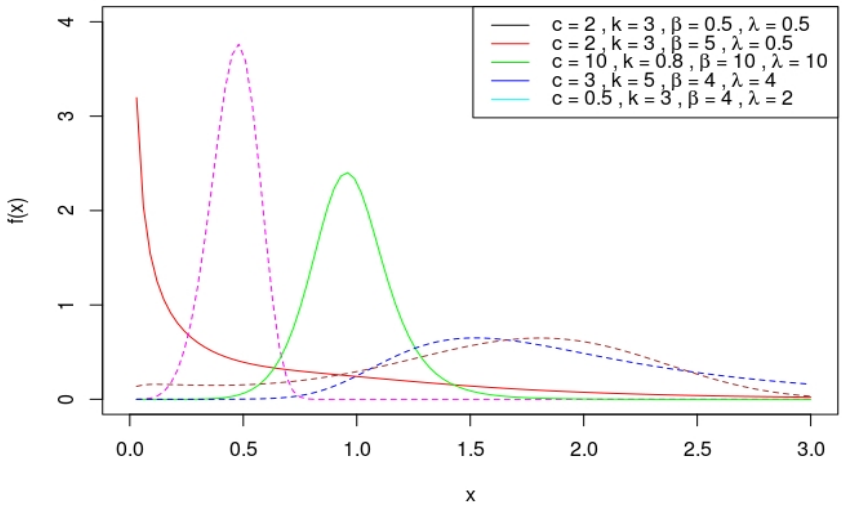


Figure 1: Plot for probability density functions of the BurrIII-Weibull distribution.

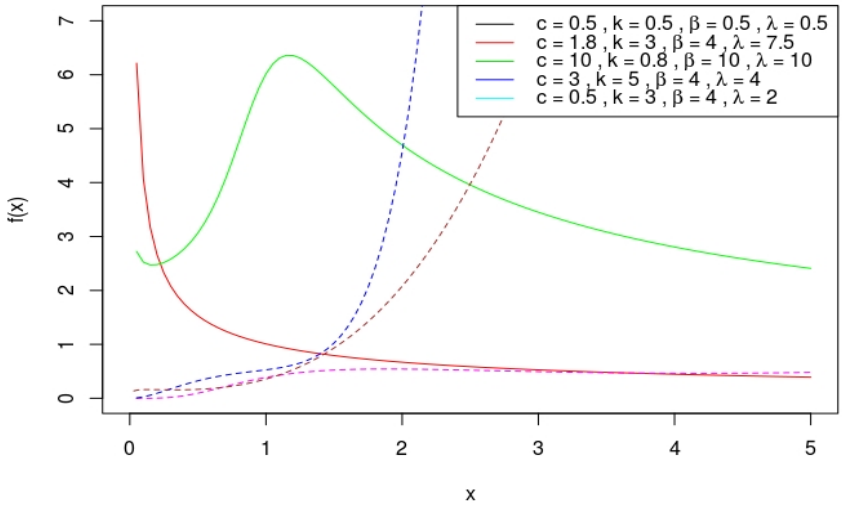


Figure 2: Plot for hazard rate functions of the BurrIII-Weibull distribution.

The plots of the probability and hazard rate functions of BW distribution for selected values of parameters are displayed in Figure 1 and Figure 2. Figure 1 shows that the BW distribution can be decreasing, approximately symmetrical and right skewed whereas Figure 2 reflects the monotonic behaviour of the hazard function for different values of parameters.

3. The Statistical Properties

In this section, some of the statistical properties of BW distribution such as quantile function, moments and order statistics are derived.

3.1 Quantile Function

The quantile function has a number of important applications, for example, it can be used to obtain the median, skewnes and kurtosis, and can also be used to generate random variables. The quantile function can be obtained by inverting $\bar{F}_{BW}(x) = 1 - u, 0 \leq u \leq 1$, where

$$\bar{F}_{BW}(x) = \left(1 - (1 + x^{-c})^{-k}\right)\left(e^{-(x/\lambda)^{\beta}}\right). \tag{10}$$

Let, $\left(1 - (1 + x^{-c})^{-k}\right)\left(e^{-(x/\lambda)^{\beta}}\right) = 1 - u$, and the corresponding quantile function is obtained by solving the non-linear equation,

$$\ln[1 - (1 + x^{-c})^{-k}] - \left(\frac{x}{\lambda}\right)^{\beta} - \ln(1 - u) = 0, \tag{11}$$

using numerical methods. Equation (11) can be used to generate random number. The quantiles for selected values of the BW distribution parameters are listed in Table 1.

3.2 Moments

Moments can be used to study the most important features and characteristics of a distribution such as central tendency, dispersion, skewness, kurtusis etc. The r^{th} moment of BW distribution is given by,

$$\begin{aligned}
E(X^r) &= \int_0^{\infty} x^r f(x)dx \\
&= \int_0^{\infty} x^r e^{-(x/\lambda)^{\beta}} \left[ck(1 + x^{-c})^{-k-1}x^{-c-1} + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}(1 - (1 + x^{-c})^{-k})\right]dx \\
&= ck\int_0^{\infty} x^{r-c-1}(1 + x^{-c})^{-k-1}e^{-(x/\lambda)^{\beta}}dx + \frac{\beta}{\lambda^{\beta}}\int_0^{\infty} x^{r+\beta-1}e^{-(x/\lambda)^{\beta}}dx \\
&- \frac{\beta}{\lambda^{\beta}}\int_0^{\infty} x^{r+\beta-1}(1 + x^{-c})^{-k}e^{-(x/\lambda)^{\beta}}dx.
\end{aligned}$$

Table 1: BW quantile for selected values

	(c,k,λ,β)				
u	(3,1,2,0.4)	(0.1,1.7,1,1)	(1.8,1.3,0.6,3)	(3,0.1,0.7,0.5)	(0.5,1.2,1,0.8)
0.1	0.00720	0.00007	0.24762	0.00023	0.014564
0.2	0.04698	0.00776	0.32549	0.00202	0.04755
0.3	0.14851	0.07691	0.38578	0.00705	0.09955
0.4	0.31794	0.20024	0.43923	0.01750	0.17508
0.5	0.50091	0.36225	0.49027	0.03678	0.28234
0.6	0.67537	0.56928	0.54192	0.070287	0.43596
0.7	0.85894	0.84254	0.59757	0.12815	0.66495
0.8	1.08687	1.2336	0.66279	0.23202	1.03706
0.9	1.46385	1.90969	0.75279	0.44542	1.78102

Let

$$\begin{aligned} A &= ck\int_0^\infty x^{r-c-1}(1+x^{-c})^{-k-1}e^{-(x/\lambda)^\beta}dx, \\ B &= \frac{\beta}{\lambda^\beta}\int_0^\infty x^{r+\beta-1}e^{-(x/\lambda)^\beta}dx\text{and} \\ C &= \frac{\beta}{\lambda^\beta}\int_0^\infty x^{r+\beta-1}(1+x^{-c})^{-k}e^{-(x/\lambda)^\beta}dx. \end{aligned}$$

Then

$$E(X^r) = A + B - C. \tag{12}$$

Consider,

$$\begin{aligned} A &= ck\sum_{m=0}^\infty \frac{(-1)^m}{\lambda^{m\beta}m!}\int_0^\infty \left(1+x^{-c}\right)^{-k-1}x^{r+m\beta-c-1}dx, \text{put } u=(1+x^{-c})^{-1}\ , \\ &= ck\sum_{m=0}^\infty \frac{(-1)^m}{\lambda^{(m\beta)}m!}\int_0^\infty u^{k+\frac{r}{c}+\frac{m\beta}{c}-1}(1-u)^{1-\frac{r}{c}-\frac{m\beta}{c}-1}du \\ &= ck\sum_{m=0}^\infty \frac{(-1)^m}{\lambda^{(m\beta)}m!}B\left(k+\frac{r}{c}+\frac{m\beta}{c},1-\frac{r}{c}-\frac{m\beta}{c}\right) \end{aligned} \tag{13}$$

Table 2: BW moments for selected values (c, k, λ, β)

Moments	(5,2.5,0.5,1.5)	(1,1,0.2,0.9)	(2,3,0.4,0.8)	(3,1.2,0.8,1.5)	(0.4,0.2,1,2)
$\mu_1^{'}$	0.44431	0.17564	0.21448	0.60806	0.14926
$\mu_2^{'}$	0.28101	0.06958	0.12327	0.50286	0.14075
$\mu_3^{'}$	0.21902	0.04434	0.12371	0.51011	0.17208
$\mu_4^{'}$	0.19687	0.03965	0.18929	0.60844	0.24533
$\mu_5^{'}$	0.19695	0.04618	0.41159	0.83411	0.39184
$\mu_6^{'}$	0.21481	0.06674	1.20954	1.29465	0.68564
SD	0.28913	0.19680	0.27797	0.36486	0.34419
CV	0.65073	1.12047	1.29601	0.60003	2.30597
CS	0.82234	2.42891	2.98560	0.87391	2.83748
Ck	3.37007	12.34781	18.5636	4.12879	11.39373

$$\begin{aligned} B &= \frac{\beta}{\lambda^\beta}\int_0^\infty x^{r+\beta-1}e^{-(x/\lambda)^\beta}dx \\ &= \frac{\beta}{\lambda^\beta}\frac{\lambda^{r+\beta}}{\beta}\Gamma\left(\frac{r+\beta}{\beta}\right), \text{using generalised gamma distribution} \qquad (14) \\ C &= \frac{\beta}{\lambda^\beta}\sum_{t=0}^{\infty}(-1)^t\binom{k+t-1}{t}\int_0^\infty x^{r+\beta-ct-1}e^{-(x/\lambda)^\beta}dx \\ &= \sum_{t=0}^{\infty}(-1)^t\binom{k+t-1}{t}\lambda^{r-ct}\Gamma\left(\frac{r+\beta-ct}{\beta}\right) \qquad (15) \end{aligned}$$

Substituting (13), (14) and (15) in (12), we get,

$$\begin{aligned} E(X^r) &= ck\sum_{m=0}^{\infty}\frac{(-1)^m}{\lambda^{(m\beta)}m!}B\left(k+\frac{r}{c}+\frac{m\beta}{c}, 1-\frac{r}{c}-\frac{m\beta}{c}\right)+\lambda^r\Gamma\left(\frac{r}{\beta}+1\right) \\ &- \sum_{t=0}^{\infty}(-1)^t\binom{k+t-1}{t}\Gamma\left(\frac{r+\beta-ct}{\beta}\right)\lambda^{r-ct} \text{ where } r<c, m\beta<c. \qquad (16) \end{aligned}$$

Where $B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt$ is the beta function and $\int_0^\infty x^{d-1}e^{-(x/a)^p}dx=\frac{\Gamma\left(\frac{d}{p}\right)}{\frac{p}{a^d}}$ is generalised gamma function. The moment generating function of the BW distribution is given by, $E\left(e^{tY}\right)=\sum_{i=0}^{\infty}\frac{t^i}{i!}E\left(Y^i\right)$ where $E\left(Y^i\right)$ is given above.

The first six moments ($\mu_1^{'}$, $\mu_2^{'}$, $\mu_3^{'}$, $\mu_4^{'}$, $\mu_5^{'}$, $\mu_6^{'}$), standard deviation (SD), coefficient of variation (CV), coefficient of skewness (CS) and coefficient of kurtosis (CK) for different selected values of the BW distribution parameters are listed in Table 2.

3.3 Conditional Moments

The r^{th} conditional moment is defined as $E(X^r/X>t)$. The r^{th} conditional moment

of the BW distribution is given by,

$$
\begin{aligned}
&E(X^r/X>t)\\
&\quad=\frac{1}{\bar{F}(t)}\int_t^\infty x^r f(x)dx\\
&\quad=\frac{1}{\bar{F}(t)}\int_t^\infty x^r e^{-\left(\frac{x}{\lambda}\right)^\beta}\left(ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{(\beta-1)}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right)dx\\
&\quad=\frac{1}{\bar{F}(t)}\int_t^\infty ck\left(1+x^{-c}\right)^{-k-1}x^{r-c-1}e^{-\left(\frac{x}{\lambda}\right)^\beta}dx+\frac{1}{\bar{F}(t)}\frac{\beta}{\lambda^\beta}\int_t^\infty x^{r+\beta-1}e^{-\left(\frac{x}{\lambda}\right)^\beta}dx\\
&\quad-\frac{1}{\bar{F}(t)}\frac{\beta}{\lambda^\beta}\int_t^\infty\left(1+x^{-c}\right)^{-k}e^{-\left(\frac{x}{\lambda}\right)^\beta}x^{r+\beta-1}dx.
\end{aligned}
$$

Let

$$
\begin{aligned}
A &= \int_t^\infty ck\left(1+x^{-c}\right)^{-k-1}x^{r-c-1}e^{-\left(\frac{x}{\lambda}\right)^\beta}dx,\\
B &= \frac{\beta}{\lambda^\beta}\int_t^\infty x^{r+\beta-1}e^{-\left(\frac{x}{\lambda}\right)^\beta}dx,\\
C &= \frac{\beta}{\lambda^\beta}\int_t^\infty\left(1+x^{-c}\right)^{-k}e^{-\left(\frac{x}{\lambda}\right)^\beta}x^{r+\beta-1}dx.
\end{aligned}
$$

Then

$$
E(X^r/X>t)=\frac{1}{\bar{F}(t)}\left(A+B-C\right). \tag{17}
$$

Consider

$$
\begin{aligned}
A &= ck\sum_{m=0}^\infty\frac{(-1)^m}{\lambda^{m\beta}m!}\int_t^\infty\left(1+x^{-c}\right)^{-k-1}x^{r+m\beta-c-1}dx, \text{put}\ \ u=\left(1+x^{-c}\right)^{-1},\\
&= k\sum_{m=0}^\infty\frac{(-1)^m}{\lambda^{m\beta}m!}\int_{(1+t^{-c})^{-1}}^1 u^{k+\frac{r}{c}+\frac{m\beta}{c}-1}\left(1-u\right)^{1-\frac{r}{c}-\frac{m\beta}{c}-1}du\\
&= k\sum_{m=0}^\infty\frac{(-1)^m}{\lambda^{m\beta}m!}B_{(1+t^{-c})^{-1}}\left(k+\frac{r}{c}+\frac{m\beta}{c},1-\frac{r}{c}-\frac{m\beta}{c}\right) \qquad (18)
\end{aligned}
$$

$$
\begin{aligned}
B &= \frac{\beta}{\lambda^\beta}\int_t^\infty x^{r+\beta-1}e^{-\left(\frac{x}{\lambda}\right)^\beta}dx, \text{let}\ \ u=\left(\frac{x}{\lambda}\right)^\beta\\
&= \lambda^r\int_{(\frac{t}{\lambda})^\beta}^\infty u^{1+\frac{r}{\beta}-1}e^{-u}du\\
&= \lambda^r\Gamma\left(\left(\frac{r}{\beta}+1\right),\left(\frac{t}{\lambda}\right)^\beta\right) \qquad (19)
\end{aligned}
$$

$$
\begin{aligned}
C &= \frac{\beta}{\lambda^{\beta}}\sum_{p=0}^{\infty}(-1)^{p}\binom{k+p-1}{p}\int_{t}^{\infty}x^{r+\beta-cp-1}e^{-\left(\frac{x}{\lambda}\right)^{\beta}}dx \\
&= \lambda^{r-cp}\sum_{p=0}^{\infty}(-1)^{p}\binom{k+p-1}{p}\int_{\left(\frac{t}{\lambda}\right)^{\beta}}^{\infty}u^{1+\frac{r}{\beta}-\frac{cp}{\beta}-1}e^{-u}du \\
&= \lambda^{r-cp}\sum_{p=0}^{\infty}(-1)^{p}\binom{k+p-1}{p}\Gamma\left(\frac{r-cp}{\beta}+1,\left(\frac{t}{\lambda}\right)^{\beta}\right) \qquad (20)
\end{aligned}
$$

Substituting (18), (19), (20) in (17), we get,

$$
\begin{aligned}
E\left(X^{r}/X>t\right) &= \frac{1}{\bar{F(t)}}\left(k\sum_{m=0}^{\infty}\frac{(-1)^{m}}{\lambda^{m\beta}m!}B_{(1+t^{-c})^{-1}}\left(k+\frac{r}{c}+\frac{m\beta}{c},1-\frac{r}{c}-\frac{m\beta}{c}\right)\right) \\
&+ \frac{1}{\bar{F(t)}}\left(\lambda^{r}\Gamma\left(\left(\frac{r}{\beta}+1\right),\left(\frac{t}{\lambda}\right)^{\beta}\right)\right) \\
&- \frac{1}{\bar{F(t)}}\left(\sum_{p=0}^{\infty}(-1)^{p}\binom{k+p-1}{p}\lambda^{r-cp}\Gamma\left(\frac{r-cp}{\beta}+1,\left(\frac{t}{\lambda}\right)^{\beta}\right)\right) \qquad (21)
\end{aligned}
$$

3.4 Mean Deviation

The amount of scatter in a population is measured to some extent by the totality of deviations from the mean and median. These are known as mean deviation about mean and as mean deviation about median and are defined as,

$$
\delta_{1}(x) = \int_{0}^{\infty}\mid x-\mu\mid f_{BW}(x)dx \text{ and } \delta_{2}(x)=\int_{0}^{\infty}\mid x-M\mid f_{BW}(x)dx
$$

respectively where $\mu=E(X)$ and M=Median(X) denote the median. The measures $\delta_{1}(x)$ and $\delta_{2}(x)$ can be calculated using the relationships,

$$
\begin{aligned}
\delta_{1}(x) &= 2\mu F_{BW}(\mu)-2\mu+2\int_{\mu}^{\infty}xf_{BW}(x)dx \qquad (22) \\
\delta_{2}(x) &= -\mu+2\int_{M}^{\infty}xf_{BW}(x)dx \qquad (23)
\end{aligned}
$$

respectively. When $r=1$ we get the mean $\mu=E(X)$. Note that $T(\mu)=\int_{\mu}^{\infty}xf_{BW}(x)dx$

and $T(M) = \int_M^\infty x f_{BW}(x)dx$, where

$$\begin{aligned} T(\mu) &= \int_\mu^\infty x f(x)dx \\ &= k\sum_{m=0}^{\infty}\frac{(-1)^m}{\lambda^{m\beta}m!}B_{(1+\mu^{-c})^{-1}}\left(k+\frac{1}{c}+\frac{m\beta}{c},1-\frac{1}{c}-\frac{m\beta}{c}\right)+\lambda\Gamma\left(\frac{1}{\beta}+1,\left(\frac{\mu}{\lambda}\right)^\beta\right) \\ &- \sum_{p=0}^{\infty}\lambda^{1-cp}(-1)^p\binom{k+p-1}{p}\Gamma\left(\frac{1-cp}{\beta}+1,\left(\frac{\mu}{\lambda}\right)^\beta\right) \end{aligned} \tag{24}$$

similarly,

$$\begin{aligned} T(M) &= \int_M^\infty x f(x)dx \\ &= k\sum_{m=0}^{\infty}\frac{(-1)^m}{\lambda^{m\beta}m!}B_{(1+M^{-c})^{-1}}\left(k+\frac{1}{c}+\frac{m\beta}{c},1-\frac{1}{c}-\frac{m\beta}{c}\right) \\ &+ \lambda\Gamma\left(\frac{1}{\beta}+1,\left(\frac{M}{\lambda}\right)^\beta\right)-\sum_{p=0}^{\infty}\lambda^{1-cp}(-1)^p\binom{k+p-1}{p}\Gamma\left(\frac{1-cp}{\beta}+1,\left(\frac{M}{\lambda}\right)^\beta\right) \end{aligned} \tag{25}$$

Consequently, the mean deviation about the mean and the mean deviation about the median reduces to

$$\delta_1(x) = 2\mu F_{BW}(\mu) - 2\mu + 2T(\mu) \quad and \quad \delta_2(x) = -\mu + 2T(M)$$

respectively.

3.5 Bonferroni and Lorenz curves

Bonferroni and Lorenz curves have applications not only in economics for the study of income and poverty, but also in other fields such as reliability, demography, insurance and medicine. Bonferroni and Lorenz curves for the BW distribution are given by,

$$\begin{aligned} B(p) &= \frac{1}{p\mu}\int_0^q x f_{BW}(x)dx = \frac{1}{p\mu}[\mu - T(q)], \quad \text{and} \\ L(p) &= \frac{1}{\mu}\int_0^q x f_{BW}(x)dx = \frac{1}{\mu}[\mu - T(q)], \end{aligned}$$

respectively, where

$$\begin{aligned} T(q) &= \int_q^\infty x f_{BW}(x)dx \\ &= k\sum_{m=0}^{\infty}\frac{(-1)^m}{\lambda^{m\beta}m!}B_{(1+q^{-c})^{-1}}\left(k+\frac{1}{c}+\frac{m\beta}{c},1-\frac{1}{c}-\frac{m\beta}{c}\right) \\ &+ \lambda\Gamma\left(\frac{1}{\beta}+1,\left(\frac{q}{\lambda}\right)^\beta\right)-\sum_{p=0}^{\infty}\lambda^{1-cp}(-1)^p\binom{k+p-1}{p}\Gamma\left(\frac{1-cp}{\beta}+1,\left(\frac{q}{\lambda}\right)^\beta\right) \quad (26) \end{aligned}$$

and $q = F^{-1}(p), 0 \le p \le 1$.

3.6 Order Statistics

The density function $f_{i:m}(x)$ of the i^{th} order statistic for i=1,2,3,...,m from independently and identically distributed random variables $X_1, X_2, ..., X_m$ following BW distribution is given by,

$$\begin{aligned} f_{i:m}(x) &= \frac{m!f_{BW}(x)}{(i-1)!(m-i)!}[F_{BW}(x)]^{i-1}[1-F_{BW}(x)]^{m-i} \\ &= \frac{m!f_{BW}(x)}{(i-1)!(m-i)!}\sum_{j=0}^{m-i}(-1)^j\binom{m-i}{j}[F_{BW}(x)]^{j+i-1} \quad (27) \end{aligned}$$

Using the binomial expansion $[1-F(x)]^{m-i} = \sum_{j=0}^{m-i}\binom{m-i}{j}(-1)^j[F(x)]^j$ and the pdf and cdf of BW distribution in (27) we have,

$$\begin{aligned} f_{i:m}(x) &= \frac{m!f_{BW}(x)}{(i-1)!(m-i)!}\sum_{j=0}^{m-i}(-1)^j\binom{m-i}{j}\left(1-\left[1-\left(1+x^{-c}\right)^{-k}\right]\left[e^{-\left(\frac{x}{\lambda}\right)^\beta}\right]\right)^{j+i-1} \\ &= \sum_{j=0}^{m-i}(-1)^j\frac{m!}{(i-1)!(m-i-j)!(j)!}\left(1-\left[1-\left(1+x^{-c}\right)^{-k}\right]\left[e^{-\left(\frac{x}{\lambda}\right)^\beta}\right]\right)^{j+i-1} \\ &\times e^{-\left(\frac{x}{\lambda}\right)^\beta}\left(ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{(\beta-1)}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right) \end{aligned}$$

The pdf of the 1^{st} and n^{th} order statistic is given by,

$$\begin{aligned} f_{1:m}(x) &= \sum_{j=0}^{m-1}(-1)^j\frac{m!}{(m-1-j)!(j)!}\left(1-\left[1-\left(1+x^{-c}\right)^{-k}\right]\left[e^{-\left(\frac{x}{\lambda}\right)^\beta}\right]\right)^{j} \\ &\times e^{-\left(\frac{x}{\lambda}\right)^\beta}\left(ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{(\beta-1)}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right) \quad (28) \\ f_{n:m}(x) &= \sum_{j=0}^{m-n}(-1)^j\frac{m!}{(n-1)!(m-n-j)!(j)!}\left(1-\left[1-\left(1+x^{-c}\right)^{-k}\right]\left[e^{-\left(\frac{x}{\lambda}\right)^\beta}\right]\right)^{j+n-1} \\ &\times e^{-\left(\frac{x}{\lambda}\right)^\beta}\left(ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1}+\frac{\beta}{\lambda^\beta}x^{(\beta-1)}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right) \quad (29) \end{aligned}$$

4. Renyi's Entropy

In this section, Renyi's entropy of the BW distribution is derived. An entropy is a measure of uncertainty or disorder of a random variable. Renyi's entropy is an extension of Shannon's entropy. In the case of BW distribution Renyi's entropy is defined to be

$$I_R(v) = \frac{1}{1-v}\ln\left(\int_0^\infty [f_{BW}(x;c,k,\lambda,\beta)]^v dx\right), v \neq 1, v > 0.$$

Renyi's entropy tends to Shannon's entropy as $v \to 1$. Note that $[f(x;c,k,\lambda,\beta)]^v = f_{BW}^v(x)$ can be written as,

$$f_{BW}^v(x) = e^{-v(x/\lambda)^\beta}\left[ck\left(1+x^{-c}\right)^{-k-1}x^{-c-1} + \frac{\beta}{\lambda^\beta}x^{\beta-1}\left(1-\left(1+x^{-c}\right)^{-k}\right)\right]^v$$

Using the expansions,

$$e^{-x} = \sum_{j=0}^{\infty}\frac{(-1)^j x^j}{j!},\ (x+y)^n = \sum_{p=0}^{n} {}^nC_p x^{n-p}y^p,\ (1-x)^p = \sum_{m=0}^{p} {}^pC_m(-1)^m x^m.$$

$$\begin{aligned}
f_{BW}^v(x) &= \sum_{j=0}^{\infty}\frac{(-1)^j v^j (\frac{x}{\lambda})^{\beta j}}{j!}\left[\frac{\beta x^{\beta-1}}{\lambda^\beta}(1-(1+x^{-c})^{-k}) + kcx^{-c-1}(1+x^{-c})^{-k-1}\right]^v \\
&= \sum_{j=0}^{\infty}\frac{(-1)^j v^j x^{\beta j}}{\lambda^{\beta j} j!}\sum_{p=0}^{v}\binom{v}{p}\left(kcx^{-c-1}(1+x^{-c})^{-k-1}\right)^{v-p}\left(\frac{\beta x^{\beta-1}}{\lambda^\beta}(1-(1+x^{-c})^{-k})\right)^p \\
&= \sum_{j=0}^{\infty}\sum_{p=0}^{v}\sum_{w=0}^{p}\frac{(-1)^{j+w}v^j\beta^p}{\lambda^{(\beta j+p\beta)}j!}(kc)^{v-p}\binom{v}{p}\binom{p}{w}x^{\beta j+cp+p-cv-v+p\beta-p} \\
&\times\ (1+x^{-c})^{kp+p-kv-v-kw}.
\end{aligned}$$

Now,

$$\begin{aligned}
\int_0^\infty f_{BW}^v(x)dx &= \sum_{j=0}^{\infty}\sum_{p=0}^{v}\sum_{w=0}^{p}\frac{(-1)^{j+w}v^j\beta^p}{\lambda^{(\beta j+p\beta)}j!}(kc)^{v-p}\binom{v}{p}\binom{p}{w} \\
&\times \int_0^\infty x^{\beta j+cp+p-cv-v+p\beta-p}(1+x^{-c})^{kp+p-kv-v-kw}dx.
\end{aligned}$$

Put $u = (1+x^{-c})^{-1}$,

$$\begin{aligned}
\int_0^\infty f_{BW}^v(x)dx &= \sum_{j=0}^{\infty}\sum_{p=0}^{v}\sum_{w=0}^{p}\frac{(-1)^{j+w}v^j\ \beta^p}{\lambda^{(\beta j+p\beta)}\ j!}\ (kc)^{v-p}\ \binom{v}{p}\ \binom{p}{w}\ \frac{1}{c} \\
&\times \int_0^\infty u^{a-1}(1-u)^{b-1}du \\
&= \sum_{j=0}^{\infty}\sum_{p=0}^{v}\sum_{w=0}^{p}\frac{(-1)^{j+w}\ v^j\ \beta^p}{\lambda^{(\beta j+p\beta)}\ j!}\ (kc)^{v-p}\ \binom{v}{p}\ \binom{p}{w}\ \frac{1}{c} \times B(a,b),
\end{aligned}$$

where

$$\begin{aligned} a &= \frac{1}{c}(p\beta + \beta j + cp - cv - v) + \frac{1}{c} - kp - p + kv + v + kw, \\ b &= \frac{1}{c}(cv + v - p\beta - \beta j - cp) - \frac{1}{c}. \end{aligned}$$

Then,

$$I_R(v) = \frac{1}{1-v} \ln \left(\sum_{j=0}^{\infty} \sum_{p=0}^{v} \sum_{w=0}^{p} \frac{(-1)^{j+w}\ v^j\ \beta^p}{\lambda^{(\beta j + p\beta)}\ j!}\ (kc)^{v-p} \binom{v}{p} \binom{p}{w} \frac{1}{c} B(a,b) \right), \quad (30)$$

for $v \neq 1$ abd $v > 0$.

5. Method of Maximum Likelihood Estimation

The most useful parametric estimation method is the maximum likelihood method. Let $x_1, x_2, ...x_n$ be a random sample of size n from BW distribution. Then the log likelihood function is given by,

$$l(c,k,\lambda,\beta) = \sum_{i=1}^{n} \ln[f(x_i,c,k,\lambda,\beta)] \quad (31)$$

The log likelihood for a single observation is given by,

$$l(c,k,\lambda,\beta) = -\left(\frac{x}{\lambda}\right)^{\beta} + \ln(ck(1+x^{-c})^{-k-1}x^{-c-1} + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}[1-(1+x^{-c})^{-k}]) \quad (32)$$

The maximum likelihood estimates can be obtained by solving the following the equations simultaneously,
$\frac{\partial l(c,k,\lambda,\beta)}{\partial c} = 0$, $\frac{\partial l(c,k,\lambda,\beta)}{\partial k} = 0$, $\frac{\partial l(c,k,\lambda,\beta)}{\partial \lambda} = 0$, $\frac{\partial l(c,k,\lambda,\beta)}{\partial \beta} = 0$ where

$$\frac{\partial l(c,k,\lambda,\beta)}{\partial c} = \frac{k\lambda^{\beta}\left(-cx^{-c-1}\ln x\left(1+x^{-c}\right)^{-k-1} - cx^{-2c-1}(-k-1)(1+x^{-c})^{-k-2}\ln x + x^{-c-1}(1+x^{-c})^{-k-1}\right) - k\beta x^{-c+\beta-1}(1+x^{-c})^{-k-1}\ln x}{\lambda^{\beta}ckx^{-c-1}(1+x^{-c})^{-k-1} + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}\left[1-(1+x^{-c})^{-k}\right]}$$

$$\frac{\partial l(c,k,\lambda,\beta)}{\partial k} = \frac{ckx^{-c-1}(1+x^{-c})^{-k-1} - cx^{-c-1}(1+x^{-c})^{-k-1}\ln(1+x^{-c}) + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}\ln(1+x^{-c})(1+x^{-c})^{-k}}{ckx^{-c-1}(1+x^{-c})^{-k-1} + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}\left[1-(1+x^{-c})^{-k}\right]}$$

$$\frac{\partial l(c,k,\lambda,\beta)}{\partial \lambda} = \frac{\beta x}{\lambda^2}\left(\frac{x}{\lambda}\right)^{\beta-1} - \frac{\beta^2 x^{\beta-1}\lambda^{-\beta-1}\left[1-\left(1+x^{-c}\right)^{-k}\right]}{ckx^{-c-1}(1+x^{-c})^{-k-1} + \frac{\beta}{\lambda^{\beta}}x^{\beta-1}\left[1-(1+x^{-c})^{-k}\right]}$$

Table 3: Simulation Results: Mean Estimates, Bias, MSE.

		I			II		
Sample Size	Parameter	Mean	Bias	MSE	Mean	Bias	MSE
n=25	c	0.3843	0.08613	0.02274	5.9395	0.43952	1.79489
	k	7.8764	-0.10006	0.22994	5.0623	0.06233	0.68445
	λ	1.1783	-0.03159	0.01312	0.89611	-0.00388	0.00353
	β	2.32138	0.297792	0.22436	3.6348	0.33482	0.47132
n=200	c	0.3557	0.05578	0.00538	5.7685	0.2685	0.7198
	k	7.9368	-0.06318	0.06149	5.06035	0.06035	0.12072
	λ	1.1776	-0.02231	0.00211	0.8961	-0.0038	0.0003
	β	2.2158	0.21584	0.05967	3.5156	0.2156	0.0854
n=400	c	0.3538	0.05388	0.004074	5.7476	0.2476	0.2950
	k	7.9723	-0.02077	0.003812	5.0564	0.05644	0.03913
	λ	1.1769	-0.02184	0.00129	0.8964	-0.0035	0.00019
	β	2.2105	0.210585	0.050837	3.5054	0.20548	0.05793
n=600	c	0.3534	0.05343	0.003653	5.6836	0.1836	0.1939
	k	7.9823	-0.01765	0.001241	5.0501	0.05018	0.03712
	λ	1.1767	-0.02158	0.00098	0.8963	-0.0034	0.00014
	β	2.2057	0.20573	0.046781	3.5039	0.2039	0.0533

$$\frac{\partial l(c,k,\lambda,\beta)}{\partial \beta} = -\left(\frac{x}{\lambda}\right)^{\beta}\ln\left(\frac{x}{\lambda}\right) + \frac{\left[1-\left(1+x^{-c}\right)\right]^{-k}\left[\lambda^{2}\left(x^{\beta-1}+\beta x^{\beta-1}\log(x)(x)\right)-\beta x^{\beta-1}\lambda^{\beta}\ln\lambda\right]}{(\lambda^{\beta})^{2}\left[ckx^{-c-1}(1+x^{-c})^{-k-1}+\frac{\beta}{\lambda^{\beta}}x^{\beta-1}\left[1-(1+x^{-c})^{-k}\right]\right]}$$

The total log likelihood function based on random sample of n observations $x_1, x_2, ...x_n$ drawn from BW distribution is given by $l^* = \sum_{i=1}^{n} l(c,k,\lambda,\beta)$ where $l(c,k,\lambda,\beta)$ is given by equation (32). Owing to the complexity of these equations, the MLEs does not have an analytical expression. However, one can use standard statistical software to solve those equations (e.g., Mathematica, R, etc.). We make use of R software to carry out this study.

6. Simulation

The accuracy and performance of the BW distribution is investigated by conducting two simulations for different parameter values and sample sizes. The simulations were repeated $N = 1000$ times each with sample sizes $n = 25, 50, 200, 400, 600$ and the true parameters values I : $c = 0.3, k = 8, \lambda = 1.2, \beta = 2$ and II : $c = 5.5, k = 5, \lambda = 0.9, \beta = 3.3$. Three quantities were computed in this simulation study: the mean, bias and mean-square error

(MSE). The mean estimate, bias and mean-square error of the MLE $\hat{\epsilon}$ of the parameter c, k, λ, β are respectively given by

$$\text{Mean} = \frac{\sum_{i=1}^{N}(\hat{\epsilon_i})}{N}, \text{ Bias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\epsilon} - \epsilon) \quad \text{and} \quad \text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\epsilon} - \epsilon)^2.$$

The mean of MLEs of the BW distribution parameters along with their respective mean square errors and bias for different sample sizes are listed in Tables 3.

7. Real Applications

This section illustrates the usefulness of the Burr III Weibull distribution using a real data set. The data set includes 101 observations which represent the failure times of Kevlar 49/epoxy strands which were subjected to constant sustained pressure at the 90 % percent stress level until all had failed. The data are: 0.01, 0.01, 0.02, 0.02, 0.02,0.03, 0.03, 0.04, 0.05, 0.06, 0.07, 0.07, 0.08, 0.09, 0.09, 0.10, 0.10, 0.11, 0.11, 0.12, 0.13, 0.18,0.19, 0.20, 0.23, 0.24, 0.24, 0.29, 0.34, 0.35, 0.36, 0.38, 0.40, 0.42, 0.43, 0.52, 0.54, 0.56, 0.60,0.60, 0.63, 0.65, 0.67, 0.68, 0.72, 0.72, 0.72, 0.73, 0.79, 0.79, 0.80, 0.80, 0.83, 0.85, 0.90, 0.92,0.95, 0.99, 1.00, 1.01, 1.02, 1.03, 1.05, 1.10, 1.10, 1.11, 1.15, 1.18,1.20, 1.29, 1.31, 1.33, 1.34,1.40, 1.43, 1.45, 1.50, 1.51, 1.52, 1.53, 1.54, 1.54, 1.55, 1.58, 1.60, 1.63, 1.64, 1.80, 1.80, 1.81,2.02, 2.05, 2.14, 2.17, 2.33, 3.03, 3.03, 3.34, 4.20, 4.69, 7.89.

We have fitted the proposed Burr III Weibull distribution (BW) to the data set and compared the proposed distribution with Weibull distribution and Burr III distribution. The values of the estimated parameters, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Consistent Akaike Information Criterion (AICC) values for the corresponding data are provided in Table 4. The BW distribution is a better model as compared to the Burr III model. We conclude that the Burr III-Weibull distribution can be comparable to the Burr III and Weibull models.

8. Conclusion

A new distribution called Burr III-Weibull(BW)distribution is proposed and its properties are studied. The BW distribution possesses increasing, decreasing and upside-down bathtub shaped failure rate function. We derived the moments, conditional moments, mean deviation, quantiles, Bonferroni and Lorentz curve etc. of the proposed distribution. Order statistics and Renyi's entropy of the proposed distribution are also obtained. Estimation of the parameters of the distribution is performed via maximum likelihood method. A simulation study is performed to validate the maximum likelihood estimator (MLE). Finally,

Table 4: Comparison of Maximum Likelihood Estimates.

Model	MLE		Log-Liklyhood	AIC	BIC	AICC
	Parameters	Estimates				
BurrIII	c	2.3858890	-98.66771	205.3354	215.7959	205.7521
-Weibull	k	2.4533820				
	λ	1.7572900				
	β	0.6791234				
BurrIII	c	1.8321566	-106.6097	217.2194	222.4497	217.3419
	k	0.5343506				
Weibull	λ	0.9899448	-102.9768	209.9536	215.1839	210.0761
	β	0.9258876				

the BW distribution is fitted to real data sets in order to illustrate its applicability and usefulness.

References

[1] Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, 13(2):215-232.

[2] Mielke, P. W. (1973). Another family of distributions for describing and analyzing precipitation data. *J. Appl. Meterol*, 12, 275-280.

[3] Gove, J. H., Ducey, M. J., Leak, W. B.,& Zhang, L.(2008). Rotated sigmoid structures in managed uneven-aged northern hardwood stands: a look at the Burr Type III distribution. *Forestry*, vol. 81, no. 2, pp. 161-176.

[4] Bebbington, M. S, Lai, C. D.,& Zitikis, R.(2007). A flexible Weibull extension. *Reliability Engineering and System Safety*, 92(6), 719-726.

[5] Zhang, T.,& Xie, M. (2011). On the upper truncated Weibull distribution and its reliability implications. *Reliability Engineering System Safety*, 96(1), 194-200.

[6] Saad, J. Almalki., & Yuan, J. (2013). A new modified Weibull distribution. *Reliability Engi- neering and System Safety*, 111, 164-170.

[7] Mokhlis, N. A.(2005). Reliability of a stress-strength model with Burr type III distributions, Communications in Statistics. *Theory and Methods*, vol. 34, no. 7, pp. 1643-1657.

[8] Burr, I. W. & Cislak, P. J.(1968). On a general system of distributions,I: its curve-shape characteristics, II: the sample median,*Journal of the American Statistical Association*, vol. 63, pp. 627-635.

[9] Johnson, N. L., Kot, S., & Balakrishnan,N.(1995). Continuous Univariate Distributions, *John Wiley and Sons, New York, NY,USA*, 2nd edition.

[10] Cooray, K. & Ananda, M. M. A.(2008). A generalization of the half-normal distribution with applications to lifetime data. *Commun Stat Theory Method*, 37, 1323-1337.

[11] Tadikamalla, P. R. (1980). A Look at the Burr and Related Distributions. *International Statistical Review*, 48, 337-344.

[12] Rodrigues, R. N. (1977). A Guide to the Burr XII Distributions.*Biometrika*, 64, 129-134.

[13] Oluyede, B. O., Huang, S. & Mdlongwa, P. (2017). The Burr XII modified Weibull distribution: model, properties and applications. *Electronic Journal of Applied Statistical Analysis*,118-145.

[14] Oluyede, B. O., Huang, S., & Yang, T. (2015). A new class of generalized modified weibull distribution with applications. *Austrian Journal of Statistics*, 44:45-68.

[15] Weibull, W. A. (1951). Statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:293-296.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 113-124.

HIV Replication- INAR(1) Model- Comparison Between MLE And CLS Estimates

Mary George K[1], K. Kalyanaraman [2] and G. Meenakshi[3]
[1] Department of Statistics, Annamalai University, Annamalai Nagar-680002
Email:joicykunjumon@rediffmail.com
[2] Department of Statistics, Kerala University, Kerala, India
[3] Department of Statistics, Annamalai University, Annamalai Nagar-680002
Email:meenakshichinna@rediffmail.com

Abstract

Different models were suggested for the viral load in Human Immunodeficiency Virus - HIV treatment. This paper discuss HIV replication model as non negative replication. That is the viral load is non-negative integer valued function. A new model approach using INAR(1) is attempted regarding this. Here we are using INAR(1) by considering the replication as a counting process. The estimates of INAR(1) is done by using Maximum Likelihood Estimate and Conditional Least square methods. A comparison is also done between these estimates. Here we check the efficiency of estimates based on the both methods. Simulations were done for arbitrary values of parameters.

Key words: HIV treatment, INAR(1) model, Maximum Likelihood Estimate, Conditional Least Square Methods.

1. Introduction

The human immunodeficiency virus, or HIV weakens ones Immune system so it can't fight off common germs, viruses, fungi, and other invaders. It is the virus that causes AIDS, Acquired Immune Deficiency Syndrome. Someone with HIV can get sick from things that do not usually affect people, and people with AIDS tend to get certain uncommon diseases and illnesses. HIV attacks and destroys a type of white blood cell: the CD4 cell, also called the T-cell. Its job is to fight disease. Acquired immune deficiency syndrome (AIDS)

is the final stage of human immunodeficiency virus (HIV) infection [5, 8]. HIV begins its life cycle when it binds to a CD4 receptor and one of two co-receptors on the surface of a CD4+ T lymphocyte. Infected T lymphocytes are eliminated with a half-life of 2-4 days from the blood of an HIV-infected person by cytotoxic HIV components. But HIV uses proteins in the cell to make a copy of itself and then kills the cell. This can go on for 10 years or more without having any symptoms. AIDS is the later stage of HIV infection. When immune system has a very low level of CD4 cells, one cannot fight off things that most people would not get sick from. People with HIV are said to have AIDS when they get certain infections or cancers, called AIDS-defining illnesses, or when their CD4 count in a blood test is less than 200. HIV is also defined by numerous opportunistic infections and cancers that occur in the presence of HIV infection.

