Gujarat Journal of Statistics and Data Science

Volume: 39 (New Series) Issue: 1 October 2023

Published by

Gujarat Statistical Association

GUJARAT JOURNAL OF STATISTICS AND DATA SCIENCE
(Formerly GUJARAT STATISTICAL REVIEW)

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Vol – 39 Issue – 1 October, 2023

CONTENTS PAGE

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Vol – 39 Issue – 1 October, 2023

Gujarat Journal of Statistics and Data Science
(Formerly Gujarat Statistical Review)
Editor in Chief: message

It's my great pleasure to bring up the next issue of this journal containing research papers of the National and International reputed authors. Professor Bikas Kumar Sinha, Managing Editor, and Professor Ashis SenGupta, Editor of this journal are extremely helpful at every stage with extreme courtesy for having the publication of this volume from beginning to end. It has been a wonderful and truly rewarding experience to work with them. I am also thankful to Dr. Parag Shah who deserves the special mention because of his continuous support with all the editorial works and extreme interest. He has been quite helpful in compiling the research papers with manuscript numbers and then making correspondence with authors, the Editor, the Managing Editor and the Editor-in-chief from time to time. I would like to acknowledge Dr. R. D. Chaudhari for his meticulous attention to detail in shaping the articles in LaTeX and maintaining the uniformity of the journal. All the contributory authors and referees were seriously involved in their respective academic activities for which we have passed through a lengthy editorial process over the last several months. I am extremely thankful to them for their academic and scientific interest in completing this task.

We the Editor in Chief, Managing Editor, and Editor of Gujarat Journal of Statistics and Data Science have taken a decision unanimously for the following two points:

(1) To publish a special volume of our journal in memory of a legendary Statistician of the world, the late Professor C. R. Rao, Ex-Director of the Indian Statistical Institute, Kolkata in October 2024. For this purpose, we are requesting and inviting all interested contributors to submit your research papers for possible publication in the special volume of this journal in the context of C. R. Rao memorial issue. It is also requested to circulate this message among your group for possible motivation. The research article must contain the works related to the late Professor Rao's research area.

(2) To include the names of two more experts in the list of Associate Editors of this journal. The following two experts are:
   (i) Dr. Ravindra Khattree, Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA.
   (ii) Dr. Indranil Ghosh, University of North Carolina, Wilmington, North Carolina, USA.

With these few words, we place this issue of Gujarat Journal of Statistics and Data Science before our readers at large. We fondly hope they will not be disappointed with this volume.

Dilip Kumar Ghosh

Editor in Chief

CHARACTERIZATION OF PROBABILITY MEASURES BASED ON Q-INDEPENDENT GENERALIZED RANDOM FIELDS

B.L.S. Prakasa Rao[1]

CR Rao Advanced Institute of Mathematics, Statistics and

Computer Science, Hyderabad 500046, India

Received: 07 September 2022 / Accepted: 15 August 2023

Abstract: Prakasa Rao (*Studia Sci. Math. Hungar.*, **11** (1976) 277-282) studied a characterization of probability distributions for linear functions of independent generalized random fields. These results are extended to Q-independent generalized random fields. It is known that independence of random variables implies Q-independence of them but the converse is not true.

Key words : Generalized random field; Characteristic functional; Q-independence; Gaussian characteristic functional.

MSC2020: Primary 60G60.

1 Introduction

Rao (1971) obtained some characterizations of probability distributions on the real line through linear functions of independent real-valued random variables. Some of these results were extended to linear functions of independent generalized random fields in Prakasa Rao (1976). Kagan and Szekely (2016) introduced the concept of Q-independence for real-valued random variables. A characterization of probability distributions for Q-independent random elements was presented in a collection of articles in Prakasa Rao (2016, 2017, 2018a,b,c) and for Q-independent random variables taking values in a locally compact Abelian group by Feldman (2017). We now extend the results in Prakasa Rao (1976) to Q-independent generalized random fields. It is known that independence of random variables implies their Q-independence but the converse is not true.

[1]E-mail address: blsprao@gmail.com

2 Preliminaries

Let $\mathcal{X}$ be the space of all real-valued functions $\phi(x) = \phi(x_1, \ldots, x_n)$ of n variables which are infinitely differentiable and have bounded supports. A sequence of functions $\{\phi_m, m \geq 1\}$ in $\mathcal{X}$ is said to converge to zero if there exists a constant a such that ϕ_m vanishes for $||x|| \geq a$, and if , for every integer $q \geq 1$, the sequence $\{\phi_m^{(q)}, m \geq 1\}$ converges uniformly to zero where $||x|| = (x_1^2 + \ldots + x_n^2)^{1/2}$ and $\phi^{(q)}$ denotes the q-th derivative of ϕ.Any continuous linear functional on $\mathcal{X}$ is called a *generalized function.*

A functional Φ defined on $\mathcal{X}$ is said to be a *random functional* if for every $\phi \in \mathcal{X}$ there is associated a real-valued random variable $\Phi(\phi)$. In other words, for every set of m elements $\phi_i, 1 \leq i \leq m$ in $\mathcal{X}$, one can specify the probability that

$$a_i \leq \Phi(\phi_i) \leq b_i, 1 \leq i \leq m$$

for $-\infty < a_i < b_i < \infty, 1 \leq i \leq m$ and these probability distributions are consistent. The random fuctional Φ is said to be *linear* if for any two elements $\phi, \psi \in \mathcal{X}$, and for any two real numbers α, β,

$$\Phi(\alpha\phi + \beta\psi) = \alpha\Phi(\phi) + \beta\Phi(\psi)$$

almost surely. A random functional Φ is said to be *continuous* if the convergence of the functions ϕ_{kj} to $\phi_j, 1 \leq j \leq m$ as $k \to \infty$ in $\mathcal{X}$ implies that for every bounded continuous function $f(x_1, \ldots, x_m)$,

$$\lim_{k\to\infty} \int_{R^m} f(x_1, \ldots, x_m) P_k(dx) = \int_{R^m} f(x_1, \ldots, x_m) P(dx)$$

where P is the probability measure corresponding to the random vector $(\Phi(\phi_1), \ldots, \Phi(\phi_m))$ and P_k is the probability measure corresponding to the random vector $(\Phi(\phi_{k1}), \ldots, \Phi(\phi_{km}))$.

Any continuous linear random functional on $\mathcal{X}$ is called a *generalized random function.* If the space $\mathcal{X}$ consists of functions of one variable, then the corresponding random functional is called a *generalized random process.* If the space $\mathcal{X}$ consists of functions of several variables, then the corresponding random functional is called a *generalized random field.*

Let Φ and Ψ be two generalized random fields on $\mathcal{X}$. The generalized random fields Φ and Ψ are said to be *independent* if the set of random variables $\{\Phi(\phi), \phi \in \mathcal{X}\}$ is independent of the set of random variables $\{\Psi(\phi), \phi \in \mathcal{X}\}$. This notion can be extended to any finite number of generalized random fields in an obvious manner.

Let Φ be a generalized random field. The functional

$$L(\phi) = E[e^{i\Phi(\phi)}], \phi \in \mathcal{X}$$

is called the *characteristic functional* of the generalized random field Φ. It can be shown that $L(0) = 1, L(-\phi) = L(\bar{\phi}), L(\phi)$ is continuous in ϕ and positive definite. Conversely, if $L(.)$ is a positive-definite continuous functional on $\mathcal{X}$ such that $L(0) = 1$, it can be shown that there exists a generalized random field Φ on $\mathcal{X}$ whose characteristic functional is $L(.)$. Furthermore the correspondence between the characteristic functionals $L(.)$ and the generalized random fields Φ on $\mathcal{X}$ is one to one.

Let $\Phi_1, \ldots, \Phi_k$ be generalized random fields on the space $\mathcal{X}$. The joint characteristic functional of the k-dimensional generalized random field $(\Phi_1, \ldots, \Phi_k)$ is defined by

$$L_{\Phi_1,\ldots,\Phi_k}(\phi_1, \ldots, \phi_k) = E[\exp\{i\Phi_1(\phi_1) + \ldots + i\Phi_k(\phi_k)\}], \phi_j, 1 \leq j \leq k \in \mathcal{X}.$$

If the generalized random fields are independent, then it can be shown that

$$L_{\Phi_1,\ldots,\Phi_k}(\phi_1, \ldots, \phi_k) = L_1(\phi_1) \ldots L_k(\phi_k)$$

where $L_j(.)$ is the characteristic functional of Φ_j for $1 \leq j \leq k$.

A generalized random field Φ on $\mathcal{X}$ is said to be *Gaussian* if its characteristic functional is of the form

$$L(\phi) = \exp(i\ m(\phi) - \frac{1}{2}B(\phi, \phi)), \phi \in \mathcal{X}$$

where $m(.)$ is a generalized function and $B(\phi, \psi) = E[\Phi(\phi)\Phi(\psi)], \phi, \psi \in \mathcal{X}$. It is said to be *degenerate* if its characteristic functional is of the form

$$L(\phi) = \exp(i\ m(\phi)), \phi \in \mathcal{X}$$

where $m(.)$ is a generalized function.

We refer the reader to Gelfand and Vilenkin (1964) for more details on generalized random fields and generalized random processes.

Denote by Δ_h the finite difference operator

$$\Delta_h f(\phi) = f(\phi + h) - f(h).$$

a function $f(\phi)$ defined on $\mathcal{X}$ is called a *polynomial* if

$$\Delta_h^{n+1} f(\phi) = 0$$

for some integer $n \geq 1$ and for all $\phi, h \in \mathcal{X}$. The minimal integer n for which this equality holds is called the *degree of the polynomial* $f(.)$ defined on $\mathcal{X}$.

Let $\Phi_i, 1 \leq i \leq k$ be generalized random fields on $\mathcal{X}$. Let $\beta_i, 1 \leq i \leq k$ be nonzero real numbers. We define the process $\beta_1\Phi_1 + \ldots + \beta_k\Phi_k$ to be the process for which to every $\phi \in \mathcal{X}$ corresponds the random variable $\Phi_1(\beta_1\phi) + \ldots + \Phi_k(\beta_k\phi)$.

Let $\Phi_1, \ldots, \Phi_k$ be generalized random fields. We say that they are *Q-independent* if their joint characteristic functional can be represented in the form

$$(2.\ 1) \qquad L_{\Phi_1,\ldots,\Phi_k}(\phi_1, \ldots, \phi_k) = \Pi_{i=1}^k L_{\Phi_i}(\phi_i) \exp(q(\phi_1, \ldots, \phi_k)), \phi_1, 1 \leq i \leq k \in \mathcal{X}$$

where $q(\phi_1, \ldots, \phi_k)$ is a continuous polynomial on the space $\mathcal{X}^k$ and $q(0, \ldots, 0) = 0$.

Suppose that $\Phi_i, i = 1, 2, 3$ are independent Gaussian random fields. Then it is obvious that $\eta_1 = \Phi_1 + \Phi_2$ and $\eta_2 = \Phi_1 + \Phi_3$ are not independent random fields. However they are Q-independent. This can be seen by computing the joint characteristic functional of the bivariate generalized random field (η_1, η_2) and the characteristic functionals of the generalized random fields η_1 and η_2.

The following result is a consequence of the Marcinkeiwicz theorem (cf. Marcinkiewicz (1938)) for real-valued random variables.

Theorem 2.1: Let $f(y)$ be the characteristic functional of a generalized random field Φ on $\mathcal{X}$. If

$$f(y) = \exp[P(y)], y \in \mathcal{X},$$

where $P(y)$ is a continuous polynomial in $y \in \mathcal{X}$, then $P(y)$ is a polynomial of degree less than or equal to 2 and $f(y)$ is the characteristic functional of a Gaussian random field which could be degenerate.

Proof : By the definition of the characteristic functional of the generalized random field Φ on $\mathcal{X}$, it follows that

$$E[\exp(i\Phi(y))] = \exp[P(y)], y \in \mathcal{X}.$$

Hence

$$E[\exp(it\Phi(y))] = E[\exp(i\Phi(ty))] = \exp[P(ty)]$$

for any real $t \in R$. Since the function on the left side of the above equation is the characteristic function of the real valued random variable $\Phi(y)$ it follows that the function on the right side of the equation has to be a polynomial of degree less than or equal to 2 in ty by the classical Marcinkeiwicz theorem. Choosing $t = 1$, it follows that $P(y)$ is a polynomial of degree less than or equal to 2 in y which in turn implies that the generalized random field is either degenerate or is Gaussian.

We now prove a theorem dealing with functional equations on the space $\mathcal{X}$ which is of independent interest. Proof of the theorem is similar to that when the space $\mathcal{X}$ is the set of real numbers (cf. Kagan et al. (1973)). Our presentation is similar to that in Feldman (2017) when the space $\mathcal{X}$ is a locally compact Abelian group. We present the detailed proof for completeness.

Theorem 2.2 : Let $\mathcal{X}$ be the space of infinitely differentiable functions. Consider the functional equation

$$\sum_{j=1}^{n} \psi_j(u + b_j v) = P(u) + Q(v) + R(u, v), u, v \in \mathcal{X} \tag{2. 2}$$

where $b_1, \ldots, b_n$ are nonzero real numbers with $b_i \neq b_j, 1 \leq j \leq n$ and $\psi_j(u), 1 \leq j \leq n, P(u), Q(v)$ are functions on $\mathcal{X}$ and $R(u, v)$ is a polynomial on $\mathcal{X} \times \mathcal{X}$. Then $P(u)$ is a polynomial on $\mathcal{X}$.

Proof: We use the finite difference method for proving the theorem. Let h_1 be an arbitrary element of $\mathcal{X}$. Define $k_1 = -b_n^{-1} h_1$. Then $h_1 + b_n k_1 = 0$. Substitute $u + h_1$ for u and $v + k_1$ for v in the equation (2.2). Subtracting the equation (2.2) from the resulting equation, it follows that

$$\sum_{j=1}^{n-1} \Delta_{\ell_{1j}} \psi_j(u + b_j v) = \Delta_{h_1} P(u) + \Delta_{k_1} Q(v) + \Delta_{(h_1,k_1)} R(u, v), u, v \in \mathcal{X} \tag{2. 3}$$

where $\ell_{1j} = h_1 + b_j k_1 = (b_j - b_n)k_1, j = 1, \ldots, n-1$. Let h_2 be an arbitrary element of $\mathcal{X}$. Let $k_2 = -b_{n-1}^{-1} h_2$. Then $h_2 + b_{n-1} k_2 = 0$. Substitute $u + h_2$ for u and $v + k_2$ for v in the equation (2.3). Subtracting equation (2.3) from the resulting equation, it follows that

$$\sum_{j=1}^{n-2} \Delta_{\ell_{2j}} \Delta_{\ell_{1j}} \psi_j(u + b_j v) \;=\; \Delta_{h_2} \Delta_{h_1} P(u) + \Delta_{k_2} \Delta_{k_1} Q(v) \tag{2. 4}$$

$$+\Delta_{(h_2,k_2)}\Delta_{(h_1,k_1)}R(u,v), u, v \in \mathcal{X},$$

where $\ell_{2j} = h_2 + b_j k_2 = (b_j - b_{n-1})k_2, j = 1, \ldots, n-2$. Following similar arguments. we get the equation

(2. 5)

$$\begin{aligned}\Delta_{\ell_{n-1,1}}\Delta_{\ell_{n-2,1}}\ldots\Delta_{\ell_{1,1}}\psi_1(u+b_1v) &= \Delta_{h_{n-1}}\Delta_{h_{n-2}}\ldots\Delta_{h_1}P(u)\\ &\quad +\Delta_{k_{n-1}}\Delta_{k_{n-2}}\ldots\Delta_{k_1}Q(v)\\ &\quad +\Delta_{(h_{n-1},k_{n-1})}\Delta_{(h_{n-2},k_{n-2})}\ldots\Delta_{(h_1,k_1)}R(u,v),\end{aligned}$$

for $u, v \in \mathcal{X}$, where h_m are arbitrary elements in $\mathcal{X}, k_m = -b_{n-m+1}^{-1}h_m, m = 1, 2, \ldots, n-1, \ell_{mj} = h_m + b_j k_m = (b_j - b_{n-m+1}k_m, j = 1, 2, \ldots n-m$. Let h_n be an arbitrary element of $\mathcal{X}$. Let $k_n = -b_1^{-1}h_n$. Then $h_n + b_1 k_n = 0$. Substitute $u + h_n$ for u and $v + k_n$ for v in the equation (2.5). Subtracting the equation (2.5) from the resulting equation, we get that

(2. 6)
$$\begin{aligned}&\Delta_{h_n}\Delta_{h_{n-1}}\ldots\Delta_{h_1}P(u) + \Delta_{k_n}\Delta_{k_{n-1}}\ldots\Delta_{k_1}Q(v)\\ &+\Delta_{(h_n,k_n)}\Delta_{(h_{n-1},k_{n-1})}\ldots\Delta_{(h_1,k_1)}R(u,v) = 0, u, v \in \mathcal{X}.\end{aligned}$$

Let h_{n+1} be an arbitrary element of $\mathcal{X}$. Substitute h_{n+1} for u in the equation (2.6). Subtracting the equation (2.6) from the resulting equation, we obtain that

(2. 7)
$$\begin{aligned}&\Delta_{h_{n+1}}\Delta_{h_n}\Delta_{h_{n-1}}\ldots\Delta_{h_1}P(u)\\ &+\Delta_{(h_n,k_n)}\Delta_{(h_{n-1},k_{n-1})}\ldots\Delta_{(h_1,k_1)}R(u,v) = 0, u, v \in \mathcal{X}.\end{aligned}$$

Observe that, if h and k are arbitrary elements of the space $\mathcal{X}$, it follows that

(2. 8)
$$\Delta_{(h,k)}^{\ell+1}R(u,v) = 0, u, v \in \mathcal{X}$$

for some integer $\ell \geq 0$ since $R(u,v)$ is a polynomial in (u,v) by hypothesis. Since $h_m, m = 1, \ldots, n+1$ are arbitrary elements of the space $\mathcal{X}$, let us choose $h_1 = \ldots = h_{n+1} = h \in \mathcal{X}$ in the equation (2.7) and apply the operator $\Delta_{(h,k)}^{\ell+1}$ to both sides of the resulting equation. Applying the equation (2.8) now leads to the equation

(2. 9)
$$\Delta_h^{\ell+n+2}P(u) = 0, u, h \in \mathcal{X}.$$

Hence the function $P(u)$ is a polynomial of degree at most $\ell + n + 1$.

Remarks: Let ℓ be the degree of the polynomial $R(u, v)$ in Theorem 2.2. Following the methods in Kagan et al. (1973), it can be shown that the degree of the polynomial $P(u)$ in Theorem 2.1 does not exceed $\max(n, \ell)$ where n is the number of functions in the left side of the functional equation (2.2).

Two generalized random fields Φ and Ψ are said to be "determined up to a Gaussian generalized random field" if there exist a generalized random field Λ such that $\Phi = \Psi + \Lambda$ almost surely. They are said to be determined up to "translation" if there exists a generalized function m such that $\Phi = \Psi + m$ almost surely.

3 Main Results

We now prove a theorem characterizing generalized random fields up to Gaussian factors.

Theorem 3.1: Let $\Phi_i, 0 \leq i \leq 3$ be four Q-independent generalized random fields on $\mathcal{X}$ and let

$$\begin{aligned} \Psi_1 &= \Phi_0 + \Phi_1 + \Phi_2 + \Phi_3 \\ \Psi_2 &= \beta_0\Phi_0 + \beta_1\Phi_1 + \beta_2\Phi_2 + \beta_3\Phi_3 \end{aligned} \tag{3. 1}$$

where $\beta_i, 0 \leq i \leq 3$ are non-zero real numbers such that $\beta_i \neq \beta_j, 0 \leq i \neq j \leq 3$. Further suppose that the joint characteristic functional $H(\phi, \psi)$ of (Ψ_1, Ψ_2) does not vanish. If $L_i(\phi)$ and $M_i(\phi)$ are two alternate possible characteristic functionals of the generalized random field $\Phi_i, 0 \leq i \leq 3$, then

$$L_j(\phi) = M_j(\phi) \exp(i\, m_j(\phi) - \frac{1}{2} B_j(\phi, \phi)), 0 \leq j \leq 3 \tag{3. 2}$$

for some generalized functions $m_j(\phi), 0 \leq j \leq 3$ and for some continuous bilinear Hermitian functionals $B_j(\phi, \psi), 0 \leq j \leq 3$.

Proof: Let $\Gamma_i, 0 \leq i \leq 3$ be Q-independent generalized random fields on $\mathcal{X}$ such that the two-dimensional generalized random field (Σ_1, Σ_2) where

$$\Sigma_1 = \Gamma_0 + \Gamma_1 + \Gamma_2 + \Gamma_3 \tag{3. 3}$$

$$\Sigma_2 \quad = \quad \beta_0\Gamma_0 + \beta_1\Gamma_1 + \beta_2\Gamma_2 + \beta_3\Gamma_3$$

has the same joint characteristic functional $H(\phi, \psi)$ as that of (Ψ_1, Ψ_2). Let $L_i(.)$ and $M_i(.), 0 \leq i \leq 3$ be the characteristic functionals of Φ_i and $\Gamma_i, 0 \leq i \leq 3$ respectively. From the Q-independence of the generalized random fields $\Phi_i, 0 \leq i \leq 3$, it follows that

$$H(\phi, \psi) = \Pi_{i=0}^{3} M_i(\phi + \beta_i\psi) \exp(P_1(\phi, \psi)), \phi, \psi \in \mathcal{X}$$

for some polynomial $P_1(\phi, \psi)$. From the Q-independence of the generalized random fields $\Gamma_i, 0 \leq i \leq 3$, it follows that

$$H(\phi, \psi) = \Pi_{i=0}^{3} L_i(\phi + \beta_i\psi) \exp(P_2(\phi, \psi)), \phi, \psi \in \mathcal{X}$$

for some polynomial $P_2(\phi, \psi)$. Hence

$$\begin{aligned} H(\phi, \psi) &= \Pi_{i=0}^{3} M_i(\phi + \beta_i\psi) \exp(P_1(\phi, \psi)) \\ &= \Pi_{i=0}^{3} L_i(\phi + \beta_i\psi) \exp(P_2(\phi, \psi)), \phi, \psi \in \mathcal{X}. \end{aligned} \tag{3. 4}$$

Since $H(\phi, \psi) \neq 0$ for all $\phi, \psi \in \mathcal{X}$ by hypothesis, the equation given above implies that $L_i(\phi + \beta_i\psi) \neq 0, 0 \leq i \leq 3$ and $M_i(\phi + \beta_i\psi) \neq 0, 0 \leq i \leq 3$ for all $\phi, \psi \in \mathcal{X}$. Let

$$J_i(\phi) = \log \frac{L_i(\phi)}{M_i(\phi)}, 0 \leq i \leq 3$$

where the logarithm is taken to be the continuous branch with $J_i(0) = 0$. The equation (3.4) implies that

$$\sum_{i=0}^{3} J_i(\phi + \beta_i\psi) = P_1(\psi, \phi) - P_2(\psi, \phi), \phi, \psi \in \mathcal{X} \tag{3. 5}$$

where $P_1(.,.)$ and $P_2(.,.)$ are polynomials. Since $\beta_i \neq \beta_j, 0 \leq i \neq j \leq 3$ and $\beta_j \neq 0$, applying arguments similar to those in the proof of Lemma 1.5.1 in Kagan et al. (1973), it follows that the functions $J_i(\phi), i = 0, \ldots, 3$ are polynomials in ϕ on $\mathcal{X}$. Hence there exists polynomials $f_j(\phi)$ such that

$$L_j(\phi) = M_j(\phi) \exp[f_j(\phi)], \phi \in \mathcal{X}, 0 \leq j \leq 3. \tag{3. 6}$$

Note that the functional $L_j(,)$ on the left side of the equation (3.6) is a characteristic functional and it is non-vanishing by the equation (3.4). Hence the function on the right side of the equation is also a non-vanishing characteristic functional which in turn implies that

the functional $\exp[f_j(\phi)], \phi \in \mathcal{X}$ is a characteristic functional by the one-to-one correspondence between the probability measures and the characteristic functionals on the space $\mathcal{X}$. An application of the Marcinkeiwicz lemma (cf. Theorem 2.1) implies that the degree of the polynomial $f_j(.)$ can not exceed two. It can be shown that

$$L_j(\phi) = M_j(\phi) \exp(i\, m_j(\phi) - \frac{1}{2} B_j(\phi, \phi)), 0 \leq j \leq 3 \tag{3. 7}$$

for some generalized functions $m_j(\phi), 0 \leq j \leq 3$ and for some continuous bilinear Hermitian functional $B_j(\phi, \psi), 0 \leq j \leq 3$ by arguments similar to those in Prakasa Rao (1976), p.281.

The following theorem can be proved by arguments similar to those given above. We omit the details.

Theorem 3.2: Let $\Phi_i, 0 \leq i \leq 2$ be four Q-independent generalized random fields on $\mathcal{X}$ and let

$$\begin{aligned} \Psi_1 &= \Phi_0 + \Phi_1 + \Phi_2 \\ \Psi_2 &= \beta_0 \Phi_0 + \beta_1 \Phi_1 + \beta_2 \Phi_2 \end{aligned} \tag{3. 8}$$

where $\beta_i, 0 \leq i \leq 2$ are non-zero real numbers such that $\beta_i \neq \beta_j, 0 \leq i \neq j \leq 2$. Further suppose that the joint characteristic functional $H(\phi, \psi)$ of (Ψ_1, Ψ_2) does not vanish. If $L_i(\phi)$ and $M_i(\phi)$ are two alternate possible characteristic functionals of the generalized random field $\Phi_i, 0 \leq i \leq 2$, then

$$L_j(\phi) = M_j(\phi) \exp(i\, m_j(\phi)), 0 \leq j \leq 2 \tag{3. 9}$$

for some generalized functions $m_j(\phi), 0 \leq j \leq 2$.

Acknowledgment : Work on this paper was supported by the scheme "INSA Senior Scientist" at the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India.

References

Feldman, G. (2017) Characterization theorems for Q-independent random variables with values in a locally compact Abelian group, *Aequat. Math.*, **91**, 949-967.

Gelfand, I.M. and Vilenkin, N. Ya. (1964) *Generalized Functions*, Vol. 4, Academic Press, New York.

Kagan, A.M., Linnik, Yu.V., and Rao, C.R. (1973) *Characterization Problems in Mathematical Statistics*, Wiley, New York.

Kagan, A.M. and Szekely, G.J. (2016) An analytic generalization of independence and identical distributiveness, *Statist. Probab. Lett.*, **110**, 244-248.

Marcinkiewicz, J. (1938) Sur une propriete de la loi Gauss, *Math. Zeit.*, **44**, 612-618.

Prakasa Rao, B.L.S. (1976) On a property of generalized random fields, *Studia Sci. Math. Hungar.*, **11**, 277-282.

Prakasa Rao, B.L.S. (2016) Characterization of probability distributions through linear forms of Q-conditionally independent random variables, *Sankhya Series A*, **78-A**, 221-230.

Prakasa Rao, B.L.S. (2017) Characterization of probability measures on Hilbert spaces via Q-independence, *J. Indian Stat. Assoc.*, **55**, 95-106.

Prakasa Rao, B.L.S. (2018a) Characterization of probability measures on locally compact Abelian groups via Q-independence, *Acta Math. Szeged*, **84**, 705-711.

Prakasa Rao, B.L.S. (2018b) Characterization of probability distributions through Q-independence, *Theory Probab. Appl.*, **62**, 335-338.

Prakasa Rao, B.L.S. (2018c) On the Skitovitch-Darmois-Ramachandran-Ibragimov theorem for linear forms of Q-independent random sequences, *Studia Sci. Math. Hungar.*, **55**, 353-363.

Rao, C.R. (1971) Characterization of probability laws through linear functions, *Sankhya, Series A*, **33**, 255-259.

A new family of continuous distributions and its properties

K. Jayakumar and K. K. Sankaran[1]
Department of Statistics, University of Calicut, Kerala-673 635, India.

Received: 27 May 2022 / Revised: 27 February 2023 / Accepted: 20 July 2023

Abstract

In this paper, we introduce a new family of continuous distributions called the Kumaraswamy discrete Linnik generalized family of distributions. Our proposed family of distributions is a mass collection of family of distributions such as Kumaraswamy discrete Mittag-Leffler generalized family of distributions, Kumaraswamy Marshall-Olkin generalized family of distributions, Kumaraswmay truncated negative binomial generalized family of distributions, etc. In particular, we study Kumaraswamy truncated discrete Mittag-Leffler exponential (Kw-DML-E) distribution in detail. The Kw-DML-E distribution contains Kumaraswamy Marshall-Olkin exponential distribution, Kumaraswamy generalized exponential distribution, Marshall-Olkin generalized exponential distribution, Marshall-Olkin exponential distribution, generalized exponential distribution and exponential distribution as special case. The density function of Kw-DML-E is symmetrical or right skewed and has constant, increasing or decreasing, hazard rate. We derive explicit expression for the moments, generating functions and quantiles. Two characterizations of Kw-DML-E distribution are obtained. The method of maximum likelihood is used to estimate the model parameters. The existence and uniqueness of maximum likelihood estimates are proved. Simulation studies are also performed. An application to a real data set is presented to illustrate the potentiality of our proposed model.

Keywords: Discrete Linnik distribution, Discrete Mittag-Leffler distribution, Exponential distribution, Kumaraswamy distribution, Marshall-Olkin family of distributions, Maximum likelihood.

1. Introduction

By various methods, new parameters can be introduced to expand families of distributions for added flexibility or to construct covaraite model. The addition of parameters has been proved useful in exploring skewness and tail properties, and also for improving the goodness-of-fit of the generated family. Introduction of a scale parameter leads to accelerate life model and taking powers of survival function introduces a parameter that leads to proportional hazards model. Also, the extended distributions have attracted several statisticians to develop new models because the analytical and computational facilities available in programming softwares such as Mathcad, Mapple, MathLab and R can easily tackle the problems involved in computing special functions in these extended distributions.

Marshall and Olkin (1997) introduced a new family of distributions by adding a parameter to a family of distributions. They started with a parent survival function $\overline{F}(x)$ and considered a family

[1]Corresponding author: Sree Narayana College Nattika, Kerala-680 566, India. email:snsankaran08@gmail.com

of survival functions given by

$$\overline{G}(x;p) = \frac{p\overline{F}(x)}{F(x) + p\overline{F}(x)}, \quad p > 0 \quad x \in \mathbb{R}. \tag{1}$$

They described the motivation for the family of distributions (1) as follows:
Let X_1, X_2, ... be a sequence of independent and identically distributed (i.i.d.) random variables with survival function $\overline{F}(x)$. Let

$$U_N = \min(X_1, X_2, ..., X_N), \tag{2}$$

where N is the geometric random variable with probability mass function (pmf) $P(N = n) = p(1-p)^{n-1}$, for $n = 1, 2, \ldots$ and $0 < p < 1$ and independent of $X_i's$. Then the random variable U_N has the survival function given by (1). If $p > 1$ and N is a geometric random variable with pmf of the form $P(N = n) = \frac{1}{p}(1-\frac{1}{p})^{n-1}$, $n = 1, 2, \ldots$, then the random variable $V_N = \max(X_1, X_2, \ldots, X_N)$ also has the survival function as (1).

If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with distribution in the family (1), and if N has a geometric distribution on $\{1, 2, \ldots\}$, then $\min(X_1, \ldots, X_N)$ and $\max(X_1, \ldots, X_N)$ have distributions in the family. The extreme value distributions are limiting distribution for extreme, and as such they are sometimes useful approximations. In practice, a random variable of interest may be the extreme of only finite, possibly random, number N of random variables. When N has a geometric distribution, the random variable has a particular nice stability property, not unlike that of extreme value distributions. The geometric-extreme stability property of $G(x;p)$ is rather remarkable, and it depends upon the fact that a geometric sum of i.i.d. geometric random variables has a geometric distribution. This partially explains why random minimum stability cannot be expected if the geometric distribution is replaced by some other distribution on $\{1, 2, \ldots\}$. For more discussion on geometric-extreme stability, see Arnold et al. (1986) and Marshall and Olkin (1997).

Pillai and Jayakumar (1995) introduced a class of discrete distributions containing geometric and it is named as discrete Mittag-Leffler (DML) distribution, since it arises as a discrete analogue of the well-known continuous Mittag-Leffler distribution introduced by Pillai (1990). The DML distribution is the distribution of $Z = X_1 + X_2 + \ldots + X_N$ where X_i's are i.i.d. Sibuya(α) random variables and N is geometric independent of X_i's. Note that the Sibuya random variable represents the number of trails till the first success in a sequence of independent Bernoulli trails, where the probability of success varies with trail, and for the k^{th} trail, probability of success equals $\frac{k}{n}$. The probability generating function (pgf) of DML distribution is

$$H(s) = \frac{1}{1 + c(1-s)^{\alpha}}, \quad c > 0,\ 0 < \alpha \leq 1,\ |s| \leq 1.$$

Note that, when $\alpha = 1$, $H(s)$ is the pgf of geometric distribution. Also Pillai and Jayakumar (1995) showed that DML distribution is infinitely divisible, geometrically infinitely divisible and belongs to discrete class L. For the application of DML distribution, see also Huillet (2016).

Sankaran and Jayakumar (2016) considered the distribution of U_N in (2) when X_i's are i.i.d. random variables having cumulative distribution function (cdf) $F(x)$ and N is truncated DML, independent of X_i's, with parameters α and c. They showed that the survival function of U_N is

$$\bar{G}(x;\alpha,c) = \frac{1 - F^{\alpha}(x)}{1 + cF^{\alpha}(x)}. \tag{3}$$

Like the Marshall-Olkin family of distributions, the family of distributions generated through truncated DML distribution is also geometric extreme stable. Note that, when Marshall-Olkin method is applied to F^α, then the resulting survival function coincides with (3). The family of distributions generated through truncated DML distribution is a generalization of Marshall-Olkin family of distributions in the sense that it reduces to the Marshall-Olkin family when $\alpha = 1$ and $c = \frac{1-p}{p}$.

Nadarajah, Jayakumar and Ristić (2013) introduced a family of life time models, using truncated negative binomial distribution with pmf given by

$$P(N = n) = \frac{\alpha^\theta}{1-\alpha^\theta} \begin{pmatrix} \theta+n-1 \\ \theta-1 \end{pmatrix} (1-\alpha)^n,\ n = 1, 2, \ldots,$$

where $\alpha > 0$ and $\theta > 0$. The authors showed that the random minimum, $U_N = \min(X_1, X_2, \ldots, X_N)$ has the survival function of the form

$$\overline{G}(x; \alpha, \theta) = \frac{\alpha^\theta}{1-\alpha^\theta}[(F(x) + \alpha\overline{F}(x))^{-\theta} - 1], \tag{4}$$

when X_i's are i.i.d. random variables having distribution function $F(x)$ and N is truncated negative binomial with parameter α and θ. Note that if $\alpha \to 1$ then $\overline{G}(x; \alpha, \theta) \to \overline{F}(x)$. The family of distributions given in (4) is a generalization of Marshall-Olkin family of distributions, in the sense that when $\theta = 1$, (4) reduces to (1).

In recent years, heavy-tailed models have been used in a variety of fields, such as mathematical finance, financial economics and statistical physics. In the framework of integer valued distributions, the discrete stable is a well known heavy-tailed law originally suggested by Steutel and van Harn (1979). Jayakumar and Sreenivas (2003) generalized the concept of Poisson mixtures to discrete stable mixtures and showed that, the distributions on Z_+ that can be approximated by mixtures of discrete Linnik distributions are discrete stable mixtures. Christoph and Schreiber (1998) emphasized that the discrete stable law may be seen as a special case of discrete Linnik law studied in Devroye (1993). Hence owing to the extra parameter, the discrete Linnik is a heavy-tailed distribution family which is more flexible than the discrete stable. Discrete Linnik distribution is a rich family of distributions which includes many important distributions. It belongs to the class of discrete self decomposable distributions. When $\nu = 1$, we get DML distribution and for $\alpha = 1$, it coincides with negative binomial distribution. For $\alpha = 1$ and $\nu = 1$, we get the geometric distribution.

The pgf of discrete Linnik distribution with parameters α, c and ν is

$$H(s) = \begin{cases} \left(\frac{1}{1+c(1-s)^\alpha}\right)^\nu & \text{for} \quad 0 < \nu < \infty \\ \\ e^{-c(1-s)^\alpha} & \text{for} \quad \nu = \infty. \end{cases}$$

Jayakumar and Sankaran (2018) introduced a new family of distributions with parameters α, c and ν having survival function

$$\bar{G}(x, \alpha, c, \nu) = \frac{(1+c)^\nu - [1 + cF^\alpha(x)]^\nu}{[(1+c)^\nu - 1][1 + cF^\alpha(x)]^\nu}. \tag{5}$$

Note that the survival function in (5) is the survival function of U_N in (2), where X_i's are i.i.d. random variables with cdf $F(x)$ and N is truncated discrete Linnik distribution with parameters

α, c and ν and N is independent of X_i's. It can be seen that the family of distributions generated through truncated negative binomial and truncated discrete Linnik are not extreme stable.

Kumaraswamy (1980) introduced a probability distribution for handling double bounded random processes with varied hydrological applications. The cumulative distribution function (cdf) of Kumaraswamy distribution is given by

$$F(x) = 1 - \{1 - x^a\}^b; \quad a > 0,\ b > 0,\ x \in [0, 1]. \tag{6}$$

The beta distribution also provides the premier family of continuous distribution on bounded support which has been utilized extensively in statistical theory and practice (see Nadarajah and Gupta (2007)). Gupta and Nadarajah (2004) provides a comprehensive account of the theory and applications of beta family of distributions. Like beta distribution, Kumaraswamy distribution also originally defined on the unit interval [0, 1] but easily extended to any finite interval and can take an amazingly great variety of forms. Thus it can be fitted practically to any data representing a phenomenon in almost any field of applications. One interpretation for integer-valued a and b through maxima and minima of i.i.d. random components is by Jones(2009). If we assuming that $a = m$ and $b = n$ are positive integers, we can find, x^m is the cdf of the maximum of i.i.d. standard uniform variables, with the corresponding survival function $1-x^m$. Thus, the quantity $\{1-x^m\}^n$ in (6) is the minimum of n such random variables, with G being the corresponding cdf. This property discussed in Jones (2009), motivated the name minimax for this distribution. Kozubowski and Podgórski (2018) extended this interpretation to the general Kumaraswamy distribution using the result of min/max of i.i.d. components with random number to the relevant pgf.

In this paper, we propose a new family of distributions, by minimax of distributions generated through truncated discrete Linnik distribution. The main motivation of this paper are:

1. To introduce a new class of univariate distributions as a generalization of families of distributions such as Kumaraswamy discrete Mittag-Leffler G family of distributions, Kumaraswamy truncated negative binomial G family of distributions, Kumaraswamy Marshall-Olkin G family of distributions introduced by Alizadeh et al. (2015), Kumaraswamy-G family of distributions introduced by Cordeiro and de Castro (2011), exponentiated Marshall-Olkin family of distributions introduced by Dias et al. (2016), etc.
2. According to the choice of baseline distribution, the shape of the density function can be symmetrical, left skewed, right skewed and reversed-J shaped. Also the hazard function can be constant, increasing, decreasing, upside-down bathtub, bathtub and S-shaped.
3. To introduce and study one member of minimax geometric-extreme stable distribution namely Kumaraswamy discrete Mittag-Leffler exponential distribution.

This paper is organized as follows. We introduce Kumaraswamy discrete Linnik G (Kw-DL-G) family of distributions in Section 2 and discuss its various sub models. In Section 3, a sub model of Kw-DL-G, namely, Kumaraswamy discrete Mittag-Leffler -G family is obtained. As a special case, Kumaraswamy discrete Mittag-Leffler exponential distribution (Kw-DML-E), a new generalization of exponential distribution is introduced in Section 4. The shape properties of density and hazard function are studied. It can be seen that Kw-DML-E distribution contains Kumaraswamy Marshall-Olkin exponential distribution, Kumaraswamy generalized exponential distribution, Kumaraswamy exponential distribution, Marshall-Olkin generalized exponential distribution, Marshall-Olkin exponential distribution, generalized exponential distribution and exponential distribution. In Section 5, some structural properties of Kw-DML-E distribution such as moments and generating

function, quantiles, unimodality, and stochastic ordering are studied. A method of generation of Kw-DML-E random variables is discussed in this section. It can be seen that the generation of the random variables is simple. Two characterizations of Kw-DML-E distribution are obtained in Section 6. Estimation of the modal parameters by maximum likelihood is performed in Section 7. The existence and uniqueness of maximum likelihood estimates are proved. Simulation studies are also carried out in order to establish the consistency property of the maximum likelihood estimates of our proposed model. An application to a real data set to illustrate the potentiality of the new family is presented in Section 8. It can be seen that Kw-DML-E distribution performs well compared to seven well known distributions. The paper is concluded in Section 9.

2. Kumaraswamy discrete Linnik G family of distributions

Let $X \sim$ truncated discrete Linnik G family of distribution with cdf

$$G(x) = \frac{1}{(1-p^{\nu})} \frac{[p+(1-p)G^{\alpha}(x)]-p^{\nu}}{[p+(1-p)G^{\alpha}(x)]^{\nu}} \tag{7}$$

We define the cdf of Kumaraswamy discrete Linnik G (Kw-DL-G) family of distributions as

$$F(x) = 1-\left\{1-\left[\frac{1}{(1-p^{\nu})} \frac{[p+(1-p)G^{\alpha}(x)]-p^{\nu}}{[p+(1-p)G^{\alpha}(x)]^{\nu}}\right]^{a}\right\}^{b}, \tag{8}$$

where $a>0,\ b>0$ and $0<p<1$ are the additional parameters. For each baseline G, the Kw-DL-G cdf is given by (8). It can be seen that equation (8) provides a class of wider family of continuous distributions. It includes the Kumaraswamy discrete Mittag-Leffler G family of distributions, Kumaraswamy truncated negative binomial G family of distributions, Kumaraswamy Marshall-Olkin G family of distributions, Kumaraswamy G family of distributions, etc. Some special cases of the Kw-DL-G models are presented in Table 1.

Sl.No.	a	b	p	α	ν	$G(x)$	Reduced model
1	-	-	-	-	1	$G(x)$	The Kw-discrete Mittag-Leffler G family of distributions
2	-	-	-	1	-	$G(x)$	The Kw-negative binomial G family of distributions
3	-	-	-	1	1	$G(x)$	The Kw- Marshall-Olkin G family of distributions
4	-	-	1	1	1	$G(x)$	The Kw- G family of distributions
5	1	1	-	-	-	$G(x)$	The discrete Linnik G family of distributions
6	1	1	-	-	1	$G(x)$	The discrete Mittag-Leffler G family of distributions
7	1	1	-	1	-	$G(x)$	The negative binomial G family of distributions
8	1	1	-	1	1	$G(x)$	The Marshall-Olkin family of distributions
9	-	1	-	1	1	$G(x)$	The exponentiated Marshall-Olkin family of distributions
10	1	-	1	1	1	$G(x)$	The proportional reversed hazard rate model
11	-	1	1	1	1	$G(x)$	The proportional hazard rate model
12	1	1	1	1	1	$G(x)$	$G(x)$

Table 1: Some special cases of Kw-DL-G distribution.

The density function corresponding to (8) is given by

$$f(x) = ab\alpha(1-p)g(x)G^{\alpha-1}(x)\left[(p+(1-p)G^{\alpha}(x))(1-\nu)+\nu p^{\nu}\right] \left[\left(\frac{1}{1-p^{\nu}}\right)^{a}\frac{[p+(1-p)G^{\alpha}(x)-p^{\nu}]^{a-1}}{[p+(1-p)G^{\alpha}(x)]^{\nu(a+1)}}\right]\left[1-\left\{\frac{1}{(1-p^{\nu})}\frac{[p+(1-p)G^{\alpha}(x)]-p^{\nu}}{[p+(1-p)G^{\alpha}(x)]^{\nu}}\right\}^{a}\right]^{b-1} \tag{9}$$

The equation (9) will be most tractable when the cdf $G(x)$ and the pdf $g(x)$ have simple analytic expressions.

The hazard rate function corresponding to $F(x)$ in (8) is

$$\begin{aligned} h(x) &= ab\alpha(1-p)g(x)G^{\alpha-1}(x)\left[(p+(1-p)G^{\alpha}(x))(1-\nu)+\nu p^{\nu}\right] \\ &\quad \left[\left(\frac{1}{1-p^{\nu}}\right)^{a}\frac{[p+(1-p)G^{\alpha}(x)-p^{\nu}]^{a-1}}{[p+(1-p)G^{\alpha}(x)]^{\nu(a+1)}}\right]\frac{1}{\left[1-\left\{\frac{1}{(1-p^{\nu})}\frac{[p+(1-p)G^{\alpha}(x)]-p^{\nu}}{[p+(1-p)G^{\alpha}(x)]^{\nu}}\right\}^{a}\right]}. \end{aligned} \tag{10}$$

3. Kumaraswamy truncated discrete Mittag-Leffler G (Kw-DML-G) distribution

For analytical tractability, let $\nu = 1$ in (8). Then we obtain the distribution function of Kumaraswamy truncated discrete Mittag-Leffler G family of distributions as

$$F(x) = 1-\left\{1-\left[\frac{G^{\alpha}(x)}{p+(1-p)G^{\alpha}(x)}\right]^{a}\right\}^{b}. \tag{11}$$

From (11), the pdf of Kw-DML-G distribution is

$$f(x) = ab\alpha p\ g(x)\frac{G^{\alpha a-1}(x)}{[p+(1-p)G^{\alpha}(x)]^{a+1}}\left\{1-\left[\frac{G^{\alpha}(x)}{p+(1-p)G^{\alpha}(x)}\right]^{a}\right\}^{b-1}. \tag{12}$$

The corresponding hazard rate function is given by

$$h(x) = \frac{ab\alpha p\ g(x)G^{\alpha a-1}(x)}{[p+(1-p)G^{\alpha}(x)][(p+(1-p)G^{\alpha}(x))^{a}-G^{\alpha a}(x)]}. \tag{13}$$

Note that when $\alpha = 1$, in (11) we obtain, Kumaraswamy Marshall-Olkin G family of distributions introduced and studied by Alizadeh et al. (2015).

Now, our study focuss on one member of Kw-DML-G distribution, namely, Kumaraswamy truncated discrete Mittag-Leffler exponential distribution, in detail.

4. A new generalization of exponential distribution

4.1. Distribution function

Let X follows exponential distribution with parameter $\lambda > 0$ having cdf $G(x) = 1-e^{-\lambda x}$ and pdf $g(x) = \lambda e^{-\lambda x}$. Hence from (11), the distribution function of the random variable X is given by

$$\begin{aligned} F(x) &= 1-\left\{1-\left[\frac{(1-e^{-\lambda x})^{\alpha}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}}\right]^{a}\right\}^{b} \\ &= 1-\frac{\eta_1\Gamma(b+1)\Gamma(aj+k)e^{-m\lambda x}}{\Gamma(b+j+1)\Gamma aj} \end{aligned} \tag{14}$$

where

$$\eta_1 = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\frac{(1-p)^{k}(-1)^{j+l+m}\Gamma(k+1)\Gamma(a\alpha(j+1)+\alpha l)}{j!k!l!m!\Gamma(k-l+1)}$$

by using generalized binomial expansion.

4.2. Probability density function

The pdf of the new distribution is given by

$$\begin{aligned} f(x;a,b,\alpha,p,\lambda) &= \frac{ab\alpha p\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a+1}}\left\{1-\left[\frac{(1-e^{-\lambda x})^{\alpha}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}}\right]^{a}\right\}^{b-1} \\ &= \eta_1\eta_2 e^{-(m+1)\lambda x} \end{aligned} \tag{15}$$

where

$$\eta_2 = \frac{ab\alpha\lambda p\Gamma b\Gamma a(j+1)+k+1}{\Gamma(b-j)\Gamma[a(j+1)+1]\Gamma a\alpha(j+1)+\alpha l-m}$$

using generalized binomial expansion.

We refer to this new distribution as Kumaraswamy truncated discrete Mittag-Leffler exponential (Kw-DML-E) distribution with parameters a, b, α, p and λ. We write it as Kw-DML-E$(a, b, \alpha, p, \lambda)$.

The shape of the density function is described analytically. The critical points of the density function of Kw-DML-E model is the roots of the equation $\frac{\partial \log f(x)}{\partial x}$=0 which yields

$$\begin{aligned} 0 &= -\lambda+(a\alpha-1)\frac{e^{-\lambda x}}{1-e^{-\lambda x}}-\frac{a(b-1)\alpha p\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]\{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a}-(1-e^{-\lambda x})^{a\alpha}\}} \\ &\quad -\frac{(a+1)(1-p)\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]}. \end{aligned} \tag{16}$$

There may be more than one roots of the above equation. Let $k(x)=\frac{\partial^2 \log f(x)}{\partial x^2}$: We have

$$\begin{aligned} k'(x) &= -\frac{(a\alpha-1)\lambda^2 e^{-\lambda x}}{(1-e^{-\lambda x})^2}-\frac{a(b-1)\alpha\lambda p e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]([p+(1-p)(1-e^{-\lambda x})^{\alpha}]-(1-e^{-\lambda x})^{a}\alpha)} \\ &\quad -\frac{(a+1)(1-p)\alpha\lambda^2 e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}}+\frac{(a+1)(1-p)(\alpha-1)\alpha\lambda^2 e^{-2\lambda x}(1-e^{-\lambda x})^{\alpha-2}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}} \\ &\quad -\frac{(a+1)(1-p)^2\alpha^2\lambda^2 e^{-2\lambda x}(1-e^{-\lambda x})^{2(\alpha-1)}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^2} \end{aligned} \tag{17}$$

If $x=x_0$ is a root of $\frac{\partial \log f(x)}{\partial x}=0$, then it corresponds to a local maximum if $\frac{\partial^2 \log f(x)}{\partial x^2}>0$ for all $x<x_0$ and $\frac{\partial^2 \log f(x)}{\partial x^2}<0$ for all $x>x_0$. It corresponds to a local minimum if $\frac{\partial^2 \log f(x)}{\partial x^2}<0$ for all $x<x_0$ and $\frac{\partial^2 \log f(x)}{\partial x^2}>0$ for all $x>x_0$. It refers the inflexion point if either $\frac{\partial^2 \log f(x)}{\partial x^2}>0$ for all $x\neq x_0$ or $\frac{\partial^2 \log f(x)}{\partial x^2}<0$ for all $x\neq x_0$.

The graph of $f(x)$ for different values of the parameters is given in Figure 1.

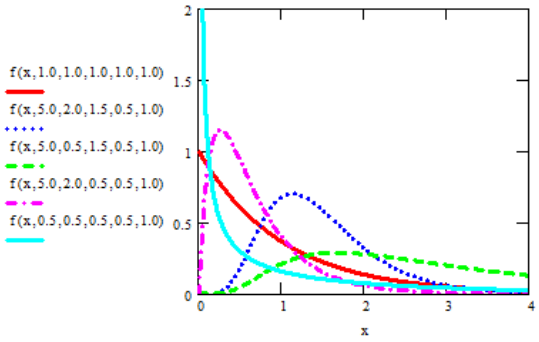


Figure 1: Probability density function of Kw-DML-E distribution for different parameter values.

Some distributions that arises as special case of the Kw-DML-E$(a, b, \alpha, p, \lambda)$ distribution are given below.

Case I: $\alpha = 1$

$$f(x; a, b, 1, p, \lambda) = \frac{abp\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{a-1}}{[p + (1 - p)(1 - e^{-\lambda x})]^{a+1}} \left\{1 - \left[\frac{(1 - e^{-\lambda x})}{[p + (1 - p)(1 - e^{-\lambda x})]}\right]^a\right\}^{b-1}.$$

This is Kumaraswamy Marshall-Olkin exponential distribution studied in George and Thobias (2018).

Case II: $p = 1$

$$f(x; a, b, \alpha, 1, \lambda) = ab\alpha\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{a\alpha-1} \left\{1 - (1 - e^{-\lambda x})^{a\alpha}\right\}^{b-1}.$$

This is Kumaraswamy generalized exponential distribution studied in Mohammed (2014).

Case III: $\alpha = 1,\ p = 1$

$$f(x; a, b, 1, 1, \lambda) = ab\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{a-1}[1 - (1 - e^{-\lambda x})^a]^{b-1}.$$

This is Kumaraswamy exponential distribution.

Case IV: $a = 1,\ b = 1$

$$f(x; 1, 1, \alpha, p, \lambda) = \frac{\alpha p\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{\alpha-1}}{[p + (1 - p)(1 - e^{-\lambda x})^\alpha]^2}.$$

This is Marshall-Olkin generalized exponential distribution studied in Ristić and Kundu (2015).

Case V: $a = 1,\ b = 1,\ \alpha = 1$

$$f(x; 1, 1, 1, p, \lambda) = \frac{p\lambda e^{-\lambda x}}{[p + (1 - p)(1 - e^{-\lambda x})]^2},$$

which is the Marshall-Olkin exponential distribution studied in Marshall and Olkin (1997).

Case VI: $a = 1,\ b = 1,\ p = 1$

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

which is the generalized exponential distribution introduced and studied by Gupta and Kundu (1999).

Case VI: $a = 1,\ b = 1,\ \alpha = 1,\ p = 1$

$$f(x; 1, 1, 1, 1, \lambda) = \lambda e^{-\lambda x},$$

which is the exponential distribution.

So Kw-DML-E$(a, b, \alpha, p, \lambda)$ distribution is a rich family of distributions that contains many existing distributions.

4.3. Hazard rate

The hazard rate is given by

$$h(x) = \frac{ab\alpha p\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]\{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a}-(1-e^{-\lambda x})^{a\alpha}\}} \tag{18}$$

The critical point of $h(x)$ is the roots of the equation $\frac{\partial \log h(x)}{\partial x} = 0$, which yields

$$\begin{aligned} 0 \;=\; & -\lambda + \frac{(a\alpha-1)\lambda e^{-\lambda x}}{(1-e^{-\lambda x})} - \frac{\alpha\lambda(1-p)e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}} \\ & - \frac{a\alpha\lambda\left\{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a-1}(1-p)(1-e^{-\lambda x})^{\alpha-1}-(1-e^{-\lambda x})^{a\alpha-1}\right\}}{\{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a}-(1-e^{-\lambda x})^{a\alpha}\}}. \end{aligned} \tag{19}$$

There may be more than one roots for (19). Let $\tau(x) = \frac{\partial^2 log[h(x)]}{\partial x^2}$. We have

$$\begin{aligned} \tau(x) = \; & -\frac{(a\alpha-1)\lambda^2 e^{-\lambda x}}{(1-e^{-\lambda x})^2} + \frac{\alpha\lambda^2(1-p)e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]} - \\ & \frac{\alpha(\alpha-1)\lambda^2(1-p)e^{-2\lambda x}(1-e^{-\lambda x})^{\alpha-2}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]} + \frac{\alpha^2\lambda^2(1-p)^2e^{-2\lambda x}(1-e^{-\lambda x})^{2\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^2} - \\ & \frac{a\alpha\lambda}{A}\{(a-1)(1-p)^2e^{-\lambda x}(1-e^{-\lambda x})^{2(\alpha-1)} \quad [p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a-2} + \\ & (\alpha-1)(1-p)\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-2}[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a-1} - \lambda(a\alpha-1)e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-2}\} - \\ & \frac{a\alpha\lambda}{A^2}\{(1-p)(1-e^{-\lambda x})^{\alpha-1}[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a-1} - \quad (1-e^{-\lambda x})^{a\alpha+1}\} \\ & \{a(1-p)\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}[p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a-1} - a\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a\alpha-1}\}. \end{aligned}$$

where $A = [p+(1-p)(1-e^{-\lambda x})^{\alpha}]^{a} - (1-e^{-\lambda x})^{a\alpha}$.
If $x = x_0$ is a root of (19), then it refers to a local maximum if $\tau(x) > 0$ for all $x < x_0$ and $\tau(x) < 0$ for all $x > x_0$. It corresponds to a local minimum if $\tau(x) < 0$ for all $x < x_0$ and $\tau(x) > 0$ for all $x > x_0$. It gives an inflexion point if either $\tau(x) > 0$ for all $x \neq x_0$ or $\tau(x) < 0$ for all $x \neq x_0$.
The graph of $h(x)$ for different values of the parameters are given in Figure 2.

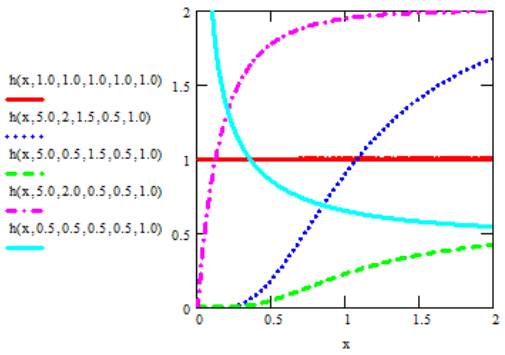


Figure 2: Hazard rate function of Kw-DML-E distribution for different parameter values.

5. General properties of the Kw-DML-E distribution

5.1. Moments and generating function

Let $X \sim$ Kw-DML-E$(a, b, \alpha, p, \lambda)$ for $r = 1,\ 2,\ 3, \ldots$, the rth moment is given by

$$\begin{aligned} E(X^r) &= \int_0^\infty x^r f(x)dx \\ &= \frac{\eta_1\eta_2\ r!}{[\lambda(1+m)]^{r+1}} \end{aligned}$$

In particular, the mean of Kw-DML-E$(a, b, \alpha, p, \lambda)$ is

$$\mu = \frac{\eta_1\eta_2}{[\lambda(1+m)]^2}.$$

The moment generating function is given by

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \frac{\eta_1\eta_2}{[\lambda(1+m)-t]}. \end{aligned}$$

5.2. Simulation and Quantiles

The Kw-DML-E distribution is easily simulated by inverting the cdf. Let U has a unform $U(0, 1)$ distribution, then

$$1 - \left\{1 - \left[\frac{(1-e^{-\lambda x})^\alpha}{1-(1-p)(1-(1-e^{-\lambda x})^\alpha)}\right]^a\right\}^b = U$$

which yields

$$X = \frac{-1}{\lambda}\log\left\{1 - \left[\frac{p[1-(1-U)^{\frac{1}{b}}]^{\frac{1}{a}}}{1-(1-p)[1-(1-U)^{\frac{1}{b}}]^{\frac{1}{a}}}\right]^{\frac{1}{\alpha}}\right\}. \tag{20}$$

In addition, the q^{th} quantile x_q of Kw-DML-E distribution is given by

$$x_q = \frac{-1}{\lambda}\log\left\{1 - \left[\frac{p[1-(1-q)^{\frac{1}{b}}]^{\frac{1}{a}}}{1-(1-p)[1-(1-q)^{\frac{1}{b}}]^{\frac{1}{a}}}\right]^{\frac{1}{\alpha}}\right\}. \tag{21}$$

$0 < q < 1$.

In particular, the median of Kw-DML-E distribution is given by

$$\text{Median} = \frac{-1}{\lambda}\log\left\{1 - \left[\frac{\mathrm{p}[1-(0.5)^{\frac{1}{\mathrm{b}}}]^{\frac{1}{\mathrm{a}}}}{1-(1-\mathrm{p})[1-(0.5)^{\frac{1}{\mathrm{b}}}]^{\frac{1}{\mathrm{a}}}}\right]^{\frac{1}{\alpha}}\right\}. \tag{22}$$

The shape of the parameters a, b, α, p, λ on the skewness and kurtosis can be based on quantile measures. There are many heavy distributions for which these measures are infinite. So, it becomes uninformative precisely when it needs to be. The Bowley's skewness is based on quantiles:

$$S = \frac{Q(3/4) + Q(1/4) - 2Q(1/2)}{Q(3/4) - Q(1/4)},$$

and the Moors' kurtosis is based on octiles:

$$K = \frac{Q(7/8) - Q(5/8) + Q(3/8) - Q(1/8)}{Q(6/8) - Q(2/8)},$$

where $Q(.)$ represents the quantile function of X. These measures are less sensitive to outliers and they exist even for distributions without moments. Skewness measures the degree of the long tail and kurtosis is a measure of the degree of peakedness. When the distribution is symmetric, $S = 0$ and when the distribution is left(or right) skewed, $S < 0$ (or $S > 0$). As K increases, the tail of the distribution becomes heavier. We compute mean, median, variance, skewness and kurtosis numerically using **R** software and presented in Table 2.

		Mean	Median	Variance	Skewness	Kurtosis
	$a = 0.5$	0.0401	0.0020	0.0167	7.8957	99.3075
	$a = 1.0$	0.1137	0.0299	0.0505	4.6048	33.4564
$b = 2.0$	$a = 1.5$	0.1939	0.0834	0.0879	3.4870	18.9717
$\alpha = 0.5$	$a = 2.0$	0.2783	0.1481	0.1213	2.9885	13.6498
	$a = 5.0$	0.6723	0.5321	0.2841	1.7539	4.7844
	$a = 10.0$	1.1198	0.9904	0.4165	1.2943	2.6842
	$b = 1.0$	0.6588	0.3554	0.5935	2.6200	10.1723
	$b = 1.5$	0.3944	0.2120	0.2568	2.8013	11.9250
$a = 2.0$	$b = 2.0$	0.2734	0.1481	0.1242	2.9017	13.0970
$\alpha = 0.5$	$b = 5.0$	0.0802	0.0493	0.0126	2.6689	12.6204
	$b = 10.0$	0.0373	0.0223	0.0020	2.7028	8.1505
	$\alpha = 0.1$	0.0207	0.0000	0.0098	11.0613	188.5540
	$\alpha = 0.2$	0.0701	0.0071	0.0336	5.9125	53.8973
$a = 2.0$	$\alpha = 0.5$	0.2734	0.1481	0.1242	2.9017	13.0970
$b = 2.0$	$\alpha = 0.7$	0.4099	0.2778	0.1764	2.3332	8.5485
	$\alpha = 1.0$	0.5950	0.4636	0.2368	1.9061	5.7999
	$\alpha = 2.0$	1.0600	0.9393	0.3464	1.4039	3.3196

Table 2: Mean, Median, Variance, Skewness and Kurtosis of Kw-DML-E distribution for some parameter values when $p = 0.5$ and $\lambda = 1.0$.

From Table 2, we can see that Kw-DML-E distribution is positively skewed and under dispersed. Also the distribution is leptokurtic. When a and α is increasing, mean,median and variance are increasing while when b is increasing, mean, median and variance are decreasing.

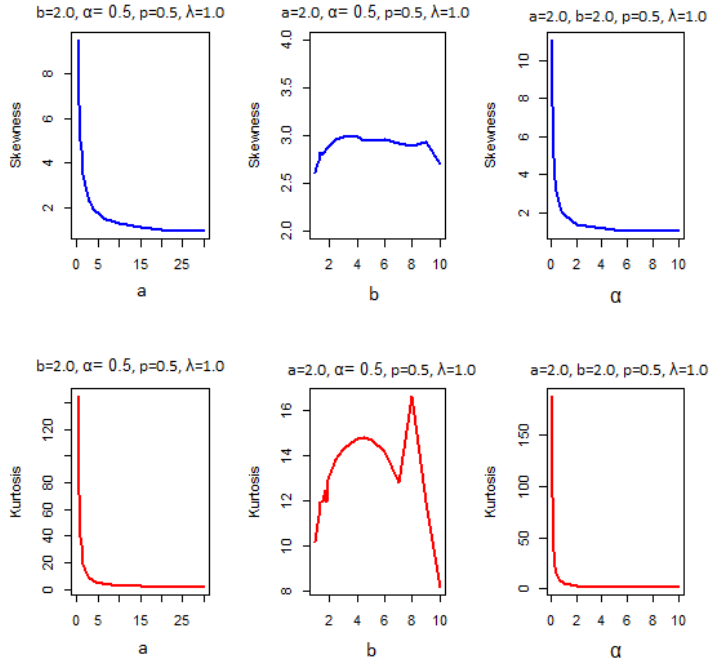


Figure 3: Skewness and kurtosis plots of the Kw-DML-E distribution for some parameter values.

