ISSN 2454-7395(online)

# STATISTICS AND APPLICATIONS

FOUNDED 1998

## Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 21, N0. 1, 2023 (New Series)

# Society of Statistics, Computer and Applications

**Society of Statistics, Computer and Applications**
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA

# Statistics and Applications

ISSN 2454-7395(online)

## Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 21, No. 1, 2023 (New Series)

# Statistics and Applications

Volume 21, No. 1, 2023 (New Series)

**Murari Singh**, Formerly at International Centre for Agricultural Research in the Dry Areas, Amman, Jordan; mandrsingh2010@gmail.com
**Nripes Kumar Mandal**, Flat No. 5, 141/2B, South Sinthee Road, Kolkata-700050; mandalnk2001@yahoo.co.in
**P. Venkatesan**, Professor Computational Biology SRIHER, Chennai, Adviser, CMRF, Chennai; venkaticmr@gmail.com
**Pritam Ranjan**, Indian Institute of Management, Indore - 453556; MP, India; pritam.ranjan@gmail.com
**Ramana V. Davuluri**, Department of Biomedical Informatics, Stony Brook University School of Medicine, Health Science Center Level 3, Room 043 Stony Brook, NY 11794-8322, USA; ramana.davuluri@stonybrookmedicine.edu; ramana.davuluri@gmail.com
**S. Ejaz Ahmed**, Faculty of Mathematics and Science, Mathematics and Statistics, Brock University, ON L2S 3A1, Canada; sahmed5@brocku.ca
**Sanjay Chaudhuri**, Department of Statistics and Applied Probability, National University of Singapore, Singapore -117546; stasc@nus.edu.sg
**Sat N. Gupta**, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC -27412, USA; sngupta@uncg.edu
**Saumyadipta Pyne**, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
**Snigdhansu Chatterjee**, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
**T.V. Ramanathan**; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
**Tapio Nummi**, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
**Tathagata Bandyopadhyay**, Indian Institute of Management Ahmedabad, Gujarat; tathagata.bandyopadhyay@gmail.com, tathagata@iima.ac.in
**Tirupati Rao Padi**, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
**V. Ramasubramanian**, ICAR-IASRI, Library Avenue, PUSA, New Delhi – 110012;

## CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 21, No. 1, 2023 (New Series)

1. Statistical Model for Brand Loyalty and Switching — 1-9
*Kumaraswamy Kandukuri and Bhatracharyulu N. Ch.*

2. Calibration Estimator in Two Stage Sampling Using Double Sampling Approach when Study Variable is Inversely Related to Auxiliary Variable — 11-22
*Ankur Biswas, Kaustav Aditya, U.C. Sud and Pradip Basak*

3. Regular Group Divisible Designs Using Symmetric Groups — 23-26
*Shyam Saurabh and Kishore Sinha*

4. New Intervention Based Exponential Model with Real Life Data Applicability — 27-40
*Vilayat Ali Bhat and Sudesh Pundir*

5. Estimation of AUC of Mixture ROC Curve in the Presence of Measurement Errors — 41-49
*G. Siva and R. Vishnu Vardhan*

6. Evaluating Batsman using Survival Analysis — 51-62
*Parag Shah, R. D. Chaudhari and M. N. Patel*

7. Discrete Harris Extended Weibull Distribution and Applications — 63-79
*Sophia P. Thomas, Lishamol Tomy and K.K. Jose*

8. Nonparametric Prediction Intervals for Future Order Statistics and *k*-Record Values — 81-98
*Laji Muraleedharan and Manoj Chacko*

9. Empirical Mode Decomposition Based Ensemble Hybrid Machine Learning Models for Agricultural Commodity Price Forecasting — 99-112
*Pankaj Das, Girish Kumar Jha and Achal Lama*

10. Estimation of Area under the Multi-Class ROC for Non-Normal Data — 113-121
*Arunima S. Kannan and R. Vishnu Vardhan*

11. Parameter Estimation of Generalized Exponential Distribution using Variations in Methods of Ranked Set Sampling — 123-141
*Vyomesh Nandurbarkar and Ashok Shanubhogue*

12. Statistical Modeling of Temperature in Krishna District using Copula Analysis — 143-159
*A. Rajini and C. Jayalakshmi*

13. On the Bivariate Generalized Chen Distribution — 161-177
*R. M. Mandouh*

14. Properties of Partial Product Processes 179-185
*Babu D., Sutha M., Neelakandan R. and Elumalai P.*

15. Asymptotic Results for Generalized Runs in Higher Order Markov Chains 187-205
*Anuradha*

16. Randomized Response in Combination with Direct Response for Estimation of Incidence Parameters of Two Sensitive Qualitative Features 207-215
*Opendra Singh*

17. Conway-Maxwell Poisson Distribution: Some New Results and Minimum Variance Unbiased Estimation 217-230
*Jahnavi Merupula and V.S.Vaidyanathan*

18. Single Server Poisson Queueing Model with Additve Exponential Service Time Distribution 231-244
*Ch. Ganapathi Swamy, K. Srinivasa Rao and S. Govinda Rao*

19. The Direct Method for the Optimal Solution of a Transportation Problem 245-249
*Priyanka Malviya, Sushma Jain and Rina Agrawal*

Statistics and Application {ISSN 2454-7395(online)}
Volume 21, No. 1, 2023 (New Series), pp 1-9

# Statistical Model for Brand Loyalty and Switching

**Kumaraswamy Kandukuri and Bhatracharyulu N.Ch.**
*Department of Statistics, University College of Science, Osmania University, Hyderabad – 7*

Received: 26 May 2021; Revised: 20 December 2021; Accepted: 01 January 2022

## Abstract

Consumer's decision on purchase of an item or brand depends on many factors. Some stochastic distributions and markov chain models were used to analyze the consumers' purchase behavior. In this paper, a statistical linear model is constructed to study the repeated purchase behavior based on the performance measures of various brands. The proposed model will generate transformation matrix, which is used to estimate the repeat purchase of brands depending on polarized index and market share values. The model is also illustrated with a suitable example.

*Key words:* Brand loyalty; Dirichlet model; Least squares estimate; Markov model; Repeat rate.

## 1. Introduction

Scientifically, the customer choice behaviors can be broadly classified as the state of decisions and actions that influence the purchase pattern. The buying pattern of consumers begins coherently with factors attitudes towards the brands in a particular product category, made up of various components attracting the preferred brand over all other brands. The study reflects how changes are in consumer attitudes incorporate the purchase behavior; changes in consumer's attitudes with the various factors can be viewed in terms of probabilities.

Lipstein (1965) constructed a statistical analytical model to study the consumer behavior on advertising effect in marketplace. Colombo and Morrison (1989) focused on marketing strategies for brand switching model with classes of consumers: hardcore loyal and potential switchers. Dirichlet probabilistic model used to study the behavior of consumer purchasing pattern of brand choice and purchase incidence in Abel *et al*. (1980), Bound (2009), and Rungie and Goodhardt (2004).

## 2. Statistical model for repeated purchase

**Definition 1 (Repeat rate):** Let the purchase occasion is the event when a shopper makes a purchase from any one of the brand categories. The proportion of buyers of a particular brand at the last purchase occasion and also buys the same brand in the next purchase occasion. It is an intuitive measure of loyalty and it records how much a brand hangs onto its buyers.

Corresponding Author: Kumaraswamy Kandukuri
E-mail: kumaraswami.kandukuri@gmail.com

**Definition 2 (Market share):** The market share of a brand is **t**he proportion of total purchases of that particular brand to total purchases of all brands.

**Definition 3 (Polarized index):** The brand Polarized index of a particular brand is defined as $\phi_i = (\rho_i - \mu_i)/\ (1 - \mu_i)$.

**Definition 4 (Dirichlet model):** The Dirichlet model is a probability distribution function that describes the distribution of consumer purchases of each brand within a product category over time. The data set is multivariate, contains numerous brands, and the model counts the number of transactions and is discrete and instructive, *i.e.*, counts are integer type, non-negative because purchases are whole numbers. As a result, it cannot be non-negative.

**Notations:**
Total no. of purchases of $i^{\text{th}}$ brand is $\alpha_{i.}$
Total no. of purchases of all brands $S = \Sigma_i \alpha_{i.}$
Market share of $i^{\text{th}}$ brand is $\mu_i = \alpha_i / S$.
Repeat rate is $\rho_i$.
Loyalty polarized index of brand '*i*' is $\phi_i$.

**Inter relationships / properties of the Dirichlet model:**
1. $\mu_i = \alpha_i / S$
2. $0 \leq \mu_i \leq 1$
3. $\Sigma_i \mu_i = 1$.
4. $\phi_i = (\rho_i - \mu_i)/\ (1 - \mu_i)$ and
5. $0 \leq \phi_i \leq 1$

To analyze consumer switching behaviors across various brands in a specific product category, a basic comprehensive intuitive model is necessary. The brand performance measures such as penetration, repeat rate, market share, and polarize index *etc*. are estimated using likelihood theory.

Let us assume there are '*k*' brands in the competitive environment for an item. Let $p_{ij}$ be the transition probability indicates the loyalty / disloyalty of customers. When $i = j$, the $p_{ii}$ denotes the loyalty probability for $i^{\text{th}}$ brand, *i.e.*, who, after being convinced by market share pressure, sticks with the same brand and when $i \neq j$, the $p_{ij}$ denotes disloyal customers who switch the brand *i* to *j*. Let the brand loyalty polarized index for the $i^{th}$ brand be $\phi_i$ and market share of the brand be $\mu_i$, where $0 \leq \phi_i,\ \mu_i \leq 1$, and $\Sigma_i \mu_i = 1$ $(i = 1, 2, \ldots, k)$. The transition probability matrix $P = ((p_{ij}))$ be the transition probability matrix defined in terms of polarized index $\phi_i$ and market share $\mu_j$ as

$$\left.\begin{array}{ll} p_{ii} = \ \phi_i + (1 - \phi_i)\mu_j & \text{for } i = j \\ p_{ij} = (1 - \phi_i)\mu_j & \text{for } i \neq j. \end{array}\right\} \qquad (1)$$

The repeated stationary purchase probabilities can be estimated using Chapman-Kolmogorov equation which is presented below.

**Theorem 1:** The stationary probability for $j^{th}$ brand is $\pi_j = R_j / \sum R_j$ where $R_j = \mu_j / (1 - \phi_j)$, $\phi_j$ is the polarized index and $\mu_j$ is the market share of $j^{\text{th}}$ brand satisfying (1).

**Proof:** Let $\phi_i$ be the loyalty polarization index and let $\mu_i$ be the market share of $i^{\text{th}}$ brand. Let ***P*** be the transition probability matrix constructed using (1). Let $\pi_0 = \mu$ be the initial probabilities for the brands.

For $j = 2$ brands, we have the Chapman-Kolmogorov equation $\pi_j = \pi_{j-1}.\ P$

$$[\pi_1 \quad \pi_2] = [\pi_1 \quad \pi_2]\begin{bmatrix} \phi_1 + (1-\phi_1)\mu_1 & (1-\phi_1)\mu_2 \\ (1-\phi_2)\mu_1 & \phi_2 + (1-\phi_2)\mu_2 \end{bmatrix} \tag{2}$$

It provides

$$\pi_1 = \pi_1[\phi_1 + (1-\phi_1).\mu_1] + \pi_2[(1-\phi_2).\mu_1]$$

$$\Rightarrow \{\pi_1(1-\phi_1) + \pi_2(1-\phi_2)\}\mu_1 = \pi_1(1-\phi_1) \tag{i}$$

Similarly,

$$\pi_2 = \pi_1[(1-\phi_1).\mu_2] + \pi_2[\phi_2 + (1-\phi_2).\mu_2$$

$$\Rightarrow \{\pi_1(1-\phi_1) + \pi_2(1-\phi_2)\}\mu_2 = \pi_2(1-\phi_2) \tag{ii}$$

from the equations, (i)/(ii)

$$\frac{\mu_1}{\mu_2} = \frac{\pi_1(1-\phi_1)}{\pi_2(1-\phi_2)}$$

$$\Rightarrow \frac{\pi_1}{\pi_2} = \frac{\dfrac{\mu_1}{(1-\phi_1)}}{\dfrac{\mu_2}{(1-\phi_2)}}$$

therefore,

$$\pi_1 = \frac{\left(\dfrac{\mu_1}{(1-\phi_1)}\right)}{\left(\dfrac{\mu_1}{(1-\phi_1)} + \dfrac{\mu_2}{(1-\phi_2)}\right)}$$

and

$$\pi_2 = \frac{\left(\dfrac{\mu_2}{(1-\phi_2)}\right)}{\left(\dfrac{\mu_1}{(1-\phi_1)} + \dfrac{\mu_2}{(1-\phi_2)}\right)}$$

for $j = 3$ brands we have,

$$[\pi_1 \quad \pi_2 \quad \pi_3] = [\pi_1 \quad \pi_2 \quad \pi_3]\begin{bmatrix} \phi_1 + (1-\phi_1)\mu_1 & (1-\phi_1)\mu_2 & (1-\phi_1)\mu_3 \\ (1-\phi_2)\mu_1 & \phi_2 + (1-\phi_2)\mu_2 & (1-\phi_2)\mu_3 \\ (1-\phi_3)\mu_1 & (1-\phi_3)\mu_2 & \phi_3 + (1-\phi_3)\mu_3 \end{bmatrix}$$

$$\pi_1 = \pi_1[\phi_1 + (1-\phi_1).\mu_1] + \pi_2[(1-\phi_2).\mu_1] + \pi_3[(1-\phi_3).\mu_1]$$

$$\pi_1[(1-\phi_1)(1-\mu_1)] = [\pi_2(1-\phi_2) + \pi_3(1-\phi_3)]\mu_1$$

$$\pi_1[(1-\phi_1)(1-\mu_1)] = [(1-\pi_1-\pi_3)(1-\phi_2) + (1-\pi_1-\pi_2)(1-\phi_3)]\mu_1$$

$$\Rightarrow \{\pi_1(1-\phi_1)+\pi_2(1-\phi_2)+\pi_3(1-\phi_3)\}\mu_1 = \pi_1(1-\phi_1) \quad \text{(a)}$$

similarly,

$$\{\pi_1(1-\phi_1)+\pi_2(1-\phi_2)+\pi_3(1-\phi_3)\}\mu_2 = \pi_2(1-\phi_2) \quad \text{(b)}$$

$$\{\pi_1(1-\phi_1)+\pi_2(1-\phi_2)+\pi_3(1-\phi_3)\}\mu_3 = \pi_3(1-\phi_3) \quad \text{(c)}$$

we have

$$(a)/(b) \Rightarrow \frac{\pi_1}{\pi_2} = \frac{\mu_1}{(1-\phi_1)} \div \frac{\mu_2}{(1-\phi_2)} \quad \text{(*a)}$$

$$(b)/(c) \Rightarrow \frac{\pi_2}{\pi_3} = \frac{\mu_2}{(1-\phi_2)} \div \frac{\mu_3}{(1-\phi_3)} \quad \text{(*b)}$$

and

$$(c)/(a) \Rightarrow \frac{\pi_3}{\pi_1} = \frac{\mu_3}{(1-\phi_3)} \div \frac{\mu_1}{(1-\phi_1)} \quad \text{(*c)}$$

From any of the two equations above, $\pi_1 : \pi_2 : \pi_3 = \frac{\mu_1}{(1-\phi_1)} : \frac{\mu_2}{(1-\phi_2)} : \frac{\mu_3}{(1-\phi_3)}$

therefore, $\pi_1 = \dfrac{\left(\frac{\mu_1}{(1-\phi_1)}\right)}{\left(\frac{\mu_1}{(1-\phi_1)}+\frac{\mu_2}{(1-\phi_2)}+\frac{\mu_3}{(1-\phi_3)}\right)}$, $\pi_2 = \dfrac{\left(\frac{\mu_2}{(1-\phi_2)}\right)}{\left(\frac{\mu_1}{(1-\phi_1)}+\frac{\mu_2}{(1-\phi_2)}+\frac{\mu_3}{(1-\phi_3)}\right)}$ and

$$\pi_3 = \frac{\left(\frac{\mu_3}{(1-\phi_3)}\right)}{\left(\frac{\mu_1}{(1-\phi_1)}+\frac{\mu_2}{(1-\phi_2)}+\frac{\mu_3}{(1-\phi_3)}\right)}.$$

In general, it can be expressed for $t = k$ brands

$$\pi_j = \frac{R_j}{\sum_{j=1}^{k} R_j}; \text{ where } R_j = \frac{\mu_j}{(1-\phi_j)}. \quad (3)$$

## 3. Analysis for the repeated purchase model

1. The repeated rate is an intuitive measure for brand loyalty, the higher repeated rate indicates larger loyal customers. The polarization index is also a measure of loyalty and where the repeat rate is standardized for market share.
2. When it is extended to '*k*' brands, the stationary probabilities for $j^{\text{th}}$ brand are

$$\pi_j = \frac{\left(\frac{\mu_j}{(1-\phi_j)}\right)}{\sum_{j=1}^{k}\left(\frac{\mu_j}{(1-\phi_j)}\right)}; \; j = 1,2,...,k \quad (4)$$

3. The empirical evidence indicated that brand loyalty transmits relatively slow and the market-share may change from purchase-to-purchase scenario. It can be noted that that

brand loyalty is assumed to be constant over some time horizon and market share is time-dependent.

4. The repeated purchase probabilities for the $j^{th}$ brand at time $t$ = 1, 2, 3, …, $M$ is

$$y_{j,t} = \phi_j y_{j,t-1} + \mu_{j,t}(1-\phi_j)\, y_{j,t-1} + \varepsilon_{j,t} \tag{5}$$

when all the brands are having equal loyalty $\phi_j = \phi$ Then the least squares estimate of $\phi$ can be obtained as

$$\varphi = \frac{\sum_{j=1}^{K}\sum_{t=2}^{M}\left(y_{jt-1} - \mu_{jt}\right)\left(y_{jt} - \mu_{jt}\right)}{\sum_{j=1}^{K}\sum_{t=2}^{M}\left(y_{jt-1} - \mu_{jt}\right)^2} \tag{6}$$

It can be noted that, when $\phi = 0$, there is no loyalty *i.e.*, when all the consumers switch frequently then $y_{j,t} = \mu_{j,t}$, and when $\phi = 1$, there is complete loyalty *i.e.*, when all consumers repeatedly purchase the same brand then $y_{j,t} = y_{j,t-1}$.

5. Consider the two brand cases *i.e.*, $k$ = 2 in a competitive market environment. Then

$$S = \sum_{t=2}^{M}\left\{y_{1,t} - \varphi_1 y_{1,t} - \mu_{1,t} + \mu_{1,t}(\varphi_1 y_{1,t-1} + \varphi_2 y_{2,t-1})\right\}^2 +$$

$$\sum_{t=2}^{M}\left\{y_{2,t} - \varphi_2 y_{2,t} - \mu_{2,t} + \mu_{2,t}(\varphi_1 y_{1,t-1} + \varphi_2 y_{2,t-1})\right\}^2$$

The resulting normal equations are,

$$\frac{\partial S}{\partial \varphi_1} = 0 \Rightarrow \varphi_1 \sum_{t=2}^{M} y_{1,t-1}^2\left[\mu_{1,t}^2 + \mu_{2,t}^2 - 2\mu_{1,t} + 1\right] + \varphi_2 \sum_{t=2}^{M} y_{1,t-1} y_{2,t-1}\left[\mu_{1,t}^2 + \mu_{2,t}^2 - \mu_{1,t} - \mu_{2,t}\right]$$

$$= \sum_{t=2}^{M} y_{1,t-1}\left[\mu_{1,t}\left(\mu_{1,t} - y_{1,t}\right) + \mu_{2,t}\left(\mu_{2,t} - y_{2,t}\right) - \left(\mu_{1,t} - y_{1,t}\right)\right]$$

$$\frac{\partial S}{\partial \varphi_2} = 0 \Rightarrow \varphi_2 \sum_{t=2}^{M} y_{2,t-1}^2\left[\mu_{1,t}^2 + \mu_{2,t}^2 - 2\mu_{2,t} + 1\right] + \varphi_1 \sum_{t=2}^{M} y_{1,t-1} y_{2,t-1}\left[\mu_{1,t}^2 + \mu_{2,t}^2 - \mu_{1,t} - \mu_{2,t}\right]$$

$$= \sum_{t=2}^{M} y_{2,t-1}\left[\mu_{1,t}\left(\mu_{1,t} - y_{1,t}\right) + \mu_{2,t}\left(\mu_{2,t} - y_{2,t}\right) - \left(\mu_{2,t} - y_{2,t}\right)\right]$$

The resulting solution to the normal equations $C\phi = B$, and $\phi = C^{-1} B$, where

$$C = \begin{bmatrix} \sum_{t=2}^{M} y_{1,t-1}^2(\mu_{1,t}^2 + \mu_{2,t}^2 - 2\mu_{1,t} + 1) & \sum_{t=2}^{M} y_{1,t-1} y_{2,t-1}(\mu_{1,t}^2 + \mu_{2,t}^2 - \mu_{1,t} - \mu_{2,t}) \\ \sum_{t=2}^{M} y_{1,t-1} y_{2,t-1}(\mu_{1,t}^2 + \mu_{2,t}^2 - \mu_{1,t} - \mu_{2,t}) & \sum_{t=2}^{M} y_{2,t-1}^2(\mu_{1,t}^2 + \mu_{2,t}^2 - 2\mu_{2,t} + 1) \end{bmatrix}, \text{and}$$

$$B = \begin{bmatrix} \sum_{t=2}^{M} y_{1,t-1}\left[\mu_{1,t}\left(\mu_{1,t} - y_{1,t}\right) + \mu_{2,t}\left(\mu_{2,t} - y_{2,t}\right) - \left(\mu_{1,t} - y_{1,t}\right)\right] \\ \\ \sum_{t=2}^{M} y_{2,t-1}\left[\mu_{1,t}\left(\mu_{1,t} - y_{1,t}\right) + \mu_{2,t}\left(\mu_{2,t} - y_{2t,}\right) - \left(\mu_{2,t} - y_{2,t}\right)\right] \end{bmatrix}$$

**Example 1:** Let B$_1$, B$_2$, B$_3$ and B$_4$ are four competitive brands for an item in the market with loyalties, market shares and with transition probability matrix *P* are

**Table 1: Loyalty and market shares for brands**

| | B$_1$ | B$_2$ | B$_3$ | B$_4$ |
|---|---|---|---|---|
| Polarized Index ($\phi_i$) | 0.30 | 0.20 | 0.50 | 0.60 |
| Market Share ($\mu_i$) | 0.30 | 0.10 | 0.40 | 0.20 |

The repeated purchase stationary probabilities are evaluated as

$$\pi_j = [\mu_j / (1 - \phi_j)] / \Sigma\, [\mu_j / (1 - \phi_j)].$$

**Table 2: Evaluation of repeated purchase stationary probabilities**

| | B$_1$ | B$_2$ | B$_3$ | B$_4$ |
|---|---|---|---|---|
| Polarized Index ($\phi_i$) | 0.30 | 0.20 | 0.50 | 0.60 |
| Market Share ($\mu_i$) | 0.30 | 0.10 | 0.40 | 0.20 |
| $R_j = \mu_j / (1-\phi_j)$ | 0.42857 | 0.125 | 0.8 | 0.5 |
| $\pi_j = R_j / \sum R_j$ | 0.2312 | 0.0675 | 0.4316 | 0.2697 |

The loyalty transition probabilities $p_{ij}$ can be evaluated

$p_{11}$ = 0.30 + (1-0.30).(0.30) = 0.51; $p_{22}$ = 0.20 + (1-0.20).(0.10) = 0.28;
$p_{33}$ = 0.50 + (1-0.50).(0.40) = 0.70; $p_{44}$ = 0.60 + (1-0.60).(0.20) = 0.68;

The disloyalty transition probabilities $p_{ij}$ can be evaluated

$p_{12}$ = (1-0.30).(0.10)=0.07; $p_{13}$ = (1-0.30).(0.40)=0.28; $p_{14}$ = (1-0.30).(0.20) = 0.14;
$p_{21}$ = (1-0.20).(0.30)=0.24; $p_{23}$ = (1-0.20).(0.40)=0.32; $p_{24}$ = (1-0.20).(0.20) = 0.16;
$p_{31}$= (1-0.50).(0.30)=0.15; $p_{32}$ = (1-0.50).(0.10)=0.05; $p_{34}$ = (1-0.50).(0.20) = 0.1;
$p_{41}$ = (1-0.60).(0.30)=0.12; $p_{42}$= (1-0.60).(0.10)=0.04; $p_{43}$ = (1-0.60).(0.40) = 0.16;

The resulting transition matrix is

$$P = \begin{bmatrix} 0.51 & 0.07 & 0.28 & 0.14 \\ 0.24 & 0.28 & 0.32 & 0.16 \\ 0.15 & 0.05 & 0.70 & 0.10 \\ 0.12 & 0.04 & 0.16 & 0.68 \end{bmatrix}$$

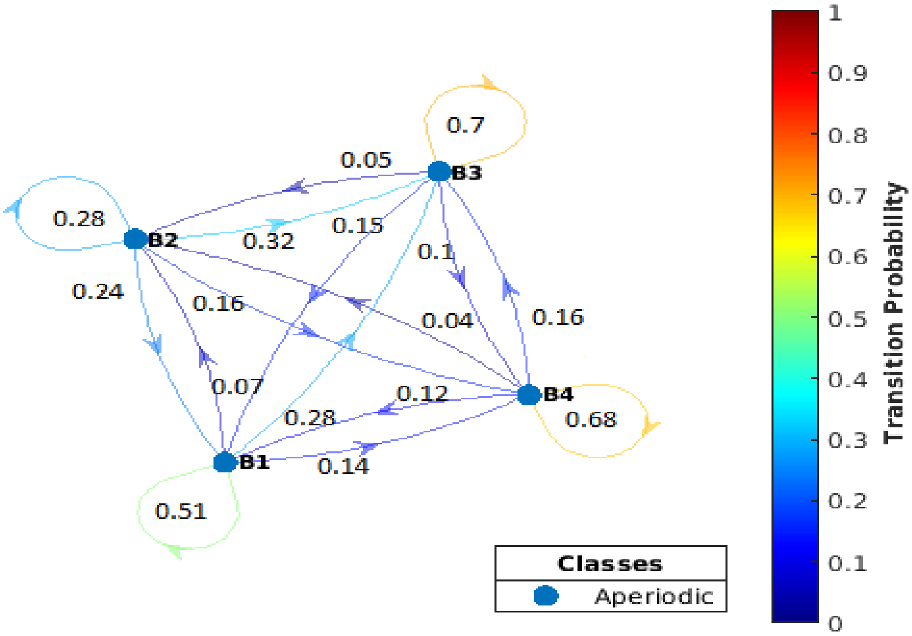


**Figure 1: Markov diagram**

It can be noted and stated that the Markov property exhibits while switching from one brand to the other, a customer is keeping in his or her memory only the loyalty of the brand he was using just before using the current brand and not keeping in memory the loyalty of all previously used brands, all states (brands) are communicated, implying that brands are essential, that consumers observe repeated purchases and switching among brands, so the transition probability matrix is irreducible. The states of the transition probability matrix are recurrent and have a periodicity of 1. As a consequence, the Markov chain is an ergodic (regular) Markov chain, culminating in a unique stationary distribution. The expected first hitting times (the first arrival from starting one point to the other point after how many transitions or purchase occasion *i.e.*, shifting one brand to the other after how many time transitions) for each state are

**Table 3: Expected first hitting times**

| B$_1$ | B$_2$ | B$_3$ | B$_4$ |
|---|---|---|---|
| - | 18.7143 | 4.0625 | 8.1964 |
| 6.0 | - | 3.8839 | 8.0179 |
| 6.75 | 19.2857 | - | 8.7679 |
| 7.25 | 19.7857 | 5.1339 | - |

Let $\pi_0$ = [0.30 0.10 0.40 0.20] be the vector of initial brands purchase probabilities.

The subsequent repeat purchase frequency rates can be evaluated as follows: $\pi_j = \pi_{j-1}. P$.

**Table 4: The iterative repeated purchase stationary probabilities**

| Iteration | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ |
|---|---|---|---|---|
| 1 | 0.2610 | 0.0770 | 0.4280 | 0.2340 |
| 2 | 0.2439 | 0.0706 | 0.4348 | 0.2508 |
| 3 | 0.2366 | 0.0686 | 0.4353 | 0.2594 |
| 4 | 0.2336 | 0.0679 | 0.4344 | 0.2641 |
| 5 | 0.2323 | 0.0677 | 0.4335 | 0.2666 |
| 6 | 0.2317 | 0.0675 | 0.4328 | 0.2680 |
| 7 | 0.2315 | 0.0675 | 0.4323 | 0.2687 |
| 8 | 0.2313 | 0.0675 | 0.4320 | 0.2692 |
| 9 | 0.2313 | 0.0675 | 0.4318 | 0.2694 |
| 10 | 0.2312 | 0.0674 | 0.4317 | 0.2696 |
| 11 | 0.2312 | 0.0674 | 0.4317 | 0.2696 |
| 12 | 0.2312 | 0.0674 | 0.4316 | 0.2697 |
| 13 | 0.2312 | 0.0674 | 0.4316 | 0.2697 |
| 14 | 0.2312 | 0.0674 | 0.4316 | 0.2697 |

For the given initial brand shares, polarized Index and market shares of the four brands, and the subsequent stationary brand shares can be evaluated by successive application of Chapman Kolmogorov equation. The consumer's' propensity to choose a loyal brand in a long period of time with purchase probabilities attained with values as the equilibrium states together $\pi = [0.2312, 0.0674, 0.4316, 0.2697]$ (or alternatively the equilibrium purchase probabilities can be obtained from the theorem) *i.e.*, the brand $B_3$ is more likely to be repeatedly purchased with the probability 0.4316 among the other competitive brands in the market. In other words, the concentration of the repeated purchase of the brands is directly proportional to the steady state probabilities. (The higher the steady state probability value indicates more likely to be repeated purchases and vice-versa).

## 4. Discussion

The brand performance indicators of the Dirichlet model are used to construct a more detailed statistical model for the transition probability matrix. In brand switching and repeat purchase analysis, this model is used to investigate Markovian characteristics. The least square principle is used to estimate the transition model parameters. The stationary probabilities are derived for each state of the Markov chain and analysis is illustrated with an example.

## Acknowledgements

The authors express their thankfulness to the Editor and the Reviewers for their valuable suggestions and comments that helped to improve the earlier version of the manuscript.

## References

Abel P. Jeuland, Frank M. Bass. and Gordon P. Wright (1980). A multibrand stochastic model compounding heterogeneous Erlang timing and multinomial choice processes. *Operations Research*, **28,** 255-277.

Aypar U. and Tuncay C. (2014). Analysis of brand loyalty with Markov Chains. *First International Joint Symposium on Business Administration – Challenges for Business Administrators in the New Millennium, Turkey*, 583 – 591.

Bound J. (2009). User's Guide to DIRICHLET. *Marketing Bulletin*,1-20

Goodhardt G. J., Ehrenberg A. S. C. and Chatfield C. (1984). The Dirichlet: A comprehensive model of buying behaviour. *Journal of the Royal Statistical Society Series A*, **147**, 621-655.

Kashyap M. P., and Bhattacharjee D. (2015). An empirical of brand switching behavior of rural and urban consumer: A Markovian approach. *Proceedings of International Conference on Frontiers in Mathematics*, 108-112.

Lipstein B. A. (1965). Mathematical model of consumer behavior. *Journal of Marketing Research*, **2**, 259-265.

Richard A. Colombo, and Donald G. Morrison (1989). Note-A Brand Switching Model with Implications for Marketing Strategies. *Marketing Science*, INFORMS, **8**, 89-99.

Rungie C. and Goodhardt G. (2004). Calculation of theoretical brand performance measures from the parameters of the Dirichlet model. *Marketing Bulletin*, **15**, Technical Note 2, 1-19.

Sulaimon, Mutiu. O. and Oyenekan Dotun F. (2015). Application of Markov chain in forecasting: A study of customers' brand loyalty for mobile phones. *European Academic Research*, **3**, 7190 – 7205.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 11-22

# Calibration Estimator in Two Stage Sampling Using Double Sampling Approach when Study Variable is Inversely Related to Auxiliary Variable

**Ankur Biswas[1], Kaustav Aditya[1], U.C. Sud[2] and Pradip Basak[3]**
[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Former Director of ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[3]*Department of Agricultural Statistics, Uttar Banga Krishi Vishwavidyalaya, West Bengal*

Received: 05 April 2021; Revised: 08 December 2021; Accepted: 20 January 2022

## Abstract

The calibration approach is a popular technique for incorporating auxiliary information for estimation of population parameters in survey sampling. In general, the Calibration Approach assumes the availability of population-level auxiliary information. On the contrary, in large scale surveys, it is often the case that population-level data on auxiliary variable is not available, but it is relatively inexpensive to collect. In the present article, in case of non-availability of population-level relatively inexpensive data on auxiliary variable under two stage sampling, we developed product type calibration estimator of the finite population total using double sampling approach along with the sampling variance and variance estimator. The study variable is assumed to be inversely related with the auxiliary variable. Proposed product type calibration estimator was evaluated through a simulation study which showed that the proposed product type calibration estimator was performing efficiently over traditional Narain-Horvitz-Thompson type expansion estimator as well as product estimator of the finite population total in case of two stage sampling involving two phases at both the stages.

*Key words*: Auxiliary information; Calibration; Design weights; Product estimator; Simulation; Double sampling.

## 1. Introduction

The calibration approach was originally suggested by Deville and Särndal (1992). It is a most widely used techniques combining auxiliary information for estimation of unknown finite population parameters of the character under study efficiently. In calibration approach, initial design weights would be converted to calibrated weights which is product of a calibration factor with the initial design weight. Following Deville and Särndal (1992), plenty of work has been carried out in the calibration estimation *i.e.* Singh *et al.* (1998, 1999), Wu and Sitter (2001), Sitter and Wu (2002), Kott (2006), Estevao and Särndal (2006), *etc*. (see Kim and Park (2010) and Särndal (2007) for comprehensive review of calibration approach).

In various medium to large scale surveys, two stage sampling is followed since at most situations it is very often the case that the sampling frame is often unavailable or it

Corresponding Author: Ankur Biswas
Email: ankur.biswas@icar.gov.in

could be too expensive to construct one. Under this sampling design, first, groups of elements are selected which are called as primary stage units (PSU) and, then, a sample of basic elements which are called as secondary stage units (SSU) are selected from each selected PSU. For example, in agricultural surveys, villages can be selected as PSU and farmers can be selected as SSU. Sukhatme *et al.* (1984) suggested several estimators of the finite population parameters using auxiliary information in two stage sampling. Särndal *et al.* (1992) considered three different situations concerning availability of complex auxiliary information under two stage sampling and discussed extensively on ratio and regression estimators under such situations. The calibration estimation under availability of complex auxiliary information under two stage sampling has been discussed by several authors such as Aditya *et al.* (2016a, 2016b), Mourya *et al.* (2016), Aditya *et al.* (2017), Basak *et al.* (2017), Salinas *et al.* (2018) and Biswas *et al.* (2020) etc.

In surveys, it is often the case that there exist certain auxiliary variables which are inversely related to the character under study. For example, in household based surveys, the marketable surplus is inversely related to family consumption of seed, feed etc. In the past, the product estimator (Murthy, 1964) was used as an efficient alternative to the traditional estimators. In such a situation, the usual methodology for calibration estimation may not fit in. Sud *et al.* (2014 a, b) and Biswas *et al.* (2020) proposed calibration estimation procedures for finite population total under uni-stage equal probability sampling and two stage sampling respectively, when a character under study is inversely related to the available auxiliary variable.

Generally, in calibration approach, it is assumed that population-level auxiliary information is available. On the contrary, population-level data on auxiliary variable is not available in practice, but relatively inexpensive to collect. Under this scenario, double sampling approach serves as a feasible solution for the estimation of finite population parameters. Double sampling has generated extensive research interests. For example Rao (1973), Hidiroglou *et al.* (2009), Haziza *et al.* (2011), Sinha *et al.* (2016), Arnab (2017), *etc*. In this present study, in case of non-availability of population-level relatively inexpensive data on auxiliary variable under two stage sampling, an attempt has been made to develop calibration estimation procedure for estimation of finite population total using double sampling approach when character under study is inversely related to the available auxiliary variable. In Section 2, we give a brief of the product type calibration estimators of finite population total under two stage sampling as proposed by Biswas *et al.* (2020). In Section 3, calibration estimators have been proposed in case of two stage sampling using double sampling approach when there was unavailability of population level auxiliary information at the SSU level and the character under study is inversely related to auxiliary variable. The statistical properties of the proposed estimators are studied empirically through a simulation study. Section 4 provides the technical details of the simulation study and simulation results. Concluding remarks are given in Section 5.

## 2. Calibration estimators under two stage sampling when character under study is inversely related to available auxiliary information

In this section, first, we briefly describe two stage sampling design along with two different calibration estimators under two stage sampling under the assumption that the character under study is inversely related to available auxiliary information as proposed by Biswas *et al.* (2020).

Let, the finite population under consideration and the corresponding character under study is denoted by *U* and *Y*. Population *U* is grouped into *N* different PSUs such that $U_I = \{1,...,i,...,N\}$ and *i*th PSU consists of *Mi* SSUs such that $U_i = \{1,...,k,...,M_i\}$, $i \in U_I$. Thus, we have $U = \bigcup_{i=1}^{N} U_i$ and total number of SSUs in *U* is $M_0 = \sum_{i=1}^{N} M_i$. Under two stage sampling, a sample of *n* PSUs ($s_I$) is drawn from $U_I$ at stage one. First and second order inclusion probabilities at the PSU level are $\pi_{Ii} = P(i \in s_I)$ and $\pi_{Iij} = P(i, j \in s_I)$. A second stage sample ($s_i$) of *mi* SSUs is drawn from $U_i$ provided, at the first stage, the ith PSU ($U_i$) is selected. First and second order inclusion probabilities at the SSU level are $\pi_{k/i} = P(k \in s_i / i \in s_I)$ and $\pi_{kl/i} = P(k, l \in s_i / i \in s_I)$. In the second stage of sampling, invariance and independence property is followed. The final sample of SSUs is denoted as, $s = \bigcup_{i=1}^{s_I} s_i$. Let, $y_{ik}$ denotes the observation of the study variable from *k*th SSU in *i*th PSU and it is observed for all the sampled SSUs. The parameter of interest is the population total $t_y = \sum_{i=1}^{N}\sum_{k=1}^{M_i} y_{ik} = \sum_{i=1}^{N} t_{yi}$, where $t_{yi} = \sum_{k=1}^{M_i} y_{ik}$ = *i*th PSU total. Usual Narain-Horvitz-Thompson estimator for population total is given by

$$\hat{t}_{y\pi} = \sum_{i=1}^{n} a_{Ii} \sum_{k=1}^{m_i} \left( a_{k/i} y_{ik} \right) = \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} y_{ik} \tag{1}$$

where, the design weights are given as

$$a_{ik} = a_{Ii}.a_{k/i},\ a_{Ii} = 1/\pi_{Ii},\ \forall i \in s_I \text{ and } a_{k/i} = 1/\pi_{k/i},\ \forall k \in s_i \text{ and } i \in s_I.$$

Biswas *et al.* (2020) proposed product type calibration estimators of population total for two situations under two stage sampling design as per Särndal *et al.* (1992) as mentioned below:

Case 1: Population level complete auxiliary information is available at the SSU level.
Case 2: Population level auxiliary information is available only for the selected Primary Stage Units (PSU).

The product type calibration estimators of population total (Biswas *et al.*, 2020) under these two cases of two stage sampling as stated above were given by

$$\hat{t}_{yCP1} = \sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{1ik} y_{ik} = \left( \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} y_{ik} \right) \left( \sum_{i=1}^{N}\sum_{k=1}^{M_i} x_{ik}^{-1} \right) \Bigg/ \left( \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} x_{ik}^{-1} \right) \text{ and}$$

$$\hat{t}_{yCP2} = \sum_{i=1}^{n} a_{Ii} \sum_{k=1}^{m_i} w_{2ik} y_{ik} = \sum_{i=1}^{n} \left[ a_{Ii} \left( \sum_{k=1}^{m_i} a_{k/i} y_{ik} \right) \left( \sum_{i=1}^{M_i} x_{ik}^{-1} \right) \Bigg/ \left( \sum_{i=1}^{m_i} a_{k/i} x_{ik}^{-1} \right) \right]$$

where,

$$w_{1ik} = a_{ik} \left( \sum_{i=1}^{N}\sum_{k=1}^{M_i} x_{ik}^{-1} \Bigg/ \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} x_{ik}^{-1} \right),\ \forall k = 1, 2, ..., m_i \text{ and}$$

$$w_{2ik} = a_{k/i}\left(\sum_{k=1}^{M_i} x_{ik}^{-1} \Big/ \sum_{k=1}^{m_i} a_{k/i}\, x_{ik}^{-1}\right),\ \forall k = 1,2,...,m_i\ .$$

## 3. Proposed calibration estimator using double sampling approach in both stages of two stage sampling

The double sampling was first given by Neyman (1938) which is generally used when the information on auxiliary variable is lacking, but comparatively low-cost to obtain. Information on auxiliary variable shall be obtained by selecting a larger preliminary sample. Further, sub-sample is taken to observe the charater under study. In the present study, we have developed calibration estimators using double sampling approach in two stage sampling for the situations when of population level auxiliary information ($x_{ik}$) was unavailabile at SSU level.

Biswas *et al.* (2020) developed product type calibration estimator under two stage sampling assuming the SSU level auxiliary variable is inversely related to the characteristics under study and $\sum_{i=1}^{N}\sum_{k=1}^{M_i} x_{ik}^{-1}$ is already known. Under the present situation, it is assumed that a correct value of $\sum_{i=1}^{N}\sum_{k=1}^{M_i} x_{ik}^{-1}$ is unavailable since there was there was unavailability of population level auxiliary information ($x_{ik}$). We consider double sampling approach under this scenario. First, a large first phase sample ($s_I'$) of $n'$ PSUs is selected from the population of $N$ PSUs ($U_I$) following a sampling design $p_I'(.)$. The design weight for $i^{\text{th}}$ PSU is given by $a_{Ii}' = 1/\pi_{Ii}'$, where $\pi_{Ii}' = P(i \in s_I')$ is the known first phase first order inclusion probability of $i^{\text{th}}$ PSU. Under SRSWOR, $\pi_{Ii}' = n'/N$. In the second stage of the first phase sampling, from each of the $i^{\text{th}}$ selected PSU, $i \in s_I'$, a sub-sample ($s_i'$) of $m_i'$ SSUs is selected from $M_i$ population SSUs ($U_i$) by a sampling design $p_{k/s_I'}(.)$. The design weight for $k^{\text{th}}$ SSU provided $i^{\text{th}}$ PSU is already selected can be given as $a_{k/i}' = 1/\pi_{k/i}'$, where $\pi_{k/i}' = P(k \in s_i' / i \in s_I')$ is first phase inclusion probability of $k^{\text{th}}$ SSU and under SRSWOR it is given by $\pi_{k/i}' = m_i'/M_i$. The observation on auxiliary variable $x_{ik}$ is taken from the $k^{\text{th}}$ SSU in $i^{\text{th}}$ PSU.

In the second phase, a smaller sub-sample ($s_I$) of $n$ PSUs is drawn from $s_I'$ by a sampling design $p_I(.)$. The design weight of the $i^{\text{th}}$ PSU is $a_{Ii/s_I'} = 1/\pi_{Ii/s_I'}$, where $\pi_{Ii/s_I'} = P(i \in s_I / s_I')$ is the second phase conditional inclusion probability of $i^{\text{th}}$ PSU, given $s_I'$, and under SRSWOR, $\pi_{Ii/s_I'} = n/n'$. In the second stage of second phase sampling, from the $i^{\text{th}}$ selected PSU, $i \in s_I$, a smaller sub-sample $s_i$ of size $m_i$ SSUs is selected from $m_i'$ first phase SSUs by a sampling design $p_{k/s_I}(.)$. The sampling weight for $k^{\text{th}}$ SSU is given by

$a_{k/s_i',i} = 1/\pi_{k/s_i',i}$, where $\pi_{k/s_i',i} = P(k \in s_i / k \in s_i', i \in s_I)$ is the second phase conditional inclusion probability of $k^{\text{th}}$ SSU and under SRSWOR it is given by $\pi_{k/s_i',i} = m_i / m_i'$. The observations on the character under study, $y_{ik}$, and auxiliary variable, $x_{ik}$, are taken from the $k^{\text{th}}$ sampled SSUs in $i^{\text{th}}$ selected PSU.

In this study, an attempt has been made to improve the traditional design weighted Narain-Horvitz-Thompson (NHT) (Narain, 1951; Horvitz and Thompson, 1952) type expansion estimator for population total ($t_y$) under two stage sampling following double sampling at both the stages which is given by

$$\hat{t}_{y\pi} = \sum_{i=1}^{n} a_{Ii} \sum_{k=1}^{m_i} a_{k/i} y_{ik} = \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} y_{ik} \qquad (2)$$

where, $a_{ik} = a_{Ii}\, a_{k/i} = \left(a_{Ii}'\, a_{i/s_I'}\right)\left(a_{k/i}'\, a_{k/s_i'i}\right)$ is the total sampling weight of $k^{\text{th}}$ SSU in $i^{\text{th}}$ selected PSU in the second phase sample, which reduces to $a_{ik} = \left(\frac{N}{n'}\frac{n'}{n}\right)\left(\frac{M_i}{m_i'}\frac{m_i'}{m_i}\right) = \frac{NM_i}{nm_i}$ under SRSWOR at all stages and phases.

Proposed calibration estimator of the population total ($t_y$) in case of two stage sampling following double sampling approach is given by

$$\hat{t}_{yCPd} = \sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ikd}\, y_{ik} \qquad (3)$$

where $w_{ikd}$ is the calibration weight under double sampling corresponding to the total sampling weight $a_{ik}$.

We obtained calibration weights $w_{ikd}$ by minimizing the Chi-square type distance function $\sum_{i=1}^{n}\sum_{k=1}^{m_i} \frac{\left(w_{ikd} - a_{ik}\right)^2}{a_{ik} q_{ik}}$ subject to the constraint $\sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ikd}\, x_{ik}^{-1} = \sum_{i=1}^{n'}\sum_{k=1}^{m_i'} a_{ik}'\, x_{ik}^{-1}$, where, $a_{ik}' = a_{Ii}'\, a_{k/i}'$. Using Lagrangian multiplier technique, the new calibrated weight is given by

$$w_{ikd} = a_{ik} + a_{ik} q_{ik} x_{ik}^{-1} \left[ \frac{\sum_{i=1}^{n'}\sum_{k=1}^{m_i'} a_{ik}'\, x_{ik}^{-1} - \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik}\, x_{ik}^{-1}}{\sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik} q_{ik} x_{ik}^{-2}} \right] \quad \forall\, k = 1,2,\ldots,m_i,\ i \in s_I . \qquad (4)$$

Using the results of the Equation (4) in (3) and considering $q_{ik} = x_{ik}$, we have, therefore, proved the following result.

**Theorem 1:** Following double sampling approach under two stage sampling, the proposed product type calibration estimator of population total is given as

$$\hat{t}_{yCPd} = \sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ikd}\, y_{ik} = \frac{\left(\sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik}\, y_{ik}\right)\left(\sum_{i=1}^{n'}\sum_{k=1}^{m'_i} a'_{ik}\, x_{ik}^{-1}\right)}{\left(\sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik}\, x_{ik}^{-1}\right)}, \tag{5}$$

where, proposed calibration weights corresponding to respective design weights are

$$w_{ikd} = a_{ik}\left(\sum_{i=1}^{n'}\sum_{k=1}^{m'_i} a'_{ik}\, x_{ik}^{-1} \Big/ \sum_{i=1}^{n}\sum_{k=1}^{m_i} a_{ik}\, x_{ik}^{-1}\right), \ \forall\, k = 1,2,\ldots,m_i,\ i \in s_I\ .$$

**Corollary 1**: Under SRSWOR at both the stages of two stage sampling, the proposed product type calibration estimator reduces to

$$\hat{t}_{yCPd} = \left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{k=1}^{m_i} y_{ik}\right)\left(\frac{N}{n'}\sum_{k=1}^{n'}\frac{M_i}{m'_i}\sum_{k=1}^{m'_i} x_{ik}^{-1}\right)\Big/\left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{k=1}^{m_i} x_{ik}^{-1}\right). \tag{6}$$

Usual product estimator of two stage sampling using double sampling approach is given by

$$\hat{t}_{yPd} = \left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{k=1}^{m_i} y_{ik}\right)\left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{n_i}\sum_{k=1}^{n_i} x_{ik}\right)\Big/\left(\frac{N}{n'}\sum_{i=1}^{n'}\frac{M_i}{m'_i}\sum_{k=1}^{m'_i} x_{ik}\right). \tag{7}$$

The approximate sampling variance of the proposed estimator $\hat{t}_{yCPd}$ is given by

$$\begin{aligned} AV(\hat{t}_{yCPd}) = {} & \sum_{i=1}^{N}\sum_{j=1}^{N}\Delta_{Iij}\frac{t_{E_{i1}}}{\pi_{Ii}}\frac{t_{E_{j1}}}{\pi_{Ij}} + \sum_{i=1}^{N}\frac{1}{\pi_{Ii}}\sum_{k=1}^{M_i}\sum_{l=1}^{M_i}\Delta_{kl/i}\frac{E_{k/i}}{\pi_{k/i}}\frac{E_{l/i}}{\pi_{l/i}} \\ & + R_1^2\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta'_{Iij}\frac{t_{x^{-1}i}}{\pi'_{Ii}}\frac{t_{x^{-1}j}}{\pi'_{Ij}} + \sum_{i=1}^{N}\frac{1}{\pi'_{Ii}}\sum_{k=1}^{M_i}\sum_{l=1}^{M_i}\Delta'_{kl/i}\frac{x_{ik}^{-1}}{\pi'_{k/i}}\frac{x_{il}^{-1}}{\pi'_{l/i}}\right] \end{aligned} \tag{8}$$

where, $\Delta'_{Iij} = (\pi'_{Iij} - \pi'_{Ii}\pi'_{Ij})$, $\Delta'_{kl/i} = \pi'_{kl/i} - \pi'_{k/i}\pi'_{l/i}$, $t_{x^{-1}i} = \sum_{k=1}^{M_i} x_{ik}^{-1}$ and

$$R_1 = \left(\sum_{i=1}^{N}\sum_{k=1}^{M_i} y_{ik}\right)\Big/\left(\sum_{i=1}^{N}\sum_{k=1}^{M_i} x_{ik}^{-1}\right).$$

Under SRSWOR design at all the stages and phases, approximate variance of the proposed product type calibration estimator reduces to

$$\begin{aligned} AV(\hat{t}_{yCPd}) = N^2\Bigg[ & \left(\frac{1}{n'}-\frac{1}{N}\right)S_{by}^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left\{S_{by}^2 + R_1^2 S_{bx^{-1}}^2 - 2R_1 S_{byx^{-1}}\right\} \\ & + \frac{1}{nN}\sum_{i=1}^{N} M_i^2\left\{\left(\frac{1}{m'_i}-\frac{1}{M_i}\right)S_{iy}^2 + \left(\frac{1}{m_i}-\frac{1}{m'}\right)\left(S_{iy}^2 + R_1^2 S_{ix^{-1}}^2 - 2R_1 S_{iyx^{-1}}\right)\right\}\Bigg] \end{aligned} \tag{9}$$

where,

$$\bar{Y}_{N.}=\frac{1}{N}\sum_{i=1}^{N}M_i\bar{Y}_{i.},\ \bar{Y}_{i.}=\frac{1}{M_i}\sum_{k=1}^{M_i}y_{ik},\ \bar{X}_{(-1)\,\mathrm{N}.}=\frac{1}{N}\sum_{i=1}^{N}M_i\bar{X}_{(-1)i.},\ \bar{X}_{(-1)i.}=\frac{1}{M_i}\sum_{k=1}^{M_i}x_{ik}^{-1},$$

$$S_{by}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(M_i\bar{Y}_{i.}-\bar{Y}_{N.}\right)^2,\qquad S_{bx^{-1}}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(M_i\bar{X}_{(-1)i.}-\bar{X}_{(-1)\,\mathrm{N}.}\right)^2,$$

$$S_{byx^{-1}}=\frac{1}{N-1}\sum_{i=1}^{N}\left(M_i\bar{Y}_{i.}-\bar{Y}_{N.}\right)\left(M_i\bar{X}_{(-1)i.}-\bar{X}_{(-1)\,\mathrm{N}.}\right),\ S_{iy}^2=\frac{1}{M_i-1}\sum_{k=1}^{M_i}\left(y_{ik}-\bar{Y}_{i.}\right)^2,$$

$$S_{ix^{-1}}^2=\frac{1}{M_i-1}\sum_{i=1}^{M_i}\left(x_{ik}^{-1}-\bar{X}_{(-1)i.}\right)^2\ \text{ and }\ S_{iyx^{-1}}=\frac{1}{M_i-1}\sum_{i=1}^{M_i}\left(y_{ik}-\bar{Y}_{i.}\right)\left(x_{ik}^{-1}-\bar{X}_{(-1)i.}\right).$$

Following Särndal *et al.* (1992), the estimator of variance of proposed product type calibration estimator can be written as

$$\hat{V}(\hat{t}_{yCPd})=\sum_{i=1}^{n}\sum_{j=1}^{n}d_{Iij}\frac{\hat{t}_{E_{i1}}}{\pi_{Ii}}\frac{\hat{t}_{E_{j1}}}{\pi_{Ij}}+\sum_{i=1}^{n}\frac{1}{\pi_{Ii}}\sum_{k=1}^{m_i}\sum_{l=1}^{m_i}d_{kl/i}\frac{e_{k/i}}{\pi_{k/i}}\frac{e_{l/i}}{\pi_{l/i}}$$
$$+\hat{R}_1^2\left[\sum_{i=1}^{n'}\sum_{j=1}^{n'}d'_{Iij}\frac{\hat{t}_{x^{-1}i}}{\pi'_{Ii}}\frac{\hat{t}_{x^{-1}j}}{\pi'_{Ij}}+\sum_{i=1}^{n'}\frac{1}{\pi'_{Ii}}\sum_{k=1}^{m'_i}\sum_{l=1}^{m'_i}d'_{kl/i}\frac{x_{ik}^{-1}}{\pi'_{k/i}}\frac{x_{il}^{-1}}{\pi'_{l/i}}\right]\tag{10}$$

where, $\hat{R}_1=\left(\sum_{i=1}^{n}a_{Ii}\sum_{k=1}^{m_i}a_{k/i}\,y_{ik}\right)\Big/\left(\sum_{i=1}^{n}a_{Ii}\sum_{k=1}^{m_i}a_{k/i}\,x_{ik}^{-1}\right),\ \hat{t}_{x^{-1}i}=\sum_{k=1}^{m_i}\frac{x_{ik}^{-1}}{\pi_{k/i}}$ ,

$$d'_{Iij}=\frac{(\pi'_{Iij}-\pi'_{Ii}\pi'_{Ij})}{\pi'_{Iij}}\ \text{ and }\ d'_{kl/i}=\frac{\pi'_{kl/i}-\pi'_{k/i}\pi'_{l/i}}{\pi'_{kl/i}}.$$

Under SRSWOR design at all the stages and phases, it reduces to

$$\hat{V}(\hat{t}_{yCPd})=N^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\hat{S}_{by}^2+\left(\frac{1}{n}-\frac{1}{n'}\right)\left\{\hat{S}_{by}^2+\hat{R}_1^2\hat{S}_{bx^{-1}}^2-2\hat{R}_1\hat{S}_{byx^{-1}}\right\}\right.$$
$$\left.+\frac{1}{nN}\sum_{i=1}^{n}M_i^2\left\{\left(\frac{1}{m'_i}-\frac{1}{M_i}\right)\hat{S}_{iy}^2+\left(\frac{1}{n_i}-\frac{1}{m'_i}\right)\left(\hat{S}_{iy}^2+\hat{R}_1^2\hat{S}_{ix^{-1}}^2-2\hat{R}_1\hat{S}_{iyx^{-1}}\right)\right\}\right]\tag{11}$$

where, $\hat{R}_1=\left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{k=1}^{m_i}y_{ik}\right)\Big/\left(\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{k=1}^{m_i}x_{ik}^{-1}\right),\quad \hat{S}_{by}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{y}_{i.}-\bar{y}_{n.}\right)^2,$

$$\hat{S}_{bx^{-1}}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{x}_{(-1)i.}-\bar{x}_{(-1)n.}\right)^2,\quad \hat{S}_{byx^{-1}}=\frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{y}_{i.}-\bar{y}_{n.}\right)\left(M_i\bar{x}_{(-1)i.}-\bar{x}_{(-1)n.}\right),$$

$$\hat{S}_{iy}^2=\frac{1}{m_i-1}\sum_{i=1}^{m_i}\left(y_{ik}-\bar{y}_{i.}\right)^2,\quad \hat{S}_{ix^{-1}}^2=\frac{1}{m_i-1}\sum_{i=1}^{m_i}\left(x_{ik}^{-1}-\bar{x}_{(-1)i.}\right)^2,$$

$$\hat{S}_{iyx^{-1}}=\frac{1}{m_i-1}\sum_{i=1}^{m_i}\left(y_{ik}-\bar{y}_{i.}\right)\left(x_{ik}^{-1}-\bar{x}_{(-1)i.}\right),\ \bar{y}_{i.}=\frac{1}{m_i}\sum_{k=1}^{m_i}y_{ik},\quad \bar{y}_{n.}=\frac{1}{n}\sum_{i=1}^{n}M_i\bar{y}_{i.},$$

$$\bar{x}_{(-1)i.}=\frac{1}{m_i}\sum_{k=1}^{m_i}x_{ik}^{-1},\quad \bar{x}_{(-1)n.}=\frac{1}{n}\sum_{i=1}^{n}M_i\bar{x}_{(-1)i.}.$$

## 4. Simulation study

In order to evaluate the statistical performance of proposed product type calibration estimators, a simulation study was carried out following double sampling approach in two stage sampling. SRSWOR is used for sample selection at both stages and the size of the PSU and the corresponding SSUs were assumed to be fixed. First, a finite population of 5000 units were generated in the similar way of Biswas *et al.* (2020). Let, number of PSU, *N*=50 and PSU size, $M_i$=100. The finite population was generated from the following model as

$$y_k = \beta x_k^{-1} + e_k,\ k = 1,\dots,M_0 \tag{12}$$

where, $M_0 = \sum_{i=1}^{N} M_i$ .

The distribution of auxiliary variable was considered as normal distribution as $x_k \sim N(5,\ 1)$ and the random errors, $e_k$, $k\ =\ 1,\dots,M_0$, are taken from normal distribution as $e_k \sim N(0,\ \sigma^2 x_k^{-1})$. The value of β has been fixed as 20. Four different values for $\sigma^2$ as 0.25, 1.0, 2.0 and 5.0 are taken. In this way, four sets of population have been generated denoted by **Set 1**, **2**, **3** and **4**, with different values of correlation coefficient between *Y* and *X* as -0.91, -0.85, -0.78 and -0.64 respectively. Then, from each of the study population sets, we have selected a total of 10000 different samples of different sizes of double samples under two stage sampling were drawn from the populations sets as given below

| | |
|---|---|
| $n'$ =10, $m'_i$ =20, $n$ =4, $m_i$ =6 | $n'$ =10, $m'_i$ =25, $n$ =4, $m_i$ =8 |
| $n'$ =15, $m'_i$ =25, $n$ =6, $m_i$ =8 | $n'$ =15, $m'_i$ =30, $n$ =6, $m_i$ =10 |
| $n'$ =20, $m'_i$ =30, $n$ =8, $m_i$ =10 | $n'$ =20, $m'_i$ =40, $n$ =8, $m_i$ =12 |
| $n'$ =25, $m'_i$ =40, $n$ =10, $m_i$ =12 | $n'$ =25, $m'_i$ =50, $n$ =10, $m_i$=15 |

Developed product type calibration estimators as well as all other usual estimators of population total using double sampling approach under two stage sampling were evaluated based on two measures *viz.* percentage Relative Bias (%RB) and percentage Relative Root Mean Squared Error (%RRMSE) of any estimator of the population parameter θ as given by

$$RB(\hat{\theta}) = \frac{1}{S}\sum_{i=1}^{S}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)\times 100 \quad and \quad RRMSE(\hat{\theta}) = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2}\times 100.$$

where, $\hat{\theta}_i$ are the estimates of population parameter $\theta$ for the character under study obtained at $i^{\text{th}}$ sample in the simulation study.

Table 1, 2, 3 and 4 present the results of the simulation study under population Set 1, 2, 3 & 4 in terms of %RB and %RRMSE of the proposed product type calibration estimator ($\hat{t}_{yCPd}$) and usual NHT estimator ($\hat{t}_{y\pi}$) and product estimator ($\hat{t}_{yPd}$) of the population total under two stage sampling design using double sampling approach. These estimators have been calculated assuming that the complete auxiliary information ($x_{ik}$) was not available at the SSU level in the population and auxiliary variable is inversely related to the character under study.

**Table 1: Comparison of all the estimators with respect to %RB and %RRMSE under two stage sampling in case of population Set 1 having correlation coefficient ($\rho$) as -0.91 using double sampling approach**

| **Sample size** $(n'\_m'_i\_n\_m_i)$ | **% RB** | | | **% RRMSE** | | |
|---|---|---|---|---|---|---|
| | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ |
| 10_20_4_6 | 0.014 | -0.010 | -0.152 | 4.896 | 1.963 | 2.559 |
| 10_25_4_8 | 0.052 | 0.001 | -0.092 | 4.261 | 1.723 | 2.256 |
| 15_25_6_8 | 0.011 | 0.004 | -0.083 | 3.532 | 1.402 | 1.839 |
| 15_30_6_10 | -0.005 | -0.003 | -0.058 | 3.101 | 1.248 | 1.639 |
| 20_30_8_10 | -0.005 | 0.007 | -0.039 | 2.676 | 1.083 | 1.432 |
| 20_40_8_12 | 0.019 | 0.009 | -0.037 | 2.434 | 0.933 | 1.258 |
| 25_40_10_12 | -0.004 | 0.003 | -0.027 | 2.178 | 0.811 | 1.117 |
| 25_50_10_15 | -0.022 | -0.004 | -0.033 | 1.950 | 0.716 | 0.996 |

**Table 2: Comparison of all the estimators with respect to %RB and %RRMSE under two stage sampling in case of population Set 2 having correlation coefficient ($\rho$) as -0.85 using double sampling approach**

| **Sample size** $(n'\_m'_i\_n\_m_i)$ | **% RB** | | | **% RRMSE** | | |
|---|---|---|---|---|---|---|
| | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ |
| 10_20_4_6 | 0.058 | 0.026 | -0.123 | 5.286 | 2.756 | 3.195 |
| 10_25_4_8 | 0.109 | 0.048 | -0.056 | 4.499 | 2.406 | 2.785 |
| 15_25_6_8 | -0.046 | -0.047 | -0.122 | 3.694 | 1.939 | 2.248 |
| 15_30_6_10 | -0.030 | -0.011 | -0.081 | 3.297 | 1.730 | 2.014 |
| 20_30_8_10 | -0.003 | -0.027 | -0.061 | 2.847 | 1.490 | 1.739 |
| 20_40_8_12 | 0.021 | 0.035 | -0.005 | 2.585 | 1.346 | 1.570 |
| 25_40_10_12 | -0.008 | 0.016 | -0.017 | 2.313 | 1.173 | 1.386 |
| 25_50_10_15 | -0.041 | -0.021 | -0.045 | 2.065 | 1.040 | 1.219 |

**Table 3: Comparison of all the estimators with respect to %RB and %RRMSE under two stage sampling in case of population Set 3 having correlation coefficient ($\rho$) as -0.78 using double sampling approach**

| **Sample size** $(n'\_m'_i\_n\_m_i)$ | **% RB** | | | **% RRMSE** | | |
|---|---|---|---|---|---|---|
| | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ |
| 10_20_4_6 | -0.095 | -0.008 | -0.167 | 5.699 | 3.569 | 3.892 |
| 10_25_4_8 | -0.018 | -0.002 | -0.135 | 4.952 | 3.127 | 3.410 |
| 15_25_6_8 | 0.002 | -0.013 | -0.087 | 4.012 | 2.506 | 2.743 |
| 15_30_6_10 | -0.020 | 0.021 | -0.047 | 3.612 | 2.231 | 2.448 |
| 20_30_8_10 | 0.025 | -0.002 | -0.042 | 3.095 | 1.945 | 2.137 |
| 20_40_8_12 | -0.031 | -0.001 | -0.054 | 2.853 | 1.762 | 1.930 |
| 25_40_10_12 | -0.016 | 0.001 | -0.040 | 2.537 | 1.543 | 1.687 |
| 25_50_10_15 | 0.012 | 0.006 | -0.020 | 2.243 | 1.359 | 1.509 |

**Table 4: Comparison of all the estimators with respect to %RB and %RRMSE under two stage sampling in case of population Set 4 having correlation coefficient ($\rho$) as -0.64 using double sampling approach**

| Sample size $(n'\_m'_i\_n\_m_i)$ | % RB | | | % RRMSE | | |
|---|---|---|---|---|---|---|
| | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ | $\hat{t}_{y\pi d}$ | $\hat{t}_{yCPd}$ | $\hat{t}_{yPd}$ |
| 10_20_4_6 | 0.048 | 0.080 | -0.094 | 6.791 | 5.230 | 5.425 |
| 10_25_4_8 | -0.015 | 0.052 | -0.083 | 5.971 | 4.612 | 4.791 |
| 15_25_6_8 | -0.019 | -0.012 | -0.100 | 4.786 | 3.679 | 3.804 |
| 15_30_6_10 | 0.003 | 0.032 | -0.030 | 4.228 | 3.293 | 3.422 |
| 20_30_8_10 | -0.020 | -0.020 | -0.083 | 3.745 | 2.884 | 2.988 |
| 20_40_8_12 | -0.031 | 0.001 | -0.056 | 3.372 | 2.565 | 2.671 |
| 25_40_10_12 | 0.001 | -0.016 | -0.049 | 2.991 | 2.275 | 2.367 |
| 25_50_10_15 | 0.005 | -0.009 | -0.035 | 2.694 | 2.051 | 2.129 |

From Table 1 it is notable that the proposed product type calibration estimator of the finite population total was giving consistently least amount %RB compared to their usual NHT and product estimator using double sampling approach for the Poulation Set 1 where correlation coefficient ($\rho$) was -0.91. Here, it is assumed that the auxiliary information was unavailable at SSU level and auxiliary variable is inversely related with the character under study. It was also seen that the proposed product type calibration estimator of the population total is always more efficient than the NHT and product estimators, since %RRMSE of the proposed product type calibration estimator is always least at different sample size combinations. It can also be seen that the %RRMSE of the proposed product type calibration estimator was decreasing with increase of sample sizes, thus, it provides a consistent estimator of the finite population total. Similar trend in simulation results can be observed in Table 2, 3 and 4, where Population Set 2, 3 and 4 are considered for simulation in which correlation coefficient ($\rho$) were -0.85, -0.78 and -0.64 respectively. Close look of Table 2, 3 and 4 reveals that %RRMSE of the proposed product type calibration estimator was decreasing with the increase in the amount of negative correlation.

## 5. Conclusions and way forward

In general, the Calibration Approach assumes the availability of population-level auxiliary information. On the contrary, in large scale surveys involving two stage sampling, it is often the case that population-level data on auxiliary variable is not available in practice, but relatively inexpensive to collect. In the present article, in case of non-availability of population-level relatively inexpensive data on auxiliary variable in two stage sampling, product type calibration estimator (Equation 5) of the finite population total has been proposed using double sampling approach when there exist inverse relation between auxiliary variable and chracter under study. In order to study the statistical performance of proposed product type calibration estimator as compared to existing estimators of population total of character under study, a simulation study was conducted. The simulation results also suggests that the proposed product type calibration estimators using double sampling approach in two stage sampling, performs better than usual Narain-Horvitz-Thompson estimator (Equation 2) and product estimator (Equation 7) of the finite population total with respect to %RB and %RRMSE. In future, investigation may be carried out to extent the work in case of different type of varying probability sampling schemes in oder to improve several well-known estimators *viz.* Narain-Horvitz-Thompson estimator, Rao, Hartley and Cochran (1962) estimator etc.

**Acknowledgement**

Authors are profoundly thankful to the editors and the anonymous reviewer for their productive comments which led to improvements in the article.

## References

Aditya, K., Sud, U. C., Chandra, H. and Biswas, A. (2016a). Calibration Based Regression Type Estimator of the Population Total under Two Stage Sampling Design. *Journal of Indian Society of Agricultural Statistics*, **70**, 19-24.

Aditya, K., Sud, U. C. and Chandra, H. (2016b). Calibration Approach based Estimation of Finite Population Total under Two Stage Sampling. *Journal of the Indian Society of Agricultural Statistics*, **70**, 219–226.

Aditya K., Biswas A., Gupta, A. K. and Chandra, H. (2017). District-level crop yield estimation using calibration approach. *Current Science*, **112**, 1927-1931.

Arnab, R. (2017). *Survey Sampling Theory and Applications*, Academic Press, Oxford.

Basak P., Sud, U. C. and Chandra, H. (2017). Calibration Estimation of Regression Coefficient for Two-stage Sampling Design. *Journal of the Indian Society of Agricultural Statistics*, **71**, 1–6.

Biswas, A., Aditya, K., Sud U. C., and Basak, P. (2020). Product type Calibration Estimation of Finite Population Total under Two Stage Sampling. *Journal of the Indian Society of Agricultural Statistics*, **74**, 23–32.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Estevao, V. M. and Sarndal, C. E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*, **74**, 127–147.

Haziza, D., Hidiroglou, M. A. and Rao, J. N. K. (2011). Comparison of variance estimators in two-phase sampling: An empirical investigation. *Pakistan Journal of Statistics*, **27**, 477-492.

Hidiroglou, M. A., Rao, J. N. K. and Haziza, D. (2009). Variance estimation in two-phase sampling. *Australian and New Zealand Journal of Statistics*, **51**, 127-141.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47,** 663-685.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, **78**, 21-39.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133-142.

Mourya, K. K., Sisodia, B. V. S. and Chandra, H. (2016). Calibration approach for estimating finite population parameter in two stage sampling. *Journal of Statistical Theory and Practice*, **10**, 550-562.

Murthy, M. N. (1964). Product method of estimation. *Sankhya*, **26A**, 69–74.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Aricultural Statistics*, **3**, 169-174.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **33**, 101-116.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **60**, 125- 133.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). On a simp,e procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, **24**, 482-491.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99-119.

Salinas, V. I., Sedory, S. A. and Singh, S. (2018). Calibrated estimators in two-stage sampling. *Communications in Statistics - Theory and Methods*, **48**, 1449-1469. DOI: 10.1080/03610926.2018.1433850.

Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of the general regression estimator: Higher level calibration approach. *Survey Methodology*, **24**, 41-50.

Singh, S., Horn, S., Choudhury, S. and Yu, F. (1999). Calibration of the estimators of variance. *Australian and New Zealand Journal of Statistics*, **41**, 199-212.

Sinha N., Sisodia, B. V. S., Singh, S. and Singh, S. K. (2016). Calibration approach estimation of mean in stratified sampling and stratified double sampling. *Communications in Statistics - Theory and Methods*, **46**, DOI: 10.1080/03610926.2015.1091083.

Sitter, R. R. and Wu, C. (2002). Efficient estimation of quadratic finite population functions. *Journal of the American Statistical Association*, **97**, 535-543.

Sud, U. C., Chandra, H. and Gupta, V. K. (2014a). Calibration based product estimator in single and two phase sampling. *Journal of Statistical Theory and Practice*, **8**, 1-11.

Sud, U. C., Chandra, H. and Gupta, V. K. (2014b). Calibration approach based regression type estimator for inverse relationship between study and auxiliary variable. *Journal of Statistical Theory and Practice*, **8**, 707-721.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok. C. (1984). *Sampling Theory with Applications*. Indian Society of Agricultural Statistics, New Delhi.

Wu, C. and Sitter, R. R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.

Statistics and Applications {ISSN 2454 –7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 23 – 26

# Regular Group Divisible Designs Using Symmetric Groups

**[1]Shyam Saurabh and [2]Kishore Sinha**
*[1]Ranchi University, Ranchi, India*
*[2]Formerly at Birsa Agricultural University, Ranchi, India and*
*#201 Maitry Residency, Kalkere Main Road, Bangalore – 560043*

Received: 05 November 2021; Revised: 29 November 2021; Accepted: 20 January 2022

## Abstract

Two regular group divisible designs with parameters: $v = 30$, $b = 60$, $r = 8$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = 5$, $n = 6$ and $v = 36$, $b = 90$, $r = 10$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = n = 6$ in the range of $r, k \leq 10$ are obtained from generalized Bhaskar Rao designs over a symmetric group of order 6.

*Key words*: Regular group divisible designs; Generalized Bhaskar Rao designs; Symmetric groups.

**MSC:** 62K10; 05B05

## 1. Introduction

Saurabh and Sinha (2021) obtained a new regular group divisible (*RGD*) design with parameters: $v = b = 39$, $r = k = 9$, $\lambda_1 = 0$, $\lambda_2 = 2$, $m = 13$, $n = 3$ by replacing the group entries of *BGW* (13, 9, 6; $D_3$) by suitable permutation matrices of order 3. Here we have used the method of Gibbons and Mathon (1987) for the construction of group divisible designs. As a particular case we obtain two RGD designs with parameters: $v = 30$, $b = 60$, $r = 8$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = 5$, $n = 6$ and $v = 36$, $b = 90$, $r = 10$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = n = 6$ in the range of $r, k \leq 10$. These designs may be considered new as these are not found in the tables of Clatworthy (1973) and Sinha (1991) but included in Saurabh and Sinha (2021).

A generalized Bhaskar Rao design *GBRD* ($v$, $b$, $r$, $k$, $\lambda$; $G$) over a group $G$ is a $v \times b$ array with entries from $G \cup \{0\}$such that:

1. each row has exactly $r$ group element entries;
2. each column has exactly $k$ group element entries;
3. for each pair of distinct rows $(x_1, x_2, \dots, x_b)$ and $(y_1, y_2, \dots, y_b)$, the multi–set $\{x_i y_i^{-1}: i = 1, 2, \dots, b;\ x_i, y_i \neq 0\}$ contains each group element exactly $\lambda/|G|$ times.

A generalized Bhaskar Rao design *GBRD* ($v, b, r, k, \lambda; G$) with $v = b$ and $r = k$ is known as a *balanced generalized Weighing matrix BGW* ($v, k, \lambda; G$).

A *RGD design* is an arrangement of $v = mn$ elements in $b$ blocks such that:

Corresponding Author: Kishore Sinha
Email: kishore.sinha@gmail.com

(i) each block contains $k$ ($< v$) distinct elements;

(ii) each element occurs $r$ times;

(iii) the elements can be divided into $m$ groups each of size $n$, any two distinct elements occurring together in $\lambda_1$ blocks if they belong to the same group, and in $\lambda_2$ blocks if they belong to the different groups;

(iv) $r - \lambda_1 > 0$ and $rk - v\lambda_2 > 0$.

Let **N** be the incidence matrix of a RGD design then the structure of $\mathbf{NN}^T$ is given as: $\mathbf{NN}^T = (r - \lambda_1)(\mathbf{I}_m \otimes \mathbf{I}_n) + (\lambda_1 - \lambda_2)(\mathbf{I}_m \otimes \mathbf{J}_n) + \lambda_2(\mathbf{J}_m \otimes \mathbf{J}_n)$ where $\mathbf{A}\otimes\mathbf{B}$ denotes the Kronecker product of two matrices **A** and **B**. For details on RGD designs, see Clatworthy (1973) and Saurabh *et al.* (2021).

*Notations*: $\mathbf{I}_n$ is the identity matrix of order $n$, $\mathbf{J}_v$ is the $v \times v$ matrix all whose entries are 1 and $\mathbf{A}^T$ is the transpose of matrix **A**. $S_n$ and $D_n$ denote symmetric and dihedral groups with orders $n!$ and $2n$ respectively. For $n = 3$, $S_n$ is isomorphic to the dihedral group $D_n$.

## 2. Two new RGD designs in the range of $r, k \leq 10$

Gibbons and Mathon (1987) gave the following method for the construction of GD designs from GBRD ($v, b, r, k, \lambda; G$):

Replacing the elements of a group $G$ of order $g$ by the corresponding $g$ x $g$ permutation matrices and 0 entry by $g$ x $g$ null matrix in GBRD ($v, b, r, k, \lambda; G$), we obtain a GD design with parameters: $v^* = vg,\ b^* = bg, r^* = r,\ k^* = k, \lambda_1 = 0, \lambda_2 = \lambda/g\,, m = v, n = g.$ (1)

In the above method, Palmer and Seberry (1988) used permutation group of order 6 and dihedral groups of order 8 and 12 while Sarvate and Seberry (1998) used elementary abelian groups for the construction of GD designs.

Following Palmer and Seberry (1988): The existence of a GBRD ($v, b, r, k, \lambda; S_3$) implies the existence of a GD design with parameters:

$$v^* = 6v,\ b^* = 6b, r^* = r,\ k^* = k, \lambda_1 = 0, \lambda_2 = \lambda/6\,, m = v, n = 6. \quad (2)$$

The above construction procedure may be generalized for any symmetric / dihedral groups but no series of GBRD ($v, b, r, k, \lambda; S_n/ D_n$) is available for $n > 3$. Using GBRD (5, 10, 8, 4, 6; $S_3$) and GBRD (6, 15, 10, 4, 6; $S_3$) from Abel *et al.* (2004) in (2), we obtain the following RGD designs:

**Design 1:** Consider a symmetric group $S_3 = \langle r, s\colon r^3 = s^2 = e, sr = r^2 s\rangle = \{e, r, r^2, s, sr, sr^2\}$.

The following is a GBRD (5, 10, 8, 4, 6; $S_3$):

$$\mathbf{A} = \begin{bmatrix} e & s & r & 0 & e & e & r^2 & e & 0 & r^2s \\ e & e & s & r & 0 & r^2s & e & r^2 & e & 0 \\ 0 & e & e & s & r & 0 & r^2s & e & r^2 & e \\ r & 0 & e & e & s & e & 0 & r^2s & e & r^2 \\ s & r & 0 & e & e & r^2 & e & 0 & r^2s & e \end{bmatrix}.$$

Replacing 0 by a null matrix of order 6 and the group elements $e, r, r^2, s, sr = r^2s, sr^2 = rs$ by the 6 × 6 permutation matrices $\mathbf{I}_6$, $\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$,

$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$ respectively in *A*,

we obtain a (0, 1) – matrix **N** of order $30 \times 60$. Then $\mathbf{NN}^T = \text{circ}\,(8I_6, J_6, J_6, J_6, J_6) = 8\mathbf{I}_{30} - \mathbf{I}_5 \otimes \mathbf{J}_6 + \mathbf{J}_5 \otimes \mathbf{J}_6$. Also each column sum of **N** is 4. Hence **N** represents a RGD design with parameters: $v = 30$, $b = 60$, $r = 8$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = 5$, $n = 6$.

**Design 2:** Further consider the following GBRD (6, 15, 10, 4, 6; $S_3$):

$$\mathbf{B} = \begin{bmatrix} e & e & e & e & e & e & e & e & e & e & 0 & 0 & 0 & 0 & 0 \\ rs & s & e & r & r^2 & r^2s & 0 & 0 & o & 0 & e & e & e & e & 0 \\ r & s & rs & 0 & 0 & 0 & r^2 & e & r^2s & 0 & s & r^2 & r & 0 & e \\ e & 0 & 0 & r & rs & 0 & r^2 & s & 0 & r^2s & r^2 & r^2s & 0 & r & r \\ 0 & e & 0 & rs & 0 & r^2 & r^2s & 0 & s & r & r^2 & 0 & r & e & r^2 \\ 0 & 0 & r & 0 & s & r^2 & 0 & e & rs & r^2s & 0 & e & s & r^2 & s \end{bmatrix}.$$

Replacing the group elements $e, r, r^2, s, sr = r^2s, sr^2 = rs$ by $6 \times 6$ matrices given as above and 0 by a null matrix of order 6 in **B**, we obtain a (0, 1) – matrix **N** of order $36 \times 90$. Then $\mathbf{NN}^T = \text{circ}\,(10\mathbf{I}_6, \mathbf{J}_6, \mathbf{J}_6, \mathbf{J}_6, \mathbf{J}_6, \mathbf{J}_6) = 10\mathbf{I}_{36} - \mathbf{I}_6 \otimes \mathbf{J}_6 + \mathbf{J}_6 \otimes \mathbf{J}_6$. Also each column sum of **N** is 4. Hence **N** represents a RGD design with parameters: $v = 36$, $b = 90$, $r = 10$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = n = 6$.

## Acknowledgement

The authors are thankful to the referee for valuable suggestions.

## References

Abel, R. J. R., Combe, D. and Palmer, W. D. (2004). Generalized Bhaskar Rao designs and dihedral groups. *Journal of Combinatorial Theory*, *Series A* **106**, 145–157.

Clatworthy, W. H. (1973). Tables of two–associate–class partially balanced designs. *National Bureau of Standards* (*U.S.*), *Applied Mathematics*, *Series* **63**.

Gibbons, P. B. and Mathon, R. (1987). Construction methods for Bhaskar Rao and related designs. *Journal of the Australian Mathematical Society (Series A)*, **42**, 5–30.

Palmer, W. D. and Seberry, J. (1988). Bhaskar Rao designs over small groups. *Ars Combinatoria*, **26(A)**, 125–148.

Sarvate, D. G. and Seberry, J. (1998). Group divisible designs, GBRSDS and Generalized Weighing matrices. *Utilitas Mathematica*, **54**, 157– 174.

Saurabh, S., Sinha, K. and Singh, M. K. (2021). Unifying constructions of group divisible designs. *Statistics and Applications*, **19**, 125–140.

Saurabh, S. and Sinha, K. (2021). A new regular group divisible design. *Examples and Counter Examples*, doi.org/10.1016/j.exco.2021.100029.

Saurabh, S. and Sinha, K. (2021). A list of new partially balanced designs. Under Preparation.

Sinha, K. (1991). A list of new group divisible designs. *Journal of Research of the National Institute of Standards and Technology*, **96**, 1–3.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 27–40

# New Intervention Based Exponential Model with Real Life Data Applicability

**Vilayat Ali Bhat and Sudesh Pundir**
*Department of Statistics, Pondicherry University, Kalapet, Puducherry-605014, INDIA.*

Received: 07 September 2021; Revised: 03 January 2022; Accepted: 13 February 2022

## Abstract

In this paper, the new extension of the extended Exponential model named inverted intervened Exponential distribution has been proposed. To explore the model, the essential statistical properties have been presented in this study, the parametric estimation was also carried out by using the method of maximum likelihood estimation ($MLE$) technique. Moreover, the reliability characterization has been given which includes the mathematical functions of the reliability, hazard rate, aging intensity, and mean residual life. Also, the Rényi and Shannon entropy measures have been derived. Monte Carlo simulation study by employing the acceptance-rejection algorithm was performed to judge the performance of maximum likelihood estimates ($MLE_s$) based on the calculated results of absolute average bias (*Abias*) and mean square error ($MSE$) of the parametric estimates. Lastly, the model applicability checkup was also done by analyzing real data set.

*Key words:* Intervened model; Entropy; Monte Carlo simulation; Model applicability.

**AMS Subject Classifications:** 62K05, 05B05

## 1. Introduction

In literature, the traditional models such as Exponential, Normal, Rayleigh, Gamma, Weibull, etc. are the basic fundamental models in statistical theory. From the past few decades, many developments have been observed in the form of modifications and generalizations to develop more flexible distributions for data analysis purposes. In history, one could observe the most exploiting and frequently used distribution in the field of reliability and survival analysis among them being the Exponential model for reference see Balakrishnan (2019). However, the disadvantage of the Exponential model is meant due to the constant hazard rate, as there arise situations where it is observed the model requirement for increasing, decreasing, bath-tub shaped hazard rate situations as well, to model the failure data. In this context, the successful efforts of the researchers who developed different types of models to cope with these situations to some extent. It gives a clear picture that every new technique has added several types of flexible distributions in statistical theory. Since a few years ago, a new concept intervention was introduced in the distribution theory, and it was Shanmugam (1985) who made a noble attempt to develop a discrete intervention based Poisson model, later which laid to a beginning new intervention-based model development in

Corresponding Author: Sudesh Pundir
E-mail: sudeshpundir19@gmail.com

the statistical literature. A similar attempt on continuous Exponential distribution by Shanmugam *et al.* (2002) developed the intervened Exponential model ($I_vED$), the outstanding medical applications of the model motivated us to develop a new extension of the model named as inverted intervened Exponential model ($II_vD$). The cumulative density function (cdf) of the newly developed model along with its probability density function (pdf) are given by:

$$F_{II_vD}(y;\Theta) = \begin{cases} \frac{\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{(\rho-1)} & \rho \neq 1 \\ \left(\frac{1-(\delta-\eta)y}{\eta y}\right) e^{-(1-\delta y)/\eta y} & \rho = 1 \end{cases} \tag{1}$$

and,

$$f_{II_vD}(y;\Theta) = \begin{cases} \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{(\rho-1)\eta y^2} & \rho \neq 1 \\ \frac{(1-\delta y)}{\eta^2 y^3}\, e^{-(1-\delta y)/\eta y} & \rho = 1 \end{cases} \tag{2}$$

where $0 < y < \frac{1}{\delta}$, and the desired parametric space of the model is denoted by $\Theta = \{(\rho, \delta, \eta) : \rho > 0, \delta > 0, \eta > 0\}$ containing $\eta$ as the rate parameter, $\rho$ being the intervention parameter, and the parameter $\delta$ is treated as the truncation point of the model. Further, the graphical illustration of the proposed model based on the desired set of parametric values for pdf is shown below:

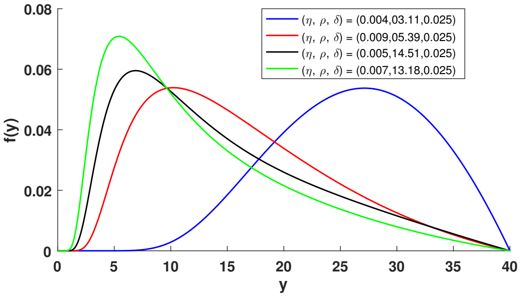


**Figure 1: PDF plot**

It could be easily predicted from graphical behavior that different shapes for pdf are exhibited on the selected set of parameters.

## 2. Statistical properties

We attempted in this section, to provide the mathematical derivation of the different statistical properties that would help to understand the nature of $II_vD$. The mathematical expressions of the obtained results include mean ($\mu_y$), median ($M_d$), and the variance ($\sigma_y^2$) of the model. The other results mentioned in subsections are the mean deviations, $r^{th}$ order moment expressions about the origin, the mean, and the different generating functions for moments. So, to begin this section, the mean of the distribution obtained is as follows:

$$\mu_y = \frac{1}{\rho-1}\left\{e^{\delta/\rho\eta}\Gamma(0,\delta/\rho\eta) - e^{\delta/\rho}\Gamma(0,\delta/\rho)\right\} \tag{3}$$

The variance of the $II_vD$ is given by,

$$\sigma_y^2 = \frac{1}{\rho(\rho-1)\eta}\left\{e^{\delta/\rho\eta}\Gamma(-1,\delta/\rho\eta) - e^{\delta/\rho}\Gamma(-1,\delta/\rho)\right\} - (\mu_y)^2 \tag{4}$$

Now to find the median of $II_vD$ mathematically, we make use of the definition of median as given below:

$$\int_{M_d}^{1/\delta} f_{II_vD}(y;\Theta)dy = 1/2$$

$$\tfrac{1}{(\rho-1)\eta} \int_{M_d}^{1/\delta} \left\{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}\right\}(1/y^2)dy = 1/2$$

$$\rho e^{(\delta M_d-1)/\rho\eta M_d} - e^{(\delta M_d-1)/\eta M_d} = (\rho-1)/2$$

Note: $\Gamma(c,t) = \int_t^\infty y^{c-1}e^{-y}dy$ is the upper incomplete gamma function.

## 2.1. Mean deviations

In statistics, two well-known measures that are used to measure the scatteredness present among the data are called the mean deviations about the mean, and another one is considered as mean deviations about the median. Henceforth, these two measures are represented by $D_{\mu_y}$ and $D_{M_d}$ respectively. The mathematical derivation for these two measures is given in the following theorem.

**Theorem 1:** If a random variable $(r.v.)$ $Y \sim II_vD(\rho,\delta,\eta)$, then the derived expression for $D_{\mu_y}$ and $D_{M_d}$ for the proposed model are as:

$$(i) \quad D_{\mu_y} = \left\{\mu_y F_{II_vD}(\mu_y) - \frac{e^{\delta/\rho\eta}}{(\rho-1)\eta}\Gamma(0,1/\mu_y\rho\eta) + \frac{e^{\delta/\eta}}{(\rho-1)\eta}\Gamma(0,1/\mu_y\eta)\right\}$$

$$(ii) \quad D_{M_d} = (\mu_y - M_d) + 2\left\{M_d F_{II_vD}(M_d) - \frac{e^{\delta/\rho\eta}}{(\rho-1)\eta}\Gamma(0,1/M_d\rho\eta) + \frac{e^{\delta/\eta}}{(\rho-1)\eta}\Gamma(0,1/M_d\eta)\right\}$$

**Proof:** $(i)$ For, any $r.v.$ $Y$ the mean deviation about mean is given by

$$\begin{aligned} D_{\mu_y} &= 2\left\{\mu_y F_{II_vD}(\mu_y) - \int_0^{1/\delta} y f_{II_vD}(y;\Theta)dy\right\} \\ &= 2\left\{\mu_y F_{II_vD}(\mu_y) - \int_0^{\mu_y} \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{(\rho-1)\eta y}dy\right\} \\ &= \left\{\mu_y F_{II_vD}(\mu_y) - \frac{e^{\delta/\rho\eta}}{(\rho-1)\eta}\Gamma(0,1/\mu_y\rho\eta) + \frac{e^{\delta/\eta}}{(\rho-1)\eta}\Gamma(0,1/\mu_y\eta)\right\} \end{aligned}$$

Hence, completes proof for part first.

$(ii)$ Again, for a continuous, and non-negative $r.v.$, $Y \sim II_vD(\rho,\delta,\eta)$, we can write the mathematical expression for median deviation as,

$$D_{M_d} = \mu_y - M_d + 2\left\{M_d F_{II_vD}(M_d) - \int_0^{M_d} y f_{II_vD}(y;\Theta)dy\right\}$$

$$= \quad (\mu_y - M_d) + 2\left\{M_d F_{II_vD}(M_d) - \int_0^{M_d} \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{(\rho-1)\eta y}dy\right\}$$

$$= \quad (\mu_y - M_d) + \left\{M_d F_{II_vD}(M_d) - \frac{e^{\delta/\rho\eta}}{(\rho-1)\eta}\Gamma(0,1/M_d\rho\eta) + \frac{e^{\delta/\eta}}{(\rho-1)\eta}\Gamma(0,1/M_d\eta)\right\}$$

This completes proof for part $(ii)$. □

### 2.2. Moments and moments generating functions

Here in this subsection, we shall derive the expression for $r^{th}$ moments about the origin and the moments about mean, and the generating functions for moments in the following subsequent theorems,

**Theorem 2:** If $Y$ be any non-negative $r.v.$ possessing $II_vD$, then the moments about the origin and the mean are given by:

$$(i) \quad \mu_r' = \frac{1}{(\rho-1)\,\eta^r}\left\{\frac{e^{\delta/\rho\eta}}{\rho^r}\Gamma\left(1-r,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-r,\delta/\eta\right)\right\},\ \ r=1,2...n.$$

$$(ii) \quad \mu_r = \frac{1}{(\rho-1)}\sum_{n=0}^{r}{}^{r}C_n\frac{(-\mu)^{r-n}}{\eta^n}\left\{\frac{e^{\delta/\rho\eta}}{\rho^n}\Gamma\left(1-n,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-n,\delta/\eta\right)\right\};\ \ r=1,2...n.$$

**Proof:** (i) For a random variable $Y\sim II_vD(\rho,\delta,\eta)$, the expression for $r^{th}$ moment about origin is,

$$\begin{aligned}\mu_r' &= E(y^r)=\frac{1}{(\rho-1)\eta}\int_0^{1/\delta}y^{r-2}\left\{e^{-(1-\delta y)/\rho\eta y}-e^{-(1-\delta y)/\eta y}\right\}dy\\ &= \frac{1}{(\rho-1)\,\eta^r}\left\{\frac{e^{\delta/\rho\eta}}{\rho^r}\Gamma\left(1-r,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-r,\delta/\eta\right)\right\}\end{aligned}$$

where $r=0,1,...,n$.
(ii) Again, for a random variable $Y\sim II_vD(\rho,\delta,\eta)$, the expression for $r^{th}$ moment about mean is

$$\begin{aligned}\mu_r &= E(y-\mu_y)^r=\frac{1}{(\rho-1)\eta}\int_0^{1/\delta}(y-\mu_y)^r\frac{\left\{e^{-(1-\delta y)/\rho\eta y}-e^{-(1-\delta y)/\eta y}\right\}}{y^2}dy\\ &= \frac{1}{(\rho-1)}\sum_{n=0}^{r}{}^{r}C_n\frac{(-\mu)^{r-n}}{\eta^n}\left\{\frac{e^{\delta/\rho\eta}}{\rho^n}\Gamma\left(1-n,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-n,\delta/\eta\right)\right\}\end{aligned}$$

where $r=0,1,...,n$. □

**Theorem 3:** If $Y$ be any non-negative $r.v.$ possessing $II_vD$, then the generating functions for moments are given by:
$(i)M_y(t)=\frac{1}{(\rho-1)\eta}\sum_{r=0}^{\infty}\frac{t^r}{\eta^r r!}\left\{\frac{e^{\delta/\rho\eta}}{\rho^r}\Gamma\left(1-r,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-r,\delta/\eta\right)\right\}$,is the moment generating function.
$(ii)\phi_y(t)=\frac{1}{(\rho-1)\eta}\sum_{r=0}^{\infty}\frac{t^r}{\eta^r r!}\left\{\frac{e^{\delta/\rho\eta}}{\rho^r}\Gamma\left(1-r,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-r,\delta/\eta\right)\right\}$, is the characteristic function.
$(iii)K_y(t)=\log\left[\frac{1}{(\rho-1)\eta}\sum_{r=0}^{\infty}\frac{t^r}{\eta^r r!}\left\{\frac{e^{\delta/\rho\eta}}{\rho^r}\Gamma\left(1-r,\delta/\rho\eta\right)-e^{\delta/\eta}\Gamma\left(1-r,\delta/\eta\right)\right\}\right]$, is the cumulant generating function.

