Gujarat Journal of Statistics and Data Science

Volume: 40 (New Series) Issue: 1 October 2024

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 – 40 Issue – 1 October, 2024

CONTENTS

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Vol – 40 Issue – 1 October, 2024

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

I express my deep pleasure to bring up the next issue of this Gujarat Journal of Statistics and Data Science containing the research papers of reputed authors from National and International reputed experts. Professor Bikas Kumar Sinha, Managing Editor, and Professor Ashis SenGupta, Editor of this journal are extremely for me by helping 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. Dr. Parag Shah deserves special thanks 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. 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 inclusion of the following two Associate editors in the Editorial board which we did:
(1) Dr. Ravindra Khattree, Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA.
(2) Dr. Indranil Ghosh, University of North Carolina, Wilmington, North Carolina, USA.

In the advisory board of Gujarat Journal of Statistics and Data Science, there was no representation from Agricultural Statistics, so to fulfill this need, we included the name of Dr. Alok Shrivastava, Professor and Head, Department of Agricultural Statistics, Navsari University, Navsari, Gujarat, India.

With these few words, we place this issue of the Gujarat Journal of Statistics and Data Science before our readers at large. We fondly hope they will not be disappointed with this volume. I am really happy in bringing up the journal in time with research papers of the National and International reputed authors.

Dilip Kumar Ghosh
Editor in Chief

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

On a characterization of probability distributions based on maxima of independent or max-independent random variables

B.L.S. Prakasa Rao[1]

1. *CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India*

Received: Jan 2024 / Accepted: Sept 2024

Abstract

Kotlarski (1978) proved a result on identification of the distributions of independent random variables X, Y and Z from the joint distribution of the bivariate random vector (U, V) where $(U, V) = (\max(X, Z), \max(Y, Z))$. We extend this result to the case

$$(U, V) = (\max(X, aZ_1, bZ_2), \max(Y, cZ_1, dZ_2)$$

where X, Y, Z_1, Z_2 are independent or max-independent random variabkes, Z_1 and Z_2 are identically distributed and a, b, c, d are known positive constants.

Keywords: Kotlarski's lemma; Identifiability; Characterization; Max-independent; Independent; Maxima.

MSC 2020: Primary 62E10

1 Introduction

Let X_0, X_1 and X_2 be independent random variables. Define $Y_1 = \max(X_0, X_1)$ and $Y_2 = \max(X_0, X_2)$. It is of interest to know whether the joint distribution of (Y_1, Y_2) determines the individual distributions of X_0, X_1 and X_2 uniquely. It is known that the random variable Y_1 alone can not determine the distributions of X_0 and X_1 uniquely unless X_0 and X_1 are identically distributed random variables (cf. Prakasa Rao (1992), Section 7.3). Kotlarski (1978) and Klebanov (1973) obtained characterizations for probability distributions through maxima or minima of independent random variables. Prakasa Rao (2024) discussed characterizations of probabilty distributions based on maxima or minima of some families of dependent random variables. We now discuss extension of the result in Kotlarski (1978) leading to characterizations of probability distributions through maxima for some classes of independent or max-independent random variables.

2 Identifiability by maxima for independent random variables

The following result is due to Kotlarski (1978).

Theorem 2.1: (Identifiability by maxima) Suppose X_0, X_1, X_2 are independent random variables. Define $Y_1 = \max(X_0, X_1)$ and $Y_2 = \max(X_0, X_2)$. Then the joint distribution of (Y_1, Y_2) uniquely determines the distributions of the independent random variables X_0, X_1 and X_2 provided the supports of their distribution functions are the same.

For a proof of Theorem 2.1, see Kotlarski (1978) (cf. Prakasa Rao (1992), Theorem 2.2.1, p.24).

We will now generalize Theorem 2.1 in analogy with the results of Li and Zheng (2019) for linear functions of independent random variables.

Theorem 2.2: Let U and V be random variables defined by the relations

$$U = \max(X, aZ_1, bZ_2);\ V = \max(Y, cZ_1, dZ_2)$$

where X, Y, Z_1 and Z_2 are independent random variables, Z_1, Z_2 are identically distributed and a, b, c, d are known positive constants. Further suppose that the distribution functions of X, Y, Z_1 have the same support $R = (-\infty, \infty)$. Then, the joint distribution function of (U, V) uniquely determines the distributions of X, Y, Z_1 if $a = b$ or if $a \neq b$ but the distribution function of Z_1 is differentiable and the derivative is continuous.

Proof: Let $G(t_1, t_2)$ be the joint distribution function of (U, V). Suppose that (N, M, S_1, S_2) is another set of independent random variables with S_1 and S_2 identically distributed with the same support as that of X and such that the joint distribution of the random vector

$$(\max(N, aS_1, bS_2), \max(M, cS_1, dS_2))$$

is the same as that of the random vector (U, V). Let $F_X(.)$ denote the distribution function of the random variable X. It can be seen that, for any$-\infty < t_1, t_2 < \infty$,

$$\begin{aligned} G(t_1, t_2) &= P(\max(X, aZ_1, bZ_2) \leq t_1, \max((Y, cZ_1, dZ_2) \leq t_2) \qquad (2.\ 1)\\ &= P(X \leq t_1, aZ_1 \leq t_1, bZ_2 \leq t_2; Y \leq t_2, cZ_1 \leq t_2, dZ_2 \leq t_2)\\ &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \frac{t_1}{a}, Z_1 \leq \frac{t_2}{c}, Z_2 \leq \frac{t_1}{b}, Z_2 \leq \frac{t_2}{d})\\ &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \min(\frac{t_1}{a}, \frac{t_2}{c}), Z_2 \leq \min(\frac{t_1}{b}, \frac{t_2}{d}))\\ &= F_X(t_1)F_Y(t_2)F_{Z_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{Z_1}(\min(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned}$$

since X, YZ_1, Z_2 are independent random variables and Z_1, Z_2 are identically distributed. By a similar argument it follows that

$$\begin{aligned} G(t_1,t_2) &= P(\max(N, aS_1, bS_2) \leq t_1, \max(M, cS_1, dS_2) \leq t_2) \\ &= F_N(t_1)F_M(t_2)F_{S_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{S_1}(\min(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned} \tag{2. 2}$$

Hence

$$\begin{aligned} &F_X(t_1)F_Y(t_2)F_{Z_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{Z_1}(\min(\frac{t_1}{b}, \frac{t_2}{d})) \\ &= F_N(t_1)F_M(t_2)F_{S_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{S_1}(\min(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned} \tag{2. 3}$$

for all $-\infty < t_1, t_2 < \infty$. Then, for all $t \in R$, define

$$\eta_1(t) = \frac{F_X(t)}{F_M(t)}, \eta_2(t) = \frac{F_Y(t)}{F_N(t)} \text{ and } \eta_3(t) = \frac{F_{Z_1}(t)}{F_{S_1}(t)}.$$

Equation (2.3) implies that

$$\eta_1(t_1)\eta_2(t_2)\eta_3(\min(\frac{t_1}{a}, \frac{t_2}{c}))\eta_3(\min(\frac{t_1}{b}, \frac{t_2}{d})) = 1, t_1, t_2 \in R. \tag{2. 4}$$

Let $t_1 \to \infty$ in equation (2.4). From the properties of the distribution functions, it follows that

$$\eta_2(t_2)\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d}) = 1, t_2 \in R.$$

Hence

$$\eta_2(t_2) = [\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d})]^{-1}, t_2 \in R. \tag{2. 5}$$

Let $t_2 \to \infty$ in equation (2.4). From the properties of the distribution functions again, it follows that

$$\eta_1(t_1)\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b}) = 1, t_1 \in R$$

which implies that

$$\eta_1(t_1) = [\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b})]^{-1}, t_1 \in R. \tag{2. 6}$$

Combining the equations (2.4)-(2.6), it follows that

$$\eta_3(\min(\frac{t_1}{a}, \frac{t_2}{c}))\eta_3(\min(\frac{t_1}{b}, \frac{t_2}{d})) = \eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b})\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d}), t_1, t_2 \in R. \tag{2. 7}$$

Fix $t_1 \in R$ and let $t_2 \to \infty$. Then the expression on the left side of equation (2.7) tends to 1 and the expression on the right side tends to

$$\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b})$$

by the properties of the distribution functions. Hence

$$\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b}) = 1, t_1 \in R. \tag{2. 8}$$

This in turn implies that

$$\eta_1(t_1) = 1, t_1 \in R \tag{2. 9}$$

from (2.6). A similar analysis shows that

$$\eta_2(t_2) = 1, t_2 \in R \tag{2. 10}$$

by fixing $t_2 \in R$ and letting $t_1 \to \infty$. In particular, it follows that

$$F_X(t) = F_M(t), t \in R$$

and

$$F_Y(t) = F_N(t), t \in R.$$

Furthermore, equation (2.8) implies that

$$\log \eta_3(\frac{t}{a}) + \log \eta_3(\frac{t}{b}) = 0, t \in R. \tag{2. 11}$$

Let $\zeta(t) = \log \eta_3(t), t \in R$. Equation (2.11) implies that

$$\zeta(\frac{t}{a}) + \zeta(\frac{t}{b}) = 0, t \in D$$

or equivalently

$$\zeta(u) = -\zeta(\lambda u), u \in R$$

where $\lambda = \frac{a}{b}$. Suppose that $\lambda = 1$. Then it follows that $\zeta(u) = 0, u \in R$ which in turn implies that $\eta_3(t) = 1, t \in R$ and hence $F_{Z_1}(t) = F_{S_1}(t), t \in R$. If $\lambda \neq 1$, applying Lemma 2 in Li and Zheng (2019), it follows that $\zeta(u) = 0, u \in R$ under the additional condition of differentiability of the function $\zeta(u)$ and continuity of its derivative which implies that $F_{Z_1}(t) = F_{S_1}(t), t \in R$.