Figure 3 also depicts plots for the skewness and kurtosis coefficients related to additional parameters. In the figure, a and α behaves alike as decreasing while b fluctuate in an interval. These plots indicate that both measures can be very sensitive on these shape parameters. Thus, indicating the importance of the proposed distribution.

5.3. Unimodality

The pdf of the Kw-DML-E model is either decreasing or unimodal. In order to investigate the critical points of density function, its first derivative with respect to x is

$$f'(x) = \frac{ab\alpha p\lambda^2(1-e^{-\lambda x})^{a\alpha-1}}{[p+(1-p)(1-e^{-\lambda x})\alpha]^{a+1}} D^{b-1}(x)\left[1-\frac{e^{-\lambda x}}{1-e^{-\lambda x}}+\frac{(a+1)\alpha(1-p)e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-2}}{[p+(1-p)(1-e^{-\lambda x})\alpha]}\right]+ \frac{a^2b(b-1)\alpha^2p\lambda^2e^{-2\lambda x}(1-e^{-\lambda x})^{\alpha(2a+1)-2}}{[p+(1-p)(1-e^{-\lambda x})\alpha]^{2(a+1)}} D^{b-2}(x)=0, \quad (23)$$

where $D(x)=1-\left[\frac{(1-e^{-\lambda x})^{\alpha}}{p+(1-p)(1-e^{-\lambda x})^{\alpha}}\right]^a$.

There may be more than one root to (23). If $x=x_0$ is a root of (23), then it corresponds to a local maximum if $f'(x)>0$ for all $x<x_0$ and $f'(x)<0$ for all $x<x_0$. It corresponds to a local minimum if $f'(x)<0$ for all $x<x_0$ and $f'(x)>0$ for all $x>x_0$. It corresponds to a point of inflexion if either $f'(x)>0$ for all $x\neq x_0$ or $f'(x)<0$ for all $x\neq x_0$.

5.4. Stochastic ordering

Stochastic orders have been used during the last forty years, at an accelerated rate, in many diverse areas of probability and statistics. Such areas include reliability theory, survival analysis, queueing theory, biology, economics, insurance and actuarial science (see, Shaked and Shanthikumar (2007)). Let X and Y be two random variables having distribution functions F and G respectively, and denote by $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ their respective survival functions, with corresponding pdf's f, g. The random variable X is said to be smaller than Y in the:
(i) stochastic order (denoted as $X \leq_{st} Y$) if $\bar{F}(x) \leq \bar{G}(x)$ for all x;
(ii)likelihood ratio order (denoted as $X \leq_{lr} Y$) if f(x)/g(x) is decreasing in $x \geq 0$;
(iii) hazard rate order (denoted as $X \leq_{hr} Y$)if $\bar{F}(x)/\bar{G}(x)$ is decreasing in $x \geq 0$;
(iv) reversed hazard rate order (denoted as $X \leq_{rhr} Y$) if F(x)/G(x) is decreasing in $x \geq 0$. The four stochastic orders defined above are related to each other, as the following implications (see, Shaked and Shanthikumar (2007)):

$$X \leq_{rhr} Y \Leftarrow X \leq_{lr} Y \Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{st} Y. \tag{24}$$

Let $X \sim$ Kw-DML-E$(a, b_1, \alpha, p, \lambda)$ and $Y \sim$ Kw-DML-E$(a, b_2, \alpha, p, \lambda)$. If $b_2 < b_1$, then

$$\frac{f_X(x)}{f_Y(x)} = \frac{b_1 \left\{1 - \left[\frac{(1-e^{-\lambda x})^\alpha}{p+(1-p)(1-e^{-\lambda x})^\alpha}\right]^a\right\}^{b_1-1}}{b_2 \left\{1 - \left[\frac{(1-e^{-\lambda x})^\alpha}{p+(1-p)(1-e^{-\lambda x})^\alpha}\right]^a\right\}^{b_2-1}}$$

Since $b_2 < b_1$,

$$\begin{aligned}\frac{d}{dx}\left[\frac{f_Y(x)}{f_X(x)}\right] &= \frac{b_1}{b_2}\frac{a(b_2-b_1)\alpha p\lambda e^{-\lambda x}}{(1-e^{-\lambda x})[p+(1-p)(1-e^{-\lambda x})^\alpha]}\left[\frac{(1-e^{-\lambda x})^\alpha}{p+(1-p)(1-e^{-\lambda x})^\alpha}\right]^a \\ &\quad \left\{1 - \left[\frac{(1-e^{-\lambda x})^\alpha}{p+(1-p)(1-e^{-\lambda x})^\alpha}\right]^a\right\}^{b_1-b_2-1} \\ &< 0\end{aligned}$$

Hence $f_X(x)/f_Y(x)$ is decreasing in x. That is $X \leq_{rhr} Y$. The remaining statements follow from the implication (24).

6. Characterization of Kw-DML-E distribution

Characterization of distributions are important to many researchers in the applied field. An investigator will be vitally interested to know if their model fits the requirements of a particular distribution. To this end, one will depend on the characterizations of this distribution which provide conditions under which a given model has the particular distribution. Various characterizations of distributions have been established in many different directions in the literature (see Gupta (2009)).

6.1. Characterization based on truncated moments

Here the characterization results will employ the result due to Glanzel(1987). The advantage of the characterizations given here is that cdf F need not have a closed form and are given in terms of an integrand depends on the solution of a first order differential equation, which can serve as a

bridge between probability and differential equation. First we state the characterization theorem due to Glanzel(1987).

Theorem 6.1 : *Let $(\Omega, \mathcal{F}, \boldsymbol{P})$ be a probability space and let H=[a,b] be an interval for some $a < b$ ($a = -\infty,\ b = \infty$ might as well be defined). Let $X : \Omega \to H$ be a continuous random variable with the distribution function F and let h and g be two real functions defined on H such that*

$$\frac{E[g(X)|X \geq x]}{E[h(X)|X \geq x]} = \eta(x), \quad x \in H,$$

is defined with some real function η. Assume that $h,\ g \in C^1(H),\ \eta \in C^2(H)$ and F is twice continuously differentiable and strictly monotone function on the set H. Finally, assume that the equation $\eta h = g$ has no real solution in the interior of H. Then F is uniquely determined by the functions $h,\ g$ and η, particularly

$$F(x) = \int_a^x C \left| \frac{\eta'(u)}{\eta(u)h(u) - g(u)} \right| e^{-s(u)} du,$$

where the function s is a solution of the differential equation $s' = \frac{\eta' h}{\eta h - g}$ and C is a constant, chosen to make $\int_H dF = 1$.

The stability property of Theorem 6.1 results to special task in statistical practice such as the estimation of the parameters of discrete distributions. Since the function triplet is not uniquely determined, it is often possible to chosen η as a linear function. In some cases, we can take $h(x) = 1$, which reduce the condition of Theorem 6.1 to $E[g(X)|X \geq x] = \eta(x),\ x \in H$. However, adding an extra function will give much more flexibility, when its application is concerned.

Proposition 6.1: Let $X : \Omega \to (0, \infty)$ be a continuous random variable and let $g(x) = \left\{1 - \left[\frac{(1-e^{-\lambda x})^\alpha}{1-(1-p)[1-(1-e^{-\lambda x})^\alpha]}\right]^a\right\}^{1-b} [(1 - e^{-\lambda x})^\alpha]^{1-a}$ and $h(x) = g(x)[1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^a$ for $x > 0$. The random variable X has pdf (15) if and only if the function η defined in Theorem 6.1 has the form

$$\eta(x) = \frac{1}{2}\left\{1 + [1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^{-a}\right\}, \quad x > 0.$$

Proof: Let X have density (15), then

$$(1 - F(x)E[g(X)|X \geq x] = \frac{b(1-p)}{p}\left\{[1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^{-a} - 1\right\},\ x > 0,$$

and

$$(1 - F(x)E[h(X)|X \geq x] = \frac{b(1-p)}{2p}\left\{[1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^{-2a} - 1\right\},\ x > 0.$$

Then

$$\eta(x) = \frac{1}{2}\left\{\frac{1 + [1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^a}{[1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^a}\right\},$$

and finally

$$\eta(x)g(x) - h(x) = \frac{1}{2}g(x)\left\{1 - [1 - (1 - p)(1 - (1 - e^{-\lambda x})^\alpha)]^{-a}\right\} < 0,\ x > 0.$$

Conversely, if $\eta(x) = \frac{1}{2}\left\{1 + [1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-a}\right\}$, then

$$\begin{aligned} s'(x) &= \frac{\eta'(x)g(x)}{\eta(x)g(x) - h(x)} \\ &= \frac{-\,a(1-p)\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-(a+1)}}{1-[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-a}},\ x>0 \end{aligned}$$

and

$$s(x) = -\log\left\{1-[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-a}\right\}, \quad x>0.$$

So, in view of Theorem 6.1, X has density function (15).

Corollary 6.1: Let $X : \Omega \to (0,\infty)$ be a continuous random variable and let h be as in Proposition 1. Then pdf of X is (15) if and only if there exist a function g and η defined in Theorem 6.1 satisfying the differential equation

$$\frac{\eta'(x)h(x)}{\eta(x)h(x)-g(x)} = \frac{-\,a(1-p)\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-(a+1)}}{1-[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-a}},\ x>0.$$

The general solution of the differential equation in Corollary 6.1 is

$$\eta(x) = \frac{\int a(1-p)\alpha\lambda e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-(a+1)}[g(x)]^{-(a+1)}h(x)dx + D}{1-[1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})]^{-a}},$$

where D is a constant. One set of appropriate functions is given in Proposition 1 with $D = \frac{1}{2}$. However, it should be also noted that there are other triplets (h, g, η) satisfying the conditions of Theorem 6.1.

6.2. Characterization based on single function of the random variable

In this subsection, we obtain a characterization of Kw-DML-E distribution using the following theorem of Hamedani et al. (2014).

Theorem 6.2 : $1 - F(x) = [a\phi(x) + b]^c$ *if and only if*

$$E[\phi(X)|X \geq x] = \frac{1}{c+1}\left[c\phi(x) - \frac{b}{a}\right],\ \alpha < x < \beta,$$

where $a \neq 0, b, c > 0$ *are finite constants.*

Corollary 6.2: By taking $a = -1$, $b = 1$, $c = b$, $\phi(x) = \left[\frac{(1-e^{-\lambda x})^{\alpha}}{1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})}\right]^a$ and $(\alpha, \beta) = (0,\infty)$ in Theorem 6.2, we get

$$F(x) = 1 - \left\{1 - \left[\frac{(1-e^{-\lambda x})^{\alpha}}{1-(1-p)(1-(1-e^{-\lambda x})^{\alpha})}\right]^a\right\}^b,\ x>0.$$

7. Maximum likelihood estimation

Several approaches for parameter estimation have been proposed in the literature, but maximum likelihood method is the most commonly employed. We consider estimation of the unknown parameters of Kw-DML-E distribution by the method of maximum likelihood. Let $x_1,\ x_2, \ldots,\ x_n$

be observed values from the Kw-DML-E distribution with parameters a, b, α, p and λ. The log-likelihood function for $(a, b, \alpha, p, \lambda)$ is given by

$$\begin{aligned}\log\ L &= n\ \log(a\ b\ \alpha\ p\ \lambda) - \sum_{i=1}^{n}\lambda x_i + (a\alpha-1)\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) + \\ &\quad (b-1)\sum_{i=1}^{n}\left\{1-\left[\frac{1-e^{-\lambda x_i}}{1-(1-p)(1-(1-e^{-\lambda x_i})^{\alpha})}\right]^{a}\right\} - \\ &\quad (a+1)\sum_{i=1}^{n}\log[1-(1-p)(1-(1-e^{-\lambda x_i})^{\alpha})].\end{aligned}$$

The derivatives of the log-likelihood function with respect to the parameters a, b, α, p and λ are given by respectively,

$$\frac{\partial \log\ L}{\partial a} = \frac{n}{a} + \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - (b-1)\sum_{i=1}^{n}\left(\frac{B}{A}\right)^{a}\left[\frac{\log(B/A)}{1-(B/A)^{a}}\right] - \sum_{i=1}^{n}\log(A), \tag{25}$$

$$\frac{\partial \log\ L}{\partial b} = \frac{n}{b} + \sum_{i=1}^{n}\log\left[1-\left(\frac{B}{A}\right)^{a}\right], \tag{26}$$

$$\begin{aligned}\frac{\partial \log\ L}{\partial \alpha} &= \frac{n}{\alpha} + a\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - (b-1)\sum_{i=1}^{n}\left[\frac{B^{a}}{A^{a+1}}\right]\frac{p\log(1-e^{-\lambda x_i})}{[1-(B/A)^{a}]} - \\ &\quad \sum_{i=1}^{n}\frac{(a+1)(1-p)B\log(1-e^{-\lambda x_i})}{A}\end{aligned} \tag{27}$$

$$\frac{\partial \log\ L}{\partial p} = \frac{n}{p} + a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{A^{2}[1-(B/A)^{a}]} - \sum_{i=1}^{n}\frac{(a+1)(1-B)}{A}, \tag{28}$$

$$\begin{aligned}\frac{\partial \log\ L}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n}x_i + (a\alpha-1)\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}} - a(b-1)\alpha p\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}(1-e^{-\lambda x_i})^{\alpha-1}}{A^{a+1}[1-(B/A)^{a}]} - \\ &\quad (a+1)\alpha(1-p)\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}(1-e^{-\lambda x_i})^{\alpha-1}}{A},\end{aligned} \tag{29}$$

where $A = 1-(1-p)[1-(1-e^{-\lambda x_i})^{\alpha}]$ and $B = (1-e^{-\lambda x_i})^{\alpha}$.
The maximum likelihood estimates of $(a, b, \alpha, p, \lambda)$, say $(\hat{a}, \hat{b}, \hat{\alpha}, \hat{p}, \hat{\lambda})$ are the simultaneous solutions of the equation $\frac{\partial \log\ L}{\partial a} = 0$, $\frac{\partial \log\ L}{\partial b} = 0$, $\frac{\partial \log\ L}{\partial \alpha} = 0$, $\frac{\partial \log\ L}{\partial p} = 0$ and $\frac{\partial \log\ L}{\partial \lambda} = 0$. Maximization of the likelihood function can be performed by using *nlm* or *optim* in **R** statistical package.

Now, we study the existence and uniqueness of the maximum likelihood estimates when the other parameters are known (given).

Theorem 7.1: *Let $f_1(a; b, \alpha, p, \lambda, x)$ denote the function on the right-hand-side (RHS) of equation (25), where b, α, p, λ are the true value of the parameters. Then there exists a solution for $f_1(a; b, \alpha, p, \lambda, x)$=0, for $\sum_{i=1}^{n}\log(A) > \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})$ and is unique when $b < 1$.*

Proof: We have

$$f_1(a; b, \alpha, p, \lambda, x) = \frac{n}{a} + \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - (b-1)\sum_{i=1}^{n}\left(\frac{B}{A}\right)^{a}\left[\frac{\log(B/A)}{1-(B/A)^{a}}\right] - \sum_{i=1}^{n}\log(A).$$

Then

$$\lim_{a\to 0} f_1(a;b,\alpha,p,\lambda,x) = \infty + \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - 0 - \sum_{i=1}^{n}\log(A) = \infty$$

Also

$$\lim_{a\to \infty} f_1(a;b,\alpha,p,\lambda,x) = 0 + \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - 0 - \sum_{i=1}^{n}\log(A) < 0$$

if and only if $\sum_{i=1}^{n}\log(A) > \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})$. Hence there exist at least one solution for $a \in (0,\infty)$ when $\sum_{i=1}^{n}\log(A) > \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})$.
To establish uniqueness part, the first derivative of $f_1(a;b,\alpha,p,\lambda,x)$ is

$$\begin{aligned}\frac{\partial f_1(a;b,\alpha,p,\lambda,x)}{\partial a} &= -(b-1)\left(\frac{B}{A}\right)^a\left[\frac{\log(B/A)^2}{1-(B/A)^a}\right] - (b-1)\left(\frac{B}{A}\right)^{2a}\left[\frac{\log(B/A)^2}{[1-(B/A)^a]^2}\right] \\ &< 0\end{aligned}$$

when $b<1$. So there exist a solution for $f_1(a;b,\alpha,p,\lambda,x)=0$ when $\sum_{i=1}^{n}\log(A) > \alpha\sum_{i=1}^{n}\log(1-e^{-\lambda x_i})$ and the root $\hat{a}$ is unique when $b<1$.

Theorem 7.2: *Let $f_2(b;a,\alpha,p,\lambda,x)$ denote the function on RHS of equation (26), where a,α,p,λ are the true value of the parameters. Then there exist a unique solution for $f_2(b;a,\alpha,p,\lambda,x) = 0$ for $b\in(0,\infty)$.*

Proof: We have

$$f_2(b;a,\alpha,p,\lambda,x) = \frac{n}{b} + \sum_{i=1}^{n}\log\left[1-\left(\frac{B}{A}\right)^a\right].$$

Now

$$\lim_{b\to 0} f_2(b;a,\alpha,p,\lambda,x) = \infty + \sum_{i=1}^{n}\log\left[1-\left(\frac{B}{A}\right)^a\right] = \infty.$$

On the other hand

$$\lim_{b\to \infty} f_2(b;a,\alpha,p,\lambda,x) = 0 + \sum_{i=1}^{n}\log\left[1-\left(\frac{B}{A}\right)^a\right] < 0.$$

Therefore, there exist at least one root say $\hat{b}\in(0,\infty)$ such that $f_2(\hat{b};a,\alpha,p,\lambda,x)=0$. To show the uniqueness part, the first derivative of $f_2(b;a,\alpha,p,\lambda,x)$ is

$$\begin{aligned}\frac{\partial f_2(b;a,\alpha,p,\lambda,x)}{\partial b} &= -\frac{n}{b^2} \\ &< 0.\end{aligned}$$

Hence, there exist a solution for $f_2(b;a,\alpha,p,\lambda,x)=0$ and the root, $\hat{b}$ is unique.

Theorem 7.3: *Let $f_3(\alpha;a,b,p,\lambda,x)$ denote the function on RHS of equation (27), where a,b,p,λ are the true value of the parameters. Then there exist a solution for $f_3(\alpha;a,b,p,\lambda,x)=0$ for $\alpha\in(0,\infty)$.*

Proof: We have

$$\begin{aligned}f_3(\alpha;a,b,p,\lambda,x) &= \frac{n}{\alpha} + a\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - (b-1)\sum_{i=1}^{n}\left[\frac{B^a}{A^{a+1}}\right]\frac{p\log(1-e^{-\lambda x_i})}{[1-(B/A)^a]} - \\ &\quad \sum_{i=1}^{n}\frac{(a+1)(1-p)B\log(1-e^{-\lambda x_i})}{A}.\end{aligned}$$

Then

$$\lim_{\alpha\to 0} f_3(\alpha;a,b,p,\lambda,x) = \infty + a\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - (b-1)\sum_{i=1}^{n} p\log(1-e^{-\lambda x_i}) - (a+1)(1-p)\log(1-e^{-\lambda x_i}) = \infty.$$

Also

$$\lim_{\alpha\to\infty} f_3(\alpha;a,b,p,\lambda,x) = 0 + a\sum_{i=1}^{n}\log(1-e^{-\lambda x_i}) - 0 - 0 - 0 < 0.$$

Therefore, there exist at least one root say $\hat{\alpha}\in(0,\infty)$ such that $f_3(\hat{\alpha};a,b,p,\lambda,x)=0$.

Theorem 7.4: *Let $f_4(p;a,b,\alpha,\lambda,x)$ denote the function on RHS of equation (28), where a,b,α,λ are the true value of the parameters. Then there exist a solution for $f_4(p;a,b,\alpha,\lambda,x)=0$ when $n+a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{1-B^a} < \sum_{i=1}^{n}(a+1)(1-B)$, $p\in(0,1)$.*

Proof: We have

$$f_4(p;a,b,\alpha,\lambda,x) = \frac{n}{p} + a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{A^2[1-(B/A)^a]} - \sum_{i=1}^{n}\frac{(a+1)(1-B)}{A}.$$

Then

$$\lim_{p\to 0} f_4(p;a,b,\alpha,\lambda,x) = \infty + \infty - \sum_{i=1}^{n}\frac{(a+1)(1-B)}{(1-e^{-\lambda x})^{\alpha}} = \infty.$$

On the other hand

$$\lim_{p\to 1} f_4(p;a,b,\alpha,\lambda,x) = n + a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{1-B^a} - \sum_{i=1}^{n}(a+1)(1-B) < 0$$

if and only if $n+a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{1-B^a} < \sum_{i=1}^{n}(a+1)(1-B)$.
Hence, there exist a solution for $f_4(p;a,b,\alpha,\lambda,x)=0$ when $n+a(b-1)\sum_{i=1}^{n}\frac{B(1-B)}{1-B^a} < \sum_{i=1}^{n}(a+1)(1-B)$.

Theorem 7.5: *Let $f_5(\lambda;a,b,\alpha,p,x)$ denote the function on RHS of equation (29), where a,b,α,p are the true value of the parameters. Then there exist a solution for $f_5(\lambda;a,b,\alpha,p,x)=0$, for $\lambda\in(0,\infty)$.*

Proof: We have

$$\begin{aligned} f_5(\lambda;a,b,\alpha,p,x) &= \frac{n}{\lambda} - \sum_{i=1}^{n}x_i + (a\alpha-1)\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}}{1-e^{-\lambda x_i}} - a(b-1)\alpha p\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}(1-e^{-\lambda x_i})^{\alpha-1}}{A^{a+1}[1-(B/A)^a]} - \\ &\quad (a+1)\alpha(1-p)\sum_{i=1}^{n}\frac{x_i e^{-\lambda x_i}(1-e^{-\lambda x_i})^{\alpha-1}}{A}. \end{aligned}$$

Then

$$\lim_{\lambda\to 0} f_5(\lambda;a,b,\alpha,p,x) = \infty - \sum_{i=1}^{n}x_i + \infty - 0 - 0 = \infty.$$

Also

$$\lim_{\lambda\to\infty} f_5(\lambda;a,b,\alpha,p,x) = 0 - \sum_{i=1}^{n}x_i + 0 - 0 - 0 < 0.$$

Therefore there exist at least one root say $\hat{\lambda} \in (0, \infty)$ such that $f_5(\lambda; a, b, \alpha, p, x) = 0$.

The normal approximation of the maximum likelihood estimates of the parameters can be adopted for constructing approximate confidence intervals and for testing hypotheses on the parameters $(a, b, \alpha, p, \lambda)$. Under conditions that are fulfilled for the parameters in the interior of the parameter space and applying the usual large sample approximation, it can be shown that $\sqrt{n}(\theta - \hat{\theta})$ can be approximated by a multivariate normal distribution with zero means and variance-covariance matrix $\mathbf{K}^{-1}(\theta)$, where $\mathbf{K}(\theta)$ is the unit expected information matrix.
As n tends to infinity, we have the asymptotic result

$$\mathbf{K}(\theta) = \lim_{n \to \infty} \frac{1}{n} \mathbf{I}(\theta)$$

where $\mathbf{I}(\theta)$ is the observed Fisher information matrix. Since $\mathbf{K}(\theta)$ involves the unknown parameter of θ, we may replace it with the MLE $\hat{\theta}$. Thus, the average matrix estimated at $\hat{\theta}$, say $\frac{1}{n}\mathbf{I}(\theta)$, can be used to estimate $\mathbf{K}(\theta)$. The estimated multivariate normal distribution can thus be used to construct approximate confidence intervals for the unknown parameters and for the hazard rate and survival function.

7.1. Simulation

In this section, we asses the performance of the Kw-DML-E$(a, b, \alpha, p, \lambda)$ distribution by conducting simulation for different sample sizes and parameter values. For analytical tractability, let $\lambda = 1.0$ in (20) and we use (20) to generate random samples from the Kw-DML-E distribution with parameters a, b, α and p. The different sample sizes considered in the simulation are $n = 50, 100, 200$ and 500. We used *nlm* package in **R** software to find the estimates. We repeated the process 1000 times and report the average estimates and associated mean squared errors listed in Table 3.

		$p = 0.5$			$p = 0.8$	
n	$\hat{a}(\hat{SE}(\hat{a}))$	$\hat{b}(\hat{SE}(\hat{b}))$	$\hat{\alpha}(\hat{SE}(\hat{\alpha}))$	$\hat{a}(\hat{SE}(\hat{a}))$	$\hat{b}(\hat{SE}(\hat{b}))$	$\hat{\alpha}(\hat{SE}(\hat{\alpha}))$
	a=5.0	b=2.0	α=1.5	a=5.0	b=0.5	α=1.5
50	4.872(0.2667)	1.314(6.4290)	1.73(0.0130)	4.773(0.0518)	0.573(1.0565)	1.84(0.0285)
100	5.228(0.1650)	2.607(1.1612)	1.64(0.0070)	4.784(0.0386)	0.580(0.8413)	1.44(0.0150)
200	4.769(0.0926)	1.503(1.2006)	1.40(0.0034)	5.088(0.0025)	0.432(0.1847)	1.56(0.0078)
500	5.235(0.0516)	2.224(0.7637)	1.43(0.0007)	5.009(0.0001)	0.506(0.0705)	1.50(0.0045)
	a=5.0	b=2.0	α=0.5	a=5.0	b=0.5	α=0.5
50	5.991(0.1086)	1.831(2.3542)	0.77(0.2345)	4.778(0.6018)	0.573(1.0565)	0.46(0.4050)
100	4.834(0.0016)	1.776(1.4201)	0.35(0.1704)	4.784(0.5186)	0.513(0.8413)	0.57(0.3160)
200	5.639(0.0284)	1.813(0.9373)	0.45(0.1025)	5.088(0.3256)	0.480(0.8124)	0.54(0.1098)
500	5.414(0.0016)	1.913(0.5324)	0.53(0.0804)	5.009(0.0984)	0.506(0.5630)	0.48(0.0931)
	a=0.5	b=2.0	α=1.5	a=0.5	b=0.5	α=1.5
50	0.626(0.3036)	2.191(0.6450)	1.40(0.1060)	0.482(1.317)	0.538(0.4186)	1.55(0.5070)
100	0.566(0.1095)	1.936(0.1916)	1.53(0.0840)	0.525(0.9801)	0.498(0.3290)	1.44(0.3018)
200	0.443(0.0854)	2.264(0.0114)	1.45(0.0651)	0.512(0.8456)	0.452(0.1853)	1.57(0.0807)
500	0.506(0.0202)	2.185(0.0029)	1.56(0.0421)	0.486(0.5412)	0.548(0.1290)	1.49(0.0402)
	a=0.5	b=2.0	α=0.5	a=0.5	b=0.5	α=0.5
50	0.528(0.0131)	2.375(0.5540)	0.42(0.8080)	0.483(0.7020)	0.4814(0.1041)	0.75(1.0125)
100	0.534(0.0026)	1.761(0.1432)	0.49(0.7077)	0.515(0.6090)	0.538(0.0913)	0.45(1.0027)
200	0.487(0.0025)	1.329(0.0648)	0.57(0.4024)	0.537(0.4756)	0.514(0.0601)	0.66(0.9012)
500	0.512(0.0004)	2.378(0.0154)	0.55(0.1546)	0.540(0.2351)	0.521(0.0402)	0.45(0.7002)

Table 3: Simulation results for some different values of the parameters a, b, α and p when $\lambda = 1$. From Table 3, we can see that as the sample size increases, the estimated values are close to the assumed values and the mean squared error decreases, which indicates the consistency property of the MLEs.

8. Application to real data

Here we present an application to a real data set for illustrating the potentiality of the new distribution. The data set is originally reported by Bader and Priest (1982) which represents the strength measured in GPa for single carbon fibers and impregnated at gauge lengths of 1, 10, 20 and 50 mm. Here, we consider the data set of single fibers of 20 mm in gauge with a sample of size 63. The data set is:

1.901	2.132	2.203	2.228	2.257	2.350	2.361	2.396	2.397	2.445	2.454	2.474	2.518	2.522
2.525	2.532	2.575	2.614	2.616	2.618	2.624	2.659	2.675	2.738	2.740	2.856	2.917	2.928
2.937	2.937	2.977	2.996	3.030	3.125	3.139	3.145	3.220	3.223	3.235	3.243	3.264	3.272
3.294	3.332	3.346	3.377	3.408	3.435	3.493	3.501	3.537	3.554	3.562	3.628	3.852	3.871
3.886	3.971	4.024	4.027	4.225	4.395	5.020.							

Descriptive statistics of the data is presented in Table 4.

Min	Median	Mean	Max	SD	Skewness	Kurtosis
1.901	2.996	3.059	5.020	0.621	0.633	3.286

Table 4: Descriptive statistics of carbon fibers data.

The distribution of the data is positively skewed and leptokurtic. We compare the fit of the Kw-DML-E distribution with the following continuous life time distributions:

1. Kumaraswamy Exponential (Kw-E) distribution having the pdf

$$f(x) = ab\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a-1}(1-(1-e^{-\lambda x})^{a})^{b-1}; \quad a,b,\lambda > 0.$$

2. Kumaraswamy Marshall-Olkin Exponential (Kw-MO-E) distribution having the pdf

$$f(x) = \frac{ab\lambda e^{-\lambda x}(1-p)(1-e^{-\lambda x})^{a-1}}{(1-pe^{-\lambda x})^{a+1}}\left\{1-\left[\frac{1-e^{-\lambda x}}{1-pe^{-\lambda x}}\right]^{a}\right\}^{b-1}; a,b,p,\lambda > 0.$$

3. Marshall-Olkin Kumaraswamy Exponential (MO-Kw-E) distribution having the pdf

$$f(x) = \frac{abp\lambda e^{-\lambda x}(1-e^{-\lambda x})^{a-1}[1-(1-e^{-\lambda x})^{a}]^{b-1}}{\{1-(1-p)[1-(1-e^{-\lambda x})^{a}]^{b}\}^{2}}; a,b,p,\lambda > 0.$$

The values of the log-likelihood (-log L) and AIC (Akaike Information Criterion) are calculated for the eight distributions in order to verify which distribution fits better to the data. The better distribution corresponds to smaller -log L and AIC. Here, $AIC = -2\log L + 2k$ where L is the likelihood function evaluated at the maximum likelihood estimates and k is the number of parameters.

We apply the Crammer-von Mises(W^*) and Anderson-Darling (A^*) statistic for formal goodness-of-fit to verify which distribution fits better to this data. In general, the smaller the values of the statistics W^* and A^*, shows better the fit to the data. Let $G(x;\theta)$ be the cdf, where the form of G is known but θ (a 4-dimensional parameter vector, say) is unknown. We calculate the statistics W^* and A^* as follows:(i) Compute $\psi_i = G(x_i;\hat{\theta})$, where the $x_i's$ are in ascending order; (ii) Compute $x_i = \phi^{-1}(\psi_i)$, where $\phi(.)$ is the normal cdf and $\phi^{-1}(.)$ its inverse.; (iii) Compute $u_i = \phi\{(x_i-\bar{x})/s_x\}$, where $\bar{y} = n^{-1}\sum_{i=1}^{n} x_i$ and $s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{(x)})^2$; (iv) Calculate

$$W^2 = \sum_{i=1}^{n}\left\{u_i - \frac{(2i-1)}{2n}\right\}^2 + \frac{1}{12n}$$

and

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\{(2i-1)\log(u_i) + (2n+1-2i)\log(1-u_i)\};$$

(v) Modify W^2 into $W^* = W^2(1+0.5/n)$ and A^2 into $A^* = A^2(1+0.75/n+2.25/n^2)$. For further details, see Chen and Balakrishnan (1995).

The values of estimates,$-\log L$, AIC, W^* and A^* for all models are listed in Table 5.

Model	Parameters	$-\log L$	AIC	W^*	A^*	$K-S$	p-value
Kw-E	$\hat{b}=1.523$ $\hat{\lambda}=1.540$	56.3546	118.7092	0.0651	0.3410	0.0815	0.7962
Kw-MO-E	$\hat{a}=69.755$ $\hat{b}=1.497$ $\hat{p}=0.474$ $\hat{\lambda}=1.560$	56.3624	120.7248	0.0658	0.3433	0.0828	0.7803
MO-Kw-E	$\hat{a}=5181.418$ $\hat{b}=0.585$ $\hat{p}=0.488$ $\hat{\lambda}=3.830$	56.2235	120.4470	0.0648	0.3425	0.0813	0.7985
Kw-DML-E	$\hat{a}=47.785$ $\hat{b}=2.833$ $\hat{\alpha}=0.5711$ $\hat{p}=0.069$ $\hat{\lambda}=0.919$	54.2107	118.4214	0.0610	0.3218	0.0782	0.8348

Table 5: Parameter estimates and goodness-of-fit statistics for various models fitted to carbon fibers data.

From the Table 5, we can see that, Kw-DML-E distribution has smallest $-\log L$, AIC, W^*, A^* and $K-S$ values. Also Kw-DML-E distribution has highest p-value. Hence, the new model, that is Kw-DML-E distribution, yields a better fit than the other models, for this data set.

To test the null hypothesis H_0 : Kw-DML-E versus H_1 : Kw-MO-E or equivalently $H_0 : \alpha = 1$ versus $H_1 : \alpha \neq 1$, we use likelihood ratio test statistic whose value is 2.1517(p-value =0.1424). As a result, the null model Kw-MO-E is rejected in favor of alternative model Kw-DML-E at any level > 0.1424.

The fitted density and the empirical cdf plot of the Kw-DML-E distribution model are presented in Figure 4. The figure indicates a satisfactory fit of the Kw-DML-E distribution.

9. Conclusion

In this paper, we have proposed a new class of continuous distributions, namely Kumaraswamy discrete Linnik G family of distributions. The new class of distributions contain Kumaraswamy discret Mittag-Leffler G family of distributions, Kumaraswamy truncated negative binomial G family of distributions, Kumaraswamy Marshall-Olkin G family of distributions, Kumaraswamy G family of distributions, families of distributions generated through truncated discrete Linnik distribution, families of distributions generated through truncated discrete Miitag-Leffler distribution, families of distributions generated through truncated negative binomial distribution, Marshall-Olkin family of distributions, etc. In particular, we study a sub model of Kumaraswamy discrete Mittag-Leffler G distribution, namely, Kumaraswamy discrete Mittag-Leffler exponential (Kw-DML-E) distribution in detail. We study the shape properties of the density function and hazard function. The explicit expression for the moments, generating functions and quantiles are derived. The stochastic ordering property studied. Two characterizations of Kw-DML-E distribution is obtained. The maximum likelihood method is employed for estimating the model parameters and its existence

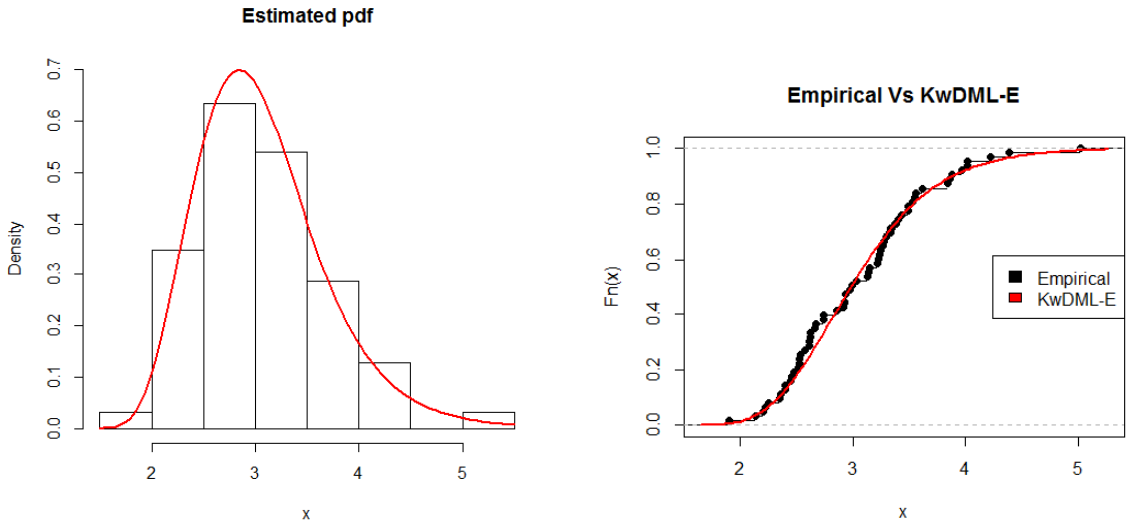


Figure 4: Plots of the estimated pdf and cdf of the Kw-DML-E model for the carbon fibers data.

and uniqueness are proved. Simulation studies are also carried out. The obtained results are validated using a real data set and it is shown that the Kw-DML-E distribution provides a better fit than Kumaraswamy exponential, Kumaraswamy Marshall-Olkin exponential and Marshall Olkin Kumaraswamy exponential distribution. Hence, the proposed model will attract wider applications in several areas such as engineering, hydrology, economics, survival and lifetime data among others.

References

1. Alizadeh, A.,Tahir, M.H., Cordeiro, G.M., Mansoor, M., Zubair, M., Hamedani, G.G. (2015). The Kumaraswamy Marshall-Olkin family of distributions. *Journal of Egyptian Mathematical Society*, 23, 546–557.
2. Arnold, B.C., Robertson, C.A., Yeh, H.C. (1986). Some properties of Pareto type distribution. *Sankhya-A*, 48, 404–408.
3. Badar, M.G., Priest, A.M. (1982). Statistical aspects of fiber and bundle strength in hybrid composites in: Hayashi, T.,Kawata, K., Umekawa, S.(Eds.)*Progress in Science and Engineering Composites*, ICCM-IV, Tokyo, 1982, 1129–1136.
4. Chen, G., Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit. *Journal of Quality Technology*, 27, 154–161.
5. Christoph, G., Schreiber, K. (1998) The generalized Linnik distributions. *Advances in Stochastic Models for Reliability, Quality and Safety*, Birkhuser, Boston, 3–18.
6. Cordeiro, G.M., de Castro, M. (2011). A new family of generalized distribution. *Journal of Statistical Computation and Simulation*, 81, 883–893.
7. Devroye, L. (1993) A triptych of discrete distribution related to the stable law. *Statistics and Probability Letters*, 18, 349–351.
8. Dias, C.R.B., Corderio, G.M., Alizadeh, A., Marinho, P.R.D., Coelho, H.F.C (2016). Exponentiated Marshall Olkin family of distributions. *Journal of Statistical Distributions and Applications*, 3, 1–21

9. Glanzel, W. (1987). A characterization theorem based on truncated moments and its application to some distribution families. *Mathematical Statistics and Probability Theory*, B.D. Reidel Publishing Company, 75–84.
10. George, R., Thobias, S. (2019). Kumaraswamy Marshall-Olkin exponential distribution. *Communications in Statistics - Theory and Methods*, 48, 1920–1937.
11. Gupta, A. K., Nadarajah, S. (2004). Handbook of beta distribution and its applications. *Marcel Dekker*, New York.
12. Gupta, R.C. (2009). Some characterization results based on residual entropy function. *Journal of Statistical Theory and Applications*, 8, 45–59.
13. Gupta, R.D., Kundu, D. (1999). Generalized exponential distribution. *Australian and New Zealand Journal of Statistics*, 41, 173–188.
14. Hamedani, G.G., Javanshiri, Z., Maadooliat, M., Yazdani, A. (2014). Remarks on characterizations of Malinowska and Szynal. *Applied Mathematics and Computation*, 246, 377–388.
15. Huillet, T.E. (2016). On Mittag-Leffler distributions and related stochastic processes. *Journal of Computational and Applied Mathematics*, 296, 181–211.
16. Jayakumar, K., Sreenivas, P.C. (2003). On discrete stable laws. *Advances and Applications in Statistics*, 3, 255-266.
17. Jayakumar, K., Sankaran, K.K. (2019). Discrete Linnik Weibull distribution. *Communications in Statistics- Simulation and Computation*, 48, 3092–3117.
18. Jones, M.C. (2009). Kumaraswamy distribution: A beta type distribution with some tractability advantages. *Statistical Methodology*, 6, 70–81.
19. Kozubowski, T.J., Podgórski. (2018). Kumaraswamy distribution and random extrema. *The Open Statistics and Probability Journal*, 09, 18–25.
20. Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46, 79–86.
21. Marshall, A.W., 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.
22. Mohammed, B.E. (2014). Statistical properties of Kumaraswamy-Generalized exponentiated exponential distribution. *International Journal of Computer Applications*, 94, 1–8.
23. Nadarajah, S., Gupta, A.K. (2007). A compound beta distribution with applications in finance. *Statistical Methods and Applications*, 16, 69–83.
24. Nadarajah, S., Jayakumar, K., Ristić, M.M. (2013). A new family of lifetime models. *Journal of Statistical Computation and Simulation*, 83, 1389–1404.
25. Pillai, R.N. (1990). On Mittag-Leffler functions and related distributions. *Annals of the Institute of Statistical Mathematics*, 42, 157–161.
26. Pillai, R.N., Jayakumar, K. (1995). Discrete Mittag-Leffler distributions. *Statistics and Probability Letters*, 23, 271-274.
27. Ristić, M.M., Kundu, D. (2015). Marshall-Olkin generalized exponential distribution. *Metron*, 73, 317–333.
28. Sankaran, K.K., Jayakumar, K. (2016). A new extended uniform distribution. *International Journal of Statistical Distributions and Applications*, 2, 35–41.
29. Shaked, M., Shanthikumar, J.G. (2007). Stochastic Orders. *Springer*, New York, 2007.
30. Steutel, F.W., van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7, 893–899.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Dual to A Two-Parameter Ratio-Product-Ratio Estimator Using Auxiliary Information in Sample Surveys

Housila P. Singh[1], Priyanka Malviya[2] and Rajesh Tailor[3]

1. *S. S. In Statistics, Vikram University, Ujjain*
2. *S. S. In Statistics, Vikram University, Ujjain*
3. *S. S. In Statistics, Vikram University, Ujjain*

Received: 20 Jan 2022 / Revised: 21 December 2022 / Accepted: 21 July 2023

Abstract

Following the methodology in the studies of Srivenkataramana (1980), Bandyopadhyaya (1980), Ray and Sahai (1980), Singh and Espejo (2003) and Chami, Singh and Thomas (2012) we have suggested a dual to a two-parameter ratio-product-ratio estimator for a finite population mean using simple random sampling without replacement (SRSWOR) scheme. Expressions for bias and mean squared error of the suggested estimator are derived upto the first degree of approximation. Regions of preference have been derived under which the proposed family of estimators is better than the sample mean, ratio, dual to ratio, product and dual to product estimators. We carry out an empirical study demonstrating that the suggested estimator out performs the traditional estimators.

Keywords: Study Variate, Auxiliary Variable, Bias, Mean Squared error, Empirical Study.

1 Introduction

In many survey situations, information on auxiliary variable is always available along with the study variable. It is well known fact that the use of auxiliary information at the estimation stage provides efficient estimators for population mean of the study variable. If the correlation between the study and auxiliary variables is positive (high), the ratio method of estimation is quite effective. On the other hand if this correlation is negative (high), the product method of estimation is employed. Further if the relation between the study variable y and the auxiliary variable x is a straight line passing through the neighbourhood of the origin and the variance of y about this line is proportional to the auxiliary variable x, the ratio estimator is as good as regression estimator. There are number of situations in which the regression line does not pass through the neighbourhood of the origin. In such cases ratio estimator does not perform equally well as that of regression estimator. Keeping this fact in view various authors including Srivastava (1967,1971), Walsh (1970), Reddy (1973), Gupta (1978), Sahai (1979), Srivastava (1980), Adhvaryu and Gupta (1983), Kothawala and Gupta (1988), Gupta and Kothawala (1990), Singh and Nigam (2020) and others have made their efforts to formulate ratio and product estimators in order to provide better alternatives. We further note that in sampling theory the prior knowledge about $C = \rho\frac{C_y}{C_x}$ has played very important role in providing these better alternatives for population mean where ' ρ ' is the correlation coefficient between the study variable y and the auxiliary variable x, C_y and C_x are coefficients of variation of y and x respectively. Use of knowledge of ' C ' was first introduced by Gupta (1978)

while discussing the higher degree ratio and product estimators. Adhvaryu and Gupta (1983) suggested its use in composite estimator where weights depend on ' C ' and have determined the range in which this value differs from optimum still the estimator remains better conventional ratio estimator. Further their results have been extended for the composite product estimator. Kothawala and Gupta in a series papers have again use this parameter ' $\mathrm{C}' = \rho\frac{C_y}{C_x}$ in various composite estimators for higher order approximation and have given exclusive tables in Kothawala (1989) thesis. A new alternative estimator for population mean of the study variable y using information on auxiliary variable x has been proposed along with its properties.

Consider a finite population $\Omega = (\Omega_1, \Omega_{2,\ldots,}\Omega_N)$ comprising of N identifiable and distinct units. Let (y, x) be the study and auxiliary variates respectively taking values $(y_i, x_i), i = 1, 2, \ldots, N$. Let $\left(\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i\right)$ be the population means of (y,x) respectively. For estimating the population mean $\bar{Y}$ of y, a simple random sample (SRS) of size n is drawn without replacement (WOR) scheme. Let $\left(\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$ be the sample means of (y,x) respectively and unbiased estimators of the population means $(\bar{Y}, \bar{X})$. Further let $s_y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ be the unbiased estimators of population variances / mean squares $S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$ and $S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$ respectively. Furthermore, we define the coefficient of variation of y and x as $C_y = \frac{S_y}{\bar{Y}}$ and $C_x = \frac{S_x}{\bar{X}}$ and $C = \rho\frac{C_y}{C_x}$, where $\rho = S_{yx}/(S_y, S_x)$ is the correlation coefficient between y and x; and $S_{yx} = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)$. For estimating the population mean $\bar{Y}$, the conventional ratio and product estimators are respectively defined by

$$\bar{y}_R = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right) \text{ (ratio estimator)}, \tag{1.1}$$

$$\bar{y}_p = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right) \text{ (product estimator)}. \tag{1.2}$$

To the first degree of approximation, the mean squared errors (MSEs) of $\bar{y}_R$ and $\bar{y}_P$ are respectively given by

$$\mathrm{MSE}\left(\bar{y}_R\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + C_x^2(1-2C)\right] \tag{1.3}$$

$$\mathrm{MSE}\left(\bar{y}_P\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + C_x^2(1+2C)\right], \tag{1.4}$$

where $f = \frac{n}{N}$ is the sampling fraction. The variance/MSE of the sample mean $\bar{y}$ under SRSWOR scheme is given by

$$\mathrm{MSE}(\bar{y}) = \mathrm{Var}(\bar{y}) = \frac{(1-f)}{n}S_y^2 = \frac{(1-f)}{n}\bar{Y}^2 C_y^2. \tag{1.5}$$

Comparing the mean squared errors of $\bar{y}, \bar{y}_R$ and $\bar{y}_P$, Murthy (1964) and Sahai and Ray (1980) have proved that the ratio estimator $\bar{y}_R$, sample mean $\bar{y}$, and product estimator $\bar{y}_P$ are more efficient when $C > \frac{1}{2}, -\frac{1}{2} \leq C \leq \frac{1}{2}$ and $C < -\frac{1}{2}$, respectively. Srivenkataramana (1980) and Bandyopadhyay (1980) suggested the dual to ratio and product estimators for $\bar{Y}$ respectively as

$$\bar{y}_{Rd} = \bar{y}\frac{\bar{x}^*}{\bar{X}}, \tag{1.6}$$

$$\bar{y}_{Pd} = \bar{y}\frac{\bar{X}}{\bar{x}^*}, \tag{1.7}$$

where $\bar{x}^* = \frac{(N\bar{X}-n\bar{x})}{(N-n)} = (1+g)\bar{X} - g\bar{x}$, with $g = \frac{n}{N-n} = \frac{f}{(1-f)}$.
To the first degree of approximation, the MSEs of $\bar{y}_{Rd}$ and $\bar{y}_{Pd}$ are respectively given by

$$\mathrm{MSE}\left(\bar{y}_{Rd}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + gC_x^2(g-2C)\right], \tag{1.8}$$

$$\mathrm{MSE}\left(\bar{y}_{Pd}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + gC_x^2(g+2C)\right]. \tag{1.9}$$

The dual to ratio estimator $\bar{y}_{Rd}$ is more efficient than $\bar{y}$ if

$$C > \frac{g}{2} \tag{1.10}$$

while dual to product estimator $\bar{y}_{Pd}$ is better than $\bar{y}$ if

$$C < -\frac{g}{2}. \tag{1.11}$$

Thus the sample mean $\bar{y}$ is to be preferred over $\bar{y}_{Rd}$ and $\bar{y}_{Pd}$ if

$$-\frac{g}{2} \leq C \leq \frac{g}{2}. \tag{1.12}$$

In this paper we have suggested a dual to a ratio -product-ratio estimator using auxiliary information for estimating population mean. Properties of suggested estimator are studied under large sample approximation. We have obtained the conditions under which the suggested estimator has smaller mean squared error than the sample mean, dual to ratio and product estimators and the conventional ratio and product estimators.

2 The Suggested dual to a two parameter Ratio-Product-Ratio Estimators

For estimating the population mean $\bar{y}$ of the main variable y, using the transformation $x_i^* = (1+g)\bar{X} - gx_i, \quad i = 1, 2, \ldots, N$.

We propose a two parameter dual to a ratio-product-estimator for population mean $\bar{Y}$ of y as
$\bar{y}^*_{(\eta,\delta)} = \eta\left[\frac{\delta\bar{x}^*+(1-\delta)\bar{X}}{(1-\delta)\bar{x}^*+\delta\bar{X}}\right]\bar{y} + (1-\eta)\left[\frac{(1-\delta)\bar{x}^*+\delta\bar{X}}{\delta\bar{x}^*+(1-\delta)\bar{X}}\right]\bar{y}$ (2.1)

where (η, δ) are real constants,

$$\bar{x}^* = \{(1+g)\bar{X} - g\bar{x}\}$$

such that
$E\left(\bar{x}^*\right) = \bar{X}$, where $g = \frac{n}{N-n}$.
Aim of this paper is to derive values for the constants (η, δ) such that the bias and/or the MSE of $\bar{y}^*_{(\eta,\delta)}$ is minimal.

It is to be mentioned that $\bar{y}^*_{(\eta,\delta)} = \bar{y}^*_{(1-\eta,1-\delta)}$, that is, the estimator $\bar{y}^*_{(\eta,\delta)}$ is invariant under a point reflection through the point $(\eta, \delta) = \left(\frac{1}{2}, \frac{1}{2}\right)$. In the point of symmetry $(\eta, \delta) = \left(\frac{1}{2}, \frac{1}{2}\right)$, the estimator boils down to the sample mean $\bar{y}$; that is, we have $\bar{y}^*_{\left(\frac{1}{2},\frac{1}{2}\right)} = \bar{y}$. Indeed, on the whole line $\delta = \frac{1}{2}$ our suggested estimator boils down to the sample mean estimator, that is, $\bar{y}^*_{\left(\eta,\frac{1}{2}\right)} = \bar{y}$. Further, we note that the estimator $\bar{y}_{(\eta,\delta)}$ reduces to dual to ratio estimator $\bar{y}_{Rd} = \bar{y}\left(\frac{\bar{x}^*}{\bar{X}}\right)$ for $(\eta, \delta) = (1, 1)$ while for

$(\eta, \delta) = (0, 1)$ it reduces to dual to product estimator $\bar{y}_{Pd} = \bar{y}\left(\frac{\bar{X}}{\bar{x}^*}\right)$. Its simplicity (essentially just using convex combinations and/or a dual to ratio estimator of convex combinations) and that all the three known estimators $(\bar{y}, \bar{y}_{Rd}, \bar{y}_{Pd})$ can be derived from it by selecting appropriate parameters and the reasons why we study the estimator in (2.1) and compare it to the estimators $(\bar{y}, \bar{y}_{Rd}, \bar{y}_{Pd}, \bar{y}_R, \bar{y}_P)$. To obtain the bias and MSE of the proposed estimator $\bar{y}^*_{(\eta,\delta)}$, we write

$$\bar{y} = \bar{Y}(1 + e_0), \bar{x} = \bar{X}(1 + e_1)$$

such that

$E(e_0) = E(e_1) = 0$ and $E(e_0^2) = \frac{(1-f)}{n}C_y^2, \quad E(e_1^2) = \frac{(1-f)}{n}C_x^2,$
$E(e_0 e_1) = \frac{(1-f)}{n}\rho C_y C_x,$

Expressing $\bar{y}^*_{(\eta,\delta)}$ in terms of (e_0, e_1) we have

$$\bar{y}^*_{(\eta,\delta)} = \bar{Y}(1 + e_0)\left[\eta\frac{(1-g\delta e_1)}{\{1-g(1-\delta)e_1\}} + (1-\eta)\frac{\{1-g(1-\delta)e_1\}}{(1-g\delta e_1)}\right]. \tag{2.2}$$

We assume that

$|e_1| < \min\left\{\left\{\frac{1}{|\delta g|}, \frac{1}{|(1-\delta)g|}\right\}\right\},$

and therefore we can expand $(1 - \delta g e_1)^{-1}$ and $(1 - (1-\delta)g e_1)^{-1}$ as a series in powers of e_1. We get up to $O(e_1^3)$

$$\bar{y}^*_{(\eta,\delta)} = \bar{Y}(1 + e_0)[1 - g(1-2\eta)(1-2\delta)e_1 + g^2(1-2\delta)(\eta-\delta)e_1^2 + 0(e_1^3)]. \tag{2.3}$$

It is assumed that the sample is large enough to make $|e_1|$ so small that contributions from powers of e_1 of degree higher than two are negligible. So neglecting terms of $e's$ having power greater two, we have

$$\left(\bar{y}^*_{(\eta,\delta)} - \bar{Y}\right) \cong \bar{Y}\left[e_0 - g(1-2\eta)(1-2\delta)e_1 - g(1-2\eta)(1-2\delta)e_0 e_1 + g^2(1-2\delta)(\eta-\delta)e_1^2\right]. \tag{2.4}$$

Taking expectations on both sides of (2.4) and inserting $C = \rho\left(\frac{C_y}{C_x}\right)$, we obtain the bias of $\bar{y}^*_{(\eta,\delta)}$ to order $O(n^{-1})$ as

$$\begin{aligned} B\left(\bar{y}^*_{(\eta,\delta)}\right) &= E\left(\bar{y}^*_{(\eta,\delta)} - \bar{Y}\right) \\ &= \frac{(1-f)}{n}(1-2\delta)g\bar{Y}C_x^2[(\eta-\delta)g - (1-2\eta)C]. \end{aligned} \tag{2.5}$$

Equating (2.5) to zero, we obtain

$$\delta = \frac{1}{2} \text{ or } \delta = \frac{(2\eta C - C + \eta g)}{g} \tag{2.6}$$

The suggested estimator $\bar{y}^*_{(\eta,\delta)}$, inserted with the values of δ from (2.6), becomes an (approximately) unbiased estimator for the population mean $\bar{Y}$.

Mean Squared Error of $\bar{y}^*_{(\eta,\delta)}$

Squaring both sides of (2.4) and neglecting terms of e's having power greater than two, we have

$$\left\{\bar{y}^*_{(\eta,\delta)} - \bar{Y}\right\}^2 = \bar{Y}^2\left[e_0^2 - 2g(1-2\eta)(1-2\delta)e_0 e_1 + g^2(1-2\eta)^2(1-2\delta)^2 e_1^2\right] \tag{2.7}$$

Taking expectation on both sides of (2.7) we get the MSE of $\bar{y}^*_{(\eta,\delta)}$ to terms of order $0\left(n^{-1}\right)$ as

$$\mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2 + g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right] \quad (2.8)$$

which is minimum when

$$(1-2\eta)(1-2\delta) = \frac{C}{g}. \quad (2.9)$$

Putting (2.9) in (2.8) we get the minimum MSE of $\bar{y}^*_{(\eta,\delta)}$ as

$$\begin{aligned}\mathrm{MSE}_{\min}\left(\bar{y}^*_{(\eta,\delta)}\right) &= \frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 - C^2C_x^2\right), \\ &= \frac{(1-f)}{n}S_y^2\left(1-\rho^2\right) = \mathrm{MSE}\left(\bar{y}^{*(0)}_{(\eta,\delta)}\right) \quad (2.10)\end{aligned}$$

which is equal to the approximate MSE of the regression estimator
$\bar{y}_{lr} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$,
where $\hat{\beta} = \frac{s_{yx}}{s_x^2}$ is the estimate of the population regression coefficient $\beta = \frac{S_{yx}}{S_x^2}$ of y on x,

$$s_{yx} = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \quad \text{and} \quad s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Now we state the following theorem:

Theorem-2.1- Up to terms of order $0\left(n^{-1}\right)$,

$$\mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) \geq \frac{(1-f)}{n}S_y^2\left(1-\rho^2\right)$$

with equality holding if

$$(1-2\eta)(1-2\delta) = \frac{C}{g}$$

3 Comparison of MSEs and Choice of Parameters

In this section we compare $\mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right)$ in (2.8) with MSE of sample mean $\bar{y}$, dual to ratio estimator $\bar{y}_{Rd}$, dual to product estimator $\bar{y}_{Pd}$, ratio estimator $\bar{y}_R$ and product estimator $\bar{y}_P$.

Comparing the MSE of Sample Mean $\bar{y}$ to our Proposed Estimator $\bar{y}^*_{(\eta,\delta)}$

From (1.5) and (2.8) we have

$$\mathrm{MSE}(\bar{y}) - \mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \bar{Y}^2\frac{(1-f)}{n}C_x^2 g(1-2\eta)(1-2\delta)[2C - g(1-2\eta)(1-2\delta)]$$

which is non-negative if

$$(1-2\eta)(1-2\delta)[2C - g(1-2\eta)(1-2\delta)] > 0$$

therefore, either

(i) $\eta > \frac{1}{2}, \delta > \frac{1}{2}$ and $C > \frac{g(1-2\eta)(1-2\delta)}{2}$
(ii) $\eta < \frac{1}{2}, \delta > \frac{1}{2}$ and $C < \frac{g(1-2\eta)(1-2\delta)}{2}$
(iii) $\eta > \frac{1}{2}, \delta < \frac{1}{2}$ and $C < \frac{g(1-2\eta)(1-2\delta)}{2}$, or
(iv) $\eta < \frac{1}{2}, \delta < \frac{1}{2}$ and $C > \frac{g(1-2\eta)(1-2\delta)}{2}$.

Combining the conditions (i) to (iv) with the condition $-\frac{g}{2} \leq C \leq \frac{g}{2}$, we obtained the following explicit ranges:

(i) if $0 < C \leq \frac{g}{2}$ and $\delta > \frac{1}{2}$, then $\frac{1}{2} < \eta < \frac{(2\delta g+2C-g)}{2g(2\delta-1)}$; (from (i))
(ii) if $0 < C \leq \frac{g}{2}$ and $\delta < \frac{1}{2}$, then $\frac{(2\delta g+2C-g)}{2g(2\delta-1)} < \eta < \frac{1}{2}$; (from (iv))
(iii) if $-\frac{g}{2} \leq C < 0$ and $\delta > \frac{1}{2}$, then $\frac{(2\delta g+2C-g)}{2g(2\delta-1)} < \eta < \frac{1}{2}$; (from (ii))
(iv) if $-\frac{g}{2} \leq C < 0$ and $\delta < \frac{1}{2}$, then $\frac{1}{2} < \eta < \frac{(2\delta g+2C-g)}{2g(2\delta-1)}$. (from (iii))

It is to be mentioned that the case $C = 0 \Rightarrow \rho = 0$, and thus the sample mean estimator $\bar{y}$ is the estimator with minimal MSE.

Comparing the MSE of the dual to Ratio Estimator $\bar{y}_{Rd}$ to the Suggested Estimator $\bar{y}^*_{(\eta,\delta)}$

When $C > \frac{g}{2}$, the dual to ratio estimator $\bar{y}_{Rd}$ is to be used instead of sample mean $\bar{y}$ or dual to product estimator $\bar{y}_{Pd}$. Here we are interested in finding the ranges of η and δ, where the suggested estimator $\bar{y}^*_{(\eta,\delta)}$ works better than the dual to ratio estimator $\bar{y}_{Rd}$.

From (1.8) and (2.8) we have

$$\mathrm{MSE}\left(\bar{y}_{Rd}\right) - \mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \bar{Y}^2 \frac{(1-f)}{n} C_x^2 \left[g^2\left\{1-(1-2\eta)^2(1-2\delta)^2\right\} - 2Cg\{1-(1-2\eta)(1-2\delta)\}\right]$$

which is greater than zero if

$$g\{1-(1-2\eta)(1-2\delta)\}[\{(1+(1-2\eta)(1-2\delta)\}g - 2C] > 0$$

that is, if

$$[2\eta\delta - \eta - \delta][C - g - g(2\eta\delta - \eta - \delta)] > 0$$

therefore,

$$\text{(i)}\ \frac{C}{g} - 1 > (2\eta\delta - \eta - \delta) > 0 \text{ or}$$
$$\text{(ii)}\ \frac{C}{g} - 1 < (2\eta\delta - \eta - \delta) < 0.$$

Hence from (i), where
$\frac{C}{g} > 1 \Rightarrow C > g$, we have the following.
(i) if $\delta < \frac{1}{2}$, then $\frac{(C-g+\delta g)}{g(2\delta-1)} < \eta < \frac{\delta}{(2\delta-1)}$;
(ii) if $\delta > \frac{1}{2}$, then $\frac{\delta}{(2\delta-1)} < \eta < \frac{(C-g+\delta g)}{g(2\delta-1)}$.
Further from (ii), where $\frac{1}{2} < \frac{C}{g} < 1 \Rightarrow \frac{g}{2} < C < g$, we obtain the following.
(i) if $\delta < \frac{1}{2}$, then $\frac{\delta}{(2\delta-1)} < \eta < \frac{(\delta g+C-g)}{g(2\delta-1)}$,
(ii) if $\delta > \frac{1}{2}$, then $\frac{(\delta g+C-g)}{g(2\delta-1)} < \eta < \frac{\delta}{(2\delta-1)}$.

Comparing the MSE of dual to Product estimator $\bar{y}_{Pd}$ to the Proposed estimator $\bar{y}^*_{(\eta,\delta)}$

It is shown earlier that, for $C < -\frac{g}{2} \Rightarrow \frac{C}{g} < -\frac{1}{2}$, the dual to product estimator $\bar{y}_{Pd}$ is preferred to the sample mean $\bar{y}$ and dual to the ratio estimator $\bar{y}_{Rd}$.

Taking the difference of (1.9) and (2.8) we have

$$\mathrm{MSE}\left(\bar{y}_{Pd}\right) - \mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \bar{Y}^2\frac{(1-f)}{n}gC_x^2\left[g\left\{1-(1-2\eta)^2(1-2\delta)^2\right\} + 2C\{1+(1-2\eta)(1-2\delta)\}\right]$$

which is positive if

$$\{1+(1-2\eta)(1-2\delta)\{[g\{1-(1-2\eta)(1-2\delta)\}+2C] > 0$$

i.e. if

$$(1+2\eta\delta-\eta-\delta)[C-g(2\eta\delta-\eta-\delta)] > 0$$

It follows from the above inequality that

(i) $\frac{C}{g} > (2\eta\delta-\eta-\delta) > -1$ (if both the factors in the above inequality are positive)

Or

(ii) $\frac{C}{g} < (2\eta\delta-\eta-\delta) < -1$ (if both the factors in the above inequality are non-negative)

Noting that we are only interested in $\frac{C}{g} < -\frac{1}{2}$, we get from (i)

$$-\frac{1}{2} > \frac{C}{g} > (2\eta\delta-\eta-\delta) > -1$$
$$\Rightarrow -1 < \frac{C}{g} < -\frac{1}{2};$$

and the range for η and δ, where these inequalities hold are explicitly given by the following two cases:

(i) if $\delta < \frac{1}{2}$, then $\frac{(\delta g+C)}{g(2\delta-1)} < \eta < \frac{(\delta-1)}{(2\delta-1)}$,

(ii) if $\delta > \frac{1}{2}$, then $\frac{(\delta-1)}{(2\delta-1)} < \eta < \frac{(\delta g+C)}{(2\delta-1)}$.