**Proof:** (i) Let the $r.v.$ $Y \sim II_vD(\rho, \delta, \eta)$, then $M_y(t)$ is derived by

$$\begin{aligned} M_y(t) &= E(e^{ty}) = \int_0^{1/\delta} e^{ty} f_{II_vD}(y;\Theta)dy \\ &= \frac{1}{(\rho-1)\,\eta} \sum_{r=0}^{\infty} \frac{t^r}{\eta^r r!} \left\{ \frac{e^{\delta/\rho\eta}}{\rho^r} \Gamma\left(1-r, \delta/\rho\eta\right) - e^{\delta/\eta} \Gamma\left(1-r, \delta/\eta\right) \right\} \end{aligned}$$

$(ii)$ To prove the characteristic function the same procedure has to be repeated that we used to derive the moment generating function, but the only change is instead of $t$ we have to proceed with $\iota t$.

$(iii)$ Let the $r.v.$ $Y \sim II_vD(\rho, \delta, \eta)$, then $K_y(t)$ is defined by

$$\begin{aligned} K_y(t) &= \log\left\{M_y(t)\right\} \\ &= \log\left[ \frac{1}{(\rho-1)\,\eta} \sum_{r=0}^{\infty} \frac{t^r}{\eta^r r!} \left\{ \frac{e^{\delta/\rho\eta}}{\rho^r} \Gamma\left(1-r, \delta/\rho\eta\right) - e^{\delta/\eta} \Gamma\left(1-r, \delta/\eta\right) \right\} \right] \end{aligned}$$

□

## 3. Reliability properties

The probability measurement of any component or a system, that will not fail before time t to perform its complete operation is called the reliability of the system. Mathematically, it is calculated as:

$$R_{II_vD}(y;\Theta) = Pr.(Y > y) = 1 - Pr.(Y \leq y)$$

Thus, for a $r.v.$ $Y \sim II_vD(\rho, \delta, \eta)$ the derived reliability function is obtained as

$$R_{II_vD}(y;\Theta) = \begin{cases} 1 - \frac{\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{(\rho-1)} & \rho \neq 1 \\ 1 - \left(\frac{1-(\delta-\eta)y}{\eta y}\right) e^{-(1-\delta y)/\eta y} & \rho = 1 \end{cases} \tag{5}$$

If $\hat{\rho}$, $\hat{\delta}$ and $\hat{\eta}$ are the $MLE_s$, then by the in-variance property the reliability estimate are given by

$$\hat{R}_{II_vD}(y;\hat{\Theta}) = \begin{cases} 1 - \frac{\hat{\rho} e^{-(1-\hat{\delta} y)/\hat{\rho}\hat{\eta} y} - e^{-(1-\hat{\delta} y)/\hat{\eta} y}}{(\hat{\rho}-1)} & \rho \neq 1 \\ 1 - \left(\frac{1-(\hat{\delta}-\hat{\eta})y}{\hat{\eta} y}\right) e^{-(1-\hat{\delta} y)/\hat{\eta} y} & \rho = 1 \end{cases} \tag{6}$$

The hazard rate for $II_vD$, which we will denote by $h_{II_vD}(y)$ is the ratio of pdf and the $R_{II_vD}(y)$ as given below:

$$h_{II_vD}(y;\Theta) = \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{\eta y^2 \left[(\rho-1) - \left\{\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}\right\}\right]} \tag{7}$$

The graphical plot of hazard function for different set of parametric values is shown in Figure 2.

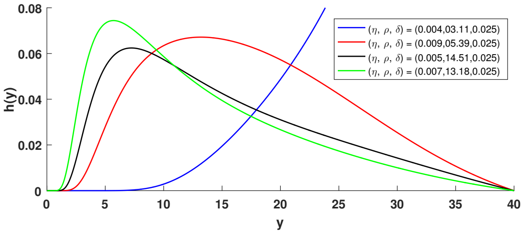


**Figure 2: Hazard plot**

The hazard rate in the reverse direction of time is called reverse hazard rate which we denote by $h^r_{II_vD}(y)$, this measure is obtained by taking the ratio of pdf and cdf and the obtained expression is

$$h^r_{II_vD}(y;\Theta) = \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{\eta y^2\left[\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}\right]} \tag{8}$$

The famous reliability measure called aging intensity $(A.I)$ developed by Jiang *et al.* (2003) are used for quantitative aging measurement purposes, as aging representation for the system by uni-modal hazard rate is difficult because of its varying trends observed in the form of constant, increasing and decreasing hazard rates. The $A.I$ for a $r.v.$ $Y \sim II_vD(\rho,\delta,\eta)$, denoted by $L_y$ is give as

$$A.I = \frac{e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}}{\eta y\left[\log\left(\rho-1\right) - \log\left\{\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}\right\}\right]\left[\rho e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y}\right]} \tag{9}$$

## 3.1. Mean residual life function

The mean residual life (MRL) function having a variety of applications in different branches of statistical and applied sciences, to define this measure, suppose a system/component functions without fail up to time $y \geq 0$, then the residual life is counted as the working hours of the system beyond time $y$ until it fails, and the conditional $r.v.$ $Y - y|Y > y$ is used to define this measure Finkelstein (2008).
For non-negative $r.v.$ $Y \sim II_vD(\rho,\delta,\eta)$, the MRL function denoted by $m_{II_vD}(y,\Theta)$ is derived as

$$m_{II_vD}(y;\Theta) = E\left[Y - y|Y > y\right] = \frac{1}{R_{II_vD}(y;\Theta)}\int_y^{1/\delta} R_{II_vD}(y;\Theta)dy.$$

$$= \frac{1}{(\rho-1)R_{II_vD}(y;\Theta)}\int_y^{1/\delta}\left\{(\rho-1) - \rho e^{(1-\delta y)/\rho\eta y} + e^{(1-\delta y)/\eta y}\right\}dy.$$

$$= \tfrac{1}{(\rho-1)R_{II_vD}(y;\Theta)}\left\{\tfrac{\left(\rho-\rho e^{\delta/\rho\eta}-e^{\delta/\eta}-1\right)(1-\delta y)}{\delta} - \tfrac{\left(e^{\delta/\rho\eta}+e^{\delta/\eta}\right)\log(\delta y)}{\eta} + \textstyle\sum_{r=2}^{\infty}\tfrac{(-1)^{r+2}}{r!\eta^r(1-r)}\left(\delta^{r-1}-y^{1-r}\right)\left(\rho^{1-r}e^{\delta/\rho\eta}-e^{\delta/\eta}\right)\right\}$$

## 4. Entropy measures

Entropy measurements are useful to determine, how much the random variable's distribution varies in terms of its level of variations, and the two important measures to address this variation are given by Rényi entropy and Shannon entropy (Refer Rényi, A. (1961, January) and Shannon (1948)).

### 4.1. Reńyi entropy

The Rényi entropy metric for a non-negative *r.v.* $Y \sim II_vD(\rho, \delta, \eta)$ of order $\vartheta$ is given by

$$H_R(\vartheta) = \frac{1}{1-\vartheta}\log\left[\int_0^{1/\delta}\left\{f_{II_vD}(y;\Theta)\right\}^{\vartheta} dy\right]; \quad \vartheta \geq 0, \vartheta \neq 1 \tag{10}$$

$$= \frac{1}{1-\vartheta}\log\left[\frac{1}{(\rho-1)^{\vartheta}\eta^{\vartheta}}\sum_{r=0}^{\vartheta}\frac{\binom{\vartheta}{r}(-1)^{r+2}e^{\psi}}{r!\psi^{2\vartheta-1}}\Gamma\left(2\vartheta-1,\psi\delta\right)\right]$$

where $\psi = \frac{(\rho-1)}{\rho\eta}\left(r + \frac{\vartheta}{(\rho-1)}\right)$ and $f_{II_vD}(y,\Theta)$ is the pdf given in equation (2), when $\rho \neq 1$.

### 4.2. Shannon entropy

In this subsection, we will derive the expression for Shannon measure of entropy for a non-negative *r.v.* $Y \sim II_vD(\rho, \delta, \eta)$, the derivation steps for this extend concept of Reńyi entropy are given by

$$H_{II_vD}(y) = -\int_0^{1/\delta} f_{II_vD}(y;\Theta)\log\left\{f_{II_vD}(y;\Theta)\right\} dy.$$

Now, substitute the density function $f_{II_vD}(y;\Theta)$, given in equation (2), when $\rho \neq 1$, and solve the integral we get

$$= \textstyle\sum_{r=1}^{\infty}\frac{\rho B\left(r+\frac{1}{(\rho-1)},2\right)}{r(\rho-1)^2} + \frac{(\rho+1)}{\rho} - 2\sum_{r=1}^{\infty}\frac{(-1)^{r+1}\eta^r\left(\rho^{r+1}-1\right)\Gamma(r+1)}{r(\rho-1)\delta^r} + \frac{\rho B\left(\frac{1}{(\rho-1)},2\right)\log\{(\rho-1)\eta\}}{(\rho-1)^2} - 2\log\delta$$

## 5. Order statistics

In the field of reliability, order statistics finds massive applications in life testing experiments for understanding system characterization of system. Let a random sample of size $n$ be taken as $Y = (y_1,\ y_2,\ \ldots\ ,y_n)$ be drawn from $II_vD(\rho, \delta, \eta)$. Then the life of $(n-i+1)$ components out-of-n *i.i.d* systems based on ordered random sample $y_{(1:n)} \leq y_{(2:n)} \leq \ldots \leq y_{(n:n)}$ are given by $y_{i:n}$; $(i = 1, 2, ..., n)$. Thus for $II_vD(\rho, \delta, \eta)$ the $i^{th}$ order statistics density function of $y_{(i:n)}$; $1 \leq i \leq n$ are given as

$$f_{i:n}(y,\Theta) = M_1[F_{II_vD}(y)]^{i-1}[1-F_{II_vD}(y)]^{n-i}f_{II_vD}(y). \tag{11}$$

Also, the pdf of $(i,j)^{th}$ order statistics density for $(y_{(i:n)}, y_{(j:n)})$; $1 \leq i \leq j \leq n$ are as

$$f_{i:j:n}(y_i,y_j) = M_2[F_{II_vD}(y_i)]^{i-1}[F_{II_vD}(y_j) - F_{I_vD}(y_i)]^{j-i-1}[1-F_{II_vD}(y_j)]^{n-j}f_{II_vD}(y_i)f_{II_vD}(y_j). \tag{12}$$

where $F(.)$, $f(.)$ is the cdf, pdf of $II_vD$ defined in (1) and (2), and the constants $M_1$ and $M_2$ are given by

$$M_1 = \textstyle\frac{n!}{(i-1)!(n-i)!} \text{ and } M_2 = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}$$

The smallest observation of ordered sample is called first-order statistic given by $y_{(1)} = min.(y_{(1)},\ y_{(2)},\ \ldots\ ,y_{(n)})$, the largest observation is called the $n^{th}$ order statistic, and the middle observation is called the median order given by $y_{m+1}$

### 5.1. Order statistic density function of $II_vD$

Let $y_{(1)}$, $y_{(2)}$, ... ,$y_{(n)}$ be $i.i.d$ ordered random sample from $II_vD$ then according to equations (1) and (2) we can write the first order statistics density ($f_{1:n}(.)$) on substituting $i = 1$, in equation (11), the $n^{th}$ order statistics density ($f_{n:n}(.)$) by substituting $i = n$ in equation (11) and the median order statistics density denoted by ($f_{m+1:n}(.)$);[$m = \frac{n}{2}$] are given below:

$$f_{1:n}(y) = n\left[1 - F_{II_vD}(y_{(1)})\right]^{n-1} f_{II_vD}(y_{(1)})$$

$$f_{1:n}(y) = \frac{n}{\eta(\rho-1)^n y_{(1)}^2}\left[(\rho-1) - \left\{\rho e^{-(1-\delta y_{(1)})/\rho\eta y_{(1)}} - e^{-(1-\delta y_{(1)})/\eta y_{(1)}}\right\}\right]^{n-1}\left[e^{-(1-\delta y_{(1)})/\rho\eta y_{(1)}} - e^{-(1-\delta y_{(1)})/\eta y_{(1)}}\right] \quad (13)$$

Similarly,

$$f_{n:n}(y) = n\left[F_{II_vD}(y)\right]^{n-1} f_{II_vD}(y_{(n)})$$

$$f_{n:n}(y) = \frac{n}{\eta(\rho-1)^n y_{(n)}^2}\left[\rho e^{-(1-\delta y_{(n)})/\rho\eta y_{(n)}} - e^{-(1-\delta y_{(n)})/\eta y_{(n)}}\right]^{n-1}\left[e^{-(1-\delta y_{(n)})/\rho\eta y_{(n)}} - e^{-(1-\delta y_{(n)})/\eta y_{(n)}}\right] \quad (14)$$

and,

$$f_{m+1:n}(y) = \frac{(2m+1)!}{(m!)^2}\left[F_{II_vD}(\bar{y})\right]^m\left[1 - F_{II_vD}(\bar{y})\right]^m f_{II_vD}(\bar{y}). \quad (15)$$

### 5.2. Joint order statistics density of $II_vD$

The joint pdf of $II_vD$ is obtained by using the pdf and cdf in (12) as shown below:

$$\begin{aligned} f_{i:j:n}(y_{(i)},\quad y_{(j)}) = {} & \frac{M_2}{\eta^2(\rho-1)^n y^4}\left[\rho e^{-(1-\delta y_{(i)})/\rho\eta y_{(i)}} - e^{-(1-\delta y_{(i)})/\eta y_{(i)}}\right]^{i-1} \\ . \quad & \left[\left\{\rho e^{-(1-\delta y_{(i)})/\rho\eta y_{(i)}} - e^{-(1-\delta y_{(i)})/\eta y_{(i)}}\right\} - \left\{\rho e^{-(1-\delta y_{(j)})/\rho\eta y_{(j)}} - e^{-(1-\delta y_{(j)})/\eta y_{(j)}}\right\}\right]^{j-i-1} \\ . \quad & \left[(\rho-1) - \left\{\rho e^{-(1-\delta y_{(j)})/\rho\eta y_{(j)}} - e^{-(1-\delta y_{(j)})/\eta y_{(j)}}\right\}\right]^{n-j}\left[e^{-(1-\delta y_{(i)})/\rho\eta y_{(i)}} - e^{-(1-\delta y_{(i)})/\eta y_{(i)}}\right] \\ . \quad & \left[e^{-(1-\delta y_{(j)})/\rho\eta y_{(j)}} - e^{-(1-\delta y_{(j)})/\eta y_{(j)}}\right] \end{aligned}$$

## 6. Stochastic ordering

Stochastic ordering measurement for lifetime distributions has vital importance in reliability theory, nicely discussed by Shaked and Shantikumar (2007). Let the $r.v.'s$ $Y_1$ and $Y_2$ possessing the $II_vD$ with pdf's $f_{Y_1}(y)$, $f_{Y_2}(y)$, and cdf's $F_{Y_1}(y)$, $F_{Y_1}(y)$ respectively. Then one would say $Y_1$ is smaller than $Y_2$ according to the stochastic ordering measurements given below:

[1] Stochastic order ($Y_1 \leq_{st} Y_2$), if $F_{Y_1}(y) \geq F_{Y_2}(y)$ for all y.
[2] Hazard rate order ($Y_1 \leq_{hr} Y_2$), if $H_{Y_1}(z) \geq H_{Y_2}(y)$ for all y.
[3] Mean residual life order ($Y_1 \leq_{MRL} Y_2$), if $m_{Y_1}(y) \geq m_{Y_2}(y)$ for all y.
[4] Likelihood ratio order ($Y_1 \leq_{LR} Y_2$), if $\frac{f_1(y)}{f_2(y)}$ decreasing in y.

Hence the following implication is revealed according to the above orderings
$Y_1 \leq_{LR} Y_2 \Rightarrow Y_1 \leq_{hr} Y_2 \Rightarrow Y_1 \leq_{MRL} Y_2$ and $Y_1 \leq_{hr} Y_2 \Rightarrow Y_1 \leq_{st} Y_2$.
Following theorem illustrate likelihood ratio ordering for $II_vD$ w.r.t the strongest likelihood.

**Theorem 4:** Let $Y_1 \sim II_vD(\delta_1, \rho_1, \eta_1)$, and $Y_2 \sim II_vD(\delta_2, \rho_2, \eta_2)$. If $\delta_1 = \delta_2 = \delta$, $(\rho_1 > \rho_2) > 1$, and $(\eta_1 > \eta_2)$ then $(Y_1 \leq_{lr} Y_2)$, $(Y_1 \leq_{st} Y_2)$, $(Y_1 \leq_{hr} Y_2)$, and $(Y_1 \leq_{MRL} Y_2)$.

**Proof:** To prove the result, the ratio of probability densities is

$$\frac{f_{Y_1}(y;\Theta_1)}{f_{Y_2}(y;\Theta_2)} = \frac{(\rho_2-1)\eta_2}{(\rho_1-1)\eta_1}\frac{\left\{e^{-(1-\delta_1 y)/\rho_1\eta_1 y} - e^{-(1-\delta_1 y)/\eta_1 y}\right\}}{\left\{e^{-(1-\delta_2 y)/\rho_2\eta_2 y} - e^{-(1-\delta_2 y)/\eta_2 y}\right\}}$$

Then,

$$\frac{d}{dy}\log\left\{\frac{f_{Y_1}(y;\Theta_1)}{f_{Y_2}(y;\Theta_2)}\right\} = \frac{\left[\{\eta_2 B\left(A_1 - \rho_1 A_2\right)\} - \{\eta_1 A\left(B_1 - \rho_2 B_2\right)\}\right]}{y^2 AB}$$

where, $A = \left\{e^{-(1-\delta_1 y)/\rho_1\eta_1 y} - e^{-(1-\delta_1 y)/\eta_1 y}\right\}$, $B = \left\{e^{-(1-\delta_2 y)/\rho_2\eta_2 y} - e^{-(1-\delta_2 y)/\eta_2 y}\right\}$,
$A_1 = e^{-(1-\delta_1 y)/\rho_1\eta_1 y}$, $A_2 = e^{-(1-\delta_1 y)/\eta_1 y}$, $B_1 = e^{-(1-\delta_2 y)/\rho_2\eta_2 y}$, and $B_2 = e^{-(1-\delta_2 y)/\eta_2 y}$.
Hence, If $\delta_1 = \delta_2 = \delta$, $(\rho_1 > \rho_2)$, and $(\eta_1 > \eta_2)$ then $\frac{d}{dy}\log\left\{\frac{f_{Y_1}(y;\Theta_1)}{f_{Y_2}(y;\Theta_2)}\right\} \leq 0$, which implies that $(Y_1 \leq_{lr} Y_2)$, $(Y_1 \leq_{st} Y_2)$, $(Y_1 \leq_{hr} Y_2)$, and $(Y_1 \leq_{MRL} Y_2)$. □

## 7. Stress-strength reliability

In this section, we study system reliability estimation under stress strength modeling, which possesses a cluster of applications, particularly in engineering statistics. Let $Y_1$ be the strength of the system subjected to stress $Y_2$. The system fails, when $Y_2 > Y_1$ (stress > strength), and functions smoothly, when $Y_1 > Y_2$ (stress < strength). Then the system reliability is measured by using the formula $R = Pr.\ (Y_1 > Y_2)$.
For two independent $r.v.'s$, $Y_1 \sim II_vD(\delta, \rho_1, \eta_1)$, and $Y_2 \sim II_vD(\delta, \rho_2, \eta_2)$, having the same parameter $\delta$. For the given, pdf of $Y_1$ and cdf of $Y_2$ the stress-strength reliability function $R$ is derived by

$$F_{II_vD}(y;\Theta_2) = \frac{\rho_2 e^{-(1-\delta y)/\rho_2\eta_2 y} - e^{-(1-\delta y)/\eta_2 y}}{(\rho_2-1)} \qquad \rho_2 \neq 1 \tag{16}$$

and,

$$f_{II_vD}(y;\Theta_1) = \frac{e^{-(1-\delta y)/\rho_1\eta_1 y} - e^{-(1-\delta y)/\eta_1 y}}{(\rho_1-1)\eta_1 y^2} \qquad \rho_1 \neq 1 \tag{17}$$

Therefore, possible derived cases are given by:
**Case (i):** when $\rho_1 \neq 1$, and $\rho_2 \neq 1$.

$$\begin{aligned} R &= \int_0^{1/\delta}\left\{\int_0^y f_{y_2}(y)dy\right\} f_{y_1}(y)dy = \int_0^{1/\delta} F_{Y_2}(y) f_{Y_1}(y)dy \\ &= \int_0^{1/\delta}\left\{\frac{\rho_2 e^{-(1-\delta y)/\rho_2\eta_2 y} - e^{-(1-\delta y)/\eta_2 y}}{(\rho_2-1)}\right\}\left\{\frac{e^{-(1-\delta y)/\rho_1\eta_1 y} - e^{-(1-\delta y)/\eta_1 y}}{(\rho_1-1)\eta_1 y^2}\right\} dy \\ &= \frac{\eta_2^2}{\rho_2-1}\left\{\frac{\rho_2^3}{\left(\rho_1\eta_1+\rho_2\eta_2\right)\left(\eta_1+\rho_2\eta_2\right)} - \frac{1}{\left(\rho_1\eta_1+\eta_2\right)\left(\eta_1+\eta_2\right)}\right\} \end{aligned}$$

**Case (ii):** when $\rho_1 \neq 1$, and $\rho_2 = 1$.

$$R = \frac{\eta_2}{(\rho_1 - 1)} \left\{ \frac{\rho_1^2 \eta_1}{(\rho_1\eta_1 + \eta_2)^2} + \frac{\rho_1}{(\rho_1\eta_1 + \eta_2)} - \frac{\eta_1}{(\eta_1 + \eta_2)^2} - 1 \right\}$$

**Case (iii):** when $\rho_1 = 1$, and $\rho_2 \neq 1$.

$$R = \frac{\eta_2^2}{(\rho_2 - 1)} \left\{ \frac{\rho_2^3}{(\eta_1 + \rho_2\eta_2)^2} - \frac{1}{(\eta_1 + \eta_2)^2} \right\}$$

**Case (iv):** when $\rho_1 = 1$, and $\rho_2 = 1$.

$$R = \frac{\eta_2^2 (3\eta_1 + \eta_2)}{(\eta_1 + \eta_2)^3}$$

## 8. Estimation of the parameters

Let us consider a random sample of $n$ observations, say $y_1$, $y_2$, ... ,$y_n$ drawn from $II_vD$ with desired defined parametric space $\Theta = (\rho, \delta, \eta)^T$ consisting $k \times 1$ vector of parameters. Then the completer data log-likelihood of the model when $\rho \neq 1$ is given by

$$\log L = \sum_{i=1}^{n} \log \left\{ e^{-(1-\delta y)/\rho\eta y} - e^{-(1-\delta y)/\eta y} \right\} - n \log(\rho - 1) - n \log(\eta) - \sum_{i=1}^{n} \log(y_i^2)$$

Let us take, $V_1 = e^{-(1-\delta y)/\rho\eta y}$, and $V_2 = e^{-(1-\delta y)/\eta y}$, then we re-write the above equation as

$$\log L = \sum_{i=1}^{n} \log \{V_1 - V_2\} - n \log(\rho - 1) - n \log(\eta) - \sum_{i=1}^{n} \log(y_i^2) \tag{18}$$

Now, the partial derivative for the above equation (18) with respect to the parameters $\rho$, $\delta$, and $\eta$ are obtained as:

$$\frac{\partial \log L}{\partial \rho} = \sum_{i=1}^{n} \frac{[1 - \delta y_i] V_1}{[V_1 - V_2] \rho^2 \eta y_i} - \frac{n}{(\rho - 1)} \tag{19}$$

$$\frac{\partial \log L}{\partial \delta} = \sum_{i=1}^{n} \frac{V_1 - \rho V_2}{\rho\eta[V_1 - V_2]} - 0 - 0 \tag{20}$$

$$\frac{\partial \log L}{\partial \eta} = \sum_{i=1}^{n} \frac{(1 - \delta y_i)[V_1 - \rho V_2]}{[V_1 - V_2] \rho\eta^2 y_i} - \frac{n}{\eta} \tag{21}$$

Equating the partial derivatives given in equations (19), (20), and (21) to zero, i,e $\frac{\partial \log L}{\partial \rho} = 0$, $\frac{\partial \log L}{\partial \delta} = 0$, and $\frac{\partial \log L}{\partial \eta} = 0$, we get $\hat{\rho}$, $\hat{\delta}$, and $\hat{\eta}$ as the $MLE_s$ of the parameters space $\Theta = \{(\rho, \delta, \eta) > 0\}$. Since the equation (19), (20) and (21) does not reveal the explicit solution, to get the parametric solution of the equations, one can counter this situation by employing the Newton Raphson algorithm. However, log-likelihood maximization could be done by using *nlm* or *optim* function in R-software.

The first-order derivatives of the log-likelihood equation of $II_vD(\rho,\delta,\eta)$ are defined in equations (19), (20), and, (21). The continuity of these partial derivatives reflects the second order partial derivatives of the log-likelihood equation does exist. If we denote the $MLE_s$ of the parametric space, $\Theta = \{(\rho,\delta,\eta) > 0\}$ by $\hat{\Theta} = \left\{(\hat{\rho},\hat{\delta},\hat{\eta}) > 0\right\}$, then the Fisher information matrix is given by

$$I\left(\Theta\right) = -E\begin{bmatrix} \frac{\partial^2 \log L}{\partial \rho^2} & \frac{\partial^2 \log L}{\partial \rho \partial \delta} & \frac{\partial^2 \log L}{\partial \rho \partial \eta} \\ \frac{\partial^2 \log L}{\partial \delta \partial \rho} & \frac{\partial^2 \log L}{\partial \delta^2} & \frac{\partial^2 \log L}{\partial \delta \partial \eta} \\ \frac{\partial^2 \log L}{\partial \eta \partial \rho} & \frac{\partial^2 \log L}{\partial \eta \partial \delta} & \frac{\partial^2 \log L}{\partial \eta^2} \end{bmatrix} \tag{22}$$

The second order partial derivatives of $I\left(\Theta\right)$ are given by

$$\frac{\partial^2 \log L}{\partial \rho^2} = \sum_{i=1}^{n} \frac{\left(1-\delta y_i\right)^2\left[V_1-V_2\right]V_1 - \left(1-\delta y_i\right)V_1\left\{2\rho\eta\left[V_1-V_2\right]y_i + \left(1-\delta y_i\right)V_1\right\}}{\left[\left(V_1-V_2\right)\rho^2\eta y_i\right]^2} + \frac{n}{(\rho-1)^2} \tag{23}$$

$$\frac{\partial^2 \log L}{\partial \delta^2} = \sum_{i=1}^{n} \frac{\left[V_1-V_2\right]\left[V_1-\rho^2V_2\right] - \left[V_1-\rho V_2\right]^2}{\left[\left(V_1-V_2\right)\rho\eta\right]^2} \tag{24}$$

$$\frac{\partial^2 \log L}{\partial \eta^2} = \sum_{i=1}^{n} \frac{\left(1-\delta y_i\right)^2\left[V_1-V_2\right]\left[V_1-\rho^2V_2\right] - \left(1-\delta y_i\right)\left[V_1-\rho V_2\right]\left\{2\rho\eta\left[V_1-V_2\right]y_i + \left(1-\delta y_i\right)\left[V_1-\rho V_2\right]\right\}}{\left[\left(V_1-V_2\right)\rho\eta^2 y_i\right]^2} - \frac{n}{\eta^2} \tag{25}$$

$$\frac{\partial^2 \log L}{\partial \delta \partial \rho} = \sum_{i=1}^{n} \frac{V_1\left[V_1-V_2\right]\left\{\left(1-\delta y_i\right) - \rho\eta y_i\right\} - V_1\left[V_1-\rho V_2\right]\left(1-\delta y_i\right)}{\left[\left(V_1-V_2\right)\eta\right]^2\rho^3 y_i} \tag{26}$$

$$\frac{\partial^2 \log L}{\partial \eta \partial \rho} = \sum_{i=1}^{n} \frac{V_1\left[V_1-V_2\right]\left(1-\delta y_i\right)^2 - V_1\left(1-\delta y_i\right)\left\{\rho\eta\left[V_1-V_2\right]y_i + \left(1-\delta y_i\right)\left[V_1-\rho\right]\right\}}{\left[\left(V_1-V_2\right)y_i\right]^2\left(\rho\eta\right)^3} \tag{27}$$

$$\frac{\partial^2 \log L}{\partial \eta \partial \delta} = \sum_{i=1}^{n} \frac{\eta\left[V_1-V_2\right]\left[V_1-\rho^2V_2\right]\left(1-\delta y_i\right) - \left[V_1-\rho V_2\right]\left\{\rho\eta^2\left[V_1-V_2\right]y_i - \left[V_1-\rho V_2\right]\right\}}{\left[\left(V_1-V_2\right)\rho\eta^2\right]^2 y_i} \tag{28}$$

It is difficult to obtain the expectation of second-order partial derivative expressions. Thus, in this situation, one can use the alternative measure called observed Fisher information matrix given by

$$I\left(\hat{\Theta}\right) = -\begin{bmatrix} \frac{\partial^2 \log L}{\partial \rho^2} & \frac{\partial^2 \log L}{\partial \rho \partial \delta} & \frac{\partial^2 \log L}{\partial \rho \partial \eta} \\ \frac{\partial^2 \log L}{\partial \delta \partial \rho} & \frac{\partial^2 \log L}{\partial \delta^2} & \frac{\partial^2 \log L}{\partial \delta \partial \eta} \\ \frac{\partial^2 \log L}{\partial \eta \partial \rho} & \frac{\partial^2 \log L}{\partial \eta \partial \delta} & \frac{\partial^2 \log L}{\partial \eta^2} \end{bmatrix}_{(\rho,\delta,\eta)=(\hat{\rho},\hat{\delta},\hat{\eta})} \tag{29}$$

The inverse of the observed Fisher information matrix $I\left(\hat{\Theta}\right)$, will give diagonal elements as variances whereas the off-diagonal elements represents the co-variances of the matrix. The approximate $(1-\sigma)$ 100% confidence intervals for all the three parameters of $II_vD$ i,e $\rho$, $\delta$, and $\eta$ are $\hat{\rho} \pm \psi_{\sigma/2}\sqrt{V(\hat{\rho})}$, $\hat{\delta} \pm \psi_{\sigma/2}\sqrt{V(\hat{\delta})}$, and $\hat{\eta} \pm \psi_{\sigma/2}\sqrt{V(\hat{\eta})}$ respectively. where, $V(\hat{\rho})$, $V(\hat{\delta})$, and $V(\hat{\eta})$ are variances given in diagonal elements of $I(\Theta)^{-1}$ and the upper $(\sigma/2)$ percentile of a standard normal distribution is denoted by $\psi_{\sigma/2}$.

## 9. Simulation

In this section, a Monte Carlo simulation study with 1000 repetitions has been performed through R-software, to illustrate the theoretical findings of the proposed model. Since data generation has been done by employing the acceptance-rejection algorithm due to the complexity of the quantile function. The performance of the parametric space $\Theta$ with different sample sizes $n$ = (25, 75, 125, 175, 250, 400) are checked by observing the calculated $AAbias$ and the $MSE$ of the estimated parameters. The output result of the simulation are summarized in Table 1, given below:

**Table 1: Simulated results of parameters for different sample sizes**

| $(\delta, \eta, \rho)$ | $n$ | $AAbias$ | | | $MSE$ | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\delta}$ | $\hat{\eta}$ | $\hat{\rho}$ | $\hat{\delta}$ | $\hat{\eta}$ | $\hat{\rho}$ |
| $(0.53, 0.92, 0.94)$ | 025 | 0.13926 | 0.36362 | 36.61545 | 0.01939 | 0.13222 | 1340.691 |
| | 075 | 0.05185 | 0.25571 | 02.96885 | 0.00269 | 0.06539 | 08.81406 |
| | 125 | 0.03573 | 0.22752 | 01.26163 | 0.00128 | 0.05177 | 01.59170 |
| | 175 | 0.02620 | 0.20220 | 00.74678 | 0.00069 | 0.04088 | 00.55769 |
| | 250 | 0.02049 | 0.18039 | 00.59632 | 0.00042 | 0.03254 | 00.35560 |
| | 400 | 0.01488 | 0.16094 | 00.50596 | 0.00022 | 0.02590 | 00.25599 |
| $(1.53, 0.92, 1.04)$ | 025 | 0.13811 | 0.32654 | 21.63234 | 0.01907 | 0.10663 | 467.9582 |
| | 075 | 0.05757 | 0.21819 | 02.32618 | 0.00331 | 0.04761 | 05.41112 |
| | 125 | 0.03954 | 0.19292 | 01.01383 | 0.00156 | 0.03722 | 01.02784 |
| | 175 | 0.02690 | 0.16254 | 00.61841 | 0.00072 | 0.02642 | 00.38243 |
| | 250 | 0.02142 | 0.13128 | 00.46323 | 0.00046 | 0.01724 | 00.21458 |
| | 400 | 0.01614 | 0.12249 | 00.40460 | 0.00026 | 0.01500 | 00.16370 |
| $(0.53, 1.92, 0.94)$ | 025 | 0.30424 | 0.77793 | 37.09144 | 0.09256 | 0.60518 | 1375.775 |
| | 075 | 0.11171 | 0.52213 | 03.76642 | 0.01248 | 0.27262 | 14.18596 |
| | 125 | 0.07156 | 0.44599 | 01.42634 | 0.00512 | 0.19891 | 02.03445 |
| | 175 | 0.05285 | 0.39587 | 00.70007 | 0.00279 | 0.15671 | 00.49010 |
| | 250 | 0.04294 | 0.36643 | 00.59156 | 0.00184 | 0.13427 | 00.34995 |
| | 400 | 0.02577 | 0.31186 | 00.45712 | 0.00066 | 0.09726 | 00.20896 |

It is easily noticed in Table 1 while increasing the sample size the $AAbias$ and $MSE$ are reducing. Hence, this admits the consistency property of the parametric space of our model.

## 10. Applications

This section is about the model applicability checkup on a real-life data basis. The data set has been analyzed. In this study, the performance of newly developed $II_vD$ is compared with existing distributions like the Exponential distribution ($ED$), Inverse Exponential distribution ($IED$), generalized Exponential distribution ($GED$), and the generalized inverse Exponential distribution ($GIED$). The best model is chosen having minimum value of Akaike information criteria ( defined as, $AIC$ = - 2 $\log L(y, \Theta)$ + $2k$ ), Bayesian information criteria ( defined as, $BIC$ = - 2 $\log L(y, \Theta)$ + $k \log(n)$ ), Hannan Quinn information criteria (defined as, $HQIC$ = - $2 \log L(y, \Theta)$ + $2k \log[\log(n)]$ ), and the goodness of fit tests, that includes Cramér-von Mises ($C_{vm}$) test, Anderson Darling ($A_n$) test, Kolmogorov Smirnov ($KS$) statistic respectively. The constant $k$ denotes the number of parameters in the model.

For a given real-life data set, the results of different information criteria, the goodness of fit measures, and the *p-value* are reported in Table 2. The $II_vD$ is compared with existing models whose probability density functions are given by

$$\begin{aligned} ED = f(y;\delta) &= \delta e^{-\delta y} && (30)\\ IED = f(y;\delta) &= \frac{\delta}{y^2} e^{-\delta/y} && (31)\\ GED = f(y;\delta,\eta) &= \delta\eta e^{-\delta y}\left(1-e^{-\delta y}\right)^{\eta-1} && (32)\\ GIED = f(y;\delta,\eta) &= \frac{\delta\eta}{y^2} e^{-\delta/y}\left(1-e^{-\delta/y}\right)^{\eta-1} && (33) \end{aligned}$$

The given data set is taken from the paper published by Ahmed M. A. (2021), which represents the lifetime (in hours) of traditional lights, for 50 devices. The data are: 0.913, 0.786, 0.860, 0.904, 0.971, 0.616, 0.961, 0.789, 0.817, 0.722, 0.956, 0.835, 0.853, 0.692, 0.850, 0.677, 0.898, 0.965, 0.820, 0.964, 0.865, 0.947, 0.798, 0.746, 0.926, 0.709, 0.615, 0.747, 0.931, 0.913, 0.895, 0.745, 0.839, 0.766, 0.690, 0.531, 0.838, 0.846, 0.876, 0.817, 0.719, 0.907, 0.915, 0.879, 0.890, 0.865, 0.869, 0.772, 0.933, 0.875.

**Table 2: Results of information measures and goodness of fit tests**

| $Models$ | $Part - I$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\hat{\delta}$ | $\hat{\eta}$ | $\hat{\rho}$ | $\log L$ | $AIC$ | $BIC$ | $HQIC$ |
| $II_vD$ | 1.02014 | 0.02359 | 7.68817 | 50.175 | -94.35043 | -88.61436 | -92.16610 |
| $GIED$ | 7.68390 | 7145.98 | - | 45.414 | -86.82726 | -83.00322 | -85.37104 |
| $GED$ | 8.67100 | 838.120 | - | 34.645 | -65.29067 | -61.46663 | -63.83445 |
| $IED$ | 0.81630 | - | - | -40.742 | 83.48415 | 85.39617 | 84.21226 |
| $ED$ | 1.20440 | - | - | -40.699 | 83.39836 | 85.31039 | 84.12647 |
| $Models$ | $Part - II$ | | | | | | |
| | $\hat{\delta}$ | $\hat{\eta}$ | $\hat{\rho}$ | $C_{vm}$ | $A_n$ | $KS$ | *p-value* |
| $II_vD$ | 1.02014 | 0.02359 | 7.68817 | 0.02439 | 0.19464 | 0.05978 | 0.9941 |
| $GIED$ | 7.68390 | 7145.98 | - | 0.09202 | 0.57741 | 0.11345 | 0.5405 |
| $GED$ | 8.67100 | 838.120 | - | 0.34397 | 2.05467 | 0.16876 | 0.1159 |
| $IED$ | 0.81630 | - | - | 0.24371 | 1.47237 | 0.56858 | $1.82 \times 10^{-14}$ |
| $ED$ | 1.20440 | - | - | 0.18330 | 1.11543 | 0.50322 | $2.01 \times 10^{-11}$ |

From Table 2, it is well observed that the $II_vD$ fits best as it has a minimum value for all the information criteria ( $AIC$, $BIC$, and $HQIC$ ) as well as the goodness of fit tests, and a higher *p-value*.

## 11. Conclusion

This manuscript presents an intervention-based model called inverted intervened Exponential distribution. The graphical plots based on a different set of parameters for pdf and hazard rate are shown, pdf having different shapes where the hazard rate function has upside down and exponentially increasing shapes, that could be useful to model different types of failure data. The essential statistical and reliability properties are derived. The parameters have been estimated by using the method of maximum likelihood estimation. A Monte Carlo simulation study has been done, where it is observed that both bias and mean square error for all the parameter decreases while increasing the sample size. The real-life data set have been analyzed and it is predicted that the values of all the information measures and the different goodness of fit tests for the proposed distribution are very less, with a higher *p-value* as compared to the existing models, which ensures the real-life applicability of the model.

## Acknowledgements

The first author is very grateful to Pondicherry University for providing him with a research fellowship to carry out this work. Also, the authors are grateful to the editor of the journal and the anonymous reviewers for their valuable suggestions.

## References

Ahmed, M. A. (2021). On the identified power Topp-Leone distribution: properties, simulation, and applications. *Thailand Statistician*, **19**, 838-854.

Balakrishnan, K. (2019). *Exponential Distribution: Theory, Methods and Applications.* CRC Press. url: https://books.google.co.in/books?id=aTmDDwAAQBAJ

Finkelstein, M. (2008). *Failure Rate Modelling for Reliability and Risk.* Springer Science & Business Media. doi: 10.1007/978-1-84800-986-8

Jiang, R. Ji, P. and Xiao, X. (2003). Aging property of unimodal failure rate models. *Reliability Engineering & System Safety*, **79**, 113-116. doi:10.1016/S0951-8320(02)00175-8

Rényi, A. (1961, January). On measures of entropy and information. *In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (pp. 547-561). University of California Press.

Shaked, M. and Shanthikumar, J. G. (2007). Stochastic orders. *Springer*, New York. doi: 10.1007/978-0-387-34675-5

Shanmugam, R. (1985). An intervened Poisson distribution and Its medical application. *Biometrics*, **41**, 1025-1029. doi:10.2307/2530973

Shanmugam, R. Bartolucci, A. A. and Singh, K. P. (2002). The analysis of neurologic studies using an extended exponential model. *Biometrics*, **59**, 81-85. doi: 10.1016/S0378-4754(01)00395-0

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 41–49

# Estimation of AUC of Mixture ROC Curve in the Presence of Measurement Errors

**G. Siva and R. Vishnu Vardhan**
*Department of Statistics, Ramanujan School of Mathematical Sciences,*
*Pondicherry University, Puducherry, India*

Received: 17 December 2021; Revised: 15 March 2022; Accepted: 18 March 2022

## Abstract

Receiver Operating Characteristic (ROC) curve is one of the widely used classification tool and its applications can be seen in diversified fields of science and engineering. In this work, we made an attempt to examine the influence of measurement errors on the AUC of a mixture ROC curve. A bias corrected estimator is proposed and derived. The proposed work is supported with real dataset and simulation studies and results show that the proposed bias corrected estimator helps in correcting the AUC with minimum bias and minimum mean square error.

*Key words:* Mixture ROC curve; Area under the curve; Measurement errors.

## 1. Introduction

Over the years, classification problems have gained a lot of attention in terms of theoretical development and practical applications in various disciplines. To handle such problems, one of the classification tool is the Receiver Operating Characteristic (ROC) curve, originated during World War II for analyzing the radar images. In diagnostic medicine, ROC curve is widely used for evaluating the test's performance and also useful in comparing diagnostic tests by means of Area under the Curve (AUC) and Sensitivities. It is a unit square graphical plot between false positive rate (1-specificity) and true positive rate (sensitivity) at various threshold values. The AUC of an ROC curve plays an important role in assessing the performance of a diagnostic test(s) and also measures the ability of a biomarker to distinguish between two groups.

Measurement error (ME) problems are among the oldest in the history of statistics and can be of great practical and economic importance. It is the difference between a measured quantity and its true value. In diagnostic medicine, markers are subject to substantial measurement errors which may be attributed to instruments used in the laboratory, knowledge of the technicians, biological variability, temporal changes in subjects, *etc.* Shear *et al.* (1987) has taken the measurements of systolic and diastolic blood pressure on children's, which were used as forecasters of future hypertension. Carracio *et al.* (1995) has done a study on the children to predict the presence or absence of bacterial menengitis using cerebrospinal

Corresponding Author: R. Vishnu Vardhan
Email: vrstatsguru@gmail.com

fluid. Since the outcome of the test in identifying the menengitis is attributed to either laboratory equipment or technician, which may lead to the phenomenon of observing errors in the measured quantities. With the above two examples, it can be understood that in most of the data collections which purely depend upon the laboratory equipments, technicians etc. there are high chances of having errors in the measurements. For more examples on ME, readers can look into Begg and Greene (1983), Begg and McNeil (1988), Berbaum *et al.* (1989), Buonaccorsi (2010) and Fuller (2009).

In ROC analysis, the most popular one is the Bi-normal ROC model, where the two populations assumed to follow normal distribution. The estimation of AUC and its measures have been addressed by several authors and a few to mention are Hanley and McNeil (1982), Faraggi and Reiser (2002), Zhou *et al.* (2009), Vishnu Vardhan and Sarma (2010). However, when the data is exposed to measurement errors the estimation of AUC will be a problem of interest. Because as the measurements are deviated from their true value, it leads to produce spurious AUC. Hence, the AUC has to be corrected by means of an estimator. The seminal work on providing an estimator to correct the AUC was addressed by Coffin and Sukhatme (1996). They showed that in the presence of measurement errors, the AUC will be biased downwards and also came out with a bias corrected estimator that corrects the AUC. In similar lines, Faraggi (2000) and Reiser (2000) have worked on estimating the confidence intervals for the AUC in the presence of measurement error. Tosteson *et al.* (2005) studied the effect of measurement errors on AUC of an ROC curve by expressing the magnitude of the measurement error as a ratio of two variances; graphical and simulated environment were presented to show the effect of ME.

The above methodologies works well only when the knowledge on class labels is known. Even though the class labels are known, in most of the practical situations we may get observed with bi-modal or multi-model patterns within each known population. In such scenarios the existing binormal structure and correction of AUC in the presence of measurement error may not feasible to execute.

In this work, we proposed a Mixture ROC model which takes into the account of a possible mean differences between populations. Let us assume that two sub components are identified in diseased population and defined as $D_1$ and $D_2$. Now, we take into the account of the following possible mean differences such as $\mu_{D_1} - \mu_H$ and $\mu_{D_2} - \mu_{D_1}$ ($\mu_{D_2} \geq \mu_{D_1} \geq \mu_H$) and the same is shown in Figure 1. In Section 3, the same scenario is illustrated using OGTT dataset.

In section 2, we present the methodology of mixture ROC curves and its correction in measurement error. In section 3, a real dataset is considered to assess the performance of the proposed methodology and in section 4, monte carlo simulations are performed to compare the MSE of estimated and bias corrected estimator of the true AUC values. This has helped to examine how the bias and MSE of the estimators are influenced by measurement errors at different sample sizes.

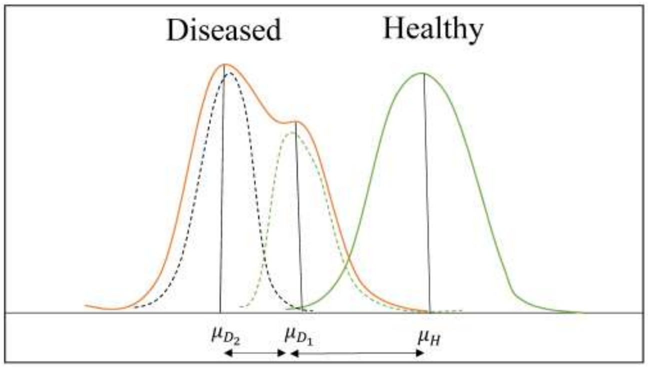


**Figure 1: Hypothetically overlapping density curves of Healthy and Diseased populations**

## 2. Methodology

### 2.1. Mixture receiver operating characteristic curve with measurement error

Let us consider the data where the class labels of subjects are known. In most of the cases, we directly start with developing a classifier rule. But there are chances of having several subgroups in each of the known populations. For instance, consider the oral glucose tolerance test (OGTT) data, where the subjects disease status is defined. However, on investigating the diseased population, it resulted with a bi-modal pattern. This indicates that there are two sub populations with in the diseased population (Figure 2).

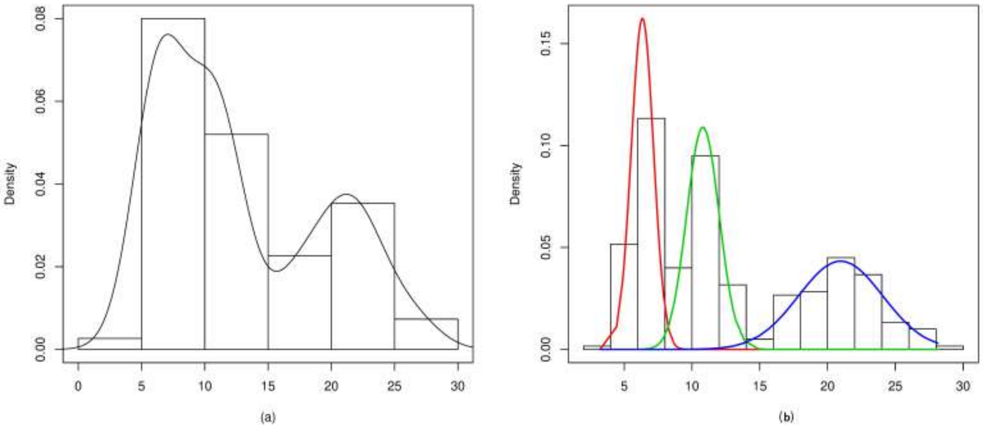


**Figure 2: (a) The overall density plot of OGTT, (b) Plot after identifying the components in the OGTT data set.**

Let us consider a binary classified data (Healthy, $H$ and Diseased, $D$) where the $D$ population consists of two sub populations within it. The identification of two sub populations ($D_1$ and $D_2$) will be done through EM algorithm. Let $\mu_H, \mu_{D_1}, \mu_{D_2}$ and $\sigma^2_H, \sigma^2_{D_1}, \sigma^2_{D_2}$ are the means and variances of three populations, respectively.

The expressions for the False Positive Rate (1-specificity) and True Positive Rate (sensitivity) in the mixture form is defined as

$$FPR = x(c) = \lambda_1\, x(c_1) + \lambda_2\, x(c_2) \tag{1}$$

$$TPR = y(c) = \lambda_1\, y(c_1) + \lambda_2\, y(c_2) \tag{2}$$

here, $\lambda_1$ and $\lambda_2$ are the mixing proportions; $c_1$ and $c_2$ are threshold values for the pairs $(D_1, H)$ and $(D_2, D_1)$.

By definition, we write

$$x(c_1) = \Phi\left(\frac{\mu_H - c_1}{\sigma_H}\right) \quad ; \quad x(c_2) = \Phi\left(\frac{\mu_{D_1} - c_2}{\sigma_{D_1}}\right) \tag{3}$$

$$y(c_1) = \Phi\left(\frac{\mu_{D_1} - c_1}{\sigma_{D_1}}\right) \quad ; \quad y(c_2) = \Phi\left(\frac{\mu_{D_2} - c_2}{\sigma_{D_2}}\right) \tag{4}$$

The expressions for $c_1$ and $c_2$ will take the following form

$$c_1 = \mu_H - \sigma_H \Phi^{-1}[x(c_1)] \quad ; \quad c_2 = \mu_{D_1} - \sigma_{D_2} \Phi^{-1}[x(c_2)] \tag{5}$$

where $\Phi^{-1}$ is the inverse cumulative distribution function of normal. The mixture ROC expression is derived by substituting (5) in (2) and is given in (6)

$$ROC = \lambda_1 \left[\Phi\left(\frac{\mu_{D_1} - \mu_H}{\sigma_{D_1}} + \frac{\sigma_H}{\sigma_{D_1}}\Phi^{-1}[x(c_1)]\right)\right] + \lambda_2 \left[\Phi\left(\frac{\mu_{D_2} - \mu_{D_1}}{\sigma_{D_2}} + \frac{\sigma_{D_1}}{\sigma_{D_2}}\Phi^{-1}[x(c_2)]\right)\right] \tag{6}$$

In general, if the diseased component has '$p$' sub populations then (6) can be rewritten as

$$ROC(c) = \sum_{i=1}^{p} \lambda_i \left[\Phi\left(A_i + B_i\ \Phi^{-1}[FPR]\right)\right] \tag{7}$$

where $\sum\limits_{i=1}^{p} \lambda_i = 1$; $A_i = \dfrac{\mu_i - \mu_{i-1}}{\sigma_i}$; $B_i = \dfrac{\sigma_{i-1}}{\sigma_i}$

### 2.2. Corrected bias approximation

Let us define $X_1, X_2, \ldots, X_m \overset{iid}{\sim} N(\mu_H, \sigma_H^2)$ , $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu_{D_1}, \sigma_{D_1}^2)$ and $Z_1, Z_2, \ldots, Z_k \overset{iid}{\sim} N(\mu_{D_2}, \sigma_{D_2}^2)$, then the AUC expression for mixture ROC curve is given as

$$mAUC = \theta = \lambda_1 \Phi\left(\frac{\mu_{D_1} - \mu_H}{\sqrt{\sigma_{D_1}^2 + \sigma_H^2}}\right) + \lambda_2 \Phi\left(\frac{\mu_{D_2} - \mu_{D_1}}{\sqrt{\sigma_{D_2}^2 + \sigma_{D_1}^2}}\right)$$

If the observations in $H, D_1$ and $D_2$ are observed with measurement errors then we define

$$x_i = X_i + u_i, \quad i = 1, 2, \ldots, m; \quad u_i \sim iid\ N(0, \sigma_u^2)$$

$$y_i = Y_i + v_j, \quad j = 1, 2, \ldots, n; \quad v_i \sim iid\ N(0, \sigma_v^2)$$

$$z_k = Z_k + \gamma_k, \quad k = 1, 2, \ldots, l; \;\; \gamma_k \sim iid\; N(0, \sigma_\gamma^2)$$

we assume $u_i, v_j, z_k, X_i, Y_j$ and $Z_k$ are all independent. The natural estimator of $\theta$ is

$$m\hat{AUC} = \hat{\theta} = \lambda_1\, \hat{\theta}_1 + \lambda_2\, \hat{\theta}_2$$

where $\hat{\theta}_1 = \Phi\left(\dfrac{\hat{\mu}_{D_1} - \hat{\mu}_H}{\sqrt{s_{D_1}^2 + s_H^2}}\right)$, $\hat{\theta}_2 = \Phi\left(\dfrac{\hat{\mu}_{D_2} - \hat{\mu}_{D_1}}{\sqrt{s_{D_2}^2 + s_{D_1}^2}}\right)$

here $s_H^2$, $s_{D_1}^2$ and $s_{D_2}^2$ are the sample variances. Using Taylor series expansion, it can be shown that $E(\hat{\theta}) = \theta + O(1)$. Since, the observations are measured with errors, the resulting area estimates i.e., AUC's will be biased downward. By adopting the methodology of Coffin and Sukhathme (1996), the expressions for $\hat{\theta_1}$ and $\hat{\theta_2}$ are

$$\begin{aligned} E(\hat{\theta}_1) &\approx P(Y > X + \delta_1) = \int\!\!\int [1 - G_Y(s+t)] f_X(s) f_{\delta_1}(t)\, dt\, ds \\ &\approx \theta_1 - \frac{1}{2} Var(\delta_1) \int g_Y^T(s) f_X(s) ds \\ E(\hat{\theta}_2) &\approx P(Z > Y + \delta_2) = \int\!\!\int [1 - G_Z(s+t)] f_Y(s) f_{\delta_2}(t)\, dt\, ds \\ &\approx \theta_2 - \frac{1}{2} Var(\delta_2) \int g_Z^T(s) f_Y(s) ds \end{aligned}$$

where $\delta_1 = u - v \sim N(0, \sigma_u^2 + \sigma_v^2)$ and $\delta_2 = v - \gamma \sim N(0, \sigma_v^2 + \sigma_\gamma^2)$, here $G_Y(.)$, $G_Z(.)$ are distribution functions of $Y, Z$ and $f_{\delta_1}(.), f_{\delta_2}(.)$ are density functions of $\delta_1$, $\delta_2$. Thus, the approximate bias in using $\hat{\theta}_1$ and $\hat{\theta}_2$ to estimate $\theta$ will be

$$\begin{aligned} -B_1 &= -\frac{1}{2} Var(\delta_1) \int g_Y^T(s) f_X(s) ds \\ &= -\frac{\frac{1}{2}(\sigma_u^2 + \sigma_v^2)}{\sqrt{2\pi}\tau_{XY}^2} \left(\frac{\mu_{D_1} - \mu_H}{\tau_{XY}}\right) exp\left\{-\frac{1}{2}\left(\frac{\mu_{D_1} - \mu_H}{\tau_{XY}}\right)^2\right\} \\ -B_2 &= -\frac{1}{2} Var(\delta_2) \int g_Z^T(s) f_Y(s) ds \\ &= -\frac{\frac{1}{2}(\sigma_v^2 + \sigma_\gamma^2)}{\sqrt{2\pi}\tau_{YZ}^2} \left(\frac{\mu_{D_2} - \mu_{D_1}}{\tau_{YZ}}\right) exp\left\{-\frac{1}{2}\left(\frac{\mu_{D_2} - \mu_{D_1}}{\tau_{YZ}}\right)^2\right\} \end{aligned}$$

where $\tau_{XY} = \sqrt{\sigma_H^2 + \sigma_{D_1}^2}, \tau_{YZ} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2}$, then the bias corrected estimator for $\theta$ in the mixture form is defined as

$$mAUC_{corr} = \theta^* = \lambda_1\, \theta_1^* + \lambda_2\, \theta_2^* \tag{8}$$

where $\theta_1^* = \hat{\theta}_1 + \hat{B}_1$ and $\theta_2^* = \hat{\theta}_2 + \hat{B}_2$. Using the unbiased estimates $\hat{\sigma}_u^2$, $\hat{\sigma}_v^2$ and $\hat{\sigma}_\gamma^2$, the estimated value of $B_1$ and $B_2$ will be

$$\hat{B}_1 = \frac{(\hat{\sigma}_u^2 + \hat{\sigma}_v^2)}{2\sqrt{2\pi}(s_H^2 + s_{D_1}^2 - \hat{\sigma}_u^2 - \hat{\sigma}_v^2)} \left(\frac{\hat{\mu}_{D_1} - \hat{\mu}_H}{\sqrt{s_H^2 + s_{D_1}^2 - \hat{\sigma}_u^2 - \hat{\sigma}_v^2}}\right) exp\left\{-\frac{1}{2}\left(\frac{\hat{\mu}_{D_1} - \hat{\mu}_H}{s_H^2 + s_{D_1}^2 - \hat{\sigma}_u^2 - \hat{\sigma}_v^2}\right)^2\right\}$$

$$\hat{B}_2 = \frac{(\hat{\sigma}_v^2 + \hat{\sigma}_\gamma^2)}{2\sqrt{2\pi}(s_{D_1}^2 + s_{D_2}^2 - \hat{\sigma}_v^2 - \hat{\sigma}_\gamma^2)} \left(\frac{\hat{\mu}_{D_2} - \hat{\mu}_{D_1}}{\sqrt{s_{D_1}^2 + s_{D_2}^2 - \hat{\sigma}_v^2 - \hat{\sigma}_\gamma^2}}\right) exp\left\{-\frac{1}{2}\left(\frac{\hat{\mu}_{D_2} - \hat{\mu}_{D_1}}{s_{D_1}^2 + s_{D_2}^2 - \hat{\sigma}_v^2 - \hat{\sigma}_\gamma^2}\right)^2\right\}$$

The confidence intervals (CI) for corrected AUC measures are obtained using

$$\widehat{mAUC}_{corr} \pm Z_{\left(1-\frac{\alpha}{2}\right)}\, S.E(\widehat{mAUC}_{corr})$$

## 3. Real data set

The OGTT dataset (Lasko *et al.*, 2005) consists of 21 samples of Healthy and a mixture of Diseased individuals. In order to show the measurement error in the data, random error observations are generated $N(0, 1.2)$ and added to the original samples. This is done to mimic the situation where the actual data is affected with ME.

Along with the accuracy measures, it's bias and MSE's are obtained and presented in table (1). From the results, it is shown that by adding error observations to the original data, the accuracy measure is affected and biased downwards (i.e., from $\theta = 0.94626$ to $\hat{\theta} = 0.91641$). In such situation, the proposed bias corrected estimator helps in achieving the true accuracy and which has minimum bias and minimum MSE when compared with the estimated accuracy. The ROC curves are drawn for the original dataset (True ROC) and after adding error observations to the data (ROC with ME). From Figure 3, it is clearly seen that errors in measurement will affect the shape of the ROC curve and it is downwards than the true ROC curve.

**Table 1: Bias and MSE of estimated and corrected estimator of AUC of OGTT dataset**

| | $\hat{\theta}$<br>(True AUC) | $\hat{\theta}_{ME}$<br>(Uncorrected AUC) | Bias | MSE | $\hat{\theta}^*$<br>(Corrected AUC) | Bias | MSE |
|---|---|---|---|---|---|---|---|
| Mixture ROC | 0.94626 | 0.91641 | -0.02985 | 0.00089 | **0.94061** | -0.00565 | **0.00003** |

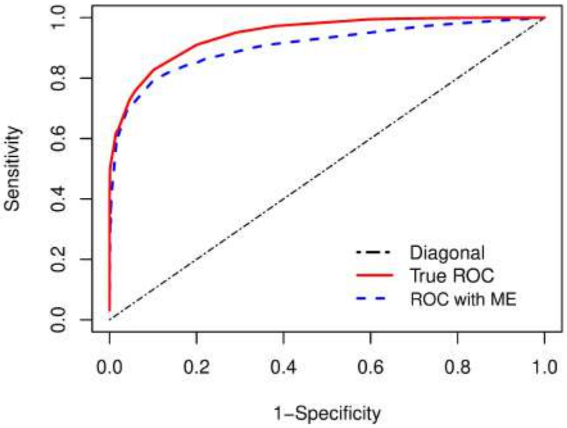


**Figure 3: True and contaminated ROC (with ME) curves for OGTT dataset**

## 4. Simulation studies

Monte Carlo simulations are carried out to illustrate the behavior of the proposed bias corrected estimator in the mixture ROC forms when the observations are measured with error.

In Table 2, two sets of means and variances are considered along with the initial values for mixing proportions. Set A and set B has unequal and equal variances, respectively. To show the influence of measurement errors in the data, the error component, $\epsilon \sim N(0, 1.9)$ is added to set A and B & AUC's are estimated (before and after correction). In each population, random samples of size $n = \{25, 50, 100, 200\}$ were generated using the parameter values listed in Table 2.

**Table 2: Considered parameters for simulation studies**

| Sets | $\lambda_1$ | $\lambda_2$ | $\mu_H$ | $\mu_{D_1}$ | $\mu_{D_2}$ | $\sigma_H$ | $\sigma_{D_1}$ | $\sigma_{D_2}$ |
|---|---|---|---|---|---|---|---|---|
| A | 0.5 | 0.5 | 29.3 | 32.5 | 35.2 | 1.0 | 1.5 | 2.0 |
| B | 0.5 | 0.5 | 29.3 | 32.5 | 35.2 | 1.5 | 1.5 | 1.5 |

The estimated and bias corrected AUC values along with its bias and mean square errors at various sample sizes are presented in Table 3.

**Table 3: The Bias, MSE of the estimated and bias-corrected estimator of AUC**

| Sets | $\hat{\theta}$ | n | $\hat{\theta}_{ME}$<br>(CI$_L$,CI$_U$) | Bias | MSE | $\hat{\theta}^*$<br>(CI$_L$,CI$_U$) | Bias | MSE |
|---|---|---|---|---|---|---|---|---|
| A | 0.91099 | 25 | 0.83777<br>(0.82228,0.85326) | -0.07322 | 0.01855 | 0.94898<br>(0.92649,0.97146) | 0.03799 | 0.00144 |
| | | 50 | 0.85403<br>(0.83667,0.87139) | -0.05696 | 0.01356 | 0.93654<br>(0.91489,0.95820) | 0.02555 | 0.00065 |
| | | 100 | 0.86416<br>(0.84844,0.87988) | -0.04683 | 0.01212 | 0.92650<br>(0.90554,0.94746) | 0.01551 | 0.00024 |
| | | 200 | 0.86922<br>(0.84916,0.88929) | -0.04177 | 0.01101 | 0.91561<br>(0.88365,0.94756) | 0.00461 | 0.00002 |
| B | 0.91673 | 25 | 0.83488<br>(0.81679,0.85298) | -0.08149 | 0.01508 | 0.93858<br>(0.91483,0.96234) | 0.02221 | 0.00049 |
| | | 50 | 0.86306<br>(0.84866,0.87746) | -0.03150 | 0.00917 | 0.93614<br>(0.90054,0.97174) | 0.01977 | 0.00039 |
| | | 100 | 0.87897<br>(0.86447,0.89347) | -0.05331 | 0.00819 | 0.92754<br>(0.90604,0.94904) | 0.01117 | 0.00012 |
| | | 200 | 0.88487<br>(0.86705,0.90268) | -0.03740 | 0.00682 | 0.90702<br>(0.87457,0.93947) | -0.00935 | 0.00009 |

From the results, it is understood that the area estimates ($\hat{\theta}_{ME}$) are biased downward at each sample size. Using the proposed mixture of bias corrected approximation, it is observed that the bias corrected estimator of AUC's ($\hat{\theta}^{*}$) are closer to the true AUC's ($\hat{\theta}$) values and has minimum MSE when compared with the estimated AUC's ($\hat{\theta}$). Using the proposed methodology of bias corrected approximation in mixture ROC, we can obtain the reliable estimates of AUC's in the presence of measurement errors.

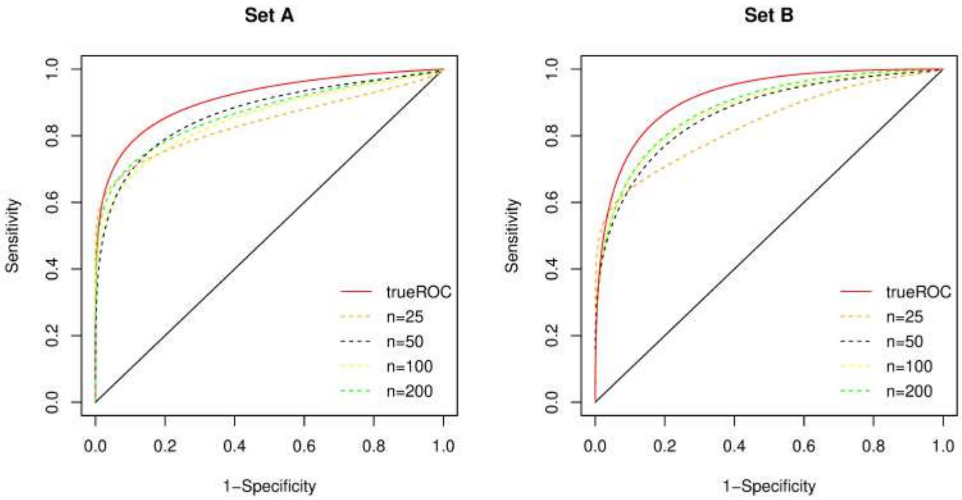


**Figure 4: The true and estimated ROC curves at various sample sizes**

The graphical representation of the true mixture ROC curve and the estimated mixture ROC curves (errors in the data) at various sample sizes is presented in Figure 4. From this graphical ROC plots also it is understood that, the resulting area estimates are downward in the presence of measurement errors.

## 5. Summary

In this paper, we made an attempt to address the problem of measurement errors in estimating the AUC of mixture normal ROC model. A bias corrected approximation has been defined in the mixture form. The methodology is supported by a OGTT dataset and monte carlo simulation studies. Results indicates that the proposed bias corrected estimator provides the corrected AUC's and it will be closer to the true AUC values with minimum bias and minimum MSE.

## References

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, **39**, 207–215.

Begg, C. B. and McNeil, B. J. (1988). Assessment of radiologic tests: control of bias and other design considerations. *Radiology*, **167**, 565–569.

Berbaum, K. S., Dorfman, D. D. and Franken Jr, E. A. (1989). Measuring observer performance by ROC analysis: indications and complications. *Investigative Radiology*, **24**, 228–233.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications.* Chapman & Hall/CRC Press.

Carraccio, C., Blotney, K. and Franken, E. A. (1989). Cerebrospinal fluid analysis in systematically children without central nervous system disease. *Pediatrics*, **96**, 48–51.

Cheam, A. S., and McNicholas, P. D. (2016). Modelling receiver operating characteristic curves using Gaussian mixtures. *Computational Statistics & Data Analysis*, **93**, 192-208.

Coffin, M. and Sukhatme, S. (1997). Receiver operating characteristic studies and measurement errors. *Biometrics*, **53**, 823–837.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Gonen, M. (2013). Mixtures of receiver operating characteristic curves. *Academic Radiology*, **20**, 831–837.

Faraggi, D. (2000). The effect of random measurement error on receiver operating characteristic (ROC) curves. *Statistics in Medicine*, **19**, 61–70.

Fuller, W. A. (2009). *Measurement Error Models.* John Wiley & Sons.

McClish, D. K. (1992). Combining and Comparing Area Estimates across Studies or Strata. *Medical Decision Making*, **12**, 274–279.

Thomas A Lasko et al. (2005). The use of receiver operating characteristics curves in biomedical informatics. *Journal of Biomedical Informatics*, **38**, 404–415.

Perkins, N. J., Schisterman, E. F. and Vexler, A. (2009). Generalized ROC curve inference for a biomarker subject to a limit of detection and measurement error. *Statistics in Medicine*, **28**, 1841–1860.

Reiser, B. (2000). Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine*, **19**, 2115–2129.

Shear et al. (1987). Designation of children with high blood pressure—considerations on percentile cut points and subsequent high blood pressure: the Bogalusa Heart Study. *American Journal of Epidemiology*, **125**, 73–84.

Tosteson et al. (2005). Measurement error and confidence intervals for ROC curves. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **47**, 409–416.

Vardhan, R. V., and Sarma, K. V. S. (2010). Estimation of the area under the ROC curve using confidence intervals of mean. *ANU Journal of Physical Sciences*, **2**, 29-39.

Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical Methods in Diagnostic Medicine.* John Wiley & Sons.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 51–62

# Evaluating Batsman using Survival Analysis

**Parag Shah**[1], **R. D. Chaudhari**[2] **and M. N. Patel**[3]
[1]*Department of Statistics, H L College of Commerce, Gujarat, INDIA*
[2]*Department of Statistics, M G Science Institute, Gujarat, INDIA*
[3]*Department of Statistics, Gujarat University, Gujarat, INDIA*

Received: 06 June 2021; Revised: 17 January 2022; Accepted: 19 March 2022

## Abstract

Batsman has always dominated the cricket arena. Lot of research has been done to measure the performance of a batsman. The performance of a batsman has been usually measured using batting average or strike rate. Some researchers have suggested a constant-hazard model to obtain the probability of a batsman being dismissed on their current score. There are studies that tries to examine the survival ability of batsman using a probabilistic model. We propose generalized exponential distribution as the best fit to the runs scored by a batsman. Survival probabilities and conditional survival probability of a batsman using this distribution gives the more accurate chance of a batsman to survive on crease. We have calculated these survival probabilities and conditional survival probabilities for ICC top 10 batsman against top cricket playing nation. This study can be used by team managements to pick up the team, decide batting order as per the opponent team and match situation. It can also be beneficial to the betting industry as individual batsman score can be predicted using these survival probabilities.

*Key words:* Conditional survival probability exponential distribution; generalized exponential distribution; survival probability; Weibull distribution.

**AMS Subject Classifications:** 62K05, 05B05

## 1. Introduction

Cricket is becoming one of the most popular sports of the today's world. Given the data-rich nature of the sport, numerous studies have used metrics to measure the performance of batsman, bowler, fielder and captain. During the past few years or more lot of work and research papers have been published which measured the performance of the players and their predictions. Many researchers have focussed their study on the most entertaining element and key factor of cricket *i.e.*, batting.

A detail study regarding research directions in cricket was considered by Swartz (2017). The use of stochastic dominance rules demonstrated by Damodaran (2006) to analyse the batting performance of Indian cricketers in ODI cricket. Shah (2017) has defined a new batting and bowling measure. He has defined batting average considering the quality of bowler he is facing and similarly bowling average considering the quality of batsman he is

Corresponding Author: Parag Shah
Email: pbshah@hlcollege.edu

bowling against. Shah and Patel (2018) have ranked captains based on several parameters using Principal Component Analysis. Also, they have included weighted average method to rank captains based on $z$ score of performance of team, individual performance of captain as batsman and bowler.

Elderton and Wood (1945) shows that geometric model can be used for batsman's scores. Bracewell and Ruggiero (2009) suggested 'Ducks $n$ runs'distribution for scores of zero to overcome inability of geometric distribution under inflated number of scores of zero.

As fall of wickets leads to the loss of resources of the batting side, so stability of a batsman on the pitch would help a team to win the match provided evidently, he should have scored runs as quickly as possible. Thus, to know how much time a batsman can survive or how many balls a batsman can face on the cricket pitch while batting might be very useful to arrange the batting order of a team in 20-20 or ODI cricket based on the match situation.

Survival analysis provides the survival ability of an individual where the outcome variable is the time until the occurrence of a particular event of interest. The survival time or time to an event of interest can be measured in hours, months, years, *etc.*, in which the objects or subjects are followed over a specified period of time to pinpoint the event of interest occurs. It is widely used in medical, clinical trial, actuarial science, *etc.*, but now a day the application of survival analysis becomes very much useful in sport (especially in cricket).

Kimber and Hansford (1993) demonstrated utility of non-parametric models for estimating hazard of player's. Stevenson and Brewer (2017) proposed Bayesian approach and hierarchical inference for player's hazard. They considered Bayesian survival analysis of batsmen.

For estimating adjusted batting average of player's a product limit estimator is also used. Das (2011) used such a method using generalized geometric distribution. A survival rate criterion is considered by van Staden (2010) for evaluating the performance of batsmen. Saikai and Bhattacharjee (2018) examined survival ability of batsmen in IPL 2012.

So, in this paper, we have examined three distributions namely: exponential, Weibull and generalised exponential for fitting the runs scored by the batsmen and found the generalised exponential distribution as a best fit. Also, the batting average of a batsman is compared with the mean of generalised exponential distribution, which almost comes out to be close to each other.Using generalised exponential distribution, we have obtained survival probabilities of the batsman. This probability will give the chance that a batsman remains on the crease and scores particular runs. This survival probabilities can be used by team managements or captains to decide the batting order as per the match situations and opposition team. We have also computed conditional probabilities for each batsman for surviving for $b$ runs given that he has survived for $a$ runs. This will be useful to make prediction of scores of the batsman and team scores during the live match.

## 2. Material and methods

Data of runs scored by batsmen up to April 2020 was taken from www.espncricinfo.com. Also, the top teams and top batsmen as per International Cricket Council (ICC) ranking for ODI of April 2020 are considered. In this paper, we first took the innings-by-innings scores of ICC top 10 batsmen and fitted various distributions like exponential, Weibull and generalised exponential using R programming language.

## 3. Results and discussion

From Table 1, it can be seen that generalised exponential distribution is the best fit to the runs scored by the batsman as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are least compared to other two exponential and Weibull distribution.

**Table 1: Fitting three distributions on runs scored by top 10 batsmen**

| | Exponential dist. | | Weibull dist. | | Generalised Exponential dist. | |
|---|---|---|---|---|---|---|
| Batsman | AIC | BIC | AIC | BIC | AIC | BIC |
| Kohli | 2035.28 | 2038.756 | 1939.482 | 1946.435 | 1900.618 | 1907.57 |
| Rohit | 1821.798 | 1825.178 | 1693.453 | 1700.212 | 1660.305 | 1667.064 |
| Babar | 621.0408 | 623.3174 | 621.5612 | 626.1145 | 617.4151 | 620.9684 |
| Taylor | 1729.64 | 1733.016 | 1673.154 | 1679.905 | 1648.36 | 1655.111 |
| Du Plessis | 1129.563 | 1132.476 | 1128.508 | 1134.333 | 1122.768 | 1128.594 |
| Warner | 1111.585 | 1114.381 | 1104.288 | 1109.879 | 1100.065 | 1105.656 |
| Williamson | 1265.706 | 1268.675 | 1257.981 | 1263.921 | 1244.698 | 1250.638 |
| Root | 1146.419 | 1149.339 | 1135.751 | 1141.591 | 1123.992 | 1129.832 |
| Finch | 1123.977 | 1126.78 | 1034.221 | 1039.829 | 998.7167 | 1004.325 |
| De Kock | 1105.748 | 1108.543 | 1093.063 | 1098.654 | 1082.108 | 1087.7 |

The actual batting average and average (mean) as per the generalised exponential distribution were compared for all top batsmen. From Table 2, it can be seen that both averages were very close in majority of the batsmen. This again validates the suitability of the generalised exponential distribution.

The probability density function and cumulative distribution function of generalised exponential distribution is given by

$$f(x;\alpha,\lambda) = \begin{cases} \alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}, & x>0, \alpha>0, \lambda>0; \\ 0, & \text{Otherwise.} \end{cases} \tag{1}$$

and

$$F(x;\alpha,\lambda) = (1-e^{-\lambda x})^{\alpha} \tag{2}$$

Its mean and variance are given as:

$$\mu = \frac{\psi(\alpha+1)-\psi(1)}{\lambda} \tag{3}$$

**Table 2: Comparison of actual batting average with the mean and sd of runs using generalised exponential distribution**

| Batsman | Country | Actual Bating Avg. | Mean | sd |
|---|---|---|---|---|
| Kohli | India | 59.33 | 65.54 | 95.42 |
| Rohit | India | 49.27 | 53.73 | 81.80 |
| Babar | Pakistan | 54.17 | 55.50 | 64.68 |
| Taylor | New Zealand | 48.44 | 53.15 | 73.16 |
| Du Plessis | South Africa | 47.47 | 48.66 | 56.05 |
| Warner | Australia | 45.8 | 46.14 | 55.16 |
| Williamson | New Zealand | 47.48 | 48.44 | 59.78 |
| Root | England | 51.05 | 53.42 | 66.88 |
| Finch | Australia | 41.02 | 40.68 | 64.72 |
| De Kock | South Africa | 44.65 | 44.96 | 56.98 |

$$\sigma^2 = \frac{\psi'(1) - \phi'(\alpha + 1)}{\lambda^2} \tag{4}$$

where,

$$\psi(\cdot) = \text{diagamma function } = \frac{d \log \Gamma(\cdot)}{d(\cdot)}$$

and

$$\psi'(\cdot) = \text{trigamma function which is derivative of diagamma function.}$$

Survival analysis is defined as a set of methods for analysing data where the outcome variable is the time until the occurrence of a particular event of interest. The event could be death due to cancer, occurrence of a disease, relief from a severe back pain, *etc.*, Let us take an example to explain mathematical definition of survival function. Suppose the actual survival time of an individual (say) $t$ which can be regarded as the value of a variable $T$ (*i.e.*, associated with the survival time). It can take any non-negative value. The different values that $T$ can take have a probability distribution, so the variable $T$ can be considered as a random variable. Now for the random variable $T$, the probability distribution function of $T$ can be defined as $F(t)$ and it is given by

$$F(t) = P(T < t) = \int_0^t f(x)\, dx$$

which represents the probability that the survival time is less than some value $t$. Now the survival function is defined as the probability that the survival time is greater than or equal to $t$. Usually, it is denoted by $S(t)$ and given by

$$S(t) = P(T \geq t) = 1 - \int_0^t f(x)\, dx$$

The survival function for generalised exponential distribution is given by

$$S(t) = 1 - (1 - e^{-\lambda x})^{\alpha}, t > 0. \tag{5}$$

Therefore, the survival function can be used to represent the probability that an individual survives from the time origin to sometime beyond $t$. The survival time or time to an event of interest can be measured in days, weeks, years, *etc.*, in which the objects or subjects are followed over a specified period of time to pinpoint the event of interest occurs. Though it's uses in medical, clinical trial, actuarial science, *etc.*, are hefty, but still the application of survival analysis in sport (especially in cricket) is limited.

We have calculated the survival probability of ICC top 10 batsmen using equation (5) and presented in Table 3 and its graph is shown in Fig. 1. We can say that Babar, Virat, Rohit and Taylor have the high survival probability of getting good runs. This is also depicted by their ICC rankings. Survival probabilities of Warner, du Plessis, de Kock suggest that the probability decreases compared to other batsmen, which suggest that they get a start but are unable to convert into big score. And Finch has the lowest probability which says that he gets out early. Virat and Babar have highest probability of getting half century. Similarly, Virat has the highest probability of scoring century among all other top batsmen. This also suggests that Virat converts a good start into half century and century, which is confirmed by the number of centuries he has scored.

**Table 3: Survival probabilities of ICC top 10 batsmen**

| | **Runs** | | | | |
|---|---|---|---|---|---|
| **Batsman** | **10** | **30** | **50** | **80** | **100** |
| **Virat** | 0.6694 | 0.4822 | 0.3738 | 0.2678 | 0.218 |
| **Rohit** | 0.6169 | 0.4262 | 0.3205 | 0.2209 | 0.1756 |
| **Babar** | 0.7616 | 0.527 | 0.3796 | 0.2384 | 0.1764 |
| **Taylor** | 0.6686 | 0.4582 | 0.3373 | 0.2234 | 0.1725 |
| **du Plessis** | 0.7458 | 0.4945 | 0.3411 | 0.2007 | 0.1422 |
| **Warner** | 0.718 | 0.4681 | 0.3206 | 0.1878 | 0.1328 |
| **Williamson** | 0.7101 | 0.4714 | 0.3306 | 0.2014 | 0.1465 |
| **Root** | 0.7188 | 0.4918 | 0.3557 | 0.227 | 0.1704 |
| **Finch** | 0.5524 | 0.3578 | 0.256 | 0.1649 | 0.1256 |
| **de Kock** | 0.6838 | 0.4438 | 0.3065 | 0.1824 | 0.1318 |

Similarly, survival probability against world's top teams as per ICC ranking April 2020 is calculated in Table 4. We can see that Virat's survival chances at initial score and after that are highest against South Africa and least against England. Rohit has highest initial survival chances against England and lowest against South Africa. While the chance of getting half-century or century is highest against Australia and England and lowest against New Zealand. Babar has the lowest chance of scoring big score against India. This way we

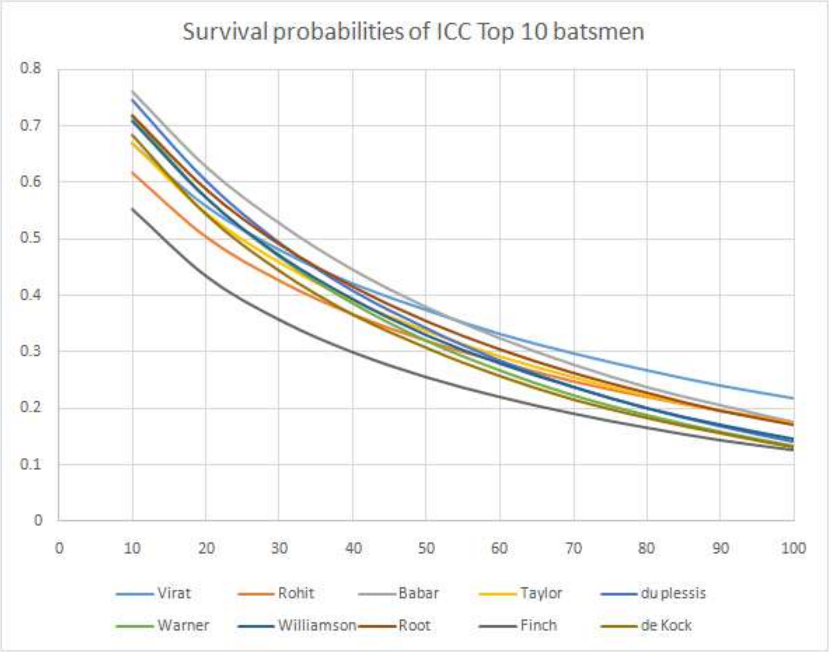


**Figure 1: Graph of survival probabilities of ICC top 10 batsmen**

can conclude about individual batsman scoring probabilities against specific teams. We can identify that against which team batsman gets out early or scores big after the start.

This survival probabilities of each batsman can be useful for team selection against a particular team. It can also be useful to predict the individual batsman score and team score in on-going match. Batting order can be decided by captains considering the match situation and survival probabilities of batsmen.

**Table 4: Survival probabilities of ICC top 10 against ICC top 7 teams**

| Batsman | Runs | India | New Zealand | Australia | England | South Africa | Sri Lanka | Pakistan |
|---|---|---|---|---|---|---|---|---|
| Virat | 10 | NA | 0.6922 | 0.6747 | 0.5659 | 0.71 | 0.6923 | 0.5805 |
| | 30 | NA | 0.4948 | 0.4672 | 0.3872 | 0.5238 | 0.5019 | 0.4067 |
| | 50 | NA | 0.3784 | 0.3471 | 0.2914 | 0.412 | 0.3894 | 0.3123 |
| | 80 | NA | 0.2647 | 0.233 | 0.2023 | 0.3 | 0.2785 | 0.223 |
| | 100 | NA | 0.2118 | 0.1814 | 0.1619 | 0.2466 | 0.2263 | 0.1818 |
| Rohit | 10 | NA | 0.6955 | 0.6661 | 0.7502 | 0.5205 | 0.5369 | 0.6448 |
| | 30 | NA | 0.3932 | 0.472 | 0.5134 | 0.3227 | 0.3851 | 0.4082 |
| | 50 | NA | 0.2302 | 0.36 | 0.3669 | 0.2225 | 0.3044 | 0.3274 |
| | 80 | NA | 0.1053 | 0.2519 | 0.2282 | 0.1361 | 0.2275 | 0.2198 |
| | 100 | NA | 0.0629 | 0.2019 | 0.1679 | 0.1001 | 0.1914 | 0.1715 |
| Babar | 10 | 0.85 | 0.4554 | 0.9999 | 0.8866 | 0.9798 | 0.917 | NA |

Continued on next page

**Table 4 – continued from previous page**

| Batsman | Runs | India | New Zealand | Australia | England | South Africa | Sri Lanka | Pakistan |
|---|---|---|---|---|---|---|---|---|
| | 30 | 0.4254 | 0.3163 | 0.8029 | 0.5676 | 0.7803 | 0.6785 | NA |
| | 50 | 0.1839 | 0.2454 | 0.4339 | 0.3326 | 0.4531 | 0.4733 | NA |
| | 80 | 0.0484 | 0.1798 | 0.1212 | 0.1406 | 0.0999 | 0.2635 | NA |
| | 100 | 0.0195 | 0.1496 | 0.0478 | 0.0779 | 0.0218 | 0.1758 | NA |
| Taylor | 10 | 0.7133 | NA | 0.7744 | 0.6994 | 0.6835 | 0.4393 | 0.6253 |
| | 30 | 0.4786 | NA | 0.4516 | 0.4744 | 0.4063 | 0.2903 | 0.4881 |
| | 50 | 0.3393 | NA | 0.2613 | 0.3422 | 0.2539 | 0.2158 | 0.4115 |
| | 80 | 0.2101 | NA | 0.1145 | 0.2184 | 0.1294 | 0.1489 | 0.3346 |
| | 100 | 0.1545 | NA | 0.0659 | 0.1641 | 0.0834 | 0.1192 | 0.2967 |
| du Plessis | 10 | 0.8714 | 0.8572 | 0.8961 | 0.5054 | NA | 0.8521 | 0.918 |
| | 30 | 0.62 | 0.5255 | 0.6295 | 0.2868 | NA | 0.5838 | 0.501 |
| | 50 | 0.4309 | 0.3007 | 0.418 | 0.1813 | NA | 0.3922 | 0.2152 |
| | 80 | 0.2456 | 0.1245 | 0.2171 | 0.0973 | NA | 0.213 | 0.0531 |
| | 100 | 0.1678 | 0.0683 | 0.1385 | 0.0656 | NA | 0.1412 | 0.0204 |
| Warner | 10 | 0.7846 | 0.8116 | NA | 0.762 | 0.6217 | 0.7137 | 0.8632 |
| | 30 | 0.5233 | 0.5315 | NA | 0.4007 | 0.4023 | 0.4211 | 0.5911 |
| | 50 | 0.3564 | 0.3474 | NA | 0.2048 | 0.2825 | 0.257 | 0.3928 |
| | 80 | 0.2032 | 0.1834 | NA | 0.0736 | 0.1753 | 0.1252 | 0.2082 |
| | 100 | 0.1403 | 0.1197 | NA | 0.0371 | 0.1297 | 0.078 | 0.1355 |
| Williamson | 10 | 0.6784 | NA | 0.6155 | 0.7133 | 0.8075 | 0.608 | 0.8682 |
| | 30 | 0.4152 | NA | 0.395 | 0.4931 | 0.5323 | 0.3736 | 0.6086 |
| | 50 | 0.269 | NA | 0.2756 | 0.3614 | 0.352 | 0.2494 | 0.4157 |
| | 80 | 0.1453 | NA | 0.1693 | 0.236 | 0.1896 | 0.1433 | 0.2303 |
| | 100 | 0.0975 | NA | 0.1246 | 0.1799 | 0.1257 | 0.1007 | 0.1544 |
| Root | 10 | 0.7811 | 0.89 | 0.5191 | NA | 0.9096 | 0.7101 | 0.6775 |
| | 30 | 0.5495 | 0.6197 | 0.319 | NA | 0.6223 | 0.5051 | 0.4559 |
| | 50 | 0.3999 | 0.4093 | 0.2183 | NA | 0.3891 | 0.3823 | 0.3283 |
| | 80 | 0.2541 | 0.2116 | 0.1319 | NA | 0.1805 | 0.2624 | 0.2098 |
| | 100 | 0.1893 | 0.1346 | 0.0964 | NA | 0.1061 | 0.2072 | 0.158 |
| Finch | 10 | 0.6064 | 0.3253 | NA | 0.4719 | 0.7489 | 0.783 | 0.5939 |
| | 30 | 0.3993 | 0.1675 | NA | 0.3287 | 0.4254 | 0.4909 | 0.434 |
| | 50 | 0.2855 | 0.1003 | NA | 0.2551 | 0.2425 | 0.3097 | 0.346 |
| | 80 | 0.1844 | 0.0506 | NA | 0.1868 | 0.1046 | 0.1558 | 0.2606 |
| | 100 | 0.14 | 0.0329 | NA | 0.1553 | 0.0597 | 0.0987 | 0.22 |
| de Kock | 10 | 0.9299 | 0.5633 | 0.5832 | 0.8564 | NA | 0.8404 | 0.467 |

Continued on next page

**Table 4 – continued from previous page**

| Batsman | Runs | India | New Zealand | Australia | England | South Africa | Sri Lanka | Pakistan |
|---|---|---|---|---|---|---|---|---|
| | 30 | 0.6939 | 0.346 | 0.3389 | 0.6303 | NA | 0.5579 | 0.2943 |
| | 50 | 0.4805 | 0.2338 | 0.2147 | 0.4641 | NA | 0.3628 | 0.2084 |
| | 80 | 0.2611 | 0.1379 | 0.1141 | 0.2936 | NA | 0.1876 | 0.1335 |
| | 100 | 0.1706 | 0.0989 | 0.0762 | 0.2164 | NA | 0.1204 | 0.1017 |

## 4. Conditional survival analysis

In this section we have considered the conditional survival probability of batsman.

Suppose that the batsman has survived at score $a$ then the probability of surviving at score $b$, $b > a$ is called conditional survival probability.

Mathematically it is defined as

$$S(b|a) = P(X > b|X > a) = \frac{P(X > b)}{P(X > a)}.$$

In case of generalised exponential distribution, it is given by

$$S(b|a) = \frac{1 - (1 - e^{-\lambda b})^{\alpha}}{1 - (1 - e^{-\lambda a})^{\alpha}} \tag{6}$$

Conditional survival probabilities using generalised exponential distribution are calculated using equation (6) for ICC top 10 batsmen as per ICC ranking April 2020 and given in Table 5. This probability describes the ability of batsman to survive for some additional scores during the on-going play. We can observe that when Virat scores 10 runs, the probability of scoring a century is highest among all the players. This shows that when Virat gets his eye set, the chance of converting it into big score is highest. The probability of scoring a century once the batsmen has scored 50 runs is quite high for Rohit, followed closely by Virat, Taylor, Finch and Root. This is also reflected in the number of centuries these batsmen have scored. Once they spend some time on crease, they score high in the match. Low probabilities for a batsman also reflect that the batsman throws away his wicket after getting his eye set. This can be a good point for the coaches to guide the player to play long. These probabilities will be more useful for predicting individual scores and team scores. This is very important probabilities for the team and also for the betting industry.

**Table 5: Conditional probabilities $P(\text{score} > b/\text{score} > a)$ of ICC top 10 batsmen**

| Batsman | a | b | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 80 | 100 |
| Virat | 10 | 1 | 0.7204 | 0.5584 | 0.4 | 0.3257 |
| | 30 | | 1 | 0.743 | 0.4987 | 0.3884 |
| | 50 | | | 1 | 0.6711 | 0.5227 |

Continued on next page

**Table 5 – continued from previous page**

| Batsman | a | b | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 80 | 100 |
| | 80 | | | | 1 | 0.7789 |
| Rohit | 10 | 1 | 0.6908 | 0.5957 | 0.358 | 0.2846 |
| | 30 | | 1 | 0.752 | 0.5182 | 0.412 |
| | 50 | | | 1 | 0.6891 | 0.5479 |
| | 80 | | | | 1 | 0.795 |
| Babar | 10 | 1 | 0.692 | 0.4984 | 0.313 | 0.2317 |
| | 30 | | 1 | 0.7202 | 0.4523 | 0.3348 |
| | 50 | | | 1 | 0.628 | 0.4648 |
| | 80 | | | | 1 | 0.7401 |
| Taylor | 10 | 1 | 0.6827 | 0.5045 | 0.3342 | 0.258 |
| | 30 | | 1 | 0.7363 | 0.4877 | 0.3766 |
| | 50 | | | 1 | 0.6624 | 0.5114 |
| | 80 | | | | 1 | 0.7722 |
| du Plessis | 10 | 1 | 0.663 | 0.4574 | 0.2691 | 0.1906 |
| | 30 | | 1 | 0.6899 | 0.4059 | 0.2875 |
| | 50 | | | 1 | 0.5884 | 0.4168 |
| | 80 | | | | 1 | 0.7083 |
| Warner | 10 | 1 | 0.6519 | 0.4466 | 0.2615 | 0.185 |
| | 30 | | 1 | 0.685 | 0.4011 | 0.2837 |
| | 50 | | | 1 | 0.5856 | 0.4142 |
| | 80 | | | | 1 | 0.7074 |
| Williamson | 10 | 1 | 0.6638 | 0.4655 | 0.2836 | 0.2063 |
| | 30 | | 1 | 0.7014 | 0.4272 | 0.3107 |
| | 50 | | | 1 | 0.6091 | 0.443 |
| | 80 | | | | 1 | 0.7273 |
| Root | 10 | 1 | 0.6842 | 0.4948 | 0.3158 | 0.2371 |
| | 30 | | 1 | 0.7232 | 0.4617 | 0.3466 |
| | 50 | | | 1 | 0.6383 | 0.4792 |
| | 80 | | | | 1 | 0.7507 |

Continued on next page

**Table 5 – continued from previous page**

| Batsman | a | b | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 80 | 100 |
| Finch | 10 | 1 | 0.6478 | 0.4633 | 0.2985 | 0.2273 |
| | 30 | | 1 | 0.7153 | 0.4607 | 0.3509 |
| | 50 | | | 1 | 0.6441 | 0.4906 |
| | 80 | | | | 1 | 0.7616 |
| de Kock | 10 | 1 | 0.649 | 0.4483 | 0.268 | 0.1927 |
| | 30 | | 1 | 0.6907 | 0.4129 | 0.2969 |
| | 50 | | | 1 | 0.5978 | 0.4299 |
| | 80 | | | | 1 | 0.7192 |

We have also calculated conditional probabilities of ICC top 10 batsmen against ICC top teams. These probabilities will evaluate the performance of a batsman against a particular team. In Table 6, we have presented conditional probabilities of 5 batsmen against top 5 teams for particular runs only.