Antiretroviral Therapy (ART) effectively prevents HIV disease progression. A low level viral replication persists during ART. A stable viral replication count study prior to initiating treatment will help the patient and pattern of treatment. Understanding the mechanisms that maintain the HIV on ART is vital to developing strategies to eradicate the infection and / or to prevent viral replication. The viral replication is well suppressed on ART for 2 to 13 years. To determine if the mechanisms that maintain HIV proviruses in lymph nodes and whether there is ongoing exchange of infected cells, a complete study of viral replication is needed. These study of analysis is possible to quantify the rapidity of HIV replication. The viral replication during ART is different for different people depending on their age, health condition etc. Most individuals infected with HIV will progress to AIDS, if not treated. However, there is a tiny group of patients who develop AIDS very slowly or never at all. These patients are called non-progressors and many seem to have a genetic difference which prevents the virus from attaching to certain immune receptors. HIV begins its life cycle when it binds to a CD4 receptor and one of two co-receptors on the surface of a CD4+ T lymphocyte. The replication of HIV one is a multistage process. Each step is crucial to successful replication and is therefore potential target of entire retroviral drugs. Step one is infection of suitable host cells. Such as a CD4+T of lymphocytes. Entry of HIV into the cell requires the presence of certain receptors on the cell surface. This is called binding or fusion. ART prevents or reduce this process to a great extend. This paper intended to estimate a general tendency of the amount of viral replication during this period. The viral replication shows an increasing tendency. This paper developed by considering count of viral replication as Integer and pattern of replication follows an exponential distribution. The conclusions of this paper will help to predict the viral replication count and hence help the health departments to take necessary precaution steps.

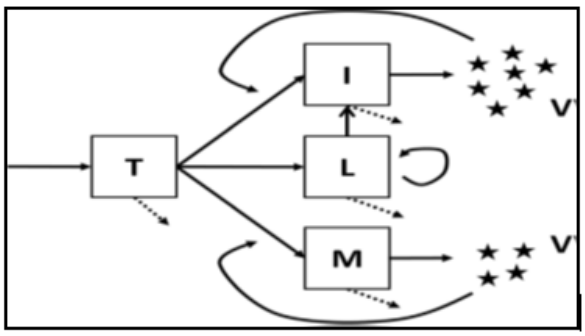
2. Review of Literature

Statistical analyses and modeling have contributed greatly to our understanding of the pathogenesis of HIV-1 infection. Various statistical methods, nonlinear mixed-effects models in particular, have been applied to model the CD4 and viral load trajectories. HIV infect CD4+T cells which play a central role in immune regulation. Many models developed for the replication of the virus and mortality of cells occurring in the lymphatic system [1].

The distribution considered in this paper is supported by Gaver D and P. Lewis (1980) introduced First order Auto regressive gamma sequences and point process by showing that there is an innovation process $\{\epsilon_n\}$ such that the sequence of random variables $\{X_n\}$ generated by the linear, additive first-order autoregressive scheme $X_n = \rho X_{n-1} + \epsilon_n$ are marginally distributed as gamma (λ, k) variables if $0 \leqq p \leqq 1$,[2]. The generalized new model for pth-order autoregressive processes with exponential marginal distributions EAR(p) is developed by A. J. Lawrance P. A. W. Lewis (1980), and an earlier model for first order moving average exponential processes is extended to qth-order, given an EMA(q) process. A mixed process, EARMA(p,q), incorporating aspects of both EAR(p) and EMA(q) correlation structures is then developed. The model of order limited to one is used in this paper according to the situation of HIV replication,[3].

The viral replication pattern shows an integer values only. M. A. ALOSH AND A. A. ALZAID, King Saud University (1987) derived a simple model for a stationary sequence of integervalued random variables is referred to as the integervalued autoregressive of order one (INAR(1)) process. The model is suitable for counting processes in which an element of the process at time t can be either the survival of an element of the process at time $t1$ or the outcome of an innovation process. The correlation structure and the distributional properties of the INAR(1) model are similar to those of the continuousvalued AR(1) process. Several methods for estimating the parameters of the model are discussed,[4].

In the simplest and earliest models of viral infection Perelson (2013), Uninfected target cells- T, Infected cells- I, Free virus-V, Target cells are assumed to be produced at constant rate λ, Die at rate dT per cell, A simple mass action infection term -βVT. This generates productively infected cells- I, which are lost at rate δ, larger than dT, to reflect viral effects in shortening the infected cell lifespan. Finally, free viruses are produced by infected cells at constant rate p per cell, and are cleared from circulation at rate c per virus.



$$\frac{dT}{dt} = \lambda - dTT - \beta VT$$
$$\frac{dI}{dt} = \beta VT - \delta I$$
$$\frac{dV}{dt} = pI - cV$$

This simple model was shown to be able to describe the kinetics of acute HIV infection.

Uninfected cells (T) can become infected by virus (V) to generate productively infected cells (I), long-lived infected cells (M) or latently infected cells (L). Latent infected cells may divide, which leaks to the productively infected class as latent cells are activated into cells producing virus. Dashed arrows indicate removal (death) of uninfected and infected cells, which occur at different rates. Equation (1) in the text considers only the uninfected cells, productively infected cells and virus. The combination of rug block, the production of infectious virus V and lead to the production of non-infectious virus.

R. Lakshmajayam and G.Meenakshi (2014) This paper introduces a new model for HIV replication. It is also designed to find out the HIV count of the succeeding period in the plasma.[6] R. Lakshmajayam and G.Meenakshi (2015) [7] Explained the determination of average HIV replication in the blood plasma using truncated logistic model.[6, 7] By considering law of mass action in the viral replication, in addition to the usual components of CD4+ , a truncated logistic distribution is suggested. The distribution is also used for future viral load prediction. Here Viral dynamic model is considered as function of the infected CD 4 + T cells, Blanket CD4 + T Cells, Lysing CD4+ T cells, Non Infected CD4 +T cells and viral load. The model of HIV replication for the future period in the viral dynamic study is give by

$$[P(H_2(t)/P(H_1(t) > 0] = \frac{BDe^{\frac{H_1(t)}{C}} + A}{De^{\frac{H_1(t)}{C}} + 1}$$

Where $H_1(t) = Y(t) + R_2(t) - \nu(t) + e, Y(t) = H(t) - \alpha[X(t)], R_2(t) = \delta_4 p R_1(t)$, $X(t)$ is powerful virus, $R_1(t)$- the number of virus replicated at the current period, $h(t)$- is existing virus in the plasma, $\nu(t)$ is the number of blanket cell death (cleared viron). Along with modified function of plasma virus by using the law of mass action that is $H_1(t) = [H(t)\alpha(x(t))]\delta_4 p R_1(t) - \delta_2 R(t)]$, where δ_2 death rate of blanket CD4+T cells, δ_4- death rate of lysing CD4+T cells, p- the number of virus released from the each lysing CD4 + T cells, α- rate of powerful virus. And $H_1(t) = C[H(t)A][B - H(t)], C > 0$, where $A = \alpha(x(t))$ is existing virus in the plasma, $B = \delta_4 p R_1(t)$-is maximum number of viral replication at the current period. $\delta_2 R(t)$ is the function of $H(t)$. D is a fixed constant. Viral replication for various counts and the corresponding probability changes are also noted in the above mentioned paper. The additions for the paper of Alan S Perelson is given in the paper of Jessica M. Conwaya,b and Alan S. Perelson (2015). Conditional least square estimator derived based on the paper of D. M. Simarmata, F. Novkanizaa), and Y. Widyaningsih (2017). Simulations were done in this paper based on the paper of Ogban, G.I. and Lebedev, K.A. (2016). Meenakshi and Lakshmi Priya (2017) suggested a new model as an extension of the above model by considering an additive concept. The viral load and CD 4+ T count examination is very much expensive experiment which is essential for the treatment of HIV. The prediction of viral load can be done using the mixed effect model suggested in the paper. Assuming α probability that virus becomes infectious, n_1, n_2, n_3- number of cells released from during interval $(0, t_1), (t_1, t_2), (t_2, t_3)$ respectively. The predictive distribution of largest HIV replication and corresponding depletion density is given by

$$g(x, y) = we^{\nu}(a/b)^{\nu}e^{-y}$$

Where $a = (1 - \tau\theta), \tau^2$ is the variation among the replication over the period, $y = y_n, y_1 < y_2 < \ < y_n$ are number of viral replication over the nth period of a HIV infected person, $\theta > 0$ is the parameter of the exponential distribution of a prior density of parameter μ, the average viral replication. a And b are from the distribution of the random variable CD4+T cell depletion, with density function of the form $f(x) = \dfrac{a^x e^{-x}}{b^x}$ with $a < b$ and $w = \dfrac{e^{y/\tau}}{\tau d}$, a and b are integers and y_n is the largest HIV replication over the period. Considering above papers, the process of replication is an integer valued one. By analysing the process pattern of replication, the process can be treated as a renewal process. The patients under ART had their own replication time in study period which should be modelled by a renewal process or counting process. The paper of Kuritzkes DR, Perelson AS, Ribeiro RM (2017) consider the influence of integrate inhibitors also. The nature and pattern study of HIV is done from the paper of Million Wesenu Demissie, J Biom (2018) which deals with the true cases of HIV in Uganda. William R Mc Manus, John M Coffin, Mary F Kearney (2019) worked on the cellular proliferation of HIV in Lymph nodes, which helped in the pattern of replication of HIV used in this paper.

Assuming that the replications of inter events of intervals are independent, the process

can be considered as a counting process. A counting process is said to have stationary increments if the distribution of number of events which occur in any interval depends only on the length of the interval. By studying the characteristics of HIV replications, the process of replication in inter events of intervals can be considered as a counting process. Each viral replication constitutes independent counting process. Modelling the replication series into an first order integer- valued Auto Regressive process [INAR(1)] as the count data recorded by a non-negative integer (0,1,2,.). The replication of viruses is always integers. Hence it is convenient if we represent the replication count like integers. Gaver and Lewis (1980) extended the process by the sequence $X_n = \rho X_{n-1} + \epsilon_n$ by identifying the innovation sequence ϵ_n in such a manner that for $0 \leq p \leq 1$. Where, ρ is the rate of replication (probability of evolution) per virus, X_n is the number of viral replication at time $t = n$, $\{\epsilon_n\}$ will be a sequence of independently and identically distributed (i.i.d.) random variables with exponential distribution having parameter $1/\mu$, as in the case of counting process. Hence, $Exp(1/\mu), \mu > 0, E(\epsilon_n) = \mu$, the average viral production $V(\epsilon_n) = \mu_2$

$$\begin{aligned} X_0 &= \epsilon_0 \\ X_n &= \rho X_{n-1} \quad with \quad probability \quad \rho \\ &= \rho X_{n-1} + \epsilon_n \quad with \quad probability \quad 1-\rho \end{aligned}$$

$\{X_n\}$ have mixed exponential distribution. The process giving rise to the sequence $\{X_n\}$ is named as Exponential Auto Regressive Process of First Order EAR(1) [2, 3]. Now,

$$X_n = \rho X_{n-1} + \epsilon_n = \sum_{j=0}^{\infty} \rho^j \epsilon_{n-j}$$

In this first order Auto regressive process certain assumptions were made as follows

1. The process $\{X_n\}$ is stationary, as the ratio of time intervals are continuous.

2. $\{\epsilon_n\}$ have an exponential distribution.

3. $\{X_n\}$ the sequence is additive in nature

2.1 Mean and Variance

$$\begin{aligned}
E(X_n) &= \rho E(X_{n-1}) + E(\epsilon_n) \\
&= \rho E(X_{n-1}) + \mu \\
E(X_n) &= \frac{\mu}{1-\rho} \\
V(X_n) &= \rho^2 V(X_{n-1}) + V(\epsilon_n) \\
V(X_n) &= \frac{\mu^2}{1-\rho^2} \\
X_n &\sim Exp[1/\frac{\mu}{1-\rho}]
\end{aligned}$$

3. Maximum Likelihood Estimator

$$\begin{aligned}
X_n &= \rho X_{n-1} + \epsilon_n \\
f(X_0, X_1, X_2, X_3, X_n) &= f(X_n/X_{n-1}, X_{n-2}, .X_0).f(X_{n-1}, X_{n-2}, .X_0) \\
&= f(X_n/X_{n-1}, X_{n-2}, .X_0).f(X_{n-1}/X_{n-2}, X_{n-3}, .X_0).f(X_{n-2}, X_{n-3}, ..X_0). \\
&\ldots\ldots\ldots\ldots\ldots \\
&\ldots\ldots\ldots\ldots\ldots \\
&= \prod_{t=1}^{n} f(X_t/X_{t-1}) f(X_0) \\
&= \prod_{t=1}^{n} \frac{\mu}{1-\rho} e^{-\frac{\mu}{(1-\rho)}t} f(X_0) \\
&= \frac{\mu^n}{(1-\rho)^n} exp\{-\sum_{t=1}^{n} \frac{\mu}{(1-\rho)}t\} f(X_0)
\end{aligned}$$

$L(X_n, \epsilon_n) = f(X_1, X_2, X_3, , \epsilon_n)$ the joint density function will be same as that above as X_n is linear function of ϵ_n

$$= \prod_{t=1}^{n} f(X_t, \epsilon_t)$$

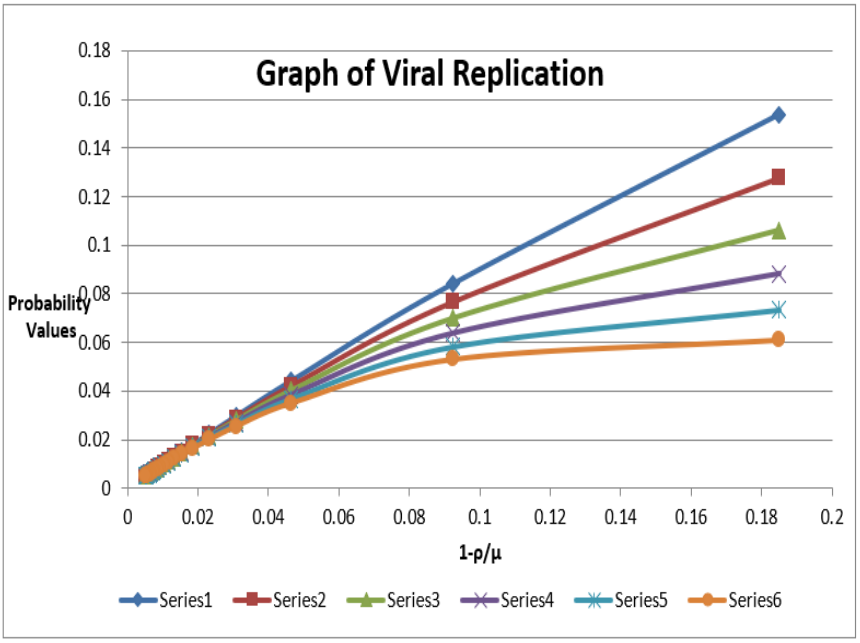
$$
\begin{aligned}
L(X_n, \epsilon_n) &= \frac{\mu^n}{(1-\rho)^n} exp\{-\sum_{t=1}^{n} \frac{\mu}{(1-\rho)} t\} f(X_0) \\
\log L &= n \log \mu - n \log (1-\rho) - \sum_{t=1}^{n} \frac{\mu}{(1-\rho)} t + \log f(X_0) \\
&= n \log \mu - n \log (1-\rho) - \frac{\mu}{(1-\rho)} . \frac{n(n+1)}{2} + \log f(X_0) \\
\frac{\partial \log L}{\partial \mu} &= \frac{n}{\mu} - \frac{1}{(1-\rho)} . \frac{n(n+1)}{2} \\
\frac{\partial^2 \log L}{\partial \mu^2} &= \frac{-n}{\mu^2} < 0, \quad Maximum \\
MLE \quad of \quad \mu &= \frac{2(1-\rho)}{(n+1)} \\
\frac{\partial \log L}{\partial \rho} &= \frac{-n}{(1-\rho)} + \frac{\mu}{(1-\rho)^2} . \frac{n(n+1)}{2} \\
\frac{\partial^2 \log L}{\partial \rho^2} &= \frac{-n}{(1-\rho)^2} - \frac{2\mu n(n+1)}{2(1-\rho)^3} < 0, \quad Maximum \\
MLE \quad of \quad \rho &= 1 - \frac{(n+1)}{2} \mu
\end{aligned}
$$

Simulations were done by taking different values for μ and ρ. For different time intervals, and for different values of μ and ρ, some of the plots are as follows.

4. Estimation using Conditional Least Square Method

The method of estimation using CLS is more suitable than the ordinary MLE method, since the current value is depending on the previous values [10]. The CLS method is used to find the parameter estimation by minimizing the square of difference between the Conditional Expectation of X_t given X_{t-1} with assumption $\epsilon_t \sim Exp(\mu)$ and the series is

μ	$\frac{(1-\rho)}{\mu}$	**Series 1**	**Series 2**	**Series 3**	**Series 4**	**Series 5**	**Series 6**
5	0.185	0.153754	0.127786	0.106203	0.088266	0.073358	0.060968
10	0.0925	0.084328	0.076877	0.070085	0.063893	0.058248	0.053102
20	0.04625	0.04416	0.042164	0.040258	0.038439	0.036701	0.035042
30	0.030833	0.029897	0.028989	0.028109	0.027256	0.026428	0.025626
40	0.023125	0.022596	0.02208	0.021575	0.021082	0.0206	0.020129
50	0.0185	0.018161	0.017828	0.017501	0.01718	0.016866	0.016556
60	0.015417	0.015181	0.014949	0.01472	0.014495	0.014273	0.014055
70	0.013214	0.013041	0.01287	0.012701	0.012534	0.012369	0.012207
80	0.011563	0.01143	0.011298	0.011168	0.01104	0.010913	0.010788
90	0.010278	0.010173	0.010069	0.009966	0.009864	0.009763	0.009663
100	0.00925	0.009165	0.00908	0.008997	0.008914	0.008832	0.008751
110	0.008409	0.008339	0.008269	0.0082	0.008131	0.008063	0.007995
120	0.007708	0.007649	0.00759	0.007532	0.007474	0.007417	0.00736
130	0.007115	0.007065	0.007015	0.006965	0.006916	0.006867	0.006818
140	0.006607	0.006564	0.00652	0.006477	0.006435	0.006392	0.00635
150	0.006167	0.006129	0.006091	0.006054	0.006016	0.005979	0.005943
160	0.005781	0.005748	0.005715	0.005682	0.005649	0.005617	0.005584
170	0.005441	0.005412	0.005382	0.005353	0.005324	0.005295	0.005266
180	0.005139	0.005113	0.005086	0.00506	0.005034	0.005009	0.004983



till $t = n$, the time interval

$$E(X_t/X_{t-1}) = \rho X_{t-1} + \mu$$

$$\sum_{t=1}^{n} [X_t - E(X_t/X_{t-1})]^2 = \sum_{t=1}^{n} [X_t - \rho X_{t-1} - \mu]^2$$

$$\frac{\partial L}{\partial \mu} = 2. \sum_{t=1}^{n} [X_t - \rho X_{t-1} - \mu] \times -1$$

$$\frac{\partial L}{\partial \mu} = 0, \quad 2. \sum_{t=1}^{n} X_t + \rho \sum_{t=1}^{n} X_{t-1} + n\mu = 0$$

$$\mu = \frac{1}{n} \sum_{t=1}^{n} X_t - \frac{\rho}{n} \sum_{t=1}^{n} X_{t-1}$$

$$\frac{\partial L}{\partial \rho} = 0, \quad 2 \times - \sum_{t=1}^{n} X_t X_{t-1} + \rho \sum_{t=1}^{n} X_{t-1}^2 + \mu \sum_{t=1}^{n} X_{t-1}$$

$$\rho = \frac{\sum_{t=1}^{n} X_t X_{t-1} - \mu \sum_{t=1}^{n} X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2}$$

$$\rho = \frac{\sum_{t=1}^{n} X_t X_{t-1} - (\frac{1}{n} \sum_{t=1}^{n} X_t - \frac{\rho}{n} \sum_{t=1}^{n} X_{t-1}) \sum_{t=1}^{n} X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2}$$

$$\rho = \frac{\sum_{t=1}^{n} X_t X_{t-1} - \frac{1}{n}\sum_{t=1}^{n} X_t . \sum_{t=1}^{n} X_{t-1} - \frac{\rho}{n}(\sum_{t=1}^{n} X_{t-1})^2}{\sum_{t=1}^{n} X_{t-1}^2}$$

$$\text{Estimate} \quad of \quad \rho = \frac{\sum_{t=1}^{n} X_t X_{t-1} - \frac{1}{n}\sum_{t=1}^{n} X_t . \sum_{t=1}^{n} X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2 - \frac{1}{n}(\sum_{t=1}^{n} X_{t-1})^2}$$

$$\text{Estimate} \quad of \quad \mu = \frac{1}{n}[\frac{\sum_{t=1}^{n} X_t \sum_{t=1}^{n} {X_{t-1}}^2 - \sum_{t=1}^{n} X_t X_{t-1} \sum_{t=1}^{n} X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2 - \frac{1}{n}(\sum_{t=1}^{n} X_{t-1})^2}]$$

If the number of viral replication at initial stage is known, the parameters μ and ρ can be easily derived for the nth time epoch.

5. Conclusion

This paper intended to give a contribution in the direction of viral replication using INAR(1) process. The main objective of this paper is to propose a new approach to the viral replication by considering the process as a counting process. Using the distribution we can find the number of viruses in active position according to the change of time. CLS estimate also can be used to estimate the viral progress, when we know the previous or initial number viruses in active position, we can predict the number of viruses when the duration of time suggested. Both estimates together will help to identify the condition of patient and help what type of treatment is to be adopted in future.

References

[1] Gaver D. and Lewis P. (1980) First order Auto regressive gamma sequences and point proces, *Adv.Appl.Prob* 12,727-745.

[2] Lowerance A. J. (1980), The mixed exponential solution for the First order Auto Regressive Model, *Journal of Applied Probability* 17, 546-552

[3] Lawerance A. J. and Lewis P. A. W. (1980), The Exponential AutoregressiveMoving Average Earma(P,Q) *ProcessJournal of Royal Statistical Society Statistical Methodology Series B*, 42 (2),150-161

[4] Alosh M. A. and Alzaid A. A., (1987) First-Order Integer-Valued Autoregressive (INAR(1)) PROCESS, *Journal Of Time Series Analysis* Vol. 8, No.3 pp 261-275.

[5] Alan S Perelson and Ruy M Ribeiro Perelson and Ribeiro (2013), Modeling the within-host dynamics of HIV Infection(2013) *BMC Biology* 11:96.

[6] Lakshmajayam R. and Meenakshi G.(2014) HIV Replication Model for The Succeeding Period Of Viral Dynamic Studies In Aids Clinical Trials *International Journal Of Mathematics And Statistics Invention (IJMSI)* 2321 4767.

[7] Lakshmajayam R. and Meenakshi G.(2015) A Statistical Model For Hiv Replication In The Infected CD4 +T - Cells *International Journal of Recent Scientific Research* Vol. 6, Issue, 6, pp.4858-4868.

[8] Ioannis Andrianakis ,Ian R. Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein and Richard G. White (2015) Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A tutorial and a case study on HIV in Uganda. pcbi. 1003968.

[9] Jessica M. Conwaya,b and Alan S. Perelson (2015) Post-treatment control of HIV infection; *Theoretical Biology and Biophysics, Los Alamos NationalLaboratory, Los Alamos*, NM 87545.

[10] D. M. Simarmata, F. Novkanizaa), and Y. Widyaningsih (2016) A Time Series Model: First-order Integer-valued Autoregressive (INAR(1)) *International Symposium on Current Progress in Mathematics and Sciences 2016 (ISCPMS 2016)* AIP Conf. Proc. 1862, 030157-1030157-8.

[11] Ogban, G.I. and Lebedev, K.A. (2016) Mathematical Study of the Dynamics of the Development of HIV. *Journal of Applied Mathematics and Physics*, 4, 66-72.

[12] Meenakshi G . and Lakshmi Priya S. (2017) A statistical modeling for viral replication in the CD4 +T cells dynamic by bayesian methodology *International Journal of Statistics and applied Mathematics* ISSN: 2456-1452 Maths 2017; 2(2): 43-49.

[13] Kuritzkes D. R., Perelson A. S. and Ribeiro R.M. (2017) Treatment with integrase inhibitor suggests a new interpretation of HIV RNA decay curves that reveals a subset of cells with slow integration. *PLoS Pathogens* 13(7): e1006478.

[14] Million Wesenu Demissie and Biom J. (2018) Statistical Modeling for the Survival of HIV/AIDS Patients Treated with Highly Active Anti-Retroviral Therapy (HAART): A Case Study at Dilchora Hospital, Dire Dawa, Ethiopia *Journal of Biometrics & Biostatistics* 9:5 DOI: 10.4172/2155-6180.1000416

[15] WMcManus W.R., Bale M.J., Spindler J., Wiegand A., Musick A., Patro S.C., Sobolewski M.D., Musick V.K., Anderson E.M., Cyktor J.C., Halvas E.K., Shao W., Wells D., Wu X., Keele B.F., Milush J.M., Hoh R., Mellors J.W.2, Hughes S.H., Deeks S.G., Coffin J.M., Kearney M.F. (2019) HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. *The Journal of Clinical Investigation* https:/doi.org/10.1172/JCI126714.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 125-154.

A Study on Some Generalizations of Gompertz Distribution and Gompertz-Makeham Distribution

Rani Sebastian and Aleesha Sebastian
Department of Statistics, St.Thomas College, Thrissur, Kerala, India
Email:ranikanjiram@yahoo.com

Abstract

A data of 30 observations representing the times of failures and running times for samples of devices from an eld-tracking study of a larger system was considered. The aim was to find the distribution which better fits the data among the four distibutions, namely, Gompertz, Gompertz-Makeham, Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham. After choosing possible models, estimates of the parameters for each distribution are estimated by Maximum Likelihood Estimation. These estimates are then used to visualize the density functions to get an idea of the fit. The AIC values and the log likelihood function estimated at the MLEs are computed. After computing these values, the Kolmogorov-Smirnov test to choose which model fits the best to the data is performed.

Key words: Gompertz distibution,Gompertz-Makeham distribution, Kumaraswamy Gompertz distribution,Kumaraswamy Gompertz-Makeham distribution.

1. Introduction

Modelling of interrelationship among naturally occurring phenomena is made possible by the use of distribution function and their properties. Because of this, considerable effort has been expended in the development of large classes of standard probability distributions along with relevant statistical methodologies. Here, we discuss about four distributions, namely, the two-parameter Gompertz distibution, three-parameter Gompertz-Makeham distribution, four-parameter Kumaraswamy Gompertz distribution and the five-parameter Kumaraswamy Gompertz-Makeham distribution.

The Gompertz distribution plays an important role in modeling survival times, human mortality and actuarial tables. According to the literature, the Gompertz distribution was formulated by Gompertz (1825) to fit mortality tables. The properties of the Gompertz distribution have been studied by many authors in recent years. Pollard and Valkovics (1992) were the first to study this distribution thoroughly. However, their results are true only in the case when the initial level of mortality is very close to zero. Many authors have contributed to the studies of statistical methodology and characterization of this model. Read (1983) discussed about the Gompertz distribution in general. Makany (1991) introduced a theoretical basis for Gompertz's curve. Rao and Damaraju (1992) discussed about the new better-than-used and other concepts for this distribution. Franses (1994) studied about fitting of a Gompertz curve. Chen (1997) developed an exact confidence interval and an exact joint confidence region for the parameters of the Gompertz distribution. Wu and Lee(1999) discussed the characterization of the mixtures of Gompertz distributions by conditional expectation of order statistics. Garg et al. (1970) and Johnson et al. (1995) studied the properties of the Gompertz distribution and obtained the maximum likelihood estimates for the parameters. Saraçoglu et al. (2009) discussed the stress-strength reliability in Gompertz case. El-Gohary et al. (2013) proposed the generalized Gompertz distribution. Minimol and Thomas (2014) obtained characterization of Gompertz and inverted Gompertz distributions based on certain properties of generalized record values. Sebastian (2017) discussed the generalization of Gompertz distribution and their applications.

A different version of Gompertz distribution which is called Gompertz-Makeham distribution was introduced by Makeham (1860). He introduced a constant (Makeham terms) that describe the age independent mortality and has received considerable attention in the literature. The Gompertz-Makeham family has been studied by Bailey(1978) and an expression using the Lambert W function for the quantile function was given by Jodrá (2009). The unimodality and the relationship between median value and the mean residual life time of the Gompertz-Makeham distribution was studied by Norström (1997).

The Kumaraswamy distribution as defined by Poondi Kumaraswamy (1980) has been identified as a viable alternative to Beta distribution because they both have the same basic shape properties (unimodal, uniantimodal, increasing, decreasing, monotone or constant). In spite of the fact that the Kumaraswamy distribution was introduced in 1980, further theoretical research on the distribution was not developed until very recently. Garg(2008) obtained the distribution of order statistics from Kumaraswamy distribution. Nadarajah (2008) further contributed to the studies on Kumaraswamy distribution. Jones(2009) discussed about Kumaraswamy distribution as a beta-type distribution with some tractability advantages. Cordeiro and Castro(2011)combined the works of Eugene et al.(2002) and Jones(2009) to construct a new family of generalized distri-

butions based on the Kumaraswamy distribution. According to Cordeiro and Castro, the Kumaraswamy-Generalized distribution has an advantage over the class of generalized beta distributions, since it does not involve any special function. Mitnik (2013) showed that Kumaraswamy variables exhibit closeness under exponentiation and under linear transformation and derived an expression for the moments of the general form of the distribution. The four-parameter Kumaraswamy Gompertz distribution and the estimation of the model parameters is discussed by Silva et al.(2015). The five-parameter generalized version of the Gompertz-Makeham distribution called Kumaraswamy Gompertz-Makeham is discussed by Chukwu and Ogunde(2016). Wang and Yu (2017) studied the point estimation and derived the interval estimation for the Kumaraswamy distribution and inverse estimators for the parameters of the Kumaraswamy distribution.

The objective is to study the generalizations of Gompertz distribution and Gompertz-Makeham distribution. In section 2, we discuss about the four distributions, namely, Gompertz, Gompertz-Makeham, Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham. The characteristic properties and Maximum Likelihood Estimation of these distributions are studied. The Kumaraswamy distribution and Kumaraswamy Generalized distributions are also discussed. In section 3, we compare the four distribution under study using a real life data. The comparison is done with the help of R software.

2. Gompertz Distribution

In probability and statistics, the Gompertz distribution is a continuous probability distribution, named after Benjamin Gompertz, a British actuary.It was formulated by Gompertz (1825) in connection with human mortality and actuarial tables. It plays an important role in modeling survival times, human mortality and actuarial data. It is often applied to describe the distribution of adult lifespans by demographers. More recently, computer scientists have also started to model the failure rates of computer codes by the Gompertz distribution. According to Jaheen (2003), Gompertz distribution has been used as a growth model, especially in epidemiological and biomedical studies. Gompertz distributions can be viewed as extensions of the exponential distributions because exponential distributions are limits of sequences of Gompertz distributions. Like the Weibull, the Gompertz distribution is characterized by two parameters.

2.1 Characteristic Properties

A continuous random variable X is said to have Gompertz distribution with parameters a and b if its pdf is given by,

$$f(x) = be^{ax}e^{-\frac{b}{a}(e^{ax}-1)} \ , \ x \geqslant 0 \ ; \ a, b > 0 \tag{1}$$

where, a is the shape parameter and b is the scale parameter.
Some of the typical Gompertz density functions for different values of a and $b = 1$ are depicted in Figure 1. It is quite evident that the Gompertz distribution is a positively skewed distribution.

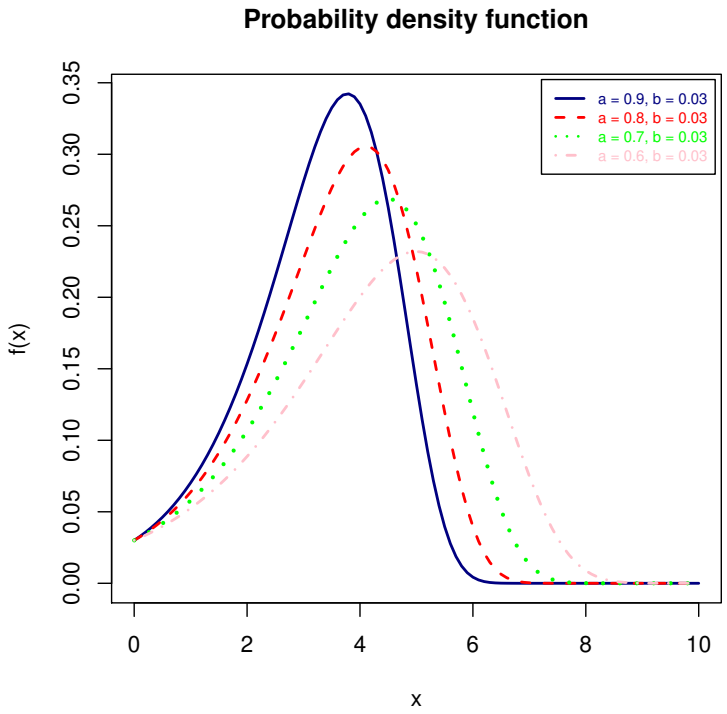


Figure 1: Plots of the probability density function of Gompertz distribution for $b = 0.03$ and different values of a

If X has a Gompertz distribution with probability density function as in equation (1), then

$$Y = \frac{b}{a}(e^{aX} - 1)$$

has a standard exponential distribution.

The cumulative distribution function of Gompertz distribution is,

$$F(x) = 1 - e^{-\frac{b}{a}(e^{ax}-1)} \tag{2}$$

Corresponding survival function is,

$$\begin{aligned} S(x) &= P(X > x) \\ &= e^{-\frac{b}{a}(e^{ax}-1)} \end{aligned} \tag{3}$$

Hazard rate function of Gompertz distribution is,

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= be^{ax} \end{aligned} \tag{4}$$

The plots of cumulative distibution function and hazard rate function of Gompertz distribution for different parameter values is shown in Figure 2.

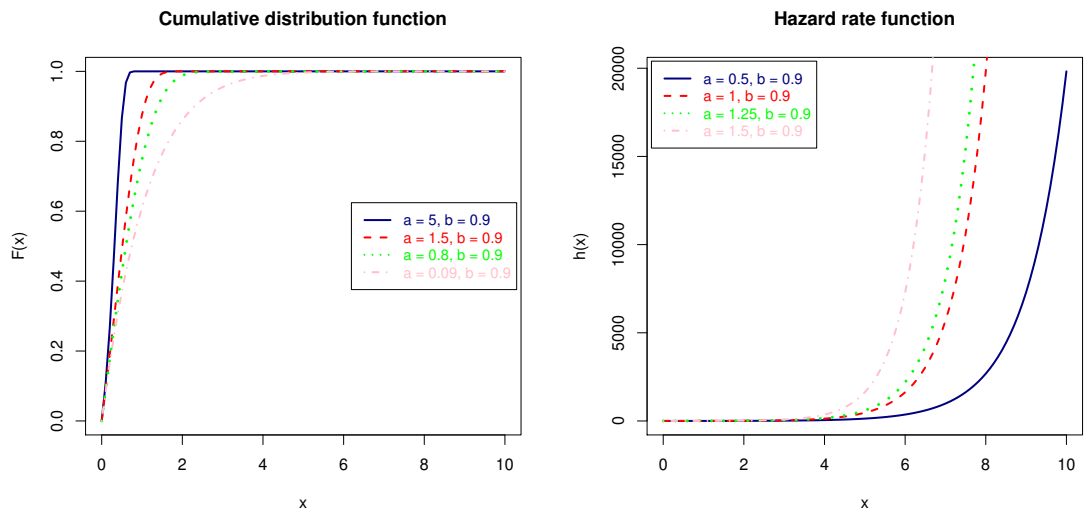


Figure 2: Plots of cdf and hazard rate function of Gompertz distribution

The hazard rate function of Gompertz distribution increases exponentially over time. When $a \to 0$, Gompertz distribution will tend to an exponential distribution with constant hazard rate. In the pdf of Gompertz distibution, when $a < 0$ (> 0), the hazard function decreases (increases) from $exp(a)$, and when $a = 0$, it reduces to the constant hazard function of an exponential distribution.

The inverse transform method is the most simple and straightforward procedure to generate samples drawn from a given probability distribution when its quantile function

exists in closed form. The quantile function of a continuous random variable is defined as the inverse of its cumulative distribution function. The quantile function $x_p = Q(p) = F^{-1}(p)$, for $0 < p < 1$, of the Gompertz distribution is obtained from (2). It follows that the quantile function x_p is,

$$x_p = \frac{1}{a}\ ln\left[1 - \frac{a}{b}\ ln(1-p)\right] \tag{5}$$

In particular the median of the Gompertz distribution can be written as,

$$Md(x) = M_d = \frac{1}{c}\ ln\left[1 - \frac{a}{b}\ ln(1-0.5)\right] \tag{6}$$

If a random variable X is distributed as Gompertz with parameters a and b, then its n^{th} moment around zero can be expressed as,

$$E(X^n) = \frac{\lambda e^{\frac{b}{a}}}{a}\int_1^\infty \frac{1}{a^n} e^{-\frac{b}{a}x}\ [ln(x)]^n dx \tag{7}$$

On simplification, we get;

$$E(X^n) = \frac{n!}{a^n}\ e^{\frac{b}{a}}\ E_1^{n-1}\left(\frac{b}{a}\right) \tag{8}$$

where,

$$\begin{aligned} E_s^n(z) &= \frac{1}{n!}\int_1^\infty [ln(x)]^n x^{-s} e^{-zx} dx \\ E_n(x) &= \int_1^\infty \frac{e^{-xt}}{t^n} dt \ \ ; n > 0,\ Re(x) > 0 \end{aligned}$$

and

$$E_s^0(z) = E_s(z),$$

is the generalized integro-exponential function.

The variance, skewness and kurtosis measures can now be calculated using the following relations:

$$\begin{aligned} Var(X) &= E(X^2) - E^2(X) \\ Skewness(X) &= \frac{E(X^3) - 3E(X)E(X^2) + 2E^3(X)}{Var^{\frac{3}{2}}(X)} \\ Kurtosis(X) &= \frac{E(X^4) - 4E(X)E(X^3) + 6E(X^2)E^2(X) - 3E^4(X)}{Var^2(X)} \end{aligned}$$

2.2 Estimation of Statistical Inference

In the view of estimating the parameters of the Gompertz distribution, we employ the method of the Maximum Likelihood Estimation.
Let $\underline{x} = (x_1, x_2, ..., x_n)$ be a random sample of 'n' independently and identically distributed random variables each having a Gompertz distribution defined in equation (2.). Then, the likelihood function $L(\underline{x}, a, b)$ is given by,

$$L(\underline{x}, a, b) \quad = \quad b^n \; e^{a\sum_{i=1}^{n} x_i} e^{-\frac{b}{a}\sum_{i=1}^{n}(e^{ax_i}-1)} \tag{9}$$

Let $L = L(\underline{x}, a, b)$.

$$logL \quad = \quad nlogb + a\sum_{i=1}^{n} x_i - \frac{b}{a}\sum_{i=1}^{n}(e^{ax_i} - 1) \tag{10}$$

Differentiating $logL$ with respect to a and b gives;

$$\frac{\partial logL}{\partial a} \quad = \quad \sum_{i=1}^{n} x_i + \frac{b}{a^2}\sum_{i=1}^{n}(e^{ax_i} - 1) - \frac{b}{a}\sum_{i=1}^{n} e^{ax_i}x_i \tag{11}$$

$$\frac{\partial logL}{\partial b} \quad = \quad \frac{n}{b} - \frac{1}{a}\sum_{i=1}^{n}(e^{ax_i} - 1) \tag{12}$$

The MLEs $\hat{a}$ and $\hat{b}$ are obtained by solving the nonlinear equations;

$$\frac{\partial logL}{\partial a} = 0 \; , \; \frac{\partial logL}{\partial b} = 0.$$

The maximization of log-likelihood function in equation (10) requires numerical methods. Iterative method such as Newton-Raphson may be used for parameter estimation.

3. Gompertz-Makeham Distribution

Gompertz Makeham distribution was introduced by another British actuary, Makeham. The fit to actuarial data provided by Gompertz distribution was examined by Makeham (1860) and he observed with specific examples that the fit could be improved with the modification now known as the Gompertz-Makeham distribution.

According to Finch (1990), the GompertzMakeham model provides a better fit empirical mortality distribution between the ages 30 and 85 years. The Gompertz-Makeham distribution gives a very good approximation of real demographic data. This explains why it is of interest to use the Gompertz-Makeham distribution for different approximations relating to life length theory. It gives very good approximations to empirical distributions

of life length not only for human populations but also for differnt biological arts.
The basic reason for making approximations of real demographic data with use of the Gompertz-Makeham distribution is that there are many different professions that have great use of these kind of approximations of life table data. One of the professions that have great use of the Gompertz-Makeham distribution are the insurance companies. The Gompertz-Makeham distribution would give them better possibilities to determine insurances that better explains the mortality among people for both accidents and natural deaths and it would be very helpful when the fees of the insurances are decided.

The Gompertz-Makeham distribution isn't only useful for approximating life lengths for human populations. It might even be of importance to use it in many different biological ways, for example, plant biology has great use of the Gompertz-Makeham distribution. Also, life lengths for different crops are an application where the Gompertz-Makeham distribution can be of importance to use. The possibility to study life lengths of differnt crops might even give the possibility to choose a treatment that in some sense raise the quality of these crops.

3.1 Characteristic Properties

A continuous random variable X is said to have a Gompertz-Makeham distribution with parameters a, b and λ if its pdf is given by,

$$f(x) = [\lambda + be^{ax}]e^{-\lambda x - \frac{b}{a}(e^{ax}-1)} \tag{13}$$

where, $x \geqslant 0$ and $a, b, \lambda > 0$.

The pdf of Gompertz distribution is represented in Figure 3 using different parameter values.
The cumulative distribution function of Gompertz-Makeham distribution is given by,

$$F(x) = 1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)} \tag{14}$$

Corresponding survival function is,

$$\begin{aligned} S(x) &= P(X > x) \\ &= e^{-\lambda x - \frac{b}{a}(e^{ax}-1)} \end{aligned} \tag{15}$$

Hazard rate function of Gompertz-Makeham distribution is,

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= [\lambda + be^{ax}] \end{aligned} \tag{16}$$

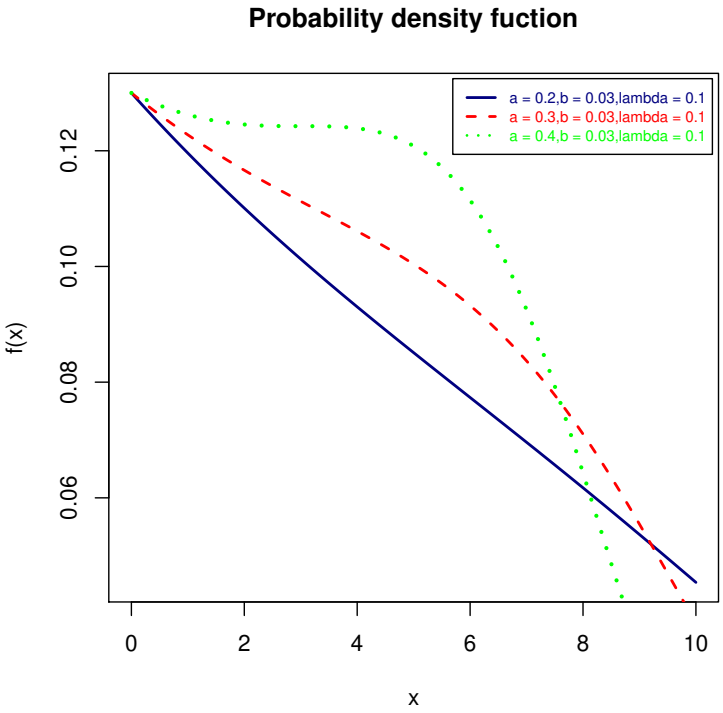


Figure 3: Plots of the probability density function of Gompertz-Makeham distribution for $b = 0.03,\ \lambda = 0.03$ and different values of a

The plots of cumulative distibution function and hazard rate function of Gompertz-Makeham distribution for different parameter values is shown in Figure 3.