For any given $C^* = \frac{C}{g}$, we mention that the two regions obtained here are symmetric through $(\eta,\delta) = \left(\frac{1}{2},\frac{1}{2}\right)$. We also mention that the parameters (η,δ) which yield an asymptotically optimum estimator (AOE) [see equation (2.9)], which for a fixed C^* lie on a hyperbola, are contained in these regions. In situation (ii), where $C^* < -1 \Rightarrow C^* < -\frac{1}{2}$, the following range for η and δ can be obtained:

(i) if $\delta < \frac{1}{2}$, then $\frac{(\delta-1)}{(2\delta-1)} < \eta < \frac{(\delta g+C)}{g(2\delta-1)}$.

(ii) if $\delta > \frac{1}{2}$, then $\frac{(\delta g+C)}{g(2\delta-1)} < \eta < \frac{(\delta-1)}{(2\delta-1)}$.

We mention that for $g = -C$, the dual to product estimator $\bar{y}_{pd}$ yields the same minimum MSE as our suggested estimator $\bar{y}^*_{(\eta,\delta)}$ on the hyperbola given by (2.10).

Comparing the MSE of the Ratio Estimator $\bar{y}_R$ to the suggested Estimator $\bar{y}^*_{(\eta,\delta)}$

From (1.3) and (2.8) we have

$$\mathrm{MSE}\left(\bar{y}_R\right) - \mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \bar{Y}^2\frac{(1-f)}{n}C_x^2\left[1-g^2(1-2\eta)^2(1-2\delta)^2-2C+2Cg(1-2\eta)(1-2\delta)\right]$$

which is non-negative if
$[1-g(1-2\eta)(1-2\delta)][1+g(1-2\eta)(1-2\delta)-2C]>0$
i.e.if

$$\left\{\begin{array}{c}\text{either}\quad \frac{(2C-1)}{g}<\theta<\frac{1}{g}\\ \text{or}\quad \frac{1}{g}<\theta<\frac{(2C-1)}{g}\end{array}\right\}$$

or equivalently,

$$\min\cdot\left\{\tfrac{1}{g},\tfrac{(2C-1)}{g}\right\}<\theta<\max\cdot\left\{\tfrac{1}{g},\tfrac{(2C-1)}{g}\right\} \tag{3.1}$$

where $\theta=(1-2\eta)(1-2\delta)$.

Thus the proposed dual to ratio- product-ratio estimator $\bar{y}^{*}_{(\eta,\delta)}$ is more efficient than the ordinary ratio estimator $\bar{y}_R$ as long as the condition (3.1) is satisfied.

Comparing the MSE of the Product Estimator $\bar{y}_P$ to the Suggested Estimator $\bar{y}^{*}_{(\eta,\delta)}$

From (1.4) and (2.8) we have

$$\text{MSE}\left(\bar{y}_P\right)-\text{MSE}\left(\bar{y}^{*}_{(\eta,\delta)}\right)=\bar{Y}^2\tfrac{(1-f)}{n}C_x^2\left[1+2C-g^2\theta^2+2C\theta g\right]>0$$

if $(1+g\theta)[1-g\theta+2C]>0$

i.e. if

$$\left\{\begin{array}{c}\text{either}\quad -\frac{1}{g}<\theta<\frac{(2C+1)}{g}\\ \text{or}\quad \frac{(2C+1)}{g}<\theta<-\frac{1}{g}\end{array}\right\}$$

or alternatively,

$$\min\cdot\left\{-\frac{1}{g},\frac{(2C+1)}{g}\right\}<\theta<\max\cdot\left\{-\frac{1}{g},\frac{(2C+1)}{g}\right\}. \tag{3.2}$$

Thus the suggested estimator $\bar{y}^{*}_{(\eta,\delta)}$ is better than the usual product estimator $\bar{y}_P$ as long as the condition (3.2) holds good.

Comparing the MSE of the Chami et.al. (2012) Two-Parameter Ratio-Product-Ratio Estimator $\bar{y}_{(\eta,\delta)}$ to the Proposed Estimator

For estimating the population mean $\bar{Y}$ of the study variable y, Chami et.al.(2012) suggested the following two-parameter ratio-product-ratio estimator :

$$\bar{y}_{(\eta,\delta)}=\eta\left[\frac{(1-\delta)\bar{x}+\delta\bar{x}}{\delta\bar{x}+(1-\delta)\bar{X}}\right]\bar{y}+(1-\eta)\left[\frac{\delta\bar{x}+(1-\delta)\bar{X}}{(1-\delta)\bar{x}+\delta\bar{X}}\right]\bar{y}, \tag{3.3}$$

where (η,δ) are same as defined for the estimator $\bar{y}^{*}_{(\eta,\delta)}$ at (2.1).

To the first degree of approximation, the MSE of $\bar{y}_{(\eta,\delta)}$ is given by

$$\text{MSE}\left(\bar{y}_{(\eta,\delta)}\right)=\bar{Y}^2\tfrac{(1-f)}{n}\left[C_y^2+\theta C_x^2(\theta-2C)\right], \tag{3.4}$$

where $\theta=(1-2\eta)(1-2\delta)$.

From (2.8) and (3.4) we have

$$\mathrm{MSE}\left(\bar{y}_{(\eta,\delta)}\right) - \mathrm{MSE}\left(\bar{y}^*_{(\eta,\delta)}\right) = \bar{Y}^2 \frac{(1-f)}{n} C_x^2 \theta \left[\theta - 2C - g^2\theta + 2gC\right]$$

which is non-negative if

$$\theta\left[\theta\left(1-g^2\right) - 2C(1-g)\right] > 0$$

i.e. if $\theta(1-g)[\theta(1+g) - 2C] > 0$

i.e. if $\theta[\theta(1+g) - 2C] > 0, N > 2n \Rightarrow f < \frac{1}{2}$ (a condition which is usually met in survey situations) i.e. if

$$\left.\begin{array}{ll} \text{either} & \theta > \frac{2C}{(1+g)}, \theta > 0 \\ \text{or} & \theta < \frac{2C}{(1+g)}, \theta < 0 \end{array}\right\} \tag{3.5}$$

Thus the proposed dual to a two-parameter ratio-product-ratio estimator $\bar{y}^*_{(\eta,\delta)}$ is more efficient than the estimator $\bar{y}_{(\eta,\delta)}$ due to Chami et.al. (2012) as long as the condition (3.5) is satisfied.

4 Unbiased Asymptotically Optimum Estimator

From (2.6) and (2.9) the values of the constants η and δ can be derived for which the suggested estimator $\bar{y}^*_{(\eta,\delta)}$ becomes at least up to first order approximation an unbiased AOE. We derive a line with (recall that on this line the proposed estimator $\bar{y}^*_{(\eta,\delta)}$ always boils down to the sample mean estimator $\bar{y})\delta = \frac{1}{2}, \frac{C}{g} = 0.$

or a "curve" $\left\{\eta^*\left(C^*\right), \delta^*\left(C^*\right), C^*\right\} \in R^3$ in the parameter space with

$$\left.\begin{array}{l} \eta^*\left(C^*\right) = \frac{1}{2}\left[1 \pm \sqrt{\frac{C^*}{(2C^*+1)}}\right] \\ \delta^*\left(C^*\right) = \frac{1}{2}\left[1 \pm \sqrt{C^*\left(2C^*+1\right)}\right] \end{array}\right\} \tag{4.1}$$

where $C^* = \frac{C}{g}$.

Substitution of (4.1) in (2.1) yields an unbiased AOE for population mean $\bar{Y}$ as

$$\bar{y}^*\left(C^*\right) = \bar{y}^*_{\eta^*(C),\delta^*(C)} = \frac{\left[2\left(c^*+1\right)\bar{X}^2 - 2\left(c^*-1\right)\bar{x}^{*2} + \left(2c^{*2} - c^* - 1\right)\left(\bar{X} - \bar{x}^*\right)^2\right]}{\left[4\bar{X}\bar{x}^* - \left(2c^{*2} - c^* - 1\right)\left(\bar{X} - \bar{x}^*\right)^2\right]}. \tag{4.2}$$

The denominator will vanish if

$$\begin{aligned} &\left[4\bar{x}^*\bar{X} - \left(2C^{*2} + C^* - 1\right)\left(\bar{X} - \bar{x}^*\right)^2\right] = 0, \\ &C^* = \frac{1}{4}\left[-1 \pm \sqrt{9 - \frac{32\bar{x}^*\bar{X}}{\left(\bar{X} - \bar{x}^*\right)^2}}\right]. \end{aligned} \tag{4.3}$$

It can be easily proved that to the first degree of approximation, the bias and MSE of $\bar{y}^*\left(C^*\right)$ are given by

$$\begin{aligned} B\left(\bar{y}^*\left(C^*\right)\right) &= 0, \\ \mathrm{MSE}\left(\bar{y}^*\left(C^*\right)\right) &= \frac{(1-f)}{n} S_y^2\left(1-\rho^2\right). \end{aligned} \tag{4.4}$$

Thus the estimator $\bar{y}^*\left(C^*\right)$ at (4.2) is unbiased AOE.

5 Outlook

The following estimators:

$\bar{y}_{h1} = \bar{y}\left(\frac{\bar{x}^*}{\bar{X}}\right)^h$,
$\bar{y}_{h2} = \bar{y}\frac{\bar{X}}{\bar{X}+h\left(\bar{x}^*-\bar{X}\right)}$
$\bar{y}_{h3} = \bar{y}\left\{2-\left(\frac{\bar{x}^*}{\bar{X}}\right)^h\right\}$,
$\bar{y}_{h4} = \bar{y}\left[h\left(\frac{\bar{X}}{\bar{x}^*}\right)+(1-h)\left(\frac{\bar{x}^*}{\bar{X}}\right)\right]$,
$\bar{y}_{h5} = \bar{y}\frac{\left\{\bar{X}+h\left(\bar{x}^*-\bar{X}\right)\right\}}{\bar{X}}$, etc.

can be considered as a generalization to the dual to ratio and product estimators reported by Srivenkataramana (1980) and Bandyopadhyaya (1980), where ' h ' is a suitably chosen constants.

To the first degree of approximation, the biases and MSEs of the estimators $\bar{y}_{h1}$ to $\bar{y}_{h5}$ are respectively given by

$$B\left(\bar{y}_{h1}\right) = \frac{(1-f)}{n}\bar{Y}hg^2C_x^2\left[\frac{(h-1)}{2}-C^*\right] \tag{5.1}$$

$$B\left(\bar{y}_{h2}\right) = \frac{(1-f)}{n}\bar{Y}hg^2C_x^2\left[h+C^*\right], \tag{5.2}$$

$$B\left(\bar{y}_{h3}\right) = \frac{(1-f)}{n}\bar{Y}hg^2C_x^2\left[C^*-\frac{(h-1)}{2}\right], \tag{5.3}$$

$$B\left(\bar{y}_{h4}\right) = \frac{(1-f)}{n}\bar{Y}g^2C_x^2\left[h\left(2C^*+1\right)-C^*\right], \tag{5.4}$$

$$B\left(\bar{y}_{h5}\right) = -\frac{(1-f)}{n}\bar{Y}hg^2C_x^2C^*, \tag{5.5}$$

$$\mathrm{MSE}\left(\bar{y}_{h1}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2+hg^2C_x^2\left(h-2C^*\right)\right], \tag{5.6}$$

$$\mathrm{MSE}\left(\bar{y}_{h2}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2+hg^2C_x^2\left(h+2C^*\right)\right], \tag{5.7}$$

$$\mathrm{MSE}\left(\bar{y}_{h3}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2+hg^2C_x^2\left(h+2C^*\right)\right], \tag{5.8}$$

$$\mathrm{MSE}\left(\bar{y}_{h4}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2+g^2(2h-1)C_x^2\left\{(2h-1)+2C^*\right\}\right], \tag{5.9}$$

$$\mathrm{MSE}\left(\bar{y}_{h5}\right) = \frac{(1-f)}{n}\bar{Y}^2\left[C_y^2+hg^2C_x^2\left(h-2C^*\right)\right], \tag{5.10}$$

The MSEs of estimators $\bar{y}_{h1}$ to $\bar{y}_{h5}$ given by (5.6) to (5.10) respectively are minimized for

$$h = C^*,$$
$$h = -C^*,$$
$$h = -C^*,$$
$$h = \frac{(1-C^*)}{2},$$
$$h = C^*.$$

Thus the resulting common minimum MSE is given by

$$\mathrm{MSE}\left(\bar{y}_{h1(c^*)}\right)=\mathrm{MSE}\left(\bar{y}_{h2(-c^*)}\right)=\mathrm{MSE}\left(\bar{y}_{h3(-c^*)}\right)=\mathrm{MSE}\left(\bar{y}_{h4\left(\frac{1-c^*}{2}\right)}\right)=\mathrm{MSE}\left(\bar{y}_{h5(c^*)}\right)=\frac{(1-f)}{n}S_y^2\left(1-\rho^2\right)$$

Thus, the estimators ($\bar{y}_{h1}$ to $\bar{y}_{h5}$) and our suggested estimator $\bar{y}^{*(0)}_{(\eta,\delta)}$ (optimum estimator in the proposed class of estimators $\bar{y}^{*}_{(\eta,\delta)}$ at (2.1)) at (2.10) are equally efficient up to terms of order $0\left(n^{-1}\right)$ having the minimal possible MSE for this type of estimators.

6 Empirical Study

To examine the merits of the suggested class of estimators $\bar{y}^{*}_{(\eta,\delta)}$ relative to $\bar{y}, \bar{y}_{Rd}, \bar{y}_{Pd}, \bar{y}_R$ and $\bar{y}_P$ we considered the following data sets.

Population -I [Source: Hossain et.al. (2003)]

$\mathrm{N}=20, \quad \mathrm{n}=8, \quad \bar{Y}=101.1, \quad \bar{X}=58.8,$

$$C_y=0.873, \quad C_x=0.745, \quad \rho=0.41.$$

Population-II [Source: Steel and Torrie (1960 P. 282)]

Y: Log of leaf burn in secs

X : Clorine percentage

$$\mathrm{N}=30, \quad \mathrm{n}=6, \quad \bar{Y}=0.6860, \quad \bar{X}=0.8077,$$
$$C_y=0.700123, \quad C_x=0.7493, \quad \rho=-0.4996.$$

We have calculated the percent relative efficiency (PRE) of the developed class of estimators $\bar{y}^{*}_{(\eta,\delta)}$ with respect to $\bar{y}, \bar{y}_{Rd}, \bar{y}_{Pd}, \bar{y}_R$ and $\bar{y}_p$ using the following PRE's expressions:

$$\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)=\frac{C_y^2}{\left[C_y^2+g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right]}\times 100 \quad (6.1)$$

$$\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Rd}\right)=\frac{\left[C_y^2+gC_x^2(g-2C)\right]}{\left[C_y^2+g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right]}\times 100 \quad (6.2)$$

$$\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)=\frac{\left[C_y^2+gC_x^2(g+2C)\right]}{\left[C_y^2+g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right]}\times 100 \quad (6.3)$$

$$\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{R}\right)=\frac{\left[C_y^2+C_x^2(1-2C)\right]}{\left[C_y^2+g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right]}\times 100 \quad (6.4)$$

$$\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{P}\right)=\frac{\left[C_y^2+C_x^2(1+2C)\right]}{\left[C_y^2+g(1-2\eta)(1-2\delta)C_x^2\{g(1-2\eta)(1-2\delta)-2C\}\right]}\times 100 \quad (6.5)$$

Note1: We have computed the value of $\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right), \quad PRE\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Rd}\right)$ and $\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{R}\right)$ for Population-I as the correlation coefficient between the study variate y and the auxiliary variable x is positive.

Note 2: As the correlation coefficient between the study variate y and the auxiliary variable x is negative therefore we have computed the values of $\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right), \mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)$ and $\mathrm{PRE}\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{p}\right)$ for Population-II.

Note 3: It is to be mentioned for tables that in tables they have been developed for various values of C, C_x, C_y and ρ but have not included in the paper to save spaces.

Results are depicted in Tables 6.2, 6.4 and 6.5.

Table 6.1-PRE's of $\bar{y}, \bar{y}_{Rd}$, and $\bar{y}_R$ with respect to $\bar{y}$ for Population-I

Estimator	$\bar{y}$	$\bar{y}_R$	$\bar{y}_{Rd}$
$\mathrm{PRE}(.,\bar{y})$	100.00	97.23	115.27

Table 6.2- Range of η for given values δ under which the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}, \bar{y}_R$ and $\bar{y}_{Rd}$ for Population-I

For given values of δ	Range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}$	Range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}_R$	Range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}_{Rd}$	Common range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}, \bar{y}_R$ and $\bar{y}_{Rd}$
3.00	$\eta < 0.87$	$0.48 < \eta < 0.88$	$0.60 < \eta < 0.77$	$0.60 < \eta < 0.77$
2.75	$\eta < 0.91$	$0.48 < \eta < 0.92$	$0.61 < \eta < 0.80$	$0.61 < \eta < 0.92$
2.50	$\eta < 0.96$	$0.48 < \eta < 0.98$	$0.63 < \eta < 0.84$	$0.63 < \eta < 0.84$
2.25	$\eta < 1.03$	$0.47 < \eta < 1.05$	$0.64 < \eta < 0.89$	$0.64 < \eta < 0.89$
2.00	$\eta < 1.12$	$0.47 < \eta < 1.14$	$0.67 < \eta < 0.95$	$0.67 < \eta < 0.95$
1.75	$\eta < 1.24$	$0.47 < \eta < 1.27$	$0.70 < \eta < 1.04$	$0.70 < \eta < 1.04$
1.50	$\eta < 1.43$	$0.46 < \eta < 1.46$	$0.75 < \eta < 1.18$	$0.75 < \eta < 1.18$
1.25	$\eta < 1.74$	$0.45 < \eta < 1.78$	$0.83 < \eta < 1.40$	$0.83 < \eta < 1.40$
1.00	$\eta < 2.35$	$0.43 < \eta < 2.42$	$1.00 < \eta < 1.85$	$1.00 < \eta < 1.85$
0.75	$\eta < 4.21$	$0.35 < \eta < 4.36$	$1.50 < \eta < 3.21$	$1.50 < \eta < 3.21$
0.45	$-18.03 < \eta < 1.25$	$-18.79 < \eta < 1.25$	$-13.03 < \eta < 4.50$	$-13.03 < \eta < -4.50$
0.25	$-3.21 < \eta < 0.50$	$-3.36 < \eta < 0.65$	$-2.21 < \eta < -0.50$	$-2.21 < \eta < -0.50$
0.00	$-1.35 < \eta < 0.50$	$-1.43 < \eta < 0.57$	$-0.8530 < \eta < 0.00$	$-0.8530 < \eta < 0.00$

Table 6.3- PRE's of $\bar{y}, \bar{y}_{Pd}$, and $\bar{y}_P$ with respect to $\bar{y}$ for Population-II

Estimator	$\bar{y}$	$\bar{y}_P$	$\bar{y}_{Pd}$
$\mathrm{PRE}(.,\bar{y})$	100.00	105.510	113.254

Table 6.4- Range of η for given values δ under which the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}, \bar{y}_P$ and $\bar{y}_{Pd}$ for Population-II

For given values of δ	Range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}$	Range ofη for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}_p$	Range of $\eta$for given values ofδ in which $\bar{y}^*_{(\eta,\delta)}$ better than $\bar{y}_{Pd}$	Common range of η for given values of δ in which $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}, \bar{y}_P$ and $\bar{y}_{Pd}$
-3.00	$\eta < 0.68$	$0.57 < \eta < 0.61$	$0.57 < \eta < 0.60$	$0.57 < \eta < 0.60$
-2.75	$\eta < 0.69$	$0.58 < \eta < 0.62$	$0.58 < \eta < 0.61$	$0.58 < \eta < 0.61$
-2.50	$\eta < 0.71$	$0.58 < \eta < 0.63$	$0.58 < \eta < 0.62$	$0.58 < \eta < 0.62$
-2.25	$\eta < 0.73$	$0.59 < \eta < 0.64$	$0.59 < \eta < 0.64$	$0.59 < \eta < 0.64$
-2.00	$\eta < 0.75$	$0.60 < \eta < 0.65$	$0.6 < \eta < 0.65$	$0.6 < \eta < 0.65$
-1.75	$\eta < 0.78$	$0.61 < \eta < 0.67$	$0.61 < \eta < 0.66$	$0.61 < \eta < 0.66$
-1.50	$\eta < 0.81$	$0.62 < \eta < 0.69$	$0.63 < \eta < 0.68$	$0.63 < \eta < 0.68$
-1.25	$\eta < 0.86$	$0.64 < \eta < 0.72$	$0.64 < \eta < 0.71$	$0.64 < \eta < 0.71$
-1.00	$\eta < 0.91$	$0.66 < \eta < 0.75$	$0.67 < \eta < 0.75$	$0.67 < \eta < 0.75$
-0.75	$\eta < 0.99$	$0.70 < \eta < 0.8$	$0.7 < \eta < 0.80$	$0.7 < \eta < 0.80$
-0.50	$\eta < 1.121$	$0.75 < \eta < 0.88$	$0.75 < \eta < 0.88$	$0.75 < \eta < 0.88$
-0.25	$\eta < 1.33$	$0.83 < \eta < 1$	$0.83 < \eta < 1$	$0.83 < \eta < 1$
-0.00	$\eta < 1.74$	$0.99 < \eta < 1.25$	$1 < \eta < 1.24$	$1 < \eta < 1.24$
0.25	$\eta < 2.99$	$1.49 < \eta < 2$	$1.5 < \eta < 1.9$	$1.5 < \eta < 1.9$

Table 6.5- PRE's of the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ with respect to $\bar{y}, \bar{y}_{Rd}$ and $\bar{y}_R$ for Population-I.

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{R}\right)$
	0.60	115.27	100	118.55
	0.65	119.33	103.52	122.73
	0.70	120.05	104.15	123.47
	0.75	117.32	101.78	120.66
	0.77	115.34	100.06	118.62
	0.62	116.12	100.74	119.43
	0.65	118.44	102.75	121.82
	0.70	120.19	104.26	123.61
	0.75	119.10	103.32	122.49
	0.80	115.34	100.06	118.62
	0.63	115.71	100.38	119.00
	0.65	117.26	101.73	120.60
	0.70	119.75	103.89	123.17
	0.75	120.05	104.15	123.47
	0.80	118.13	102.48	121.49
	0.83	115.98	100.62	119.29
	0.65	115.81	100.47	119.11
	0.70	118.77	103.04	122.16
	0.75	120.13	104.22	123.56
	0.80	119.78	103.91	123.19
	0.85	117.74	102.14	121.09
	0.88	115.77	100.44	119.07
	0.67	115.49	100.19	118.78
	0.70	117.26	101.73	120.60
	0.75	119.33	103.52	122.73
	0.80	120.19	104.27	123.61
	0.85	119.78	103.91	123.19
	0.90	118.13	102.48	121.49
	0.95	115.34	100.06	118.62
	0.70	115.27	100	118.55
	0.75	117.69	102.10	121.04
	0.80	119.33	103.52	122.73
	0.85	120.13	104.22	123.55
	0.90	120.05	104.15	123.47
	0.95	119.10	103.32	122.49
	1.00	117.32	101.78	120.66
	1.04	115.34	100.06	118.62

Table 6.5 continued

η	δ	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_R\right)$
	0.75	115.27	100	118.55
	0.80	117.26	101.73	120.60
	0.85	118.77	103.04	122.16
	0.90	119.75	103.89	123.17
	0.95	120.19	104.27	123.61
	1.00	120.05	104.15	123.47
	1.05	119.36	103.55	122.76
	1.10	118.13	102.48	121.49
	1.15	116.39	100.97	119.70
	1.17	115.56	100.25	118.85
	0.84	115.49	100.19	118.78
	0.85	115.81	100.47	119.11
	0.90	117.26	101.73	120.60
	0.95	118.44	102.75	121.82
	1.00	119.33	103.52	122.73
	1.05	119.92	104.03	123.33
	1.10	120.19	104.27	123.61
	1.15	120.14	104.22	123.56
	1.20	119.78	103.91	123.19
	1.25	119.10	103.32	122.49
	1.30	118.13	102.48	121.49
	1.35	116.87	101.39	120.19
	1.40	115.34	100.06	118.62
	1.00	115.27	100	118.55
	1.05	116.32	100.91	119.64
	1.10	117.26	101.73	120.60
	1.15	118.08	102.44	121.44
	1.20	118.77	103.04	122.16
	1.25	119.33	103.52	122.73
	1.30	119.75	103.89	123.17
	1.35	120.04	104.14	123.46
	1.40	120.19	104.27	123.61
	1.45	120.19	104.27	123.62
	1.50	120.05	104.15	123.47
	1.55	119.78	103.91	123.19
	1.60	119.36	103.55	122.76
	1.65	118.81	103.07	122.19
	1.70	118.13	102.48	121.49
	1.75	117.32	101.78	120.66
	1.80	116.39	100.97	119.70
	1.85	115.34	100.06	118.62

Table 6.5 continued

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{R}\right)$
	0.75	115.27	100	118.55
	0.80	117.26	101.73	120.60
	0.85	118.77	103.04	122.16
	0.90	119.75	103.89	123.17
	0.95	120.19	104.27	123.61
	1.00	120.05	104.15	123.47
	1.05	119.36	103.55	122.76
	1.10	118.13	102.48	121.49
	1.15	116.39	100.97	119.70
	1.17	115.56	100.25	118.85
	0.84	115.49	100.19	118.78
	0.85	115.81	100.47	119.11
	0.90	117.26	101.73	120.60
	0.95	118.44	102.75	121.82
	1.00	119.33	103.52	122.73
	1.05	119.92	104.03	123.33
	1.10	120.19	104.27	123.61
	1.15	120.14	104.22	123.56
	1.20	119.78	103.91	123.19
	1.25	119.10	103.32	122.49
	1.30	118.13	102.48	121.49
	1.35	116.87	101.39	120.19
	1.40	115.34	100.06	118.62
	1.00	115.27	100	118.55
	1.05	116.32	100.91	119.6
	1.10	117.26	101.73	120.60
	1.15	118.08	102.44	121.44
	1.20	118.77	103.04	122.16
	1.25	119.33	103.52	122.73
	1.30	119.75	103.89	123.17
	1.35	120.04	104.14	123.46
	1.40	120.19	104.27	123.61
	1.45	120.19	104.27	123.62
	1.50	120.05	104.15	123.47
	1.55	119.78	103.91	123.19
	1.60	119.36	103.55	122.76
	1.65	118.81	103.07	122.19
	1.70	118.13	102.48	121.49
	1.75	117.32	101.78	120.66
	1.80	116.39	100.97	119.70
	1.85	115.34	100.06	118.62

Table 6.5 continued

η	δ	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_R\right)$
	1.50	115.27	100	118.55
	1.55	115.81	100.47	119.11
	1.60	116.32	100.91	119.6
	1.65	116.81	101.33	120.14
	1.70	117.26	101.73	120.60
	1.75	117.69	102.10	121.04
	1.80	118.08	102.44	121.44
	1.85	118.44	102.75	121.82
	1.90	118.77	103.04	122.16
	1.95	119.07	103.29	122.46
	2.00	119.33	103.52	122.73
	2.05	119.56	103.72	122.97
	2.10	119.75	103.89	123.17
	2.15	119.92	104.03	123.33
	2.20	120.04	104.14	123.46
	2.25	120.13	104.22	123.55
	2.30	120.19	104.27	123.61
	2.35	120.21	104.28	123.63
	2.40	120.19	104.27	123.62
	2.45	120.14	104.23	123.56
	2.50	120.05	104.15	123.47
	2.55	119.93	104.05	123.35
	2.60	119.78	103.91	123.19
	2.65	119.59	103.74	122.99
	2.70	119.36	103.55	122.76
	2.75	119.10	103.32	122.49
	2.80	118.81	103.07	122.19
	2.85	118.49	102.79	121.86
	2.90	118.13	102.48	121.49
	2.95	117.74	102.14	121.09
	3.00	117.32	101.78	120.66
	3.05	116.87	101.39	120.19
	3.10	116.39	100.97	119.70
	3.15	115.88	100.53	119.18
	3.20	115.34	100.06	118.62
	-13.03	115.28	100.00	118.56
	-13.00	115.34	100.06	118.62
	-12.80	115.77	100.44	119.07
	-12.60	116.19	100.80	119.49
	-12.40	116.58	101.14	119.90

Table 6.5 continued

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{R}\right)$
	-12.20	116.96	101.47	120.29
	-12.00	117.32	101.78	120.66
	-11.90	117.49	101.93	120.84
	-11.70	117.82	102.21	121.18
	-11.50	118.13	102.48	121.49
	-11.30	118.42	102.73	121.79
	-11.10	118.68	102.96	122.07
	-11.00	118.81	103.07	122.19
	-10.80	119.05	103.28	122.44
	-10.60	119.26	103.46	122.66
	-10.40	119.46	103.63	122.86
	-10.20	119.63	103.78	123.04
	-10.00	119.78	103.91	123.19
	-9.80	119.90	104.02	123.32
	-9.60	120.01	104.11	123.43
	-9.40	120.09	104.18	123.51
	-9.20	120.15	104.24	123.58
	-9.00	120.19	104.27	123.62
	-8.80	120.21	104.28	123.63
	-8.60	120.20	104.28	123.62
	-8.40	120.17	104.25	123.59
	-8.20	120.12	104.20	123.54
	-8.00	120.04	104.14	123.46
	-7.80	119.94	104.05	123.36
	-7.60	119.82	103.95	123.24
	-7.40	119.68	103.83	123.09
	-7.20	119.52	103.68	122.92
	-7.00	119.33	103.52	122.73
	-6.80	119.12	103.34	122.52
	-6.60	118.90	103.14	122.28
	-6.40	118.64	102.93	122.02
	-6.20	118.37	102.69	121.74
	-6.00	118.08	102.44	121.44
	-5.80	117.77	102.17	121.12
	-5.60	117.44	101.88	120.78
	-5.40	117.08	101.57	120.42
	-5.20	116.71	101.25	120.04
	-5.00	116.32	100.91	119.64
	-4.80	115.92	100.56	119.22
	-4.60	115.49	100.191	118.78
	-4.50	115.27	100	118.55

Table 6.5 continued

η	δ	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{Rd}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{R}\right)$
	-2.20	115.34	100.06	118.62
	-2.10	116.39	100.97	119.70
	-2.00	117.32	101.78	120.66
	-1.90	118.13	102.48	121.49
	-1.80	118.81	103.07	122.19
	-1.70	119.36	103.55	122.77
	-1.60	119.78	103.91	123.18
	-1.50	120.05	104.15	123.47
	-1.40	120.19	104.27	123.62
	-1.30	120.19	104.27	123.61
	-1.20	120.04	104.14	123.46
	-1.10	119.75	103.89	123.17
	-1.00	119.33	103.52	122.73
	-0.90	118.77	103.04	122.16
	-0.80	118.08	102.44	121.44
	-0.70	117.26	101.73	120.60
	-0.60	116.32	100.91	119.63
	-0.50	115.27	100	118.55
	-0.85	115.34	100.06	118.62
	-0.80	116.39	100.97	119.70
	-0.75	117.32	101.78	120.66
	-0.70	118.13	102.48	121.49
	-0.65	118.81	103.07	122.19
	-0.60	119.36	103.55	122.76
	-0.55	119.78	103.91	123.19
	-0.50	120.05	104.15	123.47
	-0.45	120.19	104.27	123.62
	-0.40	120.19	104.27	123.61
	-0.35	120.04	104.14	123.46
	-0.30	119.75	103.89	123.17
	-0.25	119.33	103.52	122.73
	-0.20	118.77	103.04	122.16
	-0.15	118.08	102.44	121.44
	-0.10	117.26	101.73	120.60
	-0.05	116.32	100.91	119.64
	0	115.27	100	118.55

Table 6.6- PRE's of the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ with respect to $\bar{y}$, $\bar{y}_{Pd}$ and $\bar{y}_p$ for Population-II.

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{P}\right)$
	0.58	126.52	101.75	136.14
	0.60	130.54	104.99	140.47
	0.62	132.82	106.82	142.92
	0.64	133.15	107.09	143.28
	0.66	131.52	105.77	141.52
	0.68	128.06	102.99	137.79
	0.58	125.10	100.61	134.61
	0.60	129.30	103.98	139.12
	0.62	132.07	106.22	142.11
	0.64	133.23	107.15	143.36
	0.66	132.69	106.71	142.78
	0.68	130.48	104.94	140.40
	0.70	126.77	101.95	136.41
	0.59	125.82	101.19	135.39
	0.60	127.83	102.81	137.55
	0.62	130.98	105.34	140.94
	0.64	132.82	106.82	142.92
	0.66	133.23	107.15	143.36
	0.68	132.18	106.31	142.23
	0.70	129.75	104.35	139.61
	0.72	126.08	101.39	135.67
	0.60	126.18	101.48	135.77
	0.62	129.56	104.19	139.41
	0.64	131.91	106.09	141.94
	0.66	133.12	107.06	143.24
	0.68	133.10	107.05	143.22
	0.70	131.87	106.05	141.89
	0.72	129.49	104.14	139.33
	0.74	126.08	101.39	135.67
	0.60	124.34	100	133.79
	0.62	127.83	102.81	137.56
	0.64	130.54	104.99	140.47
	0.66	132.36	106.45	142.42
	0.68	133.20	107.13	143.33
	0.70	133.04	106.99	143.15
	0.72	131.87	106.05	141.89
	0.74	129.75	104.35	139.61
	0.76	126.77	101.95	136.41
	0.77	124.10	100.52	134.49

Table 6.6 continued

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{P}\right)$
	0.62	125.82	101.19	135.39
	0.64	128.73	103.53	138.52
	0.66	130.98	105.34	140.94
	0.68	132.49	106.55	142.56
	0.70	133.20	107.13	143.33
	0.72	133.10	107.05	143.22
	0.74	132.18	106.31	142.23
	0.76	130.48	104.94	140.40
	0.78	128.06	102.99	137.79
	0.80	124.99	100.52	134.49
	0.63	125.09	100.61	134.6
	0.65	127.83	102.81	137.56
	0.67	130.07	104.61	139.9
	0.69	131.75	105.96	141.7
	0.71	132.82	106.82	142.92
	0.73	133.25	107.17	143.38
	0.75	133.04	106.99	143.15
	0.77	132.18	106.31	142.23
	0.79	130.71	105.12	140.64
	0.81	128.65	103.47	138.44
	0.83	126.08	101.39	135.67
	0.65	125.28	100.76	134.81
	0.67	127.68	102.68	137.38
	0.69	129.69	104.30	139.55
	0.71	131.29	105.59	141.27
	0.73	132.43	106.50	142.49
	0.75	133.09	107.03	143.21
	0.77	133.26	107.17	143.39
	0.79	132.93	106.91	143.03
	0.81	132.11	106.25	142.15
	0.83	130.82	105.21	140.76
	0.85	129.08	103.81	138.89
	0.87	126.94	102.09	136.59
	0.89	124.43	100.07	133.89
	0.67	124.72	100.31	134.20
	0.69	126.86	102.03	136.51
	0.71	128.73	103.53	138.52
	0.73	130.31	104.80	140.22
	0.75	131.57	105.81	141.57
	077	132.49	106.55	142.56

Table 6.6 continued

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{P}\right)$
	0.79	133.06	107.01	143.17
	0.81	133.26	107.17	143.39
	0.83	133.10	107.05	143.22
	0.85	132.58	106.62	142.66
	0.87	131.70	105.92	141.71
	0.89	130.48	104.94	140.40
	0.91	128.94	103.69	138.74
	0.93	127.10	102.22	136.77
	0.95	124.99	100.52	134.49
	0.70	124.34	100	133.79
	0.72	126.18	101.48	135.77
	0.74	127.83	102.81	137.55
	0.76	129.29	103.98	139.12
	0.78	130.54	104.99	140.47
	0.80	131.57	105.81	141.57
	0.82	132.36	106.45	142.42
	0.84	132.91	106.89	143.01
	0.86	133.20	107.13	143.33
	0.88	133.25	107.16	143.38
	0.90	133.04	106.99	143.15
	0.92	132.58	106.62	142.66
	0.94	131.87	106.05	141.89
	0.96	130.92	105.29	140.88
	0.98	129.75	104.35	139.61
	1.00	128.36	103.23	138.12
	1.02	126.77	101.95	136.41
	1.04	124.99	100.52	134.49
	0.75	124.34	100	133.79
	0.77	125.82	101.19	135.39
	0.79	127.19	102.29	136.86
	0.81	128.44	103.29	138.21
	0.83	129.56	104.19	139.41
	0.85	130.54	104.99	140.47
	0.87	131.38	105.66	141.37
	0.89	132.07	106.22	142.11
	0.91	132.61	106.65	142.69
	0.93	132.99	106.95	143.09
	0.95	133.20	107.13	143.33
	0.97	133.26	107.17	143.39
	0.99	133.15	107.09	143.28

Table 6.6 continued

η	δ	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_P\right)$
	1.01	132.88	106.87	142.99
	1.03	132.46	106.53	142.53
	1.05	131.87	106.05	141.89
	1.07	131.13	105.46	141.10
	1.09	130.25	104.75	140.15
	1.11	129.22	103.92	139.04
	1.13	128.06	102.99	137.79
	1.15	126.77	101.95	136.41
	1.17	125.36	100.82	134.89
	1.18	124.62	100.22	134.09
	0.84	124.72	100.31	134.20
	0.86	125.82	101.19	135.39
	0.88	126.86	102.03	136.51
	0.90	127.83	102.81	137.55
	0.92	128.73	103.53	138.52
	0.94	129.56	104.19	139.41
	0.96	130.31	104.80	140.22
	0.98	130.98	105.34	140.94
	1.00	131.57	105.81	141.57
	1.02	132.07	106.22	142.11
	1.04	132.49	106.55	142.56
	1.06	132.82	106.82	142.92
	1.08	133.06	107.01	143.17
	1.10	133.20	107.13	143.33
	1.12	133.27	107.17	143.39
	1.14	133.27	107.15	143.36
	1.16	133.10	107.05	143.22
	1.18	132.88	106.87	142.99
	1.20	132.58	106.62	142.66
	1.22	132.18	106.31	142.23
	1.24	131.69	105.92	141.71
	1.26	131.13	105.46	141.10
	1.28	130.48	104.94	140.40
	1.30	129.75	104.35	139.61
	1.32	128.94	103.69	138.74
	1.34	128.06	102.99	137.79
	1.36	127.10	102.22	136.77
	1.38	126.08	101.40	135.67
	1.40	124.99	100.52	134.49

Table 6.6 continued

η	δ	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^{*}_{(\eta,\delta)}, \bar{y}_{P}\right)$
	1.00	124.34	100	133.79
	1.02	125.09	100.61	134.61
	1.04	125.82	101.19	135.39
	1.06	126.52	101.75	136.14
	1.08	127.19	102.29	136.86
	1.10	127.83	102.81	137.55
	1.12	128.44	103.30	138.21
	1.14	129.01	103.76	138.82
	1.16	129.56	104.20	139.41
	1.18	130.07	104.61	139.96
	1.20	130.54	104.99	140.47
	1.22	130.98	105.34	140.94
	1.24	131.38	105.66	141.37
	1.26	131.75	105.96	141.76
	1.28	132.07	106.22	142.11
	1.30	132.36	106.45	142.42
	1.32	132.61	106.65	142.69
	1.34	132.82	106.82	142.92
	1.36	132.99	106.95	143.09
	1.38	133.12	107.06	143.24
	1.40	133.20	107.13	143.33
	1.42	133.25	107.17	143.38
	1.44	133.26	107.17	143.39
	1.46	133.23	107.15	143.36
	1.48	133.15	107.09	143.28
	1.50	133.04	106.99	143.15
	1.52	132.88	106.87	142.99
	1.54	132.69	106.71	142.78
	1.56	132.46	106.53	142.53
	1.58	132.18	106.31	142.23
	1.60	131.87	106.05	141.89
	1.62	131.52	105.77	141.52
	1.64	131.13	105.46	141.10
	1.66	130.71	105.12	140.64
	1.68	130.25	104.75	140.15
	1.70	129.75	104.35	139.61
	1.72	129.22	103.92	139.04
	1.74	128.66	103.47	138.44
	1.76	128.06	102.99	137.79

Table 6.6 continued

η	δ	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_{Pd}\right)$	PRE $\left(\bar{y}^*_{(\eta,\delta)}, \bar{y}_P\right)$
	1.78	127.43	102.48	137.12
	1.80	126.77	101.95	136.41
	1.82	126.08	101.40	135.67
	1.84	125.36	100.82	134.89
	1.86	124.62	100.22	134.09
	1.50	124.34	100	133.79
	1.52	124.72	100.31	134.20
	1.54	125.09	100.61	134.61
	1.56	125.46	100.90	135.00
	1.58	125.82	101.19	135.39
	1.60	126.18	101.48	135.77
	1.62	126.52	101.75	136.14
	1.64	126.86	102.03	136.51
	1.66	127.19	102.29	136.86
	1.68	127.52	102.55	137.21
	1.70	127.83	102.81	137.55
	1.72	128.14	103.06	137.88
	1.74	128.44	103.30	138.21
	1.76	128.73	103.53	138.52
	1.78	129.02	103.76	138.83
	1.80	129.29	103.98	139.12
	1.82	129.56	104.20	139.41
	1.84	129.82	104.41	139.69
	1.86	130.07	104.61	139.96
	1.88	130.31	104.80	140.22
	1.90	130.54	104.99	140.47
	1.92	130.77	105.17	140.71
	1.94	130.98	105.34	140.94
	1.96	131.19	105.51	141.16
	1.98	131.38	105.66	141.37
	1.20	117.93	94.85	126.89

We have computed the ranges of η for given values of δ in which the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is more efficient than $\bar{y}$, $\bar{y}_R$ and $\bar{y}_{Rd}$ for Population-I.

We have also given the common range of η under which the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than the estimators $\bar{y}$,$\bar{y}_R$ and $\bar{y}_{Rd}$. Findings are shown in Table 6.2. It is observed from Table 6.2 that the length of the common range of η increases as the value of $\delta, \delta \geq (0.75)$ (i.e. $\delta > \frac{1}{2}\Big)$ while it decreases when $(\delta \leq 0.45)$ (i.e. $\delta < \frac{1}{2}\Big)$.

Table 6.1 shows that the dual to ratio estimator $\bar{y}_{Rd}$ is more efficient than usual unbiased estimator $\bar{y}$ and the ratio estimator $\bar{y}_R$. The performance of the ratio estimator $\bar{y}_R$ is even poor than the usual unbiased estimator $\bar{y}$.

It is observed from Tables 6.1 and 6.5 that for selected values of (η, δ) the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}, \bar{y}_R$ and $\bar{y}_{Rd}$. Largest gain in efficiency is observed by using the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ over the ratio estimator $\bar{y}_R$ followed by $\bar{y}$. However their is marginal gain in efficiency by using proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ over the dual to ratio estimator $\bar{y}_{Rd}$. Thus there is enough scope

of selecting the values of η (for given δ) vice versa for obtaining estimators better than $\bar{y}, \bar{y}_R$ and $\bar{y}_{Rd}$.

Table 6.4 presents the ranges of η for given values of δ under which the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than $\bar{y}$, $\bar{y}_P$ and $\bar{y}_{Pd}$. The common ranges of η are also demonstrated in Table 6.4. It is observed that the length of the common range of η increases as the values of $|\delta|$ decrease upto i.e. $\delta \geq (-0.25)$ while it also increases when δ goes beyond -0.25 i.e. $\delta \geq (-0.25)$.

Table 6.6 exhibits that the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ is more efficient than the estimators $\bar{y}, \bar{y}_P$ and $\bar{y}_{Pd}$. Largest gain in efficiency is observed by using the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ over $\bar{y}_P$ followed by $\bar{y}$. We add here that the suggested class of estimators $\bar{y}^*_{(\eta,\delta)}$ is better than the dual to product estimator $\bar{y}_{Pd}$ with moderate gain in efficiency.

Table 6.3 shows that the dual to product estimator $\bar{y}_{Pd}$ is more efficient than the estimators $\bar{y}$ and $\bar{y}_P$ with considerable gain in efficiency.

Comparing the entries of Table 6.3 and 6.5 it is observed that there is sufficient flexibility in the values of η for given values of δ, for obtaining estimators better than the estimators $\bar{y}, \bar{y}_P$ and $\bar{y}_{Pd}$.

7 Conclusion

This article presents a dual to two parameter ratio-product-ratio estimator for estimating the population mean $\bar{Y}$ of the study variable y. Expressions of bias and mean squared error are obtained up to first order of approximation. The optimum condition is obtained under which the suggested class of estimators has the minimum mean squared error. It is observed that the dual to ratio and the dual to product estimators investigated by Srivenkataramana (1980) and Bandyopadhyay (1980) are members of the suggested class of estimators. The biases and mean squared errors of usual unbiased estimator $\bar{y}$, the dual to ratio and the dual to product estimators $\bar{y}_{Rd}$, $\bar{y}_{Pd}$ can be easily obtained from the bias and mean squared error expressions of the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ just by putting the appropriate values of the constants (η, δ). It is the niceness of the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$. We have also obtained the preference regions of the envisaged estimator $\bar{y}^*_{(\eta,\delta)}$ under which it is better than the usual unbiased estimator $\bar{y}$, ratio estimator $\bar{y}_R$, the dual to ratio estimator $\bar{y}_{Rd}$, product estimator $\bar{y}_P$ and the dual to product estimator $\bar{y}_{Pd}$. To see the performance of the class of estimators $\bar{y}^*_{(\eta,\delta)}$ over $\bar{y}, \bar{y}_R$, $\bar{y}_{Rd}$,$\bar{y}_P$ and $\bar{y}_{Pd}$ we have conducted an empirical study. We have found that under realistic regions of η for given values of δ , the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ is more efficient than the estimators $\bar{y}, \bar{y}_R$ $\bar{y}_{Rd}$ $\bar{y}_P$ and $\bar{y}_{Pd}$. Thus the proposed class of estimators $\bar{y}^*_{(\eta,\delta)}$ is recommended for its use in practice.

Acknowledgement

Authors are thankful to the learned refree and the Editor-in-Chief Professor D. K. Ghosh for their valuable suggestions regarding improvement of the paper.

References

Adhvaryu, D. and Gupta, P. C., On some alternative sampling strategies using auxiliary information. Metrika 1983, 30, 217-226.

Bandyopadhyay, S. Improved ratio and product estimators. Sankhya 1980,12, C, 142, 45-49.

Chami, S. P., Singh, B. and Thomas, D. A Two parameter Ratio-Product-Ratio estimator using auxiliary information. ISRN Probability and Statistics 2012, 1-15.

Gupta, P.C. and Kothawala, N.H. A study of second order approximation for some product type estimators. Jour. Ind. Soc. Agril. Stat.1990, 42, 2,171-185.

Gupta, P.C., On some quadratic and higher degree ratio and product estimator. J.Ind. Soc. Agril. Statist.1978, 30, 71-80.

Hossain M. I., Rahman M. S., Ahmed M. S. Second order peoperties of some estimators under double sampling. Stat. Transit 2003, 6:543-554

Kothawala, N.H., and Gupta, P.C., A study of second order approximation of some ratio type strategies. Biom. Jour.1988, 30, 3, 369-377

Murthy, M. N. Product Method of estimation, Sankhya A, 1964, 26, 69-74.

Ray S. K. and Sahai, A., (1980). Efficient families of ratio and product-type estimators, Biometrika,1980, 67, 1, 211–215.

Reddy, V. N., On ratio and product method of estimation. Sankhya B 1973,35, 307-316.

Sahai, A. Ray, S. K. An efficient estimator using auxiliary information. Metrika 1980, 22, 271-275.

Sahai, A., An efficient variant of the product and ratio estimators. Statistica Neerlandica 1979, 33, 27-35.

Singh, H. P. and Ruiz Espejo, M. On Linear regression and ratio-product estimation of a finite population mean. The Statistician 2003, 52, 1, 59-67.

Singh, H. P., Tailor, R. and Tailor, R. Estimation of finite population mean in two-phase sampling with known coefficient of variation of an auxiliary character. Statistica 2012, 72,1, 111-126.

Singh, H.P. and Nigam, P. A general class of dual to ratio estimators. Pak.Jour.Stat.Oper.Res. 2020,16,3, 421-431.

Srivastava, S.K., An estimator using auxiliary information. Bull. Cal. Statist. Assoc. 1967,16, 121-132.

Srivastava, S.K., Generalised estimator for mean of a finite population using multiauxiliary information. Jour. Amer. Statist. Assoc.1971, 66, 404-407.

Srivastava, S.K., A class of estimators using auxiliary information in sample surveys. Canad. Jour. Statist.1980, 8, 253-254.

Srivenkataramana, T. A dual to ratio estimator in sample surveys. Biometrika, J.1980, 67,1, 199-204.

Steel, R. G. D., Torrie, J. H., Principles and Procedures of Statistics. McGraw–Hill Book Co. 1960

Walsh, J. E., Generalisation of ratio estimate for population total. Sankhya A 1970, 33, 99-106.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Impact of Ageing on the Performance of Cricketers: Evidence from Indian Premier League

Purna Chandra Padhan[1], Hemanta Saikia[2] and Dibyojyoti Bhattacharjee[3]

1. *XLRI School of Management, Jamshedpur*
2. *Department of Agricultural Statistics, Assam Agricultural University, Jorhat*
3. *Department of Statistics, Assam University, Silchar*

Received: 20 January 2022 / Revised: 24 November 2022 / Accepted: 22 March 2023

Abstract

In the Indian Premier League (IPL), a franchisee-based cricket tournament played in India, it has been found that there are as many as 18 superlative international cricketers who participated in all the seasons of the tournament from 2013 to 2018. The paper attempts to find the impact of ageing on the performance of all these 18 cricketers. Accordingly, a measure is proposed for quantifying the performance of the cricketers, based on all the different traits of the game *i.e.*, batting, bowling, fielding and wicket keeping, for each of the seasons. Following this several econometric models with performance measure of the selected cricketers across various season as the dependent variable and age as the independent variable is fitted. The study finds that - out of the different econometric models the linear fixed effect model, where intercepts vary from player to player but the slope is constant across individual players provides the most reasonable fit. The study finds that aggregate ageing has an impact on the performance of the players in Twenty20 cricket, but the rate at which ageing influences the cricketers varies. However, the quadratic models failed to fit the data which was unexpected.

Keywords: Cricket Analytics, Econometric Models, Fixed Effect Model, Performance Measurement.

1 Introduction

The individual ageing process of human beings is a natural biological progression. The natural process of getting older has a significant impact on everyone. The effect of ageing on sports persons is no exception. In most sports, if we look at the career of a thriving sportsman- in the initial days he/she is recognized in the sports arena as a good performer. With time, the player gradually improvises himself/herself to reach his physical pinnacle and also the highest level of performance in his/her career. Then with maturity, the skill and experience of the player increases and the player performs at the highest level successfully. Following this at some stage of his career, the ageing of the player has its impact on the performance level [4]. This indicates that some quadratic econometric models might fit the data better. The skill and experience are not complemented sufficiently by the physical ability and therefore the performance level shows a downward trend. This is the time when players think of retiring from sports or are not considered any further by the selectors to represent their team or country. Several players at this stage of their career change their roles and become coaches or mentors or TV commentators or settle down in other professions. But there are players who do not abide by such average laws and have a different story to tell. Some players start their career with a dramatic entry and then after some initial success gets lost

once and for all. There are some others who keep playing successfully as if the age for them is just a number. An interesting discussion in this regard appears in Saikia, Bhattacharjee and Mukherjee [16].

As per the demand of sports like hockey, soccer, basketball, etc., the physical aspect of players dominates their performance. It has been widely reported as Relative Age Effect (RAE) in sports. Relative age effect in sports is a worldwide phenomenon and it exists in many, but not all, competitive sports [12]. This phenomenon has been widely conveyed now and then in different newspapers as well as sports magazines with respect to many sports. It is well understood that the performance of the players reaches uttermost at a given age and then shows a decreasing trend. However, the determination of the peak performing age of the players is always a difficult task. Moreover, the level of performance of the players in a team game can vary depending on the strength of the opponents, different playing conditions, etc.

Cricket is a team game. So, the physical, and mental ability and technical skills isolated the better-performing players from the other team members. With the development of franchisee-based cricket tournament *viz*. Indian Premier League (IPL), Big Bash League (BBL), Bangladesh Premier League (BPL), Sri Lankan Premier League (SLPL), Caribbean Premier League (CPL), etc. both retired players and emerging talented players get a platform to perform and learn from each other. The IPL is an extensive playing tournament which was initiated by the Board of Control for Cricket in India (BCCI) in the year 2008. As of May 2023, sixteen seasons of the IPL is played so far. In all these IPL seasons, it is true that some of the young players were able to draw the attention of spectators through their performances. However, the performances of senior players are also noteworthy in the IPLs. The best example is that of Shane Warne, Sachin Tendulkar, Shane Watson, Adam Gilchrist, Mahendra Singh Dhoni etc.

It is believed that the senior players mark a massive impression in the IPLs usually due to their tactical skills. This impression helped to make them everyday cricketers unlike the ones who fire only for a while [8]. The skill sets of senior players are so comprehensive that they can crack the code in no time [21] despite having less experience in the Twenty20 format of cricket. The performances of these senior players remind us that there is no substitute for skill if anyone has to succeed in any format of the game of cricket. However, as a player gets older, he becomes less fit for the game and more prone to injury [10]. It might be due to the age effect, the demand for these players has steadily diminished. Therefore, they were no longer playing for their respective teams [11]. This explicit discussion set the background of the study which aims to examine "Does really age affects the performance of players in Twenty20 cricket?"

2 Review of Literature

The effect of age has already been studied by various authors in different kinds of sports. Most of these studies revealed linear, curvilinear or exponential trends when modelling the effects of ageing [22]. A linear trend was found when evaluating the effects of ageing on freely chosen walking speed [7]. Schulz and Curnow [18] examined the age of peak performance in a broad range of athletic events. They noted that the absolute levels of peak performance among super athletes have increased vividly but the stability of peak performance cannot be ascertained. The curvilinear trend was observed when investigating competitions in indoor rowing events [19] and in freestyle swimming performances, it was found to decrease exponentially [20]. The percentage decline in masters' super athletes with increasing age in track and field performance was examined by Baker *et al*. [2]. They found that track running records declined with ageing in a curvilinear fashion as $y = 1 - \frac{exp(T-T_0)}{\tau}$, whereas in field events it declined in a linear way as $y = \alpha[T - T_0]$. Also, they reported that the decline with ageing was greater for females and longest-running events. In the case of baseball, Fair [4] estimated the effects of age using nonlinear fixed effect regression and found that ageing effects are larger for pitchers than for batters. The peak age of performance for professional baseball pitchers and batters are 26 and 28 years respectively. In

many sports, the classification system (*i.e.* senior, young, peak performance age, etc. through cut-off values) based on biological age is difficult to organize [12]. The effect of ageing for professional football players using performance-ageing curves was evaluated by Young and Weckman [22]. Addona and Yates [1] examined the relative age effect (RAE) in the National Hockey League (NHL). An analysis of master athletes in running, swimming and cycling was performed by Ransdell *et al.* [14] by age group and gender. They also examined how physiological, sociological and psychological factors affect master-level athletes' performance in the USA. Lehto [9] investigated the age-related changes in endurance performance among male amateur marathon runners from 1979 to 2014. He found a quadratic relationship of running time t as a function of age x. The fitted quadratic model indicates that the marathon performance of the average runner improves up to age 34.3 ($\pm$ 2.6) years, thereafter the performance starts to decline. A similar study was performed by Radek [13] to evaluate peak performance age in track and field athletes. The study includes a total of 6314 athletes (3474 male and 2840 female) from the World Championships, European Championships and Olympic Games. He found that the peak performance age for males and females are 25 and 26 respectively.

Though a number of measures of performances were stated in the existing literature surrounding the game of cricket, the impact of ageing on players' feats did not enrich our search. However, a study was found led by Hazra and Biswas [6] compared the mental skill ability of cricket players per different age-level categories. They implemented a one-way analysis of variance (ANOVA) and revealed that age may be the predictor of the mental skill ability of cricket players. This finding supports the usual belief that physical, mental ability and technical skills separate the best-performing players in cricket from the rest of the sports. In cricket, it is difficult to generalize a relation between ageing and performance, as cricketers with different expertise attain their peak performance at different ages. Batsmen tend to reach their peak in their late twenties after their technique matured through experience and conversely, fast bowlers often reach their peak in between early to mid-twenties when they are at the height of their physical capacity. Other bowlers, mostly spinners, even fast bowlers who can "swing" the ball, are most effective in the later part of their career [15].

3 Data and Methodology

Out of all the franchisee-based cricket tournaments mentioned above, IPL is the eldest and most popular tournament in terms of games as well as seasons. A large number of well-recognized players from different countries participate in this tournament. The season-wise data for each player who played in IPLs are available on *www.espncricinfo.com*. Having this data set available, one could easily identify the players who had participated in several seasons of IPL played so far along with their age information. However, there are only 18 players who participated unceasingly from the 2013 to 2018 IPL seasons. These 18 players are included in the study to examine the age effect on their performance.

Performance Measurement

The performances of these selected players are quantified using the different traits of the game (*e.g.* batting, bowling, fielding and wicket-keeping). To quantify the batting performance of the players, the factors considered are batting average, strike rate, and average percentage of contribution to the team total. For the bowling performance of the players, the factors considered are bowling average, economy rate and bowling strike rate. Similarly, for fielding the factors are run-outs, catches taken and missed chances of catch and run-out are also considered. For wicket-keeping the factors are catches taken, stampings executed and bye runs conceded. These factors under each of the traits (*i.e.* batting, fielding,

bowling and wicket keeping) are normalized and then multiplied by weights and combined in a linear fashion as in equation (1), to get the composite performance score.

The performance measure of the i^{th} player is given by,

$$S_i = S_{i1} + S_{i2} + \delta_i \tag{1}$$

where

$$\delta_i = \begin{cases} S_{i3}^{a_i} + S_{i4}^{1-a_i} - 1, & \text{if } i^{th} \text{ player is either a bowler or wicket keeper} \\ 0, & \text{if } i^{th} \text{ player is neither a bowler nor wicket keeper} \end{cases}$$

where a_i is an indicator variable with

$$a_i = \begin{cases} 1, & \text{if } i^{th} \text{ player is a bowler} \\ 0, & \text{if } i^{th} \text{ player is a wicket keeper} \end{cases}$$

with S_{i1} = Performance score for batting, S_{i2}= Performance score for fielding, S_{i3}= Performance score for bowling and S_{i4} = Performance score for wicket keeping.

The performance scores of S_{i1}, S_{i2}, S_{i3} and S_{i4} for batting ($k = 1$), fielding ($k = 2$), bowling ($k = 3$) and wicket keeping ($k = 4$) are computed as

$$S_{ik} = \Sigma_{j=1}^{n_k} w_{jk} Y_{ijk}; \text{ where } n_k\text{=3 for } k\text{=1, 2, 3 and 4} \tag{2}$$

The details of normalization and weight determination of the factors can be seen in Saikia, Bhattacharjee and Radhakrishnan [17]. On getting the values of S_{i1}, S_{i2}, S_{i3} and S_{i4} the performance score of the players is computed using equation (1) and then converted into corresponding performance index (P_i) which is given by

$$P_i = \frac{S_i}{Max(S_i)} \tag{3}$$

The performance score based on P_i for each player is a number lying between 0 and 1. The higher the value of P_i better the player's performance.

Modeling of Age Effect on Performance

Case-I: In order to examine the age effect on the performance of the cricketers, a standard static linear panel regression model and a quadratic panel regression model are estimated. The models are expressed as

$$P_{it} = \beta_1 + \beta_2 A_{it} + \epsilon_{it}, \ i = 1, 2, ..., N \text{ and } t = 1, 2, ..., T \tag{4}$$

$$P_{it} = \beta_1 + \beta_2 A_{it} + \beta_3 A_{it}^2 + \epsilon_{it} \tag{5}$$

where $N = 18$ cross section players, T = five years of IPLs and P_{it} represents the performance of the i^{th} player at time t. Again A_{it} represents the age of the player in time t, β_1 is unobserved independent variable *i.e.* independent of i and t, represents the squared term of the age of the players in time, ϵ_{it} is the idiosyncratic error component varies with i and t. The model (4) is estimated applying pooled *OLS*, which is desirable if β_1 does not vary across time and cross-section, with an independent variable, like age, introduced in the model. In the pooled *OLS*, the intercepts and slopes are the same over time and with individual players. When we fit this model, it has been observed that there is no impact of age on

the performance of the players. The coefficient of age (0.0049) is found statistically insignificant due to the high *p*-value of 0.1711. The reason might be due to the possibility of heteroscedasticity. In this context, we have applied white cross-section standard errors and covariance to correct the standard error of *OLS* and solve the heteroscedasticity problem. The result shows slight improvement having *p*-value of 0.0316. However, along with the positive sign of the coefficient of age, with a low R^2 value of 0.0176, *F* statistics of 1.899 and its *p*-value of 0.1711 conclude that age does not affect performance significantly. Alternatively, the model (5) is not found to be statistically significant with a very low R^2 value of 0.02, *F* statistics of 0.98 and its *p*-value of 0.38 (*cf.* Appendix-A).

One of the problems of pooled *OLS* is that the estimated coefficients will be unbiased and efficient if (i) E(ϵ_{it}) = 0 and (ii) E(A_{it}, ϵ_{it}) = 0 *i.e.* A_{it} are weekly exogenous. However, if $corr(\epsilon_{is}, \epsilon_{it}) \neq 0$ with $s \neq t$, which is very likely due to the fact that the individual (players) are repeatedly observed, then the *OLS* will be biased and shall not be efficient. If errors are correlated, *OLS* is still unbiased but inefficient. *OLS* is biased and inconsistent if the covariates are correlated with the error term. As a result, one may go for either the Fixed Effect or Random Effect model. In the present equation, there may be some unobserved individual factors, other than Age, which might affect performance. Then we may have to decompose our error term ϵ_{it} as

$$\epsilon_{it} = \alpha_i + u_{it}$$

where u_{it} is *iid i.e.* u_{it} has a zero mean and homoscedastic and not serially correlated. In this case, we may have to rely on the *FE* or *RE* model to estimate the model.

Case-II: Let us think of a linear as well as quadratic fixed effect model, where the intercepts vary, although the slope is constant across individual players. This might be reasonable as the players are heterogeneous in nature. The fixed-effects (*FE*) models are more suitable as we are interested in analysing the impact of age variables that vary over time and it's impact on performance. If α_i are individual intercept which is fixed for a given N denoting the average performance of all players individually irrespective of age, then the models can be expressed as

$$P_{it} = \beta_{1i} + \beta_2 A_{it} + \epsilon_{it} \tag{6}$$

$$P_{it} = \beta_{1i} + \beta_2 A_{it} + \beta_3 A_{it}^2 + \epsilon_{it} \tag{7}$$

In these fixed effect models, consistency does not require that the individual intercepts, whose coefficients are the β_1's and ϵ_{it} are uncorrelated. Only $E(A_{it}, \epsilon_{it}) = 0$ must hold. The parameter can be estimated using *OLS*, *LSDV*, Within Estimator, and First Difference Estimator Method. In equation (6), it assumes that all the players are having equal physical ability and no random event is influencing their performance. By fitting the model (5), applying *OLS* the coefficient of age (-0.0157) which is negatively related to the performance signifies that age effect exists on the performance of the players assisted by *p*-value (0.0288). Applying further diagnostic checking of the model, the results show quite satisfactory. Re-estimating the *FEM* with White cross-section standard errors and covariance, *p*-value the coefficient of age is now 0.0005, R^2 value is 0.368, *F* statistics is 2.88 with *p*-value of 0.0005. The common intercept of 0.777 implies how much the performance of a player is different from the common intercept value. Since the same intercept β_1 is now statistically significant with values of 0.0003, it indicates that the performance of the players will obviously be different as players are different with respect to their individual characteristics. However, the model (7) is again not found to be statistically significant with R^2 value of 0.37, *F* statistics is 2.74 with *p*-value of 0.09 (*cf.* Appendix-A).