From Table 6, we can see that Virat has the highest probability of scoring a half-century or a century given that he scores 10 runs *i.e.* gets a start against South Africa followed by New Zealand, England and Australia. Also, Virat has capability of converting half-century into huge score like century is highest among all other batsmen. Rohit loves to score big against Australia and then England compared to other countries once he gets the start, which is reflected in the conditional probabilities. Babar loves to play against New Zealand compared to other countries with more than 50% chance of making half-century or even century from a start he gets. But he has the lowest probability of scoring against India. So, even if he gets a start, he is not able to convert into big scores against India. Taylor's probabilities suggest that once he gets his eye on the ball, he loves to score big against England and India. Du Plessis has the highest scoring probability against India compared to other teams.

## 5. Conclusion

Survival probabilities and conditional probabilities using generalised exponential distribution gives more accurate chance of survival compared to exponential and Weibull distributions. These probabilities can be used as a new measure for evaluating batsman as it gives the ability of a batsman to survive on crease. This is the measure that evaluates batsman during the live match and at every run he scores. Conditional survival probabilities can be advantageous to the team managements to decide the batting order or change it during match depending on the match situations and the opponent team. From our study we conclude that among all top batsmen, Virat and Rohit have higher survival rate and even potential of making big scores like half-century and century.

**Table 6: Conditional probabilities $P(\text{score} > b|\text{score} > a)$ of top 5 batsmen against top 5 teams**

| a | b | Country | Batsman | | | | |
|---|---|---|---|---|---|---|---|
| | | | Virat | Rohit | Babar | Taylor | du Plessis |
| 10 | 50 | India | NA | NA | 0.2163 | 0.4756 | 0.4945 |
| | | New Zealand | 0.5466 | 0.3309 | 0.5389 | NA | 0.3508 |
| | | Australia | 0.5145 | 0.5405 | 0.4352 | 0.3374 | 0.4665 |
| | | England | 0.515 | 0.4891 | 0.3751 | 0.4894 | 0.3588 |
| | | South Africa | 0.5802 | 0.3582 | 0.4624 | 0.3714 | NA |
| 10 | 100 | India | NA | NA | 0.023 | 0.2166 | 0.1926 |
| | | New Zealand | 0.306 | 0.0904 | 0.3285 | NA | 0.0797 |
| | | Australia | 0.2689 | 0.3031 | 0.0479 | 0.0151 | 0.1545 |
| | | England | 0.2861 | 0.2239 | 0.0879 | 0.2347 | 0.1297 |
| | | South Africa | 0.3473 | 0.1613 | 0.0223 | 0.122 | NA |
| 50 | 100 | India | NA | NA | 0.1062 | 0.4554 | 0.3895 |
| | | New Zealand | 0.5598 | 0.2733 | 0.6097 | NA | 0.2272 |
| | | Australia | 0.5227 | 0.5607 | 0.1101 | 0.2523 | 0.3312 |
| | | England | 0.5556 | 0.4577 | 0.2342 | 0.4796 | 0.3616 |
| | | South Africa | 0.5985 | 0.4504 | 0.0482 | 0.3284 | NA |

## 6. Future scope

This study can be applied for test match cricket and T20 cricket. It can be also be used in football, hockey to calculate survival probabilities and conditional survival probabilities of goals by a team or a goal-keeper. Also, various other multivariate techniques like logistic regression, principal component analysis can be applied to the data.

## Acknowledgements

The authors express their gratefulness to the honorable reviewers and the chief editor for making useful suggestions that helped in improving the content and the presentation of the content.

## References

Bracewell, P. J. and Ruggiero, K. (2009). A parametric control chart for monitoring individual batting performances in cricket. *Journal of Quantitative Analysis in Sports*, **5**.

Damodaran, U. (2006): Stochastic dominance and analysis of ODI batting performance: the Indian cricket team 1989-2005. *Journal of Sports Science and Medicine*, **5**, 503–508.

Das, S. (2011). On generalized geometric distributions: Application to modelling scores in cricket and improved estimation of batting average in light of not out innings. *Bangalore: Working Paper Series, Indian Institute of Management Bangalore.*

Elderton, W. and Wood, G. H. (1945). Cricket scores and geometrical progression. *Journal of the Royal Statistical Society*, **108**, 12–40.

Kimber, A. C. and Hansford, A. R. (1993). A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **156**, 443–455.

Saikai, H. and Bhattacharjee, D. (2018). Survival ability of Indian and overseas batsmen on the cricket pitch in Indian premier league. *MOJ Sports Medicine*, **2(4)**, 113-116.

Shah, P. (2017). New performance measure in Cricket. *IOSR Journal of Sports and Physical Education*, **4**, 28-30.

Shah, Parag and Patel, M. N. (2018). Ranking the cricket captains using principal component analysis. *International Journal of Physiology, Nutrition and Physical Education*, **3**, 477–483

Stevenson, O. G. and Brewer, B. J. (2017). Bayesian survival analysis of batsmen in test cricket. *Journal of Quantitative Analysis in Sports*, **13(1)**, 25-36.

Swartz, T. B. (2017). Research directions in cricket. *Handbook of Statistical Methods an analysis in sports.* Chapman and Hall.

van Staden, P. J., Meiring, A. T., Steyn, J. A., *et. al.* (2010). Meaning batting averages in cricket. *South African Statistical Journal Proceedings: Peer-reviewed Proceedings of the 52nd Annual Conference of the South African Statistical Association for 2010 (SASA 2010): Congress.*

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 63–79

# Discrete Harris Extended Weibull Distribution and Applications

**Sophia P. Thomas**[1], **Lishamol Tomy**[2] **and K. K. Jose** [3]
[1]*Department of Statistics, St.Thomas College, Pala, Kerala, India*
[2]*Department of Statistics, Deva Matha College, Kuravilangad, Kerala, India*
[3]*School of Mathematics, Statistics and Data Analytics*
*Mahatma Gandhi University, Kottayam, Kerala, India*

Received: 03 October 2021; Revised: 24 December 2021; Accepted: 19 March 2022

## Abstract

In this paper, we introduce a new family called Discrete Harris Extended (DHE) family of distributions and study its properties. It is shown that the new family is a generalization of discrete Marshall-Olkin family of distributions. In particular, we study the discrete version of Harris Extended Weibull distribution in detail. We give some selected special distributions from DHE family. We derive some basic distributional properties such as probability generating function, moments, hazard rate and quantiles of the DHEW distribution. Estimation of the parameters is done using maximum likelihood method and a simulation study is conducted to verify the performance. By using the method of maximum likelihood estimation we obtain the estimates of the proposed model parameters with respect to two discrete data sets.

*Key words:* Discrete Harris Extended Weibull distribution; Infinite divisibility; Marshall-Olkin family of distributions; Maximum likelihood.

**AMS Subject Classifications:** 60E05, 62E15, 62F10

## 1. Introduction

In the literature, there are several methods to obtain a discrete distribution from a continuous distribution: the discretization method based on the survival function (Nakagawa and Osaki, 1975), the discretization method based on an infinite series (Good, 1953; Kulasekera and Tonkyn, 1992; Kemp, 1997), the discretization method based on the hazard function (Stein, 1984), the compound two-phase method (Chakraborty, 2015), the discretization method based on reverse hazard function (Ghosh *et al.*, 2013), among many others.

The traditional discrete distributions (geometric, Poisson, etc.) have limited applicability as models for reliability, failure times, counts, etc. This has led to the development of new discrete distributions based on popular continuous models for reliability, failure times, etc. Of these, the most popular is the discrete Weibull distribution which was introduced by Nakagawa and Osaki (1975) and studied by Stein and Dattero (1984), and

Corresponding Author: Sophia P. Thomas
Email: sophiathomas92@gmail.com

Khan *et al.*(1989). Gómez-Déniz (2010) developed a new generalization of the geometric distribution using Marshall-Olkin scheme. Discrete modified Weibull distribution proposed by Nooghabi *et al.* (2011) is the discrete analogue of the modified Weibull distribution in Lai *et al.* (2003). Nekoukhou and Bidram (2015) proposed exponentiated discrete Weibull distribution as a discrete analog of the exponentiated Weibull distribution of Mudholkar and Srivastava (1993). Sandhya and Prasanth (2012, 2013) have considered generalisations of geometric and discrete uniform distributions invoking the approach of Marshal and Olkin (1997), while Sandhya and Prasanth (2016) has developed another generalisation of the discrete uniform distribution by adding two parameters to it, generalizing the Marshal-Olkin scheme itself. Recently, Jayakumar and Sankaran (2018) have introduced a new discrete family of distributions using truncated discrete Mittag-Leffler distribution and studied its properties.

In this paper, we identify some members of the DHE family of distributions using the discretization method of Nakagawa and Osaki (1975). Our work mainly focuses on the DHEW distribution. This distribution is generated by discretizing the Harris extended Weibull (HEW) distribution of Batsidis and Lemonte (2014) with survival function (sf)

$$\bar{G}(x) = \left(\frac{\lambda e^{-k(\eta x)^{\beta}}}{1-\bar{\lambda}e^{-k(\eta x)^{\beta}}}\right)^{1/k} \tag{1}$$

The HEW probabiity density function (pdf) is given by

$$g(x) = \frac{\lambda^{1/k}\beta\eta^{\beta}(x)^{\beta-1}e^{-(\eta x)^{\beta}}}{[1-\bar{\lambda}e^{-k(\eta x)^{\beta}}]^{1+\frac{1}{k}}} \quad ; \quad x>0, \quad (\lambda,\ \eta,\ k,\ \beta)>0, \quad \bar{\lambda}=1-\lambda. \tag{2}$$

Here, $\lambda > 0$, $k > 0$, and $\beta > 0$ are shape parameters, $\eta > 0$ is the scale parameter.

The HEW distribution has many applications specially in quality control and reliability; see Jose *et al.* (2018). This distribution is a suitable competitor for gamma and Weibull distributions. But, sometimes, it is impossible or inconvenient to measure the life length of a device on a continuous scale. In practice, we come across situations where lifetimes are recorded on a discrete scale. For example, on/off switching devices, bulb of photocopier machine, to and fro motion of spring devices, *etc.*, are some obvious such situations. In the last two decades, standard discrete distributions like geometric and negative binomial have been employed to model lifetime data. However, there is a need to find more flexible discrete distributions to fit various types of data.

The rest of the paper is organized as follows. Discretization of continuous family of distributions is discussed in Section 2. In Section 3, we introduce the DHE family of distributions and study its properties. In Section 4, it is shown that the DHE family of distributions is a rich class and identify some members of this family. Section 5 is devoted to the study of various properties of the DHEW distribution. In Section 6, we discuss the method of maximum likelihood estimation of parameters of the distribution and a simulation study is conducted to verify the performance. Two real data sets are analyzed to illustrate the suitability of the proposed model and the results are presented in Section 7. Concluding remarks are given in the last section.

## 2. Discretization of continuous family of distributions

The general approach of discretizing a continuous variable is to introduce the greatest integer function of $X$ namely, $[X]$ (the greatest integer less than or equal to X till it reaches the integer), in order to introduce grouping on a time axis.

Let the continuous failure time $X$ has the sf, $\bar{Q}(x) = P(X > x)$ and $Y = [X]$; be the discrete random variable obtained by grouping the continuous failure time into unit intervals, then by Roy (2003) the probability mass function (pmf) of Y can be written as

$$\begin{aligned} P(Y = y) = P(y \leq X < y + 1) &= P(X > y) - P(X > y + 1) \\ &= \bar{Q}_x(y) - \bar{Q}_x(y + 1), \qquad y = 0, 1, 2, ... \end{aligned} \tag{3}$$

where $\bar{Q}_x(y) = P(X > y)$.

Using (3) many researchers have developed discrete distributions corresponding to existing continuous distributions. For more details refer Nakagawa and Osaki (1975), Krishna and Pundir (2009), Chakraborty and Chakravarty (2012), Seethalekshmi *et al.* (2016), Gillariose *et al.* (2021).

## 3. Discrete Harris extended family of distributions

Let $F(x)$ be the baseline cumulative distribution function (cdf) of a random variable $X$ and let $\overline{F}(x)$ be the survival function (sf) of a distribution. Then the Harris family has the survival probabilities

$$\bar{Q}(x) = \left[ \frac{\lambda \bar{F}(x)^k}{1 - \bar{\lambda} \bar{F}(x)^k} \right]^{1/k} \tag{4}$$

Now the probability mass function (pmf) of the new family is

$$\begin{aligned} p_Y(x) &= \bar{Q}(x) - \bar{Q}(x+1) \\ &= \lambda^{1/k} \left\{ \frac{\bar{F}(x)}{[1 - \bar{\lambda}\bar{F}(x)^k]^{1/k}} - \frac{\bar{F}(x+1)}{[1 - \bar{\lambda}\bar{F}(x+1)^k]^{1/k}} \right\}, \quad x = 0, 1, 2, ... \end{aligned} \tag{5}$$

where, $\lambda, k > 0$, $\bar{\lambda} = 1 - \lambda$. We denote this family of distribution by DHE$(\lambda, k)$ family. Note that , when $k = 1$, the distribution with pmf (5) reduces to discrete Marshall-Olkin distribution discussed in Supanekar and Shirke (2015). Let $R(x)$ be the hazard rate function (hrf) of DHE family of the discrete random variable X, then

$$\begin{aligned} R(x) &= \frac{p_Y(x)}{\bar{Q}(x)} \\ &= 1 - \frac{\bar{F}(x)[1 - \bar{\lambda}\bar{F}(x+1)^k]^{1/k}}{\bar{F}(x)[1 - \bar{\lambda}\bar{F}(x+1)^k]^{1/k}} \end{aligned} \tag{6}$$

### 3.1. Probability generating function, moments and quantiles

The probability generating function (pgf) of (5) is given by

$$P_Y(s) = 1 + \lambda^{1/k}(s-1)\sum_{x=1}^{\infty} s^{x-1}\frac{\bar{F}(x)}{\left[1-\bar{\lambda}\bar{F}(x)^k\right]^{1/k}} \tag{7}$$

Mean and Variance of the random variable X is given by

$$E(X) = \lambda^{1/k}\sum_{x=1}^{\infty}\frac{\bar{F}(x)}{\left[1-\bar{\lambda}\bar{F}(x)^k\right]^{1/k}} \tag{8}$$

$$V(X) = \lambda^{1/k}\sum_{x=1}^{\infty}(2x-1)\frac{\bar{F}(x)}{\left[1-\bar{\lambda}\bar{F}(x)^k\right]^{1/k}} - \left\{\lambda^{1/k}\sum_{x=1}^{\infty}\frac{\bar{F}(x)}{\left[1-\bar{\lambda}\bar{F}(x)^k\right]^{1/k}}\right\}^2 \tag{9}$$

Quantiles $q_m$ and Median of DHE family are

$$q_m = \left[F^{-1}\left(1-(1-m)\left(\lambda+\bar{\lambda}(1-m)^k\right)^{-1/k}\right)-1\right] \tag{10}$$

Median is given by

$$Median = \left[F^{-1}\left(1-\left(2^k\lambda+\bar{\lambda}\right)^{-1/k}\right)-1\right] \tag{11}$$

where[.] denote the integer part.

## 4. Some members of DHE family of distributions

In this section, we give some selected special distributions from DHE family. The selected models are DHE exponential, DHE Uniform, DHE Fréchet, DHE Burr type XII, DHE Lomax and DHE Lindley.

### 4.1. DHE exponential(DHEE) distribution

Consider the sf of exponential distribution with parameter $\theta$ is given by $\bar{F}(x) = e^{-\theta x}$. Let $p = e^{-\theta}$, $0 < p < 1$. Then the probability mass function (pmf), survival function (sf), hazard rate function (hrf) of the DHEE distribution using equation (5) are respectively given by

$$p_x = \frac{\lambda^{1/k}p^x}{\left[1-\bar{\lambda}p^{kx}\right]^{1/k}} - \frac{\lambda^{1/k}p^{x+1}}{\left[1-\bar{\lambda}p^{x+1}\right]^{1/k}}, \quad x = 0,1,2,...$$

$$\bar{Q}(x) = \frac{\lambda^{1/k}p^x}{[1-\bar{\lambda}p^{kx}]^{1/k}}$$

$$R(x) = 1 - \frac{[1-\bar{\lambda}p^{kx}]^{1/k}}{[1-\bar{\lambda}p^{k(x+1)}]^{1/k}}\ p$$

For $k = 1$, the distribution reduces to generalized geometric distribution obtained by discretizing the generalized exponential distribution of Marshall-Olkin (1997).

### 4.2. DHE Uniform (DHEU) distribution

Let X $\sim$ U$(0,a)$ follows Uniform distribution with parameter $a$. Then the sf of $X$ is given by $\bar{F}(x) = 1 - \frac{x}{a}$. Then the pmf, sf, hrf of the DHEU distribution using (5) are respectively given by

$$p_x = \frac{\lambda^{1/k}(a-x)}{\left[a^k - \bar{\lambda}(a-x)^k\right]^{1/k}} - \frac{\lambda^{1/k}(a-x-1)}{\left[a^k - \bar{\lambda}(a-x-1)^k\right]^{1/k}}, \quad x = 1, 2, ..., a$$

$$\bar{Q}(x) = \frac{\lambda^{1/k}(a-x)}{\left[a^k - \bar{\lambda}(a-x)^k\right]^{1/k}}$$

$$R(x) = 1 - \frac{(a-x-1)\left[a^k - \bar{\lambda}(a-x)^k\right]^{1/k}}{(a-x)\left[a^k - \bar{\lambda}(a-x-1)^k\right]^{1/k}}$$

This distribution is obtained and studied by Prasanth and Sandhya (2016).

### 4.3. DHE Fr$\acute{e}$chet (DHEF) distribution

Consider the survival function of Fr$\acute{e}$chet distribution with parameter $\alpha$ and $\beta$ is given by $\bar{F}(x) = 1 - e^{-(\frac{\alpha}{x})^\beta}$. Let $p = e^{-\alpha^\beta}$, $0 < p < 1$. Then the pmf, sf, hrf of the DHEF distribution using equation (5) are respectively given by

$$p_x = \frac{\lambda^{1/k}(1 - p^{-(\frac{1}{x})^\beta})}{\left[1 - \bar{\lambda}1 - p^{-(\frac{1}{x})^\beta}\right]^{1/k}} - \frac{\lambda^{1/k}(1 - p^{-(\frac{1}{x+1})^\beta})}{\left[1 - \bar{\lambda}(1 - p^{-(\frac{1}{x+1})^\beta})\right]^{1/k}}, \quad x = 0, 1, 2, ...$$

$$\bar{Q}(x) = \frac{\lambda^{1/k}(1 - p^{-(\frac{1}{x})^\beta})}{\left[1 - \bar{\lambda}1 - p^{-(\frac{1}{x})^\beta}\right]^{1/k}}, \quad x = 0, 1, 2, ...$$

$$R(x) = 1 - \frac{(1 - p^{-(\frac{1}{x+1})^\beta})\left[1 - \bar{\lambda}(1 - p^{-(\frac{1}{x})^\beta})\right]^{1/k}}{(1 - p^{-(\frac{1}{x})^\beta})\left[1 - \bar{\lambda}(1 - p^{-(\frac{1}{x+1})^\beta})\right]^{1/k}}$$

### 4.4. DHE Burr type XII(DHEBXII) and Lomax (DHELX) distributions

Consider the survival function of Burr type III distribution with parameter $c$ and $b$ is given by $\bar{F}(x) = (1 + x^c)^{-b}$. Let $p = e^{-b}$, $0 < p < 1$. Then the pmf, sf, hrf of the DHEBXII distribution using equation (5) are respectively given by

$$p_x = \frac{\lambda^{1/k} p^{log(1+x^c)}}{\left[1 - \bar{\lambda} p^{klog(1+x^c)}\right]^{1/k}} - \frac{\lambda^{1/k} p^{log(1+(x+1)^c)}}{\left[1 - \bar{\lambda} p^{klog(1+(x+1)^c)}\right]^{1/k}}, \quad x = 0, 1, 2, ...$$

$$\bar{Q}(x) = \frac{\lambda^{1/k} p^{log(1+x^c)}}{\left[1 - \bar{\lambda} p^{klog(1+x^c)}\right]^{1/k}}$$

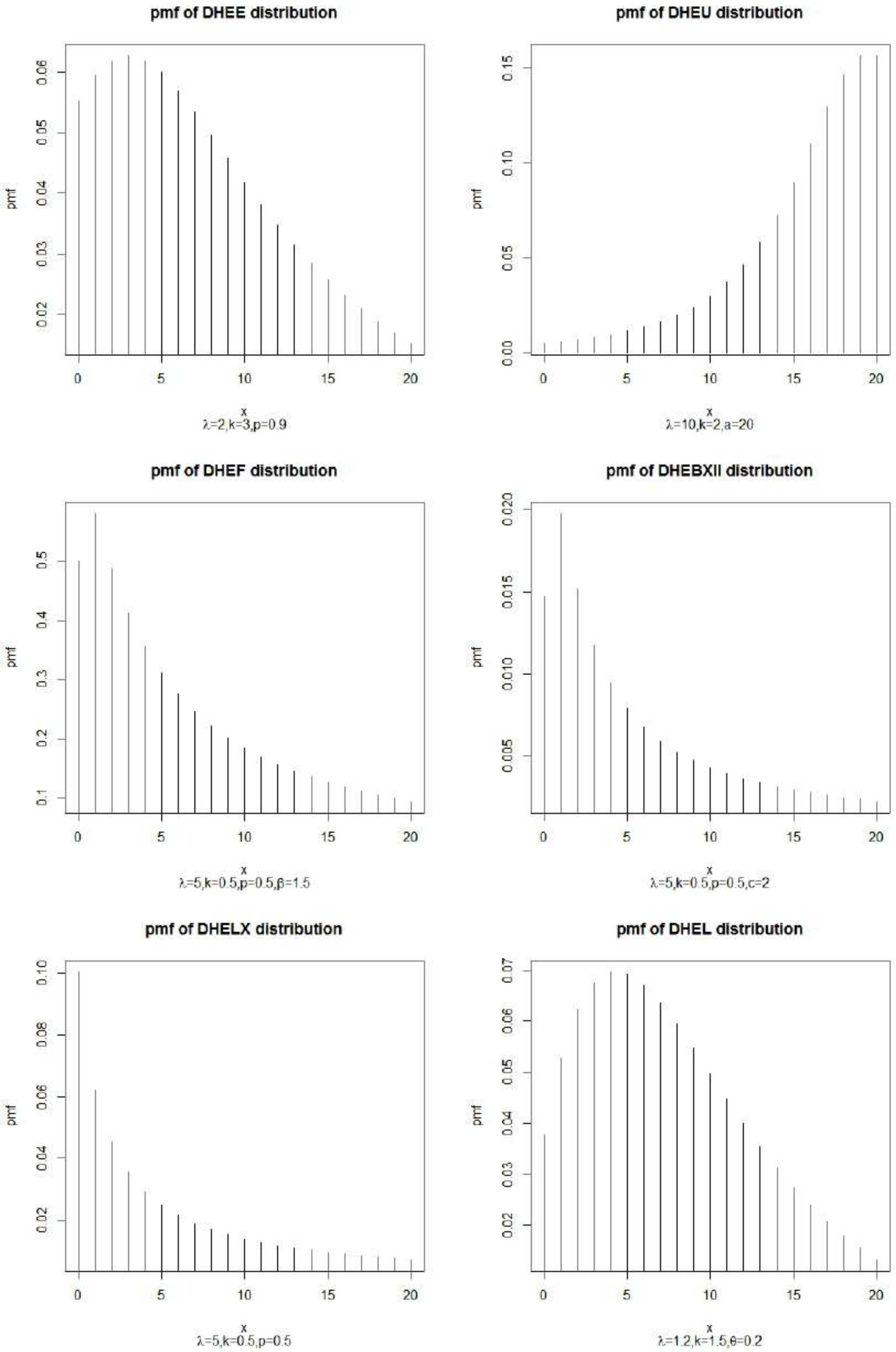


**Figure 1: pmf of discrete HE family of distributions**

$$R(x) = \frac{p^{log(1+(x+1)^c)}\left[1-\bar{\lambda}p^{klog(1+x^c)}\right]^{1/k}}{p^{log(1+x^c)}\left[1-\bar{\lambda}p^{klog(1+(x+1)^c)}\right]^{1/k}}$$

When $c = 1$, the DHEBXII distribution becomes DHELX distribution.

### 4.5. DHE Lindley (DHEL) distribution

Consider the survival function of Lindley distribution with parameter $\theta$ is given by $\bar{F}(x) = \frac{1+\theta+\theta x}{1+\theta}e^{-\theta x}$. Then the pmf, sf, hrf of the DHEL distribution using equation (5) are respectively given by

$$\begin{aligned} p_x &= \frac{\lambda^{1/k}(1+\theta+\theta x)e^{-\theta x}}{[(1+\theta)^k - \bar{\lambda}(1+\theta+\theta x)^k e^{-k\theta x}]^{1/k}} \\ &- \frac{\lambda^{1/k}(1+\theta+\theta(x+1))e^{-\theta(x+1)}}{[(1+\theta)^k - \bar{\lambda}(1+\theta+\theta(x+1))^k e^{-\theta k(x+1)}]^{1/k}}, \quad x = 0, 1, 2, ... \end{aligned}$$

where, $(\lambda, k, \theta) > 0$

$$\bar{Q}(x) = \frac{\lambda^{1/k}(1+\theta+\theta x)e^{-\theta x}}{[(1+\theta)^k - \bar{\lambda}(1+\theta+\theta x)^k e^{-k\theta x}]^{1/k}}$$

$$R(x) = 1 - \frac{[(1+\theta)^k - \bar{\lambda}(1+\theta+\theta x)^k e^{k\theta x}]^{1/k}[1+\theta+\theta(x+1)]}{[(1+\theta)^k - \bar{\lambda}(1+\theta+\theta(x+1))^k e^{-k\theta(x+1)}]^{1/k}[1+\theta+\theta x]}e^{-\theta} \tag{12}$$

We can obtain discrete half-logistic, discrete half-normal and discrete Rayleigh distribution as members of new family of distributions, defined in (5), by substituting respective distribution function. In the next section, we study discrete HEW distribution in detail. Figure 1 displays possible shapes of the selected discrete Harris extended models.

## 5. Discrete Harris extended Weibull(DHEW) distribution

The sf of Weibull distribution with scale parameter $\eta$ and shape parameter $\beta$ is given by

$$\bar{F}(x) = e^{-(\eta x)^{\beta}}; \quad x > 0, \quad \eta > 0, \quad \beta > 0$$

Let $e^{-\eta^{\beta}} = p; 0 < p < 1$. Hence the sf of the resulting discrete distribution is given by

$$\bar{Q}(x) = \frac{\lambda^{1/k}p^{x^{\beta}}}{[1-\bar{\lambda}p^{kx^{\beta}}]^{1/k}}; x = 0, 1, 2, ... \tag{13}$$

$$p_x = \frac{\lambda^{1/k}p^{x^{\beta}}}{\left[1-\bar{\lambda}p^{kx^{\beta}}\right]^{1/k}} - \frac{\lambda^{1/k}p^{(x+1)^{\beta}}}{\left[1-\bar{\lambda}p^{(x+1)^{\beta}}\right]^{1/k}}, \quad x = 0, 1, 2, ...$$

We call the random variable X, with sf (13), as DHEW distribution with parameters $\lambda > 0$,$k > 0$,$0 < p < 1$, $\beta > 0$ and denote it by DHEW $(\lambda, k, p, \beta)$. Many properties of the continuous HEW distribution also hold for DHEW $(\lambda, k, p, \beta)$. Figure 2 displays possible

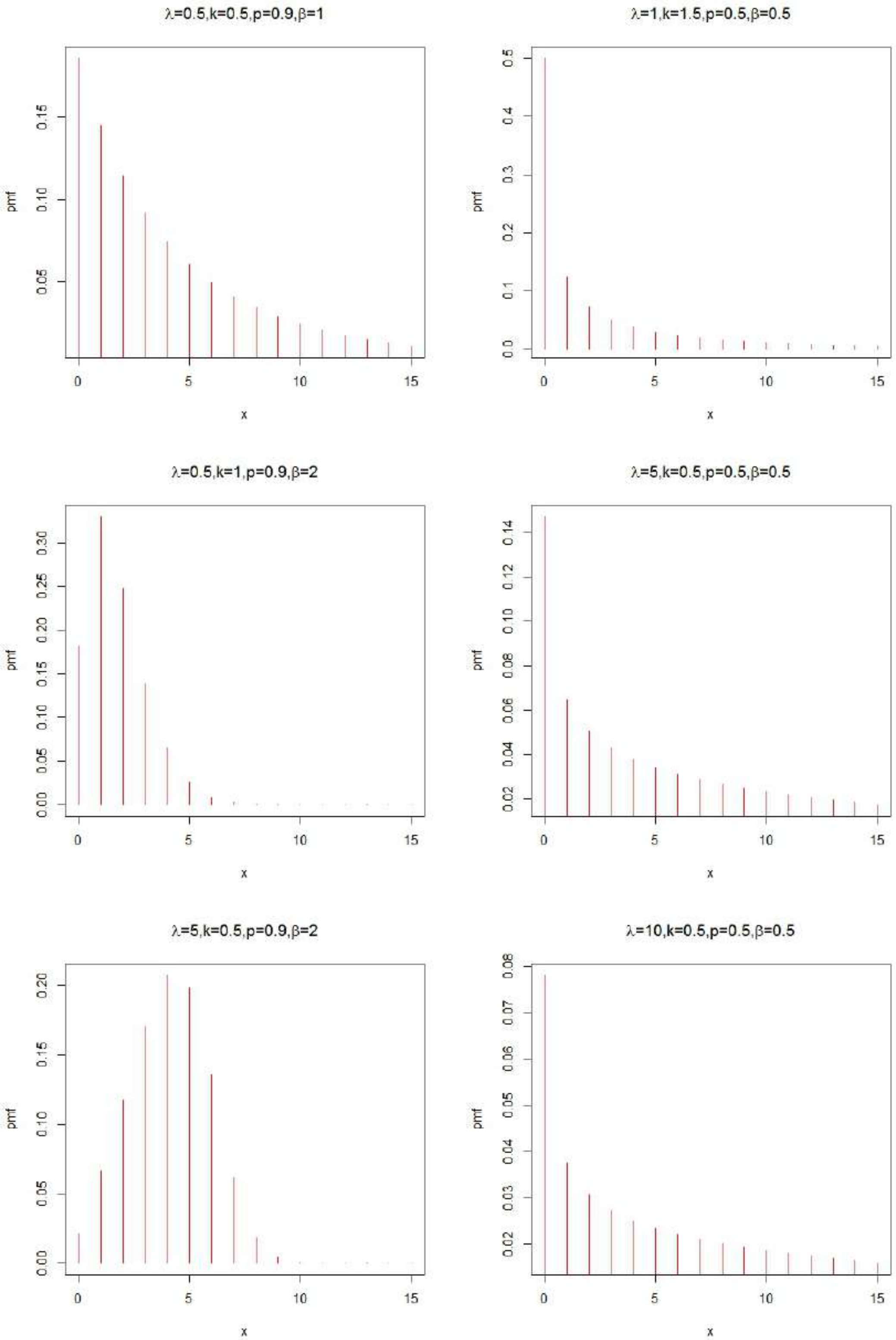


**Figure 2: pmf of DHEW distribution for various values of parameters**

shapes of the pmf of the DHEW distribution. The pmf can be increasing, decreasing and upside-down bathtub shaped. The hazard rate is given by

$$R(x) = 1 - \frac{p^{(x+1)^\beta}\left[1 - \bar{\lambda}p^{kx^\beta}\right]^{1/k}}{p^{x^\beta}\left[1 - \bar{\lambda}p^{k(x+1)^\beta}\right]^{1/k}} \tag{14}$$

Figure 3 displays possible shapes of the hrf of DHEW distribution for selected values of the parameters $\lambda$,$k > 0$,$p$ and $\beta > 0$ respectively. Obviously, from figure it is clear that the hrf can be increasing, decreasing, bathtub and upside-down bathtub shaped.

### 5.1. Special sub-models

Some discrete distributions that are special cases of DHEW distribution are:

(1) When $k = 1$, we obtain

$$p_x = \frac{\lambda[p^{x^\beta} - p^{(x+1)^\beta}]}{[\lambda + (1-\lambda)(1-p^{x^\beta})][\lambda + (1-\lambda)(1-p^{(x+1)^\beta})]}$$

which is considered as the discrete version of Marshall-Olkin Weibull distribution.

(2) When $\lambda = 1$,$k = 1$, we obtain discrete Weibull distribution of Nakagawa and Osaki(1975).In addition $\beta = 1$ geometric distribution is achieved.

(3) If $\beta = 2$, then the pmf reduce to

$$P(X = x) = p_x = \frac{\lambda^{1/k}p^{x^2}}{\left[1 - \bar{\lambda}p^{kx^2}\right]^{1/k}} - \frac{\lambda^{1/k}p^{(x+1)^2}}{\left[1 - \bar{\lambda}p^{(x+1)^2}\right]^{1/k}}, \quad x = 0, 1, 2, ...$$

which defines discrete version of Harris Extended Rayleigh distribution.

(4) If $\beta = 2$ and $\lambda = 1$,

$$p_x = \frac{\lambda[p^{x^2} - p^{(x+1)^2}]}{[\lambda + (1-\lambda)(1-p^{x^2})][\lambda + (1-\lambda)(1-p^{(x+1)^2})]}$$

which is the discrete version of Marshall-Olkin Rayleigh distribution. Moreover with $k = 1$, we get discrete Rayleigh distribution of Roy (2014).

### 5.2. Probability generating function, quantiles, mean and variance

The pgf of DHEW$(\lambda, k, p, \beta)$ is given by

$$P_X(s) = 1 + \lambda^{1/k}(s-1)\sum_{x=1}^{\infty} s^{x-1}\frac{p^{x^\beta}}{\left[1 - \bar{\lambda}p^{kx^\beta}\right]^{1/k}}$$

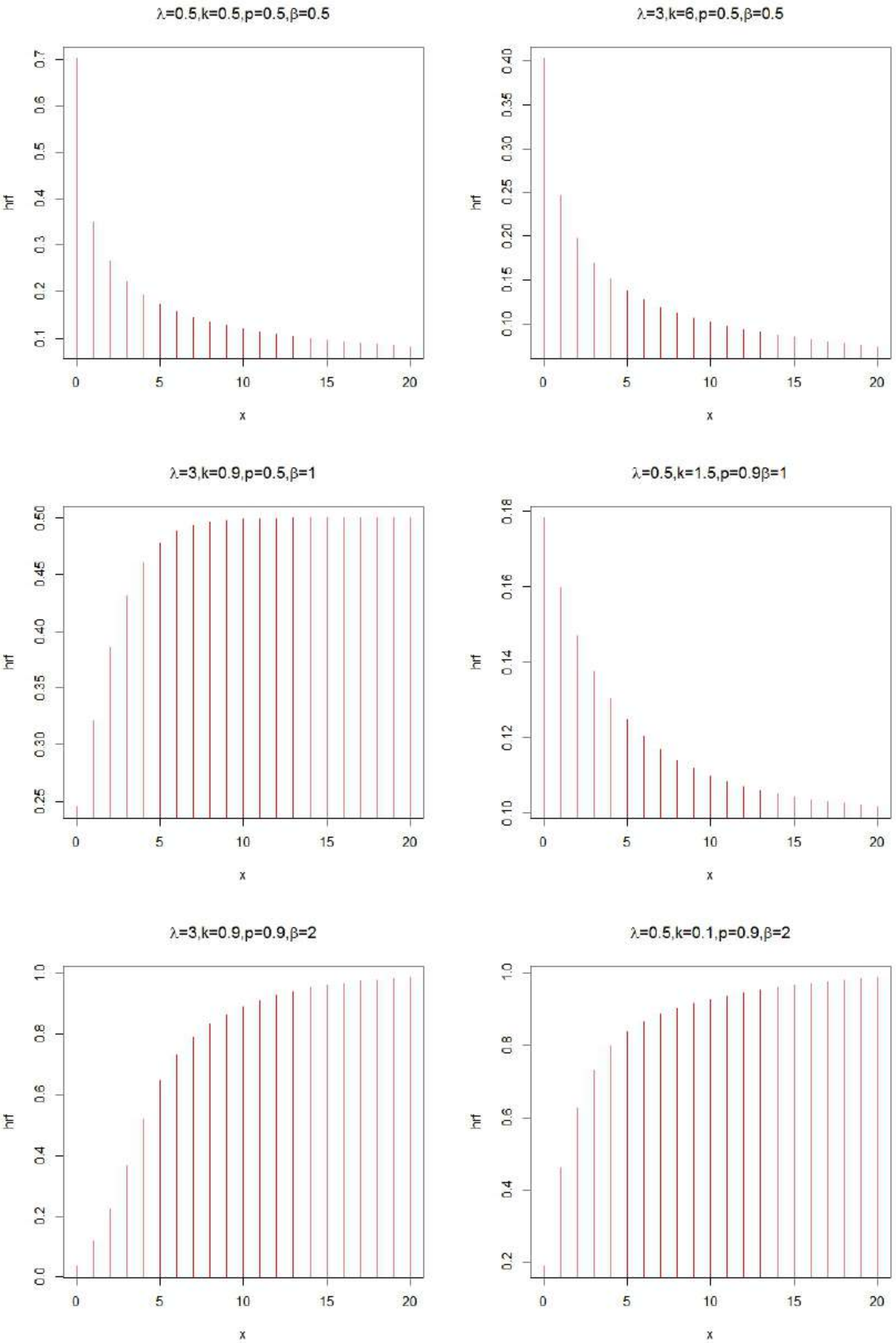


**Figure 3: hrf of DHEW distribution for various values parameters**

The $m^{th}$ quantile of DHEW distribution is denoted by $q_m$ and is given by

$$q_m = \left\{ \frac{\log p \ \log[\bar{\lambda} + \lambda(1-m)^{-k}]}{k} \right\}^{1/\beta} - 1$$

In particular, the Median is

$$Median = \left[ \frac{\log p \ \log(\bar{\lambda} + 2^k \lambda)}{k} \right]^{1/\beta} - 1$$

The expression for mean and variance of DHEW$(\lambda, k, p, \beta)$ is given by

$$E(X) = \lambda^{1/k} \sum_{x=1}^{\infty} \frac{p^{x^\beta}}{\left[1 - \bar{\lambda} p^{kx^\beta}\right]^{1/k}} \tag{15}$$

and

$$V(X) = \lambda^{1/k} \sum_{x=1}^{\infty} (2x-1) \frac{p^{x^\beta}}{\left[1 - \bar{\lambda} p^{kx^\beta}\right]^{1/k}} - \left[ \lambda^{1/k} \sum_{x=1}^{\infty} \frac{p^{x^\beta}}{\left[1 - \bar{\lambda} p^{kx^\beta}\right]^{1/k}} \right]^2 \tag{16}$$

The mean and variance of a DHEW$(\lambda, k, p, \beta)$ distribution for different values of parameters are calculated numerically in Table 1 using the expression (15) and (16). From the Table 1, we can see that depending on the values of parameters, the mean of the distribution can be equal, smaller or greater than the variance. Hence DHEW models are appropriate for modelling both over and under dispersed data.

### 5.3. Infinite divisibility

According to Steutel and van Harn (2004, pp. 56) if $p_x$, $x \in N_0$ is infinitely divisible, then $p_x < e^{-1}$ for all $x \in N$. However, *e.g.*, in a DHEW(0.25, 0.15, 0.9, 2) distribution, we see that $p_1 = 0.4493 > e^{-1} = 0.367$. Therefore, in general, DHEW$(\lambda, k, p, \beta)$ distribution is not infinitely divisible. In addition, since the class of self decomposable and stable distributions, in their discrete concept, are subclass of infinitely divisible distributions, we can conclude that DHEW distribution can be neither self decomposable nor stable, in general.

## 6. Estimation

To apply the method of maximum likelihood for estimating $\lambda$, $k$, $p$ and $\beta$ assume that $X_1$, $X_2$, ..., $X_n$ is a random sample of size $n$ from DHEW distribution. The log-likelihood function is

$$L = \frac{n}{k} log\lambda + \sum_{i=1}^{n} log \left[ \frac{p^{x_i^\beta}}{[1 - \bar{\lambda} p^{kx_i^\beta}]^{1/k}} - \frac{p^{(x_i+1)^\beta}}{[1 - \bar{\lambda} p^{k(x_i+1)^\beta}]^{1/k}} \right] \tag{17}$$

Hence, the likelihood equations are,

$$\frac{\partial L}{\partial \lambda} = \frac{n}{k\lambda} + \sum_{i=1}^{n} \frac{[V_{\lambda,k,\beta}(x_i) - V_{\lambda,k,\beta}(x_i+1)]}{k m_{\lambda,k,\beta}(x_i)} \tag{18}$$

**Table 1: The mean(standard deviation) of DHEW for different parameters**

| | $p \longrightarrow$<br>$\beta \downarrow$ | 0.25 | 0.5 | 0.75 |
|---|---|---|---|---|
| $k = 0.5$ | 0.50 | 1.2787(1.2759) | 2.555(5.3605) | 11.3245(31.4465) |
| $\lambda = 0.5$ | 0.75 | 1.16958(0.5932) | 1.70562(1.6561) | 3.9703(5.5288) |
| | 3.50 | 1.1111(0.3141) | 1.2992(0.4580) | 1.5951(0.5143) |
| $k = 0.5$ | 0.50 | 2.2751(3.0416) | 7.0255(12.4002) | 37.9023(72.2195) |
| $\lambda = 1.5$ | 0.75 | 1.36(1.2370) | 3.2516(3.2333) | 9.2204(10.542) |
| | 3.50 | 1.1644(0.48) | 1.61491(0.4883) | 1.8937(0.4887) |
| $k = 1$ | 0.50 | 1.4074(1.6471) | 3.1233(6.8603) | 14.6716(40.145) |
| $\lambda = 0.5$ | 0.75 | 1.2357(0.7304) | 1.8981(2.0028) | 4.6081(6.6489) |
| | 3.50 | 1.1428(0.3498) | 1.3335(0.4762) | 1.6196(0.5244) |
| $k = 1$ | 0.50 | 2.0718(2.6700) | 6.1911(10.918) | 33.0475(6.3592) |
| $\lambda = 1.5$ | 0.75 | 1.5943(1.1210) | 3.0204(2.9410) | 8.4713(9.5965) |
| | 3.50 | 1.333(0.4713) | 1.6005(0.4909) | 1.8749(0.4721) |
| $k = 3$ | 1.00 | 1.3808 (0.7000) | 2.1324(1.4620) | 4.3543(3.5536) |
| $\lambda = 1.5$ | 2.00 | 1.2899(0.4634) | 1.6347(0.6195) | 2.2640(0.8981) |
| | 3.50 | 1.2854(0.4515) | 1.5613(0.4970) | 1.8496(0.4648) |

$$\frac{\partial L}{\partial k} = \frac{-n}{k^2\lambda} + \sum_{i=1}^{n} \frac{\bar{\lambda}[W_{\lambda,k,\beta}(x_i+1) - W_{\lambda,k,\beta}(x_i)]}{m_{\lambda,k,\beta}(x_i)} \tag{19}$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\bar{\lambda} logp[U_{\lambda,k,\beta}(x_i+1) - U_{\lambda,k,\beta}(x_i)]}{m_{\lambda,k,\beta}(x_i)} \tag{20}$$

$$\frac{\partial L}{\partial p} = \sum_{i=1}^{n} \frac{\bar{\lambda}[(x_i+1)^{\beta} V_{\lambda,k,\beta}(x_i+1) - (x_i)^{\beta} V_{\lambda,k,\beta}(x_i)]}{m_{\lambda,k,\beta}(x_i)} \tag{21}$$

where,

$$m_{\lambda,k,\beta}(x) = \frac{p^{x^{\beta}}}{[1-\bar{\lambda}p^{kx^{\beta}}]^{1/k}} - \frac{p^{(x+1)^{\beta}}}{[1-\bar{\lambda}p^{k(x+1)^{\beta}}]^{1/k}}$$

$$V_{\lambda,k,\beta}(x) = p^{x^{\beta}} \left(\frac{1}{1-\bar{\lambda}p^{kx^{\beta}}}\right)^{\frac{1}{k}-1} p^{kx^{\beta}}$$

$$W_{\lambda,k,\beta}(x) = p^{x^{\beta}} \left(\frac{1}{1-\bar{\lambda}p^{kx^{\beta}}}\right)^{\frac{1}{\beta}} log\left(\frac{1}{1-\bar{\lambda}p^{kx^{\beta}}}\right) p^{kx^{\beta}} log(p^{x^{\beta}})$$

$$U_{\lambda,k,\beta}(x) = p^{x^{\beta}} \left(\frac{1}{1-\bar{\lambda}p^{kx^{\beta}}}\right)^{\frac{1}{k}-1} p^{kx^{\beta}} x^{\beta} logx + \left(\frac{1}{1-\bar{\lambda}p^{kx^{\beta}}}\right)^{\frac{1}{k}} p^{x^{\beta}} x^{\beta} logx$$

The solutions of likelihood equations (18)-(21) provide the maximum likelihood estimators (MLEs) of $\theta = (\lambda, k, p, \beta)^T$, say $\hat{\theta} = (\hat{\lambda}, \hat{k}, \hat{p}, \hat{\beta})^T$, which can be obtained by a numerical method such as the four variable Newton -Raphson type procedure.

### 6.1. Simulation study

Here we study the performance of the MLEs of the model parameters of DHEW distribution using Monte Carlo simulation for various sample sizes and for selected parameter values. We have taken the parameter values as $\lambda$= 1,$\beta$= 0.5, $k$= 0.2and $p$= 0.8 and generated random samples of size $n = 30$, 50 and 60 respectively. The MLEs of $\lambda$,$\beta$ $k$ and $p$ are determined by maximizing the log-likelihood function using the nlm package of R software based on each generated samples. This simulation is repeated 1000 times and the average estimates of bias and MSE are computed and presented in Table 2. It can be seen that, as the sample size increases, the bias tends to zero and MSE decreases.

**Table 2: Simulation results related to the paramters of the DHEW distribution**

| Sample size | Estimates | Average bias | MSE |
|---|---|---|---|
| 30 | 0.6079 | -0.3920 | 0.9131 |
| | 0.2916 | -0.2083 | 0.1060 |
| | 0.1193 | -0.0806 | 0.0238 |
| | 0.4630 | -0.3369 | 0.2700 |
| 50 | 0.9081 | -0.0918 | 0.0917 |
| | 0.4548 | -0.0451 | 0.0228 |
| | 0.1828 | -0.0171 | 0.0048 |
| | 0.7273 | -0.0726 | 0.0582 |
| 60 | 0.9990 | -0.0009 | 0.0009 |
| | 0.5004 | 0.0005 | 0.0002 |
| | 0.2021 | 0.0021 | 0. 0043 |
| | 0.8002 | 0.0002 | 3.8699e-05 |

## 7. Application

In this section, we illustrate the flexibility of the proposed distribution using two real data sets. Maximum likelihood estimation is used to obtain the parameter estimates of the models(using R software). We compare the fit of the DHEW distribution with the following discrete life time distributions.

(a) Exponentiated discrete Weibull (EDW) distribution (Nekoukhou and Bidram 2015) having pmf

$$P(X=x) = (1-p^{(x+1)^{\alpha}})^{\beta} - (1-p^{x^{\alpha}})^{\beta};\ \ 0<p<1,\ \alpha>0,\ \beta>0,\ x=0,1,2,...$$

(b) The pmf of the discrete Gamma (DG) distribution, which has been used first by Yang (1994) and recently considered by Chakraborty and Chakravarty (2012), is given by

$$P(X=x) = \frac{\gamma(\alpha,\beta(x+1)) - \gamma(\alpha,\beta x)}{\Gamma(\alpha)},\ \alpha>0,\ \beta>0$$

where,$\gamma(\alpha,x) = \int_0^x t^{\alpha-1}e^{-t}dt$ denotes the incomplete gamma function.

**Table 3: Aarset data**

| Time of failure | 0 | 1 | 2 | 3 | 6 | 7 | 11 | 12 | 18 | 21 | 32 | 36 | 40 | 45 | 46 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of failures | 2 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| Time of failure | 47 | 50 | 55 | 60 | 63 | 67 | 72 | 75 | 79 | 82 | 83 | 84 | 85 | 86 | |
| No. of failures | 1 | 1 | 1 | 1 | 2 | 4 | 1 | 1 | 1 | 2 | 1 | 3 | 5 | 2 | |

**Table 4: Fitted estimates for Aarset data**

| Distribution | MLEs | AIC | K-S |
|---|---|---|---|
| DHEW | $(\hat{\lambda}, \hat{k}, \hat{\beta}, \hat{p}) = (2.9, 0.30, 0.1248, 0.0403)$ | **484.777** | **0.1739** |
| EDW | $(\hat{\alpha}, \hat{\beta}, \hat{p}) = (13.0059, 0.2517, 0.2675)$ | 509.864 | 0.2194 |
| DW | $(\hat{\beta}, \hat{p}) = (1.0228, 0.9805)$ | 487.2202 | 0.1867 |

(c) A generalization of discrete Rayleigh (GDR) distribution of Roy (2004) having pmf

$$P(X = x) = (1 - p^{(x+1)^2})^{\gamma} - (1 - p^{x^2})^{\gamma}; \quad 0 < p < 1, \ \alpha > 0, \ x = 0, 1, 2, ...$$

(d) Discrete Weibull(DW) distribution (Nakagawa and Osaki 1975) having pmf

$$P(X = x) = p^{x^{\alpha}} - p^{(x+1)^{\alpha}}; \quad 0 < p < 1, \ \alpha > 0, \ x = 0, 1, 2, ...$$

The values of the K-S (Kolmogrov- Smirnov) statistic and AIC (Akaike Information Criterion with correction) are calculated for the four distributions in order to verify which distribution fits better to the data. The better distribution corresponds to smaller values of -log L, K-S statistic and AIC as well as larger p-value. Here, $AIC = -2LogL + 2k$, where, $L$ is the likelihood function evaluated at the maximum likelihood estimates, $k$ is the number of parameters and n is the sample size.

### 7.1. Discrete Aarset data

Aarset (1987) data consist of the failure times (in weeks) of 50 devices put on a life test. The TTT (Total Time on Test) plot for this data shows that the hazard rate has a bathtub-shape. The data set is given in Table 3.

The MLE of parameters of the models and the measures AIC and K-S statistic are given in Table 4. From Table4, we can see that AIC, K-S statistic are smallest for DHEW with AIC=484.77 and K-S statistic value=0.1739. Hence DHEW model gives a better fit to the data.

### 7.2. Discrete Karlis and Xekalaki data

In this section, the DHEW model will be examined for a real data set which is given by Karlis and Xekalaki (2001) on the numbers of fires in Greece for the period from 1 July 1998 to 31 August of the same year. This data set consists of 123 observations and are presented

**Table 5: Numbers of fires in Greece**

| Numbers | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 15 | 16 | 20 | 43 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 16 | 13 | 14 | 9 | 11 | 13 | 8 | 4 | 9 | 6 | 3 | 4 | 6 | 4 | 1 | 1 | 1 |

**Table 6: Fitted estimates for discrete Karlis and Xekalaki data**

| Distribution | Estimated Parameters | AIC | K-S |
|---|---|---|---|
| DHEW | $(\hat{\lambda}, \hat{k}, \hat{\beta}, \hat{p}) = (8.5913, 0.9718, 0.6705, 0.4393)$ | **693.5843** | **0.128** |
| EDW | $(\hat{\alpha}, \hat{\beta}, \hat{p}) = (1.1573, 1.0511, 0.8449)$ | 694.1897 | 0.1285 |
| GDR | $(\hat{\alpha}, \hat{p}) = (0.3934, 0.9924)$ | 694.6178 | 0.1467 |
| DG | $(\hat{\alpha}, \hat{\beta}) = (0.7525, 0.1543)$ | 749.7162 | 0.2683 |

in Table 5. Only fires in forest districts are considered. Bakouch et al. (2014) considered these data to indicate the potentiality of discrete Lindley (DL) distribution in data modeling and compared it with Poisson, geometric and discrete gamma (DG) distributions.

The MLEs of parameters of the models and the measures AIC and K-S statistic are given in Table 6. The MLEs and K-S test statistic values of the DG distribution, given in this table, are directly reported from Table 7 of Bakouch et al. (2014). From Table 6, we can see that AIC, K-S statistic are smallest for DHEW with AIC=693.5843 and K-S statistic value=0.128. Hence DHEW model gives a better fit to this data.

## 8. Conclusion

In this paper, we have introduced a new family of discrete Harris extended distributions. This family is a generalization of discrete Marshall-Olkin family of distributions. We obtained generalizations of discrete exponential, discrete uniform, discrete Weibull and many other discrete distributions using this family. As an illustration, we have studied discrete Harris extended Weibull distribution in detail. From the results presented here, it can be seen that the generalized discrete Harris extended Weibull distribution introduced in this paper appears to be more suitable for modeling many real data sets and is a better alternative to some existing distributions.

## Acknowledgements

We are grateful to the Managing Editor and Chair Editor for their valuable suggestions and comments which helped in improving the paper.

## References

Abouammoh, A. M., Alshangiti, A. M. and Ragab, I. E. (2015). A new generalized Lindley distribution. *Journal of Statistical Computation and Simulation*, **85**, 3662-3678.

Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*. **36**, 106-108.

Batsidis, A. and Lemonte, A. J. (2014). On the Harris extended family of distributions. *Statistics: A Journal of Theoretical and Applied Statistics*, **49**, 1400-1421.

Bakouch, H. S., Jazi, M. A. and Nadarajah, S. (2014). A new discrete distribution. *Statistics*, **48**, 200-240.

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions-a survey of methods and constructions. *Journal of Statistical Distributions and Applications*, **2**, 1-30.

Chakraborty, S. and Chakravarty, D. (2012). Discrete gamma distributions: Properties and parameter estimations. *Communications in Statistics-Theory and Methods*, **41**, 3301-3324.

Ghosh, T., Roy, D. and Chandra, N. K. (2013). Reliability approximation through the discretization of random variables using reversed hazard rate function. *International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering, International Science Index*, **7**, 96 -100.

Gillariose, J., Balogun, O. S., Almetwally , E. M., Sherwani, R. A. K., Jamal, F. and Joseph, J. (2021). On the Discrete Weibull Marshall Olkin family of distributions: Properties, characterizations, and applications. *Axioms*, **10**, 287.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237-264.

G*ó*mez-D*é*niz E. (2010). Another generalization of the geometric distribution.*Test*, **19**, 399-415.

Jayakumar, K. and Sankaran, K. K. (2018). A generalization of discrete Weibull distribution. *Communications in Statistics - Theory and Methods*, **47**, 6064-6078

Jose, K. K., Tomy L. and Thomas, S. P. (2018). On a generalization of the Weibull istribution and its application in quality control. *Stochastics and Quality Control*, **33** , 113-124.

Kemp, A. W.(1997). Characterizations of a discrete normal distribution. *Journal of Statistical Planning and Inference*, **63**, 223- 229.

Karlis, D. and Xekalaki, E. (2001). *On Some Discrete Valued Time Series Models Based on Mixtures and thinning*, in Proceedings of the Fifth Hellenic-European Conference on Computer Mathematics and Its Applications, E. A. Lipitakis Ed., 872-877.

Krishna, H. and Pundir, P. S., (2009). Discrete burr and discrete Pareto distributions. *Statistical Methodology*, **6**, 177-188.

Khan, M. S. A., Khalique, A. and Abouammoh, A. M. (1989). On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability*, **38** , 348-350.

Kulasekera, K. and Tonkyn D. W. (1992). A new discrete distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation*, **21**, 499-518.

Lai, C. D., Xie M. and Murthy D. N. P. (2003). A modified Weibull distribution. *IEEE Transactions on Reliability*, **52**, 33-37.

Marshall A. W. and Olkin I. (1997). A new method for adding a parameter to a family of distributions with Application to the Exponential and Weibull Families. *Biometrika*, **84**, 641-652.

Mudholkar, G. S. and Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure rate data. *IEEE Transactions on Reliability*, **42**, 299-302.

Nakagawa, T. and Osaki, S. (1975). Discrete Weibull distribution. *IEEE Transactions on Reliability*, **24**, 300-301.

Nekoukhou, V. and Bidram, H. (2015). The exponentiated discrete Weibull distribution. *SORT*, **39**, 127-146.

Stein, W. E. and Dattero, R. (1984). A new discrete Weibull distribution. *IEEE Transactions on Reliability*, **33**, 196-197.

Roy, D. (2004). Discrete Rayleigh distribution. *IEEE Transactions on Reliability*, **53**, 255-260.

Sandhya, E. and Prasanth, C. B. (2012). *A Generalized Geometric Distribution.* In Proceedings Of International Conference on Frontiers of Statistics and Application and $32^{nd}$ Annual Conference of Indian Society for Probability and Statistics, Department of Statistics, Podichery University, Dec-2012, Bonfring publication, 261-269, ISBN 978-93-82338-78-9

Sandhya, E. and Prasanth, C. B. (2013). Marshall-Olkin discrete uniform distribution. *Journal of Probability*, **2014**, 1-10.

Sandhya, E. and Prasanth, C. B. (2016). A generalized discrete uniform distribution. *Journal of Statistics Applications and Probability*, **1**, 109-121.

Seethalekshmi, V., Sebastian, S. and Joseph J. (2016). Frechet distribution on integers: Properties and estimation. *International Journal of Computer and Mathematical Sciences*, **5**, 2347-8527.

Stein, W. E. and Dattero, R. (1984). A new discrete Weibull distribution. *IEEE Transactions on Reliability*, **33**, 196-197.

Supanekar, S. R. and Shirke, D. T. (2015). A new discrete family of distributions. *ProbStat Forum*, **8**, 83-94.

Satheesh, S., Sandhya, E. and Sherly, S. (2006). A generalization of stationary AR(1) schemes. *Statistical Methods*, **8**, 213-225.

Steutel, F. W. and Van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line*. New York: Marcel Dekker.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 81–98

# Nonparametric Prediction Intervals for Future Order Statistics and $k$-Record Values

**Laji Muraleedharan and Manoj Chacko**
*Department of Statistics, University of Kerala, Trivandrum, India*

Received: 03 February 2022; Revised: 25 April 2022; Accepted: 26 May 2022

## Abstract

In this paper, we obtain distribution-free prediction intervals for future order statistics based on an observed sequence of $k$-record values. The Prediction intervals for future $k$-record values based on observed order statistics and prediction intervals of future record values based on observed $k$-record values are also derived in a similar manner. The coverage probabilities of the derived intervals are exact and independent of the parent distribution. Finally, two real data sets are used to illustrate the proposed methodologies developed in this paper.

*Key words*: Prediction intervals; Order statistics; Record values; $k$-Record values.

**AMS Subject Classifications:** Primary:62G30; Secondary: 62E15

## 1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ arising from a population with absolutely continuous cumulative distribution function (cdf) $G(x)$ and probability density function (pdf) $g(x)$. By arranging the random sample in an increasing order of magnitude as $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$, the order statistics of the sample can be obtained. The $ith$ order statistic of the sample $X_1, X_2, \ldots, X_n$ is then $X_{i:n}$. Order statistics have wide range of applications in many fields including industry, reliability analysis and material strength. For more discussions regarding the order statistics, one may refer to Arnold *et al.* (1992) and David and Nagaraja (2003). One major application of order statistics in the study of reliability of systems is the following. A system is called a *k-out-of-m* system if it consists of $m$ components and the system functions satisfactorily if at least $k\,(\leq m)$ components function. If the lifetimes of the components are independently distributed, then the lifetime of the system coincides with that of the $(m-k+1)\,th$ order statistic of the lifetime of the components. Thus, order statistics play a key role in studying the lifetimes of such systems.

The cdf of the $ith$ order statistic $X_{i:n}$ based on a random sample of size $n$ from a continuous population with cdf $G(x)$ and pdf $g(x)$ is given by (see, Arnold *et al.*,1992)

$$F_{i:n}(x) = \sum_{r=i}^{n} \binom{n}{r} \left[G(x)\right]^r \left[\bar{G}(x)\right]^{n-r}, \; -\infty < x < \infty. \tag{1}$$

Corresponding Author: Laji Muraleedharan
Email: lajikmr@gmail.com

The pdf corresponding to the cdf (1) is given by

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)} [G(x)]^{i-1} \left[\bar{G}(x)\right]^{n-i} g(x), \ -\infty < x < \infty, \tag{2}$$

where $\bar{G} = 1 - G$ and $B(.,.)$ denotes the complete beta function.

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed (iid) random variables with an absolutely continuous cdf $G(x)$ and pdf $g(x)$. If an observation $X_j$ exceeds all of its previous observations, that is, $X_j > X_i$ for every $i < j$, then it is referred to as an upper record value. Thus $X_1$ is the first upper record value by definition. Similarly, the lower record values can be defined. Many authors have studied the record values of iid random variables as well as their features in the literature. Arnold *et al.* (1998), Ahsanullah (1995) and the literature referenced therein can be used to have a more in-depth look at this topic.

Since Chandler (1952) brought up the idea of record values for the first time in the literature, there has been a significant growth in the study of record values. Record values have many statistical applications, such as modelling and inference involving data pertaining to mining, sports, industry, seismology, life testing and so on. Interested Surveys are given in Glick (1978), Gulati and Padgett (1994), Ahsanullah (1995), Arnold *et al.* (1998), Nagaraja (1988) and the literature cited therein.

One of the challenges in dealing with problems involving inference with record data is that the expected waiting time for consecutive records after the first is infinite. Such an issue does not arise if we use the $k$-records proposed by Dziubdziela and Kopociński (1976). We use the following formal definition of $k$-record values given by Arnold *et al.* (1998).

For a fixed positive integer $k$, the upper $k$-record times $\tau_{n(k)}$ and the upper $k$-record values $U_{n(k)}$ are defined as follows.
Define $\tau_{1(k)} = k$ and $U_{1(k)} = X_{1:k}$ then for $n > 1$,

$$\tau_{n(k)} = min\left\{i : i > \tau_{n-1(k)}, \ X_i > X_{\tau_{n-1(k)}-k+1:\tau_{n-1(k)}}\right\}.$$

Then the sequence of upper $k$-record values $\left\{U_{n(k)}, n \geq 1\right\}$ is defined as

$$U_{n(k)} = X_{\tau_{n(k)}-k+1:\tau_{n(k)}}.$$

The cdf of the $n$th upper $k$-record value $U_{n(k)}$ for $n \geq 1$ is given by (see, Arnold *et al.*,1998)

$$F_{n(k)}(x) = 1 - \left[\bar{G}(x)\right]^k \sum_{i=1}^{n-1} \frac{\left[-k \log \bar{G}(x)\right]^i}{i!}, \quad -\infty < x < \infty. \tag{3}$$

The pdf corresponds to the cdf (3) is given by

$$f_{n(k)}(x) = \frac{k^n}{\Gamma(n)} \left[-\log \bar{G}(x)\right]^{n-1} \left[\bar{G}(x)\right]^{k-1} g(x), \ -\infty < x < \infty, \tag{4}$$

where $\Gamma(\cdot)$ denotes the complete gamma function. Similarly, we can define the lower $k$-record values as well.

For a fixed positive integer $k$, the sequence of lower $k$-record times $\left\{\tau^*_{n(k)}\right\}$ and lower $k$-record value $L_{n(k)}$ are defined as follows. Let
$\tau^*_{1(k)} = k$ and $L_{1(k)} = X_{k:k}$ then for $n > 1$,

$$\tau^*_{n(k)} = min\left\{j : j > \tau^*_{n-1(k)},\ X_j < X_{k:\tau^*_{n-1(k)}}\right\}.$$

Now the sequence of lower $k$-record values $\left\{L_{n(k)}, n \geq 1\right\}$ is defined by

$$L_{n(k)} = X_{k:\tau^*_{n(k)}}.$$

Recently, the $k$-records data has shown an increased trend in problems involving statistical inference and future event prediction. Chacko and Muraleedharan (2018) have obtained the Bayesian and maximum likelihood estimators for the parameters of a generalized exponential distribution based on $k$-record values. The same problem was discussed by Muraleedharan and Chacko (2019) for Gompertz distribution. The recurrence relation for the single and product moment of Gompertz distribution and its characterization based on $k$-records were studied by Minimol and Thomas (2014). The Bayesian estimation of parameters for a Gumbel distribution and the one sample prediction of future $k$-record values under the Bayesian frame work were studied by Malinowska and Szynal (2004). The best linear unbiased predictor (BLUP) for future $k$-record value based on $k$-records arising from a normal distribution was discussed by Chacko and Mary (2013) whereas the same problem for a generalized Pareto distribution was discussed by Muraleedharan and Chacko (2022). Paul and Thomas (2015) established some properties of upper $k$-record values which characterize the Weibull distribution and has derived the BLUP for the model. Deheuvels and Nevzorov (1994) studied the limiting behaviour of $k$-record values such as strong laws of large numbers, central limit theorems, functional laws of the iterated logarithm and strong invariance principles *etc.*

In statistical inference, predicting future events based on the current knowledge is a fundamental problem. It can be expressed in a variety of ways and in various settings. There are two different sorts of prediction problems. The one sample prediction problem is that the event to be predicted comes from the same sequence of events, whereas the two sample prediction problem is when the event to be predicted comes from a different independent sequence of events.

Several authors have considered prediction problem involving record values and order statistics. Hsieh (1997) developed the explicit expression for the prediction intervals for future Weibull order statistics. Al-Hussaini and Ahmad (2003) obtained the Bayesian prediction bounds for future record values from a general class of distributions. Prediction of distribution-free confidence intervals based on record values, order statistics and progressively type II censored samples were extensively discussed by Ahmadi and Balakrishnan (2005, 2008, 2010), Ahmadi *et al.* (2010) and Guilbaud (2004) respectively. In this paper, we consider the two sample distribution-free prediction intervals for order statistics and $k$-record values.

This paper is structured as follows. In Section 2, we use the observed $k$- record values to derive the prediction intervals and the corresponding prediction coefficient for future order statistics. In Section 3, based on the observed order statistics, we obtain the prediction intervals and its coefficient for the future $k$-record values. In Section 4, we consider the interval prediction of future record values based on observed $k$-record values. In Section 5, two real data sets are used to exemplify the proposed approaches presented in this paper and finally some concluding remarks are made in Section 6.

## 2. Prediction of order statistics based on *k*-record values

In this section, we consider the two-sided prediction intervals for an order statistic from the future sample based on the observed $k$-record values. Let $\left\{R_{i(k)}, i \geq 1\right\}$ be a sequence of observed upper (lower) $k$-record values arising from a population with absolutely continuous cdf $G(x)$. Suppose we are interested in obtaining an interval of the form $\left(R_{s(k)}, R_{t(k)}\right)$, $1 \leq s < t$, for the $rth$ order statistic $Y_{r:n}$, $1 \leq r \leq n$, of the future sample of size $n$ arising from the same population such that

$$P\left(R_{s(k)} \leq Y_{r:n} \leq R_{t(k)}\right) = 1 - \alpha.$$

Then the interval $\left(R_{s(k)}, R_{t(k)}\right)$ is called a $100\,(1-\alpha)\,\%$ prediction interval with prediction coefficient $(1-\alpha)$ for the future order statistic $Y_{r:n}$. In this section, we derive such two-sided prediction intervals for $Y_{r:n}$ with coverage probabilities that are free of the parent distribution function $G$.

### 2.1. Prediction of order statistics based on upper *k*-record values

Let $\{Y_i, i \geq 1\}$ be a sequence of iid random variables having an absolutely continuous cdf $G(x)$ and pdf $g(x)$. In the following theorem, we establish the prediction intervals for future order statistics based on the observed sequence of upper $k$-record values.

**Theorem 1:** Let $\left\{U_{i(k)}, i \geq 1\right\}$ be a sequence of observed upper k-record values arising from a population with absolutely continuous cdf $G$ and pdf $g$. Let $Y_{1:n} \leq Y_{2:n} \leq \cdots \leq Y_{n:n}$ be the order statistics from a future random sample of size $n$ arising from the same cdf $G$. Then $\left(U_{s(k)}, U_{t(k)}\right)$, for $1 \leq s < t$, is a prediction interval for the $rth$ order statistic $Y_{r:n}$, for $1 \leq r \leq n$, whose coverage probability is free of $G$ and is given by

$$\alpha_{1(k)}\left(s,t;r,n\right) = r\binom{n}{r}\sum_{i=s}^{t-1}\sum_{j=0}^{r-1}\binom{r-1}{j}\frac{(-1)^j\,k^i}{(n+k+j+1-r)^{i+1}}. \tag{5}$$

**Proof:** For a given real number $v$ and for $1 \leq s < t$, we have

$$\begin{aligned} P\left(U_{s(k)} \leq v\right) &= P\left(U_{s(k)} \leq v, U_{t(k)} < v\right) + P\left(U_{s(k)} \leq v, U_{t(k)} \geq v\right) \\ &= P\left(U_{t(k)} < v\right) + P\left(U_{s(k)} \leq v \leq U_{t(k)}\right). \end{aligned}$$

Hence

$$P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) = P\left(U_{s(k)} \leq v\right) - P\left(U_{t(k)} < v\right). \tag{6}$$

By using (3), (6) can be expressed as

$$P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) = \left[\bar{G}(v)\right]^k \sum_{i=s}^{t-1} \frac{\left[-k\log\bar{G}(v)\right]^i}{i!}. \tag{7}$$

Now for $1 \leq s < t$, and using the conditioning arguments, we can write (7) as

$$\begin{aligned}
\alpha_{1(k)}(s,t;r,n) &= P\left(U_{s(k)} \leq Y_{r:n} \leq U_{t(k)}\right) \\
&= \int_{-\infty}^{\infty} P\left(U_{s(k)} \leq Y_{r:n} \leq U_{t(k)} | Y_{r:n} = v\right) f_{r:n}(v)dv \\
&= \int_{-\infty}^{\infty} P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) f_{r:n}(v)dv \\
&= \sum_{i=s}^{t-1} \frac{n!}{i!\,(r-1)!\,(n-r)!} \int_{-\infty}^{\infty} \left[-k\log\bar{G}(v)\right]^i \left[\bar{G}(v)\right]^{n+k-r} \left[G(v)\right]^{r-1} \\
&\times \; g(v)dv. \qquad (8)
\end{aligned}$$

Taking $y = -k\log\bar{G}(v)$ and applying the binomial expansion, (8) reduces to the following

$$\begin{aligned}
\alpha_{1(k)}(s,t;r,n) &= \frac{r}{k}\binom{n}{r}\sum_{i=s}^{t-1}\sum_{j=0}^{r-1}\frac{(-1)^j}{i!}\binom{r-1}{j}\int_{y=0}^{\infty} y^i \exp\left[-\left(\frac{n+k+j+1-r}{k}\right)y\right]dy \\
&= r\binom{n}{r}\sum_{i=s}^{t-1}\sum_{j=0}^{r-1}\binom{r-1}{j}\frac{(-1)^j k^i}{(n+k+j+1-r)^{i+1}}. \qquad (9)
\end{aligned}$$

Hence the proof. □

If $n, r$ and the desired confidence level $\alpha_0$ are supplied, we can choose $s$ and $t$ so that $\alpha_{1(k)}(s,t;r,n)$ surpasses $\alpha_0$ . Since $\alpha_{1(k)}(s,t;r,n)$ is a step function, the confidence coefficient may not equal to $\alpha_0$ but may be set to a value somewhat higher than $\alpha_0$. Furthermore, the choice of $s$ and $t$ is not unique. So, for a given confidence level $\alpha_0$, $r$ and $n$, we would like to construct a prediction interval whose expected length as short as possible among all prediction intervals with the same level. First, notice that the two-sided prediction intervals exist for a given $\alpha_0$, $r$ and $n$ if and only if, for large $m$,

$$P\left(U_{1(k)} \leq Y_{r:n} \leq U_{m(k)}\right) \geq \alpha_0.$$

We have evaluated $\alpha_{1(k)}(s,t;r,n)$ for $n = 20, 30$ and some selected values of $(s,t)$ and $r$ for $k = 2$ and $k = 3$ and the values are presented in Table 1. It can be observed that the prediction coefficient is increasing in $r$ when the other parameters $(s,t)$ and $n$ are fixed and achieves reasonable prediction coefficient value when $r$ close to $n$. It is also observed that for fixed $n$, $r$ and $k$, the prediction coefficient $\alpha_{1(k)}(s,t;r,n)$ is decreasing in $s$ and increasing in $t$.

### 2.2. Prediction of order statistics based on lower $k$-record values

In this subsection, we consider the prediction intervals for future order statistics on the basis of the observed lower $k$-record values. If $L_{n(k)}$ denotes the *nth* lower $k$-record value, then the cdf of $L_{n(k)}$ is given by

$$F^{*}_{n(k)}(x) = [G(x)]^{k} \sum_{s=1}^{n-1} \frac{[-k \log G(x)]^{s}}{s!}, \quad -\infty < x < \infty. \tag{10}$$

The pdf corresponds to the cdf (10) is given by

$$f^{*}_{n(k)}(x) = \frac{k^{n}}{\Gamma(n)} \left[-\log G(x)\right]^{n-1} \left[G(x)\right]^{k-1} g(x), \ -\infty < x < \infty. \tag{11}$$

Now we can establish the following theorem for the interval prediction of future order statistics based on the observed sequence of lower $k$-record values.

**Theorem 2:** Suppose the conditions of Theorem 1 hold and let $\left\{L_{i(k)}, i \geq 1\right\}$ be the sequence of observed lower k-record values emerging from the population. Then $\left(L_{t(k)}, L_{s(k)}\right)$, for $1 \leq s < t$, is a prediction interval for the *rth* order statistic $Y_{r:n}$, for $1 \leq r \leq n$, whose prediction coefficient is free of $G$ and is given by

$$\alpha_{2(k)}\left(s, t; r, n\right) = r\binom{n}{r} \sum_{i=s}^{t-1} \sum_{j=0}^{n-r} \binom{n-r}{j} \frac{(-1)^{j}\, k^{i}}{(r+k+j)^{i+1}}. \tag{12}$$

**Proof:** The proof is similar to that of Theorem 1 and thus omitted. □

**Remark 1:** Since $\alpha_{2(k)}\left(s, t; r, n\right) = \alpha_{1(k)}\left(s, t; n-r+1, n\right)$, we can use Table 1 for evaluating (12).

### 2.3. Prediction of order statistics based on upper and lower $k$-record values jointly

In certain studies such as meteorological studies, the upper and lower $k$-record values are observed simultaneously. In such studies, when predicting the order statistics from a future sample, it is preferable to examine both the upper and lower $k$-record values together.

**Theorem 3:** Suppose the conditions of Theorem 1 hold; let $L_{s(k)}$ and $U_{t(k)}$ denote the *sth* lower k-record and *tth* upper k-record values respectively. Then $\left(\text{Ł}_{s(k)}, U_{t(k)}\right)$ is a prediction interval for the *rth* order statistic $Y_{r:n}$, for $1 \leq r \leq n$, with coverage probability free of $G$ and is given by

$$\alpha_{3(k)}\left(s, t; r, n\right) = r\binom{n}{r} \left[\sum_{i=0}^{s-1} \sum_{j=0}^{n-r} \frac{(-1)^{j} \binom{n-r}{j} k^{i}}{(j+k+r)^{i+1}} + \sum_{i=0}^{t-1} \sum_{j=0}^{r-1} \frac{(-1)^{j} \binom{r-1}{j} k^{i}}{(n+j+k+1-r)^{i+1}}\right] - 1. \tag{13}$$

**Proof:** For a fixed real number $v$ and $1 \leq s < t$, we can express

$$P\left(L_{s(k)} \leq v \leq U_{t(k)}\right) = [G(v)]^{k} \sum_{i=0}^{s-1} \frac{[-k \log G(v)]^{i}}{i!} + \left[\bar{G}(v)\right]^{k} \sum_{i=0}^{t-1} \frac{\left[-k \log \bar{G}(v)\right]^{i}}{i!} - 1. \tag{14}$$

Now for $1 \leq s < t$, and using the conditioning arguments, we can write (14) as

$$\begin{aligned}
\alpha_{3(k)}\left(s,t;r,n\right) &= P\left(L_{s(k)} \leq Y_{r:n} \leq U_{t(k)}\right) \\
&= \int\limits_{-\infty}^{\infty} P\left(U_{s(k)} \leq Y_{r:n} \leq U_{t(k)} | Y_{r:n} = v\right) f_{r:n}(v)dv \\
&= \int\limits_{-\infty}^{\infty} P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) f_{r:n}(v)dv \\
&= \sum_{i=1}^{s-1} \frac{n!}{i!\,(n-r)!\,(r-1)!} \int\limits_{-\infty}^{\infty} \left[-k\log G(v)\right]^{i} \left[\bar{G}(v)\right]^{n-r} \left[G(v)\right]^{k+r-1} g(v)dv \\
&+ \sum_{i=1}^{t-1} \frac{n!}{i!\,(n-r)!\,(r-1)!} \int\limits_{-\infty}^{\infty} \left[-k\log \bar{G}(v)\right]^{i} \left[\bar{G}(v)\right]^{k+n-r} \left[G(v)\right]^{r-1} g(v)dv - 1 \\
&= r\binom{n}{r} \sum_{i=1}^{s-1}\sum_{j=0}^{n-r} \frac{(-1)^{j}\binom{n-r}{j}}{i!k} \int\limits_{y=0}^{\infty} y^{i}\left(e^{-\frac{y}{k}}\right)^{j+k+r} dy + r\binom{n}{r} \sum_{i=1}^{t-1}\sum_{j=0}^{r-1} \frac{(-1)^{j}\binom{r-1}{j}}{i!k} \\
&\times \int\limits_{z=0}^{\infty} z^{i}\left(e^{-\frac{z}{k}}\right)^{n+j+k+1-r} dz - 1 \\
&= r\binom{n}{r}\left[\sum_{i=0}^{t-1}\sum_{j=0}^{r-1} \frac{(-1)^{j}\binom{n-r}{j}k^{i}}{(j+k+r)^{i+1}} + \sum_{i=0}^{t-1}\sum_{j=0}^{r-1} \frac{(-1)^{j}\binom{r-1}{j}k^{i}}{(n+j+k+1-r)^{i+1}}\right] - 1.
\end{aligned}$$

Hence the proof. □

Table 2 provides the values of $\alpha_{3(k)}\left(s,t;r,n\right)$ for $n = 10, 20$ and 30 and some selected values of $(s,t)$ and $r$ for $k = 2$ and $k = 3$. We can see that the prediction coefficient improves when the intervals are constructed upper and lower $k$-record values jointly. It is also observed that for fixed $n$, $r$ and $k$, the prediction coefficient $\alpha_{3(k)}\left(s,t;r,n\right)$ is non-decreasing in $s$ and $t$.

## 3. Prediction of future $k$-record values based on order statistics

Suppose we are interested in obtaining an interval for the $rth$ future $k$-record value $R_{r(k)}$ (upper or lower) based on the observed order statistics of size $n$ of the form $\left(X_{s:n}, X_{t:n}\right), 1 \leq s < t \leq n$, such that

$$P\left(X_{s:n} \leq R_{r(k)} \leq X_{t:n}\right) = 1 - \alpha.$$

Then we refer the interval $\left(X_{s:n}, X_{t:n}\right)$ as a $100\left(1-\alpha\right)\%$ prediction interval with prediction coefficient $(1-\alpha)$ for the $rth$ future $k$- record value $R_{r(k)}$. In this section, we derive such two-sided prediction intervals with coverage probabilities being free of the parent distribution.

### 3.1. Prediction of upper *k*- record values based on order statistics

In this subsection, we wish to predict the $rth$ future upper $k$-record value $U_{r(k)}$ based on the observed order statistics.

**Theorem 4:** Let $Y_{1:n} \leq Y_{2:n} \leq \cdots \leq Y_{n:n}$ be the observed order statstics arising from a random sample of size $n$ from a population with absolutely continuous cdf $G$ and pdf $g$ respectively. Then $(Y_{s:n}, Y_{t:n})$, for $1 \leq s < t \leq n$, is a prediction interval for the $rth$ future upper k-record value $U_{r(k)}$ arising from the same population whose coverage probability is free of $G$ and is given by

$$\alpha_{4(k)}\left(s,t;r,n\right) = \sum_{i=s}^{t-1}\sum_{j=0}^{i}\binom{n}{i}\binom{i}{j}\frac{\left(-1\right)^{j}k^{r}}{\left(n+j+k-i\right)^{r}}. \tag{15}$$

**Proof:** For any real number $v$ and $1 \leq s < t \leq n$, by using (1), we obtain the following

$$P\left(Y_{s:n} \leq v \leq Y_{t:n}\right) = \sum_{i=s}^{t-1}\binom{n}{i}\left[G(v)\right]^{i}\left[\bar{G}(v)\right]^{n-i}. \tag{16}$$

Now for $1 \leq s < t \leq n$, and using the conditioning arguments, we can write

$$\begin{aligned}
\alpha_{4(k)}\left(s,t;r,n\right) &= P\left(Y_{s:n} \leq U_{r(k)} \leq Y_{t:n}\right) \\
&= \int_{-\infty}^{\infty} P\left(X_{s:n} \leq v \leq X_{t:n} | U_{r(k)} = v\right) f_{r(k)}(v)dv \\
&= \int_{-\infty}^{\infty} P\left(X_{s:n} \leq v \leq X_{t:n}\right) f_{r(k)}(v)dv \\
&= \sum_{i=s}^{t-1}\binom{n}{i}\frac{k^{r}}{(r-1)!}\int_{-\infty}^{\infty}\left[-\log\bar{G}(v)\right]^{r-1}\left[\bar{G}(v)\right]^{n+k-i-1}\left[G(v)\right]^{i}g(v)dv \\
&= \sum_{i=s}^{t-1}\sum_{j=0}^{i}\binom{n}{i}\binom{i}{j}\frac{k^{r}\left(-1\right)^{j}}{(r-1)!}\int_{-\infty}^{\infty}y^{r-1}\exp\left[-\left(n+j+k-i\right)y\right]dy \\
&= \sum_{i=s}^{t-1}\sum_{j=0}^{i}\binom{n}{i}\binom{i}{j}\frac{\left(-1\right)^{j}k^{r}}{\left(n+j+k-i\right)^{r}}.
\end{aligned}$$

Hence the proof. □

For a given confidence level $\alpha_0$ and specified $r$, we would like to construct prediction interval whose expected length as short as possible among all prediction intervals with the same confidence level. First observe that, for a given $\alpha_0$ and $r$, the two-sided prediction interval exists if and only if

$$P\left(X_{1:n} \leq U_{r(k)} \leq X_{n:n}\right) \geq \alpha_0.$$

Table 3 represents the values of $\alpha_{4(k)}\left(s,n;r,n\right)$ for $n = 10, 20, 30, 35$ and $40$ and some selected values of $s$ and $r$ for $k = 2$ and $k = 3$. We can observe that $\alpha_{4(k)}\left(s,n;r,n\right)$ is decreasing in $r$ and $s$ but improves with $n$ and $k$.

### 3.2. Prediction of lower $k$-record values based on order statistics

For predicting lower $k$-record values, we consider an interval $(X_{s:n}, X_{t:n})$, for $1 \leq s < t \leq n$, based on the observed order statistics. Analogous to the result presented for upper $k$-record values, we obtain the following theorem.

**Theorem 5:** Suppose the conditions of Theorem 4 hold; then $(Y_{s:n}, Y_{t:n})$, for $1 \leq s < t \leq n$, is a prediction interval for the $rth$ future lower k-record value $L_{r(k)}$ arising from the same population whose coverage probability is free of $G$ and is given by

$$\alpha_{5(k)}\left(s, t; r, n\right) = \sum_{i=s}^{t-1} \sum_{j=0}^{n-i} \binom{n}{i} \binom{n-i}{j} \frac{\left(-1\right)^{j} k^{r}}{\left(i+j+k\right)^{r}}. \tag{17}$$

**Proof:** Proof is similar to that of Theorem 4 and hence omitted. □

**Remark 2:** Note that if

$$\alpha_{4(k)}\left(s, t; r, n\right) = \sum_{i=s}^{t-1} \binom{n}{i} \psi_{k}\left(i, j, r; n\right)$$

then

$$\alpha_{5(k)}\left(s, t; r, n\right) = \sum_{i=s}^{t-1} \binom{n}{i} \psi_{k}\left(n-i, j, r; n\right),$$

where

$$\psi_{k}\left(i, j, r; n\right) = \sum_{j=0}^{i} \binom{i}{j} \frac{\left(-1\right)^{j} k^{r}}{\left(n+j+k-i\right)^{r}}. \tag{18}$$

Thus we can use Table 3 for evaluating (17) by making a simple modification.

## 4. Prediction of future record values based on $k$-record values

Let $\left\{R_{i(k)}, i \geq 1\right\}$ be a sequence of observed $k$-record values arising from a population with absolutely continuous cdf $G(x)$. Suppose we are interested in obtaining an interval for the $rth$ future record value $R_r$ of the form $\left(R_{m(k)}, R_{n(k)}\right), 1 \leq m < n$, such that

$$P\left(R_{n(k)} \leq R_r \leq R_{n(k)}\right) = 1 - \alpha.$$

Then we refer the interval $\left(R_{m(k)}, R_{n(k)}\right)$ as a $100\left(1-\alpha\right)\%$ prediction interval with prediction coefficient $(1-\alpha)$ for the future record value $R_r$. In this section, we derive such two-sided prediction intervals for $R_r$ with coverage probabilities being free of the parent distribution $G$.

### 4.1. Prediction of future upper record values based on upper $k$-record values

In this subsection, we wish to predict the $rth$ future upper record value based on the observed sequence of upper $k$-record values.

**Theorem 6:** Let $\left\{U_{i(k)}, i \geq 1\right\}$ be a sequence of observed upper k- record values arising from a population with absolutely continuous cdf $G$. Then $\left(U_{s(k)}, U_{t(k)}\right)$, for $1 \leq s < t$, is a prediction interval for the $rth$ future upper record value $U_r$ arising from the same population with the corresponding prediction coefficient is given by

$$\alpha_{6(k)}\left(s, t; r\right) = \sum_{j=s}^{t-1} \binom{j+r-1}{j} \frac{k^j}{(1+k)^{j+r}}. \tag{19}$$

**Proof:** For a given real number $v$ and for $1 \leq s < t$, we can express

$$P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) = \left[\bar{G}(v)\right]^k \sum_{j=s}^{t-1} \frac{\left[-k \log \bar{G}(v)\right]^j}{j!}. \tag{20}$$

Now for $s < t$, and using the conditioning arguments, we can write (20) as

$$\begin{aligned}
\alpha_{6(k)}\left(s, t; r\right) &= P\left(U_{s(k)} \leq U_r \leq U_{t(k)}\right) \\
&= \int_{-\infty}^{\infty} P\left(U_{s(k)} \leq U_r \leq U_{t(k)} | U_r = v\right) f_{r(1)}(v) dv \\
&= \int_{-\infty}^{\infty} P\left(U_{s(k)} \leq v \leq U_{t(k)}\right) f_{r(1)}(v) dv \\
&= \sum_{j=s}^{t-1} \frac{1}{j!\,(r-1)!} \int_{-\infty}^{\infty} \left[-k \log \bar{G}(v)\right]^j \left[-\log \bar{G}(v)\right]^{r-1} \left[\bar{G}(v)\right]^k g(v) dv \\
&= \sum_{j=s}^{t-1} \binom{j+r-1}{j} \frac{k^j}{(1+k)^{j+r}}.
\end{aligned}$$

Hence the proof. □

Let $W$ denote a negative binomial random variable counting the number of trials needed to get $rth$ success where the success probability $p = 1/\left(k+1\right)$. Then, the expression in (19) can be viewed as the probability that the $rth$ success occurs between $sth$ and $t-1$ trials; that is, it represents $P\left(s \leq W < t\right)$, and hence $\alpha_{6(k)}\left(s, t; r\right)$ can be directly computed from negative binomial cdf using common statistical packages.

## 4.2. Prediction of future lower record values based on lower *k*-record values

In this section, we construct the prediction intervals for future lower record value based on the observed sequence lower $k$- record values. Analogous to the result presented for upper record values, we obtain the following theorem.

**Theorem 7:** Suppose the conditions of Theorem 6 hold, let $\left\{L_{n(k)}, n \geq 1\right\}$ be a sequence of observed lower $k$-record values arising from a population. Then $\left(L_{t(k)}, L_{s(k)}\right)$, for $1 \leq s < t$, is a prediction interval for the $rth$ future lower record value $L_r$ arising from the same population with the corresponding prediction coefficient is given by (19).

**Proof:** Proof is similar to that of Theorem 6 and hence omitted. □

### 4.3. Prediction of upper record value based on lower and upper *k*-record values jointly

There are some situations wherein upper and lower $k$-record values are observed jointly, just as in the case of weather data. In these cases, it would be better to consider the upper and lower $k$-record values jointly to predict the future upper record value of a future sample.

**Theorem 8:** Let $\left\{L_{i(k)}, i \geq 1\right\}$ and $\left\{U_{i(k)}, i \geq 1\right\}$ respectively denote the observed sequences of lower and upper $k$-record values arising from a population with absolutely continuous cdf $G$. Then $\left(L_{s(k)}, U_{t(k)}\right)$, for $1 \leq s < t$, is a prediction interval for the $rth$ future upper record value $U_r$ of the future random sample arising from the same population with the corresponding prediction coefficient, denoted by $\alpha_{7(k)}\left(s,t;r,n\right)$ being free of $G$; it can be expressed as

$$\alpha_{7(k)}\left(s,t;r\right) = \sum_{j=1}^{s-1} \frac{\theta_k\left(j,r\right)}{j!\left(r-1\right)!} + \alpha_{6(k)}\left(0,t;r\right) - 1, \tag{21}$$

where

$$\theta_k\left(j,r\right) = \int\limits_0^1 y^k\left(-k\log y\right)^j\left[-\log\left(1-y\right)\right]^{r-1}dy \tag{22}$$

and $\alpha_{6(k)}\left(0,t,r\right)$ is defined by (19).

**Proof:** For a given real number $v$ and for $1 \leq s < t$, we obtain

$$\begin{aligned} P\left(L_{s(k)} \leq v \leq U_{t(k)}\right) &= P\left(L_{s(k)} \leq v\right) - P\left(U_{t(k)} \leq v\right) \\ &= \sum_{j=0}^{s-1}\frac{\left[-k\log G(v)\right]^j}{j!}\left[G(v)\right]^k + \sum_{j=0}^{t-1}\frac{\left[-k\log \bar{G}(v)\right]^j}{j!}\left[\bar{G}(v)\right]^k - 1. \end{aligned} \tag{23}$$

Now for $1 \leq s < t$, and using the conditioning arguments, we can write (23) as

$$\begin{aligned} \alpha_{7(k)}\left(s,t;r\right) &= P\left(L_{s(k)} \leq U_r \leq U_{t(k)}\right) \\ &= \int\limits_{-\infty}^{\infty} P\left(L_{s(k)} \leq U_r \leq U_{t(k)} | U_r = v\right) g_r(v)dv \\ &= \int\limits_{-\infty}^{\infty} P\left(L_{s(k)} \leq v \leq U_{t(k)}\right) g_r(v)dv \\ &= \sum_{j=0}^{s-1}\frac{1}{j!\left(r-1\right)!}\int\limits_{-\infty}^{\infty}\left[-k\log G(v)\right]^j\left[-\log\bar{G}(v)\right]^{r-1}\left[G(v)\right]^k g(v)dv \\ &+ \sum_{j=0}^{t-1}\frac{1}{j!\left(r-1\right)!}\int\limits_{-\infty}^{\infty}\left[-k\log\bar{G}(v)\right]^j\left[-\log\bar{G}\right]^{r-1}\left[\bar{G}(v)\right]^k g(v)dv - 1. \end{aligned} \tag{24}$$

Taking $y = G(v)$ in the first integral and $z = -k \log \bar{G}(v)$ in the second integral of (24) and then evaluating, we obtain

$$\alpha_{7(k)}\left(s,t;r\right) = \sum_{j=0}^{s-1} \frac{\theta_k\left(j,r\right)}{j!\left(r-1\right)!} + \sum_{j=0}^{t-1} \frac{k^j}{\left(1+k\right)^{j+r}} \binom{j+r-1}{j} - 1, \tag{25}$$

where $\theta_k\left(j,r\right)$ is defined in (22). Therefore the prediction interval for the $rth$ upper record value $U_r$ from the future sequence is $\left(L_{s(k)}, U_{t(k)}\right)$ whose prediction coefficient is free of the parent distribution $G$ and is given by

$$\alpha_{7(k)}\left(s,t;r\right) = \sum_{j=1}^{s-1} \frac{\theta_k\left(j,r\right)}{j!\left(r-1\right)!} + \alpha_{6(k)}\left(0,t;r\right) - 1. \tag{26}$$

Hence the proof. □

Tables 4 and 5 represent the values of $\alpha_{7(k)}\left(s,t;r\right)$ when $r = 1$ and $r = 2$ for $k = 2$ and $k = 3$ with $1 \leq s \leq 7$ and $4 \leq t \leq 7$.

## 5. Illustration using real data

**Example 1:** We use the data set given in Arnold *et al.*(1998, pp.49-50) which represent the average July temparatures (in degrees centigrade) of Neuenberg, Switzerland, during the period 1864-1993, and extract the *2*- record values to illustrate the prediction methods described for predicting future order statistics. Ahmadi and Balakrishnan (2011) used the same data set for predicting future order statistics based on observed ordinary record values. The first order autocorrelation for the data set at the first three lags are 0.022, -0.007 and -0.076 respectively. This small amount of autocorrelation shows that the data is independent in nature. The upper and lower *2*- record values extracted from the data set are obtained as given below.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $U_{m(2)}$ | 19.0 | 19.7 | 20.1 | 21.0 | 21.4 | 21.7 | 22 | 22.1 | 22.3 | 22 |
| $L_{m(2)}$ | 20.1 | 19 | 18.4 | 17.4 | 17.2 | 16.2 | 15.8 | 15.6 | - | - |

Based on the observed upper and lower 2-record values and by using Table 2, we obtain the prediction intervals of future order statistics with prediction coefficient at least 90% for $n = 10, 20$ and 30 are presented in the following table.

| $(n,r)$ | $(s,t)$ | $(L_s,U_t)$ | $\alpha_{3(2)}\left(s,t;r,n\right)$ | $(n,r)$ | $(s,t)$ | $(L_s,U_t)$ | $\alpha_{3(2)}\left(s,t;r,n\right)$ |
|---|---|---|---|---|---|---|---|
| $(10,6)$ | $(5,7)$ | $(17.2,22)$ | 0.9705 | $(20,15)$ | $(4,5)$ | $(17.4,21.4)$ | 0.9456 |
| $(10,8)$ | $(3,8)$ | $(18.4,22.1)$ | 0.9279 | $(30,10)$ | $(7,4)$ | $(15.8,21)$ | 0.9726 |
| $(20,5)$ | $(7,4)$ | $(15.8,21)$ | 0.9433 | $(30,20)$ | $(8,7)$ | $(15.6,22)$ | 0.9900 |

When comparing the results in Table 2 to those of Ahmadi and Balakrishnan (2010), we see that when upper and lower record values are evaluated jointly, the interval prediction coefficient increases with lower values of $k$.