The Cumulative hazard function of Gompertz-Makeham distribution is,

$$\begin{aligned} H(x) &= -ln(\overline{F}(x)) \\ &= \lambda x + \frac{b}{a}(e^{ax} - 1) \end{aligned} \tag{17}$$

A closed-form expression for the quantile function of the Gompertz Makeham distribution is expressed in terms of the principal branch of the Lambert W function. Thus, the quantile function of the Gompertz Makeham distribution can be expressed explicitly in terms of the Lambert W function.

The Lambert W function is defined as the solution to the equation,

$$W(z)e^{W(z)} = z \tag{18}$$

where, z is a complex number.

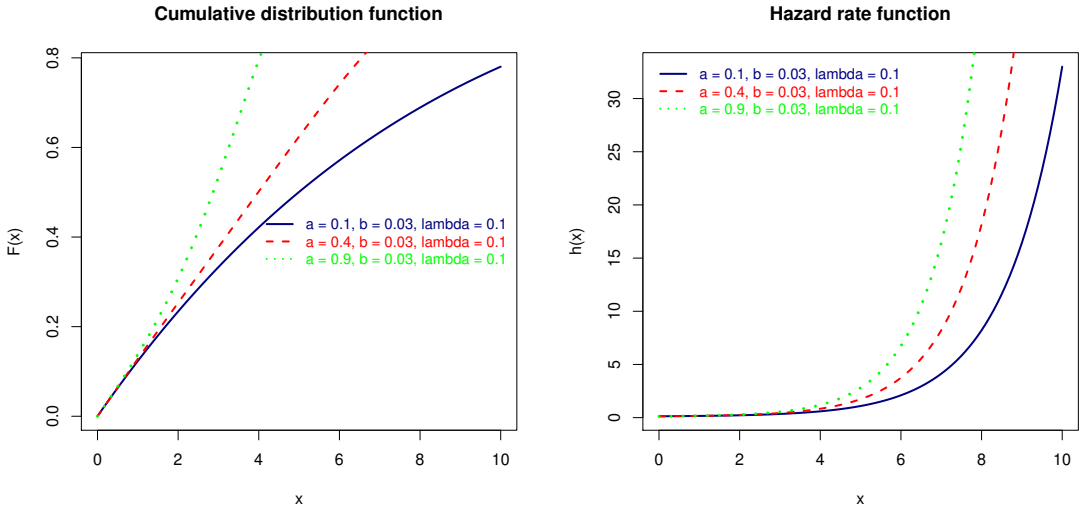


Figure 4: Plots of cdf and hazard rate function of Gompertz-Makeham distribution

Equation (18) has only one real solution for $z \geq 0$, this being $W_0(z)$
(called the principal branch). The quantile function $x_p = Q(p) = F^{-1}(p)$,
for $0 < p < 1$, of the Gompertz-Makeham distribution is obtained from equation (14). It follows that the quantile function x_p is,

$$x_p = \frac{b}{a\lambda} - \frac{1}{\lambda} ln(1-p) - \frac{1}{a} W_0\left(\frac{b}{\lambda} e^{\frac{b}{\lambda}} (1-p)^{-\frac{a}{\lambda}} \right) \tag{19}$$

where, W_0 denotes the principal branch of the Lambert W function.

In particular the median of the Gompertz-Makeham distribution can be written as,

$$Md(x) = M_d = \frac{b}{a\lambda} - \frac{1}{\lambda} ln(1-0.5) - \frac{1}{a} W_0\left(\frac{b}{\lambda} e^{\frac{b}{\lambda}} (1-0.5)^{-\frac{a}{\lambda}} \right) \tag{20}$$

3.2 Estimation of Statistical Inference:

In the view of estimating the parameters of Gompertz-Makeham distribution, we employ the method of Maximum Likelihood Estimation.
Let $\underline{x} = (x_1, x_2, ..., x_n)$ be a random sample of 'n' independently and identically distributed random variables each having a Gompertz-Makeham distribution defined in equation (2.13). Then, the likelihood function $L(\underline{x}, a, b, \lambda)$ is given by,

$$L(\underline{x}, a, b, \lambda) = \prod_{i=1}^{n} (\lambda + be^{ax_i}) e^{-\lambda \sum_{i=1}^{n} x_i - \frac{b}{a} \sum_{i=1}^{n} e^{ax_i} - 1} \tag{21}$$

Let $L = L(\underline{x}, a, b, \lambda)$.

$$logL = \sum_{i=1}^{n} log(\lambda + be^{ax_i}) - \lambda \sum_{i=1}^{n} x_i - \frac{b}{a} \sum_{i=1}^{n} (e^{ax_i} - 1) \tag{22}$$

Differentiating $logL$ with respect to a, b and λ gives;

$$\frac{\partial logL}{\partial a} = b \sum_{i=1}^{n} \frac{x_i e^{ax_i}}{\lambda + be^{ax_i}} + \frac{b}{a^2} \sum_{i=1}^{n} (e^{ax_i} - 1) - \frac{b}{a} \sum_{i=1}^{n} x_i e^{ax_i} \tag{23}$$

$$\frac{\partial logL}{\partial b} = \sum_{i=1}^{n} \frac{e^{ax_i}}{\lambda + be^{ax_i}} - \frac{1}{a} \sum_{i=1}^{n} (e^{ax_i} - 1) \tag{24}$$

$$\frac{\partial logL}{\partial \lambda} = \sum_{i=1}^{n} \frac{1}{\lambda + be^{ax_i}} - \sum_{i=1}^{n} x_i \tag{25}$$

The MLEs $\hat{a}$, $\hat{b}$ and $\hat{\lambda}$ are obtained by solving the nonlinear equations;

$$\frac{\partial logL}{\partial a} = 0 \ , \ \frac{\partial logL}{\partial b} = 0, \ \frac{\partial logL}{\partial \lambda} = 0.$$

The maximization of log likelihood function in equation (22) requires numerical methods. Iterative method such as Newton-Raphson may be used for parameter estimation.

4. Kumaraswamy Distribution

The Kumaraswamy distribution introduced by Poondi Kumaraswamy (1980) is a two-parameter continuous probability distribution with double-bounded support. The Kumaraswamy distribution was originally conceived to model hydrological phenomena, and has been used for this but also for other purposes. It is very similar to the Beta distribution, and can thus assume a strikingly large variety of shapes and be used to model many random processes and uncertainties.

Compared to the Beta distribution, the Kumaraswamy distribution has the key advantage of a closed-form cumulative distribution function. This makes it much better suited than the Beta distribution for computation-intensive activities like simulation modeling and the estimation of models by simulation-based methods. However, in spite of the fact that the Kumaraswamy distribution was introduced in 1980, further theoretical research on the distribution was not developed until very recently.

Many lifetime distribution models have successfully served as population models for risk

analysis and reliability mechanisms. The Kumaraswamy distribution is one of these distributions which is particularly useful to many natural phenomena whose outcomes have lower and upper bounds or bounded outcomes in the biomedical and epidemiological research. Due to its beta-type and better than beta distribution by its explicit expression of quantile function, the Kumaraswamy distribution has received considerable attention in the literature. Its hazard function can be unimodal, uniantimodal, increasing, decreasing and constant. This shows that the Kumaraswamy distribution can model a wide variety of data sets.

4.1 Characteristic properties

A continuous random variable X is said to have a Kumaraswamy distribution with parameters r, u if its pdf is given by,

$$f(x) = rux^{r-1}(1-x^r)^{u-1} \tag{26}$$

where, $0 < x < 1$ and $r, u > 0$. Here, both r and u are shape parameters.

The pdf of Kumaraswamy distribution is represented in Figure 5 using different parameter values.
The cumulative distribution function of Kumaraswamy distribution is,

$$F(x) = 1 - (1-x^r)^u \tag{27}$$

Corresponding survival function is,

$$\overline{F}(x) = (1-x^r)^u \tag{28}$$

Hazard rate function of Kumaraswamy distribution is,

$$h(x) = r\ u\ x^{r-1}\ (1-x^r)^{-1} \tag{29}$$

The plots of cumulative distibution function and hazard rate function of Kumaraswamy distribution for different parameter values is shown in Figure 6.

The Cumulative hazard function of Kumaraswamy distribution is,

$$\begin{aligned} H(x) &= -ln(\overline{F}(x)) \\ &= -u\ ln(1-x^r) \end{aligned} \tag{30}$$

The quantile function $x_p = Q(p) = F^{-1}(p)$, for $0 < p < 1$, of the Kumaraswamy distribution is obtained from (27). It follows that the quantile function x_p is,

$$x_p = [1-(1-p)^{\frac{1}{u}}]^{\frac{1}{r}} \tag{31}$$

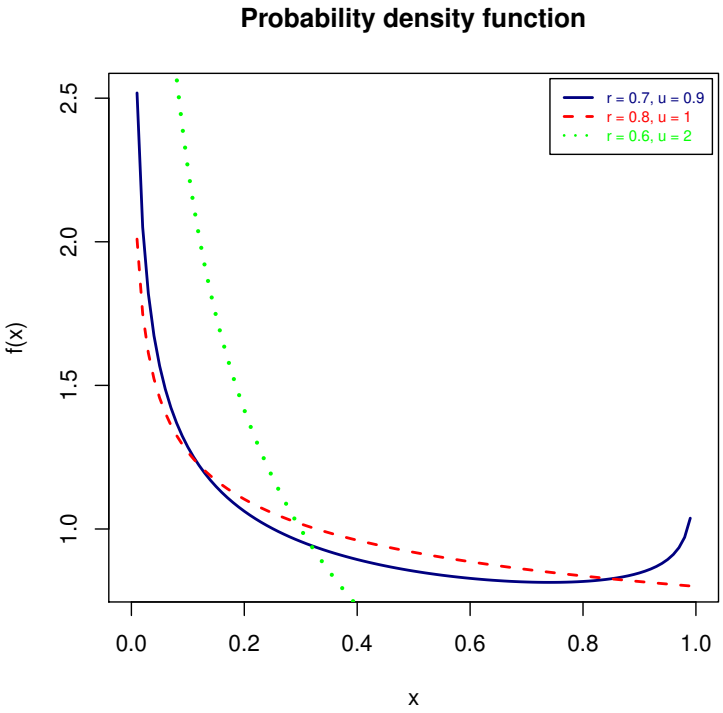


Figure 5: Plots of probability density function of Kumaraswamy distribution for different values of r and u

In particular the median of the Kumaraswamy distribution can be written as,

$$Md(x) = M_d = [1 - (1 - 0.5)^{\frac{1}{u}}]^{\frac{1}{r}} \tag{32}$$

If a random variable X is distributed as Kumaraswamy with parameters r and u, then its n^{th} moment around zero can be expressed as,

$$E(X^n) = u\ B\left(1 + \frac{n}{r}, u\right) \tag{33}$$

They exist for all $n > -r$. In particular,

$$\begin{aligned} E(X) &= u\ B\left(1 + \frac{1}{r}, u\right) \\ Var(X) &= u\ B\left(1 + \frac{2}{r}, u\right) - \left[u\ B\left(1 + \frac{1}{r}, u\right)\right]^2 \end{aligned} \tag{34}$$

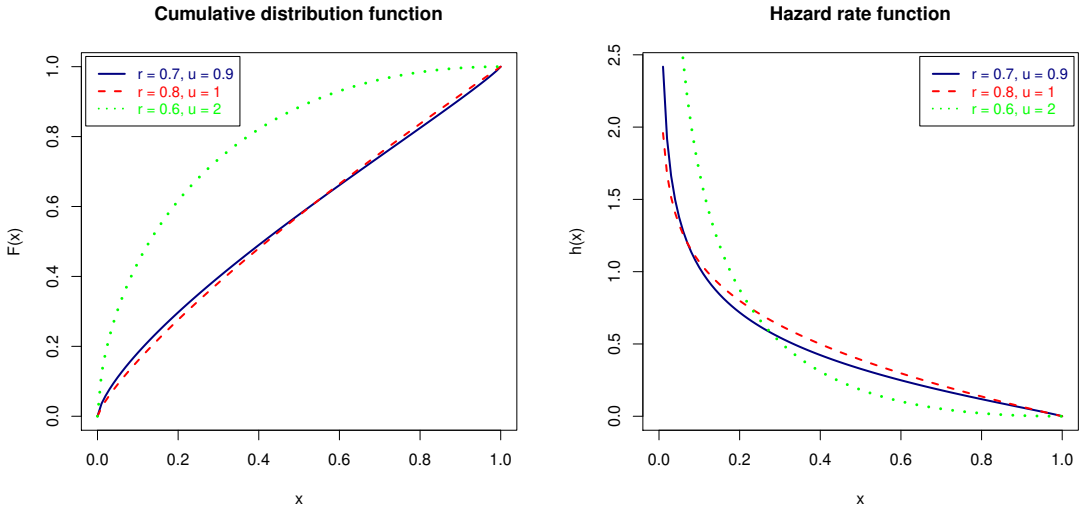


Figure 6: Plots of cdf and hazard rate function of Kumaraswamy distribution

5. Kumaraswamy Generalized Distribution

Combining the work of Eugene et al. (2002) and Jones (2004) lead to the construction of a new class of Kumaraswamy generalized distribution. From an arbitrary parent cumulative density function, $G(x)$, the cumulative density function, $F(x)$ of the Kumaraswamy G distribution is defined by,

$$F(x) = 1 - (1 - G(x)^r)^u \tag{35}$$

where, $r, u > 0$.

Correspondingly, the density function of this family is given by,

$$f(x) = r\ u\ g(x)\ G(x)^{r-1}\ (1 - G(x)^r)^{u-1} \tag{36}$$

For every given baseline cumulative function G, this defines a family of distributions. Clearly, the Kumaraswamy G distribution for $r = u = 1$ is the baseline distribution. The Kumaraswamy G distribution has same parameters of the G distribution plus two additional shape parameters $r > 0$ and $u > 0$. The Kumaraswamy G family of densities in equation (36) allows for greater flexibility of its tails and can be widely applied in many areas of biology and engineering.

6. Kumaraswamy Gompertz Distribution

The probability density function of the four-parameter Kumaraswamy Gompertz Distribution is defined from equation (36) by taking $g(x)$ to be equal to the pdf defined in equation (1).

6.1 Characteristic properties

A continuous random variable X is said to have a Kumaraswamy Gompertz distribution with parameters a, b, r and u if its pdf is given by,

$$\begin{aligned} f(x) &= rube^{ax}e^{-\frac{b}{a}(e^{ax}-1)}[1-e^{-\frac{b}{a}(e^{ax}-1)}]^{r-1} \\ &\quad [1-[1-e^{-\frac{b}{a}(e^{ax}-1)}]^r]^{u-1} \end{aligned} \tag{37}$$

where, $x > 0$ and $a, b, r, u > 0$.

The pdf of Kumaraswamy Gompertz distribution with different parameter values is represented in Figure 7.

The cumulative distribution function of Kumaraswamy Gompertz distribution is given by,

$$F(x) = 1-[1-[1-e^{-\frac{b}{a}(e^{ax}-1)}]^r]^u \tag{38}$$

Corresponding survival function is,

$$\begin{aligned} S(x) &= P(X > x) \\ &= [1-[1-e^{-\frac{b}{a}(e^{ax}-1)}]^r]^u \end{aligned} \tag{39}$$

Hazard rate function of Kumaraswamy Gompertz distribution is,

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= \frac{rube^{ax}e^{-\frac{b}{a}(e^{ax}-1)}[1-e^{-\frac{b}{a}(e^{ax}-1)}]^{r-1}}{1-[1-e^{-\frac{b}{a}(e^{ax}-1)}]^r} \end{aligned} \tag{40}$$

The plots of cumulative distibution function and hazard rate function of Kumaraswamy Gompertz distribution for different parameter values is shown in Figure 8.

The quantile function of a continuous random variable is defined as the inverse of its cumulative distribution function.

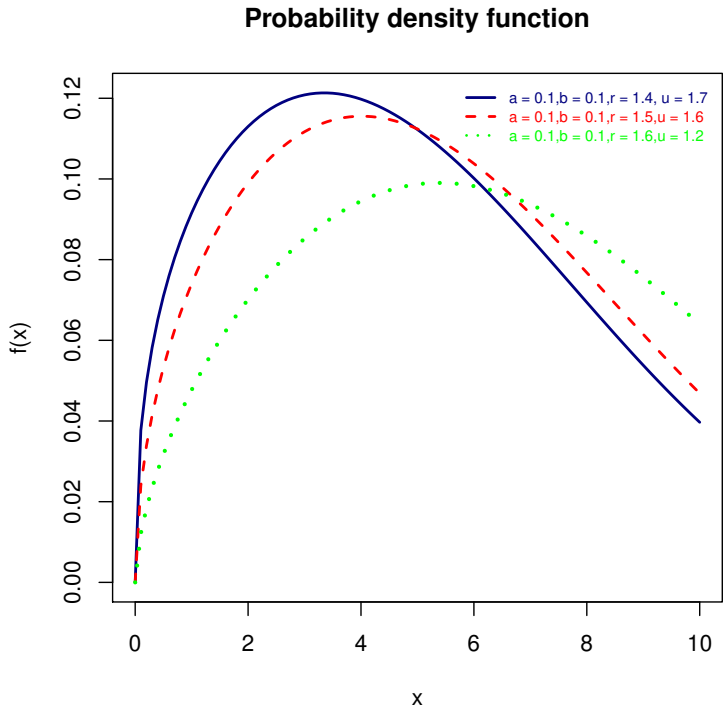


Figure 7: Plots of the probability density function of Kumaraswamy Gompertz distribution for different values of a, b, r and u

The quantile function $x_p = Q(p) = F^{-1}(p)$, for $0 < p < 1$, of the Kumaraswamy Gompertz distribution is obtained from equation (37). It follows that the quantile function x_p is,

$$x_p = \frac{1}{a}\, ln[1 - \frac{a}{b}\, ln(1 - [1 - (1 - p)^{\frac{1}{u}}]^{\frac{1}{r}})] \tag{41}$$

In particular the median of the Gompertz distribution can be written as,

$$Md(x) = M_d = \frac{1}{a}\, ln[1 - \frac{a}{b}\, ln(1 - [1 - (1 - 0.5)^{\frac{1}{u}}]^{\frac{1}{r}})] \tag{42}$$

6.2 Estimation of Statistical Inference

In the view of estimating the parameters of Kumaraswamy Gompertz distribution, we employ the method of Maximum Likelihood Estimation.

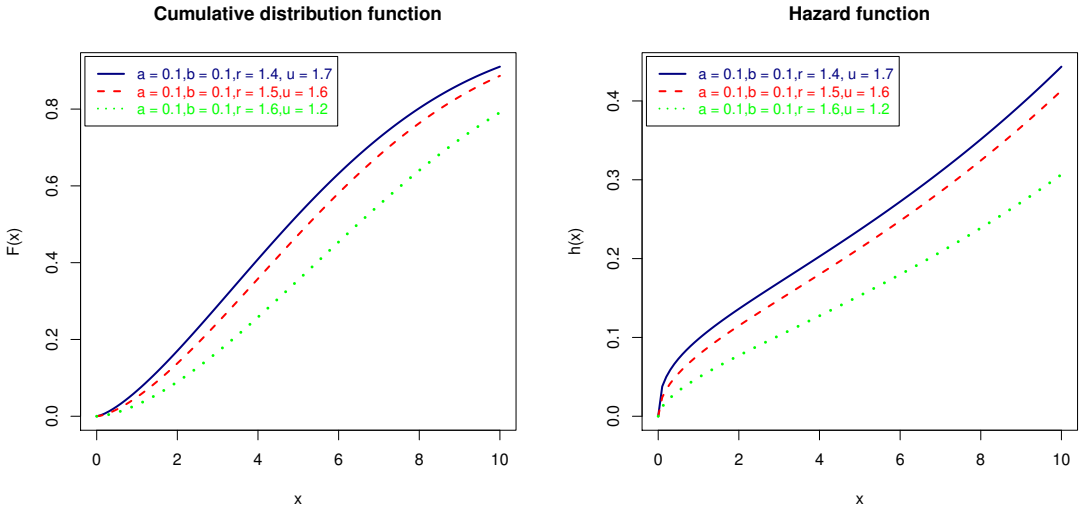


Figure 8: Plots of cdf and hazard rate function of Kumaraswamy Gompertz distribution

Let $\underline{x} = (x_1, x_2, ..., x_n)$ be a random sample of 'n' independently and identically distributed random variables each having a Kumaraswamy Gompertz distribution defined in equation (37).

Let $\xi = -\frac{b}{a}(e^{ax_i} - 1)$.

Then, the likelihood function $L(\underline{x}, a, b, r, u)$ is given by,

$$\begin{aligned} L(\underline{x}, a, b, r, u) &= (r\ u)^n \prod_{i=1}^{n} b\ e^{ax_i} \prod_{i=1}^{n} e^{\xi} \prod_{i=1}^{n} (1 - e^{\xi})^{r-1} \\ &\quad \prod_{i=1}^{n} [1 - (1 - e^{\xi})^r]^{u-1} \end{aligned} \tag{43}$$

Let $L = L(\underline{x}, a, b, r, u)$.

$$\begin{aligned} logL &= n[logr + logu] + \sum_{i=1}^{n} log(be^{ax_i}) + (r-1) \sum_{i=1}^{n} log(1 - e^{\xi}) + \\ &\quad \sum_{i=1}^{n} \xi + (u-1) \sum_{i=1}^{n} log(1 - (1 - e^{\xi})^r) \end{aligned} \tag{44}$$

Differentiating $logL$ with respect to a, b, r and u gives;

$$\begin{aligned}\frac{\partial logL}{\partial a} &= \sum_{i=1}^{n}\frac{bx_ie^{ax_i}}{be^{ax_i}}+\sum_{i=1}^{n}\left[\frac{b}{a^2}(e^{ax_i}-1)-\frac{b}{a}e^{ax_i}x_i\right]- \\ &\quad (r-1)\sum_{i=1}^{n}\frac{e^{\xi}}{1-e^{\xi}}\left[\frac{b}{a^2}(e^{ax_i}-1)-\frac{b}{a}e^{ax_i}x_i\right]+ \\ &\quad (u-1)\sum_{i=1}^{n}\frac{re^{\xi}(1-e^{\xi})^{r-1}}{1-(1-e^{\xi})^{r}}\left[\frac{b}{a^2}(e^{ax_i}-1)-\frac{b}{a}e^{ax_i}x_i\right] \\ &= \sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\left[\left(\frac{b}{a^2}(e^{ax_i}-1)-\frac{b}{a}e^{ax_i}x_i\right)\right. \\ &\quad \left.\left(1-\frac{(r-1)e^{\xi}}{(1-e^{\xi})}+\frac{r(u-1)e^{\xi}(1-e^{\xi})^{r-1}}{1-(1-e^{\xi})^{r}}\right)\right] \qquad (45)\end{aligned}$$

$$\begin{aligned}\frac{\partial logL}{\partial b} &= \sum_{i=1}^{n}\frac{e^{ax_i}}{be^{ax_i}}-\frac{1}{a}\sum_{i=1}^{n}(e^{ax_i}-1)+\frac{(r-1)}{a}\sum_{i=1}^{n}\frac{e^{\xi}(e^{ax_i}-1)}{1-e^{\xi}} \\ &\quad -\frac{r(u-1)}{a}\sum_{i=1}^{n}\left[\frac{e^{\xi}(1-e^{\xi})^{r-1}(e^{ax_i}-1)}{1-(1-e^{\xi})^{r}}\right] \\ &= \sum_{i=1}^{n}\left[\frac{e^{ax_i}-\frac{1}{a}(e^{ax_i}-1)be^{ax_i}}{be^{ax_i}}\right]+ \\ &\quad \sum_{i=1}^{n}\frac{e^{\xi}(e^{ax_i}-1)}{a(1-e^{\xi})}\left[(r-1)-\frac{r(u-1)(1-e^{\xi})^{r}}{1-(1-e^{\xi})^{r}}\right] \qquad (46)\end{aligned}$$

$$\begin{aligned}\frac{\partial logL}{\partial r} &= \frac{n}{r}+\sum_{i=1}^{n}log(1-e^{\xi})+ \\ &\quad (u-1)\sum_{i=1}^{n}\frac{(1-e^{\xi})^{r}log(1-e^{\xi})}{1-(1-e^{\xi})^{r}} \\ &= \frac{n}{r}+\sum_{i=1}^{n}log(1-e^{\xi})\left[1-\frac{(u-1)(1-e^{\xi})^{r}}{1-(1-e^{\xi})^{r}}\right] \qquad (47)\end{aligned}$$

$$\frac{\partial logL}{\partial u} = \frac{n}{u}+\sum_{i=1}^{n}log[1-(1-e^{\xi})^{r}] \qquad (48)$$

The MLEs $\hat{a}$, $\hat{b}$, $\hat{r}$ and $\hat{u}$ are obtained by solving the nonlinear equations;

$$\frac{\partial logL}{\partial a}=0\ ,\ \frac{\partial logL}{\partial b}=0,\ \frac{\partial logL}{\partial r}=0,\ \frac{\partial logL}{\partial u}=0$$

The maximization of log likelihood function in equation (44) requires numerical methods. Iterative method such as Newton-Raphson may be used for parameter estimation.

7. Kumaraswamy Gompertz-Makeham Distribution

Kumaraswamy Gompertz-Makeham distribution is quite flexible and can have increasing, decreasing and bathtub-shaped failure rate function depending on its parameters making it effective in modeling survival data and reliability problems. The probability density function of the five-parameter Kumaraswamy Gompertz Distribution is defined from equation (36) by taking $g(x)$ to be equal to the pdf defined in equation (13).

7.1 Characteristic properties

A continuous random variable X is said to have a Kumaraswamy Gompertz-Makeham distribution with parameters a, b, r, u and λ if its pdf is given by,

$$\begin{aligned} f(x) &= ru[\lambda + be^{ax}]e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}[1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r-1} \\ &\quad [1 - [1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r}]^{u-1} \end{aligned} \tag{49}$$

where, $x \geq 0$ and $a, b, r, u, \lambda > 0$.

Plots of probability density function of Kumaraswamy Gompertz-Makeham distribution for different values of a, b, r, u and λ is shown in Figure 9.

Corresponding survival function is,

$$\overline{F}(x) = [1 - [1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r}]^{u} \tag{50}$$

Hazard rate function of Kumaraswamy Gompertz-Makeham distribution is,

$$h(x) = \frac{ru[\lambda + be^{ax}]e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}[1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r-1}}{1 - [1 - e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r}} \tag{51}$$

The plots of cumulative distibution function and hazard rate function of Kumaraswamy Gompertz-Makeham distribution for different parameter values is shown in Figure 10.

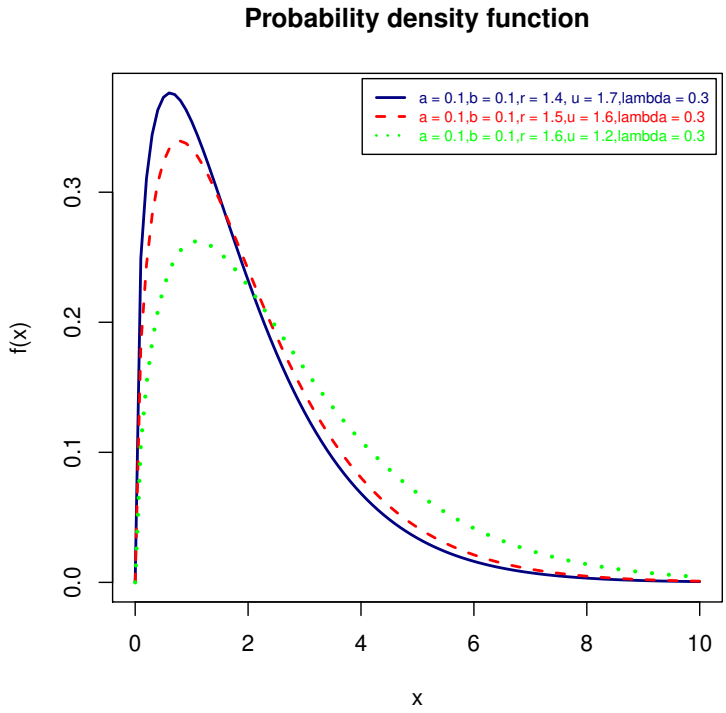


Figure 9: Plots of probability density function of Kumaraswamy Gompertz-Makeham distribution for different values of a, b, r, u and λ

7.2 Estimation of Statistical Inference

In the view of estimating the parameters of Kumaraswamy Gompertz-Makeham distribution, we employ the method of Maximum Likelihood Estimation.
Let $\underline{x} = (x_1, x_2, ..., x_n)$ be a random sample of 'n' independently and identically distributed random variables each having a Kumaraswamy Gompertz-Makeham distribution defined in equation (49).
Let $\xi = -\lambda x - \frac{b}{a}(e^{ax} - 1)$.

Then, the likelihood function $L(\underline{x}, a, b, r, u, \lambda)$ is given by,

$$L(\underline{x}, a, b, r, u, \lambda) = (ru)^n \prod_{i=1}^{n}(be^{ax_i} + \lambda) \prod_{i=1}^{n} e^{\xi} \prod_{i=1}^{n}(1 - e^{\xi})^{r-1} \prod_{i=1}^{n}[1 - (1 - e^{\xi})^r]^{u-1} \quad (52)$$

Let $L = L(\underline{x}, a, b, r, u, \lambda)$.

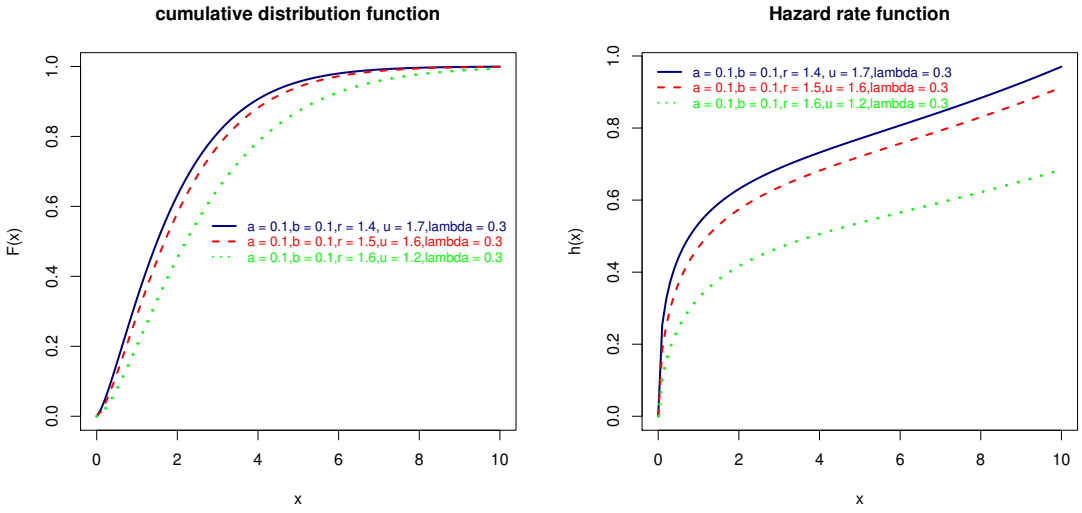


Figure 10: Plots of cdf and hazard rate function of Kumaraswamy Gompertz-Makeham distribution

$$\begin{aligned} logL &= n[logr + logu] + \sum_{i=1}^{n} log(be^{ax_i} + \lambda) + \sum_{i=1}^{n} \xi + \\ &\quad (r-1)\sum_{i=1}^{n} log(1-e^{\xi}) + (u-1)\sum_{i=1}^{n} log[1-(1-e^{\xi})^r] \end{aligned} \tag{53}$$

Differentiating $logL$ with respect to a, b, r, u and λ gives;

$$\begin{aligned} \frac{\partial logL}{\partial a} &= \sum_{i=1}^{n}\left[\frac{bx_ie^{ax_i}}{be^{ax_i}+\lambda} + b\left(\frac{e^{ax_i}-1}{a^2} - \frac{x_ie^{ax_i}}{a}\right) + \right. \\ &\quad \frac{b(r-1)e^{\xi}}{1-e^{\xi}}\left(-\frac{e^{ax_i}-1}{a^2} + \frac{x_ie^{ax_i}}{a}\right) - \\ &\quad \left.\frac{br(u-1)e^{\xi}(1-e^{\xi})^{r-1}}{1-(1-e^{\xi})^r}\left(-\frac{e^{ax_i}-1}{a^2} + \frac{x_ie^{ax_i}}{a}\right)\right] \\ &= \sum_{i=1}^{n}\left[\frac{be^{\xi}}{1-e^{\xi}}\left(\frac{x_ie^{ax_i}}{a} - \frac{e^{ax_i}-1}{a^2}\right)\left((r-1) - \frac{r(u-1)(1-e^{\xi})^r}{1-(1-e^{\xi})^r}\right) + \right. \\ &\quad \left.\left(\frac{b}{be^{ax_i}+\lambda}\right)\left[x_ie^{ax_i} + (be^{ax_i}+\lambda)\left(\frac{e^{ax_i}-1}{a^2} - \frac{x_ie^{ax_i}}{a}\right)\right]\right] \end{aligned} \tag{54}$$

$$\begin{aligned}
\frac{\partial logL}{\partial b} &= \sum_{i=1}^{n} \frac{e^{ax_i} - \frac{1}{a}(e^{ax_i}-1)(be^{ax_i}+\lambda)}{be^{ax_i}+\lambda} + \frac{r-1}{a}\sum_{i=1}^{n}\frac{e^{\xi}(e^{ax_i}-1)}{1-e^{\xi}} - \\
&\quad \frac{r(u-1)}{a}\sum_{i=1}^{n}\frac{e^{\xi}(1-e^{\xi})^{r-1}(e^{ax_i}-1)}{1-(1-e^{\xi})^{r}} \\
&= \sum_{i=1}^{n}\left[\frac{e^{ax_i} - \frac{1}{a}(e^{ax_i}-1)(be^{ax_i}+\lambda)}{be^{ax_i}+\lambda}\right] + \\
&\quad \sum_{i=1}^{n}\frac{e^{\xi}(e^{ax_i}-1)}{a(1-e^{\xi})}\left[(r-1) - \frac{r(u-1)(1-e^{\xi})^{r}}{1-(1-e^{\xi})^{r}}\right] \qquad (55) \\
\frac{\partial logL}{\partial r} &= \frac{n}{r} + \sum_{i=1}^{n} log(1-e^{\xi}) - (u-1)\sum_{i=1}^{n}\frac{(1-e^{\xi})^{r}log(1-e^{\xi})}{1-(1-e^{\xi})^{r}} \\
&= \frac{n}{r} + \sum_{i=1}^{n} log(1-e^{\xi})[1 - \frac{(u-1)(1-e^{\xi})^{r}}{1-(1-e^{\xi})^{r}}] \qquad (56) \\
\frac{\partial logL}{\partial u} &= \frac{n}{u} + \sum_{i=1}^{n} log[1-(1-e^{\xi})^{r}] \qquad (57)
\end{aligned}$$

The MLEs $\hat{a}$, $\hat{b}$, $\hat{r}$, $\hat{u}$ and $\hat{\lambda}$ are obtained by solving the nonlinear equations;

$$\frac{\partial logL}{\partial a} = 0\ ,\ \frac{\partial logL}{\partial b} = 0,\ \frac{\partial logL}{\partial r} = 0,\ \frac{\partial logL}{\partial u} = 0,\ \frac{\partial logL}{\partial \lambda} = 0$$

The maximization of the log likelihood function in equation (2.53) requires numerical methods. Iterative method such as Newton-Raphson can be used for parameter estimation.

8. Application

To compare the four distributions, namely: Gompertz, Gompertz-Makeham, Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham, we fit these distributions to an uncensored data set. The data represents the times of failures and running times for samples of devices from an eld-tracking study of a larger system. The data has 30 observations and it is as follows:
2.75, 0.13, 1.47, 0.23, 1.81, 0.30, 0.65, 0.10, 3.00, 1.73, 1.06, 3.00, 3.00, 2.12, 3.00, 3.00, 3.00, 0.02, 2.61, 2.93, 0.88, 2.47, 0.28, 1.43, 3.00, 0.23, 3.00, 0.80, 2.45, 2.66. Computations are done with the help of R software. Now, consider the two-parameter Gompertz distibution ($G(a,b)$), three-parameter Gompertz-Makeham distribution ($GM(a,b,\lambda)$), four-parameter

Kumaraswamy Gompertz distribution ($KG(a,b,r,u)$) and five-parameter Kumaraswamy Gompertz-Makeham distribution ($KGM(a,b,r,u,\lambda)$) with corresponding densities:

$$G : f_G(x,a,b) = be^{ax}e^{-\frac{b}{a}(e^{ax}-1)} \ ; \ x \geqslant 0$$

$$GM : f_{GM}(x,a,b,\lambda) = [\lambda + be^{ax}]e^{-\lambda x - \frac{b}{a}(e^{ax}-1)} \ ; \ x \geqslant 0$$

$$\begin{aligned} KG : f_{KG}(x,a,b,r,u,\lambda) &= rube^{ax}e^{-\frac{b}{a}(e^{ax}-1)}[1-e^{-\frac{b}{a}(e^{ax}-1)}]^{r-1} \\ &\quad [1-[1-e^{-\frac{b}{a}(e^{ax}-1)}]^r]^{u-1} \ ; \ x \geqslant 0 \end{aligned}$$

$$\begin{aligned} KGM : f_{KGM}(x,a,b,r,u,\lambda) &= ru[\lambda + be^{ax}]e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}[1-e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^{r-1} \\ &\quad [1-[1-e^{-\lambda x - \frac{b}{a}(e^{ax}-1)}]^r]^{u-1} \ ; \ x \geqslant 0 \end{aligned}$$

The following table gives the descriptive statistics of the data:

Min	Q_1	Median	Mean	Q_3	Max	Kurtosis	Skewness
0.02	0.6875	1.9650	1.77	2.9820	3	1.4537	-0.2699

To summarize, it is important to use the visual representations as well as formal statistical tests to decide which model is the best. The K-S test and AIC value only measure which of the presented models is the best; they do not say if any of the models are even a good fit to the data. Visual representations like p-p plots, help make sure that the specified models adequately fit the data. Once it is verified that a models fits the data by looking at the p-p plot, the K-S test and AIC value will be used to decide which model is the best for the data.

The p-p plot for the four distributions - Gompertz, Gompertz-Makeham, Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham is shown in Figure 11 and 12 The p-p plot for Kumaraswamy Gompertz distibution shows that it is unfit for the given data set. Also, it may be noted that the p-p plot of Kumaraswamy Gompertz-Makeham distribution indicates a better fit for our data.

The Maximum Likelihood Estimates of the unknown parameters of each distribution with respect to the given data are obtained using Non-Linear Minimization. With these obtained estimates, the models are compared using the density plot. The Maximum Likelihood Estimates of the distribution parameters for $G(a,b)$, $GM(a,b,\lambda)$, $KG(a,b,r,u)$ and $KGM(a,b,r,u,\lambda)$ and the statistics $l(\hat{\theta})$ is given in the following table:

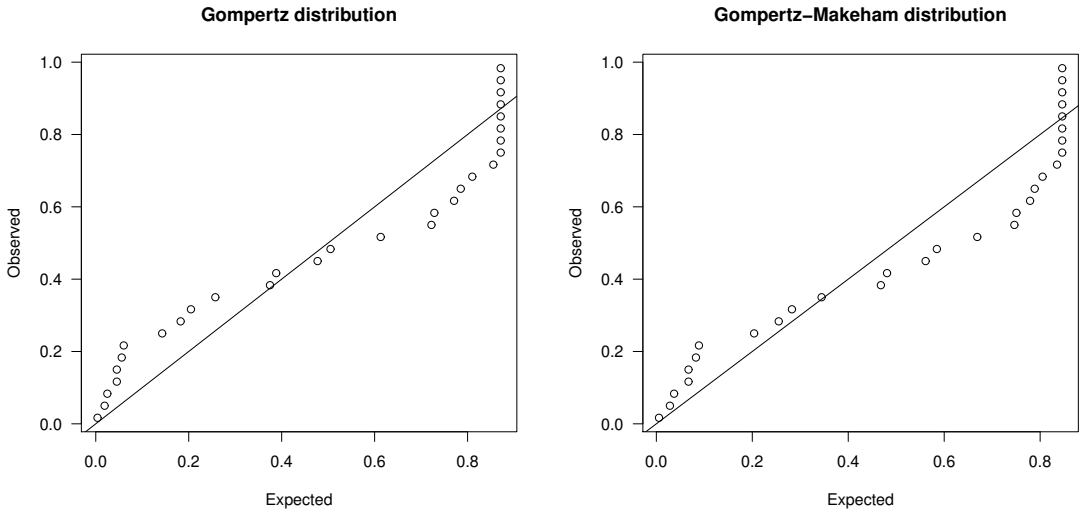


Figure 11: p-p plot of Gompertz and Gompertz-Makeham distribution

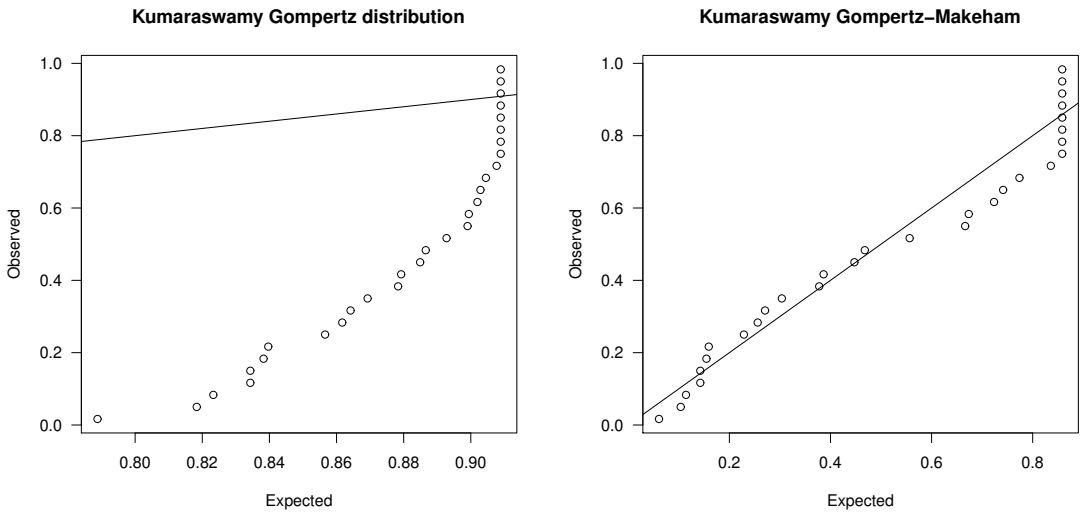


Figure 12: p-p plot of Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham distribution

Model	Maximum Likelihood Estimates				
$G(a,b)$	0.7402285	0.1847689	-	-	-
$GM(a,b,\lambda)$	-1.934506×10^{-3}	-1.203726×10^{2}	1.206476×10^{2}	-	-
$KG(a,b,r,u)$	1.099991	2.089319×10^{-6}	3.623456×10^{-2}	1.999525	-
$KGM(a,b,r,u,\lambda)$	1.4781023	0.2024657	0.2475147	0.1222549	1.0616082

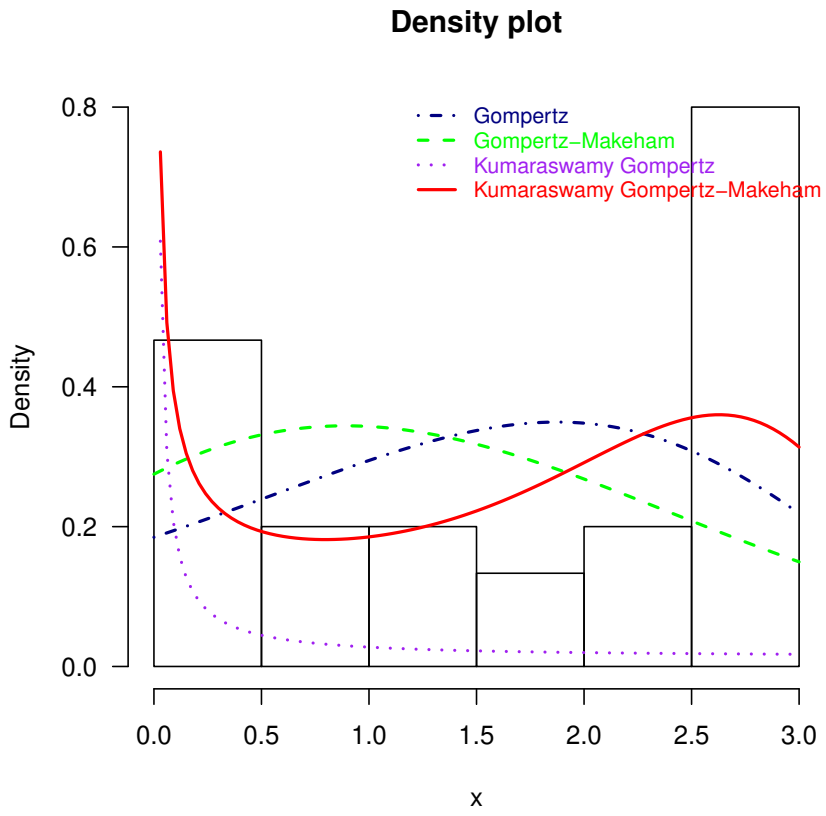


Figure 13: Density plots of the distributions along with histogram for the given data set.