This completes the proof of Theorem 2.2.

We will now investigate the same problem when the constants $a > 0, b < 0, c > 0$ and $d < 0$.

Theorem 2.3: Let U and V be random variables defined by the relations

$$U = \max(X, aZ_1, bZ_2); \; V = \max(Y, cZ_1, dZ_2)$$

where X, Y, Z_1 and Z_2 are independent random variables, Z_1, Z_2 are identically distributed and $a > 0, b < 0, c > 0, d < 0$, are known constants. Further suppose that the distribution functions of X, Y, Z_1 have the same support $R = (-\infty, \infty)$. Then, the joint distribution function of (U, V) are connected by the distributions of X, Y, Z_1 and the distributions of M, N, S_1 by the equations through the equations (2.12) and (2.13) given below:

$$F_X(t) = F_M(t) \frac{F_{S_1}(\frac{t}{a})(1 - F_{S_1}(\frac{t}{b}))}{F_{Z_1}(\frac{t}{a})(1 - F_{Z_1}(\frac{t}{b}))} \tag{2. 12}$$

and

$$F_Y(t) = F_N(t)\frac{F_{S_1}(\frac{t}{c})(1 - F_{S_1}(\frac{t}{d}))}{F_{Z_1}(\frac{t}{c})(1 - F_{Z_1}(\frac{t}{d}))}. \tag{2. 13}$$

Proof: We follow the same notation as given in the proof of Theorem 2.2. It easy to see that that the joint distribution of The bivariate random vector (U, V) is given by

$$\begin{aligned} G(t_1, t_2) &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \frac{t_1}{a}, Z_1 \leq \frac{t_2}{c}, Z_2 \geq \frac{t_1}{b}, Z_2 \geq \frac{t_2}{d}) \\ &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \min(\frac{t_1}{a}, \frac{t_2}{c}), Z_2 \geq \max(\frac{t_1}{b}, \frac{t_2}{d})) \\ &= F_X(t_1)F_Y(t_2)F_{Z_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))(1 - F_{Z_1}(\max(\frac{t_1}{b}, \frac{t_2}{d})) \\ &= F_M(t_1)F_N(t_2)F_{S_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))(1 - F_{S_1}(\max(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned}$$

for all $t_1, t_2 \in R$. Define $\eta_1(t), \eta_2(t)$ and $\eta_3(t)$ as defined in the proof of Theorem 2.2. Then it follows that

$$\eta_1(t_1)\eta_2(t_2)\eta_3(\min(\frac{t_1}{a}, \frac{t_2}{c}))\frac{(1 - F_{Z_1}(\max(\frac{t_1}{b}, \frac{t_2}{d}))}{(1 - F_{S_1}(\max(\frac{t_1}{b}, \frac{t_2}{d}))} = 1, t_1, t_2 \in R. \tag{2. 14}$$

Let $t_2 \to \infty$. Then, it follows that,

$$\eta_1(t_1)\eta_3(\frac{t_1}{a})\frac{(1 - F_{Z_1}(\frac{t_1}{b}))}{(1 - F_{S_1}(\frac{t_1}{b}))} = 1 \tag{2. 15}$$

observing that $\frac{t_2}{d} \to -\infty$ as $t_2 \to \infty$. Letting $t_1 \to \infty$, it follows that

$$\eta_2(t_2)\eta_3(\frac{t_2}{c})\frac{(1 - F_{Z_1}(\frac{t_2}{d}))}{(1 - F_{S_1}(\frac{t_2}{d}))} = 1. \tag{2. 16}$$

These equations in turn show that

$$F_X(t) = F_M(t)\frac{F_{S_1}(\frac{t}{a})(1 - F_{S_1}(\frac{t}{b}))}{F_{Z_1}(\frac{t}{a})(1 - F_{Z_1}(\frac{t}{b}))} \tag{2. 17}$$

and

$$F_Y(t) = F_N(t)\frac{F_{S_1}(\frac{t}{c})(1 - F_{S_1}(\frac{t}{d}))}{F_{Z_1}(\frac{t}{c})(1 - F_{Z_1}(\frac{t}{d}))}. \tag{2. 18}$$

Remarks: Theorem 2.2 will continue to hold if the support R is replaced by $R_+ = [0, \infty)$ and Theorem 2.2 is not a consequence of Theorem 2.1. Suppose X, Y, Z_1 and Z_2 are independent random variables and a, b, c, d are positive constants.Note that, if

$$U = \max(X, aZ_1, bZ_2) = \max(aZ_1, \max(X, bZ_2)),$$

then

$$\frac{U}{a} = \max(Z_1, \max(\frac{X}{a}, \frac{bZ_2}{a})).$$

Similarly, if

$$V = \max(Y, cZ_1, dZ_2) = \max(cZ_1, \max(Y, dZ_2)),$$

then

$$\frac{V}{c} = \max(Z_1, \max(\frac{Y}{c}, \frac{dZ_2}{c})).$$

Observe that the random variables $Z_1, \max(\frac{X}{a}, \frac{bZ_2}{a})$, and $\max(\frac{Y}{c}, \frac{dZ_2}{c})$ are not independent and Theorem 2.1 is not applicable. In particular, the joint distribution of $(U/a, V/c)$ may not determine the distributions of X, Z_1 and Z_2.

3 Identifiability by maxima for max-independent random variables

Definition: A finite collection of random variables $X_1, \ldots, X_n$ is said to be *max-independent* if there exists a function $\beta(x_1, \ldots, x_n)$ such that

$$F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)\beta(x_1, \ldots, x_n), x_i \in R, 1 \leq i \leq n$$

where $F(x_1, \ldots, x_n)$ is the joint distribution of $(X_1, \ldots, X_n)$, $F_i(x)$ is the distribution function of X_i for $1 \leq i \leq n$ and $\beta(x_1, \ldots, x_n)$ is a function taking values in the interval $(0, 1]$ such that $\beta(x_1, \ldots, x_n) \to 1$ if $x_i \to \infty$ for some $i, 1 \leq i \leq n$ (cf. Prakasa Rao (2023)). The function $\beta(x_1, \ldots, x_n)$ is called the *generator* of the random sequence $X_1, \ldots, X_n$.

Examples of sequence of max-independent random variables which are not independent are given in Prakasa Rao (2023). We will now generalize Theorem 2.2 to max-independent random variables.

Theorem 3.1: Let X, Y, Z_1 and Z_2 be max-independent random variables with generator $\beta(x_1, x_2, x_3, x_4)$. Define the random variables U and V by the relations

$$U = \max(X, aZ_1, bZ_2);\ V = \max(Y, cZ_1, dZ_2)$$

where Z_1, Z_2 are identically distributed and a, b, c, d are known positive constants. Further suppose that the distribution functions of X, Y, Z_1 have the same support $R = (-\infty, \infty)$.. Then, the joint distribution function of (U, V) uniquely determines the distributions of X, Y, Z_1 if $a = b$ or if $a \neq b$ but the distribution function of Z_1 is differentiable and the derivative is continuous.