Case-III: However, there is no reason to assume that all the players are having equal physical ability. Therefore, the intercept term shall vary in equation (6) from player to player. Here one can think of fitting random effect models where the intercept varies, and the slope is constant across individual players. The random effect model to examine age effect (A) on performance (P) are defined as

$$P_{it} = \beta_{1i} + \beta_2 A_{2it} + \beta_{3it} + \epsilon it \tag{8}$$

$$P_{it} = \beta_{1i} + \beta_2 A_{2it} + \beta_3 A_{it}^2 + \beta_{4it} + \epsilon it \tag{9}$$

If $\beta_{1i} = \beta_1 + \delta_t$, then the new random effect model of equation (8) shall be

$$P_{it} = \beta_1 + \beta_2 A_{2it} + \beta_{3it} + w_{it}, \text{ where } w_{it} \sim \text{iid } N(0, \sigma_\epsilon^2) \tag{10}$$

The β_1's are random variables with the same version. The value of β_1 is specific to individual players. The β_1's for different players are different having a mean of zero and the distributions are assumed to be normal. The overall mean is captured by β_0 and β_1 is time-invariant. To estimate the parameter *OLS* can be applied provided ϵ_{it} is homoscedastic and not serially correlated. But if ϵ_{it} is not homoscedastic then one can apply *GLS* and if it has serial correlation then *FGLS* can be applied (Parks, 1967). However, the problem of *FGLS* is that it is implemented and performs well only when $T > N$ and underestimated *SE*s in finite sample. The solution to this problem could be to apply Panel Corrected Standard Error (*PCSE*) by Beck and Katz [3]. So *PCSE* is better than *FGLS*.

In equation (8), the variations across players are assumed to be random and have some influence on the predictor "age". By fitting this model with *OLS*, we confirmed that age does not significantly affect the performance of the players. Since the *p*-value of age coefficient (0.0004) is 0.9197. We have explored all four *PCSE* specifications to correct the serial correlation problem and re-estimate the *t*-statistics. Applying the standard errors and covariance the *p*-values for, cross-section *SUR* (*PCSE*) is 0.93, for cross-section weights (*PCSE*) is 0.92, for period *SUR* (*PCSE*) is 0.93 and for period weights (*PCSE*) is 0.923. This confirms that age does not statistically significantly affect performance. However, as stated random effect may not be suitable when T (= 6) $< N$ (=18), as in this case, thus we are getting an absurd result. Alternatively, the model (9) is again not found to be statistically significant with R^2 value of 0.004, *F* statistics is 0.214 with *p*-value of 0.808 (*cf.* Appendix-A).

Further, which model is better to examine the age effect on performance, in both equation (6) and equation (10), the age effect is found to be contradictory. The solution should be apparently judged by the statistical significance of the estimated regression coefficients. Here we would like to use *Durbin-Wu-Hausman* test, which is also popularly known as Hausman [5] test to choose an appropriate model for our data between fixed effect and random effect model. Moreover, pooled *OLS* regression yields inconsistent coefficient estimates when the true model is the fixed effects model. So, to test the presence of subject-specific fixed effects, it is common to perform Hausman test.

The null hypothesis (H_0) of this test is that the preferred model for our data is a random effect model, which is consistent as well as efficient against the alternative hypothesis (H_1) the preferred model is a fixed effect model that is at least as consistent. Under the null hypothesis (H_0), the *Durbin-Wu-Hausman* test statistic is defined as

$$H = (\beta_2 - \beta_1)'(Var(\beta_1) - Var(\beta_2))^+(\beta_2 - \beta_1) \tag{11}$$

where '+' denotes the *Moore–Penrose pseudo-inverse* [1]. The test statistic asymptotically follows the

[1] In linear algebra, a ***pseudo-inverse*** A^+ of a matrix A is a generalization of the inverse matrix. The most widely known type of matrix pseudo-inverse is the ***Moore-Penrose Inverse***, which was independently described by E. H. Moore in 1920, Arne Bjerhammarin 1951, and Roger Penrose in 1955.

chi-squared distribution with the degrees of freedom equal to the rank of matrix $(Var(\beta_1) - Var(\beta_2))$. This test rejects the null hypothesis (H_0), as the *p*-value corresponding to χ^2 statistic (8.281) is 0.004 (< 0.05) and so we do not have robust and sufficient evidence to consider the random effect model for our data. Therefore, we shall proceed with the fixed effect model (*cf.* Case-II) to examine the age effect on the performance of the players.

4 Regression Results and Discussion

Using equation (6), the estimated regression coefficients under the fixed effect model are shown in Table 1. The *OLS* method is being used to estimate the intercept and coefficients.

Variable	Coefficient	Std.Error	*t*-statistics	*p*-value
C	0.777061	0.208181	3.732618	0.0003
AGE	-0.015733	0.007072	-2.224832	0.0286
Fixed Effects				
Ab de Villiars-C	0.040830			
A Rahane-C	-0.090964			
A Rayudu-C	-0.080678			
C Gayle-C	0.184728			
G Gambhir-C	-0.017734			
K Pollard-C	0.032356			
D Karthik-C	-0.065031			
M Pandey-C	-0.144187			
MS Dhoni-C	0.058574			
P Patel-C	-0.099320			
R Sharma-C	-0.019837			
R Uthappa-C	-0.034303			
S Dhawan-C	-0.035526			
S Raina-C	0.078206			
S Samson-C	-0.216253			
V Kohli-C	-0.027457			
Y Pathan-C	0.255233			
Y Singh-C	-0.181361			

Table 1: The estimated parameters of fixed effect model

In Table 1, the value of 'C' is known as the differential intercept coefficient, which represents the average performance of all the players across all six seasons of IPLs (*i.e.* 2013 to 2018 years). The *p*-value corresponding to the 'Age' coefficient is 0.028 (< 0.05) which reflects a significant age effect on the performance of the players. In addition, the negative sign of the age coefficient value indicates that the performance of the players decreases as age increases with all players considered together.

Apart from that some fixed effects coefficients for individual players can be seen in Table 1. If we add these fixed effect coefficients value to the average performance of all the players (*i.e.* C) then we shall get the expected average performance of each player estimated from the model. For example, let us consider the fixed effect coefficient value of the player AB de Villiars which is 0.040830. Now 0.040830+0.777061

$= 0.817891$ is the estimated average performance of AB de Villiars for all six years. Similarly, for A Rahane the estimated average performance based on the fitted model is $-0.090964+0.777061 = 0.686096$. The negative value of the fixed effect coefficient indicates that the performance of the player has decreased over the last six seasons of IPLs. On the contrary, a positive value implies an improvement in the payer's performance over the period considered.

The quadratic models are defined in equations (5), (7) and (9), *viz.* Pooled *OLS* (*POLS*), Fixed Effect Model (*FEM*) and Random Effect Model (*REM*) significantly fit the data when age variable is included in the models as quadratic terms. Though in terms of the sign of the coefficient Age^2 provides a negative coefficient indicating that as age increases, performance declines. However, as the coefficient associated with Age^2 is not statistically significant, so we can't have a definite conclusion. Since we took into account only six IPL seasons (*i.e.* 2013 to 2018) hence the coefficient associated with Age^2 is not significant. Thus, the claim made by Fair [4] does not seem to hold for this data set. This might be because of the fact that the number of years considered for the study is not sufficient to capture the quadratic pattern in the performance of the players against ageing.

5 Conclusion

This study tries to examine the age effect on the performance of the players in Twenty20 cricket. The performances of the players are quantified through different traits of the game (*i.e.* batting, fielding, bowling and wicket-keeping). Considering the performance as the dependent variable and the age of the players as the independent variable, we tried to fit several regression models to examine the age effect. Out of all the models discussed in the methodology, the fixed effect regression model is found to attain the objective of the study in the best fashion.

The results obtained from the fixed effect regression model confirmed that in Twenty20 cricket age has a significant impact on the performance of the players over time. Also, the performances of all the players are not affected in a similar fashion with time. Every player has reached their peak performance at different ages. However, the determination of peak performance age for individual players in cricket is not an easy task. The level of performance of the players in a team game is varied depending on the strength of the opponents and different playing conditions. This can be considered as further scope of this research. Meanwhile, as an outcome of the study, it can be stated that the age effect indeed exists in Twenty20 cricket too.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable suggestions to improve the quality of work.

Authors Contribution

All authors contributed equally in the manuscript.

References

[1] V Addona and P A Yates. "A closer look at the relative age effect in the National Hockey League." In: *Journal of Quantitative Analysis in Sports* 6 (2010), pp. 1–19. DOI: 10.2202/1559-0410.1227.

[2] A B Baker, Y Q Tang, and M J Turner. "Percentage decline in masters superathlete track and field performance with aging." In: *Experimental Aging Research* 29 (2003), pp. 47–65. DOI: 10.1080/03610730303706.

[3] N Beck and J N Katz. "What To Do (And What Not To Do) With Time-Series Cross-Section Data." In: *American Political Science Review* 89 (1995), pp. 634–647. DOI: 10.2307/2082979.

[4] R Fair. "Estimated age effects in baseball." In: *Journal of Quantitative Analysis in Sports* 4 (2008), pp. 1–39. DOI: 10.2202/1559-0410.1074.

[5] J A Hausman. "Specification tests in Econometrics." In: *Econometrica* 46 (1978), pp. 1251–1271. DOI: 10.2307/1913827.

[6] M S Hazra and S Biswas. "A study on mental skill ability of different age level cricket players." In: *International Journal of Physiology, Nutrition and Physical Education* 3 (2018), pp. 1177–1180.

[7] J E Himann et al. "Age-related changes in speed of walking." In: *Medicine and Science in Sports and Exercise* 20 (1988), pp. 121–127. DOI: 10.1249/00005768-198820020-00010.

[8] *IPL is here to stay*. Sportsstar. May 23, 2009. URL: https://sportstar.thehindu.com/magazine/ipl-is-here-to-stay/article29706115.ece (visited on 05/17/2023).

[9] N Lehto. "Effects of age on marathon finishing time among male amateur runners in Stockholm marathon 1979-2014." In: *Journal of Sport and Health Science* 5 (2015), pp. 349–354. DOI: 10.1016/j.jshs.2015.01.008.

[10] C Lucifora and R Simmons. "Superstar Effects in Sport: Evidence from Italian Soccer." In: *Journal of Sports Economics* 4 (2003), pp. 35–55. DOI: 10.1177/1527002502239657.

[11] S Mitra. "The IPL: India's foray into world sports business." In: *Sports in Society* 13 (2010), pp. 1314–1333. DOI: 10.1080/17430437.2010.534294.

[12] J Musch and S Grondin. "Unequal Competition as an Impediment to Personal Development: A Review of the Relative Age Effect in Sport." In: *Developmental Review* 21 (2001), pp. 147–167. DOI: 10.1006/drev.2000.0516.

[13] V Radek. "Identification of peak performance age in track and field athletes." In: *International Journal of Performance Analysis in Sports* 14 (2014), pp. 238–251. DOI: 10.1080/24748668.2014.11868718.

[14] L B Ransdell, J Vener, and J Huberty. "Masters athletes: An analysis of running, swimming and cycling performance by age and gender." In: *Journal of Exercise Science and Fitness* 7 (2009), S61–S73. DOI: 10.1016/S1728-869X(09)60024-1.

[15] H Saikia and D Bhattacharjee. "On classification of all-rounders of the Indian Premier League (IPL): A bayesian approach." In: *Vikalpa: The Journal for Decision Makers* 36 (2011), pp. 51–66. DOI: 10.1177/0256090920110404.

[16] H Saikia, D Bhattacharjee, and D Mukherjee. *Cricket Performance Management: Mathematical Formulation and Analysis*. 1st ed. Singapore: Springer Nature, 2019. DOI: 10.1007/978-981-15-1354-1.

[17] H Saikia, D Bhattacharjee, and U K Radhakrishnan. "A new model for player selection in Cricket." In: *International Journal of Performance Analysis in Sports* 16 (2016), pp. 373–388. DOI: 10.1080/24748668.2016.11868893.

[18] R Schulz and C Curnow. "Peak performance and age among super athletes: Track and field, swimming, baseball, tennis and golf." In: *Journal of Gerontology: Psychological Sciences* 43 (1988), pp. 113–120. DOI: 10.1093/geronj/43.5.p113.

[19] K Seiler, W Spirduso, and J Martin. "Gender differences in rowing performance and power with aging." In: *Medicine and Science in Sports and Exercise* 30 (1998), pp. 121–127. DOI: 10.1097/00005768-199801000-00017.

[20] H Tanaka and D Seals. "Age and gender interactions in physiological functional capacity: insight from swimming performance." In: *Journal of Applied Physiology* 82 (1997), pp. 846–851. DOI: 10.1152/jappl.1997.82.3.846.

[21] *The IPL XI*. ESPNcricinfo. May 30, 2011. URL: https://www.espncricinfo.com/story/indian-premier-league-2011-the-ipl-xi-517150 (visited on 05/17/2023).

[22] W A Young and G R Weckman. "Evaluating the effects of aging for professional football players in combine events using performance-aging curves." In: *International Journal of Sports Science and Engineering* 2 (2008), pp. 131–143.

Appendix

	POLS			FEM			REM		
$Variables$	C	Age	Age^2	C	Age	Age^2	C	Age	Age^2
Coefficients	0.0225	0.0154	-0.0002	0.1640	0.0267	-0.0007	-0.0890	0.0283	-0.0005
SE	0.5263	0.0367	0.0006	0.8719	0.0591	0.0010	0.6091	0.0422	0.0007
t-statistic	0.0428	0.4191	-0.2870	0.1881	0.4525	-0.7241	-0.1462	0.6715	-0.6722
p-value	0.9660	0.6760	0.7747	0.8512	0.6520	0.4709	0.8841	0.5033	0.5029
R-square		0.02			0.37			0.004	
Adj.*R*-square		0.00			0.24			-0.015	
F-statistics		0.98			2.74			0.214	
p-value		0.38			0.09			0.808	
Hausman test					8.268				
p-value					0.016				

Table 2: (Appendix-A) Panel data regression results using equations (5), (7) and (9)

Players	Age-IPL2018	PS-IPL2013	PS-IPL2014	PS-IPL2015	PS-IPL2016	PS-IPL2017	PS-IPL2018
Ab de Villiars	33	0.3296	0.3778	0.3499	0.3330	0.2482	0.3897
A Rahane	29	0.2773	0.2826	0.3150	0.2662	0.2453	0.2286
A Rayudu	32	0.2141	0.2865	0.2120	0.1947	0.1514	0.3348
C Gayle	38	0.8325	0.2175	0.2946	0.5583	0.2154	0.3011
G Gambhir	36	0.2648	0.2515	0.1887	0.2448	0.3318	0.1119
K Pollard	30	0.6482	0.3983	0.5761	0.1863	0.2777	0.1738
D Karthik	32	0.2752	0.2812	0.1102	0.1845	0.3032	0.3331
M Pandey	28	0.2201	0.2937	0.1573	0.1948	0.3317	0.1923
MS Dhoni	36	0.3660	0.4425	0.1979	0.2276	0.2231	0.3944
P Patel	32	0.2519	0.2222	0.1989	0.1316	0.2590	0.2180
R Sharma	30	0.3119	0.6656	0.2560	0.2708	0.2253	0.2178
R Uthappa	32	0.2758	0.4120	0.2291	0.2138	0.3253	0.2157
S Dhawan	32	0.3305	0.2987	0.2011	0.2241	0.3195	0.2904
S Raina	31	0.3406	0.4842	0.5181	0.2041	0.6244	0.2698
S Samson	23	0.2458	0.2872	0.1534	0.1713	0.2954	0.2765
V kohli	29	0.3731	0.3232	0.5318	0.7068	0.5338	0.6571
Y Pathan	35	0.3731	0.3232	0.5318	0.7068	0.5338	0.6571
Y Singh	36	0.6084	0.6737	0.4206	0.1709	0.6317	0.0829

Table 3: (Appendix-B) Age of the players in IPL 2018 and Performance Score (PS) of the players from IPL 2013 to IPL 2018 computed from raw data using the equation (1)

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Design-Model Based Approach to Composite Estimators for Small Area Estimation and their Sensitivity Intervals of Weights

Piyush Kant Rai [1], Shiwani Tiwari [2] and Alka [3]

1. *Dept. of Statistics, Banaras Hindu University, Varanasi, U.P., India, Email: raipiyush5@gmail.com*
2. *Dept. of Statistics, Banaras Hindu University, Varanasi, U.P., India, Email: shiwanitiwari15@gmail.com*
3. *Dept. of Mathematics and Statistics, Banasthali University, Rajasthan, India, Email: singhalka2889@gmail.com*

Received: 18 January 2022 / Revised: 24 August 2022 / Accepted: 16 December 2022

Abstract

The concept of composite estimator plays a very important role in the development of small area estimation (SAE) techniques and it has very wide applications as available in the literature. The computational aspects and optimality of weights associated with the genesis of composite estimators are very precarious issues. Due to related issues of composite estimators, its applicability is hampered and questionable in many real-life situations, especially in domain estimation and particularly in small area estimation. In the present article, model-based composite estimators for the small area proposed by Pandey and Kathuria (1995) have been utilized to derive the sensitivity interval on weights with their performance regarding efficiency for the estimators. In addition, some more weighing intervals and bounds are also developed which guarantees the superiority of composite estimators as compare to their either component estimators.

Keywords: Small Area Estimation, Model-Based Composite Estimators, Optimum Weights, Sensitivity Interval, Synthetic Estimates.

1 Introduction

The fundamental objectives of sample surveys have long been used as cost-effective means for data collection. It is also pertinent to accept that the general purpose surveys will often not achieve adequate precision of the estimator for subpopulations (often called domains or areas) of interest. Domains may be geographical based areas such as districts, counties or states. They may also be cross classifications of a small geographic area and a specific demographic, social or economic subgroups. A domain is regarded as small if the domain-specific sample is not large enough to produce a direct estimate with adequate reliable precision. Alternative estimation procedures based on "borrowing strength" from other related small areas in order to increase the effective sample size for estimation and hence the accuracy of the resulting estimates is called indirect method of estimation for e.g. synthetic estimation. According to Gonzalez (1973), "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub areas on the assumption that the small areas have the same characteristics as larger area, we identify these estimates as synthetic estimators".

Direct estimators outperform the synthetic estimators if the sample size is large for the small area while synthetic estimator is efficient as compared to direct estimators in case of small sample size for the small

area provided larger area resamples with a smaller area in properties. Thus, an alternative approach of composite estimators evolved which is weighted sum of two component estimators which can have mean square error (MSE) smaller in comparison with MSE of either component estimator if a proper weighing technique is used Schaible (1979). Computation of the optimal weight for the composite estimators is generally insolvable and very tricky problem in small area estimation. It is usually not possible to get the optimal weight for the composite estimator because the expression for the optimal MSE of the estimators are difficult to evaluate under considered sampling plan.

Holt et al. (1979) has developed model-based approach for the estimation of small area total and it was used by Pandey and Kathuria (1995) to obtained the composite estimators for small area estimation. They also derived the expressions for optimum weights and the corresponding optimum MSE's of the proposed model-based composite estimators. The generalized class of composite estimator is also developed and analyze by Tikkiwal and Ghiya (2000), which include simple ratio, ratio-synthetic among many other group of such estimators through proper convex combination of weights. Later, Rai and Pandey (2013) analyze the efficiency of generalized composite estimators using two auxilary variables under different weighting schemes for different domains and recommended the application of two auxiliry varibles over single one. Tikkiwal and Rai (2009) proposed composite estimators and their sensitivity intervals of weights for the small area estimation. Moretti and Whitworth (2019) used sample size dependent composite estimators in spatial microsimulation approaches for small area estimation.

To take care of the absence of optimum weights, we have also obtained the sensitivity interval of involved weights in the form of better performance interval with a view to retaining superiority for the composite estimator under different models developed by Holt et al.(1979).

2 Notations and Terminologies

A finite population of size N is divided into P mutually exclusive small areas, labeled $i = 1, 2, 3, ..., P$ for which estimates are required. Further, within each small area there are Q identifiable subgroups labeled $j = 1, 2, 3, ..., Q$. This labeling of units gives a complete cross classification into PQ cells with N_{ij} population individual in the $(i, j)^{th}$ cell which is assumed to be known from previous censuses or any other data sources.

Consider, Y_{ijk} be the value of k^{th} unit in i^{th} small area belonging to j^{th} group. Y_{ij} and $\bar{Y}_{ij}$ be the total and mean respectively for the population unit belonging to the $(i, j)^{th}$ cell. Also, let $Y_{i.}$ and $\bar{Y}_{i.}$ be the total and mean respectively for the character under study of i^{th} small area in the population whereas $Y_{..}$ and $\bar{Y}_{..}$ are the total and mean respectively for the character under study in the whole population. Let N, N_{ij}, $N_{.j}$ and $N_{i.}$ are the total number of population units, size of population belonging to $(i, j)^{th}$ cell, marginals size of j^{th} group and i^{th} small area respectively. Corresponding sample sizes are denoted as $n, n_{ij}, n_{.j}$ and $n_{i.}$ with an assumption that $n_{ij} > 0$ for all i and j.

Further, let $y_{ij} = \sum_{k \in s_{ij}} y_{ijk}$, $\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k \in s_{ij}} y_{ijk}$ be the sample total and sample mean for the character under study of $(i, j)^{th}$ cell respectively. Whereas, $\bar{y}_{i.} = \frac{1}{n_{i.}} \sum_j \sum_{k \in s_{ij}} y_{ijk}$ and $\bar{y}_{.j} = \frac{1}{n_{.j}} \sum_i \sum_{k \in s_{ij}} y_{ijk}$ are the sample means for the character under study for i^{th} small area and j^{th} group respectively. Moreover, $\bar{y}_{..} = \frac{1}{n} \sum_i \sum_j \sum_{k \in s_{ij}} \bar{y}_{i.}$ is the overall sample mean for the character under study. Holt et al. (1979) have given four different models for the estimation of the mean of small area. These

are:

$$\text{Model I} : y_{ijk} = \mu + e_{ijk},\ \text{var}(e_{ijk}) = \sigma_1^2; \tag{2.1}$$
$$\text{Model II} : y_{ijk} = \alpha_i + e_{ijk},\ \text{var}(e_{ijk}) = \sigma_2^2; \tag{2.2}$$
$$\text{Model III} : y_{ijk} = \beta_j + e_{ijk},\ \text{var}(e_{ijk}) = \sigma_3^2; \tag{2.3}$$
$$\text{Model IV} : y_{ijk} = \mu_{ij} + e_{ijk},\ \text{var}(e_{ijk}) = \sigma_4^2 \tag{2.4}$$

where $E(e_{ijk}) = 0$, $i = 1, 2, ...P$, $j = 1, 2, ...Q$, $k = 1, 2, ...N_{ij}$. Here, μ_{ij} is an effect due to interaction between i^{th} small area and the j^{th} group whereas α_i and β_j are the effects due to i^{th} small area and the j^{th} group respectively in which k^{th} observation lies.

Consider the following estimators for small domain total based on the above models

$$\hat{Y}_{i(1)} = n_{i.}\bar{y}_{i.} + (N_{i.} - n_{i.})\bar{y}_{..} \tag{2.5}$$
$$\hat{Y}_{i(2)} = N_{i.}\bar{y}_{i.} \tag{2.6}$$
$$\hat{Y}_{i(3)} = \sum_j n_{ij}\left(\bar{y}_{ij} - \bar{y}_{.j}\right) + \sum_j N_{ij}\bar{y}_{.j} \tag{2.7}$$
$$\hat{Y}_{i(4)} = \sum_j N_{ij}\bar{y}_{ij} \tag{2.8}$$

and their corresponding MSE's are given as

$$MSE(\hat{Y}_{i(1)}) = \left[\frac{(N_{i.} - n_{i.})^2}{n} + N_{i.} - n_{i.}\right]\sigma_1^2 \tag{2.9}$$
$$MSE(\hat{Y}_{i(2)}) = N_{i.}^2\left(1 - \frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}} \tag{2.10}$$
$$MSE(\hat{Y}_{i(3)}) = \sum_j \frac{N_{ij} - n_{ij}}{n_{ij}}(N_{ij} - n_{ij} + n_{.j})\sigma_3^2 \tag{2.11}$$
$$MSE(\hat{Y}_{i(4)}) = \sum_j \frac{N_{ij}^2}{n_{ij}}\left(1 - \frac{nij}{N_{ij}}\right)\sigma_4^2 \tag{2.12}$$

respectively. Pandey and Kathuria (1995) have given the model-based composite estimators for small domain total based on the estimators developed in Eqs. (2.5) to (2.8) using models I to IV as follows:

$$\hat{Y}_{ic(1)} = \delta_1\hat{Y}_{i(1)} + (1 - \delta_1)\hat{Y}_{i(2)} \tag{2.13}$$
$$\hat{Y}_{ic(2)} = \delta_2\hat{Y}_{i(1)} + (1 - \delta_2)\hat{Y}_{i(3)} \tag{2.14}$$
$$\hat{Y}_{ic(3)} = \delta_3\hat{Y}_{i(3)} + (1 - \delta_3)\hat{Y}_{i(4)} \tag{2.15}$$
$$\hat{Y}_{ic(4)} = \delta_4\hat{Y}_{i(2)} + (1 - \delta_4)\hat{Y}_{i(4)} \tag{2.16}$$

where $\delta_1, \delta_2, \delta_3, \delta_4$ are the chosen weights. It is already discussed in the literature that estimation of the weights is not so easy for researchers. Here, we have provided some alternative ways to get the bounds and intervals for the weights in the absence of their optimum values and to get better composite estimates. The details are provided in the next sections.

3 Proposed Performance Interval for Weights

One way to get the weights of composite estimators is to use MSE's of the estimators provided and it is free from complexity of its computation and existence. The composite estimator would perform better if

MSE of composite estimator is always less than or equal to MSE of its individual component. This property of composite estimator provides a performing interval for weights. Let us obtain the interval of weights for composite estimator $\hat{Y}_{ic(1)}$ defined in Eq.(2.13) under the condition that $MSE(\hat{Y}_{ic(1)}) \leq MSE(\hat{Y}_{i(1)})$ implies

$$MSE\left[\delta_1\hat{Y}_{i(1)} + (1-\delta_1)\hat{Y}_{i(2)}\right] \leq MSE\left(\hat{Y}_{i(1)}\right)$$
$${\delta_1}^2 MSE\left(\hat{Y}_{i(1)}\right) + (1-\delta_1)^2 MSE\left(\hat{Y}_{i(2)}\right) + 2\delta_1(1-\delta_1)Cov\left(\hat{Y}_{i(1)}, \hat{Y}_{i(2)}\right) \leq MSE\left(\hat{Y}_{i(1)}\right).$$

After neglecting the covariance term, this can be written as

$${\delta_1}^2(MSE(\hat{Y}_{i(1)}) + MSE(\hat{Y}_{i(2)})) - 2\delta_1 MSE(\hat{Y}_{i(2)}) + (MSE(\hat{Y}_{i(2)}) - MSE(\hat{Y}_{i(1)})) \leq 0.$$

This equation is quadratic in terms of δ_1, after solving the equation we get two possible values such that

$$\delta_1 = \frac{MSE(\hat{Y}_{i(2)}) + MSE(\hat{Y}_{i(1)})}{MSE(\hat{Y}_{i(2)}) + MSE(\hat{Y}_{i(1)})} = 1 \text{ or } \delta_1 = \frac{MSE(\hat{Y}_{i(2)}) - MSE(\hat{Y}_{i(1)})}{MSE(\hat{Y}_{i(2)}) + MSE(\hat{Y}_{i(1)})} \tag{3.1}$$

Since, 1 is the highest value of composite weight δ_1. So, we can evaluate the lower limit for performance interval of weight by taking

$$\delta_1 \geq \frac{MSE(\hat{Y}_{i(2)}) - MSE(\hat{Y}_{i(1)})}{MSE(\hat{Y}_{i(2)}) + MSE(\hat{Y}_{i(1)})} \tag{3.2}$$

On putting the expressions of $MSE(\hat{Y}_{i(1)})$ and $MSE(\hat{Y}_{i(2)})$ in Eq.(3.2), we get

$$\delta_1 \geq \frac{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}} - \left[\frac{(N_{i.}-n_{i.})^2}{n} + N_{i.} - n_{i.}\right]\sigma_1^2}{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}} + \left[\frac{(N_{i.}-n_{i.})^2}{n} + N_{i.} - n_{i.}\right]\sigma_1^2} \tag{3.3}$$

The above Eq.(3.3) gives the lower bound for the performance interval of weights for $\hat{Y}_{ic(1)}$. If the value of population variance is not available then the usual practice is to replace it by its unbiased estimate. Now to get the upper bound, we consider the inequality $MSE(\hat{Y}_{ic(1)}) \leq MSE(\hat{Y}_{i(2)})$ and solving further we get,

$$MSE\left(\delta_1\hat{Y}_{i(1)} + (1-\delta_1)\hat{Y}_{i(2)}\right) \leq MSE(\hat{Y}_{i(2)})$$

after neglecting covariance term

$$\begin{aligned}
&{\delta_1}^2 MSE(\hat{Y}_{i(1)}) + (1-\delta_1)^2 MSE(\hat{Y}_{i(2)}) \leq MSE(\hat{Y}_{i(2)}) \\
&{\delta_1}^2\left(MSE(\hat{Y}_{i(1)}) + MSE(\hat{Y}_{i(2)})\right) - 2\delta_1 MSE(\hat{Y}_{i(2)}) \leq 0 \\
&{\delta_1}^2\left(MSE(\hat{Y}_{i(1)}) + MSE(\hat{Y}_{i(2)})\right) \leq 2\delta_1 MSE(\hat{Y}_{i(2)})
\end{aligned} \tag{3.4}$$

On solving further, the upper limit of composite weight δ_1 and its performance interval is obtained as

$$\delta_1 \leq \frac{2\,MSE(\hat{Y}_{i(2)})}{MSE(\hat{Y}_{i(1)}) + MSE(\hat{Y}_{i(2)})} \tag{3.5}$$

After substituting the values from Eqs.(2.5) and (2.6) in Eq.(3.5) we get

$$\delta_1 \leq \frac{2\,N_{i.}^2(1-\frac{ni.}{N_{i.}})\frac{\sigma_2^2}{n_{i.}}}{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}} + \left[\frac{(N_{i.}-n_{i.})^2}{n} + N_{i.} - n_{i.}\right]\sigma_1^2}. \tag{3.6}$$

Now combining Eq.(3.3) and Eq.(3.6), the obtained performance interval for weight δ_1 is

$$\frac{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}}-\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2}{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}}+\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2}\leq\delta_1\leq \frac{2\,N_{i.}^2(1-\frac{ni.}{N_{i.}})\frac{\sigma_2^2}{n_{i.}}}{N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}}+\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2}. \tag{3.7}$$

In the similar way, the interval of weight for composite estimator $\hat{Y}_{ic(2)}$ can be obtained by solving $MSE(\hat{Y}_{ic(2)})\leq MSE(\hat{Y}_{i(1)})$ and $MSE(\hat{Y}_{ic(2)})\leq MSE(\hat{Y}_{i(3)})$, $\hat{Y}_{ic(3)}$ by solving inequalities $MSE(\hat{Y}_{ic(3)})\leq MSE(\hat{Y}_{i(3)})$ and $MSE(\hat{Y}_{ic(3)})\leq MSE(\hat{Y}_{i(4)})$ and for composite estimator $\hat{Y}_{ic(3)}$ by solving $MSE(\hat{Y}_{ic(4)})\leq MSE(\hat{Y}_{i(2)})$ and $MSE(\hat{Y}_{ic(4)})\leq MSE(\hat{Y}_{i(4)})$ and are given as

$$\frac{\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2-\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2}{\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2+\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2}\leq\delta_2\leq \frac{2\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2}{\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2+\left[\frac{(N_{i.}-n_{i.})^2}{n}+N_{i.}-n_{i.}\right]\sigma_1^2} \tag{3.8}$$

$$\frac{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2-\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2}{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2+\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2}\leq\delta_3\leq \frac{2\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2}{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2+\sum_j\frac{N_{ij}-n_{ij}}{n_{ij}}(N_{ij}-n_{ij}+n_{.j})\,\sigma_3^2} \tag{3.9}$$

and

$$\frac{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2-N_{i.}^2(1-\frac{ni.}{N_{i.}})\frac{\sigma_2^2}{n_{i.}}}{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2+N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}}}\leq\delta_4\leq\frac{2\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2}{\sum_j\frac{N_{ij}^2}{n_{ij}}\left(1-\frac{nij}{N_{ij}}\right)\sigma_4^2+N_{i.}^2\left(1-\frac{ni.}{N_{i.}}\right)\frac{\sigma_2^2}{n_{i.}}} \tag{3.10}$$

respectively. It should be noted that, if the component estimators are biased, the composite estimator will have minimum MSE than either of the component estimators under these performance interval of weights $\delta_i, i=1,2,3,4$ [see Royall (1978)]. The width of these performance interval is one. However, if the component estimators are independent and either of them is unbiased with estimable variance then optimum value of weights can be estimated in a straightforward manner.

4 Sensitivity Interval for Weights

The optimum weights and their corresponding optimum MSE expressions involved variance and other population parameters terms which is not known. Thus, the concept of optimality gone in such situations and population parameter terms are replaced by their unbiased estimates to compute the weights and MSE accordingly. Here, we considered the optimum weights $\delta_i^*, i=1,2,3,4$ and corresponding optimum MSE computed by Pandey and Kathuria (1995). The optimum weight and corresponding optimum MSE

under composite model defined in Eq.(2.13) is given by

$$\delta_1^* = \frac{\sigma_2^2}{n_{i.}\left[\left(\frac{1}{n_{i.}}+\frac{1}{n}\right)\sigma_2^2+(\alpha_i-\bar{\alpha})^2\right]} \tag{4.1}$$

and $$MSE^*(\hat{Y}_{ic(1)}) = (N_{i.}-n_{i.})\frac{N_{i.}}{n_{i.}}\left[\frac{\left(\frac{1}{n}+\frac{1}{N_{i.}}\right)\sigma_2^2+(\alpha_i-\bar{\alpha})^2}{\left(\frac{1}{n}+\frac{1}{n_{i.}}\right)\sigma_2^2+(\alpha_i-\bar{\alpha})^2}\right]\sigma_2^2. \tag{4.2}$$

Further, the optimum weights of the model-based composite estimators defined in Eqs. (2.14), (2.15) and (2.16) are

$$\delta_2^* = \frac{\sum_j \frac{(N_{ij}-n_{ij})^2\sigma_3^2}{n.j}}{\left[\sum_j (N_{ij}-n_{ij})\frac{(N_{ij}-n_{ij}+n.j)}{n.j}+(N_{i.}-n_{i.})\frac{(N_{i.}-n_{i.}+n)}{n}\right]\sigma_3^2+\left[(N_{ij}-n_{ij})\left(\beta_j-\bar{\beta}\right)\right]^2} \tag{4.3}$$

$$\delta_3^* = \frac{\sum_j (N_{ij}-n_{ij})^2\left(\frac{1}{n_{ij}}-\frac{1}{n.j}\right)\sigma_4^2}{\sum_j (N_{ij}-n_{ij})^2\left(\frac{1}{n_{ij}}-\frac{1}{n.j}\right)\sigma_4^2+\left[\sum_j (N_{ij}-n_{ij})(\mu_{ij}-\bar{\mu}_{.j})\right]^2} \tag{4.4}$$

$$\delta_4^* = \frac{\sum_j (N_{ij}-n_{ij})\left(\frac{N_{ij}}{n_{ij}}-\frac{N_{i.}}{n_{i.}}\right)\sigma_4^2}{\sum_j (N_{ij}-n_{ij})\left(\frac{N_{ij}}{n_{ij}}-\frac{N_{i.}}{n_{i.}}\right)\sigma_4^2+(\theta_i-\bar{\theta}_i)^2} \tag{4.5}$$

and their corresponding optimum MSE's are given by

$$MSE^*(\hat{Y}_{ic(2)}) = \frac{\left[(N_{i.}-n_{i.})\frac{(N_{i.}-n_{i.}+n)}{n}\sum_j \frac{(N_{ij}-n_{ij})^2\sigma_3^2}{n.j}+\left[(N_{ij}-n_{ij})\left(\beta_j-\bar{\beta}\right)\right]^2\right]\sigma_3^2}{\left\{\sum_j (N_{ij}-n_{ij})\frac{(N_{ij}-n_{ij}+n.j)}{n.j}\right\}\left[\sum_j (N_{ij}-n_{ij})\frac{(N_{ij}-n_{ij}+n.j)}{n.j}+(N_{i.}-n_{i.})\frac{(N_{i.}-n_{i.}+n)}{n}\right]\sigma_3^2+\left[(N_{ij}-n_{ij})\left(\beta_j-\bar{\beta}\right)\right]^2} \tag{4.6}$$

$$MSE^*(\hat{Y}_{ic(3)}) = \frac{\left[\left\{\sum_j (N_{ij}-n_{ij})\frac{(N_{ij}-n_{ij}+n.j)}{n.j}\sigma_4^2\right\}\left\{\sum_j (N_{ij}-n_{ij})^2\left(\frac{1}{n_{ij}}-\frac{1}{n.j}\right)\right\}+\left\{\sum_j (N_{ij}-n_{ij})(\mu_{ij}-\bar{\mu}_{.j})\right\}^2\sum_j \frac{N_{ij}(N_{ij}-n_{ij})}{n}\right]}{\sum_j (N_{ij}-n_{ij})^2\left(\frac{1}{n_{ij}}-\frac{1}{n.j}\right)\sigma_4^2+\left[\sum_j (N_{ij}-n_{ij})(\mu_{ij}-\bar{\mu}_{.j})\right]^2} \tag{4.7}$$

$$MSE^*(\hat{Y}_{ic(4)}) = \frac{\left(N_{i.}\left(\frac{N_{i.}-n_{i.}}{ni.}\right)\right)\sum_j (N_{ij}-n_{ij})\left(\frac{N_{ij}}{n_{ij}}-\frac{N_{i.}}{n_{i.}}\right)\sigma_4^2+(\theta_i-\bar{\theta}_i)^2\left(\sum_j N_{ij}\left(\frac{N_{ij}-n_{ij}}{nij}\right)\right)\sigma_4^2}{\sum_j (N_{ij}-n_{ij})\left(\frac{N_{ij}}{n_{ij}}-\frac{N_{i.}}{n_{i.}}\right)\sigma_4^2+(\theta_i-\bar{\theta}_i)^2}, \tag{4.8}$$

where, θ_i and $\bar{\theta}_i$ are $\sum_j N_{ij}\mu_{ij}$ and $\frac{N_{i.}}{n_{i.}}\sum_j N_{ij}\mu_{ij}$ respectively. In the absence of optimum weights, the sensitivity interval or bounds of involved weights are derived for the above composite estimators.

Let us define the proportional inflation I in the variance of the composite estimators, resulting from

the use of some weight δ other than the optimum weight δ^* and is given by

$$\begin{aligned}
I =& \frac{MSE\left(\hat{Y}_{ic(1)}\right) - MSE^*\left(\hat{Y}_{ic(1)}\right)}{MSE^*\left(\hat{Y}_{ic(1)}\right)} \\
=& \frac{\left[\delta_1^2 MSE\left(\hat{Y}_{i(1)}\right) + (1-\delta_1)^2 MSE\left(\hat{Y}_{i(2)}\right)\right] - \left[\delta_1^{*2} MSE\left(\hat{Y}_{i(1)}\right) + (1-\delta_1^*)^2 MSE\left(\hat{Y}_{i(2)}\right)\right]}{\left[\delta_1^{*2} MSE\left(\hat{Y}_{i(1)}\right) + (1-\delta_1^*)^2 MSE\left(\hat{Y}_{i(2)}\right)\right]} \\
=& \frac{\left(\delta_1^2 - \delta_1^{*2}\right) MSE\left(\hat{Y}_{i(1)}\right) + \left[\left(1-\delta_1^2\right) - \left(1-\delta_1^{*2}\right)\right] MSE\left(\hat{Y}_{i(2)}\right)}{\delta_1^{*2} MSE\left(\hat{Y}_{i(1)}\right) + \left((1-\delta_1^*)^2\right) MSE\left(\hat{Y}_{i(2)}\right)} \\
=& \frac{\left(\frac{1-\delta_1}{1-\delta_1^*}\right)^2 \left[\left(\frac{\delta_1}{1-\delta_1}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)\right] - \left[\left(\frac{\delta_1^*}{1-\delta_1^*}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)\right]}{\left[\left(\frac{\delta_1^*}{1-\delta_1^*}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)\right]} \\
=& \left[\frac{1-\delta_1}{1-\delta_1^*}\right]^2 I^{'} - 1
\end{aligned} \tag{4.9}$$

where $I^{'}$ is given as

$$I^{'} = \frac{\left(\frac{\delta_1}{1-\delta_1}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)}{\left(\frac{\delta_1^*}{1-\delta_1^*}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)}. \tag{4.10}$$

As,

$$\begin{aligned}
& I \geq 0 \Rightarrow \left(\frac{1-\delta_1}{1-\delta_1^*}\right)^2 I^{'} - 1 \geq 0 \Rightarrow I^{'} \geq \left(\frac{1-\delta_1^*}{1-\delta_1}\right)^2 \\
\Rightarrow & \frac{\left(\frac{\delta_1}{1-\delta_1}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)}{\left(\frac{\delta_1^*}{1-\delta_1^*}\right)^2 MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)} \geq \left(\frac{1-\delta_1^*}{1-\delta_1}\right)^2 \\
\Rightarrow & (\delta_1 + \delta_1^*) \geq \frac{2MSE\left(\hat{Y}_{i(2)}\right)}{MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(2)}\right)}.
\end{aligned} \tag{4.11}$$

Similarly, the sensitivity interval of involved weights for other three composite estimators defined in Eqs. (2.14), (2.15) and (2.16) are obtained as

$$(\delta_2 + \delta_2^*) \geq \frac{2MSE\left(\hat{Y}_{i(3)}\right)}{MSE\left(\hat{Y}_{i(1)}\right) + MSE\left(\hat{Y}_{i(3)}\right)} \tag{4.12}$$

$$(\delta_3 + \delta_3^*) \geq \frac{2MSE\left(\hat{Y}_{i(4)}\right)}{MSE\left(\hat{Y}_{i(3)}\right) + MSE\left(\hat{Y}_{i(4)}\right)} \tag{4.13}$$

$$(\delta_4 + \delta_4^*) \geq \frac{2MSE\left(\hat{Y}_{i(4)}\right)}{MSE\left(\hat{Y}_{i(2)}\right) + MSE\left(\hat{Y}_{i(4)}\right)} \tag{4.14}$$

respectively. These expressions also provide a bound for weights interlinked with optimum weights with the availability of MSE's or their estimates.

5 An Empirical Study

In this section, an empirical study is carried out to evaluate the performance interval and the optimum value of weights $(\delta_1, \delta_2, \delta_3, \delta_4)$ for different composite estimators derived in defined in Eqs. (2.13), (2.14), (2.15) and (2.16) under different models. For this study, a hypothetical data set is taken from the website https://www.scribbr.com/statistics/anova-in-r/ which contains the observed yield of a crop. The experiment is designed to observe the effect of three factors i.e. three different fertilizer $(P = 1, 2, 3)$ at four levels (i.e. variation in soil in four different blocks $Q = 1, 2, 3, 4$). Hence, the complete dataset is divided into 4 blocks comprising 24 rows each and 3 small domains consisting 32 observations each. The total observation consists of $N = 96$ units whereas each $(i, j)^{th}$ cell contains 8 observations i.e. $k = 1, 2, \cdots, 8$ for $i \in P$ and $j \in Q$. The probability samples of sizes $n = 20, 40, 60$ are drawn from randomly selected 5, 10 and 15 rows respectively using SRSWR sampling scheme. The values of performance interval derived in Eqs. (3.7), (3.8), (3.9), (3.10) and optimum weights derived in Eqs.(4.11), (4.12), (4.13), (4.14) are calculated using R-software and are presented in Table 1.

From Table 1, we can observe that the performance interval for weights at different sample sizes vary for different composite estimators. It can be noted that, as the sample size increases, the value of optimum weights is getting close to zero, this means there is little room for error in an estimate of the optimum weights if the composite estimator is to outperform either component estimators. Also, the length of sensitivity interval is obtained smallest for the large sample sizes. From Table 2, we can observe that the calculated sensitivity interval reduced the length of obtained performance interval for all the composite estimators, which is more reasonable and justifiable.

Sample Size	$\hat{Y}_{ic(1)}$		$\hat{Y}_{ic(2)}$	
	Performance Interval	δ_1^*	Performance Interval	δ_2^*
20	(0.14861863, 1.14861863)	0.00003685	(0.07879330, 1.07879330)	0.000012343
40	(0.31611261, 1.31611261)	0.00001333	(0.43497616, 1.43497616)	0.000009837
60	(0.25925601, 1.25925601)	0.00000691	(0.26898933, 1.26898933)	0.000003120
	$\hat{Y}_{ic(3)}$		$\hat{Y}_{ic(4)}$	
	Performance Interval	δ_3^*	Performance Interval	δ_4^*
20	(0.93511119, 1.93511119)	0.000001312	(0.92561194, 1.92561194)	0.000000183
40	(0.72576161, 1.72576161)	0.000001825	(0.78505402, 1.78505402)	0.000000022
60	(0.84929089, 1.84929089)	0.000000244	(0.85218129, 1.85218129)	0.000000052

Table 1: Performance interval and optimum values of weights

Sample Size	$\hat{Y}_{ic(1)}$	$\hat{Y}_{ic(2)}$
20	(1.14858178, 1.14861863)	(1.07878095, 1.07879330)
40	(1.31609927, 1.31611261)	(1.43496633, 1.43497616)
60	(1.25924909, 1.25925601)	(1.26898621, 1.26898933)
	$\hat{Y}_{ic(3)}$	$\hat{Y}_{ic(4)}$
20	(1.93510987, 1.93511119)	(1.92561176, 1.92561194)
40	(1.72575979, 1.72576161)	(1.78505399, 1.78505402)
60	(1.84929064, 1.84929089)	(1.85218123, 1.85218129)

Table 2: Sensitivity interval of weights

6 Conclusion

The application of composite estimators for the estimation of domain parameters is restricted because of the problem related to the estimation of weights and its optimum value. The situation is more difficult in the case of small area estimation using composite estimators. The work done in this article motivates the use of composite estimators efficiently using different weighting schemes proposed under the model given by Holt et al.(1979). It will also be very helpful to develop a small area estimation methodology based on composite estimation along with bounds and intervals of weights which will be proven its superiority over other constituent estimators.

References

[1] Gonzalez, M. E. (1973). Use and Evaluation of Synthetic Estimates. In *Proceedings of the Social Statistics Section*, 33-36, American Statistical Society.

[2] Holt, D., Smith, T. M. F., and Tomberlin, T. J. (1979): A Model-Based Approach to Estimation for Small Subgroups of a Population. *Journal of the American Statistical Association*, 74, 405-410.

[3] Moretti, A., and Whitworth, A. (2019). Evaluations of Small Area Composite Estimators Based on the Iterative Proportional Fitting Algorithm. *Communications in Statistics-Simulation and Computation*, 1-18.

[4] Pandey, P. S., and Kathuria, O. P. (1995): Some Composite Estimators for Small Area Estimation. *Journal of the Indian Society of Agricultural Statistics*, 47(3), 262-272.

[5] Rai, P. K. and Pandey, K. K. (2013). Synthetic Estimators Using Auxiliary Information in Small Domains. *Statistics in Transition new series*, 1(14), 31-44.

[6] Royall, R. M. (1978). Prediction Models in Small Area Estimation. Presented at the NIDA-NCHS Workshop on Synthetic Estimates, Princeton, N. J.

[7] Schaible, W. L. (1979). A Composite Estimator for Small Area Statistics. *In Synthetic Estimates for Small Areas: Statistical Workshop Papers, Edited by: Steinberg, J.*. Rockville, Md.: National Institute on Drug Abuse. (NIDA Research Monograph 24), 36-53.

[8] Tikkiwal, G. C., and Rai, P. K. (2009): A Composite Estimator for Small Domains and Its Sensitivity Interval for Weights α. *Statistics In Transition*, 10(2), 269-275.

[9] Tikkiwal, G. C., and Ghiya, A. (2000). A Generalized Class of Synthetic Estimators with Application to Crop Acreage Estimation for Small Domains. *Journal of Mathematical Methods in Biosciences*, 42(7), 865-876.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

MLE and Inference Under Multiply Progressive Type-II Censoring for Exponential Life-time Model

M N Patel [1], R D Chaudhari[2]

1. *Retired Professor, Department of Statistics, School of Sciences, Gujarat University, Ahmedabad-380009, Email: mnpatel.stat@gmail.com*
2. *M G Science Institute, Ahmedabad-380009, Gujarat, India*

Received: 15 February 2022 / Revised: 17 January 2023 / Accepted: 22 July 2023

Abstract

Censoring occurs commonly in life testing experiments. This paper introduces a mixture of multiply Type-II censoring and progressive Type-II censoring schemes, called multiply progressive Type-II censoring, for life-testing or reliability experiments. We use the maximum likelihood method to obtain point and interval estimators of the mean lifetime, reliability, and hazard rate under exponential distribution with this censoring scheme. We also present the associated expressions of the expected total test time, which will be useful for experimental planning, in case of this censoring scheme and the complete sample without censoring. Real-life example is considered to illustrate the methods of inference developed here. A simulation study is carried out to check the performance of the estimators as the multiply Type-II progressive censoring scheme, and the parameter vary..

Keywords: Maximum likelihood estimation, confidence interval, reliability characteristics, mean squared error, simulation, expected test termination time.

1 Introduction

Exponential distribution plays an important role in life-testing and reliability studies. It is a very widely used life-time model for which statistical methods were extensively developed. Many authors have contributed to the methodology of this distribution. For example, Sukhatme [34], Epstein ([10], [9]), Epstein and Sobel ([11], [12]), Lawless [20], Patel and Gajjar ([29], [30]), Patel ([28], [26]).

In failure data analysis, life testing experiments often deal with censored samples. An experimenter may terminate the life test before all n units on the test fail to save time and cost. Two types of censoring viz: Type-I censoring and Type-II censoring generally recognized. Multiply Type-II censoring is a generalization of Type-II censoring. Multiply Type-II censored samples may arise in life testing experiments when failure times of some units on the test were not observed due to mechanical or experimental difficulties. Alternatively, in the situation where some units failed between two points of observation with exact times of failure of these units unobserved. Balasubramanian and Balakrishnan [4] obtained MLE and BLUE in the two-parameter exponential distribution. Patel [27], Patel and Patel [24], and Shah and Patel ([33], [32]) considered the inference for exponential, Pareto, Rayleigh and geometric life-time models under multiply Type-II censoring. They have considered the multiply Type-II censoring scheme as follows: Suppose n items are placed on life test and only failure times of $r_1^{th}, r_2^{th}, \cdots, r_k^{th}$ failures are made available for analysis, when $1 \leq r_1 \leq r_2 \leq \cdots \leq r_k \leq n$.

The other version of the multiply Type-II censoring scheme is prescribed as follows: In the life-testing experiment with n items on the test, the first r, last s, and middle l observations are censored, and

remaining failure times are observed ($r + s + l < n$). Balakrishnan [1] considered such a multiply Type-II censoring scheme to obtain the approximate maximum likelihood estimators of location and scale parameters of an exponential distribution. Some more references related to such multiply Type-II censoring schemes are Fei and Cong [13], Upadhyay et al. [37], and Shah and Patel [33]. Recently, Gadhvi [15] utilized a joint multiply Type-II censoring scheme to estimate the parameters of two exponential distributions. Multiply Type-II censoring is a frequently practiced censoring scheme, particularly, if one follows a Type-II censoring scheme, but since there are not enough time and human resources to record the failure time of each subject, only several failures and the number of failures between them are recorded. This frequently happens in follow-up studies in epidemiology and reliability, etc.

If an experimenter desires to remove live units at points other than the final termination point of a life test, the Type-II censoring or multiply Type-II censoring scheme will not be of use to the experimenter. Type-II censoring and multiply Type-II censoring schemes do not allow for units to be removed from the life test before the final termination point. However, this allowance will be desirable, as in the case of accidental breakage of test units, in which the loss of units at points other than the termination point may be unavoidable. Intermediate removal may also be desirable when a compromise is sought between time consumption and the observation of some extreme values. These lead us to the area of progressive Type-II censoring. Progressive Type-II censoring schemes are also considered in clinical trials. Here, the drop-out of patients may be caused by migration, lack of interest or by ethical decisions. These reasons may be regarded as random withdrawals during the experiment.

Some early works on progressive censoring can be found in Cohen [7], Mann [22], and Thomas and Wilson [35]. The statistical inference on the parameters of lifetime distributions under progressive Type-II censoring has been studied by several authors, including Wingo [39], Cohen and Norgaard [8], Gibbons and Vance [17], and Wong [40]. Patel and Gajjar ([29], [30]) have used k-stage grouped and ungrouped progressive censoring schemes with changing failure rates for exponential life-time distribution. Viveros and Balakrishnan [38] obtain exact confidence intervals for extreme value parameters and exponential distributions based on progressively censored data. Patel and Patel [25] applied a progressive censoring scheme in the case of a discrete life-time model. Xie et al. [42] applied a progressive Type-II censoring scheme to derive exact inferences and obtain an optimal scheme for a simple step-stress model. Gajjar and Patel [16] discussed estimation for a mixture of exponential distributions based on progressively Type-II censored sample. Hofmann et al. [18] indicated that the Type-II progressive censoring schemes significantly improve conventional Type-II censoring in many situations.

This scheme is generalized by Balakrishnan and Sandhu [3], in which, initially some failures are not observed, then after progressive censored failure times are observed, such scheme is known as general progressive Type-II censoring. Such a scheme is used by Fernandez [14] for an exponential life-time model. Barot and Patel [5] have used the general progressive censoring scheme to obtain reliability indexes for cold standby systems. Readers can refer to the book written by Balakrishnan and Aggarwala [2] for more details on the methods and applications of progressive censoring.

By combining multiply Type-II censoring and progressive Type-II censoring schemes a scheme is developed as a mixture of both the censoring schemes, which we shall call multiply progressive Type-II censoring scheme.

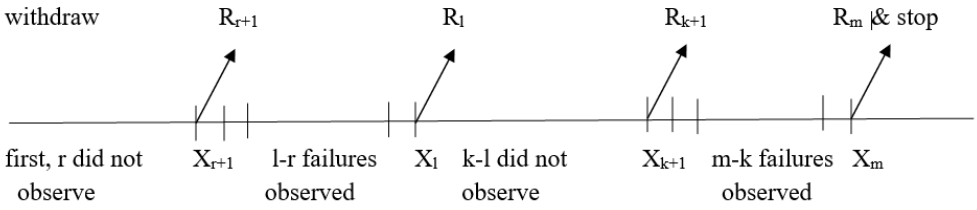
The scheme is described as follows:

Suppose n items are placed on a life testing experiment, and first r failures are not observed. Then $(r+1)^{th}, (r+2)^{th}, \cdots, l^{th}$ failure times $X_{r+1} < X_{r+2} < \cdots < X_l$ are made available, and at each of these failure times $R_{r+1}, R_{r+2}, \cdots, R_l$ units are removed respectively from the test randomly from the available survival units. Then again $(k-l)$ failures are not observed. Then after $(k+1)^{th}, (k+2)^{th}, \cdots, m^{th}$ failure times $X_{k+1} < X_{k+2} < \cdots < X_m$ are made available and at each

of these failure times again $R_{k+1}, R_{k+2}, \cdots, R_m$ units are removed respectively from the test, randomly from the available survival units. Finally, the experiment is terminated at m^{th} failure. Obviously, here $R_m = n - \sum_{j=r+1}^{l} R_j - \sum_{j=k+1}^{m-1} R_j - m$ is the remaining all survival items. As this scheme is a combination of multiply Type-II censoring and progressive Type-II censoring schemes, it allows the experimenter to remove units from a life test at various stages during the experiment. Hence, it becomes useful in clinical trials and life-testing experiments. The advantage of this censoring scheme is the experimenter can decide the value of r, l and m as a compromise between a shorter experimental time and a higher chance to observe failures. Thus, the scheme assures us not only to get a sufficient number of observed failure times for efficiency of statistical inference but also to control the total time on the test.

One can see that the multiply Type-II censoring is a special case of the multiply progressive Type-II censoring when $R_{r+1} = R_{r+2} = \cdots = R_l = R_{k+1} = R_{k+2} = \cdots = R_{m-1} = 0$. For $r = 0$, $k = l$ and $R_{r+1} = R_{r+2} = \cdots = R_l = R_{k+1} = R_{k+2} = \cdots = R_{m-1} = 0$, the scheme reduces to ordinary Type-II censoring scheme.

The scheme is visualized in the following figure:



Patel [25] considered multiply Type-II censoring schemes in a two-stage progressive censoring scheme and used it for estimating the parameters of exponential model with changing failure rates. No further work has been found on such a multiply progressive Type-II censoring scheme, which motivates us to consider such a censoring scheme.

In this paper, we use multiply progressive Type-II data when the life time of each experimental unit follows an exponential distribution. The aim of this paper is twofold. First, to obtain point and interval estimates of the reliability characteristics using frequentist approach. The second aim of this paper is to consider the expected termination time of the test when the data are multiply progressive Type-II censored.

The rest of this paper is organized as follows. Section 2 presents the multiply progressively Type-II censoring scheme and the likelihood function. Section 3 is concerned with maximum likelihood (ML) estimation. The asymptotic property of the ML estimator is considered. The confidence interval for the parameter based on the ML estimator is derived. In Section 4, the expected test termination time is derived and compared it with the expected test termination time obtained under the complete sample without censoring. Numerical studies and conclusion are presented in Section 5. Section 6 gives simulation studies. The paper is concluded in Section 7.

2 Model

Suppose that the life-time X of a unit is assumed to follow an exponential distribution Exp (θ) with mean $\theta > 0$. The probability density function and cumulative distribution function are given, respectively as

$$f(x, \theta) = \frac{1}{\theta} \exp(-x/\theta) \tag{1}$$

and

$$F(x, \theta) - 1 - \exp(-x/\theta), \quad x > 0, \quad \theta > 0. \tag{2}$$

The reliability function and hazard function of the distribution are obtained, respectively as

$$R(t) = \exp(-t/\theta) \tag{3}$$

and

$$h(t) = 1/\theta,\ \mathrm{t}\ >\ 0. \tag{4}$$

Suppose that n randomly selected units from an Exp (θ) population, θ being unknown, are put on test under a multiply progressive Type-II censoring scheme as discussed in Section 1. The likelihood function of θ is then

$$L = C\left[F(x_{r+1})\right]^r \prod_{i=r+1}^{l}\left[f(x_i)\left\{1-F(x_i)\right\}^{R_i}\right]\left[F(x_{k+1})-F(x_l)\right]^{k-l}\prod_{i=k+1}^{m}\left[f(x_i)\left\{1-F(x_i)\right\}^{R_i}\right] \tag{5}$$

where constant

$$C = \binom{n}{r}(n-r)\prod_{j=r+2}^{l}\left(n-\sum_{i=r+1}^{j-1}R_i-j+1\right)\binom{n-l}{k-l}(n-k)\prod_{j=k+2}^{m}\left(n-\sum_{i=k+1}^{j-1}R_i-j+1\right)$$

By (1) and (2), the likelihood becomes

$$\begin{aligned} L &= C\left(1-\exp(-x_{r+1}/\theta)\right)^r \prod_{i=r+1}^{l}\left[\frac{1}{\theta}\exp(-x_i/\theta)(\exp(-x_i/\theta))^{R_i}\right]\times \\ &\quad \left[\exp(-x_l/\theta)-\exp(-x_{k+1}/\theta)\right]^{k-l}\prod_{i=k+1}^{m}\left[\frac{1}{\theta}\exp(-x_i/\theta)(\exp(-x_i/\theta))^{R_i}\right] \\ &= C\theta^{-(l-r+m-k)}\exp\left\{-T/\theta\right\}\left(1-\exp\left(-(x_{k+1}-x_l)/\theta\right)\right)^{k-l}\left(1-\exp\left(-(x_{r+1}/\theta)\right)\right)^r \end{aligned} \tag{6}$$

where,

$$T = \sum_{i=r+1}^{l}x_i(1+R_i)+\sum_{i=k+1}^{m}x_i(1+R_i)+(k-l)x_l$$

3 Maximum Likelihood Estimation

The maximum likelihood estimator (MLE) of θ, denoted by $\hat{\theta}$ can be obtained by solving the equation

$$\frac{\partial \log L}{\partial \theta} = 0,\ \text{provided}\ \frac{\partial^2 \log L}{\partial \theta^2} < 0 \tag{7}$$

Using (6)

$$\begin{aligned} LogL &= \log C-(l-r+m-k)\log\theta-(T/\theta)+(k-l)\log\{1-\exp\left(-(x_{k+1}-x_l)/\theta\right)\} \\ &\quad +r\log\{1-\exp\left(-(x_{r+1}/\theta)\right)\} \end{aligned}$$

$$\begin{aligned} \frac{\partial \log L}{\partial \theta} &= \frac{-(l-r+m-k)}{\theta}+\frac{T}{\theta^2}-\left(\frac{k-l}{\theta^2}\right)\frac{(x_{k+1}-x_l)\exp(-(x_{k+1}-x_l)/\theta)}{1-\exp(-(x_{k+1}-x_l)/\theta)} \\ &\quad +\left(\frac{r}{\theta^2}\right)\frac{x_{r+1}\exp(-x_{r+1}/\theta)}{1-\exp(-x_{r+1}/\theta)} \end{aligned} \tag{8}$$

From (7) and (8) we have the likelihood equation

$$\begin{aligned}\theta &= \frac{T - \frac{(k-l)(x_{k+1}-x_l)\exp(-(x_{k+1}-x_l)/\theta)}{1-\exp(-(x_{k+1}-x_l)/\theta)} - \frac{rx_{r+1}\exp(-x_{r+1}/\theta)}{1-\exp(-x_{r+1}/\theta)}}{l-r+m-k} \\ &= \psi(\theta), \text{ a function of } \theta\end{aligned} \tag{9}$$

From equation (9), the MLE of θ can be obtained by using some suitable numerical iterative procedure such as the Newton-Raphson method.

From the MLE of θ, one can obtain the MLE of reliability and hazard function at time t as

$$\hat{R}(t) = \exp(-t/\hat{\theta}) \tag{10}$$

$$\hat{h}(t) = 1/\hat{\theta} \tag{11}$$

The Fisher information contained about θ is given by

$$I_x(\theta) = E(-\frac{\partial^2 \log L}{\partial \theta^2}) \tag{12}$$

where

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \theta^2} &= \frac{l-r+m-k}{\theta^2} - \frac{2T}{\theta^3} + \frac{2(k-l)(x_{k+1}-x_l)\exp(-(x_{k+1}-x_l)/\theta)}{\theta^3\{1-\exp(-(x_{k+1}-x_l)/\theta)\}} \\ &\quad - \frac{(k-l)(x_{k+1}-x_l)^2\exp(-(x_{k+1}-x_l)/\theta)}{\theta^4\{1-\exp(-(x_{k+1}-x_l)/\theta)\}^2} + \frac{2rx_{r+1}\exp(-(x_{r+1}/\theta))}{\theta^3\{1-\exp(-(x_{r+1}/\theta))\}} \\ &\quad - \frac{rx_{r+1}^2\exp(-x_{r+1}/\theta)}{\theta^4\{1-\exp(-x_{r+1}/\theta)\}^2}.\end{aligned} \tag{13}$$

The exact expression for the expectation of the above is difficult to obtain. However, in practice, we would need the estimate of variance, for which, Cohen [6] recommended it using the approximation

$$E(-\frac{\partial^2 \log L}{\partial \theta^2}) \approx -\frac{\partial^2 \log L}{\partial \theta^2}\Big]_{\theta=\widehat{\theta}} \tag{14}$$

Proposition: Asymptotically $\hat{\theta}$ has a normal distribution with mean θ and variance

$$1/E(-\frac{\partial^2 \log L}{\partial \theta^2}).$$

One may refer to Kendall and Stuart [19] or Rao [31] for the proof.