The density plots of the distributions under study is shown in Figure 13. From the figure, we can see that the Kumaraswamy Gompertz-Makeham distribution fits better compared to the other distributions. Also, it may be noted that the Kumaraswamy Gompertz distribution do not fit the data well.

A method for comparing among two or more models is the Akaike Information Criterion (AIC) which is a measure of the relative goodness of fit for statistical model. The AIC is

calculated as negative two times the value of the log-likelihood function evaluated at the maximum likelihood estimates plus two times the number of parameters or,

$$AIC = -2\ l(\hat{\theta}) + 2p$$

where, $l(\hat{\theta})$ is the log-likelihood function evaluated at the maximum likelihood estimates and p is the number of parameters.

To choose among the best models, the AIC values and the values of $l(\hat{\theta})$ are calculated for each of the models, and then the models are ranked by this criterion. The model with the lowest AIC value and the largest value of $l(\hat{\theta})$ is considered the best. The values of Akaike Information Criterion (AIC) and log-likelihood function obtained is given in the following table:

Model	$l(\hat{\theta})$	AIC
$G(a, b)$	-41.34595	86.6919
$GM(a, b, \lambda)$	-43.98049	93.96097
$KG(a, b, r, u)$	-123.6019	255.2038
$KGM(a, b, r, u, \lambda)$	**-37.12753**	**84.25507**

From the above table, we can see that the Kumaraswamy Gompertz-Makeham distribution has the minimum AIC value and the maximum value of $l(\hat{\theta})$. Therefore, Kumaraswamy Gompertz-Makeham distribution provides the better fit among the compared distributions for the given data.

Most statistical methods assume an underlying model in the derivation of their results. However, when we presume that the data follow a specific model, we are making an assumption. If such a model does not hold, then the conclusions from such analysis may be invalid. Although density plotting and other graphical methods can guide the choice of the parametric distribution, one cannot of course be sure that the proper model has been selected. Hence model validation is still necessary to check whether we have achieved the goal of choosing the right model.

One method that can be used to check model appropriateness or the goodness of fit is the Kolmogorov-Smirnov (K-S) test. The K-S test is a nonparametric goodness-of-fit test and is used to determine whether an underlying probability distribution differs from a hypothesized distribution. KS test can be used to compare a sample with a reference probability model. The K-S test statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference model.

To study the goodness-of-fit of the models under study, we compute the Kolmogorov-Smirnov statistic between the empirical distribution function and the fitted distribution function when the parameters are obtained by method of maximum likelihood. The results of the K-S test are shown in the table below:

Model	D-value	p-value
$G(a,b)$	0.18892	0.2346
$GM(a,b,\lambda)$	0.21295	0.1316
$KG(a,b,r,u)$	0.78879	6.328×10^{-15}
$KGM(a,b,r,u,\lambda)$	**0.14158**	**0.5845**

From the above table, since the fitted KGM has the highest p-value , it is clear that the estimated Kumaraswamy Gompertz-Makeham model provides good fit to the given data set. Also, it is seen that the Kumaraswamy Gompertz distribution is unfit for the given data.

9. Conclusion

A data of 30 observations representing the times of failures and running times for samples of devices from an eld-tracking study of a larger system was considered. The aim was to find the distribution which better fits the data among the four distibutions, namely, Gompertz, Gompertz-Makeham, Kumaraswamy Gompertz and Kumaraswamy Gompertz-Makeham. Ultimately, in order to fit a probability distribution to a data set, it is important to first get an understanding of the structure of the data. This will help in the specification of a model. After choosing possible models, estimates of the parameters for each distribution are estimated by Maximum Likelihood Estimation. These estimates are then used to visualize the density functions to get an idea of the fit. The AIC values and the log likelihood function estimated at the MLEs are computed. After computing these values, the Kolmogorov-Smirnov test to choose which model fits the best to the data is performed. Also, p-p plots for the given data set are constructed.

Using the AIC values and to compare all four distributions, it is found that Kumaraswamy Gompertz-Makeham distribution has the lowest AIC value and maximum value of $l(\hat{\theta})$. The K-S test also shows similar result, the Kumaraswamy Gompertz-Makeham distribution being the one with the lowest value of test statistic and maximum p-value in comparison with the other three distributions. The density plot and p-p plot also favours the same distribution. Therefore, from all the results above, it can be concluded that the Kumaraswamy Gompertz-Makeham distribution provides significantly better fit to the given data set.

References

[1] Bailey, R. (1978). Limiting forms of makeham model and their use for survival analysis of transplant studies,*In Biometrics* Vol. 34, pp. 725726.

[2] Chen, Z. (1997). Parameter estimation of the gompertz population*Biomet rical Journal* 39(1), 117124.

[3] Chukwu, A., Ogunde, A. (2016). On kumaraswamy gompertz makeham distribution*American Journal of Mathematics and Statistics* 6(3),122127.

[4] Cordeiro, G. M., de Castro, M. (2011). A new family of generalized distributions*Journal of statistical computation and simulation* 81(7), 883898.

[5] da Silva, R. C., Sanchez, J. J., Lima, F., Cordeiro, G. (2015). The kumaraswamy gompertz distribution*Journal of Data Science* 13(2), 241259.

[6] El-Gohary, A., Alshamrani, A., Al-Otaibi, A. N. (2013). The generalized gompertz distribution*Applied Mathematical Modelling*37(1-2), 13 24.

[7] Eugene, N., Lee, C., Famoye, F. (2002). Beta-normal distribution and its applications*Communications in Statistics-Theory and methods* 31(4), 497512.

[8] Finch, C. E., Pike, M. C., Witten, M. (1990). Slow mortality rate ac

celerations during aging in some animals approximate that of humans*Science* 249(4971), 902905.

[9] Franses, P. H. (1994). Fitting a gompertz curve*Journal of the Operational Research Society* 45(1), 109113.

[10] Garg, M. (2008). On distribution of order statistics from kumaraswamy distribution*Kyungpook mathematical journal* 48(3).

[11] Garg, M. L., Rao, B. R., Redmond, C. K. (1970). Maximum-likelihood estimation of the parameters of the gompertz survival function*Journal of the Royal Statistical Society, Series C (Applied Statistics)* 19(2), 152159.

[12] Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies*in a letter to francis baily, esq. frs c. Philosophical transactions of the Royal Society of London*(115), 513583.

[13] Jaheen, Z. F. (2003). Bayesian prediction under a mixture of two-component gompertz lifetime model. 12(2), 413426.

[14] Jodra, P. (2009). A closed-form expression for the quantile function of the gompertz-makeham distribution*Mathematics and Computers in Simulation* 79(10), 30693075.

[15] Johnson, N. L., Kotz, S., Balakrishnan, N. (1995). Continuous univariate distributions, vol. 2. john wiley sons. Inc., 75.

[16] Jones, M. (2009). Kumaraswamys distribution: A beta-type distribution with some tractability advantages*Statistical Methodology* 6(1), 70 81.

[17] Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes*Journal of Hydrology* 46(1-2), 79 88.

[18] Makany, R. (1991). A theoretical basis for gompertzs curve*Biometrical Journal* 33(1), 121128.

[19] Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables

[20] Journal of the Institute of Actuaries 8(6), 301310.

[21] Minimol, S., Thomas, P. Y. (2014). On characterization of gompertz distri bution by properties of generalized record values*Journal of Statistical Theory and Applications* 13(1), 3845.

[22] Mitnik, P. A. (2013). New properties of the kumaraswamy distribution. *Communications in Statistics-Theory and Methods* 42(5), 741755.

[23] Nadarajah, S. (2008). On the distribution of kumaraswamy*Journal of Hydrology* 348, 568569.

[24] Norstrom, F. (1997). The gompertz-makeham distribution. Pollard, J. H., Valkovics, E. J. (1992). The gompertz distribution and its applications*Genus* 1528.

[25] Rao, B. R., Damaraju, C., Rao, P. F. R., Damaraju, C. (1992). New better than used and other concepts for a class of life distributions. *Biometrical Journal* 34(8), 919935.

[26] Read, C. B. (1983). Gompertz distribution*Encyclopedia of statistical sciences*

[27] Saracoglu, B., Kaya, M. F., Abd-Elfattah, A. (2009). Comparison of estimators for stress-strength reliability in the gompertz case*Hacettepe Journal of Mathematics and Statistics* 38(3), 339349.

[28] Sebastian, R. (2017). Generalization of gompertz distribution and its appli cations in reliability and time series*The Kerala Statistical Association* 28, 6879.

[29] Wang, B. X., Wang, X. K., Yu, K. (2017). Inference on the ku maraswamy distribution*Communications in Statistics-Theory and*

Methods 46(5), 20792090.

[30] Wu, J. W., Lee, W. C. (1999). Characterization of the mixtures of gom pertz distributions by conditional expectation of order statistics*Bio metrical Journal: Journal of Mathematical Methods in Biosciences* 41(3), 371381.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 155-164.

Prevalence and Risk Factors of Non-Alcoholic Steatohepatitis and Cardiovascular disease in Type 2 Diabetes Mellitus Childbearing Women

Ramya M. S.[1] **and Prasanth C. B.**[2]
[1]Department of Statistics, Annamalai University, India
Email:msramyaotp@gmail.com
[2]Department of Statistics, Sree Kerala Varma College, Thrissur, Kerala, India
Email:cbpwarrier@gmail.com

Abstract

Prevalence of non-alcoholic steatohepatitis is increasing in parallel with the rising prevalence of type 2 diabetes and obesity around the world. Current evidence strongly suggests that women who have had type 2 Diabetes Mellitus are at greater risk of Cardio Vascular Disease later in life. A definite model for predicting absolute risk of Cardio Vascular Disease later in women who had gestational diabetes is not yet available.

Non-alcoholic steatohepatitis (NASH) is commonly associated with type 2 diabetes mellitus (DM). Prevalence of NASH in type 2 DM has not been well studied and there is an epidemic rise in type 2 DM in developed and developing countries. Its association with chronic liver disease in the form of NASH makes it an important health problem. The world wide prevalence of non-alcoholic fatty liver disease(FLD) is estimated to have more in adults. Non-alcoholic fatty liver disease (NAFLD)is often associated with insulin resistance and is strongly associated with type 2 diabetes mellitus and obesity. NAFLD patients are at risk of progressing to NASH and ultimately cirrhosis; they are also at higher risk of cardiovascular diseases (CVD), including coronary heart disease and stroke. In this paper we are trying to analyse the risk of new CVD event in child bearing women with diabetic and non diabetic. Baseline characteristics were age, obesity, diabetic level, hypertension, cholesterol level etc. and trying to derive a statistical model for estimating the association of many such components with blood sugar. Finally the significance of the model statistically tested. wepropose to expand our study to prove whether NASH is an independent risk factor for CVD later.

Key words: Diabetes Mellitus, Non Alcoholic Fatty Liver Disease, Non Alcoholic Steato hepatitis, cardiovascular diseases, Estimate.

1. Introduction

Non-alcoholic fatty liver disease (NAFLD) is a condition in which excess fat is stored in your liver. This build-up of fat is not caused by heavy alcohol use. When heavy alcohol use causes fat to build up in the liver, this condition is called alcoholic liver disease. Two types of NAFLD are simple fatty liver and non-alcoholic steatohepatitis (NASH). Simple fatty liver and NASH are two separate conditions. People typically develop one type of NAFLD or the other, although sometimes people with one form are later diagnosed with the other form of NAFLD. Simple fatty liver, also called non-alcoholic fatty liver (NAFL), is a form of NAFLD in which you have fat in your liver but little or no inflammation or liver cell damage. Simple fatty liver typically does not progress to cause liver damage or complications. NASH is a form of NAFLD in which you have hepatitis inflammation of the liverand liver cell damage, in addition to fat in your liver. Inflammation and liver cell damage can cause fibrosis, or scarring, of the liver. NASH may lead to cirrhosis or liver cancer. Experts are not sure why some people with NAFLD have NASH while others have simple fatty liver.

Baseline information was obtained via a questionnaire on general information, physical examination (height, weight, and blood pressure), laboratory tests (triglycerides, total cholesterol, blood glucose, aspartate aminotransferase (AST) or serum glutamic-oxaloacetic transaminase (SGOT) and alanine aminotransferase (ALT) or Serum glutamic pyruvic transaminase (SGPT), Body mass index (BMI) and waist circumference (WC). Prevalence of NAFLD and NASH in our study of type 2DM patients is high and increases with multiple components of metabolic syndrome. (Metabolic syndrome is a cluster of conditions that occur together, increasing your risk of heart disease, stroke and type 2 diabetes. These conditions include increased blood pressure, high blood sugar, excess body fat around the waist, and abnormal cholesterol or triglyceride levels). BMI is obtained by dividing body weight in kilograms by height in meters squared. In developed countries, subjects with a $BMI \geq 25kg/m^2$ are defined as overweight and those with a $BMI \geq 30kg/m^2$ are defined as obese, and there are good associations and positive predictive effects between BMI and obesity-related chronic diseases. Both obesity and NAFLD are closely related to T2DM.

Heredity is not a factor for NASH. As DM attained by heredity, there is no chance for NAFLD. It is a life style disease. TG LEVEL INCREASES means there is a high risk of CVD.

As SGPT and SGOT level are normal levels in almost data, even though there exists chance of fatty liver.

NAFLD is more common in people who have certain conditions, including obesity and

conditions that may be related to obesity, such as type 2 diabetes. Researchers have found NAFLD in 40 to 80 percent of people who have type 2 diabetes and in 30 to 90 percent of people who are obese. In research that tested for NAFLD in people who were severely obese and undergoing bariatric surgery, more than 90 percent of the people studied had NAFLD.

The majority of people with NAFLD have simple fatty liver, and people with simple fatty liver typically dont develop complications. NASH can lead to complications, such as cirrhosis and liver cancer. People with NASH have an increased chance of dying from liver-related causes. If NASH leads to cirrhosis, and cirrhosis leads to liver failure, you may need a liver transplant to survive.

Cardiovascular disease (CVD). It is a general term for a disease of the heart or blood vessels. Blood flow to the heart, brain or body is reduced because of a blood clot (thrombosis) or a build-up of fatty deposits inside an artery, leading to hardening and narrowing of the artery (atherosclerosis).

The three main types of CVD are: coronary heart disease, stroke & peripheral arterial disease. Coronary heart disease this occurs when your heart blood supply is blocked or interrupted by a build-up of fatty substances (called atheroma) in the coronary arteries. The coronary arteries are two major blood vessels that supply the heart with blood. Stroke-A stroke is a serious medical condition that occurs when the blood supply to the brain is disturbed. Peripheral arterial disease is also known as peripheral vascular disease, occurs when there is a blockage in the arteries to your limbs (usually your legs).

2. Review Of Literature

NasrinAmiri et.al., (2017) conducted the study, Type 2 diabetes mellitus and non-alcoholic fatty liver disease: a systematic review and meta-analysis. The findings indicated that the overall prevalence of NAFLD among type 2 diabetes mellitus patients is significantly higher. It can be concluded that type 2 diabetes mellitus patients should be managed to prevent NAFLD.

Alessandro Mantovaniet.al, (2018) studied Non-alcoholic Fatty Liver Disease and Risk of Incident Type 2 Diabetes: A Meta-analysis and found that NAFLD is significantly associated with a twofold increased risk of incident diabetes. However, the observational design of the eligible studies does not allow for proving causality.

Hagstromet.al, (2019) studied Cardiovascular risk factors in non-alcoholic fatty liver disease. It concluded that Patients with NAFLD are at an increased risk for CVD compared to matched controls, but histological parameters do not seem to independently predict this risk.

Carrie et.al.,(2018) conducted a study on, The Association Between Non-alcoholic Fatty Liver Disease and Cardiovascular Disease Out comes found that the causal relationship of

CVD and NAFLD remains under investigation, but the strong bidirectional association between CVD and NAFLD warrants clinical intervention in patients with NAFLD to modify metabolic risk factors, including T2DM, dyslipidaemia, hypertension, and obesity.

Mathews et.al., (2018) conducted a study Nonalcoholic steatohepatitis, obesity, and cardiac dysfunction. It summarises Obesity is a major factor in the development of non-alcoholic fatty liver disease (NAFLD) and its progression to steatohepatitis. Patients with NAFLD have a significant increase in cardiovascular disease risk. For biopsy- proven NASH, vitamin E and pioglitazone are the recommended medical treatments in addition to lifestyle modification.

Rashmee and and Gagan (2017) studied Non-alcoholic fatty liver disease and cardiovascular risk and concluded that Non-alcoholic fatty liver disease (NAFLD) is often associated with insulin resistance and is strongly associated with type 2 diabetes mellitus and obesity. NAFLD is now recognized as a risk factor for poor cardiovascular outcomes including mortality and morbidity from major vascular events. As a whole, NAFLD patients may benefit from more careful surveillance and early treatment interventions.

Dharmalingam M, Yamasandhi P G (2018) conducted a study on alcoholic fatty liver disease and Type 2 diabetes mellitus and concluded that T2DM and NAFLD have a common association. The increasing prevalence makes it a public health problem.

Bhatt KN et.al., (2017) conducted study Prevalence of nonalcoholic fatty liver disease in type 2 diabetes mellitus and its relation with insulin resistance in South Gujarat Region. Study revealed a high incidence of NAFLD in Type 2 diabetes patients stressing the need for early screening.

Claudio Tana et.al, (2019) conducted study on Cardiovascular Risk in Non-Alcoholic Fatty Liver Disease: Mechanisms and Therapeutic Implications conclude that all cardiometabolic risk factors should be carefully and routinely screened among patients with NAFLD, and that disease management should be focused on both specific lifestyle modifications and aggressive risk factors modification, which would not only reduce the risk of liver disease progression, but may also provide benefit by reducing the risk of developing cardiac complications.

Jonathan M. Hazlehurst, et.al (2016) studied Non-alcoholic fatty liver disease and diabetes and concluded that Diabetes and NAFLD are reciprocal risk factors and when they are occurtogether, an increasing body of data demonstrates that diabetes is more difficult to manage and that NAFLD is more likely to progress.

Giovanni Targher, et.al (2005) conducted study on Nonalcoholic Fatty Liver Disease and Risk of Future Cardiovascular Events Among Type 2 Diabetic Patients and suggest that the metabolicsyndrome predicts incident cardiovascular disease (CVD), so it is possible to hypothesize that NAFLD patients might portend a greater CVD risk and that NAFLD itself might confer a CVD risk above that associated with individual metabolic syndrome risk factors.

Dyson JK, et.al (2014) studied Non-alcoholic fatty liver disease: a practical approach

to treatment concluded that Lifestyle interventions aimed at weight loss and increased activity are essential for all patients with NAFLD and if sustained are effective in the treatment of NAFLD.

Carrie R. Wong, et.al (2018) conducted study on The Association Between Nonalcoholic Fatty Liver Disease and Cardiovascular Disease Outcomes and concluded that The causal relationship of CVD and NAFLD remains under investigation, but the strong bidirectional association between CVD and NAFLD warrants clinical intervention in patients with NAFLD to modify metabolic risk factors, including T2DM, dyslipidemia, hypertension, and obesity.

3. Methodology & Data Analysis

The proposed research is to develop a statistical model which specifies the significance of diabetes mellitus in the risk factor analysis of NASH and CVD. From the literatures we can see that there are some relations between diabetes mellitus, NASH and CVD.

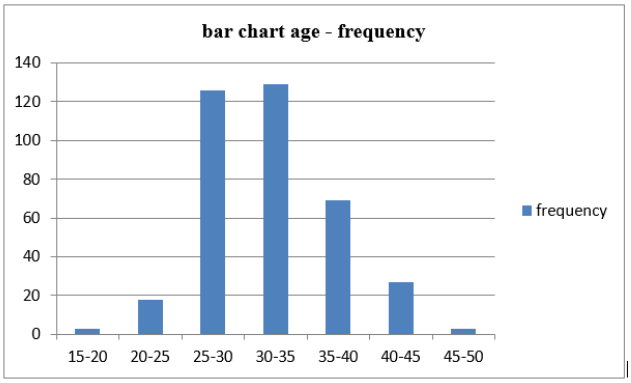
Here, in the initial stage of the research work a regression model is formulated to show the effect of age, weight, waist circumference etcon blood sugar. Also, discussed about the various risk factors which cause NAFLD and CVD.

375 child bearing diabetic women were screened from various hospitals in the 2 districts - Palakkad and Malappuram - at Kerala to make a real data analysis.135 (36%) of them with diabetic as a hereditary component. 219 of them were reported with fatty liver. That is 59% of carrying ladies were with FLD. Initially the analysis of DM with related components, but will be more effective for further future analysis. Age, height, weight, waist circumference, the presence of comorbidities (diabetes, arterial hypertension), SGOT, SGPT, Total cholesterol, triglycerides (TG) and body mass index (BMI) were evaluated. Using the real values of above factors, a frequency table is formulated and a bar diagram is drawn based on the table. The level of association is shown through a dendrogram. A regression model is obtained to predict the effect of various baseline characteristics with CVD and NAFLD on blood sugar. Based on the abdominal ultra sound scanning images, fatty liver is confirmed. Blood sugars, SGOT, SGPT, Total cholesterol, triglycerides (TG) were obtained from bio chemistry reports. Some of the patients having control level of SGPT, SGOT have symptoms of fatty liver shown in their Ultra Sound Scanning. Central obesity as measured by weight circumferences (WC) and SGPT levels were significantly higher in people with fatty liver.

Table 1: **Frequency Table Age of people**

Age	Frequency
15-20	3
20-25	18
25-30	126
30-35	129
35-40	69
40-45	27
45-50	3
total	375

Figure 1: **Bar chart age frequency**



From the bar chart we can analysed that the child bearing women have more number of type2 diabetic patients in the age 35-40.

Table 2 shows the correlation between the different base line characteristics of study. Correlation Coefficient greater than 0.5 shows high risk of having NAFLD and CVD. High positive correlation indicates that it is a risk factor for NAFLD. The correlation between SGPT, SGOT AND TG indicates as increasing level of SGPT,SGOT enzymes affect the functioning of liver and causes to liver damage. As TG level increases there is a high risk of CVD. The high correlation of hypertension also indicates that high risk of CVD.

4. Dendrogram

The dendrogram is a multilevel hierarchy where clusters at one level are joined together to form the clusters at the next levels. This makes it possible to decide the level at which

Table 2: **Correlation Coefficient between the components**

Risk Factor	Correlation Coefficient
Weight & Waist Circumference	0.69
Waist Circumference & TG	0.588
SGOT & TG	0.709
SGPT & BMI	0.597
Weight & BMI	0 .87
Waist Circumference & BMI	0.597
SGPT & TG	0.693
Systolic Pressure & Diastolic Pressure	0.876

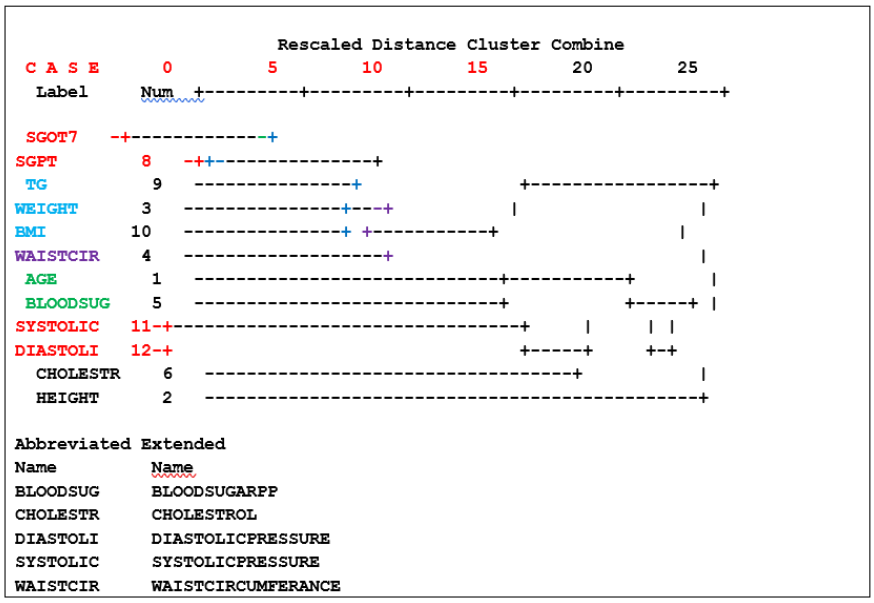
to cut the tree for generating suitable groups of a data objects. It is a type of tree diagram showing hierarchical clustering relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node.

Figure 2 indicates the dendrogram of the average linkage data variables. In the dendrogram it is found that the labels SGOT and SGPT with label numbers 7and 8 has first level of association with the risk of NAFD if their values is between 0-5. SGOT and SGPT affect the predicted variable in the same way. That means these two components are equally affected in the predicted variable NAFLD. The labels systolic and diastolic pressure with label number 11 and 12 also hasfirst level of association. Therefore, there exist a high correlation between these risk factors and CVD. The labels SGPT, SGOT, with label number 7and 8 have second level of association if their values is between 5-10.TG, WEIGHT, BMI with label number9,3,10 will also have second level of association. There is also a positive correlation. That is these variables affect directly in predicted variables NAFLD and CVD. Also, the labels WEIGHT, BMI AND WAISTCIR with label number 3,10and 4 have third level of association. That means these three variables also affect the predicted variables.

5. Conclusion

We discussed the risk of new CVD and NASH event in child bearing women with dia-

Figure 2: **Dendrogram using Average Linkage (Between Groups)**



betic and non-diabetic. Out of 375 child bearing diabetic women, 135 (36%) of them with diabetic as a hereditary component. 219 of them were reported with fatty liver. That is 59% of carrying ladies were with FLD. A regression analysis is conducted on the observations taken from child bearing women having diabetics to evaluate the effect of the various risk factors of NASH and CVD on blood sugar and a statistical model is derived. From the analysis the relation between predictor variables such as age, weight, waist circumference etc. with the dependent variable blood sugar is expressed as a linear regression model. The variance inflation factor (VIF) is calculated to estimate multicollinearity. From the evaluation it is found that the VIF is less than 10 which indicates there is no multicollinearity in the observed data. Hence the model is nicely fitted for such data. The proposed research objective is to develop a statistical model which specifies the significance of diabetes mellitus in the risk factor analysis of NASH and CVD and we are trying to analyze the significance of the same in the coming days.

References

[1] Alessandro Mantovani, Christopher D. Byrne, EnzoBonora and Giovanni Targher(2018). Nonalcoholic Fatty Liver Disease and Risk of Incident Type 2 Diabetes: A Meta- analysis.Diabetes Care; 41(2): 372-382.

[2] Bhatt KN, Pranav V, Dipika Y, Dharmesh N, Radhika N, Arvind S (2017). Prevalence of nonalcoholic fatty liver disease in type 2 diabetes mellitus and its relation with

insulin resistance in South Gujarat Region. J Mahatma Gandhi Inst Med Sci 2017; 22:8-11.

[3] Carrie R. Wong, and Joseph K. Lim, (2018). The Association Between Nonalcoholic Fatty Liver Disease and Cardiovascular Disease Outcomes. Clinical Liver Disease, (Hoboken)v.12(2); PMC6385911. doi: 10.1002/cld.721.

[4] Claudio Tana, Stefano Ballestri, Fabrizio Ricci, Angelo Di Vincenzo, AndreaTicinesi, SabinaGallina, Maria Adele Giamberardino, FrancescoCipollone, RichardSutto, Roberto Vettor, ArturFedorowski, and TizianaMeschi (2019). Cardiovascular Risk in Non-Alcoholic Fatty Liver Disease: Mechanisms and Therapeutic Implications.Int. J. Environ. Res. Public Health 2019, 16, 3104.

[5] Dharmalingam M, Yamasandhi P G (2018). Nonalcoholic fatty liver disease and Type 2 diabetes mellitus. Indian J EndocrMetab ;22:421-8.

[6] Dyson J K,Q M Anstee,S McPherson((2014). Non-alcoholic fatty liver disease: a practical approach to treatment. Frontline Gastroenterology 2014; 5:277286.

[7] Giovanni Targher, LorenzoBertolini, FelicePoli, Stefano Rodella, LucaScala, Roberto Tessari, LucianoZenari, and Giancarlo Falezza (2005). Nonalcoholic Fatty Liver Disease and Risk of Future Cardiovascular Events Among Type 2 Diabetic Patients. Diabetes 54:35413546, 2005.

[8] Hagstrm H, Nasr P, Ekstedt M, Hammar U, Stl P, Askling J, Hultcrantz R, Kechagias S(2019). Cardiovascular risk factors in non-alcoholic fatty liver disease.Liver Int;39(1):197-204. doi: 10.1111/liv.13973.

[9] Jonathan M.Hazlehurst, ConorWoods, Thomas ,Marjot, Jeremy F. Cobboldb, Jeremy W. Tomlinsona(2016). Non-alcoholic fatty liver disease and diabetes. Metabolism clinical andexperimental.65,1096-1108.

[10] Mathews SE, Kumar RB and ShuklaAP(2018). Nonalcoholicsteatohepatitis, obesity, and cardiac dysfunction. 25(5):315-320. doi: 10.1097/MED.0000000000000432.

[11] NasrinAmiri Dash Atan, Mehdi Koushki, MortezaMotedayen, Majid-Dousti,FatemehSayehmiri, RezaVafaee, MohsenNorouzinia, and Reza Gholami(2017). Type 2 diabetes mellitus and non-alcoholic fatty liver disease: a systematic review and meta-analysis.GastroenterolHepatol Bed Bench. Winter; 10(Suppl1): S1S7.

[12] Rashmee Patil and Gagan K Sood(2017). Non-alcoholic fatty liver disease and cardiovascular risk. World J Gastrointest Pathophysiol. 2017 May 15; 8(2): 5158.doi: 10.4291/wjgp.v8.i2.51. PMCID: PMC5437502. Shunquan Wu, FuquanWu,Yingying Ding, Jun Hou, JingfengBi & Zheng Zhang(2016). Association of non-alcoholic fatty

liver disease with major adverse cardiovascular events: A systematic review and meta-analysis. Scientific Reports, volume 6, Article number: 33386.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 165-182.

Robust Estimation For Multivariate Zero-Inflated Poisson Regression

Jeena Joseph and Pinky Peter
Department of Statistics, St.Thomas College, Thrissur, Kerala -1, India
Email:sony.jeena@gmail.com

Abstract

In this paper, we introduce zero-inflated Poisson (ZIP) and multivariate zero-inflated Poisson (MZIP) distributions and the corresponding regression models. The EM method is used to obtain estimates of the regression coefficients. A detailed review on Robust estimation for ZIP regression are also given. Along with it, we propose a robust estimation for MZIP regression and conduct a simulation study. In this paper the method adopted for robust estimation is to accommodate, rather than eliminate, outliers and use a robust estimation methodology that minimizes their effect on estimation of the model that is followed by the vast majority of data.

Key words:Zero-inflated Poisson, zero-inflated Poisson, Robust estimation.

1. Introduction

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or predictors). A linear regression analysis is used for investigating and modelling the linear relationship between dependent variable and independent variables by fitting a linear equation to observed data. Here the response variable must be continuous random variable which is assumed to have normal distribution with constant variance and The explanatory variables can be either quantitative or qualitative variables. Typically, maximum likelihood estimation is used for fitting such models. But when outliers are present in the data then the maximum likelihood estimator can become

unstable. In this situation we adopt robust regression. It is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. They are not overly effected by violations of assumptions by the underlying data-generating process.

A simple linear regression model is a model with a single regressor x that has a relationship with a response y that is a straight line. This simple linear regression model is obtained as follows,

$$y = \beta_0 + \beta_1 x + \epsilon$$

where the intercept β_0 and the slope β_1 are unknown constants and $\epsilon$is a random error component. The errors are assumed to have mean zero, unknown variance σ^2 and are uncorrelated.

A multiple linear regression model is a model with two or more regressors that has a relationship with a response y which describes a hyperplane. The model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

is called a multiple linear regression model with k regressors. The parameters β_j, j = 1, 2, ..., k, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables x_j.

The k^th order polynomial model in one variable is given by

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \epsilon$$

If $x_j = x^j$; $j = 1, 2, ...k$ then the model is multiple linear regression model in k explanatory variables $x_1, x_2, ..., x_k$. The polynomial model can be used in those situations where the relationship between study and explanatory variables are curvilinear.

A logistic regression model is one form of a generalized linear model where the response variable takes on the values either 0 or 1. The general form of the logistic regression model is given by

$$y_i = E(y_i) + \epsilon_i$$

where

$$E(y_i) = \pi_i = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)}$$

A Poisson regression model is an another form of a generalized linear model where the response variable follow Poisson distribution. A random variable Y is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function (pmf) is given by;

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, if y = 0, 1, 2, ..., \lambda > 0, \quad (1)$$

The Poisson distribution is often used to model rare events and having random variables with non-negative integer values. The data can be modeled through simple Poisson regression as given below,

$$log(\mu_i) = \theta' X$$

The Zero Inflated Poisson (ZIP) regression model, well described by Lambert (1992), is a simple mixture model for count data with excess zeros. The model is a combination of a Poisson distribution and a degenerated distribution at zeros. If the observed counts are not univariate, we consider multivariate zero-inflated Poisson regression model. Section 2 consists of a review on zero-inflated Poisson distribution and its properties, zero-inflated Poisson regression modelling procedures and its robust estimation. Section 3 consists of a study on multivariate Zero-Inflated Poisson (MZIP) distribution and MZIP regression models including estimation of parameters. In section 4, we present the robust estimation procedure for MZIP model. Section 5 does simulation study of MZIP model in a contaminated data in R. Section 6 gives a brief conclusion.

2. Zero Inflated Poisson Regression Model

Zero-inflated distributions are used to model count data that have many zero counts, i.e. the observations of counts in which the number of zero events is higher than predicted by a Poisson model. Thus the ZIP distribution employs two components. The first component is governed by a degenerate distribution which is localized at 0. The second component is governed by a traditional Poisson distribution that generates non-negative counts including zeros.

Specifically if Y are independent random variables having a ZIP distribution, the zeros are assumed to arise in two ways corresponding to distinct underlying states. The first state occurs with probability p and produces only zeros, while the other state occurs with probability 1- p and leads to standard Poisson counts with mean λ and a chance for further zeros. In general the zeros from the first state are called structural zeros and those from the Poisson distribution are called sampling zeros. For example, in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease, or simply because no disease spores have landed on it. This is the distinction between structural zeros, which are inevitable, and sampling zeros, which occur by chance.

This two-state process gives a simple two-component mixture distribution.

$$f(y) = pf_1(y) + (1-p)f_2(y)$$

where f_1 is a degenerated distribution of zeros and f_2 is Poisson distribution with parameter

λ. The pmf can be written as

$$\begin{aligned} f(y;\lambda,p) &= p+(1-p)e^{-\lambda}, if y=0, \lambda>0 \\ &= (1-p)\frac{e^{-\lambda}\lambda^{y}}{y!}, if y=1,2,...,0\leq p\leq 1 \end{aligned} \tag{2}$$

We denote this symbolically as $Y \sim ZIP(\lambda, p)$. It is clear that this reduces to standard Poisson when p = 0. For positive values of p we have zero inflation.

2.1 Properties

- $E(Y) = \lambda$
- $V(Y) = \frac{p}{1-p}\lambda^2 + \lambda$
- $M_Y(t) = p + (1-p)e^{\lambda(e^t-1)}$
- $\Phi_Y(t) = p + (1-p)e^{\lambda(e^{it}-1)}$

2.2 Estimation of Parameter of ZIP distribution

Maximum likelihood estimation is used for estimating the parameters of ZIP distribution.
In the case of homogeneous sample the log-likelihood function can be written as

$$l(\lambda,p) = n_0 log[p+(1-p)e^{-\lambda}] + \sum_{j=1}^{J} n_j log[(1-p)\frac{e^{-\lambda}\lambda^j}{j!}]$$

where J is the largest observed count value, n_j is the frequency of each possible count value, $j = 0, 1, ..., J$; n_0 is the number of observed zeros and $\sum_{j=1}^{J} n_j = n$ the total number of observations or the sample size.
The maximum likelihood estimates (MLE) $\hat{\lambda}$and $\hat{p}$ are obtained as the values of λ and p which maximize $l(\lambda, p)$. The above likelihood function can be written as

$$l(\lambda,p) = n_0 log[p+(1-p)e^{-\lambda}] + log(1-p)\sum_{j=1}^{J} n_j - \lambda\sum_{j=1}^{J} n_j + log\lambda\sum_{j=1}^{J} jn_j - j!\sum_{j=1}^{J} n_j$$

Then we have

$$\frac{\partial l(\lambda,p)}{\partial\lambda} = \frac{-n_0(1-p)e^{-\lambda}}{p+(1-p)e^{-\lambda}} - \sum_{j=1}^{J} n_j + \frac{\sum_{j=1}^{J} jn_j}{\lambda} \tag{3}$$

$$\frac{\partial l(\lambda,p)}{\partial p} = \frac{n_0(1-e^{-\lambda})}{p+(1-p)e^{-\lambda}} - \frac{\sum_{j=1}^{J} n_j}{1-p} \tag{4}$$

Equating (3)and (4)to zero, we get

$$\frac{n_0(1-p)e^{-\lambda}}{p+(1-p)e^{-\lambda}}+\sum_{j=1}^{J}n_j=\frac{\sum_{j=1}^{J}jn_j}{\lambda} \tag{5}$$

$$\frac{n_0(1-e^{-\lambda})}{p+(1-p)e^{-\lambda}}=\frac{\sum_{j=1}^{J}n_j}{1-p} \tag{6}$$

Solving (5) and (6), we have

$$\frac{e^{-\lambda}\sum_{j=1}^{J}n_j}{1-e^{-\lambda}}+\sum_{j=1}^{J}n_j=\frac{\sum_{j=1}^{J}jn_j}{\lambda} \tag{7}$$

That is,

$$\frac{\lambda}{1-e^{-\lambda}}=\frac{\sum_{j=1}^{J}jn_j}{\sum_{j=1}^{J}n_j} \tag{8}$$

This means that $\hat{\lambda}$ satisfies the equation

$$\hat{\lambda}=\frac{\sum_{j=1}^{J}j\times n_j(1-e^{-\lambda})}{\sum_{j=1}^{J}n_j} \tag{9}$$

Note that this does not depend on p or n_0. From (6) we have,

$$n_0(1-e^{-\lambda})(1-p)=[p+(1-p)e^{-\lambda}]\sum_{j=1}^{J}n_j \tag{10}$$

Solving this, we get

$$\hat{p}=\frac{n_0-(n_0+\sum_{j=1}^{J}n_j)e^{-\lambda}}{\sum_{j=1}^{J}n_j+n_0-(\sum_{j=1}^{J}n_j+n_0)e^{-\lambda}} \tag{11}$$

ie,

$$\hat{p}=\frac{n_0-ne^{-\lambda}}{n(1-e^{-\lambda})} \tag{12}$$

where $n=n_0+\sum_{j=1}^{J}n_j$

From above equations for $\hat{\lambda}$ and $\hat{p}$ it is clear that these values again contains unknown parameters. Therefore we use the following steps to obtain $\hat{\lambda}$ and $\hat{p}$.

1. Fit a standard Poisson model to obtain an initial value $\lambda^(0)$ for λ.
2. Use an iterative scheme for $\hat{\lambda}$

$$\hat{\lambda}(m+1)=\frac{[1-e^{\hat{\lambda}(m+1)}]\sum_{j=1}^{J}n_j\times j}{\sum_{j=1}^{J}n_j} \tag{13}$$

The iterations are continued until $|\hat{\lambda}^{(m+1)}-\hat{\lambda}^{(m)}|\leq\epsilon$, a small quantity, where $\hat{\lambda}^{(m)}$ and $\hat{\lambda}^{(m+1)}$ are the estimates of λ at m^{th} and $(m+1)^{th}$ iteration respectively.

3. Obtain $\hat{p}$ by substituting $\hat{\lambda}$ given by the final iterations of step (2) of (11).

The above algorithm is a particular case of Expectation Maximization (EM) algorithm introduced by Dempster (1977), which is a general methodology for maximum likelihood estimation for finite mixture models and latent variable models.