Proof: Let $G(t_1, t_2)$ be the joint distribution function of (U, V). Suppose that (N, S_1, S_2) is another set of max-independent random variables with the generator $\beta(x_1, x_2, x_3, x_4)$ such that the joint distribution of the random vector

$$(\max(N, aS_1, bS_2), \max(M, cS_1, dS_2))$$

is the same as that of the random vector (U, V). Let $F_X(.)$ denote the distribution function of the random variable X. It can be seen that, for any$-\infty < t_1, t_2 < \infty$,

$$\begin{aligned} G(t_1,t_2) &= P(\max(X, aZ_1, bZ_2) \leq t_1, \max((Y, cZ_1, dZ_2) \leq t_2) \\ &= P(X \leq t_1, aZ_1 \leq t_1, bZ_2 \leq t_2; Y \leq t_2, cZ_1 \leq t_2, dZ_2 \leq t_2) \\ &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \frac{t_1}{a}, Z_1 \leq \frac{t_2}{c}, Z_2 \leq \frac{t_1}{b}, Z_2 \leq \frac{t_2}{d}) \\ &= P(X \leq t_1, Y \leq t_2, Z_1 \leq \min(\frac{t_1}{a}, \frac{t_2}{c}), Z_2 \leq \min(\frac{t_1}{b}, \frac{t_2}{d})) \\ &= F_X(t_1)F_Y(t_2)F_{Z_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{Z_1}(\min(\frac{t_1}{b}, \frac{t_2}{d}))\beta(t_1, t_2, \min(\frac{t_1}{a}, \frac{t_2}{c}), \min(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned} \tag{3. 1}$$

since $X, Y Z_1, Z_2$ are max-independent random variables,Z_1, Z_2 are identically distributed with generator $\beta(x_1, x_2, x_3, x_4)$. By a similar argument it follows that

$$\begin{aligned} G(t_1,t_2) &= P(\max(N, aS_1, bS_2) \leq t_1, \max((M, cS_1, dS_2) \leq t_2)) \\ &= F_N(t_1)F_M(t_2)F_{S_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{S_1}(\min(\frac{t_1}{b}, \frac{t_2}{d}))\beta(t_1, t_2, \min(\frac{t_1}{a}, \frac{t_2}{c}), \min(\frac{t_1}{b}, \frac{t_2}{d})). \end{aligned} \tag{3. 2}$$

Hence

$$\begin{aligned} & F_X(t_1)F_Y(t_2)F_{Z_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{Z_1}(\min(\frac{t_1}{b}, \frac{t_2}{d}))\beta(t_1, t_2, \min(\frac{t_1}{a}, \frac{t_2}{c}), \min(\frac{t_1}{b}, \frac{t_2}{d})) \\ &= F_N(t_1)F_M(t_2)F_{S_1}(\min(\frac{t_1}{a}, \frac{t_2}{c}))F_{S_1}(\min(\frac{t_1}{b}, \frac{t_2}{d}))\beta(t_1, t_2, \min(\frac{t_1}{a}, \frac{t_2}{c}), \min(\frac{t_1}{b}, \frac{t_2}{d})) \end{aligned} \tag{3. 3}$$

for all $-\infty < t_1, t_2 < \infty$. Then, for all $t \in R$,, define

$$\eta_1(t) = \frac{F_X(t)}{F_M(t)}, \eta_2(t) = \frac{F_Y(t)}{F_N(t)}, \text{and } \eta_3(t) = \frac{F_{Z_1}(t)}{F_{S_1}(t)}.$$

Equation (3.3) implies that

$$\eta_1(t_1)\eta_2(t_2)\eta_3(\min(\frac{t_1}{a}, \frac{t_2}{c}))\eta_3(\min(\frac{t_1}{b}, \frac{t_2}{d})) = 1, t_1, t_2 \in R. \tag{3. 4}$$

Let $t_1 \to \infty$ in equation (3.4). From the properties of the distribution functions, it follows that

$$\eta_2(t_2)\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d}) = 1, t_2 \in R.$$

Hence

$$\eta_2(t_2) = [\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d})]^{-1}, t_2 \in R. \tag{3. 5}$$

Let $t_2 \to \infty$ in equation (3.4). From the properties of the distribution functions again, it follows that

$$\eta_1(t_1)\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b}) = 1, t_1 \in R$$

which implies that

$$\eta_1(t_1) = [\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b})]^{-1}, t_2 \in R. \tag{3. 6}$$

Combining the equations (3.4)-(3.6), it follows that

$$\eta_3(\min(\frac{t_1}{a}, \frac{t_2}{c}))\eta_3(\min(\frac{t_1}{b}, \frac{t_2}{d})) = \eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b})\eta_3(\frac{t_2}{c})\eta_3(\frac{t_2}{d}), t_1, t_2 \in R. \tag{3. 7}$$

Fix $t_1 \in R$ and let $t_2 \to \infty$. Then the expression on the left side of equation (3.6) tends to 1 and the expression on the right side tends to

$$\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b}).$$

Hence

$$\eta_3(\frac{t_1}{a})\eta_3(\frac{t_1}{b}) = 1, t_1 \in R. \tag{3. 8}$$

This in turn implies that

$$\eta_1(t_1) = 1, t_1 \in R. \tag{3. 9}$$

A similar analysis shows that

$$\eta_2(t_2) = 1, t_2 \in R \tag{3. 10}$$

by fixing $t_2 \in R$ and letting $t_1 \to \infty$. In particular, it follows that

$$F_X(t) = F_M(t), t \in R$$

and

$$F_Y(t) = F_N(t), t \in R.$$

Furthermore, equation (3.8) implies that

$$\log \eta_3(\frac{t}{a}) + \log \eta_3(\frac{t}{b}) = 1, t \in R. \tag{3. 11}$$

Let $\zeta(t) = \log \eta_3(t), t \in R$. Equation (3.11) implies that

$$\zeta(\frac{t}{a}) + \zeta(t)(\frac{t}{b}) = 0, t \in R$$

or equivalently

$$\zeta(u) = -\zeta(\lambda u), u \in R$$

where $\lambda = \frac{a}{b}$. Suppose that $\lambda = 1$. Then it follows that $\zeta(u) = 0, u \in R$ which in turn implies that $\eta_3(t) = 1, t \in R$ and hence $F_{Z_1}(t) = F_{S_1}(t), t \in R$. If $\lambda \neq 1$, applying Lemma 2 in Li and Zheng (2019), it follows that $\zeta(u) = 0, u \in D^*$ under the additional condition of differentiability of the function $\zeta(u)$ and continuity of its derivative which implies that $F_{Z_1}(t) = F_{S_1}(t), t \in R$.

This completes the proof of Theorem 3.1.

Remarks: Theorem 3.1 will continue to hold if the support R is replaced by the support $[0, \infty)$.

Acknowledgment: This work was supported under the scheme “INSA Honorary Scientist" at the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India.

References

Klebanov, L. 1973. Reconstructing the distributions of the components of a random vector from distributions of certain statistics, *Mathematical Notes*, 13:71-72.

Kotlarski, I. 1978. On some characterization in probability by using minima and maxima of random variables, *Aequationes Mathematicae*, 17:77-82.

Li, Siran., and Zheng, Xunjie. 2019. A generalization of Lemma 1 in Kotlarski (1967). Preprint.

Prakasa Rao, B.L.S. 1992. *Identifiability in Stochastic Models: Characterization of Probability Distributions*. New York: Academic Press.

Prakasa Rao , B.L.S. 2023. On some characterizations of probability distributions based on maxima or minima of some families of dependent random variables, *Jour. Indian Soc. Probab. and Statist.*, https://doi.org/10.10007/s41096-024-00180-1.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

On the Existence and Non-existence of Robust Optimal Covariate Designs in the Presence of Neighbor Effects: RBDs

Sobita Sapam [1] and Bikas Kumar Sinha [2]

1. *Dhanamanjuri University, Imphal, Manipur*
2. *Indian Statistical Institute, Kolkata*

Received: Feb 2023 / Accepted: July 2024

Abstract

There has been a steady flow of research publications in the narrow area of 'optimal covariate designs'. The context deals with quantitative non-stochastic covariates - each lying in the closed interval $[-1, 1]$. Starting with an entirely homogeneous set-up of experimental units, the literature extends into many directions - covering differential treatment effects, block effects, row-and column-effects and so on. However, the central problem has been - most efficient estimation of covariate effects parameters - in the presence of so many nuisance parameters. It is no wonder that Springer Monograph [**Das-2015**] is available with summarization of results on existence and constructional aspects of such optimal covariate designs in diverse experimental settings.

An entirely new set-up has surfaced up in the context of 'neighboring designs' while studying such optimality issues. In this paper we will focus on that and primarily we will supplement some results in [**Sapam-2021**] while dealing with RBDs with neighbor effects.

Keywords: Designs with Neighbor-Effects, RBDs, LSDs, Optimal designs for covariates effects.