Thus, the asymptotic variance of MLE of reliability and hazard function at time t can be obtained respectively as

$$\begin{aligned}V(\hat{R}(t)) &= \left(\frac{\partial R(t)}{\partial \theta}\right)^2 V(\hat{\theta})\Big]_{\theta=\hat{\theta}} \\ &= \exp(-2t/\theta)\frac{t^2}{\theta^4}V(\hat{\theta})\Big]_{\theta=\hat{\theta}}\end{aligned} \tag{15}$$

and

$$V(\hat{h}(t)) = \left(\frac{\partial h(t)}{\partial \theta}\right)^2 V(\hat{\theta})\Bigg]_{\theta=\hat{\theta}}$$
$$= \frac{V(\hat{\theta})}{\theta^4}\Bigg]_{\theta=\hat{\theta}} \tag{16}$$

where

$$V(\hat{\theta}) = \frac{1}{E\left(\frac{-\partial^2 LogL}{\partial \theta^2}\right)} \tag{17}$$

In order to examine the existence of MLE, we have considered the second derivatives and observed that the second derivatives at the solution of the likelihood equation is negative which is verified by utilizing Visual Basic programming. Hence, standard errors of MLEs of reliability characteristics are calculated and shown in Section 5 and in Section 6.

Note that in the case of a complete sample, the MLE of the parameter θ is the sample mean $\overline{x}$, and its variance is θ^2/n.

The (1-α)100% asymptotic confidence interval for θ can be developed as

$$\hat{\theta} \mp Z_{\alpha/2}\sqrt{V(\hat{\theta})} \tag{18}$$

where $Z_{\alpha/2}$ is the upper $(\alpha/2)^{th}$-percentile of standard normal distribution.

For various sample sizes and censoring schemes, the standard errors are compared with that of based on a complete sample by using a simulation study in Section 5.

4 Expected Test Termination Time

For certain types of censoring schemes, such as, Type-II censoring, multiply Type-II censoring, progressive Type-II censoring, and multiply progressively Type-II censoring schemes, the test termination time is not fixed. In practical applications, it is often useful to have an idea of the duration of a time of whole life test. Therefore, it is important to compute the expected time required to complete a life test.

Tse and Yuen [36] computed the expected experiment times for the lifetimes of Weibull distributed under Type-II progressive censoring with random removals. Wu and Chang [41] described a comparison of the expected test termination time of the test based on exponential progressive Type-II censored data with random removals and tests based on uncensored complete sample data. Lin et al. [21] obtained the expected total test time based on Type-II progressively hybrid censored data with Weibull lifetimes.

This section we will present the expressions of the expected total test times (ETT) under progressive Type-II censoring and a complete sample without censoring. For progressively Type-II censoring, the ETT for the experiment is given by the expectation of the m^{th} order statistic X_m. We consider the following theorem to derive an explicit expression for the expectation of X_m,

Theorem 1. *Let* $\{X_{r+1}, X_{r+2}, \cdots, X_l, X_{k+1}, X_{k+2}, \cdots, X_m\}$ *denote a multiply progressive Type-II censored sample (ordered failure times) with censoring scheme* $(R_{r+1}, R_{r+2}, \cdots, R_l, R_{k+1}, R_{k+2}, \cdots, R_m)$. *The generalized spacing*

$$
\begin{aligned}
Z_{r+1} &= (n-r)X_{r+1} \\
Z_{r+2} &= (n-r-R_{r+1}-1)(X_{r+2}-X_{r+1}) \\
Z_l &= (n-r-R_{r+1}-R_{r+2}-\cdots-R_{l-1}-l+r+1)(X_l-X_{l-1}) \\
&\vdots \quad \vdots \\
Z_{k+1} &= (n-\sum_{j=r+1}^{l} R_j - k)X_{k+1} \\
Z_{k+2} &= (n-\sum_{j=r+1}^{l} R_j - k - R_{k+1}-1)(X_{k+2}-X_{k+1}) \\
&\vdots \quad \vdots \\
Z_m &= (n-\sum_{j=r+1}^{l} R_j - k - \sum_{j=k+1}^{m-1} R_j - m + k + 1)(X_m - X_{m-1})
\end{aligned}
$$

are independent random variables with $Z_{r+2}, \cdots, Z_l, Z_{k+1}, \cdots, Z_m$ *being exponential variates with mean* θ *and* $\frac{Z_{r+1}}{n-r} = X_{r+1}$ *being* $(r+1)^{th}$ *usual order statistic from a sample of size* n *from the exponential distribution with mean* θ. $\frac{Z_{k+1}}{n-\sum_{j=r+1}^{l} R_j - k} = X_{k+1}$ *being distributed as* $(k+1)^{th}$ *usual order statistic after* l^{th} *order statistic from exponential distribution with mean* θ.

The proof can be developed in a similar manner from Theorem 2.6 and Theorem 3.4, given by Balakrishnan and Aggarwala [2].

Here we can write

$X_m = X_{r+1} +$ (a linear combination of independent exponential random variables Z_{r+2} to Z_l) + $X_{k+1} +$ (a linear combination of independent exponential random variables Z_{k+2} to Z_m).

Hence, we have the following results:

$$
\begin{aligned}
E(X_{r+1}) &= \theta \sum_{i=1}^{r+1} \frac{1}{n-i+1} \\
E(X_j) &= E(X_{r+1}) + \theta \sum_{i=r+2}^{j} \frac{1}{n-\sum_{w=r+1}^{i-1} R_w - (i-1)} \\
&\qquad \text{for } j = r+2, r+3, \cdots, l \\
E(X_{k+1}) &= E(X_l) + \theta \sum_{i=r+1}^{k+1} \frac{1}{n-\sum_{w=r+1}^{l} R_w - (i-1)} \\
E(X_j) &= E(X_{k+1}) + \theta \sum_{i=k+2}^{j} \frac{1}{n-\sum_{w=r+1}^{l} R_w - \sum_{w=k+1}^{i-1} R_w - (i-1)} \\
&\qquad \text{for } j = k+2, k+3, \cdots, m.
\end{aligned}
$$

Thus, the expected test termination time $E(X_m)$ becomes

$$
E(X_m) = \theta \left[\begin{array}{l} \sum_{i=1}^{r+1} \frac{1}{n-i+1} + \sum_{i=r+2}^{l} \frac{1}{n-\sum_{w=r+1}^{i-1} R_w - i + 1} + \sum_{i=l+1}^{k+1} \frac{1}{n-\sum_{w=r+1}^{l} R_w - i + 1} + \\ \sum_{i=k+2}^{m} \frac{1}{n-\sum_{w=r+1}^{l} R_w - \sum_{w=k+1}^{i-1} R_w - i + 1} \end{array} \right] \tag{19}
$$

The expected test termination time in case of a complete sample reduces to

$$E(X_n) = \theta \sum_{i=1}^{n} \frac{1}{n-i+1} \tag{20}$$

To compare them, we compute the ratio of the two expected test termination times given in (19) and (20), which is defined as

$$RETT = \frac{E(X_m) \text{ under MPC for a sample of size } n}{E(X_n) \text{ under complete sampling for a sample of size } n} \tag{21}$$

Here RETT does not depend on the parameter θ. Suppose that an experimenter wants to observe the failure of at least $l - r + m - k$ units when the test is considered under multiply progressive Type-II censoring, then the RETT provides important information in determining whether the test termination time can be shortened significantly if a much larger sample of n units is used and the test is stopped at m^{th} failure observed in multiply progressive Type-II censoring. It is obvious that comparing these two expected test times analytically is very difficult. So, we will be calculating them numerically for various values of n,m, l, r, and k.

In the next section the simulation study is carried out to compute RETT for different sample sizes and censoring schemes.

5 Real example

Nelson ([23], 105, Table 1.1) presents the time to breakdown of an insulating fluid between electrodes at a voltage of 34 KV (minutes) for 19 specimens. As a numerical illustration, we have generated a multiply progressive Type-II censored sample of a lifetime based on this data set, for which exponential distribution appears to be adequate. The generated data is presented below with the censoring pattern.

Multiply progressively Type-II censored sample:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	-	-	0.96	2.78	3.16	4.15	-	6.50	8.01	8.27	31.75	33.91
R_i	-	-	1	0	0	1	-	1	0	1	1	2

As per our notations here $n = 19, r = 2, l = 6, k = 7$ and $m = 12$.

The maximum likelihood estimates of θ, $R(t)$, and $h(t)$ along with their standard errors(SE) and 95% confidence intervals(CI) are obtained as follow:

Reliability characteristics	MLE	Standard error	95% confidence interval
θ	6.92288	1.99904	(3.00476, 10.84100)
$R(t = 1.5)$	0.80519	0.05038	(0.70645, 0.90393)
$h(t)$	0.14445	0.04171	(0.06270, 0.22620)

Table 1: MLE, standard error and confidence interval for reliability characteristics

The expected test termination time under multiply censoring is 9.51293, whereas based on the complete sample, it is 24.56058, and RETTT = 0.38733.

Here we note that the expected test termination time under multiply censoring is smaller than that of under a complete sample.

6 Simulation

In this section, we present some experimental results, mainly to observe how the different methods behave for different sample sizes and for different censoring schemes. Maximum likelihood estimators for reliability characteristics like mean life-time, reliability at time $t = 1.5$, and hazard rate are computed along with their standard errors. 95% confidence intervals for the three reliability characteristics are also obtained. We compare the performances of the estimators under different cases with respect to their MSEs. We also compare the expected test termination time under different censoring schemes and under a complete sample. All the experiments have been performed on a Window 7 PC using Visual Basic.

We consider the value of parameter $\theta = 1.5, m = (12, 15, 18)$ for sample size $n = 20$ (i.e. 60%, 75% and 90% of n) and $m = (12, 15, 20, 25)$ for sample size $n = 30$ (i.e. 40%, 50%, 70% and 85% of n). Several combinations of r, l and k are considered appropriately as per the selected value of m. Different progressive censoring schemes with the two extreme censoring schemes of $(0^{m-1}, n-m)$ and $(n-m, 0^{m-1})$ are included in every case. The censoring schemes with withdrawals at early stages of failures, at later stages of failures, and withdrawals alternatively at failure stages are also considered. For simplicity in notation, we have used the notation as $(0^3, 20)$ for the progressive censoring scheme (0, 0, 0, 20). For usual Type-II censoring, the censoring scheme will be $(0^{m-1}, n-m)$. In order to evaluate the effect of differing levels of censoring as compared to the two extreme censoring schemes, some specifically chosen progressive censoring schemes are also included.

For each case, we simulated 1000 progressively Type-II censored samples from the exponential lifetime distribution with mean θ. To generate a sample based on multiply progressive Type-II censoring described in the Section 1, the following algorithm is used.

1 Simulate $r+1$ independent exponential variates $Z_1, Z_2, \cdots, Z_{r+1}$ with mean θ.

2 Set $X_{r+1} = \sum_{i=1}^{r+1} \frac{Z_i}{n-i+1}$

3 Simulate l-r-1 independent exponential variates Z_{r+2}, Z_{r+3} , ..., Z_l with mean θ.

4 Set $X_s = X_{r+1} + \sum_{j=r+2}^{s} \frac{Z_j}{n-\sum_{i=r+1}^{j-1} R_i - j+1}$, $s = r+2, r+3, \cdots, l$

5 Simulate $Z_{l+1}, Z_{l+2}, \cdots, Z_{k+1}$ exponential variates with mean θ.

6 Set $Xs = X_{k+1} + \sum_{j=k+2}^{s} \frac{Z_j}{n-\sum_{w=r+1}^{l} R_w - \sum_{w=k+1}^{j-1} R_w - j+1}$, $s = k+2, k+3, \cdots, m$

Then $\{X_{r+1}, X_{r+2}, \cdots, X_l, X_{k+1}, X_{k+2}, \cdots, X_m\}$ is the required multiply progressive Type-II right censored sample from the exponential distribution with mean θ, $\theta > 0$.

Table 2-5 report the average values of the 1000 simulated results. Expected test termination time under the different censoring schemes and for the complete sample are shown in Table 6 and Table 7. Conclusions are made from these tables.

From the results obtained in Table 2 to 5, we observe the following conclusions.

The maximum likelihood estimates of reliability characteristics are quite closed to their actual values under multiply progressive Type-II censoring, usual Type-II censoring and complete sample schemes.

For a given value of n, the standard error and length of the asymptotic confidence interval decrease as m or the observed number of failures $(l - r + m - k)$ increases.

The comparison of estimators obtained under different censoring schemes may not be highly justified. MSE and length of confidence interval do not change much with a change in the censoring schemes for a fixed number of observed life times. But the withdrawals made at early stages of failures increase the

r,l,k,m	Scheme	Parameter	MLE	SE	CI
	Complete sample	θ	1.48304	0.33974	(0.81715, 2.14893)
2,6,7,12	(0^4,1^4,4)	θ	1.47991	0.44459	(0.60851, 2.35131)
		R(t)	0.36292	0.11051	(0.14632, 0.57951)
		h(t)	0.67572	0.20300	(0.27784, 1.07359)
	(0^4, 0^3,8,0)	θ	1.47991	0.44459	(0.60851, 2.35131)
		R(t)	0.36292	0.11051	(0.14324, 0.57951)
		h(t)	0.67572	0.20300	(0.27784, 1.07359)
	(8,0^7)	θ	1.48010	0.44490	(0.60810, 2.35210)
		R(t)	0.36297	0.11057	(0.14625, 0.57968)
		h(t)	0.67563	0.20308	(0.27758, 1.07367)
	(1,0,1,0,1,0,1,0,4)	θ	1.47993	0.44461	(0.60848, 2.35137)
		R(t)	0.36293	0.11051	(0.14632, 0.57953)
		h(t)	0.67571	0.20300	(0.27782, 1.07360)
Type-II, m= 12	(0^{11},8)	θ	1.47990	0.44443	(0.60882, 2.35099)
		R(t)	0.36292	0.11047	(0.14640, 0.57944)
		h(t)	0.67572	0.20293	(0.27799, 1.07345)
2,7,9,15	(1,0,1,0^2,1,0^2,1,0,1)	θ	1.47707	0.39433	(0.70419, 2.24996)
		R(t)	0.36221	0.09820	(0.16974, 0.55468)
		h(t)	0.67701	0.18074	(0.32276, 1.03127)
	(0^2,2, 0^3,2,0^3,1)	θ	1.47707	0.39433	(0.70419, 2.24996)
		R(t)	0.36221	0.09820	(0.16974, 0.55468)
		h(t)	0.67701	0.18074	(0.32276, 1.03127)
	(5,0^{10})	θ	1.47701	0.39458	(0.70365, 2.25038)
		R(t)	0.36220	0.09827	(0.16960, 0.55480)
		h(t)	0.67704	0.18087	(0.32254, 1.03154)
	(0^8,5,0^2)	θ	1.47711	0.39426	(0.70436, 2.24986)
		R(t)	0.36222	0.09818	(0.16979, 0.55466)
		h(t)	0.67700	0.18070	(0.32282, 1.03117)

Table 2: Averaged values of MLE, SE and CI for reliability characteristics (n=20)

standard error as well as the length of the confidence interval, compared to the withdrawals at some last stages of failures.

From the results presented in Table 6 and 7, we notice the following:

For given values of n and m, the expected test termination time is much smaller in the case of multiply progressive Type-II censoring for all types of censoring schemes compared to the test based on a complete sample. That is though MSE is smaller in the case of the test based on a complete sample compared to a multiply Type-II progressive censoring scheme, the experimenter should prefer to multiply Type-II progressive censoring, if he wants to decrease the time of the test and hence cost of the experiment.

RETT increased rapidly and approached to 1 for withdrawals at early stages of failures compared to the withdrawals at some later stages of failures. It suggests that there is not much gain in shortening the experiment time in these cases compared to the test based on a complete sample.

The choice of higher withdrawals at later stages of failures gives a smaller expected test termination time compared to the other types of withdrawals.

As m increases, the expected test termination time gradually increases for multiply Type-II progressive

r, l, k, m	Scheme	Parameter	MLE	SE	CI
	Complete sample	θ	1.48304	0.33974	(0.81715, 2.14893)
2,6,7,15	$(0^7,1^5)$	θ	1.47746	0.39422	(0.70479, 2.25012)
		$R(t)$	0.36231	0.09815	(0.16994, 0.55468)
		$h(t)$	0.67684	0.18060	(0.32287, 1.03081)
	$(1,0,1,0,1,0,1,0^{4,}\ 1)$	θ	1.47747	0.39423	(0.70477, 2.25017)
		$R(t)$	0.36231	0.09815	(0.16994, 0.55469)
		$h(t)$	0.67683	0.18060	(0.32286, 1.03081)
	$(0^{10},5,0\ )$	θ	1.47746	0.39422	(0.70479, 2.25012)
		$R(t)$	0.36231	0.09815	(0.16994, 0.55468)
		$h(t)$	0.67684	0.18060	(0.32287, 1.03081)
	$(5,0^{11})$	θ	1.47751	0.39428	(0.70471, 2.25030)
		$R(t)$	0.36232	0.09816	(0.16993, 0.55471)
		$h(t)$	0.67682	0.18061	(0.32282, 1.03082)
Type-II, $m = 15$	$(0^{14},5)$	θ	1.47746	0.39410	(0.70502, 2.24990)
		$R(t)$	0.36231	0.09812	(0.17000, 0.55462)
		$h(t)$	0.67684	0.18054	(0.32298, 1.03070)
1,7,8,18	$(1,0^5,1,0^9)$	θ	1.47779	0.35783	(0.77644, 2.17915)
		$R(t)$	0.36239	0.08907	(0.18782, 0.53697)
		$h(t)$	0.67668	0.16385	(0.35553, 0.99783)
	$(1,\ 0^{14,}\ 1)$	θ	1.47779	0.35783	(0.77644, 2.17915)
		$R(t)$	0.36239	0.08907	(0.18782, 0.53697)
		$h(t)$	0.67668	0.16385	(0.35553, 0.99783)
	$(0^6,2,0^9)$	θ	1.47779	0.35780	(0.77645, 2.17914)
		$R(t)$	0.36239	0.08907	(0.18782, 0.53697)
		$h(t)$	0.67668	0.16385	(0.35553, 0.99783)
Type-II, $m = 18$	$(0^{17},2)$	θ	1.47781	0.35780	(0.77652, 2.17910)
		$R(t)$	0.36240	0.08906	(0.18784, 0.53695)
		$h(t)$	0.67668	0.16383	(0.35556, 0.99779)

Table 3: Averaged values of MLE, SE and CI for reliability characteristics (n = 20).

censoring schemes, which must be true.

The expected test termination time under the usual Type-II censoring scheme is smaller than that of under a multiply Type-II progressive censoring scheme, and the test is based on a complete sample.

7 Conclusion

In this paper, we discussed multiply progressive Type-II censoring schemes and discussed the statistical inference based on exponential life-time data. We compared the performance of MLE of the reliability characteristics and expected test termination time for the exponential life-time model when the data are multiply progressive Type-II censored. The biases of MLE of the reliability characteristics are smaller for all types of censoring schemes of multiply progressive Type-II censoring as well as for the test based on a complete sample. The standard errors as well as lengths of the confidence intervals decrease for the censoring schemes with withdrawals at some last stages of failures compared to the early stages of failures. One important point that should be mentioned here is that, though MSE is smaller in the case of a test based on a complete sample, ETT is much smaller in the case of multiply progressive Type-II censoring

r, l, k, m	Scheme	Parameter	MLE	SE	CI
	Complete sample	θ	1.48264	0.27521	(0.94323, 2.02206)
2,6,7,12	(0^4,1^4,14)	θ	1.49609	0.44962	(0.61483, 2.37735)
		$R(t)$	0.36692	0.11056	(0.15022, 0.58362)
		$h(t)$	0.66841	0.20088	(0.27469, 1.06213)
	(8,0^7,10)	θ	1.49612	0.44965	(0.61480, 2.37743)
		$R(t)$	0.36693	0.11056	(0.15022, 0.58363)
		$h(t)$	0.66834	0.20088	(0.27467, 1.06213)
	(1^4,0^4 ,14)	θ	1.49601	0.44963	(0.61482, 2.37738)
		$R(t)$	0.36692	0.11056	(0.15022, 0.58362)
		$h(t)$	0.66841	0.20088	(0.27468, 1.06213)
	(2,0,2,0,2,0,2,0,10)	θ	1.49607	0.44963	(0.61482, 2.37738)
		$R(t)$	0.36692	0.11056	(0.15022, 0.58362)
		$h(t)$	0.66841	0.20088	(0.27468, 1.06213)
Type-II, $m = 12$	(0^{11},18)	θ	1.49617	0.44961	(0.61494, 2.37741)
		$R(t)$	0.36694	0.11055	(0.15026, 0.58362)
		$h(t)$	0.66837	0.20085	(0.27470, 1.06204)
2,6,7,15	(0^7,1^4,11)	θ	1.48951	0.39731	(0.71077, 2.26824)
		$R(t)$	0.36530	0.09813	(0.17297, 0.55763)
		$h(t)$	0.67136	0.17908	(0.32036, 1.02236)
	(1,0,1,0,1,0,1,0^4,11)	θ	1.48951	0.39732	(0.71077, 2.26825)
		$R(t)$	0.36530	0.09813	(0.17297, 0.55763)
		$h(t)$	0.67136	0.17908	(0.32036, 1.02236)
	(0^{10},5,10)	θ	1.48951	0.39731	(0.71077, 2.26824)
		$R(t)$	0.36530	0.09813	(0.17297, 0.55763)
		$h(t)$	0.67136	0.17908	(0.32036, 1.02236)
	(3^4,0^7,3)	θ	1.48955	0.39735	(0.71074, 2.26837)
		$R(t)$	0.36531	0.09813	(0.17297, 0.55765)
		$h(t)$	0.67134	0.17909	(0.32033, 1.02235)
	(15,0^{11},)	θ	1.48959	0.39740	(0.71069, 2.26849)
		$R(t)$	0.36532	0.09814	(0.17296, 0.55767)
		$h(t)$	0.67133	0.17910	(0.32023, 1.02236)
Type-II, $m = 15$	(0^{14},15)	θ	1.48958	0.39731	(0.71085, 2.26832)
		$R(t)$	0.36532	0.09812	(0.17300, 0.55763)
		$h(t)$	0.67133	0.17906	(0.32037, 1.02229)

Table 4: Averaged values of MLE, SE and CI for reliability characteristics (n = 30).

r, l, k, m	Scheme	Parameter	MLE	SE	CI
	Complete sample	θ	1.48264	0.27521	(0.94323, 2.02206)
2,8,10,20	(1^3, 0^3,1^3,0^6,4)	θ	1.48628	0.34085	(0.81822, 2.15434)
		$R(t)$	0.36450	0.08436	(0.19915, 0.52985)
		$h(t)$	0.67282	0.15430	(0.37040, 0.97525)
	(0^{10}, 2^5,0)	θ	1.48629	0.34084	(0.81825, 2.15433)
		$R(t)$	0.36450	0.08436	(0.19916, 0.52984)
		$h(t)$	0.67282	0.15429	(0.37041, 0.97523)
	(10,0^{14},0)	θ	1.48624	0.34093	(0.81802, 2.15446)
		$R(t)$	0.36450	0.08438	(0.19910, 0.52988)
		$h(t)$	0.67284	0.15434	(0.37033, 0.97535)
Type-II $m = 20$	(0^{19},10)	θ	1.48644	0.34081	(0.81844, 2.15444)
		$R(t)$	0.36454	0.08434	(0.19932, 0.52985)
		$h(t)$	0.67275	0.15425	(0.37042, 0.97508)
1,9,12,25	(2,0^7,2,0^{11},1)	θ	1.48565	0.30321	(0.89137, 2.07993)
		$R(t)$	0.36434	0.07508	(0.21799, 0.51149)
		$h(t)$	0.67311	0.13737	(0.40385, 0.94236)
	(5,0^{20})	θ	1.48564	0.30325	(0.89127, 2.08001)
		$R(t)$	0.36434	0.07509	(0.21717, 0.51151)
		$h(t)$	0.67311	0.13740	(0.40382, 0.94241)
	(0^6,1^2,0^{11},1,2)	θ	1.48565	0.30321	(0.89137, 2.07993)
		$R(t)$	0.36434	0.07508	(0.21719, 0.51149)
		$h(t)$	0.67311	0.13737	(0.40385, 0.94236)
Type-II $m = 25$	(0^{24},5)	θ	1.48571	0.30311	(0.89161, 2.07981)
		$R(t)$	0.36436	0.07505	(0.21726, 0.51146)
		$h(t)$	0.67308	0.13732	(0.40393, 0.94223)

Table 5: Averaged values of MLE, SE and CI for reliability characteristics (n = 30).

r, l, k, m	Scheme	$E(X_m)$	$E(X_n)$	RETT
2,6,7,12	(0^4,1^4,4)	1.53823	5.32432	0.28891
	(0^7, 8,0)	2.61762	5.32432	0.49163
	(8,0^7)	4.42129	5.32502	0.83029
	(1,0,1,0,1,0,1,0,4)	1.68903	5.32439	0.31723
Type-II, m= 12	(0^{11},8)	1.30214	5.32431	0.24457
2,6,7,15	(0^7,1^5)	2.67484	5.31550	0.50322
	(1,0,1,0,1,0,1,0^4, 1)	3.28682	5.31555	0.61834
	(0^{10},5,0)	3.17319	5.31550	0.59697
	(5,0^{11})	4.81873	5.31568	0.90651
2,7,9,15	(1,0,1,0^2,1,0^2,1,0,1)	3.09076	5.31412	0.58161
	(0^2,2, 0^3,2,0^3,1)	3.15789	5.31412	0.59763
	(5,0^{10})	4.81713	5.31391	0.90651
	(0^8,5,0^2)	3.69999	5.31425	0.69624
Type-II, m=15	(0^{14},5)	1.94199	5.31553	0.36539
1,7,8,18	(1,0^5,1,0^9)	5.08684	5.31672	0.95676
	(1, 0^{14}, 1)	3.75682	5.31672	0.70661
	(0^6,2,0^9)	5.03461	5.31674	0.94694
Type-II, m=18	(0^{17},2)	3.10006	5.31679	0.58307

Table 6: Expected test termination time for different censoring schemes and complete sample and their ratio for n = 20.

r, l, k, m	Scheme	E(X_m)	E(X_n)	RETT
2,6,7,12	(0^4,1^4,14)	0.79280	5.97685	0.13264
	(8,0^7,10)	1.08065	5.97697	0.18080
	(1^4,0^4 ,14)	0.86069	5.97690	0.14400
	(2,0,2,0,2,0,2,0,10)	0.95579	5.97690	0.15991
Type-II, m=12	(0^{11},18)	0.74791	5.97720	0.12513
2,6,7,15	(0^7,1^4,11)	1.07254	5.95056	0.18024
	(1,0,1,0,1,0,1,0^4, 11)	1.17311	5.95058	0.19714
	(0^{10},5,10)	1.05035	5.95056	0.17651
	(3^4,0^7,3)	2.27540	5.95075	0.38237
	(15,0^{11})	4.77672	5.95088	0.80269
Type-II, m=15	(0^{14},15)	1.00809	5.95086	0.16940
2,8,10,20	(1^3, 0^3,1^3,0^6,4)	2.33922	5.93766	0.39396
	(0^{10}, 2^5,0)	3.33900	5.93770	0.56234
	(10,0^{14},0)	5.26587	5.93751	0.88688
Type-II, m=20	(0^{19},10)	1.58458	5.93832	0.26684
1,9,12,25	(2,0^7,2,0^{11},1)	4.13626	5.93516	0.69691
	(5,0^{20})	5.63486	5.93511	0.95172
	(0^6,1^2,0^{11},1,2)	3.19346	5.93516	0.53806
Type-II, m=25	(0^{24},5)	2.54302	5.93540	0.42845

Table 7: Expected test termination time for different censoring schemes and complete sample and their ratio for n = 30.

for all the types of censoring schemes. This is definitely one major advantage of a multiply progressive Type-II censoring scheme for saving time and cost of the experiment.

From this study, once again, we can see that there is always a trade-off between (i) total time on the test, (ii) saving experimental units, and (iii) efficiency in estimation. The computation formulas and results provided in this paper give a guideline on planning an experiment to compromise these three concerns. Further investigation on obtaining optimal censoring schemes for given values (n, r, l, k, m) would be of interest in experimental planning.

Acknowledgements

The authors are highly thankful to the Editor in Chief, the Associate Editor, and the anonymous referees for their fruitful comments, which immensely helped to improve the quality and presentation of the article.

References

[1] N. Balakrishnan. "On the maximum likelihood estimation of the location and scale parameters of exponential distribution based on multiply Type-II censored samples." In: *Journal of Applied Statistics* 17(1) (1990), pp. 61–65.

[2] N. Balakrishnan and R. Aggarwala. *Progressive censoring: Theory Methods and Application.* Boston: Birkhauser Publishers, 2000.

[3] N. Balakrishnan and R. A. Sandhu. "Best Linear Unbiased and Maximum Likelihood Estimation for Exponential Distributions under General progressive Type-II Censored Samples." In: *Sankhya-B* 58 (1996), pp. 1–9.

[4] K. Balasubramanian and N. Balakrishnan. "Estimation for one and two parameter exponential distributions under multiply Type –II censoring." In: *Statistical Papers* 33 (1992), pp. 203–216.

[5] D. R. Barot and M. N. Patel. "Bayesian estimation of reliability indexes for cold standby system Under general progressive type –II censored data." In: *International journal of quality and reliability management.* 31(3) (2014), pp. 311–343.

[6] A. C. Cohen. "Maximum likelihood estimation in Weibull distribution based on complete and censored samples." In: *Technometrics* 7 (1965), pp. 579–588.

[7] A. C. Cohen. "Progressively censored samples in life testing." In: *Technometrics* 5 (1963), pp. 327–329.

[8] A. C. Cohen and N. J. Norgaard. "Progressively censored sampling in the three parameter gamma distribution." In: *Technometrics* 19 (1977), pp. 333–340.

[9] B. Epstein. "Estimation of parameters of two parameter exponential distribution from censored samples." In: *Technometrics* 12 (1960), pp. 399–407.

[10] B. Epstein. "Truncated life tests in exponential case." In: *Annals of Mathematical Statistics* 2 (1954), pp. 554–564.

[11] B. Epstein and M. Sobel. "Life testing." In: *Journal of American Statistical Association* 48 (1953), pp. 486–502.

[12] B. Epstein and M. Sobel. "Some problems relevant to life testing from an exponential distribution." In: *Ann Math. Stat.* 25 (1954), pp. 373–381.

[13] H. Fei and F. Kong. "Interval estimations for one-and two-parameter exponential distributions under multiply Type-II censoring." In: *Communi. in Statistics- Theo. & Meth.* 23 (1994), pp. 1717–1733.

[14] A. J. Fernandez. "On Estimating Exponential Parameters with General Type-II Progressive Censoring." In: *Journal of Statistical Planning and Inference* 121 (2004), pp. 136–147.

[15] N. K. Gadhvi. "Estimation for two exponential life time models under joint multiply Type-II censoring." In: *Journal of the Egyptian Mathematical Society* 29(1) (2021), pp. 1–16.

[16] K. A. Gajjar and M. N. Patel. "Estimation for a mixture of exponential distributions based on progressively Type-II censored sample." In: *International Jr. Agricult. Stat. Sci.* 4(1) (2008), pp. 169–176.

[17] D. I. Gibbons and Vance L. C. "Estimators for the 2-parameter Weibull distribution with progressively censored samples." In: *IEEE Transactions on Reliability* 32 (1983), pp. 95–99.

[18] G. Hofmann et al. "An asymptotic approach to progressive censoring." In: *J. Stat. Planing and Inference* 32 (2005), pp. 95–99.

[19] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics-2*. London: Charles Griffin and Co., 1973.

[20] J. F. Lawless. *Statistical models and methods for life time data.* New York: John Wiley and Sons, 1973.

[21] C. T. Lin, H.K.T. Ng, and P. S. Chan. "Statistical Inference of Type-II Progressively Hybrid Censored Data with Weibull Lifetimes." In: *Communi. in Statistics Theo. & Meth.* 38(10) (2009), pp. 1710–1729.

[22] N. R. Mann. "Best linear invariant estimation for Weibull parameters under progressive censoring." In: *Technometrics* 13 (1971), pp. 521–533.

[23] W. B. Nelson. *Applied life data analysis*. New York: John Wiley and Sons, 1973.

[24] M. W. Patel and M.N. Patel. "Estimation of two parameter geometric distribution under Type-II and multiply Type-II censoring." In: *Statistical Methods* 8(2) (2006), pp. 194–205.

[25] M. W. Patel and M.N. Patel. "Progressively Type-II censored samples from geometric life time model." In: *Journal of probability and statistical science* 5(1) (2007), pp. 81–95.

[26] M.N. Patel. "Bayesian prediction in exponential distribution with random sample size under multiply Type-II censoring." In: *Revista evista Invesigacion operacional* 41(3) (2020), pp. 389–399.

[27] M.N. Patel. "MLE for exponential model with changing failure rates based on two stage progressively multiply Type-II censored samples." In: *Journal of Probability and Statistical Science* 4(2) (2006), pp. 221–232.

[28] M.N. Patel. "Progressively censored samples from mixture of two exponential distributions with changing parameters." In: *IAPQR Transactions* 23(2) (1998), pp. 83–93.

[29] M.N. Patel and A.V. Gajjar. "Maximum likelihood estimation in compound exponential failure model with changing failure rates based on Type-I progressively censored and group censored samples." In: *Communi. in Statistics, Theo. & Meth.* 21(10) (1992), pp. 2899–2908.

[30] M.N. Patel and A.V. Gajjar. "Some results on maximum likelihood estimators of exponential distribution under Type-I progressive censoring with changing failure rates." In: *Communi. in Statistics, Theo. & Meth.* 24(9) (1995), pp. 2421–2435.

[31] C. R. Rao. *Linear Statistical Inference and Its Applications.* New York: John Wiley & Sons, 1975.

[32] J. B. Shah and M. N. Patel. "Bayes estimation under asymmetric loss function from multiply Type-II censored Rayleigh data." In: *Jr. of Indian Statistical Association* 46(1) (2008), pp. 67–78.

[33] J. B. Shah and M. N. Patel. "Estimator of parameter in a Pareto distribution from multiply type- II censored data." In: *International Jr. Agricult. Stat. Sci.* 3(2) (2007), pp. 581–588.

[34] P. V. Sukhatme. "Tests of Significance for samples of the chi-square population with two degrees of freedom." In: *Ann. Eugen.* 8 (1937), pp. 52–56.

[35] D. R. Thomas and W.M. Wilson. "Linear order statistic estimation for the two- parameter Weibull and extreme value distributions from Type-II progressively censored samples." In: *Technometrics* 14 (1972), pp. 679–691.

[36] S. K. Tse and H. K. Yuen. "Expected experiment times for the Weibull distribution under progressive censoring with random removals." In: *Journal of Applied Statistics* 25 (1998), pp. 75–83.

[37] S.K. Upadhyay, U. Singh, and V. Shastri. "Estimation of exponential parameters under multiply Type-II censoring." In: *Communi. in Statistics, Simu. & Compu.* 25(3) (1996), pp. 801–815.

[38] R. Viveros and N. Balakrishnan. "Interval estimation of parameters of life from progressively censored data." In: *Technometrics* 36(1) (1994), pp. 84–91.

[39] D. R. Wingo. "Solution of the three-parameter Weibull equations by constrained modified quasilinearization (progressively censored samples)." In: *IEEE Trans. on Reliab.* 22 (1973), pp. 96–102.

[40] J. Y. Wong. "Simultaneously estimating the three Weibull parameters from progressively censored samples." In: *Microelectronics and Reliability* 33 (1993), pp. 2217–2224.

[41] J. Wu S and C. T. Chang. "Parameter Estimations Based on Exponential Progressive Type-II Censored Data with Binomial Removals." In: *Information and Management Sciences* 13(3) (2002), pp. 37–46.

[42] Q. Xie, N. Balakrishnan, and D. H. Han. "Exact inference and optimal censoring scheme for a simple step-stress model under progressive Type-II censoring." In: *In advances in mathematical and statistical modeling*. Ed. by Jan Fagerberg, David C. Mowery, and Richard R. Nelson. Boston: Birkhäuser, 2008. Chap. 9, pp. 107–137.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Calibration Estimation of Population Mean using Log Function in Stratified Sampling

Menakshi Pachori [1], Neha Garg [2], Anant Patel [3] and Rajesh Tailor [4]

1. *School of Sciences, Indira Gandhi National Open University, New Delhi-110068, India*
2. *Corresponding Author, School of Sciences, Indira Gandhi National Open University, New Delhi-110068, India*
3. *School of Sciences, Indira Gandhi National Open University, New Delhi-110068, India*
4. *School of Studies in Statistics, Vikram University, Ujjain-456010 (M.P.), india*

Received: 12 February 2022 / Revised: 10 April 2023 / Accepted: 16 July 2023

Abstract

In this paper, a modified calibration estimator for the population mean in stratified random sampling using log mean of the auxiliary variable has been derived, and the result so obtained is being extended in case of stratified double sampling. The simulation study is carried out to check the performance of the suggested estimators over the existing estimators based on the generated artificial population. The result reveals that the proposed estimators are more efficient as compared with the estimator given by [9], [10] and [12].

Keywords: Auxiliary Information, Calibration Estimation, Stratified Sampling, Stratified Double Sampling, Log, Mean.

1 Introduction

In the recent years, the calibration estimation has become very popular. It is defined as a technique for adjusting weights for estimating population parameters based on auxiliary information in survey sampling. Calibration Estimation has gained prominence after Deville and Sarndal (1992) defined the calibration estimation as a procedure of minimizing a distance function subject to calibration constraints. Following [2], many other researchers such as [5], [11], [6], [8], [7], [1], [3], [4], etc, have defined calibration estimators using different calibration constraints under various sampling schemes.

The purpose of this study is to define a new calibration estimator for population mean using log function of mean of the auxiliary variable as a calibration constraint, under stratified and stratified double random sampling schemes. The simulation study has been presented and the results of the proposed estimator in comparison with the estimators suggested by [10] and [12] based on the artificial datasets is generated from exponential and Chi-square distributions.

Singh, Horn and Yu (1998)

The calibration estimator for population mean $\bar{Y}$ under stratified random sampling given by [9] is

$$\bar{y}_{sh} = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{1}$$

where Ω_h are the new calibrated weights obtained by minimizing the Chi-square distance measure $\sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{Q_h W_h}$, subject to the calibration constraint:

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h \tag{2}$$

where

$$\Omega_h = W_h \ + (W_h Q_h \bar{x}_h) \left[\frac{(\sum_{h=1}^{L} W_h \bar{X}_h - \sum_{h=1}^{L} W_h \bar{x}_h)}{(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2)} \right] \tag{3}$$

Singh (2003)

Similarly the calibration estimator for population mean $\bar{Y}$ under stratified random sampling given by [10] is

$$\bar{y}_s = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{4}$$

subject to the two calibration constraints:

$$\sum_{h=1}^{L} \Omega_h = 1 \tag{5}$$

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h \tag{6}$$

Minimization of Chi-square distance measure, given by [10] subject to the calibration constraints as mentioned above, the calibrated weight is given as:

$$\Omega_h = W_h \ + \left[\frac{(W_h Q_h \bar{x}_h)(\sum_{h=1}^{L} W_h Q_h) - (W_h Q_h)(\sum_{h=1}^{L} W_h Q_h \bar{x}_h)}{(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2)(\sum_{h=1}^{L} W_h Q_h) - (\sum_{h=1}^{L} W_h Q_h \bar{x}_h)^2} \right] (\bar{X} - \sum_{h=1}^{L} W_h \bar{x}_h) \tag{7}$$

Tracy, Singh and Arnab (2003)

The calibration estimator for population mean $\bar{Y}$ under the stratified random sampling defined by [12] is given as:

$$\bar{y}_{tr} = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{8}$$

subject to the two calibration constraints:

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h \tag{9}$$

$$\sum_{h=1}^{L} \Omega_h s_{hx}^2 = \sum_{h=1}^{L} W_h S_{hx}^2 \tag{10}$$

Minimizing Chi-square distance measure, given by [12] subject to the calibration constraints as mentioned above, the calibrated weights are obtained as:

$$\begin{aligned} \Omega_h = W_h + W_h Q_h \bar{x}_h & \frac{(\sum_{h=1}^{L} W_h Q_h s_{hx}^4)(\sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h)) - (\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2)(\sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2))}{(\sum_{h=1}^{L} W_h Q_h s_{hx}^4)(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2) - (\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2)^2} \\ + W_h Q_h s_{hx}^2 & \frac{(\sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2))(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2) - (\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2)(\sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h))}{(\sum_{h=1}^{L} W_h Q_h s_{hx}^4)(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2) - (\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2)^2} \end{aligned} \tag{11}$$

2 Proposed Calibration Estimator

Calibration Estimator under Stratified Sampling

Considering the population of size N which has been divided into L homogeneous subgroups called strata consisting of N_h units in h^{th} stratum such that $\sum_{h=1}^{L} N_h = N$. A sample of size n_h is drawn from the h^{th} stratum using Simple Random Sampling Without Replacement (SRSWOR), where n is the required sample size, and $\sum_{h=1}^{L} n_h = n$.

Suppose y_{hi} and x_{hi} is the i^{th} unit of the study and auxiliary variables, respectively, in the h^{th} stratum for i= 1, 2, ..., n_h and h = 1, 2, . . . , L. $\sum_{h=1}^{L} W_h = \frac{N_h}{N}$ is the stratum weight and $\sum_{h=1}^{L} f_h = \frac{n_h}{N}$ is the sample fraction.

The proposed calibration estimator for stratified random sampling using log of mean is given as

$$\bar{y}_L = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{12}$$

where the calibration weights are such chosen in order to minimize the Chi-square distance measure given as

$$\sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{W_h Q_h} \tag{13}$$

subject to the following calibration constraint

$$\sum_{h=1}^{L} \Omega_h \log \bar{x}_h = \sum_{h=1}^{L} W_h \log \bar{X}_h \tag{14}$$

The Lagrange function is given as

$$L = \sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{W_h Q_h} - 2\lambda(\sum_{h=1}^{L} \Omega_h \log \bar{x}_h - \sum_{h=1}^{L} W_h \log \bar{X}_h) \tag{15}$$

where λ is the Lagrange's multiplier. To find the optimum value of Ω_h we differentiate the Lagrange function given in equation (15) with respect to Ω_h and equate it to zero. Thus the calibration weight is obtained as

$$\Omega_h = W_h + \lambda(W_h Q_h \log \bar{x}_h). \tag{16}$$

Here λ is determined by substituting the value of Ω_h from equation (16) to equation (14), so this leads to a calibrated weight as

$$\Omega_h = W_h + (W_h Q_h \log \bar{x}_h) \left[\frac{(\sum_{h=1}^{L} W_h(\log \bar{X}_h - \log \bar{x}_h))}{(\sum_{h=1}^{L} W_h Q_h (\log \bar{x}_h)^2} \right] \tag{17}$$

Thus, on substituting the value of Ω_h from equation (17) in equation (12), we get the proposed calibrated estimator as

$$\bar{y}_L = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}_L \left[\sum_{h=1}^{L} W_h(\log \bar{X}_h - \log \bar{x}_h) \right] \tag{18}$$

where

$$\hat{\beta}_L = \left[\frac{(\sum_{h=1}^{L} W_h Q_h \log \bar{x}_h \bar{y}_h)}{(\sum_{h=1}^{L} W_h Q_h (\log \bar{x}_h)^2} \right] \tag{19}$$

Calibration Estimator under Stratified Double Sampling

The result attained above is now extended in case of stratified double sampling. For this scheme a preliminary sample of size m_h units as a first phase sample is drawn by using SRSWOR and a subsample of n_h units is drawn from the preliminary sample of size m_h units by SRSWOR. Let $\bar{x}_h^* = \frac{1}{m_h} \sum_{i=1}^{m_h} x_{hi}$ be the first phase sample mean and $\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$ and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ are the second phase sample means of auxiliary variable and study variable, respectively. Thus, the proposed calibration estimator in case of stratified double sampling are given as:

$$\bar{y}_{L\_d} = \sum_{h=1}^{L} \Omega_h^* \bar{y}_h \tag{20}$$

where the calibration weights Ω_h^* are so chosen in order to minimize the Chi-square distance measure

$$\sum_{h=1}^{L} \frac{(\Omega_h^* - W_h)^2}{W_h Q_h} \tag{21}$$

subject to the following calibration constraint given as

$$\sum_{h=1}^{L} \Omega_h^* \log \bar{x}_h = \sum_{h=1}^{L} W_h \log \bar{x}_h^* \tag{22}$$

where $W_h = \frac{N_h}{N}$ are the known stratum weights. The Lagrange function is given as

$$L = \sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{W_h Q_h} + 2\lambda (\sum_{h=1}^{L} \Omega_h^* \log \bar{x}_h - \sum_{h=1}^{L} W_h \log \bar{x}_h^*) \tag{23}$$

where λ is the Lagrange's multiplier. In order to find the optimum value of Ω_h^*, we differentiate the Lagrange function with respect to Ω_h^* and equate it to zero. Thus the calibration weight obtained is given as

$$\Omega_h^* = W_h + \lambda (W_h Q_h \log \bar{x}_h) \tag{24}$$

Here λ is obtained by substituting the value of Ω_h^* from equation (24) to equation (22), thus this leads to a calibrated weight as

$$\Omega_h^* = W_h + (W_h Q_h \log \bar{x}_h) \left[\frac{(\sum_{h=1}^{L} W_h (\log \bar{x}_h^* - \log \bar{x}_h)}{(\sum_{h=1}^{L} W_h Q_h \log \bar{x}_h^2)} \right] \tag{25}$$

Thus, substituting the value of Ω_h^* from equation (25) in equation (20), the proposed calibrated estimator attained is

$$\bar{y}_{L\_d} = \sum_{h=1}^{L} W_h \bar{y}_h + \left[\hat{\beta}_{L\_d} \sum_{h=1}^{L} W_h (\log \bar{x}_h^* - \log \bar{x}_h) \right] \tag{26}$$

where

$$\hat{\beta}_{L\_d} = \left[\frac{(\sum_{h=1}^{L} W_h Q_h \log \bar{x}_h \bar{y}_h)}{(\sum_{h=1}^{L} W_h Q_h \log \bar{x}_h^2)} \right] \tag{27}$$

For different values of Q_h, we can obtain the different forms of the calibration estimators defined in equations (18) and equation (26).

3 Simulation Study

A simulation study was carried out on a real data and two artificially generated data sets. Following the principle of Proportional allocation, the random samples are drawn using SRSWOR from each stratum. A simulated study is done by generating 25,000 samples in R-software.

Real Data

The real data used here is the wheat population for the year 2017 (http://data.icrisat.org/dld/src/crops.html), which has been divided into three strata of unequal sizes. The X variable is the area of the field (in ha) and Y variable is the yield (kg per ha). Here, population of size N=109 is divided in h=3 strata where the stratum sizes are N1=37, N2=26 and N3=46 (such that N1+N2+N3=37+26+46 = 109), respectively. 25,000 samples of sizes n=10, 15, 20, 25 and 30 are generated under stratified sampling. In case of stratified double sampling, the first phase sample of size m=50 is drawn.

Artificial Data

To assess the performance of the proposed calibrated estimators, a simulation study is carried out by generating a finite population of size N=4000 for h=3 (strata) where the stratum sizes are N1=1000, N2=1200 and N3=1800 (such that N1+N2+N3=1000+1200+1800 = 4000), respectively. 25,000 samples of sizes n=100, 200, 300, 400 and 500 are generated under stratified sampling. In case of stratified double sampling, the first phase sample size m=1000 and second phase sample sizes n=100, 200, 300, 400 and 500 are considered. The values of the auxiliary variable X are generated using Exponential and Chi-Square distributions with different values of the parameters for each stratum and the variable Y is generated using the following models:

Exponential Distribution

1^{st} strata: $X_1 = Exp(1000, 15)$ *and* $Y_1 = 100 + (\beta_1 * X_1) + \varepsilon_l$ where $\beta_1 = 1$ *and* $\varepsilon_l \sim N(0, 2)$
2^{nd} strata: $X_2 = Exp(1200, 16)$ *and* $Y_1 = 200 + (\beta_2 * X_2) + \varepsilon_2$ where $\beta_2 = 2$ *and* $\varepsilon_2 \sim N(0, 3)$
3^{rd} strata: $X_3 = Exp(1800, 20)$ *and* $Y_1 = 300 + (\beta_3 * X_3) + \varepsilon_3$ where $\beta_3 = 3$ *and* $\varepsilon_3 \sim N(0, 4)$

Chi-Square Distribution

1^{st} strata: $X_1 = \chi^2(1000, 15)$ *and* $Y_1 = 50 + (\beta_1 * X_1) + \varepsilon_l$ where $\beta_1 = 0.25$ *and* $\varepsilon_l \sim N(0, 4)$

2^{nd} strata: $X_2 = \chi^2(1200, 16)$ *and* $Y_1 = 100 + (\beta_2 * X_2) + \varepsilon_2$ where $\beta_2 = 0.50$ *and* $\varepsilon_2 \sim N(0, 5)$

3^{rd} strata: $X_3 = \chi^2(1800, 17)$ *and* $Y_1 = 150 + (\beta_3 * X_3) + \varepsilon_3$ where $\beta_3 = 0.75$ *and* $\varepsilon_3 \sim N(0, 6)$

The performance of the suggested estimators is measured using percentage absolute relative bias (%ARB) and percentage relative root mean squared error (%RRMSE) in case of both sampling schemes which are computed as:

$$\%ARB(\bar{y}_\alpha) = \frac{1}{25000}\sum_{i=1}^{25000}\left|\frac{(\bar{y}_{i\alpha}-\bar{Y})}{\bar{Y}}\right| \times 100; \alpha = tr, sh, s, L, dtr, dsh, ds, dL \tag{28}$$

$$\%RRMSE(\bar{y}_\alpha) = \sqrt{\frac{1}{25000}\sum_{i=1}^{25000}\left(\frac{(\bar{y}_{i\alpha}-\bar{Y})}{\bar{Y}}\right)^2} \times 100 \ ; \ \alpha = tr,\ sh, s, L, dtr,\ \ dsh,\ \ ds,\ \ dL \tag{29}$$

The results attained in case of stratified sampling and stratified double sampling are given in the following Tables:

Table 1: Percentage Absolute Relative Bias for Real Population under Stratified Sampling

Q_h	Sample size (n)	$\mathbf{ARB}\,(\bar{y}_{tr})$	$\mathbf{ARB}\,(\bar{y}_{sh})$	$\%\mathbf{ARB}\,(\bar{y}_s)$	$\mathbf{ARB}\,(\bar{y}_L)$
$Q_h = 1$	10	0.6236	1.3740	1.2499	1.1641
	15	8.6440	0.8157	0.4425	0.5883
	20	8.9973	0.4625	0.2051	0.4181
	25	9.6235	0.4519	0.1430	0.3491
	30	9.3793	0.3391	0.1110	0.2791
$Q_h = \frac{1}{\bar{x}_h}$	10	1.7931	1.6933	1.2887	1.1655
	15	9.0027	1.0062	0.4605	0.5849
	20	9.4757	0.6071	0.2164	0.4152
	25	10.0638	0.5910	0.1488	0.3472
	30	9.7563	0.4479	0.1139	0.2776
$Q_h = \frac{1}{\log \bar{x}_h}$	10	0.7633	1.4205	1.2604	1.1642
	15	8.7424	0.8467	0.4464	0.5877
	20	9.1127	0.4860	0.2075	0.4175
	25	9.7343	0.4748	0.1441	0.3488
	30	9.4698	0.3570	0.1116	0.2788

Table 2: Percentage Relative Root Mean Squared Error for Real Population under Stratified Sampling

Q_h	Sample size (n)	$\%\mathbf{RRMSE}\,(\bar{y}_{tr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{sh})$	$\%\mathbf{RRMSE}\,(\bar{y}_s)$	% RMSE $(\bar{y}_L)$
$Q_h = 1$	10	173.462	12.622	7.747	5.264
	15	70.591	9.607	5.584	3.898
	20	53.614	8.09	5.016	3.265
	25	54.328	7.075	3.566	2.889
	30	45.61	6.283	2.965	2.566

Continued on next page

Table 2 – continued from previous page

Q_h	Sample size (n)	$\%\mathbf{RRMSE}\,(\bar{y}_{tr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{sh})$	$\%\mathbf{RRMSE}\,(\bar{y}_s)$	% RMSE $(\bar{y}_L)$
$Q_h = \frac{1}{\bar{x}_h}$	10	215.329	13.634	7.792	5.278
	15	77.17	10.156	5.595	3.899
	20	59.682	8.517	5.019	3.265
	25	57.276	7.442	3.574	2.89
	30	53.568	6.589	2.97	2.566
$Q_h = \frac{1}{\log \bar{x}_h}$	10	182.882	12.777	7.754	5.266
	15	71.694	9.696	5.585	3.898
	20	54.559	8.16	5.016	3.265
	25	54.796	7.136	3.567	2.889
	30	46.479	6.334	2.966	2.566

Table 3: Percentage Absolute Relative Bias for Exponential Population under Stratified Sampling

Q_h	Sample size (n)	$\%\mathbf{ARB}\,(\bar{y}_{tr})$	$\%\mathbf{ARB}\,(\bar{y}_{sh})$	$\%\mathbf{ARB}\,(\bar{y}_s)$	% ARB $(\bar{y}_L)$
$Q_h = 1$	100	2.9946	1.2831	0.9324	0.3787
	200	4.4038	1.2254	0.4781	0.1884
	300	3.962	1.3506	0.2973	0.1191
	400	4.0945	1.0256	0.2066	0.0831
	500	3.6429	0.7785	0.1541	0.0626
$Q_h = \frac{1}{\bar{x}_h}$	100	3.0531	1.2181	0.9541	0.3896
	200	4.4437	1.208	0.4845	0.1953
	300	3.9976	1.3464	0.301	0.1242
	400	4.1103	1.0297	0.2077	0.0866
	500	3.6675	0.7833	0.1548	0.0655
$Q_h = \frac{1}{\log \bar{x}_h}$	100	2.9722	1.3061	0.931	0.3753
	200	4.3891	1.2325	0.4786	0.1864
	300	3.9495	1.3528	0.2977	0.1176
	400	4.0902	1.0246	0.2074	0.0821
	500	3.6365	0.777	0.1548	0.0618

Table 4: Percentage Relative Root Mean Square Error for Exponential Population under Stratified Sampling

Q_h	Sample size (n)	$\%\mathbf{RRMSE}\,(\bar{y}_{tr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{sh})$	$\%\mathbf{RRMSE}\,(\bar{y}_s)$	% RRMSE $(\bar{y}_L)$
$Q_h = 1$	100	153.551	42.372	9.289	3.493
	200	64.977	36.24	6.485	2.443
	300	95.341	17.964	5.165	1.954
	400	52.67	14.472	4.424	1.677
	500	44.794	11.924	3.871	1.467
$Q_h = \frac{1}{\bar{x}_h}$	100	155.419	42.439	10.071	3.602
	200	67.065	36.299	6.981	2.53
	300	95.675	18.037	5.55	2.027
	400	52.784	14.518	4.751	1.677
	500	45.956	11.957	4.154	1.467
$Q_h = \frac{1}{\log \bar{x}_h}$	100	152.999	42.355	9.042	3.452
	200	64.307	36.222	6.322	2.411
	300	95.242	17.941	5.037	1.927
	400	52.639	14.457	4.315	1.677
	500	44.396	11.915	3.777	1.467

Table 5: Percentage Absolute Relative Bias for Chi-Square Population under Stratified Sampling

Q_h	Sample size (n)	$\%\mathbf{ARB}\,(\bar{y}_{tr})$	$\%\mathbf{ARB}\,(\bar{y}_{sh})$	$\%\mathbf{ARB}\,(\bar{y}_s)$	% ARB $(\bar{y}_L)$
$Q_h = 1$	100	0.062	0.4329	0.1154	0.0806
	200	0.594	0.1832	0.0642	0.0398
	300	0.2283	0.2913	0.0454	0.0292
	400	0.2053	0.0102	0.0216	0.0163
	500	0.0056	0.0545	0.0166	0.0129
$Q_h = \frac{1}{\bar{x}_h}$	100	0.0707	0.4243	0.114	0.0788
	200	0.5878	0.1762	0.0632	0.0388
	300	0.2353	0.2882	0.0446	0.0284
	400	0.2087	0.0121	0.0212	0.0159
	500	0.0013	0.0548	0.0162	0.0126
$Q_h = \frac{1}{\log \bar{x}_h}$	100	0.0648	0.4299	0.1147	0.0799
	200	0.5919	0.1807	0.0637	0.0394
	300	0.2309	0.2901	0.0451	0.0289

Continued on next page

Table 5 – continued from previous page

Q_h	Sample size (n)	$\%\mathbf{ARB}\,(\bar{y}_{tr})$	$\%\mathbf{ARB}\,(\bar{y}_{sh})$	$\%\mathbf{ARB}\,(\bar{y}_{s})$	% ARB $(\bar{y}_{L})$
	400	0.2065	0.0109	0.0214	0.0161
	500	0.0032	0.0547	0.0164	0.0128

Table 6: Percentage Relative Root Mean Square Error for Chi-Square Population under Stratified Sampling

Q_h	Sample size (n)	$\%\mathbf{RRMSE}\,(\bar{y}_{tr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{sh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{s})$	% RRMSE $(\bar{y}_{L})$
$Q_h = 1$	100	59.165	38.037	3.253	1.083
	200	45.969	44.751	2.302	0.761
	300	45.009	21.453	1.839	0.61
	400	42.838	18.043	1.568	0.519
	500	39.389	14.754	1.38	0.458
$Q_h = \frac{1}{\bar{x}_h}$	100	59.225	38.031	3.212	1.065
	200	45.975	44.744	2.271	0.749
	300	44.95	21.44	1.813	0.6
	400	42.757	18.035	1.546	0.511
	500	39.43	14.747	1.36	0.45
$Q_h = \frac{1}{\log \bar{x}_h}$	100	59.18	38.035	3.238	1.076
	200	45.97	44.748	2.291	0.757
	300	44.988	21.448	1.829	0.606
	400	42.809	18.04	1.56	0.516
	500	39.404	14.751	1.373	0.455

Table 7: Percentage Absolute Relative Bias for Real Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{ARB}\,(\bar{y}_{dtr})$	$\%\mathbf{ARB}\,(\bar{y}_{dsh})$	$\%\mathbf{ARB}\,(\bar{y}_{ds})$	% ARB $(\bar{y}_{dL})$
$Q_h = 1$	50; 10	1.6119	1.3687	1.0952	1.0207
	50; 15	6.5088	0.6447	0.3683	0.4581
	50; 20	7.232	0.3792	0.1783	0.2953
	50; 25	5.9368	0.2757	0.0884	0.2222
	50; 30	6.0201	0.2317	0.0882	0.1789
$Q_h = \frac{1}{\bar{x}_h}$	50; 10	3.1666	1.6701	1.1279	1.0206

Continued on next page

Table 7 – continued from previous page

Q_h	Sample size (m;n)	$\%\mathbf{ARB}\,(\bar{y}_{dtr})$	$\%\mathbf{ARB}\,(\bar{y}_{dsh})$	$\%\mathbf{ARB}\,(\bar{y}_{ds})$	% ARB $(\bar{y}_{dL})$
	50; 15	6.781	0.7997	0.3812	0.4549
	50; 20	7.5947	0.4832	0.1857	0.2931
	50; 25	6.2457	0.3665	0.094	0.2209
	50; 30	6.3262	0.2982	0.0901	0.1778
$Q_h = \frac{1}{\log \bar{x}_h}$	50; 10	1.8252	1.4127	1.1041	1.0205
	50; 15	6.5759	0.67	0.371	0.4575
	50; 20	7.3259	0.3961	0.1799	0.2949
	50; 25	6.0144	0.2906	0.0895	0.2219
	50; 30	6.0889	0.2426	0.0886	0.1787

Table 8: Percentage Relative Root Mean Square Error for Real Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{RRMSE}\,(\bar{y}_{dtr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{dsh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{ds})$	% RRMSE $(\bar{y}_{dL})$
$Q_h = 1$	50; 10	185.036	12.225	7.681	5.283
	50; 15	83.319	8.851	5.173	3.954
	50; 20	52.734	7.133	3.942	3.364
	50; 25	47.448	5.976	3.36	2.99
	50; 30	41.435	5.022	2.836	2.665
$Q_h = \frac{1}{\bar{x}_h}$	50; 10	194.478	13.181	7.723	5.294
	50; 15	88.354	9.356	5.18	3.955
	50; 20	58.202	7.492	3.944	3.364
	50; 25	48.988	6.256	3.361	2.991
	50; 30	47.5	5.246	2.838	2.665
$Q_h = \frac{1}{\log \bar{x}_h}$	50; 10	185.52	12.371	7.688	5.285
	50; 15	84.041	8.932	5.173	3.954
	50; 20	53.717	7.192	3.941	3.364
	50; 25	47.639	6.022	3.36	2.99
	50; 30	42.454	5.06	2.836	2.665

Table 9: Percentage Absolute Relative Bias for Exponential Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{ARB}\,(\bar{y}_{dtr})$	$\%\mathbf{ARB}\,(\bar{y}_{dsh})$	$\%\mathbf{ARB}\,(\bar{y}_{ds})$	% ARB $(\bar{y}_{dL})$
$Q_h = 1$	1000; 100	3.2578	0.8599	0.8234	0.345
	1000; 200	3.5903	0.9484	0.2886	0.1196
	1000; 300	3.8258	0.825	0.1759	0.0735
	1000; 400	2.7085	0.5316	0.0956	0.0444
	1000; 500	1.7879	0.427	0.0794	0.0319
$Q_h = \frac{1}{\bar{x}_h}$	1000; 100	3.3057	0.8082	0.8399	0.3547
	1000; 200	3.577	0.9316	0.2879	0.1243
	1000; 300	3.8497	0.8244	0.1739	0.0766
	1000; 400	2.7306	0.5324	0.0925	0.0462
	1000; 500	1.7919	0.4313	0.0782	0.0332
$Q_h = \frac{1}{\log \bar{x}_h}$	1000; 100	3.2355	0.8781	0.8236	0.342
	1000; 200	3.5941	0.9551	0.2912	0.1183
	1000; 300	3.8171	0.8256	0.1779	0.0727
	1000; 400	2.7025	0.5316	0.0974	0.0439
	1000; 500	1.7883	0.4256	0.0803	0.0316

Table 10: Percentage Relative Root Mean Square Error for Exponential Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{RRMSE}\,(\bar{y}_{dtr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{dsh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{ds})$	% RRMSE $(\bar{y}_{dL})$
$Q_h = 1$	1000; 100	152.986	34.659	8.839	3.337
	1000; 200	60.775	24.613	5.898	2.235
	1000; 300	53.079	14.609	4.448	1.684
	1000; 400	63.375	11.159	3.556	1.349
	1000; 500	34.654	9.045	2.926	1.094
$Q_h = \frac{1}{\bar{x}_h}$	1000; 100	153.006	34.755	9.598	3.442
	1000; 200	61.867	24.68	6.359	2.315
	1000; 300	53.903	14.676	4.779	1.746
	1000; 400	64.237	11.201	3.819	1.401
	1000; 500	35.014	9.073	3.139	1.152
$Q_h = \frac{1}{\log \bar{x}_h}$	1000; 100	153.011	34.633	8.599	3.297
	1000; 200	60.418	24.592	5.747	2.205
	1000; 300	52.793	14.588	4.338	1.661

Continued on next page

Table 10 – continued from previous page

Q_h	Sample size (m;n)	$\%\mathbf{RRMSE}\,(\bar{y}_{dtr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{dsh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{ds})$	% RRMSE $(\bar{y}_{dL})$
	1000; 400	63.074	11.146	3.468	1.33
	1000; 500	34.531	9.037	2.854	1.094

Table 11: Percentage Absolute Relative Bias for Chi-Square Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{ARB}\,(\bar{y}_{dtr})$	$\%\mathbf{ARB}\,(\bar{y}_{dsh})$	$\%\mathbf{ARB}\,(\bar{y}_{ds})$	% ARB $(\bar{y}_{dL})$
$Q_h = 1$	1000; 100	0.669	0.5784	0.1257	0.0826
	1000; 200	0.423	0.1409	0.0282	0.0236
	1000; 300	0.2107	0.0585	0.0263	0.017
	1000; 400	0.0369	0.0557	0.0072	0.0088
	1000; 500	0.3538	0.019	0.0093	0.0072
$Q_h = \frac{1}{\bar{x}_h}$	1000; 100	0.6806	0.5687	0.1236	0.0805
	1000; 200	0.4179	0.1389	0.0278	0.0229
	1000; 300	0.1951	0.0565	0.0257	0.0165
	1000; 400	0.0292	0.0542	0.0071	0.0085
	1000; 500	0.3558	0.0184	0.0091	0.007
$Q_h = \frac{1}{\log \bar{x}_h}$	1000; 100	0.6732	0.575	0.1248	0.0817
	1000; 200	0.4212	0.1402	0.028	0.0233
	1000; 300	0.2051	0.0577	0.026	0.0168
	1000; 400	0.0341	0.0552	0.0071	0.0087
	1000; 500	0.3545	0.0187	0.0092	0.0071

Table 12: Percentage Relative Root Mean Square Error for Chi-Square Population under Stratified Double Sampling

Q_h	Sample size (m;n)	$\%\mathbf{RRMSE}\,(\bar{y}_{dtr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{dsh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{ds})$	% RRMSE $(\bar{y}_{dL})$
$Q_h = 1$	1000; 100	51.981	37.613	3.164	1.063
	1000; 200	47.139	24.501	2.088	0.704
	1000; 300	67.436	20.02	1.617	0.551
	1000; 400	45.65	14.253	1.293	0.45
	1000; 500	36.251	10.51	1.058	0.376
$Q_h = \frac{1}{\bar{x}_h}$	1000; 100	52.489	37.606	3.125	1.046

Continued on next page

Table 12 – continued from previous page

Q_h	Sample size (m;n)	$\%\mathbf{RRMSE}\,(\bar{y}_{dtr})$	$\%\mathbf{RRMSE}\,(\bar{y}_{dsh})$	$\%\mathbf{RRMSE}\,(\bar{y}_{ds})$	% RRMSE $(\bar{y}_{dL})$
	1000; 200	47.131	24.492	2.06	0.694
	1000; 300	67.244	20.011	1.594	0.543
	1000; 400	45.595	14.247	1.275	0.45
	1000; 500	36.152	10.505	1.044	0.376
$Q_h = \frac{1}{\log \bar{x}_h}$	1000; 100	52.139	37.61	3.15	1.057
	1000; 200	47.135	24.498	2.078	0.7
	1000; 300	67.367	20.017	1.609	0.548
	1000; 400	45.629	14.25	1.287	0.45
	1000; 500	36.216	10.508	1.053	0.376

4 Conclusion

It is clear from Tables 1 to 12 that the values of Percentage absolute relative bias $\%\ ARB$ and Percentage Relative Root Mean Squared Error $\%\ RRMSE$ of the suggested estimators are less the the estimators given by [9], [10] and [12] for both stratified and stratified double sampling schemes. The values of the $\%\ RRMSE$ also decrease as the values of sample size increase for a real as well as both artificial datasets. The result of the simulation study indicates that the proposed estimators of the population mean in case of stratified random sampling and stratified double sampling using logarithmic mean of the auxiliary variable in calibration constraint are found to be more efficient as compared to the estimators of [9], [10] and [12] based on the considered datasets.