2.3 Zero-Inflated Poisson Regression Model

Zero-inflated Poisson (ZIP) regression model, well described by Lambert (1992), is a simple mixture model for count data with excess zeros. The model is a combination of a Poisson distribution and a degenerated distribution at zeros. This model is useful when the data is over dispersed with many zero counts. Here the zero counts are divided into two parts, certain zeros and the random zeros. The certain zero is modelling using logistic regression analysis and the random zeros and positive counts are modelling using Poisson regression analysis. In ZIP regression, the response vector is $y = (y_1, y_2, ..., y_n)^T$, where y_i is the observed value of the random variable Y_i. The Yi s are assumed independent, where

$$Y_i = \begin{cases} 0, & \text{with probability } p_i \\ \text{Poisson } \lambda_i, & \text{with probability } (1 - p_i), \end{cases} \tag{14}$$

Moreover, the parameters $p = (p_1, ..., p_n)^T$ and $\lambda = (\lambda_1, ..., \lambda_n)^T$ are modelled through canonical link generalized linear models (GLM) as $logit(p) = G_\gamma$ and $log(\lambda) = B\beta$, where γ and β are regression parameters, and G and B are corresponding design matrices that pertain to the probability of zero state and Poisson mean, respectively. The log-likelihood function for this model can be written as

$$l(\gamma, \beta; y) = \sum_{y_i=0} log[G_i^T\gamma + exp(e^{-B_i^T\beta})] + \sum_{y_i>0} (y_i B_i^T\beta - exp(e^{B_i^T\beta}) - \sum_{y_i>0} log(y_i!) - \sum_{i=1}^{n} (1 + e^{(G_i^T\gamma)}) \tag{15}$$

2.4 Robust Estimation for Zero-Inflated Poisson Regresion

The zero-inflated Poisson regression model is a special case of finite mixture models that is useful for count data containing many zeros. Typically, maximum likelihood (ML) estimation is used for fitting such models. However, it is well known that ML estimator is highly sensitive to the presence of outliers and can become unstable when the mixture components are poorly separated. In this case we use Robust estimation, an estimation technique which is insensitive to small departures from the idealized assumptions which have been used to optimize the algorithm. Here we establish an alternative robust estimation approach, robust expectation-solution (RES) estimation. And the proposed RES algorithm is a modification of the EM algorithm with the property of robustness.
In ZIP models, as in other mixture models, the EM algorithm is a particularly convenient

approach for computing MLE (e.g. Lambert, 1992). This algorithm is set up by introducing missing data into the problem. In particular, suppose we knew which zeroes came from the degenerate distribution (the zero state); and which came from the non-degenerate distribution (the non-zero state). That is, suppose we could observe $z_i = 1$when y_i is from zero state, and $z_i = 0$when y_i is from non-zero state. Then the log-likelihood for the complete data (y, z) would be

$$l^c(\gamma,\beta;y,z) = \sum [Z_i G_i^T\gamma - log(1+e^{G_i^T\gamma}] + \sum (1-z_i)[y_i\beta_i^T\beta - e^{B_i^T\beta} - log(y_i!] = l_\gamma^c(\gamma;y,z) + l_\beta^c(\beta;y,z) \tag{16}$$

where$z = (z_1, z_2, ..., z_n)^T$ This log-likelihood is easy to maximize, because $l_\gamma^c(\gamma;y,z)$ and $l_\beta^c(\beta;y,z)$ can be maximized separately with respect to γ and β respectively, via standard calculations. With the EM algorithm, the log-likelihood of model is maximized iteratively by alternating between estimating z_i by its conditional expectation under the current estimates of (γ,β) (E step) and then,with the z_i fixed at their expected values from the E-step, maximizing $l^c(\gamma,\beta;y,z)$ (M step), until the estimated (γ,β) converges and iteration stops.

In more detail, the EM algorithm begins with starting values $\theta^{(0)} = (\gamma^{(0)T}, \beta^{(0)T})^T$ and proceeds iteratively via the following three steps until convergence.

E step: Estimate z_i by its conditional mean $z_i^{(r)}$ under the current estimates $\gamma^{(r)}$ and $\beta^{(r)}$

$$\begin{aligned} z_i^{(r)} &= P(zero\ \ state|y_i,\gamma^{(r)},\beta^{(r)}) \\ &= \frac{P(y_i|zero\ \ state)P(zero\ \ state)}{P(y_i|zero\ \ state)P(zero\ \ state)+P(y_i|Poisson\ \ state)P(Poisson\ \ state)} \end{aligned}$$

M step for γ:Find $\gamma^{(r+1)}$ by maximizing $l_\gamma^c(\gamma;y,z)$. This can be accomplished by fitting a binomial logistic regression of $z^{(r)}$on design matrix G with binomial denominator equal to one. It is equivalent to solving the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}[z_i^{(r)} - logit^{-1}(G_i^T\gamma)]G_i = 0 \tag{17}$$

M step for β:Find $\beta^{(r+1)}$ by maximizing $l_\beta^c(\beta;y,z)$. It is equivalent to solving the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}(1-z_i^{(r)})[y_i - e^{B_i^T\beta}]B_i = 0 \tag{18}$$

In the RES approach, we propose to replace the estimating functions from the M step of the EM algorithm with robustified estimating functions. Thus we change from EM algorithm to RES algorithm. Essentially we propose to down weight observations that fall in the extreme upper and lower tail of the Poisson distribution in the estimating function. Specifically, we suggest that $\gamma^{(r+1)}$ and $\beta^{(r+1)}$ be found by solving the following equations:

$$\frac{1}{n}\sum_{i=1}^{n}\omega(G_i)[z_i^{(r)} - logit^{-1}(G_i^T\gamma)]G_i = 0 \tag{19}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1-z_i^{(r)})\omega(B_i)[\psi_c(y_i)-o_i(\beta,c)]B_i=0 \tag{20}$$

where

$$\psi_c(y)=\begin{cases} j_1, & y<j_1 \\ y, & y\epsilon[j_1,j_2] \\ j_2, & y>j_2 \end{cases} \tag{21}$$

where j_1 and j_2 being the c and (1 - c) quantiles of the non-degenerate Poisson component, respectively; and

$$\begin{aligned} o_i(\beta,c) &= E[\psi_c(Y_i)|Y_i\sim Poisson(\lambda_i=e^{B_i\beta})] \\ &= j_1P(Y_i<j_1)+\lambda_iP(j_1-1\leq Y_i\leq j_2)+j_2P(Y_i>j_2) \end{aligned}$$

where probabilities are computed based on the Poisson component density. Here,$\omega(.)$ is a function to down weight large leverage points. A simple choice for $\omega(G_i)$ that we use throughout this paper is $\sqrt{1-h_i}$, where h_i is the i^{th} diagonal element of $H = G(G^TG)^{-1}G^T$, with a similar definition for $\omega(B_i)$. The choice of upper and lower quantile, in $\psi_c(.)$ controls the trade-off between robustness and efficiency. Here we take c = 0.01, a value that has been chosen to be small to guard against the occurrence of a small number of truly anomalous (or even erroneous) observations rather than to eliminate data that come from a real, non-trivial component of the mixture (i.e., a third latent class underlying the data that arises with non-negligible probability).

2.5 Influence function(IF)

The IF is a useful and popular tool for quantifying the degree of robustness of a statistic by measuring the potential effect of an additional observation. The classical ML estimating equations for $\theta=(\gamma^T,\beta^T)^T$ can be written as joint equations

$$\frac{1}{n}\sum_{i=1}^{n}[E_\theta(z_i|y_i)-logit^{-1}(G_i^T\gamma)]G_i=0 \tag{22}$$

$$\frac{1}{n}\sum_{i=1}^{n}[1-E_\theta(z_i|y_i)][y_i-e^{B_i^T\beta}]B_i=0 \tag{23}$$

where the expectation is with respect to z_i given $Y_i = y_i$. The IF of $\hat{\beta}_{MLE}$, the MLE with respect to β for the ZI model, quantifies the influence of one additional observation y_j drawn from model (15). The function is given by

$$IF_{MLE}(y_j)=[1-E_\theta(z_j|y_j)][y_j-e^{B_j^T\beta}][-E_\theta(\frac{\partial}{\partial\beta^T}[1-E_\theta(z_j|y_j)][y_j-e^{B_j^T\beta}])B_j]^{-1}B_j \tag{24}$$

As can be seen in $IF_{MLE}(y_j)$ the influence of an outlier on the MLE is proportional to the score function and is, therefore, unbounded in general. The estimating functions underlying the RES method are:

$$\frac{1}{n}\sum_{i=1}^{n}\omega(G_i)[E_\theta(z_i|y_i) - logit^{-1}(G_i^T\gamma)]G_i = 0 \tag{25}$$

$$\frac{1}{n}\sum_{i=1}^{n}[1 - E_\theta(z_i|y_i)]\Psi(\beta, y_i) = 0 \tag{26}$$

where $\Psi(\beta, y_i) = \omega(B_i)[\psi_c(y_i) - o_i(\beta, c)]B_i$.
Then the IF of $\hat{\beta}_{RES}$ is

$$IF_{RES}(y_j) = [1 - E_\theta(z_j|y_j)][-E_\theta(\frac{\partial}{\partial\beta^T}[1 - E_\theta(z_j|y_j)][\Psi(\beta, y_i)])]^{-1}\Psi(\beta, y_i) \tag{27}$$

The IF of RES estimator is bounded because the estimating function Ψ is bounded. Therefore, $\hat{\beta}_{RES}$ is as called as B-robust. Similarly, $\hat{\gamma}_{RES}$ is B-robust because of the boundedness of the estimating function (25).

2.6 Simulation Study

Here we compared ML and RES estimation methods for ZIP data with constant p with and without contamination in y in each data set were randomly selected to be replaced by $y + 15$. True values of p and β were specified as listed in Table 1 and were chosen to make the non-degenerate components mean large (μ ranging between 2.78 and 20 over the values of the covariate vector $x_i = (x_{1i}, ..., x_{4i})^T$ and to give a moderate level of ZI (20 percent). And Bias, mean square error (MSE) and empirical size of a nominally 0.05 - level Wald test for equality with the true value were calculated for each model parameter. In addition, we provided the MSE for $\zeta = \frac{1}{n}\sum_{i=1}^{n}(1 - p_i)\mu_i$, the average marginal mean according to he model.Here the tuning quantile c was set to 0.01.
Generally speaking, these results favour the RES approach over the ML estimation. In the absence of contamination RES performs slightly worse than ML for n=100 and essentially the same for the larger sample size. It should be kept in mind that the degree of contamination here is fairly extreme. Both the proportion (5 per cent) and magnitude $(y + 15)$of outliers here are quite large. Under these extreme circumstances, the Wald tests under RES estimation perform reasonably well, and seem to retain some value as inferential tools. In contrast, the tests under ML estimation have been completely undermined.
To examine the effect of more moderate degrees of contamination, we ran simulations similar in design to these but with 5 per cent of the responses increased by 7 rather than 15. The results from those simulations are similar to those from the bottom half of Table 2, with smaller but still quite substantial improvements in bias and MSE achieved using the RES method.

Because these results are, as one might expect, intermediate to those in the top and bottom halves of Table 2, we do not report them in detail here for the sake of brevity.
Also from the simulation results it is observed that increasing sample size from n=100 to 200 has the expected effect of decreasing MSE for all parameters across these two methods.

3. Multivariate Zero-Inflated Poisson (MZIP) distribution

If the observed counts are not univariate, we need to consider the multivariate distributions. If an excess of zero events occurs, the multivariate Zero-inflated Poisson model should be considered instead. This MZIP model is a mixture of multidimensional degenerate distribution at point zero and a traditional multivariate Poisson (MPoi) distribution as follows. Let Y be a random variable following MZIP distribution,

$$Y = \begin{cases} (0, 0, ..., 0), & \text{with probability } \varphi \\ \text{MPoi}(y_1, y_2, ..., y_n), & \text{with probability } (1-\varphi), \end{cases} \tag{28}$$

where the observation follows a degenerate distribution at (0, 0, ..., 0) with probability φ and a MPoi distribution with probability $1-\varphi$ where $\varphi \epsilon [0, 1]$. It is a reasonable model, since the MZIP distribution is mainly used for situations in which most defect counts are 0.

3.1 Multivariate Zero-Inflated Poisson regression Models

In this model, there are m log-link functions and an extra link function logit(φ) for the Bernoulli probability φ,

$$ln\lambda_j' = \sum_{r-1}^{p_1} x_r \beta_{rj}, j = 1, 2, ..., m$$

$$logit(\varphi) = ln(\frac{\varphi}{1-\varphi}) = \sum_{r=1}^{p_2} g_r \tau_r$$

where $\lambda$and φ are not necessarily dependent on the same covariates. We have p_1 parameter $\beta' s$ for each λ and p_2 parameter $\tau' s$ for φ. In summary, in the MZIP model, there are two possible states, the perfect-state and the non-perfect state. If the observed counts are all zeros, it is possible the system is in either the perfect or the non-perfect state. However, if the observed counts are not all zeros, we know the system must be in the non-perfect state.

3.2 EM Algorithm for MZIP Regression Models

The estimations for the MZIP regression coefficient matrix β and vector τ are also obtained by the MLE method. For $n_1 + n_2$ observations following MZIP distribution, the

log-likelihood function would be

$$lnL(\tau,\beta,\lambda_0)=\sum_{i=1}^{n} ln[\frac{e^{\sum g\tau}}{1+e^{\sum g\tau}}+\frac{1}{1+e^{\sum g\tau}}e^{(m-1)\lambda_0-\sum_{j=0}^{m}e^{\sum_{r=1}^{p_1}x_{ir}\beta_{rj}}}]$$
$$+\sum_{i=1}^{n_2}[ln\frac{1}{1+e^{\sum g\tau}}+(m-1)\lambda_0-\sum_{j=0}^{n}\sum_{r=1}^{p_1}x_{ir}\beta_{rj}+ln\phi(\vec{y_i})]$$

where we assume n_1 is the number of observations $\vec{y}=(0,0,...,0)$ and n_2 is the number of observations $\vec{y}\neq(0,0,...,0)$. For the first term of function, we cannot get a simple first-order derivative. The computational difficulties make it impossible to solve the maximum of this likelihood function by the NR method. Hence, we utilize the EM algorithm to maximize the likelihood.

The EM algorithm is performed by calculating the pseudo-values based on the current estimates obtained from the n^{th} iteration, and using that pseudo-values to maximize the lower bound on the log-likelihood to obtain a new setting of estimates, and iterating between the above two steps until some converge criterion is satisfied. The EM algorithm is set up by introducing "missing data" into the problem. In particular we choose latent a variable w which is a two-point distribution:w = 1 the system is in perfect state(the observation follows degenerate distribution at (0,0,...,0)); w = 0 the system is in non-perfect state (the observation follows multi- variate Poisson (MPoi) distribution). Then the joint distribution function of $(w,\vec{y})$ would be

$$f(w,\vec{y}|\varphi,\vec{\lambda})=\begin{cases}\varphi^w[(1-\varphi)e^{-\sum\lambda_j}]^{1-w}, & \vec{y}=(0,0,...,0)\\ (1-\varphi)e^{-\sum\lambda_j}\phi(\vec{y}), & \vec{y}\neq(0,0,...,0)\end{cases} \tag{29}$$

where we assume the w and $\vec{y}$ are not independent.The EM method can be organized as follows. Given arbitrary initial values for $(\varphi^{(0)},\vec{\beta}^{(0)},\lambda_0^{(0)})$, we can find the conditional function and the mean of the variable w in the E step, and get the MLE for the joint distribution to update the estimates. We repeat the EM-method until the convergence conditions are satisfied.

E - step:

$$f(w|\vec{y},\varphi^{(t)},\beta^{(n)},\lambda_0^{(t)})$$

M - step:

$$(\varphi^{(t+1)},\vec{\beta}^{(t+1)},\lambda_0^{(t+1)})=argmax\sum_{i=1}^{n}E_{w|\vec{y},\varphi^{(t)},\beta^{(n)},\lambda_0^{(t)}}[lnf(\varphi,\beta,\lambda_0|w,\vec{y})]$$

In the $(t+1)^{th}$ E - step, we want to find the conditional distribution of w given the data and the t^{th} estimates for β,λ_0,τ.

The conditional distribution of w is

$$f(w|\vec{y},\varphi,\phi)=\frac{f(w,\vec{y}|\varphi,\phi)}{f(\vec{y}|\varphi,\phi)}$$

For convenience, we can organize the conditional distribution of w as two kinds of distributions given different observations; when the random variables of $\vec{y}$ are all zero, w is a Bernoulli distribution; when the random variables are not all zero, the system can only be in the non-perfect state, where the w is a degenerate distribution at 0.

$$w|(\vec{y},\varphi,\phi) \sim \begin{cases} Bernoulli(p), & \vec{y} = (0,0,...,0) \\ Degenerate(0), & \vec{y} \neq (0,0,...,0) \end{cases}$$

where the Bernoulli random variable has the possibility of $p = \frac{\varphi}{\varphi+(1-\varphi)exp(-\lambda_j)}$ with the system in the perfect state. The conditional expectation can be derived from the distribution

$$E_{w_i|\vec{y_i},\varphi_i,\beta_i,\lambda_{i0}}(w_i) = \begin{cases} \frac{\varphi_i}{\varphi_i+(1-\varphi_i)exp(-\lambda_{ij})}, & \vec{y} = (0,0,...,0) \\ 0, & \vec{y} \neq (0,0,...,0) \end{cases}$$

In the $(t+1)^{th}$ M-step, we find the maximum value of expectation of the log- likelihood function for the above joint distribution.
The log-likelihood function is

$$\begin{aligned} lnL(\tau,\beta,\lambda_0) &= \sum_{i=1}^{n_1} w_i ln\varphi_i + (1-w_i)(ln(1-\varphi_i) - \sum_{j=0}^{m}\lambda_{ij}) \\ &\quad + \sum_{i=1}^{n_2} ln(1-\varphi_i) + lnf(\vec{y_i}|\lambda_i) \\ &= \sum_{i=1}^{n_1}[w_i \sum_{r=1}^{p_2} g_{ir}\tau_r - ln(1+e^{\sum_{r=1}^{p_2} g_{ir}\tau_r}) \\ &\quad + (1-w_i)[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_2} x^{ir}\beta_{rj}}]] \\ &\quad + \sum_{i=1}^{n_2}[-ln(1+e^{\sum_{r=1}^{p_2} g_{ir}\tau_r}) + lnf(\vec{y_i}|\lambda_i,\beta)] \end{aligned}$$

where the $f(\vec{y_i}|\lambda_i,\beta)$ is the MPoi distribution function having the probability mass function as follows

$$f(y) = exp(-\sum_{j=0}^{m}\lambda_j)\phi(y_1,y_2,...,y_m)$$

where

$$\phi(y_1,y_2,...,y_m) = \sum_{z_0=0}^{min(y)}(\prod_{j=1}^{m}\frac{\lambda_j^{y_j-z_0}}{(y_j-z_0)!})\frac{\lambda_0^{z_0}}{z_0!}$$

for convenience, We rewrite the above log-likelihood as the sum of two functions L_1 and L_2

$$L_1(\tau) = \sum_{i=1}^{n_1}(w_i \sum_{r=1}^{p_2} g_{ir}\tau_r) - \sum_{i=1}^{n_1+n_2} ln(1 + e^{\sum_{r=1}^{p_2} g_{ir}\tau_r})$$

$$L_2(\beta, \lambda_0) = \sum_{i=1}^{n_1}(1 - w_i)[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_2} x^{ir}\beta_{rj}}] + \sum_{i=1}^{n_2} f(\vec{y_i}|\lambda_i, \beta)$$

L_1 is a function of τ and L_2 is a function of β and λ_0. By the NR method we should derive the first derivatives V and second derivatives J of the likelihood function, where

$$V = [V_1, V_2]'$$

and

$$\begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}$$

Vectors V_1 and V_2 are the first derivatives of L_1 and L_2, respectively. They have following items

$$\frac{\partial lnE(L_1)}{\partial \tau_{r_2}} = \sum_{i=1}^{n_1} E[w_i] g_{ir_2} - \sum_{i=1}^{n_1+n_2} g_{ir_2}\varphi_i$$

$$\frac{\partial lnE(L_2)}{\partial \beta_{r_{ij}}} = \sum_{i=1}^{n_1}(E[w_i] - 1)x_{ir}\lambda'_{ij} + \sum_{i=1}^{n_2} x_{ir}\lambda'_{ij}(\varphi_{ij}(1) - 1)$$

$$\frac{\partial lnE(L_2)}{\partial \lambda_0} = \sum_{i=1}^{n_1}(1 - E[w_i])(m-1) + n_2(m-1) + \sum_{i=1}^{n_2}(\phi_{ir_1}) - \sum_{j=1}^{m}\phi_{ij}$$

$r_1 = (1, 2, ..., p_1), j = (1, 2, ..., m)$
With the NR method, we repeatedly update the estimates until they satisfy the given criteria. After obtaining the $(t+1)^{th}$ estimates, we check the convergence criterion first. If it is satisfied, we can stop; if not, we continue to do the EM algorithm.

4. Robust estimation for multivariate zero-inflated Poisson regression

Here we propose an algorithm which is a modification of the EM algorithm with the property of robustness. Here also we choose latent a variable w which is a two-point distribution:w = 1 the system is in perfect state(the observation follows degenerate distribution at (0,0,...,0)); w = 0 the system is in non-perfect state (the observation follows MPoi distribution).Then the joint distribution function of $(w, \vec{y})$ would be (29).

The proposed algorithm begins with starting values for $(\tau^{(0)}, \vec{\beta}^{(0)}, \lambda_0^{(0)})$ and proceeds iteratively via the following steps until convergence. **E - step:**

$$f(w|\vec{y}, \tau^{(t)}, \beta^{(n)}, \lambda_0^{(t)})$$

M - step for τ : Find γ^{t+1} by maximizing $L_(\tau)$. This can be accomplished by fitting a binomial logistic regression of $w(r)$ on design matrix G with binomial denominator equal to one. It is equivalent to solving the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}[w_i^(r) - logit^{-1}(\sum_{r=1}^{p_2} g_{ir}\tau_r)]g_{ir} = 0 \tag{30}$$

M - step for $\beta and \lambda_0$**:** Find $\beta^(r+1)$ and $\lambda_0^(r+1)$ by maximizing $L_2(\beta, \lambda_0)$.It is equivalent to solving the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}(1 - w_i^{(r)})[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_2} x^{ir}\beta_{rj}}] + \sum_{i=1}^{n_2} lnf(\vec{y_i}|\lambda_0, \beta) = 0 \tag{31}$$

Now we replace the estimating functions (30) and (31) from the M step of the EM algorithm with robustified estimating functions. For this we downweight observations that fall in the extreme upper and lower tail of the multivariate Poisson distribution in the estimating function. Specifically, we suggest that $\tau^{(t+1)}$,$\beta^{(\vec{t+1})}$ and $\lambda_0^{(t+1)}$ be found by solving the following equations

$$\frac{1}{n}\sum_{i=1}^{n} w(g_i)[w_i^(r) - logit^{-1}(\sum_{r=1}^{p_2} g_{ir}\tau_r)]g_{ir} = 0 \tag{32}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - w_i^{(r)})w(x_i)[(m-1)\lambda_0 - \sum_{j=1}^{m} e^{\sum_{r=1}^{p_2} x^{ir}\beta_{rj}}] + \sum_{i=1}^{n_2} lnf(\vec{y_i}|\lambda_0, \beta) = 0 \tag{33}$$

where

$$f(\vec{y_i}|\lambda_0, \beta) = exp(-\sum_{j=0}^{m}\lambda_j)\phi(y_1, y_2, ..., y_m)$$

where

$$\phi(y_1, y_2, ..., y_m) = \sum_{z_0=0}^{min(y)}(\prod_{j=1}^{m}\frac{\lambda_j^{y_j - z_0}}{(y_j - z_0)!})\frac{\lambda_0^{z_0}}{z_0!}$$

and the function $\phi(y_1, y_2, ..., y_m)$ takes the value zero for outliers present in the data.
Here, $w(.)$ is a function to down weight large leverage points. A simple choice for $w(g_i)$ that we use throughout this paper is $\sqrt{1-h_i}$ where h_i is the i^{th} diagonal element of $H = g(g^Tg)^{-1}g^T$, with a similar definition for $w(xi)$.

Table 1: The 4-D zero-inflated Poisson observations

(y_1, y_2, y_3, y_4)	(y_1, y_2, y_3, y_4)	(y_1, y_2, y_3, y_4)	(y_1, y_2, y_3, y_4)
(0,0,0,0)	(1,0,0,0)	(0,0,0,0)	(0,0,1,2)
(0,0,0,0)	(0,0,0,0)	(0,3,0,0)	(0,0,1,0)
(0,0,0,0)	(1,0,0,0)	(0,0,0,0)	(0,0,1,2)
(0,1,0,0)	(1,0,0,0)	(0,0,0,0)	(0,0,0,0)
(0,0,0,0)	(1,0,0,0)	(0,0,0,0)	(0,0,0,0)
.	.	.	.
.	.	.	.
.	.	.	.
(0,0,0,0)	(0,0,0,0)	(0,0,0,0)	(0,0,0,0)
(0,1,0,0)	(0,1,0,0)	(0,0,1,0)	(0,0,0,0)
(0,4,0,2)	(0,0,0,0)	(0,0,0,0)	(0,0,0,0)

5. Simulation Study

Here we simulate 100 4-dimensional $y's$ and the data set look like the one shown in Table . Based o the link functions, we need the parameters τ , β, and the covariance λ_0 to generate the desired data sets.The parameters we used are as following where there are 4 $\tau's$ and 4 $\beta's$in each link function. We repeat the simulations 100 times in R to get 100 different count samples. $\tau = [0.9, 0.8, 0.4, 0.5], \lambda_0 = 0.64$, $\beta =$

$$\begin{pmatrix} 1.70 & 0.54 & 0.31 & 0.09 \\ 0.54 & 1.20 & 0.43 & 0.78 \\ 0.31 & 0.43 & 1.40 & 0.31 \\ 0.09 & 0.78 & 0.31 & 1.6 \end{pmatrix}$$

The simulation studies are conducted in R to study the performance of maximum likelihood estimation and the proposed estimation methods for MZIP data in the presence of outliers, i.e. with contamination in y. In the contaminated scenario, 5 per cent of the response y in the data set were randomly selected to be replaced by $y + 15$. Since solving log(L) analytically is difficult, we use maxLik package in R software to find the mles. Table 2 and Table 3 give the maximum likelihood estimates of parameters with bias obtained by EM algorithm. Table 4 and 5 give the estimates of parameters of the MZIP model with bias obtained by the proposed technique. From these tables we observe that there are some sort of improvements in the estimates because less bias in certain estimates of parameters.

Table 2: MLE estimates for τ and λ_0 with bias using EM algorithm

τ_1	τ_2	τ_3	τ_4	λ_0
0.89(-0.09)	0.199(0.601)	0.39(0.01)	0.499(0.001)	0.6399(0.0001)

Table 3: MLE estimates for β with bias using EM algorithm

j	1	2	3	4
β_{j1}	1.68(0.020)	0.5399(0.001)	0.309(0.001)	0.08(0.01)
β_{j2}	0.5403(-0.003)	1.19(0.01)	0.429(-0.001)	0.77(0.01)
β_{j3}	0.309(0.001)	0.429(-0.001)	1.39(0.01)	0.312(-0.002)
β_{j4}	0.08(0.01)	0.779(0.001)	0.311(-0.001)	1.5(0.1)

Table 4: MLE estimates for τ and λ_0 with bias using the proposed algorithm

τ_1	τ_2	τ_3	τ_4	λ_0
0.899(-0.099)	0.799(0.001)	0.399(0.001)	0.5001(-0.0001)	0.6401(-0.0001)

Table 5: MLE estimates for β with bias using the proposed algorithm

j	1	2	3	4
β_{j1}	1.699(0.001)	0.5399(0.001)	0.3100(0.09)	0.0899(0.0001)
β_{j2}	0.539(-0.009)	1.1995(0.01)	0.4302(-0.002)	0.779(-0.019)
β_{j3}	0.309(0.001)	0.429(0.001)	1.39(0.01)	0.309(-0.002)
β_{j4}	0.089(0.001)	0.779(-0.019)	0.309(0.001)	1.599(0.001)

6. Conclusion

In this paper, we introduce The ZIP and MZIP distributions and the corresponding regression models. The EM method is used to obtain estimates of the regression coefficients. A detailed review on Robust estimation for ZIP regression are also given. Along with it, we propose a robust estimation for MZIP regression and conduct a simulation study. In this paper the method adopted for robust estimation is to accommodate, rather than eliminate, outliers and use a robust estimation methodology that minimizes their effect on estimation of the model that is followed by the vast majority of data. On the other hand, it needs some research study to obtain a more accurate estimating function to down weight the extreme observations with increased number of variables.

References

[1] Hall, D. B., and Shen, J. (2010). Robust estimation for zero-inflated poisson regression. Scandinavian Journal of Statistics, 37 (2), 237252.

[2] Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in r. Computational Statistics, 26 (3), 443458.

[3] Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics, 34 (1), 114.

[4] Mallows, C. L. (1975). On some topics in robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ .

[5] McCullagh, P.and JA. (81). Nelder. 1989. *Generalized linear models.* London, chapman hall.

[6] Mian, R.and Paul, S. (2016). Estimation for zero-inflated over-dispersed count data model with missing response. Statistics in medicine, 35 (30), 56035624.

[7] Ridout, M., Demetrio, C. G. and Hinde, J. (1998). Models for count data with many zeros. *In Proceedings of the xixth international biometric conference* (Vol. 19, pp. 179 - 192).

[8] Rousseeuw, F. H. E. R. P., Hampel, F., Ronchetti, E. and Stahel, W. (1986). *Robust statistics: the approach based on influence functions.* Wiley, New YorkHuber PJ (1964) Robust estimation of a location parameter. Ann Math Stat, 35 , 73101Huber.

[9] Shen, J. (2006). Robust estimation and inference in finite mixtures of generalized linear models (Unpublished doctoral dissertation). University of Georgia.

[10] Wang, Y. (2017). *Multivariate zero-inflated poisson regression.*

[11] Yang, J., Xie, M. and Goh, T. N. (2011). Outlier identification and robust parameter estimation in a zero-inflated poisson model. *Journal of Applied Statistics*,38 (2), 421 - 430.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 183-192.

Compound Extended Gamma Density and Connection to Inverse Gaussian

Dhannya P. Joseph
Department of Statistics, K. E. College Mannanam
Kerala, India
Email:dhannyapj@gmail.com

Abstract

Extended form of gamma density is considered. Pathway model of Mathai [A pathway to matrix-variate gamma and normal densities, Linear Algebra and its Applications, 396 (2005)317-328] is utilized for the extension. Normalizing constants are evaluated by using inverse Mellin transform techniques. A connection between extended gamma density and extended form of inverse Gaussian is derived. Some extended mixtures are discussed.

Key words: Pathway model, extended gamma, $H-$function.

1. Introduction

In probability theory, the gamma density is a two-parameter family of continuous probability density with support on $(0, \infty)$. Its probability density function is given by

$$f(x; m, p) = \frac{m^p}{\Gamma(p)} x^{p-1} \mathrm{e}^{-mx}, \tag{1}$$

for $x > 0$, where $p > 0$ is the shape and $m > 0$ is the scale parameter. In this paper this density is given a generalization or extension with the help of pathway model of Mathai (2005). In order to create a distributional pathway for proceeding from one functional form to another a pathway parameter α is introduced and a pathway model is created. For the real scalar case the pathway model is the following, for $x > 0$

$$g_1(x) = k_1 x^{\gamma-1}[1 - a(1-\alpha)x^{\delta}]^{\frac{\eta}{1-\alpha}} \tag{2}$$

$a > 0,\ \delta > 0,\ 1 - a(1-\alpha)x^{\delta} > 0,\ \gamma > 0,\ \eta > 0$, where k_1 is the normalizing constant and α is the pathway parameter. For $\alpha < 1$ the model remains as a generalized type-1 beta model in the real case. When $\alpha > 1$ we may write $1 - \alpha = -(\alpha - 1)$ so that (2) assumes the form of generalized type-2 beta model, given by

$$g_2(x) = k_2 x^{\gamma-1}[1 + a(\alpha-1)x^{\delta}]^{-\frac{\eta}{\alpha-1}},\ x > 0, \tag{3}$$

where k_2 is the normalizing constant. When $\alpha \to 1$ the forms in (2) and (3) reduce to generalized gamma form

$$g_3(x) = k_3 x^{\gamma-1}\mathrm{e}^{-a\eta x^{\delta}},\ x > 0, \tag{4}$$

where k_3 is the normalizing constant, see Mathai & Haubold (2007) and Mathai & Moschopoulos (1991). More families are available when the variable is allowed to vary over the real line.

In order to extend the gamma density we use the model in (3). Let us take the extended generalized models as

$$g_\beta(x) = k_\beta x^{\gamma}\mathrm{e}^{-ax^{\delta}}[1 + b(\beta-1)x^{-\rho}]^{-\frac{1}{\beta-1}}, x > 0, a > 0, b > 0, \delta > 0, \rho > 0, \beta > 1, \tag{5}$$

$$g_\alpha(x) = k_\alpha x^{\gamma}[1 + a(\alpha-1)x^{\delta}]^{-\frac{1}{\alpha-1}}\mathrm{e}^{-bx^{-\rho}}, x > 0, a > 0, b > 0, \delta > 0, \rho > 0, \alpha > 1, \tag{6}$$

and

$$g(x) = kx^{\gamma}\mathrm{e}^{-ax^{\delta}}\mathrm{e}^{-bx^{-\rho}}, x > 0, a > 0, b > 0, \delta > 0, \rho > 0. \tag{7}$$

where k_β, k_α and k are the normalizing constants. Note that

$$\lim_{\beta\to 1} g_\beta(x) = \lim_{\alpha\to 1} g_\alpha(x) = g(x). \tag{8}$$

If the generalized gamma is a limiting or stable situation then the pathway model gives all the paths leading to this stable or limiting distribution. We also give interpretations in terms of special functions (Mathai (1993), Mathai & Saxena (1978) or Kilbas & Saigo (2004)) in statistical distribution theory.

2. Evaluation Of The Normalizing Constant Of The Extended Gamma Density

In order to find the normalizing constant, we have evaluate the total integral. Let

$$k_\beta^{-1} = \int_0^{\infty} x^{\gamma}\mathrm{e}^{-ax^{\delta}}[1 + b(\beta-1)x^{-\rho}]^{-\frac{1}{\beta-1}}\mathrm{d}x,\ a > 0, b > 0, \delta > 0, \rho > 0, \beta > 1. \tag{9}$$

Here the integrand can be taken as a product of positive integrable functions and then we can apply statistical distribution theory to evaluate this integral. Let x_1 and x_2 be real

scalar independently distributed random variables having densities

$$f_1(x_1) = \begin{cases} c_1 x_1^{\gamma+1} \mathrm{e}^{-a{x_1}^{\delta}}, & 0 < x_1 < \infty,\ \gamma + 2 > 0,\ a > 0 \\ 0, \text{ elsewhere} \end{cases} \tag{10}$$

and

$$f_2(x_2) = \begin{cases} c_2[1 + (\beta - 1)x_2{}^{\rho}]^{-\frac{1}{\beta-1}}, 0 < x_2 < \infty,\ \rho > 0,\ \beta > 1 \\ 0, \text{ elsewhere.} \end{cases} \tag{11}$$

where c_1 and c_2 are normalizing constants. For the method to be discussed we need not evaluate the normalizing constants explicitly. Let us transform x_1 and x_2 to $u = x_1 x_2$ and $v = x_1$. Then from standard procedures in statistical distribution theory the marginal density of u is given by

$$\begin{aligned} g_u(u) &= \int_v \frac{1}{v} f_1(v) f_2(\frac{u}{v}) \mathrm{d}v \\ &= c_1 c_2 \int_0^{\infty} v^{\gamma} \mathrm{e}^{-av^{\delta}} [1 + b(\beta - 1)v^{-\rho}]^{-\frac{1}{\beta-1}} \mathrm{d}v, \text{ where } b = u^{\rho}. \end{aligned} \tag{12}$$

Let us evaluate the density through moments. We will take the h^{th} moment for $h = s - 1$ where s is a complex variable. This is done so that we can apply the theory of Mellin and inverse Mellin transforms to evaluate the density in explicit form.

$$\begin{aligned} E(u^{s-1}) &= E(x_1^{s-1})E(x_2^{s-1}), \text{ due to statistical independence of } x_1 \text{ and } x_2 \\ &= \frac{c_1 c_2 \Gamma(\frac{\gamma+1+s}{\delta})\Gamma(\frac{s}{\rho})\Gamma(\frac{1}{\beta-1} - \frac{s}{\rho})}{\delta \rho a^{\frac{\gamma+s+1}{\delta}} (\beta - 1)^{\frac{s}{\delta}} \Gamma(\frac{1}{\beta-1})}, \end{aligned} \tag{13}$$

$\Re(\gamma + s + 1) > 0,\ \Re(s) > 0,\ \Re(\frac{1}{\beta-1} - \frac{s}{\rho}) > 0.$

Here we have the moments available for a complex parameter s. In this case we can apply the theory of inverse Mellin transform to determine the density of u uniquely. Looking at the $(s-1)^{th}$ moment as the Mellin transform of the corresponding density and then taking the inverse Mellin transform we get the density of u,

$$g_u(u) = \frac{c_1 c_2}{\delta \rho a^{\frac{\gamma+1}{\delta}} \Gamma(\frac{1}{\beta-1})} \frac{1}{2\pi i} \int_L \Gamma(\frac{\gamma+1}{\delta} + \frac{s}{\delta})\Gamma(\frac{s}{\rho})\Gamma(\frac{1}{\beta-1} - \frac{s}{\rho}) \big(u a^{\frac{1}{\delta}} (\beta - 1)^{\frac{1}{\rho}}\big)^{-s} \mathrm{d}s. \tag{14}$$

This contour integral can be written as an H-function (Mathai & Saxena (1978)). That is,

$$g_u(u) = \frac{c_1 c_2}{\delta \rho a^{\frac{\gamma+1}{\delta}} \Gamma(\frac{1}{\beta-1})} H_{1,2}^{2,1}\left[a^{\frac{1}{\delta}} (b(\beta - 1))^{\frac{1}{\rho}} \Big|_{(0,\frac{1}{\rho}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})}^{(1-\frac{1}{\beta-1},\frac{1}{\rho})} \right],\ b = u^{\rho}, \tag{15}$$

where $H-$ function is defined as

$$H_{p,q}^{m,n}\big[z\big|_{(b_1,\beta_1),,\ldots,(b_q,\beta_q)}^{(a_1,\alpha_1),\ldots,(a_p,\alpha_p)}\big] = \frac{1}{2\pi i} \int_L \phi(s)\ z^{-s} \mathrm{d}s, \tag{16}$$

where

$$\phi(s) = \frac{\left\{\prod_{j=1}^{m}\Gamma(b_j+\beta_j s)\right\}\left\{\prod_{j=1}^{n}\Gamma(1-a_j-\alpha_j s)\right\}}{\left\{\prod_{j=m+1}^{q}\Gamma(1-b_j-\beta_j s)\right\}\left\{\prod_{j=n+1}^{p}\Gamma(a_j+\alpha_j s)\right\}},$$

where $\alpha_j,\ j=1,2,...,p$ and $\beta_j,\ j=1,2,...,q$ are real positive numbers, $a_j,\ j=1,2,...,p$ and $b_j,\ j=1,2,...,q$ are complex numbers, L is a contour separating the poles of $\Gamma(b_j+\beta_j s),\ j=1,2,...,m$ from those of $\Gamma(1-a_j-\alpha_j s),\ j=1,2,...,n$. Comparing equations (12) and (15)

$$\begin{aligned} k_\beta^{-1} &= \int_0^\infty x^\gamma \mathrm{e}^{-ax^\delta}[1+b(\beta-1)x^{-\rho}]^{-\frac{1}{\beta-1}}\mathrm{d}x \\ &= \frac{1}{\delta\rho a^{\frac{\gamma+1}{\delta}}\Gamma(\frac{1}{\beta-1})}H_{1,2}^{2,1}\left[a^{\frac{1}{\delta}}(b(\beta-1))^{\frac{1}{\rho}}\middle|_{(0,\frac{1}{\rho}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})}^{(1-\frac{1}{\beta-1},\frac{1}{\rho})}\right]. \end{aligned} \tag{17}$$

In a similar way we can evaluate the normalizing constants of the functions $g_\alpha(x)$ and $g(x)$. For evaluating the normalizing constant of $g_\alpha(x)$, we take $f_1(x_1)$ as

$$f_1(x_1) = \begin{cases} m_1 x_1^{\gamma+1}[1+a(\alpha-1)x_1{}^\delta]^{-\frac{1}{\alpha-1}}, & 0<x_1<\infty,\ \gamma+2>0,\ a>0,\ \delta>0 \\ 0,\ \text{elsewhere} \end{cases} \tag{18}$$

and $f_2(x_2)$ as

$$f_2(x_2) = \begin{cases} m_2\mathrm{e}^{-x_2{}^\rho}, 0<x_2<\infty,\ \rho>0,\ \beta>1 \\ 0,\ \text{elsewhere}, \end{cases} \tag{19}$$

and proceed as before we get the normalizing constant k_α of $g_\alpha(x)$, where

$$\begin{aligned} k_\alpha^{-1} &= \int_0^\infty x^\gamma[1+a(\alpha-1)x^\delta]^{-\frac{1}{\alpha-1}}\mathrm{e}^{-bx^{-\rho}}\mathrm{d}x \\ &= \frac{1}{\delta\rho(a(\alpha-1))^{\frac{\gamma+1}{\delta}}\Gamma(\frac{1}{\alpha-1})}H_{1,2}^{2,1}\left[(a(\alpha-1))^{\frac{1}{\delta}}b^{\frac{1}{\rho}}\middle|_{(0,\frac{1}{\rho}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})}^{(1-\frac{1}{\beta-1}+\frac{\gamma+1}{\delta},\frac{1}{\delta})}\right],\ b=u^\rho. \end{aligned} \tag{20}$$

For evaluating the normalizing constant of $g(x)$, we take the limit $\beta\to 1$ in (17) or $\alpha\to 1$ in (20). By taking limits we apply Stirling's approximations for gamma functions, see for example Mathai (1993), given by

$$\Gamma(z+a)\to(2\pi)^{\frac{1}{2}}z^{z+a-\frac{1}{2}}\mathrm{e}^{-z}, \text{for } |z|\to\infty \text{ and } a \text{ is bounded} \tag{21}$$

to the gamma ratios in the $H-$ functions in equation (17) or (20) and we obtain

$$\begin{aligned} k^{-1} &= \int_0^\infty x^\gamma \mathrm{e}^{-ax^\delta}\mathrm{e}^{-bx^{-\rho}}\mathrm{d}x \\ &= \frac{1}{\delta\rho a^{\frac{\gamma+1}{\delta}}}H_{0,2}^{2,0}\left[a^{\frac{1}{\delta}}b^{\frac{1}{\rho}}\middle|_{(0,\frac{1}{\rho}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})}\right],\ b=u^\rho, \end{aligned} \tag{22}$$

3. Inverse Gaussian As A Particular Case

Note that one form of the inverse Gaussian probability density function is given by

$$h(x) = cx^{-\frac{3}{2}}\mathrm{e}^{-\frac{\lambda}{2}(\frac{x}{\mu^2}+\frac{1}{x})},\ \mu \neq 0,\ x > 0,\ \lambda > 0,$$

where c is the normalizing constant (see Mathai (1993)). Put $\gamma = -\frac{3}{2},\ \delta = 1,\ \rho = 1,\ a = \frac{\lambda}{2\mu^2},\ b = \frac{\lambda}{2}$ in $g(x)$ given in equation (7), we can see that $g(x)$ will become the inverse Gaussian density. Since $g_\beta(x)$ and $g_\alpha(x)$ are the extensions of $g(x)$, we can say that $g_\beta(x)$ and $g_\alpha(x)$ are the extended forms of inverse Gaussian density.