1 Introduction

The key references to this paper are (i) Springer Monograph on Optimal Covariate Designs [**Das-2015**] and (ii) a recent article [**Sapam-2021**]. There are a good number of papers in this fascinating topic. As far as we can trace it out, [**Lopes-1982a**], [**Lopes-1982b**] discussed about optimal covariate designs in CRD set-ups. Twenty years later, [**Das-2003**] introduced a fundamental resourceful matrix W in the CRD set-up that simplified the whole approach so much so that the formulations and solutions in the cases of RBDs and LSDs were nicely visualized. Subsequently, there was a continuous flow of research articles in this direction. Quite appropriately, a research monograph was in sight and it was duly published [**Das-2015**].

The article by [**Sapam-2021**] focuses on such covariate designs when Neighbor-Effects are incorporated in the models. While discussing generalizations of results in the presence of Row-Neighbor and Column-Neighbor Effects, the authors hinted at a question in "Remark 3: [It is tempting to conjecture that for any given layout of an RBD, there is at least one X-matrix available satisfying all the properties stipulated]". That is our starting point. We focus our attention on that query.

It might be appropriate to briefly discuss the notion of neighbor designs. Block designs with one-sided, two-sided and four-sided neighbor effects have been studied in considerable details in the literature. Some of the relevant references are [**Varghese-2014**], [**Azais-1993**], [**Jaggi-2003**], [**Jaggi-2018**], [**Sapam-2019a**], [**Sapam-2019b**]. As desired by one referee, we describe the concept of neighbor designs with reference to an RBD with $b = 4, v = 5$ shown below:

	4	5	1	2	3	
5	1	2	3	4	5	1
1	2	3	4	5	1	2
2	3	4	5	1	2	3
3	4	5	1	2	3	4
	1	2	3	4	5	

Table 1: RBD (b=4, v=5)

The blocks (rows) are circular that is, the border plot of treatment 1 at the left end is neighbor of the treatment 5 at the right end of the block 1. In the same manner, the treatment 1 at the top left-corner is neighbor of treatment 4 at the bottom- left corner of Column 1, that means circular column-wise. In the experimental situations where a treatment has neighbor effects from its left and right adjacent plots it is known as two- sided neighbor effects whereas if a treatment has neighbor effects from the adjacent left -and -right plot as well as from the adjacent top and bottom plots, it is said to be four-sided neighbor effects. In the above RBD in Table 1, every adjacent pair of distinct treatments has the concurrence μ, say, equal to 4. It defines cirular balanced under the assumption of circular blocks rows. [**Bailey-2003**] gives the details of circular neighbor balanced designs under different block sizes. We will deal with such neighbor-effects designs in this study of optimal estimation of covariate effects.

Not to obscure the essential steps of reasoning and, moreover, to ease out the readers' understanding about the subject matter, we will discuss some basics in optimal covariates' designs.

(1) Basically, we are dealing with multivariate linear regression models involving, say, k regressors. Specifically, in its simple form, the model envisaged is :

$$y_{ij} = \mu + \sum_{j=1}^{k} \beta_j x_{ij} + e_{ij}, i = 1, 2, ..., n. \tag{1}$$

We are thus referring to a basic model with k covariates and the beta-coefficients need to be estimated most efficiently, assuming that $[-1 \leq x_{ij} \leq 1]$.

(2) On this model, we gradually impose treatment variations [CRD Model], block [row] variations [RBD/BIBD etc], row-column [LSD/MOLS] variations and so on.

For the simplest model in (1), regression parameters are most efficiently estimated when the elements $(x_{ij})'s$ satisfy the conditions of mutual orthogonality involving the elements 1 and -1. [**Lopes-1982a**], [**Lopes-1982b**] studied the nature of optimal covariate designs under CRD set-up. It is in this context that [**Das-2003**] introduced the W-matrix which greatly simplified the understanding of the optimal designs for most of the design set-ups. In a CRD model with $n = vr$ involving v treatments, each with r replications, the matrix W refers to a matrix of order $v \times r$ whose elements are $x'_{ij}s$.

The study of optimal covariate designs rests on extensive uses of Hadamard Matrices and hence the orders of the RBDs and LSDs are conveniently chosen as multiples of 4. Also optimality rests on the ability to construct W-matrices using only the elements $+1, -1$ and satisfying a series of conditions listed below. For an expository review on Hadamard Matrices, we refer to [**Hedayat-1978**].

(i) each row sum of w-values (as block totals) is 0;

(ii) each treatment position sum of w- values is 0;

(iii) sum of w- values in the positions of all the left-neighbors of
a given treatment is 0 and that holds for each treatment;

(iv) sum of w- values in the positions of all the right-neighbors of a given treatment is 0 and that holds for each treatment.

In the literature, a few more relevant conditions are also stated for meeting other desirable properties. For the time being, we are focusing on the existence of only one W-matrix for a given RBD design layout. When the elements of the W-matrix satisfy the conditions laid down in (iii) and (iv) above, we refer to such a matrix as being 'robust'. The sense of robustness is with respect to the presence of neighbor effects. In the sequel, we will use X and W interchangeably, without making any distinction.

Our aim in this paper is to examine the conjecture stated in [**Sapam-2021**]. It turns out that for any given set of values of b and v, there are at least two distinct RBD design layouts having underlying X-matrices in the two opposite directions ! To start with, we study a few special cases in Sections 2-3-4, before treating the general case in Section 5.

2 RBDs with v = b = 4

In an RBD with v=b=4, there are $(4!)^4$ choices of design layouts - though plenty of them are 'permutation invariant' in subsets. We consider one subset of them viz., 4!=24 RBDs starting with the first three blocks in the natural order of the treatments i.e., treatment-levels $1, 2, 3, 4$. Then only the last row [i.e., 4^{th} row] is made to be composed of a permutation of the treatment levels $1, 2, 3, 4$. This results into $4! = 24$ distinct design layouts. Our objective is to check the existence and non-existence of X-matrices satisfying the conditions $(i) - (iv)$ laid down above. We claim that there are two distinct design layouts among these 24 choices whose X-matrices are in opposite directions. For one of them, we can construct an X-matrix satisfying all the conditions. For another, we argue that there does not exist any such X-matrix satisfying all the conditions laid down above.

We start with the first choice of RBD (v=b=4) listed below in Table 2.

1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

Table 2: RBD (v=b=4)

Clearly, in this choice we can find out the following X-matrix in Table 3 satisfying all the four conditions given above.

1	-1	1	-1
-1	1	-1	1
1	-1	1	-1
-1	1	-1	1

Table 3: X-matrix of RBD in Table 1

Next we make another choice of the design layout shown in Table 4 below.

For this choice we argue that there does not exist any underlying X-matrix satisfying all the conditions mentioned above. In this context, we consider the patterns of left-neighbor (LN) and right-neighbor (RN) effects of the treatments in the below table [Table 5].

Given the above, we will now argue closely towards non-existence of an underlying X-matrix satisfying the stated conditions. For convenience, we will use the following notations for x-values in this particular case of $b = v = 4$, as shown in $Table\ 6$ below.

1	2	3	4
1	2	3	4
1	2	3	4
2	1	3	4

Table 4: The second choice of RBD (v=b=4)

Treatments	LN	RN
1	4 4 4 2	2 2 2 3
2	1 1 1 4	3 3 3 1
3	2 2 2 1	4 4 4 4
4	3 3 3 3	1 1 1 2

Table 5: LN and RN treatments of the RBD in Table 3

We now analyse the x-values in Table 6 with reference to the conditions stated above, particularly, (iii) and (iv). We note in passing that each x-value is confined to $+1/-1$ only. Moreover, $\sum p_i = \sum q_i = \sum r_i = \sum s_i = 0$ is a necessary condition to be satisfied by these elements. Meaningful conditions from (iii) and (iv) suggest :

$$q_1 + q_2 + q_3 + r_4 = 0 \tag{2}$$

$$p_1 + p_2 + p_3 + s_4 = 0 \tag{3}$$

$$q_1 + q_2 + q_3 + p_4 = 0 \tag{4}$$

$$p_1 + p_2 + p_3 + q_4 = 0 \tag{5}$$

Therefore, it turns out that $p_4 = r_4; q_4 = s_4$. WOLG, we set $p_4 = r_4 = +1$ and $q_4 = s_4 = -1$. From the above, we then infer that

$p_1 + p_2 + p_3 = +1;\ q_1 + q_2 + q_3 = -1.$

And these now suggest for Treatment level 1 and Treatment level 2 :

$p_1 + p_2 + p_3 + p_4 = +2; q_1 + q_2 + q_3 + q_4 = -2.$

Thus, eventually, we run into a contradiction regarding the choice of the x-values. This establishes the points we are making.