Acknowledgements

The authors are highly thankful to the Editor-in-Chief Prof. D. K. Ghosh and the anonymous referees for their constructive suggestions.

References

[1] E P Clement and E I Enang. "On the efficiency of ratio estimator over the regression estimator." In: *Communications in Statistics-Theory and Methods* 46 (2017), pp. 5357–5367.

[2] J C Deville and C E Särndal. "Calibration estimators in survey sampling." In: *Journal of the American statistical Association* 87 (1992), pp. 376–382.

[3] N Garg and M Pachori. "Calibration estimation of population mean in stratified sampling using coefficient of skewness." In: *International Journal of Agricultural and Statistical Sciences* 15 (2019), pp. 211–219.

[4] N Garg and M Pachori. "Use of coefficient of variation in calibration estimation of population mean in stratified sampling." In: *Communications in Statistics-Theory and Methods* 49 (2020), pp. 5842–5852.

[5] J M Kim, E A Sungur, and T Y Heo. "Calibration approach estimators in stratified sampling." In: *Statistics & probability letters* 77 (2007), pp. 99–103.

[6] N Koyuncu and C Kadilar. "Calibration estimator using different distance measures in stratified random sampling." In: *International Journal of Modern Engineering Research* 3 (2013), pp. 415–419.

[7] N Koyuncu and C Kadilar. "Calibration weighting in stratified random sampling." In: *Communications in Statistics-Simulation and Computation* 45 (2016), pp. 2267–2275.

[8] A M Mouhamed, A A El-Sheikh, and H A Mohamed. "A new calibration estimator of stratified random sample." In: *Applied Mathematical Sciences* 9 (2015), pp. 1735–44.

[9] S Singh. "Estimation of variance of the general regression estimator: Higher level calibration approach." In: *Survey methodology* 24 (1998), pp. 41–50.

[10] S Singh. "Golden jubilee year 2003 of the linear regression estimator." In: *Working paper at St. Cloud State University, St. Cloud, MN, USA* (2003).

[11] S Singh and R Arnab. "On calibration of design weights." In: *Metron* 69 (2011), pp. 185–205.

[12] D S Tracy, S Singh, and R Arnab. "Note on calibration in stratified and double sampling." In: *Survey Methodology* 29 (2003), pp. 99–104.

A Statistical Assessment of Air Quality Index on Vegetation using Geo-Spatial Data

K. Muralidharan[1], Shrey Pandya[2], Manthan Daraji[3], Kalpesh Mali[4], Rutvik Panchal[5], Paresh Chaudhari[6]

Department of Statistics, Faculty of Science,
The Maharaja Sayajirao University of Baroda, Vadodara, 390002. India

Email: [1] muralikustat@gmail.com, [2]shreypandya24@rediffmail.com, [3] darjimanthan1@gmail.com, [4] kalpeshbmali1999@gmail.com, [5] panchal.rutvik99@gmail.com, [6] paresh.chaudhari18799@gmail.com,

Received: 14 March 2022 / Revised: 21 January 2023 / Accepted: 16 August 2023

Abstract. Air is the most essential element for existence of life. With urbanization and industrialization air pollution had increased and had become a major global issue for development and growth. It is impacting the health of humans, flora, and fauna of our Biodiversity. In many studies, it is discussed that air pollution hurts the lants and ecosystem. An approach is made to analyze the impact, compare, and model greenness using a geospatial data considering two sites namely Ahmedabad city and Nal-Sarovar of Gujarat state. We observe that the average NDVI of both the sites was not the same. Based on the statistical spatial analysis further models for each location were built to predict NDVI using PM10, CO, LST, SMI, and NDVI. Also, for Ahmedabad city out of 17 places 7 had reliability more than 0.50 while that of Nal-Sarovar out of 13 places 7 models had reliability greater than 0.50. This study also raises the concern that NDVI had a positive relationship with all variables under study except NDMI. This gives us a new dimension to investigate the impact of AQI on vegetation on the selected area more preciously and concisely.

Keywords: Spatial data, NDVI, PM10, CO, LST, SMI, NDMI

1. Introduction

Air is the mixture of several gases like N_2, O_2, O_3, CO_2, CH_4, SO_2, N_2O, etc., with development in urbanization and industrialization these gases have also increased and have crossed the limits and are now present with a high concentration in the atmosphere which leads to air pollution. Major sources of air pollution in the cities around the globe is emission from vehicles, thermal power plants, industrial units, and domestic (Salah and Ghada, 2014; Salah, 2011). In recent years these unplanned and uncontrolled development has led to ecological imbalance, and ultimately, ecological collapse of all the hazards that our ecology is prone to in today's environmental scenario, air pollution has become a major concern around the globe (Gheorghe

and Ion, 2011). It causes several health problems, and it has been linked with illnesses and deaths from heart or lung diseases, also has affected the climatic cycle of the environment and has resulted in an increase in temperature every year (Lim et al., 2010).

Many studies across the earth are conducted where it has been shown how air pollution is impacting vegetation in the surroundings. Pollutants present in the air hurt plants either directly (toxicity) or indirectly (changing soil pH) (Alseroury,2017)**.** The particulate matters have a negative mechanical effect. They cover the leaf blade reducing light penetration and blocking the opening of stomata. These impediments strongly influence the process of photosynthesis and has resulted in a sharp decline in it (Emberson et al., 2001)**.** These results in the deterioration of the health and the greenness of the plants, thus necessitating the monitoring and intervention air pollution. Air Quality Index (AQI) is one such measure that helps us to signify the severity of air for all biotics on planet earth. The AQI is a way of showing changes in the amount of pollution in the air (Lim et al., 2009). With the increase in transportation and industrialization, air pollution has also increased and has shown its adverse effect on biotic and non-biotic objects all around the globe. To look around a large area for a discrete time point over a larger area earlier was difficult, but now with the evolution of modern technology many alternative approaches are possible for modelling and decision making, which has made certain tasks smooth in estimating the AQI. The use of modern technology such as remote sensing and GIS can provide geospatial distribution of pollutants over time and location, and digital records and maps are acquired as output for the same (Agrawal *et al.*, 2003; Hurlock and Stutz, 2004). With the innovation of these technologies, it has become possible and easy to monitor and detect changes on land and water (Sohrabiniaa and Khorshiddoust, 2007). Many qualitative types of research are conducted in which the impact of air pollution is measured while some studies have used spatial approach to estimate the AQI. However, the object of this study is to estimate the AQI and NDVI using spatial data quantitatively.

The article is organized in the following way: Section 2 offers the materials and methods needed for the study. The methodology and study area are briefly mentioned in this section. The details about the presentation of data are shown in Section 3. The statistical analysis and findings are presented in Sections 4 and Section 5 respectively. The last section offers some conclusions.

2. Materials and Methods

2.1. *Methodology*

Two different sources of data are taken into the consideration to conduct this study i.e spatial and non-spatial data. Multispectral Landsat-8 Optical Land Imager (OLI) spatial data with a resolution of 30m was acquired from the USGS website. This was used for generating various indices like vegetation index, moisture index, and Soil index, while non-spatial data of pollutants and AQI is acquired from the websites for stations defined with the boundaries of the study area. With the help of different sources of collected data an approach is made to understand the relationship between air pollutants and spatial indices generated from spatial data and established a model which can help in predicting NDVI. To achieve this, our objective methodology given in figure 1 was strictly followed.

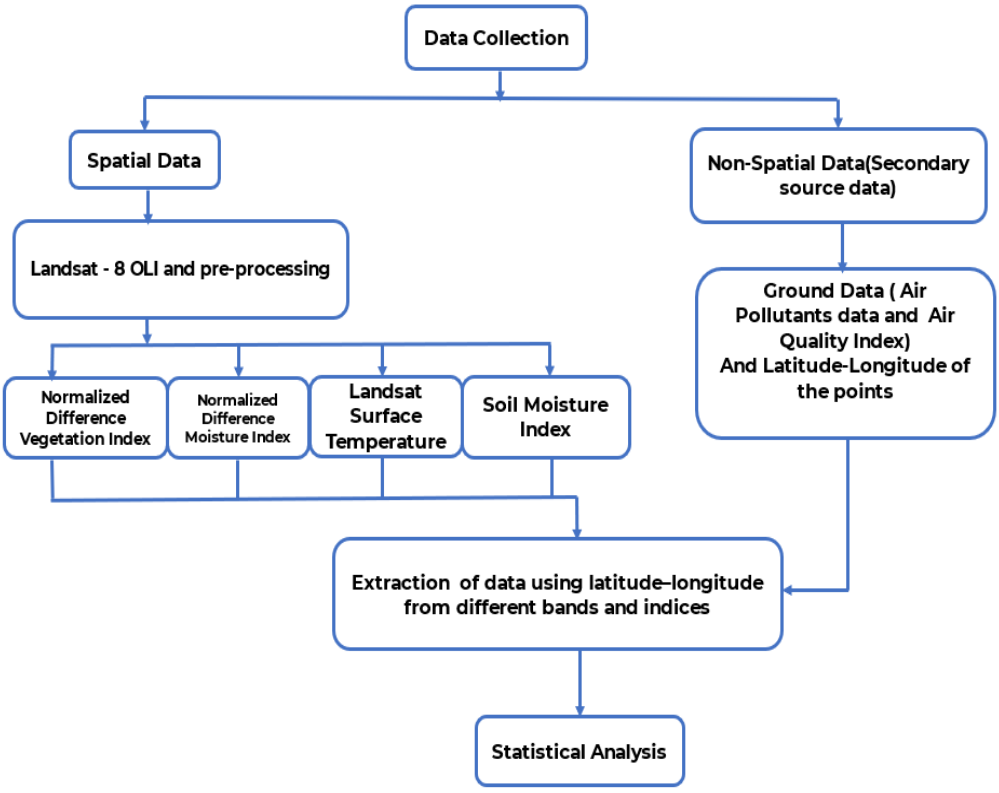


Fig. 1: Methodology of the study

2.2 *Study Area*

Since the study focusses on the vegetation status of two different locations in Gujarat, we used the point of reference as same for both the location. Ahmedabad is one of Gujrat's largest and fastest-growing cities with a population of over 7.5 million, located on the banks of the Sabarmati River. It is located at 23.03°N 72.58°E spanning an area of 205 km². According to the WHO urban air quality database and several international and Indian studies, Ahmadabad is identified as one of the Air polluted cities in Gujarat. The second site is situated in the west of Ahmedabad near Sanand village is Nal-Sarovar which is about 64 km away from Ahmedabad city. It is one of the largest wetland and bird sanctuaries in Gujarat, also surrounded by vegetation diversity as well. Overall, 30 locations were selected of which 17 were from different areas of the city and 13 were from the nearby areas of wetland as shown in Figure 2. The names of this selected location are given in table 1. Locations with names are presented in figure 2 of the study area map and A and B represent the sites.

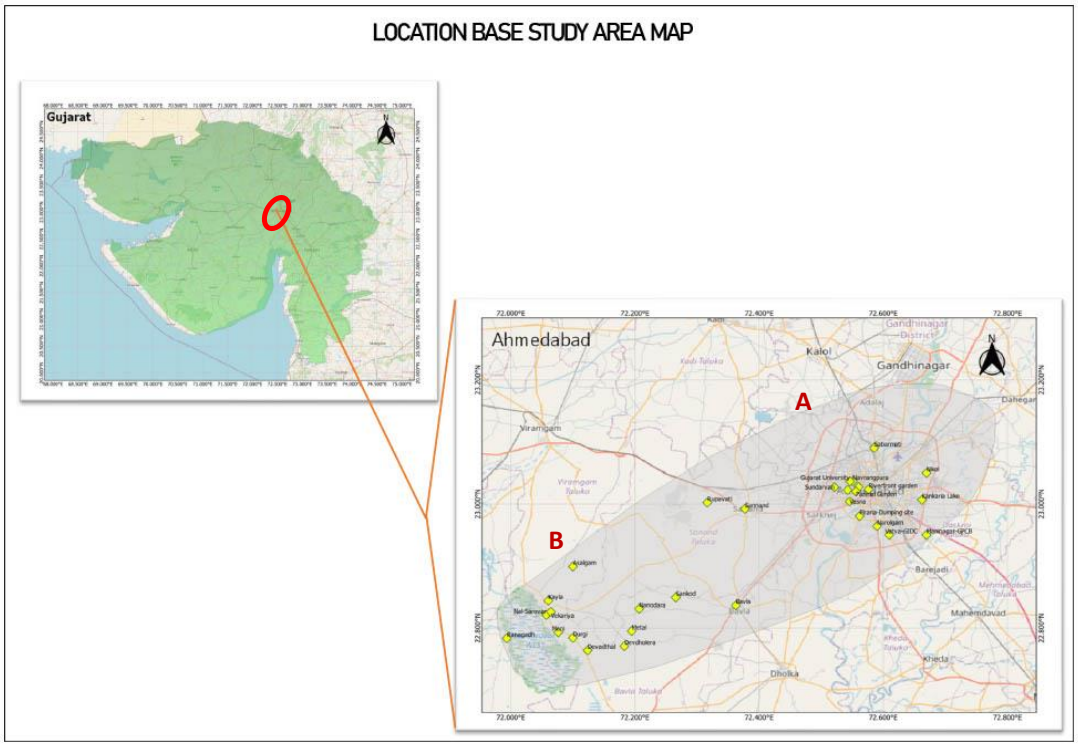


Figure 2: *Study area map*

Table 1: *locations name concerning site*

Site	Location's name
Ahmedabad City (A)	Kankaria Lake, Law Garden, Maninagar, Vatva, Pirana, Sabarmati River Front, Parimal Garden, Navrangpura, Gujarat University, Nikol, Vasna, Riverfront Garden, Ambavadi, Narol-Gam, Sundarvan, Bavla, Sanand
Nal-Sarovar (B)	Nal-Sarovar, Ranagadh, Kayla, Devadthal, Durgi, Devdholera, Nanodara, Meni, Asalgam, Sankod, Vekeriya, Metal, Rupavati

3. Data presentation

3.1 *Remote sensing data acquisition*

Remote sensing data used for this study is the Landsat 8 Optical Land Imager (OLI) level-1 imaginary, acquired for free through the United States Geological Survey (USGS) of Earth Resources Observation and Science Center (EROS) from January 2017 to March 2021. For this study bi-monthly data was considered with no or less cloud coverage for which cloud coverage was kept as 10%. Overall, we had 30 spatiotemporal data scenes. The acquired data of the study area was already geometrically corrected and further, radiometric correction of multi spectral imagery was done of acquired data by converting digital numbers (DNs) to the spectral radiance by the proposed method by Landsat 8 OLI level-1 guideline pdf. Processed data consists of 11 bands ranging from 0.433 µm-12.51 µm, which comprises visible bands, NIR, SWIR, and TRIS bands. All 13 bands comprise different resolutions, the visible band has a resolution of 30 meters while NIR, SWIR, and TRIS have a resolution of 100 meters and the panchromatic band has a resolution of 15 meters. Data for the same can be acquired after every 15 days. Table 1 presents the specifications of these bands.

Table 2: *L-8 OLI level-1 band description*

Band Number	Band Description	Wavelength (µm)
1	Costal Aerosol	0.43 – 0.453
2	Blue	0.450 – 0.515
3	Green	0.525 – 0.600
4	Red	0.630 – 0.680
5	Near-Infrared (NIR)	0.845 – 0.885
6	Short Wave Infrared (SWIR-1)	1.560 – 1.660
7	SWIR-2	2.100 – 2.300

8	Panchromatic	0.500 – 0.680
9	Cirrus	1.36-1.38
10	Thermal Infrared (TIRS) 1	10.30 – 11.30
11	Thermal Infrared (TIRS) 1	11.50 – 12.51

3.2 *Spectral indices*

All materials are prone to some electromagnetic spectrum, which is reflected and the rest is absorbed by the material. Considering the same mechanism and based on the understanding of reflecting and absorption these spectral indices are calculated. Some of the spectral indices used in our study are Normalized Difference Vegetation Index (NDVI), Soil Moisture Index (SMI), and Normalized Difference Moisture Index (NDMI). Their analytical significance is described below.

3.2.1 *Normalized Difference Vegetation Index (NDVI)*

Vegetation absorbs green light while reflecting red light more and near-infrared (NIR) less. Due to this reason red and NIR are used to calculate NDVI (J. E. George et al., 2017; Yuan et al., 2007). Formula to get NDVI is given as follows:

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (1)$$

To make values easily interpretable normalization and standardization is done by taking the difference between both the bands in the numerator and summing them up in the denominator, and hence, the values will range between -1 and +1. The negative values indicate the presence of non-vegetative areas like water bodies, urban areas, Baran land, snow, and cloud while the positive values indicate vegetations and that could be anything like dense forests, sparse vegetation, auricular land, etc.

3.2.2 *Land Surface Temperature (LST)*

This index indicates the temperature of the ground surface. Thermal bands numbers 10 and 11 of Landsat image are used to get LST. The pixels of images consist of Digital Numbers (DN) which can't give information about the temperature. Since reflection of electromagnetic waves differs for each material as a result of rising temperature, the DN value is converted to radiance. Radiance is the measure of how much light the sensor can see the object, which is reflected

from the object under consideration (Stathopoulou et al.,2007). The equation of atmosphere radiance is given as follows:

$$L_\lambda = M_L + Q_{CAL} * A_L \quad (2)$$

where L_λ: Top of atmosphere radiance (watts/ (m2 $*$ srad $*\mu$), M_L: Band multiplicative rescaling factor from metadata, A_L: Band additive rescaling factor from metadata, and Q_{CAL}: Quantized and calibrated standard product values (DN). Now we have the radiance values of the Thermal Infrared bands of the data which should be converted to satellite temperature. It is the measure of radiation traveling from the top of the atmosphere to satellite. The satellite brightness temperature [5] is calculated by,

$$T = \frac{K_1}{\ln\left(\frac{K_2}{L_\lambda}\right)+1} \quad (3)$$

where $K_1$and K_2 are the coefficients calculated by the effective wavelength of the satellite sensor which is given in the metadata file and L_λ is the radiance value. To get LST some other parameters like NDVI, Proportion of vegetation, and land surface emissivity should be used and can be found in George et al. (2017).

3.2.3 *Normalized Difference Moisture Index (NDMI)*

Normalized difference moisture index indicates the presence of moisture content in vegetation (Josef et al. ,2020; Rahman and Victor,2019). This can be obtained using the NIR and SWNIR bands and its formula is given by

$$NDMI = \frac{NIR-SWNIR}{NIR+SWNIR} \quad (4)$$

3.3 *Ground Data*

Data of various pollutants and AQI (Air Quality Index) was collected for January 2019 to April 2021 from predefined stations. Parameters like PM10, PM2.5, O3, NO2, CO, SO2 were taken on ground stations. These parameters were then considered for calculating the AQI (Air Quality Index) and were consider to further to understand the relationship with vegetation index and build a model.

4. Statistical Analysis

Data analysis using statistical principles and methods form the basis of any measurement-based decision-making in any branch of science and management. Below we use various statistical tools to estimate and model the NDVI and AQI. We also propose various suggestions and recommendations for policy makers working in environmental science and climate management issues.

4.1 *Two-sample t-test*

Two sample t-test is a statistical test that is used for the comparison of the mean of the two samples when the standard deviation is not known and assume it to be equal. In this case, the formula for a two-sample t-test is given as

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S^2\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}} \qquad (5)$$

Accordingly, our proposed tests of hypothesis will be to test $H_0: \mu_0 = \mu_1$ versus $H_1: \mu_0 > \mu_1$, where μ_0 is the mean NDVI of Nal-Sarovar and μ_1 is the mean NDVI of Ahmedabad city. The specificity of the alternative hypothesis is due to the fact that the NDVI of Nal-Sarovar is expected to be larger than the other city area as discussed previously.

4.2 *Multiple Linear Regression*

Multiple linear regression is the extended version of simple linear regression. In this, at a time more than one exploratory variable is considered and the statistical model that is constructed is shown in the following way:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \qquad (6)$$

where, Y = NDVI, β_0 = Intercept of the model, $\beta_{i's}$ = slops; i= 1,2,3,4,5, $x_1 = \mathrm{PM10}$, $x_2 = \mathrm{CO}$, $x_3 = \mathrm{LST}$, $x_4 = \mathrm{SMI}$, $x_5 = \mathrm{NDMI}$, and $\varepsilon = \text{error term}$.

In this study, we have tried to build-up a multiple liner model that tries to quantify the relation between image extracted vegetation parameter with that of air pollutants continuously measured at ground stations (Mozumder et al.,2012).

5. Result and findings

At the initial stage visualization of ground data was done to understand the trend of AQI considering the lower and upper limit specified by the Central pollution control board (CPCB) (lower limit = 60 and upper limit = 100) for AQI as shown in figure 3.

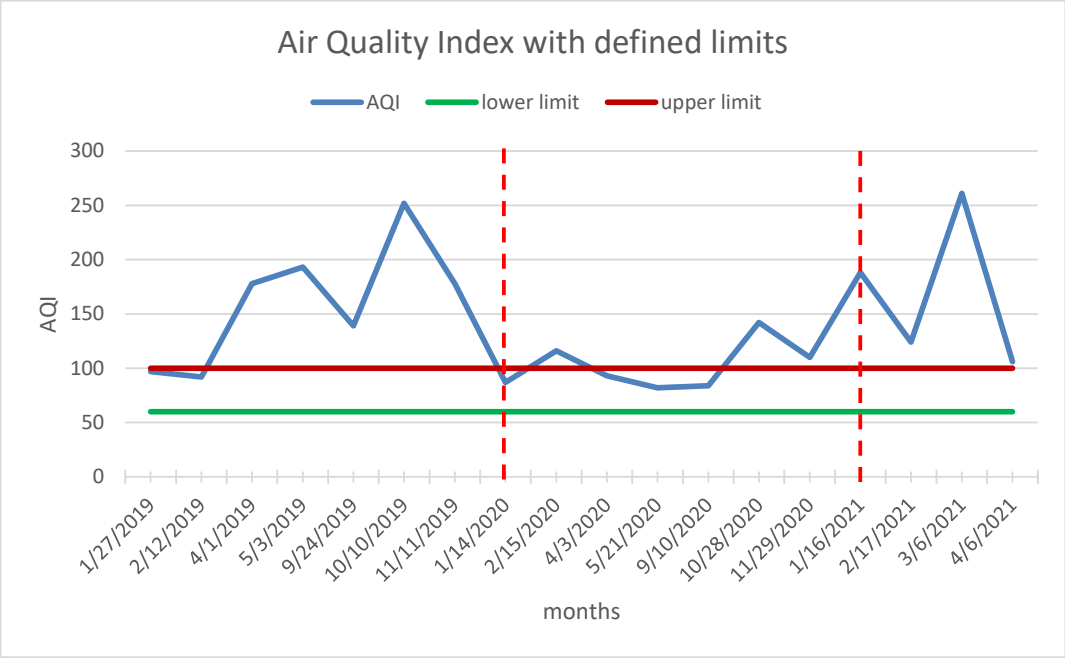


Figure 3: *AQI line chart from January 2019 to April 2021*

The variation with an upward trend is observed in values of AQI for the year 2019 till October after which a sharp fall is observed and has gone below 100. During the year 2020, it remained below 100 and again as at the end of the year it started to rise. One of the possible reasons could be the relaxation in the lockdown due to COVID19 given by the government. In the year 2021 more variation with an upwards trend is observed during the initial days and started dipping thereafter possibly due to the second wave OVID19. One of the reasons for this is because the pattern of cases strictly followed an exponential distribution.

Further to test if there is a significant difference in the air quality in the two sites discussed, a two-sample t-test was applied on the NDVI average measure. We found that the hypothesis is rejected 5% level of significance (p-value is 6.6178E-6) and conclude that the average NDVI of Nal-Sarovar is more than that of Ahmedabad city. In many types of research,

approaches are made to predict the parameters using information extracted from the spatial data. In our study before building a model to predict the NDVI from the spatial data, it is essential to observe the relationship exists between the sources of data (spatial and non-spatial source of data) in terms of its variables, we obtain, the correlation matrix as shown in figure 4. It is observed that the variables NDVI, PM10, CO, LST had shown a moderate positive correlation.

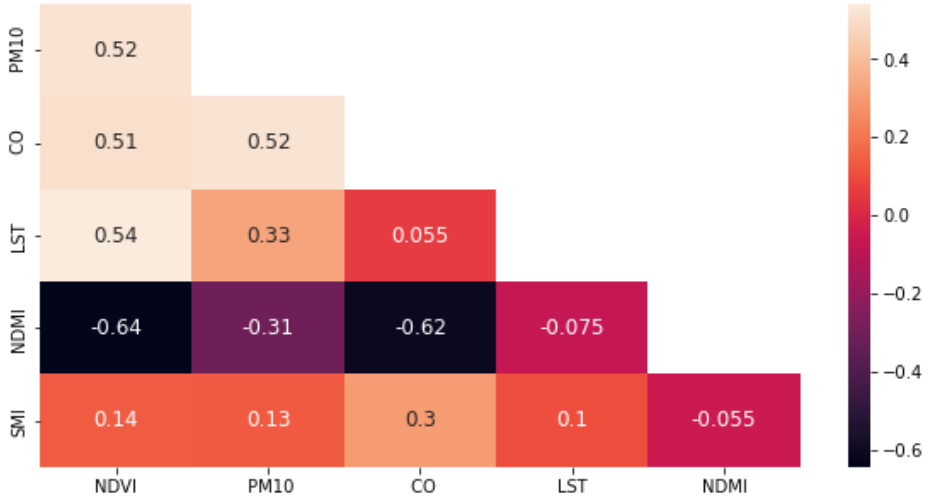


Figure 4: *Heat map with correlation values between the variables*

We observed that response variable NDVI had a moderate correlation with LST, PM10 and CO. However, the variable NDMI had a low or high negative correlation with all other variables. Based on the correlation matrix, a multiple linear regression models for NDVI for each location were built. The model reliability is measured using the coefficient of determination R^2. Models concerning two sites with respect to reliability are shown in table 3 and table 4 respectively for Ahmedabad city and Nal-Sarovar sites.

Table 3: *Models of 17 locations of Ahmedabad city*

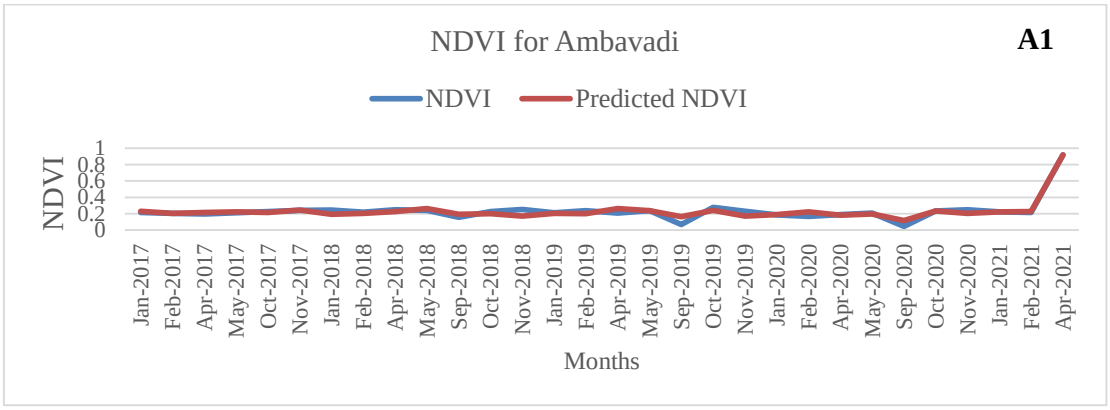
Sr. No	Location	Equation	R-Square
1	Ambavadi	NDVI =0.1309 +0.000253*PM10 + 0.000137*CO + 0.000324*LST+0.00544*SMI + 0.4739*NDMI	0.92
2	Pirana	0.5452 - 0.000024*PM10 - 0.00117*CO + 0.0000183*LST + 0.046*SMI - 0.9340*NDMI	0.64
3	Maninagar	0.2141 - 0.0001008*PM10 + 0.0004196*CO + 0.0009842*LST + 0.071*SMI - 0.4272*NDMI	0.61

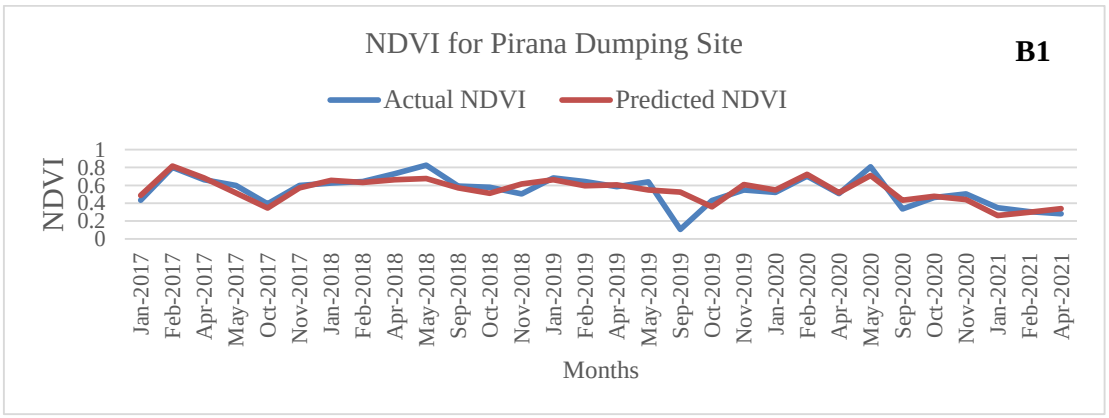
4	Navrangpura	0.3967 - 0.00052*PM10 + 0.000271*CO + 0.0000158*LST + 0.1076*SMI -0.6771*NDMI	0.54
5	Bavla	0.3646 +0.00054*PM10 -0.0004*CO -0.0003*LST + 0.0774*SMI -0.6952*NDMI	0.53
6	Sundarvan	0.1743 + 0.0000794*PM10 + 0.000614*CO +0.000199*LST +0.0583*SMI-0.2268*NDMI	0.51
7	Vasna	-0.01421 +0.000334*PM10 -0.00016*CO + 0.000262*LST+0.13174*SMI + 0.13018*NDMI	0.50
8	Parimal Garden	0.2798 - 0.00029*PM10 - 0.0006*CO + 0.000398*LST + 0.1915*SMI -1.14*NDMI	0.45
9	Sanand	0.1391+0.000228*PM10 + 0.000285*CO - 0.00004*LST +0.04381*SMI +0.4351*NDMI	0.40
10	Gujarat University	0.2109 - 0.0000091*PM10 + 0.000487*CO + 0.0000605*LST-0.01144*SMI -0.2438*NDMI	0.37
11	Nikol	0.08925 +0.00045*PM10 - 0.0002*CO + 0.00027*LST+0.05786*SMI + 0.3024*NDMI	0.36
12	Vatva	0.1180 + 0.000209*PM10 + 0.000227*CO + 0.000197*LST + 0.028*SMI + 0.3497*NDMI	0.33
13	Sabarmati River Front	0.184+ 0.0002534*PM10 - 0.000106*CO + 0.0000964*LST + 0.046*SMI + 0.01686*NDMI	0.33
14	Law Garden	0.3964 - 0.0004269*PM10 + 0.0000693*CO + 0.000837*LST + 0.1496*SMI - 0.32005*NDMI	0.28
15	Kankaria Lake	0.126833 + 0.000044*PM10 + 0.000228*CO – 0.0000140*LST + 0.0328*SMI + 0.1642*NDMI	0.25
16	Riverfront Garden	0.2022 +0.0000715*PM10 +0.0000382*CO + 0.000494*LST+0.0391*SMI +0.478*NDMI	0.20
17	Narol-Gam	0.2589 +0.00055*PM10 - 0.00014*CO + 0.0000057*LST-0.0155*SMI + 0.1459*NDMI	0.18

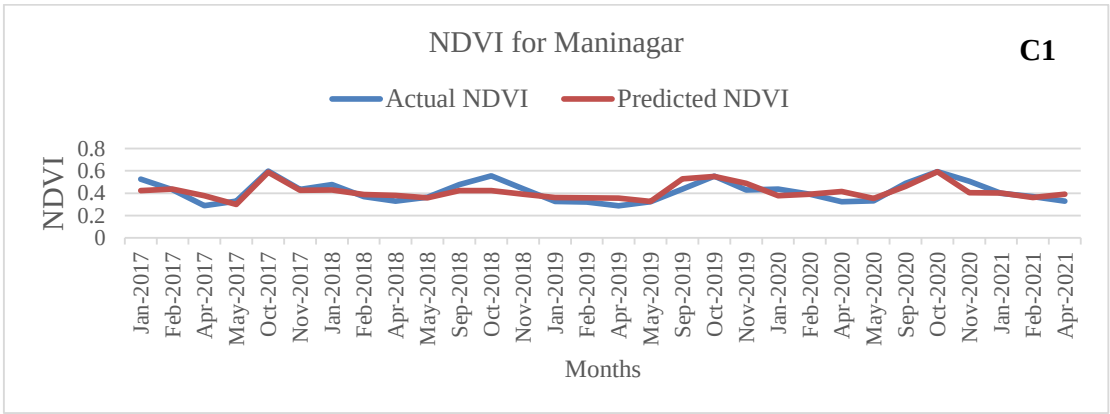
Table 4: *Models of 13 locations of Nal-Sarovar*

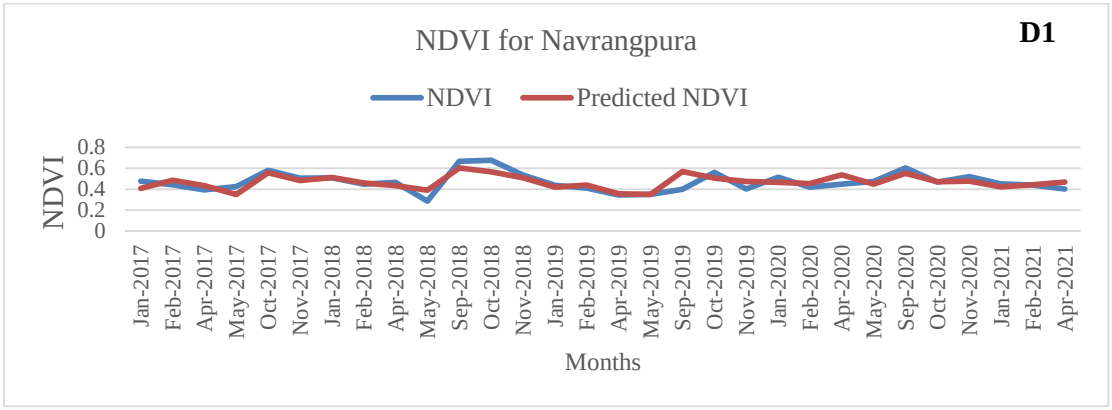
Sr. No	Location	Equation	R-Square
1	Sankod	0.4272 + 0.000225*PM10 -0.0000217*CO -0.000246*LST + 0.0659 *SMI -0.8769*NDMI	0.82
2	Devdholera	0.4396 + 0.00007*PM10 – 0.0006*CO + 0.000374*LST + 0.0948*SMI -0.8327*NDMI	0.80
3	Metal	0.4140 + 0.000296*PM10 -0.000129*CO +0.000145*LST +0.04218*SMI – 0.9293*NDMI	0.78
4	Meni	0.4119 -0.0000905*PM10 -0.000164*CO + 0.000173*LST +0.0675*SMI -0.8984*NDMI	0.74
5	Kayla	0.1226 -0.0006356*PM10 +0.0024*CO + 0.0004603*LST +0.01898*SMI + 0.7958*NDMI	0.53
6	Nal-Sarovar	0.1609 +0.0000853*PM10 +0.00065*CO + 0.00031*LST+0.065*SMI – 0.0778*NDMI	0.51
7	Nanodara	0.2935 -0.000106*PM10 -0.000027*CO + 0.001257*LST +0.06732*SMI -0.77359*NDMI	0.50
8	Vekeriya	0.3152 -0.00077*PM10 +0.00208*CO - 0.00136*LST+0.2417*SMI +0.5183*NDMI	0.44
9	Ranagadh	0.2031 +0.000262*PM10 +0.0003467*CO + 0.0000794*LST +0.0599*SMI + 0.05294*NDMI	0.42
10	Asalgam	-0.01539 +0.001504*PM10 -0.00155*CO +0.000408*LST +0.08975*SMI -0.2409*NDMI	0.34
11	Durgi	0.2518 + 0.0000648*PM10 + 0.0000754*CO +0.000708*LST +0.0718*SMI – 0.46412*NDMI	0.30
12	Devadthal	0.1776 + 0.0000509*PM10 + 0.00112*CO + 0.000275*LST + 0.1446*SMI + 0.3495*NDMI	0.20

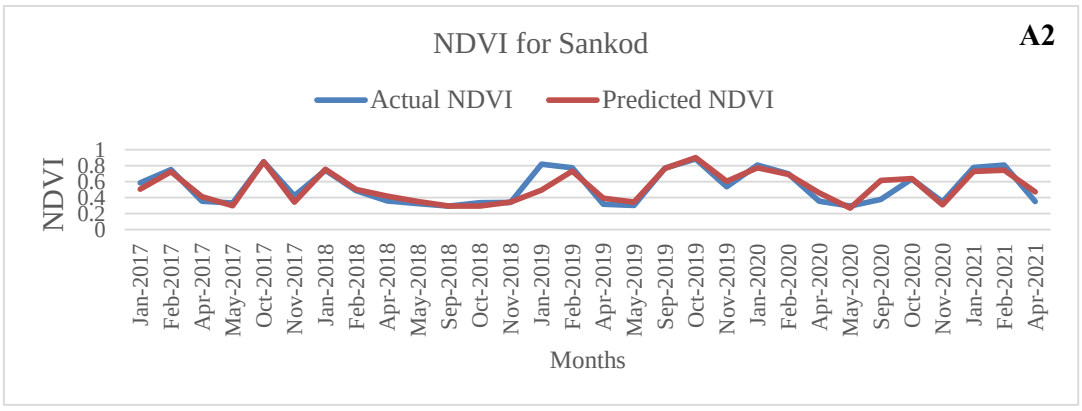
13	Rupavati	0.2651 +0.000428*PM10 +0.000186*CO -0.00012*LST +0.03396*SMI +0.1469*NDMI	0.16

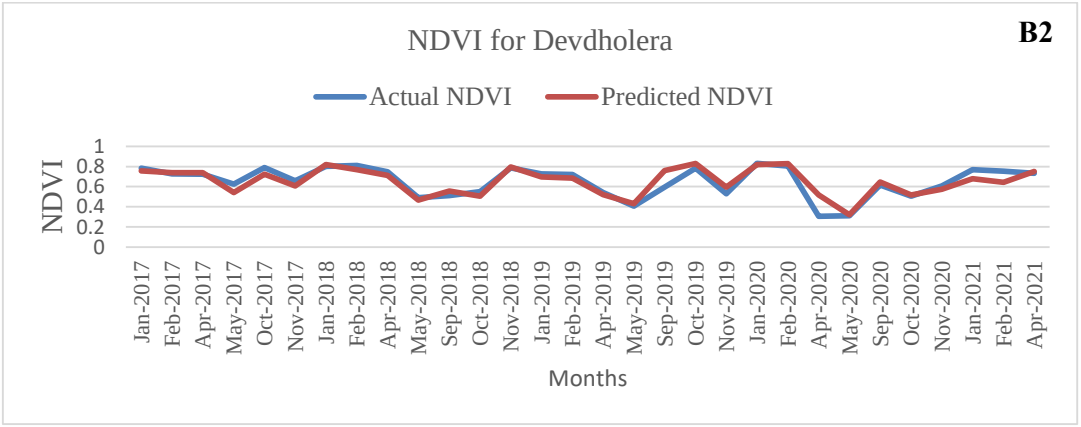


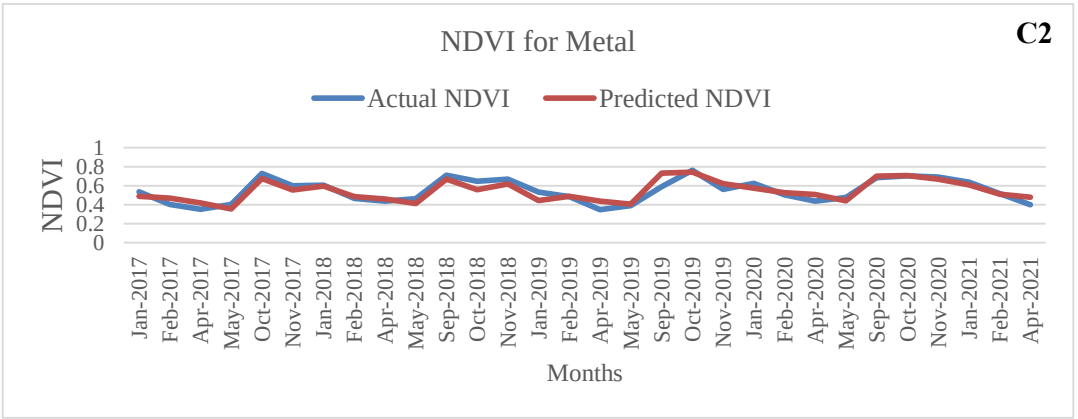





D1
NDVI for Navrangpura
NDVI
Predicted NDVI
NDVI
0.8
0.6
0.4
0.2
0
Jan-2017
Feb-2017
Apr-2017
May-2017
Oct-2017
Nov-2017
Jan-2018
Feb-2018
Apr-2018
May-2018
Sep-2018
Oct-2018
Nov-2018
Jan-2019
Feb-2019
Apr-2019
May-2019
Sep-2019
Oct-2019
Nov-2019
Jan-2020
Feb-2020
Apr-2020
May-2020
Sep-2020
Oct-2020
Nov-2020
Jan-2021
Feb-2021
Apr-2021
Months


A2
NDVI for Sankod
Actual NDVI
Predicted NDVI
NDVI
1
0.8
0.6
0.4
0.2
0
Jan-2017
Feb-2017
Apr-2017
May-2017
Oct-2017
Nov-2017
Jan-2018
Feb-2018
Apr-2018
May-2018
Sep-2018
Oct-2018
Nov-2018
Jan-2019
Feb-2019
Apr-2019
May-2019
Sep-2019
Oct-2019
Nov-2019
Jan-2020
Feb-2020
Apr-2020
May-2020
Sep-2020
Oct-2020
Nov-2020
Jan-2021
Feb-2021
Apr-2021
Months


B2
NDVI for Devdholera
Actual NDVI
Predicted NDVI
NDVI
1
0.8
0.6
0.4
0.2
0
Jan-2017
Feb-2017
Apr-2017
May-2017
Oct-2017
Nov-2017
Jan-2018
Feb-2018
Apr-2018
May-2018
Sep-2018
Oct-2018
Nov-2018
Jan-2019
Feb-2019
Apr-2019
May-2019
Sep-2019
Oct-2019
Nov-2019
Jan-2020
Feb-2020
Apr-2020
May-2020
Sep-2020
Oct-2020
Nov-2020
Jan-2021
Feb-2021
Apr-2021
Months


C2
NDVI for Metal
Actual NDVI
Predicted NDVI
NDVI
1
0.8
0.6
0.4
0.2
0
Jan-2017
Feb-2017
Apr-2017
May-2017
Oct-2017
Nov-2017
Jan-2018
Feb-2018
Apr-2018
May-2018
Sep-2018
Oct-2018
Nov-2018
Jan-2019
Feb-2019
Apr-2019
May-2019
Sep-2019
Oct-2019
Nov-2019
Jan-2020
Feb-2020
Apr-2020
May-2020
Sep-2020
Oct-2020
Nov-2020
Jan-2021
Feb-2021
Apr-2021
Months

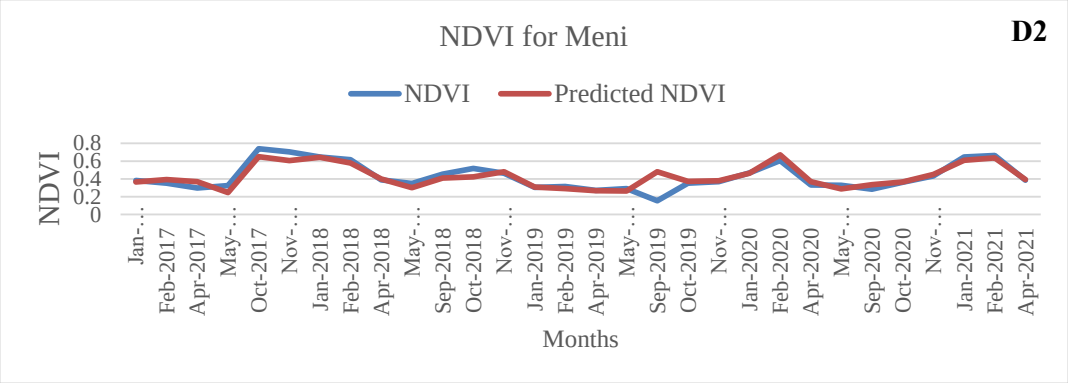


Figure 5: *Actual vs predicted NDVI for top 4 locations for both sites*

The graphical plot of Actual vs predicted NDVI for the top 4 locations for both sites are presented in Figures 5 A-D. In the above figure A1, B1, C1, and D1 represent the chart of the top four locations whose reliability is 0.92, 0.64, 0.61, and 0.54 at Ahmedabad city site respectively, while A2, B2, C2, and D2 represent the charts of top four locations whose reliability is 0.82, 0.8, 0.74, and 0.72 at the Nal-Sarovar site respectively. In most of the cases, the predicted values are closed to the actual values.

6. Conclusion

The relationship among the NDVI and other parameters was a point of concern and have shown a new dimension to carry out the study deeper into understand the change and the relation between vegetation and air pollution component over some time. Also, it is observed that at identical point in time, the mean NDVI is not the same for both the sites. Results obtained were based on shorter time frame. Further on extending study it may give encouraging results with remarkable predictability preciseness. The reliability of model is less as single factor is taken under the consideration apart from the air pollutants other factors alike soil fertility, water level,flow of wind, altitude etc can be considered for improving the defined model. Apart from linear model other models such as non-linear models, Gaussian dispersion models, model based on feature extraction techniques while constructing any deep machine learning models can be tried and tested for different conditions and locations. We believe that Geospatial data has more potential to answer some of the current ongoing discussions and questions at the global level for climate change, ecology cycle, and conservation issues.

References

Agrawal I. C., Gupta R. D., Gupta V. (2003). "GIS as modelling and decision support tool for air quality management: a conceptual framework", *6th International Conference on GIS/GPS/RS: MapIndia*, India.

Alseroury F. (2017). "Use of GIS to study the effect of air pollutants on the vegetation cover", *International Journal of Biosciences*, Vol 11, No. 6, pp 1-8.

Emberson L., Ashmore M., Murray F., Kuylenstierna J.C.I. (2001). "Impacts of Air Pollutants on Vegetation in Developing Countries", *Water Air and Soil Pollution*, Vol. 130, No. 1, pp. 107-118.

Famoso F., Wilson J., Monforte P., Brusca S., and Lulla V. (2017). "Measurement and modeling of ground-level ozone concentration in Catania, Italy using biophysical remote sensing and GIS", *International Journal of Applied Engineering Research*, Vol. 12, No. 21, pp. 10551-10562.

George J. E., Aravinth J., and Veni S. (2017). "Detection of Pollution Content in an Urban area using Landsat 8 Data", *IEEE*

Gheorghe I. F. and Ion B. (2011). "The Effects of Air Pollutants on Vegetation and the Role of Vegetation in Reducing Atmospheric Pollution", *The Impact of Air Pollution on Health, Economy, Environment and Agricultural Sources,*

Hurlock S. C., Stutz J. (2004). "GIS in air pollution research, the role of building surfaces", *ESRI International user conference*, San Diego, California, p 20-44.

Josef L., Pavel S., Daniel P., Natalia K., Jan S., Radovan H. and Premysl S., 2020, "Sentinel-2 Data in an Evaluation of the Impact of the Disturbances on Forest Vegetation", *Remote Sensing*, Vol. 12, No. 12, pp.19-14.

Lim H. S., Jafri M. Z. M., Abdullah K. (2010). "Algorithm for Air Quality Mapping Using Satellite Images", *Air Quality*

Lim H. S., MatJafri M. Z., Abdullah K. and Wong C. J. (2009). "Air Pollution Determination Using Remote Sensing Technique", *Advance in Geoscience and Remote Sensing*

Lorenz M., Clarke N., Paoletti E., Bytnerowicz A., Grulke N., Lukina N., Sase H. and Staelens J. (2010). "Air Pollution Impacts on Forests in a Changing Climate", Forest and Society – Responding to Global Drivers of Change, *International Union of Forest Research Organizations*, Vol.25, pp.55-74.

Manisalidis I., Stavropoulou E., Stavropoulos A. and Bezirtzoglou E. (2020). Environmental and Health Impacts of Air Pollution: A Review, *Front. Public Health*, Vol.8, No. 14.

Mendoza C.I.A., Teodora A.C., Torres N., Vevanco V. (2019). "Assessment of Remote Sensing Data to Model PM10 Estimation in Cities with a Low Number of Air Quality Stations: A Case of Study in Quito", *Environments*, Vol. 6, No. 7, pp 85.

Mishra M. (2019). "Poison in the air: Declining air quality in India", *lungindia*, Vol. 36, No.2, pp. 160–161.

Mozumder C., Reddy K. V., Pratap D. (2012). “Air Pollution Modeling from Remotely Sensed Data Using Regression Techniques”, *Indian Society of Remote Sensing,*Vol. 41, pp. 269-277

Rahman S. and Victor M. (2019). “Change Vector Analysis, Tasseled Cap, and NDVI-NDMI for Measuring Land Use/Cover Changes Caused by a Sudden Short-Term Severe Drought: 2011 Texas Event”, *MDPI,* Vol. 11, No. 19, pp. 22-17.

Salah A.H. (2011). “Air Quality Over Baghdad City Using Earth Observation and Landsat Thermal Data”, *Journal of Asian Scientific Research,* Vol. 1, No. 6, pp. 291-298.

Salah A.H.S. and Ghada H. (2014). “Estimation of PM10 Concentration using Ground Measurements and Landsat 8 OLI Satellite Image”, *Journal of Remote Sensing & GIS,* Vol. 3, No. 2.

Sohrabinia M., Khorshiddoust A.M. (2007). “Application of satellite data and GIS in studying air pollutants in Tehran”, *Habitat International,* Vol. 31, pp. 268-275.

Somvanshi S. S., Vashisht A., Chandra U. and Kaushik G. (2019). “Delhi Air Pollution Modelling Using Remote Sensing Technique”, *Handbook of Environmental Materials Management, Springer,* pp 1-27.

Stathopoulou, Marina, and Cartalis C., 2007, “Daytime urban heat islands from Landsat ETM+ and Corine land cover data: An application to major cities in Greece.” *Solar Energy,* Vol. 81, No. 3, pp. 358-368

Stevens C. J., Bell J. N. B., Brimblecombe P., Clark C. M., Dise N. B., Fowler D., Lovett G. M. and Wolseley P. A., 2020, “The impact of air pollution on terrestrial managed and natural vegetation”, *Philos Trans A Math Phys Eng Sci,* Vol. 378, No. 2183

Vashisht A. and Somvanshi S., 2018, “Use of remote sensing technique in Air Quality Modelling of Delhi Region”, *19th Esri India User Conference,* Gurugram, NCR

Yuan, Fei, and Bauer M. E., 2007, “Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery”, *Remote Sensing of environment,* Vol. 106, No. 3, pp. 375-386

https://greentumble.com/effect-of-pollution-on-plants/

Estimating the Finite Population Mean With Known Coefficient of Variation of Study Variable and Using Information on Auxiliary Variable under Scrambled Response Model in Presence of Non-Response

Housila P. Singh and Preeti Patidar
School of Studies in Statistics
Vikram University, Ujjain - 456010, M.P., India.

Received: 28 Jan 2022 / Revised: 24 October 2022 / Accepted: 21 July 2023

ABSTRACT

Taking motivation from Searls (1964), Hansen and Hurwitz (1946), Khare and Kumar (2011), Diana and Perri (2011), Diana et al (2014), Ahmed et al (2017) proposed an estimator for population mean of a sensitive quantitative variable with known coefficient of variation of the study variable using auxiliary information in two phase sampling scheme considering a randomization mechanism on the second call that provides privacy protection to the respondents to obtain truthful information. Ahmed et al (2017) have further suggested generalized ratio and regression- type estimators under two phase sampling scheme. In this paper we have suggested a model which is more general and more efficient than Ahmed et al (2017) model under realistic condition. We have suggested three classes of estimators under two phase sampling scheme for finite population mean exploiting the same amount of information considering a randomization mechanism on the second call. The properties of the suggested classes of estimators are studied under large sample approximation. It is observed that the mean squared errors obtained by Ahmed et al (2017) of their estimators are incorrect. So we have also obtained the correct expressions of the mean squared errors of Ahmed et al (2017) estimators. Some more efficient estimators are investigated. An empirical study is carried out to judge the performances of the suggested classes of estimators.

KEYWORDS: Quantitative sensitive variable, Study variable, Auxiliary variable, Non-response, Coefficient of variation.

AMS Classification: 62D05

1. INTRODUCTION

In human sample surveys, issues with non-response arise when people are approached via telephone, mail or direct interviews. There are many aspects in which problems with non-response arise. It depends on the essence of the information needed, whether the survey concerns general or sensitive issues of society. Typically, when surveys are performed to gather general information about people such as age, schooling, income, household size, etc., the prevalence of the non-response may be in the form of people's unavailability, not at home, unable to understand the questionnaire, etc. On the other hand, if the information needed concerns sensitive issues such as betting, money laundering, drug dependency, number of abortions, then people usually hesitate to offer true answers and ultimately fail to reply or deliver an evasive response. In such situations, to reduce the non-response bias and to estimate the population parameters, Hansen and Hurwitz (1946) introduced a procedure for sub-sampling the non-respondents, in which it is supposed that all person give full response on second call.

In the case of surveys relevant to the sensitive characteristics of the population, reducing the bias of non-response and obtaining accurate information from respondents rely on the security of their confidentiality and privacy there are some statistical procedures for the interviewee to collect true data in order to preserve privacy. These procedures are known as Randomized Response techniques (RRTs). Warner (1965) first used the RRT to estimate the proportion of population possessing a sensitive feature that requires a 'yes' or 'no' response, or to pick a response from a range of nominal categories. Later on, several authors have given contribution for improving efficiency of this procedure, among others Fox and Tracy (1986), Mangat and Singh (1990), Shabbir and Gupta (2005), Diana and Perri (2009), Singh and Gorey (2016), Singh and Tarray (2014). RRT provides quantitative response that depends on a random number from a known distribution. Using the idea of quantitative sensitive response, some authors proposed the Scrambled RRTs, for instance, see Pollock and Bek (1976), Eichorn and Hayre (1983), Diana and Perri (2010, 2011), Tarray and Singh (2016), Bar-Lev et al (2004), Saha (2007). Diana et al (2014) developed a modified version of the Hansen and Hurwitz (1946) estimator for quantitative study variable and considered a second call randomization mechanism that would provide privacy protection for respondents in order to obtain truthful information. This estimator decreases non response bias but increases variance owing to the use of Scrambled RRT in the non-response class.

It is well founded that the use of auxiliary information provides efficient estimators at the stage of estimation, for instance see, Singh H.P. (1986) and Singh S. (2003). Searls (1964) was first who suggests an estimator for population mean using the known coefficient of variation of the study variable *y*. After that, some authors including

Searls (1967), Sen (1978, 1979), Upadhaya and Singh (1984) and Singh and Katyar (1988) have used the coefficient of variation for estimating the population mean. Khare and Kumar (2009) proposed ratio, product and regression type estimators of population mean in the presence of non-response using coefficient of variation of the study character *y*. This study has been further extended by Khare and Kumar (2011) for estimating the population mean using known coefficient of variation of the study character in two phase sampling in presence of non-response. Rajyaguru and Gupta (1995, 2004, 2006) have also discussed the problem of estimation of coefficient of variation. Ahmed et al (2017) suggested an estimator for population mean of a quantitative study variable utilizing known coefficient of variation of the study variable under two-phase sampling scheme using the scrambled response to non-respondents on the second call.

In this paper we propose three classes of estimators under scrambled response model using auxiliary information under two-phase sampling scheme for population mean with known coefficient of variation of the study character in presence of non - response. The bias and mean square errors (MSEs) of the suggested classes of estimators under large sample approximation have been obtained. We have also given the correct MSE expressions of the estimators envisaged by Ahmed et al (2017). Merits of the proposed classes of estimators are evaluated through an empirical study.

1.1 BACKGROUND AND HANSEN AND HURWITZ (1946) ESTIMATOR

Consider a finite population $\Omega = (\Omega_1, \Omega_2, ..., \Omega_N)$ of size *N* (units). Let (y_i, x_i) be the values on ith unit for study character *y* and auxiliary character *x* of the population Ω. Here *y* denotes quantitative study variable of interest with unknown population mean $\bar{Y}$ and unknown population variance S_y^2 supposing that non-response occurs in y. When the population mean $\bar{X}$ of the insensitive auxiliary variable *x* is not known then we use the two-phase (or double)-sampling scheme. In the first-phase, select a large sample $s_{n'}$ of size $n'(n' < N)$ by simple random sampling without replacement (SRSWOR) to estimate population mean $\bar{X}$ and then in the second phase take a smaller sub- sample s_n of size *n* from $n'(n < n')$ by SRSWOR to estimate $(\bar{Y}, \bar{X})$. Hansen and Hurwitz (1946) proposed the following sub-sampling scheme. Suppose that from *n* sample units, a subset s_1 of size n_1 supplies information on the *y* and, the remaining $n_2 = n - n_1$ units are non-respondents. Then a sub sample s_{2r} of size $r = n_2/k$, $k > 1$, is selected from n_2 non-response units, where *r* would be an integer otherwise it must be rounded. Assume that all *r* selected units show full response on second call. Consequently, the entire population Ω is divided into two groups P_1 and P_2, where P_1 is the group of the respondents of size

N_1 that would give response on first call at the second phase, whereas P_2 is non-respondents group of size N_2 which would not respond on first call at the second phase but will cooperate on the second call. Obviously, N_1 and N_2 are unknown.

We denote $(\bar{y}, \bar{x})$ the sample means of (y, x) respectively based on a sample of size *n* drawn from population Ω using SRSWOR. Ignoring fpc term, Searls (1964) suggested an estimator for population mean $\bar{Y}$ as

$$t_1 = \frac{n}{n + C_y^2}\,\bar{y}, \tag{1.1}$$

where $C_y = \frac{S_y}{\bar{Y}}$ is the known coefficient of variation of *y*, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean.

Hansen and Hurwitz (1946) suggested an unbiased estimator of the population mean $\bar{Y}$ of the study variable *y* as

$$\bar{y}^* = \frac{n_1}{n}\,\bar{y}_1 + \frac{n_2}{n}\,\bar{y}_2, \tag{1.2}$$

where $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i$, $\bar{y}_2 = \frac{1}{r}\sum_{i=1}^{r} y_i$ and *r* is the responding units.

The variance/ MSE of $\bar{y}^*$ due to Hansen and Hurwitz (1946) is given by

$$MSE\left(\bar{y}^*\right) = Var\left(\bar{y}^*\right) = \frac{(1-f)}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2,$$

where $f = \frac{n}{N}$ is the sampling fraction, $S_{y(2)}^2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2}\left(y_i - \bar{Y}_2\right)^2$, denotes the population variance for the non-responding part of the population for the study variable *y* and $\bar{Y}_2 = \sum_{i=1}^{N_2} y_i \Big/ N_2$.

By ignoring correction factor $(1-f)$ for ease of computations, we have

$$MSE\left(\bar{y}^*\right) = Var\left(\bar{y}^*\right) = \frac{1}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2, \tag{1.3}$$

see Ahmed et al (2017, p. 8437).

Motivated by Searls (1964), Khare and Kumar (2009, 2011) envisaged an improved version of Hansen and Hurwitz (1946) estimator $\bar{y}^*$ as

$t = a\,\bar{y}^*$,

where '*a*' is the suitably chosen constant to be determined such that *MSE* of *t* is minimum.