**Example 2:** Consider the amount of annual rainfall at Los Angeles Civic Centre (LACC) during 1900-2000. Then by Ahmadi and Balakrishnan (2011), the order statistics corresponding to the data set is given by

| $r$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $Year$ | 1960 | 1958 | 1923 | 1971 | 1975 | 1947 | 1989 | 1986 | 1969 | 1963 |
| $Y_{r:n}$ | 4.85 | 5.58 | 6.67 | 7.17 | 7.21 | 7.22 | 7.35 | 7.66 | 7.74 | 7.93 |
| r | 20 | 30 | 50 | 70 | 80 | 90 | 95 | 98 | 99 | 100 |
| $Year$ | 1980 | 1941 | 1928 | 1926 | 1921 | 1937 | 1992 | 1982 | 1940 | 1977 |
| $Y_{r:n}$ | 8.96 | 11.10 | 12.66 | 18.03 | 19.66 | 23.43 | 27.36 | 31.28 | 32.76 | 33.44 |

By using Table 3, we obtain the prediction intervals of future $k$-record values with prediction coefficient at least 90% for $k = 2$ and $k = 3$ are presented in the following table.

| $(n, r)$ | $s$ | $(Y_{s:n}, Y_{n:n})$ | $\alpha_{4(2)}\,(s, n; r, n)$ | $(n, r)$ | $s$ | $(Y_{s:n}, Y_{n:n})$ | $\alpha_{4(3)}\,(s, n; r, n)$ |
|---|---|---|---|---|---|---|---|
| $(35, 4)$ | 6 | $(7.21, 33.44)$ | 0.9204 | $(20, 4)$ | 6 | $(7.21, 33.44)$ | 0.9331 |
| $(35, 4)$ | 8 | $(7.66, 33.44)$ | 0.9185 | $(40, 5)$ | 10 | $(7.93, 33.44)$ | 0.9733 |
| $(40, 4)$ | 10 | $(7.93, 33.44)$ | 0.9288 | $(40, 6)$ | 20 | $(8.96, 33.44)$ | 0.9273 |

Ahmadi and Balakrishnan (2010) also used the same data set for constructing prediction intervals for future ordinary record values.

## 6. Conclusion

In this paper, we derived the distribution-free prediction intervals for order statistics and record values based on observed $k$-record values, as well as for future $k$-record values based on observed order statistics. The coverage probabilities of these intervals are exact and independent of the parent distribution. The proposed method can be extended to develop outer and inner prediction intervals for future $k$-record intervals.

## Acknowledgement

The authors would like to thank the Chair Editor and the anonymous reviewer for their valuable comments and suggestions that significantly improved the quality of the paper.

**Table 1: The values of $\alpha_{1(k)}\,(s,t;r,n)$ for $n = 20$ and $30$ and some selected values of $s, t$ and $r$ when $k = 2$ and $k = 3$**

| $n$ | $r$ | $s$ | $k = 2$ | | | | $k = 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $t$ | | | | $t$ | | | |
| | | | 4 | 6 | 8 | 10 | 4 | 6 | 8 | 10 |
| 20 | 5 | 1 | 0.4061 | 0.4111 | 0.4113 | 0.4113 | 0.5208 | 0.5381 | 0.5392 | 0.5392 |
| | | 2 | 0.1103 | 0.1152 | 0.1154 | 0.1154 | 0.1902 | 0.2075 | 0.2086 | 0.2086 |
| | | 3 | 0.0210 | 0.0260 | 0.0261 | 0.0261 | 0.0478 | 0.0651 | 0.0661 | 0.0662 |
| | 10 | 1 | 0.6562 | 0.7089 | 0.7139 | 0.7143 | 0.6789 | 0.8104 | 0.8349 | 0.8381 |
| | | 2 | 0.3204 | 0.3732 | 0.3782 | 0.3785 | 0.4104 | 0.5419 | 0.5664 | 0.5697 |
| | | 3 | 0.1028 | 0.1556 | 0.1606 | 0.1609 | 0.1644 | 0.2958 | 0.3203 | 0.3236 |
| | 12 | 1 | 0.6895 | 0.7894 | 0.8036 | 0.8051 | 0.6274 | 0.8358 | 0.8936 | 0.9049 |
| | | 2 | 0.3927 | 0.4926 | 0.5068 | 0.5083 | 0.4277 | 0.6362 | 0.6939 | 0.7052 |
| | | 3 | 0.1468 | 0.2467 | 0.2609 | 0.2624 | 0.1951 | 0.4036 | 0.4613 | 0.4726 |
| | 15 | 1 | 0.6327 | 0.8414 | 0.8971 | 0.9074 | 0.4378 | 0.7435 | 0.8969 | 0.9501 |
| | | 2 | 0.4331 | 0.6418 | 0.6975 | 0.7078 | 0.3413 | 0.6471 | 0.8004 | 0.8537 |
| | | 3 | 0.1978 | 0.4065 | 0.4622 | 0.4725 | 0.1837 | 0.4894 | 0.6428 | 0.6960 |
| | 17 | 1 | 0.5013 | 0.7894 | 0.9112 | 0.9468 | 0.2635 | 0.5639 | 0.7937 | 0.9152 |
| | | 2 | 0.3794 | 0.6676 | 0.7893 | 0.8249 | 0.2199 | 0.5204 | 0.7502 | 0.8717 |
| | | 3 | 0.1961 | 0.4842 | 0.6059 | 0.6415 | 0.1305 | 0.4309 | 0.6607 | 0.7822 |
| | 19 | 1 | 0.2795 | 0.5792 | 0.8001 | 0.9151 | 0.0959 | 0.2817 | 0.5125 | 0.7129 |
| | | 2 | 0.2313 | 0.5309 | 0.7518 | 0.8668 | 0.0847 | 0.2705 | 0.5014 | 0.7017 |
| | | 3 | 0.1354 | 0.4351 | 0.6560 | 0.7710 | 0.0552 | 0.2410 | 0.4718 | 0.6722 |
| | 20 | 1 | 0.1384 | 0.3555 | 0.5859 | 0.7633 | 0.0342 | 0.1238 | 0.2729 | 0.4487 |
| | | 2 | 0.1195 | 0.3365 | 0.5669 | 0.7443 | 0.0309 | 0.1206 | 0.2697 | 0.4455 |
| | | 3 | 0.0749 | 0.2919 | 0.5223 | 0.6998 | 0.0211 | 0.1108 | 0.2599 | 0.4357 |
| 30 | 20 | 1 | 0.6908 | 0.8379 | 0.8637 | 0.8667 | 0.5546 | 0.8264 | 0.9214 | 0.9433 |
| | | 2 | 0.4366 | 0.5836 | 0.6095 | 0.6124 | 0.4117 | 0.6835 | 0.7785 | 0.8004 |
| | | 3 | 0.1806 | 0.3277 | 0.3535 | 0.3565 | 0.2065 | 0.4783 | 0.5733 | 0.5952 |
| | 22 | 1 | 0.6498 | 0.8527 | 0.9008 | 0.9083 | 0.4523 | 0.7661 | 0.9125 | 0.9574 |
| | | 2 | 0.4449 | 0.6477 | 0.6959 | 0.7034 | 0.3553 | 0.6691 | 0.8155 | 0.8604 |
| | | 3 | 0.2017 | 0.4045 | 0.4527 | 0.4602 | 0.1920 | 0.5059 | 0.6523 | 0.6972 |
| | 24 | 1 | 0.5727 | 0.8359 | 0.9219 | 0.9402 | 0.3323 | 0.6591 | 0.8652 | 0.9505 |
| | | 2 | 0.4213 | 0.6846 | 0.7705 | 0.7889 | 0.2741 | 0.6009 | 0.8070 | 0.8923 |
| | | 3 | 0.2086 | 0.4718 | 0.5578 | 0.5762 | 0.1590 | 0.4858 | 0.6918 | 0.7771 |
| | 26 | 1 | 0.4538 | 0.7676 | 0.9131 | 0.9575 | 0.2069 | 0.4992 | 0.7545 | 0.9013 |
| | | 2 | 0.3565 | 0.6703 | 0.8158 | 0.8602 | 0.1781 | 0.4704 | 0.7257 | 0.8725 |
| | | 3 | 0.1926 | 0.5064 | 0.6519 | 0.6963 | 0.1105 | 0.4029 | 0.6581 | 0.8049 |
| | 28 | 1 | 0.2900 | 0.6077 | 0.8312 | 0.9363 | 0.0936 | 0.2907 | 0.5402 | 0.7506 |
| | | 2 | 0.2423 | 0.5599 | 0.7834 | 0.8885 | 0.0837 | 0.2807 | 0.5303 | 0.7407 |
| | | 3 | 0.1433 | 0.4609 | 0.6844 | 0.7895 | 0.0556 | 0.2526 | 0.5022 | 0.7126 |
| | 29 | 1 | 0.1930 | 0.4699 | 0.7216 | 0.8770 | 0.0489 | 0.1777 | 0.3804 | 0.5949 |
| | | 2 | 0.1661 | 0.4429 | 0.6947 | 0.8501 | 0.0445 | 0.1733 | 0.3760 | 0.5905 |
| | | 3 | 0.1031 | 0.3800 | 0.6317 | 0.7872 | 0.0306 | 0.1594 | 0.3621 | 0.5766 |
| | 30 | 1 | 0.0911 | 0.2718 | 0.4982 | 0.6964 | 0.0166 | 0.0730 | 0.1877 | 0.3465 |
| | | 2 | 0.0808 | 0.2615 | 0.4879 | 0.6861 | 0.0154 | 0.0718 | 0.1864 | 0.3452 |
| | | 3 | 0.0529 | 0.2336 | 0.4601 | 0.6582 | 0.0109 | 0.0674 | 0.1820 | 0.3408 |

**Table 2: The values of $\alpha_{3(k)}\,(s,t;r,n)$ for $n = 10, 20$ and $30$ and some selected values of $s, t$ and $r$ when $k = 2$ and $k = 3$**

| $n$ | $r$ | $s$ | $k = 2$ | | | | | $k = 3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $t$ | | | | | $t$ | | | | |
| | | | 4 | 5 | 6 | 7 | 8 | 4 | 5 | 6 | 7 | 8 |
| 10 | 2 | 2 | 0.1569 | 0.1601 | 0.1607 | 0.1609 | 0.1609 | 0.0060 | 0.0156 | 0.0184 | 0.0192 | 0.0194 |
| | | 3 | 0.3220 | 0.3252 | 0.3259 | 0.3260 | 0.3260 | 0.0296 | 0.0392 | 0.0420 | 0.0428 | 0.0430 |
| | | 5 | 0.6527 | 0.6559 | 0.6565 | 0.6567 | 0.6567 | 0.0861 | 0.0957 | 0.0985 | 0.0993 | 0.0995 |
| | 4 | 2 | 0.3701 | 0.3909 | 0.3974 | 0.3992 | 0.3998 | 0.0192 | 0.0693 | 0.0905 | 0.0990 | 0.1022 |
| | | 3 | 0.6046 | 0.6253 | 0.6318 | 0.6337 | 0.6342 | 0.0976 | 0.1477 | 0.1689 | 0.1774 | 0.1806 |
| | | 5 | 0.8715 | 0.8922 | 0.8987 | 0.9006 | 0.9011 | 0.2125 | 0.2625 | 0.2838 | 0.2922 | 0.2955 |
| | 6 | 2 | 0.5279 | 0.5942 | 0.6237 | 0.6357 | 0.6403 | 0.0215 | 0.1390 | 0.2090 | 0.2475 | 0.2674 |
| | | 3 | 0.7275 | 0.7938 | 0.8232 | 0.8353 | 0.8399 | 0.1432 | 0.2606 | 0.3307 | 0.3692 | 0.3891 |
| | | 5 | 0.8627 | 0.9290 | 0.9585 | 0.9705 | 0.9752 | 0.2525 | 0.3699 | 0.4400 | 0.4785 | 0.4983 |
| | 8 | 2 | 0.5219 | 0.6563 | 0.7419 | 0.7915 | 0.8183 | 0.4357 | 0.1488 | 0.2772 | 0.3743 | 0.4427 |
| | | 3 | 0.6315 | 0.7659 | 0.8514 | 0.9010 | 0.9279 | 0.1007 | 0.2551 | 0.3835 | 0.4806 | 0.5490 |
| | | 5 | 0.6742 | 0.8086 | 0.8941 | 0.9437 | 0.9705 | 0.1562 | 0.3106 | 0.4389 | 0.5361 | 0.6044 |
| 20 | 2 | 2 | 0.0609 | 0.0612 | 0.0612 | 0.0612 | 0.0612 | 0.0012 | 0.0024 | 0.0025 | 0.0026 | 0.0026 |
| | | 3 | 0.1571 | 0.1571 | 0.1571 | 0.1571 | 0.1571 | 0.0062 | 0.0074 | 0.0076 | 0.0076 | 0.0076 |
| | | 4 | 0.2921 | 0.2925 | 0.2925 | 0.2925 | 0.2925 | 0.0146 | 0.0158 | 0.0158 | 0.0160 | 0.0160 |
| | 5 | 3 | 0.4351 | 0.4393 | 0.4400 | 0.4402 | 0.4402 | 0.0446 | 0.0583 | 0.0619 | 0.0627 | 0.0629 |
| | | 5 | 0.7902 | 0.7944 | 0.7952 | 0.7953 | 0.7954 | 0.1282 | 0.1419 | 0.1455 | 0.1463 | 0.1465 |
| | | 7 | 0.9433 | 0.9475 | 0.9483 | 0.9484 | 0.9485 | 0.1851 | 0.1989 | 0.2035 | 0.2033 | 0.2035 |
| | 7 | 3 | 0.5882 | 0.6006 | 0.6035 | 0.6041 | 0.6043 | 0.0865 | 0.1218 | 0.1336 | 0.1372 | 0.1382 |
| | | 5 | 0.8847 | 0.8970 | 0.9000 | 0.9006 | 0.9007 | 0.2058 | 0.2411 | 0.2529 | 0.2565 | 0.2575 |
| | | 9 | 0.9818 | 0.9942 | 0.9971 | 0.9977 | 0.9979 | 0.2770 | 0.3123 | 0.3241 | 0.3277 | 0.3288 |
| | 12 | 5 | 0.8730 | 0.9437 | 0.9729 | 0.9836 | 0.9871 | 0.2707 | 0.4027 | 0.4792 | 0.5185 | 0.5369 |
| | | 7 | 0.8836 | 0.9543 | 0.9835 | 0.9942 | 0.9977 | 0.2896 | 0.4216 | 0.4981 | 0.5374 | 0.5558 |
| | | 10 | 0.8843 | 0.9550 | 0.9842 | 0.9949 | 0.9984 | 0.2919 | 0.4239 | 0.5004 | 0.5397 | 0.5581 |
| | 15 | 4 | 0.8749 | 0.9456 | 0.9747 | 0.9854 | 0.9890 | 0.1641 | 0.3349 | 0.4699 | 0.5639 | 0.6232 |
| | | 8 | 0.8843 | 0.9550 | 0.9842 | 0.9949 | 0.9984 | 0.1836 | 0.3544 | 0.4894 | 0.5834 | 0.6427 |
| | | 12 | 0.8843 | 0.9550 | 0.9842 | 0.9949 | 0.9985 | 0.1837 | 0.3545 | 0.4894 | 0.5835 | 0.6428 |
| | 18 | 5 | 0.4278 | 0.5981 | 0.7388 | 0.8414 | 0.9091 | 0.0376 | 0.1602 | 0.2960 | 0.4284 | 0.5450 |
| | | 10 | 0.4279 | 0.5982 | 0.7389 | 0.8415 | 0.9093 | 0.0381 | 0.1608 | 0.2966 | 0.4289 | 0.5455 |
| | | 17 | 0.4279 | 0.5982 | 0.7389 | 0.8415 | 0.9093 | 0.0381 | 0.1608 | 0.2966 | 0.4289 | 0.5455 |
| 30 | 10 | 2 | 0.3290 | 0.3388 | 0.3408 | 0.3412 | 0.3413 | 0.0170 | 0.0469 | 0.0558 | 0.0582 | 0.0588 |
| | | 5 | 0.8889 | 0.8987 | 0.9007 | 0.9011 | 0.9011 | 0.2015 | 0.2314 | 0.2403 | 0.2427 | 0.2433 |
| | | 7 | 0.9726 | 0.9824 | 0.9844 | 0.9848 | 0.9848 | 0.2598 | 0.2897 | 0.2986 | 0.3010 | 0.3016 |
| | 15 | 5 | 0.9200 | 0.9594 | 0.9718 | 0.9753 | 0.9761 | 0.2872 | 0.3792 | 0.4210 | 0.4377 | 0.4438 |
| | | 10 | 0.9435 | 0.9829 | 0.9954 | 0.9988 | 0.9997 | 0.3243 | 0.4163 | 0.4581 | 0.4748 | 0.4809 |
| | | 12 | 0.9435 | 0.9829 | 0.9954 | 0.9988 | 0.9997 | 0.3245 | 0.4165 | 0.4583 | 0.4750 | 0.4810 |
| | 20 | 8 | 0.8240 | 0.9243 | 0.9710 | 0.9900 | 0.9968 | 0.2520 | 0.4155 | 0.5239 | 0.5865 | 0.6188 |
| | | 15 | 0.8240 | 0.9243 | 0.9710 | 0.9900 | 0.9969 | 0.2522 | 0.4156 | 0.5240 | 0.5867 | 0.6190 |
| | | 18 | 0.8240 | 0.9243 | 0.9710 | 0.9900 | 0.9969 | 0.2522 | 0.4156 | 0.5240 | 0.5867 | 0.6190 |
| | 25 | 10 | 0.5611 | 0.7324 | 0.8524 | 0.9254 | 0.9651 | 0.0860 | 0.2460 | 0.4026 | 0.5357 | 0.6366 |
| | | 15 | 0.5611 | 0.7324 | 0.8524 | 0.9254 | 0.9651 | 0.0860 | 0.2460 | 0.4026 | 0.5357 | 0.6366 |

**Table 3: The values of $\alpha_{4(k)}\,(s, n; r, n)$ for $n = 10, 20, 30, 35$ and $40$ and some selected values of $s$ and $r$ when $k = 2$ and $k = 3$**

| $n$ | $s$ | $k = 2$ | | | | | $k = 3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $r$ | | | | | $r$ | | | | |
| | | 4 | 5 | 6 | 7 | 8 | 4 | 5 | 6 | 7 | 8 |
| 10 | 2 | 0.7280 | 0.6016 | 0.4758 | 0.3633 | 0.2696 | 0.8841 | 0.8211 | 0.7369 | 0.6432 | 0.5481 |
| | 4 | 0.7022 | 0.5934 | 0.4734 | 0.3626 | 0.2694 | 0.8128 | 0.7902 | 0.7245 | 0.6384 | 0.5464 |
| | 6 | 0.6169 | 0.5537 | 0.4567 | 0.3561 | 0.2670 | 0.6337 | 0.6785 | 0.6616 | 0.6056 | 0.5302 |
| | 8 | 0.4093 | 0.4141 | 0.3732 | 0.3102 | 0.2433 | 0.3332 | 0.4150 | 0.4564 | 0.4590 | 0.4321 |
| 20 | 2 | 0.8569 | 0.7563 | 0.6402 | 0.5212 | 0.4098 | 0.9639 | 0.9286 | 0.8756 | 0.8070 | 0.7265 |
| | 4 | 0.8547 | 0.7560 | 0.6401 | 0.5212 | 0.4098 | 0.9558 | 0.9268 | 0.8752 | 0.8069 | 0.7265 |
| | 6 | 0.8478 | 0.7545 | 0.6398 | 0.5211 | 0.4098 | 0.9331 | 0.9196 | 0.8732 | 0.8064 | 0.7264 |
| | 8 | 0.8315 | 0.7497 | 0.6386 | 0.5209 | 0.4097 | 0.8860 | 0.9003 | 0.8663 | 0.8041 | 0.7257 |
| | 10 | 0.7992 | 0.7379 | 0.6349 | 0.5158 | 0.4094 | 0.8057 | 0.8587 | 0.8475 | 0.7965 | 0.7229 |
| | 15 | 0.5808 | 0.6128 | 0.5725 | 0.4918 | 0.3978 | 0.4346 | 0.5691 | 0.6508 | 0.6761 | 0.6550 |
| | 18 | 0.2852 | 0.3546 | 0.3791 | 0.3627 | 0.3191 | 0.1463 | 0.2325 | 0.3151 | 0.3787 | 0.4149 |
| 30 | 2 | 0.9068 | 0.8272 | 0.7262 | 0.6135 | 0.4998 | 0.9829 | 0.9618 | 0.9268 | 0.8766 | 0.8122 |
| | 4 | 0.9063 | 0.8272 | 0.7262 | 0.6135 | 0.4998 | 0.9809 | 0.9615 | 0.9267 | 0.8766 | 0.8122 |
| | 6 | 0.9048 | 0.8270 | 0.7262 | 0.6135 | 0.4998 | 0.9753 | 0.9603 | 0.9265 | 0.8766 | 0.8122 |
| | 8 | 0.9013 | 0.8263 | 0.7261 | 0.6134 | 0.4997 | 0.9631 | 0.9570 | 0.9257 | 0.8764 | 0.8121 |
| | 10 | 0.8946 | 0.8248 | 0.7257 | 0.6134 | 0.4997 | 0.9413 | 0.9498 | 0.9236 | 0.8759 | 0.8120 |
| | 15 | 0.8504 | 0.8102 | 0.7216 | 0.6123 | 0.4995 | 0.8239 | 0.8945 | 0.9012 | 0.8677 | 0.8093 |
| | 20 | 0.7307 | 0.7515 | 0.6971 | 0.6033 | 0.4965 | 0.5902 | 0.7323 | 0.8056 | 0.8181 | 0.7860 |
| | 25 | 0.4680 | 0.5596 | 0.5786 | 0.5389 | 0.4649 | 0.2628 | 0.4014 | 0.5230 | 0.6059 | 0.6423 |
| 35 | 2 | 0.9215 | 0.8499 | 0.7556 | 0.6470 | 0.5341 | 0.9873 | 0.9704 | 0.9411 | 0.8976 | 0.8397 |
| | 4 | 0.9213 | 0.8499 | 0.7556 | 0.6470 | 0.5341 | 0.9862 | 0.9702 | 0.9411 | 0.8976 | 0.8397 |
| | 6 | 0.9204 | 0.8498 | 0.7556 | 0.6470 | 0.5341 | 0.9829 | 0.9696 | 0.9410 | 0.8976 | 0.8397 |
| | 8 | 0.9185 | 0.8495 | 0.7556 | 0.6470 | 0.5341 | 0.9758 | 0.9680 | 0.9407 | 0.8975 | 0.8397 |
| | 10 | 0.9148 | 0.8488 | 0.7555 | 0.6469 | 0.5341 | 0.9630 | 0.9644 | 0.9398 | 0.8973 | 0.8397 |
| | 15 | 0.8909 | 0.8422 | 0.7539 | 0.6466 | 0.5341 | 0.8920 | 0.9364 | 0.9303 | 0.8945 | 0.8389 |
| | 25 | 0.6871 | 0.7394 | 0.7093 | 0.6294 | 0.5280 | 0.5020 | 0.6622 | 0.7642 | 0.8047 | 0.7946 |
| | 30 | 0.4219 | 0.5289 | 0.5689 | 0.5476 | 0.4852 | 0.2107 | 0.3405 | 0.4663 | 0.5639 | 0.6202 |
| 40 | 2 | 0.9327 | 0.8679 | 0.7795 | 0.6750 | 0.5636 | 0.9998 | 0.9767 | 0.9515 | 0.9135 | 0.8613 |
| | 4 | 0.9323 | 0.8679 | 0.7795 | 0.6750 | 0.5636 | 0.9998 | 0.9767 | 0.9514 | 0.9135 | 0.8613 |
| | 6 | 0.9318 | 0.8678 | 0.7795 | 0.6750 | 0.5636 | 0.9998 | 0.9762 | 0.9514 | 0.9135 | 0.8613 |
| | 8 | 0.9308 | 0.8676 | 0.7795 | 0.6750 | 0.5636 | 0.9965 | 0.9753 | 0.9512 | 0.9135 | 0.8613 |
| | 10 | 0.9288 | 0.8672 | 0.7795 | 0.6750 | 0.5636 | 0.9887 | 0.9733 | 0.9508 | 0.9134 | 0.8612 |
| | 15 | 0.9142 | 0.8639 | 0.7788 | 0.6748 | 0.5636 | 0.9458 | 0.9582 | 0.9464 | 0.9122 | 0.8610 |
| | 20 | 0.8771 | 0.8516 | 0.7754 | 0.6740 | 0.5634 | 0.8442 | 0.9108 | 0.9273 | 0.9055 | 0.8588 |
| | 25 | 0.7965 | 0.8152 | 0.7615 | 0.6694 | 0.5621 | 0.6762 | 0.8014 | 0.8677 | 0.8772 | 0.8468 |
| | 30 | 0.6437 | 0.7216 | 0.7137 | 0.6482 | 0.5537 | 0.4360 | 0.5966 | 0.7189 | 0.7830 | 0.7937 |
| | 35 | 0.3823 | 0.4991 | 0.5557 | 0.5506 | 0.4999 | 0.1718 | 0.2916 | 0.4166 | 0.5231 | 0.5943 |

**Table 4: Values of $\alpha_{7(k)}\,(s, t; 1)$ for $1 \leq s \leq 7$ and $4 \leq t \leq 7$**

| $s$ | $k = 2$ | | | | $k = 3$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $t$ | | | | $t$ | | | |
| | 4 | 5 | 6 | 7 | 4 | 5 | 6 | 7 |
| 1 | 0.1358 | 0.2016 | 0.2455 | 0.2748 | 0.0035 | 0.0527 | 0.1121 | 0.1565 |
| 2 | 0.3580 | 0.4239 | 0.4678 | 0.4970 | 0.1911 | 0.2402 | 0.2995 | 0.3440 |
| 3 | 0.5062 | 0.5720 | 0.6159 | 0.6452 | 0.3317 | 0.3808 | 0.4401 | 0.4846 |
| 4 | 0.6049 | 0.6707 | 0.7146 | 0.7339 | 0.4372 | 0.4862 | 0.5456 | 0.5901 |
| 5 | 0.6708 | 0.7366 | 0.7805 | 0.8098 | 0.5163 | 0.5654 | 0.6247 | 0.6692 |
| 6 | 0.7147 | 0.7805 | 0.8244 | 0.8537 | 0.5756 | 0.6247 | 0.6840 | 0.7285 |
| 7 | 0.7439 | 0.8098 | 0.8537 | 0.8829 | 0.6201 | 0.6692 | 0.7285 | 0.7730 |

**Table 5: Values of $\alpha_{7(k)}\,(s, t; 2)$ for $1 \leq s \leq 7$ and $4 \leq t \leq 7$**

| $s$ | $k = 2$ | | | | $k = 3$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $t$ | | | | $t$ | | | |
| | 4 | 5 | 6 | 7 | 4 | 5 | 6 | 7 |
| 1 | 0.1502 | 0.2599 | 0.3477 | 0.4160 | 0.0011 | 0.0018 | 0.0759 | 0.1538 |
| 2 | 0.3684 | 0.4781 | 0.5659 | 0.6342 | 0.1126 | 0.2115 | 0.3005 | 0.3784 |
| 3 | 0.4605 | 0.5702 | 0.6580 | 0.7263 | 0.2262 | 0.3251 | 0.4141 | 0.4919 |
| 4 | 0.5019 | 0.6117 | 0.6994 | 0.7678 | 0.2873 | 0.3862 | 0.4752 | 0.5530 |
| 5 | 0.5213 | 0.6310 | 0.7188 | 0.7871 | 0.3213 | 0.4201 | 0.5091 | 0.5870 |
| 6 | 0.5304 | 0.6401 | 0.7279 | 0.7963 | 0.3405 | 0.4393 | 0.5284 | 0.6062 |
| 7 | 0.5349 | 0.6446 | 0.7323 | 0.8006 | 0.3516 | 0.4505 | 0.5395 | 0.6173 |

## References

Ahmadi, J. and Balakrishnan, N. (2005). Distribution-free confidence intervals for quantile intervals based on current records. *Statistics and Probability Letters*, **75**, 190-202.

Ahmadi, J. and Balakrishnan, N. (2008). Nonparametric confidence intervals for quantile intervals and quantile differences based on record statistics. *Statistics and Probability Letters*, **78**, 1236-1245.

Ahmadi, J. and Balakrishnan, N. (2010). Prediction of order statistics and record values from two independent sequences. *Statistics*, **44**, 417-430.

Ahmadi, J., MirMostafaee, S. M. T. K. and Balakrishnan, N. (2010). Nonparametric prediction intervals for future record intervals based on order statistics. *Statistics and Probability Letters*, **80**, 1663-1672.

Ahmadi, J. and Balakrishnan, N. (2011). Distribution-free prediction intervals for order statistics based on record coverage. *Journal of the Korean Statistical Society*, **40**, 181-192.

Ahsanullah, M.(1995). *Record Statistics.* Nova Science Publishers, New York.

Al-Hussaini, E. K. and Ahmad, A. E. B. A. (2003). On Bayesian interval prediction of future records. *Test*, **12**, 79-99.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records.* John Wiley and Sons, New York.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics.* USA: Classics in Applied Mathematics.

Chacko, M. and Mary, M. S. (2013). Estimation and prediction based on k-record values from normal distribution. *Statistica*, **73**, 505-516.

Chacko, M. and Muraleedharan, L. (2018). Inference based on k-record values from generalized exponential distribution. *Statistica*, **78**, 37-56.

Chandler, K. (1952). The distribution and frequency of record values. *Journal of the Royal Statistical Society: Series B (Methodological)*, **14**, 220-228.

David, H. A. and Nagaraja, H.N. (2003). *Order Statistics.* Third Edition. New York: John Wiley & Sons.

Deheuvels, P. and Nevzorov, V. B. (1994). Limit laws for k-record times. *Journal of Statistical Planning and Inference*, **38**, 279-307.

Dziubdziela, W. and Kopociński, B. (1976). Limiting properties of the k-th record values. *Applicationes Mathematicae*, **2**, 187-190.

Glick, N. (1978). Breaking records and breaking boards. *The American Mathematical Monthly*, **85**, 2-26.

Guilbaud, O. (2004). Exact non-parametric confidence, prediction and tolerance intervals with progressive type-II censoring. *Scandinavian Journal of Statistics*, **31**, 265-281.

Gulati, S. and Padgett, W. J. (1994). Smooth nonparametric estimation of the distribution and density functions from record-breaking data. *Communications in Statistics-Theory and Methods*, **23**, 1259-1274.

Hsieh, H. K. (1997). Prediction intervals for Weibull order statistics. *Statistica Sinica*, **7**, 1039-1051.

Malinowska, I. and Szynal, D. (2004). On a family of Bayesian estimators and predictors for a Gumbel model based on the kth lower records. *Applicationes Mathematicae*, **1**, 107-115.

Minimol, S. and Thomas, P. Y. (2014). On characterization of Gompertz distribution by properties of generalized record values. *Journal of Statistical Theory and Applications*, **13**, 38-45.

Muraleedharan, L. and Chacko, M. (2019). Inference on Gompertz distribution based upper k- record values. *Journal of the Kerala Statistical Association,* **30**, 47-63.

Muraleedharan, L. and Chacko, M. (2022). Estimation and prediction based on k-record values from a generalized Pareto distribution. *International Journal of Mathematics and Statistics*, **23**, 59-75.

Nagaraja, H. N. (1988). Record values and related statistics-a review. *Communications in Statistics-Theory and Methods*, **17**, 2223-2238.

Paul, J. and Thomas, P. Y. (2015). On generalized upper (k) record values from Weibull distribution. *Statistica*, **75**, 313-330.

Statistics and Applications {ISSN 2454-7395(Online)}
Volume 21, No. 1, 2023 (New Series), pp 99-112

# Empirical Mode Decomposition Based Ensemble Hybrid Machine Learning Models for Agricultural Commodity Price Forecasting

**Pankaj Das[1], Girish Kumar Jha[2] and Achal Lama[1]**
[1]*ICAR- Indian Agricultural Statistics Research Institute, New Delhi, India*
[2]*ICAR- Indian Agricultural Research Institute, New Delhi, India*

Received: 12 April 2022; Revised: 20 May 2022; Accepted: 09 June 2022

## Abstract

Agricultural commodity price is very volatile in nature due to its nonlinearity and nonstationary character. The volatility behaviour of the commodity price creates a lot of problems for both producer and consumer. The steady forecast of the price may reduce the problems and increase the profit for the stakeholders. In this study, an ensemble hybrid machine learning model based on empirical mode decomposition (EMD) has been proposed to forecast the commodity price. EMD decomposes the nonstationary and nonlinear price series into different stationary intrinsic mode functions (IMF) and a final residue. Then Machine learning techniques like Artificial neural network (ANN) and Support vector regression (SVR) were used to forecast each of the decomposed components. Finally, all the forecasted values of the decomposed components were aggregated to produce the final forecast. Two R modules were developed for the application of the proposed methodology. The proposed methodology has been applied to the monthly wholesale price index of vegetables. The results indicated that the ensemble hybrid machine learning model based on empirical mode decomposition has superior performance compared to generic models.

*Key words*: Agricultural commodity price; Machine learning; Empirical mode decomposition; Nonlinearity; Nonstationary; Artificial neural network; Support vector regression.

## 1. Introduction

The scientific and effective forecasting method is helpful to correctly guide producers and policy makers to match the supply and demand of the agricultural production and facilitate the decision-making process of the government. Agricultural price forecasting is not an easy task due to its dependency on many extraneous factors. Nonlinearity and nonstationary behaviour of data series are crucial problems in the agricultural price forecasting. Agricultural commodity prices are volatile in nature due to seasonality, inelastic demand, production uncertainty *etc*. Traditionally, time series forecasting has been dominated by linear methods like ARIMA (Box and Jenkins, 1970) and nonlinear models such as SETAR, STAR, *etc*. because they are well understood and effective in many situations. These traditional methods suffer from some limitations, such as linear models focusing on linear relationships, fixed temporal dependence *etc*. and nonlinear models

Corresponding Author: Pankaj Das
Email: pankaj.iasri@gmail.com

require the specific nonlinear relation of data generating process to be known a priori. Zhang *et al*. (1998) demonstrated that traditional methods requiring the time series data to be stable or stable after being differentiated. On the other hand, Machine learning (ML) models with their flexible functional designs and powerful self-learning capabilities have recently become a great alternative for time series data forecasting. Machine learning techniques like Artificial Neural Network (ANN) and Support Vector Machine (SVM) become popular to handle the nonlinearity problem in the dataset (Qin and Chiang, 2019). Darbellay and Slama (2000) highlighted that ANN which is non-linear, nonparametric and data driven self-adaptive method, is most suitable for forecasting agricultural price series which is inherently noisy and nonlinear in nature. Levis and Papageorgiou (2005) clarified the numerous advantages of SVR for nonlinear times series forecasting. Lu *et al*. (2009) demonstrated the generalization characteristics of SVR for finding a unique solution. An *et al.* (2012) reported that empirical mode decomposition (EMD) can reveal the hidden pattern and trends of time series which can effectively assist in designing forecasting models for various applications. Sugiyama and Kawanabe (2012) stated the inability of ML techniques to counter the nonstationarity behaviour of a dataset. Huang *et al.* (1998) pointed out how EMD is capable to deal with the problem of inherent nonstationarity or nonlinearity in a time series dataset. EMD has the power to isolate the high fluctuating data into respective smaller frequency components (Mumtaz *et al*., 2019). EMD also has the capability to reduce the influence of nonlinear characteristics of the stock series (Xuan *et al*., 2020)

It has been observed in the literature that a single model is not sufficient to deal with complex real-world systems such as agricultural price data which contains unknown mixed patterns. Besides, inherent non-stationarity and nonlinearity behaviour of price series create problems in robust forecasting (Taylor and Kingsman, 1978). ML algorithms are compatible and efficient to deal with nonlinear problems (Bishop, 1995). To handle non-stationarity features, the inputs for machine learning models need to be properly pre-processed (Wang *et al*., 2017). Therefore, some multi-resolution analysis techniques are widely used in many forecasting problems. In view of these factors, there is an ever-increasing need of using hybrid models to improve the accuracy of predictions. In order to improve forecast accuracy, hybridization is a good idea because it can capture various patterns in the data concurrently. Hybrid models, combining the benefits of different models, are suggested to achieve better prediction and the decomposition approaches such as EMD enhance the performance of hybrid models (Mo *et al*., 2020). These studies led to the development of novel hybrid models which are more robust as they often compliment the advantages of the individual technique involved and improve the forecasting accuracy. In this study, novel hybrid models have been proposed by combining EMD and ML algorithms like ANN and SVR to deal with nonlinearity and non-stationarity problems in a time series data. EMD counters with non-stationarity of a time series data by decomposing into several stationary components which is nonlinear in nature. Further ANN and SVR models have been used to forecast these nonlinear decomposed components. We have used monthly wholesale price index of vegetables for practical evaluation and compared the proposed model with other models, including hybrid and individual models.

The remaining portion of the paper is organized as follows. Section 2 deals with the methodology. The data and results of the experiment are explained in the third section. The final section concludes the paper.

## 2. Materials and methods

### 2.1. Empirical mode decomposition (EMD)

The empirical mode decomposition method was introduced by Huang *et al.* in 1998. It assumes that the data have many coexisting oscillatory modes of significantly distinct frequencies and these modes superimpose on each other and form an observable time series. EMD decomposes original nonstationary and nonlinear data into a finite and small number of independent sub-series (including intrinsic mode functions and a final residue). Intrinsic Mode Function (IMF) is the finite additive oscillatory component decomposed by EMD. For example, let $y_t$ be a time series (TS) dataset at time *t* consisting of high frequency part and low frequency part. After first decomposition, original time series data results into first IMF and the residue. The decomposed TS takes the following form:

$$y_t = d_t(1) + r_t(1) \tag{1}$$

where $d_t(1)$ = high frequency part *i.e.* IMF and $r_t(1)$ = low frequency part *i.e.* residue. EMD algorithm iterates over the slow oscillation component considered as a new signal. In the next iteration, the residue $r_t(1)$ will be treated as new signal for EMD decomposition.

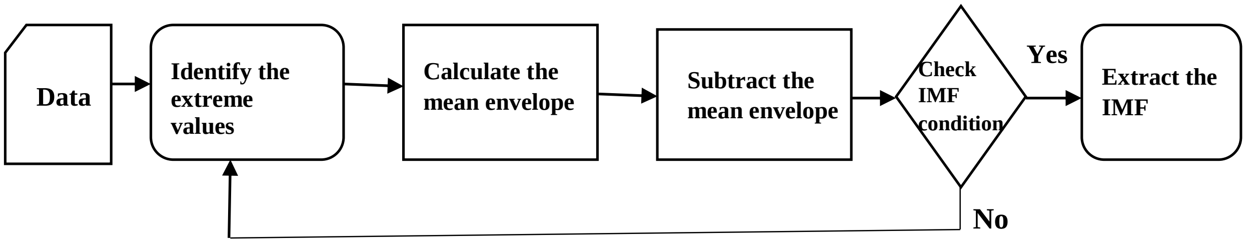
After second decomposition,

$$r_t(1) = d_t(2) + r_t(2)$$

Let us assume that $y_t$ has been decomposed into *k* numbers of IMFs and final residues after the EMD decomposition process completes. Then, $y_t$ can be expressed as follows:

$$y_t = \sum_{i=1}^{k} d_t(i) + R \tag{2}$$

Hence, original time series data = sum of IMFs + final residue (*R*). The stepwise EMD algorithm procedure is mentioned below:


Data
Identify the extreme values
Calculate the mean envelope
Subtract the mean envelope
Check IMF condition
Yes
Extract the IMF
No

All these steps come under the first iteration in the shifting process for $y_t$. The shifting process continues till we obtain an IMF. The point of termination of a sifting process is called stopping point $k$ and the iteration is called $k^{th}$ iteration. The stopping criteria of an iteration is the standard deviation ($\sigma$) between two IMF (*d*) values. The predefined threshold value ranges between 0.2 to 0.3 (Huang *et al.*, 1998). When the $\sigma$ value lies between 0.2 to 0.3, the iteration should stop. The threshold value is calculated using the given expression:

$$\text{Threshold value} = \sum_{l=0}^{T}\left[\frac{(d_t(l(k-1)-d_t(lk))^2}{d_t^2(l(k-1)}\right] \tag{3}$$

## 2.2. Artificial neural network (ANN) model

Artificial neural network is a non-linear, data driven self-adaptive approach as opposed to the traditional model based methods (Jha and Sinha, 2014). ANN can identify and learn correlated patterns between input data sets and corresponding target values. ANN imitates the learning process of the human brain and can process problems involving non-linear and complex data even if the data are imprecise and noisy. Thus, it is ideally suited for modelling of agricultural data which are known to be complex and often non-linear. Haykin (1999) stated mathematically that a neuron $k$ can be defined by the following equations:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{4}$$

$$y_k = \varphi(u_k + b_k) \tag{5}$$

Here bias ($b_k$), has the effect of increasing or lowering the net input of the activation function. $x_1, x_2, ..., x_m$ are the inputs; $w_{k1}, w_{k2,}..., w_{km}$ are the weights of the neuron $k$; $u_k$ is the linear combiner output due to input variables; $\varphi(.)$ is the activation function; $y_k$ is the output of the neuron, $w_{kj}$ the weight attached to the connection from $j^{th}$ hidden node to the output node. The backpropagation algorithm can be implemented under the following components:

1. Data should contain input-output pair $(\vec{x}_i, \vec{y}_i)$, where $\vec{x}_i = \{y_{t-d}, y_{t-d+1}, ..., y_{t-1}\}$ the input, $d$ is a user-defined parameter, which corresponds to the number of previous time-steps and $\vec{y}_i = y_t$ is the desired output. For $T$ data of $X = \{(\vec{x}_1, \vec{y}_1), ..., (\vec{x}_T, \vec{y}_T)\}$ .
2. Need a feedforward neural network. Let the parameters of the network be denoted by $\theta$. The parameters of interest in backpropagation are the weights $w_{ij}^k$, node $j$ in layer $l_k$ and node $i$ in layer $l_{k-1}$ and bias $b_i^k$ the bias for node $i$ in layer $l_k$.

Error function $E(X,\theta) = \frac{1}{2T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2$ ; where $\hat{y}_t$ are the computed output of the network on input $\vec{x}_t$ and $y_t$ is the target value for input-output pair $(\vec{x}_i, \vec{y}_i)$.

In the present study, Logistic function was used as an activation function and resilient backpropagation algorithm was used to adjust the weights in the multi-layered feed-forward network.

## 2.3. Support vector regression (SVR) model

Vapnik (1998) introduced support vector regression model by incorporating a loss function. SVR fits linear regression in the outer space through mapping input vectors into a high dimensional space. In the present study, a modified SVR model named least squares support vector regression (LS-SVR) proposed by Suykens *et al.* (2002) has been used. LS-SVR model focuses on set of linear equations instead of a quadratic programming problem in SVR model.

LS-SVR model is represented as:

$$y = w^T \varphi(x) + b \tag{6}$$

with $x \in R^T$ and $\varphi$, mapping function $R^n \to R^{n_t}$ to high dimensional feature space, bias *b* and error *e*. For a given training set $\{x_t, y_t\}_{t=1}^{T}$, optimization problem becomes as follows:

$$\min_{\{w,e,b\}} J(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{t=1}^{T} e_t^2 \tag{7}$$

where $e_t = y_t - w^t \varphi(x_t) + b$ is the fitting errors, subject to equality constraints

$$y_t = w^T \varphi(x_t) + b + e_t\,;\ t= 1, 2, ..., T. \tag{8}$$

This is a form of ridge regression. Now incorporating Lagrange multiplier $\alpha_t$

$$L(w,b,e;\alpha) = J(w,e) - \sum_{t=1}^{T} \alpha_t \{w^T \varphi(x_t) + b + e_t - y_t\} \tag{9}$$

with following conditions of optimality

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \to w = \sum_{t=1}^{T} \alpha_t \varphi(x_t) \\ \dfrac{\partial L}{\partial b} = 0 \to \sum_{t=1}^{T} \alpha_t = 0 \\ \dfrac{\partial L}{\partial e_t} = 0 \to \alpha_t = \gamma e_t, \qquad t = 1,2,...,T \\ \dfrac{\partial L}{\partial \alpha_t} = 0 \to w^T \varphi(x_t) + b + e_t - y_t = 0, \quad t = 1,2,...,T \end{cases} \tag{10}$$

$$\text{Solution will be } \begin{bmatrix} 0 & \vec{1}^T \\ 1 & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{11}$$

with $y = [y_1, ..., y_T], \vec{1} = [1, ..., 1],\ \alpha = [\alpha_1, ..., \alpha_T]$. By applying Mercer's condition

$$\Omega_{tu} = \Omega(\alpha_t, \alpha_u) = \sum_{t=1}^{T} d_t (\alpha_t - \alpha_t^{'}) - e \sum_{t=1}^{T} (\alpha_t - \alpha_t^{'}) - \frac{1}{2} \sum_{t=1}^{T} \sum_{u=1}^{T} (\alpha_t - \alpha_t^{'})(\alpha_u - \alpha_u^{'}) K(x_t, x_u) \tag{12}$$

where $d$ = scalar vector of $x$, $K(x_t, x_u)$ is the kernel function.

The final LS-SVR model can be written as:

$$y(x) = \sum_{t=1}^{T} \alpha_t K(x_t, x) + b \tag{13}$$

In the present study, least squares SVR model with Radial basis function (RBF) kernel was used for nonlinear mapping of dataset.

## 2.4. Proposed ensemble hybrid model

Ensemble method is a machine learning approach which combines multiple base models to produce an optimal predictive model. The proposed EMD-SVR/ANN consists of three steps depicted in Figure 1. In the first step, original nonlinear and nonstationary dataset is decomposed into a finite and often small numbers of independent sub-series by EMD technique. This sub-series contain $k$ intrinsic mode functions (IMFs) and a final residue. Secondly, these IMFs and residue is modelled and predicted through ANN or SVR. Then, all the forecasted values of the IMFs and residue are summed up to produce ensemble forecast for the original series. The prediction model of ANN (Equation 5) and SVR (Equation 13) were used with decomposed components as inputs. The input lags were selected based on Akaike information criterion (AIC) and Bayesian information criterion (BIC).

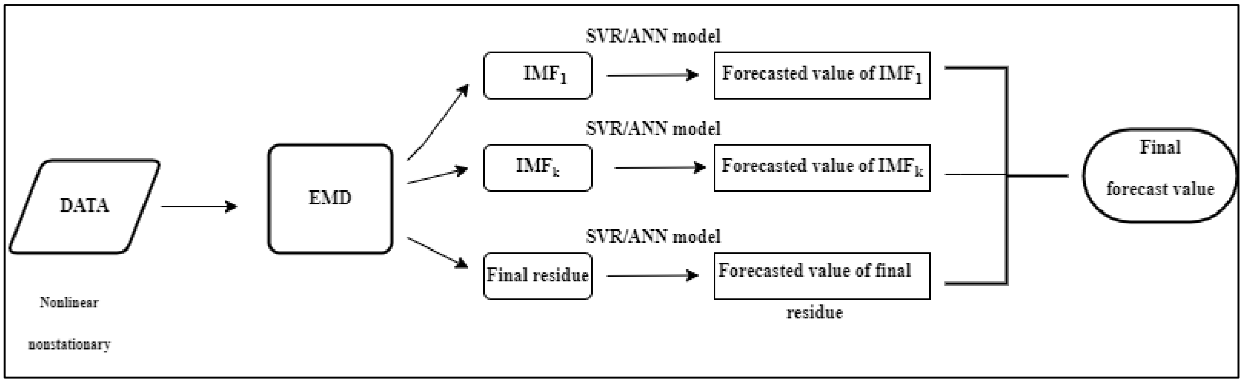


**Figure 1: EMD based ensemble hybrid machine learning model for a dataset**

## 2.5. Assessment of the fitted models

The fitted models were assessed using the performance measures like root mean squared error (RMSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE) and maximum error (ME). These performance measures can be expressed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T}} \qquad \text{MAD} = \frac{\sum_{t=1}^{T}|y_t - \hat{y}_t|}{T}$$

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{y_t} \qquad \text{ME} = \max\sum_{t=1}^{T}|y_t - \hat{y}_t|$$

where $y_t$ and $\hat{y}_t$ are the actual value and predicted value of response variable and $T$ is the number of data points. For comparing the forecasting performance, the DM (Diebold and Mariano, 1995) test was used.

## 3. Results and discussion

The complete analysis of the present study was done using RStudio. For analysis, two R packages *i.e. EMDANNhybrid* (Das *et al*., 2021) and *EMDSVRhybrid* (Das *et al*., 2021) have been developed and are available in CRAN repository. Considering the time dependency of observations, the data was divided into training and testing set for all models.

### 3.1. Data source

In the present study, monthly price index of vegetables was used to evaluate the performance of the proposed EMD-ANN and EMD-SVR models. The dataset was obtained from the Office of the Economic Advisor, Ministry of Commerce, Government of India (*https://eaindustry.nic.in/*). Figure 2 illustrates the monthly data of wholesale price index (WPI) of vegetables (January 2005 to November 2020) containing 191 data points. The descriptive statistics, stationarity test and normality of sample data were presented in Table 1. The statistics obtained through Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) test were insignificant *i.e*. null hypothesis of the unit root test cannot be refused. It indicated that the given dataset was nonstationary. Jarque-Bera test (Table 1) indicated the nonnormality of data.

**Table 1: The descriptive statistics, stationarity and normality tests of data**

| Descriptive statistics | | Stationarity test | | Normality test |
|---|---|---|---|---|
| Numbers of observations | 191 | Augmented Dickey-Fuller test (*p* value) | Phillips-Perron test (*p* value) | Jarque-Bera test (*p* value) |
| Maximum | 284.90 | 0.23 | 0.30 | <0.01 |
| Minimum | 43.70 | | | |
| Mean | 121.97 | | | |
| Standard deviation | 52.17 | | | |
| Skewness | 0.64 | | | |
| Kurtosis | -0.12 | | | |

Brock-Dechert-Scheinkman (Brock *et al*., 1996) test was used in the dataset for checking the nonlinearity of data. The results of the BDS test (Table 2) indicated that the test statistics were far bigger than the critical values. It provided a piece of evidence to reject the null hypothesis that the price series is linearly dependent. The results obtained from various tests revealed that the monthly vegetables WPI dataset is nonlinear and nonstationary in nature. These characteristics of the dataset enabled us to implement and evaluate the performance of the proposed EMD-ANN/SVR models with existing individual models.

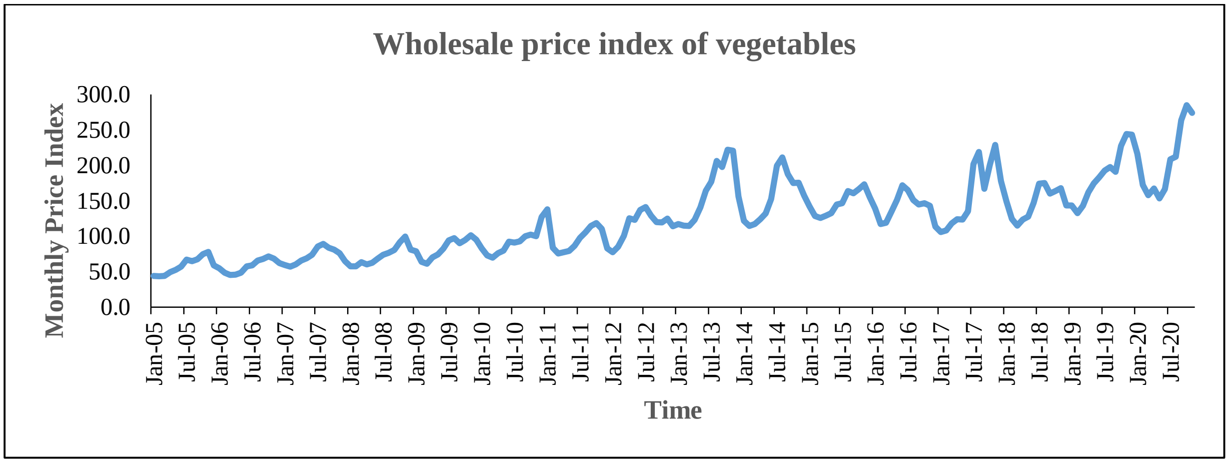


**Figure 2: Time plot of vegetables monthly price index (2004-05=100)**

**Table 2: Results of Brock- Dechert-Scheinkman (BDS) test**

| Embedding dimension | | | | Conclusion |
|---|---|---|---|---|
| 2 | | 3 | | |
| Statistics | Probability | Statistics | Probability | |
| 102.75 | < 0.001 | 172.21 | < 0.001 | Nonlinear |
| 54.88 | < 0.001 | 68.12 | < 0.001 | |
| 40.38 | < 0.001 | 42.96 | < 0.001 | |
| 37.68 | < 0.001 | 37.17 | < 0.001 | |

### 3.2. EMD decomposition

The EMD, as an adaptive decomposition technique is quite effective in extracting characteristic information from nonlinear and nonstationary time series. EMD methodology has been employed to decompose the series. Firstly, EMD algorithm finds the local extreme values (*i.e*. maxima and minima) from the dataset. Local extrema are the points where slope sign is changed. The envelopes are the curve that passing through the local extrema. The curves passing through the local minima and local maxima are known as upper envelope and lower envelope respectively. Mean envelope are the curves that is passing through mean values of local maxima and minima. The whole process of EMD decomposition has been visualized in Figure A.1. The figure gives an idea of how the EMD algorithm finds the envelopes and residue from a dataset. In the given figure red line and blue line indicates the upper and lower envelope respectively. The mean envelope and the residue are denoted using black and green lines in Figure A.1. The original series has been decomposed into four IMFs and one final residue using EMD (Figure A.2). It has been observed that the frequencies and amplitudes of IMFs were different and independent from each other. Thus, the different hidden oscillatory modes in the original datasets were separated by EMD. Each decomposed IMF contains certain characteristics of the dataset which needs to be modelled and forecasted using appropriate model.

After decomposition, it is pertinent to check the stationarity of IMFs and residue. The results (Table 3) of the test indicated that all IMFs were stationary, but the final residue was nonstationary because it contains the remaining portion of the data that cannot be decomposed by the EMD algorithm. As stationarity is one of the important assumptions for forecasting, hence, the nonstationary residue as such cannot be used in forecasting. The residue was transformed into stationary by first differencing. The stationary decomposed parts *i.e.* IMFs and the differenced residue were used for forecasting.

**Table 3: Unit root test of decomposed components of vegetables WPI dataset**

| Components | Augmented Dickey-Fuller test (*p*-value) | Augmented Dickey-Fuller test (p-value) | Remarks |
|---|---|---|---|
| $IMF_1$ | <0.01 | <0.01 | Stationary |
| $IMF_2$ | <0.01 | <0.01 | Stationary |
| $IMF_3$ | <0.01 | <0.01 | Stationary |
| $IMF_4$ | <0.01 | <0.01 | Stationary |
| Residue | 0.14 | 0.14 | Nonstationary |

### 3.3. ANN training and forecasting

ANN model was employed to different IMFs and final residue for forecasting purpose as it is capable of handling nonlinear and complex data. For this purpose, the *EMDANNhybrid* R-Package has been developed and used for analysis. Backpropagation training algorithm was used for ANN fitting. In practice, ANN with a small number of parameters namely input lags and hidden nodes often perform better for out of sample forecasting. This may be because over-fitting is a common problem in case of neural network models with a large number of parameters. In this study, we varied input lags and hidden nodes from one to five. ANN model with three input lags and four hidden nodes was found to be the most suitable model for the given dataset in terms of accuracy criterion. The other parameters like the maximum number of iterations for the neural network was fixed at 200. We averaged the results of 26 neural networks for getting the final output. The number of neural networks to be averaged was selected based on the minimum error criterion. We tried averaging 10 to 50 neural networks and obtained the best result on 26 number of neural networks. In our study, 80% of data as training set and the remaining 20% as testing set were used.

### 3.4. SVR training and forecasting

Similarly, the SVR model was also fitted to different IMFs and final residue. The each component (IMFs and residue) was modelled and forecasted using the *EMDSVRhybrid* R-package, developed under this study. The SVR model was preferred over other machine learning algorithms due to its capability to handle nonlinear systems as well as its suitability for a small sample size.

For employing the SVR model, we divided the dataset into training and testing sets. The training set is used for model building purposes whereas, the testing set allows us to understand the generalization ability of the developed model. In this study, we have used 80% of the data as a training set and the remaining 20% as a testing set. The developed SVR model for each decomposed component (IMFs and residue) was used to forecast the respective components. Then all the forecasted values of IMFs and residue were summed up to get an ensemble forecast of the data. Radial Basis Function (RBF), polynomial, linear and sigmoid kernel functions were implemented in SVR model fitting. The best result was obtained using the RBF kernel function. To overcome the problem of overfitting, 10-fold cross-validation was also done.

### 3.5. Performance comparison of fitted models

The performance (in-sample and out-sample) of the EMD-ANN and EMD-SVR model was compared with the individual ANN and SVR model (Table 4 and 5) with forecasting horizon of six months. Both the in-sample and out-sample performance of EMD-SVR was relatively superior as compared to other competing models. EMD based ANN and SVR models outperformed the individual models like ANN and SVR. The reason behind the poor performance of singular ANN and SVR models can be attributed mainly to the fact that these models could not handle the nonstationary behaviour of the given dataset. On the other hand, the hybrid models EMD-ANN and EMD-SVR performed better due to the ability to capture both nonlinearity and nonstationarity patterns of the dataset.

**Table 4: In-sample performance of fitted models**

| | ANN | SVR | EMD-ANN | EMD-SVR |
|---|---|---|---|---|
| RMSE | 10.68 | 28.25 | 5.13 | 3.15 |
| MAPE | 5.22 | 0.21 | 0.04 | 0.03 |
| MAD | 6.67 | 20.84 | 4.36 | 2.71 |
| ME | 0.87 | 98.19 | 19.63 | 15.76 |

**Table 5: Out-sample performance of fitted models**

| | ANN | SVR | EMD-ANN | EMD-SVR |
|---|---|---|---|---|
| RMSE | 44.39 | 54.39 | 25.36 | 23.69 |
| MAPE | 0.26 | 0.24 | 0.15 | 0.10 |
| MAD | 41.33 | 44.80 | 23.12 | 17.12 |
| ME | 69.81 | 115.83 | 67.23 | 68.74 |

Further DM test was employed to assess the accuracy of the EMD-SVR model compared to the EMD-ANN model. The null hypothesis of the DM test was that both models have the same accuracy. Results of Table 6 clearly indicated that the EMD-SVR model was superior to the EMD-ANN model in terms of all the criteria. The test also indicated that the forecasting performance of both EMD-ANN and EMD-SVR gave better results compared to the individual ANN and SVR models. The novelty of the proposed ensemble approach is that it can handle nonlinear and nonstationary data which is difficult for the existing time series methods. Our empirical findings suggest that the proposed EMD-SVR and EMD-ANN models can be considered as an alternative tool for volatile agricultural price series forecasting.

**Table 6: Results of Diebold-Mariano (DM) test**

| Hypothesis | DM value | *p* value | Remarks |
|---|---|---|---|
| $H_0$: The accuracy of both EMD-SVR and EMD-ANN is same.<br>$H_1$: The accuracy of EMD-SVR is superior to EMD-ANN. | 5.48 | <0.01 | The accuracy of EMD-SVR is superior to EMD-ANN. |
| $H_0$: The accuracy of both EMD-SVR and SVR is same.<br>$H_1$: The accuracy of EMD-SVR is superior to SVR. | 3.02 | <0.01 | The accuracy of EMD-SVR is superior to SVR. |
| $H_0$: The accuracy of both EMD-ANN and ANN is same.<br>$H_1$: The accuracy of EMD-ANN is superior ANN. | 6.23 | <0.01 | The accuracy of EMD-ANN is superior ANN. |

## 4. Conclusion

In this study, EMD based hybrid machine learning models for forecasting have been proposed to deal with inherent nonlinearity and non-stationarity behaviour in a time series dataset. Single models fail to capture both aforementioned characteristics in a dataset. To deal with these problems, one decomposition technique namely EMD and two machine learning algorithms *i.e*. ANN and SVR were combined to formulate two hybrid models. The hybrid models, EMD-ANN and EMD-SVR are capable to deal with the nonlinearity and non-stationarity in a dataset. The performance of hybrid models was also evaluated with a real dataset. The empirical results clearly demonstrated the superior forecast accuracy of the proposed hybrid models (EMD-ANN and EMD-SVR) as compared to the individual ANN and SVR model.

## Acknowledgements

The authors are thankful to the editor and the anonymous reviewer for their valuable suggestions which helped in the improvement of the paper. The authors are also thankful to the Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi for providing necessary facilities for carrying out the present research work.

## References

An, X., Jiang, D., Zhao, M. and Liu, C. (2012). Short time prediction of wind power using EMD and chaotic theory. *Communication in Nonlinear Science and Numerical Simulation*, **17**, 1036-1042.

Bishop, M. C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

Brock, W. A., Scheinkman. J. A., Dechert, W. D. and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, **15**, 197-235.

Darbellay, G. A. and Slama, M. (2000). Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting*, **16**, 71-83.

Das, P., Jha, G. K., Lama, A., Parsad, R. and Mishra, D. (2020). Empirical mode decomposition based support vector regression for agricultural price forecasting. *Indian Journal of Extension Education*, **56**, 7-12.

Das, P., Lama, A. and Jha, G. K. (2021). R Package *EMDSVRhybrid*. (https://CRAN.R-project.org/package=EMDSVRhybrid).

Das, P., Lama, A and Jha, G. K. (2021). Package *EMDANNhybrid*. https://CRAN.R-project.org/package=EMDANNhybrid.

DES (Directorate of Economics and Statistics) *Cost of Cultivation of Principal Crops in India* (*various issues*), Government of India, New Delhi.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **84**, 427–431.

Diebold, F. X. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253-265.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Ontario.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. L., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proceeding of the Royal Society London A*, **454**, 909-995.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, **55**, 163–172.

Jha, G. K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India, *Neural Computing and Applications*, **3-4**, 563-571.

Levis, A. and Papageorgiou, L. (2005). Customer demand forecasting via support vector regression analysis. *Chemical Engineering Research and Design*, **83**, 1009-1018.

Lu, C. J., Lee, T. S. and Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector machine. *Decision Support Systems*, **47**, 115-125.

Mo, Y., Li, Q., Karimian, H., Fang, S., Tang, B. and Chen, G. (2020). A novel framework for daily forecasting of ozone mass concentrations based on cycle reservoir with regular jumps neural networks. *Atmospheric Environment*, **220**, 117072. https://doi.org/10.1016/j.atmosenv.2019.117072.

Mumtaz, A., Prasad, R., Xiang, R., Zaher, Y. and Yaseen, M. (2020). Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of Hydrology*, **584**, 124647. https://doi.org/10.1016/j.jhydrol.2020.124647.

Qin, S. and Chiang, L. (2019). Advances and opportunities in machine learning for process data analytics. *Computers and Chemical Engineering*, **126**, 465-473.

Sugiyama, M. and Kawanabe, M. (2012). *Machine Learning in Non-Stationary Environments-Introduction to Covariate Shift Adaptation*. 2nd Edition, MIT Press, Cambridge, Massachusetts, London, England.

Taylor, S. J. and Kingsman, B. G. (1978). Non-stationarity in sugar prices. *Journal of the Operational Research Society*, **29**, 971–980.

Vapnik, N. V. (1998). *Statistical Learning Theory*. 1st Edition, Wiley-Interscience, New York.

Wang, D., Wei, S., Luo, H., Yue, C. and Grunder, O. (2017). A novel hybrid model for air quality index forecasting based on two phase decomposition technique and modified extreme learning machine. *Science of 6e Total Environment*, **580**, 719–733.

Xuan, Y., Yu, Y. and Wu, K. (2020). Prediction of short-term stock prices based on EMD-LSTM-CSI neural network method. *5th IEEE International Conference on Big Data Analytics (ICBDA),* (pp. 135-139). Xiamen, China.

Zhang, G. P., Zhang, B. E. and Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, **14**, 35-62.

## Appendix A

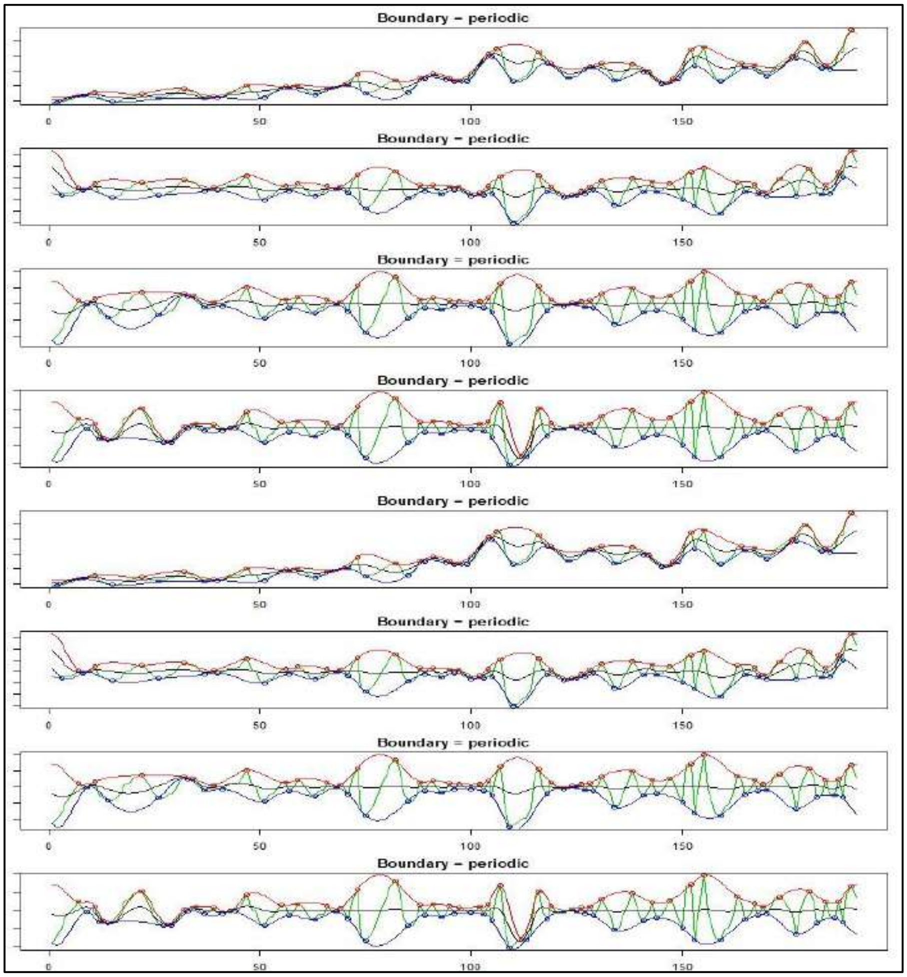


**Figure A.1: EMD process in the monthly vegetables WPI dataset**

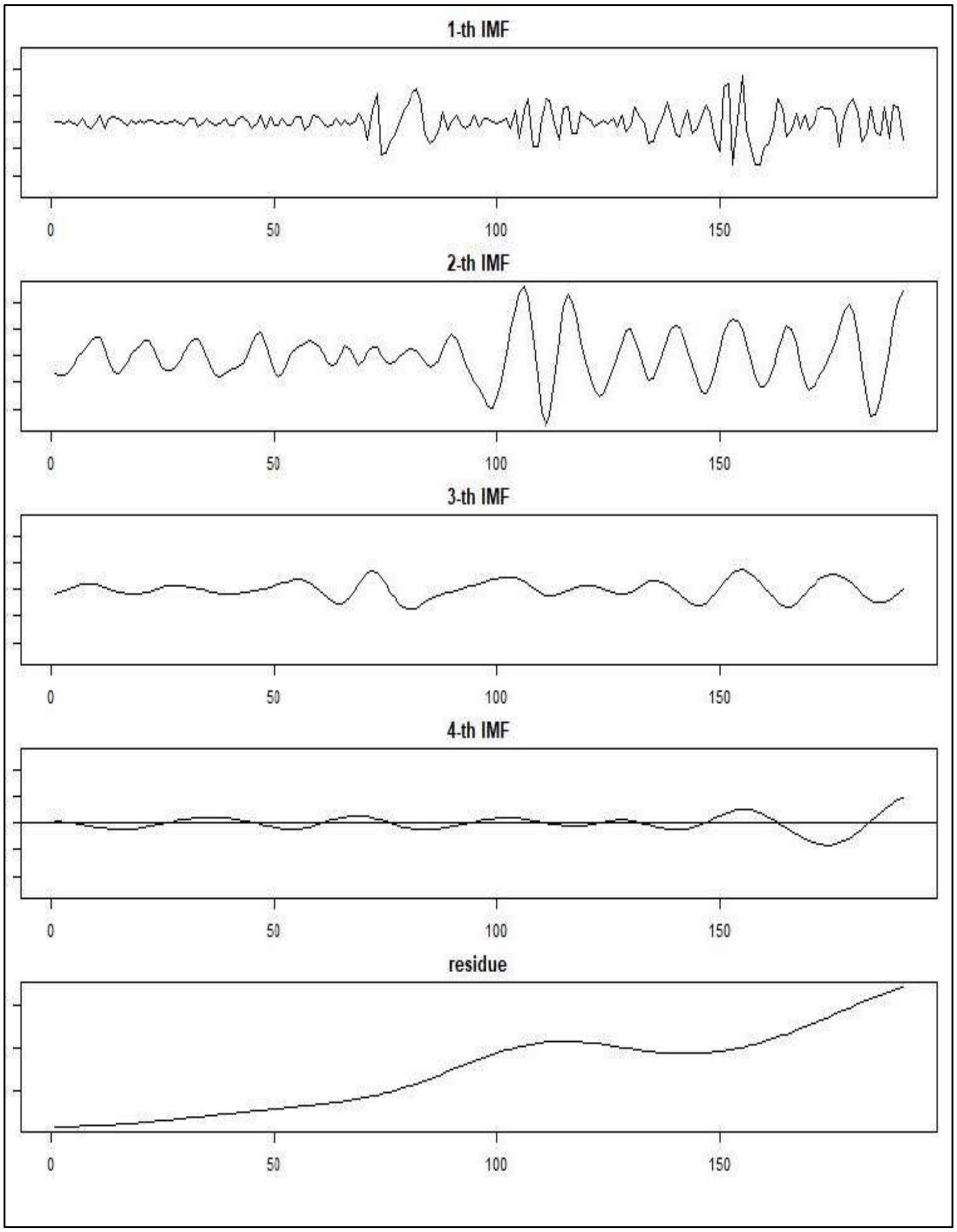


**Figure A.2: Decomposed components of monthly vegetables WPI dataset**

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 113–121

# Estimation of Area under the Multi-Class ROC for Non-Normal Data

**Arunima S. Kannan and R. Vishnu Vardhan**
*Department of Statistics, Ramanujan School of Mathematical Sciences,
Pondicherry University, Puducherry, India*

Received: 24 March 2022; Revised: 20 May 2022; Accepted: 05 July 2022

**Abstract**

In modelling ROC curves, there are several bi-distributional ROC models available in the literature. These are developed in the context of normal and non-normal data patterns and in the framework of binary classification. However, in most of the practical data at hand may exhibit multi-model patterns or it may be of multi-class, then the existing bi-distributional ROC forms are not viable to apply and fit the curve. So, in this paper, we made an attempt to address the above mentioned situations using finite mixtures. We proposed a mixture Exponential ROC model and its measures like AUC, FPR, TPR and optimal cut-offs are derived. The methodology is supported with simulated and real data sets.

*Key words:* Area under the curve; Exponential distributions; Finite mixture; ROC curve.

**AMS Subject Classifications:** 62P10

## 1. Introduction

The Receiver Operating Characteristic (ROC) curve is a classification tool that is widely used in the field of diagnostic medicine. Classification of individuals into one of the predefined groups/populations will be based on a cut-off. For a given value of cut-off, one can define the pair of true-positive rates (TPRs) and false-positive rates (FPRs), using these the ROC curve is constructed. The summary measure of ROC, which assesses the performance of a particular diagnostic test, is the area under the curve (AUC) whose value lies between 0 and 1. Higher the AUC value, the better the diagnostic test's performance.

The initial work on the distributional approach to model the ROC curve was by Green *et al.* (1966) where data is assumed to follow the gaussian distribution. In later years, Dorfman and Alf (1968) gave the maximum likelihood estimates for the binormal ROC curve. Metz (1978) gave a detailed explanation about the basic principles of ROC curve and its measures. Estimation of the parameters of the binormal model was of prime focus by many researchers. Goddard and Hindberg (1990) proposed a ROC model that meets the criterion of non-normal data, namely the Bi-logistic ROC model. Farraggi and Reiser (2002) provided the parametric and non-parametric approach of estimating the AUC of the ROC curve. Over the years, many researchers have attempted to develop Bi-distributional

Corresponding Author: R. Vishnu Vardhan
Email: vrstatsguru@gmail.com

ROC curves; a few to mention are the Bi-Generalized Exponential ROC model by Hussain (2011), Vardhan *et al.* (2012) on Bi-Exponential and Bi-Weibull ROC model, Bi-Gamma ROC model by Hussain (2012). A detailed review of several bi-distributional ROC models was made by Balaswamy and Vishnu (2016).

In classification, one of the main issues is that in most of the data sets, we do not have the information about the group membership; there, we need to use appropriate statistical methods to figure out the homogeneous subsets. We can make the graphical depiction of unsupervised data, and it may exhibit unimodal or multi-model patterns that exist in the data. One of the most widely used methodologies that helps to sum up the multi-model patterns accompanied by their respective weights in the form of convex combination is the Finite Mixture Models (FMM). The general expression of the finite mixture distribution is given in equation (1).

$$g(x) = \sum_{i=1}^{k} \pi_i f_i(x) \tag{1}$$

where, $\pi_i$'s are the mixture proportions or mixture weights such that $\pi_i > 0$ ; $\sum_{i=1}^{k} \pi_i = 1$ and $f_i$'s are component distributions ; $i = 1, 2, ..k$.

The seminal work on mixture models using crabs data was by Pearson (1894) and a detailed study on mixture models was given by Lindsay (1995). Over the years, the practical applicability of FMM branched out to various fields like remote sensing, environmental studies, diagnostic medicine, survival analysis, social and psychological science (Peel and MacLahlan, 2000). But, most of the works reported in the literature were based on the normal distribution. However, there are several practical instances where data may not follow the normal distribution. In such situations, the existing normal mixture models do not support, hence there is a need to have mixture models for non-normal data. Here a brief review on Mixture Exponential is presented. Mendenhall and Hader (1958) estimated the parameters of mixed exponentially distributed failure time distribution. Jewell (1982) gives a detailed explanation of the mixture of exponential distributions and gives a practical algorithm for the maximum likelihood estimate. Wang and Wang (2014) proposed an EM Algorithm for the finite mixture of exponential distribution models. Literature has many applications with the use of mixture exponential distributions, recently, Polymenis (2020) used mixture of exponential distributions for assessing hazard rates from COVID-19.

This paper provides an approach to classify the non-normal data with hidden populations. Here it is assumed that the population follows a mixture of exponential distribution and derived the Mixture Exponential ROC and its measures. The rest of the paper is organized as follows. Section 2 discusses the proposed Mixture Exponential ROC. Section 3 provides numerical illustrations of the proposed methodology with simulated as well as real-life data sets. Section 4 concludes the paper with the summary.

## 2. Mixture exponential ROC

Let us assume that healthy population, $H \sim exp(\theta_0)$ and diseased population has two sub populations/mixture of populations of $D_1$ and $D_2$ such that, $D_1 \sim exp(\theta_1)$ and $D_2 \sim exp(\theta_2)$. Then the expressions for intrinsic measures of Mixture Exponential ROC (mixExp ROC) are defined as FPR of mixExp ROC (mixFPR) is given as

$$mixFPR = \pi_1 FPR_1 + \pi_2 FPR_2 \tag{2}$$

where

$$\begin{aligned} FPR_1 = P(S > t_1 \mid H) \;\; ; \;\; FPR_2 = P(S > t_2 \mid D_1) \\ FPR_1 = x(t_1) = e^{-\theta_0 t_1} \;\; ; \;\; FPR_2 = x(t_2) = e^{-\theta_1 t_2} \end{aligned} \tag{3}$$

where $\pi_i$'s are mixing proportions/weights, $t_1$ and $t_2$ are the respective cut-off values for the classification of $(H, D_1)$ and $(D_1, D_2)$ respectively. Here $FPR_1$, $FPR_2$ are the false positive rate values of $H$ and $D_1$ populations & $D_1$ and $D_2$ populations respectively. From equation (3) we can write $t_1$ and $t_2$ as

$$t_1 = -\frac{log(x(t_1))}{\theta_0} \;\; ; \;\; t_2 = -\frac{log(x(t_2))}{\theta_1} \tag{4}$$

TPR of mixExp ROC (mixTPR) is given as

$$mixTPR = \pi_1 TPR_1 + \pi_2 TPR_2 \tag{5}$$

where

$$\begin{aligned} TPR_1 = P(S > t_1 \mid D_1) \;\; ; \;\; TPR_2 = P(S > t_2 \mid D_2) \\ TPR_1 = y(t_1) = e^{-\theta_1 t_1} \;\; ; \;\; TPR_2 = y(t_2) = e^{-\theta_2 t_2} \end{aligned} \tag{6}$$

Here, $TPR_1$, $TPR_2$ are the true positive rate values of H and $D_1$ populations & $D_1$ and $D_2$ populations respectively. Substituting equation (4) in (6) we will get the mixture exponential ROC curve which be written as

$$mixROC = \pi_1 ROC_1 + \pi_2 ROC_2 \tag{7}$$

$$ROC_1 = x(t_1)^{\frac{\theta_1}{\theta_0}} \;\; ; \;\; ROC_2 = x(t_2)^{\frac{\theta_2}{\theta_1}} \tag{8}$$

By equating the pdf's of the distributions, the optimal cut-off can be obtained as

$$t_1 = \frac{log\theta_1 - log\theta_0}{\theta_1 - \theta_0} \;\; ; \; t_2 = \frac{log\theta_2 - log\theta_1}{\theta_2 - \theta_1} \tag{9}$$

accuracy can be expressed notationally as

$$mixAUC = \int_0^1 mixROC(t)dt = \pi_1 \frac{\theta_0}{\theta_0 + \theta_1} + \pi_2 \frac{\theta_1}{\theta_1 + \theta_2} \tag{10}$$

Youden's $J$ index (Youden, 1950) is another way of summarising the performance of a diagnostic test.

Youden's $J$ index is defined as

$$J = (TPR - FPR) \tag{11}$$

then maximum Youden's index is reported as

$$J = max_{(t)} \left(TPR(t) - FPR(t)\right) \tag{12}$$

where $t$ denotes the classification threshold for which $J$ is maximal. From the above equation, the optimal threshold can be estimated at the maximum Youden's Index value, since, the maximum distance between the curve and the chance line can be used to identify the optimal threshold and will be unique in nature. This optimal threshold classifies the individuals with a better accuracy and further it can be used to assign the status of the unspecified subjects/individuals. A value of $J$=1 sures that there are no false positives or false negatives, *i.e.* the test is perfect.

## 3. Numerical illustrations

### 3.1. Simulated data

Simulation studies are carried out at various parameter combinations by considering equal mixture weights. Using the parameters values given in Table 1, random samples are generated for n = (25, 50, 100, 200).

**Table 1: Initial parameters**

| Case | $\pi_1$ | $\pi_2$ | $\theta_0$ | $\theta_1$ | $\theta_2$ |
|---|---|---|---|---|---|
| I | 0.5 | 0.5 | 0.4 | 0.1 | 0.01 |
| II | 0.5 | 0.5 | 0.4 | 0.2 | 0.05 |
| III | 0.5 | 0.5 | 0.4 | 0.25 | 0.1 |
| IV | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 |

The results pertaining to each case at every sample size in Table 2 and respective ROC curves are depicted in Figure 1. The parameter values are chosen in such a way that they exhibit worst and moderate classification scenarios. The estimation of parameters of the mixture distribution is carried out using EM algorithm in R software. It is a known fact that as higher the AUC, minimum will be the overlapping region, in turn giving out better percentage of correct classification. The estimated values of $t_1$ and $t_2$ are the optimal one, which are derived using the Youden's $J$. The interpretation of $t_1$ and $t_2$ goes like this:

Let S be the values/samples generated using each parameter combination

$$\text{The individual will be classified as} = \begin{cases} H, & \text{if } S \leq t_1 \\ D_1, & \text{if } t_1 < S \leq t_2 \\ D_2, & \text{if } S > t_2 \end{cases}$$

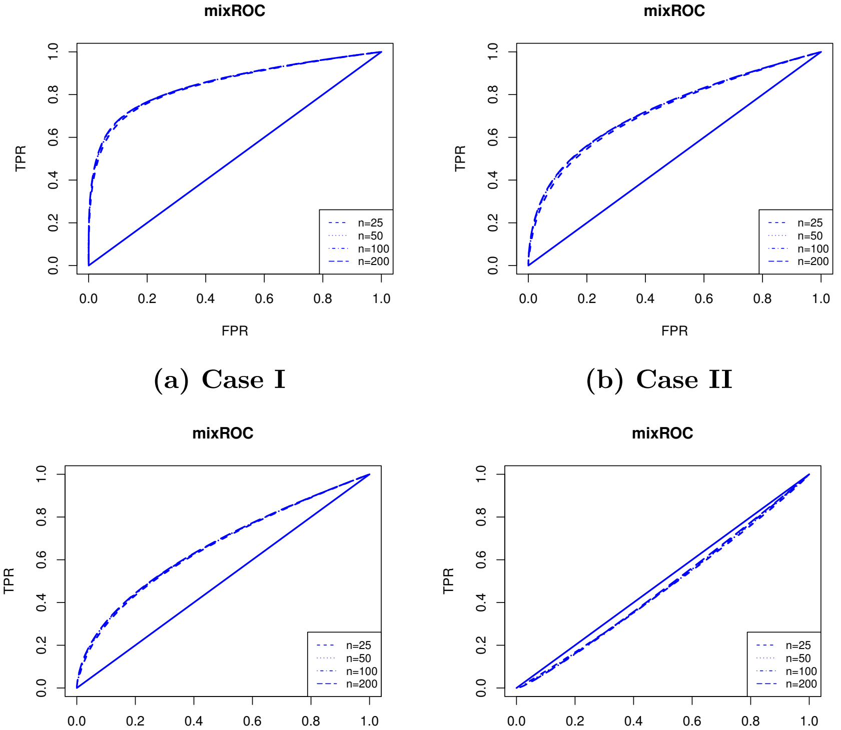
To have a better understanding of $t_1$ and $t_2$, FPR and TPR, let us consider an instance under case I from Table 2. For n=100, the $\hat{t_1} = 4.60711$; $\hat{t_2} = 25.43069$; $\widehat{mixFPR} = 0.12541$; $\widehat{mixTPR} = 0.70705$ and $\hat{J} = 0.58164$ results $\widehat{mixAUC} = 0.85340$. This means to that an individual will be classified in the following manner

$$\text{The individual will be classified as} = \begin{cases} H, \text{ if } S \leq 4.60711 \\ D_1, \text{ if } 4.60711 < S \leq 25.43069 \\ D_2, \text{ if } S > 25.43069 \end{cases}$$

**Table 2: ROC curve estimates for simulated data**

| Case | n | $\widehat{\pi_1}$ | $\widehat{\pi_2}$ | $\widehat{t_1}$ | $\widehat{t_2}$ | $\widehat{J}$ | $\widehat{mixFPR}$ | | $\widehat{mixTPR}$ | | $\widehat{mixAUC}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $\widehat{FPR_1}$ | $\widehat{FPR_2}$ | $\widehat{TPR_1}$ | $\widehat{TPR_2}$ | $\widehat{AUC_1}$ | $\widehat{AUC_2}$ |
| I | 25 | 0.50213 | 0.49787 | 4.56913 | 25.23253 | 0.57538 | 0.13001 | | 0.70539 | | 0.85036 | |
| | | | | | | | 0.16142 | 0.08043 | 0.62848 | 0.77137 | 0.79547 | 0.90595 |
| | 50 | 0.49731 | 0.50269 | 4.59730 | 25.44494 | 0.58133 | 0.12534 | | 0.70667 | | 0.85318 | |
| | | | | | | | 0.15910 | 0.07838 | 0.62913 | 0.77283 | 0.79784 | 0.90766 |
| | 100 | 0.49938 | 0.50062 | 4.60711 | 25.43069 | 0.58164 | 0.12541 | | 0.70705 | | 0.85340 | |
| | | | | | | | 0.15871 | 0.07795 | 0.62899 | 0.77323 | 0.79837 | 0.90821 |
| | 200 | 0.49976 | 0.50024 | 4.61104 | 25.45291 | 0.58233 | 0.12478 | | 0.70710 | | 0.85378 | |
| | | | | | | | 0.15800 | 0.07801 | 0.62933 | 0.77331 | 0.79917 | 0.90832 |
| II | 25 | 0.53123 | 0.46877 | 3.42247 | 9.14696 | 0.34870 | 0.21594 | | 0.56464 | | 0.72485 | |
| | | | | | | | 0.25757 | 0.15880 | 0.50218 | 0.63186 | 0.66176 | 0.79887 |
| | 50 | 0.50817 | 0.49183 | 3.41111 | 9.14908 | 0.35861 | 0.20937 | | 0.56798 | | 0.73139 | |
| | | | | | | | 0.24995 | 0.15892 | 0.50140 | 0.63015 | 0.66688 | 0.79854 |
| | 100 | 0.50680 | 0.49320 | 3.46106 | 9.23796 | 0.35864 | 0.20821 | | 0.56685 | | 0.73169 | |
| | | | | | | | 0.25005 | 0.15837 | 0.50063 | 0.62916 | 0.66671 | 0.79866 |
| | 200 | 0.49881 | 0.50119 | 3.46938 | 9.24755 | 0.36089 | 0.20754 | | 0.56843 | | 0.73312 | |
| | | | | | | | 0.24999 | 0.15823 | 0.50041 | 0.62900 | 0.66676 | 0.79885 |
| III | 25 | 0.53521 | 0.46479 | 3.08832 | 6.03442 | 0.23997 | 0.26023 | | 0.50021 | | 0.65824 | |
| | | | | | | | 0.31259 | 0.21777 | 0.48196 | 0.54426 | 0.61200 | 0.71325 |
| | 50 | 0.52410 | 0.47590 | 3.11203 | 6.04253 | 0.24393 | 0.25670 | | 0.50063 | | 0.66124 | |
| | | | | | | | 0.29402 | 0.21777 | 0.46093 | 0.54316 | 0.61165 | 0.71320 |
| | 100 | 0.51487 | 0.48513 | 3.13166 | 6.11228 | 0.24513 | 0.25448 | | 0.49960 | | 0.66246 | |
| | | | | | | | 0.28625 | 0.21780 | 0.45799 | 0.54262 | 0.61539 | 0.71332 |
| | 200 | 0.51262 | 0.48738 | 3.12554 | 6.10010 | 0.24725 | 0.25349 | | 0.50073 | | 0.66390 | |
| | | | | | | | 0.28586 | 0.21708 | 0.45705 | 0.54311 | 0.61515 | 0.71427 |
| IV | 25 | 0.56482 | 0.43518 | 2.47823 | 2.48022 | 0.02424 | 0.56973 | | 0.59397 | | 0.49789 | |
| | | | | | | | 0.58665 | 0.59789 | 0.62503 | 0.63822 | 0.49900 | 0.50069 |
| | 50 | 0.54976 | 0.45024 | 2.47846 | 2.48475 | 0.01853 | 0.55706 | | 0.57559 | | 0.49997 | |
| | | | | | | | 0.61440 | 0.58862 | 0.64271 | 0.61674 | 0.49925 | 0.49956 |
| | 100 | 0.53953 | 0.46047 | 2.47643 | 2.48859 | 0.01319 | 0.57909 | | 0.59228 | | 0.49943 | |
| | | | | | | | 0.62699 | 0.61702 | 0.64770 | 0.63862 | 0.49972 | 0.50064 |
| | 200 | 0.55491 | 0.44509 | 2.48864 | 2.49903 | 0.01036 | 0.57716 | | 0.58752 | | 0.50036 | |
| | | | | | | | 0.63807 | 0.62109 | 0.65242 | 0.63512 | 0.49983 | 0.50025 |

The cut-offs $\widehat{t_1}$ and $\widehat{t_2}$, are able to provide an $\widehat{FPR}$ of 12.54% and $\widehat{TPR}$ of 70.70%. So, if there are 100 samples in the data, these two cut-offs will be able to detect the true positives upto 70% and with a wrong classification of around 12%. In total, the accuracy of $\hat{t_1}$ and $\hat{t_2}$ is around 85%. In similar lines, the other combinations can be interpreted. From Figure 1, it is clear that the area under the curve is decreasing from case I to case IV, which is indicating that the accuracy of the classification is decreasing from case I to case IV. The curve of case IV is close to the diagonal line, results the worst classification.



**(a) Case I** **(b) Case II**

**(c) Case III** **(d) Case IV**

**Figure 1: mixExp ROC curves of simulated data sets**

### 3.2. Real data

The real data set represent the survival times of 121 patients with breast cancer obtained from a large hospital in a period from 1929 to 1938 (Lee and Wang, 2003). This data set has recently been studied by Yang *et al.* (2021). The p-value for *K*-*S* test for exponential distribution for this data is 0.06024 (Test statistic, $D = 0.12031$), which indicates that the data follows exponential distribution. We have $\theta_0 = 0.4$, $\theta_1 = 0.0280$ and $\theta_2 = 0.0202$. The estimated measures of mixExp ROC curve is given in Table 3 and respective ROC curve is depicted in Figure 2. As the curve is observed between the chance line and the left top corner and also connecting to the AUC= 0.7355, this indicates a moderate amount of classification with cut-offs $\hat{t_1}$ and $\hat{t_2}$.

From Table 3, $\hat{t_1} = 20.24715$; $\hat{t_2} = 46.32893$; $\widehat{mixFPR} = 0.24871$; $\widehat{mixTPR} = 0.6628$

**Table 3: mixExp ROC curve estimates of breast cancer data**

| $\widehat{\pi_1}$ | $\widehat{\pi_2}$ | $\widehat{t_1}$ | $\widehat{t_2}$ | $\widehat{J}$ | $\widehat{mixFPR}$ | | $\widehat{mixTPR}$ | | $\widehat{mixAUC}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\widehat{FPR_1}$ | $\widehat{FPR_2}$ | $\widehat{TPR_1}$ | $\widehat{TPR_2}$ | $\widehat{AUC_1}$ | $\widehat{AUC_2}$ |
| 0.49918 | 0.50018 | 20.24715 | 46.32893 | 0.24908 | 0.24871 | | 0.6628 | | 0.7355 | |
| | | | | | 0.04951 | 0.3493 | 0.8418 | 0.3865 | 0.9458 | 0.5253 |

and results $\widehat{mixAUC} = 0.7355$. This means that an individual will be classified as follows.

$$\text{The individual will be classified as} = \begin{cases} \text{low survival rate, if } S \leq 20.24715 \\ \text{moderate survival rate, if } 20.24715 < S \leq 46.32893 \\ \text{high survival rate, if } S > 46.32893 \end{cases}$$

The cut-offs $\widehat{t_1}$ and $\widehat{t_2}$, are able to provide false positive rate of 24.87% and true positive rate of 66.28%. In total, the accuracy is around 73.55%, which indicates of a moderate classification. Further, 100 bootstrap samples are generated from the breast cancer data. The bootstrap estimates of important measures and their confidence intervals are reported in Table 4. The mixROC curves are also drawn for all the bootstrap samples and is shown in Figure 3. The curves clearly depict a moderate classification. From Table 4, it is observed

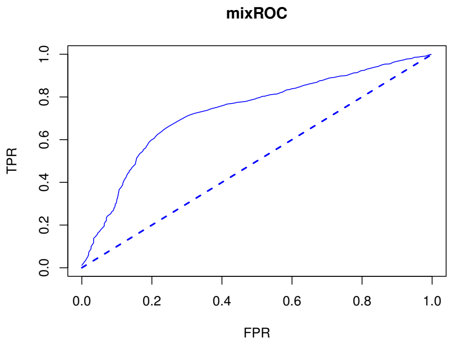


**Figure 2: mixExp ROC Curve**

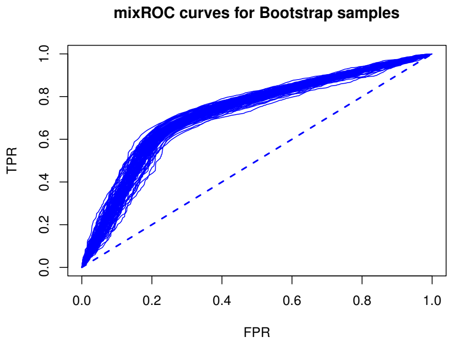


**Figure 3: mixExp ROC Curves for 100 bootstrap samples**

**Table 4: Bootstrap estimates of breast cancer data**

| Bootstrap | $\widehat{mixFPR}_{boot}$ | $\widehat{mixTPR}_{boot}$ | $\widehat{J}_{boot}$ | $\widehat{mixAUC}_{boot}$ |
|---|---|---|---|---|
| Estimates | 0.248375 | 0.647383 | 0.399009 | 0.7230622 |
| Variance | 0.000271 | 0.000207 | 0.000309 | 0.000160269 |
| 95% Lower limit | 0.2375 | 0.635805 | 0.388395 | 0.7164525 |
| 95% Upper limit | 0.257676 | 0.658599 | 0.41017 | 0.730174 |

that the cut-offs provide reasonably low FPR = 0.248375 (0.2375, 0.257676) and a good level of TPR = 0.647383 (0.635805, 0.658599). This means that if there are 100 subjects then the cut-offs will be able to detect the class/status of around 65 subjects correctly, providing an accuracy of 0.7230 (0.7164525, 0.730174). Upon conducting 100 bootstraps and constructing the 95% confidence interval the outcomes revealed an observation that the width of the confidence interval is shorter indicating consistent estimates. Further the results of the bootstrap matches closely to the results in Table 3.