Remark 1

It would be interesting to examine if any of the other 4! design layouts also exhibit the non existence feature(s) of the underlying X- matrices! Surprisingly, indeed there are quite a few of them [16 in number] in the non-existence category. All these 16 design layouts are shown in Table 7. The remaining 8 design layouts [including the RBD shown above] exhibit affirmative solutions i.e., each one possesses an X-matrix satisfying the necessary conditions. The design layouts as also the solutions are shown in Table 8.

$1(p_1)$	$2(q_1)$	$3(r_1)$	$4(s_1)$
$1(p_2)$	$2(q_2)$	$3(r_2)$	$4(s_2)$
$1(p_3)$	$2(q_3)$	$3(r_3)$	$4(s_3)$
$2(q_4)$	$1(p_4)$	$3(r_4)$	$4(s_4)$

Table 6: x-values for the second choice of RBD (v=b=4)

Sl. no.	RBDs	Sl.no.	RBDs
1.	1 2 3 4 1 2 3 4 1 2 3 4 1 2 4 3	9.	1 2 3 4 1 2 3 4 1 2 3 4 3 1 2 4
2.	1 2 3 4 1 2 3 4 1 2 3 4 1 3 2 4	10.	1 2 3 4 1 2 3 4 1 2 3 4 3 1 4 2
3.	1 2 3 4 1 2 3 4 1 2 3 4 1 3 4 2	11.	1 2 3 4 1 2 3 4 1 2 3 4 3 2 4 1
4.	1 2 3 4 1 2 3 4 1 2 3 4 1 4 2 3	12.	1 2 3 4 1 2 3 4 1 2 3 4 3 4 2 1
5.	1 2 3 4 1 2 3 4 1 2 3 4 2 1 3 4	13.	1 2 3 4 1 2 3 4 1 2 3 4 4 1 3 2
6.	1 2 3 4 1 2 3 4 1 2 3 4 2 3 1 4	14.	1 2 3 4 1 2 3 4 1 2 3 4 4 2 1 3
7.	1 2 3 4 1 2 3 4 1 2 3 4 2 4 1 3	15.	1 2 3 4 1 2 3 4 1 2 3 4 4 2 3 1
8.	1 2 3 4 1 2 3 4 1 2 3 4 2 4 3 1	16.	1 2 3 4 1 2 3 4 1 2 3 4 4 3 1 2

Table 7: RBDs (v=b=4) having no X-matrices

SI no.	RBDs	corresponding X-matrix
1.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
2.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	1 4 3 2	-1 1 -1 1
3.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	2 1 4 3	1 -1 1 -1
4.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	2 3 4 1	1 -1 1 -1
5.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	3 2 1 4	-1 1 -1 1
6.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	3 4 1 2	-1 1 -1 1
7.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	4 1 2 3	1 -1 1 -1
8.	1 2 3 4	1 -1 1 -1
	1 2 3 4	-1 1 -1 1
	1 2 3 4	1 -1 1 -1
	4 3 2 1	1 -1 1 -1

Table 8: RBDs (v=b=4) having X- matrices

3 RBDs with v=b=6

In the study on existence of optimal X- matrices, we generally ask for much more than one such matrix. Consequently, it is necessary that $b = v = 0(mod4)$.

However in this query about existence/non existence of just one matrix, we may deal with $b = v = 0(mod2)$.

We now consider two RBDs with v=b=6 such that one has the X- matrix satisfying the four conditions mentioned above and we argue that for the other design, no X-matrix is available. The layout of the two RBDs are shown in the Tables 9 and 10 and the corresponding X-matrix of the design in Table 9 is exhibited in Table 11.

1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6

Table 9: RBD (v=b=6): first choice

1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
2	1	3	4	5	6

Table 10: RBD (v=b=6): second choice

1	-1	1	-1	1	-1
-1	1	-1	1	-1	1
1	-1	1	-1	1	-1
-1	1	-1	1	-1	1
1	-1	1	-1	1	-1
-1	1	-1	1	-1	1

Table 11: X-matrix of RBD (v=b=6) in Table 9

Remark 2

We claim that for the RBD in Table 10, there does not exist any X-matrix satisfying the stated conditions. The details of our arguments are developed below. As in the case of the design layout with $b = v = 4$, we use letter symbols for the elements of an X-matrix.

We analyse the conditions (i) - (iv) stated above with reference to the x- values of Table 12. $\sum a_i = \sum b_i = \sum c_i = \sum d_i = \sum e_i = \sum f_i = 0$ for all i = 1,...,6 is a necessary condition to be satisfied

$1(a_1)$	$2(b_1)$	$3(c_1)$	$4(d_1)$	$5(e_1)$	$6(f_1)$
$1(a_2)$	$2(b_2)$	$3(c_2)$	$4(d_2)$	$5(e_2)$	$6(f_2)$
$1(a_3)$	$2(b_3)$	$3(c_3)$	$4(d_3)$	$5(e_3)$	$6(f_3)$
$1(a_4)$	$2(b_4)$	$3(c_4)$	$4(d_4)$	$5(e_4)$	$6(f_4)$
$1(a_5)$	$2(b_5)$	$3(c_5)$	$4(d_5)$	$5(e_5)$	$6(f_5)$
$2(b_6)$	$1(a_6)$	$3(c_6)$	$4(d_6)$	$5(e_6)$	$6(f_6)$

Table 12: x-values for the second choice of RBD (v=b=6)

by the these x- values. Moreover, from the conditions (iii) and (iv), we derive:

$$f_1 + f_2 + f_3 + f_4 + f_5 + b_6 = 0 \tag{6}$$

$$f_1 + f_2 + f_3 + f_4 + f_5 + f_6 = 0 \tag{7}$$

$$b_1 + b_2 + b_3 + b_4 + b_5 + c_6 = 0 \tag{8}$$

$$b_1 + b_2 + b_3 + b_4 + b_5 + a_6 = 0 \tag{9}$$

$$c_1 + c_2 + c_3 + c_4 + c_5 + c_6 = 0 \tag{10}$$

$$d_1 + d_2 + d_3 + d_4 + d_5 + d_6 = 0 \tag{11}$$

Therefore we can claim that $b_6 = f_6$ and $a_6 = c_6$. WOLG we set $a_6 = c_6 =$ +1 and $b_6 = f_6 =$ -1. Now from the above relations (6) to (11) we then infer that
$f_1 + f_2 + f_3 + f_4 + f_5 =$ +1;
$b_1 + b_2 + b_3 + b_4 + b_5 =$ -1. In addition to this, the subtotal of the x- values of the first five terms of treatment 4 is then,
$(d_1 + d_2 + d_3 + d_4 + d_5) =$ +1 and $d_6 =$ -1
or $(d_1 + d_2 + d_3 + d_4 + d_5) =$ -1 and $d_6 =$ +1
In the same way, for the treatment 5, either
$(e_1 + e_2 + e_3 + e_4 + e_5) =$ +1 and $e_6 =$ -1
or $(e_1 + e_2 + e_3 + e_4 + e_5) =$ -1 and $e_6 =$ +1
should be satisfied.
Then these suggest for the Treatment level 1 and the Treatment level 2:
$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 = +2$
and $b_1 + b_2 + b_3 + b_4 + b_5 + b_6 = -2$
Thus eventually, we run into a contradiction regarding the choice of x- values.

4 RBD (v=4, b=6)

We further consider RBDs where the number of treatments and number of blocks are unequal with RBD (v=4, b=6). The following two RBDs shown in Tables 13 and 14 are taken up for dealing with the existence / non existence of X-matrix as in the above cases. For the RBD in Table 13, we have existence result and it is shown in Table 15.

Remark 3

For the RBD in Table 14, there does not exist any X-matrix satisfying the stated conditions, and our detailed arguments are developed below.

The arguments are very similar to those for RBD (v=4, b=4). Now the x-values for Table 14 are denoted by p_i, q_i, r_i, s_i, for i=1,2,...,6 as shown in Table 16.
We analysed the conditions (i)- (iv) mentioned above with reference to the x-values of RBD in Table 14.

1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

Table 13: The first choice of RBD (v=4, b=6)

1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
2	1	3	4

Table 14: The second choice of RBD (v=4, b=6)

$\sum p_i = \sum q_i = \sum r_i = \sum s_i = 0$ is a necessary condition; further, conditions (iii) and (iv) also suggest that

$$s_1 + s_2 + s_3 + s_4 + s_5 + q_6 = 0$$
$$s_1 + s_2 + s_3 + s_4 + s_5 + s_6 = 0$$
$$q_1 + q_2 + q_3 + q_4 + q_5 + p_6 = 0$$
$$q_1 + q_2 + q_3 + q_4 + q_5 + r_6 = 0$$

which implies that $p_6 = r_6$ and q_6= s_6. WOLG we set $p_6 = r_6$ =+1 and $q_6 = s_6 =$ -1; then we can infer that $s_1 + s_2 + s_3 + s_4 + s_5 = +1$;
$q_1 + q_2 + q_3 + q_4 + q_5 = -1$
Now for Treatment level 1 and Treatment level 2, the total of x-values are:
$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = +2$ and
$q_1 + q_2 + q_3 + q_4 + q_5 + q_6 = -2$ respectively,
which is a contradiction to the choice of x-values stated in the conditions.