The integral k^{-1} in equation (22) can be used to evaluate the moments of inverse Gaussian density. Also reaction rate probability integral in nuclear reaction rate theory, Krätzel integrals in applied analysis and the like will become special cases of the integral k^{-1} (see Krätzel (1979), Haubold & Kumar (2008), Mathai & Haubold (1988)).

4. Laplace Transform Of Extended Generalized Gamma

We have

$$g_\beta(x) = k_\beta x^\gamma \mathrm{e}^{-ax^\delta}[1 + b(\beta-1)x^{-\rho}]^{-\frac{1}{\beta-1}}.$$

Put $\delta = 1$ in $g_\beta(x)$, then the Laplace transform of $g_\beta(x)$ is given by

$$\begin{aligned} L_{g_\beta}(t) &= k_\beta \int_0^\infty x^\gamma \mathrm{e}^{-x(a+t)}[1 + b(\beta-1)x^{-\rho}]^{-\frac{1}{\beta-1}}\mathrm{d}x \\ &= c_\beta^{-1}\Big(\frac{a}{a+t}\Big)^{\gamma+1} H_{1,2}^{2,1}\left[(a+t)(b(\beta-1))^{\frac{1}{\rho}}\middle|{}^{(1-\frac{1}{\beta-1},\frac{1}{\rho})}_{(0,\frac{1}{\rho}),(\gamma+1,1)}\right], \end{aligned} \tag{23}$$

where $c_\beta = H_{1,2}^{2,1}\left[a(b(\beta-1))^{\frac{1}{\rho}}\middle|{}^{(1-\frac{1}{\beta-1},\frac{1}{\rho})}_{(0,\frac{1}{\rho}),(\gamma+1,1)}\right]$.

5. Connection To Extended Form Of Gamma

If $\rho < 0$ and $\delta = 1$ in $g_\beta(x)$ of equation (5), then the integral k_β^{-1} becomes

$$k_{1_\beta}^{-1} = \int_0^\infty x^\gamma \mathrm{e}^{-ax}[1 + b(\beta-1)x^\mu]^{-\frac{1}{\beta-1}}\mathrm{d}x,\ a > 0, b > 0, \mu = -\rho > 0, \beta > 1. \tag{24}$$

Note that this will be the Laplace transform of the extended form of the gamma density. Extended in the sense that when we take the limit as $\beta \longrightarrow 1$, it will be the Laplace transform of the generalized gamma density. The Laplace transform of this density also provides

the moment generating function of extended gamma density thereby the moment generating functions of extended forms of Weibull, chisquare, Reyleigh, Maxwell-Boltzmann, exponential and other densities in this general class. Observe that the structure in (24) is different from those considered so far because the exponents in both the factors are positive. In the previous cases we had one exponent positive and the other negative. In order to evaluate this integral let us consider the transformation $u = \frac{x_1}{x_2}$ and $v = x_2$, where x_1 and x_2 are real scalar independently distributed random variables. Let the densities of x_1 and x_2 be

$$h_1(x_1) = \begin{cases} c_3 \mathrm{e}^{-x_1}, & 0 < x_1 < \infty \\ 0, & \text{elsewhere} \end{cases} \tag{25}$$

and

$$h_2(x_2) = \begin{cases} c_4 x_2^{\gamma-1}[1 + b(\beta-1)x_2{}^{\mu}]^{-\frac{1}{\beta-1}}, x_2 > 0,\ b > 0,\ \gamma > 0,\ \mu > 0,\ \beta > 1 \\ 0, \text{elsewhere}, \end{cases} \tag{26}$$

where c_3 and c_4 are the normalizing constants. Then the density of u, denoted by $g_u(u)$, is given by

$$\begin{aligned} g_u(u) &= \int_0^\infty v g_1(uv) g_2(v) \mathrm{d}v \\ &= c_3 c_4 \int_0^\infty v \mathrm{e}^{-uv} v^{\gamma-1}[1 + b(\beta-1)v^{\mu}]^{-\frac{1}{\beta-1}} \mathrm{d}v. \end{aligned} \tag{27}$$

Here

$$\begin{aligned} E(u^{s-1}) &= E(\frac{x_1}{x_2})^{s-1} = E(x_1)^{s-1} E(x_2)^{1-s}, \text{because of independance} \\ &= c_3 c_4 \frac{\Gamma(s)\Gamma(\frac{\gamma+1}{\mu} - \frac{s}{\mu})\Gamma(\frac{1}{\beta-1} - \frac{\gamma}{\mu} - \frac{1}{\mu} + \frac{s}{\mu})}{\mu(b(\beta-1))^{\frac{\gamma+1-s}{\mu}}\Gamma(\frac{1}{\beta-1})}, \end{aligned} \tag{28}$$

for $\Re(s) > 0,\ \Re(\gamma - s + 1) > 0,\ \Re(\frac{1}{\beta-1} - \frac{\gamma}{\mu} - \frac{1}{\mu} + \frac{s}{\mu}) > 0$.
Taking inverse Mellin transform and comparing with (27) (with $u = a$)

$$\begin{aligned} I_{1_\beta} &= \int_0^\infty \mathrm{e}^{-av} v^{\gamma}[1 + b(\beta-1)v^{\mu}]^{-\frac{1}{\beta-1}} \mathrm{d}v \\ &= \frac{1}{\mu(b(\beta-1))^{\frac{\gamma+1}{\mu}}\Gamma(\frac{1}{\beta-1})} H_{1,2}^{2,1}\left[\frac{a}{b^{\frac{1}{\mu}}(\beta-1)^{\frac{1}{\mu}}}\middle|_{(0,1),(\frac{1}{\beta-1}-\frac{\gamma}{\mu}-\frac{1}{\mu},\frac{1}{\mu})}^{(1-\frac{\gamma+1}{\mu},\frac{1}{\mu})}\right]. \end{aligned} \tag{29}$$

Now put $\rho = -1$ in $f_\alpha(x)$ of equation (**??**), then the integral I_α becomes

$$I_{1_\alpha} = \int_0^\infty x^{\gamma} \mathrm{e}^{-bx}[1 + a(\alpha-1)x^{\delta}]^{-\frac{1}{\alpha-1}} \mathrm{d}x,\ a > 0, b > 0, \delta > 0, \alpha > 1. \tag{30}$$

Note that this integral will be similar to I_{1_β}.

6. Extended Gamma Mixtures

6.1 Poisson-Extended Generalized Gamma

One of the first to consider inverse Gaussian as a mixing tool was Holla (1966). He assumed that the mean parameter θ of a Poisson law followed the inverse Gaussian distribution and arrived at the compound law which come to be known as the Poisson-inverse Gaussian law. Here we assume that the parameter θ of the Poisson law

$$g_{x|\theta}(x|\theta) = \frac{\theta^{\delta x}\mathrm{e}^{-\theta^\delta}}{x!},\ \theta > 0,\ \delta > 0,\ x = 0, 1, 2\cdots, \tag{31}$$

following the extended generalized gamma model

$$\pi_\theta(\theta) = k_{\beta_1}\theta^\gamma \mathrm{e}^{-a\theta^\delta}[1 + b(\beta-1)\theta^{-\delta}]^{-\frac{1}{\beta-1}},\ \theta > 0,\ a > 0,\ \beta > 1,\ b > 0,\ \delta > 0, \tag{32}$$

where

$$k_{\beta_1} = d_1^{-1}\delta^2 a^{\frac{\gamma+1}{\delta}}\Gamma(\frac{1}{\beta-1}), \tag{33}$$

where

$$d_1 = H_{1,2}^{2,1}\left[(ab(\beta-1))^{\frac{1}{\delta}}\middle|\begin{matrix}(1-\frac{1}{\beta-1},\frac{1}{\delta})\\ (0,\frac{1}{\delta}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})\end{matrix}\right]. \tag{34}$$

Then we can find out the Poisson-extended generalized gamma law and is given by

$$\begin{aligned} p_1(x) &= k_{\beta_1}\int_0^\infty g_{x|\theta}(x|\theta)\pi_\theta(\theta)\mathrm{d}\theta \\ &= \frac{k_{\beta_1}}{x!}\int_0^\infty \theta^{\gamma+\delta x}\mathrm{e}^{-(a+1)\theta^\delta}[1 + b(\beta-1)\theta^{-\delta}]^{-\frac{1}{\beta-1}}\mathrm{d}\theta \\ &= d_1^{-1}\frac{a^{\frac{\gamma+1}{\delta}}}{x!(a+1)^{\frac{\gamma+\delta x+1}{\delta}}}H_{1,2}^{2,1}\left[((a+1)b(\beta-1))^{\frac{1}{\delta}}\middle|\begin{matrix}(1-\frac{1}{\beta-1},\frac{1}{\delta})\\ (0,\frac{1}{\delta}),(\frac{\gamma+\delta x+1}{\delta},\frac{1}{\delta})\end{matrix}\right],\ x = 0, 1, 2\cdots \end{aligned} \tag{35}$$

6.2 Exponential-extended generalized gamma

Bhattacharya & Kumar (1986) proposed a model for life time distribution by compounding the exponential distribution with the inverse Gaussian law, to obtain the exponential-inverse Gaussian law. Here we are going to find the exponential-extended generalized gamma law. Consider the exponential density with parameter λ

$$g_{x|\lambda}(x|\lambda) = \lambda\mathrm{e}^{-\lambda x},\ \lambda > 0,\ x > 0, \tag{36}$$

and assume that the parameter λ has a prior extended generalized gamma density $\pi_\lambda(\lambda)$

$$\pi_\lambda(\lambda) = k_{\beta_2}\lambda^{\gamma}\mathrm{e}^{-a\lambda}[1+b(\beta-1)\lambda^{-\rho}]^{-\frac{1}{\beta-1}},\ \rho>0,\ a>0,\ \beta>1,\ b>0,\ \lambda>0 \tag{37}$$

where

$$k_{\beta_2} = d_2^{-1}\rho a^{\gamma+1}\Gamma(\frac{1}{\beta-1}), \tag{38}$$

where

$$d_2 = H_{1,2}^{2,1}\left[a(b(\beta-1))^{\frac{1}{\rho}}\middle|\begin{matrix}(1-\frac{1}{\beta-1},\frac{1}{\rho})\\ (0,\frac{1}{\rho}),(\gamma+1,1)\end{matrix}\right]. \tag{39}$$

Then we get the exponential-extended generalized gamma law as

$$\begin{aligned} p_2(x) &= k_{\beta_2}\int_0^\infty g_{x|\lambda}(x|\lambda)\pi_\lambda(\lambda)\mathrm{d}\lambda \\ &= k_{\beta_2}\int_0^\infty \lambda^{\gamma+1}\mathrm{e}^{-\lambda(a+x)}[1+b(\beta-1)\lambda^{-\rho}]^{-\frac{1}{\beta-1}}\mathrm{d}\lambda \\ &= d_2^{-1}\frac{a^{\gamma+1}}{(a+x)^{\gamma+2}}H_{1,2}^{2,1}\left[(a+x)(b(\beta-1))^{\frac{1}{\delta}}\middle|\begin{matrix}(1-\frac{1}{\beta-1},\frac{1}{\rho})\\ (0,\frac{1}{\rho}),(\gamma+2,1)\end{matrix}\right],\ x>0. \end{aligned} \tag{40}$$

6.3 Gamma-extended generalized gamma

Consider the generalized gamma density

$$g_{x|\mu}(x|\mu) = \frac{\delta\mu^{\frac{\nu+1}{\delta}}}{\Gamma(\frac{\nu+1}{\delta})}x^{\nu}\mathrm{e}^{-\mu x^{\delta}},\ x>0,\ \delta>0,\ \mu>0\ \nu>0 \tag{41}$$

and assume that the parameter μ has a prior extended generalized gamma density $\pi_\mu(\mu)$

$$\pi_\mu(\mu) = k_{\beta_2}\mu^{\gamma}\mathrm{e}^{-a\mu}[1+b(\beta-1)\mu^{-\rho}]^{-\frac{1}{\beta-1}},\ \rho>0,\ a>0,\ \beta>1,\ b>0,\ \mu>0 \tag{42}$$

Then we get the gamma-extended generalized gamma law as

$$\begin{aligned} p_3(x) &= k_{\beta_2}\int_0^\infty g_{x|\mu}(x|\mu)\pi_\mu(\mu)\mathrm{d}\mu \\ &= d_2^{-1}\delta x^{\nu}\frac{a^{\gamma+1}}{\Gamma(\frac{\nu+1}{\delta})(a+x^{\delta})^{\gamma+1+\frac{\nu+1}{\delta}}}H_{1,2}^{2,1}\left[(a+x^{\delta})(b(\beta-1))^{\frac{1}{\rho}}\middle|\begin{matrix}(1-\frac{1}{\beta-1},\frac{1}{\rho})\\ (0,\frac{1}{\rho}),(\gamma+1+\frac{\nu+1}{\delta},1)\end{matrix}\right], \\ &\qquad x>0. \end{aligned} \tag{43}$$

6.4 Gaussian-extended generalized gamma

Sankaran (1968) proposed the Normal-inverse Gaussian model. Consider the normal density

$$g_{x|\sigma}(x|\sigma) = \frac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{x^2}{2\sigma^2}},\ \sigma>0,\ -\infty<x<\infty, \tag{44}$$

and assume that the parameter σ has a prior extended generalized gamma density

$$\pi_\sigma(\sigma) = k_{\alpha_1}\sigma^\gamma[1+a(\alpha-1)\sigma^\delta]^{-\frac{1}{\alpha-1}}\mathrm{e}^{-b\sigma^{-2}},\ \sigma>0,\ a>0,\ b>0,\ \delta>0. \tag{45}$$

where

$$k_{\alpha_1} = 2d_3^{-1}\delta(a(\alpha-1))^{\frac{\gamma+1}{\delta}}\Gamma(\frac{1}{\alpha-1}), \tag{46}$$

where

$$d_3 = H_{1,2}^{2,1}\left[(a(\alpha-1))^{\frac{1}{\delta}}(b)^{\frac{1}{2}}\Big|_{(0,\frac{1}{2}),(\frac{\gamma+1}{\delta},\frac{1}{\delta})}^{(1-\frac{1}{\beta-1}+\frac{\gamma+1}{\delta},\frac{1}{\delta})}\right]. \tag{47}$$

Now we get the Normal-extended generalized gamma law as

$$\begin{aligned} p_4(x) &= k_\alpha\int_0^\infty g_{x|\sigma}(x|\sigma)\pi_\sigma(\sigma)\mathrm{d}\sigma \\ &= k_\alpha\int_0^\infty \sigma^{\gamma-1}\mathrm{e}^{-\frac{x^2}{2}\sigma^{-2}}[1+a(\alpha-1)\sigma^\delta]^{-\frac{1}{\alpha-1}}\mathrm{d}\sigma \\ &= d_3^{-1}(a(\alpha-1))^{\frac{1}{\delta}}H_{1,2}^{2,1}\left[(a(\alpha-1))^{\frac{1}{\delta}}(\frac{x^2}{2}+b)^{\frac{1}{2}}\Big|_{(0,\frac{1}{2}),(\frac{\gamma}{\delta},\frac{1}{\delta})}^{(1-\frac{1}{\beta-1}+\frac{\gamma}{\delta},\frac{1}{\delta})}\right], \\ &\qquad -\infty<x<\infty. \end{aligned} \tag{48}$$

7. Conclusions

By using the pathway parameter we can go to extended densities where the standard densities become particular cases. Through the pathway parameter α or β we can go from one functional form to another. If the standard densities are considered as stable then many unstable and chaotic behavior showing densities are covered using the extended model considered here. The graphs explained here show the behavior of the extended densities and the standard densities. These densities have wide range of applications in physics and statistics.

References

[1] S. K. Bhattacharya & S. Kumar (1986). Bayesian Life Estimation with an inverse Gaussian prior. *South African statistical journal*, 20, 37-43.

[2] H. J. Haubold & D. Kumar, D. (2008). Extension of thermonuclear function through the pathway model including Maxwell-Boltzmann and Tsallis distributions. *Astro Particles Physics*, 29, 70–76.

[3] M. S. Holla (1966). On a Poisson-inverse Gaussian distribution. *Metrica*, 11, 115–121.

[4] A. A. Kilbas & M. Saigo (2004). Analytic Methods and Special Functions, *H-transforms: Theory and Applications. Ser.*, 9.

[5] E. Krätzel (1979). Integral transformations of Bessel type. *Generalized functions and Operational Calculus*,148–465. Proc. Conf. Varna, , Bulg. Acad. Sci. Sofia.

[6] A. M. Mathai (1993). *A Handbook of Generalized Special Functions for Statistical and Physical Sciences*. Oxford University Press, Oxford.

[7] A. M. Mathai (2005). A Pathway to matrix-variate gamma and normal densities, *Linear Algebra and its Applications*, 396, 317–328.

[8] A.M. Mathai & H. J. Haubold (1988). *Modern Problems in Nuclear and Neutrino Astrophysics*. Akademic-Verlag Berlin.

[9] A.M. Mathai & P. G. Moschopoulos (1991).On Multivariate Gamma. *Journal of Multivariate Analysis*, 39, 135–153.

[10] A. M. Mathai & S. B. Provost (2006).On $q-$Logistic and Related Models. *IEEE Transactions on Reliability*, 55(2), 237–244.

[11] A. M. Mathai & R. K. Saxena (1978). *The H-function with Applications in Statistics and Other Disciplines*. Wiley Halsted, New York, London and Sidney.

[12] M. Sankaran (1968). Mixtures by the inverse Gaussian distribution. *Sankhya*, 30B, 455–458.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 193-206.

Robust Multivariate Analysis of Variance

Sajesh T A and Arya Damodharan
Department of Statistics, St.Thomas College, Thrissur, Kerala -1.
Email:sajesh.t.abraham@gmail.com, aryadamodharan@gmail.com

Abstract

The Wilks' Lambda Statistic (likelihood ratio test, LRT) is a commonly used tool for inference about the mean vectors of several multivariate normal populations. It is well known that the Wilk's Lambda Statistic which is based on classical estimates of population mean and dispersion matrices is extremely sensitive to influence of outliers. This study is an attempt to propose a robust version of multivariate analysis of variance by replacing the classical Wilk's Lambda statistic into a robust one through substituting the classical estimates by a highly robust and efficient kurtosis estimates and discuss about its properties. The approximate distribution of the proposed robust statistic was derived using simulations and its fitting is examined using QQ – plots. Monte Carlo simulations were used to investigate the efficiencies of the proposed method. The level of significance and power of the proposed robust MANOVA are compared with that of classical MANOVA using the size -power curve. Finally, the method is applied to well-known real data sets to evaluate its performance.

Key words: Robustness, Kurtosis estimates, Wilks' Lambda, MANOVA.

1. Introduction

One-way multivariate analysis of variance (MANOVA) deals with testing the null hypothesis of equal mean vectors across the g considered groups. The setup is similar to that of the one-way univariate analysis of variance (ANOVA) but the inter-correlations of the independent variables are taken into account, i.e. the variables are considered multivariate. Under the classical assumptions that all groups arise from multivariate normal distributions, many test statistics are discussed in the literature, one of the most widely used being the likelihood ratio test. This test statistic is better known as Wilk's Lambda

in MANOVA. The Wilks' Lambda is reported as part of the test output in almost all statistical packages. However, this measure which uses the classical normal theory as well as the inference based on it can be adversely affected by outliers present in the data. The non-robustness of the Wilk's Lambda statistic in the context of variable selection in linear discriminant analysis was demonstrated in Todorov (2007a).

The effect of outliers on the quality of the hypothesis test based on the classical Wilk's Lambda statistics will be illustrated in the example and the simulation study in the coming sections. Therefore we propose to use robust estimators instead of the classical ones for computing Wilk's Lambda statistic. The non-robustness of the normal theory based test statistic has led many other authors also to search for alternatives. For this purpose we will use the Kurtosis estimator of Pena and Prieto (2001) which is a highly robust estimator of location and scatter. Since the distribution of the robust Wilk's Lambda statistic based on kurtosis differs from the classical one it is necessary to find a good approximation for this distribution. We construct an approximate distribution based on a Monte Carlo study and examined its accuracy. The adaptations of the kurtosis estimator for computing the common covariance matrix is also summarized in the coming sections.

Monte Carlo simulations are used to evaluate the performance of the proposed test statistic under various distributions in terms of the simulated significance levels, its power functions and robustness. The power of the robust and classical statistic is compared using size-power curves, for the construction of which no knowledge about the distribution of the statistic is necessary (see Davidson and McKinnon, 1998). In this study we also describe the design of the simulation study and its results and presented the illustrative example.

2. The Robust Wilk'S Lambda Statistic

Let $x_{k1}, x_{k2}, \ldots, x_{kn_k}$ be n_k independent and identically distributed p-dimensional observations from a continuous p-variate distribution with distribution function $F_k(u)$ where $k = 1, 2, \ldots, g$ and the number of groups $g \geq 2$. If all g distributions are exactly the same but only their locations differ we have

$$F_k(u) = F(u - \mu_k)$$

Then the hypothesis we want to test is that all F_k are identical, hence

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_g$$

against the alternative hypothesis

$$H_a \colon \mu_i \neq \mu_j \text{for at least one } i \neq j$$

Under the classical assumptions that all groups arise from multivariate normal distributions, the most widely used test statistic is the Wilk's Lambda (the likelihood ratio test).

The Wilks' Lambda statistic is the ratio of the within generalized dispersion to the total generalized dispersion. The within generalized dispersion is the determinant of the within-group sums of squares and cross-products matrix **W** and the total generalized dispersion is the determinant of the total sums of squares and cross-products matrix **T** (see Johnson and Wichern, 2002). The statistic

$$\Lambda_{Wilks} = \frac{det(\mathbf{W})}{det(\mathbf{T})} \tag{1}$$

takes values between zero and one (here $det\mathbf{(A)}$ means the determinant of **A**).

In order to obtain a robust procedure with high breakdown point for inference about the means in the one-way MANOVA model we construct a robust version of the Wilk's Lambda statistic by replacing the classical estimators by the kurtosis estimators. The Kurtosis estimator is introduced by Pena and Prieto (2001) looks for a subset of h observations which optimize (maximize or minimize) the kurtosis coefficient. The method is affine equivariant, and it shows a very satisfactory practical performance, especially for large sample space dimensions and concentrated contamination. The method also produces good robust estimates for the covariance matrix, with low bias.

We start by finding estimates of the group means m_k^0 and the common covariance matrix C_{kur} based on the kurtosis estimate. The proposed estimators become

$$\begin{aligned} m_k^0 &= \frac{\sum_i^u x_{ik}}{|U|} \\ C_{kur} &= \frac{\sum_i^u (x_{ik} - m_k^0)(x_{ik} - m_k^0)'}{(|U| - 1)} \end{aligned} \tag{2}$$

where U is the set of all observations is not labeled as outliers, —U— denotes the number of observations in this set. Using this obtained estimates m_k and C_{kur} ijn (2) we can calculate the robust distances (Pena and Prieto, 2001) as

$$RD_{ik} = \sqrt{(x_{ik} - m_k^0)' C_{kur} (x_{ik} - m_k^0)} \tag{3}$$

With these robust distances we can define a weight for each observation $x_{ik}, i = 1, \ldots, n_k and k = 1, \ldots, g$ by using the weight function defined as With these weights we can calculate the final estimates, namely the group means m_k the within-groups sum of squares and cross-products matrix W_R, the between-groups sum of squares and cross-products matrix B_R and the total sum of squares and cross-products matrix $T_R = W_R + B_R$ which are necessary for constructing the robust Wilks' Lambda Λ_R statistic as defined in equation (1).

$$\begin{aligned} m_k &= \frac{\sum_{i=1}^{n_k} w_{ik}x_{ik}}{v_k} \\ W_R &= \sum_{k=1}^{g}\sum_{i=1}^{n_k} w_{ik}(x_{ik}-m_k)(x_{ik}-m_k)' \\ B_R &= \sum_{k=1}^{g} v_k(m_k-m)(m_k-m)' \\ T_R &= \sum_{k=1}^{g}\sum_{i=1}^{n_k} w_{ik}(x_{ik}-m)(x_{ik}-m)' = W_R + B_R \end{aligned} \tag{4}$$

where v_k are the sums of the weights group k for $k = 1, \ldots, g$ and v is the total sum of weights: $v_k = \sum_{i=1}^{n_k} w_{ik}$ and $v = \sum_{i=1}^{n_k} v_{ik}$. Substituting these estimates of the matrices $\mathbf{W}$ and $\mathbf{T}$ into equation (1) we obtain a robust version of the test statistic Λ given by

$$\Lambda_R = \frac{det(W_R)}{det(T_R)}. \tag{5}$$

For computing the kurtosis and related estimators the kurtosis algorithm of Pena and Prieto will be used as implemented in the Matlab software.

3. The Approximate Distribution Of Robust Wilk'S Lambda

The distribution of Λ_{Wilks} is considered by Anderson (1958) as a ratio of two Wishart distributions but it is so complicated that except for some special cases it is hardly usable in practice. One of the most popular approximations is Bartlett's χ^2 approximation given by

$$-(n-1-(p+g)/2)ln\Lambda_{Wilks} \approx \chi^2_{p(g-1)} \tag{6}$$

where $n = \sum_{i=1}^{g} n_i$. Analogously to this χ^2 approximation of the classical statistic we can assume for Λ_R the following approximation:

$$L_R = ln\Lambda_R \approx d\chi^2_q \tag{7}$$

and then express the multiplication factor d and the degrees of freedom of the χ^2distribution q through the expectation and variance of L_R

$$\begin{aligned} E[L_R] &= dq \\ Var[L_R] &= 2d^2q \\ d &= E[L_R]\frac{1}{q} \end{aligned} \tag{8}$$

$$q = 2\frac{E[L_R]^2}{Var[L_R]}$$

Since it is not possible to obtain the mean and variance of the robust Wilks' Lambda statistic Λ_R analytically, they will be determined by simulation. These values will be used to approximate the true distribution of Λ_R

For a given dimension p, number of groups g and sample sizes of each group $n_i = 1, 2, ..., g$, samples $\boldsymbol{X_k} = \boldsymbol{(x_1, ..x_n)}$ of size $n = n_i$ from the standard normal distribution will be generated, i.e.x_i $N(0,_p)$. For each sample the robust Wilk's Lambda statistic Λ_R based on the kurtosis estimator will be calculated. After performing m=3000 trials, mean and variance of Λ_R will be obtained as

$$ave(\Lambda_R) = \frac{1}{m}\sum_{k=1}^{m}\Lambda_R^k$$

$$var(\Lambda_R) = \frac{1}{m-1}\sum_{k=1}^{m}(\Lambda_R^k - ave(\Lambda_R))^2$$

Substituting these values into equation (9) we can obtain estimates for the constants d and q which in turn will be used in equation (8) to obtain the approximate distribution of the robust Wilk's Lambda statistic Λ_R.

Now we will investigate the accuracy of this approximation. For several values of the dimension p, the number of groups g and the sample sizes$n_i; i = 1, 2, ..., g$, of each group,$m = 3000$ samples from standard normal distribution will be generated and for each of them Λ_R will be calculated. The empirical distribution of these 3000 statistics will be compared to the approximate distribution given by equation (8) by $QQ - plots$, some of which are shown in Figure 1 and 2 for two groups. It is seen from the plots that the approximation is very precise for large and small sample sizes. The following are QQ - plots for the robust Wilk's Lambda statistics Λ_R for two groups and several values for p and $n = \sum n_k$.

4. Monte Carlo Simulations

In this section a Monte Carlo study is undertaken to assess the performance of the proposed statistic. The assessment of the performance of any test statistics involves two measures – the attained significance level and the power of the test. Additionally we will investigate the behavior of the robust statistic in the presence of outliers and will compare the results to the classical Wilks' Lambda statistic.

4.1 Significance Levels

First we study the attained significance level (i.e., Type I error rate or size) of the proposed robust test. We will consider several dimensions $p = 2, 4, 6, 8, 10$, numbers of

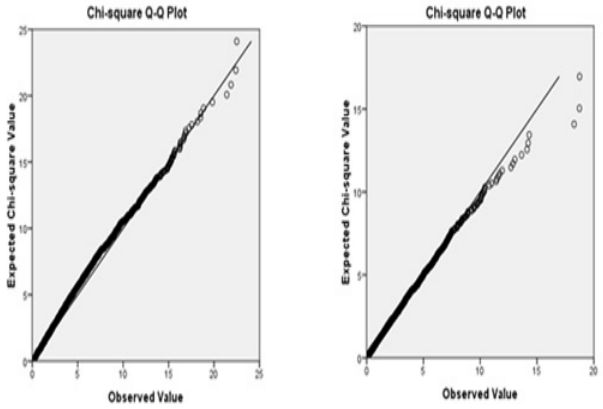


Figure 1: $p = 6, n_1 = 30, n_2 = 30$ and $p = 2, n_1 = 100, n_2 = 100$.

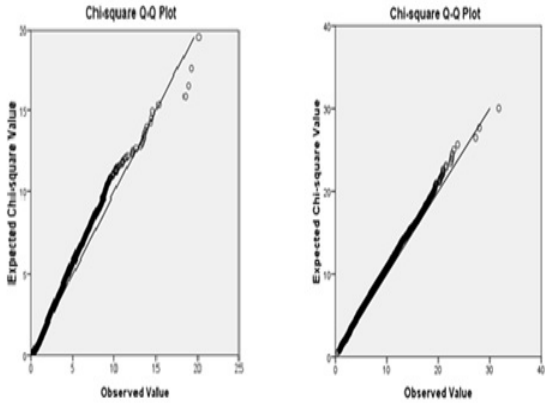


Figure 2: $p = 4, n_1 = 50, n_2 = 20$ and $p = 8, n_1 = 100, n_2 = 10$.

Table 1: Selected group sizes for the simulation study.

Two groups (n_1,n_2)	Three groups (n_1,n_2,n_3)
(10, 10)	(10, 10, 10)
(20, 20)	(20, 20, 20)
(30,30)	(30, 30, 30)
(50, 50)	(50, 50, 50)
(100, 100)	(100, 100, 100)
(200, 200)	(20, 20, 10)
(20, 10)	(30, 30, 10)
(30, 10)	(50,50,20)
(50, 10)	(100,50,20)

groups $g = 2, 3$ and sample sizes $n_k; k = 1, 2, ..., g$. Equal as well as unequal group sizes are investigated. The sample sizes for two and three groups are selected as shown in Table 1. Only the cases where, $p > 2n_i$ for all $i = 1, \ldots, g$ were considered, since otherwise the kurtosis estimate is not computable.

Under the null hypothesis H_0 in one-way MANOVA we assume that the observations come from identical multivariate distributions, i.e. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$.Since the considered statistics are affine equivariant, without loss of generality we can assume each location vector $\mu_i$to be null vector, i.e. $\mu_i = (0, \ldots, 0)^t$ and the covariance matrix to be $_p$. Thus we generate $n = \sum_{i=1}^{g} n_i$ p -variate vectors distributed as $N(0, I_p)$ and calculate the classical statistic Λ_{Wilks} and the robust version based on kurtosis estimates Λ_R.This is repeated $m = 3000$ and the percentages of values of the test statistics above the appropriate critical value of the corresponding approximate distribution are taken as an estimate of the true significance level. The classical Wilk's Lambda is compared to the Bartlett approximation given by equation (2) while the kurtosis based Wilk's Lambda is compared to the approximate distribution given in equation (3). The true significance levels α are taken to be 0.10,0.05 and 0.01 which together with the number of replications $m = 3000$.

In Table 2 the results for two groups are shown. It is clearly seen that the difference between the actual cutoff and the nominal value is very small, i.e. the approximations are capable to keep the significance levels across all investigated combinations of dimension p and sample sizes. The results for two groups are similar.

4.2 Power Comparison

In order to assess the power of the robust Wilk's Lambda statistics we will generate

Table 2: Significance levels of test statistics Λ_{wilks} and Λ_R for multivariate normal distributions in the case of two groups for several values of the dimension p and the sample size $n = n_1 + n_2$ (N=3000 Monte Carloreplications,true significance level $\alpha = 0.1, 0.05\Lambda 0.01$)

			α=0.1		α=0.05		α=0.01	
p	n_1	n_2	Λ_{wilks}	Λ_R	Λ_{wilks}	Λ_R	Λ_{wilks}	Λ_R
11	10	10	0.097	0.092	0.050	0.045	0.010	0.010
2	20	20	0.099	0.097	0.053	0.040	0.009	0.010
2	30	30	0.100	0.079	0.050	0.040	0.007	0.010
2	50	50	0.101	0.076	0.046	0.040	0.006	0.010
2	100	100	0.093	0.08	0.043	0.040	0.010	0.010
2	200	200	0.109	0.099	0.057	0.040	0.009	0.010
2	20	10	0.095	0.102	0.047	0.040	0.010	0.010
2	30	10	0.097	0.081	0.050	0.050	0.013	0.010
2	50	20	0.102	0.081	0.055	0.050	0.013	0.010
2	100	10	0.099	0.094	0.049	0.040	0.010	0.010
4	10	10	0.098	0.097	0.049	0.050	0.009	0.010
4	20	20	0.094	0.083	0.046	0.050	0.009	0.010
4	30	30	0.104	0.086	0.051	0.050	0.011	0.010
4	50	50	0.107	0.092	0.058	0.040	0.011	0.010
4	100	100	0.105	0.099	0.051	0.050	0.012	0.010
4	200	200	0.088	0.108	0.042	0.040	0.012	0.010
4	20	10	0.097	0.113	0.048	0.050	0.009	0.012
4	30	10	0.088	0.099	0.046	0.040	0.010	0.010
4	50	20	0.097	0.130	0.053	0.048	0.011	0.010
4	100	10	0.088	0.100	0.050	0.043	0.009	0.008
6	20	20	0.098	0.087	0.047	0.045	0.009	0.010
6	30	30	0.099	0.080	0.054	0.051	0.013	0.010
6	50	50	0.106	0.070	0.052	0.046	0.013	0.010
6	100	100	0.109	0.120	0.056	0.045	0.011	0.008
6	200	200	0.108	0.110	0.051	0.043	0.010	0.008
6	50	20	0.109	0.09	0.052	0.049	0.010	0.009
8	20	20	0.102	0.100	0.051	0.041	0.012	0.010
8	30	30	0.101	0.070	0.049	0.047	0.011	0.010
8	50	50	0.112	0.100	0.055	0.045	0.012	0.010
8	100	100	0.096	0.100	0.049	0.057	0.011	0.010
8	200	200	0.092	0.100	0.047	0.042	0.007	0.010
8	50	20	0.095	0.100	0.047	0.047	0.009	0.010
10	30	30	0.100	0.100	0.049	0.053	0.012	0.010
10	50	50	0.097	0.100	0.052	0.043	0.010	0.010
10	100	100	0.098	0.100	0.050	0.050	0.013	0.010
10	200	200	0.112	0.100	0.055	0.050	0.015	0.010

data under an alternative hypothesis (H_a : $notall\mu_i, i = 1, \ldots, g$) and will examine the frequency of incorrectly failing to reject H_0 (i.e. the frequency of type II errors). The same combinations of dimensions p, number of groups g and sample sizes $n_i, i = 1, \ldots, g$ as in the experiments for studying the significance levels will be used. There are infinitely many possibilities for selecting H_a but for the purpose of the study we will use the following fixed alternatives : all groups $\pi_j, j = 1, \ldots, g$ come from multivariate normal distribution with the same spherical covariance matrix I_p; the mean of the first group is the origin, the mean of the second group is at distance $d = 1$ along the first coordinate, the mean of the third group is at distance $d = 1$ along the second coordinate and so on. More precisely, the data sets are generated from the following p-dimensional normal distributions, where each group $\pi_j, j = 1, \ldots, g$, has a different mean μ_j and all of them have the same covariance matrix I_p

$$\pi_j \sim N_p(\mu_j, I_p), j = 1, \ldots, g \tag{9}$$

with $\mu_1 = (0, 0, ..., 0)^t; \mu_2 = (d, 0, ..., 0)^t; \mu_3 = (0, 0, d, ..., 0)^t; ...; \mu_g = (0, 0, ...d, 0)^t$.
The classical and the robust test statistics are computed and the rejection frequency (out of m=3000 runs) where the statistics exceeds its appropriate critical value is the estimate for the specific configuration.

The power of the two statistics can be visually compared by simulating size-power curves under fixed alternatives, as proposed by Davidson and McKinnon (1998). Constructing size-power plots does not require knowledge of the asymptotic distribution of the test statistic. For other recent applications of the size-power plots see Siani and de Peretti (2006) and Gelper and Croux (2007). The size-power curves are simulated in the following way:
i. First $m = 3000$ data sets under the null hypothesis are generated. For each of them the test statistics are computed and the obtained values are sorted in increasing order. The i^{th} value of this ordered sequence is denoted by θ_i. If the critical value is chosen as θ_i then the quantity $s_i = ((m - i))/((m + 1))$ equals the size of the test.

ii. After that m = 3000 data sets are generated under the fixed alternative hypothesis and for each of them the test statistics are computed. For a certain critical value θ_i the power of the test $f_i$is estimated by the fraction of test statistics that exceed θ_i.

iii. The pairs $(s_i, f_i), i = 1, \ldots, m$ representing the power vs. the size of the test are plotted as size-power curves.

The size-power curve should lie above the 45° line, the larger the distance between the curve and the 45° line the better. The most interesting part of the size-power curve is the region where the size ranges from zero to 0.2 since in practice a significance level above 20% is never used. In Figure 3 the size-power curves for several values of the dimension p and the sample sizes n_i in case of two groups are shown. The results for three groups

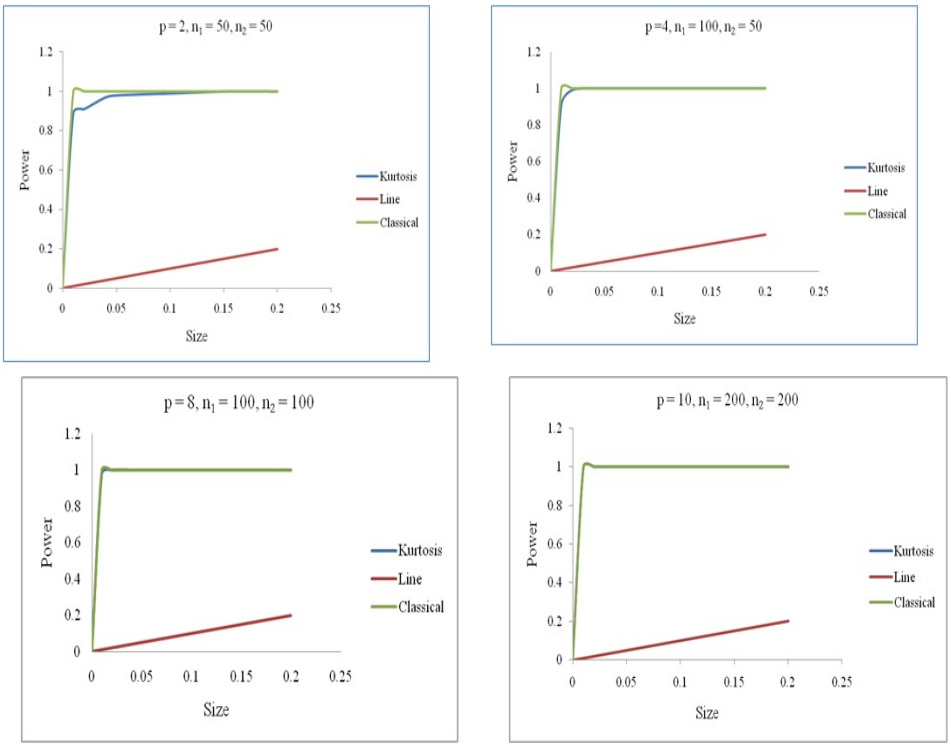


Figure 3: Size-power curves for the Wilk's Lambda statistic Λ_{Wilks}(green line) and the robust Wilk's Lambda statistic Λ_R (blue line) for two groups and several values for p and $n = \sum n_k$. The 45° line is represented by red line.

are similar. It is clearly seen that in all of the investigated combinations of dimensions p and sample sizes both curves are far above the 45° line with the line of the robust statistic being slightly below or almost equal to the classical one. Thus the loss of power for the robust statistic is acceptable throughout the investigated range of dimensions and sample sizes.

4.3 Robustness comparisons

Now we will investigate the robustness of the one-way MANOVA hypothesis test based on the proposed robust version of the Wilk's Lambda statistic Λ_R. For this purpose we will generate data sets under the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$ and will contaminate them by adding outliers. More precisely the data will be generated from the following contamination model

$$\pi_j \sim (1-\varepsilon)N_p(0, I_p) + \varepsilon N_p(\widehat{\mu_j}, 0.25^2 I_p)$$

Table 3: Oslo-transect data: Names of the lithology groups.The last column shows the number of objects in each class.

Lithological group			#
1	CAMSED	Cambro-Silurian sedimentary rocks	98
2	GNEISS_O	Precambrian gneisses - Oslo	89
3	GNEISS_R	Precambrian gneisses - Randsfjord	32
4	MAGM	Magmatic rocks of the Oslo Rift	113

$$\widehat{\mu_j} = (vQ_p, ..., vQ_p)^t where Q_p = \sqrt{\frac{\chi^2_{p,.001}}{p}},$$

where $\varepsilon = 0.1$ and$v = 5$. By adding to vQ_p each component of the outliers we guarantee a comparable shift for different dimensions p (see Rocke and Woodruff, 1996). The same combinations of dimensions p, numbers of groups g and sample sizes $n_i, i = 1, \ldots, g$, as in the experiments for studying the significance levels will be used.

Again we generate $n_i = \sum_{i=1}^{g} n_i$ p-variate vectors and calculate the classical statistic Λ_{Wilks} and the robust version based on kurtosis estimates Λ_R. This is repeated m = 3000 and the percentages of values of the test statistics above the appropriate critical value of the corresponding approximate distribution are taken as an estimate of the true significance level. We represent the results for two groups graphically in Figure 4. Similar results are obtained for three groups also.

The difference between the actual cutoff based on $\Lambda_R$and the nominal value remains acceptably small for the different combinations of dimension p and sample size $n_j, j = 1, \ldots, g$. Furthermore this difference is much smaller compared to the classical Wilk's Lambda statistic Λ_{Wilks}.

4.4 Real Life Example

We will now illustrate the application of the proposed robust statistic with the Oslo-transect data (see Reimann et al., 2007, and the references therein). Samples of different plant species were collected along a 120 km transect running through the city of Oslo, Norway, and the concentrations of 25 chemical elements for the sample materials are reported. The factors that influenced the observed element concentrations in the sample materials were investigated. For our example we will consider only the lithology as a factor. This factor has four levels which are listed in Table 3. The last column shows the number of objects in each group.