## 4. Summary

In this paper, we proposed an ROC model that follows exponential distribution with multi-class classification. Here we considered situation like (i) if we come across multi-model patterns in the diseased population and (ii) if there are more than two categories in the data. The proposed model addresses the above two situations and is dealt using the concept of finite mixtures. The model so constructed is named as *Mixture Exponential ROC Curve*. The measures such as mixAUC, mixFPR, mixTPR and optimal cut-offs are derived and supported with numerical illustrations. With respect to simulations, we tried to present the behaviour of the proposed ROC model by constructing the worst and moderate cases. Further the numerical illustrations is extended with breast cancer dataset. It is noticed that there were two sub populations in the diseased population. The overall AUC is observed to be 73.5 and optimal thresholds are 20.24 and 46.32. To summarize the work, and mixture exponential ROC model is proposed, and for the non-normal and multi-class data this model can be applied.

## Acknowledgements

The authors are thankful to the referees for their valuable comments and suggestions which helped in improvising the paper.

## References

Balaswamy, S. and Vishnu, R. V. (2016). An anthology on parametric ROC models. *Research Reviews:Journal of Statistics*, **5**, 32-46.

Dorfman, D. D. and Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory - A direct solution. *Psychometrika*, **33**, 117-124.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, **21**, 3093-3106.

Goddard, M. and Hindberg, I. (1990). Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine*, **9**, 325-337.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* John Wiley & Sons.

Hussain, E. (2011). The ROC curve model from generalized-exponential distribution. *Pakistan Journal of Statistics and Operation Research*, **7**, 283-303.

Hussain, E. (2012). The bi-gamma ROC curve in a straightforward manner. *Journal of Basic & Applied Sciences*, **8**, 309-314.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, **10**, 479–484.

Lee, E. T. and Wang, J. (2003). *Statistical Methods for Survival Data Analysis.* $3^{rd}$ Edition, John Wiley & Sons.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, **5**, 1–163.

Mendenhall, W. and Hader, R. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504-520.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283–298.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, **185**, 71–110.

Peel, D. and MacLahlan, G. (2000). *Finite Mixture Models.* John Wiley & Sons.

Polymenis, A. (2020). An application of a mixture of exponential distributions for assessing hazard rates from covid-19. *Journal of Population Therapeutics and Clinical Pharmacology*, **27**, 58–63.

Vardhan, R. V., Pundir, S., and Sameera, G. (2012). Estimation of area under the ROC curve using exponential and weibull distributions. *Bonfring International Journal of Data Mining*, **2**, 52–56.

Wang, Y. and Wang, J. (2014). The EM Algorithm for The Finite Mixture of Exponential Distribution Models. *International Journal of Contemporary Mathematical Sciences*, **9**, 57-64.

Yang, Y., Tian, W., and Tong, T. (2021). Generalized mixtures of exponential distribution and associated inference. *Mathematics*, **9**, 1-22.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, **3**, 32-35.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 123–141

# Parameter Estimation of Generalized Exponential Distribution using Variations in Methods of Ranked Set Sampling

**Vyomesh Nandurbarkar**[1] **and Ashok Shanubhogue**[2]
[1]*Department of Mathematical Sciences*
*Charotar University of Science and Technology, Changa*
[2] *Department of Statistics*
*Hon. Director, UGC- Human Resource Development Center*
*Sardar Patel University, V.V. Nagar*

Received: 30 January 2022; Revised: 06 June 2022; Accepted: 10 July 2022

**Abstract**

In this study, we estimate the parameters of the generalized exponential distribution using moving extreme ranked set sampling (MERSS) and ranked set sampling (RSS). Under both sampling schemes, we obtain expressions for likelihood functions and derive maximum likelihood equations and the Fisher information matrices. We numerically compute the ML estimates. We compare these estimates with estimates obtained by simple random sampling(SRS) and ranked set sampling using mean square error. Based on simulation studies, we demonstrate that the RSS scheme is more efficient for small set sizes than MERSS and SRS for shape and scale parameters.

*Key words:* Ranked set sampling; Moving extremes ranked set sampling; Maximum likelihood estimator; Fisher information number; Mean square error.

**AMS Subject Classifications:** 62F07, 62F10, 62D05

## 1. Ranked set sampling

When sampling environmental and ecological data, there may be situations in which it is difficult to measure (or quantify) a selected unit with appropriate accuracy, but ranking a few selected units based on the characteristic of interest can be simple. As an example, if one wishes to estimate the mean height of trees, then measuring the height of the sampled trees might be challenging, but there are relatively easy methods to rank small sets of trees based on eye observation of their heights. Rank set sampling (RSS) was developed by McIntyre (1952) as an improvement on random sampling in situations such as these, where some ranking of units may be simple.

A ranked set sample is obtained by randomly selecting $m^2$ units from an infinite population. These units are then partitioned randomly into $m$ equal samples of $m$ units each. The units in the sample are ranked by judgment or visual inspection or by cheap (low cost) way, or by using auxiliary variables but without actual measurements. The unit with the lowest rank is measured for variable $X$ of interest and the remaining units are discarded; the second sample of $m$ units is ranked without actual measurements. As a result, the second-lowest unit in this set is measured for variable $X$ of interest, while the rest are discarded. Once the largest unit in the last sample of size $m$ has been measured, the entire procedure is repeated $h$ times, yielding $n = mh$ measured units from $m^2h$ selected units. Compared to a simple random sample, this ranked set sample also represents the entire population and is spread throughout the population.

When sampling units are selected without replacement, the estimation of population mean using RSS, and comparison of RSS estimate to SRS estimate, are discussed in Patil et al. (1994).

There are different modifications of the methods of RSS available in the literature. Al-Odat and Al-Saleh (2001) introduced the concept of varied size RSS, which is known as moving extreme ranked set sampling. The described scheme is as given below.

### 1.1. Moving extreme ranked set sampling

Step 1. Select $m$ simple random samples of size $1, 2, 3, \ldots, m$ respectively.

Step 2. Order the sampling units of each of the samples by eye or by some other relatively inexpensive method, without actual measurements.

Step 3. Measure accurately the maximum order observation from the first set, the maximum order observation from the second set. The process continues in this way until the maximum order observation from the last $m$th sample is measured.

Step 4. Repeat the Step 1. to Step 3. and then measure the minimum order observation instead maximum order observation.

Step 5. The prcedure described above is one cycle.The entire cycle can be repeated $h$ times to obtain a MERSS of size $n = 2mh$

Hence, this scheme is more simple to implement than RSS. Recently Wangxue et al. (2019), has implemented the MERSS scheme to estimate parameters of Pareto distribution, and He et al. (2021) has implemented the MERSS scheme to estimate parameters of log-logistic distribution and Chen et al. (2021) discussed estimation of location parameter using maximum likelihood under MERSS scheme.

## 2. Generalized exponential distribution

Consider a continuous random variable $X$ which follows two-parameter generalized exponential distribution having cumulative distribution function and probability diensity function respectively as,

$$H(x; \alpha, \beta) = (1 - e^{-\beta x})^{\alpha}, \quad \alpha, \beta, x > 0 \tag{1}$$

and

$$h(x;\alpha,\beta)=\alpha\beta(1-e^{-\beta x})^{\alpha-1}e^{-\beta x} \quad \alpha,\beta,x>0. \tag{2}$$

Here $\alpha$ is the shape parameter, and $\beta$ is the scale parameter.

As seen in (1), the distribution is of the type $[F(\cdot)]^{\alpha}$ where $F(\cdot)$ is a cumulative distribution function of exponential distribution with scale parameter $\beta$. The distribution (1) is introduced and studied in detail by Gupta and Kundu (1999). According to Gupta and Kundu (2001), while fitting distribution for positive lifetime data, the generalized exponential distribution is used as an alternative to the two-parameter Weibull distribution and two parameter gamma distribution.

The mean and variance of the distribution with density function given in (2) are

$$E(X)=\frac{1}{\beta}(\psi(\alpha+1)-\psi(1)) \tag{3}$$

and

$$V(X)=-\frac{1}{\beta^2}(\psi'(\alpha+1)-\psi'(1)), \tag{4}$$

where $\psi(\cdot)$ is the digamma function and $\psi'(\cdot)$ is the derivative of $\psi(\cdot)$. The skewness and kurtosis both are independent of the scale parameter and they are decreasing function of the shape parameter $\alpha$.

Consider the transformation $Y=\beta X$ in (2) so that the probability density function of $Y$ is

$$f(y;\alpha)=\alpha(1-e^{-y})^{\alpha-1}e^{-y}, \quad y>0 \tag{5}$$

Let $Y_1,Y_2,\ldots,Y_n$ be a random sample on $Y$. Then the pdf of $r^{\text{th}}$ order statistic $Y_{r:n},(r=1,2,\ldots,n)$ is

$$f_{r:n}(y)=\frac{1}{B(r,n-r+1)}F^{r-1}(y)\left[1-F(y)\right]^{n-r}f(y), \tag{6}$$

where $B(p,q)=\dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \quad p>0,q>0.$

$$\begin{aligned}
f_{r:n}(y)&=\frac{1}{B(r,n-r+1)}\sum\nolimits_{i=0}^{n-r}(-1)^i\binom{n-r}{i}\alpha\left[1-e^{-y}\right]^{(r+i)\alpha-1}e^{-y},\\
&=\frac{1}{B(r,n-r+1)}\sum\nolimits_{i=0}^{n-r}(-1)^i\binom{n-r}{i}\alpha(r+i)\frac{\left[1-e^{-y}\right]^{(r+i)\alpha-1}}{r+i}e^{-y},\\
&=\frac{1}{B(r,n-r+1)}\sum\nolimits_{i=0}^{n-r}\frac{(-1)^i\binom{n-r}{i}}{r+i}f(y;\alpha(r+i)).
\end{aligned}$$

For $r = 1, 2, \ldots, n$, the first order and second order moments are

$$\begin{aligned} E(Y_{r:n}) &= \frac{1}{B(r, n-r+1)} \sum\nolimits_{i=0}^{n-r} \frac{(-1)^i \binom{n-r}{i}}{r+i} \left[-\psi(1) + \psi(\alpha(r+i)+1)\right] \\ &= \frac{1}{B(r, n-r+1)} \sum\nolimits_{i=0}^{n-r} \frac{(-1)^i \binom{n-r}{i}}{r+i} \left[\gamma + \psi(\alpha(r+i)+1)\right] \end{aligned} \tag{7}$$

and

$$\begin{aligned} E(Y_{r:n}^2) = \frac{1}{B(r, n-r+1)} \sum\nolimits_{i=0}^{n-r} \frac{(-1)^i \binom{n-r}{i}}{r+i} \Bigg\{ & \Big[\psi((r+i)\alpha+1) + \gamma\Big]^2 \\ & -\psi'((r+i)\alpha+1) + \pi^2/6 \Bigg\}, \end{aligned} \tag{8}$$

where $\gamma = -\psi(1) = 0.577215\ldots$ and $\psi'(1) = \pi^2/6$.

Our discussion assumes that the sampling units are ranked without error for the characteristic of interest. In section (3) we discuss the estimation of shape and scale parameters of distribution given in (1) using MERSS scheme. In section (4) we discuss estimation of shape and scale parameters given in (1) using RSS scheme. Section (6) presents our findings.

## 3. The MERSS sample

Let $X_{j(11)}, X_{j(21)}, X_{j(22)}, X_{j(31)}, X_{j(32)}, X_{j(33)}, \ldots X_{j(m1)}, X_{j(m2)}, \ldots, X_{j(mm)}$ and $X'_{j(11)}, X'_{j(21)}, X'_{j(22)}, X'_{j(31)}, X'_{j(32)}, X'_{j(33)}, \ldots X'_{j(m1)}, X'_{j(m2)}, \ldots, X'_{j(mm)}$ be independent random variables all having the same distribution given in (1) at cycle $j = 1, 2, \ldots, h$ In the case of perfect ranking, for $j = 1$,
$X_{1(1:1)} = \max\{X_{1(11)}\}$,
$X_{1(2:2)} = \max\{X_{1(21)}, X_{1(22)}\}$,
$X_{1(3:3)} = \max\{X_{1(31)}, X_{1(32)}, X_{1(33)}\}, \ldots,$
$X_{1(i:i)} = \max\{X_{1(i1)}, X_{1(i2)}, \ldots, X_{1(ii)}\}$ ,
denote the $i-$th order statistic from random sample of size $i, i = 1, 2, \ldots, m$ for cycle 1 and
$X'_{1(1:1)} = \min\{X'_{1(11)}\}$,
$X'_{1(1:2)} = \min\{X'_{1(21)}, X'_{1(22)}\}$,
$X'_{1(1:3)} = \min\{X'_{1(31)}, X'_{1(32)}, X'_{1(33)}\}, \ldots,$
$X'_{1(1:i)} = \min\{X'_{1(i1)}, X'_{1(i2)}, \ldots, X'_{1(ii)}\}$
denote the first order statistic from random sample of size $i,\ i = 1, 2, \ldots, m$ for cycle 1.
Considering $\{X_{1(1:1)}, X_{1(2:2)}, X_{1(3:3)}, \ldots, X_{1(m:m)}\}$ as $MERSS_{\text{Maximum}}$ , and
$\{X'_{1(1:1)}, X'_{1(1:2)}, X_{1(1:3)}, \ldots, X'_{1(1:m)}\}$as $MERSS_{\text{Minimum}}$, the MERSS sample (Cycle 1) of size $2m$ is

$$\left\{ X_{1(1:1)}, X_{1(2:2)}, X_{1(3:3)}, \ldots, X_{1(m:m)}; X'_{1(1:1)}, X'_{1(1:2)}, X'_{1(1:3)}, \ldots, X'_{1(1:m)} \right\}$$

For simplicity let $u_{ji}$ represents the observed values of $X_{j(i:i)}$ and $v_{ji}$ represents the observed values of $X'_{j(1:i)}$, then the probability density functions of

$MERSS_{\text{maximum}}$ and $MERSS_{\text{minimum}}$ respectively are

$$\begin{aligned} f_{X_{j(i:i)}}(u_{ji}) &= i\left[H(u_{ji})\right]^{i-1} h(u_{ji}) \\ &= i\left[(1-e^{-\beta u_{ji}})^{\alpha}\right]^{i-1} \alpha\beta\left(1-e^{-\beta u_{ji}}\right)^{\alpha-1} e^{-\beta u_{ji}} \\ &= \alpha\beta i\left(1-e^{-\beta u_{ji}}\right)^{i\alpha-1} e^{-\beta u_{ji}} \end{aligned} \tag{9}$$

and

$$\begin{aligned} f_{X'_{j(1:i)}}(v_{ji}) &= i\left[1-H(v_{ji})\right]^{i-1} h(v_{ji}) \\ &= i\left[1-(1-e^{-\beta v_{ji}})^{\alpha}\right]^{i-1} \alpha\beta\left(1-e^{-\beta v_{ji}}\right)^{\alpha-1} e^{-\beta v_{ji}} \\ &= \alpha\beta i\left(1-e^{-\beta v_{ji}}\right)^{\alpha-1}\left[1-(1-e^{-\beta v_{ji}})^{\alpha}\right]^{i-1} e^{-\beta v_{ji}} \end{aligned} \tag{10}$$

From (7) and (8), the first order and second order moments are,

$$\mu_{j(i:i)} = E\{X_{j(i:i)}\} = \frac{1}{\beta}\left[\psi(i\alpha+1)+\gamma\right] \tag{11}$$

$$\mu^2_{j(i:i)} = E\{X^2_{j(i:i)}\} = \frac{1}{\beta^2}\left[\{\psi(i\alpha+1)+\gamma\}^2 - \psi'(i\alpha+1)+\pi^2/6\right] \tag{12}$$

$$\begin{aligned} \mu^2_{j(1:i)} = E\{X'^2_{j(1:i)}\} = \frac{1}{\beta^2} i \textstyle\sum_{k=0}^{i-1} \frac{(-1)^k\binom{i-1}{k}}{k+1} &\Big\{\left[\psi((k+1)\alpha+1)+\gamma\right]^2 \\ &-\psi'((k+1)\alpha+1)+\pi^2/6\Big\} \end{aligned} \tag{13}$$

$$\begin{aligned} \mu^2_{j(1:i)} = E\{X'^2_{j(1:i)}\} = \frac{1}{\beta^2} i \textstyle\sum_{k=0}^{i-1} \frac{(-1)^k\binom{i-1}{k}}{k+1} &\Big\{\left[\psi((k+1)\alpha+1)+\gamma\right]^2 \\ &-\psi'((k+1)\alpha+1)+\pi^2/6\Big\} \end{aligned} \tag{14}$$

### 3.1. Likelihood function

The likelihood function for $MERSS_{\text{minimum}}$ is,

$$\begin{aligned} L_{MERSS,\text{minimum}} &= \prod_{j=1}^{h}\prod_{i=1}^{m} f_{X'_{j(1:i)}}(v_{ji}) \\ &= \prod_{j=1}^{h}\prod_{i=1}^{m} \alpha\beta i\left(1-e^{-\beta v_{ji}}\right)^{\alpha-1}\left[1-(1-e^{-\beta v_{ji}})^{\alpha}\right]^{i-1} e^{-\beta v_{ji}} \end{aligned}$$

and the likelihood function for $MERSS_{\text{maximum}}$ is,

$$\begin{aligned} L_{MERSS,\text{maximum}} &= \prod_{j=1}^{h}\prod_{i=1}^{m} f_{X_{j(i:i)}}(u_{ji}) \\ &= \prod_{j=1}^{h}\prod_{i=1}^{m} \alpha\beta i\left(1-e^{-\beta u_{ji}}\right)^{i\alpha-1} e^{-\beta u_{ji}} \end{aligned}$$

Thereforethe likelihood function under the scheme of MERSS is,

$$\begin{aligned} L_{MERSS} &= \prod_{j=1}^{h}\prod_{i=1}^{m}\left\{f_{X'_{j(1:i)}}(v_{ji})\right\}\left\{f_{X_{j(i:i)}}(u_{ji})\right\} \\ &= \prod_{j=1}^{h}\prod_{i=1}^{m}\left[\alpha\beta i\left(1-e^{-\beta v_{ji}}\right)^{\alpha-1}\left\{1-(1-e^{-\beta v_{ji}})^{\alpha}\right\}^{i-1} e^{-\beta v_{ji}}\right] \\ &\qquad\qquad \left[\alpha\beta i\left(1-e^{-\beta u_{ji}}\right)^{i\alpha-1} e^{-\beta u_{ji}}\right] \end{aligned}$$

and log-likelihood,
$\log L_{MERSS}$

$$\begin{aligned} &= \sum_{j=1}^{h}\sum_{i=1}^{m}\left[\log\left\{f_{X'_{j(1:i)}}(v_{ji})\right\}+\log\left\{f_{X_{j(i:i)}}(u_{ji})\right\}\right] \\ &= \sum_{j=1}^{h}\sum_{i=1}^{m}\left[\log\left\{\alpha\beta i\left(1-e^{-\beta v_{ji}}\right)^{\alpha-1}\left\{1-(1-e^{-\beta v_{ji}})^{\alpha}\right\}^{i-1} e^{-\beta v_{ji}}\right\}\right. \\ &\qquad\qquad \left.+\log\left\{\alpha\beta i\left(1-e^{-\beta u_{ji}}\right)^{i\alpha-1} e^{-\beta u_{ji}}\right\}\right] \\ &= \sum_{j=1}^{h}\sum_{i=1}^{m}\left[\left\{\log i+(\alpha-1)\log\left(1-e^{-\beta v_{ji}}\right)+(i-1)\log\left(1-(1-e^{-\beta v_{ji}})^{\alpha}\right)\right.\right. \\ &\qquad\qquad \left.\left. -\beta v_{ji}\right\}+\left\{\log i+(i\alpha-1)\log\left(1-e^{-\beta u_{ji}}\right)-\beta u_{ji}\right\}\right]+C \qquad (15) \end{aligned}$$

where $C = 2mh\left\{\log\alpha+\log\beta\right\}$

### 3.2. ML estimates

Differentiating (15) with respect to $\alpha$ we get,

$$\begin{aligned} \frac{\partial \text{log} L_{MERSS}}{\partial\alpha} = \frac{2mh}{\alpha} + \sum_{j=1}^{h}\sum_{i=1}^{m}\left[\log\left(1-e^{-\beta v_{ji}}\right)\left\{\frac{1-i\left(1-e^{-\beta v_{ji}}\right)^{\alpha}}{1-(1-e^{-\beta v_{ji}})^{\alpha}}\right\}\right. \\ \left.+\left\{i\text{log}\left(1-e^{-\beta u_{ji}}\right)\right\}\right] \qquad (16) \end{aligned}$$

Differentiating (15) with respect to $\beta$ we get,

$$\frac{\partial \log L_{MERSS}}{\partial \beta}$$

$$= \frac{2mh}{\beta} + \sum_{j=1}^{h}\sum_{i=1}^{m}\Bigg[ v_{ji}\left\{ \frac{e^{-\beta v_{ji}}}{(1-e^{-\beta v_{ji}})} \frac{(\alpha-1)-(i\alpha-1)\left(1-e^{-\beta v_{ji}}\right)^{\alpha}}{(1-(1-e^{-\beta v_{ji}})^{\alpha})} - 1 \right\}$$
$$ - u_{ji}\left\{ \frac{(i\alpha-1)e^{-\beta u_{ji}}}{(1-e^{-\beta u_{ji}})} - 1 \right\} \Bigg] \tag{17}$$

We observe that the equations (16) and (17) can not be solved simultaneously to get closed form solution for $\alpha$ and $\beta$. Therefore, we solve numerically these equations simulataneously using R software (Henningsen and Toomet (2011)).

### 3.3. The observed Fisher information

Differentiating (16) with respect to $\alpha$ we get

$$\frac{\partial^2 \mathrm{log} L_{MERSS}}{\partial \alpha^2} = -\frac{2mh}{\alpha^2} - \sum_{j=1}^{h}\sum_{i=1}^{m}\left[ \frac{(i-1)(1-e^{-\beta v_{ji}})^{\alpha}\left\{\log(1-e^{-\beta v_{ji}})\right\}^2}{(1-(1-e^{-\beta v_{ji}})^{\alpha})^2} \right] \tag{18}$$

Differentiating (17) with respect to $\beta$ we get

$$\frac{\partial^2 \log L_{MERSS}}{\partial \beta^2} = -\frac{2mh}{\beta^2} + \sum_{j=1}^{h}\sum_{i=1}^{m}\Bigg[ \left\{ \frac{e^{-\beta v_{ji}} v_{ji}}{(1-e^{-\beta v_{ji}})} \right\}^2$$
$$\left\{ \frac{(\alpha-1)-(\alpha i(i+1)-\alpha-1)\left(1-e^{-\beta v_{ji}}\right)^{\alpha}}{(1-(1-e^{-\beta v_{ji}})^{\alpha})^2} \right\}$$
$$ - \left\{ \frac{(i\alpha-1)e^{-\beta u_{ji}} u_{ji}^2}{(1-e^{-\beta u_{ji}})^2} \right\} \Bigg] \tag{19}$$

and differentiating (16) with respect to $\beta$ we get

$$\frac{\partial^2 \mathrm{log} L_{MERSS}}{\partial \alpha \partial \beta} = \sum_{j=1}^{h}\sum_{i=1}^{m}\Bigg[ \left\{ \frac{e^{-\beta vji} v_{ji}}{1-e^{-\beta v_{ji}}} \right\}$$
$$\left\{ \frac{1-i\left(1-e^{-\beta v_{ji}}\right)^{\alpha}}{1-(1-e^{-\beta v_{ji}})^{\alpha}} - \frac{\alpha(i-1)\left(1-e^{-\beta v_{ji}}\right)^{\alpha}\log\left(1-e^{-\beta v_{ji}}\right)}{(1-(1-e^{-\beta v_{ji}})^{\alpha})^2} \right\}$$
$$ - \left\{ i\frac{e^{-\beta uji} u_{ji}}{1-e^{-\beta u_{ji}}} \right\} \Bigg] \tag{20}$$

We can compute numerically the elements of the observed Fisher information matrix and Variance -Covariance matrix for $\theta = (\alpha, \beta)$

$$I(\hat{\theta}) = \begin{bmatrix} -\dfrac{\partial^2 \mathrm{log} L_{MERSS}}{\partial \alpha^2} & -\dfrac{\partial^2 \mathrm{log} L_{MERSS}}{\partial \alpha \partial \beta} \\ -\dfrac{\partial^2 \mathrm{log} L_{MERSS}}{\partial \beta \partial \alpha} & -\dfrac{\partial^2 \log L_{MERSS}}{\partial \beta^2} \end{bmatrix}_{(\hat{\alpha},\hat{\beta})}$$

and $I^{-1}(\hat{\theta})$ respectively.

### 3.4. Fisher information

In this section, we obtain the Fisher information under the MERSS scheme. Following Azzalini (1996), the sample based on $MERSS_{\text{maximum}}$ and sample based on $MERSS_{\text{minimum}}$ are independent, therefore under certain regularity conditions the Fisher Information of MERSS scheme is given by

$$I_{MERSS}(\alpha, \beta) = I_{MERSS,\text{maximum}}(\alpha, \beta) + I_{MERSS,\text{minimum}}(\alpha, \beta) \tag{21}$$

The components of matrix $I_{MERSS,\text{maximum}}(\alpha, \beta)$ are
$\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$.

Where,

$$I_{11} = -E\left[\frac{\partial^2 \mathrm{log} L_{MERSS,\max}}{\partial \alpha^2}\right] = \frac{hm}{\alpha^2} \tag{22}$$

$$\begin{aligned} I_{22} &= -E\left[\frac{\partial^2 \log L_{MERSS,\max}}{\partial \beta^2}\right] \\ &= \frac{hm}{\beta^2} + \sum_{j=1}^{h}\sum_{i=1}^{m} \frac{(i\alpha)(i\alpha-1)}{\beta^2}\left[\textstyle\sum_{k=0}^{\infty}(-1)^k \dbinom{i\alpha-3}{k}\dfrac{2}{(k+2)^3}\right] \end{aligned} \tag{23}$$

$$\begin{aligned} I_{21} &= -E\left[\frac{\partial^2 \mathrm{log} L_{MERSS,\max}}{\partial \beta \partial \alpha}\right] \\ &= -\frac{\alpha}{\beta}\sum_{j=1}^{h}\sum_{i=1}^{m} i^2 \left[\textstyle\sum_{k=0}^{\infty}(-1)^k \dbinom{i\alpha-2}{k}\dfrac{2}{(k+2)^2}\right] \end{aligned} \tag{24}$$

The components of matrix $I_{MERSS,\text{minimum}}(\alpha, \beta) = \begin{bmatrix} I'_{11} & I'_{12} \\ I'_{21} & I'_{22} \end{bmatrix}$

**Case:** $i = 2$

$$\begin{aligned}I'_{11} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\alpha^2}\right]\\ &= \frac{h}{\alpha^2}\left[1 - 4\sum_{k=0}^{\infty}(-1)^k \frac{1}{(k+2)^3}\right]\end{aligned} \tag{25}$$

$$\begin{aligned}I'_{22} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta^2}\right]\\ &= \frac{h}{\beta^2} + \frac{2h\alpha(\alpha-1)}{\beta^2}\\ &\qquad \frac{1}{(\alpha-2)}\left[\{\psi(\alpha-1)-\psi(1)\}^2 - \{\psi'(\alpha-1)-\psi'(1)\}\right]\\ &\qquad - \frac{1}{2(\alpha-1)}\left[\{\psi(2\alpha-1)-\psi(1)\}^2 - \{\psi'(2\alpha-1)-\psi'(1)\}\right]\\ &\qquad - \frac{1}{(\alpha-1)}\left[\{\psi(\alpha)-\psi(1)\}^2 - \{\psi'(\alpha)-\psi'(1)\}\right]\\ &\qquad - \frac{1}{(2\alpha-1)}\left[\{\psi(2\alpha)-\psi(1)\}^2 - \{\psi'(2\alpha)-\psi'(1)\}\right]\\ &\qquad\qquad + 2h\left[\alpha E'_1 - E'_2\right],\end{aligned} \tag{26}$$

where

$$\begin{aligned}E'_1 &= 2\alpha\sum_{k=0}^{\infty}\Bigg[\frac{1}{\alpha(k+2)-2}\Big(\{\psi(\alpha(k+2)-1)-\psi(1)\}^2 - \{\psi'(\alpha(k+2)-1)-\\ &\quad \psi'(1)\}\big) - \frac{2}{\alpha(k+2)-1}\Big(\{\psi(\alpha(k+2))-\psi(1)\}^2 - \{\psi'(\alpha(k+2))-\psi'(1)\}\Big)\\ &\quad + \frac{1}{\alpha(k+2)}\Big(\{\psi(\alpha(k+2))-\psi(1)\}^2 - \{\psi'(\alpha(k+2))-\psi'(1)\}\Big)\Bigg]\\ E'_2 &= \frac{2\alpha}{2\alpha-1}\left[\{\psi(2\alpha)-\psi(1)\}^2 - \{\psi'(2\alpha)-\psi'(1)\}\right]\\ &\qquad - \left[\{\psi(2\alpha+1)-\psi(1)\}^2 - \{\psi'(2\alpha+1)-\psi'(1)\}\right]\end{aligned}$$

$$\begin{aligned}I'_{12} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta\partial\alpha}\right]\\ &= -h\left[E'_3 - \{E'_4 + E'_5\}\right]\end{aligned}$$

Where

$$E_3' = \frac{2\alpha}{\alpha-1}\left(\{\psi(\alpha)-\psi(1)\}\right) - \frac{2\alpha}{2\alpha-1}\left(\{\psi(2\alpha)-\psi(1)\}\right) - \mu_{j(1:2)}$$

$$E_4' = \frac{2\alpha}{2\alpha-1}\left(\{\psi(2\alpha)-\psi(1)\}\right) - \left(\{\psi(2\alpha+1)-\psi(1)\}\right)$$

$$E_5' = \frac{2\alpha^2}{\beta}\sum_{l=1}^{\infty}\sum_{k=0}^{\infty}\frac{1}{l}\left[\frac{1}{(\alpha(k+2)+l)^2} - \frac{1}{(\alpha(k+2)+l-1)^2}\right]$$

**Case:** $i = 3$

$$\begin{aligned} I_{11}' &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\alpha^2}\right] \\ &= \frac{5h}{2\alpha^2} \end{aligned} \tag{27}$$

$$\begin{aligned} I_{22}' &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta^2}\right] \\ &= \frac{h}{\beta^2} + \frac{3h\alpha(\alpha-1)}{\beta^2}\frac{1}{(\alpha-1)}\left[\{\psi(\alpha)-\psi(1)\}^2 - \{\psi'(\alpha)-\psi'(1)\}\right] \\ &\qquad - \frac{2}{(2\alpha-1)}\left[\{\psi(2\alpha)-\psi(1)\}^2 - \{\psi'(2\alpha)-\psi'(1)\}\right] \\ &\qquad - \frac{1}{(3\alpha-1)}\left[\{\psi(3\alpha)-\psi(1)\}^2 - \{\psi'(3\alpha)-\psi'(1)\}\right] \\ &\qquad - \left[\{\psi(\alpha+1)-\psi(1)\}^2 - \{\psi'(\alpha+1)-\psi'(1)\}\right] \\ &\qquad + \frac{1}{\alpha}\left[\{\psi(2\alpha+1)-\psi(1)\}^2\{\psi'(2\alpha+1)-\psi'(1)\}\right] \\ &\qquad - \frac{1}{3\alpha}\left[\{\psi(3\alpha+1)-\psi(1)\}^2\{\psi'(3\alpha+1)-\psi'(1)\}\right] \\ &\qquad + 2h\left[\alpha E_1' - E_2'\right], \end{aligned} \tag{28}$$

where

$$E_1' = \frac{3\alpha}{\beta^2}\left[\frac{1}{2\alpha-2}\left[\{\psi(2\alpha-1)-\psi(1)\}^2-\{\psi'(2\alpha-1)-\psi'(1)\}\right]\right.$$
$$-\frac{2}{2\alpha-1}\left[\{\psi(2\alpha)-\psi(1)\}^2-\{\psi'(2\alpha)-\psi'(1)\}\right]$$
$$\left.+\frac{1}{2\alpha}\left[\{\psi(2\alpha+1)-\psi(1)\}^2-\{\psi'(2\alpha+1)-\psi'(1)\}\right]\right]$$
$$E_2' = \frac{3\alpha}{\beta^2}\left[\frac{1}{2\alpha-2}\left[\{\psi(2\alpha-1)-\psi(1)\}^2-\{\psi'(2\alpha-1)-\psi'(1)\}\right]\right.$$
$$-\frac{1}{3\alpha-2}\left[\{\psi(3\alpha-1)-\psi(1)\}^2-\{\psi'(3\alpha-1)-\psi'(1)\}\right]$$
$$-\frac{1}{2\alpha-1}\left[\{\psi(2\alpha)-\psi(1)\}^2-\{\psi'(2\alpha)-\psi'(1)\}\right]$$
$$\left.+\frac{1}{3\alpha-1}\left[\{\psi(3\alpha)-\psi(1)\}^2-\{\psi'(3\alpha)-\psi'(1)\}\right]\right]$$

$$I_{12}' = -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta\partial\alpha}\right]$$
$$= -2h\left[E_3'-\{E_4'+E_5'\}\right],$$

where

$$E_3' = \frac{3\alpha}{\beta}\left[\frac{1}{\alpha-1}\left[\{\psi(\alpha)-\psi(1)\}\right]-\frac{2}{2\alpha-1}\left[\{\psi(2\alpha)-\psi(1)\}\right]\right.$$
$$\left.+\frac{1}{3\alpha-1}\left[\{\psi(3\alpha)-\psi(1)\}\right]-\mu_{j(1:3)}\right]$$

$$E_4' = \frac{3\alpha}{\beta}\left[\frac{1}{2\alpha-1}\left[\{\psi(2\alpha)-\psi(1)\}\right]-\frac{1}{3\alpha-1}\left[\{\psi(3\alpha)-\psi(1)\}\right]\right.$$
$$\left.-\frac{1}{2\alpha}\left[\{\psi(2\alpha+1)-\psi(1)\}\right]+\left[\{\psi(3\alpha)-\psi(1)\}\right]\right]$$

$$E_5' = \frac{3\alpha^2}{\beta}\sum_{l=1}^{\infty}\frac{1}{l}\left[\frac{1}{(2\alpha+l)^2}-\frac{1}{(2\alpha+l-1)^2}\right]$$

In general,

$$\begin{aligned} I'_{11} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\alpha^2}\right] \\ &= \frac{hm}{\alpha^2} + \sum_{j=1}^{h}\sum_{i=4}^{m}\frac{i(i-1)}{\alpha^2}\sum_{k=0}^{i-3}(-1)^k\binom{i-3}{k}\frac{2}{(k+2)^3} \end{aligned} \tag{29}$$

$$\begin{aligned} I'_{22} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta^2}\right] \\ &= \frac{hm}{\beta^2} + \sum_{j=1}^{h}\sum_{i=4}^{m}\frac{i\alpha(\alpha-1)}{\beta^2} \\ &\qquad \sum_{k=0}^{i-1}(-1)^k\binom{i-1}{k}\frac{1}{(\alpha(k+1)-2)}\left[\{\psi(\alpha(k+1)-1)-\psi(1)\}^2\right. \\ &\qquad\qquad -\{\psi'(\alpha(k+1)-1)-\psi'(1)\}] \\ &\qquad\qquad -\frac{1}{(\alpha(k+1)-1)}\left[\{\psi(\alpha(k+1))-\psi(1)\}^2\right. \\ &\qquad\qquad\qquad -\{\psi'(\alpha(k+1))-\psi'(1)\}] \\ &- (i-1)\alpha\sum_{j=1}^{h}\sum_{i=1}^{m}[\alpha E'_1 - E'_2] \end{aligned} \tag{30}$$

$$\begin{aligned} I'_{12} &= -E\left[\frac{\partial^2 \log L_{MERSS,\min}}{\partial\beta\partial\alpha}\right] \\ &= -\sum_{j=1}^{h}\sum_{i=4}^{m}[E'_3 - (i-1)\{E'_4 + E'_5\}] \end{aligned} \tag{31}$$

where

$$\begin{aligned} E'_1 = \frac{i\alpha}{\beta^2}\sum_{k=0}^{i-3}(-1)^k\binom{i-3}{k}&\left[\frac{1}{\alpha(k+2)-1}\{\psi(\alpha(k+2))-\psi(1)\}^2\right. \\ &-\{\psi'(\alpha(k+2))-\psi'(1)\}+\frac{2}{\alpha(k+2)}\{\psi(\alpha(k+2)+1)-\psi(1)\}^2 \\ &-\{\psi'(\alpha(k+2)+1)-\psi'(1)\}+\frac{1}{\alpha(k+2)+1}\{\psi(\alpha(k+2)+2) \\ &\left.-\psi(1)\}^2-\{\psi'(\alpha(k+2)+2)-\psi'(1)\}\right] \end{aligned}$$

$$E_2' = \frac{i\alpha}{\beta^2}\sum_{k=0}^{i-2}(-1)^k\binom{i-2}{k}\left[\frac{1}{\alpha(k+2)-1}\left\{\psi(\alpha(k+2))-\psi(1)\right\}^2\right.$$
$$-\left\{\psi'(\alpha(k+2))-\psi'(1)\right\}+\frac{1}{\alpha(k+2)}\left\{\psi(\alpha(k+2)+1)-\psi(1)\right\}^2$$
$$\left.-\left\{\psi'(\alpha(k+2)+1)-\psi'(1)\right\}\right]$$
$$E_3' = \frac{i\alpha}{\beta}\sum_{k=0}^{i-1}(-1)^k\binom{i-1}{k}\frac{\psi(\alpha(k+1))-\psi(1)}{\alpha(k+1)-1}-\mu_{j(1:i)}$$
$$E_4' = \frac{i\alpha}{\beta}\sum_{k=0}^{i-2}(-1)^k\binom{i-2}{k}\left[\frac{\psi(\alpha(k+2))-\psi(1)}{\alpha(k+2)-1}-\frac{\psi(\alpha(k+2)+1)-\psi(1)}{\alpha(k+2)}\right]$$
$$E_5' = \frac{i\alpha^2}{\beta}\sum_{l=1}^{\infty}\sum_{k=0}^{i-3}(-1)^k\binom{i-3}{k}\frac{1}{l}\left[\frac{1}{(\alpha(k+2)+l)^2}-\frac{1}{(\alpha(k+2)+l-1)^2}\right]$$

We note that above expectations exist for $\alpha > 2$ as $\psi(\cdot)$ and $\psi'(\cdot)$ exist and finite.

## 4. Ranked set sample

Let $X_{j1}, X_{j2}, \ldots, X_{jm}, X_{j(m+1)}, X_{j(m+2)}, \ldots, X_{jm^2}$ be independent random variables having the same distribution given in (1) of cycle $j = 1, 2, \ldots, h$. Then from $i$-th set $\{X_{j((i-1)m+1)}, X_{j((i-1)m+2)j},$
$\ldots, X_{j(im)}\}$, $X_{ji(i:m)}, i = 1, 2, \ldots, m$ denote $i-$th order statistic assuming error free rankings. Let $x_{ji}$ denote observed value of $X_{ji(i:m)}$, then the pdf of $i$-th order statistic,

$$\begin{aligned} f_{X_{ji(i:m)}}(x_{ji}) &= \frac{1}{B(i,m-i+1)}F^{i-1}(x_{ji})(1-F(x_{ji}))^{m-i}f(x_{ji}) \\ &= \frac{1}{B(i,m-i+1)}\left((1-e^{-\beta x_{ji}})^{\alpha}\right)^{i-1}\left(1-(1-e^{-\beta x_{ji}})^{\alpha}\right)^{m-i} \\ &\qquad \alpha\beta(1-e^{-\beta x_{ji}})^{\alpha-1}e^{-\beta x_{ji}}, \alpha>0, \beta>0 \end{aligned} \tag{32}$$

From (7) and (8) the first order and second order moments of $X_{ji(i:m)}$ respectively are $\mu_{ji(i:m)} = E\{X_{ji(i:m)}\} = \frac{1}{\beta}E\{Y_{(i:m)}\}$, and $\mu^2_{ji(i:m)} = E\{X^2_{ji(i:m)}\} = \frac{1}{\beta^2}E\{Y^2_{(i:m)}\}$. Then the log-likelihood function is

$$\begin{aligned} \log L_{RSS}(\alpha,\beta) &= \sum_{j=1}^{h}\sum_{i=1}^{m}\log f_{X_{ji(i:m)}}(x_{ji}) \\ &= C_4 + mh(\log\alpha+\log\beta) \\ &\quad + \sum_{j=1}^{h}\sum_{i=1}^{m}\Big\{(i\alpha-1)\log\left(1-e^{-\beta x_{ji}}\right) \\ &\qquad +(m-i)\log\left(1-(1-e^{-\beta x_{ji}})^{\alpha}\right)-\beta x_{ji}\Big\}, \end{aligned} \tag{33}$$

where $C_4 = mh \log\left(\frac{1}{B(i, m-i+1)}\right)$.

### 4.1. The likelihood equations

Differentiating (33) with respect to $\alpha$

$$\frac{\partial \log L_{RSS}(\alpha,\beta)}{\partial\alpha} = \frac{mh}{\alpha} + \sum_{j=1}^{h}\sum_{i=1}^{m} \log(1-e^{-\beta x_{ji}}) \left\{ \frac{i - m(1-e^{-\beta x_{ji}})}{(1-(1-e^{-\beta x_{ji}})^{\alpha})} \right\} \tag{34}$$

Differentiating (33) with respect to $\beta$

$$\frac{\partial \log L_{RSS}(\alpha,\beta)}{\partial\beta} = \frac{mh}{\beta} + \sum_{j=1}^{h}\sum_{i=1}^{m} x_{ji}\left[\frac{e^{-\beta x_{ji}}}{(1-e^{-\beta x_{ji}})}\left\{\frac{i-m(1-e^{-\beta x_{ji}})^{\alpha}}{1-(1-e^{-\beta x_{ji}})^{\alpha}} - 1\right\} - 1\right] \tag{35}$$

We observe that the equations (34) and (35) can not be solved simultaneously to get a closed-form solution for $\alpha$ and $\beta$. Therefore, we solve numerically these equations simulataneously using R software (Henningsen and Toomet (2011)).

Differentiating (34) with respect to $\alpha$

$$\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\alpha^2} = -\frac{mh}{\alpha^2} - \sum_{j=1}^{h}\sum_{i=1}^{m} \left\{ \frac{(m-i)(1-e^{-\beta x_{ji}})\left(\log(1-e^{-\beta x_{ji}})\right)^2}{\left(1-(1-e^{-\beta x_{ji}})^{\alpha}\right)^2} \right\} \tag{36}$$

Differentiating (35) with respect to $\beta$

$$\begin{aligned}\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\beta^2} = -\frac{mh}{\beta^2} - \sum_{j=1}^{h}\sum_{i=1}^{m} x_{ji}^2 \frac{e^{-\beta x_{ji}}}{(1-e^{-\beta x_{ji}})^2}\Bigg[\left\{\frac{i-m(1-e^{-\beta x_{ji}})^{\alpha}}{1-(1-e^{-\beta x_{ji}})^{\alpha}} - 1\right\} \\ + e^{-\beta x_{ji}}\frac{\alpha(m+i)(1-e^{-\beta x_{ji}})^{\alpha}}{\left(1-(1-e^{-\beta x_{ji}})^{\alpha}\right)^2}\Bigg]\end{aligned} \tag{37}$$

and differentiating (34) with respect to $\beta$

$$\begin{aligned}\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\alpha\partial\beta} = \sum_{j=1}^{h}\sum_{i=1}^{m} \frac{x_{ji}e^{-\beta x_{ji}}}{(1-e^{-\beta x_{ji}})}\Bigg[\left\{\frac{i-m(1-e^{-\beta x_{ji}})^{\alpha}}{(1-(1-e^{-\beta x_{ji}})^{\alpha})}\right\} \\ - \frac{\alpha(m-i)(1-e^{-\beta x_{ji}})^{\alpha}\log(1-e^{-\beta x_{ji}})}{\left(1-(1-e^{-\beta x_{ji}})^{\alpha}\right)^2}\Bigg]\end{aligned} \tag{38}$$

We can numerically compute elements of the observed Fisher information matrix,

$$I(\hat{\theta}) = \begin{bmatrix} -\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\alpha^2} & -\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\alpha\partial\beta} \\ -\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\beta\partial\alpha} & -\frac{\partial^2 \log L_{RSS}(\alpha,\beta)}{\partial\beta^2} \end{bmatrix}_{(\hat{\alpha},\hat{\beta})}$$

and variance-covariance matrix $I^{-1}(\hat{\theta})$

## 5. Comparing information of MERSS and RSS schemes

In this section, we compare the information of MERSS and RSS schemes respectively to estimate shape and scale parameters . Let $\hat{\alpha}_{\text{RSS}}$, $\hat{\beta}_{\text{RSS}}$ and $\hat{\alpha}_{\text{MERSS}}$, $\hat{\beta}_{\text{MERSS}}$ are the estimates of shape and scale parameters under RSS and MERSS schemes respectively. Assuming $\dfrac{\hat{\alpha}_{\text{RSS}}}{\hat{\alpha}_{\text{MERSS}}} \to 1$ and $\dfrac{\hat{\beta}_{\text{RSS}}}{\hat{\beta}_{\text{MERSS}}} \to 1$ then,

**RSS:**

$$I_\alpha(RSS) = \frac{mh}{\alpha^2} + \sum_{j=1}^{h}\sum_{i=1}^{m} \frac{(m-i)}{B(i,m-i+1)\alpha^2} \sum_{k=0}^{m-i-2} (-1)^k \binom{m-i-2}{k} \frac{2}{(k+i+1)^3} \tag{39}$$

$$\begin{aligned} I_\beta(RSS) = \frac{mh}{\beta^2} + (i\alpha-1)\sum_{j=1}^{h}\sum_{i=1}^{m} \frac{\alpha}{B(i,m-i+1)\beta^2} \sum_{s=0}^{m-i}(-1)^s \binom{m-i}{s} \times \\ \sum_{k=0}^{\infty}(-1)^k \binom{(i+s)\alpha-3}{k} \frac{2}{(k+2)^3} \\ + \alpha(m-i)\Bigg[\frac{\alpha^2}{B(i,m-i+1)\beta^2} \sum_{s=0}^{m-i-2}(-1)^s \binom{m-i-2}{s} \times \\ \sum_{k=0}^{\infty}(-1)^k \binom{(i+s+1)\alpha-3}{k} \frac{2}{(k+3)^3} \\ - \frac{\alpha}{B(i,m-i+1)\beta^2} \sum_{s=0}^{m-i-1}(-1)^s \binom{m-i-1}{s} \times \\ \sum_{k=0}^{\infty}(-1)^k \binom{(i+s+1)\alpha-3}{k} \frac{2}{(k+2)^3}\Bigg] \end{aligned} \tag{40}$$

**MERSS:**

$$I_\alpha(MERSS) = \frac{2hm}{\alpha^2} + \sum_{j=1}^{h}\sum_{i=1}^{m} \frac{i(i-1)}{\alpha^2} \sum_{k=0}^{i-3}(-1)^k \binom{i-3}{k} \frac{2}{(k+2)^3} \tag{41}$$

and $I_\beta(MERSS)$ is obtained in section (3.4). We note that $I_\beta(MERSS)$ exists and finite for $\alpha > 2$.

The computations are,

When sample size of RSS is $2mh$ and sample size of MERSS is $2mh$, then

$$\begin{aligned}
&\text{For } (m=3)\\
&I_\alpha(RSS) = h\left[\frac{12.6553}{\alpha^2} + \frac{60}{\alpha^2}\sum_{k=0}^{\infty}\frac{1}{(k+6)^3}\right] = \frac{13.6390h}{\alpha^2}\\
&I_\beta(RSS) = h\frac{36.47049}{\beta^2}, \quad I_\beta(MERSS) = h\frac{207.6518}{\beta^2} \quad \alpha = 2.5\\
&I_\beta(RSS) = h\frac{42.74969}{\beta^2}, \quad I_\beta(MERSS) = h\frac{97.40603}{\beta^2} \quad \alpha = 3.0\\
&\text{For}(m=4)\\
&I_\alpha(RSS) = h\left[\frac{158.0665}{\alpha^2} + \frac{112}{\alpha^2}\sum_{k=0}^{\infty}\frac{1}{(k+8)^3}\right] = \frac{159.0577h}{\alpha^2}\\
&I_\beta(RSS) = h\frac{65.12496}{\beta^2}, \quad I_\beta(MERSS) = h\frac{287.7424}{\beta^2} \quad \alpha = 2.5\\
&I_\beta(RSS) = h\frac{76.34466}{\beta^2}, \quad I_\beta(MERSS) = h\frac{147.3634}{\beta^2} \quad \alpha = 3.0
\end{aligned}$$

$$\left[\frac{I_\alpha(RSS)}{I_\alpha(MERSS)}\right]_{m=3} = 1.6416, \quad \left[\frac{I_\alpha(RSS)}{I_\alpha(MERSS)}\right]_{m=4} = 12.8073$$

For $\alpha = 2.5$

$$\left[\frac{I_\beta(RSS)}{I_\beta(MERSS)}\right]_{m=3} = 0.17563, \quad \left[\frac{I_\alpha(RSS)}{I_\alpha(MERSS)}\right]_{m=4} = 0.22633$$

For $\alpha = 3.0$

$$\left[\frac{I_\beta(RSS)}{I_\beta(MERSS)}\right]_{m=3} = 0.43888, \quad \left[\frac{I_\alpha(RSS)}{I_\alpha(MERSS)}\right]_{m=4} = 0.51807$$

We note that information for the shape parameter under the scheme of the RSS is greater than information for the shape parameter under the scheme of MERSS, and information for the scale parameter under the scheme of MERSS is greater than information for the scale parameter under the scheme of RSS.

### 5.1. Simulated output

In this section, we generate 100000 random numbers from the generalized exponential distribution given in (1) using known values of parameters $(\alpha, \beta)$ taking $\alpha = 0.8, \alpha = 1.2, \alpha = 1.6, \alpha = 1.8, \alpha = 2.2, \alpha = 3.8$ and $\beta = 0.25$. At each cycle we obtain MERSS sample, by selecting $m^2 + m$ units, SRS sample, by selecting $2m$ units, and RSS sample, by selecting

$4m^2$ units. The ML estimates $(\hat{\alpha}, \hat{\beta})$ are calculated numerically and the comparisons are determined by the mean squared error of estimates, when sample size of SRS, MERSS and RSS are equal to $2mh$ respectively. We define the efficiency of sampling scheme as,

$$E_{\text{Sampling}_1,\text{Sampling}_2}(\hat{\alpha}) = \frac{MSE(\hat{\alpha})_{\text{Sampling}_2}}{MSE(\hat{\alpha})_{\text{Sampling}_1}}$$

and

$$E_{\text{Sampling}_1,\text{Sampling}_2}(\hat{\beta}) = \frac{MSE(\hat{\beta})_{\text{Sampling}_2}}{MSE(\hat{\beta})_{\text{Sampling}_1}}$$

## 6. Conclusion

1. In this paper, we estimate the parameters of the generalized exponential distribution using moving extreme ranked sampling, ranked set sampling, and simple random sampling. Then we compared the estimates using the mean squared errors of the estimates.

2. From the simulation study, it is found that under the RSS scheme we get smaller MSE as compared to MSE obtained for the MERSS scheme and SRS scheme for both shape and scale parameters. (Annexure Table A2)

## Acknowledgements

The authors would like to thank the editor and referees for carefully reading the paper and for constructive suggestions and for their help in improving the paper.

## References

Al-Odat, M. T. and Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, **10**, 137–146.

Azzalini, A. (1996). *Statistical Inference Based on the Likelihood.* CRC Press.

Chen, W. X., Long, C. X., Yang, R., and Yao, D. (2021). Maximum likelihood estimator of the location parameter under moving extremes ranked set sampling design. *Acta Mathematicae Applicatae Sinica, English Series*, **37**, 101–108.

Gupta, R. D. and Kundu, D. (1999). Generalized exponential distributions. *Australian & New Zealand Journal of Statistics*, **41**, 173–188.

Gupta, R. D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **43**, 117–130.

He, X. F., Chen, W. X., and Yang, R. (2021). Log-logistic parameters estimation using moving extremes ranked set sampling design. *Applied Mathematics-A Journal of Chinese Universities*, **36**, 99–113.

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, **26**.

McIntyre, G. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385.

Patil, G., Sinha, A., and Taillie, C. (1994). *5 Ranked Set Sampling*, pages 167–200. Elsevier.

Wangxue, C., Yang, R., Long, C., and Yao, D. (2019). Pareto parameters estimation using moving extremes ranked set sampling. *Statistical Papers*, **62**, 1195–1211.

## ANNEXURE

**Table A1: Biases and MSE's of ML estimates of shape and scale parameters**

| Sampling | $(m, h)$ | $(\alpha, \beta)$ | $\hat{\alpha}$ | MSE($\hat{\alpha}$) | Bias($\hat{\alpha}$) | $\hat{\beta}$ | MSE($\hat{\beta}$) | Bias($\hat{\beta}$) |
|---|---|---|---|---|---|---|---|---|
| SRS | (3,3) | (0.8,0.25) | 0.9487606 | 0.1736299 | 0.1487606 | 0.2956023 | 0.0133619 | 0.0456023 |
| RSS | | | 0.8817586 | 0.0584896 | 0.0817586 | 0.2733177 | 0.0055097 | 0.0233177 |
| MERSS | | | 0.7964535 | 0.0717147 | -0.0035465 | 0.2423628 | 0.0071007 | -0.0076372 |
| SRS | (3,3) | (1.2,0.25) | 1.445111 | 0.4235251 | 0.2451109 | 0.2874173 | 0.0102904 | 0.0374173 |
| RSS | | | 1.336743 | 0.1457796 | 0.1367427 | 0.2681817 | 0.0039607 | 0.0181817 |
| MERSS | | | 1.202511 | 0.1835160 | 0.0025113 | 0.2452994 | 0.0062704 | -0.0047006 |
| SRS | (3,3) | (1.6,0.25) | 2.022529 | 1.1205389 | 0.4225288 | 0.2816831 | 0.0075248 | 0.0316831 |
| RSS | | | 1.806412 | 0.3668359 | 0.2064123 | 0.2668984 | 0.0036996 | 0.0168984 |
| MERSS | | | 1.634325 | 0.5443902 | 0.0343248 | 0.2458775 | 0.0057707 | -0.0041225 |
| SRS | (3,3) | (1.8,0.25) | 2.320982 | 2.1188691 | 0.5209824 | 0.2832089 | 0.0086354 | 0.0332089 |
| RSS | | | 1.994951 | 0.4090394 | 0.1949514 | 0.2623048 | 0.0028807 | 0.0123048 |
| MERSS | | | 1.797398 | 0.5843733 | -0.0026018 | 0.2401457 | 0.0046996 | -0.0098543 |
| SRS | (3,3) | (2.2,0.25) | 2.741057 | 2.1544205 | 0.5410568 | 0.2753242 | 0.0061387 | 0.0253242 |
| RSS | | | 2.505455 | 0.7952925 | 0.3054547 | 0.2648720 | 0.0029338 | 0.0148720 |
| MERSS | | | 2.167485 | 0.8985339 | -0.0325153 | 0.2383081 | 0.0046774 | -0.0116919 |
| SRS | (3,3) | (3.8,0.25) | 5.034128 | 10.224889 | 1.2341283 | 0.2764764 | 0.0054923 | 0.0264764 |
| RSS | | | 4.414143 | 3.255205 | 0.6141435 | 0.2632917 | 0.0024231 | 0.0132917 |
| MERSS | | | 3.895133 | 4.580507 | 0.0951328 | 0.2400658 | 0.0040669 | -0.0099342 |

### Table A1(Continued): Biases and MSE's of ML estimates of shape and scale parameters

| Sampling | $(m,h)$ | $(\alpha,\beta)$ | $\hat{\alpha}$ | MSE($\hat{\alpha}$) | Bias($\hat{\alpha}$) | $\hat{\beta}$ | MSE($\hat{\beta}$) | Bias($\hat{\beta}$) |
|---|---|---|---|---|---|---|---|---|
| SRS | (4,2) | (0.8,0.25) | 0.9653919 | 0.1879596 | 0.1653919 | 0.2988921 | 0.0166438 | 0.0488921 |
| RSS | | | 0.8787682 | 0.0488045 | 0.0787682 | 0.2753540 | 0.0054171 | 0.0253540 |
| MERSS | | | 0.7710374 | 0.0690403 | -0.0289626 | 0.2380351 | 0.0083744 | -0.0119649 |
| SRS | (4,2) | (1.2,0.25) | 1.523014 | 0.5874794 | 0.3230142 | 0.2970299 | 0.0125121 | 0.0470299 |
| RSS | | | 1.312745 | 0.1347717 | 0.1127447 | 0.2664989 | 0.0038060 | 0.0164989 |
| MERSS | | | 1.172552 | 0.1804458 | -0.0274481 | 0.2371569 | 0.0060172 | -0.0128431 |
| SRS | (4,2) | (1.6,0.25) | 1.990566 | 1.0587583 | 0.3905659 | 0.2806856 | 0.0085753 | 0.0306856 |
| RSS | | | 1.786667 | 0.3209858 | 0.1866674 | 0.2664005 | 0.0033632 | 0.0164005 |
| MERSS | | | 1.556896 | 0.4012962 | -0.0431041 | 0.2366258 | 0.0056646 | -0.0133742 |
| SRS | (4,2) | (1.8,0.25) | 2.300448 | 1.5793808 | 0.5004485 | 0.2871264 | 0.0087790 | 0.0371264 |
| RSS | | | 2.033651 | 0.4139100 | 0.2336511 | 0.2685046 | 0.0033391 | 0.0185046 |
| MERSS | | | 1.761394 | 0.6294839 | -0.0386061 | 0.2354936 | 0.0057293 | -0.0145064 |
| SRS | (4,2) | (2.2,0.25) | 2.840629 | 3.684262 | 0.6406294 | 0.2808774 | 0.0070519 | 0.0308774 |
| RSS | | | 2.427666 | 0.690291 | 0.2276657 | 0.2619086 | 0.0028063 | 0.0119086 |
| MERSS | | | 2.190393 | 1.325949 | -0.0096071 | 0.2392440 | 0.0054996 | -0.0107560 |
| SRS | (4,2) | (3.8,0.25) | 5.512445 | 30.686737 | 1.7124454 | 0.2776625 | 0.0069002 | 0.0276625 |
| RSS | | | 4.361663 | 2.803481 | 0.5616628 | 0.2609717 | 0.0022263 | 0.0109717 |
| MERSS | | | 3.758171 | 4.772074 | -0.0418289 | 0.2353198 | 0.0043514 | -0.0146802 |

### Table A2: Efficiency of estimators of shape and scale parameters under MERSS and RSS schemes

| $(\alpha,\beta)$ | $(m,h)$ | $E_{\text{MERSS,SRS}}(\hat{\alpha})$ | $E_{\text{MERSS,SRS}}(\hat{\beta})$ | $E_{\text{RSS,MERSS}}(\hat{\alpha})$ | $E_{\text{RSS,MERSS}}(\hat{\beta})$ | $E_{\text{RSS,SRS}}(\hat{\alpha})$ | $E_{\text{RSS,SRS}}(\hat{\beta})$ |
|---|---|---|---|---|---|---|---|
| (0.8,0.25) | (3,3) | 2.421120 | 1.881772 | 1.226110 | 1.288763 | 2.968560 | 2.425159 |
| | (4,2) | 2.722462 | 1.987462 | 1.414630 | 1.545919 | 3.851276 | 3.072456 |
| (1.2,0.25) | (3,3) | 2.307837 | 1.641107 | 1.258859 | 1.583154 | 2.905243 | 2.598127 |
| | (4,2) | 3.255711 | 2.079389 | 1.338900 | 1.580977 | 4.359071 | 3.287467 |
| (1.6,0.25) | (3,3) | 2.058338 | 1.303967 | 1.484016 | 1.559817 | 3.054605 | 2.033950 |
| | (4,2) | 2.638346 | 1.513840 | 1.250199 | 1.684289 | 3.298458 | 2.549744 |
| (1.8,0.25) | (3,3) | 3.625883 | 1.837476 | 1.428648 | 1.631409 | 5.180110 | 2.997674 |
| | (4,2) | 2.509009 | 1.532299 | 1.520823 | 1.715822 | 3.815759 | 2.629152 |
| (2.2,0.25) | (3,3) | 2.397706 | 1.312417 | 1.129816 | 1.594315 | 2.708966 | 2.092406 |
| | (4,2) | 2.778585 | 1.282257 | 1.920855 | 1.959733 | 5.337259 | 2.512882 |
| (3.8,0.25) | (3,3) | 2.232261 | 1.350488 | 1.407133 | 1.678387 | 3.141089 | 2.266642 |
| | (4,2) | 6.430482 | 1.585743 | 1.702196 | 1.954543 | 10.945941 | 3.099403 |

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 143–159

# Statistical Modeling of Temperature in Krishna District using Copula Analysis

**A. Rajini[1] and C. Jayalakshmi[2]**
*[1]Department of Mathematics and Statistics, Bhavan's Vivekananda College of Science, Humanities and Commerce, Sainikpuri, Secunderabad, Telangana, India*
*[2]Department of Statistics, Osmania University, Hyderabad, Telangana, India*

Received: 12 February 2022; Revised: 20 May 2022; Accepted: 18 June 2022

## Abstract

Examining the relationship between temperature and precipitation in the Krishna district and forecast temperature is the purpose of study. Krishna is a district in Andhra Pradesh Plateau region, and it was chosen because it is a densely populated area with significant towns and ports. The district's climate is tropical, with sweltering summers and mild winters. From 1901 to 2019, data was obtained from the Indian Meteorological Department in Pune. Data from 1901 to 1996 was used for training, while data from 1997 to 2019 was used for testing. Through Copula analysis, a model is built keeping in view the relationship between Temperature and Precipitation. The Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Square Error (NRMSE) for the best model in Krishna were determined to be 0.03 and 3.823 for the month of May, which has the highest temperature and precipitation dependency when compared to other months. A similar analysis is carried out for the months in which dependence is significant. It is found that five months interdependency coefficient is insignificant. The data was analyzed using R-software, and IBM SPSS statistics version 25 and the results were interpreted. The best Copula does not have to be the same for different datasets. Based on AIC and BIC criteria, the best Copula for Krishna was Gaussians Copula for the month of April and July, Rotated Gumbel 90 Copula for the month of May and September, Rotated Tawn type 2 270 Copula for the month of June, Rotated Gumbel 270 Copula for the month of August and Rotated Clayton 90 Copula for the month of October. Temperature simulated data was found to be very close to testing results. This article examines how Copula modelling can be used to predict temperature, which helps in planning agriculture and trading commodities. So far, this type of analysis and model fitting is not found in the literature for Krishna district in Andhra Pradesh. The temperature in this location may be accurately predicted using our fitted models.

*Key words:* Temperature; Precipitation; Copula analysis; Mean absolute percentage error; AIC; BIC.

## 1. Introduction

The atmospheric conditions like temperature, air pressure and moisture vary from one place to another. Due to changes in the climate, it is tough to predict drought and heavy rains at any time in any corner. Weather forecasts are essential warnings as they help protect life and wealth. Forecasted Temperature helps in planning agriculture and trading of

Corresponding Author: A. Rajini
Email: rajini.ststs@bhavansvc.ac.in

commodities. Temperature is a critical parameter in farming varieties of vegetables, fruits and pulses. Hence, there is a need to carry out continuous research on the influencing factors (Temperature and Precipitation) to meet the demand for an increased population. It has been established in the literature of research studies that Temperature and Precipitation have a deep interdependency.

Some research has been conducted in this direction. Dzupire *et al.* (2020), have used Copula analysis to identify interdependency patterns between Temperature and Precipitation. Lazoglou and Anagnostopoulou (2019) developed a joint distribution for the above two factors using Copula in the Mediterranean region. Similar studies have been carried out by Bezak *et al.*(2018) in Slovenia *et al.* (2020) in China, and Shaukat *et al.* (2020) in Pakistan. Mesbahzadeh *et al.* (2019) modelled Temperature and Precipitation for the Arid region using Copula analysis. In their study, Pandey *et al.* (2018) modelled interdependency between Rainfall and Temperature using Copula. This study was carried out in Agartala (humid region) and Bikaner (Arid region). Zscheischler *et al.* (2017), have inferred from their study that environmental change, Precipitation and Temperature are the significant factors that affect the nature of vulnerability influencing harvest. Crop yields are firmly vulnerable to outrageous atmospheres like a dry spell, floods, and warmth waves. Vergara *et al.* (2008) have examined how catastrophe risk modelling can be used in agriculture as a planning tool to predict the frequency and severity of future weather-related catastrophic events, allowing crop insurance firms and policymakers to better prepare for the financial effect of natural disasters. Zhang and Singh (2007) analyzed hydrology for examining the random factors dependence structure modelled independently by marginal distribution individually; Copula approach has been broadly utilized. Olesen and Bindi (2002) discussed and analyzed Temperature and Precipitation influence the duration of expanding season and plant creation (leaf territory and the photosynthetic productivity), respectively. A lot of literature is available on the impacts of Temperature and Precipitation on crop output. Therefore, we can understand the correlation between Precipitation and Temperature that keeps changing in different time periods. Nelsen (2007) demonstrated the factors between dependence structures Copulas are intended based on uniform marginal values.

These studies tried to forecast rainfall in humid, arid and Mediterranean environments. According to a literature survey, there were few pieces of research on the Plateau region and no temperature predictions using Copula analysis were provided. Our study aims to develop a bi-variate model for monthly average Temperature and Precipitation that can be used to simulate Temperature in the selected regions (Krishna district of Andhra Pradesh, which is a Plateau region). Copula analysis was shown to be the most appropriate methodology in this direction. The study's goal is to develop a Copula model that can be used to estimate Temperature in a specific region.

## 2. Data and methodology

The study used historical monthly average temperature and monthly average precipitation for 119 years, covering the period from 1901 to 2019, collected by the Indian Meteorological Department for the Krishna district. This district is chosen because it is a densely populated area with significant towns and ports. The district's climate is tropical, with extremely hot summers and mild winters.

### 2.1. Copula methodology

To work with Copulas, one must be familiar with probability and quantile transformations. As the present work is carried out with the help of packages, a detailed discussion on the hidden procedures is not presented in detail. A brief description is given for the sake of the reader.

Quantile transform: If $U$~$U($0, 1$)$ has a standard uniform distribution, then $P(F^{-1}(U) \leq x) = F(x)$ it denotes the generalized inverse.

Probability transform: If $X$ has distribution function F with continuous univariate distribution function, then $F(x)\sim U(0,1)$.

Sklar's theorem is a valuable theorem in a Copula environment. It claims that if $F$ has a joint distribution with marginals, then there exists a unique Copula $C$ such that for all $x_1, x_2, \dots\dots, x_d \in R\ \ F(x_1, x_2, \dots\dots, x_d) = C\big(F_1(x_1), \dots\dots F_d(x_d)\big)$.

Co-monotonicity and Counter-monotonicity Copulas are two essential Copulas. The terms Co-monotonicity and Counter-monotonicity refer to perfect positive and negative dependence, respectively. Intuitively, if a Copula exists and has neither positive nor negative dependence structures, it must be somewhere in the middle. Therefore, every Copula $C(u_1,\ u_2, \dots\dots\ u_d)$ has bounds:

$$max\left(\textstyle\sum_{i=1}^{d} u_i + 1 - d, 0\right) \leq C(u) \leq min(u_1,\ u_2, \dots\dots\ u_d)$$

and is called Frechet bounds for Copula. As a result, the Frechet upper bound Copula is comonotonicity, while the Frechet Lower bound Copula is counter-monotonicity. Fundamental Copulas identify three sorts of dependent structures; definitions of implicit and explicit Copulas are another method of viewing Copulas. Sklar's theorem is used to extract implicit Copulas from well-known multivariate distributions, but they don't have to result in closed-form expressions. Explicit Copulas, in contrast to implicit Copulas, form closed-form expressions and have Yield Copulas as mathematical structures.

### 2.2. Bivariate Copula

We limit ourselves to the bivariate situation in this study and highlight the significant properties of $d$-dimensional Copulas that are relevant to the current work. We have,

$$C:\ [0,\ 1]^2\ \rightarrow [0,1], (u,v) = C(u,v)$$

with properties

1. For all $u, v\ \in\ [0,1]$it holds:
$$C(u\,,0) = C(0,v) = 0\ and\ C(u,1) =\ u\ and\ C(1,v) =\ v$$

2. For all $u_1, u_2, v_1, v_2\ \in\ [0,1]$ with $u_1 \leq u_2\ and\ v_1 \leq\ v_2$ it holds:
$$C(u_2, v_2) -\ C(u_2, v_1) -\ C(u_1, v_2) + C(u_1, v_1)\ \geq 0$$

is called a bivariate Copula function.

Let $X$ and $Y$ denote temperature and precipitation, which are continuous in nature, with CDF (Cumulative distribution function) $F_X(x) = P(X \leq x)$and $G_Y(y) = P(Y \leq y)$ respectively.

By the definition of Sklar (1973), the joint probability function is given by

$$P(X \leq x, Y \leq y) = C\big(F(x), F(y)\big)$$

where $C$ is an unique function and is known as Copula *i.e.*, $C(u, v) = P(U \leq u, V \leq v)$ is the distribution of $(U,V) = (F(X), G(Y))$ whose marginal distributions are $U[0,1]$. As contended by Joe (1997) and Nelsen (2007), $C$ portrays the dependence between $(X, Y)$. In literature, many Copula families are accessible whose parameters control the intensity of dependence of the variables $(X, Y)$.

Once the parameters of different Copula are estimated, selecting the Copula which can represent the structure of dependency between the interested variables is very important. Few criteria like Aldrian and Black Information Criteria, are available in the literature to identify the best Copula. Information criteria are received here because they can portray the tradeoff between bias (precision) and variance (intricacy) in model development. To measure the relative goodness of fit of a statistical model we use the Akaike information criterion (AIC). It is defined as

$$AIC = 2k - 2\ ln(L)$$

here $k$ is the Copula parameters; $L$ is the optimized value of the likelihood function of the Copula.

The Bayesian information criterion (BIC) was evolved by Schwarz using Bayesian formalism. It is defined as

$$BIC = -2\ ln(L) + k\ ln(N)$$

here $N$ represents the sample size.

## 3. Analysis

### A. Descriptive statistics of temperature and precipitation

The Krishna district's climate is tropical consisting of sweltering summers and mild winters. A clear seasonal cycle has been observed considering the monthly mean Temperature in Krishna from 1901 to 2019, as shown in Figure 1. Similarly, it is observed that there is a seasonal cycle in the monthly Precipitation in Krishna from 1901 to 2019. Figure 2 displays the average monthly precipitation. Table 1 exhibits the Descriptive statistics of Precipitation.

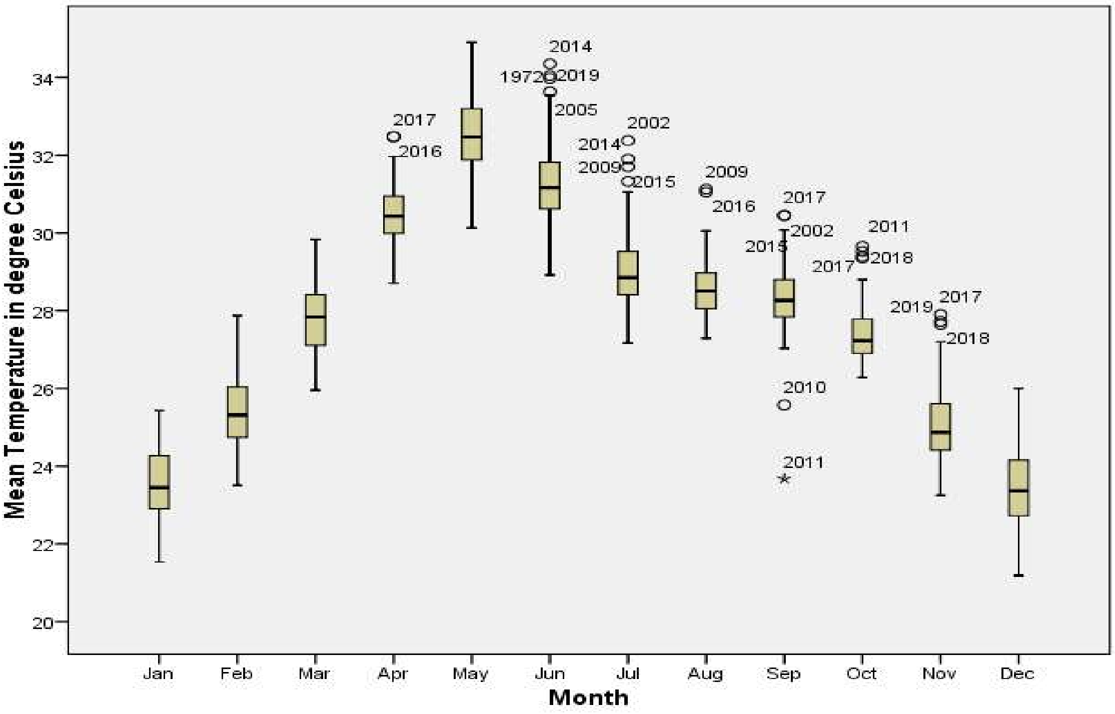


**Figure 1: Mean Temperature in Krishna, Andhra Pradesh from 1901 to 2019**

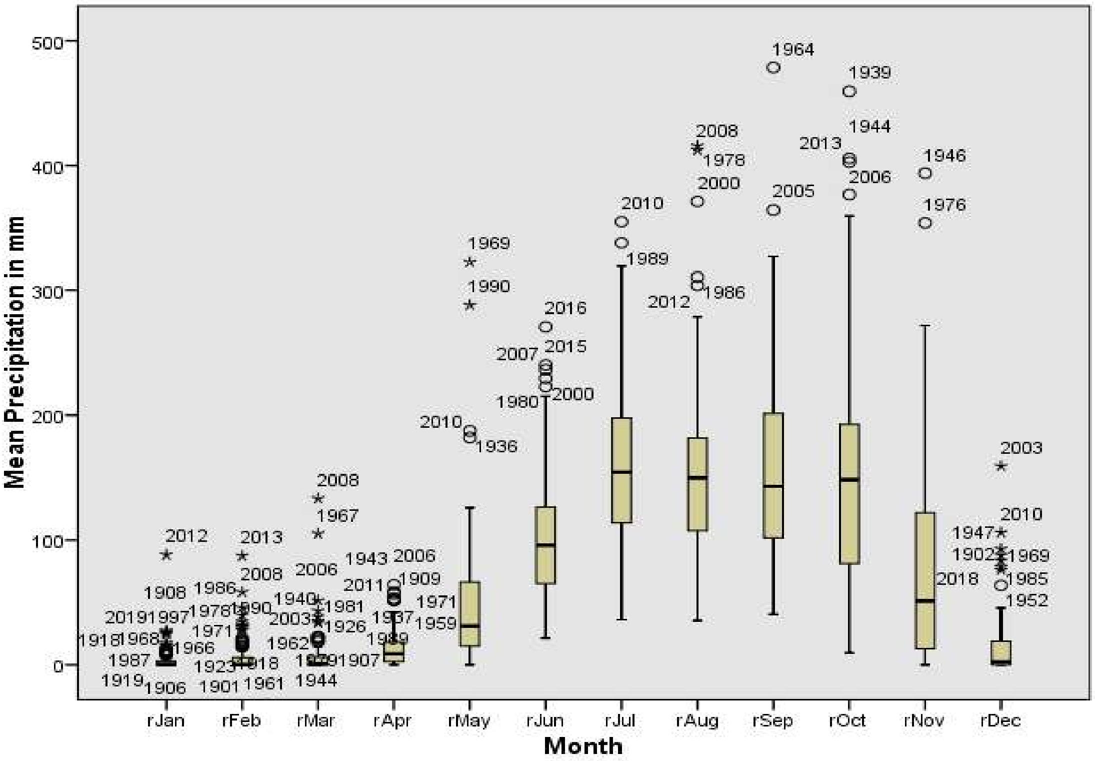


**Figure 2: Mean Precipitation (month wise) in Krishna, Andhra Pradesh from 1901 to 2019**

**Table 1: Descriptive statistics of mean temperature and mean precipitation in Krishna from 1901 to 2019**

| Month | Temperature in Degree Celsius | | | | | Precipitation in mm | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Skewness | Kurtosis | Min | Max | Mean | Skewness | Kurtosis |
| Jan | 21.53 | 25.43 | 23.55 | 0.12 | –0.60 | 0.00 | 88.10 | 3.81 | 6.17 | 49.40 |
| Feb | 23.50 | 27.88 | 25.42 | 0.33 | –0.40 | 0.00 | 87.20 | 6.48 | 3.31 | 14.06 |
| Mar | 25.95 | 29.83 | 27.82 | 0.17 | –0.66 | 0.00 | 133.30 | 6.69 | 5.15 | 31.18 |
| Apr | 28.71 | 32.48 | 30.47 | 0.24 | –0.11 | 0.00 | 64.20 | 13.89 | 1.45 | 1.58 |
| May | 30.12 | 34.90 | 32.56 | 0.12 | –0.02 | 0.00 | 322.67 | 46.59 | 2.84 | 11.57 |
| Jun | 28.91 | 34.35 | 31.28 | 0.47 | 0.54 | 21.26 | 270.67 | 102.31 | 0.91 | 0.69 |
| Jul | 27.17 | 32.38 | 29.04 | 1.15 | 1.83 | 36.25 | 355.00 | 160.77 | 0.60 | 0.25 |
| Aug | 27.29 | 31.13 | 28.58 | 0.88 | 1.16 | 35.52 | 415.80 | 157.66 | 1.16 | 2.06 |
| Sep | 23.68 | 30.45 | 28.34 | –1.09 | 6.76 | 40.30 | 478.71 | 156.25 | 1.18 | 2.52 |
| Oct | 26.28 | 29.65 | 27.39 | 0.99 | 0.97 | 9.39 | 459.49 | 152.17 | 0.87 | 0.94 |
| Nov | 23.25 | 27.90 | 25.09 | 0.61 | 0.01 | 0.00 | 393.89 | 75.31 | 1.46 | 2.20 |
| Dec | 21.18 | 26.00 | 23.48 | 0.52 | –0.22 | 0.00 | 159.10 | 14.14 | 2.96 | 10.74 |

## B. The association between precipitation and temperature in Krishna district

As the sample data shows a non-Gaussian distribution, the Kendall's tau correlation coefficient is utilized to ascertain the relationship between a month to month Temperature and Precipitation. A negative association has been observed between Precipitation and Temperature during April – October (at the 5% significant level), as given in Table 2.

**Table 2: Correlation analysis between temperature and precipitation in Krishna (1901 to 2019)**

| | **January** | **February** | **March** | **April** | **May** | **June** |
|---|---|---|---|---|---|---|
| **Kendal's Correlation Coefficient** | 0.147 | 0.022 | – 0.112 | – **0.305** | – **0.342** | – **0.285** |
| ***p* - Value** | 0.124 | 0.736 | 0.082 | **0.001** | **0.001** | **0.001** |
| | **July** | **August** | **September** | **October** | **November** | **December** |
| **Kendal's Correlation Coefficient** | – **0.232** | – **0.173** | – **0.289** | –**0.178** | –0.035 | 0.07 |
| ***p* - Value** | **0.001** | **0.005** | **0.001** | **0.004** | 0.577 | 0.271 |

## C. Estimation of parameters

Initially, a suitable distribution was fitted to Temperature and Precipitation data using R-Software. The best fitted distributions and their parameter estimates are found using this software. Minimum values of AIC and BIC criteria indicate the best fitted distribution. Chi-

Square test is used to determine a variable is likely to come from specified distribution or not. The Maximum Likelihood technique is utilized to estimate the parameters of the best fitted distribution. Table 3 presents the obtained results of Temperature and Precipitation in Krishna. From Table 3, we can observe that all *p*-values are greater than 0.05 (5% level of significance) that is large *p*-values indicate that we can accept the null hypothesis and conclude that data was drawn from a population with the specified distribution.

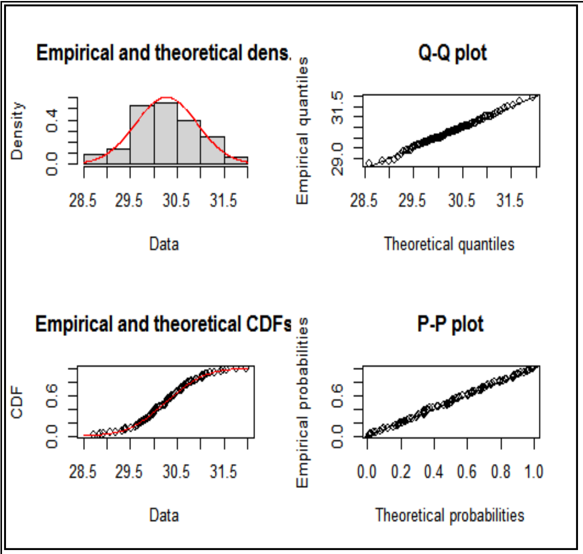
**Table 3: Temperature and precipitation parameters estimates**

| Month | Temperature | | | | | | Precipitation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Distribution** | **Parameter** | **Estimate** | ***p* - Value** | **AIC** | **BIC** | **Distribution** | **Parameter** | **Estimate** | ***p* - Value** | **AIC** | **BIC** |
| **Apr** | Normal | $\mu$ | 30.265 | 0.432 | 186.47 | 191.49 | Exponential | $\mu$ | 14.541 | 0.945 | 671.20 | 673.71 |
| | | $\sigma$ | 0.659 | | | | | | | | | |
| **May** | Normal | $\mu$ | 32.34 | 0.064 | 247.32 | 252.43 | Exponential | $\mu$ | 47.142 | 0.075 | 624.10 | 926.65 |
| | | $\sigma$ | 0.871 | | | | | | | | | |
| **Jun** | Logistic | $\mu$ | 31.079 | 0.166 | 259.82 | 264.95 | Gamma | $\mu$ | 96.733 | 0.062 | 996.89 | 1002.02 |
| | | $\sigma$ | 0.515 | | | | | $\sigma$ | 0.478 | | | |
| **Jul** | Normal | $\mu$ | 28.776 | 0.278 | 188.34 | 193.47 | Gamma | $\mu$ | 154.48 | 0.097 | 1054.65 | 1059.78 |
| | | $\sigma$ | 0.632 | | | | | $\sigma$ | 0.393 | | | |
| **Aug** | Reverse Gumbel | $\mu$ | 28.112 | 0.131 | 161.00 | 166.13 | Gamma | $\mu$ | 147.78 | 0.068 | 1062.25 | 1067.38 |
| | | $\sigma$ | 0.481 | | | | | $\sigma$ | 0.433 | | | |
| **Sep** | Reverse Gumbel | $\mu$ | 27.927 | 0.072 | 156.59 | 161.72 | Inverse Gaussian | $\mu$ | 155.5 | 0.826 | 1070.33 | 1075.45 |
| | | $\sigma$ | 0.469 | | | | | $\sigma$ | 0.037 | | | |
| **Oct** | Normal | $\mu$ | 27.137 | 0.305 | 124.65 | 129.78 | Weibull | $\mu$ | 171.84 | 0.071 | 1125.17 | 1130.3 |
| | | $\sigma$ | 0.453 | | | | | $\sigma$ | 1.754 | | | |

The empirical density function, a simple modification and improvement of the usual histogram, is defined, and its properties are studied. A CDF is the Cumulative Distribution Function. The CDF essentially allows you to plot a feature of the data in order from least to greatest and see the whole feature as if it is distributed across the data set.

Probability plots are the best way to determine whether the data follow a particular distribution. If data follow the straight line on the graph, the distribution fits the data. The Q-Q plot (Quantile – Quantile plot) and P–P plot (Probability–Probability plot) are graphical tools used to determine how well a given data set fits a specific probability distribution that we are testing. Q-Q plot and P–P plot are used to assess how closely two datasets agree, where the two cumulative distribution functions are plotted against each other. If the data points fall along the straight line, we can conclude the data follow that specified distribution.

The graphs from Figure 3 display the frequency curve, cumulative frequency curve, Q-Q and P-P Plots of best fitted distribution of temperature.



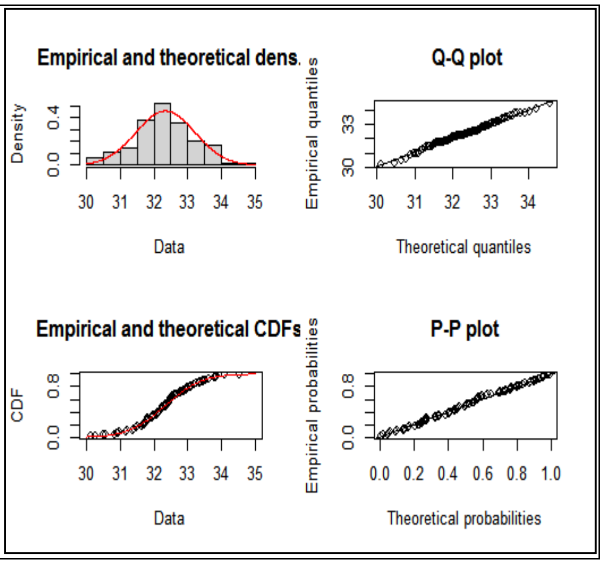
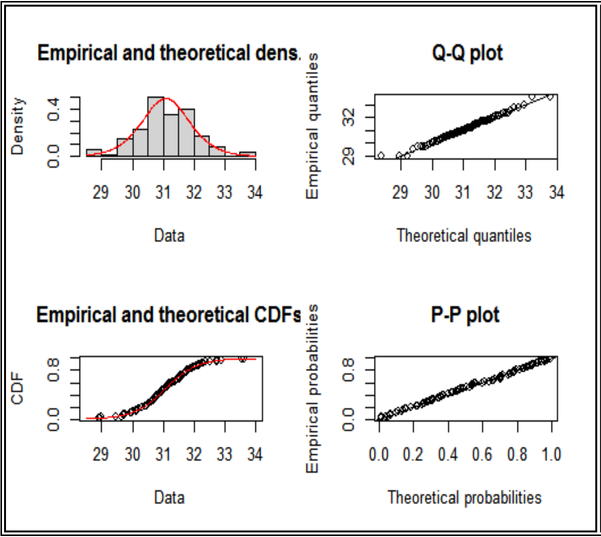


**Figure 3(a): Fitted normal distribution of April**

**Figure 3(b): Fitted normal distribution of May**



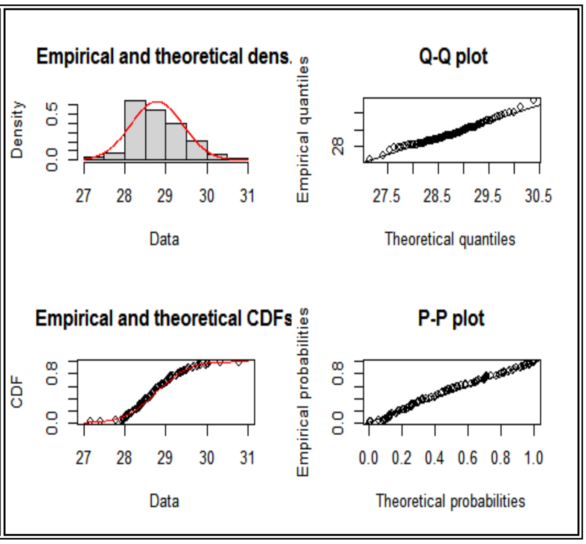


**Figure 3(c): Fitted normal distribution of June**

**Figure 3(d): Fitted normal distribution of July**

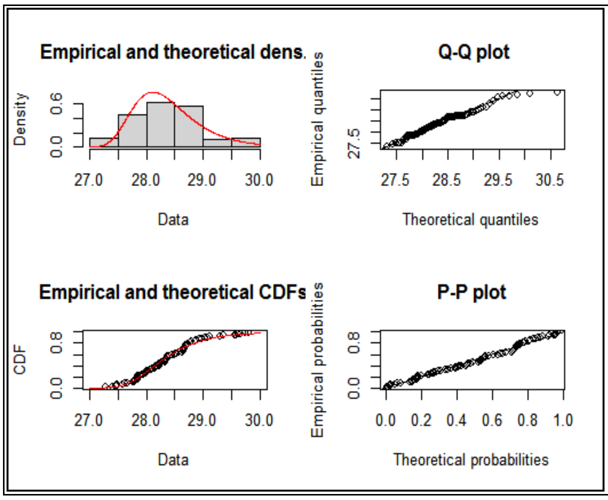


**Figure 3(e): Fitted Reverse Gumbel distribution of August**

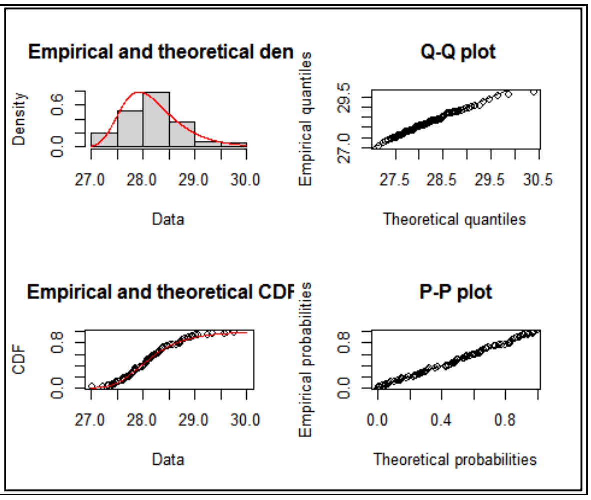


**Figure 3(f): Fitted Reverse Gumbel distribution of September**

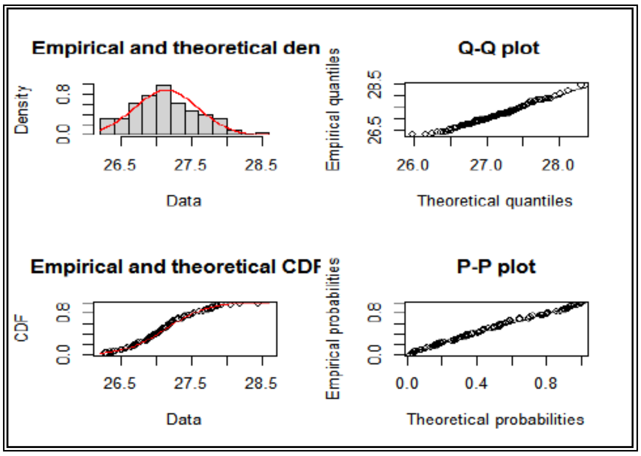


**Figure 3(g): Fitted normal distribution of October**

Q-Q plots and P-P plots pertaining to precipitation have been shown in Figure 4

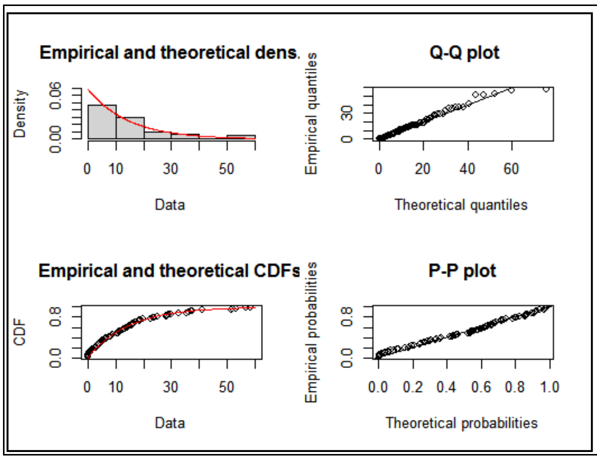


**Figure 4(a): Fitted Exponential distribution of April**

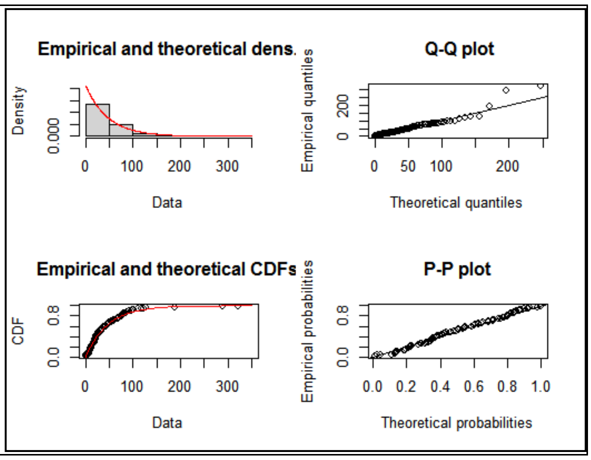
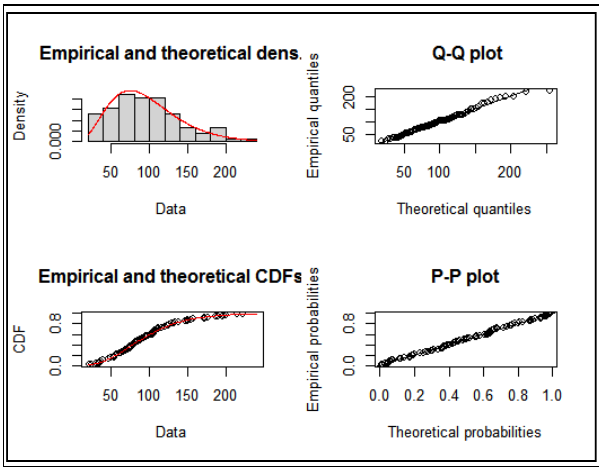


**Figure 4(b): Fitted Exponential distribution of May**



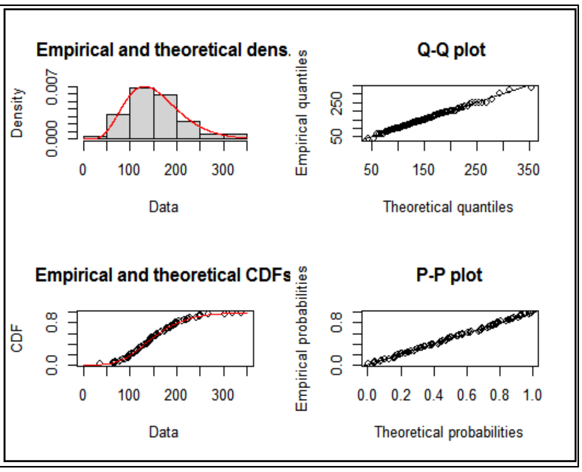
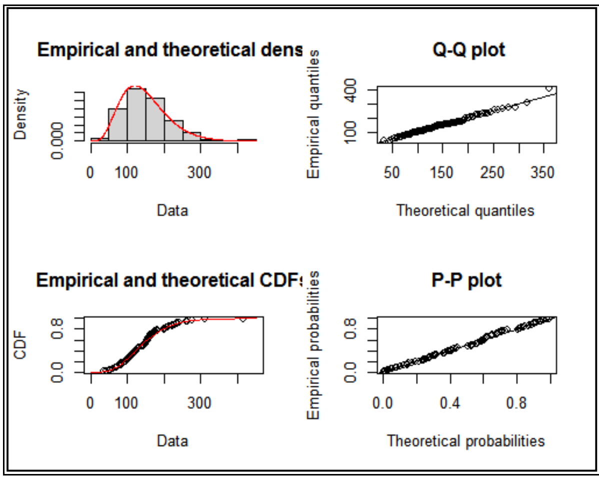


**Figure 4(c): Fitted Gamma distribution of June**

**Figure 4(d): Fitted Gamma distribution of July**



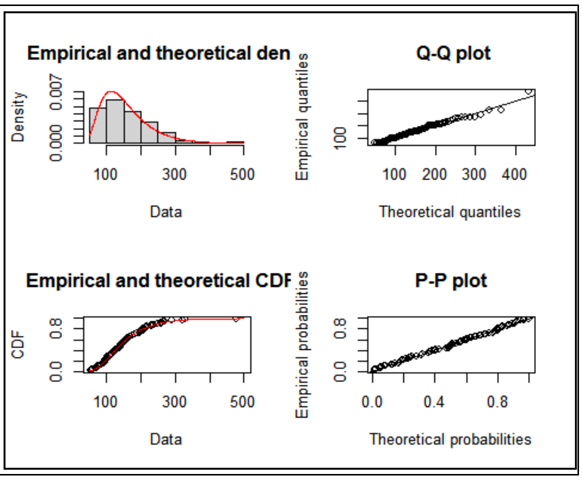


**Figure 4(e): Fitted Gamma distribution of August**

**Figure 4(f): Fitted Inverse Gaussian distribution of September**

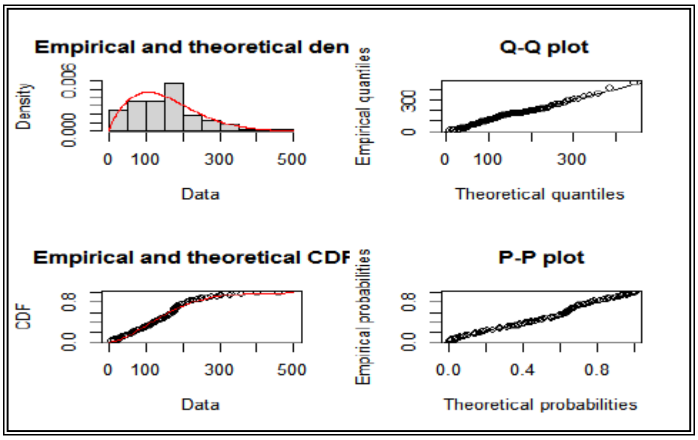


**Figure 4(g): Fitted Weibull distribution of October**

The graphs shown in Figures 3 and 4 indicate that the fitted distributions of Temperature and Precipitation of the Krishna district are very close to the observed data in each month.