Remark 4

The choice of the design layout in Table 14 suggests a non-existence result in the framework of $b = 6, v = 4$. As in the case of an RBD with $b = v = 4$, here also we have 24 possible RBD layouts by variations of

1	-1	1	-1
-1	1	-1	1
1	-1	1	-1
-1	1	-1	1
1	-1	1	-1
-1	1	-1	1

Table 15: X-matrix of RBD (v=4, b=6) in Table 13

$1(p_1)$	$2(q_1)$	$3(r_1)$	$4(s_1)$
$1(p_2)$	$2(q_2)$	$3(r_2)$	$4(s_2)$
$1(p_3)$	$2(q_3)$	$3(r_3)$	$4(s_3)$
$1(p_4)$	$2(q_4)$	$3(r_4)$	$4(s_4)$
$1(p_5)$	$2(q_5)$	$3(r_5)$	$4(s_5)$
$2(q_6)$	$1(p_6)$	$3(r_6)$	$4(s_6)$

Table 16: x-values for the RBD (v=4, b=6) of Table 14

the treatment allocations in the last block. It would be interesting to examine how many of these layouts result in non-existence!

5 RBD(b = 2p, v = 2q, p and q being positive integers)

Having understood the existence/non-existence results in the particular frameworks of selected values of b and v, we now venture into the general case. Towards the non-existence result, we start with the $RBD(b, v)$ design layout as

$$D_{b\times v} = \begin{bmatrix} 1 & 2 & 3 & \cdots & v \\ 1 & 2 & 3 & \cdots & v \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & 2 & 3 & \cdots & v \\ 2 & 1 & 3 & \cdots & v \end{bmatrix}$$

Further, let

$$X_{b\times v} = \left[\ ((x_{ij}))\ \right]$$

be the usual matrix of associated covariate-values.

This suggests : x_{ij} = covariate value at (i, j) position, $1 \leq i \leq b, 1 \leq j \leq v$. Note that we necessarily require $\sum_{i=1}^{b} x_{ij}$=0 for each j, $1 \leq j \leq v$.

Further, for orthogonality of left neighbor (LN) and right neighbor (RN) effects, we need additional conditions as stipulated in (iii) - (iv) in Section 1.

For j = 1, 2, 3 and v, we derive the conditions:

(i) RN of 1 implies $\sum_i x_{i2} - x_{b2} + x_{b3}$=0

(ii) LN of 2 implies $\sum_i x_{i1} - x_{b1} + x_{bv}$=0

(iii) LN of 3 implies $\sum_i x_{i2} - x_{b2} + x_{b1}$=0

(iv) RN of v implies $\sum_i x_{i1} - x_{b1} + x_{b2}$=0.

Now from these (i) to (iv) conditions, we can see that $x_{b1} = x_{b3}$ and $x_{b2} = x_{bv}$.

WOLG we may set $x_{b1} = x_{b3}$= +1 and $x_{b2} = x_{bv}$= -1.

Then clearly we deduce from (iii) and (iv) above that $\sum_i x_{i2} = x_{b2} - x_{b1}$= - 2 and $\sum_i x_{i1} = x_{b1} - x_{b2}$= +2 respectively. And these are both contradictions to our stated conditions.

Towards existence result, as in the particular cases, we may start with the standard RBD for treatment allocations in the natural order for each block and follow the allocations of the x-values as $+1$'s and -1's alternately as are shown in the particular cases. In matrix notation, in this case, the solution matrix is represented as the matrix product $H_2 \otimes J_{p\times q}$ which is a succession of the Hadamard matrix of order 2 i.e., of H_2 matrix in every row and column covering the matrix J of order $p \times q$.

$$H_2 \otimes J_{p\times q} = \begin{bmatrix} H_2 & H_2 & \cdots & H_2 \\ H_2 & H_2 & \cdots & H_2 \\ \vdots & \vdots & \ddots & \vdots \\ H_2 & H_2 & \cdots & H_2 \\ H_2 & H_2 & \cdots & H_2 \end{bmatrix}$$

6 Concluding remark

The present study was meant to examine robustness [with respect to presence of neighbor effects] of optimal covariates designs in RBD set-ups. The query was raised in [**Sapam-2021**]. It transpires from the current study that the choice of the RBD layout is very crucial for the robustness to hold. Further studies wrt higher order design layouts such as LSDs, MOLSDs etc may be revealing as well.

Acknowledgements

The first author acknowledges financial support from Women Scientist Project No. **DST/WOS-A/PM-27/2021, DST, New Delhi**. She also thanks her Mentor Professor KK Singh Meitei for providing all facilities towards successfully pursuing her research in the broad area of DoE and also for arranging academic visits of Professor Sinha to the Manipur University for collaborative research.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Bivariate Dynamic Cumulative Past Entropy in the Context of Weighted Models

Rohini, S. Nair [1], E. I. Abdul Sathar [2]

1. *Assistant Statistician, Cost of Cultivation Scheme, University of Kerala, Thiruvananthapuram-695 581, India.*
2. *Professor, Department of Statistics, University of Kerala, Thiruvananthapuram-695 581, India.*

Received: Sept 2023 / Accepted: May 2024

Abstract

The cumulative past entropy (CPE), introduced by [5] is viewed as a measure of uncertainty. [9] proposed the bivariate of form DCPE, namely bivariate dynamic cumulative past entropy (BDCPE) and discussed some of its properties. In this article, we introduced the weighted form of BDCPE and study some of its properties. We also look into the problem of extending weighted CPE for conditionally specified models. We derived some features of conditional distributions. It is shown that the proposed measure uniquely determines the distribution function, together with certain characterization results. Additionally, we suggested a non-parametric estimator for the newly developed measure.

Keywords: Life distributions, Shannon entropy, Past lifetime, Vector valued hazard rate, Bivariate distribution.

1 Introduction

An important measure of uncertainty associated with a random variable X is the notion of entropy, introduced by [17]. If X is a non-negative random variable having an absolutely continuous distribution function *F(x)* with probability density function *f(x)*, then the Shannon's entropy is defined as

$$H(X) = -\int_{0}^{\infty} f(x)\log f(x)dx. \tag{1}$$

$H(X)$ measures expected uncertainty in *f(x)* about the predictability of an outcome of X.

Even though Shannon's entropy finds applications in many areas of research, [16] identified some limitations of the use of (1) in measuring the randomness of certain systems and thereby to overcome those limitations, introduced an alternative measure of uncertainty that extends Shannon entropy to random variables with continuous distributions called cumulative residual entropy (CRE), which relates to uncertainty on the future lifetime of a system.

Motivated by the salient features of CRE, [5] proposed a dual concept of CRE called cumulative past entropy (CPE) and is defined as

$$\bar{\varepsilon}(X) = -\int_0^{\infty} F(x)\log F(x)\ dx. \tag{2}$$

The entropy (2) measures the uncertainty about the inactivity time of X.

[5] introduced the concept of CPE for past lifetime called dynamic CPE (DCPE). For a non-negative random variable X representing the lifetime of a component, the dynamic cumulative past entropy (DCPE) is the CPE associated with the random variable $X|X<t$, and is defined as

$$\bar{\varepsilon}(X;t) = -\int_0^t \frac{F(x)}{F(t)}\log\frac{F(x)}{F(t)}\ dx,\ t>0. \tag{3}$$

For more properties, applications of (3), one may refer to [1] and [6].

The concept of weighted distributions was introduced by [15] in connection with modelling statistical data, where the usual practice of employing standard distributions is not found appropriate in some cases. However, in certain applied areas, such as reliability theory or mathematical neurobiology, it is desirable to deal with shift-dependent information measures. In recent years, this concept has been applied in many areas of statistics, such as analysis of family size, world life population study, renewal theory, biomedical, statistical ecology, reliability modelling etc.

Associated to a random variable X with probability density function f and to a non-negative real function w, we can define the weighted random variable X^w with density function

$$f^w(x) = \frac{w(x)f(x)}{E(w(X))},$$

where we assume $0 < E(w(X)) < \infty$. When $w(x) = x$, X^w is called the length (or size) biased random variable.