The *MSE* of *t* is given by

$$MSE(t)=\left[V(t)+(B(t)^2)\right]=\left[a^2V\left(\bar{y}^*\right)+(a-1)^2\bar{Y}^2\right]=\bar{Y}^2\left[a^2\left(1+\frac{V\left(\bar{y}^*\right)}{\bar{Y}^2}\right)-2a+1\right].$$

Using (1.3) in the above expression, we have

$$MSE(t)=\bar{Y}^2\left[a^2\left(1+\frac{1}{n}C_y^2+\frac{W_2(k-1)}{n}\frac{S_{y(2)}^2}{\bar{Y}^2}\right)-2a+1\right],$$

where $C_y^2=\frac{S_y^2}{\bar{Y}^2}$ is the square of the coefficient of variation of the study variable *y*.

Setting $\frac{\partial MSE(t)}{\partial a}=0$, we get the optimum value of '*a*' for which the $MSE(t)$ is minimum as

$$a_{opt}=\left(1+\frac{1}{n}C_y^2+\frac{W_2(k-1)}{n}\frac{S_{y(2)}^2}{\bar{Y}^2}\right)^{-1}.$$

Under the assumption that $\frac{S_y^2}{\bar{Y}^2}$ and $\frac{S_{y(2)}^2}{\bar{Y}^2}$ do not differ significantly, Khare and Kumar (2009, 2011) and Ahmed et al (2017) approximated $\frac{S_y^2}{\bar{Y}^2}\cong\frac{S_{y(2)}^2}{\bar{Y}^2}=C_y^2$.

Replacing $W_2=\frac{N_2}{N}$, by its unbiased estimator $\hat{W}_2=\frac{n_2}{n}$ and $\frac{S_{y(2)}^2}{\bar{Y}^2}by C_y^2$ in 'a_{opt}', Khare and Kumar (2009, 2011) and Ahmed et al (2017) derived an estimator of a_{opt} as

$$\hat{a}_{opt}=\left[1+\frac{C_y^2}{n}\left\{1+\frac{n_2}{n}(k-1)\right\}\right]^{-1}.$$

Thus the improved version of Hansen and Hurwitz (1946) for population mean $\bar{Y}$ of *y* is given by

$$t_2=\left[1+\frac{C_y^2}{n}\left\{1+\frac{n_2}{n}(k-1)\right\}\right]^{-1}\bar{y}^*. \qquad (1.4)$$

which has the MSE

$$MSE(t_2)=\left[(1-A)\frac{S_y^2}{n}+(1-2B)\frac{W_2(k-1)}{n}S_{y(2)}^2\right], \qquad (1.5)$$

where $A=\frac{C_y^2}{n}\left\{1-W_2^2(k-1)^2\right\}$ and $B=\frac{C_y^2}{n}\left\{1+W_2(k-1)\right\}$.

Here we mention that Khare and Kumar (2009, 2011) assumed that $\frac{S_y^2}{\bar{Y}^2}$ and $\frac{S_{y(2)}^2}{\bar{Y}^2}$ do not differ significantly, therefore they have approximated $\frac{S_y^2}{\bar{Y}^2} \cong \frac{S_{y(2)}^2}{\bar{Y}^2} = C_y^2$, which is not appropriate because $S_{y(2)}^2$ is either greater than S_y^2 or smaller than S_y^2 i.e. $S_{y(2)}^2 > S_y^2$ or $S_{y(2)}^2 < S_y^2$. Thus the assumption $S_{y(2)}^2 = \delta S_y^2$ is more flexible and appropriate such that $\delta > 0$.

Keeping this in view, we suggest the following estimator for the population mean $\bar{Y}$ as

$$t_3 = \left[1 + \frac{C_y^2}{n}\left\{1 + \frac{n_2}{n}(k-1)\delta\right\}\right]^{-1} \bar{y}^*, \tag{1.6}$$

The MSE of t_3 is given by

$$MSE(t_3) = \left[(1 - A_\delta)\frac{S_y^2}{n} + (1 - 2B_\delta)\frac{W_2(k-1)}{n} S_{y(2)}^2\right], \tag{1.7}$$

where $A_\delta = \frac{C_y^2}{n}\left\{1 - W_2^2(k-1)^2\delta^2\right\}$ and $B_\delta = \frac{C_y^2}{n}\left\{1 + W_2(k-1)\delta\right\}$.

1.2 SUGGESTED IMPROVED MODEL

Motivated by Diana et al (2014) and Ahmed et al (2017) we have suggested the following linear scrambled RR model.

Let Z be the scrambled response and (V_1, V_2) be two independent random variables unrelated to the study variable y, with known means (μ_{v_1}, μ_{v_2}) and variances $(\sigma_{v_1}^2, \sigma_{v_2}^2)$.

Let $Z_\eta = V_1 Y + \eta V_2$, (1.8)

$E_R(Z_\eta) = \mu_{v_1} Y + \eta\mu_{v_2}$, (1.9)

and $V_R(Z_\eta) = \sigma_{v_1}^2 Y^2 + \eta^2\sigma_{v_2}^2$, (1.10)

where (E_R, V_R) are expectation and variance under randomization device and η is suitably chosen scalar.

In order to increase trust in the respondents about their privacy security, it is presumed that the interviewer is fully unaware of the numbers created by the respondents from the scrambling distributions V_1 and V_2.

Let $\hat{y}_{\eta i}$ be the transformation of RR of the i^{th} unit whose expectation under the randomization mechanism coincides with the true response y_i as

$$\hat{y}_{\eta i} = \frac{z_{\eta i} - \eta\mu_{v_2}}{\mu_{v_1}}, \tag{1.11}$$

and

$E_R(\hat{y}_{\eta i}) = y_i$.

The variance of $\hat{y}_{\eta i}$ is given by

$$V_R(\hat{y}_{\eta i}) = \frac{\sigma_{v_1}^2 y_i^2 + \eta^2 \sigma_{v_2}^2}{\mu_{v_1}^2} = \phi_{\eta i} \quad . \tag{1.12}$$

Following Diana et al (2014) we propose the following unbiased estimator for $\bar{Y}$ as

$$\hat{\bar{y}}_{\eta}^{*} = w_1 \bar{y}_1 + w_2 \hat{\bar{y}}_{\eta 2}^{'}, \tag{1.13}$$

where $\hat{\bar{y}}_{\eta 2}^{'} = \dfrac{\sum_{i=1}^{r} \hat{y}_{\eta i}}{r}$.

Ignoring the finite population correction (fpc) term, the variance of $\hat{\bar{y}}_{\eta}^{*}$ is given by

$$Var(\hat{\bar{y}}_{\eta}^{*}) = \frac{1}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2 + \frac{k}{nN} \sum_{i=1}^{N_2} \phi_{\eta i}, \tag{1.14}$$

where $\dfrac{1}{N_2} \sum_{i=1}^{N_2} \phi_{\eta i} = \left\{ \dfrac{\sigma_{v_1}^2 \mu_{2,y} + \eta^2 \sigma_{v_2}^2}{\mu_{v_1}^2} \right\}$, and $\mu_{2,y} = S_{y(2)}^2 + \bar{Y}_2^2$. (1.15)

We note that the unknown $\mu_{2,y}$ can be calculated by two possible methods, one of which is to use good guesses from previous work or pilot surveys, and otherwise the sample estimate must include details on the second moment taking into account its sensitive nature, see Diana and Perri (2010), Diana et al (2014) and Ahmed et al (2017).

We mention that if we set $\eta = 1$ in the model (1.8), it reduces to the Ahmed et al (2017) model

$$Z = V_1 Y + V_2, \tag{1.16}$$

and hence the estimator $\hat{\bar{y}}_{\eta}^{*}$ reduces to the estimator

$$\hat{\bar{y}}^{*} = w_1 \bar{y}_1 + w_2 \hat{\bar{y}}_2^{'}, \tag{1.17}$$

which is due to Diana et al (2014), $\hat{\bar{y}}_2 = \dfrac{1}{r} \sum_{i=1}^{r} \hat{y}_i$ with $\hat{y}_i = \dfrac{z_i - \mu_{v_2}}{\mu_{v_1}}$.

Putting $\eta = 1$ in (1.14) we get the $Var(\hat{\bar{y}}^{*})$ as

$$Var(\hat{\bar{y}}^{*}) = \frac{1}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2 + \frac{k}{nN} \sum_{i=1}^{N_2} \phi_i, \tag{1.18}$$

where $\phi_i = \dfrac{\sigma_{v_1}^2 y_i^2 + \sigma_{v_2}^2}{\mu_{v_1}^2}$.

It is to be mentioned that Diana et al (2014) have made a tradeoff between efficiency and confidentiality by selecting a suitable scrambled response among different models because efficiency and confidentiality walk in opposite direction. So it is hard to keep both of these on a desired level for a fixed sample size. These lead authors to make an effort to improve efficiency at a fixed level of confidentiality, (see Ahmed et al (2017, p.8439)). To fulfill this objective, we have made an effort to formulate some estimators of population mean of a quantitative study variable using known coefficient of variation of the study variable which is more efficient than Diana et al (2014) and Ahmed et al (2017).

From (1.14) and (1.18) we have

$$Var(\hat{\bar{y}}^{*}) - Var(\hat{\bar{y}}_{\eta}^{*}) = \frac{kN_2\sigma_{v_2}^2}{nN\mu_{v_1}^2}\left(1-\eta^2\right),$$

which is non-negative if $\left(1-\eta^2\right) > 0$

i.e. if $|\eta| < 1.$ (1.19)

Expression (1.19) shows that the propounded estimator $\hat{y}_{\eta}^{*}$ is always better than Diana et al (2014) estimator $\hat{\bar{y}}^{*}$ as long as the condition $|\eta| < 1$ is satisfied.

Thus to obtain the better estimates from the propounded model we will put the restriction on η as $|\eta| < 1$ i.e.

$$Z_{\eta} = V_1 Y + \eta V_2, \tag{1.20}$$

where η is a scalar such that $|\eta| < 1.$

Following the same procedure as adopted by Searls (1964) we define an improved version of $\hat{\bar{y}}_{\eta}^{*}$ for the population mean $\bar{Y}$ as

$$\hat{t}_s = \hat{M}_{0\delta}\,\hat{\bar{y}}_{\eta}^{*} = \left[1 + \frac{C_y^2}{n}\left\{1 + \frac{n_2}{n}(k-1)\delta\right\} + \frac{k}{nN}\frac{S_{\eta r}^2}{\bar{Y}^2}\right]^{-1}\hat{\bar{y}}_{\eta}^{*}. \tag{1.21}$$

The bias and MSE of $\hat{t}_s$ are respectively given by

$$B(\hat{t}_s) = -\bar{Y}\left[B_{\eta}^{*} - \frac{1}{n^2}\left\{C_y^4\left\{1 + 2W_2(k-1)\delta\right\} + \frac{k^2}{N^2}\frac{S_{\eta r}^4}{\bar{Y}^2} + 2\frac{kC_y^2 S_{\eta r}^2}{N\bar{Y}^2}\left\{1 + W_2(k-1)\delta\right\}\right\}\right], \tag{1.22}$$

$$MSE(\hat{t}_s) = \left[\left(1 - A_{\eta}^{*}\right)\frac{S_y^2}{n} + \left(1 - 2B_{\eta}^{*}\right)\frac{W_2(k-1)}{n}S_{y(2)}^2 + \frac{kS_{\eta r}^2}{nN}\left(1 - \frac{k}{nN}\frac{S_{\eta r}^2}{\bar{Y}^2}\right)\right], \tag{1.23}$$

where $A_{\eta}^{*} = \left[\frac{C_y^2}{n}\left\{1 - W_2^2(k-1)^2\delta^2\right\} + \frac{2k}{nN}\frac{S_{\eta r}^2}{\bar{Y}^2}\right]$ and $B_{\eta}^{*} = \left[\frac{C_y^2}{n}\left\{1 + W_2(k-1)\delta\right\} + \frac{k}{nN}\frac{S_{\eta r}^2}{\bar{Y}^2}\right].$

If we set $(\eta, \delta)=(1,1)$ in (1.21), the estimator $\hat{t}_s$ reduces to the estimator due to Ahmed et al (2017):

$$\hat{\bar{y}}^{**}=\left[1+\frac{C_y^2}{n}\left\{1+\frac{n_2}{n}(k-1)\right\}+\frac{k}{nN}\frac{S_r^2}{\bar{Y}^2}\right]^{-1}\hat{\bar{y}}^{*}, \tag{1.24}$$

where $\hat{\bar{y}}^{*}=\frac{n_1}{n}\bar{y}_1+\frac{n_2}{n}\hat{\bar{y}}_2'$ with $\hat{\bar{y}}_2'=\frac{1}{r}\sum_{i=1}^{r}\hat{y}_i$, $S_r^2=\sum_{i=1}^{N_2}\phi_i=N_2\frac{\left(\sigma_{v_1}^2\mu_{2,y}+\sigma_{v_2}^2\right)}{\mu_{v_1}^2}$.

Putting $(\eta, \delta)=(1,1)$ in (1.22) and (1.23) respectively we get the bias and MSE of the Ahmed et al (2017) estimator $\hat{\bar{y}}^{**}$ as

$$B\left(\hat{\bar{y}}^{**}\right)=-\bar{Y}\left[B^{*}-\frac{1}{n^2}\left\{C_y^4\left\{1+2W_2(k-1)\right\}+\frac{k^2}{N^2}\frac{S_r^4}{\bar{Y}^2}+2\frac{kC_y^2S_r^2}{N\bar{Y}^2}\left\{1+W_2(k-1)\right\}\right\}\right], \tag{1.25}$$

$$MSE\left(\hat{\bar{y}}^{**}\right)=\left[\left(1-A^{*}\right)\frac{S_y^2}{n}+\left(1-2B^{*}\right)\frac{W_2(k-1)}{n}S_{y(2)}^2+\frac{kS_r^2}{nN}\left(1-\frac{k}{nN}\frac{S_r^2}{\bar{Y}^2}\right)\right], \tag{1.26}$$

where $A^{*}=\left[\frac{C_y^2}{n}\left\{1-W_2^2(k-1)^2\right\}+\frac{2k}{nN}\frac{S_r^2}{\bar{Y}^2}\right]$ and $B^{*}=\left[\frac{C_y^2}{n}\left\{1+W_2(k-1)\right\}+\frac{k}{nN}\frac{S_r^2}{\bar{Y}^2}\right]$.

Further we define the following estimators for population mean $\bar{Y}$ as

$$\hat{t}_{s1}=\left[1+\frac{C_y^2}{n}\left\{1+\frac{n_2}{n}(k-1)\delta\right\}\right]^{-1}\hat{\bar{y}}_{\eta}^{*}\ . \tag{1.27}$$

and

$$\hat{t}_{s2}=\left(\frac{n}{n+C_y^2}\right)\hat{\bar{y}}_{\eta}^{*}. \tag{1.28}$$

The bias and MSE of $\hat{t}_{s1}$ and $\hat{t}_{s2}$ are respectively given by

$$B\left(\hat{t}_{s1}\right)=-B_{\delta}\left(1-B_{\delta}\right)\bar{Y}, \tag{1.29}$$

$$B\left(\hat{t}_{s2}\right)=-\left(\frac{C_y^2}{n+C_y^2}\right)\bar{Y}, \tag{1.30}$$

$$MSE\left(\hat{t}_{s1}\right)=\left[\left(1-A_{\delta}\right)\frac{S_y^2}{n}+\left(1-2B_{\delta}\right)\frac{W_2(k-1)}{n}S_{y(2)}^2+\left(1-2B_{\delta}\right)\frac{kS_{\eta r}^2}{nN}\right]. \tag{1.31}$$

$$MSE\left(\hat{t}_{s2}\right)=\left[\frac{S_y^2}{\left(n+C_y^2\right)}+\frac{n}{\left(n+C_y^2\right)^2}\left\{W_2(k-1)S_{y(2)}^2+\frac{kS_{\eta r}^2}{N}\right\}\right]\ ,$$

$$\cong\left[\left(1-\frac{C_y^2}{n}\right)\frac{S_y^2}{n}+\left(1-2\frac{C_y^2}{n}\right)\frac{W_2(k-1)}{n}S_{y(2)}^2+\left(1-2\frac{C_y^2}{n}\right)\frac{kS_{\eta r}^2}{nN}\right]. \tag{1.32}$$

1.4 EFFICIENCY COMPARISON

From (1.23), (1.31) and (1.32) we have

$$MSE(\hat{t}_{s1}) - MSE(\hat{t}_s) = \frac{2k^2 S_{\eta r}^4}{n^2 N^2 \bar{Y}^2} \geq 0 \text{ (under the assumption that } S_{y(2)}^2 = \delta S_y^2 \text{ with } \delta > 0) \quad (1.33)$$

$$MSE(\hat{t}_{s2}) - MSE(\hat{t}_{s1}) = \frac{C_y^2}{n^2}\left[\frac{S_y^2 C_y^2}{(n+C_y^2)} + C_y^2\left\{W_2(k-1)S_{y(2)}^2 + \frac{kS_{\eta r}^2}{N}\right\}\left\{\frac{C_y^2(3n+2C_y^2)}{(n+C_y^2)^2} + 2W_2(k-1)\delta\right\}\right]$$

≥ 0 (under the assumption that $S_{y(2)}^2 = \delta S_y^2$). (1.34)

Expressions (1.33) and (1.34) give the inequality

$$MSE(\hat{t}_s) \leq MSE(\hat{t}_{s1}) \leq MSE(\hat{t}_{s2}). \quad (1.35)$$

It follows that the proposed estimator $\hat{t}_s$ is better than $\hat{t}_{s1}$ and $\hat{t}_{s2}$.

From (1.14) and (1.32) we have

$$MSE(\hat{t}_{s2}) - Var(\hat{\bar{y}}_\eta^*) = -\frac{C_y^2}{n(n+C_y^2)}\left[S_y^2 + \frac{(2n+C_y^2)}{(n+C_y^2)}\left\{W_2(k-1)S_{y(2)}^2 + \frac{kS_{\eta r}^2}{N}\right\}\right] \leq 0$$

which gives the inequality

$$MSE(\hat{t}_{s2}) \leq Var(\hat{\bar{y}}_\eta^*). \quad (1.36)$$

Further from (1.19) we note that

$$Var(\hat{\bar{y}}_\eta^*) \leq Var(\hat{\bar{y}}^*) \text{ provided } |\eta| < 1. \quad (1.37)$$

Combining (1.35), (1.36) and (1.37) we have the inequality:

$$MSE(\hat{t}_s) \leq MSE(\hat{t}_{s1}) \leq MSE(\hat{t}_{s2}) \leq Var(\hat{\bar{y}}_\eta^*) \leq Var(\hat{\bar{y}}^*) \text{ provided } |\eta| < 1. \quad (1.38)$$

It follows from (1.38) that the propounded estimator $\hat{t}_s$ is more efficient than $\hat{t}_{s1}$, $\hat{t}_{s2}$, $\bar{y}_\eta^*$ and Diana's et al (2014) estimator $\hat{\bar{y}}^*$ provided $|\eta| < 1$. However the propounded estimator $\hat{t}_s$ is always more efficient than $\hat{t}_{s1}$, $\hat{t}_{s2}$ and $\hat{\bar{y}}_\eta^*$. Here we note that the proposed estimator $\hat{t}_s$ depends upon the parameter $\bar{Y}$ under investigation which prevents the practical utility of the estimator $\hat{t}_s$ while the estimators $\hat{t}_{s1}$ and $\hat{t}_{s2}$ are free from such restrictions therefore the estimators $(\hat{t}_{s1}, \hat{t}_{s2})$ can be used in practice without any restriction.

Further we note from (1.14) and (1.23) that

$$Var(\hat{\bar{y}}_\eta^*) - MSE(\hat{t}_s) = \left[\frac{S_y^2}{n}A_\eta^* + \frac{2W_2(k-1)}{n}S_{y(2)}^2 B_\eta^* + \frac{k^2 S_{\eta r}^4}{n^2 N^2 \bar{Y}^2}\right]. \quad (1.39)$$

Under the assumption $S_{y(2)}^2 = \delta S_y^2, \delta > 0,$ (1.39) reduces to

$$Var\left(\hat{\bar{y}}_\eta^*\right) - MSE\left(\hat{t}_s\right) = \frac{S_y^2}{n}\left[A_\eta^* + 2W_2(k-1)B_\eta^*\right] + \frac{k^2 S_{\eta r}^4}{n^2 N^2 \bar{\bar{Y}}^2},$$

which is non-negative if

$$\frac{C_y^2}{n^2}\left\{1 + W_2(k-1)\delta\right\}^2 + \frac{2kS_{\eta r}^2}{nN\bar{\bar{Y}}^2}\left\{1 + W_2(k-1)\delta\right\} > 0,$$

i.e if $\left\{1 + W_2(k-1)\delta\right\}^2 > 0$ and $\left\{1 + W_2(k-1)\delta\right\} > 0$, both conditions are satisfied for $k > 1$.

This shows that the proposed estimator $\hat{t}_s$ performs better than the suggested estimator $\hat{\bar{y}}_\eta^*$ and hence it is more efficient than Diana et al (2014) estimator $\hat{\bar{y}}^*$.

Further from (1.39) we have that $MSE\left(\hat{t}_s\right) \le Var\left(\hat{\bar{y}}_\eta^*\right)$ if

$$1 < k < \left(1 + 1/W_2\delta\right), \tag{1.40}$$

Under this condition, the estimator $\bar{y}_\eta^*$ (and hence Diana et al (2014) estimator $\hat{\bar{y}}^*$) may be improved by replacing a better estimator $\hat{t}_s$ for population mean $\bar{\bar{Y}}$ than $\hat{\bar{y}}^*$ in case of non-response in sample survey. However the estimator $\hat{t}_s$ may also be more efficient than $\hat{\bar{y}}^*$ beyond the range $k > \left(1 + 1/W_2\delta\right)$.

2. SOME MODELS FOR CONSIDERATION

To examine the performance of the estimators we consider four known scrambled RR models of additive, multiplicative and mixed nature. These models are derived from the general linear scrambled randomized response model earlier considered by Diana and Perri (2010) are showed in the following scheme.

Table 2.1: The members of scrambled response model.

S.No.	Authors	Model	V_1	V_2
1.	Pollock and Bek (1976)-type	$M_1 : Z = Y + \eta U_2$	1	U_2
2.	Eichhorn and Hayre (1983)-type	$M_2 : Z = U_1 Y$	U_1	0
3.	Saha (2007)-type	$M_3 : Z = U_1(Y + \eta U_2)$	U_1	$U_1 U_2$
4.	Diana and Perri (2011)-type	$M_4 : Z = \varphi(Y + \eta U_2) + (1-\varphi)U_1 Y$	$\varphi + (1-\varphi)U_1$	φU_2

where φ is a scalar in the interval $(0,1)$.

We use $S_{\eta r(j)}, j=1,2,3,4$ instead of $S_{\eta r}^2$ to get the MSEs of suggested estimators under model $M_j, j=1to4$ respectively, where

$$S_{\eta r(1)}^2 = N_2\eta^2\sigma_{u_2}^2, \quad S_{\eta r(2)}^2 = N_2\left(\frac{\sigma_{u_1}^2\mu_{2,y}}{\mu_{u_1}^2}\right),$$

$$S_{\eta r(3)}^2 = N_2\left\{\frac{\sigma_{u_1}^2\mu_{2,y} + \mu_{u_1}^2\eta^2\sigma_{u_2}^2 + 2\sigma_{u_1}^2\eta\mu_{u_2}\bar{Y}_2 + \sigma_{u_1}^2\eta^2\left(\mu_{u_2}^2+\sigma_{u_2}^2\right)}{\mu_{u_1}^2}\right\},$$

$$S_{\eta r(4)}^2 = \left[\frac{(1-\varphi)^2\sigma_{u_1}^2\mu_{2,y} + \eta^2\varphi^2\sigma_{u_2}^2}{\left(\varphi+(1-\varphi)\mu_{u_1}\right)^2}\right].$$

The MSE of the propounded estimators under different scrambled RR models can be derived by putting $S_{\eta r(1)}^2$, $S_{\eta r(2)}^2, S_{\eta r(3)}^2$ and $S_{\eta r(4)}^2$ in place of $S_{\eta r}^2$ in expression respective of MSE.

If we set $\eta=1$ in $S_{\eta r(j)}, j=1to4$ and we use $S_{r(j)}$ instead of S_r^2 to get suggested estimators under model $M_j^*, j=1to4$ respectively, where

$$S_{r(1)}^2 = N_2\sigma_{u_2}^2,\ S_{r(2)}^2 = N_2\left(\frac{\sigma_{u_1}^2\mu_{2,y}}{\mu_{u_1}^2}\right),$$

$$S_{r(3)}^2 = N_2\left\{\frac{\sigma_{u_1}^2\mu_{2,y} + \mu_{u_1}^2\sigma_{u_2}^2 + 2\sigma_{u_1}^2\mu_{u_2}\bar{Y}_2 + \sigma_{u_1}^2\left(\mu_{u_2}^2+\sigma_{u_2}^2\right)}{\mu_{u_1}^2}\right\},$$

$$S_{r(4)}^2 = \left[\frac{(1-\varphi)^2\sigma_{u_1}^2\mu_{2,y} + \varphi^2\sigma_{u_2}^2}{\left(\varphi+(1-\varphi)\mu_{u_1}\right)^2}\right],$$

$M_1^*:\ Z = Y + U_2 \rightarrow$ Additive model due to Pollock and Bek(1976),

$M_2^*:\ Z = U_1Y \rightarrow$ Multiplicative model due to Eichorn and Hayre (1983),

$M_3^*:\ Z = U_1(Y+U_2) \rightarrow$ Mixed model 1 due to Saha (2007),

$M_4^*:\ Z = \varphi(Y+U_2) + (1-\varphi)U_1Y \rightarrow$ Mixed model 2 due to Diana and Perri (2011).

The MSE's of the estimators under scrambled randomized response models $M_j^*, j=1to4$ can be obtained by inserting $S_{r(j)}^2, j=1to4$ in place of S_r^2 in expression respective of MSE.

3. THE PROPOSED CLASS OF ESTIMATORS

We define an estimator for $\bar{Y}$ as

$$\hat{\bar{y}}_j^{**} = Q_j \hat{\bar{y}}_\eta^{*} \ , \ j = 1,2,3 \ ; \tag{3.1}$$

where $Q_1 = \left(\frac{n}{n + C_y^2} \right)$, $Q_2 = \left[1 + \frac{C_y^2}{n} \left\{ 1 + \frac{n_2}{n} (k-1)\delta \right\} \right]^{-1}$

and $Q_3 = \left[1 + \frac{C_y^2}{n} \left\{ 1 + \frac{n_2}{n} (k-1)\delta \right\} + \frac{k S_{\eta r}^2}{n N \bar{Y}^2} \right]^{-1}$ with $S_{\eta r}^2 = N_2 \left\{ \frac{\sigma_{v_1}^2 \mu_{2,y} + \eta^2 \sigma_{v_2}^2}{\mu_{v_1}^2} \right\}$.

We consider the situation, where non response occurs only on study variable *y* and the complete information on *x* is available in the second-phase sample of size *n* and also the population mean $\bar{X}$ of the non-sensitive auxiliary variable *x* is not known .

Reference is made here that, when proposing the estimator of the population mean $\bar{Y}$, Okafor and Lee (2000), Diana et al (2014) and Ahmed et al (2017) used only the information on the second phase sample mean $\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$ and on the first-phase sample mean $\bar{x}' = \sum_{i=1}^{n'} x_i \Big/ n'$. However, one can also obtain the unbiased estimator $\bar{x}^* = \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}_2$, of $\bar{X}$ (without any extra effort, while in the process of obtaining $\bar{y}^* = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_2$, the unbiased estimator of the population mean $\bar{Y}$, see Diana et al (2014), Ahmed et al (2017) and Singh and Kumar (2008,2009). Thus in this situation we have two unbiased estimators $\bar{x}^*$ and $\bar{x}$, of the population mean $\bar{X}$ of the non-sensitive auxiliary variable *x* at second phase.

Let $u = \bar{x}^* / \bar{x}'$, $v = \bar{x} / \bar{x}'$. Whatever be the sample chosen, let (*u*,*v*) adopt values in a bounded, closed convex subset, *S*, of the two dimensional real space enclosing the point '(1,1)'. Let $h(u,v)$ be a function of (*u*,*v*) such that $h(1,1) = 1$ and it satisfies the following conditions:

1. In *S*, the function $h(u,v)$ is continuous and bounded.
2. The first and second order partial derivatives of $h(u,v)$ exist and are continuous and bounded in *S*.

With this background and motivated by Srivastava (1971), we define the class of estimators of the population mean , $\bar{Y}$, of the study variable *y* as

$$\hat{\bar{y}}_{hd}^{(j)} = \hat{\bar{y}}_j^{**} h(u,v), \ j = 1,2,3. \tag{3.2}$$

Any parametric function $h(u,v)$ satisfying the above condition can be considered as an estimator of the population mean $\overline{Y}$. The class of such estimators is very vast. Some members of class of estimators $\hat{\bar{y}}_{hd}$ are:

$$\hat{\bar{y}}_{hd(1)}^{(j)} = \hat{\bar{y}}_j^{**} u^{\alpha_j} v^{\beta_j},$$

$$\hat{\bar{y}}_{hd(2)}^{(j)} = \hat{\bar{y}}_j^{**} \left[2 - u^{\alpha_j} v^{\beta_j}\right],$$

$$\hat{\bar{y}}_{hd(3)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j u + (1-\alpha) v \beta_j\right],$$

$$\hat{\bar{y}}_{hd(4)}^{(j)} = \hat{\bar{y}}_j^{**} \left[w_{1j} u^{\alpha_j} + w_{2j} v^{\beta_j}\right], \quad w_{1j} + w_{2j} = 1,$$

$$\hat{\bar{y}}_{hd(5)}^{(j)} = \hat{\bar{y}}_j^{**} \exp\left[\alpha_j (u-1) + \beta_j (v-1)\right],$$

$$\hat{\bar{y}}_{hd(6)}^{(j)} = \hat{\bar{y}}_j^{**} \frac{\left\{1 + \alpha_j (u-1)\right\}}{\left\{1 + \beta_j (v-1)\right\}},$$

$$\hat{\bar{y}}_{hd(7)}^{(j)} = \hat{\bar{y}}_j^{**} \left[1 + \alpha_j (u-1) + \beta_j (v-1)\right],$$

$$\hat{\bar{y}}_{hd(8)}^{(j)} = \frac{\hat{\bar{y}}_j^{**}}{\left[1 + \alpha_j \left(u^{\beta_j} v^{\gamma_j} - 1\right)\right]},$$

$$\hat{\bar{y}}_{hd(9)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j u + \left(1 - \alpha_j\right) u^2\right] v^{\beta_j},$$

$$\hat{\bar{y}}_{hd(10)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j v + \left(1 - \alpha_j\right) v^2\right] u^{\beta_j},$$

$$\hat{\bar{y}}_{hd(11)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j u + \left(1 - \alpha_j\right) u^2\right] \exp\left(\frac{\beta_j (v-1)}{(v+1)}\right),$$

$$\hat{\bar{y}}_{hd(12)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j v + \left(1 - \alpha_j\right) v^2\right] \exp\left(\frac{\beta_j (u-1)}{(u+1)}\right),$$

$$\hat{\bar{y}}_{hd(13)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j \exp\left(\frac{(u-1)}{(u+1)}\right) + \left(1 - \alpha_j\right) \exp\left(\frac{\beta_j (v-1)}{(v+1)}\right)\right],$$

$$\hat{\bar{y}}_{hd(14)}^{(j)} = \hat{\bar{y}}_j^{**} \left[\alpha_j \exp\left(\frac{(v-1)}{(v+1)}\right) + \left(1 - \alpha_j\right) \exp\left(\frac{\beta_j (u-1)}{(u+1)}\right)\right],$$

$$\hat{\bar{y}}_{hd(15)}^{(j)} = \hat{\bar{y}}_j^{**} u^{\alpha_j} \exp\left(\frac{\beta_j (v-1)}{(v+1)}\right),$$

$$\hat{\bar{y}}_{hd(16)}^{(j)} = \hat{\bar{y}}_j^{**} v^{\alpha_j} \exp\left(\frac{\beta_j (u-1)}{(u+1)}\right),$$

$$\hat{\bar{y}}_{hd(17)}^{(j)} = \hat{\bar{y}}_j^{**} \left[w_{1j} \exp\left(\frac{\alpha_j (u-1)}{(u+1)}\right) + w_{2j} \exp\left(\frac{\beta_j (v-1)}{(v+1)}\right)\right], \; w_{1j} + w_{2j} = 1,$$

$$\hat{\bar{y}}_{hd(18)}^{(j)} = \hat{\bar{y}}_j^{**} \left[1 + \alpha_j (u-1) + \beta_j (v-1)\right]^{-1}$$

etc . with $j=1,2,3$; where α_j and β_j , $j=1,2,3$; are suitably chosen constants to be determined such that the MSEs of the estimators are minimum.

The bias and MSE of the class of estimators, $\hat{\bar{y}}_{hd}^{(j)}$, exist since the number of possible samples is finite and we assume that the function is bounded. Expanding $h(u,v)$ about the point $(u,v)=(1,1)$ by a second order Taylor's series, we obtain

$$\hat{\bar{y}}_{hd}^{(j)}=\hat{\bar{y}}_j^{**}\left[h(1,1)+(u-1)h_1(1,1)+(v-1)h_2(1,1)+\frac{1}{2}\left\{(u-1)\frac{\partial}{\partial u}+(v-1)\frac{\partial}{\partial v}\right\}^2 h(u^*,v^*)\right], \quad (3.3)$$

where $u^*=1+\theta(u-1), v^*=1+\theta(v-1), 0<\theta<1,\ h_1(1,1)=\left.\frac{\partial h(u,v)}{\partial u}\right|_{(1,1)}$ and $h_2(1,1)=\left.\frac{\partial h(u,v)}{\partial v}\right|_{(1,1)}$.

We write $\hat{\bar{y}}^*=\bar{Y}(1+\hat{e}_0^*),\ \bar{x}=\bar{X}(1+e_1),\ \bar{x}'=\bar{X}(1+e_1'),\ \bar{x}^*=\bar{X}(1+e_1^*)$

such that

$E(\hat{e}_0^*)=E(e_1)=E(e_1')=E(e_1^*)=0$ and ignoring fpc terms

we have

$$E(\hat{e}_0^{*2})=\left(\frac{1}{n}C_y^2+\xi C_{y(2)}^2+\frac{k}{nN}C_{\eta r}^2\right),\ E(e_1'^2)=\frac{1}{n'}C_x^2,\ E(e_1^{*2})=\left(\frac{1}{n}C_x^2+\xi C_{x(2)}^2\right),\ E(e_1^2)=\frac{1}{n}C_x^2$$

$$E(\hat{e}_0^*e_1')=\frac{1}{n'}C_{yx}=\frac{1}{n'}\rho_{yx}C_yC_x,\qquad E(\hat{e}_0^*e_1^*)=\left(\frac{1}{n}C_{yx}+\xi C_{yx(2)}\right),\qquad E(e_1e_1')=\frac{1}{n'}C_x^2,$$

$$E(\hat{e}_0^*e_1)=\frac{1}{n}C_{yx}=\frac{1}{n}\rho_{yx}C_yC_x,\quad E(e_1^*e_1')=\frac{1}{n'}C_x^2,\ E(e_1e_1^*)=\frac{1}{n}C_x^2,$$

where $C_x^2=\frac{S_x^2}{\bar{X}^2},\ C_y^2=\frac{S_y^2}{\bar{Y}^2},\ C_{x(2)}^2=\frac{S_{x(2)}^2}{\bar{X}^2},\ C_{y(2)}^2=\frac{S_{y(2)}^2}{\bar{Y}^2},\ C_{\eta r}^2=\frac{S_{\eta r}^2}{\bar{Y}^2},\ C_{yx}=\frac{S_{yx}}{\bar{Y}\bar{X}},$

$$C_{yx(2)}=\frac{S_{yx(2)}}{\bar{Y}\bar{X}},\ C_{yx}=\rho_{yx}C_yC_x,\ C_{yx(2)}=\rho_{yx(2)}C_{y(2)}C_{x(2)},\ \rho_{yx}=\frac{S_{yx}}{S_yS_x},\ \rho_{yx(2)}=\frac{S_{yx(2)}}{S_{y(2)}S_{x(2)}},$$

$$\bar{X}_2=\frac{1}{N_2}\sum_{i=1}^{N_2}x_i\ ,S_x^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})^2\ ,S_{x(2)}^2=\frac{1}{N_2-1}\sum_{i=1}^{N}(x_i-\bar{X}_2)^2\ ,$$

$$S_{yx}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y}),\ S_{yx(2)}=\frac{1}{N_2-1}\sum_{i=1}^{N}(x_i-\bar{X}_2)(y_i-\bar{Y}_2),$$

$$\xi=\frac{W_2(k-1)}{n}\ \text{and}\ \lambda^*=\left(\frac{1}{n}-\frac{1}{n'}\right).$$

The MSE of $\hat{\bar{y}}_{hd}^{(j)}$ is given by

$$MSE(\hat{\bar{y}}_{hd}^{(j)})=E(\hat{\bar{y}}_{hd}^{(j)}-\bar{Y})^2=E(\hat{\bar{y}}_j^{**}h(u,v)-\bar{Y})^2=E(Q_j\hat{\bar{y}}^*h(u,v)-\bar{Y})^2$$

$$= E\Bigg[Q_j \bar{Y}\left(1+\hat{e}_0^*\right)\Big\{ h(1,1) + (u-1)h_1(1,1) + (v-1)h_2(1,1) + \frac{1}{2}(u-1)^2 h_{11}(1,1)$$

$$+ \frac{1}{2}(v-1)^2 h_{22}(1,1) + (u-1)(v-1)h_{12}(1,1)\Big\} - \bar{Y}\Bigg]^2$$

Noting that $h(1,1)=1$, we have

$$MSE\left(\hat{\bar{y}}_{hd}^{(j)}\right) = \bar{Y}^2\Big[Q_j\Big\{1+\hat{e}_0^* + (u-1)h_1(1,1) + \hat{e}_0^*(u-1)h_1(1,1) + (v-1)h_2(1,1) + \hat{e}_0^*(v-1)h_2(1,1)$$

$$+ \frac{1}{2}(u-1)^2 h_{11}(1,1) + \frac{1}{2}(v-1)^2 h_{22}(1,1) + (u-1)(v-1)h_{12}(1,1)\Big\} - 1\Big]^2 \quad ,$$

$$= \bar{Y}^2 E\Big[Q_j^2\Big\{1 + 2\hat{e}_0^* + 2\Big[\left(e_1^* - e_1^{'}\right) + \left(e_1^{'2} - e_1^* e_1^{'}\right)\Big]h_1(1,1) + 2\Big[\left(e_1 - e_1^{'}\right) + \left(e_1^{'2} - e_1 e_1^{'}\right)\Big]h_2(1,1)$$

$$+ \hat{e}_0^{*2} + \left(e_1^* - e_1^{'}\right)^2 h_1^2(1,1) + \left(e_1 - e_1^{'}\right)^2 h_2^2(1,1) + 4\left(\hat{e}_0^* e_1^* - \hat{e}_0^* e_1^{'}\right)h_1(1,1)$$

$$+ 4\left(\hat{e}_0^* e_1 - \hat{e}_0^* e_1^{'}\right)h_2(1,1) + 2\left(e_1^* e_1 - e_1^{'} e_1 - e_1^* e_1^{'} + e_1^{'2}\right)h_1(1,1)h_2(1,1)$$

$$+ 2\left(e_1^* e_1 - e_1^{'} e_1 - e_1^* e_1^{'} - e_1^{'2}\right)h_{12}(1,1) + \left(e_1^* - e_1^{'}\right)^2 h_{11}(1,1) + \left(e_1 - e_1^{'}\right)^2 h_{22}(1,1)\Big\}$$

$$- 2Q_j\Big\{1 + \hat{e}_0^* + \Big[\left(e_1^* - e_1^{'}\right) + \left(e_1^{'2} - e_1^* e_1^{'}\right)\Big]h_1(1,1) + \Big[\left(e_1 - e_1^{'}\right) + \left(e_1^{'2} - e_1 e_1^{'}\right)\Big]h_2(1,1)$$

$$+ \left(\hat{e}_0^* e_1^* - \hat{e}_0^* e_1^{'}\right)h_1(1,1) + \left(\hat{e}_0^* e_1 - \hat{e}_0^* e_1^{'}\right)h_2(1,1) + + \frac{1}{2}\left(e_1^* - e_1^{'}\right)^2 h_{11}(1,1)$$

$$+ \frac{1}{2}\left(e_1 - e_1^{'}\right)^2 h_{22}(1,1) + \left(e_1^* e_1 - e_1^{'} e_1 - e_1^* e_1^{'} - e_1^{'2}\right)h_{12}(1,1)\Big\} + 1\Big]. \qquad (3.4)$$

Taking expectation of both sides of (3.4) we get the $MSE\left(\hat{\bar{y}}_{hd}^{(j)}\right)$ as

$$MSE\left(\hat{\bar{y}}_{hd}^{(j)}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) + Q_j^*\Big[\left\{\lambda^* C_x^2 + \xi C_{x(2)}^2\right\}\left\{h_1^2(1,1) + h_{11}(1,1)\right\}$$

$$+ \lambda^* C_x^2\left\{h_2^2(1,1) + h_{22}(1,1)\right\} + 2\lambda^* C_x^2\left\{h_1(1,1)h_2(1,1) + h_{12}(1,1)\right\}$$

$$+ 4\left\{\lambda^* C_{yx} + \xi C_{yx(2)}\right\}h_1(1,1) + 4\lambda^* C_{yx} h_2(1,1)\Big] - 2Q_j^{**}\Big[\left\{\lambda^* C_{yx} + \xi C_{yx(2)}\right\}h_1(1,1) \qquad (3.5)$$

$$+ \lambda^* C_{yx} h_2(1,1) + \frac{1}{2}\left\{\lambda^* C_x^2 + \xi C_{x(2)}^2\right\}h_{11}(1,1) + \lambda^* C_x^2 h_{12}(1,1) + \frac{1}{2}\lambda^* C_x^2 h_{22}(1,1)\Big],$$

where

$$MSE\left(\hat{\bar{y}}_1^{**}\right) = MSE\left(\hat{t}_{s2}\right) = \left[\frac{S_y^2}{\left(n + C_y^2\right)} + \frac{n}{\left(n + C_y^2\right)^2}\left\{W_2(k-1)S_{y(2)}^2 + \frac{kS_{\eta r}^2}{N}\right\}\right], \qquad (3.6)$$

$$MSE\left(\hat{\bar{y}}_2^{**}\right) = MSE\left(\hat{t}_{s1}\right) = \left[\left(1 - A_\delta\right)\frac{S_y^2}{n} + \left(1 - 2B_\delta\right)\frac{W_2(k-1)}{n}S_{y(2)}^2 + \left(1 - 2B_\delta\right)\frac{kS_{\eta r}^2}{nN}\right], \qquad (3.7)$$

$$MSE\left(\hat{\bar{y}}_3^{**}\right) = MSE\left(\hat{t}_s\right) = \left[\left(1 - A_\eta^*\right)\frac{S_y^2}{n} + \left(1 - 2B_\eta^*\right)\frac{W_2(k-1)}{n}S_{y(2)}^2 + \frac{kS_{\eta r}^2}{nN}\left(1 - \frac{kC_{\eta r}^2}{nN}\right)\right], \qquad (3.8)$$

$$Q_1^* = \frac{n^2}{\left(n+C_y^2\right)^2} \cong \left(1-2\frac{C_y^2}{n}\right),\ Q_1^{**} = \frac{n}{\left(n+C_y^2\right)} \cong \left(1-\frac{C_y^2}{n}\right),$$

$Q_2^* = \left[1-2\beta_\delta\right],\ Q_2^{**} = \left[1-\beta_\delta\right],\ Q_3^* = \left[1-2\beta_\eta^*\right]$, and $Q_3^{**} = \left[1-\beta_\eta^*\right]$.

PARTICULAR CASE – To illustrate our general outcomes we deliberate the following class of estimators for population mean $\bar{Y}$ as

$$\hat{\bar{y}}_{hd(1)}^{(j)} = \hat{\bar{y}}_j^{**} u^{\alpha_j} v^{\beta_j} \tag{3.9}$$

where α_j *and* β_j are suitably chosen constants and $j=1,2,3$.

It is observed from (3.5) that $MSE\left(\hat{\bar{y}}_{hd}^{(j)}\right)$ depends on the derivatives $h_1(1,1)$, $h_2(1,1)$, $h_{11}(1,1), h_{12}(1,1)$ and $h_{22}(1,1)$. So to obtain the MSE of the estimator $\hat{\bar{y}}_{hd(1)}^{(j)}$, $j=1,2,3$, we have the values of $h_1(1,1), h_2(1,1), h_{11}(1,1), h_{12}(1,1)$ and $h_{22}(1,1)$ for the function $h(u,v) = u^{\alpha_j}\, v^{\beta_j}$ as obtained below:

$$h_1(1,1) = \left.\frac{\partial h(u,v)}{\partial u}\right|_{(1,1)} = \alpha_j, \qquad h_2(1,1) = \left.\frac{\partial h(u,v)}{\partial v}\right|_{(1,1)} = \beta_j,$$

$$h_{11}(1,1) = \left.\frac{\partial^2 h(v,w)}{\partial u^2}\right|_{(1,1)} = \alpha_j\left(\alpha_j - 1\right), \qquad h_{12}(1,1) = \left.\frac{\partial^2 h(v,w)}{\partial u \partial v}\right|_{(1,1)} = \alpha_j\beta_j,$$

$$h_{22}(1,1) = \left.\frac{\partial^2 h(v,w)}{\partial v^2}\right|_{(1,1)} = \beta_j\left(\beta_j - 1\right) \qquad \text{and} \qquad h(1,1) = 1,$$

Putting the values of $h_1(1,1)$, $h_2(1,1)$, $h_{11}(1,1), h_{12}(1,1)$ and $h_{22}(1,1)$ in (3.5) we get the MSE of $\hat{\bar{y}}_{hd(1)}^{(j)}$ as

$$MSE\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) + \bar{Y}^2\left[\alpha_j^2 A_{1(j)} + \beta_j^2 A_{2(j)} + 2\alpha_j\beta_j A_{3(j)} - 2\alpha_j A_{4(j)} - 2\beta_j A_{5(j)}\right], \tag{3.10}$$

where $A_{1(j)} = Q_{(j)}^* A_x$, $A_{2(j)} = \lambda^* Q_{(j)}^* C_x^2$, $A_{3(j)} = \lambda^* Q_{(j)}^* C_x^2$, $A_{4(j)} = \frac{1}{2}\left(Q_{d(j)}^* A_x - 2Q_{(j)}^* A_{yx}\right)$

$A_{5(j)} = \frac{1}{2}\lambda^*\left(Q_{d(j)}^* C_x^2 - 2Q_{(j)}^* C_{yx}\right)$, $Q_{(j)}^* = \left(2Q_j^* - Q_j^{**}\right)$, $Q_{d(j)}^* = \left(Q_j^* - Q_j^{**}\right)$,

$A_x = \left\{\lambda^* C_x^2 + \xi C_{x(2)}^2\right\}, A_{yx} = \left\{\lambda^* C_{yx} + \xi C_{yx(2)}\right\}$.

The $MSE\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)$ at (3.10) is minimum when

$$\left.\begin{aligned} \alpha_j &= \frac{\left(A_{2(j)}A_{4(j)} - A_{3(j)}A_{5(j)}\right)}{\left(A_{1(j)}A_{2(j)} - A_{3(j)}^2\right)} = \alpha_{j0}(say) \\ \beta_j &= \frac{\left(A_{1(j)}A_{5(j)} - A_{3(j)}A_{4(j)}\right)}{\left(A_{1(j)}A_{2(j)} - A_{3(j)}^2\right)} = \beta_{j0}(say) \end{aligned}\right\};\ j = 1,2,3 \tag{3.11}$$

Thus the resulting minimum MSE of $\hat{\bar{y}}_{hd(1)}^{(j)}$ is given by

$$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)=MSE\left(\hat{\bar{y}}_{j}^{**}\right)-\bar{Y}^{2}\frac{\left(A_{2(j)}A_{4(j)}^{2}-2A_{3(j)}A_{4(j)}A_{5(j)}+A_{1(j)}A_{5(j)}^{2}\right)}{\left(A_{1(j)}A_{2(j)}-A_{3(j)}^{2}\right)};\ j=1,2,3. \quad (3.12)$$

Now we state the following theorem.

THEOREM 3.1- The MSE of $\hat{\bar{y}}_{hd(1)}^{(j)}$ is greater than or equal to minimum MSE of $\hat{\bar{y}}_{hd(1)}^{(j)}$ i.e.

$$MSE\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)\geq MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)$$

i.e. $$MSE\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)\geq\left[MSE\left(\hat{\bar{y}}_{j}^{**}\right)-\bar{Y}^{2}\frac{\left(A_{2(j)}A_{4(j)}^{2}-2A_{3(j)}A_{4(j)}A_{5(j)}+A_{1(j)}A_{5(j)}^{2}\right)}{\left(A_{1(j)}A_{2(j)}-A_{3(j)}^{2}\right)}\right];\ j=1,2,3$$

with equality holding if

$\alpha_j=\alpha_{j0},\ \beta_j=\beta_{j0}$, $j=1,2,3$;

where α_{j0} and β_{j0} are given by (3.11).

Further putting $j=1,2,3$ in (3.10) we get the MSE's of the estimators $\hat{\bar{y}}_{hd(1)}^{(1)}=\hat{\bar{y}}_{1}^{**}u^{\alpha_1}v^{\beta_1}$

$\hat{\bar{y}}_{hd(1)}^{(2)}=\hat{\bar{y}}_{2}^{**}u^{\alpha_2}v^{\beta_2}$ and $\hat{\bar{y}}_{hd(1)}^{(3)}=\hat{\bar{y}}_{3}^{**}u^{\alpha_3}v^{\beta_3}$

with $\hat{\bar{y}}_{1}^{**}=Q_1\hat{\bar{y}}^{*}$, $\hat{\bar{y}}_{2}^{**}=Q_2\hat{\bar{y}}^{*}$, $\hat{\bar{y}}_{3}^{**}=Q_3\hat{\bar{y}}^{*}$ respectively as

$$MSE\left(\hat{\bar{y}}_{hd(1)}^{(1)}\right)=MSE\left(\hat{\bar{y}}_{1}^{**}\right)+\bar{Y}^{2}\left[\alpha_1^{2}A_{1(1)}+\beta_1^{2}A_{2(1)}+2\alpha_1\beta_1A_{3(1)}-2\alpha_1A_{4(1)}-2\beta_1A_{5(1)}\right], \quad (3.13)$$

$$MSE\left(\hat{\bar{y}}_{hd(1)}^{(2)}\right)=MSE\left(\hat{\bar{y}}_{2}^{**}\right)+\bar{Y}^{2}\left[\alpha_2^{2}A_{1(2)}+\beta_2^{2}A_{2(2)}+2\alpha_2\beta_2A_{3(2)}-2\alpha_2A_{4(2)}-2\beta_2A_{5(2)}\right], \quad (3.14)$$

$$MSE\left(\hat{\bar{y}}_{hd(1)}^{(3)}\right)=MSE\left(\hat{\bar{y}}_{3}^{**}\right)+\bar{Y}^{2}\left[\alpha_3^{2}A_{1(3)}+\beta_3^{2}A_{2(3)}+2\alpha_3\beta_3A_{3(3)}-2\alpha_3A_{4(3)}-2\beta_3A_{5(3)}\right], \quad (3.15)$$

where $A_{1(1)}=Q_{(1)}^{*}A_x$, $A_{2(1)}=\lambda^{*}Q_{(1)}^{*}C_x^{2}$, $A_{3(1)}=\lambda^{*}Q_{(1)}^{*}C_x^{2}$, $A_{4(1)}=\frac{1}{2}\left(Q_{d(1)}^{*}A_x-2Q_{(1)}^{*}A_{yx}\right)$,

$A_{5(1)}=\frac{1}{2}\lambda^{*}\left(Q_{d(1)}^{*}C_x^{2}-2Q_{(1)}^{*}C_{yx}\right)$, $Q_{(1)}^{*}=\left(2Q_1^{*}-Q_1^{**}\right)$, $Q_{d(1)}^{*}=\left(Q_1^{*}-Q_1^{**}\right)$,

$A_{1(2)}=Q_{(2)}^{*}A_x$, $A_{2(2)}=\lambda^{*}Q_{(2)}^{*}C_x^{2}$, $A_{3(2)}=\lambda^{*}Q_{(2)}^{*}C_x^{2}$, $A_{4(2)}=\frac{1}{2}\left(Q_{d(2)}^{*}A_x-2Q_{(2)}^{*}A_{yx}\right)$,

$A_{5(2)}=\frac{1}{2}\lambda^{*}\left(Q_{d(2)}^{*}C_x^{2}-2Q_{(2)}^{*}C_{yx}\right)$, $Q_{(2)}^{*}=\left(2Q_2^{*}-Q_2^{**}\right)$, $Q_{d(2)}^{*}=\left(Q_2^{*}-Q_2^{**}\right)$,

$A_{1(3)}=Q_{(3)}^{*}A_x$, $A_{2(3)}=\lambda^{*}Q_{(3)}^{*}C_x^{2}$, $A_{3(3)}=\lambda^{*}Q_{(3)}^{*}C_x^{2}$, $A_{4(3)}=\frac{1}{2}\left(Q_{d(3)}^{*}A_x-2Q_{(3)}^{*}A_{yx}\right)$,

$A_{5(3)}=\frac{1}{2}\lambda^{*}\left(Q_{d(3)}^{*}C_x^{2}-2Q_{(3)}^{*}C_{yx}\right)$, $Q_{(3)}^{*}=\left(2Q_3^{*}-Q_3^{**}\right)$, $Q_{d(3)}^{*}=\left(Q_3^{*}-Q_3^{**}\right)$,

$Q_{(1)}^{*}=\frac{n\left(n-C_y^{2}\right)}{\left(n+C_y^{2}\right)^{2}}\cong\left(1-3\frac{C_y^{2}}{n}\right)$, $Q_{d(1)}^{*}=\frac{nC_y^{2}}{\left(n+C_y^{2}\right)^{2}}\cong-\frac{C_y^{2}}{n}$,

$Q_{(2)}^{*}=\left[1-3\beta_{\delta}\right]$, $Q_{d(2)}^{*}=-\beta_{\delta}$, $Q_{(3)}^{*}=\left[1-3\beta_{\eta}^{*}\right]$, and $Q_{d(3)}^{*}=-\beta_{\eta}^{*}$.

Minimization of (3.13), (3.14) and (3.15) yields the optimum values of $(\alpha_1,\beta_1),(\alpha_2,\beta_2)$ and (α_3,β_3)

$$\left.\alpha_{10}=\frac{\left(A_{2(1)}A_{4(1)}-A_{3(1)}A_{5(1)}\right)}{\left(A_{1(1)}A_{2(1)}-A_{3(1)}^2\right)}\ and\ \beta_{10}=\frac{\left(A_{1(1)}A_{5(1)}-A_{3(1)}A_{4(1)}\right)}{\left(A_{1(1)}A_{2(1)}-A_{3(1)}^2\right)}\right\}, \tag{3.16}$$

$$\left.\alpha_{20}=\frac{\left(A_{2(2)}A_{4(2)}-A_{3(2)}A_{5(2)}\right)}{\left(A_{1(2)}A_{2(2)}-A_{3(2)}^2\right)}\ and\ \beta_{20}=\frac{\left(A_{1(2)}A_{5(2)}-A_{3(2)}A_{4(2)}\right)}{\left(A_{1(2)}A_{2(2)}-A_{3(2)}^2\right)}\right\}, \tag{3.17}$$

$$\left.\alpha_{30}=\frac{\left(A_{2(3)}A_{4(3)}-A_{3(3)}A_{5(3)}\right)}{\left(A_{1(3)}A_{2(3)}-A_{3(3)}^2\right)}\ and\ \beta_{30}=\frac{\left(A_{1(3)}A_{5(3)}-A_{3(3)}A_{4(3)}\right)}{\left(A_{1(3)}A_{2(3)}-A_{3(3)}^2\right)}\right\}\ . \tag{3.18}$$

Thus the resulting minimum MSE's of $\hat{\bar{y}}_{hd(1)}^{(1)}$, $\hat{\bar{y}}_{hd(1)}^{(2)}$ and $\hat{\bar{y}}_{hd(1)}^{(3)}$ are respectively given by

$$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(1)}\right)=MSE\left(\hat{\bar{y}}_1^{**}\right)-\bar{Y}^2\frac{\left(A_{2(1)}A_{4(1)}^2-2A_{3(1)}A_{4(1)}A_{5(1)}+A_{1(1)}A_{5(1)}^2\right)}{\left(A_{1(1)}A_{2(1)}-A_{3(1)}^2\right)}, \tag{3.19}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(2)}\right)=MSE\left(\hat{\bar{y}}_2^{**}\right)-\bar{Y}^2\frac{\left(A_{2(2)}A_{4(2)}^2-2A_{3(2)}A_{4(2)}A_{5(2)}+A_{1(2)}A_{5(2)}^2\right)}{\left(A_{1(2)}A_{2(2)}-A_{3(2)}^2\right)}, \tag{3.20}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(3)}\right)=MSE\left(\hat{\bar{y}}_3^{**}\right)-\bar{Y}^2\frac{\left(A_{2(3)}A_{4(3)}^2-2A_{3(3)}A_{4(3)}A_{5(3)}+A_{1(3)}A_{5(3)}^2\right)}{\left(A_{1(3)}A_{2(3)}-A_{3(3)}^2\right)}. \tag{3.21}$$

Expressions (3.19), (3.20) and (3.21) clearly show that the proposed estimators $\hat{\bar{y}}_{hd(1)}^{(1)}$, $\hat{\bar{y}}_{hd(1)}^{(2)}$ and $\hat{\bar{y}}_{hd(1)}^{(3)}$ are more efficient than the estimators $\hat{\bar{y}}_1^{**}$, $\hat{\bar{y}}_2^{**}$ and $\hat{\bar{y}}_3^{**}$ respectively.

3.1 SPECIAL CASE I- Putting $\beta_j=0$ in (3.9) we get a class of estimators of the population mean $\bar{Y}$ as

$$\hat{\bar{y}}_{\alpha j}=\hat{\bar{y}}_j^{**}u^{\alpha_j};\ j=1,2,3. \tag{3.22}$$

Putting $\beta_j=0$ in (3.10) we get the MSE of $\hat{\bar{y}}_{\alpha j}$ as

$$MSE\left(\hat{\bar{y}}_{\alpha j}\right)=MSE\left(\hat{\bar{y}}_j^{**}\right)+\bar{Y}^2\left[\alpha_j^2A_{1(j)}-2\alpha_jA_{4(j)}\right], \tag{3.23}$$

which is minimum when

$$\alpha_j=\frac{A_{4(j)}}{A_{1(j)}}=\alpha_{j0}^{*}(say)\ . \tag{3.24}$$

Thus the resulting minimum MSE of $\hat{\bar{y}}_{\alpha j}$ is given by

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha j}\right)=MSE\left(\hat{\bar{y}}_j^{**}\right)-\bar{Y}^2\frac{A_{4(j)}^2}{A_{1(j)}}\ . \tag{3.25}$$

Now, we arrived at the following theorem.

THEOREM 3.2- To the first degree of approximation,

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha j}\right) \geq MSE\left(\hat{\bar{y}}_{j}^{**}\right) - \bar{Y}^{2}\frac{A_{4(j)}^{2}}{A_{1(j)}}, \tag{3.26}$$

with equality holding if

$$\alpha_j = \frac{A_{4(j)}}{A_{1(j)}},\ j = 1,2,3.$$

Putting $j = 1,2,3$ in (3.23) we get the MSE's of the estimators $\hat{\bar{y}}_{\alpha 1} = \hat{\bar{y}}_{1}^{**}u^{\alpha_1}$, $\hat{\bar{y}}_{\alpha 2} = \hat{\bar{y}}_{2}^{**}u^{\alpha_2}$ and $\hat{\bar{y}}_{\alpha 3} = \hat{\bar{y}}_{3}^{**}u^{\alpha_3}$ (due to Ahmed et al (2017) for $\eta = 1 and\, \delta = 1$) respectively as

$$MSE\left(\hat{\bar{y}}_{\alpha 1}\right) = MSE\left(\hat{\bar{y}}_{1}^{**}\right) + \bar{Y}^{2}\left[\alpha_1^{2}A_{1(1)} - 2\alpha_1 A_{4(1)}\right], \tag{3.27}$$

$$MSE\left(\hat{\bar{y}}_{\alpha 2}\right) = MSE\left(\hat{\bar{y}}_{2}^{**}\right) + \bar{Y}^{2}\left[\alpha_2^{2}A_{1(2)} - 2\alpha_2 A_{4(2)}\right], \tag{3.28}$$

$$MSE\left(\hat{\bar{y}}_{\alpha 3}\right) = MSE\left(\hat{\bar{y}}_{3}^{**}\right) + \bar{Y}^{2}\left[\alpha_3^{2}A_{1(3)} - 2\alpha_3 A_{4(3)}\right]\ . \tag{3.29}$$

which are respectively minimized for

$$\alpha_1 = A_{4(1)}/A_{1(1)} = \alpha_{10}^{*}, \tag{3.30}$$

$$\alpha_2 = A_{4(2)}/A_{1(2)} = \alpha_{20}^{*}, \tag{3.31}$$

$$\alpha_3 = A_{4(3)}/A_{1(3)} = \alpha_{30}^{*}. \tag{3.32}$$

Thus the resulting minimum MSE of $\hat{\bar{y}}_{\alpha 1}$, $\hat{\bar{y}}_{\alpha 2}$ and $\hat{\bar{y}}_{\alpha 3}$ are respectively given by

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 1}\right) = MSE\left(\hat{\bar{y}}_{1}^{**}\right) - \bar{Y}^{2}\frac{A_{4(1)}^{2}}{A_{1(1)}}, \tag{3.33}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 2}\right) = MSE\left(\hat{\bar{y}}_{2}^{**}\right) - \bar{Y}^{2}\frac{A_{4(2)}^{2}}{A_{1(2)}}, \tag{3.34}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 3}\right) = MSE\left(\hat{\bar{y}}_{3}^{**}\right) - \bar{Y}^{2}\frac{A_{4(3)}^{2}}{A_{1(3)}}. \tag{3.35}$$

From (3.12) and (3.25) we have

$$MSE_{\min}\left(\hat{\bar{y}}_{\alpha j}\right) - MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right) = \bar{Y}^{2}\frac{\left(A_{1(j)}A_{5(j)} - A_{3(j)}A_{4(j)}\right)^{2}}{A_{1(j)}\left(A_{1(j)}A_{2(j)} - A_{3(j)}^{2}\right)};\ j = 1,2,3 \tag{3.36}$$

which is non-negative.

Thus the proposed class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$ is more efficient than $\hat{\bar{y}}_{\alpha j}$, $j = 1,2,3$.