## D. Identification of bi-variate Copula

Using these fitted best distributions as Marginal distributions, we find a best fitted joint distribution, for each month, using Copula Analysis Technique, to estimate and forecast the Temperature. Maximum Likelihood Estimation is utilized to find parameter(s) estimates of fitted best Bi-variate Copula distribution. The best Copula distribution is determined using the minimum AIC and BIC criteria.

**Table 4: Copula distribution parameter estimates**

| **Month** | **Bivariate Copula** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Distribution** | **Parameter** | **AIC** | **BIC** | **MAPE** | **RMSE** | **NRMSE** |
| **April** | Gaussians | $\theta$ =–0.472 | –18.019 | –15.508 | 0.026 | 0.962 | 3.178 |
| **May** | Rotated Gumbel 90 degrees | $\theta$ =–1.573 | –30.998 | –28.444 | 0.03 | 1.236 | 3.823 |
| **June** | Rotated Tawn type 2 270 degrees | $\theta$ =–1.925<br>$\omega_1$ = 0.539 | –27.879 | –22.75 | 0.035 | 1.392 | 4.477 |
| **July** | Gaussian | $\theta$ = –0.532 | –26.528 | –23.963 | 0.026 | 0.955 | 3.317 |
| **August** | Rotated Gumbel 270 degrees | $\theta$ =–1.570 | –32.369 | –29.805 | 0.024 | 0.830 | 2.924 |
| **September** | Rotated Gumbel 90 degrees | $\theta$ =–1.626 | –35.227 | –32.663 | 0.024 | 0.873 | 3.097 |
| **October** | Rotated Clayton 90 degrees | $\theta$ =–0.450 | –7.529 | –4.965 | 0.02 | 0.674 | 2.485 |

The error of estimation is calculated by using Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Normalized Root Mean Square Error (NRMSE) has been presented in Table 4. The above table indicates that MAPE, RMSE and NRMSE are less than 5% for all the months. It is observed that in Krishna the average MAPE, RMSE and NRMSE for all months under consideration are 0.026, 0.989 and 3.329 respectively. It is implied that the variation in the observed Temperature data can be explained by these models with approximately 97.4% accuracy. Further, Table 5 shows that there is a similar relation between estimated values of Temperature and Precipitation as exhibited by the observed data.

**Table 5: Relation between temperature and precipitation**

| Month | Apr | May | Jun | July | Aug | Sep | Oct |
|---|---|---|---|---|---|---|---|
| **Observed data relation** | – 0.305 | – 0.342 | – 0.285 | – 0.232 | – 0.173 | – 0.289 | – 0.178 |
| **Estimated data relation** | – 0.318 | – 0.445 | – 0.317 | – 0.261 | – 0.215 | – 0.371 | – 0.213 |

## 4. Prediction

The MAPE, RMSE and NRMSE values being less than 5% indicates that the fitted best Bi-Variate Copula can be used for estimation and/or prediction of Temperature. Therefore, Temperature values for the Krishna district are estimated using Gaussian for the months of April and July, Rotated Gumbel 90 degrees for the months of May and September, Rotated Tawn type 2 270 degrees for the month of June, Rotated Gumbel 270 degrees for the month of August and Rotated Clayton 90 degree for the month of October Copula distributions for training as well as testing data sets.

Figure 5 presents the observed and predicted values for the testing period.

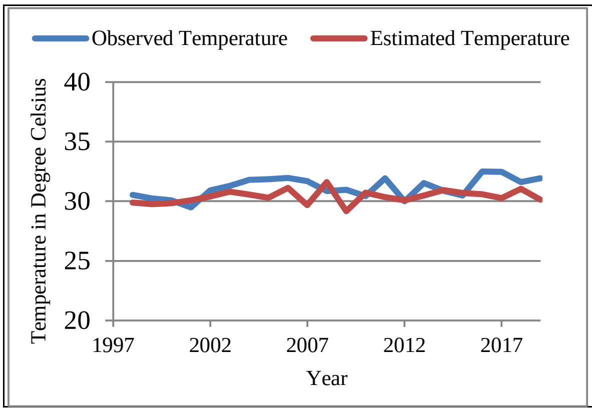


**Figure 5(a): Krishna test data of April**

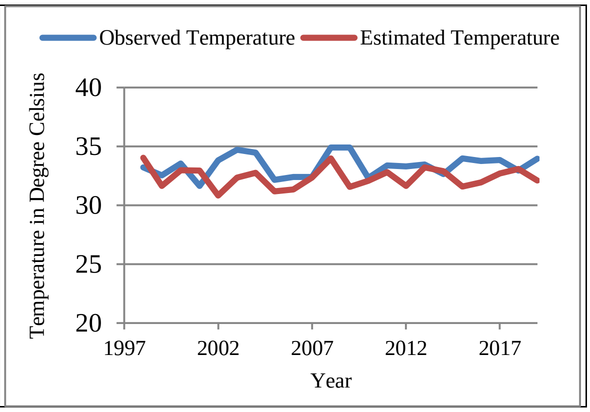


**Figure 5(b): Krishna test data of May**

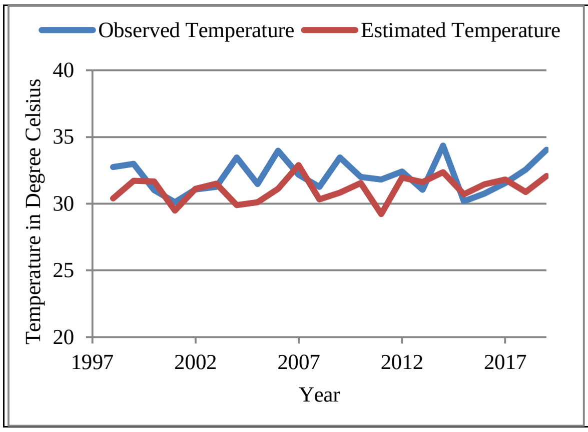


**Figure 5(c): Krishna test data of June**

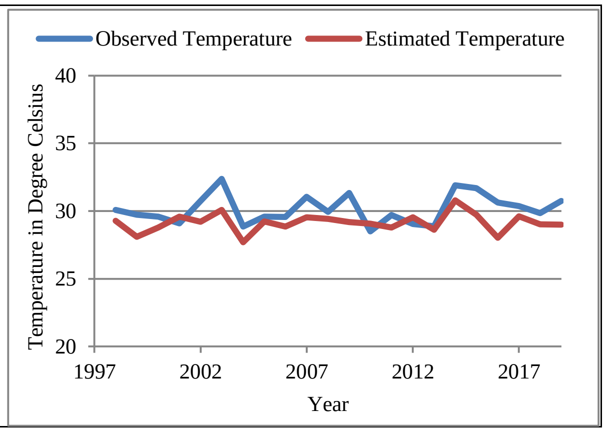
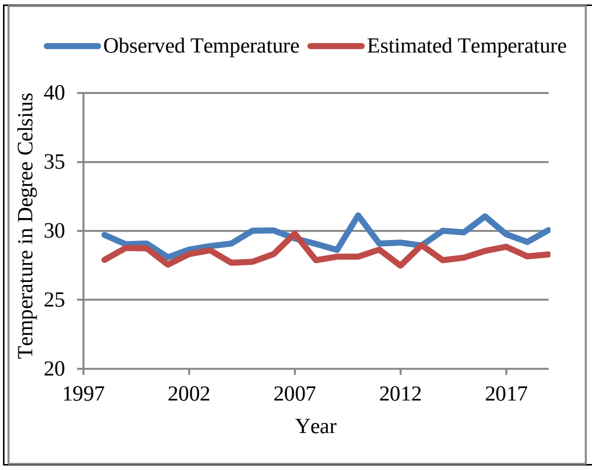


**Figure 5(d): Krishna test data of July**



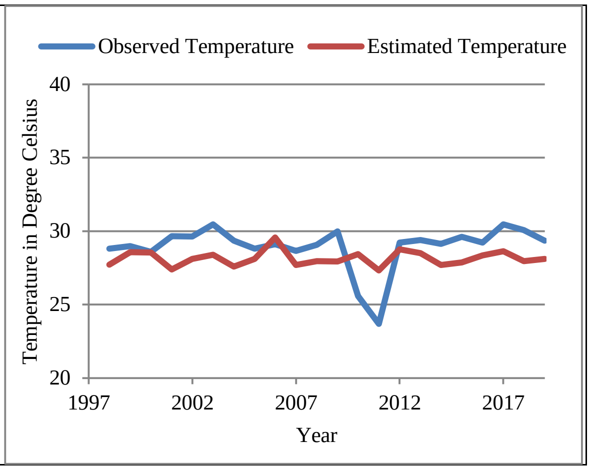


**Figure 5(e): Krishna test data of August**

**Figure 5(f): Krishna test data of September**

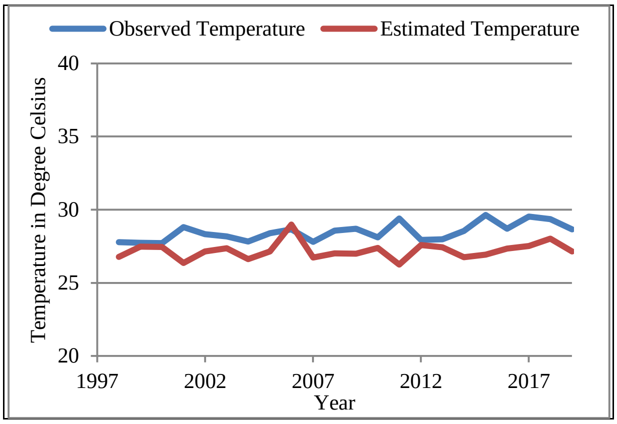


**Figure 5(g): Krishna test data of October**

The above graphs show that using the best fitted Copula distribution of April to October there is a reasonably good agreement in the patterns between observed and predicted Temperature values.

## 5. Conclusion

In this study, we have followed a novel approach to predict Temperature for Krishna district in Andhra Pradesh from April to October, by using Copula analysis. After the complete analysis, we could draw the following conclusions.

Using 80% of the collected data, initially, we identified the best fitted Probability distributions to the variables Temperature and Precipitation, separately. These distributions, in general, are different for different districts and months. The Probability density functions of these distributions are listed in Annexure 1.

Using the above identified best component distributions, we could identify the best fitted Joint Copula model that could predict month-wise Temperature for Krishna district in

Andhra Pradesh from April to October. As the climatic conditions change from region to region, the identified best Copula models are not the same for all the months in this district. The Probability density functions of best fitted Joint Copula distribution are listed in Annexure 2.

In order to establish the model adequacy of the best fitted Copulas, we computed AIC, BIC, MAPE, RMSE and NRMSE for the Krishna district from April to October, where ever the dependency is significant. It is observed that all the values of AIC and BIC are the least, all the values of MAPE are less than 5% and all the values of NRMSE are less than or around 5%. This establishes that the fitted models are adequate.

As the fitted models are adequate, using them we forecasted the values of Temperature for the time points in the testing data period. In the Krishna district, for all the months (where ever models are fitted), we could find a close agreement between observed and forecasted testing data.

Hence, it can be concluded that the identified best Copula models can be used for the prediction of future data points. Forecasted Temperature helps in planning agriculture and trading of commodities. This forecasting helps in deciding whether a crop has to be irrigated or not, should use fertilizers and whether it is a right time to complete harvesting *etc*.

## References

Bezak, N., Zabret, K. and Sraj, M. (2018). Application of Copula functions for rainfall interception modelling. *Water*, **10**, 995.

Dzupire, N. C., Ngare, P. and Odongo, L. (2020). A Copula based bi-variate model for temperature and rainfall processes. *Scientific African*, **8**, 1-10.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC press.

Lazoglou, G.and Anagnostopoulou, C. (2019). Joint distribution of temperature and precipitation in the Mediterranean, using the Copula method. *Theoretical and Applied Climatology*, **135**, 1399-1411.

Mesbahzadeh, T., Miglietta, M. M., Mirakbari, M., Soleimani Sardoo, F.and Abdolhoseini, M. (2019). Joint modeling of precipitation and temperature using Copula theory for current and future prediction under climate change scenarios in arid lands (Case Study, Kerman Province, Iran). *Advances in Meteorology*, 1-15.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science and Business Media.

Olesen, J. E. and Bindi, M. (2002). Consequences of climate change for European agricultural productivity, land use and policy. *European Journal of Agronomy*, **16**, 239-262.

Pandey, P. K., Das, L., Jhajharia, D. and Pandey, V. (2018). Modelling of interdependence between rainfall and temperature using Copula. *Modeling Earth Systems and Environment*, **4**, 867-879.

Shaukat, M. H., Hussain, I., Faisal, M., Al-Dousari, A., Ismail, M., Shoukry, A. M., Elashkar, E.E. and Gani, S. (2020). Monthly drought prediction based on ensemble models. *PeerJ*, **8**, e9853.

Sklar, A. (1973). Random variables, joint distribution functions, and Copulas. *Kybernetika*, **9**, 449-460.

Vergara, O., Zuba, G., Doggett, T. and Seaquist, J. (2008). Modeling the potential impact of catastrophic weather on crop insurance industry portfolio losses. *American Journal of Agricultural Economics*, **90**, 1256-1262.

Yu, R., Yang, R., Zhang, C., Špoljar, M., Kuczyńska-Kippen, N. and Sang, G. (2020). A vine Copula-based modeling for identification of multivariate water pollution risk in an interconnected river system network. *Water*, **12**, 2741.

Zhang, L. and Singh, V. P. (2007). Bivariate rainfall frequency distributions using Archimedean Copulas. *Journal of Hydrology*, **332**, 93-109.

Zscheischler, J., Orth, R. and Seneviratne, S. I. (2017). Bivariate return periods of temperature and precipitation explain a large fraction of European crop yields. *Biogeosciences*, **14**, 3309-3320.

## ANNEXURE 1

### Density functions of the identified best fitted component (Temperature and Precipitation) distribution

### Normal distribution

Normal distribution is a two-parameter distribution function and the parameterization of the normal distribution given in the function is

$$f(x/\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\text{; where } -\infty < x < \infty,\ -\infty < \mu < \infty \text{ and } \sigma > 0$$

Here μ and σ are mean and standard deviation of the distribution respectively.

### Reverse Gumbel distribution

For positive skewed data the suitable distribution is Reverse Gumbel distribution. The probability density function of Reverse Gumbel distribution is

$$f(x/\mu,\sigma)\ =\ \frac{1}{\sigma}exp\left[-\left(\frac{x-\mu}{\sigma}\right)-exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right];$$

where $-\infty < x < \infty,\ -\infty < \mu < \infty$ and $\sigma > 0$

Here μ and σ are mean and standard deviation of the distribution respectively.

### Logistic distribution

The Logistic distribution is suitable for moderate kurtosis data. The probability density function is given by

$$f(x/\ \mu,\sigma)\ =\ \frac{1}{\sigma}\left[exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right]\left[1+exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right]^{-2}$$

where $-\infty < x < \infty,\ -\infty < \mu < \infty$ and $\sigma > 0.$

### Weibull distribution

Weibull distribution is a two-parameter distribution function and the parameterization of the Weibull distribution given in the function is

$$f(x\ /\ \mu,\sigma)\ =\ \frac{\sigma x^{\sigma-1}}{\mu^{\sigma}} exp\left(-\frac{x}{\mu}\right)^{\sigma}; \text{ where } x > 0,\ \mu > 0 \text{ and } \sigma > 0$$

Here $\mu$ and $\sigma$ are mean and standard deviation of the distribution respectively.

### Exponential distribution

Exponential distribution is a one parameter distribution function and the parameterization of the Exponential distribution given in the function is

$$f(x/\mu)\ =\ \frac{1}{\mu} exp\left(-\frac{x}{\mu}\right); \text{ Where } x > 0,\ \mu > 0$$

Here $\mu$ is mean of the distribution respectively.

### Gamma distribution

Gamma distribution is a one parameter distribution function and the parameterization of the Gamma distribution given in the function is

$$f(x/\mu,\sigma)\ =\ \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{x^{\left({}^{1}/_{\sigma^2}\right)-1}\, exp(\sigma^2\mu)}{\gamma\left({}^{1}/_{\sigma^2}\right)};$$

where $x > 0$, $\mu > 0$ and $\sigma > 0$, $\mu$ and $\sigma$ are mean and standard deviation of the distribution respectively.

### Inverse Gaussian distribution

Inverse Gaussian distribution pdf is,

$$f(x/\ \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} exp\left[-\frac{1}{2\mu^2\sigma^2 y}(x-\mu)^2\right]$$

for $x > 0, \mu > 0$ and $\sigma > 0.$

## ANNEXURE 2

## Density functions of identified best joint Copula distribution

### Gaussian Copula

Gaussian Copula is defined as,

$$C(u,v) = \frac{1}{\sqrt{(1-\theta^2)}}\ e^{\left(\frac{\theta^2(u^2+v^2)-2\theta uv}{2\,(1-\,\theta^2)}\right)}$$

where, $-1 < \theta < 1$

Here, the Gaussian Copula parameter $\theta$ is given by $\theta = Sin\left[\frac{\pi}{2}\ \tau\right]$

**Tawn Copula**

The Tawn Copula is defined as

$$C(u, v) = \ (u, v)^{A(\alpha)} \text{ ; with } \alpha = \frac{ln(u)}{ln(uv)}$$

The Pickand function of Tawn Copula is given by

$$A(t) = \ (1 - \omega_2)(1 - t) + \ (1 - \omega_1)t + \ \left[\left(\theta_1(1 - t)\right)^{\alpha} + \ (\theta_2\ t)^{\alpha}\right]^{1/\theta}$$

where, *t ∈ [0,1], 0 ≤ω₁, ω₂ ≤ 1* and *θ∈ [0,∞),* the Tawn type 1 and type 2 refers to $\omega_1 = 1$ or $\omega_2 = 1$

Here, $\theta$ and $\omega_1$ are the two parameters of Tawn Type 2 Copula, $\omega_2$ = 1, and the Parameter $\theta$ is given by $\tau = 1 - \theta^{-1}$ .

$$\lambda_u = \omega_1 + \ 1 - \ \left(\omega_1^{\theta} + \ 1\right)^{\frac{1}{\theta}} \ \text{ or } \ \lambda_u = \omega_2 + \ 1 - \ \left(\omega_2^{\theta} + \ 1\right)^{\frac{1}{\theta}}$$

where Rotated 270 Copula means,

$$C^{270}(u, v) = C(v, 1 - u)$$

**Rotated Gumble Copula**

Rotated Gumble Copula is defined as,

$$C(u, v) = e^{\left(-[(-\,ln\,u)^{\theta} + (-\,ln\,v)^{\theta}]^{\frac{1}{\theta}}\right)}$$

Here, the Gumbel parameter $\theta$ *(*≥1*)* is given by $\hat{\theta} = \frac{1}{1 - \tau}$ and $\tau$ is the correlation between the variables. Here Rotated 270 Copula means,

$$C^{90}(u, v) = C(1 - u, v)$$

**Clayton Copula**

Clayton Copula is defined as,

$$C(u, v) = max\left(\left(u^{-\theta} + \ v^{-\theta} - \ 1\right)^{-\frac{1}{\theta}}, 0\right)$$

Here, the Clayton Copula parameter θ is given by $\tau = \frac{\theta}{\theta + 2}$ and *θ∈ [*-1, ∞*)*

where Rotated 90 Copula means,

$$C^{90}(u, v) = C(1 - u, v)$$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 161–177

# On the Bivariate Generalized Chen Distribution

**R. M. Mandouh**
*Department of Mathematical Statistics*
*Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt*

Received: 13 September 2021; Revised: 21 June 2022; Accepted: 07 July 2022

## Abstract

A new model of bivariate distributions is presented in this paper. The model introduced here is of the Marshall–Olkin type. The joint survival function, the joint probability density function and the joint hazard function of the bivariate generalized Chen (BGCh) distribution are obtained. The maximum likelihood and Bayesian methods are used to estimate the unknown parameters of the BGCh distribution. Numerical methods are required to calculate the desired estimates.

*Key words:* Marshall–Olkin type; Generalized Chen (BGCh) distribution; The maximum likelihood and Bayesian methods; MCMC.

**AMS Subject Classifications:** 62K05, 05B05

## 1. Introduction

A suitable parametric model is often of interest in the analysis of survival data, as it provides insight into the characteristics of the failure times and hazard functions that may not be available with non-parametric methods. The Weibull distribution is one of the most commonly used families for modeling such data. However, only monotonically increasing and decreasing hazard functions can be generated from the classic two-parameter Weibull distribution. As such this two-parameter model is inadequate when the true hazard shape is of bathtub nature. Models with bathtub-shaped hazard rate are needed in reliability analysis and decision making when the complete life cycle of the system is to be modeled. Many authors have proposed models with bathtub-shaped failure rates. For example, Smith and Bain (1975) proposed the exponential power distribution. Mudholkar and Srivastava (1993) suggested the exponentiated Weibull distribution. Chen (2000) provided a two-parameter lifetime distribution with bathtub shape or increasing failure function, now known as Chen distribution. Xie *et al.* (2002) modified the Chen distribution to include a scale parameter named modified Weibull extension and also referred to as the generalized Chen distribution. They discussed the parameters' estimation using maximum likelihood method. For more generalizations and modifications of Weibull distribution, see Murthy *et al.* (2004) and Pham and Lai (2007).

Bivariate lifetime data arise frequently in many practical problems and in these situations it is important to consider different bivariate models that could be used to model

Corresponding Author: R. M. Mandouh
Email: rshmndoh@cu.edu.eg

such bivariate lifetime data. There are a number of papers dealing with bivariate models of type of Marshal-Olkin. For example, Sarhan and Balakrishnan (2007) introduced a bivariate distribution using exponential and generalized exponential distributions, now known as Sarhan-Balakrishnan bivariate (SBBV) distribution. Although, they derived several interesting properties of this distribution, the marginal distributions of SBBV distribution are not in known forms. Kundu *et al.* (2012) modified the SBBV distribution to include a scale parameter and discussed the estimation of parameters using maximum likelihood method. Kundu and Gupta followed the idea using the generalized exponential to introduce the bivariate generalized exponential (BVGE) distribution so that the marginal distributions are generalized exponential distributions. They derived several interesting properties of this distribution and discussed the maximum likelihood estimation of the unknown parameters. Also, they re-analyzed a real data set that was analyzed by Meintanis (2007) and concluded that the BVGE distribution provides a better fit than the bivariate Marshall-Olkin distribution. Sarhan (2019) noted that none of the marginal distributions of the SBBV and the BVGE provide a bathtub shape of the hazard function and this lack of the bathtub property limits the application of these distributions. Thus he introduced a new bivariate distribution named the bivariate generalized Rayleigh (BVGR) distribution. The BVGR distribution has generalized Rayleigh marginal distributions. The hazard rate functions of the marginals of the BVGR can be either increasing or decreasing or bathtub shaped, and with this property the BVGR distribution has wider applicability than other distributions. Sarhan (2019) investigated several interesting properties of this distribution and estimated the unknown parameters by using the maximum likelihood and Bayes methods. Many authors discussed the Marshal-Olkin idea for different distributions; see for example; El-Gohary *et al.* (2015), Kundu and Gupta (2017), Azizi *et al.* (2019), Muhammed (2019) and others.

Using the idea of Marshal-Olkin, we propose a new bivariate generalized Chen (BGCh) distribution. The BGCh distribution has generalized Chen marginal distributions. The joint survival function, the joint probability density function and the joint hazard function of the BGCh distribution are obtained. The maximum likelihood and Bayesian methods are used to estimate the unknown parameters of the BGCh distribution. Numerical methods are required to calculate these estimates.

## 2. The bivariate generalized Chen distribution

In this section, we define a new bivariate distribution, shortly denoted by BGCh. We start with the joint survival function of the distribution and then we derive the corresponding joint probability density function.

### 2.1. The joint survival function

Chen (2000) introduced a two-parameter lifetime distribution with either bathtub-shaped or increasing failure rate with the survival function

$$S_{Ch}(t) = exp(\lambda(1 - e^{(t)^{\beta}})), \quad t \geq 0, \lambda \text{ and } \beta > 0.$$

and the corresponding probability density function

$$f_{Ch}(t) = \lambda\beta(t)^{\beta-1}exp((t)^{\beta} + \lambda(1 - e^{(t)^{\beta}})), \quad t \geq 0, \lambda \text{ and } \beta > 0.$$

Xie *et al.* (2002) modified the Chen distribution to include a scale parameter named the generalized Chen distribution. The survival function of the univariate generalized Chen (GCh) distribution is

$$S_{GCh}(t) = exp(\lambda\alpha(1 - e^{(t/\alpha)^{\beta}})), \quad t \geq 0, \lambda, \alpha \text{ and } \beta > 0. \tag{1}$$

with probability density function (pdf)

$$f_{GCh}(t) = \lambda\beta(t/\alpha)^{\beta-1} exp((t/\alpha)^{\beta} + \lambda\alpha(1 - e^{(t/\alpha)^{\beta}})), \quad t \geq 0, \lambda, \alpha \text{ and } \beta > 0. \tag{2}$$

Now, suppose that $T_j, j = 1, 2, 3$ are independent random variables with $T_i$ having GCh distributions with scale parameters $\alpha$, and $\lambda_j, j = 1, 2, 3$ and shape parameter $\beta$; i.e. $T_i \sim$ GCh$(\alpha, \beta, \lambda_j), j = 1, 2, 3$. Define $X_i = min(T_i, T_3), i = 1, 2$. Then one can say that the vector $(X_1, X_2)$ follows the bivariate generalized Chen distribution with scale parameters $\alpha$,and $\lambda_j, j = 1, 2, 3$ and shape parameter $\beta$. We will denote it by BGCh$(\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$ and to simplify we write $\lambda_{123} = \lambda_1 + \lambda_2 + \lambda_3$ and $\lambda_{i3} = \lambda_i + \lambda_3, i = 1, 2$.

**Theorem 1:** Let $(X_1, X_2)$ follows BGCh$(\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$, then the joint survival function of $(X_1, X_2)$ for $x_1 > 0, x_2 > 0$, is

$$\begin{aligned} S_{X_1,X_2}(x_1, x_2) &= P(X_1 > x_1, X_2 > X_2) \\ &= P(T_1 > x_1, T_2 > x_2, T_3 > x_3) \\ &= \prod_{i=1}^{3} exp(\lambda_i\alpha(1 - e^{(x_i/\alpha)^{\beta}})), \end{aligned} \tag{3}$$

where $x_3 = \max\{x_1, x_2\}$.
Also, the joint survival function of $(X_1, X_2)$ can be written as

$$\begin{aligned} S_{X_1,X_2}(x_1, x_2) &= \prod_{i=1}^{3} S_{GCh}(x_i; \alpha, \beta, \lambda_i) \\ &= \begin{cases} S_{GCh}(x_1; \alpha, \beta, \lambda_1) S_{GCh}(x_2; \alpha, \beta, \lambda_{23}) & \text{if } x_1 < x_2 \\ S_{GCh}(x_2; \alpha, \beta, \lambda_2) S_{GCh}(x_1; \alpha, \beta, \lambda_{13}) & \text{if } x_2 < x_1 \\ S_{GCh}(x; \alpha, \beta, \lambda_{123}) & \text{if } x_1 = x_2 = x. \end{cases} \end{aligned} \tag{4}$$

## 2.2. The joint probability density function

The following theorem gives the joint probability density function of the BGCh distribution.

**Theorem 2:** Let $(X_1, X_2)$ follows BGCh$(\alpha, \beta, \lambda_1, \lambda_2, \lambda_3\ )$, then the joint pdf of $(X_1, X_2)$ takes the form

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} f_1(x_1, x_2) & \text{if } 0 < x_1 < x_2 < \infty \\ f_2(x_1, x_2) & \text{if } 0 < x_2 < x_1 < \infty \\ f_3(x) & \text{if } 0 < x_1 = x_2 = x < \infty. \end{cases} \tag{5}$$

where

$$\begin{aligned} f_1(x_1, x_2) &= \lambda_1\lambda_{23}\beta^2(x_1/\alpha)^{(\beta-1)}(x_2/\alpha)^{(\beta-1)} e^{(x_1/\alpha)^{\beta} + (x_2/\alpha)^{\beta}} e^{\lambda_1\alpha(1-e^{(x_1/\alpha)^{\beta}}) + \lambda_{23}\alpha(1-e^{(x_2/\alpha)^{\beta}}))} \\ &= f_{GCh}(x_1; \alpha, \beta, \lambda_1) f_{GCh}(x_2; \alpha, \beta, \lambda_{23}), \end{aligned}$$

$$\begin{aligned} f_2(x_1,x_2) &= \lambda_{13}\lambda_2\beta^2(x_1/\alpha)^{(\beta-1)}(x_2/\alpha)^{(\beta-1)}e^{(x_1/\alpha)^\beta+(x_2/\alpha)^\beta}e^{\lambda_{13}\alpha(1-e^{(x_1/\alpha)^\beta})+\lambda_2\alpha(1-e^{(x_2/\alpha)^\beta})} \\ &= f_{GCh}(x_1;\alpha,\beta,\lambda_{13})f_{GCh}(x_2;\alpha,\beta,\lambda_2), \end{aligned}$$

and

$$\begin{aligned} f_3(x) &= \lambda_3\beta(x/\alpha)^{(\beta-1)}e^{(x/\alpha)^\beta}e^{\lambda_{123}\alpha(1-e^{(x/\alpha)^\beta})} \\ &= \frac{\lambda_3}{\lambda_{123}}f_{GCh}(x;\alpha,\beta,\lambda_{123}). \end{aligned}$$

**Proof**: The forms of $f_1(.,.)$ and $f_2(.,.)$ can be obtained simply by differentiating $S_{X_1,X_2}(x_1,x_2)$ in (4) with respect to $x_1$ and $x_2$ for $x_1 < x_2$ and $x_2 < x_1$, respectively. The form of $f_3(x)$ can not obtained in the same way but it can be derived by using the following identity:

$$\int_0^\infty\int_0^{x_2} f_1(x_1,x_2)dx_1dx_2 + \int_0^\infty\int_0^{x_1} f_2(x_1,x_2)dx_2dx_1 + \int_0^\infty f_3(x)dx = 1$$

which completes the proof of the theorem.

**Proposition 1:** Let $(X_1,X_2)$ follows BCh$(\beta,\lambda_1,\lambda_2,\lambda_3)$, then the joint pdf of $(X_1,X_2)$ takes the form

$$g_{X_1,X_2}(x_1,x_2) = \begin{cases} g_1(x_1,x_2) & \text{if } 0 < x_1 < x_2 < \infty \\ g_2(x_1,x_2) & \text{if } 0 < x_2 < x_1 < \infty \\ g_3(x) & \text{if } 0 < x_1 = x_2 = x < \infty. \end{cases} \tag{6}$$

where

$$\begin{aligned} g_1(x_1,x_2) &= \lambda_1\lambda_{23}\beta^2(x_1)^{(\beta-1)}(x_2)^{(\beta-1)}e^{(x_1)^\beta+(x_2)^\beta}e^{\lambda_1(1-e^{(x_1)^\beta})+\lambda_{23}(1-e^{(x_2)^\beta})} \\ &= g_{Ch}(x_1;\beta,\lambda_1)g_{Ch}(x_2;\beta,\lambda_{23}), \end{aligned}$$

$$\begin{aligned} g_2(x_1,x_2) &= \lambda_{13}\lambda_2\beta^2(x_1)^{(\beta-1)}(x_2)^{(\beta-1)}e^{(x_1)^\beta+(x_2)^\beta}e^{\lambda_{13}(1-e^{(x_1)^\beta})+\lambda_2(1-e^{(x_2)^\beta})} \\ &= g_{Ch}(x_1;\beta,\lambda_{13})g_{Ch}(x_2;\beta,\lambda_2), \end{aligned}$$

and

$$\begin{aligned} g_3(x) &= \lambda_3\beta(x)^{(\beta-1)}e^{(x)^\beta}e^{\lambda_{123}(1-e^{(x)^\beta})} \\ &= \frac{\lambda_3}{\lambda_{123}}g_{Ch}(x;\beta,\lambda_{123}). \end{aligned}$$

**Proof**: The result is obtained immediately from Theorem 2 upon setting $\alpha = 1$.

The BGCh distribution has both a singular part and an absolutely continuous part similar to Marshal-Olkin's bivariate exponential distribution, Sarhan and Balakrishnan bivariate distribution, the bivariate generalized exponential introduced by Kundu and Gupta (2009) and the bivariate generalized Rayleigh distribution provided by Sarhan (2019). The function $f_{X_1,X_2}$ $(.,.)$ may be considered to be a density function for the BGCh distribution if it is understood that the first two terms are densities with respect to two-dimensional Lebesgue measure and the third term is a density function with respect to one dimensional Lebesgue measure, see Bemis *et al.* (1972). It is well known that although in one dimension the practical use of a distribution with this property is unusual, but they do arise quite naturally in higher dimensions, see Marshal and Olkin (1967).

In many practical situations it may happen that $X_1$ and $X_2$ both are continuous random variables, but $X_1 = X_2$ has a positive probability. The BGCh distribution may be used as a competing risk model or a shock model similar to the bivariate Marshall-Olkin model. Marshal and Olkin (1967) has examples in this connection.. The following theorem provides the explicit forms of the absolute continuous and the singular parts of the BGCh distribution.

**Theorem 3:** If $(X_1, X_2)$ follows BGCh$(\alpha, \beta, \lambda_1, \lambda_2, \lambda_3)$, then

$$S_{X_1,X_2}(x_1, x_2) = \frac{\lambda_3}{\lambda_{123}} S_s(x_1, x_2) + \frac{\lambda_{12}}{\lambda_{123}} S_a(x_1, x_2).$$

For $x = max(x_1, x_2)$ we get,

$$S_s(x_1, x_2) = e^{\lambda_{123}\alpha(1-e^{(x)^\beta})},$$

and

$$S_a(x_1, x_2) = \frac{\lambda_{123}}{\lambda_{12}} \prod_{i=1}^{3} e^{\lambda_i \alpha(1-e^{(x_i/\alpha)^\beta})} - \frac{\lambda_3}{\lambda_1 2} e^{\lambda_{123}\alpha(1-e^{(x)^\beta})},$$

here $S_s(.,.)$ and $S_a(.,.)$ are the singular and the absolutely continuous parts, respectively.

**Proof**: The joint survival function $S_{X_1,X_2}(x_1, x_2)$ can be written as

$$S_{X_1,X_2}(x_1, x_2) = P(X_1 > x_1, X_2 > x_2|A)P(A) + P(X_1 > x_1, X_2 > x_2|\acute{A})P(\acute{A})$$

Let $A = \{T_3 < T_1\} \cap \{T_3 < T_2\} \equiv \{X_1 = X_2\}$, therefore

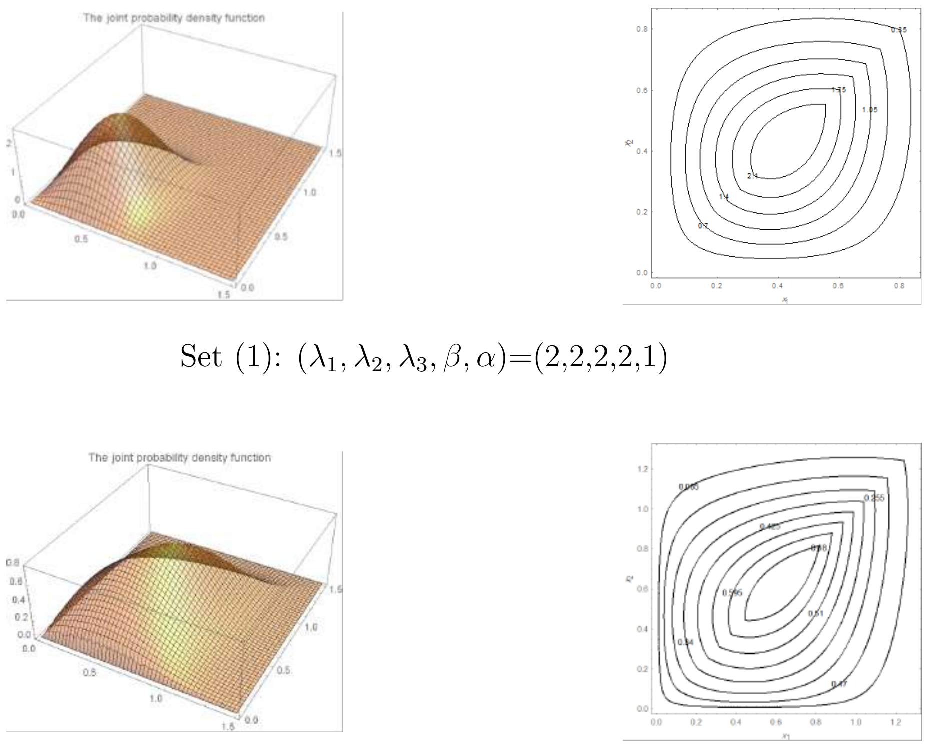
$$P(A) = \int_0^\infty \lambda_3 \beta (x/\alpha)^{(\beta-1)} e^{(x/\alpha)^\beta} e^{\lambda_{123}\alpha(1-e^{(x/\alpha)^\beta})} dx = \frac{\lambda_3}{\lambda_{123}}$$

and

$$\begin{aligned} S_s(x_1, x_2) &= P(X_1 > x_1, X_2 > x_2|A) \\ &= \frac{\lambda_{123}}{\lambda_3} \int_0^\infty \lambda_3 \beta (x/\alpha)^{(\beta-1)} e^{(x/\alpha)^\beta} e^{\lambda_{123}\alpha(1-e^{(x/\alpha)^\beta})} dx \\ &= e^{\lambda_{123}\alpha(1-e^{(x/\alpha)^\beta})}. \end{aligned}$$

Once $P(A)$ and $S_s(x_1, x_2)$ are obtained, the function $S_a(x_1, x_2)$ can be obtained by subtraction.

Different shapes of the joint pdf and corresponding contours for different sets of parameters values are provided in Figure 1.



Set (1): $(\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$=(2,2,2,2,1)

Set (2): $(\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$=(0.5,0.5,1,1.5,1)

**Figure 1: The joint probability density function of the BGCh distribution and corresponding contour**

### 2.3. The joint hazard rate function

Using the relation between the joint pdf of $(X_1, X_2)$ and the joint survival function of $(X_1, X_2)$, one can obtain the joint hazard rate function of $(X_1, X_2)$ according to the relation

$$h_{X_1,X_2}(x_1, x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{S_{X_1,X_2}(x_1, x_2)}.$$

Here we use the forms (4) and (6) to obtain the joint hazard rate function. In Figure 2 we provide the surface plots of the joint hazard rate function and corresponding contours for different values of the parameters.

## 3. Statistical properties

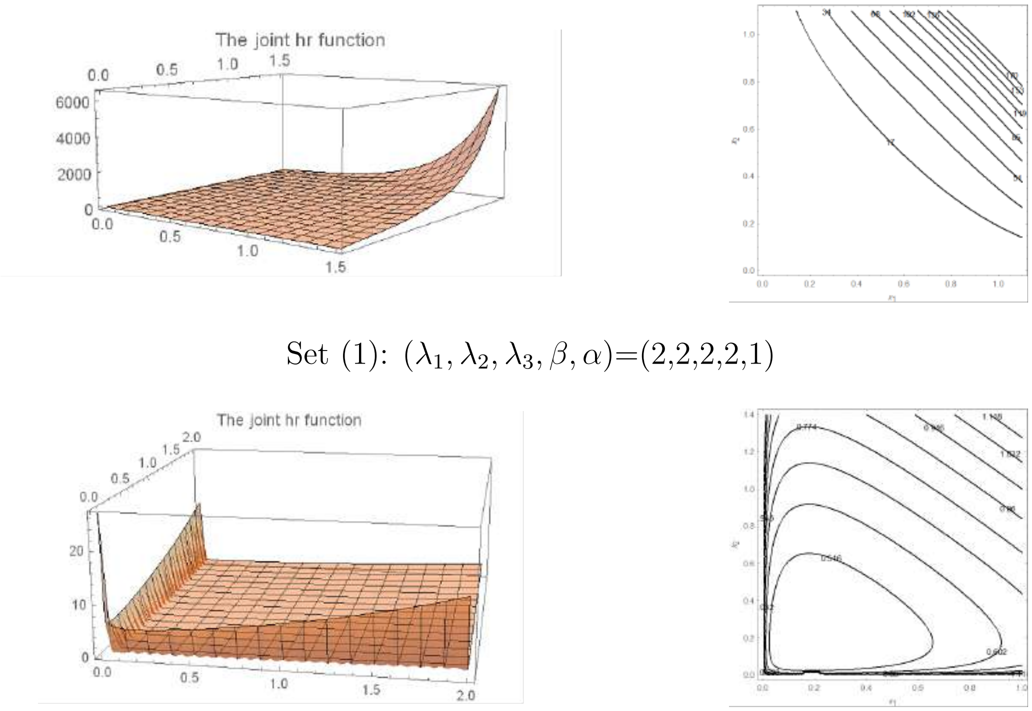
### 3.1. Marginal distributions

One can easily verify that the marginal distribution of $X_i, i = 1, 2$, follows GCh$(\beta, \alpha, \lambda_i)$. For this, we first derive the marginal survival function of $X_i$, say $S_{X_i}(x)$, as follows

$$S_{X_i}(x) = P(X_i > x) = P(min(T_i, T_3) > x) = P(T_i > x, T_3 > x)$$

and since $T_i, i = 1, 2$ and $T_3$ are independent random variables, then

$$S_{X_i}(x) = P(T_i > x)P(T_3 > x) = S_{X_i}(x; \beta, \alpha, \lambda_{i3}) = e^{\lambda_{i3}\alpha(1-e^{(x/\alpha)^{\beta}})} \quad (7)$$



Set (1): $(\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$=(2,2,2,2,1)

Set (2): $(\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$=(0.2,0.2,0.8,0.8,1)

**Figure 2: The joint hazard rate function of the BGCh distribution and corresponding contour**

Using (7), the marginal pdf of $X_i$ is

$$f_{X_i}(x) = \lambda_{i3}\beta(x/\alpha)^{(\beta-1)}e^{(x/\alpha)^{\beta}}e^{\lambda_{i3}\alpha(1-e^{(x/\alpha)^{\beta}})}, \tag{8}$$

and the marginal hazard rate function (hrf) of $X_i$ is

$$h_{X_i}(x) = \lambda_{i3}\beta(x/\alpha)^{(\beta-1)}e^{(x/\alpha)^{\beta}}. \tag{9}$$

Xie *et al.* (2002) noted that the hrf depends only on the shape parameter $\beta$ and they observed that: when $\beta > 1$, the hrf has an increasing shape and when $\beta < 1$, the hrf has a bathtub shape. Shapes of the pdf and hrf of $X_i$ for different values of $\beta, \alpha$ and $\lambda_{i3}$ are provided in Figure 3 . Also, Xie *et al.* (2002) showed that the GCh distribution can be used in modeling bathtub-shaped failure rate univariate lifetime data. Hence, we expect the BGCh distribution can be used in modeling bathtub-shaped failure rate bivariate lifetime data.

### 3.2. Conditional distributions

Having obtained the marginal pdf of $X_1$ and $X_2$, one can derive the conditional probability density function. The following theorem provides the conditional pdf of $X_1$ given $X_2 = x_2$, say $f_{X_1|X_2}(x_1|x_2)$.

**Theorem 4:** If $(X_1, X_2)$ follows BGCh$(\beta, \alpha, \lambda_1, \lambda_2, \lambda_3)$, then the conditional pdf of $X_1$ given

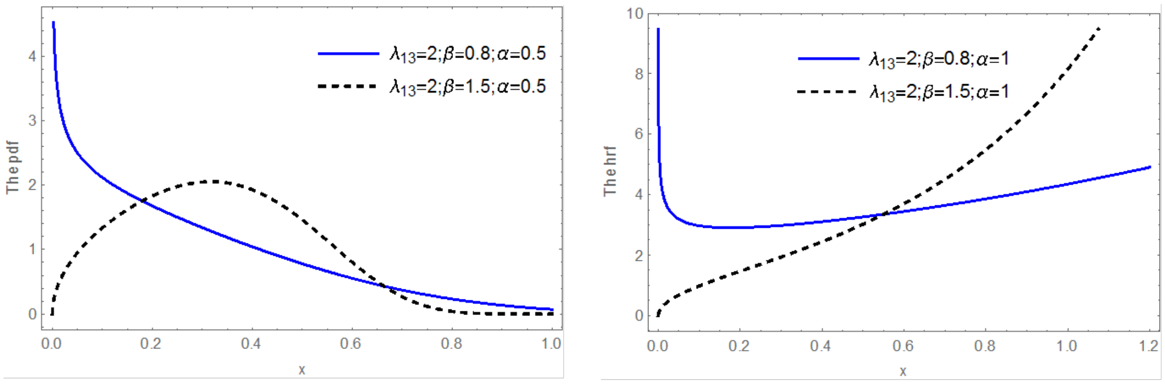


**Figure 3: The probability density and hazard rate functions of the marginal distribution of $X_1$**

$X_2 = x_2$ is

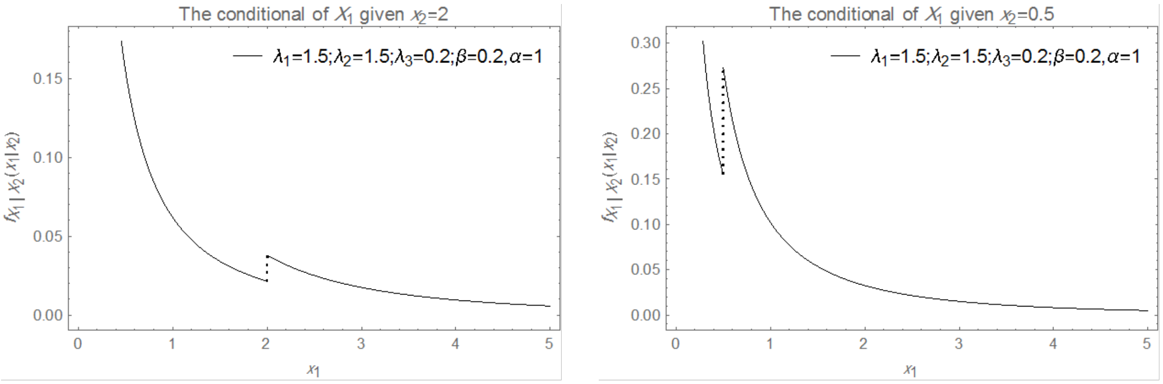
$$f_{X_1|X_2}(x_1, x_2) = \begin{cases} f_1(x_1|x_2) & \text{if } x_1 < x_2 \\ f_2(x_1|x_2) & \text{if } x_2 < x_1 \\ f_3(x_1|x_2) & \text{if } x_1 = x_2 = x, \end{cases} \tag{10}$$

where

$$\begin{aligned} f_1(x_1|x_2) &= \lambda_1\beta(x_1/\alpha)^{(\beta-1)}e^{(x_1/\alpha)^\beta}e^{\lambda_1\alpha(1-e^{(x_1/\alpha)^\beta})}, \\ f_2(x_1|x_2) &= (\lambda_{12}\lambda_2)/\lambda_{23}\beta(x_1/\alpha)^{(\beta-1)}e^{(x_1/\alpha)^\beta}e^{\alpha(\lambda_{13}(1-e^{(x_1/\alpha)^\beta})-\lambda_3(1-e^{(x_2/\alpha)^\beta}))}, \text{ and} \\ f_3(x_1|x_2) &= \lambda_3/\lambda_{23}e^{\lambda_1\alpha(1-e^{(x/\alpha)^\beta})}. \end{aligned}$$

**Proof**: The results of this theorem are easily derived using the definition of conditional probability and the results of Theorem 2 and the form (8). Figure 4 shows some plots of the conditional pdf's of $X_1$ given $X_2 = x_2$ for different values of $x_2(x_2 = 0.5, 1, 2)$ and different values of parameters.

Similarly, the conditional pdf of $X_2$ given $X_1 = x_1$ can be obtained in a similar manner as above. Also, one can note that if $\alpha = 1$, the conditional pdf in the case of BCh distribution can be obtained.



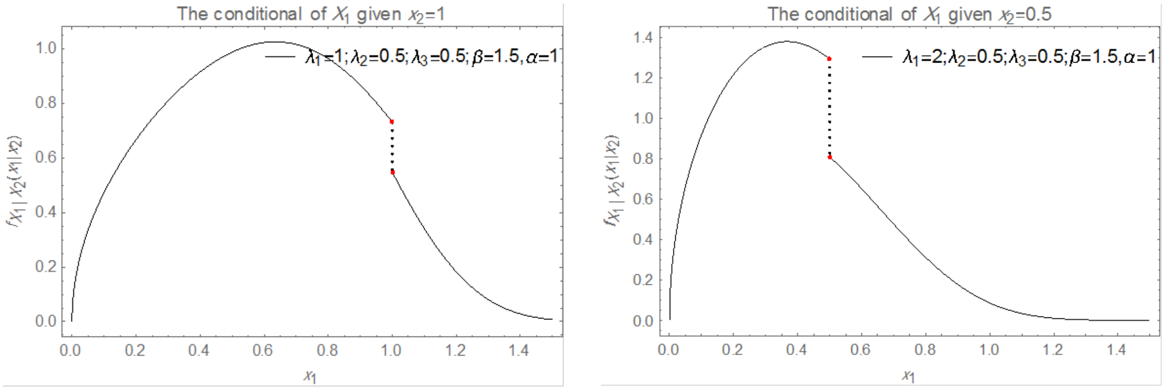


**Figure 4: The conditional probability density function of $X_1$ given $X_2 = x_2$ at different sets of the parameters**

## 4. Parameters' estimation

Suppose that $\{(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})\}$ is a random sample from BGCh $(\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$. The likelihood function for this sample is

$$L(\text{data};\theta) = \prod_{i=1}^{n} f_1(x_{1i}, x_{2i})^{I(x_{1i}<x_{2i})} f_2(x_{1i}, x_{2i})^{I(x_{1i}>x_{2i})} f_3(x_{1i}, x_{2i})^{I(x_1 i=x_{2i})}, \tag{11}$$

where $I(A)$ is an indicator function that is equal to 1 if $A$ is true and 0 otherwise and $\theta = (\lambda_1, \lambda_2, \lambda_3, \beta, \alpha)$. Substituting (5) in (11) and taking the natural logarithm, we obtain the log-likelihood function as

$$\begin{aligned}\mathcal{LL} = \sum_{i=1}^{n} & I(x_{1i} < x_{2i})\{ln(\lambda_1) + ln(\lambda_{23}) + 2ln(\beta) + (\beta-1)ln(x_{1i}/\alpha) + ln(x_{2i}/\alpha) + (x_{1i}/\alpha)^\beta \\ & + (x_{2i}/\alpha)^\beta + \lambda_1(1 - e^{(x_{1i}/\alpha)^\beta}) + (\lambda_{23})(1 - e^{(x_{2i}/\alpha)^\beta}))\} \\ & + I(x_{1i} > x_{2i})\{ln(\lambda_2) + ln(\lambda_{13}) + 2ln(\beta) + (\beta-1)(ln(x_{1i}/\alpha) + ln(x_{2i}/\alpha) + (x_{1i}/\alpha)^\beta \\ & + (x_{2i}/\alpha)^\beta + \lambda_2(1 - e^{(x_{2i}/\alpha)^\beta}) + (\lambda_{13})(1 - e^{(x_{1i}/\alpha)^\beta})\} \\ & + I(x_{2i} = x_{1i})\{ln(\lambda_3) + ln(\beta) + (\beta-1)ln(x_{1i}/\alpha) + (x_{1i}/\alpha)^\beta + (\lambda_{123})(1 - e^{(x_{1i}/\alpha)^\beta})\}.\end{aligned} \tag{12}$$

### 4.1. Maximum likelihood estimation

Here we use maximum likelihood method to estimate the unknown parameters of the BGCh distribution. For fixed $\alpha$, the likelihood equations are

$$
\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \lambda_1} &= \frac{n_1}{\lambda_1} + \frac{n_2}{\lambda_{13}} + \sum_{i=1}^{n} I(x_{1i} < x_{2i})(1 - e^{(x_{1i}/\alpha)^{\beta}}) = 0, \\
\frac{\partial \mathcal{LL}}{\partial \lambda_2} &= \frac{n_1}{\lambda_{23}} + \frac{n_2}{\lambda_2} + \sum_{i=1}^{n} I(x_{1i} > x_{2i})(1 - e^{(x_{2i}/\alpha)^{\beta}}) = 0, \\
\frac{\partial \mathcal{LL}}{\partial \lambda_3} &= \frac{n_1}{\lambda_{23}} + \frac{n_2}{\lambda_{13}} + \frac{n_3}{\lambda_3} + \sum_{i=1}^{n} \{I(x_{1i} < x_{2i})(1 - e^{(x_{2i}/\alpha)^{\beta}}) \\
&\quad + \{I(x_{1i} > x_{2i}) + I(x_{1i} = x_{2i})\}(1 - e^{(x_{1i}/\alpha)^{\beta}})\} = 0, \\
&\text{and} \\
\frac{\partial \mathcal{LL}}{\partial \beta} &= \sum_{i=1}^{n} I(x_{1i} < x_{2i})\{2/\beta + ln(x_{1i}/\alpha) + ln(x_{2i}/\alpha) + (x_{1i}/\alpha)^{\beta} ln(x_{1i}/\alpha)(1 - \lambda_1 e^{(x_{1i}/\alpha)^{\beta}}) \\
&\quad + (x_{2i}/\alpha)^{\beta} ln(x_{2i}/\alpha)(1 - \lambda_{23}) e^{(x_{2i}/\alpha)^{\beta}}\} \\
&\quad + I(x_{1i} > x_{2i})\{2/\beta + ln(x_{1i}/\alpha) + ln(x_{2i}/\alpha) + (x_{1i}/\alpha)^{\beta} ln(x_{1i}/\alpha)(1 - \lambda_{13} e^{(x_{1i}/\alpha)^{\beta}}) \\
&\quad + (x_{2i}/\alpha)^{\beta} ln(x_{2i}/\alpha)(1 - \lambda_2) e^{(x_{2i}/\alpha)^{\beta}})\} \\
&\quad + I(x_{2i} = x_{1i})\{1/\beta + ln(x_{1i}/\alpha) + (x_{1i}/\alpha)^{\beta} ln(x_{1i}/\alpha)(1 - \lambda_{123} e^{(x_{1i}/\alpha)^{\beta}})\} = 0,
\end{aligned}
\tag{13}
$$

where $n_1 = \sum_{i=1}^{n} I(x_{1i} < x_{2i})$ , $n_2 = \sum_{i=1}^{n} I(x_{1i} > x_{2i})$, and $n_3 = \sum_{i=1}^{n} I(x_{1i} = x_{2i})$. The likelihood equations (13) do not have a closed-form solution, so a numerical technique must be used to find the maximum likelihood estimates (mles) of $\lambda_1, \lambda_2, \lambda_3,$ and $\beta$. The likelihood equations may have multiple roots, Small *et al.* (2000) discussed this problem using the Hessian matrix. They showed that the likelihood equations have a unique root when the Hessian matrix of the log-likelihood is negative definite for all value of $\boldsymbol{\theta}$. This relies on maximizing the log-likelihood function. The Hessian matrix is written as

$$
T(\boldsymbol{\theta}) = \begin{pmatrix}
\mathcal{LL}_{\lambda_1\lambda_1} & \mathcal{LL}_{\lambda_1\lambda_2} & \mathcal{LL}_{\lambda_1\lambda_3} & \mathcal{LL}_{\lambda_1\beta} \\
\mathcal{LL}_{\lambda_2\lambda_1} & \mathcal{LL}_{\lambda_2\lambda_2} & \mathcal{LL}_{\lambda_2\lambda_3} & \mathcal{LL}_{\lambda_2\beta} \\
\mathcal{LL}_{\lambda_3\lambda_1} & \mathcal{LL}_{\lambda_3\lambda_2} & \mathcal{LL}_{\lambda_3\lambda_3} & \mathcal{LL}_{\lambda_3\beta} \\
\mathcal{LL}_{\beta\lambda_1} & \mathcal{LL}_{\beta\lambda_2} & \mathcal{LL}_{\beta\lambda_3} & \mathcal{LL}_{\beta\beta}
\end{pmatrix}
$$

where $\mathcal{LL}_{\theta_i\theta_j} = \frac{\partial^2 \mathcal{LL}}{\partial\theta_i \partial\theta_j}$ is the second partial derivative of the log-likelihood function with respect to the components $\theta_i$ and $\theta_j$ of $\boldsymbol{\theta}$ and $T(\hat{\boldsymbol{\theta}})$ is the Hessian matrix computed at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

**Large-sample confidence intervals:** Under regularity conditions, the mles of the parameters $\lambda_1, \lambda_2, \lambda_3,$ and $\beta$ are asymptotically normally distributed with means equal to the true values of these parameters and variances given by the inverse of the information matrix. One can approximate the expected values of the second-order derivatives of logarithms of likelihood function with the maximum likelihood estimates of the parameters as given in Cohen (1965). That is, using normality property of mles, one can construct the asymptotic confidence interval for each parameter.

### 4.2. Bayes estimation

Now, we discuss the Bayesian estimation of the unknown parameters of the BGCh distribution. For fixed $\alpha$, let the four parameters $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \beta)$ are independent random variables and follow the gamma prior distribution. That is, the joint prior pdf of $\boldsymbol{\theta}$ is

$$g_0(\boldsymbol{\theta}) \propto \lambda_1^{(a_1-1)}\lambda_2^{(a_2-1)}\lambda_3^{(a_3-1)}\beta^{(a_4-1)}e^{(-b_1\lambda_1-b_2\lambda_2-b_3\lambda_3-b_4\beta)}, \lambda_1, \lambda_2, \lambda_3, \beta > 0, \tag{14}$$

where all the hyperparameters $a_i$ and $b_i, i = 1, 2, 3, 4$ are assumed to be positive and known. The log-prior density function is

$$\mathcal{g}_0(\boldsymbol{\theta}) \propto \sum_{i=1}^{3}(a_i - 1)ln(\lambda_i) + (a_4 - 1)ln(\beta) - \sum_{i=1}^{3} b_i\lambda_i - b_4\beta. \tag{15}$$

Using (12) and (15) and applying Bayes theorem, the joint posterior probability density function of $\boldsymbol{\theta}$, given data, is

$$g(\boldsymbol{\theta}|data) = \frac{1}{K}exp(\mathcal{LL} + \mathcal{g}_0(\boldsymbol{\theta})), \tag{16}$$

where $K$ is the normalizing constant. Bayes estimators of the unknown parameters and/or of any function of the unknown parameters, say $w(\boldsymbol{\theta})$, can be obtained as follows

$$\hat{w}(data) = \frac{\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty w(\boldsymbol{\theta})exp(\mathcal{LL} + \mathcal{g}_0(\boldsymbol{\theta}))d\lambda_1 d\lambda_2 d\lambda_3 d\beta}{\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty exp(\mathcal{LL} + \mathcal{g}_0(\boldsymbol{\theta}))d\lambda_1 d\lambda_2 d\lambda_3 d\beta}. \tag{17}$$

Formula (17) involves a ratio of two multidimentional integrals and does not have analytical solution. Thus, some approximation methods were suggested to approximate these integrals and calculate the ratio of the integrals such as the methods discussed by Lindley (1980) and Tierney and Kadane (1986). These methods work well for low dimensions. In this paper we will use Markov Chain Monte Carlo (MCMC) method that work well in the case of high dimensions, see Gelman *et al.* (2003). MCMC method generates random draws from the joint posterior distribution by generating draws from an arbitrary distribution (proposal distribution) that easy to simulate from then apply an accept-reject method. Here, we use multivariate normal as a proposal distribution. The following steps can be followed to generate random draws from the joint posterior distribution (16):

1. Specify the size of the random draws we wish to generate, say m.
2. Choose an initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$.
3. For $i = 1, 2, \ldots, m$, repeated the following steps:
    (a) Generate $\boldsymbol{\theta}^*$ from the multivariate normal with mean $\boldsymbol{\theta}^{(i-1)}$ and variance-covariance $\Sigma$.
    (b) Compute the ratio $\kappa = \min\{1, \frac{g(\boldsymbol{\theta}^*|\text{data})}{g(\boldsymbol{\theta}^{(i-1)}|\text{data})}\}$.
    (c) Generate a random value from uniform distribution on (0, 1).
    (d) If $\kappa \geq$ put $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$, otherwise put $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$.

Discarding the early $m_0$ number of burn-in draws and using the remaining $m-m_0$, $\boldsymbol{\theta}^{(m_0+1)}$, $\boldsymbol{\theta}^{(m_0+2)}, \ldots, \boldsymbol{\theta}^{(m)}$, as the chosen draws from the joint posterior distribution, the Bayes estimate of $\theta_j$ is

$$\hat{\theta}_j = \sum_{i=m_0+1}^{m-m_0} \frac{\theta_j^{(i)}}{m-m_0}, j = 1, 2, 3, 4.$$

Furthermore, for $0 < \nu < 1$, one can obtain the lower and upper bounds of the $100(1-\nu)\%$ Bayesian probability interval of $\theta_j$ via $(\nu/2)100th$ and $(1-\nu/2)100th$ percentiles of the sequence of the $m-m_0$ draws; $\boldsymbol{\theta}^{(m_0+1)}, \boldsymbol{\theta}^{(m_0+2)}, \ldots, \boldsymbol{\theta}^{(m)}$.

## 5. Simulation results and applications

In this section, some simulation results and the analysis of a data set are presented.

### 5.1. Simulation results

In this section, we provide the following steps to generate a random sample of the BGCh distribution:

1. Generate $u_1, u_2$ and $u_3$ from uniform (0, 1).
2. Compute $t_1 = \alpha(ln(1 - \frac{ln(1-u_1)}{\lambda_1 \alpha})^{1/\beta}, t_1 = \alpha(ln(1 - \frac{ln(1-u_2)}{\lambda_2 \alpha})^{1/\beta}$ and $t_3 = \alpha(ln(1 - \frac{ln(1-u_3)}{\lambda_3 \alpha})^{1/\beta}$.
3. Obtain $x_1 = min(t_1, t_3)$ and $x_2 = min(t_2, t_3)$.

To obtain some simulation results for samples size (n=100) and for different parameter values, we consider three different sets of parameter values namely: (i) $\lambda_1 = \lambda_2 = \lambda_3 = \beta = 1$, (ii) $\lambda_1 = \lambda_2 = \lambda_3 = 2, \beta = 1$, and (iii) $\lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 1, \beta = 1.5$. We replicate the process 1000 times and report the average estimates and the root mean square errors (RMSEs) in Table 1. Also, we compute the Bayes estimates of the unknown parameters as mentioned in the previous section with assuming uniform priors. We simulate 10000 runs and replicate the process 1000 times. The average estimates and the RMSEs are also listed in Table 1 and one can note that results of Bayes estimates are better than mles.

### 5.2. Applications

In this section we present the analysis of a data set to discuss how the proposed distribution can be used in practice. This data represents the UEFA Champion's League Data and it was analyzed in Meintanis (2007) using the Marshall-Olkin exponential model (MO) and by Kundu and Gupta (2009) using the bivariate generalized exponential (BVGE) model, then by Sarhan (2019) using the bivariate generalized Rayleigh (BVGR) model. Kundu and Gupta (2009) reported that the BVGE model fits the data better than MO model and Sarhan (2019) reported that the BVGR model fits the data better than both the MO and the BVGE models. Here, we use the BCh model to reanalyze the same data and compare it with the three models; the MO, the BVGE and the BVGR but first we have fitted Ch$(\beta, \lambda)$ model to the marginal and the minimum of the two marginals. The mles of

**Table 1: The mles and the Bayes estimates and their RMSEs (in parentheses) of the parameters**

| Parameter value | Method | $\beta$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ |
|---|---|---|---|---|---|
| | MLE | 1.3834 | 2.6776 | 1.3080 | 1.5323 |
| | | (0.4243) | (1.6961) | (0.3732) | (0.5492) |
| (1.0, 1.0, 1.0, 1.0) | Bayes | 1.3429 | 1.4783 | 1.2716 | 1.4297 |
| | | (0.3609) | (0.4788) | (0.2993) | (0.4316) |
| | MLE | 2.9372 | 5.0508 | 2.8083 | 1.5604 |
| | | (1.9725) | (3.0754) | (0.8590) | (0.4518) |
| (1.0, 2.0, 2.0, 2.0) | Bayes | 2.3606 | 2.4705 | 2.3693 | 1.3772 |
| | | (1.3661) | (0.4714) | (0.3867) | (0.6247) |
| | MLE | 1.7625 | 4.3562 | 1.5935 | 1.4858 |
| | | (0.4276) | (3.8752) | (1.1251) | (0.4916) |
| (1.5, 0.5, 0.5, 1.0) | Bayes | 0.9269 | 0.9891 | 1.3696 | 1.3696 |
| | | (0.5764) | (0.4892) | (0.8753) | (0.3736) |

the unknown parameters, the Kolmogorov-Smirnov (K-S) distances between the empirical distribution function (EDF) and the fitted distribution function and the associated p values are reported in Table 2. Based on the p values, one can observe that Chen distribution may be used to fit $X_1, X_2$ and $min(X_1, X_2)$.

**Table 2: The mles of the parameters, the K-S test statistics and associated p-values**

| Variable | mle | K-S | $p$-value |
|---|---|---|---|
| $X_1$ | $\hat{\beta} = 0.403, \hat{\lambda} = 0.010$ | 0.013 | 0.572 |
| $X_2$ | $\hat{\beta} = 0.379, \hat{\lambda} = 0.184$ | 0.106 | 0.804 |
| $\min(X_1, X_2)$ | $\hat{\beta} = 0.389, \hat{\lambda} = 0.019$ | 0.094 | 0.899 |

Now, to test whether BCh distribution fits the data or not, we use the two-dimensional Kolomogorov-Sminrov test of goodness of fit as proposed by Peacock (1983). Using the computational environmental R peacock package, we obtain the value of test statistic as 0.2712 with p value 0.6482. Based on the p value, we cannot reject the null hypothesiss that the data came from the BCh distribution at 0.05 level of significance. For more details about multivariate Kolomogorov-Sminrov test of goodness of fit see Justel *et al.* (1997).

Hence, we have used the BCh model to analyze the bivariate data set. We use R to get mles of the unknown parameters. Table 3 shows the mles of the unknown parameters of the proposed distribution together with the values of the log-likelihood values and the Akaike information criterion (AIC=-2 LL+2k,k is the number of estimated parameters; see Akaike, 1974). The AIC suggests that the BCh distribution provides a better fit than the three models; the MO, the BVGE and the BVGR.

To indicate that a unique root for the likelihood equations exist. We use the estimates $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\beta}$ obtained with respect to the given bivariate data set. These estimates

**Table 3: The mles of the parameters, the log-likelihood values and AIC values**

| Model | mle | $\mathcal{L}$ | AIC |
|---|---|---|---|
| MO | $\hat{\lambda}_1 = 0.012, \hat{\lambda}_2 = 0.014, \hat{\lambda}_3 = 0.022,$ | -339.006 | 684.012 |
| BVGE | $\hat{\alpha}_1 = 1.351, \hat{\alpha}_2 = 0.465, \hat{\alpha}_3 = 1.153, \hat{\beta} = 0.039$ | -296.935 | 601.870 |
| BVGR | $\hat{\alpha}_1 = 0.492, \hat{\alpha}_2 = 0.166, \hat{\alpha}_3 = 0.410, \hat{\lambda} = 0.020$ | -293.357 | 594.714 |
| BCh | $\hat{\lambda}_1 = 0.026, lam\hat{b}da_2 = 0.055, \hat{\lambda}_3 = 0.048, \hat{\beta} = 1.020$ | 0.094 | 0.899 |

are obtained using nlm R package which minimize the negative of the log-likelihood function. We obtain $T(\hat{\boldsymbol{\theta}})$ as follows:

$$\begin{pmatrix} 0.0424 & -1.63E-0710^{-7} & 9.8070 & 6.0055 \\ - & 33.114 & 2.0183 & 8.427 \\ - & - & 46.709 & 11.059 \\ - & - & - & 104.0074 \end{pmatrix}$$

The eigen values of this matrix are -103.6646, -42.8617, -31.9991 and -2.2724. This indicates that $T(\hat{\boldsymbol{\theta}})$ is negative definite. Then according to Small *et al.* (2000), the likelihood equations has a unique root. For more details see Thomas and Jose (2021).

For Bayesian computations, we obtain the Bayes estimates of the unknown parameters based on the uniform priors and the gamma priors. In the case of the gamma priors, we assume that all hyperparameters equal and equal to 0.5. For the two cases, the proposal distribution is multinormal with variance covariance matrix and the choice of its value depends on the acceptance rate which is assumed such that the acceptance rate (number of accepted runs out of total runs) increases. Here, we simulate 10000 runs from the joint posterior distribution of the four parameters and the early 20% of the runs were discarded. The trace plots of the draws are plotted in Figures 5 and 6 after discarding the early 2000 draws (burn-in period). Tables 4-5 list the posterior descriptive summaries of interest such as the posterior mean, median, standard deviation and the 95% Bayesian credible intervals.

**Table 4: Summary results for the posterior parameters in the case of gamma priors (the acceptance rate is** 38.18%)

| Parameter | Mean | Median | Standard deviation | 95% credible intervals |
|---|---|---|---|---|
| $\lambda_1$ | 0.0837 | 0.0754 | 0.0408 | (0.0533, 0.1053) |
| $\lambda_2$ | 0.5379 | 0.5278 | 0.1352 | (0.4347, 0.6332) |
| $\lambda_3$ | 0.1707 | 0.1604 | 0.0620 | (0.1274, 0.2031) |
| $\beta$ | 0.2114 | 0.2114 | 0.0168 | (0.1995, 0.2227) |

## 6. Conclusion

In this paper, the bivariate generalized Chen distribution (BGCh) is proposed as a new bivariate lifetime distribution. The BGCh distribution is of Marshal-Olkin type whose marginal are generalized Chen distributions. One can observe that the BGCh distribution is

**Table 5: Summary results for the posterior parameters in the case of uniform priors (the acceptance rate is** 54.83%)

| Parameter | Mean | Median | Standard deviation | 95% credible intervals |
|---|---|---|---|---|
| $\lambda_1$ | 0.1137 | 0.1042 | 0.0533 | (0.0756, 0.1401) |
| $\lambda_2$ | 0.6255 | 0.6134 | 0.1575 | (0.5175, 0.7273) |
| $\lambda_3$ | 0.2227 | 0.2167 | 0.0779 | (0.1657, 0.2697) |
| $\beta$ | 0.2019 | 0.2015 | 0.0171 | (0.1900, 0.2127) |

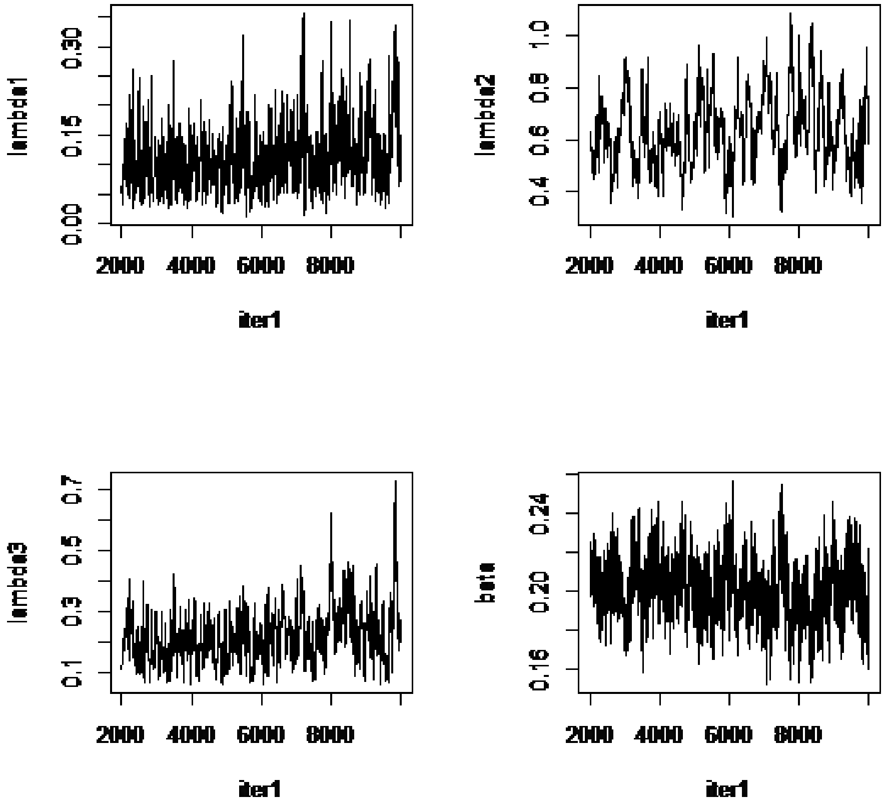


**Figure 5: The trace plot of the random draws from the joint posterior distribution in the case of gamma priors**

a singular distribution and has an absolute continuous and a singular part. Some statistical properties are investigated. The estimation of the parameters has been approached by maximum likelihood and Bayesian methods. For Bayesian method, we used the MCMC method. Numerical methods are required to calculate the desired estimates. One real data set is analyzed using the BCh distribution which showed a better fit than the MO, the BVGE and the BVGR distributions.

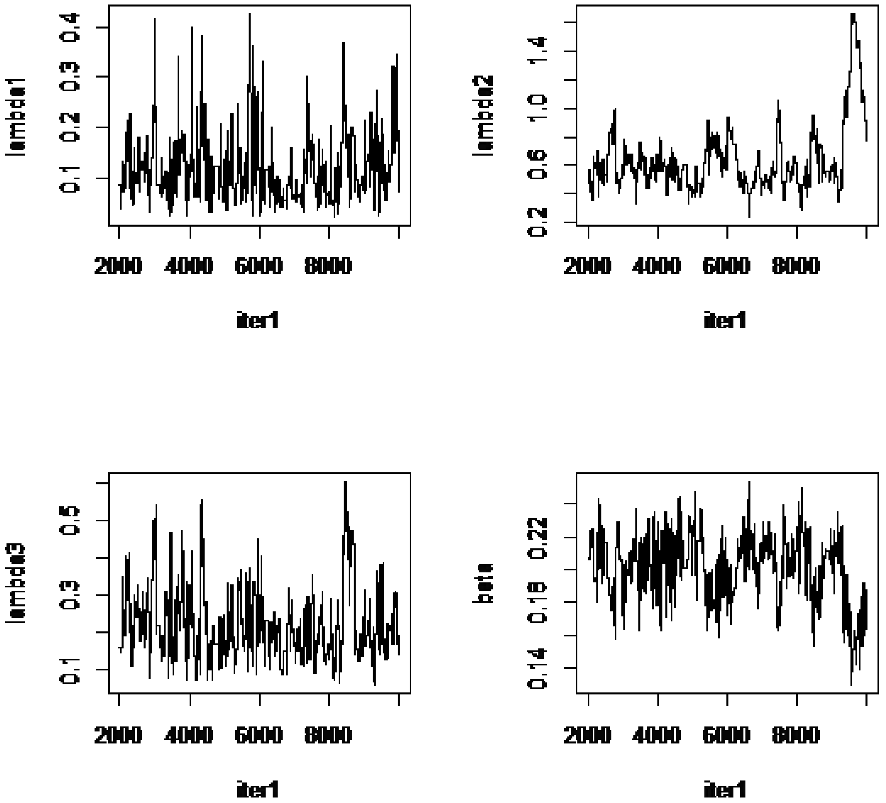


**Figure 6: The trace plot of the random draws from the joint posterior distribution in the case of uniform priors**

## Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Azizi, A., Sayyareh, A. and Panahi, H. (2019). Inference about the bivariate new extended Weibull distribution based on complete and censored data. *Communications in Statistics-Simulation and Computation*, **51**, 738–756..

Bemis, B., Bain, L. J. and Higgins, J. J. (1972). Estimation and hypothesis testing for the parameters of a bivariate exponential distribution. *Journal of the American Statistical Association*, **67**, 927–929.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, **49**, 155–161.

Cohen, A. C. (1965). Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. *Technometrics*, **7**, 579-–588.

El-Gohary, A., El-Bassiouny, A. H. and El-Morshedy, M. (2016). Bivariate exponentiated modified Weibull extension distribution. *Journal of Statistics Applications and Probability*, **5**, 67–78.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis.* 2nd Edition, New York: Chapman and Hall/CRC.

Justel, A., Pena, D. and Zamar, R. (1997). A multivariate Kolomogrov-Smirnov test of goodness of fit. *Statistics and Probability Letters*, **35**, 251–291.

Kundu, D. and Gupta, R. D. (2009). Bivariate generalized exponential distribution. *Journal of Multivariate Analysis*, **100**, 581–593.

Kundu, D. and Gupta, A. (2017). On bivariate inverse Weibull distribution. *Brazilian Journal of Probability and Statistics*, **31**, 275-–302.

Kundu, D., Sarhan, A.M. and Gupta, R.D. (2012). On Sarhan-Balakrishnan bivariate distribution. *Journal of Statistics and Probability*, **1**, 163–170.

Lindley, D. V.(1980). Approximate Bayesian methods. *Trabajos de Estadistica*, **31**, 223–245.

Marshal, A. W. and Olkin, I. A. (1967). A multivariate exponential distribution. *Journal of American Statistical Association*, **62**, 30–44.

Meintanis, S. G. (2007). Test of fit for Marshall-Olkin distributions with applications. *Journal of Statistical Planning and Inference*, **137**, 3954–3963.

Mudholkar, G. S. and Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data, *IEEE Transactions on Reliability*, **42**, 299–302.

Muhammed, H. Z. (2019). Bivariate Generalized Burr and Related Distributions: Properties and Estimation. *Journal of Data Science*, **17**, 532 –548.

Murthy, D. N., Xie, M. and Jiang, R. (2004). *Weibull Models.* John Wiley & Sons, Inc.

Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, **202**, 615-–627

Sarhan, A. M. (2019). The Bivariate Generalized Rayleigh Distribution. *Journal of Mathematical Sciences and Modelling*, **2**, 99–111.

Sarhan, A. M. and Balakrishnan, N. (2007). A new class of bivariate distributions and its mixture. *Journal of Multivariate Analysis*, **98**, 1508–1527.

Small, C. G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statistical Science*, **15**, 313-–341.

Smith, R. M. and Bain, L. J. (1975). An exponential power life-testing distribution. *Communications in Statistics-Theory and Methods*, **4**, 469–481.

Pham, H. and Lai, C. D. (2007). On recent generalizations of the Weibull distribution. *IEEE Transactions on Reliability*, **56**, 454–458.

Thomas, P Y. and Jose, J. (2021). On Weibull-Burr impounded bivariate distribution. *Japanese Journal of Statistics and Data Science*, **4**, 73–105.

Tierney, L. and Kadane, D. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association*, **81**, 82–86.

Xie, M., Tang, Y. and Goh, T. N. (2002). A modified Weibull extension with bathtub-shaped failure rate function. *Reliability Engineering and System Safety*, **76**, 279–285.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 179–185

# Properties of Partial Product Processes

**Babu D., Sutha M., Neelakandan R. and Elumalai P.**
*Department of Mathematics*
*Thanthai Periyar Government Institute of Technology, Vellore, India*

Received: 14 January 2022; Revised: 07 May 2022; Accepted: 08 July 2022

## Abstract

The partial product process is a sequence of non-negative random variables $X_1, X_2, X_3, ...$ such that the distribution function of $X_1$ is $F(x)$ and the distribution function of $X_{i+1}$ is $F(\beta_i x) \quad (i = 1, 2, 3, ...)$ where $\beta_i > 0$ are constants and $\beta_i = \beta_0 \beta_1 \beta_2 ... \beta_{i-1}$. It is a monotone process. In this paper, the probabilistic properties of the partial product process are studied and some limit theorems are also established.

*Key words:* Geometric process; Partial product process; Renewal process.

**AMS Subject Classifications:** 60K10, 90B25

## 1. Introduction

The mathematical theory of reliability has put forth a great effort to issues of life-testing, machine support, replacement, order statistics, and so on. The maintenance problems are concerned about the circumstance that emerges about the reduction of the productivity level of items or breakdown. The problem of replacement is to recognize the best policy which enables determination of ideal replacement time that is generally economical. One of the most interesting and critical topics to study in reliability is the study of maintenance problems.

A common assumption in the initial period of studying maintenance issues is that repair is perfect, a repairable framework after the repair is *as good as new.* This assumption clearly has the effect of a natal way. In practice, most repairable systems deteriorate because of the combined wear and tear impact. Barlow (1960) thusly presented a minimal repair model in which a system after the repair has the same failure rate and effective age as it was when it failed. Brown(1983) proposes an imperfect repair model, in which the repair is perfect with likelihood $p$, and the repair is minimal with a probability of $1 - p$.

Deteriorating systems have a different problem as the one portrayed above. For instance, in machine maintenance problems, after every repair, the working time of a machine will end up shorter and shorter, so the absolute working time or the existence of the machine must be limited. However, in perspective on the aging and aggregate wear, the repair time will turn out to be longer and tend to increase so that at the end the machine is non-repairable. Therefore, there is need to consider a repair replacement model for deteriorating systems, the progressive survival times are diminishing, while the consecutive repair times are expanding.

Lam (1988) first presented a Geometric Process Repair model to model a deterio-

Corresponding Author: Babu D.
Email: babudurai15@gmail.com

rating system with the above characteristics.

**Definition 1:** The sequence $\{X_n,\ n=1,2,3,\ldots\}$ of non negative independent random variables is called a geometric process, if the distribution function of $X_n$ is given by $F\left(a^{n-1}x\right)$ for $n=1,2,3,\ldots$, where $a(>0)$ is a constant.

In Geometric process, the operating times and repair times of a system are uniformly decreasing. But practically, it is not uniform. To overcome this, the partial product process was introduced by Babu et. al (2018).

**Definition 2:** Let $\{X_n,\ n=1,2,...\}$ be a sequence of non-negative independent random variables and let $F\left(x\right)$ be the distribution function of $X_1$. Then $\{X_n,\ n=1,2,...\}$ is called a partial product process, if the distribution function of $X_{i+1}$ is $F\left(\beta_i x\right) \quad (i=1,2,...)$ where $\beta_i>0$ are constants and $\beta_i=\beta_0\beta_1\beta_2...\beta_{i-1}$.

By induction, it is clear that for real $\beta_i$ $(i=1,2,...)$, $\beta_i={\beta_0}^{2^{i-1}}$. Thus, the distribution function of $X_{i+1}$ is $F\left({\beta_0}^{2^{i-1}}x\right)$ for $i=1,2,...$ .

The remainder of this paper are structured as follows. In section 2, some probabilistic properties of the partial product process are studied and in section 3, some limit theorems are established.

## 2. Probabilistic properties of partial product process

Let $F$ and $f$ be the distribution function and density function of $X_1$ respectively, and denote $E\left(X_1\right)=\lambda$ and $Var\left(X_1\right)=\sigma^2$.

Then for $i=1,2,...$, we have

$$E\left(X_{i+1}\right)=\frac{\lambda}{\beta_0^{2^{i-1}}}$$

and

$$Var\left(X_{i+1}\right)=\frac{\sigma^2}{\beta_0^{2^i}}.$$

Thus, $\beta_0,\ \lambda$ and $\sigma^2$ are three important parameters of partial product process.

Note that when $F\left(0\right)<1$, then $\lambda>0$.

Define $S_0=0$ and

$$S_n=\sum_{i=1}^{n}X_i$$

Let $\mathbb{F}_n=\sigma\left(X_1,X_2,...,X_n\right)$ be the $\sigma$- algebra generated by $\{X_i,\ i=1,2,...,n\}$.

**Theorem 1:** If $\beta_0>1$, then $\{S_n,\ n=1,2,...\}$ is a nonnegative submartingale with respect to $\mathbb{F}_n=\sigma\left(X_1,X_2,...,X_n\right)$.

**Proof:** Obviously, $\{S_n,\ n=1,2,...\}$ is a sequence of increasing non-negative random variables with

$$E\left[S_{n+1}|\mathbb{F}_n\right]=S_n+E\left[X_{n+1}\right]\geqslant S_n \tag{1}$$

Also,

$$\begin{aligned}\underset{n\geqslant 0}{Sup}\, E\left[|S_n|\right] &= \lim_{n\to\infty} E\left[S_n\right]\\ &= \lim_{n\to\infty} E\left[\sum_{i=1}^{n} X_i\right]\\ &= \lim_{n\to\infty} \sum_{i=1}^{n} E\left(X_i\right)\\ &= \lim_{n\to\infty}\left[\lambda+\sum_{i=2}^{n}\frac{\lambda}{{\beta_0}^{2^{i-2}}}\right]\\ &= \lambda\left[1+\sum_{i=2}^{\infty}\left(\frac{1}{\beta_0}\right)^{2^{i-2}}\right]<\infty, \qquad (2)\end{aligned}$$

where equation (2) is due to the fact that the series $\sum_{i=2}^{\infty}\left(\frac{1}{\beta_0}\right)^{2^{i-2}}$ is convergent by comparing with geometric series if $\frac{1}{\beta_0} < 1$.

**Theorem 2:** If $\beta_0 > 1$, there exists a random variable $S$ such that

$$S_n \xrightarrow{a.s.} S \; as \; n \to \infty.$$

**Proof:** By Theorem 1 and *Martingale Convergence Theorem*(Ross, 1996), we have $S_n \xrightarrow{a.s.} S \; as \; n \to \infty$.

**Theorem 3:** If $\beta_0 > 1$, $\{S_n,\ n = 1, 2, ...\}$ has a unique decomposition such that

$$S_n = L_n - A_n \qquad (3)$$

where $\{L_n,\ n = 1, 2, ...\}$ is a martingle, $\{A_n,\ n = 1, 2, ...\}$ is decreasing with $A_1 = 0$ and $A_n \in \mathbb{F}_{n-1}$.

**Proof:** Let $L_1 = S_1$ and $A_1 = 0$. For $n \geqslant 2$, define

$$L_n = L_{n-1} + \left(S_n - E\left[S_n|\mathbb{F}_{n-1}\right]\right), \qquad (4)$$

$$A_n = A_{n-1} + \left(S_{n-1} - E\left[S_n|\mathbb{F}_{n-1}\right]\right). \qquad (5)$$

From equations (4) and (5), we have

$$L_n - A_n = \sum_{i=2}^{n}\left(S_i - S_{i-1}\right) + S_1 - A_1 = S_n$$

and (3) follows. It is easy to check $\{L_n,\ n = 1, 2, ...\}$ and $\{A_n,\ n = 1, 2, ...\}$ satisfy the requirements. Next, we need to prove such a decomposition is unique.

Suppose $S_n = {L_n}' - {A_n}'$ is another decomposition. Then,

$$L_n - {L_n}' = A_n - {A_n}'$$

Since $A_1 = {A_1}' = 0$, it is clear that $L_1 = {L_1}'$. Also as $L_2 - {L_2}' = A_2 - {A_2}' \in \mathbb{F}_1,$, we have

$$L_2 - {L_2}' = E\left[L_2 - {L_2}'|\mathbb{F}_1\right] = L_1 - {L_1}' = 0.$$

This implies that $L_2 = {L_2}'$. Then, by induction, we can prove that $L_n = {L_n}'$ and hence $A_n = {A_n}'$. Thus the uniqueness is proved.

**Theorem 4:** If $\beta_0 > 1$, then $\{S_n,\ n = 1, 2, ...\}$ has a unique Riesz decomposition such that

$$S_n = Y_n + Z_n \tag{6}$$

where $\{Y_n,\ n = 1, 2, ...\}$ is a non-negative martingale and $\{Z_n,\ n = 1, 2, ...\}$ is a non-positive submartingale with $\lim\limits_{n\to\infty} E\left[Z_n\right] = 0$.

**Proof:** From equation (2),

$$\lim_{n\to\infty} E\left[S_n\right] = \lambda \left[1 + \sum_{i=2}^{\infty} \left(\frac{1}{\beta_0}\right)^{2^{i-2}}\right] < \infty.$$

Thus the proof is complete by Riesz decomposition theorem(Ross, 1996).

**Theorem 5:** If $\beta_0 > 1$, then $\{S_n,\ n = 1, 2, ...\}$ has a Krickeberg decomposition such that

$$S_n = L_n - M_n$$

where $\{L_n,\ n = 1, 2, ...\}$ is a non-positive submartingale and $\{M_n,\ n = 1, 2, ...\}$ is a non-positive martingale. Moreover, such a decomposition has the maximality such that for any other decomposition $S_n = {L_n}' - {M_n}'$ where ${L_n}'$ , ${M_n}'$ are nonpositive submartingale and nonpositive martingale respectively, then

$$L_n \geq {L_n}' \quad and \quad M_n \geq {M_n}'$$

**Proof:** Note that,

$$\underset{n}{Sup}\, E\left[{S_n}^{+}\right] = \underset{n}{Sup}\, E\left[S_n\right] = \lambda \left[1 + \sum_{i=2}^{\infty} \left(\frac{1}{\beta_0}\right)^{2^{i-2}}\right] < \infty.$$

Thus the proof is complete by Krickeberg decomposition(Ross, 1996).

## 3. Limit theorems for partial product process

In renewal process, Wald's equation plays a key role. The following theorem is a generalization of the Wald's equation to a *partial product process*, it is called as Wald's equation for *partial product process.*

**Theorem 6** (Wald's equation for *partial product process*)**:** Suppose that $\{X_n,\ n = 1, 2, ...\}$ forms a *partial product process* with $E\left[X_1\right] = \lambda < \infty$, then for $t > 0$, we have

$$E\left[S_{N(t)+1}\right] = \lambda E\left[1 + \sum_{n=2}^{N(t)+1} \frac{1}{{\beta_0}^{2^{n-2}}}\right], \tag{7}$$

where $N(t)$ is the counting process which represents the number of occurrences of an event up to time $t$.

**Proof:** Let $I_A$ be the indicator function of event $A$. Then $I_{\{S_{n-1}\le t\}} = I_{\{N(t)+1\ge n\}}$ and $X_n$ are independent. Accordingly, for $t > 0$, we have

$$\begin{aligned} E\left[S_{N(t)+1}\right] &= E\left[\sum_{n=1}^{N(t)+1} X_n\right] \\ &= \sum_{n=1}^{\infty} E\left[X_n I_{\{N(t)+1\ge n\}}\right] \\ &= \sum_{n=1}^{\infty} E\left[X_n\right] P\left(N\left(t\right)+1\ge n\right) \qquad (8) \\ &= \lambda E\left[1+\sum_{n=2}^{N(t)+1} \frac{1}{{\beta_0}^{2^{n-2}}}\right]. \end{aligned}$$

**Corollary 1:**

$$E\left[S_{N(t)+1}\right] \begin{cases} > \lambda\left[E\left(N\left(t\right)\right)+1\right], & 0<\beta_0<1 \\ = \lambda\left[E\left(N\left(t\right)\right)+1\right], & \beta_0 = 1 \\ < \lambda\left[E\left(N\left(t\right)\right)+1\right], & \beta_0>1. \end{cases}$$

**Proof:** Let $\beta_0 > 1$. Then,

$$\begin{aligned} \frac{1}{\beta_0}<1 \quad &\Rightarrow \quad E\left[1+\sum_{n=2}^{N(t)+1} \frac{1}{{\beta_0}^{2^{n-2}}}\right] < E\left[1+\sum_{n=2}^{N(t)+1}(1)\right] \\ &\Rightarrow \quad \lambda E\left[1+\sum_{n=2}^{N(t)+1} \frac{1}{{\beta_0}^{2^{n-2}}}\right] < \lambda E\left[N\left(t\right)+1\right] \\ &\Rightarrow \quad E\left[S_{N(t)+1}\right] < \lambda\left[E\left(N\left(t\right)\right)+1\right] \quad (\textit{by theorem } 6). \end{aligned}$$

Similarly, we can prove that $E\left[S_{N(t)+1}\right] > \lambda\left[E\left(N\left(t\right)\right)+1\right]$ if $\beta_0 < 1$. For $\beta_0 = 1$, the result is trivial.