[3] have introduced the concept of weighted entropy through

$$H^w(X) \quad = \quad -\int\limits_0^\infty x f(x) \log f(x) dx \tag{4}$$

The factor x, inside the integral on the right hand side of (4) represents a weight linearly emphasizing the occurrence of the event $\{X = x\}$. This yields a length biased shift dependent information measure assigning greater importance to larger values of X. For more properties and applications of weighted entropy, we refer to [4], [10] and [7].

Recently [9] have considered extension of CPE to bivariate setup, namely bivariate cumulative past entropy (BCPE) and is given by

$$\bar{\varepsilon}(X_1, X_2) = -\int_0^{b_1}\int_0^{b_2} F(x_1, x_2) \log F(x_1, x_2) \; dx_2 dx_1. \tag{5}$$

They have also studied BCPE for past lifetime called bivariate dynamic CPE (BDCPE), defined as

$$\bar{\varepsilon}(t_1, t_2) = -\int_0^{t_1}\int_0^{t_2} \frac{F(x_1, x_2)}{F(t_1, t_2)} \log \frac{F(x_1, x_2)}{F(t_1, t_2)} \; dx_2 dx_1. \tag{6}$$

In this paper, we consider the weighted form of BDCPE namely bivariate dynamic weighted CPE (BDWCPE) and study its various properties. The rest of the chapter is organized as follows. Section 2 includes the definition and basic properties of the BDWCPE. In section 3, we consider the behaviour of dynamic weighted CPE for conditional distributions. In section 3.1, we consider the behaviour of conditional dynamic weighted CPE for $X_j < t_j$. Section 3.2 deals with the behaviour of conditional dynamic weighted CPE for $X_j = t_j$. In section 4, a non-parametric estimator is suggested for the conditional dynamic weighted CPE and in section 4.1, we applied the estimator to a real data set.

2 Bivariate Weighted Cumulative Past Entropy

In studying the reliability aspects of multi-component system with each component having a lifetime depending on the lifetimes of the other components, multivariate life distributions are employed. Reliability characteristics in the univariate case can be extended to the corresponding multivariate version. Even though a lots of interest has been evoked on the entropy of residual and past lifetime in the univariate case, only a few works seem to have been done in higher dimensions. In a recent work, [14] have considered extension of DCRE to bivariate setup and studied its properties. Several generalizations to the concept of bivariate DCRE can be found in [18] and [14]. In various contexts, the uncertainty is not necessarily related to the future but may refer to the past. It is to be noted that the concepts in past time are more appropriate than those truncated from below when the observations are predominantly from left tail. This shows the relevance and usefulness of studying CPE when uncertainty is related to the past. For more recent works on bivariate information notions and weighted measures one may refer to [11] and [2].

In this section, we introduce the weighted form of BDCPE defined in (6). Since the past lifetime has always a finite support we restrict our attention to random variables with finite supports. Therefore, we assume that the support of (X_1, X_2) is included in $(0, b_1) \times (0, b_2)$ for some non negative real values b_1, b_2. In this section, we consider BCPE and BDCPE defined in (5) and (6) respectively for weighted models.

Definition 1. *Let $X = (X_1, X_2)$ be a bivariate random vector having the distribution function $F(x_1, x_2)$, then we define the bivariate weighted CPE (BWCPE) as*

$$\bar{\mathscr{E}}^w(X_1, X_2) = -\int_0^{b_1}\int_0^{b_2} x_1x_2\ F(x_1, x_2)\log F(x_1, x_2)\ dx_2dx_1, \tag{7}$$

provided the integral on the right hand side is finite.

If X_1 and X_2 are independent, then from (7), we get

$$\bar{\mathscr{E}}^w(X_1, X_2) = \left(\int_0^{b_2} x_2F(x_2)dx_2\right)\bar{\mathscr{E}}^w(X_1) + \left(\int_0^{b_1} x_1F(x_1)dx_1\right)\bar{\mathscr{E}}^w(X_2). \tag{8}$$

The following is an additive property of BWCPE.

The following proposition shows that BWCPE is not invariant under non-singular transformations.

Proposition 1. *Let $Y = (Y_1, Y_2)$ be a non negative bivariate random vector. If $Y_i = \varphi_i(X_i),\ i = 1, 2$ are one to one transformations with $\varphi_i(X_i)$ are differentiable function, then*

$$\bar{\mathscr{E}}^w(Y_1, Y_2) = -\int_0^{b_1}\int_0^{b_2} x_1x_2\ F(x_1, x_2)\log F(x_1, x_2) \mid J \mid dx_2dx_1, \tag{9}$$

where $J = \frac{\partial}{\partial x_1}\varphi_1(x_1)\frac{\partial}{\partial x_2}\varphi_2(x_2)$ is the Jacobian of the transformation.

If $X = (X_1, X_2)$ represents the lifetimes of two components in a system where both the components are found failed at times t_1 and t_2, respectively, then the measure of uncertainty associated with the past lifetimes of the system are important. In the following, we define bivariate dynamic WCPE (BDWCPE).

Definition 2. *For an absolutely continuous non-negative bivariate random vector $X = (X_1, X_2)$ with joint pdf $f(x_1, x_2)$ and distribution function $F(x_1, x_2)$, the bivariate dynamic WCPE (BDWCPE) is defined as*

$$\bar{\mathscr{E}}^w(t_1, t_2) = -\int_0^{t_1}\int_0^{t_2} x_1x_2\frac{F(x_1, x_2)}{F(t_1, t_2)}\log\frac{F(x_1, x_2)}{F(t_1, t_2)}\ dx_2dx_1, \tag{10}$$

which is the two dimensional extension of dynamic WCPE. If X_1 and X_2 are independent, then

$$\bar{\mathscr{E}}^w(t_1, t_2) = m_2^w(t_2)\ \bar{\mathscr{E}}^w(t_1) + m_1^w(t_1)\ \bar{\mathscr{E}}^w(t_2), \tag{11}$$

where $m_i^w(t_i) = \frac{1}{F(t_i)}\int_0^{t_i} x_iF(x_i)\ dx_i$, are the marginal weighted expected inactivity time (WEIT) of the components $X_i,\ i = 1, 2$.

3 Conditional Dynamic Weighted CPE

When we consider bivariate measures, it is necessary that the measurement of uncertainty on the basis of one component is not affected by the missing or unreliable data on the other component, and hence it is necessary to consider component-wise measures.

Conditional Dynamic Weighted CPE for X_i given $X_j < t_j$

In this section, we consider the conditional dynamic weighted CPE for X_i given $X_i < t_j$ and its properties. For a bivariate random vector $X = (X_1, X_2)$, let us consider another set of bivariate random vector $Y = (Y_1, Y_2)$ defined by $Y_i = [X_i | X_1 < t_1, X_2 < t_2],\ i = 1, 2$, which corresponds to the conditional distributions of X_i subject to the condition that the first component failed at any time during $(0, t_1)$ and the second one during $(0, t_2)$. Then the dynamic WCPE for Y, called conditional dynamic WCPE (CDWCPE) is defined as

$$\bar{\mathscr{E}}_1^{*w}(X; t_1, t_2) = -\int_0^{t_1} x_1 \frac{F(x_1, t_2)}{F(t_1, t_2)} \log \frac{F(x_1, t_2)}{F(t_1, t_2)}\, dx_1 \tag{12}$$

and

$$\bar{\mathscr{E}}_2^{*w}(X; t_1, t_2) = -\int_0^{t_2} x_2 \frac{F(t_1, x_2)}{F(t_1, t_2)} \log \frac{F(t_1, x_2)}{F(t_1, t_2)}\, dx_2. \tag{13}$$

In particular if X_1 and X_2 are independent, then $\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) = \bar{\mathscr{E}}_{X_i}^{w}(t_i),\ i = 1, 2$.

In the following, we compute $\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2)$ for some well known distributions.