We select the variables P, K, Zn and Cu that represent elements from the group of the nutrients and expect that the lithology strongly influences the take in of the plants

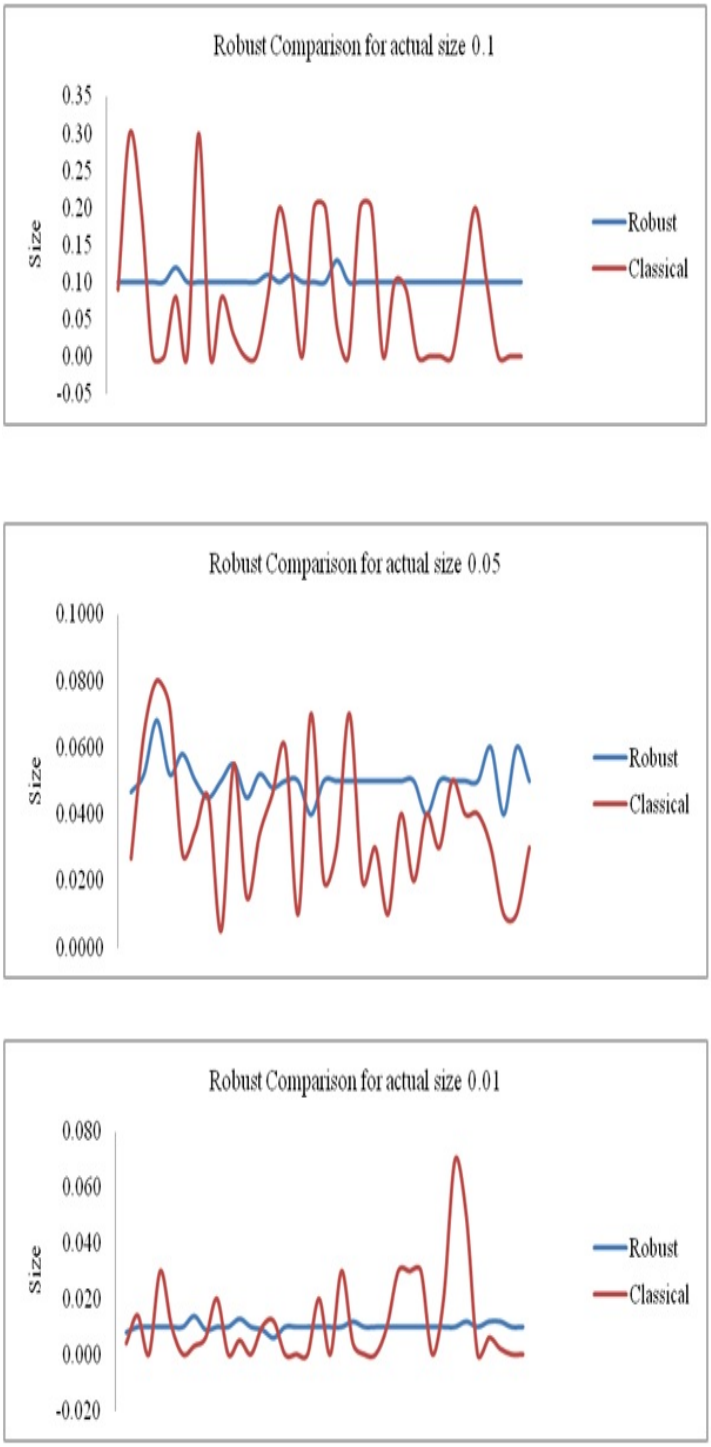


Figure 4: Robust Comparison.

when compared to the effect of the plant species themselves, i.e. we expect that the multivariate group means are significantly different. After removing the observations with missing values we remain with a data matrix of $n = 332$ rows and $p = 4$ columns. Since geo-chemical data are usually right skewed we log-transform the variables. Todorov and Filzmoser (2007) verified the non-normality and presence of outlier in the data.

Let us denote the means of the four groups by $\mu_1, \mu_2, \mu_3 and \mu_4$and perform a one-way MANOVA, testing the hypothesis $H_0 \colon \mu_1 = \mu_2 = \mu_3 = \mu_4$. The classical Wilk's Lambda statistic for this data set yields$\Lambda_{Wilks} = 0.9603$ which corresponds to a p-value of 0.3510. This suggests that the hypothesis of equal means cannot be rejected at the 10% level of significance. On the other hand the robust Wilk's Lambda statistic yields $\Lambda_R = 0.8961$ which corresponds to a p-value of 0.001495 and we can reject the null hypothesis even at the 1% level of significance.

5. Conclusion

Multivariate analysis of variance (MANOVA) is an extension of analysis of variance (ANOVA) to accommodate more than one dependent variable. The Wilks' Lambda Statistic (likelihood ratio test, LRT) is a commonly used tool for inference about the mean vectors of several multivariate normal populations. As classical estimates of population mean and dispersion are highly sensitive to outliers, classical Wilk's Lambda and classical one-way MANOVA are highly affected by outliers. Consequently classical MANOVA gives misleading results when the data contains contaminated observations. This leads to think on the robust version of MANOVA, which looks for a robust Wilk's Lambda based on highly robust and efficient estimates of population mean and dispersion matrices. There are a number of robust estimates discussed in literature and in this study we make use of a highly robust and efficient estimates introduced by Pena and Prieto (2001) to compute a robust Wilk's Lambda. The approximate distribution of the proposed robust statistic was derived using simulations and its fitting is examined using QQ – plots.

Further, Monte Carlo simulations were used to investigate the efficiencies of the proposed method and the level of significance and power of the proposed robust MANOVA are compared with that of classical MANOVA. The results of simulation study says that the estimated significance level in uncontaminated data using this robust method is approximately equal to the actual size, especially in high dimensional data sets. The size -power curve proposed by Davidson and McKinnon used here to compare the power of the proposed robust method with the classical one. Curve shows that the calculated power of the robust method is slightly below those obtained in classical. In most cases it is equal to the classical one. That means there is only an acceptable level of power loss is occurred for the proposed method in uncontaminated data sets, which shows the advantage of the robust test over classical test. In contaminated situation the robust method is more suitable

than the classical one.

Additionally, the proposed robust test is applied and investigated its performance in a real life benchmark data (Oslo-transect data). The result shows that the proposed Robust MANOVA is much less affected by the presence of outliers compared to the classical MANOVA. That is, the proposed robust MANOVA technique performs well under contaminated as well as un-contaminated situations and is a better alternative to classical test for the data containing contaminated observations. Thus we can apply this proposed Robust MANOVA in real life data sets, as they should necessarily have outliers.

References

[1] Anderson, T.W. (1958). *An Introduction Multivariate Statistical Analysis*, John Wiley & Sons Inc, New York.

[2] Davidson,R. and McKinnon, J. (1998.) Graphical methods for investigating the size and power of hypothesis tests. *Manchester Schoo*l, 66, 1-26.

[3] Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall International.

[4] D. Pena and F.J. Prieto,(2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation, *Technometrics*, Vol. 43, pp. 286-300.

[5] Reimann, C., Arnoldussen, A., Boyd, R., Finne, T.E., Koller, F., Nordgullen, O. and Englmair, P. (2007). Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients, *Science of the Total Environment*, 377, 416-433.

[6] Siani, C. and de Peretti, C. (2007). Analysing the performance of bootstrap neural tests for conditional heteroscedasticity in ARCH-M models. *Computational Statistics and Data Analysis*, 51(5), 2442-2460.

[7] Todorov, V. (2007). Robust selection of variables in linear discriminant analysis. *Statistical Methods and Applications*, 15, 395-407.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 207-216.

Generalized Gamma Model and Pathway Fractional Operator

SEEMA S. NAIR
Department of Statistics, St Gregorios College, Kottarakkara
Pulamon P.O., Kollam, Kerala 691 531, India
e-mail: seema.cms@gmail.com

Abstract

The object of this paper is to introduce a new generalized probability distribution associated with Mittag-Leffler function,which directly connect to the theory of fractional calculus. It gives an extension to the gamma type distribution. Various properties of this new distribution are investigated. Availability of probability models with thicker or thinner tails through this new density is also illustrated in this paper. Connection to fractional calculus is also established here.

Key words: Bayesian analysis, Gamma density, Mittag-Leffler function.

1. Introduction

In reaction rate theory, input-output type situations and reaction-diffusion problems in physics and chemistry, when the total derivatives are replaced by fractional derivatives the solutions automatically go in terms of Mittag-Leffler functions and their generalizations, see Haubold and Mathai (2000). The ordinary and generalized Mittag-Leffler functions interpolate between a purely exponential law and power-law like behavior of phenomena governed by ordinary kinetic equations and their fractional counterparts, see Kilbas et al. (2004), Kiryakova (2000), Mathai (2010) and Mathai et al. (2010). Among the various results presented by various researchers, the important ones deal with Laplace transform and asymptotic expansions of this function.

By the application of Laplace integral, it follows that

$$\int_0^\infty \mathrm{e}^{-ax} x^{\beta-1} E_{\alpha,\beta}(-\delta x^\alpha)\mathrm{d}x = \frac{a^{\alpha-\beta}}{a^\alpha+\delta}, \tag{1}$$

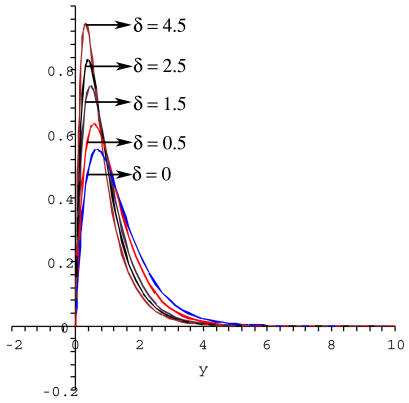
where $\Re(\alpha) > 0,\ \Re(\beta) > 0$. It is shown to be very relevant in fractional reaction-diffusion problems in physics, since it naturally occurs in the derivation of the inverse Laplace transform of the functions of the type $a^{\alpha}(d + ba^{\beta})$, where a is the Laplace transform parameter and d and b are constants. A class of Laplace transforms is examined in Mathai et al. (2006), to show that particular cases of this class are associated with production-destruction and reaction-diffusion problems in physics, study of differences of independently distributed random variables and the concept of Laplacianness in statistics, α-Laplace and Mittag-Leffler stochastic processes, the concepts of infinite divisibility and geometric infinite divisibility problems in probability theory and certain fractional integrals and fractional derivatives. A number of applications are pointed out with special reference to solutions of fractional reaction and reaction-diffusion equations and their generalizations.

Motivated from the integral in (1), we think of a statistical model out of this and it is observed that the new model will provide more flexible than the generalized gamma density. Recently, probability models with thicker or thinner tails have got more importance among the statisticians and physicists because of their vast applications in random walks, Lévi flights, financial modeling etc. In many modeling problems like solar modeling, growth-decay modeling, it is often found that the selected model is not a good fit for the experimental data because it requires a model with thicker or thinner tail than the ones available from the parametric family of distributions. In this paper we considered a new family of generalized probability distributions associated with Mittag-Leffler function (a thicker or thinner tailed model associated with gamma density) which can also be obtained as a limiting case of pathway fractional integral operator (for more details about pathway fractional integral operator one can look at the papers of Seema Nair [(2009), (2011)]). This family gives an extension to the generalized gamma family and opens up vast area of potential applications and establish connections to the current popular topics of nonextensive statistical mechanics, Tsallis statistics, superstatistics and Mittag-Leffler stochastic process, Lévi process and time series. In addition to this, the new model will then connect to fractional calculus and statistical distribution theory through the theory of special functions.

The generalized gamma Mittag-Leffler distribution is specified by the probability density of the form

$$f(x) = \begin{cases} Cx^{\beta-1}\mathrm{e}^{-ax}E_{\alpha,\beta}(-\delta x^{\alpha}), & \Re(\beta) > 0,\ \Re(\alpha) > 0,\ a > 0,\ x > 0 \\ 0, \text{ elsewhere} \end{cases} \tag{2}$$

where C is the normalizing constant and can be evaluated from the integral in (1) such that $C^{-1} = \frac{a^{\alpha-\beta}}{a^{\alpha}+\delta}$. For fixed values of a, β and for various values of δ, we can look at the graphs which give a suitable interpretation to the model in (2).



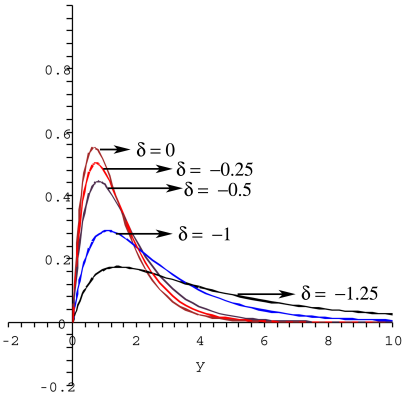


Figure 1: (a) For $a = 1.5,\ \beta = 2$ and $\delta > 0$ (b) For $a = 1.5,\ \beta = 2$ and $\delta < 0$

The following figures show a comparison between gamma density and gamma Mittag-Leffler density for different values of δ. Observe that $\delta = 0$ corresponds to the gamma density. In Figure 1 (a), $\delta = 0$ corresponds to the gamma density. When the values of δ increases from $\delta = 0$, the right tail of the new density becomes thinner and thinner compared to that of a gamma density. Similarly peakedness of the curve slowly increases. In Figure 1 (b), as the value of δ decreases the right tail of the new density becomes thicker and thicker compared to that of a gamma density. Similarly the peakedness of the curve slowly decreases. Hence when we look for a model with thicker or thinner tail while a gamma density is found to be more or less proper fit then a member from the new family of densities introduced here will become quite useful and handy to model which deviate from a gamma type model. Observe that the new density is mathematically and computationally tractable easily just like a gamma density.It may be observed that the distribution function is available in terms of series of incomplete gamma functions, given as

$$F(x) = C \sum_{k=0}^{\infty} M_k\ \gamma(\alpha k + \beta, x), \quad \alpha > 0,\ \beta > 0, \tag{3}$$

where $\gamma(\alpha k+\beta, x)$ is the incomplete gamma function and $M_k = \frac{(-\delta)^k}{\Gamma(\alpha k+\beta)}$. Similarly survival function can be evaluated by $\bar{F}(x) = 1 - F(x)$, where $F(x)$ is given in (3).
The moment generating function of (2) is given by

$$M_x(t) = \frac{b^\beta}{(b-t)^\beta} \frac{(1 - \frac{\delta}{b^\alpha})}{(1 - \frac{\delta}{(b-t)^\alpha})}, \quad |\frac{\delta}{(b-t)^\alpha}| < 1,\ (b-t) > 0. \tag{4}$$

The characteristic function can be obtained if we replace t by it, $i = \sqrt{-1}$. If we put $-t$ instead of t then we will obtain the Laplace transform. Using the Laplace transform or (moment generating function) we can easily obtain the integer moments by using the following formula:

$$\mu'_r = E(x^r) = (-1)^r \ \frac{\mathrm{d}^r}{\mathrm{d}t^r} L_f(t)|_{t=0}.$$

Thus the mean value will be obtained as

$$\mu'_1 = \frac{\beta}{b} + \frac{\delta \ \alpha}{b(b^\alpha + \delta)}. \tag{5}$$

and

$$\mu'_2 = \frac{1}{b^2}\left\{2\alpha^2(\frac{\delta}{b^\alpha+\delta})^2 + (\alpha^2 + \alpha(2\beta+1))(\frac{\delta}{b^\alpha+\delta}) + \beta(\beta+1)\right\}. \tag{6}$$

Variance

$$\mu_2 = \frac{\beta}{b^2} + \frac{\delta\alpha[(\alpha+1)b^\alpha - \delta]}{b^2(b^\alpha+\delta)^2}.$$

Arbitrary moments of the distribution can be obtained in terms of generalized Wright hyper-geometric function. That is

$$\mu'_\gamma = E(x^\gamma) = \int_0^\infty x^\gamma \ f(x)\mathrm{d}x$$

which is nothing but the Mellin transform of the function f with $\gamma = s - 1$.

$$\mu'_\gamma = \frac{(1 - \frac{\delta}{b^\alpha})}{b^\gamma} \ {}_2\Psi_1\left[\begin{matrix}(1,1),(\beta+\gamma,\alpha)\\(\beta,\alpha)\end{matrix}|\frac{\delta}{b^\alpha}\right] \tag{7}$$

where ${}_p\Psi_q(z)$ is the generalized Wright's hypergeometric function defined for $z \in C$, complex $a_i, b_j \in C$ and $\alpha_i, \beta_j \in \Re_+ = (0, \infty)$, $\alpha_i, \beta_j \neq 0; i = 1, 2, \ldots, p; j = 1, 2, \ldots, q$ by the series

$${}_p\Psi_q(z) \equiv \ {}_p\Psi_q\left[\begin{matrix}(a_i,\alpha_i)1,p\\(b_j,\beta_j)1,q\end{matrix}|z\right] = \sum_{k=0}^{\infty} \frac{\{\prod_{i=1}^p \Gamma(a_i + \alpha_i k)\} \ z^k}{\{\prod_{j=1}^q \Gamma(b_j + \beta_j k)\}k!}. \tag{8}$$

The function in (8) was introduced by Wright and is called the generalized Wright's hypergeometric function. For convergence conditions, existence of various contours and other properties see Wright [**??**] or from the theory of H-function to be discussed later. If we take γ as integers in (7) then we will obtain integer moments.

2. Estimation of parameter

In this section we have given explicit forms of the estimators of the parameters using method of moments. The method of moments is the process of equating sample and population moments and trying to solve for the unknown parameters. The motivation comes

from the fact that the sample moments are consistent estimators for the corresponding population moments. To start with, let us consider the case for $b = 1,\ \beta = 1$ in (2), then the model will become the Mittag-Leffler extension of standard exponential distribution, and has the density of the form

$$g_1(x) = (1+\delta)\mathrm{e}^{-x}E_\alpha(-\delta x^\alpha). \tag{9}$$

The moments can be obtained from (5) and (6) with the parameter value $a = 1,\ \beta = 1$.

The moment estimators of δ and α are given by

$$\hat{\alpha} = \frac{\sum(x_i-\bar{x})^2 - n\bar{x}^2}{n(\bar{x}-1)} + 1$$

and

$$\hat{\delta} = \frac{n(\bar{x}-1)^2}{\sum(x_i-\bar{x})^2 - n\bar{x}}.$$

Now, consider the Mittag-Leffler extension of standard gamma density. For that we take $b = 1$ in (2), thus the model has the following form

$$g_2(x) = (1-\delta)x^{\beta-1}\mathrm{e}^{-x}E_{\alpha,\beta}(\delta x^\alpha). \tag{10}$$

Using the same procedure as above one can obtain the estimators as

$$\hat{\beta} = \frac{\sum(x_i-\bar{x})^3 - [\sum(x_i-\bar{x})^2][\sum x_i^2 + n\bar{x}] + 2n\bar{x}[n\bar{x}^3 - \bar{x}^2]}{\sum x_i^2[4n\bar{x} - 3n - 3] + n\bar{x}^2[8n\bar{x} + 13n] + n^2\bar{x}}$$

$$\hat{\alpha} = \frac{\sum(x_i-\bar{x})^2 - n(\bar{x}-\hat{\beta})^2 - n\bar{x}}{n(\bar{x}-\hat{\beta})}$$

and

$$\hat{\delta} = \frac{(\bar{x}-\hat{\beta})}{\hat{\alpha}+\bar{x}-\hat{\beta}}.$$

If we retain all parameters $\alpha,\ \beta,\ \delta$ and a, then the analytical solution is quite difficult but the numerical solution can be obtained by using the softwares like MATLAB and Maple.

3. A Bayesian Approach to Generalized Superstatistics in Statistical Mechanics

Suppose $x \sim$ generalized gamma having density function

$$f(x|\theta) = \frac{\theta^\gamma}{\Gamma(\gamma)}\mathrm{e}^{-\theta x^\rho}x^{\gamma-1},\quad \theta > 0,\ \gamma > 0. \tag{11}$$

Assume a prior distribution for the parameter θ of the type in (2). That means $\theta \sim gammaML(\alpha, \beta, \delta)$. Now the unconditional density $g(x)$ can be obtained from the following computations.

$$\begin{aligned} g(x) &= \int_\theta f(x|\theta,\gamma)h(\theta)\mathrm{d}\theta \\ &= \frac{a^\beta(1+\frac{\delta}{a^\alpha})}{\Gamma(\gamma)}x^{\gamma-1}\int_{\theta=0}^{\infty}\theta^{\beta+\gamma-1}\mathrm{e}^{-\theta(x^\rho+a)}E_{\alpha,\beta}(-\delta\theta^\alpha)\mathrm{d}\theta \\ &= \frac{a^\beta(1+\frac{\delta}{a^\alpha})}{\Gamma(\gamma)}x^{\gamma-1}(a+x^\rho)^{-(\gamma+\beta)}{}_2\psi_1\left[\begin{matrix}(1,1),(\alpha,\beta+\gamma)\\(\alpha,\beta)\end{matrix}\middle|-\frac{\delta}{(a+x^\rho)^\alpha}\right] \end{aligned} \quad (12)$$

which is the generalized superstatistics associated with Wright function. In particular for $\alpha = 1$ we can obtain the unconditional density as

$$g(x) = \frac{(1+\frac{\delta}{a})}{a^\gamma B(\gamma,\beta)}x^{\gamma-1}(1+\frac{x^\rho}{a})^{-(\gamma+\beta)}\ {}_2F_1(\gamma+\beta,\beta,-\frac{\delta}{a+x^\rho}),\quad |\frac{\delta}{a+x^\rho}|<1, \quad (13)$$

the generalized superstatistics associated with Gauss hypergeometric function. For $\delta = 0$, It will become the superstatistics of Beck and Cohen (2003). In a physical system the parameter θ in (11) may represent the temperature so that the density $f(x|\theta)$ may represent the production of the item x at a fixed temperature θ. Then the marginal density of θ in (2) may represent the temperature distribution. Then the unconditional density of x is the distribution of the production of x over all temperature variations or averaged over the density of θ. Since a density in (2) is superimposed over the density in (11), the resulting density in (2) is called generalized superstatistics.

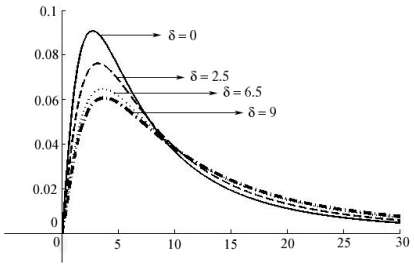


Figure 3: For $a = 5.5,\ \beta = 2,$ $\gamma = 2.5$ and $\delta > 0$

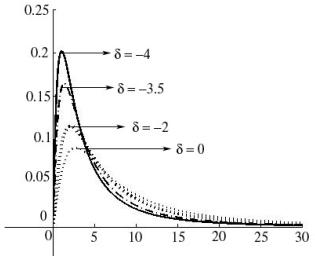


Figure 4: For $a = 5.5,\ \beta = 2,$ $\gamma = 2.5$ and $\delta < 0$

The Bayes' density/posterior density of θ can be obtained from the following formula

$$g_1(\theta|x) = \frac{f(x|\theta)h(a)}{\int h(a)f(x|\theta)\mathrm{d}a}. \quad (14)$$

$$g_1(\theta|x) = K\ \theta^{\beta+\gamma-1}\mathrm{e}^{-(a+x^\rho)\theta}E_{\alpha,\beta}(-\delta\theta^\alpha), \quad (15)$$

where $K^{-1} = (a+x^{\rho})^{-(\beta+\gamma)}{}_2\psi_1\left[\begin{matrix}(1,1),(\alpha,\beta+\gamma)\\(\alpha,\beta)\end{matrix}\middle| -\frac{\delta}{(a+x^{\rho})^{\alpha}}\right]$ and we can see that the posterior density can be obtained in the same format of the prior density.

4. Generalized Gamma Mittag-Leffler Model as a Limiting Case of Pathway Fractional Integral Operator

Recently, Seema Nair (2009) introduced new fractional integral operator called pathway fractional integral operator by means of pathway model of Mathai (2005). For the pathway parameter $q < 1$, the definition of pathway operator is as follows:

Let $f(x) \in L(a,b),\ \eta \in C,\ \ \Re(\eta) > 0,\ a > 0,\ x > 0$ and let us take the "pathway parameter" $q < 1$. Then the pathway fractional integration operator (pathway operator) is defined as

$$(P_{0+}^{(\eta,q)}f)(x) = x^{\eta}\int_0^{[\frac{x}{a(1-q)}]}[1-\frac{a(1-q)t}{x}]^{\frac{\eta}{(1-q)}}f(t)\mathrm{d}t, \tag{16}$$

where $f(t)$ is an arbitrary function. If we restricted $f(t)$ as any real-valued positive integrable scalar function of t, one can bring out a statistical model from the pathway fractional integral operator, see Seema Nair [(2011), (2009)]. Also for any positive integrable function $f(t)$, we can make interpretation to pathway fractional operator as the density of a sum of independently distributed positive random variables (Laplace convolution of two independently distributed positive random variables). To facilitate a connection to gamma Mittag-leffler density via pathway operator, let us consider the integrand in (16). From this one can obtain the gamma Mittag-Leffler density as a limiting case. That is when $q \to 1_-$, replace $f(t)$ by $t^{\beta-1}E_{\alpha,\beta}(-\delta t^{\alpha})$, then the integrand will coincide with the model in (2). Thus we can see that the new model connects fractional calculus and statistical distribution theory through the theory of special functions.

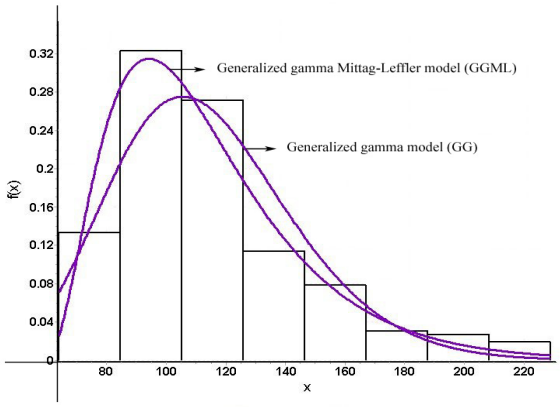
5. Application in Economic Modeling

In this section we present application of generalized gamma Mittag-Leffler density in modeling Production and Sales dataset, which is a subset of the Main Economic Indicators (MEI) database which contains predominantly monthly statistics, and associated statistical methodological information, for the 34 OECD (Organization for Economic Co-operation and Development) member countries and for selected non-member economies: Brazil; China; India; Indonesia; Russian Federation; and South Africa. The MEI database contains a wide variety statistics that can be classified as Short-Term Economic Statistics.

The Production and Sales dataset contains industrial statistics on four separate subjects: Production; Sales; Orders; and Work started. The data series presented within these

subjects have been chosen as the most relevant industrial statistics in the MEI database for which comparable data across countries is available. Most data are available monthly and are presented as an index (where the year 2005 is the base year) or as a level depending on which measure is seen as the most appropriate and/or useful in the economic analysis context. Due to differences in statistical or economic environment at country level, however, availability of data varies from one country to another.

Here we consider the monthly production of total industry for the past few years. The data are collected from the official website of Main Economic Indicators http:// dx.doi.org/10.1787/mei-data-en. The following is the graph of the data embedded with the generalized gamma Mittag-Leffler density and gamma density. With the help of the softwares MATLAB and Maple we have completed the analysis. We are not specifying any parameters here to plot the density function and the same program produced the two different graphs as shown below. We have calculated Kolmogorov-Smirnov statistic for both the models and it is observed that the maximum distance measure for our new model is less than the same for the generalized gamma model. That is, the calculated distances for generalized gamma Mittag-Leffler and generalized gamma are 0.05818 and 0.08921 respectively and the corresponding tabled value is 0.08521. With this arguments we can conclude that the new model in (2) gives a better fit to the data set considered.



The histogram embedded with GGML and GG models

References

[1] A. Erdélyi, On fractional integration and its application to the theory of Hankal transforms. *Quar. J. Math., Oxford Series,* **2**,44 (1940), 293-303.

[2] A. Erdélyi, On fractional integration and its application to the theory of Hankal transforms. *Quar. J. Math., Oxford Series,* **2**,44 (1940), 293-303.

[3] R.C. Griffiths, Characterization of infinitely divisible multivariate gamma distributions, Journal of Multivariate Analysis, 15 (1984) 13-20.

[4] A.A. Kilbas, M. Saigo and R.K. Saxena, Generalized Mittag-Leffler function and generalized fractional calculus operators. *Integral Transform. Spec. Funct.* **15**, (2004), 31-49.

[5] A.A. Kilbas, H.M. Srivastava and J.J. Trujillo, *Theory and Applications of Fractional Differential Equations*. Elsevier, Amsterdam (2006).

[6] H. Kober, On fractional integrals and derivatives. *Quar. J. Math., Oxford Series* **II**, (1940), 193-211.

[7] V. Kiryakova, Multiple (Multiindex) Mittag-Leffler functions, relations to generalized fractional calculus. *Fract. Calc. Appl. Anal.* **2**, No 4 (1999), 445-462.

[8] V. Kiryakova, Multiindex Mittag-Leffler functions, related Gelfond-Leontiev operators and Laplace type transforms. *J. Comput. Appl. Mathematics* **118**, (2000), 241-259.

[9] A.C. McBride, *Fractional Calculus and Integral Transforms of Generalized Functions*, Pitman, London (1979).

[10] O.I. Marichev, *Handbook of Integral Transforms of Higher Trascendential Functions. Theory and Algorithmic Tables*. Ellis Horwood, Chichester; Wiley, New York (1983).

[11] A.M. Mathai, A pathway to matrix-variate gamma and normal densities. *Linear Algebra and Its Applications* **396**, (2005) 317-328.

[12] A.M. Mathai, Fractional integrals in the matrix-variate cases and connection to statistical distributions. *Integral Transforms and Special Functions* **20(12)**, (2009) 871-882.

[13] A.M. Mathai, R.K. Saxena and H.J. Haubold, A Certain Class of Laplace Transforms with Applications to Reaction and Reaction-Diffusion Equations. *Astrophysics and Space Science* **305**, (2006) 283-288.

[14] R. Metzler and J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Phys. Rep.* **339**, (2000) 1-77.

[15] M. Saigo, A remark on integral operators involving the Gauss hypergeometric functions. *Math. Rep. Kyushu Univ.* **11**, (1978), 135-143.

[16] R.K. Saxena, A.M. Mathai and H.J. Haubold, Unified fractional kinetic equation and a fractional diffusion equation. *Astrophysics and Space Science* **209**, (2004) 299-310.

[17] Seema S. Nair, Pathway fractinal integration operator. *fractional Calculus and Applied Analysis* **12(3)**, (2009) 237-252.

[18] Seema S. Nair, Pathway fractional integral operator and matrix-variate functions. *Integral Transforms and Special Functions* **22(3)**, (2011) 233-244.

[19] E.M. Wright, The generalized Bessel functions of order greater than one. *Quart. J. Math. Oxford Ser.* **11**, (1940) 36-48.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 217-232.

Three Parameter Quasi Lindley Distribution

Ashlin Mathew P.M. and Sneha Robinson
Department of Statistics, St.Thomas College, Thrissur, Kerala, India
Email:ashlinmathewpm@gmail.com, sneharobinson97@gmail.com

Abstract

A Three parameter Quasi Lindley distribution (TQLD), of which the Lindley distribution (LD) is a particular case, has been introduced. Its moments, failure rate function, mean residual life function and stochastic orderings have been discussed. It is found that the expressions for failure rate function, mean residual life function, and stochastic orderings of the TQLD shows its flexibility over Lindley distribution and Exponential distribution. The maximum likelihood method has been discussed for estimating its parameters. The distribution has been fitted to data-set represents the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli to test its goodness of fit to which earlier the Three parameter Lindley distribution has been fitted by others and it is found that to almost the TQLD provides closer fits than those by the Three Parameter Lindley distribution.

Key words: Quasi Lindley distribution, moments, failure rate function, mean residual life function, stochastic ordering, estimation of parameters, goodness of fit.

1. Introduction

Lindley [8] introduced a One-parameter distribution, known as Lindley distribution, given by its probability density function(p.d.f)

$$f(x;\theta) = \frac{\theta^2}{1+\theta}(1+x)e^{-\theta x} \quad ; x>0, \theta>0 \tag{1}$$

It can be seen that this distribution is a mixture of Exponential (θ) and Gamma $(2,\theta)$ distributions. Its cumulative distribution functions has been obtained as,

$$F(x) = 1 - \frac{\theta+1+\theta x}{\theta+1}e^{-\theta x} \; ; x>0, \theta>0 \tag{2}$$

Ghitany et al[6] have discussed various properties of this distribution and showed that in many ways (1) provides a better model for some applications than the Exponential distribution. Ghitany et al[6] studied the various properties of this distribution. A discrete version of this distribution has been suggested by Deniz and Ojeda[7] having its applications in count data related to insurance. Sankaran[10] obtained the Lindley mixture of Poisson distribution. Mazucheli and Achar[9], Ghitany et al[4], Ghitany et al[5] and bakouchi et al[1] are some among others who discussed its various applications. Shanker and Mishra[12] introduced a Quasi Lindley distribution (QLD) with parameters α and θ defined by its probability density function (p.d.f)

$$f(x;\alpha,\theta)=\frac{\theta(\alpha+x\theta)}{\alpha+1}e^{-\theta x} \quad ;x>0,\theta>0,\alpha>-1 \tag{3}$$

It can easily be seen that at $\alpha=\theta$, the QLD (3) reduces to the Lindley distribution (1). Shanker and Mishra[12] have discussed its various properties and showed that this QLD is a better model than the Lindley distribution for modeling waiting and survival times data.

A new two parameter Quasi Lindley distribution of Shanker and Amanuel Habte Ghebretsadik[11] with parameters θ and α is defined by its probability density function (p.d.f)

$$f(x;\theta,\alpha)=\frac{\theta^2}{\theta^2+\alpha}(\theta+\alpha x)e^{-\theta x} \quad ;x>0,\theta>0,\alpha<-\theta^2 \tag{4}$$

The cumulative distribution function (c.d.f.) of the New QLD is obtained as

$$F(x)=1-\frac{\theta^2+\alpha+\theta\alpha x}{\theta^2+\alpha}e^{-\theta x};x>0,\theta>0,\alpha<-\theta^2 \tag{5}$$

The first four moments about origin of the New QLD obtained are

$$\mu_1'=\frac{\theta^2+2\alpha}{\theta(\theta^2+\alpha)},\mu_2'=\frac{2(\theta^2+3\alpha)}{\theta^2(\theta^2+\alpha)},\mu_3'=\frac{6(\theta^2+4\alpha)}{\theta^3(\theta^2+\alpha)},\mu_4'=\frac{24(\theta^2+5\alpha)}{\theta^4(\theta^2+\alpha)} \tag{6}$$

It can easily be verified that for $\alpha=\theta$, the moments about origin of the new QLD reduce to the respective moments of the Lindley distribution.
The central moments of the New QLD have thus been obtained as

$$\mu_2=\frac{\theta^4+4\theta^2\alpha+2\alpha^2}{\theta^2(\theta^2+\alpha)^2}, \qquad \mu_3=\frac{2(\theta^6+6\theta^4\alpha+6\theta^2\alpha^2+2\alpha^3)}{\theta^3(\theta^2+\alpha)^3}$$

,

$$\mu_4=\frac{3(3\theta^8+24\theta^6\alpha+44\theta^4\alpha^2+32\theta^2\alpha^3+8\alpha^4)}{\theta^4(\theta^2+\alpha)^4} \tag{7}$$

Shanker[13] has detailed study about Three Parameter Generalised Lindley distributionb (TPGLD) and obtained expressions for coefficient of variation, skewness, kurtosis,

index of dispersion, hazard rate function and the mean residual life function.
Shanker[13] has detailed comparative study of TPGLD and Three-Parameter Generalized Gamma distribution (TPGGD) and observed that in most of the data sets from medical science and engineering TPGGD gives better fit than TPGLD.

The p.d.f and c.d.f of a Three Parameter Lindley distribution introduced by Rama Shanker, Kamlesh Kumar Shukla, Ravi Shanker and Tekie Asehun Leonida[14] are given by

$$f(x;\theta,\alpha,\beta) = \frac{\theta^2}{\theta\alpha+\beta}(\alpha+\beta x)e^{-\theta x} \quad ; x>0, \theta>0, \beta>0, \theta\alpha+\beta>0 \tag{8}$$

$$F(x;\theta,\alpha,\beta) = 10[1+\frac{\theta\beta x}{\theta\alpha+\beta}]e^{-\theta x} \quad ; x>0, \theta>0, \beta>0, \theta\alpha+\beta>0 \tag{9}$$

There are many situations where these distributions are not suitable for modeling lifetime data from theoretical or applied point of view. Therefore, an attempt has been made here to obtain a new distribution which is flexible than these lifetime distributions for modelling lifetime data in reliability and in terms of hazard rate shapes.

2. Three Parameter Quasi Lindley Distribution

Three Parameter Quasi Lindley distribution (TQLD) with parameters θ, α and β is defined by its probability density function (p.d.f)

$$f(x;\theta,\alpha,\beta) = \frac{\theta(\alpha+\beta\theta x)e^{-\theta x}}{\alpha+\beta} \quad ; x>0, \theta>0, \alpha>0, \alpha+\beta>0 \tag{10}$$

here we can conclude that α and β are scale parameters and θ is shape parameter.

it can easily be seen that

- at $\beta = 1$ the Three Parameter Quasi Lindley distribution reduces to the Quasi Lindley distribution.
- at $\beta = 1$ and $\alpha = \theta$, it reduces to the Lindley distribution (1).
- at $\beta = 1$ and $\alpha = 0$ it reduces to Beta(2,θ).

The p.d.f (10) can be shown as a mixture of Gamma (2,θ) and Exponential (θ) distributions.

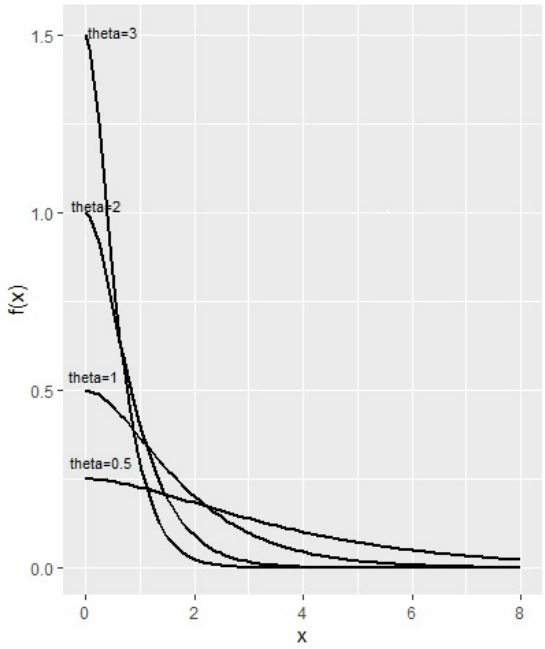


Figure 1: Density curves of TQLD with $\alpha = 1$ and $\beta = 1$

The cumulative distribution function (c.d.f) of the TQLD is obtained as

$$\begin{aligned} F(x) &= \int_0^x \frac{\theta\,(\beta\theta t + \alpha)\,\mathrm{e}^{-\theta t}}{\beta + \alpha} dt \\ &= 1 - \frac{(\beta\theta x + \alpha + \beta)e^{-\theta x}}{\alpha + \beta} \quad ; x > 0, \theta > 0, \alpha > 0, \alpha + \beta > 0 \end{aligned} \tag{11}$$

The first derivative of (10) is

$$f'(x) = \frac{\theta^2}{\alpha + \beta}(\beta(1 - x\theta) - \alpha)e^{-\theta x}$$

and $f'(x) = 0$ gives $x = \dfrac{\beta - \alpha}{\beta\theta}$. From this it follows that

for $\beta > \alpha$, $x_0 = \dfrac{\beta - \alpha}{\beta\theta}$ is the unique critical point at which f(x) is maximum. if $\beta \leq \alpha$, $f'(x) \leq 0$, i.e f(x) is decreasing in x.

Therefore, the mode of the Three Parameter Quasi Lindley distribution (TQLD) is given by

$$Mode = \begin{cases} \dfrac{\beta - \alpha}{\beta\theta} & , \beta > \alpha \\ 0 & , otherwise \end{cases}$$

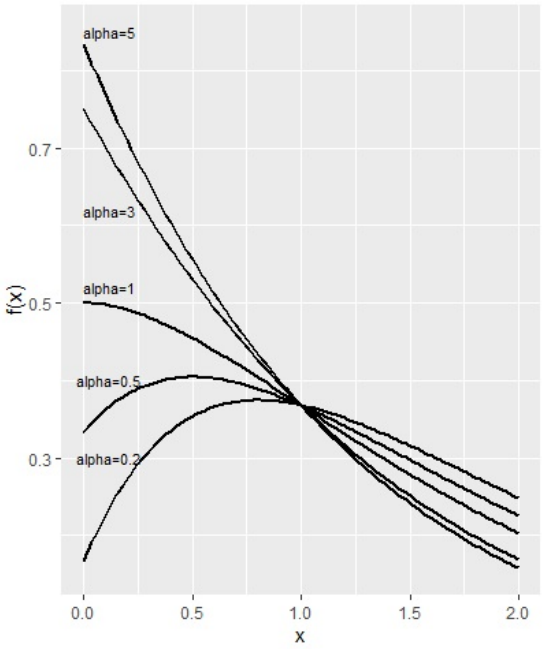


Figure 2: Density curves of TQLD with $\theta = 1$ and $\beta = 1$.

2.1 Moments and some related measures

The r^{th} moment about origin of the Three Parameter Quasi Lindley distribution (TQLD) (10) has been obtained as

$$\mu'_r = \frac{\alpha\Gamma(r+1) + \beta\Gamma(r+2)}{\theta^r(\alpha+\beta)}; r = 1, 2, 3, ... \tag{12}$$

It can easily be verified that for $\beta = 1$ and $\alpha = \theta$, the moments about origin of the Three Parameter Quasi Lindley distribution reduces to the respective moments of the Lindley distribution.

The central moments of TQLD have been obtained as

$$\mu_2 = \frac{\alpha^2 + 2\beta(2\alpha+\beta)}{\theta^2(\alpha+\beta)^2} \tag{13}$$

$$\mu_3 = \frac{18\beta^3 + 37\alpha\beta^2 + 30\beta\alpha^2 + 2(\alpha+2\beta)^3}{\theta^3(\alpha+\beta)^3} \tag{14}$$

$$\mu_4 = \frac{24(\alpha+30\beta) - 3(\alpha+2\beta)[8\alpha+32\beta-4\alpha^2-20\alpha\beta-24\beta^2+\alpha^3+8\beta^3+6\alpha^2\beta+12\alpha\beta^2]}{\theta^4(\alpha+\beta)^4} \tag{15}$$

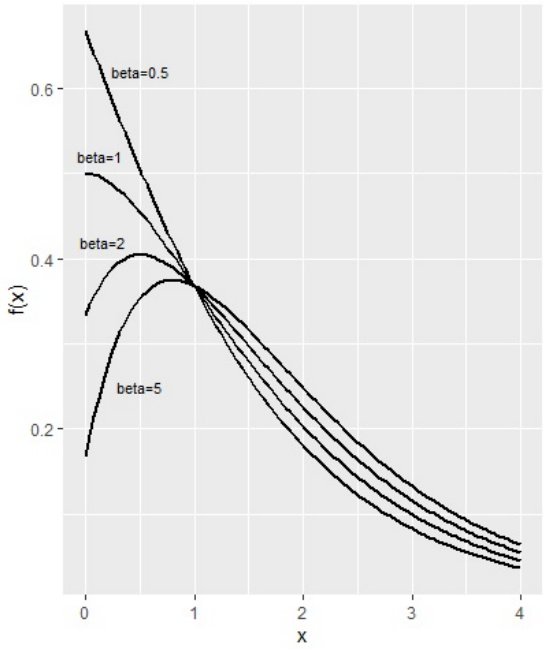


Figure 3: Density curves of TQLD with $\theta = 1$ and $\alpha = 1$.