Note that if $\beta_0 = 1$, the *partial product process* reduces to a renewal process, while corollary 1 gives $E\left[S_{N(t)+1}\right] = \lambda\left[E\left(N\left(t\right)\right)+1\right]$. This is Wald's equation for the renewal process.

**Theorem 7:** If a stochastic process $\{X_n,\ n = 1, 2, 3, ...\}$ is a *partial product process*, then

$$(1) \quad \lim_{t\to\infty} \frac{1}{t} E\left[S_{N(t)+1}\right] = 0 \quad \text{if} \ \ \beta_0 > 1.$$

$$(2) \quad \lim_{t\to\infty} \frac{1}{t} E\left[S_{N(t)+1}\right] = 1 \quad \text{if} \ \ \beta_0 = 1.$$

**Proof:** Let $\beta_0 > 1$. Then from equation (8), we have

$$\begin{aligned}
\lim_{t\to\infty} E\left[S_{N(t)+1}\right] &= \lim_{t\to\infty} \sum_{n=1}^{\infty} E\left[X_n\right] P\left(N\left(t\right)+1 \geq n\right) \\
&= \sum_{n=1}^{\infty} E\left[X_n\right] P\left(S_n < \infty\right) \\
&= \sum_{n=1}^{\infty} E\left[X_n\right](1) \\
&= \lambda\left(1+\sum_{n=2}^{\infty} \frac{1}{{\beta_0}^{2^{n-2}}}\right) < \infty.
\end{aligned}$$

Thus,

$$\lim_{t\to\infty} \frac{1}{t} E\left[S_{N(t)+1}\right] = 0.$$

This completes the proof of part (1).

Now, assume that $\beta_0 = 1$. Then,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t} E\left[S_{N(t)+1}\right] &= \lim_{t\to\infty} \frac{\lambda}{t}\left[E\left(N\left(t\right)\right)+1\right] \\
&= \lambda \lim_{t\to\infty}\left[\frac{E\left(N\left(t\right)\right)}{t}\right] \\
&= \lambda \times \frac{1}{\lambda} = 1, \qquad (9)
\end{aligned}$$

where (9) due to *Elementary Renewal Theorem.* This completes the proof of part (2) and theorem.

## 4. Conclusion

In this paper, some limit theorems and probability properties of the partial product process are established. Since it is monotone process, it can be applied to the maintenance model for a deteriorating system.

## References

Babu, D., Govindaraju, P. and Rizwan, U. (2018). Partial product processes and replacement problem. *International Journal of Current Advanced Research*, **7**, 139-142.

Barlow, R. and Hunter, L. (1960). Optimum preventive maintenance policies. *Operations Research*, **8**, 90-100.

Barlow, R. E. and Proschan, F. (1996). *Mathematical Theory of Reliability.* Society for Industrial and Applied Mathematics.

Braun, W. J., Li, W. and Zhao, Y. Q. (2005). Properties of the geometric and related processes. *Naval Research Logistics (NRL)*, **52**, 607-616.

Brown, M. and Proschan, F.(1983). Imperfect repair. *Journal of Applied Probability*, **20**, 851-859.

Finkelstein, M. S. (1993). A scale model of general repair. *Microelectronics Reliability*, **33**, 41–44.

Lam, Y. (1988). Geometric processes and replacement problem. *Acta Mathematicae Applicatae Sinica*, **4**, 366-377.

Lam, Y. (2007). *The Geometric Process and its Applications.* World Scientific.

Ross, S. M. *Stochastic Processes.* Vol. **2**. New York: Wiley, 1996.

Wang, H. (2002). A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research*, **139**, 469-489.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 187–205

# Asymptotic Results for Generalized Runs in Higher Order Markov Chains

**Anuradha**
*Department of Statistics, Lady Shri Ram College for Women*
*University of Delhi, Lajpat Nagar-IV, New Delhi – 110024, India*

Received: 06 February 2022; Revised: 29 May 2022; Accepted: 08 July 2022

**Abstract**

Consider an $m^{th}$ order Markov chain $\{X_j : j \geq -m+1\}$ taking values in $\{\mathbf{0}, \mathbf{1}\}$. Fix $k \geq 1$ and $r \geq 0$. A $r$-look-back run of length $k$, is defined as a run of $\mathbf{1}$'s, provided that there are at least $r$ trials in between the ending point of the current run and the ending point of the previous occurrence of the $r$-look-back run of length $k$. The $r$-look-back run of length $k$ encompasses the non-overlapping counting, the overlapping counting as well as the $l$-overlapping counting for $0 \leq l \leq k-1$ (defined by Aki and Hirano (2000)). We show that the waiting time for the $n^{th}$ occurrence of the $r$-look-back run of length $k$ converges in distribution to an extended Poisson distribution under the assumption that the model exhibits a strong propensity towards success. This generalizes similar results on $l$-overlapping runs of length $k$ obtained under the Markov dependent set-up. We obtain a central limit theorem for the number of $r$-look-back runs of length $k$ till the $n^{th}$ trial. Further, we show that the rate of convergence in the central limit theorem is at least a fractional power of $n$ with a logarithmic correction factor. We support our findings on the rate of convergence with some simulation results.

*Key words:* Success runs; Waiting time; Markov chain; Extended Poisson distribution; Central limit theorem; Rate of convergence.

**AMS Subject Classifications:** 60C05, 60E05, 60F05

## 1. Introduction

Let $\{X_i : i \geq 1\}$ be a sequence of $\{\mathbf{0}, \mathbf{1}\}$-valued random variables. Here $X_n$ stands for the outcome of an experiment at the $n$-th trial and $\mathbf{1}$ and $\mathbf{0}$ imply success and failure respectively of the experiment. A *run of length k* is an occurrence of $k$ $(\geq 1)$ consecutive $\mathbf{1}$'s. In the literature, there are several schemes of counting runs of length $k$; two of the most commonly used ones are (a) *Non-overlapping counting* and (b) *Overlapping counting.* In the non-overlapping counting scheme a trial can contribute to only one possible run, while in the overlapping counting scheme a trial can contribute towards the counting of more than one run. Another method of $l$-overlapping counting has been introduced by Aki and Hirano (2000) where they allow an overlap of at most $l$ successes between two consecutive runs of length $k$ where $0 \leq l < k$. It is easy to observe that when $l = 0$ and $l = k-1$, this definition is equivalent to the non-overlapping counting and the overlapping counting respectively. Han

Corresponding Author: Anuradha
Email: sarkar.anuradha@gmail.com

and Aki (2000) have extended this counting scheme to the case where $l$ assumes negative values. For $l < 0$, there is at least $|l|$ trials difference between the two runs of length $k$.

In this paper, we consider a new scheme of counting runs of length $k$. We will refer to this new scheme of counting as *look-back counting*. Let $r \geq 0$ be a fixed number. In the $r$-look-back counting scheme, the starting points (hence the ending points) of the two consecutive runs of length $k$ should be separated by at least $r$ trials in between, *i.e.*, a new run of length $k$ can be counted only after $r$ trials have elapsed since the starting point of the last counted run. Suppose that we are at the trial $i$ such that it is a starting point of a run of length $k$, *i.e.*, $X_i = X_{i+1} = \ldots = X_{i+k-1} = 1$. Now, suppose that $i'$ is the trial where the last enumerated $r$-look-back run of length $k$ started. In order to enumerate the run starting at the $i^{th}$ trial as a $r$-look-back run, we must have $i - i' > r$. The definition of $r$-look-back run of length $k$, encompasses the above definitions of overlapping runs and non-overlapping runs, in the sense that when $r = 0$ and $r = k - 1$, the $r$-look-back run of length $k$ matches exactly with the overlapping counting and the non-overlapping counting respectively. Moreover, the $l$-overlapping run of length $k$ can be identified as a $r$-look-back run of length $k$ with $l = k - r - 1$ for $0 \leq l \leq k - 1$. However, when $l$ assumes negative values the definitions do not match. To illustrate this, we quote the example from Han and Aki (2000):

$$\mathbf{11110110001111111110000111}.$$

In this example, for $k = 3$ and $r = 3$, we see that there are four 3-look-back runs of length 3 starting at trials $1, 11, 15$ and 23, while there are only three $(-1)$-overlapping runs of length 3, starting at $1, 11$ and 15. This is because the number of remaining trials (0 here) after the last run of length $k$ starting at trial 23, is less than $|l|$, in the $l$-overlapping counting of runs of length $k$ (for $l < 0$), such a run cannot be counted (see Han and Aki (2000)). But, in the look-back counting scheme, we do not put such a restriction. This will be clear from the mathematical definition given in the next section.

Practical usage of this scheme of counting can be illustrated from the following examples. In many counters for detection of cosmic rays and $\alpha$-particles, the counter records a hit (detection of a particle) whenever the frequency recorded lies in a particular region (depending on the particle under detection). We refer to the detection of the particle as a success, while the non-detection is regarded as a failure. However, if particles are detected for $k(\geq 1)$ successive time points, the counter loses its power and is locked; hence it cannot record anything in the next $r - k + 1$ $(r > k - 1)$ time points. The number of $r$-look-back runs of length $k$ is exactly the number of times the instrument loses its power and is locked. Another example is seen in the congestion model of computer networks, where a network receives packets of information from other networks and sends information back to the originating network. Each of these processes consumes certain processing resource. If the network receives packets at $k$ consecutive time points, all its resources are spent in processing the information received; as a result it can not receive any information for the next $r - k + 1$ $(r > k - 1)$ time points. In one of the models of computer networking, the packets are rejected for these time points and are required to be re-sent by the originating network at a later time point. This situation is called the congestion of the network. Here also, the number of $r$-look-back runs of length $k$ is exactly the number of times congestion occurs. In a drug administration model, observations are taken every hour for the presence or absence (success or failure) of a particular symptom, say, fever exceeding a specified temperature.

If we observe the presence of the symptom for $k$ consecutive points (hours), a drug has to be administered; however, as is the case with most drugs, once the drug is administered, we have to wait for $r$-hours for the next administration of the drug with $r < k$. But the process of the observation for the presence or absence of the symptom is continued as ever. In such a case, the number of administrations of the drug until time point $n$, is the number of $r$-look-back runs of length $k$ up to time $n$. In the first two examples, we have $r > k - 1$ while in the last example $0 \leq r \leq k - 1$.

The theory of runs plays a vital role in diverse fields of statistics, such as, non-parametric inference, statistical quality control, reliability theory *etc.*. Runs and run-related statistics have engaged researchers since the time of De Moivre (see Feller (1968)). In recent years, this field has seen tremendous growth, with researchers contributing to the theory as well as their practical applications to various disciplines. Systematic study of the theory of distributions of non-overlapping runs was initiated by Feller (1968). Feller studied the distribution of the number of non-overlapping runs up to the *n-th* trial and obtained its asymptotic distribution using the renewal theory techniques where the underlying trials were i.i.d. Bernoulli random variables. Aki (1985), Hirano (1986), Philippou and Makri (1986) *etc.* studied various run-statistics based on the non-overlapping counting of runs. Ling (1988) obtained the distribution of the number of overlapping runs of length $k$ for a sequence of $n$ i.i.d. Bernoulli trials. Aki and Hirano (1988), Godbole (1990, 1992) also studied the properties of the distribution of the number of overlapping runs up to time $n$. Hirano *et al.* (1991) obtained the probability generating function of the number of overlapping runs and also obtained the asymptotic distribution. Several generalization of the underlying model has also been considered (see Aki and Hirano (1995), Fu and Koutras (1994), Koutras (1996), Uchida and Aki (1995), Uchida (1998) and references therein). The waiting time distributions for the occurrence of runs of various types has been studied extensively by several authors (see, for example, Koutras (1996), Aki, Balakrishnan and Mohanty (1996), Balasubramanian, Viveros and Balakrishnan (1993) and references therein). Uchida (1998) has also investigated the waiting time problems for patterns under $m^{th}$ order Markov set up. Makri and Psillakis (2015) also studied $l$-overlapping runs of successes of length $k$ and obtained recurrence relations for probability mass functions for the case of Bernoulli trials ordered in line as well as in circle. For a more detailed account on the theory of runs and its applications, we refer the reader to Balakrishnan and Koutras (2002) and Makri and Psillakis (2015).

In this paper, we assume that the underlying trials form a $m^{th}$ order homogeneous Markov chain. We study the waiting time of the $n^{th}$ occurrence of the $r$-look-back run of length $k$. We show that under the assumption of the model exhibiting a strong propensity towards success, *i.e.*, the probability of getting success converges to 1 in a certain sense, the waiting time for the $n^{th}$ occurrence of $r$-look-back run of length $k$ converges to a compound Poisson distribution as $n \to \infty$. This result generalises the results of Inoue and Aki (2003) where they considered that the underlying trials are from a homogeneous Markov chain ($m = 1$). Also, we show that the number of $r$-look-back runs of length $k$ till the $n^{th}$ trial, suitably normalised, converges to a normal distribution when the underlying process is an $m^{th}$ order homogeneous Markov chain. Further, we obtain the (uniform) rate of convergence of the central limit theorem (Berry Essen type result). This result shows that the convergence rate is at least a fraction power of $n$ with a logarithmic correction factor.

In the next section we give the formal definitions and the statement of results. The third section is devoted to showing the convergence of waiting times, while in the fourth section we obtain the rate of convergence results. In the final section, we demonstrate simulation results where the underlying trials are from a homogeneous Markov chain ($m = 1$) exhibiting the rate of convergence.

## 2. Definitions and statement of theorem

Let $X_{-m+1}, X_{-m+2}, \dots, X_0, X_1, \dots$ be a sequence of stationary $m^{th}$ order $\{\mathbf{0}, \mathbf{1}\}$-valued Markov chain. It is assumed that the values of $X_{-m+1}, X_{-m+2}, \dots, X_0$ are known, *i.e.*, we are given the initial condition $\{X_0 = x_0, X_{-1} = x_1, \dots, X_{-m+1} = x_{m-1}\}$, $x_i \in \{\mathbf{0}, \mathbf{1}\}, i = -m+1, \dots, -1, 0$.

Define $N_l := \{0, 1, \dots, 2^l - 1\}$ for any $l \geq 0$. It is clear that $\{\mathbf{0}, \mathbf{1}\}^l$ and $N_l$ can be identified by the mapping $(x_0, x_1, \dots, x_{l-1}) \longrightarrow \sum_{j=0}^{l-1} 2^j x_j$. Thus, we will represent the initial condition by taking $x \in N_m = \{0, 1, \dots, 2^m - 1\}$ where $x = \sum_{j=0}^{m-1} 2^j x_j$.

We define, for any $n \geq 0$,

$$p_x := \mathbb{P}(X_{n+1} = 1 | X_n = x_0, X_{n-1} = x_1, \dots, X_{n-m+1} = x_{m-1}). \tag{1}$$

Consequently, $q_x := \mathbb{P}(X_{n+1} = 0 | X_n = x_0, X_{n-1} = x_1, \dots, X_{n-m+1} = x_{m-1}) = 1 - p_x$. We assume that $0 < p_x < 1$ for all $x \in N_m$. Define two functions $f_l, g_l : N_l \to N_l$ as

$$f_l(x) := 2x + 1 \pmod{2^l} \text{ and } g_l(x) := 2x \pmod{2^l}.$$

Note that, $f_m(x), g_m(x)$ can be interpreted as the two possible states which can be reached from the state $x$ in a single step, provided we obtain a success, failure respectively in the next trial.

Let $R_i(k, r)$ be the indicator of the event that a $r$-look-back run of length $k$ is completed at the $i^{th}$ trial. In order that $R_i(k, r) = 1$, we must have $R_{i-1}(k, r) = R_{i-2}(k, r) = \dots = R_{i-r}(k, r) = 0$. Thus, the formal definition of $R_i(k, r)$ can be given inductively as follows:

**Definition 1:** Set $R_i(k, r) = 0$ for $i \leq k - 1$ and for any $i \geq k$, define

$$R_i(k, r) := \prod_{j=1}^{r} (1 - R_{i-j}(k, r)) \prod_{j=i-k+1}^{i} X_j. \tag{2}$$

When $r = 0$, the first product should be interpreted as 1. If $R_i(k, r) = 1$, we say that a *r-look-back run of length k* has been recorded at the $i^{th}$ trial (*i.e.*, ending at trial $i$). Define

$$N_{n,k,r} := \sum_{i=1}^{n} R_i(k, r) = \sum_{i=k}^{n} R_i(k, r)$$

as the number of $r$-look-back runs of length $k$ till the $n^{th}$ trial. A sequence of stopping times is defined as follows: set $\tau_0(k, r) = 0$ and for $n \geq 1$,

$$\tau_n(k, r) := \inf\{i > \tau_{n-1}(k, r) : N_{i,k,r} = n\}. \tag{3}$$

Note, $\tau_n(k,r)$ is the waiting time for the $n^{th}$ occurrence of $r$-look-back run of length $k$.

In the sequel, we say that a random variable $\xi$ is of Poisson type with multiplicity $p$ and parameter $\alpha$, denoted by $\xi \sim \text{Poi}(p,\alpha)$, if

$$\mathbb{P}(\xi = pt) = \frac{\exp(-\alpha)\alpha^t}{t!} \qquad \text{for } t = 0, 1, \dots . \tag{4}$$

Note that, when $p = 1$ it is the usual Poisson distribution. Following Aki (1985), we say that a random variable $\xi$ follows an extended Poisson distribution of order $k$ with parameters $(\alpha_1, \alpha_2, \dots, \alpha_k)$, if its p. g. f. is given by,

$$\phi(z; \alpha_1, \alpha_2, \dots, \alpha_k) = \exp\Big(-\sum_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} \alpha_j z^j\Big).$$

It should be noted that if $\xi$ follows an extended Poisson distribution of order $k$ with parameters $(\alpha_1, \alpha_2, \dots, \alpha_k)$, then it can be represented as

$$\xi \overset{d}{=} \sum_{j=1}^{k} \xi_j$$

where $\{\xi_j : 1 \leq j \leq k\}$ are independent and $\xi_j \sim \text{Poi}(j, \alpha_j)$.

The assumption, we impose on our model, is that the system has a strong tendency towards success. We formalize this by stating that, for $x \in N_m$, $p_x$ (as a function of $n$) converges to 1 in such a way that

$$n(1 - p_x) \to \lambda_x \text{ as } n \to \infty \text{ where } \lambda_x > 0 \text{ is a positive constant.} \tag{5}$$

Our first theorem is:

**Theorem 1:** For any initial condition $x \in N_m$, if the condition (5) holds, we have

$$\begin{aligned} (i)\ \ \tau_n(k,r) - (k-r-1) - n(r+1) \ &\Rightarrow\ \sum_{i=0}^{r} \xi_i^{(1)} \qquad \text{when } k \geq r+1 \\ (ii)\ \ \tau_n(k,r) - k - (n-1)(r+1) \ &\Rightarrow\ \sum_{i=0}^{k-1} \xi_i^{(2)} \qquad \text{when } k < r+1 \end{aligned}$$

where $\left\{\xi_i^{(1)} : i = 0, 1, \dots, r\right\}$ are independent random variables with $\xi_i^{(1)} \sim \text{Poi}(k+i-r, \lambda_{2^m-1})$ for $i = 0, 1, \dots, r$ while $\left\{\xi_i^{(2)} : i = 0, 1, \dots, k-1\right\}$ are independent random variables with $\xi_i^{(2)} \sim \text{Poi}(i+1, \lambda_{2^m-1})$ for $i = 0, 1, \dots, k-1$.

From the above, when $0 \leq r \leq k-1$, the limiting distribution, $\sum_{i=0}^{r} \xi_i^{(1)}$, follows an extended Poisson distribution of order $k$ with parameters $(\overbrace{0, 0, \dots, 0}^{k-r-1}, \overbrace{\lambda_{2^m-1}, \lambda_{2^m-1}, \dots, \lambda_{2^m-1}}^{r+1})$ and when $r \geq k$, $\sum_{i=0}^{k-1} \xi_i^{(2)}$ follows an extended Poisson distribution of order $k$ with parameters $(\overbrace{\lambda_{2^m-1}, \lambda_{2^m-1}, \dots, \lambda_{2^m-1}}^{k})$. Inoue and Aki (2003) have obtained a similar result for

$l$-overlapping counting under the Markov chain ($m = 1$) set-up. It should be noted that, Inoue and Aki (2003) have counted the run of length $k$ as a $l$-overlapping run (for $l < 0$) even when the remaining number of trials after the run is completed, is less than $|l|$. Therefore, it matches with our counting scheme with $r = k - 1 + (-l)$ and hence their results can be deduced as a special case from our result. However, even if we follow the definition of Han and Aki (2000), a similar result can be established following our method.

Further, we establish a central limit theorem for $N_{n,k,r}$ and study the rate of convergence in the central limit theorem under the $m^{th}$ order Markov chain set-up. Let $\sigma_n^2 = \text{Var}(N_{n,k,r})$. We show that

**Theorem 2:** For any $r \geq 0$ and $k \geq 1$, we have

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\Big(N_{n,k,r} - \mathbb{E}(N_{n,k,r}) \leq t\sigma_n\Big) - \Phi(t)\right| = O(n^{-2/11}\log n)$$

where $O(f(n))$ is a function $g(n)$ such that $|g(n)/f(n)|$ remains bounded as $n \to \infty$ and $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Since $n^{-2/11}\log n \to 0$ as $n \to \infty$, we obtain the standard central limit theorem from Theorem 2. Further, this result gives the uniform rate at which the normalised variable $N_{n,k,r}$ converges to normality. Since the number of $l$-overlapping runs of length $k$ up to the $n^{th}$ trial is at most one less than $N_{n,k,r}$, where $r = k - l - 1$, exactly same results will hold for the number of $l$-overlapping runs of length $k$ up to the $n^{th}$ trial.

## 3. Convergence of waiting time

In this section, we prove Theorem 1. We require the following lemmas on weak convergence of discrete random variables. The first lemma is an easy consequence of the Portmanteau Theorem (Billingsley (1968) p.p. 11) and the fact that all the random variables involved are discrete in nature; hence we omit its proof.

**Lemma 1:** If $\{\xi_n : n \geq 1\}$ and $\xi$ are random variables taking values in $\mathbb{N} = \{0, 1, \dots\}$ such that for each $t \geq 0$

$$\liminf_{n\to\infty} \mathbb{P}(\xi_n = t) \geq \mathbb{P}(\xi = t),$$

then $\xi_n \Rightarrow \xi$.

**Lemma 2:** Suppose that $\{\xi_n : n \geq 0\}$ and $\{\xi_i^{(1)} : 1 \leq i \leq p_1\}$ and $\{\xi_i^{(2)} : 1 \leq i \leq p_2\}$ are random variables taking values in $\mathbb{N}$ and $\{\xi_1^{(1)}, \xi_2^{(1)}, \dots, \xi_{p_1}^{(1)}, \xi_1^{(2)}, \xi_2^{(2)}, \dots, \xi_{p_2}^{(2)}\}$ are independent. Suppose that, for each $n \geq 1$ and $t \geq 0$, $\Big\{A_n^t(u_1, u_2, \dots, u_{p_1}, v_1, v_2, \dots, v_{p_2}) :$ $u_i \geq 0$ for $1 \leq i \leq p_1, v_i \geq 0$ for $1 \leq i \leq p_2$ and $\sum_{i=1}^{p_1} u_i = t\Big\}$ is a collection of disjoint events, such that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\xi_n = t, A_n^t(u_1, u_2, \dots, u_{p_1}, v_1, v_2, \dots, v_{p_2})\Big) \geq \prod_{i=1}^{p_1} \mathbb{P}(\xi_i^{(1)} = u_i) \prod_{i=1}^{p_2} \mathbb{P}(\xi_i^{(2)} = v_i).$$

Then

$$\xi_n \Rightarrow \sum_{i=1}^{p_1} \xi_i^{(1)}.$$

Note that, we require $p_1 \geq 1$ but $p_2 \geq 0$. In one of our applications, we will take $p_2 = 0$.

**Proof:** Clearly, for any fixed $t \in \mathbb{N}$,

$$\mathbb{P}\Big(\sum_{i=1}^{p_1} \xi_i^{(1)} = t\Big) = \sum_{\substack{u_1,\ldots,u_{p_1} \in \mathbb{N} \\ v_1,\ldots,v_{p_2} \in \mathbb{N} \\ \sum_{i=1}^{p_1} u_i = t}} \prod_{i=1}^{p_1} \mathbb{P}(\xi_i^{(1)} = u_i) \prod_{i=1}^{p_2} \mathbb{P}(\xi_i^{(2)} = v_i).$$

Fix any $\epsilon > 0$ and choose $J$ so large that

$$\sum_{\substack{u_1,\ldots,u_{p_1} \in \mathbb{N} \\ v_1,\ldots,v_{p_2} \in \mathbb{N} \\ \sum_{i=1}^{p_1} u_i = t}} \prod_{i=1}^{p_1} \mathbb{P}(\xi_i^{(1)} = u_i) \prod_{i=1}^{p_2} \mathbb{P}(\xi_i^{(2)} = v_i) - \sum_{\substack{0 \leq u_1,\ldots,u_{p_1} \leq J \\ 0 \leq v_1,\ldots,v_{p_2} \leq J \\ \sum_{i=1}^{p_1} u_i = t}} \prod_{i=1}^{p_1} \mathbb{P}(\xi_i^{(1)} = u_i) \prod_{i=1}^{p_2} \mathbb{P}(\xi_i^{(2)} = v_i) < \epsilon.$$

Thus, we have,

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}(\xi_n = t) \\
&\quad \geq \liminf_{n\to\infty} \sum_{\substack{u_1,\ldots,u_{p_1} \in \mathbb{N} \\ v_1,\ldots,v_{p_2} \in \mathbb{N} \\ \sum_{i=1}^{p_1} u_i = t}} \mathbb{P}(\xi_n = t, A_n^t(u_1, u_2, \ldots, u_{p_1}, v_1, v_2, \ldots, v_{p_2})) \\
&\quad \geq \liminf_{n\to\infty} \sum_{\substack{0 \leq u_1,\ldots,u_{p_1} \leq J \\ 0 \leq v_1,\ldots,v_{p_2} \leq J \\ \sum_{i=1}^{p_1} u_i = t}} \mathbb{P}(\xi_n = t, A_n^t(u_1, u_2, \ldots, u_{p_1}, v_1, v_2, \ldots, v_{p_2})) \\
&\quad \geq \sum_{\substack{0 \leq u_1,\ldots,u_{p_1} \leq J \\ 0 \leq v_1,\ldots,v_{p_2} \leq J \\ \sum_{i=1}^{p_1} u_i = t}} \liminf_{n\to\infty} \mathbb{P}(\xi_n = t, A_n^t(u_1, u_2, \ldots, u_{p_1}, v_1, v_2, \ldots, v_{p_2})) \\
&\quad \geq \sum_{\substack{0 \leq u_1,\ldots,u_{p_1} \leq J \\ 0 \leq v_1,\ldots,v_{p_2} \leq J \\ \sum_{i=1}^{p_1} u_i = t}} \prod_{i=1}^{p_1} \mathbb{P}(\xi_i^{(1)} = u_i) \prod_{i=1}^{p_2} \mathbb{P}(\xi_i^{(2)} = v_i) \\
&\quad \geq \mathbb{P}\Big(\sum_{i=1}^{p_1} \xi_i^{(1)} = t\Big) - \epsilon.
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, by Lemma 1, the result follows. □

In the next lemma, we derive a lower bound of a particular event, defined below. This lower bound will be used in proving Theorem 1.

**Definition 2:** For any $K \geq m$, define $A_\gamma^\alpha(K)$ as the collection of all strings consisting of **0**'s and **1**'s, of length $\gamma$ and having exactly $\alpha$ **0**'s, such that

(a) the number of **1**'s before the first occurrence of **0** is at least $K$,

(b) the number of **1**'s after the last occurrence of **0** is at least $K$,

(c) the number of $\mathbf{1}$'s between any two occurrences of $\mathbf{0}$'s is at least $K$.

For the initial condition $x \in N_m$, the probability of observing any given string $s \in A^{\alpha}_{\gamma}(K)$ is given by:

$$\delta_m(\gamma, \alpha, x) := h_m(x)\Big(h_m(2^m - 2)\Big)^{\alpha}\Big(1 - p_{2^m-1}\Big)^{\alpha}\Big(p_{2^m-1}\Big)^{\gamma-m-\alpha(m+1)} \tag{6}$$

where $h_m(x) := \prod_{j=0}^{m-1} p_{f^j_m(x)}$ where $f^0_m(x) := x$ and $f^{j+1}_m(x) := f_m(f^j_m(x))$ for $x \in N_m$.

For a string $s \in A^{\alpha}_{\gamma}(K)$, let $\beta_0$ be the number of $\mathbf{1}$'s before the first occurrence of $\mathbf{0}$, $\beta_\alpha$ be the number of $\mathbf{1}$'s after the last occurrence of $\mathbf{0}$ and $\beta_i$ be the number of $\mathbf{1}$'s between the $i^{th}$ and $(i+1)^{th}$ occurrences of $\mathbf{0}$ for $i = 1, 2, \ldots, \alpha - 1$.

- Case $r \le k - 1$: Define

$$\beta'_i := (\beta_i - (k - r - 1)) \mod (r + 1) \text{ for } i = 0, 1, \ldots, \alpha.$$

Clearly $0 \le \beta'_i \le r$. Set $S^{(1)}_j(s) := \#\Big\{i : \beta'_i = j\Big\}$ for $j = 0, 1, \ldots, r$, then $\sum_{j=0}^{r} S^{(1)}_j(s) = \alpha$. Define the event

$$B_\gamma(K, u_0, u_1, \ldots, u_r) := \Big\{s \in A^{\alpha}_{\gamma}(K) : S^{(1)}_j(s) = u_j \text{ for } j = 0, 1, \ldots, r\Big\}.$$

So, we must have $\alpha = \sum_{j=0}^{r} u_j$.

- Case $r \ge k$: Define

$$\beta'_0 := \beta_0 \mod (r + 1).$$

Having specified $\beta'_0, \beta'_1, \ldots, \beta'_i$, define

$$\beta'_{i+1} := \begin{cases} \beta_{i+1} \mod (r+1) & \text{if } \beta'_i < k \\ \Big(\beta_{i+1} - (r - \beta'_i)\Big) \mod (r+1) & \text{if } \beta'_i \ge k. \end{cases}$$

Set $S^{(2)}_j(s) := \#\Big\{i : \beta'_i = j\Big\}$ for $j = 0, 1, \ldots, r$. Here also, $\sum_{i=0}^{r} S^{(2)}_j(s) = \alpha$. Define the event

$$C_\gamma(K, u_0, \ldots, u_{k-1}, v_0, \ldots, v_{r-k}) := \Big\{s \in A^{\alpha}_{\gamma}(K) : S^{(2)}_j(s) = u_j \text{ for } j = 0, 1, \ldots, k-1 \\ \text{and } S^{(2)}_j(s) = v_{j-k} \text{ for } j = k, k+1, \ldots, r\Big\}.$$

Here, we must have $\alpha = \sum_{j=0}^{k-1} u_j + \sum_{j=0}^{r-k} v_j$.

The following lemma gives a lower bound of the probability of the events defined above. When $r \le k - 1$, we choose $K = K(r) = (k - r - 1) + t_0(r + 1)$ where $t_0 \ge 1$ is so large that $K \ge m$. When $r \ge k$, we choose $K = K(r) = t_0(r + 1)$ so that $t_0(r + 1) \ge m$.

**Lemma 3:** (a) For $r \leq k-1$ and given non-negative integers $u_0, u_1, \ldots, u_r$ and $n$ such that $n \geq t_0(1+\sum_{i=0}^r u_i)$, define $\gamma(n) = (k-r-1) + n(r+1) + \sum_{i=0}^r u_i(k+i-r)$. Then for any $x \in N_m$, we have

$$\mathbb{P}_x\Big(B_{\gamma(n)}(K, u_0, u_1, \ldots, u_r)\Big) \geq \delta_m\Big(\gamma(n), \sum_{i=0}^r u_i, x\Big) \frac{\Big(n - t_0(1+\sum_{i=0}^r u_i) + \sum_{i=0}^r u_i\Big)!}{\prod_{i=0}^r u_i!\Big(n - t_0(1+\sum_{i=0}^r u_i)\Big)!}. \tag{7}$$

(b) For $r > k-1$ and given non-negative integers $u_0, u_1, \ldots, u_{k-1}, v_0, v_1, \ldots, v_{r-k}$ and $n$ such that $n \geq t_0 + t_0\sum_{i=0}^{k-1} u_i + (1+t_0)\sum_{i=0}^{r-k} v_i + 1$, define $\gamma(n) = k + (n-1)(r+1) + \sum_{i=0}^{k-1} u_i(i+1)$. Then for any $x \in N_m$, we have

$$\begin{aligned}\mathbb{P}_x\Big(C_{\gamma(n)}(K, u_0 \ldots, u_{k-1}, v_0, \ldots, v_{r-k})\Big) &\geq \delta_m\Big(\gamma(n), \sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i, x\Big)\\ &\times \frac{\Big(n-1-(t_0 + t_0\sum_{i=0}^{k-1} u_i + (1+t_0)\sum_{i=0}^{r-k} v_i) + \sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i\Big)!}{\prod_{i=0}^{k-1} u_i! \prod_{i=0}^{r-k} v_i!\Big(n-1-(t_0 + t_0\sum_{i=0}^{k-1} u_i + (t_0+1)\sum_{i=0}^{r-k} v_i)\Big)!}.\end{aligned} \tag{8}$$

**Proof:** As we have noted, the probability of any string is independent of the positions of the **0**'s and is given by (6). So, by multiplying this with the number of strings in $B_{\gamma(n)}(K, u_0, \ldots, u_r)$ and $C_{\gamma(n)}(K, u_0, \ldots, u_{k-1}, v_0, \ldots, v_{k-r})$ respectively, we get the probability of the respective events. Now we describe a method for obtaining a lower bound of the number of strings in the respective events.

(a) We define $r+2$ objects as follows: $O_0 = \mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}, O_1 = \mathbf{10}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}, \ldots, O_r = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{r}\mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}$ and $O_{r+1} = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{r+1}$.

First, we put $K$ **1**'s in the beginning of the string. Next we distribute $u_i$ objects of type $O_i$ for $i = 0, 1, \ldots, r$ and $(n - t_0(1+\sum_{i=0}^r u_i))$ objects of type $O_{r+1}$ in any way we like. It is evident, from the construction of the objects, that any arrangement given above will result in a string in $B_{\gamma(n)}(K, u_0, \ldots, u_r)$. Thus the number of arrangements of the above objects will provide a lower bound of the number of strings in $B_{\gamma(n)}(K, u_0, \ldots, u_r)$. The number of arrangements is given by $(n - t_0(1+\sum_{i=0}^r u_i) + \sum_{i=0}^r u_i)!/\Big(\prod_{i=0}^r u_i!(n - t_0(1+\sum_{i=0}^r u_i))!\Big)$. This proves part (a).

(b) Here we define the objects as follows: $O_0 = \mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}$, $O_1 = \mathbf{10}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}$, $\ldots, O_{k-1} = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{k-1}\mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}, O_k = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{k}\mathbf{0}\underbrace{\mathbf{1}\cdots\mathbf{1}}_{r-k}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}, O_{k+1} = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{k}\mathbf{10}\underbrace{\mathbf{1}\cdots\mathbf{1}}_{r-1-k}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}, \ldots, O_r = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{k}\,\underbrace{\mathbf{1}\cdots\mathbf{1}}_{r-k}\mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{1}}_{K}$ and $O_{r+1} = \underbrace{\mathbf{1}\cdots\mathbf{1}}_{r+1}$.

First, we put $K$ **1**'s in the beginning of the string. Now, we distribute $u_i$ objects of type $O_i$ for $i = 0, 1, \ldots, k-1$ and $v_i$ objects of type $O_{k+i}$ for $i = 0, 1, \ldots, r-k$ and $(n - 1 - (t_0 + t_0\sum_{i=0}^{k-1} u_i + (1+t_0)\sum_{i=0}^{r-k} v_i))$ objects of type $O_{r+1}$. Again, it is evident from the construction of the objects that any arrangement of the objects will result in a string in $C_{\gamma(n)}(K, u_0, \ldots, u_{k-1}, v_0, \ldots, v_{r-k})$. Thus a lower bound of the number of strings in $C_{\gamma(n)}(K, u_0, \ldots, u_{k-1}, v_0, \ldots, v_{r-k})$ is obtained by counting the number of such possible arrangements, which is given by $(n-1-(t_0+t_0\sum_{i=0}^{k-1}u_i+(1+t_0)\sum_{i=0}^{r-k}v_i)+\sum_{i=0}^{k-1}u_i+\sum_{i=0}^{r-k}v_i)!/(\prod_{i=0}^{k-1}u_i!\prod_{i=0}^{r-k}v_i!(n-1-(t_0+t_0\sum_{i=0}^{k-1}u_i+(t_0+1)\sum_{i=0}^{r-k}v_i))!)$. □

Now we are in a position to prove the Theorem 1.

**Proof:** (a) Fix any $t \geq 0$. We consider the collection of events $\Big\{B_{\gamma(n)}(K, u_0,\, u_1, \ldots, u_r) : u_i \geq 0, t = \sum_{i=0}^{r} u_i(k+i-r)\Big\}$. Clearly, these events are disjoint. By Lemma 2, it is enough to show that for any $x \in N_m$,

$$\liminf_{n\to\infty} \mathbb{P}_x\Big(\tau_n(k,r)-(k-r-1)-n(r+1)=t, B_{\gamma(n)}(K,u_0,u_1,\ldots,u_r)\Big) \geq \prod_{i=0}^{r}\frac{\exp(-\lambda_{2^m-1})(\lambda_{2^m-1})^{u_i}}{u_i!}.$$

It is clear that if $\omega \in B_{\gamma(n)}(K, u_0, u_1, \ldots, u_r)$, $\tau_n(k,r) = (k-r-1)+n(r+1)+\sum_{i=0}^{r}u_i(k+i-r) = (k-r-1)+n(r+1)+t$ and $\gamma(n) = (k-r-1)+n(r+1)+t$. Thus, $\Big\{\tau_n(k,r)-(k-r-1)-n(r+1)=t, B_{\gamma(n)}(K,u_0,u_1,\ldots,u_r)\Big\} = B_{\gamma(n)}(K,u_0,u_1,\ldots,u_r)$. So, by part (a) of Lemma 3, we have

$$\begin{aligned}
&\mathbb{P}_x\Big(\tau_n(k,r)-(k-r-1)-n(r+1)=t, B_{\gamma(n)}(K,u_0,u_1,\ldots,u_r)\Big)\\
&\geq \delta_m\Big(\gamma(n),\sum_{i=0}^{r}u_i,x\Big)\frac{\Big(n-t_0(1+\sum_{i=0}^{r}u_i)+\sum_{i=0}^{r}u_i\Big)!}{\prod_{i=0}^{r}u_i!\Big(n-t_0(1+\sum_{i=0}^{r}u_i)\Big)!}\\
&= \Big(1-p_{2^m-1}\Big)^{(\sum_{i=0}^{r}u_i)}\Big(p_{2^m-1}\Big)^{n(r+1)}\Big(1+o(1)\Big)\times\frac{n^{(\sum_{i=0}^{r}u_i)}}{\prod_{i=0}^{r}u_i!}\Big(1+o(1)\Big)\\
&\to \prod_{i=0}^{r}\frac{\exp(-\lambda_{2^m-1})(\lambda_{2^m-1})^{u_i}}{u_i!} \qquad \text{as } n\to\infty.
\end{aligned}$$

This establishes part (a).

For part (b), fix any $t \geq 0$. Consider the collection of events $\Big\{C_{\gamma(n)}(K, u_0, u_1, \ldots, u_{k-1}, v_0, v_1, \ldots, v_{r-k}) : u_i, v_i \geq 0, t = \sum_{i=0}^{k-1}u_i(i+1)\Big\}$. Again these events are disjoint. Further, we have $\Big\{\tau_n(k,r)-k-(n-1)(r+1)=t, C_{\gamma(n)}(K,u_0,u_1,\ldots,u_{k-1},v_0,v_1,\ldots,v_{r-k})\Big\} =$

$C_{\gamma(n)}(K, u_0, u_1, \ldots, u_{k-1}, v_0, v_1, \ldots, v_{r-k})$. Therefore, by part (b) of Lemma 3, we have

$$\begin{aligned}
&\mathbb{P}_x\Big(\tau_n(k,r) - k - (n-1)(r+1) = t, C_{\gamma(n)}(K, u_0, \ldots, u_{k-1}, v_0, \ldots, v_{r-k})\Big)\\
&\geq \frac{\Big(n-1-(t_0 + t_0\sum_{i=0}^{k-1} u_i + (1+t_0)\sum_{i=0}^{r-k} v_i) + \sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i\Big)!}{\prod_{i=0}^{k-1} u_i! \prod_{i=0}^{r-k} v_i! \Big(n-1-(t_0 + t_0\sum_{i=0}^{k-1} u_i + (t_0+1)\sum_{i=0}^{r-k} v_i)\Big)!}\\
&\qquad \times \delta_m\Big(\gamma(n), \sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i, x\Big)\\
&= \Big(1 - p_{2^m-1}\Big)^{(\sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i)} \Big(p_{2^m-1}\Big)^{n(r+1)} \Big(1 + o(1)\Big)\\
&\qquad \times \frac{n^{(\sum_{i=0}^{k-1} u_i + \sum_{i=0}^{r-k} v_i)}}{\prod_{i=0}^{k-1} u_i! \prod_{i=0}^{r-k} v_i!} \Big(1 + o(1)\Big)\\
&\to \prod_{i=0}^{k-1} \frac{\exp(-\lambda_{2^m-1})(\lambda_{2^m-1})^{u_i}}{u_i!} \prod_{i=0}^{r-k} \frac{\exp(-\lambda_{2^m-1})(\lambda_{2^m-1})^{v_i}}{v_i!} \qquad \text{as } n \to \infty.
\end{aligned}$$

This, by Lemma 2, completes the proof of the Theorem. □

**Remark 1:** Since the limiting distribution is independent of the initial condition, we can assume any distribution on the initial conditions. Suppose that $\mu$ is the probability distribution on $\{0,1\}^m$. As we have already discussed, $\mu$ can be identified as a probability measure on $N_m$ by the mapping $(x_0, x_1, \ldots, x_{m-1}) \to x = \sum_{j=0}^{m-1} 2^j x_j$ where each $x_i \in \{\mathbf{0}, \mathbf{1}\}$. Let $\mathbb{P}_\mu$ be the probability measure governing the Markov chain with initial distribution $\mu$. From theorem 1, we can easily conclude that, under $\mathbb{P}_\mu$

$$\begin{aligned}
&(a) \qquad \tau_n(k,r) - (k-r-1) - n(r+1) \;\Rightarrow\; \sum_{i=0}^{r} \xi_i^{(1)} \text{ when } r \leq k-1\\
&(b) \qquad \tau_n(k,r) - k - (n-1)(r+1) \;\Rightarrow\; \sum_{i=0}^{k-1} \xi_i^{(2)} \text{ when } r > k-1
\end{aligned}$$

by first conditioning on $x \in N_m$ and then summing over all possible values of $x \in N_m$. The random variables, $\xi_i^{(1)}$ and $\xi_i^{(2)}$ are as defined in the Theorem 1.

## 4. Central limit theorem

In this section, we prove the central limit theorem for $N_{n,k,r}$ and obtain the uniform rate of convergence for the central limit theorem. The result can be generalized for a wider class of processes; however we concentrate only on the $m^{th}$ order Markov chain set-up described in this paper.

We define two new sequences of random variables: the first one, $Y_n$, captures the sequence of **1**'s observed in last $k$ trials (going back from trial $n$) and the second one, $Z_n$, keeps track whether the end point of the last $r$-look-back run of length $k$ is within $r$ trials (going back) from the trial $n$. Both these random variables assume values in finite sets. Further, the random vector $(Y_n, Z_n)$ jointly form a homogeneous Markov chain taking values in a finite set (for sufficiently large $n$). Next, we translate the description of $r$-look-back runs of length $k$, from the original random variables $\{X_n\}$ to the new set of random vectors $\{(Y_n, Z_n)\}$. Further, the newly defined Markov chain will be a irreducible chain; hence will satisfy the properties of $\phi$-mixing sequence. This allows us to apply the central limit theorem and the rate of convergence results for the $\phi$-mixing sequence to this case to yield Theorem 2.

Define $s = \max(k, m)$. Set $X_{-m} = X_{-m-1} = \cdots = X_{-s+1} = 0$ provided $s > m$. Define a sequence of random variables $Y_n$ as follows:

$$Y_n = \sum_{j=0}^{s-1} 2^j X_{n-j}$$

for $n \geq 1$. Since $X_i \in \{0, 1\}$ for all $i$, $Y_n$ assumes values in the set $N_s$. It is clear that $Y_n$ captures the last $s$ observations $\{X_n, X_{n-1}, \ldots, X_{n-s+1}\}$. Indeed, from the binary expansion of $Y_n$, one can easily retrieve the values of $X_n$'s.

Since the sequence of random variables $X_n$ is stationary and form a $m^{th}$ order Markov chain, we have that the random variables $\{Y_i : i \geq 0\}$ form a homogeneous Markov chain with initial distribution $\delta_x$, where $\delta_x$ is the Dirac measure at $x \in N_m$, and transition probabilities given by

$$\mathbb{P}(Y_{n+1} = y_1 | Y_n = y_0) = \begin{cases} p_{\theta_m(y_0)} & \text{if } y_1 = f_s(y_0) \\ 1 - p_{\theta_m(y_0)} & \text{if } y_1 = g_s(y_0) \\ 0 & \text{otherwise} \end{cases}$$

where $\theta_m : N_s \to N_m$ is given by $\theta_m(x) = x \pmod{2^m}$.

Let $N_r' = \{0, 1, 2, \ldots, 2^{r-1}\}$. Now, we define another sequence of random variables $\{Z_n\}$ taking values in the set $N_r'$. Set $Z_n = 0$ for $n < k$. For $n \geq k$, we define

$$Z_n = \begin{cases} 2Z_{n-1} \pmod{2^r} & \text{if } Z_{n-1} > 0 \\ 1\{Y_n \pmod{2^k} = 2^k - 1\} & \text{otherwise} \end{cases}$$

where $1\{Y_n \pmod{2^k} = 2^k - 1\}$ is the indicator variable for the event $\{Y_n \pmod{2^k} = 2^k - 1\}$.

Now, for all $n \geq s$, the joint distribution of $(Y_n, Z_n)$ is Markovian since $Y_n$ is Markovian and independent of $\{Z_i : i \leq n-1\}$ and the value of $Z_n$ depends only on the values of $Z_{n-1}$ and $Y_n$. The transition probabilities are easy to compute: for $y_0, y_1 \in N_s, z_0, z_1 \in N_r'$, we have

$$\begin{aligned} &\mathbb{P}\Big((Y_{n+1}, Z_{n+1}) = (y_1, z_1) \mid (Y_n, Z_n) = (y_0, z_0)\Big) \\ &= \mathbb{P}\Big(Z_{n+1} = z_1 | Y_{n+1} = y_1, (Y_n, Z_n) = (y_0, z_0)\Big)\mathbb{P}\Big(Y_{n+1} = y_1 | (Y_n, Z_n) = (y_0, z_0)\Big) \\ &= \mathbb{P}\Big(Z_{n+1} = z_1 | Y_{n+1} = y_1, Z_n = z_0\Big)\mathbb{P}\Big(Y_{n+1} = y_1 | Y_n = y_0\Big). \end{aligned}$$

Note that $\mathbb{P}\Big(Z_{n+1} = z_1 | Y_{n+1} = y_1, Z_n = z_0\Big)$ is a deterministic function taking values in $\{0, 1\}$. We also assume that $p_x > 0$ for all $x \in N_m$. Therefore we can conclude that the sequence of random variables $\{(Y_n, Z_n) : n \geq s\}$ is a homogeneous Markov chain with transition probabilities specified by the above formula. However, not all states of $N_s \times N_r'$

are feasible for the Markov chain $\{(Y_n, Z_n) : n \geq s\}$. For example, the state $(0, 1)$ can never be reached. Therefore, we need to restrict our attention to a smaller set.

Let $\mathcal{S}$ be the collection of all the feasible states, *i.e.*, all states of $N_s \times N'_r$ which can be reached by this Markov chain. Formally, we define $\mathcal{S} := \{(y, z) : (0, 0) \rightsquigarrow (y, z)\}$ where $(y_1, z_1) \rightsquigarrow (y_2, z_2)$ implies that $(y_2, z_2)$ can be reached from $(y_1, z_1)$ (in the usual Markov chain sense). Since $1 > p_x > 0$ for all $x \in N_m$, it is easy to see that if $(0, 0) \rightsquigarrow (y, z)$, then $(y, z) \rightsquigarrow (0, 0)$. Therefore, if we restrict our attention to the set $\mathcal{S}$, we get an irreducible Markov chain. More formally, define the following stopping time:

$$\tau_S := \inf\{n \geq 1 : (Y_n, Z_n) \in \mathcal{S}\}. \tag{9}$$

Now, the process after the stopping time, *i.e.*, $\{(Y_n, Z_n) : n \geq \tau_S + 1\}$, using the strong Markov property, is a homogeneous, irreducible Markov chain with state space $\mathcal{S}$ with transition probabilities as specified. The initial distribution of this chain is given by the distribution of the random variable $(Y_{\tau_S}, Z_{\tau_S})$ starting from $x \in N_m$ *i.e.*, the measure $\mu^{(x)}$ on $\mathcal{S} \subseteq N_s \times N'_r$ where

$$\mu^{(x)}(\{y, z\}) = \mathbb{P}((Y_{\tau_S} = y, Z_{\tau_S} = z)|Y_0 = x) \text{ for } (y, z) \in \mathcal{S}.$$

Further, observe that $\tau_S \leq s$ almost surely. Indeed, suppose that $(y, z)$ is any possible value of $(Y_s, Z_s)$ which has been obtained through the observations $X_1 = x_1, X_2 = x_2, \ldots, X_s = x_s$. Now, for any $n \geq \tau_S$ with $Y_n = 0$ and $Z_n = 0$, the probability of the event $\{X_{n+1} = x_1, X_{n+2} = x_2, \ldots, X_{n+s} = x_s\}$ is positive. Clearly, in that case, $Y_{n+s} = y$ and $Z_{n+s} = z$. Therefore, $(y, z) \in \mathcal{S}$ which implies that $(Y_s, Z_s) \in \mathcal{S}$.

Note that $R_i(k, r) = 1$ if and only if $Z_i = 1$ for any $i \geq 1$. Therefore, we may define $R_i(k, r) = 1\{Z_i = 1\}$ for $i \geq 1$. Now, we claim that $\{R_i(k, r) : i \geq \tau_S + 1\}$ is a $\phi$-mixing sequence. Since $\{(Y_i, Z_i) : i \geq \tau_S + 1\}$ is an irreducible homogeneous Markov chain with finite state space, it is a $\phi$-mixing sequence with mixing coefficients given by $\phi_n = C\rho^n$ for $n \geq 1$ where $C > 0$ and $0 < \rho < 1$ are constants. Since, $R_i(k, r)$ is function of $Z_i$ only, the same mixing coefficients will satisfy the mixing condition for the sequence $\{R_i(k, r) : i \geq \tau_S + 1\}$.

However, note here that $\{R_i(k, r) : i \geq \tau_S + 1\}$ need not be a stationary sequence of random variables. Babu, Ghosh and Singh (1978) have studied the convergence rates of central limit theorem for non-stationary $\phi$-mixing sequences. We state theorem 1 of Babu, Ghosh and Singh (1978) here for the sake of completeness.

**Theorem 3:** (Babu, Ghosh and Singh) For a $\phi$-mixing sequence $\{X_n\}$, let $S_n = \sum_{i=1}^{n} X_i$, $\sigma_n^2 = \text{Var}(S_n)$ and $F_n(t) = \mathbb{P}(S_n \leq t\sigma_n)$. Suppose that

$$\mathbb{E}(X_n) = 0 \quad \text{for all } n \geq 1, \tag{10}$$

$$\sum_{n=1}^{\infty} \phi_n^{1/2} < \infty, \tag{11}$$

$$\inf_{n \geq 1} n^{-1/2}\sigma_n > 0, \tag{12}$$

and for some $c > 0$ and $M > 1$

$$\mathbb{E}(|X_n|^{2+c}) < M, \quad \text{for all } n \geq 1. \tag{13}$$

Then,

$$\sup_{t\in\mathbb{R}} \mid F_n(t) - \Phi(t) \mid = O(n^{-\gamma(c)} \log n)$$

where $\gamma(c) = 2c^\star/(6+5c^\star)$ and $c^\star = \min(1, c)$.

We will use the above result to prove the Theorem 2. Let us define, for $i \geq \tau_S + 1$,

$$R'_i(k,r) = R_i(k,r) - \mathbb{E}(R_i(k,r))$$

as the centred sequence of random variables. It is clear that the conditions (10), (11) and (13) of above result holds with $c = 1$. Since $c = 1$, we have that $\gamma(c) = 2/11$. To prove (12) we proceed as follows:

Define a sequence of stopping times in the following way: $\zeta_0 := \inf\{n > \tau_S : (Y_n, Z_n) = 0\}$ and for $i \geq 1$, set $\zeta_i := \inf\{n > \zeta_{i-1} : (Y_n, Z_n) = 0\}$. Further define a sequence of random variables, $U_0 := \sum_{j=\tau_S+1}^{\zeta_0} R'_j(k,r)$ and for $i \geq 1$, $U_i := \sum_{j=\zeta_{i-1}+1}^{\zeta_i} R'_j(k,r)$. In the following lemma, we prove independence of the collection of random variables $\{U_i : i \geq 0\}$.

**Lemma 4:** The collection of random variables $\{U_t : t \geq 0\}$ are independent. Further, $\{U_t : t \geq 1\}$ are identically distributed.

We will proceed to prove theorem 2 assuming the result of Lemma 4 and prove this lemma in the end. Define, $N(n) := \inf\{t : \zeta_t > n\}$. Next we need the following result

**Lemma 5:**

$$\frac{\mathrm{Var}(\sum_{j=\tau_S+1}^{n} R'_j(k,r))}{n} \to C_1 \text{ as } n \to \infty$$

where $C_1 > 0$ is a constant.

**Proof:** Let us define $N'_{n,k,r} = \sum_{j=\tau_S+1}^{\zeta_{N(n)}} R'_j(k,r) = U_0 + \sum_{j=1}^{N(n)} U_j$. Thus, we have,

$$\begin{aligned}
\mathbb{E}(N'_{n,k,r}) &= \mathbb{E}(U_0) + \mathbb{E}(\sum_{j=1}^{N(n)} U_j) = \mathbb{E}(U_0) + \mathbb{E}(N(n))\mathbb{E}(U_1) \\
\mathrm{Var}(N'_{n,k,r}) &= \mathrm{Var}(U_0 + \sum_{j=1}^{N(n)} U_j) \\
&= \mathrm{Var}(U_0) + \mathbb{E}(N(n))\mathrm{Var}(U_1) + \mathrm{Var}(N(n))\mathbb{E}(U_1^2).
\end{aligned}$$

The stop times $\{\zeta_t : t \geq 0\}$ represent the visits of the Markov chain to the state $(0,0)$. Thus it is a renewal event. So, $N(n)$ represents the number of renewals till time $n - \tau_S$. Since $\tau_S \leq s$, we have

$$\frac{\mathbb{E}(N(n))}{n} \to C_2 \text{ and } \frac{\mathrm{Var}(N(n))}{n} \to C_3 \text{ as } n \to \infty$$

where $C_2, C_3 > 0$ (see Feller (1968)). Thus, we have that

$$\frac{\mathrm{Var}(N'_{n,k,r})}{n} \to C_4 \text{ for some constant } C_4 > 0. \tag{14}$$

Now, we have,

$$
\begin{aligned}
&\left|\mathrm{Var}(\sum_{j=\tau_S+1}^{n} R'_j(k,r)) - \mathrm{Var}(N'_{n,k,r})\right| \\
&\leq \mathrm{Var}\left(\sum_{j=n+1}^{\zeta_{N(n)}} R'_j(k,r)\right) + 2\left|\mathrm{Cov}\left(\sum_{j=n+1}^{\zeta_{N(n)}} R'_j(k,r), N'_{n,k,r}\right)\right| \\
&\leq 4\mathbb{E}(\zeta_1^2) + 2\left(\mathrm{Var}\left(\sum_{j=n+1}^{\zeta_{N(n)}} R'_j(k,r)\right)\mathrm{Var}\left(N'_{n,k,r}\right)\right)^{1/2} \\
&\leq C_5 n^{1/2}
\end{aligned}
$$

for some constant $C_5 > 0$. This coupled with (14) proves the Lemma. □

Now, we are in a position to prove the Theorem 2.

**Proof:** For $n > s$, we have that

$$
\begin{aligned}
&\frac{N_{n,k,r} - \mathbb{E}(N_{n,k,r})}{\sqrt{\mathrm{Var}(N_{n,k,r})}} \\
&= \frac{\sum_{i=k}^{\tau_S} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} + \frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} - \frac{\sum_{i=n+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} \\
&= \frac{\sum_{i=k}^{\tau_S} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} + \frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}} \times \frac{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}}{\sqrt{\mathrm{Var}(N_{n,k,r})}} - \frac{\sum_{i=n+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} \\
&= E_1 + E_2 + E_3 + \frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}}
\end{aligned}
$$

where

$$
\begin{aligned}
E_1 &= \frac{\sum_{i=k}^{\tau_S} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} \\
E_2 &= -\frac{\sum_{i=n+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(N_{n,k,r})}} \\
E_3 &= \Big(\frac{\sigma'_n}{\sigma_n} - 1\Big)\frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}}
\end{aligned}
$$

with $\sigma_n^2 = \mathrm{Var}(N_{n,k,r})$ and $(\sigma'_n)^2 = \mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))$.

First, using Lemma 5, we show that

$$n^{-1/2}\sigma'_n \to C_6 \text{ as } n \to \infty$$

where $C_6 > 0$ is a constant. Indeed, we have

$$\begin{aligned}
&\left|(\sigma'_n)^2 - \mathrm{Var}(N'_{n,k,r})\right| \\
&\leq \ \mathrm{Var}(\sum_{j=n+1}^{\tau_S+n} R'_i(k,r)) + 2\left|\mathrm{Cov}(\sum_{j=n+1}^{\tau_S+n} R'_i(k,r), N'_{n,k,r})\right| \\
&\leq \ C_7 n^{1/2}
\end{aligned}$$

using Lemma 5 and the fact that $\tau_S \leq s$. This implies the condition (12) of Babu, Ghosh and Singh (1978) is satisfied. Hence, using their result, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r) \leq t\sigma'_n\right) - \Phi(t)\right| = O(n^{-2/11}\log n). \qquad (15)$$

To conclude the result of the Theorem 2, we need to show that for some $K > 0$,

$$\mathbb{P}(|E_1 + E_2 + E_3| > Kn^{-2/11}\log n) = O(n^{-2/11}\log n).$$

Note that, using a similar argument as above, we get

$$\frac{\sigma_n^2}{n} \to C_8 \text{ as } n \to \infty$$

where $C_8 > 0$. Thus, we have, for constants $C_9$ and $C_{10}$, $|E_1| \leq C_9 n^{-1/2}$ and $|E_2| \leq C_{10} n^{-1/2}$. Finally, again using similar arguments, $|\sigma'_n/\sigma_n - 1| \leq C_{11} n^{-1/2}$. Thus, for $n$ sufficiently large, we have,

$$\begin{aligned}
&\mathbb{P}(|E_1 + E_2 + E_3| > Kn^{-2/11}\log n) \\
&\leq \ \mathbb{P}(|E_3| > K' n^{-2/11}\log n) \\
&\leq \ \mathbb{P}\left(\left|\frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}}\right| > K'' n^{1/2-2/11}\log n\right) \\
&\leq \ \Phi(K'' n^{1/2-2/11}\log n) + 2\sup_{t\in\mathbb{R}}\left|\Phi(t) - \mathbb{P}\left(\frac{\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r)}{\sqrt{\mathrm{Var}(\sum_{i=\tau_S+1}^{\tau_S+n} R'_i(k,r))}} \leq t\right)\right| \\
&= \ O(n^{-2/11}\log n)
\end{aligned}$$

using (15) and property of the normal distribution function, where $K', K''$ are positive constants. This proves the theorem. □

Now, we prove the Lemma 4.

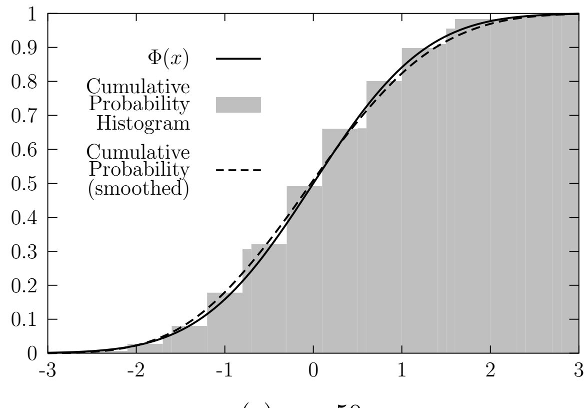
**Proof:** It is clear, from the definition of $U_0$, that $U_0$ is determined by the process $\{(Y_j, Z_j) : \tau_S + 1 \leq j \leq \zeta_0\}$, *i.e.*, $U_0$ is a $\mathcal{F}_{\zeta_0} = \sigma((Y_j, Z_j) : \tau_S + 1 \leq j \leq \zeta_0)$ measurable random variable. Further, for $i \geq 1$, the random variable $U_i$ is determined by the process $\{(Y_j, Z_j) : j > \zeta_{i-1} + 1\}$. Therefore, the sequence of random variables $\{U_i : i \geq 1\}$ are measurable with respect to the sigma algebra generated by $\{(Y_j, Z_j) : j \geq \zeta_0 + 1\} = \mathcal{F}_{\zeta_0+}$. Now, the conditional distribution of the process $\{(Y_j, Z_j) : j \geq \zeta_0 + 1\}$, given the process up to time $\zeta_0(\mathcal{F}_{\zeta_0})$, using the strong Markov property, is same as that of $\{(Y_j, Z_j) : j \geq 0\}$ with the initial condition that $(Y_0, Z_0) = (0, 0)$. Therefore, it is independent of the process up to time $\zeta_0$. Hence, it is independent of the random variables which are measurable with respect to the process $\{(Y_j, Z_j) : j \leq \zeta_0\}$. Thus, $U_0$ is independent of $\{U_i : i \geq 1\}$. Now this argument can be carried out inductively to prove the result. Further, the distribution $\{U_i : i \geq 1\}$ depends only on the initial condition $(Y_{\zeta_0}, Z_{\zeta_0}) = (0, 0)$ and transition matrix of the Markov chain. Since, for any $i \geq 1$, the sequence $\{U_j : j \geq i\}$ will have the same initial condition $((Y_{\zeta_{i-1}}, Z_{\zeta_{i-1}}) = (0, 0))$ and the same transition probabilities, we have that $\{U_i : i \geq 1\}$ are identically distributed. However, the initial condition of the sequence $\{U_i : i \geq 0\}$ is given by the distribution of $(Y_{\tau_S}, Z_{\tau_S})$. There is no reason to expect that, $(Y_{\tau_S}, Z_{\tau_S}) = (0, 0)$. Therefore, $U_0$ may have a different distribution. □

## 5. Simulation results

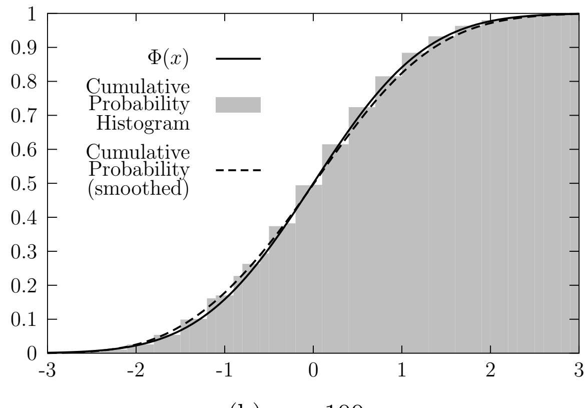
In this section, we provide some simulation results exhibiting the goodness of the approximation in the central limit theorem. These results have been obtained for a Markov chain, with the transition matrix $P$ given by

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}.$$

Simulation has been performed for $n$ number of trials where $n = 50, 100, 500$ and 1000. For $k = 4$ and $r = 2$, the values of $N_{n,4,2}$ have been computed. For each choice of $n$, the experiment is repeated 10000 times and then the mean and the variance of $N_{n,4,2}$ have been obtained. Normalizing these 10000 observations, cumulative probability histograms have been drawn for a grid of 0.1 using the computer package GNU PLOT. The smoothed version of the histogram have been plotted using bezier smoothing algorithm. From the following plots, it is indeed evident that the smoothed version of the cumulative probability histogram is a good approximation of the normal probability distribution function ($\Phi(x)$) even for value of $n$ as small as 50.

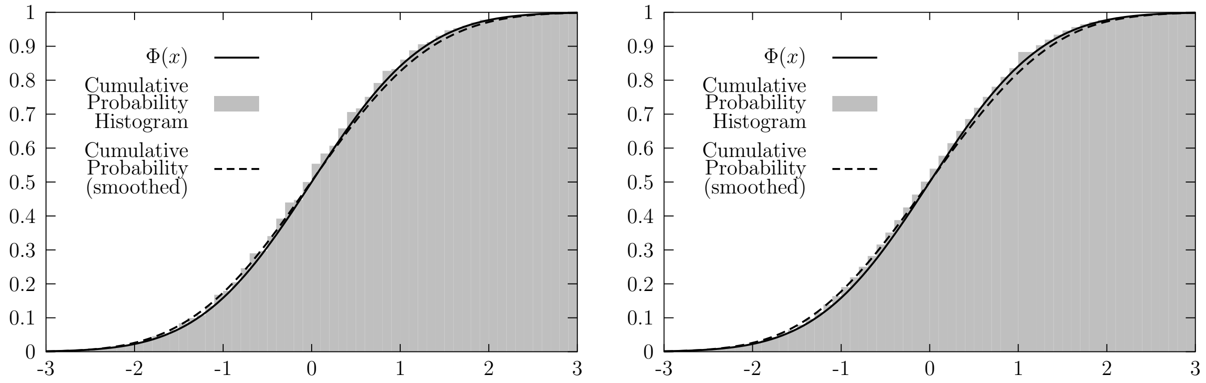


(a) $n = 50$



(b) $n = 100$

Further, we support the findings by illustrating the simulated values of the maximum difference between $\Phi(x)$ and $\mathbb{P}(N_{n,k,r} - \mathbb{E}(N_{n,k,r}) \leq \sigma_n x)$ for $-3.0 \leq x \leq 3.0$ and for various



(c) $n = 500$ (d) $n = 1000$

choices of $(k, r)$ and $n$ in Table 1. Here the underlying sequence of random variables constitute a Markov chain with transition probabilities $p_{01} = 0.4$ and $p_{11} = 0.8$. The table 1 shows a nice decay of the maximum difference as $n$ grows.

**Table 1**

| Sample size $(n)$ | $k = 4$<br>$r = 2$ | $k = 5$<br>$r = 7$ | $k = 5$<br>$r = 0$ | $k = 6$<br>$r = 5$ | $k = 6$<br>$r = 8$ | $k = 7$<br>$r = 3$ | $k = 8$<br>$r = 2$ |
|---|---|---|---|---|---|---|---|
| 50 | 0.081278 | 0.164263 | 0.060424 | 0.132263 | 0.172504 | 0.127723 | 0.148725 |
| 75 | 0.064522 | 0.134955 | 0.056294 | 0.118998 | 0.140559 | 0.107202 | 0.105870 |
| 100 | 0.050815 | 0.117124 | 0.052313 | 0.092598 | 0.123940 | 0.091438 | 0.093857 |
| 125 | 0.045183 | 0.099382 | 0.041538 | 0.074459 | 0.107547 | 0.082336 | 0.085319 |
| 150 | 0.038893 | 0.083109 | 0.039858 | 0.069004 | 0.090610 | 0.060012 | 0.076902 |
| 200 | 0.036554 | 0.075693 | 0.028927 | 0.067591 | 0.088099 | 0.059561 | 0.062446 |
| 300 | 0.028691 | 0.061773 | 0.025912 | 0.047239 | 0.066568 | 0.047434 | 0.044028 |
| 400 | 0.020599 | 0.049510 | 0.019476 | 0.041114 | 0.060933 | 0.035864 | 0.035205 |
| 500 | 0.019384 | 0.047328 | 0.018392 | 0.039172 | 0.054056 | 0.024733 | 0.025782 |
| 1000 | 0.008500 | 0.032516 | 0.009471 | 0.022737 | 0.035285 | 0.022301 | 0.009905 |

## References

Aki, S. (1985). Discrete distributions of order $k$ on a binary sequence. *Annals of the Institute of Statistical Mathematics*, **37**, 205–224.

Aki, S., Balakrishnan, N. and Mohanty, S. G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials. *Annals of the Institute of Statistical Mathematics*, **48**, 773–787.

Aki, S. and Hirano, K. (1988). Some characteristics of the binomial distribution of order $k$ and related distributions. *Statistical Theory and Data Analysis II, Proceedings of the $2^{nd}$ Pacific Area Statistical Conference,* North-Holland, Amsterdam, 211–222.

Aki, S. and Hirano, K. (1995). Joint distributions of numbers of success-runs and failures until the first $k$ consecutive successes. *Annals of the Institute of Statistical Mathematics*, **47**, 225–235.

Aki, S. and Hirano, K. (2000). Number of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order $k$. *Annals of the Institute of Statistical Mathematics*, **52**, 767–777.

Babu, G. J., Ghosh, M. and Singh. K. (1978) On rates of convergence to narmality for $\phi$-mixing processes, *Sankhyã, Series A*, **40**, 278–293.

Balasubramanian, K., Viveros, R. and Balakrishnan, N. (1993). Sooner and later waiting time problems for Markovian Bernoulli trials. *Statistics and Probability Letters*, **18**, 153–161.

Billingsley P. (1968). *Convergence of Probability Measures.* John Wiley, New York.

Feller W. (1968) *An introduction to Probability Theory and its Applications. Vol - I.* John Wiley, New York, 3rd ed.

Fu, J. C. and Koutras, M. V. (1994). Distribution Theory of runs: a Markov Chain approach. *Journal of the American Statistical Association*, **89**, 1050–1058.

Godbole, A. P. (1990). Specific formulæ for some success run distributions. *Statistics and Probability Letters*, **10**, 119–124.

Godbole, A. P. (1992). The exact and asymptotic distribution of overlapping success runs. *Communications in Statistics - Theory and Methods*, **21**, 953–967.

Han, S. and Aki, S. (2000). A unified approach to binomial-type distributions of order $k$. *Communications in Statistics - Theory and Methods*, **29**, 1929–1943.

Hirano, K. (1986). Some properties of the distributions of order $k$. in: A. N. Philippou and A. F. Horadam eds., *Fibonacci numbers and their applications*, (Reidel Dordrecht), 43–53.

Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain. *Statistica Sinica*, **3**, 313–320.

Hirano, K., Aki, S., Kashiwagi, N. and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order $k$. *Statistics and Probability Letters*, **11**, 503–509.

Inoue, K. and Aki, S. (2003). Generalized binomial and negative binomial distributions of order $k$ by the $l$-overlapping enumeration scheme. *Annals of the Institute of Statistical Mathematics*, **55**, 153–167.

Koutras, M. V. (1996). On a waiting time distribution in a sequence of Bernoulli Trials. *Annals of the Institute of Statistical Mathematics*, **48**, 789–806.

Ling, K. D. (1988). On binomial distributions of order $k$. *Statistics and Probability Letters*, **6**, 247–250.

Makri, F. S. and Psillakis, Z. M. (2015). On $l$-overlapping Runs of Ones of Length $k$ in Sequences of Independent Binary Random Variables. *Communications in Statistics - Theory and Methods*, **44**, 3865–3884.

Philippou, A. N. and Makri, F. S. (1986). Successes, runs and longest runs. *Statistics and Probability Letters*, **4**, 101–105.

Uchida, M. (1998). On number of occurrences of success runs of specified length in a higher-order two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **50**, 587–601.

Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **47**, 415–433.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 207–215

# Randomized Response in Combination with Direct Response for Estimation of Incidence Parameters of Two Sensitive Qualitative Features

**Opendra Salam**
*Department of Statistics*
*Manipur University, India*

Received: 10 July 2021; Revised: 24 February 2022; Accepted: 09 July 2022

## Abstract

We consider a situation wherein we are dealing with two sensitive qualitative features (SQlFs) say $Q_1^*$ and $Q_2^*$ with respective incidence proportions/parameters $P_1^*$ and $P_2^*$ [both unknown]. In a given sample of $n$ respondents, some respondents will be comfortable with $Q_1^*$; some will be comfortable with $Q_2^*$; some respondents will be comfortable with both the features while some others will not be comfortable with either of the two. Here 'comfortable' refers to the situation wherein the respondent is agreeable to provide 'direct (yet, truthful) response' to the sensitive feature under consideration. Therefore, there are 4 obvious categories of respondents in a random sample of any reasonable size. The same categorization holds in the entire population of respondents in an analogous manner. Our objective is to provide unbiased estimates for the incidence parameters of the two categories, based on data [of responses from all the four types of respondents] accrued from a survey.

*Key words:* Qualitative features; Sensitive features; Direct response; Randomized response; Population proportion; Binary response; Designing the survey; Combination of estimates.

## 1. Introduction

In sample survey, the study variable may be sensitive in nature; as for example, it may be related to "addiction" to a drug, being a "habitual gambler", having a history of "abortion", "extramarital affairs" and the like. For such items of information, generally, the respondents may be reluctant to provide "Direct" and yet "Truthful" responses. It is also possible that a fraction of the respondents are quite 'comfortable' with such questions and are agreeable to respond truthfully without any social embarrassment. For two such sensitive features, naturally, there are 4 types of respondent-categories, as explained in the abstract.

As is well-known, Warner (1965) introduced "Randomized Response Techniques [RRTs]" to address such questions of eliciting information on sensitive qualitative features. In 2015, there was world-wide celebration of "Fifty Years of RRT" and Handbook of Statistics, Volume 34 was published. Research continues in this fascinating area of survey sampling - theory and applications - dealing with sensitive issues. We refer to the books / book chapters on RRT by Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), Hedayat and Sinha (1991),

Corresponding Author: Opendra Salam
Email: sopendra@manipuruniv.ac.in

Chaudhuri (2011), Chaudhuri and Christofides (2013), and Mukherjee *et al.* (2018). For the work in this paper, we refer to (i) Nandy, Marcovitz and Sinha (2015), (ii) Sinha (2017), (iii) Nandy and Sinha (2020) and (iv) Salam and Sinha (2020). This last reference will be mainly used in the present study.

We contemplate a situation wherein we are dealing with two sensitive qualitative features [SQlFs] $Q_1^*$ and $Q_2^*$, with corresponding incidence proportions $P_1^*$ and $P_2^*$ respectively in the population as a whole. However, a fraction of the sampled respondents are known to be 'comfortable' with 'Direct Response' to one or the other or both of the SQlFs Our aim is to unbiasedly estimate both these proportions $P_1^*$ and $P_2^*$, based on the complete survey data.

Naturally, we have 4 different categories of responding units in the population of size $N$, as also in a randomly drawn sample of reasonably adequate size $n$. Without any loss of generality, we may start with the following table of classification of the population of $N$ respondents.

**Table 1: 2-Way classification of respondents**

| Type | Comfortable with $Q_2^*$ | Uncomfortable with $Q_2^*$ | Total size |
|---|---|---|---|
| Comfortable with $Q_1^*$ | $N(C,C)$ | $N(C,NC)$ | $N(C,.)$ |
| Uncomfortable with $Q_1^*$ | $N(NC,C)$ | $N(NC,NC)$ | $N(NC,.)$ |
| Total | $N(.,C)$ | $N(.,NC)$ | $N(.,.)$ |

In case of sample respondents, we use the obvious notations

$$n(C,C), n(C,NC), n(NC,C), n(NC,NC)$$

for the sample frequencies in the respective different categories. Under simple random sampling without replacement, it may be assumed that the sample counts in different categories are proportional to the respective population counts. Henceforth, we will assume simple random sampling of the respondents in each category.

It transpires that for respondents in the Category (C, C), we may freely address the two questions $Q_1^*$ and $Q_2^*$, independent of one another, and extract truthful responses from each of the sampled respondents. Again, for the Category (C, NC) [respectively, (NC, C)], only the question $Q_1^*$ [resp., $Q_2^*$] can be put forward directly and truthful responses may be extracted from the relevant respondents. The other question has to be handled by taking recourse to an RRT. For the Category (NC, NC), we must adopt some kind of RRT for simultaneous estimation of the underlying parameters. For this latest kind of subpopulation of respondent categories, we may take recourse to the procedures studied in recent years. Vide Salam and Sinha (2020), for example.

**Remark 1:** As a general principle, for unbiased estimation of a population proportion $\pi$ [of "Yes" responses] based on the available responses on a binary [Yes/No] response feature, it is well-known that the sample proportion $p$ [of 'Yes' responses] is an unbiased estimate for the corresponding population proportion $\pi$. Moreover, in order to combine information on the common proportion $\pi$ arising out of different/disjoint independent samples, we collect head-counts of all the ‘Yes’ responses from different sources together and divide it by the total count of respondents. Recall the formula $\hat{p} = \frac{\sum_i x_i}{\sum_i n_i}$.

**Remark 2:** From the nature of the sampled respondents in three of the four categories, it transpires that RRT provides estimate(s) of the desired proportion(s) involving the sensitive feature(s). From there, we work out estimate(s) of the number of respondents in the relevant ‘yes’ category of the sensitive feature. Recall $\hat{p}_i = x_i/n_i$ from the $i^{th}$ source so that $x_i \; = n_i\hat{p}_i$. Then we go by the technique of combination of evidences from different sources, as explained in Remark [1].

The rest of the paper is organized as follows. In Section 2, we briefly outline the general approach for tackling the problem stated above. Then, in Section 3, we consider an illustrative example to work out all the essential details. Finally, in Section 4, we provide some concluding remarks.

## 2. General approach for analysis of data

We refer to the general description of the 4 categories of respondents as laid down in Section 1.

Set $NP_i^* = N_i^*, i = 1, 2$ and, further, consider the natural and obvious decomposition of $N_i^*$ as

$$N_i^* = N_i^*(C, C) + N_i^*(C, NC) + N_i^*(NC, C) + N_i^*(NC, NC), i = 1, 2.$$

We also express the above quantities - quite meaningfully - as

$$N_1^* = N_1^*(C, .) + N_1^*(NC, C) + N_1^*(NC, NC); N_2^* = N_2^*(., C) + N_2^*(C, NC) + N_2^*(NC, NC).$$

Note that both the quantities $N_1^*(C, .)$ and $N_2^*(., C)$ are amenable to unbiased estimation by direct questionnaire method. For $N_1^*(C, .) = N_1^*(C, C) + N_1^*(C, NC)$ population units, in view of simple random sampling, we know (i) the number of “Yes” respondents among $n_1^*(C, C)$ sampled respondents wrt the Category $Q_1^*$ and also (ii) the number of “Yes” respondents among $n_1^*(C, NC)$ sampled respondents wrt the Category $Q_1^*$. Likewise, we have direct “Yes” responses for $Q_2^*$ from $n_2^*(C, C)$ respondents randomly sampled from $N_2^*(C, C)$ respondents in the reference subpopulation and also, we have direct “Yes” responses for $Q_2^*$ from $n_2^*(NC, C)$ respondents randomly sampled from $N_2^*(NC, C)$ respondents in the reference subpopulation.

These four separate count estimates of “Yes” categories are the ingredients for arriving at final estimates of $P_1^*$ and $P_2^*$.

For unbiased estimation of $N_1^*(NC, C)$ or of $N_2^*(C, NC)$, we can follow the technique as in Section 2 [Subsections 2.1, 2.2 and 2.3] of Salam and Sinha (2020) - appropriately adjusted for our purpose- for unbiased estimation of the corresponding proportions *i.e.*, $P_1^*(NC, C) = \frac{N_1^*(NC,C)}{N_1(NC,C)}$ and $P_2^*(C, NC) = \frac{N_2^*(C,NC)}{N_2(C,NC)}$. Finally, for unbiased estimation of $N_1^*(NC, NC)$ and $N_2^*(NC, NC)$, we refer to Subsection 2.4 of Salam-Sinha (2020) paper which provides formulae for simultaneous unbiased estimation of the underlying population proportions.

## 3. Worked out example

We consider a large population of $N = 30,000$ respondents - broadly classified in the 4 categories as

$$N(C, C) = 2000, N(C, NC) = 3000, N(NC, C) = 2000, N(NC, NC) = 23000.$$

This information is a priori available to the investigating agency. In a random sample of $n = 3000$ respondents, the stratified sample sizes under proportional allocation are taken as $[200, 300, 200, 2300]$.

(i) Data Type : (C, C) Simple and direct implementation of the questionnaire yields: For $Q_1^*$, freq. count of “Yes” = 85; for $Q_2^*$, it is 56.

(ii) Data Type :(C, NC) (a) Direct implementation of $Q_1^*$ yields : Freq. Count of “Yes” = 114.
(b) Implementation of technique adopted in Subsection 2.2 of Salam-Sinha (2020) paper yields: $\hat{P}_2^*(C, NC) = 0.35$.

(iii) Data Type :(NC, C) (a) Implementation of technique adopted in Subsection 2.2 of Salam-Sinha (2020) paper (Annexure 1) yields: $\hat{P}_1^*(NC, C) = 0.38$. (b) Direct implementation of $Q_2^*$ yields: Freq. Count of ”Yes” = 94.

(iv) Data Type :(NC, NC) (a) Implementation of technique adopted in Subsection 2.4 of Salam-Sinha (2020) paper yields: $\hat{P}_1^*(NC, NC) = 0.42$.
(b) Implementation of technique adopted in Subsection 2.4 of Salam-Sinha (2020) paper yields: $\hat{P}_2^*(NC, NC) = 0.33$.

**Remark 3:** The reader will find repeated use of a result from Salam-Sinha(20202) paper. One referee has suggested that it would be instructional to explain the methodology from that paper. Not to obscure the essential steps of reasoning, we will proceed through the following steps, using critical close arguments as in Salam-Sinha (2020) paper. For ready reference we reproduce the techniques and computations from the cited paper.

Now we are in a position to provide (unbiased) estimates for $P_1^*$ and $P_2^*$.

We display all the four source information on each of the two sensitive features in the following table.

**Table 2: 2-Way classification of observed / estimated number of “Yes” responses for the two features**

| Type | Comfortable with $Q_2^*$ | Uncomfortable with $Q_2^*$ | Total Figures for $(Q_1^*)$ |
|---|---|---|---|
| Comfortable with $Q_1^*$ | $(85/200, 56/200)$ | $(114/300, 105/300)$ | $(199/500)$ |
| Uncomfortable with $Q_1^*$ | $(76/200, 94/200)$ | $(966/2300, 759/2300)$ | $(1042/2500)$ |
| Total Figures for $(Q_2^*)$ | $(150/400)$ | $(864/2600)$ | $(1241/3000, 1014/3000)$ |

For the respondents who are comfortable with $Q_1^*$ we may use the notation $N_1(C, C)$ and $N_1(C, NC)$ to denote the corresponding frequency counts with respect to $Q_1^*$ . On the other hand for those who are not comfortable with $Q_1^*$ we have to use RRT technique to estimate the proportions and hence the frequency counts in the two categories corresponding to (NC,C) and (NC,NC) have to be estimated indirectly. That is where Salam-Sinha (2020) technique has been used. Likewise we can use an analogous notation for cases involving $Q_2^*$.

We may thus conclude that

$$N_1^*(C, C) = 85; N_1^*(C, NC) = 114;$$

$$\hat{N}_1^*(NC, C) = 76; \hat{N}_1^*(NC, NC) = 966.$$

Therefore, $P_1^* = (85 + 114 + 76 + 966)/3000 = 1241/300 = 41.37$ *per cent*. Likewise, for estimation of $P_2^*$, we proceed similarly and derive an estimate as $P_2^* = 1014/3000 = 33.80$ *per cent*. Based on combined evidence of sample data covering both 'C', 'NC' for $Q_1^*$ and $Q_2^*$, our estimation procedure produces estimates of proportions of $Q_1^*$ and $Q_2^*$ and we end up with $\hat{Q}_1^* = 0.41$ and $\hat{Q}_2^* = 0.34$ respectively.

## 4. Conclusion

For two Sensitive Qualitative Features along with a provision for Optional Randomization for either or both, we have considered a blend of the Randomised Response Technique and Direct Response Technique to estimate the two incidence parameters from a complex pattern of respondents' response profiles. Simple Random Sample of a reasonably large size is assumed to be available. For three or more Sensitive Qualitative Features with this kind of Optional Randomization for one or two or more of such features, it would be an interesting topic by itself to estimate all the incidence proportions. This seems to be a non-trivial generalization of our approach. Again, even for two such sensitive features with provision(s) for optional randomization, multi-category response profiles would be worth studying.

## Acknowledgments

I am extremely grateful to Professor Bikas K Sinha, Retired Professor of ISI, Kolkata, for his continued interest in this work and for going through a preliminary version of this manuscript and offering suggestions to remove obscurities.

## References

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Applications.* Marcel and Dekker.

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* CRC Press, Chapman and Hall, Taylor and Francis Group, FL, USA.

Chaudhuri, A. and Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys.* Springer, Germany.

Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research*, **40**, 169-193.

Fox, J. A. and Tracy, P. E. (1986). *A Method for Sensitive Surveys.* Sage University paper series. Quantitative Applications in the Social Sciences, 07-058. Sage Publication.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling.* New York:Wiley.

Mukherjee, S. and Chattopadhyay, (2018). *Statistical Methods in Social Science Research.* Springer.

Nandy, K. and Sinha, B. K. (2020). Block total response technique for quantitative sensitive items in a finite population. *Statistics and Applications*, **18**, $85 - 95$.

Nandy, K., Marcovitz, M. and Sinha, B. K. (2015). Eliciting information on sensitive features: Block total response technique and related inference. In *Handbook of Statistics*, **34**: Data gathering, analysis, and protection of privacy through randomized response techniques: qualitative and quantitative human traits, 317 - 328.

Opendra, S. and Sinha, B. K. Further thoughts on applications of block total response techniques in case of one or two sensitive binary features. Accepted; 2020 in press (publication scheduled for 2023 in *Thailand Statistician* )

Raghavarao, D. and Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society*, Series B, Methodological, 40-45.

Sinha, B. K. (2017). Some refinements of block total response technique in the context of RRT methodology. *Statistics and Applications*, **15**, 167-171.

Warner, S. L. (1965). Randomized response: a survey technique for eliminating Evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

## APPENDIX

In 1965, Warner introduced a Randomized Response Technique/Methodology (RRT/ RRM) and Raghavarao and Federer (1979) introduced a novel technique/methodology termed Block Total Response Technique/Methodology [BTRT/BTRM] to increase the degree of protection of confidentiality of the respondent's response. The technique is elaborated below.

Consider a collection of $v$ Regular [Non-sensitive] Qualitative Features (RQlFs) [$Q_1, Q_2, \ldots, Q_v$] and one Sensitive Qualitative Feature (SQlF) $Q^*$. Let $b$ be the number of blocks (*i.e.*, sets of questions), each containing $k$ (distinct) RQlFs and the SQlF $Q^*$. Each RQlF is replicated $r$ times in the entire collection of $b$ blocks and there is one block $B_0$ containing all RQlFs. Thus, there are $k+1$ QlFs for each block and $(b+1)$ blocks. The respondents in the sample are split into (b+1) sets of size $n^*, n^*, \ldots, n^*$, $n_0$ such that $n = (bn^* + n_0)$. That is, each of the $b$ blocks received $n^*$ respondents and block $B_0$ received $n_0$ respondents. A block of questions (which contains $k$ NSBFs and the single SBF $Q^*$) is presented to each respondent. The respondent is to provide only the Block Total Response (BTR) in terms of the overall score (*i.e.*, only the total number of yes answers) without divulging any response to any specific Qs – be it NSBF or SBF. It is believed that this BTR Technique [BTRT] will adequately protect the privacy of the respondent, and hence, the correct response to the SBF will emerge. BTRT is an alternative method of RRT to increase respondent's anonymity and enable estimation of the parameter (*i.e.*, proportion of yes respondents in the population) involving sensitive binary feature. We know from BTRT that, in every block [$B_1 to B_b$] we utilize only $k$ of the $v$ NSBFs, while the rest $(v-k)$ NSBFs are left unutilized. When $k$ is small, respondents may feel uncomfortable responding truthfully since responding to $Q^*$ is compulsory in each of the $b$ blocks. Nandy and Sinha (2020) extended the above technique by bringing variations in the block compositions as:

1. List of $k$ 'must respond' NSBFs are kept in Part $A$.

2. Remaining $(v-k)$ NSBFs and $Q^*$ (SBF) are all kept in Part $B$.

A respondent is to choose one question from $(v-k+1)$ questions in part B and mix with the questions in Part A and supply BTR without divulging the identity of the question selected from Part B. Salam-Sinha (2020) introduced a purely random choice from both parts A and B as is explained below.

Suppose there are $k_1$ RQlFs in Part $A$ and $k_2 = v-k_1+1$ RQlFs, including the sensitive question $Q^*$ in Part $B$. Arrangement of $k_1$ RQlFs in Part $A$ is the same as above. Respondent is to blend randomly selected $s_1$ RQlFs from $k_1$ RQlFs and $s_2$ from $k_2 = v-k_1+1$ RQlFs including the sensitive question $Q^*$ and supply the BTR of $(s_1+s_2)$ RQlFs possibly including the sensitive question without divulging any information about the selected questions. Let $\pi_1$ and $\pi_2$ denote the inclusion probabilities of $i^{th}$ unit $[i=1,2,\ldots,k_1]$ RQlFs from Part $A$ and inclusion probability of $i^{th}$ RQlFs $[i=1,2,\ldots,(v-k_1+1)]$ from Part $B$ [including $Q^*$] respectively. Therefore, every question in Parts $A$ and $B$ have an equal chance of inclusion *viz.*, $\pi_1 = s_1/k_1, \pi_2 = s_2/k_2$ respectively.

We have

$$\bar{x}_1 = \sum_{i=1}^{k_1} p_i(s_1/k_1) + P^*(s_2/k_2) + (\boldsymbol{\Delta} - \sum_{i=1}^{k_1} p_i)(s_2/k_2).$$

And, summing over all the $b$ blocks, we derive the single estimating equation for $P^*$ as

$$\sum_{i=1}^{b} \bar{x}_i = r\boldsymbol{\Delta}(s_1/k_1) + bP^*(s_2/k_2) + (b-r)\boldsymbol{\Delta}(s_2/k_2).$$

Therefore, the population proportion $P^*$ of the sensitive qualitative question can be estimated by using the formula

$$\hat{P}^* = \frac{\sum_{i=1}^{b} \bar{x}_i - (s_1/k_1)2\boldsymbol{\Delta} - (b-r)\boldsymbol{\Delta} \times (s_2/k_2)}{b(s_2/k_2)}.$$

Note $\Delta = \sum_{i=1}^{10} p_i$ and $\hat{\Delta} = \bar{x}_0$ which stands for the sample mean of BTRs of the $n_0$ respondents in the block $B_0$.

**Example**

Suppose we have design parameters *i.e.*, $b=5, v=10, k=4, r=2$ and $n=300$ respondents. we randomly split them into 6 sets, taking $n^*=50$ and $n_0=50$. We adopt the same block compositions as are displayed above. We have $k_1=k=4$ and we select $s_1=3$ RQlFs from part A and we also select $s_2=3$ from part B. In order to implement the above scheme for Block 1, for example, we prepare a set of 11 identical cards of the same shape [square, say] and size. At the back of the cards, we write the numbers $1,2,\ldots,10$ and the symbol (*) one card for each. The procedure is : A respondent belonging to Block 1 is to draw three cards at random from the collection of first 4 cards [1 *to* 4]. Note that this is as good as selecting one card at random and discarding the same, and thereby, taking the rest at hand! Out of the remaining 7 cards, the respondent has to select any 3. Thus he/she will have a collection of 6 cards altogether from the two sets. Next, he/she will respond truthfully to all the 6 binary $[1-0]$ features selected and arrive at the Block Total Response and only the BTR score is supposed to be reported – without divulging any details. Note that the respondent may / may not have chosen the SBF ($Q^*$). Of course, he/she must respond truthfully and provide the BTR score – even if this has been selected. Likewise, we prepare 11 cards for use of the respondents in Block 2 and so on. Of course, each time we study the block composition before using the cards to form two designated sets. For the last block $B_0$, we do not need any cards. All NSBFs are compulsory. Having implemented the data-gathering tools, we end up with 'raw' scores of each of the 300 respondents – classified into 6 distinct groups. In each group, we calculate the group average of the scores and these are called 'summary statistics'. Assume that at the end we end up with the following results:

**Table 1: Example : Summary statistics**

| Block | Total score | No. of respondents | Average score |
|---|---|---|---|
| 1 | 128 | 50 | 2.56 |
| 2 | 136 | 50 | 2.72 |
| 3 | 146 | 50 | 2.92 |
| 4 | 125 | 50 | 2.50 |
| 5 | 115 | 50 | 2.30 |
| 6 | 220 | 50 | 4.40 |

Following Nandy and Sinha (2020), we may prepare the following table.