If we set $\alpha_j = -1 and 1$ in (3.21) we get the ratio and product type estimators for $\bar{Y}$ respectively as

$$\hat{\bar{y}}_{\alpha j(-1)} = \hat{\bar{y}}_{j}^{**}\left(\frac{\bar{X}'}{\bar{X}^{*}}\right), \tag{3.37}$$

$$\hat{\bar{y}}_{\alpha j(1)} = \hat{\bar{y}}_j^{**}\left(\frac{\bar{x}^*}{\bar{x}'}\right). \tag{3.38}$$

Putting $\alpha_j = -1 and 1$ in (3.23) we get the MSE's of $\hat{\bar{y}}_{\alpha j(-1)}$ and $\hat{\bar{y}}_{\alpha j(1)}$ respectively as

$$MSE\left(\hat{\bar{y}}_{\alpha j(-1)}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) + \bar{Y}^2\left[A_{1(j)} + 2A_{4(j)}\right], \tag{3.39}$$

$$MSE\left(\hat{\bar{y}}_{\alpha j(1)}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) + \bar{Y}^2\left[A_{1(j)} - 2A_{4(j)}\right]. \tag{3.40}$$

Remark 3.1

For $j = 3, (\delta, \eta) = (1,1)$, the class of estimators $\hat{\bar{y}}_{\alpha 3} = \hat{\bar{y}}_3^{**} u^{\alpha_3}$ reduces to the estimator

$$\hat{\bar{y}}_{\alpha 3(A)} = \hat{\bar{y}}_{3(A)}^{**} u^{\alpha_3} \tag{3.41}$$

which is due to Ahmed et al (2017), where

$$\hat{\bar{y}}_{3(A)}^{**} = \left[1 + \frac{C_y^2}{n}\left\{1 + \frac{n_2}{n}(k-1)\right\} + \frac{k}{nN}\frac{S_r^2}{\bar{Y}^2}\right]^{-1} \hat{\bar{y}}^*.$$

For $\alpha_3 = -1 and 1$, the class of estimators $\hat{\bar{y}}_{\alpha 3(A)}$ reduces to the estimators respectively as

$$\hat{\bar{y}}_{\alpha 3(-1)} = \hat{\bar{y}}_{3(A)}^{**}\left(\frac{\bar{x}'}{\bar{x}^*}\right), \tag{3.42}$$

$$\hat{\bar{y}}_{\alpha 3(1)} = \hat{\bar{y}}_{3(A)}^{**}\left(\frac{\bar{x}^*}{\bar{x}'}\right). \tag{3.43}$$

The MSEs of $\hat{\bar{y}}_{\alpha 3(A)}$, $\hat{\bar{y}}_{\alpha 3(-1)}$ and $\hat{\bar{y}}_{\alpha 3(1)}$ are respectively given by

$$\begin{aligned} MSE\left(\hat{\bar{y}}_{\alpha 3(A)}\right) &= MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right) + \bar{Y}^2\left[\alpha_3^2 A_{1(3)} - 2\alpha_3 A_{4(3)}\right] \\ &= MSE\left(\hat{\bar{y}}^{**}\right) + \bar{Y}^2\left[\alpha_3\left\{\alpha_3 - B^*(\alpha_3 - 1)\right\}A_x + 2\alpha_3(1 - 3B^*)A_{yx}\right], \end{aligned} \tag{3.44}$$

$$\begin{aligned} MSE\left(\hat{\bar{y}}_{\alpha 3(-1)}\right) &= MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right) + \bar{Y}^2\left[A_{1(3)} + 2A_{4(3)}\right], \\ &= MSE\left(\hat{\bar{y}}^{**}\right) + \bar{Y}^2\left[\lambda^*\left\{(1 - 4B^*)C_x^2 - 2(1 - 3B^*)C_{yx}\right\}\right. \\ &\qquad \left. + \xi\left\{(1 - 4B^*)C_{x(2)}^2 - 2(1 - 3B^*)C_{yx(2)}\right\}\right], \end{aligned} \tag{3.45}$$

$$\begin{aligned} MSE\left(\hat{\bar{y}}_{\alpha 3(1)}\right) &= MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right) + \bar{Y}^2\left[A_{1(3)} - 2A_{4(3)}\right], \\ &= MSE\left(\hat{\bar{y}}^{**}\right) + \bar{Y}^2\left[(1 - 2B^*)\left\{\lambda^* C_x^2 + \xi C_{x(2)}^2\right\} + 2(1 - 3B^*)\left\{\lambda^* C_{yx} + \xi C_{yx(2)}\right\}\right], \end{aligned} \tag{3.46}$$

where

$$MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right) = MSE\left(\hat{\bar{y}}^{**}\right) = \left[(1 - A^*)\frac{S_y^2}{n} + (1 - 2B^*)\frac{W_2(k-1)}{n}S_{y(2)}^2 + \frac{kS_r^2}{nN}\left(1 - \frac{k}{nN}\frac{S_r^2}{\bar{Y}^2}\right)\right]. \tag{3.47}$$

3.2 SPECIAL CASE –II- Putting $\alpha_j = 0$ in (3.9) we get a class of estimators of the population mean $\bar{Y}$ as

$$\hat{\bar{y}}_{\beta j} = \hat{\bar{y}}_j^{**} v^{\beta_j};\ j = 1,2,3 \tag{3.48}$$

Putting $\alpha_j = 0$ in (3.10) we get the MSE of $\hat{\bar{y}}_{\beta j}$ as

$$MSE\left(\hat{\bar{y}}_{\beta j}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) + \bar{Y}^2\left[\beta_j^2 A_{2(j)} - 2\beta_j A_{5(j)}\right], \tag{3.49}$$

which is minimum when

$$\beta_j = \frac{A_{5(j)}}{A_{2(j)}} = \beta_{j0}^{*}(say). \tag{3.50}$$

Thus the resulting minimum MSE of $\hat{\bar{y}}_{\beta j}$ is given by

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta j}\right) = MSE\left(\hat{\bar{y}}_j^{**}\right) - \bar{Y}^2 \frac{A_{5(j)}^2}{A_{2(j)}}. \tag{3.51}$$

Now we state the following theorem.

THEOREM 3.3- To the first degree of approximation,

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta j}\right) \geq MSE\left(\hat{\bar{y}}_j^{**}\right) - \bar{Y}^2 \frac{A_{5(j)}^2}{A_{2(j)}}, \tag{3.52}$$

with equality holding if $\beta_j = \frac{A_{5(j)}}{A_{2(j)}}$, $j = 1,2,3.$

Putting $j = 1,2,3$ in (3.49) we get the MSE's of the estimators $\hat{\bar{y}}_{\beta 1} = \hat{\bar{y}}_1^{**} v^{\beta_1}$, $\hat{\bar{y}}_{\beta 2} = \hat{\bar{y}}_2^{**} v^{\beta_2}$ and $\hat{\bar{y}}_{\beta 3} = \hat{\bar{y}}_3^{**} v^{\beta_3}$ (due to Ahmed et al (2017) for $\eta = 1 and \delta = 1$) respectively as

$$MSE\left(\hat{\bar{y}}_{\beta 1}\right) = MSE\left(\hat{\bar{y}}_1^{**}\right) + \bar{Y}^2\left[\beta_1^2 A_{2(1)} - 2\beta_1 A_{5(1)}\right], \tag{3.53}$$

$$MSE\left(\hat{\bar{y}}_{\beta 2}\right) = MSE\left(\hat{\bar{y}}_1^{**}\right) + \bar{Y}^2\left[\beta_2^2 A_{2(2)} - 2\beta_2 A_{5(2)}\right], \tag{3.54}$$

$$MSE\left(\hat{\bar{y}}_{\beta 3}\right) = MSE\left(\hat{\bar{y}}_1^{**}\right) + \bar{Y}^2\left[\beta_3^2 A_{2(3)} - 2\beta_3 A_{5(3)}\right]. \tag{3.55}$$

which are respectively minimized for

$$\beta_1 = A_{5(1)} / A_{2(1)} = \beta_{10}^{*}, \tag{3.56}$$

$$\beta_2 = A_{5(2)} / A_{2(2)} = \beta_{20}^{*}, \tag{3.57}$$

$$\beta_3 = A_{5(3)} / A_{2(3)} = \beta_{30}^{*}. \tag{3.58}$$

Thus the resulting minimum MSE of $\hat{\bar{y}}_{\beta 1}$, $\hat{\bar{y}}_{\beta 2}$ and $\hat{\bar{y}}_{\beta 3}$ are respectively given by

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta 1}\right) = MSE\left(\hat{\bar{y}}_1^{**}\right) - \bar{Y}^2 \frac{A_{5(1)}^2}{A_{2(1)}}, \tag{3.59}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta 2}\right) = MSE\left(\hat{\bar{y}}_2^{**}\right) - \bar{Y}^2 \frac{A_{5(2)}^2}{A_{2(2)}}, \tag{3.60}$$

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta 3}\right) = MSE\left(\hat{\bar{y}}_3^{**}\right) - \bar{Y}^2 \frac{A_{5(3)}^2}{A_{2(3)}}. \tag{3.61}$$

From (3.12) and (3.44) we have

$$MSE_{\min}\left(\hat{\bar{y}}_{\beta j}\right)-MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(j)}\right)=\bar{Y}^2\frac{\left(A_{2(j)}A_{4(j)}-A_{3(j)}A_{5(j)}\right)^2}{A_{2(j)}\left(A_{1(j)}A_{2(j)}-A_{3(j)}^2\right)};\ j=1,2,3; \tag{3.62}$$

which is non-negative.

Thus the proposed class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$ is more efficient than $\hat{\bar{y}}_{\beta j}$, $j=1,2,3$.

If we set $\beta_j=-1 and 1$ in (3.48) we get the ratio and product type estimators for $\bar{Y}$ respectively as

$$\hat{\bar{y}}_{\beta j(-1)}=\hat{\bar{y}}_j^{**}\left(\frac{\bar{x}'}{\bar{x}}\right), \tag{3.63}$$

$$\hat{\bar{y}}_{\beta j(1)}=\hat{\bar{y}}_j^{**}\left(\frac{\bar{x}}{\bar{x}'}\right). \tag{3.64}$$

Putting $\beta_j=-1 and 1$ in (3.49) we get the MSE's of $\hat{\bar{y}}_{\beta j(-1)}$ and $\hat{\bar{y}}_{\beta j(1)}$ respectively as

$$MSE\left(\hat{\bar{y}}_{\beta j(-1)}\right)=MSE\left(\hat{\bar{y}}_j^{**}\right)+\bar{Y}^2\left[A_{2(j)}+2A_{5(j)}\right], \tag{3.65}$$

$$MSE\left(\hat{\bar{y}}_{\beta j(1)}\right)=MSE\left(\hat{\bar{y}}_j^{**}\right)+\bar{Y}^2\left[A_{2(j)}-2A_{5(j)}\right], \tag{3.66}$$

Remark 3.2

For $j=3,(\delta,\eta)=(1,1)$, the class of estimators $\hat{\bar{y}}_{\beta 3}=\hat{\bar{y}}_3^{**}v^{\beta_3}$ reduces to the estimator

$$\hat{\bar{y}}_{\beta 3(A)}=\hat{\bar{y}}_{3(A)}^{**}v^{\beta_3} \tag{3.67}$$

which is due to Ahmed et al (2017).

For $\beta_3=-1 and 1$, the class of estimators $\hat{\bar{y}}_{\beta 3(A)}$ reduces to the estimators respectively as

$$\hat{\bar{y}}_{\beta 3(-1)}=\hat{\bar{y}}_{3(A)}^{**}\left(\frac{\bar{x}'}{\bar{x}}\right), \tag{3.68}$$

$$\hat{\bar{y}}_{\beta 3(1)}=\hat{\bar{y}}_{3(A)}^{**}\left(\frac{\bar{x}}{\bar{x}'}\right). \tag{3.69}$$

The MSEs of $\hat{\bar{y}}_{\beta 3(A)}$, $\hat{\bar{y}}_{\beta 3(-1)}$ and $\hat{\bar{y}}_{\beta 3(1)}$ are respectively given by

$$\begin{aligned}MSE\left(\hat{\bar{y}}_{\beta 3(A)}\right)&=MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right)+\bar{Y}^2\left[\beta_3^2A_{2(3)}-2\beta_3A_{5(3)}\right],\\&=MSE\left(\hat{\bar{y}}^{**}\right)+\bar{Y}^2\beta_3\lambda^*\left[\left\{\beta_3-B^*\left(3\beta_3-1\right)\right\}C_x^2+2\left(1-3B^*\right)C_{yx}\right],\end{aligned} \tag{3.70}$$

$$\begin{aligned}MSE\left(\hat{\bar{y}}_{\beta 3(-1)}\right)&=MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right)+\bar{Y}^2\left[A_{2(3)}+2A_{5(3)}\right],\\&=MSE\left(\hat{\bar{y}}^{**}\right)+\bar{Y}^2\lambda^*\left[\left(1-4B^*\right)C_x^2-2\left(1-3B^*\right)C_{yx}\right],\end{aligned} \tag{3.71}$$

$$\begin{aligned}MSE\left(\hat{\bar{y}}_{\beta 3(1)}\right)&=MSE\left(\hat{\bar{y}}_{3(A)}^{**}\right)+\bar{Y}^2\left[A_{2(3)}-2A_{5(3)}\right],\\&=MSE\left(\hat{\bar{y}}^{**}\right)+\bar{Y}^2\lambda^*\left[\left(1-2B^*\right)C_x^2+2\left(1-3B^*\right)C_{yx}\right].\end{aligned} \tag{3.72}$$

We note that the MSEs of the estimators $\hat{\bar{y}}_{\alpha 3(A)}$, $\hat{\bar{y}}_{\beta 3(A)}$, $\hat{\bar{y}}_{\alpha 3(-1)}$, $\hat{\bar{y}}_{\beta 3(-1)}$ and $\hat{\bar{y}}_{\alpha 3(1)}$, $\hat{\bar{y}}_{\beta 3(1)}$ given in ((3.44), (3.70)), ((3.45),(3.71)) and ((3.46), (3.72)) respectively are correct

while the MSE's of these estimators obtained by Ahmed et al (2017, equations (19),(21),(23), (24), (25), (26), pp. 8443-8444) are not correct. Resulting the MSE's of the estimators $t_1, t_2,\ T_1^{'}, T_2^{'}$ and $T_3^{'}, T_4^{'}$ reported by Khare and Kumar (2011, equations (2.10), (3.9)and (3.10)) are also not correct.

4. EMPIRICAL STUDY

Data [Source: Khare and Kumar (2011) and Ahmed et al (2017)]

y : Number of cultivators, and x: Population of villages.

The proportion of non-respondents in the population is 25%, so we consider last 24 units of population as non-respondents. It is also assumed that $V_1 and V_2$ are scrambled variables each distributed uniformly in the interval [0, 1]. The summary statistics are:

$N = 96,\ n = 25,\ \overline{Y} = 185.22,\ \overline{X} = 1807.23,\ \overline{Y}_2 = 128.46,.\overline{X}_2 = 1571.71,\ S_y = 195.03,$

$S_x = 1921.77,\ S_{y(2)} = 97.82, S_{x(2)} = 1068.44,\ S_{yx} = 338835.88, S_{yx(2)} = 93560.01,$

$\rho = 0.904,\ \rho_2 = 0.895,\ \mu_{v1} = 0.50,\ \mu_{v2} = 0.50,\ \sigma_{v1}^2 = 0.0833,\ \sigma_{v2}^2 = 0.0833, f_h = 2.$

We have computed the MSEs (MSE's) of the suggested classes of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$, $\hat{\bar{y}}_{\alpha j}$ and $\hat{\bar{y}}_{\beta j}$, $j = 1,2,3$ and also find the MSE's of t_s, t_{s1} and t_{s2} for the given data set and findings are presented in Table 4.1

Table 4.1 demonstrates that the MSE's of the members $\hat{\bar{y}}_{hd(1)}^{(1)}$, $\hat{\bar{y}}_{hd(1)}^{(2)}$ and $\hat{\bar{y}}_{hd(1)}^{(3)}$ are less than the suggested estimators $\hat{t}_{s2}, \hat{t}_{s1}$ and $\hat{t}_s$. Further we note that the class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}; j = 1,2,3$ is more efficient than $\hat{\bar{y}}_{\eta}^{*}$ (modified Ahmed et al (2017) estimator) and Diana et al (2014) estimator $\hat{\bar{y}}^{*}$. Thus we infer that the proposed class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$ is more efficient and more flexible than the other existing estimators.

5. CONCLUSION

In this paper we have considered the problem of estimating the population mean in presence of non- response using auxiliary information when the coefficient of variation of study variable y is known. We have improved the Ahmed et al (2017) model. We have further derived an improved estimator $\hat{\bar{y}}_{\eta}^{*}$ with that of Diana et al (2014) estimator $\hat{\bar{y}}^{*}$ and found that the estimator $\hat{\bar{y}}_{\eta}^{*}$ is more efficient than the Diana et al (2014) estimator $\hat{\bar{y}}^{*}$ under very realistic condition. A class of estimators $\hat{t}_s = M\ \hat{\bar{y}}_{\eta}^{*}$ and finally obtained the estimator $\hat{t}_s$ with known Coefficient of variation C_y for population mean $\overline{Y}$ is defined. Using two different values of *M*, we have also obtained the two different estimators $\hat{t}_{s1}$

and $\hat{t}_{s2}$. It is also shown that the estimator $\hat{\bar{y}}^{**}$ due to Ahmed et al (2017) is a member of the suggested estimator $\hat{t}_s$. It has been shown that the proposed estimator $\hat{t}_s$ is more efficient than the estimators $\hat{t}_{s1}$, $\hat{t}_{s2}$, $\hat{\bar{y}}_{\eta}^{*}$ and $\hat{\bar{y}}^{*}$ under very realistic condition $|\eta| < 1$. With this discussion we inferred that the suggested estimator $\hat{t}_s$ is more efficient as well as more flexible than Ahmed et al (2017) estimator $\hat{\bar{y}}^{**}$ and Diana et al (2014) estimator $\hat{\bar{y}}^{*}$ .Making the use of non-sensitive auxiliary variable *x* we have suggested a very general class of estimators $\hat{\bar{y}}_{hd}^{(j)}$, $j = 1, 2, 3$; for estimating the population mean $\bar{Y}$. The bias and MSE of the suggested class of estimators are derived. In particular, to illustrate the result of general class of estimators $\hat{\bar{y}}_{hd}^{(j)}$, $j = 1, 2, 3$; we have obtained the bias and MSE of the class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$. The optimum condition is obtained in which the MSE of $\hat{\bar{y}}_{hd(1)}^{(j)}$ is minimum. Two subclasses of estimators $\hat{\bar{y}}_{\alpha j}$ and $\hat{\bar{y}}_{\beta j}$ from the proposed class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$, $j = 1, 2, 3$; are identified alongwith their properties. It has been shown that the proposed class of estimators $\hat{\bar{y}}_{hd(1)}^{(j)}$ is better the classes of estimators $\hat{\bar{y}}_{\alpha j}$ and $\hat{\bar{y}}_{\beta j}$, $j = 1, 2, 3$. In support of the current study, an empirical study will be carried out.

ACKNOWLEDGEMENT: Authors are grateful to the Editor in chief- Professor D.K. Ghosh and thankful to the both learned referees for their valuable suggestions regarding improvement of the paper.

REFERENCES

Ahmed, S., Shabbir, J. and Gupta, S. (2017): Use of scrambled response model in estimating the finite population mean in presence of non-response when coefficient of variation is known. *Communications in Statistics- Theory and Methods*, 46 (17), 8435-8449.

Bar- Lev, S.K., Bobivitch, E. and Boukai, B. (2004): A note on randomized response models for quantitative data. *Metrika,* 60, 255 – 260.

Diana, G. and Perri, P.F. (2009): Estimating a sensitive proportion through randomized response procedures based on auxiliary information. *Statistical Papers*, 50, 661–672.

Diana,G. and Perri, P.F. (2010): New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics*, 37(11), 1875-1890.

Diana, G. and Perri, P.F. (2011): A class of estimators for quantitative sensitive data. *Statistical Papers*, 52, 633–650.

Diana, G., Riaz, S. and Shabbir, J. (2014): Hansen and Horwitz estimator with scrambled response on the second call. *Journal of Applied Statistics*, 41(3), 596-611.

Eichhorn, B.H. and Hayre, L.S. (1983): Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference,*7(4), 307-316.

Fox, J.A. and Tracy, P.E. (1986): Randomized response: A method for sensitive survey, *Sage Publication Inc., NewburyPark, CA.*

Hansen, M. H. and Hurwitz, W. N. (1946): The problem of non-response in sample surveys. *Journal of the American Statistical Association,* 41, 517-529.

Khare, B.B. and Kumar, S. (2009): Utilization of coefficient of variation in the estimation of population mean using auxiliary character in the presence of non-response. *National Academy of Science Letter, India,* 32(7&8), 235-241.

Khare, B.B. and Kumar, S. (2011): Estimation of population mean using known coefficient of variation of the study character in the presence of non-response. *Communications in Statistics- Theory and Methods,* 40(11), 2044-2058.

Mangat, N.S. and Singh R., (1990): An alternative randomized response procedure. *Biometrika*, 77(2), 439- 442.

Pollack. K.H. and Bek, Y. (1976): A Comparison of Three Randomized Response Models For Quantitative Data. *Journal of the American Statistical Association,* 71(356), 884-886.

Rajyaguru A. and Gupta P.C. (1995): On estimation of coefficient of variation from finite population I. *Journal of Statiatical Research (Dhaka)*, 36(2), 145-156.

Rajyaguru A. and Gupta P.C. (2006): On estimation of coefficient of variation from finite population II. *Journal of Model Assisted Statistics and Application*, I, 57-66.

Rajyaguru A. and Gupta P.C. (2004): On estimation of coefficient of variation from population- III, *Journal of South Gujarat University, Surat,* 2, 157-162.

Saha, A. (2007): A simple randomized response technique in complex surveys. *Metron*, 65(1), 59-66.

Searls, DT. (1964): The utilization of known coefficient of variation in the estimation procedure. *Journal of the American Statistical Association,* 59, 1225-1226.

Searls, DT. (1967): A note on the use of an approximately known coefficient of variation. *Journal of the American Statistical Association,*21, 20-21.

Sen, A.R. (1978): Estimation of the population mean when the coefficient of variation is known. *Communications in Statistics- Theory and Methods,* 7(7), 657-672.

Sen, A.R. (1979): Sampling theory on repeated occasions with ecological applications. *Sampling Biological Populations*, 315-328.

Shabbir, J. and Gupta, S (2005): On modified randomized device of warner's model. *Pakistan Journal of Statistics and Operation Research*, 21, 123–129.

Singh, H.P (1986): A generalized class of estimators of ratio, product and mean using supplementary information on an auxiliary character in PPSWR sampling scheme. *Gujarat Statistical Review,*13 (2), 1-30.

Singh, S. (2003): Advanced sampling theory with applications: how michael "selected" amy. Dordrecht: Kluwer.

Singh, H.P. and Kumar, S. (2008): A regression approach to the estimation of finite population mean in the presence of non-response. *Australian and New Zealand Journal of Statistics,* 50(4), 395-408.

Singh, H.P. and Kumar, S. (2009): A generalized procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *SORT* (*Statistics and Operations Research Transactions*), 33(1), 71-84.

Singh, H.P. and Gorey, S.M. (2016): Efficient estimation of population mean of sensitive variable in presence of scrambled response. *Communications in Statistics-Theory and Methods,* 46(19), 9557-9565.

Singh, H.P. and Katyar, N.P. (1988): A generalized class of estimators for common parameters of two normal distribution with known coefficient of variation. *Journal of the Indian Society of Agricultural Statistics*, 40(2), 127-149.

Singh, H.P. and Tarray, T.A., (2014): An improved randomized response additive models. *Srilankan Journal of Applied Statistics,* 15(2), 131-138.

Tarray, T.A., and Singh, H.P(2016): New Scrambling Randomized Response Models. *Jordan Journal of Mathematics and Statistics* 9(1), 1-15.

Upadhyaya, L.N. and Singh , H.P. (1984): On the Estimation of Population Mean With Known Coefficient of Variation. *Biometrical Jour.* 26(8), 915-922.

Warner, S.L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association,* 60, 63-69.

Table 4.1: The MSEs of the suggested classes of estimators

MSEs	**Values of η**								
	-1	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1
$MSE(\hat{\bar{y}}_{\eta}^{*})$	1790.8975	1790.8946	1790.8925	1790.8913	1790.8909	1790.8913	1790.8925	1790.8946	1790.8975 $= MSE(\hat{\bar{y}}^{*})$
$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(1)}\right)$	1205.398	1205.396	1205.394	1205.393	1205.392	1205.393	1205.394	1205.396	1205.398
$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 1}\right)$	1206.185	1206.182	1206.18	1206.179	1206.178	1206.179	1206.18	1206.182	1206.185
$MSE_{\min}\left(\hat{\bar{y}}_{\beta 1}\right)$	1276.197	1276.194	1276.192	1276.191	1276.19	1276.191	1276.192	1276.194	1276.197
$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(2)}\right)$	1202.563	1202.561	1202.559	1202.558	1202.557	1202.558	1202.559	1202.561	1202.563
$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 2}\right)$	1203.342	1203.339	1203.337	1203.336	1203.336	1203.336	1203.337	1203.339	1203.342
$MSE_{\min}\left(\hat{\bar{y}}_{\beta 2}\right)$	1272.997	1272.994	1272.992	1272.991	1272.991	1272.991	1272.992	1272.994	1272.997
$MSE_{\min}\left(\hat{\bar{y}}_{hd(1)}^{(3)}\right)$	1206.696	1206.694	1206.692	1206.691	1206.69	1206.691	1206.692	1206.694	1206.696
$MSE_{\min}\left(\hat{\bar{y}}_{\alpha 3}\right)$	1207.461	1207.459	1207.457	1207.455	1207.455	1207.455	1207.457	1207.459	1207.461
$MSE_{\min}\left(\hat{\bar{y}}_{\beta 3}\right)$	1276.47	1276.467	1276.465	1276.464	1276.464	1276.464	1276.465	1276.467	1276.47
$MSE(\hat{t}_{s})=$ $MSE\left(\hat{\bar{y}}_{3}^{**}\right)$	1692.407	1692.405	1692.403	1692.402	1692.401	1692.402	1692.403	1692.405	1692.407
$MSE(\hat{t}_{s1})=$ $MSE\left(\hat{\bar{y}}_{2}^{**}\right)$	1698.287	1698.285	1698.283	1698.282	1698.281	1698.282	1698.283	1698.285	1698.287
$MSE(\hat{t}_{s2})=$ $MSE\left(\hat{\bar{y}}_{1}^{**}\right)$	1703.89	1703.89	1703.89	1703.89	1703.89	1703.89	1703.89	1703.89	1703.89

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Statistical Models to Find Correlates of Alcohol Use among Adolescents in India: Comparative Appraisal of Conventional and Hierarchical Models

Ashish Datt Upadhyay, Sada Nand Dwivedi, Vishnubhatla Sreenivas[1],Anju Dhawan[2]

1. *Department of Biostatistics, AIIMS, New Delhi*
2. *National Drug Dependence Treatment Center, Department of Psychiatry, AIIMS, New Delhi, India*

Received: 21 January 2022 / Revised: 23 November 2022 / Accepted: 25 July 2023

Abstract

Alcohol use is a known risk factor of many ill health conditions like neuropsychiatric disorders, and non-communicable diseases. In the presence of hierarchical structure in data, multilevel regression model is often suggested to obtain more accurate analytical results regarding epidemiological understanding. Aim of this study was to find factors associated with alcohol use among adolescents and assess comparative findings under conventional and multilevel logistic models. Under this study, data of 37033 adolescents (15–19 years old) available under National Family Health Survey-3 (2005–2006) was utilized. The performance of the models was evaluated by the log-likelihood values, area under the ROC curve, AIC and BIC. As per findings of this study, socio-demographic variables likely to be positively associated with the chance of alcohol use among adolescents were low education, employment, caste other than Non-SC/ST, advancing age and male gender. Adolescents belonging to alcohol using households were more likely to use alcohol than their counterparts. Among the state level variables, adolescents belonging to states having prevalence of alcohol users above 12.5% and that of literacy rate (10th and above) below 30.8% were more likely to use alcohol. This study revealed that low education of adolescents and higher prevalence of alcohol use in the state are important factors to encourage alcohol use among adolescents. Interestingly, although associated factors under multilevel model also remained similar, the model performance parameters showed efficacy of multilevel model in comparison to conventional logistic model.

Keywords: Alcohol Use, Adolescents, Logistic regression Multilevel Model, Median Odds Ratio, intra-class correlation, 80% interval odds ratio

1 Introduction

Alcohol use is the third leading risk factor for poor health globally. According to W.H.O., 3.3 million deaths in 2012 may be attributed to harmful use of alcohol. Harmful use of alcohol contributes 5.9 % of all deaths in the world and 5.1% of the global burden of disease measured in the form of disability-adjusted life years lost. Worldwide, 38.3% consume alcohol and on average individuals of age 15 years or above consume 6.2 litres of alcohol annually [1]. It contributes too many ill health conditions like neuropsychiatric disorders, non-communicable diseases (e.g., cardiovascular diseases, cirrhosis of the liver, cancers), infectious diseases (e.g., HIV/AIDS, tuberculosis and pneumonia) and unintentional or intentional injuries (e.g., due to road traffic crashes and violence), and suicides [2]. In India, alcohol related problems account for more than one fifth of hospital admissions, 18% of psychiatric emergencies, more than 20% of all brain injuries, and 60% of all injuries reporting for emergency care [3]. Out of nearly 70 million alcohol users in India, 12

million were alcohol dependent [4]. According to national household survey of drug abuse, prevalence of ever alcohol use was 21.4% [5]. National Family Health Survey (NFHS-3) [6] revealed that prevalence of alcohol use in males (age 15-54 years) and females (age 15-49 years) were 32% and 2.2% respectively. Also, among 15-19 years old males, its prevalence was 11% and among females its prevalence was 1%. According to Global status report on alcohol and health by WHO [1], alcohol consumption increased in India during the period from 2008 to 2012. According to it, around 26 % of total population of India is ever alcohol users. The per capita consumption of alcohol in India also increased from 1.6 to 2.2 litres from 2003-2005 to 2010-2012. The adolescents are the future of a country and they often try new things to taste the world without thinking about pros and cons. As such, they are targeted by the alcohol companies. According to census 2011, 40.7% of the Indian population was less than 20 years of age and 20.9% were in the age group 10-19 years. Therefore, adolescents are needed to be handled gingerly because they are more probable to adapt the risk behavior like alcohol use. Surprisingly, in our county, a minimal attention has so far been given on understanding the determinants of alcohol use among adolescents. Also, in the fields of medical, social, and other sciences, characteristics of individuals get easily influenced by neighborhoods like community, district or state where they reside. In case of dichotomous outcome, the conventional logistic regression model assumes that all records are independent without taking into account existing hierarchical structure in the data [7,8]. Hence, virtually required assumption of independence does not get fulfill which results into underestimation of the standard errors of the parameters. In view of this, obtained inaccurate analytical results tend to inappropriate public health implications. Therefore, this study aimed to find out factors associated with alcohol use among adolescents and their comparative appraisal between the conventional and multilevel regression models.

2 Materials and Methods

Data was extracted on adolescents of age group 15-19 years from third round of the National Family Health survey (NFHS-3)[6]. This survey was done among 29 states in 2005–06 across India. NFHS-3 mainly provides national and state level estimates of, infant and child mortality, family planning, fertility, reproductive and child health, the quality of health and family welfare services and nutrition of women and children. Multistage sampling design with two stage design in rural areas and a three stage design in urban areas was used to select 109041 households. As such, a total of 124,385 females aged 15–49 years and 74369 males aged 15–54 years were interviewed. Methodological details are available in its national level report (IIPS and Macro International. 2007, NFHS-3, 2005–06: India: Volume I & Volume II). In the survey questionnaires, there were two questions addressing self report on alcohol use. They were: (1) Do you drink alcohol? & (ii) How often do you drink alcohol: almost every day, about once a week or less often? Accordingly, the data on alcohol use was collected directly from adolescents (i.e., 15-19 years), among males as well as females.

On the basis of subject knowledge and review of literatures, a set of independent or exploratory variables were selected for the analysis. After exploratory analysis on raw forms of variables in original data, some of them were retained in their existing forms and others were modified to get meaningful results. The set of independent/ exploratory variables considered in their existing forms were: age; sex (male/female); place of residence (rural /urban); employment (yes/no); household structure (nuclear/non-nuclear); employment of adolescent (yes/no); and tobacco use of adolescent (yes/no). Further, some qualitative variables considered after appropriate modification were: adolescent's education (below secondary/ secondary and above); wealth index of household (poor or poorest/middle or high or highest), and adolescent's marital status (married/un-married or single). Also, religion and caste were pooled [9] to derive another variable religion/caste (SC-ST Hindu/other Hindu/Muslim /other religion). An adolescent was categorized as exposed to household alcohol use if any member other than him/herself uses alcohol in family. To

take into account hierarchical structure of data in analysis, some community / state level variables were included such as alcohol use prevalence in states ($\leq 12.5\%$ or $> 12.5\%$) and state literacy above 10th standard among subjects of 15 years and above ($< 30.8\%$ or $\geq 30.8\%$). Among these variables, state literacy was extracted from 2011 census data and threshold level was chosen considering the national average. However, percentage of alcohol use in states ($\leq 12.5\%$ or $> 12.5\%$) was derived by aggregating the proportion of alcohol users in each state and then states were categorized taking median proportion alcohol users among states.

Ethics Statement

In spite of used data being available in public domain for academic use (http://www.measuredhs.com), ethical clearance was obtained from institutional ethics committee.

Statistical Models

The conventional logistic regression was used to illustrate the relationship of associated variables with alcohol use among adolescents. Under this model, the probability of using alcohol by i^{th} adolescent in the j^{th} state is expressed as p_{ij} which is a function of considered associated variables [10]:

$$\log_e \left[\frac{p_{ij}}{1 - p_{ij}} \right] = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ijk}$$

Where, β_k is the coefficient of regression of the k^{th} associated factor and x_{ijk} represents i^{th} adolescent's value in j^{th} state for k^{th} associated factor; β_0 is a constant.

Intuitively, under the above model, state level characteristics are also dis-aggregated at individual level (i.e., adolescent's level) that distorts the assumptions of independence. It may obviously under estimate standard error of the regression coefficients of the variables. As a result, even a variable with a least relevance may turn up to be significantly associated with alcohol use.

Under present study, adolescents were nested within states. The variation may prevail at both levels, adolescent level and state level. To quantify variation at each level, multilevel analysis may be applied. For this, based on exploratory analysis, multilevel logistic regression random intercept model involving 2-level data structure, level-1 (adolescents) and level-2 (states) was considered [8,10]:

$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{00} + \beta x_{ij} + \delta w_j + u_j + e_{ij}$

Where,

$u_j \sim N\left(0, {\sigma_u}^2\right), \& \; e_{ij} \sim N\left(0, {\sigma_{ij}}^2\right)$

Where, p_{ij} is probability that i^{th} adolescent in j^{th} state uses alcohol.

x_{ij} and w_j are vectors of individual adolescent and state level characteristics; β and δ are vectors of estimated regression coefficients for the respective covariates.

u_j: unobserved variation at state level.

e_{ij}: error terms at individual level.

The variability unexplained by considered state level covariates is estimated under multilevel regression model by estimation of σ and given as:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_{00} + \beta x_{ij} + \delta w_j + \sigma v_j + e_{ij}$$

Where, $v_j \sim N(0, 1)$ and therefore variability estimated by the multilevel approach known as Multilevel effect is given by the term 'σv'. Accordingly, if $v_j = 1$, there will be increase in $\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right)$ by σ. On the other hand, if $v_j = -1$, then there will be decrease in $\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right)$ by σ.

In the beginning, using chi-square test, association analysis was carried out for each exploratory variable to assess its association with alcohol use among the adolescents. All variables found significant at 25% level of significance in association analysis were considered for multivariable logistic regression analysis. Further, to begin with, the possible multi-co linearity and first order interaction effect was explored during model building process. Non-occurrence of multi-co linearity was considered for a covariate with cut-off point of mean of variation inflation factor (VIF) as less than five [11]. Effect modifier was explored by two methods; first was stratified analysis with confidence interval approach, and second was multivariable regression approach. However, among the considered variables in this study, none of them was found to be either multi-collinear or effect modifier. Stepwise regression approach with entry probability of 0.10 and an exit probability of 0.15 was used under conventional multivariable logistic regression analysis. Further, a maximum likelihood approach was used for parameter estimation.

Keeping in view of comparative appraisal, multivariable multilevel logistic regression with random intercept analysis was also performed on similar set of covariates as those under multivariable conventional logistic regression. The log likelihood for this model was approximated by maximum likelihood estimation with adaptive Gaussian quadrature [12].

The area level variance was assessed by median odds ratio, 80% interval odds ratio, and intra-class correlation. The social, economic, health facilities and other characteristics of one state are likely to be different from another state. Further, all adolescents in a state share same state level characteristics. Intra-class correlation measures the proportion of total variance in the outcome that is attributable to the state level characteristics [13]. Median Odds Ratio (MOR) proposed by Larsen & Merlo (2005) [13] transforms the state level variance to the odds ratio scale. It is the median of the set of odds ratios that could be calculated by comparing the randomly chosen two adolescents, one from low risk state and other from high risk state with identical individual level covariates. If MOR is one, it implies that the area level variation is close to zero and if greater than one, there exits considerable amount of state level variation. An 80% Interval Odds Ratio (IOR-80) is another index used for describing the state level variability [13]. It is the measure of a fixed-effect which quantifies the effect of state-level variables. It is middle eighty percent of the range of all odds ratio evaluated for state level variables from each pair of all probable pairs of adolescents with identical individual-level risk factors from different states but who differ by one level in state-level risk factors. If the IOR-80 does not include unity, it precisely means that specific state level variable describes the area level variability substantially. In spite of this, if interval is wider or includes 1, it implies that area level variability is minimally explained by specific state level variable [13].

Predictive performance of the developed models was compared using indices such as Log-likelihood, Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and area under receiving operative curve analysis (ROC). Higher values of Log-likelihood and ROC and lower values of AIC and BIC indicate better fit of the model. The analysis was carried out on Stata software (version 14) and p-value less than 0.05 was considered as significant.

3 Results

In NFHS-3, out of 37025 adolescents (15-19 years), 1884 (5.09%) were alcohol users. Among these adolescents, sixty five percent were female, and more than half of them (54%) were from rural area with mean age ($\pm$SD) 17.0$\pm$1.3 years. Bivariate analysis (Tables 1(a) & 1(b)) showed that adolescents who were male (OR=7.4, 95%CI: 6.69– 8.13), educated below secondary school (OR=1.6, 95%CI: 1.47-1.79) and belong to non-nuclear family (OR=1.1, 95%CI: 1.03–1.24), were significantly more likely to use alcohol. On the other-hand, adolescents residing in a state with low level of 10^{th} standard and above literacy had higher chance of using alcohol (OR=1.4, 95%CI: 1.28-1.55). Similar finding was observed in

case of wealth status (OR=1.6, 95%CI: 1.43-1.74).

Table 1(a): Association of alcohol use among adolescents with individual level socio-economic & demographic factors

Characteristic	Alcohol Use		Un-adjusted Odds Ratio (95%C.I.)
	Yes f(%)	No f(%)	
Age(years) Mean ± SD	17.5 ± 1.3	16.9 ± 1.4	1.3 (1.26,1.37)
Sex			
Female	352 (1.5)	23596 (98.5)	1
Male	1532 (11.7)	11545 (88.3)	7.4 (6.69,8.13)
Residence			
Urban	784 (4.6)	16236 (95.4)	1
Rural	1100 (5.5)	18905 (94.5)	1.2 (1.09,1.32)
Household structure			
Nuclear	956 (4.8)	18905 (95.2)	1
Non-nuclear	928 (5.4)	16236 (94.6)	1.1 (1.03,1.24)
Wealth Index			
Richer/Richest/ Middle	1224(4.5)	26206(95.5)	1
Poorest/Poorer	660 (6.9)	8935 (93.1)	1.6 (1.43,1.74)
Caste or Religion			
Hindu(non SC/ST)	687 (3.8)	17248 (96.2)	1
Hindu(SC/ST)	567 (7.3)	7247 (92.7)	1.9 (1.75,2.20)
Muslim	80 (1.4)	5764 (98.6)	0.3 (0.28,0.44)
Others religions	539 (10.3)	4674 (89.7)	2.9 (2.57,3.26)
Education			
Secondary & above	1222 (4.4)	26371(95.6)	1
Illiterate/Primary	662 (7.1)	8761 (92.9)	1.6 (1.48,1.79)
Marital status			
Unmarried/Single	1673 (5.2)	30288 (94.8)	1
Married	211 (4.2)	4853 (95.8)	0.8 (0.68,0.91)
Employment			
Unemployed	769 (3.1)	24597 (96.9)	1
Employed	1115 (9.6)	10465 (90.4)	3.4 (3.10,3.74)
Exposure to media			
No	123 (3.9)	2983 (96.1)	1
Yes	1761 (5.2)	32140 (94.8)	1.3 (1.10,1.60)
Tobacco use			
No	592 (1.8)	31634 (98.2)	1
Yes	1292 (26.9)	3507 (73.1)	19.7 (17.7,21.82)

Table 1(b): Association of alcohol use among adolescents with household level & states level factors

Characteristic	Current Alcohol Use		Un-adjusted Odds Ratio (95%C.I.)
	Yes f(%)	No f(%)	
Family alcohol use			
No	745 (3.1)	23696 (96.9)	1
Yes	973 (9.9)	8811 (90.1)	3.5 (3.18,3.88)
Literacy rate (10th and above) in State			
$\geq$ 30.8%	710 (4.2)	16158 (95.8)	1
$<$ 30.8%	1174 (5.8)	18983 (94.2)	1.4 (1.28,1.55)
Prevalence of alcohol use in state			
$\leq$ 2.5%	568 (3.5)	18118 (96.5)	1
$>$ 12.5%	1316 (9.1)	17023 (90.9)	2.5 (2.23,2.73)

Under multivariable logistic regression analysis (Table 2(a)), the statistically significant socio-demographic, household and state level variables associated with alcohol use were gender, age, caste or religion, marital status, employment, education, tobacco use of adolescents, household member's alcohol use, and state level prevalence of alcohol use among adults, and prevalence of 10^{th} and above education. Adolescents belonging to Schedule tribe caste AOR=1.4, (95%CI: 1.24–1.64) or other religions AOR=1.9, (95%CI: 1.68–2.26) as compared to Non–Scheduled Caste or Tribe; and who were married AOR=1.4, (95%CI: 1.15–1.77) were more likely to use alcohol. However, Muslim adolescents were less likely to use alcohol as compared to Non–SC/ST AOR=0.3, (95%CI: 0.24–0.39). Further, adolescents who had education below secondary level were 40 percent more chance to use alcohol as compared to their counterparts AOR=1.4, (95%CI: 1.18–1.55). Also, adolescents who were employed, has 40 percent higher chance AOR=1.4, (95%CI: 1.27–1.64). As obvious, tobacco using adolescents had more than 8-fold higher chance AOR=8.2, (95%CI: 7.24–9.27) of using alcohol. Among household level variables (Table 2(b)), adolescents exposed to household level alcohol use had more than two-fold chance to use alcohol AOR=2.2, (95%CI: 1.84–2.48). Among state level variables, adolescents belonging to state with lower education level of 10^{th} standard and above had 20% AOR=1.2, (95%CI: 1.09–1.42) more chance to use alcohol. Further, adolescents residing in states having higher prevalence of alcohol users had two-fold higher chance AOR=2.1, (95%CI: 1.85–2.40) to use alcohol.

Table 2(a): Adjusted association of alcohol use among adolescents with individual level study variables

Characteristic	Logistic Regression Adjusted Odds Ratio (95%C.I.)	Multilevel Logistic Regression Adjusted Odds Ratio (95%C.I.)
Age (years)	1.2 (1.16,1.26)	1.2 (1.16,1.27)
Sex		
Female	1	1
Male	5.7 (4.91,6.73)	6.6 (5.58,7.78)

Caste & Religion		
Hindu(non SC/ST)	1	1
Hindu(SC/ST)	1.4 (1.23,1.63)	1.4 (1.24,1.67)
Muslim	0.3 (0.24,0.39)	0.28 (0.21,0.36)
Others religions	1.9 (1.68,2.26)	2.1 (1.72,2.55)
Education		
Secondary&above	1	1
Illiterate/Primary	1.4 (1.18,1.55)	1.3 (1.09,1.45)
Employment		
Unemployed	1	1
Employed	1.4 (1.27,1.64)	1.4 (1.25,1.62)
Tobacco use		
No	1	1
Yes	8.2 (7.24,9.27)	10.4 (9.09,11.9)
Marital status		
Unmarried/Single	1	1
Married	1.4 (1.15,1.77)	1.4 (1.13,1.77)

The results under multilevel model accounting hierarchical structure of data (Table 2(a)) revealed that there was variation in alcohol use among the states and proportion of the estimated variance in alcohol use among adolescents between states was 13% (ICC=13%, 95% CI: 7.0%-21.0%). In terms of Median odds ratio (MOR), if adolescents move to another state with a higher probability of adolescents' alcohol use, the median increase in their odds of alcohol use would be almost two-fold (MOR=1.9, 95% CI: 1.64-2.39). As evident from the table 2(a), nearly all the significantly associated factors under conventional logistic regression analysis remained significant under multilevel analysis. However, their confidence intervals became wider in multilevel model. It may be attributed to the fact that conventional logistic regression does not account the state heterogeneity. Secondly, in case of multilevel logistic regression, odds of alcohol use by adolescents having education below secondary (OR=1.2, 95% CI: 1.09–1.44) is interpreted as if we compare two adolescents with identical level of associated factors, one with below secondary education and one with secondary and above, limited to the same state, then the chance of alcohol use increased by 1.2 times for the adolescents having below secondary education. In case of state level variables (table 2(b)), this interpretation is limited to state having same level of alcohol use. The odds of alcohol use for adolescents of state with lower literacy ($<$ 30.8%), and other from its counterpart and those states possibly differ in alcohol use risks, the odds ratio for the comparison will lie between 0.31 to 3.69 with 80% probability. In case of state level adult alcohol use prevalence ($>$ 12.5%), IOR-80% was (0.62, 7.40). Since IOR-80 for state level variables were wider and contain 1, it indicates incapability of state level factors to add meaningful justification of variation in prevalence of alcohol use among adolescents in the states.

Table 2(b): Adjusted association of alcohol use among adolescents with household level & states level factors and measures of state level variation

Characteristic	Logistic Regression Adjusted Odds Ratio (95%C.I.)	Multilevel Logistic Regression Adjusted Odds Ratio (95%C.I.)
Household alcohol use		
No	1	1
Yes	2.2 (1.84, 2.48)	1.9 (1.75,2.22)
Literacy rate (10th standard and above)		
≥30.8%	1	1
< 30.8%	1.2 (1.09, 1.42)	1.1 (0.62, 1.82)
IOR-80		[0.31, 3.69]
Prevalence of alcohol use		
≤12.5%	1	1
> 12.5%	2.1 (1.85, 2.40)	2.1 (1.25, 3.65)
IOR-80		[0.62, 7.40]
Measures of state level variation		
State level Variance (95% C.I.) Full Model		0.47 (0.27, 0.84)
ICC(State level) (95% C.I.) Full model		0.13 (0.07, 0.21)
Median Odds Ratio (95% C.I.) Full Model		1.9 (1.64, 2.39)

The assessment of discriminating ability of two modelling approaches (table 3) shows that Area under Receiving Operative Curve analysis (AROC) was more under multilevel model (AROC= 92.7 95%CI: 92.21-93.24) as compared to conventional logistic model (AROC=91.5, 95%CI: 90.41-91.73). Also, the multilevel model had largest log likelihood and the smallest AIC and BIC, as compared to conventional logistic regression, which suggests a best goodness of fit in case of multilevel model.

Table 3: Comparison of models for alcohol use among adolescents in India

	Conventional Logistic Regression	Multilevel Logistic Regression
Log likelihood	-4479.3	-4290.7
AIC	8984.7	8609.4
BIC	9094.3	8727.5
Area under ROC	91.5 (90.41, 91.73)	92.7 (92.21, 93.24)

4 Discussion

A harmful alcohol use among adolescents has various bad impacts not only on their future life, but also on society in general. The present study explored the factors associated with adolescents' inclination towards

alcohol use along with comparative appraisal of the developed conventional multivariable logistic regression model and multilevel multivariable logistic regression model in terms of their predicting ability. Each of the two multivariable analyses revealed that similar factors are associated with alcohol use among adolescents. Among them, one of the important factors which increase the chance of alcohol use was house-hold level alcohol use. This is consistent with earlier studies [14-18]. It also implies the role of drinking parents which influences the children's attitude and personal beliefs towards alcohol use after they get exposed to drinking by family members [14] and also indicates the social learning about initiation of alcohol use. As a further endorsement, adolescents belonging to states where prevalence of alcohol use was higher were more likely to use alcohol. The findings under this study suggest that it is highly required to educate parents and society about ill effect of alcohol use and strong message should be communicated to them that those children are more likely to use alcohol whose family members use alcohol. It also suggests that community interventions may be required that include supportive environment, strong policy support and community participation. The male and older adolescents were more likely to use alcohol and this finding was also observed in other studies [15,19]. Interestingly, in contrary to univariable analysis, marital status of the adolescents turned to be a risk factor in multivariable regression analysis. This may be attributed to the fact that nearly 65% of adolescents were women and among them 20% were married, where as among males (35%) only 2 % percent were married. The present study also has shown that literacy plays an important role in protecting adolescents from using alcohol. The adolescents having lower literacy and belonging to state where literacy was low were more likely to use alcohol than their counter parts [15, 20]. The reason behind this may be that due to higher education, individuals and community are likely to have better understanding about bad impact of alcohol use on self as well as at community level. Tobacco use was a major correlate of alcohol use in our study. Joint use of tobacco and alcohol has been reported in India [21] and other parts of the world [17]. Earlier studies on adolescents in India have not considered caste and religion [23,24]. In present study, alcohol use was more prominent among Hindus (S.C./S.T.) and other religions as compared to Hindus (Non S.C/S.T.). On the other hand, Muslims were less likely to use alcohol. This finding is analogous with other studies [19, 22,25]. Further, studies on general population have shown same trend [26]. The prevailing low literacy among adolescents with lower socioeconomic status may be major reasons behind it. In terms of discriminating ability and other performance indicators, multilevel model emerged to be better as compared to conventional logistic model. It emerged to be better in terms of lower AIC and BIC. As obvious, confidence intervals of the estimated parameters were slightly wider under multilevel regression models than conventional logistic regression [27]. This may be attributed to the fact that conventional logistic regression does not take into account the state level heterogeneity [27]. Further, observed IOR- 80% implies that considered state level variables did not add much in the variation in alcohol use by adolescents among the states. In addition, median odds ratio also implies that in spite of considered covariates in the models, a substantial between state variability still exists. As such, more relevant state level variables may be required to explain the states' heterogeneity regarding alcohol use more closely.

5 Conclusion

This study has dealt with the application, interpretation and comparison of conventional and multilevel logistic regression models regarding determinants of alcohol use among the adolescents. To the best of our knowledge, this study is the first study on such a large national level data to examine the associated factors of alcohol use among adolescents more scientifically. In our study, multilevel model outperformed conventional model due to obvious presence of clustering in our data. This study has shown that education level of adolescents and their residing community, and prevailing alcohol use in their family and community were main predictors of alcohol use among adolescents. Keeping in view of the associated factors, the

preventive activity to curb alcohol use among them can be carried out by encouraging them regarding higher education and also changing the social norm of alcohol use among the parents and society at home as well as at public places. To generate effective public health program, since data used in this study was not collected solely for the alcohol use, further research is needed to explore the vulnerability of certain more relevant variables associated with alcohol use.

6 Acknowledgement

The authors acknowledge the All India Institute of Medical Science, New Delhi, for supporting required infrastructure during the study.

7 Conflict of interest

The authors declare no conflict of interests.

References

[1] World Health Organization, Management of Substance Abuse Unit, 2014. Global status report on alcohol and health, 2014. World Health Organization, Geneva.

[2] World Health Organization (Ed.), 2010. Global strategy to reduce the harmful use of alcohol. World Health Organization, Geneva.

[3] Prasad, R., 2009. Alcohol use on the rise in India. The Lancet 373, 17–18. https://doi.org/10.1016/S0140-6736(08)61939-X

[4] Gururaj G, Murthy P, Girish N, Benegal V, 2011. Alcohol related harm: implications for public health and policy in India, 73rd ed. NIMHANS.

[5] Pal, H., Srivastava, A., Dwivedi, S.N., Pandey, A., Nath, J., 2015. Prevalence of Drug Abuse in India through a National Household Survey 11.

[6] International Institute for Population Sciences (IIPS) and Macro International. 2007. National Family Health Survey (NFHS-3), 2005–06: India: Volume I & Volume II. Mumbai: IIPS

[7] Raudenbush, S. W., and Bryk, A. S. (1992). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.

[8] SN Dwivedi, KR Sundaram., 2000. Epidemiological models and related simulation results for understanding of contraceptive adoption in India, International Journal of Epidemiology, Volume 29, Issue 2, April 2000, Pages 300–307

[9] Pandey, A., Choe, M. K., Luther, N. Y., Sahu, D., and Chand, J. (1998). Infant and Child Mortality in India: National Family Health Survey Subject Reports Number 11. International Institute of Population Sciences: Mumbai

[10] Goldstein, H. (2011). Multilevel Statistical Models, 4th edition. New York: Edward Arnold.

[11] Garson GD (2012) Testing statistical assumptions, pp: 44-45.

[12] Pinheiro, J. C., and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics, 15(1), 58-81

[13] Larsen, K., and Merlo, J. (2005). Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. American journal of epidemiology, 161(1), 81-88.

[14] Dhupdale, N.Y., Motghare, D.D., Ferreira, A.M.A., Prasad, Y.D., 2006. Prevalence and Pattern of Alcohol Consumption in Rural Goa 31, 2.

[15] Kestila, L., Martelin, T., Rahkonen, O., Joutsenniemi, K., Pirkola, S., Poikolainen, K., Koskinen, S., 2008. Childhood and Current Determinants of Heavy Drinking in Early Adulthood. Alcohol and Alcoholism 43, 460–469. https://doi.org/10.1093/alcalc/agn018

[16] Benjet, C., Borges, G., Méndez, E., Casanova, L., Medina-Mora, M.E., 2014. Adolescent alcohol use and alcohol use disorders in Mexico City. Drug and Alcohol Dependence 136, 43–50. https://doi.org/10.1016/j.drugalcdep.2013.12.006

[17] Jackson, N., Denny, S., Ameratunga, S., 2014. Social and socio-demographic neighborhood effects on adolescent alcohol use: A systematic review of multi-level studies. Social Science & Medicine 115, 10–20. https://doi.org/10.1016/j.socscimed.2014.06.004

[18] Dada, O., Odukoya, O., Okuyemi, K., 2016. Risk perception and correlates of alcohol use among out-of-school youth in motor parks in Lagos State, Nigeria. Malawi Medical Journal 28, 19. https://doi.org/10.4314/mmj.v28i1.5

[19] Adhikari, R.P., Upadhaya, N., Pokhrel, R., Suwal, B.R., Shrestha, M.P., Subedi, P.K., n.d. Health and Social Vulnerability of Adolescents in Nepal 6.

[20] Sarangi, L., Acharya, H.P., Panigrahi, O.P., 2008. Substance abuse among adolescents in urban slums of Sambalpur. Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine 33, 265

[21] Gupta PC, Maulik PK, Pednekar MS, Saxena S. Concurrent alcohol and tobacco use among a middle-aged and elderly population in Mumbai. Natl Med J India. 2005;18(2):88

[22] Jaisoorya, T.S., Beena, K.V., Beena, M., Ellangovan, K., Jose, D.C., Thennarasu, K., Benegal, V., 2016. Prevalence and correlates of alcohol use among adolescents attending school in Kerala, India: Alcohol use among adolescents in India. Drug and Alcohol Review 35, 523–529. https://doi.org/10.1111/dar.12358

[23] Mini, S., S., S.K., K., S.K., 2017. Prevalence of alcohol use among high school students, the pattern of consumption and the physical circumstances associated with alcoholism in an urban area of Kerala, India. International Journal Of Community Medicine And Public Health 4, 738. https://doi.org/10.18203/2394-6040.ijcmph20170750

[24] Singh, A., Gupta, P., Gupta, P., Srivastava, M.R., Srivastava, M.R., Zaidi, Z.H., Zaidi, Z.H., 2017. A cross sectional study on tobacco and alcohol abuse among medical college students. International Journal Of Community Medicine And Public Health 4, 3372. https://doi.org/10.18203/2394-6040.ijcmph20173847

[25] Sumskas L, Zaborskis A, Grabauskas V. Social determinants of smoking, alcohol and drug use among Lithuanian School-aged children: results from 5 consecutive HBSC surveys, 1994-2010. Medicina (Kaunas). 2012;48(11):595-604.

[26] Subramanian, S.V., Nandy, S., Irving, M., Gordon, D., Smith, G.D., 2005. Role of socioeconomic markers and state prohibition policy in predicting alcohol consumption among men and women in India: a multilevel statistical analysis. Bulletin of the World Health Organization 10.

[27] Sanagou, M., Wolfe, R., Forbes, A., Reid, C.M., 2013. Hospital-level associations with 30-day patient mortality after cardiac surgery: a tutorial on the application and interpretation of marginal and multilevel logistic regression. BMC Medical Research Methodology 12. https://doi.org/10.1186/1471-2288-12-28

The Panel of the Reviewers in the current issue

Dr. Indranil Ghosh
Dept. of Mathematics and Statistics
University of North Carolina, Wilmington
North Carolina, USA

Dr. Sathees Kumar
Dept. of Statistics
University of Kerala
Thiruvanthapuram. Kerala, India

Dr. K. Muralidharan
Dept. of Statistics
Faculty of Science
The Maharaja Sayajirao University of Baroda
Vadodara, Gujarat, India

Dr. P. C. Gupta
Dept. of Statistics
University of Rajasthan
Jaipur, Rajasthan, India

Dr. B. V. S. Sisodiya
Retd. Professor and Head
Dept. of Agricultural Statistics
Narendradev Agricultural University,
Faizabad, UP

Dr. G. C. Tikkiwal
Retd. Professor and Head
Dept. of Statistics
Jai Narain Vyas University, Jodhpur,
Jodhpur, Rajasthan

Dr. Piyush Kant Rai
Dept. of Statistics
Banaras Hindu University
Varanasi, UP

Dr. Kaustav Aditya
Sampling Division
ICAR- I. A. S. R. I. New Delhi
Pusa Road, New Delhi 12

Dr. Sadanand Dwivedi
Dept. of Bio-Statistics
AIIMS, New Delhi

Dr. Anoop Chaturvedi
Retd. Professor
Dept. of Statistics
Allahabad University, Allahabad, UP

Dr. Avinash Dharmadhikari
Tata, Motors
Pune

Dr. Girish Chandra
Dept. of Statistics
University of Delhi
Delhi

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Memberships:

One can become the Member/Life member of the Gujarat Statistical Association by applying in the prescribed form available at the website of Gujarat Statistical Association (GSA) https://www.thegsa.in/

Subscription rates: Subscription rates for Gujarat Journal of Statistics and Data Science are as follows:

Inland	Rs. 500.00 (Inclusive of postage)
Foreign	U.S. $ 100.00 (Inclusive of postage)

Gujarat Journal of Statistics and Data Science will be available at Gujarat Statistical Association (GSA) site https://www.thegsa.in/. The members of the association, editorial board, and the authors can download the paper(s) free of cost.

Gujarat Journal of Statistics and Data Science (formerly Gujarat Statistical Review) is a peer reviewed research journal in Statistics and published twice in a year in past. Now the journal will be publishing once in a year in the beginning. The journal will publish referred original research papers, reviews, and case studied related to any branch of applied and theoretical Statistics. Book's reviews, letters to the Editor in Chief related to the article/papers of the journal are also welcome for possible publication.

Authors are encouraged to submit manuscripts in electronic form. Author should submit their manuscript as softcopy either directly to the Editor in Chief (ghosh_dkg@rediffmail.com) or through the Managing Editor (bikassinha1946@gmail.com) or Editor (amsseng@gmail.com).

How to prepare manuscript:

Author should prepare and submit their manuscripts electronically using the LaTeX package. Authors are also requested to submit Encapsulated PostScript ("eps") or .png or .jpeg or .pdf format files for figures in both electronic and non - electronic papers. Authors should ensure that laser – printed originals of these figures are of high quality and suitable for scanning. Authors are requested to submit their (1) LaTex file (.tex format) [Refer the LaTex format available at website], (2). Bibliography file (.bib format) [Refer demo.bib file available at website], and (3). pdf file of your article.

Authors are requested to download their manuscripts in the prescribed standard format available at website of Gujarat Statistical Association (GSA) https://www.thegsa.in/

Manuscript Components: Research paper must be written in English. The manuscript must be organized as following:

Title of the paper,
Author(s) name, affiliations and e-mail address,
Abstract (not more than 250 words) without reference,
Key word (not more than six),
AMS subject classification,
Complete articles which include Tables, Figure, graph (if any required),
Acknowledgement,
Conclusion,

Appendices (if required) and finally,

References.

The complete articles will be organized as appropriate number of sections, Subsections, Equations, Figures, Tables, Equations. All Figures and Tables should normally be mentioned explicitly by number and should appear in correct numerical order in the body of the text. Tables should be numbered continuously starting with 1 irrespective of their subsections etc.

The reference list and text citations should agree and be accurate. All references cited in the text must appear in the reference list, similarly, all references listed in the reference list must be cited in the text.

Acronyms and abbreviations should be spelled out the first time they are used unless they are common throughout the discipline. Avoid beginning sentences with a symbol, number, or lowercase letter.

References for research article and book should follow the following:

Research paper:

1. Ghosh, D.K. and Karmakar, P. (1988) Some series of efficiency balanced designs. Aust. Jour. Statist. 30(1), 47-51.
2. Ghosh, D.K. (1989). Construction of confounded designs for mixed factorial experiments, Jour. Statist. Planning and Inference., 23, 253- 261.

Book:

1. Das, M. N. and Giri, N. C. (1973). Design and Analysis of Experiments. Third Edition, New Age International Publishers.

It is the policy of the journal that no submission, or substantially overlapping submission, be published or be under review at another journal or conference at any time during the review process. Submission of a manuscript implies that it has been approved by all authors as well as by the responsible authorities tacitly or explicitly at the institute where the work has been carried out. The publisher and the editors will not be held legally responsible should there be any claims for compensation.

Decision after Review:

After considering the Reviewer's reports, the Editorial board members will make one of the following decisions:

(i) Accept the paper for possible publication
(ii) Request for a minor/major revision
(iii) Reject the publication of the paper

The peer-review process of the journal is a double-blind process. While submitting the revised manuscript, the authors should clearly explain point-by-point as to how the reviewer's comments have been addressed. If the authors disagree with reviewer's comments, the reason(s) should be explained. Decision of the editorial board will be final. Those papers recommended by two reviewers will be accepted for publication. The peer review process may be completed within 4 months.

Gujarat Statistical Association Body

Executive Committee

President (I/C)	Dr N D Shah
Vice President	Dr D K Ghosh
Secretary	Dr Chirag Trivedi
Joint Secretary	Dr R D Patel
Treasurer	Dr (Mrs.) C D Bhavsar

Members

Dr Rakesh Srivastav	Dr Parag Shah
Dr Ashok Shanubhogue	Dr H D Budhbhatti
Dr Aarti Rajguru	Dr L Murlidharan
Dr Sudhir Joshi	Dr A J Patel
Dr Mayuri Pandya	Dr Rajashree Mengore

Dr B B Jani (Editor: Sankhya Vignan)

Co-opted Members

Dr J B Shah	Dr Mrs. Kavita Dave

Invited Members

Dr Arnab Laha	Prof (Smt.) Subha Rani

Gujarat Journal of Statistics and Data Science

1.	Place of Publication:	Gujarat Statistical Association C/0 Department of Statistics Gujarat University, Navrangpura Ahmedabad – 380009, Gujarat, India
2.	Periodicity of Publication:	Annually
3.	Printer's Name:	Gujarat Statistical Association
4.	Address:	Gujarat Statistical Association C/0 Department of Statistics Gujarat University, Navrangpura Ahmedabad – 380009, Gujarat, India
5.	Publisher's Name:	Dilip Kumar Ghosh
6.	Nationality:	Indian
7.	Editor in Chief's Name:	Dilip Kumar Ghosh
8.	Owner's Name:	Gujarat Statistical Association

I, Dilip Kumar Ghosh declare that the particulars provided above are true to the best of my knowledge and belief.

Date: October, 31, 2023

Sd/- Dilip Kumar Ghosh

Signature of Publisher