The coefficient of variation(γ), skewness ($\sqrt{\beta_1}$) and the kurtosis (β_2) of the Three Parameter Quasi Lindley distribution have been obtained as

$$\gamma = \frac{\sqrt{\alpha^2 + 2\beta(2\alpha + \beta)}}{\alpha + 2\beta} \tag{16}$$

$$\sqrt{\beta_1} = \frac{18\beta^3 + 37\alpha\beta^2 + 30\beta\alpha^2 + 2(\alpha + 2\beta)^3}{(\alpha^2 + 2\beta(2\alpha + \beta))^{3/2}} \tag{17}$$

$$\beta_2 = \frac{\theta^3(\alpha + 2\beta)(\alpha + \beta)}{(\alpha^2 + 2\beta(2\alpha + \beta))^2} \tag{18}$$

Although, the expressions for coefficient of variation (γ), skewness ($\sqrt{\beta_1}$), and kurtosis (β_2) of the QLD of Shanker and Mishra[12] depend upon the parameter α only, the expressions for coefficient of variation(γ), skewness ($\sqrt{\beta_1}$), and kurtosis(β_2) of the TQLD depend upon all parameters θ, α and β . It can also be seen that the TQLD is positively skewed.

2.2 Hazard rate Function and Mean Residual Life Function

For a continous distribution with p.d.f. f(x) and c.d.f. F(x), The hazard rate function (also known as the failure rate function) and the mean residual life function of X are

respectively defined as,

$$h(x) = \lim_{\Delta X \to 0} \frac{P(X < x + \Delta x | X > x)}{\Delta x} \tag{19}$$

$$m(x) = \frac{1}{1 - F(x)} \int_x^\infty [1 - F(t)]dt \tag{20}$$

The corresponding hazard rate function, h(x) and the mean residual life function, m(x) of TQLD are obtained as

$$h(x) = \frac{\theta(\alpha + \beta\theta x)}{\beta\theta x + \alpha + \beta} \tag{21}$$

$$m(x) = \frac{\beta\theta x + \alpha + 2\beta}{\theta(\beta\theta x + \alpha + \beta)} \tag{22}$$

It can be easily verified that $h(0) = \dfrac{\theta\alpha}{\alpha + \beta} = f(0)$ and $m(0) = \dfrac{2\beta + \alpha}{\theta(\alpha + \beta)} = \mu_1'$. It is also obvious that h(x) is an increasing function of x, α, β and θ whereas m(x) is a decreasing function of x,α, β and θ. The graph of the hazard rate function is shown in figure 4. The failure rate function and the mean residual life function of the TQLD show its flexibility over Lindley distribution, Exponential distribution, the QLD of Shanker and Mishra[12] and A New Two Parameter Quasi Lindley distribution of Rama Shanker and Amanuel Habte Ghebretsadik[11].

2.3 Mean Deviations

The amount of scatter in a population is measured to some extent by the totality of deviations usually from mean and median. These are known as the mean deviation about the mean and the mean deviation about the median defined by

$$\delta_1(X) = \int_0^\infty |x - \mu| f(x)dx$$

and

$$\delta_2(X) = \int_0^\infty |x - M| f(x)dx$$

respectively, where $\mu = E(x)$.The measure $\delta_1(X)$ and $\delta_2(X)$ can be calculated using the relationships

$$\begin{aligned}
\delta_1(X) &= \int_0^\mu (\mu - x)f(x)dx + \int_\mu^\infty (x - \mu)f(x)dx \\
&= \mu F(\mu) - \int_0^\mu xf(x)dx - \mu[1 - F(\mu)] + \int_\mu^\infty xf(x)dx \\
&= 2\mu F(\mu) - 2\mu + 2\int_\mu^\infty xf(x)dx \\
&= 2\mu F(\mu) - 2\int_0^\mu xf(x)dx
\end{aligned} \tag{23}$$

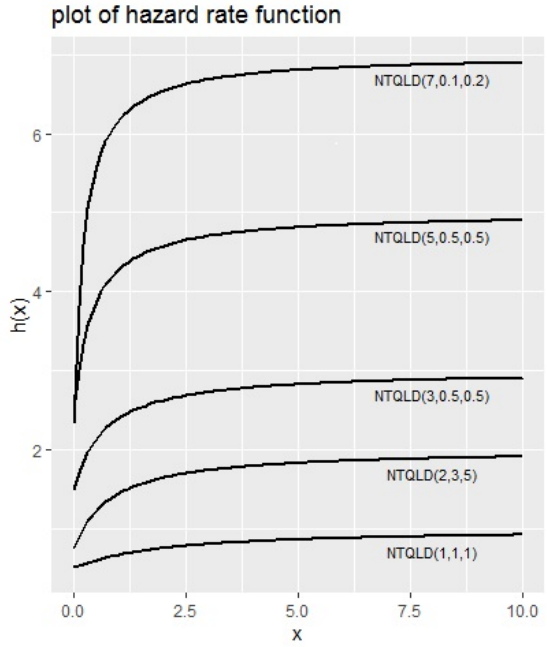


Figure 4: Hazard rate plot.

$$\begin{aligned}
\delta_2(X) &= \int_0^M (M-x)f(x)dx + \int_M^\infty (x-M)f(x)dx \\
&= MF(M) - \int_0^M xf(x)dx - M[1-F(M)] + \int_M^\infty xf(x)dx \\
&= -\mu + 2\int_M^\infty xf(x)dx \\
&= \mu - 2\int_0^M xf(x)dx \qquad (24)
\end{aligned}$$

Using p.d.f.(10) and expression for the mean of TQLD, we get

$$\int_0^\mu xf(x)dx = \mu - \frac{\left(\beta\mu^2\theta^2 + (2\beta+\alpha)\,\mu\theta + 2\beta + \alpha\right)\mathrm{e}^{-\mu\theta}}{(\beta+\alpha)\,\theta} \qquad (25)$$

$$\int_0^M xf(x)dx = \mu - \frac{\left(\beta M^2\theta^2 + (2\beta+\alpha)\,M\theta + 2\beta + \alpha\right)\mathrm{e}^{-M\theta}}{(\beta+\alpha)\,\theta} \qquad (26)$$

Then $\delta_1(x)$ $\delta_2(x)$ and can be written as

$$\delta_1(x) = \frac{(-\beta\theta^2\mu^2 - \alpha\mu\theta + 2\beta + \alpha)e^{-\theta\mu}}{\theta(\alpha+\beta)} \qquad (27)$$

$$\delta_2(x) = \frac{2\left(\beta M^2\theta^2 + (2\beta+\alpha)M\theta + 2\beta + \alpha\right)\mathrm{e}^{-M\theta}}{(\beta+\alpha)\theta} - \mu \tag{28}$$

2.4 Bonferroni And Lorenz Curves

The Bonferroni and Lorenz curves ([3]) have applications not only in economics to study income and poverty, but also in other fields like reliability, demography, insurance and medicine. The Bonferroni and Lorenz curves are defined as,

$$\begin{aligned} B(p) &= \frac{1}{p\mu}\int_0^q xf(x)dx \\ &= \frac{1}{p\mu}[\int_0^\infty xf(x)dx - \int_q^\infty xf(x)dx] \\ &= \frac{1}{p\mu}[\mu - \int_q^\infty xf(x)dx] \end{aligned} \tag{29}$$

and

$$\begin{aligned} L(p) &= \frac{1}{\mu}\int_0^q xf(x)dx \\ &= \frac{1}{\mu}[\int_0^\infty xf(x)dx - \int_q^\infty xf(x)dx] \\ &= \frac{1}{\mu}[\mu - \int_q^\infty xf(x)dx] \end{aligned} \tag{30}$$

respectively or equivalently

$$B(p) = \frac{1}{p\mu}\int_0^p F^{-1}(x)dx \tag{31}$$

and

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(x)dx \tag{32}$$

respectively, where $\mu = E(x)$ and $q = F^{-1}(p)$
Using p.d.f. (10), we get

$$\int_q^\infty xf(x)dx = \frac{\left(\beta q^2\theta^2 + (2\beta+\alpha)q\theta + 2\beta + \alpha\right)\mathrm{e}^{-q\theta}}{(\beta+\alpha)\theta} \tag{33}$$

Now using equation (33) in (29) and (30), we get

$$B(p) = \frac{1}{p}[1 - \frac{\left(\beta q^2\theta^2 + (2\beta+\alpha)q\theta + 2\beta + \alpha\right)\mathrm{e}^{-q\theta}}{\alpha+2\beta}] \tag{34}$$

and

$$L(p) = 1 - \frac{\left(\beta q^2\theta^2 + (2\beta+\alpha)q\theta + 2\beta + \alpha\right)\mathrm{e}^{-q\theta}}{\alpha+2\beta} \tag{35}$$

2.5. Stochastic Ordering

Stochastic ordering of positive continous random variables is an important tool for judging their comparative behaviour. A random variable X is said to be smaller than a random variable Y in the

1. Stochastic order ($X \leq_{st} Y$) if
   $F_X(x) \geq F_Y(x) \quad \forall x$
2. Hazard rate order ($X \leq_{hr} Y$) if
   $h_X(x) \geq h_Y(x) \quad \forall x$
3. Mean residual life order ($X \leq_{mlr} Y$) if
   $m_X(x) \leq m_Y(x) \quad \forall x$
4. Likelihood ratio order $X \leq_{lr} Y$ if
   $\frac{f_X(x)}{f_Y(x)}$ decreases in x

the following results due to Shaked and Shanthikumar[15] are well known for establishing stochastic ordering of distribution
$X \leq_{lr} Y \implies X \leq_{hr} \leq Y \implies X \leq_{mlr} \leq Y$
and
$X \leq_{hr} \leq Y \implies X \leq_{st} Y$
The TQLD is orderd w.r.to the strongest likelihood ratio ordering as shown in the following theorem.

Theorem 2.1: *Let $X \sim TQLD(\theta_1, \alpha_1, \beta_1)$ and $Y \sim TQLD(\theta_2, \alpha_2, \beta_2)$. Now under conditions*

1. $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ *and* $\theta_1 > \theta_2$
2. $\alpha_1 = \alpha_2, \beta_1 > \beta_2$ *and* $\theta_1 = \theta_2$ *and*
3. $\alpha_1 > \alpha_2, \beta_1 = \beta_2$ *and* $\theta_1 = \theta_2$

$X \leq_{lr} Y$ *and hence* $X \leq_{hr} Y, X \leq_{mlr} Y$ *and* $X \leq_{st} Y$

Proof: We have

$$\frac{f_X(x)}{f_Y(x)} = \frac{\theta_1(\alpha_1 + \beta_1\theta_1 x)e^{-\theta_1 x}}{\alpha_1 + \beta_1} \frac{\alpha_2 + \beta_2}{\theta_2(\alpha_2 + \beta_2\theta_2 x)e^{-\theta_2 x}}$$

now

$$log\frac{f_X(x)}{f_Y(x)} = log[\frac{\theta_1(\alpha_1 + \beta_1\theta_1 x)}{\theta_2(\alpha_2 + \beta_2\theta_2 x)}] + log[\frac{\alpha_2 + \beta_2}{\alpha_1 + \beta_1}] - x(\theta_1 - \theta_2)$$

and

$$\begin{aligned}\frac{d}{dx}log\frac{f_X(x)}{f_Y(x)} &= \frac{d}{dx}(log[\frac{\theta_1(\alpha_1+\beta_1\theta_1 x)}{\theta_2(\alpha_2+\beta_2\theta_2 x)}]+log[\frac{\alpha_2+\beta_2}{\alpha_1+\beta_1}]-x(\theta_1-\theta_2))\\ &= \frac{\alpha_2\beta_1\theta_1-\alpha_1\beta_2\theta_2}{(\alpha_1+\beta_1\theta_1 x)(\alpha_2+\beta_2\theta_2 x)}-(\theta_1-\theta_2) \qquad (36)\end{aligned}$$

It can be easily verified that under conditions,
$\frac{d}{dx}log\frac{f_X(x)}{f_Y(x)} < 0$. This means that $X \leq_{lr} Y$ and hence $X \leq_{hr} Y, X \leq_{mlr} Y$ and $X \leq_{st} Y$.

2.6. Order Statistics And Renyi Entropy Measure

Let $X_1, X_2, ..., X_n$ be a random sample of size n from TQLD (10). Let $X_{(1)} < X_{(2)} < ... < X_{(n)}$ denote the corresponding order statistics. The p.d.f. and the c.d.f. of the k^{th} order statistic, say $Y = X_{(k)}$ are given by

$$\begin{aligned}f_Y(y) &= \frac{n!}{(k-1)!(n-k)!}F^{k-1}(y)\{1-F(y)\}^{n-k}f(y)\\ &= \frac{n!}{(k-1)!(n-k)!}\sum_{l=0}^{n-k} {}^{n-k}C_l\ (-1)^l F^{k+l-1}(y)f(y) \qquad and\\ F_Y(y) &= \sum_{j=k}^{n} {}^{n}C_j F^j(y)\{1-F(y)\}^{n-j}\\ &= \sum_{j=k}^{n}\sum_{l=0}^{n-j} {}^{n}C_j\ {}^{n-j}C_l F^{j+l}(y)\end{aligned}$$

respectively, for k=1,2,3,...,n.

Thus, the p.d.f. and c.d.f. of k^{th} order statistics of TQLD are obtained as

$$f_Y(y) = \frac{n!}{(k-1)!(n-k)!}\frac{\theta(\alpha+\beta\theta x)}{\alpha+\beta}e^{-\theta x}\sum_{l=0}^{n-k} {}^{n-k}C_l(-1)^l[1-\frac{\beta\theta x+\alpha+\beta}{\alpha+\beta}e^{-\theta x}]^{k+l-1}$$

and

$$F_Y(y) = \sum_{j=k}^{n}\sum_{l=0}^{n-j} {}^{n}C_j\ {}^{n-j}C_l\ [1-\frac{\beta\theta x+\alpha+\beta}{\alpha+\beta}e^{-\theta x}]^{j+l}$$

An entropy of a random variable X is a measure of variatio of uncertainty. A popular entropy measure is Renyi entropy (1961). If X is a continuous random variable having probability density function f(.), then Renyi entropy is defined as

$$T_R(\gamma) = \frac{1}{1-\gamma}log\{\int f^\gamma(x)dx\}$$

where $\gamma > 0$ *and* $\gamma \neq 1$

$$\begin{aligned} T_R(\gamma) &= \frac{1}{1-\gamma} log[\int_0^\infty \frac{\theta^\gamma}{(\alpha+\beta)^\gamma}(\alpha+\beta\theta x)^\gamma e^{-\theta\gamma x} dx] \\ &= \frac{1}{1-\gamma} log[\int_0^\infty \frac{\theta^\gamma \alpha^\gamma}{(\alpha+\beta)^\gamma} \sum_{j=0}^{\infty} {}^rC_j (\frac{\beta\theta}{\alpha}x)^j e^{-\theta\gamma x} dx] \\ &= \frac{1}{1-\gamma} log[\sum_{j=0}^{\infty} {}^rC_j \frac{\theta^{\gamma-1}\beta^j \alpha^{\gamma-j}}{(\alpha+\beta)^\gamma} \frac{\Gamma(j+1)}{\Gamma(\gamma^{j+1})}] \end{aligned}$$

2.7. Stress-Strength Reliability

The stress- strength reliability describes the life of a component which has random strength that is subjected to a random stress. When the stress applied to it exceeds the strength, the component fails instantly and the component will function satisfactorily till X>Y. Therefore, R=P(Y<X) is a measure of component reliability and in statistical literature it is known as stress-strength parameter. It has wide applications in almost all areas of knowledge especially in engineering such as structures, deterioration of rocket motors, static fatigue of ceramic components, aging of concrete pressure vessels etc.
Let X and Y be independent strength and stress random variables having TQLD (10) with parameters (θ_1,α_1,β_1) and (θ_2,α_2,β_2) respectively. Then the stress-strength reliability R of TQLD can be obtained as

$$\begin{aligned} R = P(Y < X) &= \int_0^\infty f(x, \theta_1, \alpha_1.\beta_1) F(x; \theta_2, \alpha_2, \beta_2) dx \\ &= \int_0^\infty \frac{\theta_1 \left(\beta_1\theta_1 x + \alpha_1\right) \mathrm{e}^{-\theta_1 x} \left(1 - \frac{(\beta_2\theta_2 x + \beta_2 + \alpha_2)\mathrm{e}^{-\theta_1 x}}{\beta_2 + \alpha_2}\right)}{\beta_1 + \alpha_1} dx \\ &= \frac{3\beta_1 + 2\alpha_1}{4\left(\beta_1 + \alpha_1\right)} - \frac{\beta_2\theta_2}{4\left(\beta_2 + \alpha_2\right)\theta_1} \end{aligned}$$

3. Maximum Likelihood Estimate (MLE)

Let $(x_1, x_2, ..., x_n)$ be a random sample of size n from TQLD (10) by

$$L = (\frac{\theta}{\alpha+\beta})^n \prod_{i=1}^{n} (\alpha + \beta\theta x_i) e^{-n\theta\bar{x}}$$

The natural log likelihood function is thus obtained as

$$lnL = nln(\frac{\theta}{\alpha+\beta}) + \sum_{i=1}^{n} ln(\alpha + \beta\theta x_i) - n\theta\bar{x}$$

The maximum likelihood estimators (MLE) $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\beta}$ of θ, α and β are the solutions of the following non-linear equations

$$\frac{\partial L}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \frac{\beta x_i}{\alpha + \beta\theta x_i} - n\bar{x} = 0 \tag{37}$$

$$\frac{\partial L}{\partial \alpha} = \frac{-n}{\alpha + \beta} + \sum_{i=1}^{n} \frac{1}{\alpha + \beta\theta x_i} = 0 \tag{38}$$

$$\frac{\partial L}{\partial \beta} = \frac{-n}{\alpha + \beta} + \sum_{i=1}^{n} \frac{\theta x_i}{\alpha + \beta\theta x_i} = 0 \tag{39}$$

where $\bar{x}$ is the sample mean.

These three natural log likelihood equations do not seem to be solved directly. However, the Fisher's scoring method can be applied to solve these equations. We have

$$\begin{aligned}
\frac{\partial^2 lnL}{\partial \theta^2} &= \frac{-n}{\theta^2} - \sum_{i=1}^{n} \frac{\beta^2 x_i^2}{(\alpha + \beta\theta x_i)^2} \\
\frac{\partial^2 lnL}{\partial \theta \partial \alpha} &= -\sum_{i=1}^{n} \frac{\beta x_i}{(\alpha + \beta\theta x_i)^2} = \frac{\partial^2 lnL}{\partial \alpha \partial \theta} \\
\frac{\partial^2 lnL}{\partial \theta \partial \beta} &= \sum_{i=1}^{n} \frac{(\alpha + \beta\theta x_i)x_i - \beta x_i(1 + \theta x_i)}{(\alpha + \beta\theta x_i)^2} = \frac{\partial^2 lnL}{\partial \beta \partial \theta} \\
\frac{\partial^2 lnL}{\partial \alpha^2} &= \frac{n}{(\alpha + \beta)^2} - \sum_{i=1}^{n} \frac{1}{(\alpha + \beta\theta x_i)^2} \\
\frac{\partial^2 lnL}{\partial \alpha \partial \beta} &= \frac{n}{(\alpha + \beta)^2} - \sum_{i=1}^{n} \frac{\theta x_i}{(\alpha + \beta\theta x_i)^2} = \frac{\partial^2 lnL}{\partial \beta \partial \alpha} \\
\frac{\partial^2 lnL}{\partial \beta^2} &= \frac{n}{(\alpha + \beta)^2} - \sum_{i=1}^{n} \frac{\theta^2 x_i^2}{(\alpha + \beta\theta x_i)^2}
\end{aligned}$$

The following equations can be solved for MLE's $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\beta}$ of θ, α and β of TQLD (10)

$$\begin{pmatrix}
\frac{\partial^2 lnL}{\partial \theta^2} & \frac{\partial^2 lnL}{\partial \theta \partial \alpha} & \frac{\partial^2 lnL}{\partial \theta \partial \beta} \\
\frac{\partial^2 lnL}{\partial \alpha \partial \theta} & \frac{\partial^2 lnL}{\partial \alpha^2} & \frac{\partial^2 lnL}{\partial \alpha \partial \beta} \\
\frac{\partial^2 lnL}{\partial \beta \partial \theta} & \frac{\partial^2 lnL}{\partial \beta \partial \alpha} & \frac{\partial^2 lnL}{\partial \beta^2}
\end{pmatrix}
\begin{pmatrix}
\hat{\theta} - \theta_0 \\
\hat{\alpha} - \alpha_0 \\
\hat{\beta} - \beta_0
\end{pmatrix} =
\begin{pmatrix}
\frac{\partial L}{\partial \theta} \\
\frac{\partial L}{\partial \alpha} \\
\frac{\partial L}{\partial \beta}
\end{pmatrix}$$

where θ_0, α_0 and β_0 are the initial values of θ, α and β respectively. These equations are solved iteratively till sufficiently close values of $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\beta}$ are obtained.

4. Goodness of Fit

Three parameter Quasi Lindley distribution (TQLD) has been fitted to a number of lifetime data to test its goodness of fit. In this section, we present the goodness of fit of TQLD for a real lifetime data and its fit has been compared with the three-parameter Lindley distribution (TPLD), introduced by [14]. The following lifetime data has been considered for testing the goodness of fit of TQLD and TPLD.

Data Set: This data represents the survival times (in days) of 72 guinea pigs infected with virulent tubercle bacilli, observed and reported by Bjerkedal[2]

10, 33, 44, 56, 59, 72, 74, 77, 92, 93, 96, 100, 100, 102, 105, 107, 107, 108, 108, 108, 109, 112, 113, 115, 116, 120, 121, 122, 122, 124, 130, 134, 136, 139, 144, 146, 153, 159, 160, 163 163, 168, 171 172, 176, 183, 195, 196, 197, 202 213, 215, 216, 222, 230, 231, 240, 245, 251, 253, 254, 254, 278, 293, 327, 342, 347, 361, 402, 432, 458, 555

In order to compare TQLD and TPLD, values of -2ln L and K-S Statistics (Kolmogorov-Smirnov Statistics) for real life time data has been computed. The formulae for computing K-S Statistics is as follows:

$$K - S = Sup_x|F_n(x) - F_0(x)| \tag{40}$$

where k = the number of parameters, n = the sample size and $F_{n(x)}$ is the empirical distribution function.

The best distribution is the distribution which corresponds to lower value of -2ln L and K-S statistics and higher p-value.

Table 1: Ml estimates of TQLD and TPLD

Distributions	ML Estimate
	$\hat{\theta}\ \hat{\alpha}\ \hat{\beta}$
TQLD	0.0118342 -0.1762485 2.0793276
TPLD	0.01183 -0.1758485 2.793276

Table 2: AIC, BIC, K-S statistic and p value corresponding to TPLD and TQLD

Distributions	-2 ln L	AIC	BIC	K-S statistic	P-value
TQLD	856.212	862.212	869.042	0.15179	0.07245
TPLD	858.4548	864.4548	871.2848	0.18537	0.01419

It can be easily seen from above table that TQLD gives better fit than the TPLD.

5. Concluding Remarks

In this paper, we proposed Three parameter Quasi Lindley distribution (TQLD), of which the one-parameter LD is a particular case. Several properties of the TQLD such as moments, skewness and Kurtosis have been discussed. Various reliability properties such as failure rate function, mean residual life function, stochastic orderings have been obtained and discussed and shown that the TQLD is more flexible than Lindley distribution, Exponential distribution, and New QLD. The density function of the TQLD along with its cumulative distribution function and hazard rate function has been shown graphically for different values of its parameters for comparative study with TPLD. The estimation of parameters by the method of maximum likelihood have been discussed. The proposed distribution has been fitted to a data set, to test its goodness of fit to which earlier the Three parameter Lindley distribution has been fitted[14] and it is found that the TQLD provides closer fits than those by the Three Parameter Lindley distribution.

References

[1] H. S. Bakouch, B. M. Al-Zahrani, A. A. Al-Shomrani, V. A. Marchi, and F. Louzada, An extended lindley distribution, *Journal of the Korean Statistical Society*, 41(1):75-85, 2012.

[2] T. Bjerkedal et al. , Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, *American Journal of Epidemiology*, 72(1), 130 -148, 1960.

[3] C. E. Bonferroni, *Elementi di Statistca generale, (ristampa con aggiunte): anno accademico 1932/33; Bari, R. Istit. super. di scienze economiche*. Gilli, 1933.

[4] M. Ghitany and D. Al-Mutairi, Estimation methods for the discrete Poisson–Lindley distribution, *Journal of Statistical Computation and Simulation*, 79(1):1-9, 2009.

[5] M. Ghitany, F. Alqallaf, D. K. Al-Mutairi and H. Husain, A two-parameter weighted Lindley distribution and its applications to survival data, *Mathematics and Computers in simulation*, 81(6):1190-1201, 2011.

[6] M. E. Ghitany, B. Atieh, and S. Nadarajah, Lindley distribution and its Applications, *Mathematics and Computers in Simulation*, Vol.78(4):493-506, 2008.

[7] E. Gómez-Déniz and E. Calderín-Ojeda, The discrete Lindley distribution: properties and applications,*Journal of Statistical Computation and Simulation*, 81(11):1405-1416, 2011.

[8] D. V. Lindley, Fiducial distributions and Bayestheorem, *Journal of the Royal Statistical Society. Series B(Methodological)*, pages 102- 107, 1958.

[9] J. Mazucheli and J. A. Achcar, The Lindley distribution applied to competing risks lifetime data, *Computer methods and programs in biomedicine*, 104(2):188-192, 2011.

[10] M. Sankaran, The discrete poisson-lindley distribution, *Biometrics*, pages 145-149, 1970.

[11] R. Shanker and A. Amanuel, A new quasi Lindley distribution,*International Journal of Statistics and systems*,8(2):143-156, 2013.

[12] R. Shanker and A. Mishra, A quasi Lindley distribution, *African Journal of Mathematics and Computer Science Research*, 6(4):64-71, 2013.

[13] R. Shanker and A. Mishra, *Journal of Indian Statistical Association*,54(1&2):113-125, 2016.

[14] R. Shanker, K. K. Shukla, R. Shanker and A. Tekie, A three-parameter lindley Distribution, *American Journal of Mathematics and Statistics*, 7(1):15-26,2017

[15] J. G. Shanthikumar, Stochastic orders and their applications, Academic Press, 1994.

Proceedings of National Seminar in Applied Statistics
& Symposium on Stochastic Modelling - 2020, pp. 233-240.

Quasi-Bayesian Estimation of Time to Test Transform in Lomax Model

Sowbhagya S Prabhu[1] and E. S. Jeevanand[2]
[1]Department of Statistics, Nirmala College, Muvatupuzha,Kerala, India.
Email:sowbhagyasprabhu@gmail.com
[2]Department of Mathematics, Union Christian College, Aluva, Kerala, India.

Abstract

In this article, we estimate the Time To Test Transform for the Lomax function based on Quasi likelihood. Both the classical and Bayes estimators have been developed. The empirical evaluation of the estimates are done using a simulation study.

Key words: Time To Test Transform, Lomax distribution, Quasi- Bayesian estimation, Loss functions.

1. Introduction

The concept of time to test (TTT) transforms is well known for its applications in different fields of study such as reliability analysis (Lai and Xie (2006, p. 42)), econometrics (Pham and Turkkan (1994)), stochastic modelling (Vera and Lynch (2005)), tail orderings (Bartoszewicz (1995)), and ordering distributions (Kochar *et al.* (2002)). A major share of the literature on TTT is concerned with reliability problems that include characterization of ageing properties, model identification, tests of hypotheses, age replacement policies in maintenance life distributions, and defining new class q qes of life distributions. We refer the reader to Bergman ad Klefsjo (1984), Bartoszewicz (1995), Haupt and Schabe (1997), Kochar (2002), Li and Zou (2004), Ahmed *et al.* (2005), Li and Shaked (2004), Nanda and Shaked (2008), and the references therein for further details.

Time to test (TTT) transform plots are useful for analysing non-negative data. The plots help in choosing a mathematical model for the data and provide information about failure rate. Also incomplete data can be analysed and there is a theoretical basis for such

an analysis, see Barlow and Campo (1975). As TTT is useful in analysing incomplete data, we can order the distributions according to TTT of respective distributions.

The Lomax distribution also known as the Pareto distribution of second kind has, in recent years, assumed opposition of importance in the field of life testing because of its uses to fit business failure data. It has been used in the analysis of income data, and business failure data. It may describe the life time of a decreasing failure rate component as a heavy tailed alternative to the exponential distribution. Lomax distribution was introduced by Lomax (1954), Abdullah and Abdullah (2010) estimated the parameters of Lomax distribution based on generalized probability weighted moment. Zangan (1999) deals with the properties of the Lomax distribution with three parameters. Abd-Elfatth and Mandouh (2004) discussed inference for $R = PrY < X$ when X and Y are two independent Lomax random variables. Nasiri and Hosseini (2012) performs comparisons of maximum likelihood estimation (MLE) based on records and a proper prior distribution to attain a Bayes estimation (both informative and non-informative) based on records under quadratic loss and squared error loss functions. Afaq *et al.* (2014) estimates parameters of Lomax distribution using Jeffery's and extension of Jeffery's prior under different loss functions.

The pdf is given by

$$f(x:\theta,\lambda) = \frac{\theta\lambda^{\theta}}{(\lambda+x)^{\theta+1}} \qquad x,\theta,\lambda > 0 \tag{1}$$

Where θ and λ are shape and scale parameters respectively.
For the above model, the TTT simplifies to

$$\phi(t) = 1 - (1-t)^{\frac{\theta-1}{\theta}} \tag{2}$$

In the next section we obtain Quasi likelihood estimate of the TTT for the Lomax distribution (1) when the parameter λ is to be known both in classical and Bayesian framework.

2. Quasi-Likelihood Estimation

The quasi-likelihood function was introduced by Wedderburn (1974), to be used for estimating the unknown parameters in generalized linear models. The idea of quasi-likelihood weakens the assumption that we know exactly the distribution of the random component in the model, and replace it by an assumption about how the variance changes with mean. The quasi-likelihood function could be used for estimation in the same way as the usual likelihood function. Wedderburn (1974) and McCullagh (1983) showed that the maximum quasi-likelihood estimates have many properties similar to particular, the maximum

Table 1: Bias of the Maximum Quasi likelihood estimates of θ

n	θ	Bias
25	1.5	0.00644
	2.5	0.00144
	3.5	0.02689
50	1.5	0.01344
	2.5	0.08923
	3.5	0.07326
75	1.5	0.08425
	2.5	0.01044
	3.5	0.07894

quasi-likelihood estimate of the vector β (the vector of parameters in regression models) is a asymptotically normal with mean β, and asymptotic covariance may be derived in the usual fashion from the second derivative matrix of the quasi-likelihood function. Also, if the underlying distribution comes from a natural exponential family the maximum quasi-likelihood estimate maximizes the likelihood function and so it has full asymptotic efficiency; under more general distributions there is some loss of efficiency, which have been investigated by Firth (1987) and Hill and Tsai (1988). Weddwerburn defned the quasi-likelihood function as

$$Q(x,\mu) = \int_{\mu} \frac{x-\mu}{V(\mu)} d\mu + o(x) \tag{3}$$

where $\mu = E(x)$, $V(\mu) = Var(x)$, and $o(x)$ is some function of x only. The variance assumption is to $Var(x) = \phi Var(\mu)$ where the variance function $V(.)$ is assumed to be known and the parameter ϕ may be unknown. The quasi-likelihood function has properties similar to those of the log-likelihood function.
For a sample $\underline{x} = (x_1, x_2, \cdots, x_n)$ of size n from (3), the quasi-likelihood function simplifies to

$$Q(x,\theta,\lambda) = log\frac{\theta-1}{\lambda}^{n} - \frac{\theta-1}{\lambda}v \quad \text{where } v = \sum_{i=1}^{n} x_i$$

The natural exponent of $Q(x,\theta,\lambda)$ is the likelihood function and is given as

$$l(\underline{x}|\theta,\lambda) = \frac{\theta-1}{\lambda}^{n} .exp[-\frac{\theta-1}{\lambda}v] \tag{4}$$

Using (4), the maximum quasilikelihood estimate of the parameter θ denoted by $\widehat{\theta}_{MQL}$ is simplified as

$$\widehat{\theta}_{MQL} = 1 + \frac{n\lambda}{v}$$

The maximum quasi-likelihood estimator for the TTT, denoted by $\widehat{\phi}_{MQL}(t)$ can be obtained from (1.2) after replacing θ by $\widehat{\phi}_{MQL}$.

3. Quasi-Bayesian Estimation

The Bayesian approach provides the possibility for incorporating prior information about the relevant parameters. To this end the parameter θ, is considered as a random variable, having some specified distribution. Here we suggest the conjugate prior distribution for the parameters and is given by

$$g(\theta, \lambda) = \frac{\tau^r}{\Gamma r}\theta^{r-1}e^{-\theta\tau}. \tag{5}$$

Combining (4) and (5), the joint posterior density is obtained as

$$f(\theta, \lambda|\underline{x}) \propto \frac{\tau^r}{\Gamma r}\theta^{r-1}\frac{\theta-1}{\lambda}^n .exp[-\theta\tau - (\frac{\theta-1}{\lambda})v],$$

$$f(\theta|\underline{x}) = \frac{\theta^{r-1}\frac{\theta-1}{\lambda}^n .exp[-\theta\tau - (\frac{\theta-1}{\lambda})v]}{C_1(0)}, \tag{6}$$

where

$$C_1(d) = \int_0^\infty \theta^{r+d-1}\frac{\theta-1}{\lambda}^n .exp[-\theta\tau - (\frac{\theta-1}{\lambda})v]d\theta.$$

The symbol C with suffixes stands for the normalizing constants. Let $\phi(t)$ be a parameter itself denoted by ϕ for simplicity. Replacing θ in (3.3) in terms of ϕ by that (1.2), we get the posterior of the TTT as

$$f(\theta|\underline{x}) = \frac{R_\phi^{r+1}(1-\theta)^{-1}(\frac{R_\phi-1}{\lambda})^n .exp[-R_\phi\tau - (\frac{R_\phi-1}{\lambda})v]}{C_2(t,0)}$$

$$\text{where } C_2(t,d) = \int_0^1 \phi^d R_\phi^{r+1}(1-\theta)^{-1}\frac{R_\phi-1}{\lambda}^n .exp[-R_\phi\tau - [\frac{R_\phi-1}{\lambda}]v]d\phi, \tag{7}$$

with

$$R_\phi = \Big[1 - \frac{log[1-\phi(t)]}{log(1-t)}\Big]^{-1}.$$

From a theoretic view point, in order to select the bestestimator, a loss function must be specified and is used to represent a penalty associated with each of the possible estimates.

- The Quasi-Bayes estimator and the posterior risk of TTT under Squared error loss function are given by

$$\widehat{\phi}_{QBS} = E(\phi|x) = \frac{C_2(1)}{C_2(0)}$$

with risk function,

$$R(\phi, \widehat{\phi}_{QBS}) = Var(\phi|x) = \frac{C_2(1)}{C_2(0)} - \frac{C_2(1)}{C_2(0)}^2$$

- The Quasi Bayes estimator and the posterior risk of TTT under Linex loss function are given by

$$\widehat{\phi}_{QBL} = \frac{-1}{a} ln[E(e^{-a\phi}|x)] = \frac{-1}{a} lnB_2,$$

and

$$R(\phi, \widehat{\phi}_{QBL}) = lnB_2 + a\frac{C_2(1)}{C_2(0)},$$

where $B_2 = \frac{\int_0^1 \phi^d R_\phi^{r+1}(1-\theta)^{-1} fracR_\phi - 1\lambda^n .exp[-R_\phi\tau - [\frac{R_\phi - 1}{\lambda}]v]d\phi}{C_2(t,0)}$ and $C_2(t,d)$ is given in (7).

4. A Numerical Example

In the absence of real data we compare the estimates empirically by generating observations from the Lomax distribution. The samples of sizes $n = 25, 50, 75$ were generated for different values of the parameter from Lomax distribution with $\alpha = 1.5, 2.5, 3.5$ and $\beta = 1.5, 2.5, 3.5$. For the simulation study, we choose the value of $a = 0.5$, the LINEX shape parameter and set $t = 0.3$. The bias and the posterior risks (in parenthesis) for each estimates using Quasi -Bayesian estimation and MQL estimates are tabulated in the following tables.

5. Conclusion

We obtained the estimators of the parameter and TTT function of Lomax distribution. The estimators are obtained using both symmetric and asymmetric loss functions. From the above tables, we can conclude that,

1. As the sample size increases the bias and risk decreases for all values of θ.
2. Linex loss function has lesser bias than the squared error loss.

Table 2: Bias and risks (in parentheses) of the Quasi Bayesian estimate of TTT under SELF and LLF

n	θ	Trueϕ	$Bias_{MQL}$	$Bias_{QBS}$	$Bias_{QBS}$	$Bias_{QBL}$	$Bias_{QBL}$
25	1.5	0.112	0.00091	0.00054	(0.00452)	0.00031	(0.00046)
	2.5	0.193	0.00007	0.00179	(0.00125)	0.00009	(0.00052)
	3.5	0.225	0.00060	0.00044	(0.00235)	0.00431	(0.00056)
50	1.5	0.112	0.00191	0.00121	(0.00111)	0.00031	(0.00031)
	2.5	0.193	0.00396	0.00012	(0.00825)	0.00006	(0.0002)
	3.5	0.223	0.00169	0.00175	(0.00043)	0.00012	(0.00036)
75	1.5	0.112	0.01116	0.01108	(0.00023)	0.00012	(0.00023)
	2.5	0.193	0.00048	0.00025	(0.00081)	0.00014	(0.00013)
	3.5	0.223	0.00174	0.00135	(0.00134)	0.00015	(0.00012)

3. The Quasi Bayesian estimates are found to be better than the maximum Quasi likelihood estimates.

References

[1] Afaq, A., Ahmad S. P., and Ahmed A.,(2015). Bayesian Analysis of shape parameter of Lomax distribution under different loss functions, *International Journal of Statistics and Mathematics*, 2 (1), 55-65.

[2] Barlow, R. E. (1979). Geometry of the total time on test transform. *Naval Res. Logist. Quart.*,26, 393-402.

[3] Barlow, R. E. and Campo, R. (1975). Total time on test processes and applications to failure data analysis. *Reliab. and Fault Tree Analysis, SIAM, Philadelphia*, 451-481.

[4] Bartoszewicz, J. (1995). Stochastic order relations and the total time on test transform. *Statistics and Probability Letters* 22, 103-110.

[5] EL-Shahat, M.A.T. and Ismail, A.E., Quasi- Likelihood Estimation for Inverse Weibull Distribution, (under publication).

[6] Haupt, E. and Schabe, H. (1997). The TTT transformation and a new bathtab distribution model. *J. Statist. Planning Infer.* 60, 229-240.

[7] Kochar, S. C., Li, X.and Shaked,M. (2002). The total time test transform and the excess wealth stochastic orders of distribution. *Adv.Appl.Prob.*34, 826-845.

[8] Li,H. and Shaked, M. (2007). A general family of univariate stochastic orders. *J. Statist.planning Infer.* 137, 3601-3610.

[9] Li, X. and Zou, M. (2004). Preservation of stochastic orders for random minima and maxima with applications. *Naval Res.logistics* 51,332-334.

[10] Nanda, A. K. and Shaked, M.(2008). Partial ordering and aging properties of order statistics when sample size is random: a brief review. *Commun.Statist .Theory Meth.*37, 1710-1720.

[11] Pham, T. G. and Turkkan, M. (1994). The Lorenz and the scaled total-time-on-test transform curves: a unified approach. *IEEE Trans. Reliab.* 43, 76-84.

[12] Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Models and the Gauss-Newton Method, *Biometrika*, 61 (3), 439-443.

[13] S. K. Ashour and E. A. Elsherpieny, on the efficiency of quasi-Bayesian estimation (a numerical investigation), The Annual Conference, Vol. 33, ISSR, Cairo University, 1998.

[14] Vera, F. and Lynch, J. (2005). K-mart stochastic modeling using iterated total time on test transforms. In *Modern Statistical and Mathematical Methods in Reliability*, World Scientific, Singapore, 395-409.

[15] Youssuf .M. S (2009) Bayesian estimation for the Pareto parameters using Quasi Likelihood function ,Applied mathematical sciences ,vol.3,2009,no.11,509-517.

Author Index

U. H. Acharya, 79
Aleesha Sebastian, 125
Arya Damodharan, 193
Ashlin Mathew P. M., 217
Bavagosai Pratima, 23
Deepthy G. S., 97
Dhannya P. Joseph, 183
Drisya M., 65
E. V. Gijo, 79
K. Jayakumar, 35
Jeena Joseph, 165
E. S. Jeevanand, 233
Joby K. Jose, 65
K. K. Jose, 49
K. Kalyanaraman, 113
Krishnendu K., 65
A. M. Mathai, 1
Mary George K., 113
G. Meenakshi, 113
A. Mohammed Shabeer, 35
K. Muralidharan, 23
Nicy Sebastian, 97
Pinky Peter, 165
Prasanth C. B., 155
Ramya M. S., 155
Rani Sebastian, 125
Reshma Rison, 97
Sajesh T. A., 193
Seema S. Nair, 207
Shalitha Jacob, 49
Sneha Robinson, 217
Sowbhagya S. Prabhu, 233
Sudheesh K. Kattumannil, 91

About this book

This book is published as a collection of selected peer-reviewed papers presented by the participants of the National Seminar in Applied Statistics and Symposium on Stochastic Modelling, NSASSSM-2020 organized by the Department of Statistics and B.Voc Data Science, St. Thomas College, Thrissur, Kerala 680001 as part of centenary celebrations of the college & 85th birth anniversary of Professor A. M. Mathai during 04 - 06 February 2020. The conference is partially funded by DST, UGC and Mathematical and Statistical Sciences Trust. Topics covered in this book are Distribution theory, Special functions and fractional calculus, Multivariate analysis, Reliability and Survival analysis, Quality control, Bayesian inference, Population studies, Official statistics and Time series and Stochastic Modelling. This book is intended to provide an opportunity to the young Statisticians and researchers to get to know the latest development in various fields of Applied Statistics and Stochastic Modelling.

ISBN 978-81-944730-1-5