**Table 2: Data analysis : Theory**

| Block | Sample size | Expected block average (EBA) | Average Score |
|---|---|---|---|
| 1 ($B_1$) | $n^*$ | EBA(1) | $\bar{x}_1$ |
| 2 ($B_2$) | $n^*$ | EBA(2) | $\bar{x}_2$ |
| 3 ($B_3$) | $n^*$ | EBA(3) | $\bar{x}_3$ |
| 4 ($B_4$) | $n^*$ | EBA(4) | $\bar{x}_4$ |
| 5 ($B_5$) | $n^*$ | EBA(5) | $\bar{x}_5$ |
| 6 ($B_0$) | $n_0$ | $\boldsymbol{\Delta}$ | $\bar{x}_0$ |

In the above,

$$EBA(1) = [(3/4)(p_1 + p_2 + p_3 + p_4) + (3/7)[P^* + (\boldsymbol{\Delta} - p_1 - p_2 - p_3 - p_4)],$$

$$EBA(2) = [(3/4)(p_5 + p_6 + p_7 + p_8) + (3/7)[P^* + (\boldsymbol{\Delta} - p_5 - p_6 - p_7 - p_8)],$$

$$EBA(3) = [(3/4)(p_9 + p_{10} + p_1 + p_2) + (3/7)[P^* + (\boldsymbol{\Delta} - p_1 - p_2 - p_9 - p_{10})],$$

$$EBA(4) = [(3/4)(p_3 + p_4 + p_5 + p_6) + (3/7)[P^* + (\boldsymbol{\Delta} - p_3 - p_4 - p_5 - p_6)],$$

$$EBA(5) = [(3/4)(p_7 + p_8 + p_9 + p_{10}) + (3/7)[P^* + (\boldsymbol{\Delta} - p_7 - p_8 - p_9 - p_{10})],$$

$$EBA(6) = \mathbf{\Delta} = \sum_{i=1}^{10} p_i.$$

Summing over all the first five block means, we obtain the estimating equation :

$$\sum_{i=1}^{5} \bar{x}_i = (3/4)2\mathbf{\Delta} + (3/7)5P^* + (3/7) \times 3\mathbf{\Delta}$$

$\bar{x}_0$ (Sample mean of BTRs of the $n_0$ respondents in the block $B_0$) $= 4.40$

Replacing $\mathbf{\Delta}$ by its estimate $\bar{x}_0$ and derive.

$$\hat{P}^* = [\sum_{i=1}^{5} \bar{x}_i - (39/14)\bar{x}_0] \times (7/15) = 0.35.$$

An estimate of the population proportion $P^*$ of the sensitive qualitative question is 0.35.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 217–230

# Conway-Maxwell Poisson Distribution: Some New Results and Minimum Variance Unbiased Estimation

**Jahnavi Merupula and V.S.Vaidyanathan**
*Department of Statistics*
*Pondicherry University, Puducherry, India*

Received: 02 April 2022; Revised: 06 July 2022; Accepted: 12 July 2022

## Abstract

Probability distributions for count data have potential applications in medical, epidemiological, and actuarial studies. Conway-Maxwell Poisson (CMP) distribution is a two-parameter Poisson distribution that can handle over- and under-dispersed data. In this paper, some new results on the distributional properties of CMP distribution are presented. Also, minimum variance unbiased (MVU) estimator of the location parameter is derived using the complete sufficient statistic. The primary advantage of the MVU estimator is that it has a closed-form expression, unlike other existing estimators. An approximate expression for the variance of the MVU estimator is obtained, and the performance of the MVU estimator is compared with that of the ML estimator in terms of relative efficiency through simulated and real-life datasets.

*Key words:* CMP distribution; Generalized hypergeometric series; Minimum variance unbiased estimation; Power series family; Relative efficiency; Sufficient statistic.

**AMS Subject Classifications:** 60E05, 62F10

## 1. Introduction

Poisson distribution is often a natural choice among researchers to model count data. However, its applicability is restricted to situations where there is equi-dispersion of data, i.e., the mean is equal to the variance. Often, in reality, count data are over- or under-dispersed. For example, data on the word lengths in a dictionary or the number of infected spots in the leaves of a plant is under-dispersed. Alternative distributions to Poisson are available in the literature to model over- or under-dispersed data. These include mixtures of Poisson, weighted Poisson and generalized Poisson distributions. However, these distributions have more parameters and involve mathematical intricacies which limit their usage. For example, the generalized Poisson distribution does not model under-dispersion effectively due to parameter constraints. Hence, probability models having fewer parameters that can address the problem of over- or under-dispersion are of interest to study both from a theoretical and application perspective.

A two-parameter Poisson distribution capable of handling over- and under-dispersion is Conway-Maxwell Poisson (CMP) distribution introduced by Conway and Maxwell (1961).

Corresponding Author: V. S.Vaidyanathan
Email: vaidya.stats@gmail.com

The probability mass function (pmf) of CMP distribution is

$$P(X=x)=\frac{\lambda^x}{(x!)^\nu}\frac{1}{Z(\lambda,\nu)}, \quad x=0,1,2,\ldots, \quad \lambda>0, \quad \nu\geq 0 \tag{1}$$

where

$$Z(\lambda,\nu)=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}$$

is the normalizing constant. Here, $\lambda$ denotes the location parameter and $\nu$ denotes the dispersion parameter that captures the degree of over- or under-dispersion. The CMP distribution is over-dispersed for $\nu<1$, under-dispersed for $\nu>1$, and equi-dispersed for $\nu=1$. The pmf is not defined for $\nu=0$ and $\lambda\geq 1$.

Shmueli *et al.* (2005) have revisited this distribution to study its properties. A review of CMP distribution, its characterizations and applications can be found in Sellers *et al.* (2012). There has been an increased interest in research about extensions and generalizations of CMP distribution in the recent past. Cordeiro *et al.* (2012) introduced exponential-CMP distribution as a lifetime distribution by compounding an exponential distribution with a CMP distribution and explored its properties. Chakraborty and Imoto (2016) proposed a flexible four-parameter extension of CMP distribution, which encompasses Conway-Maxwell negative binomial and generalized CMP distributions, and also derived its properties. Roy *et al.* (2020) developed Conway-Maxwell negative hypergeometric distribution as a modification to negative hypergeometric distribution along with its characterizations.

Although some extensions and characterizations of CMP distribution are available, properties in terms of differential equation involving recurrent probabilities have not yet been addressed. Such properties are available for popular discrete distributions, see, for example, Boswell and Patil (1973), and the same is discussed for CMP distribution in this paper. Also, a new representation of the CMP distribution in terms of generalized hypergeometric series is given.

From an inferential point of view, existing estimators of the parameters do not have closed-form expressions and have to be computed using iterative methods. Since the CMP distribution belongs to the exponential family of distributions, in the present work, minimum variance unbiased (MVU) estimation of the location parameter is carried out using the distribution of the complete sufficient statistic. Also, an approximate expression for the variance of the estimator is obtained. The merit of using the proposed MVU estimator is highlighted through numerical illustration.

The paper is organized as follows. In Section 2, some properties of CMP distribution are listed, and two new results involving recurrent probabilities and probability generating function (pgf) are presented. The methodology to obtain MVU estimator of the location parameter is explained in Section 3. Numerical illustration to compare the performance of the MVU estimation with likelihood estimation in terms of mean absolute bias (MAB) and relative efficiency (RE) is provided in Section 4 through simulated and real-life data. Concluding remarks are given in Section 5.

## 2. Distributional properties

As mentioned in Section 1, CMP distribution can model both over- and under-dispersed data. CMP distribution encapsulates well-known distributions, including Poisson distribution ($\nu = 1$), geometric distribution ($\nu = 0, \lambda < 1$), and Bernoulli distribution ($\nu \to \infty$). From equation (1), based on $n$ independent and identically distributed (iid) samples on $X$, it can be seen that CMP distribution belongs to the exponential family of distributions. Also, CMP distribution is a member of the two-parameter power series family of distributions with pgf of the form

$$P_X(z) = \frac{Z\left(\lambda z, \nu\right)}{Z(\lambda, \nu)} \tag{2}$$

The pgf of CMP distribution can be expressed in terms of generalized hypergeometric series as (See Nadarajah, 2009)

$$P_X(z) = \frac{{}_0F_{\nu-1}(;1,\ldots,1;\lambda z)}{{}_0F_{\nu-1}(;1,\ldots,1;\lambda)} \tag{3}$$

Comparing equations (2) and (3), we get

$$Z(\lambda, \nu) = {}_0F_{\nu-1}(;1,\ldots,1;\lambda) \tag{4}$$

Using the pgf, the expected value and variance of $X$ can be obtained as

$$E(X) = \lambda \frac{\partial}{\partial \lambda} \log(Z(\lambda, \nu)) \tag{5}$$

and

$$V(X) = \lambda \frac{\partial}{\partial \lambda}\left[\lambda \frac{\partial}{\partial \lambda} \log(Z(\lambda, \nu))\right]$$

For further properties and characterizations of CMP distribution, one may refer to Nadarajah (2009), Daly and Gaunt (2016) and Li *et al.* (2019). In the sequel, two new results on CMP distribution are presented.

### 2.1. Recurrence relationship of probabilities

Boswell and Patil (1973) have shown that any discrete distribution can be characterized in terms of differential equations involving its parameters. For example, Poisson distribution with mean $\lambda$ satisfy the following recurrence relationship, namely,

$$\frac{dp_x}{d\lambda} = p_{x-1} - p_x$$

where $p_x$ is the pmf of the Poisson distribution. A similar recurrence relationship for CMP distribution is obtained below.

**Result 1:** Let $p_x$ denote the pmf of CMP distribution. Then

$$\frac{\partial p_x}{\partial \lambda} = \frac{1}{x^{\nu-1}} p_{x-1} - \frac{E(X)}{\lambda} p_x$$

**Proof:** Partially differentiating the pmf in equation (1) with respect to $\lambda$, we get,

$$\frac{\partial p_x}{\partial \lambda} = \frac{x\lambda^{x-1}Z(\lambda,\nu) - \lambda^x \frac{\partial}{\partial \lambda}Z(\lambda,\nu)}{(x!)^\nu Z^2(\lambda,\nu)}$$

Note that $\frac{\partial}{\partial \lambda}\log(Z(\lambda,\nu)) = \frac{1}{Z(\lambda,\nu)}\frac{\partial}{\partial \lambda}Z(\lambda,\nu)$. Thus

$$\begin{aligned}\frac{\partial p_x}{\partial \lambda} &= \frac{x\lambda^{x-1}}{(x!)^\nu Z(\lambda,\nu)} - \frac{\lambda^x \frac{\partial}{\partial \lambda}\log(Z(\lambda,\nu))}{(x!)^\nu Z(\lambda,\nu)} \\ &= \frac{1}{x^{\nu-1}}\frac{\lambda^{x-1}}{((x-1)!)^\nu Z(\lambda,\nu)} - \frac{\lambda^x}{(x!)^\nu Z(\lambda,\nu)}\frac{\partial}{\partial \lambda}\log(Z(\lambda,\nu)) \qquad (6)\end{aligned}$$

Using equations (1) and (5) in equation (6), we get,

$$\frac{\partial p_x}{\partial \lambda} = \frac{1}{x^{\nu-1}}p_{x-1} - \frac{E(X)}{\lambda}p_x$$

□

An illustration of the computation of the probabilities using the recurrence relation for the parameter choice $(\lambda,\nu) = (1.2, 0.5)$ is shown below. To carry out the recursive process, the values of $P(X = 0)$ and $E(X)$ need to be computed. $P(X = 0)$ is computed by substituting $x = 0$ in equation (1) and $E(X)$ is computed from equation (5) using `com.mean()` function available in `compoisson` package in R. The values are found to be 0.2096 and 1.992285, respectively. The successive probabilities are computed using the recurrence relation

$$p_x = \frac{\lambda}{E(X)}\left[\frac{1}{x^{\nu-1}}p_{x-1} - \frac{\partial p_x}{\partial \lambda}\right], \quad x = 1, 2, \ldots$$

and are tabulated below for $x = 1, 2, 3, 4$.

**Table 1: Recurrence probabilities for** $(\lambda,\nu) = (1.2, 0.5)$

| x | $\frac{\partial p_x}{\partial \lambda}$ | $p_x$ |
|---|---|---|
| 0 | - | 0.2096 |
| 1 | 0.0962 | 0.2511 |
| 2 | 0.0020 | 0.2127 |
| 3 | 0.1246 | 0.1468 |
| 4 | 0.1487 | 0.0872 |

## 2.2. CMP as Generalized Hypergeometric distribution

A discrete random variable $X$ with pmf

$$P(X = k) = C\frac{\lambda^k}{k!}\gamma_k[(\mathbf{a}); (\mathbf{c})], \quad k = 0, 1, 2, \ldots \qquad (7)$$

is said to belong to generalized hypergeometric family of distributions, provided its pgf can be expressed in terms of generalized hypergeometric series as (See Dacey, 1972)

$$P_X(z) = C_pF_q[(\mathbf{a}); (\mathbf{c}); \lambda z] \tag{8}$$

Here $C$ denotes the normalizing constant and

$$\gamma_k[(\mathbf{a}); (\mathbf{c})] = \frac{\Gamma[(\mathbf{a}+k); (\mathbf{c}+k)]}{\Gamma[(\mathbf{a}); (\mathbf{c})]}$$

with

$$\Gamma[(\mathbf{a}); (\mathbf{c})] = \frac{\Gamma(a_1)\Gamma(a_2)\ldots\Gamma(a_p)}{\Gamma(c_1)\Gamma(c_2)\ldots\Gamma(c_q)}$$

**Result 2:** CMP distribution belongs to the generalized hypergeometric family of distributions.

**Proof:** Let ${}_pF_q[(\mathbf{a}); (\mathbf{c}); t]$ denote the generalized hypergeometric series where $\mathbf{a} = (a_1, a_2, \ldots, a_p)$ and $\mathbf{c} = (c_1, c_2, \ldots, c_q)$. The pgf of CMP distribution given in equation (3) is obtained by taking $p = 0$, $q = \nu - 1$ and $\mathbf{c} = (1, 1, \ldots, 1)$. Comparing equation (3) with equation (8), we get,

$$\begin{aligned}
P(X = k) &= \frac{1}{{}_0F_{\nu-1}(; 1, \ldots, 1; \lambda)} \frac{\lambda^k}{k!} \gamma_k[; (1, \ldots, 1)] \\
&= \frac{1}{Z(\lambda, \nu)} \frac{\lambda^k}{k!} \frac{\Gamma[; (1+k)]}{\Gamma[; (1)]} \qquad \text{(using equation (4))} \\
&= \frac{1}{Z(\lambda, \nu)} \frac{\lambda^k}{k!} \left[ \underbrace{\frac{1}{\Gamma(1+k)\ldots\Gamma(1+k)}}_{(\nu-1)\text{ terms}} \right] \left[ \underbrace{\frac{1}{\Gamma(1)\ldots\Gamma(1)}}_{(\nu-1)\text{ terms}} \right]^{-1} \\
&= \frac{1}{Z(\lambda, \nu)} \frac{\lambda^k}{k!} \frac{1}{(k!)^{\nu-1}} \\
&= \frac{1}{Z(\lambda, \nu)} \frac{\lambda^k}{(k!)^{\nu}}
\end{aligned}$$

which is the pmf of the CMP distribution. Hence the result. □

## 3. Minimum variance unbiased estimation

In this section, we propose a minimum variance unbiased estimator for the location parameter $\lambda$ of the CMP distribution when $\nu$ is known. For fixed $\nu$, CMP distribution belongs to the one-parameter power series family of distributions.
From Roy and Mitra (1957), the pmf of the complete sufficient statistic $T$ of one-parameter power series family of distributions with parameter $\theta$ is given by

$$P(T = t) = \frac{A(t, n)\theta^t}{[c(\theta)]^n} \tag{9}$$

Accordingly, the MVU estimator of $\theta^r$, $r = 1, 2, \ldots$, denoted by $\delta(t, r)$ is

$$\delta(t,r) = \begin{cases} 0, & \text{if } t < r \\ \dfrac{A(t-r,n)}{A(t,n)}, & \text{if } t \geq r \end{cases} \tag{10}$$

Since CMP distribution belongs to the exponential family of distributions, for fixed $\nu$, $\sum_{i=1}^n X_i$ is a complete sufficient statistic for $\lambda$. The pmf of $T = \sum_{i=1}^n X_i$ is given by (Sellers *et al.*, 2017)

$$P(T=t) = P(t) = \frac{\lambda^t}{(t!)^\nu [Z(\lambda,\nu)]^n} \sum_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^\nu, t = 0, 1, \ldots \tag{11}$$

Comparing equation (11) with equation (9), it can be seen that

$$A(t,n) = \frac{1}{(t!)^\nu} \sum_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^\nu \tag{12}$$

Using equation (12) in equation (10) with $r = 1$, the MVU estimator of $\lambda$, namely, $\delta(t, 1) = \delta(t)$ (say) is obtained as

$$\delta(t) = \begin{cases} 0, & \text{if } t < 1 \\ \dfrac{\sum\limits_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t-1}} \binom{t-1}{x_1 \ldots x_n}^\nu}{\sum\limits_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^\nu} \, t^\nu, & \text{if } t \geq 1. \end{cases} \tag{13}$$

To verify that $\delta(t)$ is indeed an unbiased estimator of $\lambda$, we proceed as follows.

Consider the ratio of consecutive probabilities of $T$, namely,

$$\frac{P(t-1)}{P(t)} = \frac{t^\nu}{\lambda} \frac{\sum\limits_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t-1}} \binom{t-1}{x_1 \ldots x_n}^\nu}{\sum\limits_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^\nu} \tag{14}$$

Using equation (14) in equation (13) and taking expectation, we get,

$$\begin{aligned} E[\delta(t)] &= E\left[\frac{P(t-1)}{P(t)} \lambda\right] \\ &= \lambda \sum_{t=1}^\infty \frac{P(t-1)}{P(t)} P(t) \\ &= \lambda \sum_{t=1}^\infty P(t-1) \\ &= \lambda \end{aligned}$$

Thus, $\delta(t)$ is an unbiased estimator of $\lambda$.

An alternate expression for $\delta(t)$ can be obtained by using the approximation for the sums of powers of multinomial coefficients given below.

$$\sum_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^{\nu} \simeq n^{\nu t}\sqrt{\frac{K_{n\nu}}{(\pi t)^{(n-1)(\nu-1)}}}, \tag{15}$$

where

$$K_{n\nu} = \frac{n^{n(\nu-1)}}{\nu^{(n-1)}2^{(n-1)(\nu-1)}}.$$

From equation (13), an approximate expression for $\delta(t)$ is

$$\delta(t) \simeq \frac{t^{\nu}}{n^{\nu}}\sqrt{\left(\frac{t}{t-1}\right)^{(\nu-1)(n-1)}}, \quad t > 1. \tag{16}$$

An approximate expression for the variance of $\delta(t)$ in equation (16) is obtained as follows. Consider

$$\begin{aligned} V[\delta(t)] &= E[\delta^2(t)] - (E[\delta(t)])^2 \\ &\simeq E\left[\frac{t^{2\nu}}{n^{2\nu}}\left(\frac{t}{t-1}\right)^{(\nu-1)(n-1)}\right] - \lambda^2 \\ &= \left[\sum_{t=2}^{\infty}\frac{t^{2\nu}}{n^{2\nu}}\left(\frac{t}{t-1}\right)^{(\nu-1)(n-1)} P(t)\right] - \lambda^2 \\ &= \left[\sum_{t=2}^{\infty}\frac{t^{2\nu}}{n^{2\nu}}\left(\frac{t}{t-1}\right)^{(\nu-1)(n-1)} \frac{\lambda^t}{(t!)^{\nu}(Z(\lambda,\nu))^n} \sum_{\substack{x_1,x_2,\ldots,x_n \\ x_1+x_2+\ldots+x_n=t}} \binom{t}{x_1 \ldots x_n}^{\nu}\right] - \lambda^2 \\ &= \frac{1}{n^{2\nu}(Z(\lambda,\nu))^n}\sum_{t=2}^{\infty} t^{2\nu}\left(\frac{t}{t-1}\right)^{(\nu-1)(n-1)} \frac{\lambda^t}{(t!)^{\nu}}\left[n^{\nu t}\left(\frac{n^{n(\nu-1)}}{(2\pi t)^{(n-1)(\nu-1)}\nu^{n-1}}\right)^{1/2}\right] - \lambda^2 \\ &\qquad\qquad \text{(using equation (15))} \\ &= \frac{1}{n^{2\nu}(Z(\lambda,\nu))^n}\left(\frac{n^{n(\nu-1)}}{2\pi^{(n-1)(\nu-1)}\nu^{n-1}}\right)^{1/2}\sum_{t=2}^{\infty}\frac{n^{\nu t}\lambda^t t^{2\nu}}{(t!)^{\nu}}\left(\frac{1}{t-1}\right)^{(n-1)(\nu-1)} - \lambda^2. \end{aligned}$$

## 4. Comparison of MVU and likelihood estimation

In this section, the performance of the MVU estimator of $\lambda$ is compared with that of the maximum likelihood (ML) estimator through MAB and RE using simulated and real-life datasets. The likelihood function based on $n$ iid observations, namely, $\vec{x} = (x_1, x_2, \ldots, x_n)$ on $X$ having CMP distribution is given by

$$L(\lambda;\nu,\vec{x}) = \lambda^{\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}(x_i!)^{-\nu}\,[Z(\lambda,\nu)]^{-n}$$

Since $Z(\lambda,\nu)$ involve an infinite sum, a closed form expression for the likelihood estimator of $\lambda$, namely, $\hat{\lambda}_{ML}$ cannot be obtained. However, an estimate of $\hat{\lambda}_{ML}$, can be obtained using

Newton-Raphson method. `COMPoissonReg` package in R contains functions to compute the ML estimates. The RE of $\delta(t)$ with respect to $\hat{\lambda}_{ML}$ is defined as

$$RE(\delta(t), \hat{\lambda}_{ML}) = \frac{V[\hat{\lambda}_{ML}]}{V[\delta(t)]}$$

A value of RE more than one imply that $V[\delta(t)]$ is less than $V[\hat{\lambda}_{ML}]$, suggesting that $\delta(t)$ is efficient than $\hat{\lambda}_{ML}$. However, $V[\hat{\lambda}_{ML}]$ does not have a closed-form expression. Therefore, we make use of bootstrap approach to compute the RE values. The method to compute RE using bootstrap samples is given in the following steps.

1. Generate a random sample $n^*$ of size $n$ from CMP distribution for fixed $\lambda$ and $\nu$.

2. Draw $B$ bootstrap samples each of size $n$ with replacement from $n^*$.

3. For each bootstrap sample, compute $\delta(t)$ and $\hat{\lambda}_{ML}$. Denote these values as $\delta^{[b]}(t)$, $\hat{\lambda}^{[b]}_{ML}$, $b = 1, 2, \ldots, B$.

4. Using the $B$ bootstrap estimates of $\delta(t)$ and $\hat{\lambda}_{ML}$, calculate $v(\delta(t))$ and $v(\hat{\lambda}_{ML})$ defined respectively as

$$v(\delta(t)) = \frac{1}{B-1} \sum_{b=1}^{B} \left( \delta^{[b]}(t) - \delta^*(t) \right)^2$$

where

$$\delta^*(t) = \frac{1}{B} \sum_{b=1}^{B} \delta^{[b]}(t)$$

and

$$v(\hat{\lambda}_{ML}) = \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\lambda}^{[b]}_{ML} - \hat{\lambda}^*_{ML} \right)^2$$

where

$$\hat{\lambda}^*_{ML} = \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^{[b]}_{ML}$$

5. RE based on bootstrap samples is computed as the ratio of $v(\hat{\lambda}_{ML})$ to $v(\delta(t))$

### 4.1. Simulation study

A simulation study is carried out to examine the behaviour of the ML and MVU estimates by computing the MAB and RE. Random samples of sizes $n = 25, 50$ are generated from the CMP distribution by fixing the parameters $\lambda$ and $\nu$ as below.

- Case 1: $\nu = 0.2$, $\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$

- Case 2: $\nu = 2.0$, $\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$

The `COMPoissonReg` package in R (Kimberly *et al.*, 2019) is used to generate the sample observations. Case 1 corresponds to over-dispersed counts and Case 2 to under-dispersed counts. Based on the simulated observations, $\delta^{[b]}(t)$, $\hat{\lambda}_{ML}^{[b]}$ and RE are computed using the bootstrap procedure given in steps 1 to 5 of the previous section taking $B = 200$. To find $\hat{\lambda}_{ML}^{[b]}$, the in-built function `glm.cmp()` available in `COMPoissonReg` package is used. To understand the fluctuations in the above values when the sample observations change, the procedure is repeated for 100 runs. One run of the bootstrap procedure will yield a value for the RE, $v(\hat{\lambda}_{ML})$ and $v(\delta(t))$. Based on the $B$ bootstrap estimates in each run, MAB of the estimators are computed. MAB of the bootstrap estimator $L$ of the parameter $\theta$ is defined as $MAB = \dfrac{1}{B}\sum_{b=1}^{B}|L^{[b]} - \theta|$. The summary statistics of the MAB values under case 1 and 2 for $n = 25$ and 50 are presented in Table 2.

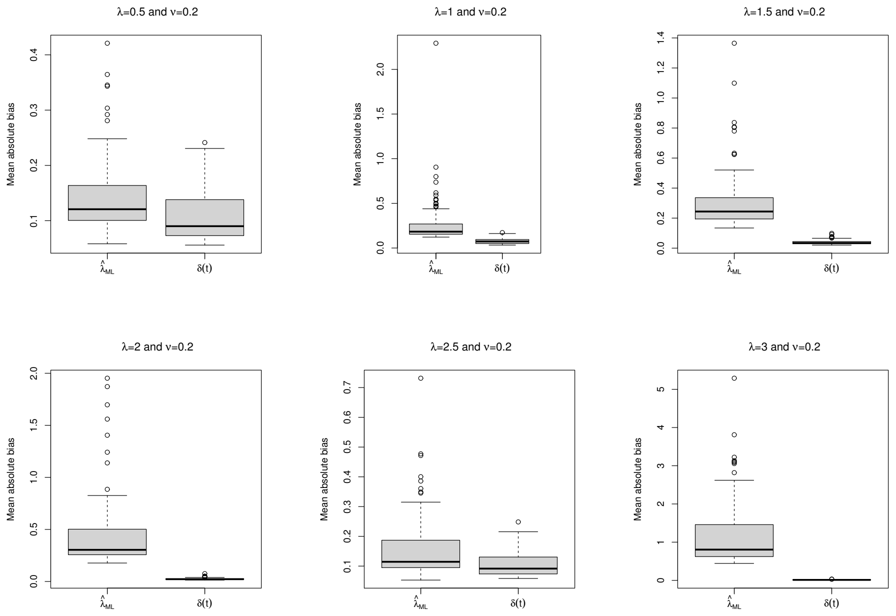
**Table 2: Summary statistics (min, $25^{th}$ quantile, median, mean, $75^{th}$ quantile, max) of MAB values of $\hat{\lambda}_{ML}$ and $\delta(t)$ under case 1 and case 2 for $n = 25$ and 50**

| | | Case 1 | |
|---|---|---|---|
| | | $\nu$=0.2 | |
| $\lambda$ | Estimator | $n = 25$ | $n = 50$ |
| 0.5 | $\hat{\lambda}_{ML}$ | (0.0750, 0.1404,0.1829,0.2504,0.2831,1.1164) | (0.0582,0.1006,0.1207,0.1460,0.1636,0.4210) |
| | $\delta(t)$ | (0.0680,0.0960,0.1196,0.1319,0.1578,0.2877) | (0.0558,0.0730,0.0900,0.1097,0.1379,0.2413) |
| 1.0 | $\hat{\lambda}_{ML}$ | (0.1680, 0.2274,0.2870,0.4468,0.5143,1.7551) | (0.1221,0.1546,0.1831,0.2734,0.2631,2.2928) |
| | $\delta(t)$ | (0.0421,0.0679,0.0802,0.0900,0.1166,0.1743) | (0.0327,0.0506,0.0734,0.0771,0.0947,0.1727) |
| 1.5 | $\hat{\lambda}_{ML}$ | (0.0421,0.0679,0.0802,0.0900,0.1166,0.1743) | (0.1343,0.1952,0.2439,0.3188,0.3358,1.3648) |
| | $\delta(t)$ | (0.0288,0.0381,0.0467,0.0519,0.0599,0.1357) | (0.0204,0.0290,0.0340,0.0390,0.0439,0.0986) |
| 2.0 | $\hat{\lambda}_{ML}$ | (0.2569,0.4035,0.5755,1.1046,1.0828,10.7535) | (0.1773,0.2574,0.3040,0.4520,0.5003,1.9523) |
| | $\delta(t)$ | (0.0187,0.0256,0.0293,0.0335,0.0348,0.0987) | (0.0135,0.0190,0.0220,0.0247,0.0269,0.0753) |
| 2.5 | $\hat{\lambda}_{ML}$ | (0.4173,0.6163,0.8842,1.9223,2.0632,24.4187) | (0.0530,0.0950,0.1145,0.1581,0.1869,0.7314) |
| | $\delta(t)$ | (0.0141,0.0186,0.0225,0.0252,0.0278,0.0625) | (0.0585,0.0739,0.0916,0.1069,0.1297,0.2484) |
| 3.0 | $\hat{\lambda}_{ML}$ | (0.5448,1.0515,1.5072,3.8200,3.2055,67.8926) | (0.4440,0.6244,0.8051,1.1763,1.4477,5.2907) |
| | $\delta(t)$ | (0.0092,0.0141,0.0176,0.0205,0.0245,0.0563) | (0.0078,0.0101,0.0126,0.0144,0.0175,0.0341) |
| | | Case 2 | |
| | | $\nu$=2.0 | |
| $\lambda$ | Estimator | $n = 25$ | $n = 50$ |
| 0.5 | $\hat{\lambda}_{ML}$ | (0.0750,0.1404,0.1829,0.2504,0.2831,1.1164) | (0.0582,0.1006,0.1207,0.1460,0.1637,0.4210) |
| | $\delta(t)$ | (0.0680,0.0960,0.1196,0.1319,0.1578,0.2877) | (0.0558,0.0730,0.0901,0.1097,0.1379,0.2413) |
| 1.0 | $\hat{\lambda}_{ML}$ | (0.0000,0.0000,0.0000,9.7e+05,1.0000,9.7e+07) | (0.2016,0.2394,0.2754,0.3733,0.4210,1.3105) |
| | $\delta(t)$ | (0.1378,0.1953,0.2286,0.2703,0.3053,1.2498) | (0.1218,0.1505,0.1869,0.2168,0.2527,0.7250) |
| 1.5 | $\hat{\lambda}_{ML}$ | (0.0000,1.0000,1.0000,1.3e+06,1.0000,7.3e+07) | (0.3160,0.3915,0.4694,0.5979,0.6645,2.0552) |
| | $\delta(t)$ | (0.1686,0.2887,0.3404,0.3877,0.4182,1.0897) | (0.1683,0.2167,0.2513,0.2860,0.3255,0.6965) |
| 2.0 | $\hat{\lambda}_{ML}$ | (1.0000,1.0000,1.0000,3.3e+06,3.0000, 1.1e+08) | (0.4564,0.5459,0.6823,1.0315,1.1313,7.3437) |
| | $\delta(t)$ | (0.2765,0.3927,0.4420,0.5051,0.5440,1.4926) | (0.1950,0.2934,0.3349,0.3841,0.4470,0.9988) |
| 2.5 | $\hat{\lambda}_{ML}$ | (1.0000,1.0000,2.0000,3.8e+06,2.3e+05,9.5e+07) | (0.5956,0.7449,0.8666,1.6464,1.7517,10.5577) |
| | $\delta(t)$ | (0.3069,0.4858,0.5898,0.6579,0.7665,1.3116) | (0.2547,0.3283,0.3813,0.4420,0.5051,1.1666) |
| 3.0 | $\hat{\lambda}_{ML}$ | (1.0000,2.0000,4.0000,6.8e+06,6.8e+05,2.1e+08) | (1.0000, 1.0000,1.0000,4.2e+05,3.0000,4.2e+07) |
| | $\delta(t)$ | (0.4241,0.5499,0.6534,0.7568,0.8361,2.5188) | (0.2375,0.3989,0.4527,0.5337,0.6380,1.4019) |

The boxplots of MAB values under both the cases for $n = 50$ are given in Table 3. It is observed from the plots and the summary statistics that the MAB values corresponding to MVU estimator are comparatively small and less dispersed than that of ML estimator under both the cases. Also, the presence of extreme values in the plots corresponding to the ML estimator indicate that the likelihood approach at times over or under estimates the parameter. The line plots displayed in Table 4 correspond to the variances of the ML and MVU estimates based on bootstrap samples for 100 runs. The x-axis in the plots denote the runs and the y-axis denote the variances of the estimates.

**Table 3: Boxplots of MABs of $\hat{\lambda}_{ML}$ and $\delta(t)$ for $n = 50$**

Case 1



Case 2

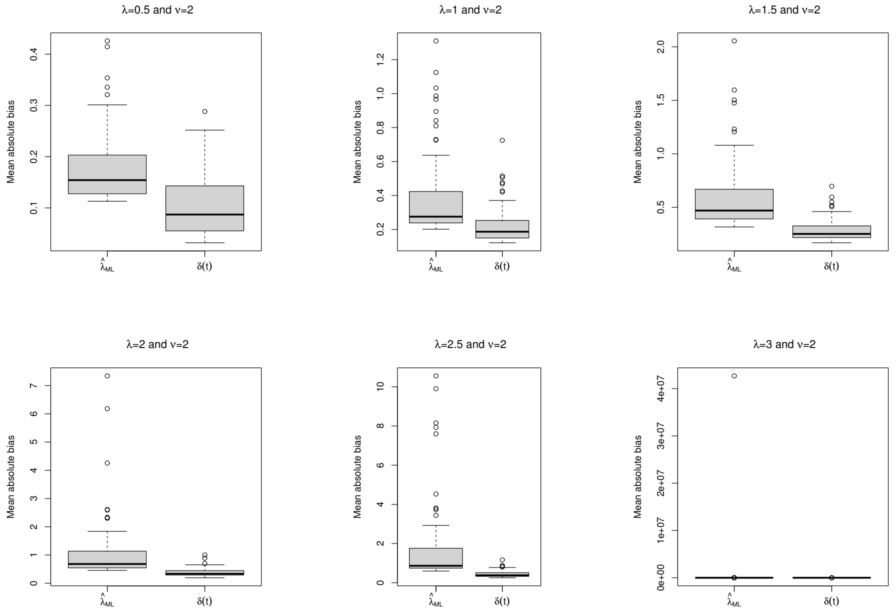
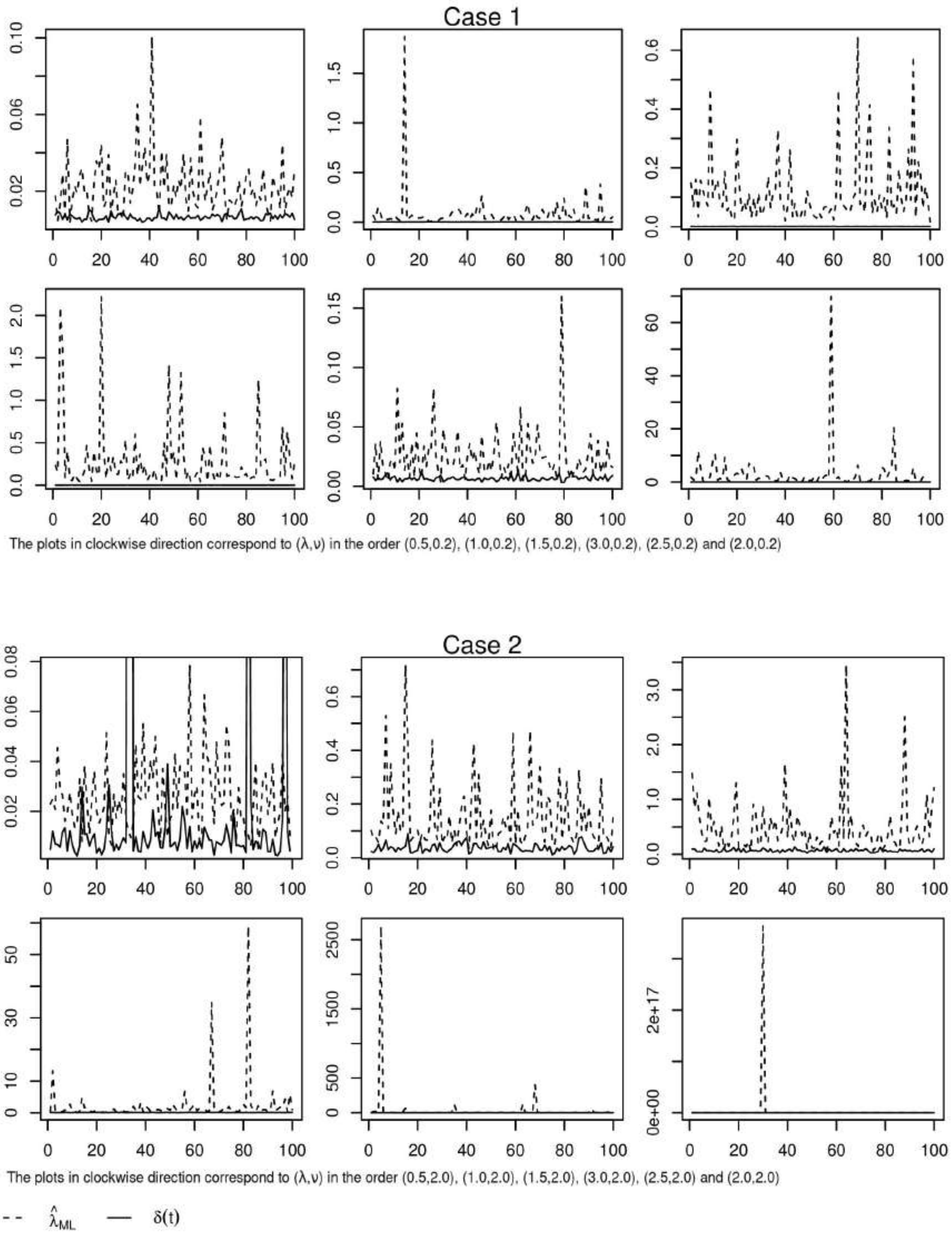


**Table 4: Plots of variances of $\hat{\lambda}_{ML}$ and $\delta(t)$ for $n = 50$**



The plots in clockwise direction correspond to (λ,ν) in the order (0.5,0.2), (1.0,0.2), (1.5,0.2), (3.0,0.2), (2.5,0.2) and (2.0,0.2)

The plots in clockwise direction correspond to (λ,ν) in the order (0.5,2.0), (1.0,2.0), (1.5,2.0), (3.0,2.0), (2.5,2.0) and (2.0,2.0)

It can be observed from the plots that the estimates of $\delta(t)$ are less dispersed compared to $\hat{\lambda}_{ML}$. Also, it is observed that the variances of the ML estimates are large when compared to that of MVU estimates. In particular, for $\nu = 2$ and $\lambda = 3$, the variance is found to be much larger than 2e+17 in some runs. However, the corresponding variances of the MVU estimates are very close to zero for all the runs.

For each of the cases, RE value is computed and the proportion of RE values greater than one in the 100 runs are obtained for $n = 25$ and 50 respectively. The proportions are given in Tables 5 and 6.

**Table 5: Proportion of RE values greater than one for Case 1**

| $\lambda$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 |
|---|---|---|---|---|---|---|
| $n = 25$ | 0.86 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| $n = 50$ | 0.88 | 0.99 | 0.99 | 1.00 | 0.86 | 1.00 |

**Table 6: Proportion of RE values greater than one for Case 2**

| $\lambda$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 |
|---|---|---|---|---|---|---|
| $n = 25$ | 0.79 | 0.94 | 0.98 | 1.00 | 0.98 | 0.97 |
| $n = 50$ | 0.88 | 0.96 | 0.98 | 0.99 | 0.99 | 0.99 |

As seen from Tables 5 and 6, the proportion of times RE values greater than one is more than 0.8, at times closer to 1, for both the cases indicating $\delta(t)$ yields estimates having smaller variance than $\hat{\lambda}_{ML}$. Thus, the simulation results indicate that the proposed MVU estimator is better than the ML in terms of MAB and variance for both over- and under-dispersed data.

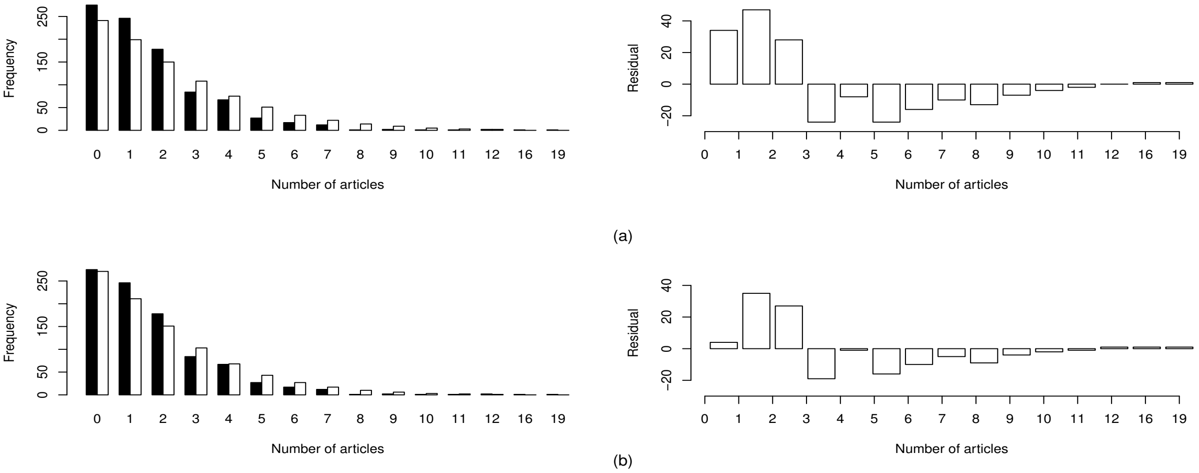
### 4.2. Real-life illustration

As an application of the proposed estimation method to real-life data, we consider the article publishing dataset given in Long (1997). The data relates to the number of articles (X) published by Ph.D. biochemists (B). The dataset is as given in Table 7.

**Table 7: Article publication data**

| X | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 16 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | 275 | 246 | 178 | 84 | 67 | 27 | 17 | 12 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |

The data is tested for equi-dispersion using `dispersiontest()` in `AER` package in R and the results indicate that the data is over-dispersed (p-value is 1.44e-06, dispersion index is 2.1889). Hence, Poisson distribution is not a suitable choice to model the data and therefore it can be modelled using CMP distribution. The dispersion parameter $\nu$ is estimated using the method of moments and is found to be $\hat{\nu} = 0.1249$. The estimate of the location parameter of $\lambda$ is obtained using the proposed MVU estimator $\delta(t)$ and the ML estimator by fixing $\nu = 0.1249$. The MVU and the ML estimates are found to be 0.8248 and 0.7809, respectively. To compute the sample variances of the estimates, bootstrap samples each of size $n = 915$ are replicated for $B = 200$ times from the data set. The corresponding sample variances are found to be 0.0001304 and 0.0021546. The plot of the observed and the expected frequencies from CMP distribution using $\delta(t)$, $\hat{\lambda}_{ML}$ and $\nu = 0.1249$ is shown in Figure 1. The corresponding residual (difference of the observed and expected frequency) plots are also presented. From the plots, it can be observed that both the estimators provide similar fits. However, the variance of the MVU estimator of $\lambda$ is smaller than that of the ML estimator suggesting that MVU estimation is efficient.



(a)

(b)

**Figure 1: Observed (■) and expected (□) frequencies of CMP distribution with (a) MVU estimate of $\lambda$ and (b) ML estimate of $\lambda$ with the corresponding residual plots**

## 5. Concluding remarks

The method of MVU estimation of the location parameter of CMP distribution proposed in this paper is simple and easy to compute. Unlike the existing estimators available in the literature, the proposed MVU estimator has a closed-form expression and does not require iterative procedures for computation. The estimator is based on the distribution of the complete sufficient statistic of the parameter. Application of the proposed estimator to simulated and real-life data reveals that the resulting estimates are less biased and efficient. Unlike the ML estimator, the proposed MVU estimator does not over or under estimate the parameter. However, to implement the proposed method, the value of the dispersion parameter $\nu$ should be known. In case it is not available, the same can be estimated using the ratio of the sample mean to the sample variance or by the method of moments.

## Acknowledgements

The authors are grateful to the anonymous referees for their suggestions and comments, which helped in improvising the paper.

## References

Boswell, M. and Patil, G. (1973). Characterization of certain discrete distributions by differential equations with respect to their parameters. *Australian Journal of Statistics*, **15**, 128–131.

Chakraborty, S. and Imoto, T. (2016). Extended Conway-Maxwell-Poisson distribution and its properties and applications. *Journal of Statistical Distributions and Applications*, **3**, 1–19.

Conway, R. and Maxwell, W. L. (1961). A queuing model with state-dependent service rate. *Journal of Industrial Engineering*, **12**, 132–136.

Cordeiro, G. M., Rodrigues, J. and de Castro, M. (2012). The exponential COM-Poisson distribution. *Statistical Papers*, **53**, 653-664.

Dacey, M. F. (1972). A family of discrete probability distributions defined by the generalized hypergeometric series. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, **34**, 243–250.

Daly, F. and Gaunt, R. (2016). The Conway-Maxwell-Poisson distribution: Distributional theory and approximation. *Latin American Journal of Probability and Mathematical Statistics*, **13**, 635–658.

Kimberly, S., Thomas, L. and Andrew, R. (2019). COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression. R package version 0.7.0. URL https://CRAN.R-project.org/package=COMPoissonReg.

Li, B., Zhang, H. and He, J. (2019). Some characterizations and properties of COM-Poisson random variables. *Communications in Statistics - Theory and Methods*, **49**, 1311–1329.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA, Sage Publications.

Nadarajah, S. (2009). Useful moment and cdf formulations for the COM–Poisson distribution. *Statistical Papers*, **50**, 617–622.

Roy, J. and Mitra, S. K. (1957). Unbiased minimum variance estimation in a class of discrete distributions. *Sankhyā: The Indian Journal of Statistics(1933-1960)*, **18**, 371–378.

Roy, S., Tripathi, R. C. and Balakrishnan, N. (2020). A Conway Maxwell Poisson type generalization of the negative hypergeometric distribution. *Communications in Statistics-Theory and Methods*, **49**, 2410–2428.

Sellers, K. F., Borle, S. and Shmueli, G. (2012). The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, **28**, 104–116.

Sellers, K. F., Swift, A. W. and Weems, K. S. (2017). A flexible distribution class for count data. *Journal of Statistical Distributions and Applications*, **4**, 1–21.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 127–142.

Statistics and Applications {ISSN 2454-7395(online)}
Volume 21, No. 1, 2023 (New Series), pp 231–244

# Single Server Poisson Queueing Model with Additive Exponential Service Time Distribution

**Ch. Ganapathi Swamy**[1], **K. Srinivasa Rao**[2] **and S. Govinda Rao**[3]
[1]*Department of Community Medicine, GSL Medical College, Rajahmundry, India*
[2]*Department of Statistics, Andhra University, Visakhapatnam, India*
[3]*Department of Statistics and Computer Applications, Agricultural College, Naira, India*

Received: 15 December 2021; Revised: 05 July 2022; Accepted: 19 July 2022

## Abstract

A stochastic process associated with queuing system is specified by the knowledge of (i) Arrival process (ii) Queue discipline (iii) Service process. Among these three, the service process is more important since it can be controlled by the operators of the system. A long with many other assumptions, it is customary to consider that the inter service time are Exponential. A generalization of it is Erlangian service time in which it is assumed that there are k-phase of service and each have identically distributed as Negative Exponential Distribution. But in many practical situations the service times are not identical. Hence in this paper we consider a queueing system with Poisson arrival having component of additive exponential service times. Using the probability generating function the system size distribution is derived. The system behaviour analyzed by deriving the system characteristics like, average number of customer in the system, the variability of system size, etc,. The waiting time distribution of the system is also derived. The sensitivity of the model with respect to the parameter is analyzed. It is observed that the system performance is influenced by the service time distribution parameters. This model includes M/M/1, M/E/1 models as particular cases for specific or timely value of parameters.

*Key words:* Queueing system; Erlangian service time; Additive exponential service times; Negative Exponential Distribution; M/M/1; M/E/1; Sensitivity analysis.

## 1. Introduction

In many of the queuing models it is customary to consider that the inter service times follows exponential distribution. In many practical situations the exponential assumption concerning service times being distributed may be rather limiting on its utility. In particular, in computer communications the service time of request is sum of two random variables namely, (1) entering (key in) time and (2) processing time. Each of these service times are exponentially distributed with different parameters say and as result of it the inter-service time between two customers follows an additive exponential distribution. Very little work has been observed regarding queueing models with additive exponential distribution. Hence, in this paper, we develop and analysis a single server queueing model with Poisson arrivals having additive exponential service times distribution.

Corresponding Author: S. Govinda Rao
Email: govinda.seepana@gmail.com

In this section, we briefly review some of the contributions in queueing models with non exponential service time in order to highlight the present work in its right perspective. Kendall (1951) used the concept of regeneration point by suitable choice of regeneration points and extracts. This method is known as embedded Markov chains. This method pioneered the M/G/1 queueing models. Keillson and Koharian (1960) developed the supplementary variable technique for analyzing the M/G/1 queueing model. This technique is very popular in analyzing the non-Markovian queueing models. Heymans (1968) considered the economic behavior of an M/G/1 queueing system that operates under the cost structure, a server start-up cost, a server shut-down cost, a cost per unit time when the server is busy and a holding cost per unit time spent in the system for each customer. The author proved that for a single server queue, there is a stationary optimal operating policy. Levy and Yechiah (1975) considered the utilization of idle time of the server in a M/G/1 for some additional work in a secondary queue. Two types of vacation policies *viz.,* M/G/1/Vs and M/G/1/Vm with exhaustive service are also studied.

Bohm (1992) considered an M/G/1 queueing model with N-policy operating. The server start up only if a queue of a prescribed length was built up. For this model, the time dependent distribution of the queue length is given by renewal arguments without resorting to integral transform techniques. Movaghar (1998) studied a queueing system where customers have strict deadlines until the beginning of their service. An analytical method is given for the analysis of a class of such queues, namely, M(n)/M/m/{rm FCFS} + {rm G} models. The principal measure of performance is the probability measure induced by the offered waiting time.

Hisashi and Brian (2001) studied the loss models in the traffic engineering of traditional telephone exchanges. These models were generalized to the loss networks, which provide models for resource-sharing in multi-service telecommunication networks. The authors introduced a generalized class of models, queueing-loss networks, which captures both queueing and loss aspects of a system. Choudhury *et. al.* (2004) considered an $M_x$/M/1 queueing model under a threshold policy with vacation process, where the server takes a sequence of vacations, till the server returns to find at least some prespecified number of customers (threshold) observed after each grand vacation.

EI-Paoumy (2008) derived the analytical solution of the queue: $M_x$/M/2/N for batch arrival system with balking, reneging and two heterogeneous servers. A modified queue discipline is used with a more general condition. The steady-state probabilities and measures of effectiveness are derived. El-Paoumy and Ismail (2009) studied $M_x/E_k$/I/N with balking and reneging queueing model in which, (i) Units arrive in batches of random size with the inter arrival times of batches following negative exponential distribution. (ii) The queue discipline is FCFS, it being assumed that the batches are pre-ordered for service purpose. (iii) The service time distribution is Erlangian with K stages. Recurrence relations connecting the various probabilities are derived. Measures of effectiveness as L and $L_q$ are deducted and some special cases are presented.

Fralix and Zwart (2010) studied a conjecture “the distribution of the number of jobs in the system of a symmetric M/G/1 queue at a fixed time is independent of the service discipline if the system starts empty”. Their arguments are based on a time-reversal argument for regenerative processes. Down *et. al.* (2011) discussed the dynamic server control in a two-class service system with abandonments. Two models are considered. In the first case, rewards are received upon service completion, and there are no abandonment costs (other than the lost opportunity to gain rewards). In the second, holding costs per customer per unit time are accrued, and each abandonment involves a fixed cost. Both cases are considered under the discounted or average reward/cost criterion. Chydzinski and Adamczyk (2019) studied Queues with the dropping function and general service time Firstly, a stability condition, more general than the well-known $\rho < 1$, is proven. Secondly, the formulas for the queue size distribution, loss ratio and mean duration of the busy period, are derived. Thirdly, numerical examples are given, including optimizations of the shape of the dropping function with regard to the combined cost of the queue size and loss ratio.

Dudin *et.al.* (2021) studied the single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities. Using the embedded Markov chain technique the probability generating function of the system size distribution under steady state condition is derived. The system performance of like the probability of system emptiness, the average no of customers in the system and in the queue, the variance of the number of customers in the system, Laplace transformation of waiting time distribution of the customers in the system, the average waiting time of the customers in the system and queue ,the variance of the waiting time distribution etc., are derived. The sensitivity of the model with respect to parameters is studied through numerical illustration. This model includes the M/M/1 model when $1/\theta_1 \to 0$ this also includes M/$E_2$/1 model if $\theta_2 \to \theta_1$.

### Additive exponential service time distribution

The additive exponential distribution was introduced as a sum of two different exponential variates. The general procedure for obtaining the probability distribution function for two independent different exponential random variables is through Jacobian transformation or inverse theorem of characteristic functions. This distribution also includes exponential if one of the parameters tends to zero. Consider two univariate continues random variables $T_1$ and $T_2$ which follow Exponential distributions with parameters $\theta_1$ and $\theta_2$ respectively. Then the addition of these two random variables $T = T_1 + T_2$ is having an Additive exponential distribution with probability density function

$$f(t) = \frac{\left(e^{-\left(\frac{t}{\theta_1}\right)} - e^{-\left(\frac{t}{\theta_2}\right)}\right)}{\theta_1 - \theta_2} \qquad \theta_1 > \theta_2 > 0; \;\; t > 0$$

### Properties of additive exponential distribution

i) If $\theta_1 \to \theta_2$ then the above probability density function gives Gamma distribution with parameters $\theta_2$ as $\theta_1 \to \theta_2$

ii) The cumulative distribution of the additive exponential distribution is,

$$= \frac{\left(e^{-\left(\frac{t}{\theta_1}\right)} - e^{-\left(\frac{t}{\theta_2}\right)}\right)}{\theta_1 - \theta_2} \qquad \theta_1 > \theta_2 > 0; \;\; t > 0$$

iii) The mean of the distribution is,

$$Mean = \theta_1 + \theta_2$$

iv) The variance of the distribution is given by,

$$\mu_2 = (\theta_1^2 + \theta_2^2)$$

v) Moment generating function of the distribution is given by,

$$M_t(x) = \frac{1}{(1 - t\theta_1)(1 - t\theta_2)}$$

vi) Characteristic function of the distribution is given by,

$$\phi_t(x) = \frac{1}{(1 - it\theta_1)(1 - it\theta_2)}$$

vii) The $r^{th}$ raw moment of the distribution is,

$$\mu_r' = \int_0^\infty t^r.b(t).dt = \frac{r!}{\theta_1 - \theta_2}\left(\theta_1^{r+1} - \theta_2^{r+1}\right)$$

viii) The $r^{th}$ cumulant of the distribution is,

$$k_r = \frac{(r-1)!}{\theta_1 - \theta_2}\left(\theta_1^{r+1} - \theta_2^{r+1}\right)$$

ix) The first four central moments of this distribution are,

$$\mu_1 = 0,\ \ \mu_2 = (\theta_1^2 + \theta_2^2),\ \ \mu_3 = 2(\theta_1^3 + \theta_2^3),\ \ \mu_4 = 9\theta_1^4 + 6\theta_1^2\theta_2^2 + 9\theta_2^4$$

x) The skewness of the distribution is,

$$= 4\frac{(\theta_1^3 + \theta_2^3)^2}{(\theta_1^2 + \theta_2^2)^3}$$

This distribution is positively skewed distribution.

xi) The kurtosis of the distribution is,

$$= 9 - 12\frac{(\theta_1\theta_2)^2}{(\theta_1^2 + \theta_2^2)}$$

## 2. Single server poisson queueing model with additive exponential service time distribution

In this section, a single server infinite capacity Poisson queueing system having FIFO discipline in which the arrivals follows a Poisson process with parameter $\lambda$ is considered. It is also assumed that the inter-service times follows an additive exponential service time distribution with parameters $\theta_1$ and $\theta_2$. The probability density function of inter-service times is,

$$f(t) = \frac{\left(e^{-\left(\frac{t}{\theta_1}\right)} - e^{-\left(\frac{t}{\theta_2}\right)}\right)}{\theta_1 - \theta_2} \qquad \theta_1 > \theta_2 > 0;\ \ t > 0 \tag{1}$$

Following the heauristic arguments of Gross and Harris (1974) the queueing model is analyzed. The embedded stochastic process X($t_i$),where, X denotes the number in the system and $t_1$, $t_2$, $t_3 \ldots$,are the successive times of completion of service. Since, $t_i$ is the completion time of the $i^{th}$ customer, then X($t_i$) is the number of customers the $i^{th}$ customer leaves behind as he departs. Since, the state space is discrete, $X_i$ represents the number of customers remaining in the system as the $i^{th}$ customer departs. Then for all n > 0 one can have.

$$X_{n+1} = \begin{cases} X_n - 1 + A_{n+1} & ; \quad X_n \geq 1 \\ A_{n+1} & ; \quad X_n = 0 \end{cases} \tag{2}$$

where $X_n$ is the number in the stem at the $n^{th}$ departure point and $A_{n+1}$ is the number of customers who arrived during the service time, $S^{n+1}$ of the $(n+1)^{th}$ customer.

The random variable $S^{n+1}$ by assumption is independent of previous service times and the length of the queue, since arrivals are Poissonian, the ransom variable $A_{n+1}$ depends only on S and not on the queue or on the time of service initiation. Then,

$$P\{A = a\} = \int_0^\infty P\{A = a|S = t\} dB(t) \tag{3}$$

$$\text{and} \qquad P\{A = a|S = t\} = \frac{e^{-\lambda t}(\lambda t)^a}{a!} \tag{4}$$

so that,

$$\begin{aligned} P\{X_{n+1} = j|X_n = i\} &= P\{A = j - i + 1\} \\ &= \begin{cases} \int_0^\infty \dfrac{e^{-\lambda t}(\lambda t)^{(j-i+1)}}{(j-i+1)!}\, dB(t) & ; \quad (j \geq i-1, i \geq 1) \\ 0 & ; \quad (j \geq i-1, i \geq 1) \end{cases} \end{aligned} \tag{5}$$

If a departing customer leaves an empty system, the system state remains zero until an arrival comes. Thus the transition probabilities for the case i=0 are identical to those for i=1. Let $p_{ij}$ denote the probability that the system size immediately after a departure point is j given that the system size after previous departure was i. $k_n$ is the probability that there are n arrivals during a service time t.

Then,

$p_{ij}$ = Pr {system size immediately after a departure point j | system size after previous departure was i}

= P$\{X_{n+1}$=j | $X_n$=i }

$$P_{ij} = \int_0^\infty \left( \frac{e^{-t\left(\lambda+\frac{1}{\theta_1}\right)} - e^{-t\left(\lambda+\frac{1}{\theta_2}\right)}}{\theta_1 - \theta_2} \right) \frac{(\lambda t)^{(j-i+1)}}{(j-i+1)!}\, dt; \quad j \geq i-1, i \geq 1 \tag{6}$$

Therefore, $k_n$ = P{n arrivals during the service time S=t}

$$k_n = \frac{\lambda^n}{\theta_1 - \theta_2} \left( \frac{\theta_1}{(\theta_1 \lambda + 1)} \right)^{n+1} - \frac{\lambda^n}{\theta_1 - \theta_2} \left( \frac{\theta_2}{(\theta_2 \lambda + 1)} \right)^{n+1} \tag{7}$$

Therefore,

$$p = [p_{ij}] = \begin{bmatrix} k_0 & k_1 & k_2 & \dots \\ k_0 & k_1 & k_2 & \dots \\ 0 & k_0 & k_1 & \dots \\ 0 & 0 & k_0 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \tag{8}$$

Assuming that the system is in steady state, and $p_{ij}$= $\pi_j$, then,

$$p = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_i k_{i-j+1} \qquad (i = 0, 1, 2, ....) \tag{9}$$

where, $\pi_j$ is the probability of j customers in the system at departure point after steady state is reached.
Let

$$K(z) = \sum_{n=0}^{\infty} k_i z^i \tag{10}$$

$$\pi(z) = \sum_{n=0}^{\infty} \pi_i z^i \qquad (|z| \leq 1) \tag{11}$$

are generating functions of $\pi_n$ and $k_n$ respectively.
Hence,

$$K(z) = \sum_{i=0}^{\infty} \frac{\lambda^i}{\theta_1 - \theta_2} \left( \frac{\theta_1}{(\theta_1 \lambda + 1)} \right)^{i+1} z^i - \sum_{i=0}^{\infty} \frac{\lambda^i}{\theta_1 - \theta_2} \left( \frac{\theta_2}{(\theta_2 \lambda + 1)} \right)^{i+1} z^i \tag{12}$$

After simplification, we get

$$K(z) = \frac{\theta_1}{(\theta_1 - \theta_2)(1 + \theta_1 \lambda (1-z))} - \frac{\theta_2}{(\theta_1 - \theta_2)(1 + \theta_2 \lambda (1-z))} \tag{13}$$

Therefore, the probability generating function of system size distribution for the M/G/1 model under consideration as,

$$\pi(z) = \frac{(1-K'(z))(1-z)K(z)}{K(z)-z} \tag{14}$$

Differentiating the equation (13) with respect to z and taking z=1 we get

$$\frac{dK(z)}{dz} = \left[\left(\frac{\theta_1}{\theta_1-\theta_2}\right)\left(\frac{\theta_1\lambda}{(1+\theta_1\lambda(1-z))^2}\right) - \left(\frac{\theta_2}{\theta_1-\theta_2)}\right)\left(\frac{\theta_2\lambda}{(1+\theta_2\lambda(1-z))^2}\right)\right]_{/z=1}$$

This implies,

$$\rho = K'(z)_{/z=1} = \lambda(\theta_1+\theta_2) \tag{15}$$

Substituting equations (13) and (15) in equation (14) we get,

$$\pi(z) = \frac{\dfrac{(1-\rho)(1-z)}{(\theta_1-\theta_2)}\left[\dfrac{\theta_1}{1+\theta_1\lambda(1-z)} - \dfrac{\theta_2}{1+\theta_2\lambda(1-z)}\right]}{\left[\left(\dfrac{1}{(\theta_1-\theta_2)}\right)\left[\dfrac{\theta_1}{1+\theta_1\lambda(1-z)} - \dfrac{\theta_2}{1+\theta_2\lambda(1-z)}\right] - z\right]} \tag{16}$$

## 3. System characteristics

In this section we derive and analyze the performance of the queueing model. The probability that there are n customers in the system at any arbitrary time is, coefficient of $z^n$,

$$p_n = A\left[\sum_{j=0}^{n/2} B(n-j)C(j) - p\sum_{j=0}^{(n/2)-1} B(n-j-1)C(j)\right], \text{where n is even}$$

$$p_n = A\left[\sum_{j=0}^{(n+1/2)} B(n-j)C(j) - p\sum_{j=0}^{(n-1/2)} B(n-j-1)C(j)\right], \text{where n is odd}$$

From the equation (16) the probability generating function of the number of customers in system is

$$\pi(z) = \frac{\dfrac{(1-\rho)(1-z)}{(\theta_1-\theta_2)}\left[\dfrac{\theta_1}{1+\theta_1\lambda(1-z)} - \dfrac{\theta_2}{1+\theta_2\lambda(1-z)}\right]}{\left[\left(\dfrac{1}{(\theta_1-\theta_2)}\right)\left[\dfrac{\theta_1}{1+\theta_1\lambda(1-z)} - \dfrac{\theta_2}{1+\theta_2\lambda(1-z)}\right] - z\right]} \tag{17}$$

Expending equation (1) and collecting the constant terms we get the probability that the system is empty as

$$P_0 = 1-\lambda(\theta_1+\theta_2) \tag{18}$$

The average number of customers in the system can be obtained as,

$$L_s = \left[\frac{d}{dz}[\pi(z)]\right]_{z=1}$$

Differentiating equation (2) and using L-Hospital rule, we get,

$$L_s = \left[\frac{[\rho - \theta\lambda^2]}{[1-\rho]}\right]$$

$$\theta_1\theta_2 = \theta, \lambda(\theta_1 + \theta_2) = \rho \tag{19}$$

The average number of customers in the queue is

$$L_q = L_s - \rho$$

$$L_q = \frac{\rho^2 - \theta\lambda^2}{1-\rho} \tag{20}$$

The variance of the number of customers in the system is given by,

$$\begin{aligned} V_s &= E(N^2 - N) + E(N) - (E(N))^2 \\ &= [\pi''(z) + \pi'(z) - [\pi'(z)]^2] \end{aligned} \tag{21}$$

Differentiating equation (1) with respect to z and using L-Hospital rule, we get the variance of the number of customers in the system as

$$V(N) = \frac{\rho - \theta\lambda^2(3 + \rho - \theta\lambda^2)}{(1-\rho)^2} \tag{22}$$

## 4. Waiting time distribution

In this section we derive the waiting time distribution of the single server Poisson arrival queueing model with additive exponential inter-service time distribution. Consider the queue discipline of the system as FIFO, following the heauristic arguments of Gross and Harris (1974) for the M/G/1 model, we derive the Laplace transformation of the waiting time distribution.

Let $B^*(s)$ be the Laplace Transformation of the inter-service time distribution and $W^*(s)$ be the Laplace transformation of the waiting time distribution. Then we have,

$$B^*(s) = \frac{1}{(s\theta_1 + 1)(s\theta_2 + 1)}$$

we have

$$K(z) = \frac{\theta_1}{(\theta_1 - \theta_2)(1 + \theta_1\lambda(1-z))} - \frac{\theta_2}{(\theta_1 - \theta_2)(1 + \theta_2\lambda(1-z))}$$

Therefore,

$$K(z) = B^*(\lambda - \lambda z) \tag{23}$$

The Laplace transformation of waiting time distribution is

$$W^*[\lambda(1-z)] = \frac{[1 - K'(1)](1-z)B^*(\lambda(1-z))}{B^*(\lambda(1-z)) - z} \tag{24}$$

where, $K'(1)$ is as given in equation (15)

Writing $\lambda(1-z)=s$, we get $z=1-\frac{s}{\lambda}$

Therefore,

$$W^*(s)=\frac{[1-K'(1)]sB^*(s)}{s-\lambda(1-B^*(s))} \tag{25}$$

From the convolution property of transformation,

$$W^*(s)=W_q^*(s).B^*(s) \tag{26}$$

where, T is the waiting time of the customer in the system and $T_q$is the time waiting time of a customer in the queue and S is the service time of the customer and $T=T_q+S$. Therefore,

$$W_q^*=\frac{[1-K'(1)]s}{s-\lambda(1-B^*(s))} \tag{27}$$

The mean waiting time of a customer in the queue is,

$$W_q=\left(\frac{d\left(W_q^*(s)\right)}{ds}\right)_{s=0}$$

$$\frac{d}{ds}\left[\frac{W_q^*(s)}{[1-K'(1)]}\right]=\frac{s-\lambda+\lambda B^*(s)-s[1+\lambda B^*(s)]}{s-\lambda[1-B^*(s)]]^2} \tag{28}$$

substituting the values of $B^*(s)$ in equation (28) and using L-Hospital rule, we get the random waiting time of a customer in the queue as

$$W_q=\frac{1}{\lambda}\left[\frac{[\theta\lambda^2-\rho]}{[1-\rho]}\right] \tag{29}$$

The waiting time of the customers in system is,

$$W_s=W_q+\rho, \quad \text{where,}\rho=\lambda(\theta_1+\theta_2)$$

Therefore,

$$W_s=\frac{\theta\lambda^2-\rho+\rho\lambda(1-\rho)}{(1-\rho)} \tag{30}$$

The variance of the waiting time of customer in the queue is,

$$V(W_q)=V_q=\left(\frac{d^2\left(W^*(s)\right)}{ds^2}\right)_{s=0}-[W_q]^2 \tag{31}$$

Therefore,

$$V_q=\frac{\rho^3(2-\rho)-2\rho\theta\lambda^2(4-\rho)-3\theta^2\lambda^4}{\lambda^2(1-\rho)^2} \tag{32}$$

**Table 1: Values of $P_0$ and $(1 - P_0)$ for different values of $\lambda$, $1/\theta_1$ and $1/\theta_2$**

| $\lambda$ | $1/\theta_1$ | $1/\theta_2$ | $P_0$ | $1 - P_0$ |
|---|---|---|---|---|
| 2 | 11 | 11 | 0.636 | 0.364 |
| 2 | 11 | 13 | 0.664 | 0.336 |
| 2 | 11 | 15 | 0.685 | 0.315 |
| 2 | 11 | 17 | 0.701 | 0.299 |
| 2 | 11 | 19 | 0.713 | 0.287 |
| 4 | 11 | 15 | 0.273 | 0.727 |
| 4 | 13 | 15 | 0.329 | 0.671 |
| 4 | 15 | 15 | 0.370 | 0.630 |
| 4 | 17 | 15 | 0.401 | 0.599 |
| 4 | 19 | 15 | 0.426 | 0.574 |
| 1 | 15 | 15 | 0.867 | 0.133 |
| 2 | 15 | 15 | 0.733 | 0.267 |
| 3 | 15 | 15 | 0.600 | 0.400 |
| 4 | 15 | 15 | 0.467 | 0.533 |
| 5 | 15 | 15 | 0.333 | 0.667 |

## 5. Sensitivity analysis

In this section, the performance of the queueing mode is discussed through a numerical illustrations. Different values of the parameter are considered for the given value of $\lambda$=1,2,3,4,5, $1/\theta_1$=11,13,15,17,19 and $1/\theta_2$=11,13,15,17,19. The probability that the system is empty and the probability the service is busy are computed and presented in Table 1.The relation between the parameters and probability of the idleness are shown in the figure 1.

From Table 1, it is observed that the probability of emptiness is highly influenced by

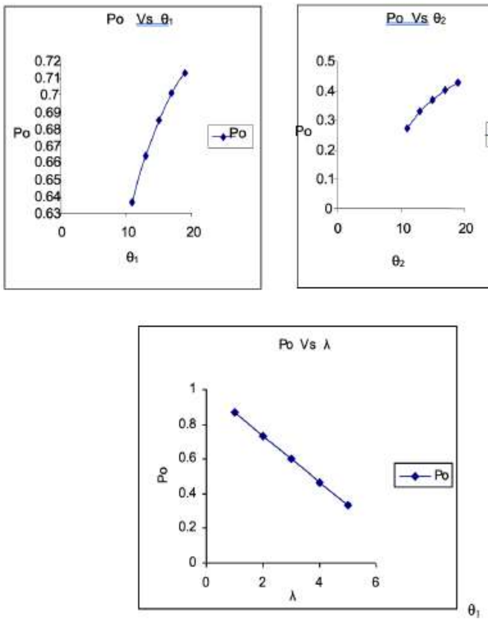


**Figure 1: Relation between probability of emptiness and input parameters**

the model parameters. As the mean arrival rate $\lambda$ varies from 1 to 5, the probability that the emptiness in the system is decreasing from 0.867 to 0.333 when other parameters are fixed at $1/\theta_1$=15 and $1/\theta_2$=15. The service time parameter $1/\theta_1$ increases from 11 to 19, the probability that the emptiness in the system increasing from 0.273 to 0.426 when other parameter are fixed at $\lambda = 4$ and $1/\theta_2 = 15$. The service time parameter $1/\theta_2$ increases from 11 to 19, the probability that the system is empty is in the system increasing from 0.636 to 0.713 when other parameter are fixed at $\lambda = 2$ and $1/\theta_1 = 11$.

For different values of the parameter the average number of customers in the system, average number of customers in the queue and the variance of the number of customers in the system are computed and presented in Table 2.The relation between the parameters and the performance measures in the figure 2. From Table 2, it is observed that the performance measures of the queueing model are significantly influenced by the parameters of the model. As the mean arrival rate $\lambda$ varies from 1 to 5, the average number of customers in the system is increasing. The same phenomenon is observed with respective the average number of customers in the queue for the given values of the other parameters.

When the parameter $1/\theta_1$ increases from 11 to 19, the average number of customers in the system is decreasing from 2.182 to 1.169 for fixed values of $\lambda$=4, $1/\theta_2$ =15. Similarly the value of average number of customers in the queue is decreasing from 1.937 to 0.773. It is observed that as $\lambda$ increases the variance of the number of customers in system is increasing from given values of the other parameters when $1/\theta_1$ is increasing the variance of the number of the customers in system is decreasing for fixed values of the other parameters. When $1/\theta_2$ is increasing the variance of the number of the customers in system is decreasing for fixed values of the other parameters.

For the different values of parameters the values of the average waiting time of customer in system, the average waiting time of customer in queue, the variance of the waiting of the customer in the queue are computed and given the Table 3. The relation between the

**Table 2: Values of $L_s$,$L_q$ and $V_s$ for different values of $\lambda$, $1/\theta_1$ and $1/\theta_2$**

| $\lambda$ | $1/\theta_1$ | $1/\theta_2$ | $L_s$ | $L_q$ | $V_s$ |
|---|---|---|---|---|---|
| 2 | 11 | 11 | 0.519 | 0.207 | 0.626 |
| 2 | 11 | 13 | 0.463 | 0.169 | 0.551 |
| 2 | 11 | 15 | 0.425 | 0.145 | 0.502 |
| 2 | 11 | 17 | 0.397 | 0.128 | 0.467 |
| 2 | 11 | 19 | 0.376 | 0.115 | 0.442 |
| 4 | 11 | 15 | 2.182 | 1.937 | 3.387 |
| 4 | 13 | 15 | 1.702 | 1.369 | 2.528 |
| 4 | 15 | 15 | 1.443 | 1.073 | 2.105 |
| 4 | 17 | 15 | 1.280 | 0.893 | 1.855 |
| 4 | 19 | 15 | 1.169 | 0.773 | 1.690 |
| 1 | 15 | 15 | 0.149 | 0.020 | 0.159 |
| 2 | 15 | 15 | 0.339 | 0.097 | 0.338 |
| 3 | 15 | 15 | 0.600 | 0.266 | 0.738 |
| 4 | 15 | 15 | 0.990 | 0.609 | 1.318 |
| 5 | 15 | 15 | 1.667 | 1.332 | 2.444 |

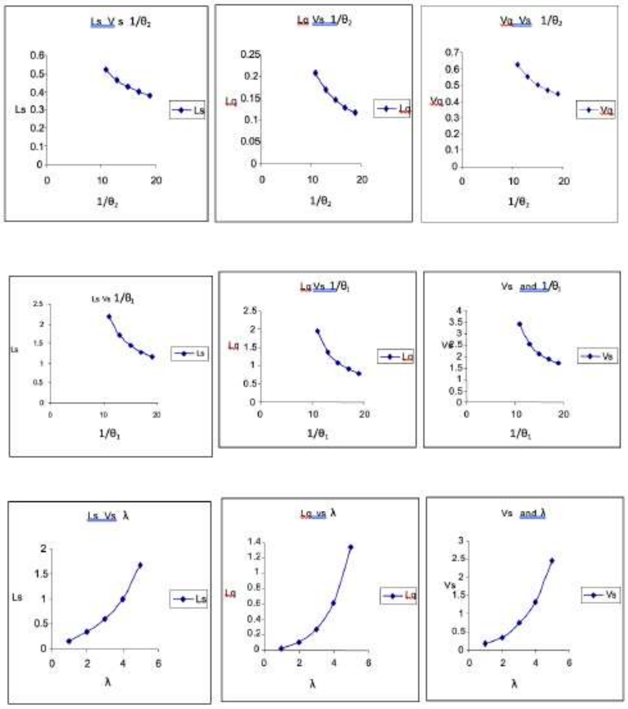


**Figure 2: Relation between probability of emptiness and input parameters**

parameters and the performance measures in the figure 3.

From the table 3 it is observed that the model parameters have a significant influence on the waiting time of the customer in the system and the queue. As the mean arrival rate $\lambda$ is increasing then the average waiting time of the customer in the queue and the average

**Table 3: Values of $W_s$,$W_q$ and $V_q$ for different values of $\lambda$, $1/\theta_1$ and $1/\theta_2$**

| $\lambda$ | $1/\theta_1$ | $1/\theta_2$ | $W_s$ | $W_q$ | $V_q$ |
|---|---|---|---|---|---|
| 2 | 11 | 11 | 0.442 | 0.078 | 0.025 |
| 2 | 11 | 13 | 0.399 | 0.064 | 0.018 |
| 2 | 11 | 15 | 0.370 | 0.005 | 0.015 |
| 2 | 11 | 17 | 0.348 | 0.049 | 0.012 |
| 2 | 11 | 19 | 0.331 | 0.044 | 0.011 |
| 4 | 11 | 15 | 1.091 | 0.364 | 0.220 |
| 4 | 13 | 15 | 0.929 | 0.258 | 0.124 |
| 4 | 15 | 15 | 0.833 | 0.203 | 0.085 |
| 4 | 17 | 15 | 0.769 | 0.170 | 0.064 |
| 4 | 19 | 15 | 0.732 | 0.149 | 0.052 |
| 1 | 15 | 15 | 0.149 | 0.015 | 0.003 |
| 2 | 15 | 15 | 0.303 | 0.036 | 0.008 |
| 3 | 15 | 15 | 0.467 | 0.067 | 0.016 |
| 4 | 15 | 15 | 0.648 | 0.114 | 0.033 |
| 5 | 15 | 15 | 0.867 | 0.200 | 0.076 |

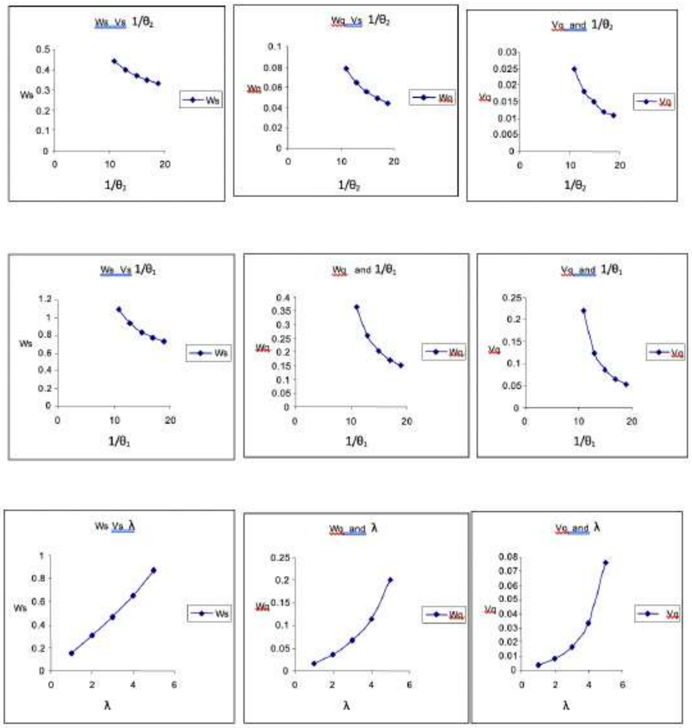


**Figure 3: Relation between $W_s$, $W_q$ and input parameters**

waiting time of customer in the system are increasing when the other parameters remain fixed. It is observed that as the parameter $1/\theta_1$ is increasing from 11 to 19, the average waiting time of the customer in the system and the average waiting time of the customer in the queue are decreasing from 1.091 to 0.732 and 0.364 to 0.149 respectively, for fixed values of other parameters. It is observed that as the parameter $1/\theta_2$ is increasing from 11 to 19, the average waiting time of the customer in the system and the average waiting time of the customer in the queue are decreasing from 0.442 to 0.331 and 0.078 to 0.044 respectively, for fixed values of other parameters. It is further observed that when the mean arrival rate $\lambda$ increases the variance of the waiting time of a customer in the system is increasing when other parameters remain fixed.

## 6. Conclusion

Developed and analyzed a single sever queueing model with Additive exponential service time distribution having Poisson arrivals. Here it is assumed that the queue discipline is FIFO. Using the embedded Markov technique the probability generating function of the queue size distribution under steady state condition is derived. The performance measures of the model like, the average number of customers in the system, the average number of customers in the queue, the probability of emptiness of the system, the probability that the server is busy, the variance of the number of the customers in system, the Laplace transformation of the waiting time distribution of a customer in the system, the average waiting time of a customer in the system, the average waiting time of customer in the queue, the variance of the waiting time of the customers are derived explicitly. The effect of the variation of the input parameter of the model on the performance measures is studied through numerical analysis. It is observed that the model parameter has significant influence on the average number of customers in the system and the average waiting time of a customer in the system and in the queue. This model also includes the M/M/I and M/$E_2$/I models as particular cases for limiting values of the parameters.

This model includes several of the earlier models as particular cases for specific or limiting values of the the parameters

If $1/\theta_1 \to 0$ then this includes M/M/1 queueing model
If $\theta_1 \to \theta_2 \to 0$ then this includes M/$E_2$/1 queueing model

The performance measures of both Exponential and Additive exponential distributions were differ.

## References

Böhm, W. (1992). A transient analysis of M/G/1 queues with N-policy. *Statistical Papers*, **33**, 151–157.

Choudhury, G., Borthakur, A. and Kalita, S. (2004). On a batch arrival Poisson queue under threshold policy with a grand vacation. *International Journal of Information and Management Sciences*, **15**, 13–27.

Chydzinski, A. and Adamczyk, B. (2019). Queues with the dropping function and general service time. *PLoS One*, **14**, 1-21.

Down, D. G., Koole, G. and Lewis, M. E. (2011). Dynamic control of a single-server system with abandonments. *Queueing Systems*, **67**, 63–90.

Dudin, A., Dudina, O., Dudin, S. and Samouylov, K. (2021). Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities. *Mathematics*, **9**, 1–17.

EI-Paoumy, M. S. (2008). On Poisson Bulk arrival queue $M^x$/M/2/N with balking, reneging and heterogeneous servers. *Applied Mathematical Sciences*, **2**, 1169–1175.

EI-Paoumy, M. S. and Ismail, M. M. (2009). On a truncated Erlang Queuing System with Bulk arrivals, Balking and Reneging *Applied Mathematical Sciences*, **3**, 1103–1113.

Fralix, B. and Zwart, B. (2011). Time-dependent properties of symmetric queues. *Queueing Systems*, **67**, 33–45.

Heyman, D. P. (1968). Optimal operating policies for M/G/1 queueing system. *Operations Research*, **16**, 362–382.

Hisashi, K. and Brian, L. M. (2001). Generalized loss models and queueing-loss networks. *International Transactions in Operations Research*, **9**, 97–112.

Keilson, J. and Kooharian, A. (1960). On time dependent queuing processes. *The Annals of Mathematical Statistics*, **31**, 104–112.

Kendall, D. G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society, Series B (Methodological)*, **13**, 151–173.

Levy, Y. and Yechiali, U. (1975). Utilization of idle time in an M/G/1 queueing system. *Management Science*, **22**, 202–211.

Movaghar, A. (1998). On queueing with customer impatience until the beginning of service. *Queueing Systems*, **29**, 337–350.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 245-249

# The Direct Method for the Optimal Solution of a Transportation Problem

**Priyanka Malviya, Sushma Jain and Rina Agrawal**
*Department of Statistics, Govt. M.V.M., Bhopal* (*M.P.*)*, India*

Received: 08 August 2021; Revised: 14 February 2022; Accepted: 24 July 2022

## Abstract

In this paper we discuss a new approach for solving both balanced and unbalanced transportation problem. The algorithm for proposed method discussed in this paper gives an initial as well as either optimal solution or near to optimal solution. Some numerical examples have been given to show the efficiency of the proposed method. Then results of the new approach are compared with the MODI method and we found that the proposed method gives either minimum or same optimal cost as compared to MODI's method and that too, in less iteration.

*Key words:* Balanced and unbalanced Transportation Problem; Basic feasible solution; Optimal solution; MODI method.

## 1. Introduction

Transportation Problem (TP) is one of the subclasses of Linear Programming Problems in which the objective is to transport various quantities of a single homogeneous commodity that are initially stored at various origins to different destinations in such a way that the total transportation cost is minimum. To achieve this objective, we must know the amount and location of available supplies and the quantities demanded. Also, we know the unit transportation cost of the commodity to be transported from various origins to destinations.

It was first studied by Hitchcock (1941) and then separately by Koopmans (1947) and finally placed in the framework of linear programming and solved by simplex method by Dantzig (1951). Since then, improved methods of solutions have been developed and the range of application has been steadily widened. It is now accepted as one of the important analytical and planning tools in business and industry. Several sorts of methods have been established for finding the optimal solution. Among them, some methods directly attain the optimal solution namely Zero Suffix Method, ASM-Method, *etc*. Also, it can be said that these methods obtain an optimal solution without disturbing degeneracy condition. They also require least iterations to reach optimality compared to the existing methods available in the literature. The degeneracy problem is also avoided by these methods. Recently, Pandian and Sudhakar proposed two different methods in 2010 and 2012 respectively for finding an optimal solution directly. However, the study on alternate optimal solutions is clearly limited in the literature of transportation except for Sudhakar, Arunnsankar and Karpagam (2012)

Corresponding Author: Priyanka Malviya
Email: deepriyanka_13@yahoo.com

who suggested a new approach for finding an optimal solution for transportation problems. Here we are proposing an easier approach for finding an optimal solution directly of the transportation problem as compared to MODI's method. Also, the proposed method is having lesser number of iterations and having easy arithmetical calculations.

## 2. Methodology

It is a simple and efficient method to obtain an optimal solution of transportation problem directly. The steps of the method are given below:

**Step 1:** Construct the transportation matrix from given transportation problem.

**Step 2:** Determine the smallest cost in the cost matrix of the transportation table. Let it be $c_{ij}$.

**Step 3:** Subtract the selected least cost $c_{ij}$ from all the remaining cost in the matrix.

**Step 4:** Compare the minimum of supply or demand whichever is minimal then allocate the minimum supply or demand at the place of minimum value of related row or column. Let the minimum of supply (or demand) corresponds to $i^{th}$ row (or $j^{th}$ column). Let $c_{ij}$ be the smallest cost in the $i^{th}$ row (or $j^{th}$ column). Allocate $x_{ij}$ = min ($a_i$, $b_j$) in the $(i, j)^{th}$ cell. If tie occurs at the place of minimum value in supply or demand, then allocate at the maximum of supply or demand is observed.

**Step 5:** After performing Step 4, delete the $i^{th}$ row (or $j^{th}$ column) for further allocation where supply from a given source is depleted (or the demand for a given destination is satisfied).

**Step 6:** Repeat Step 4 and Step 5 for the reduced transportation table until all the demands are satisfied, and all the supplies are exhausted.

## 3. Numerical Problem

**Problem 1:** Consider the following cost minimizing transportation problem (balanced case):

| | **D1** | **D2** | **D3** | **D4** | **Supply** |
|---|---|---|---|---|---|
| **S1** | 13 | 18 | 30 | 8 | **8** |
| **S2** | 55 | 20 | 25 | 40 | **10** |
| **S3** | 30 | 6 | 50 | 10 | **11** |
| **Demand** | **4** | **7** | **6** | **12** | **Total = 29** |

| | **D1** | **D2** | **D3** | **D4** | **Supply** |
|---|---|---|---|---|---|
| **S1** | 7 | 12 | 24 | 2 | **8** |
| **S2** | 49 | 14 | 19 | 34 | **10** |
| **S3** | 24 | 0 | 44 | 4 | **11** |
| **Demand** | **4** | **7** | **6** | **12** | **Total = 29** |

Following allocations are obtained by applying the proposed method:

| | **D1** | **D2** | **D3** | **D4** | **Supply** |
|---|---|---|---|---|---|
| **S1** | 4 | | | 4 | **8** |
| **S2** | | 4 | 6 | | **10** |
| **S3** | | 3 | | 8 | **11** |
| **Demand** | **4** | **7** | **6** | **12** | **Total = 29** |

The Total cost from these allocations is 412 units.

**Problem 2:** Consider the following cost minimizing transportation problem (balanced case):

| | **D1** | **D2** | **D3** | **Supply** |
|---|---|---|---|---|
| **S1** | 11 | 9 | 6 | **40** |
| **S2** | 12 | 14 | 11 | **50** |
| **S3** | 10 | 8 | 10 | **40** |
| **Demand** | **55** | **45** | **30** | **Total = 130** |

| | **D1** | **D2** | **D3** | **Supply** |
|---|---|---|---|---|
| **S1** | 5 | 3 | 0 | **40** |
| **S2** | 6 | 8 | 5 | **50** |
| **S3** | 4 | 2 | 4 | **40** |
| **Demand** | **55** | **45** | **30** | **Total = 130** |

Following allocations are obtained by applying the proposed method:

| | **D1** | **D2** | **D3** | **Supply** |
|---|---|---|---|---|
| **S1** | | **10** | **30** | **40** |
| **S2** | **50** | | | **50** |
| **S3** | **5** | **35** | | **40** |
| **Demand** | **55** | **45** | **30** | **Total = 130** |

The Total cost from these allocations is 1200 units.

**Problem 3:** Consider the following cost minimizing transportation problem (unbalanced case):

**Warehouse→**

| **Plants** | **W1** | **W2** | **W3** | **Supply** |
|---|---|---|---|---|
| **A** | 28 | 17 | 26 | **500** |
| **B** | 19 | 12 | 16 | **300** |
| **Demand** | **250** | **250** | **500** | |

**Warehouse→**

| **Plants** | **W1** | **W2** | **W3** | **Supply** |
|---|---|---|---|---|
| **A** | 28 | 17 | 26 | **500** |
| **B** | 19 | 12 | 16 | **300** |
| **C** | 0 | 0 | 0 | **200** |
| **Demand** | **250** | **250** | **500** | **Total=1000** |

Following allocations are obtained by applying the proposed method:

**Warehouse→**

| **Plants** | **W1** | **W2** | **W3** | **Supply** |
|---|---|---|---|---|
| **A** | 50 | 250 | 200 | **500** |
| **B** | | | 300 | **300** |
| **C** | 200 | | | **200** |
| **Demand** | **250** | **250** | **500** | **Total = 1000** |

The Total cost from these allocations is 15650 units.

**Problem 4:** Consider the following cost minimizing transportation problem (Degeneracy case):

| | D1 | D2 | D3 | D4 | Supply |
|---|---|---|---|---|---|
| **S1** | 3 | 7 | 6 | 4 | **5** |
| **S2** | 2 | 4 | 3 | 2 | **2** |
| **S3** | 4 | 3 | 8 | 5 | **3** |
| **Demand** | **3** | **3** | **2** | **2** | **Total = 10** |

| | D1 | D2 | D3 | D4 | Supply |
|---|---|---|---|---|---|
| **S1** | 1 | 5 | 4 | 2 | **5** |
| **S2** | 0 | 2 | 1 | 0 | **2** |
| **S3** | 2 | 1 | 6 | 3 | **3** |
| **Demand** | **3** | **3** | **2** | **2** | **Total = 10** |

Following allocations are obtained by applying the proposed method:

| | D1 | D2 | D3 | D4 | Supply |
|---|---|---|---|---|---|
| **S1** | 3 | | | 2 | **5** |
| **S2** | | | 2 | $\varepsilon_1$ | **2** |
| **S3** | $\varepsilon_2$ | 3 | | | **3** |
| **Demand** | **3** | **3** | **2** | **2** | **Total=10** |

Total transportation cost $= (3 \times 3) + (2 \times 4) + (2 \times 3) + (\varepsilon_1 \times 2) + (\varepsilon_2 \times 4) + (3 \times 3)$

$= 32 + 2\varepsilon_1 + 4\varepsilon_2$

$= 32$ as $\varepsilon_1 \to 0$ and $\varepsilon_2 \to 0$

The total transportation cost is 32 units.

## 4. Results and Comparison

Comparison of total cost of Transportation Problem of above examples between MODI method and proposed method is:

| Problem# | Type of Problem | Problem Dimension | MODI's Method | Proposed Method |
|---|---|---|---|---|
| 1 | Balanced | 3×4 | **412** | **412** |
| 2 | Balanced | 3×3 | **1320** | **1200** |
| 3 | Unbalanced | 2×3 | **15700** | **15650** |
| 4 | Degeneracy | 4×3 | **32** | **32** |

## 5. Conclusion

In this paper, a simple and more efficient method is determined to solve both the balanced and unbalanced transportation problem. Also, the proposed method gives an optimal transportation cost directly without solving the Initial Basic Feasible Solution. The new approach finds an optimal cost of the transportation problem in a very short time-period and having lesser computations as compared to MODI method. Also, the problem of degeneracy can be handled by this proposed method. Thus, our study clearly shows that the new approach is more efficient and reliable for getting an optimal solution of various types of transportation problems as compared to the well-known existing methods present in the literature.

Finally, the proposed method presented in this paper claims its wide application in solving transportation problems of higher order matrices.

## Acknowledgement

We would like to thank the staff of the Statistics Department of M.V.M. College, Bhopal for supporting this study and thankful to God and to my family specially my daughter Ojal for their support.

## References

Ahmed, M. M., Khan, A. R., Ahmed, F. and Uddin, M. S. (2016). Incessant allocation method for solving transportation problems. *American Journal of Operations Research*, **6**, 236-244.

Deshmukh, N. M. (2012). An innovative method for solving transportation problem. *International Journal of Physics and Mathematical Sciences*, **2**, 86-91.

Frederick S. Hillier, Gerald J. Lieberman, Bodhibroto Nag and Preetam Basu (2021). *Introduction to Operations Research (SIE)*, 11th Edition, McGraw Hill.

Hasan, M. K. (2012). Direct methods for finding optimal solution of a transportation problem are not always reliable. *International Refereed Journal of Engineering and Science (IRJES)*, **1**, 46-52.

Khan, A. R. (2012). *Analysis and Re-Solution of the Transportation Problem: A Linear Programming Approach*. M.Phil. Thesis, Jahangirnagar University, Savar.

Pandian, P. and Natarajan, P. (2010). A new method for finding an optimal solution for transportation problems. *International Journal of Mathematics Science and Engineering Applications (IJMSEA)*, **4**, 59-65.

Panneerselvam, R. (2010). *Operations Research*. 2nd Edition, PHI Learning Private Ltd., New Delhi, 71.

Patel, G. R. and Bhathwala, P. H. (2015). The advance method for the optimum solution of a transportation problem. *International Journal of Science and Research (IJSR)*, **4**, 703-705.

Swarup, K., Gupta, P. K. and Manmohan (2011). *Operations Research*. 15th Edition, Sultan Chand & Sons Educational Publishers, New Delhi.

Sharma, G., Abbas, H. S. and Gupta, V. K. (2012). Solving transportation problem with the help of integer programming problem. *IOSR Journal of Engineering*, **2**, 1274-1277.

Sudhakar, V. J., Arunnsankar, N. and Karpagam, T. (2012). A new approach for finding an optimal solution for transportation problems. *European Journal of Scientific Research*, **68**, 254-257.

Publisher

**Society of Statistics, Computer and Applications**

**Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA**<br>
**Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA**<br>
**Tele: 011 - 40517662**<br>
**https://ssca.org.in/**<br>
**statapp1999@gmail.com**

2023

**Printed by : Galaxy Studio & Graphics**

**Mob: +91 9818 35 2203, +91 9582 94 1203**