Example 1. *Consider the bivariate uniform distribution specified by the distribution function*

$$F(t_1, t_2) = t_1^{1+\theta \log t_2}\, t_2; \quad 0 < t_1, t_2 < 1, \quad \theta \leq 0. \tag{14}$$

Straightforward calculations, using (12), give

$$\begin{aligned}\bar{\mathscr{E}}_1^{*w}(X; t_1, t_2) &= \frac{1}{(3 + \theta \log t_2)^2} t_1^2 \left\{-3 + \theta \log t_2 + \log t_1^{1+\theta \log t_2} [-\theta \log t_2 \right.\\ &\quad \left. + \log t_1^{\theta \log t_2} (3 + \theta \log t_2)] t_1^{1+\theta \log t_2}\right\}.\end{aligned}$$

Similarly, using (13) we get

$$\begin{aligned}\bar{\mathscr{E}}_2^{*w}(X; t_1, t_2) &= \frac{1}{(3 + \theta \log t_1)^2} t_2^2 \left\{-3 + \theta \log t_1 + \log t_2^{1+\theta \log t_1} [-\theta \log t_1 \right.\\ &\quad \left. + \log t_2^{\theta \log t_1} (3 + \theta \log t_1)] t_2^{1+\theta \log t_1}\right\}.\end{aligned}$$

Example 2. *Consider the bivariate power distribution specified by the distribution function*

$$F(t_1, t_2) = t_1^{2k_1 - 1 + \theta \log t_2}\, t_2^{2k_2 - 1}; \quad \theta < 0,\ k_1, k_2 > 0,\ 0 < t_1, t_2 < 1. \tag{15}$$

Using (12), we get

$$\begin{aligned}\bar{\mathscr{E}}_1^{*w}(X; t_1, t_2) &= \frac{t_1}{(1 + 2k_1 + \theta \log t_2)^2} \left\{-t_1 (1 + 2k_1 - \theta \log t_2)\right.\\ &\quad + \log(t_1^{2k_1 - 1 + \theta \log t_2}) t_1^{2k_1 + \theta \log t_2} \left[\log t_1^{2k_1 + \theta \log t_2}\right.\\ &\quad + 2(-1 + \log t_1^{2k_1 + \theta \log t_2}) k_1\\ &\quad \left.\left. + (-1 + \log t_1^{2k_1 + \theta \log t_2}) \theta \log t_2\right]\right\}.\end{aligned} \tag{16}$$

Similarly, using (13), we get

$$\begin{aligned}\bar{\mathscr{E}}_2^{*w}(X;t_1,t_2) &= \frac{t_2}{(1+2k_2+\theta\log t_1)^2}\{-t_2(1+2k_2-\theta\log t_1) \\ &\quad + \log(t_2^{2k_2-1+\theta\log t_1})t_2^{2k_2+\theta\log t_1}\Big[\log t_2^{2k_2+\theta\log t_1} \\ &\quad + 2(-1+\log t_2^{2k_2+\theta\log t_1})k_2 \\ &\quad + (-1+\log t_2^{2k_2+\theta\log t_1})\theta\log t_1\Big]\Big\}. \end{aligned} \tag{17}$$

Now we study some characterization results of Conditional Dynamic Weighted CPE.

In the sequel we give the definitions of bivariate reversed hazard rate (BRHR) function and bivariate weighted expected inactivity time (BWEIT).

Definition 3. *For a random vector $X = (X_1, X_2)$ with distribution functions $F(t_1, t_2)$*

(i) the bivariate reversed hazard rate is defined as a vector, $\bar{h}(t_1,t_2) = (\bar{h}_1(t_1,t_2), \bar{h}_2(t_1,t_2))$, where

$$\bar{h}_i(t_i,t_j) = \frac{\partial}{\partial t_i}\log F(t_i,t_j),\ i \neq j = 1,2 \tag{18}$$

are the components of bivariate reversed hazard rate;

(ii) the bivariate weighted EIT is defined by the vector, $\bar{m}^w(t_1,t_2) = (\bar{m}_1^w(t_1,t_2), \bar{m}_2^w(t_1,t_2))$, where

$$\bar{m}_i^w(t_i,t_j) = \frac{1}{F(t_i,t_j)}\int_0^{t_i} x_i F(x_i,t_j)\,dx_i,\ i \neq j = 1,2. \tag{19}$$

which measures the the expected waiting time of the first component conditioned on the fact that both the components were failed before times t_1 and t_2, respectively.

Note that (12) can alternatively written as

$$\bar{\mathscr{E}}_1^{*w}(X;t_1,t_2) = \bar{m}_1^w(t_1,t_2)\log F(t_1,t_2) - \int_0^{t_1} x_1\frac{F(x_1,t_2)}{F(t_1,t_2)}\log F(x_1,t_2)\,dx_1. \tag{20}$$

Similarly, (13) can be written as

$$\bar{\mathscr{E}}_2^{*w}(X;t_1,t_2) = \bar{m}_2^w(t_1,t_2)\log F(t_1,t_2) - \int_0^{t_2} x_2\frac{F(t_1,x_2)}{F(t_1,t_2)}\log F(t_1,x_2)\,dx_2. \tag{21}$$

Differentiating (12) and (13) with respect to t_1 and t_2 respectively, we get in general

$$\frac{\partial}{\partial t_i}\bar{\mathscr{E}}_i^{*w}(X;t_1,t_2) = \bar{h}_i(t_1,t_2)[\bar{m}_i^w(t_1,t_2) - \bar{\mathscr{E}}_i^{*w}(X;t_1,t_2)]. \tag{22}$$

Characterization Results Based on Conditional Dynamic Weighted CPE for X_i given $X_j < t_j$

In this section, we consider some important characterization results based on conditional dynamic weighted CPE for X_i given $X_j < t_j$. In the following theorem, we show that under certain conditions, $\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2);\ i = 1, 2$ determines the distribution function uniquely.

Theorem 2. *Let $X = (X_1, X_2)$ be a non negative bivariate random variable having absolutely continuous distribution function F with respect to the Lebesgue measure. Then CDWCPE of X, defined in (12) and (13), uniquely determines the distribution function provided they are finite.*

Proof. Let X and Y be two bivariate random variables having joint distribution functions F and G, respectively. Also let, for all $t_1, t_2 \geq 0$,

$$\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) = \bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2),\ i = 1, 2.$$

Differentiating $\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2)$ and $\bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2)$ with respect to $t_i,\ i = 1, 2$ and on using the relation $\phi_i^X(t_1, t_2)\bar{m}_i^w(t_1, t_2) = t_i - \frac{\partial}{\partial t_i}\bar{m}_i^w(t_1, t_2)$, from (22) we have

$$\frac{\partial}{\partial t_i}\bar{m}_{iX}^w(t_1, t_2) = \frac{\frac{\partial}{\partial t_i}\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2)\bar{m}_{iX}^w(t_1, t_2) + t_i\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) - t_i m_{iX}^w(t_1, t_2)}{\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) - \bar{m}_{iX}^w(t_1, t_2)}$$

and

$$\frac{\partial}{\partial t_i}\bar{m}_{iY}^w(t_1, t_2) = \frac{\frac{\partial}{\partial t_i}\bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2)\bar{m}_{iY}^w(t_1, t_2) + t_i\bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2) - t_i\bar{m}_{iY}^w(t_1, t_2)}{\bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2) - \bar{m}_{iY}^w(t_1, t_2)}.$$

Let

$$\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) = \bar{\mathscr{E}}_i^{*w}(Y; t_1, t_2) = \theta_i(\underline{\mathbf{t}});\ \underline{\mathbf{t}} = (t_1, t_2)$$

and

$$\psi_i(\underline{\mathbf{t}}, k) = \frac{\frac{\partial}{\partial t_i}\theta_i(\underline{\mathbf{t}})k + t_i\theta_i(\underline{\mathbf{t}}) - t_i k}{\theta_i(\underline{\mathbf{t}}) - k},\ \underline{\mathbf{t}} = (t_1, t_2), i = 1, 2.$$

Thus we can write

$$\frac{\partial}{\partial t_i}\bar{m}_{iX}^w(t_1, t_2) = \psi_i(\underline{\mathbf{t}}, \bar{m}_{iX}^w(t_1, t_2))$$

and

$$\frac{\partial}{\partial t_i}\bar{m}_{iY}^w(t_1, t_2) = \psi_i(\underline{\mathbf{t}}, \bar{m}_{iY}^w(t_1, t_2)). \tag{23}$$

Now we show that $\bar{m}_i^w(t_1, t_2))$ uniquely determines the distribution function. Differentiating (19) with respect to t_1, we get

$$\frac{\partial}{\partial t_1}\log F(t_1, t_2) = \frac{1}{\bar{m}_1^w(t_1, t_2)}\left(t_1 - \frac{\partial}{\partial t_1}\bar{m}_1^w(t_1, t_2)\right)$$

Now integrating the above equation on (t_1, b_1), we get

$$F(t_1, t_2) = F_2(t_2)\exp\left\{-\int_{t_1}^{b_1}\frac{1}{\bar{m}_1^w(x, t_2)}\left(x - \frac{\partial}{\partial x}\bar{m}_1^w(x, t_2)\right)dx\right\} \tag{24}$$

Similarly for $i = 2$,

$$F(t_1, t_2) = F_1(t_1)\exp\left\{-\int_{t_2}^{b_2}\frac{1}{\bar{m}_2^w(t_1, x)}\left(x - \frac{\partial}{\partial x}\bar{m}_2^w(t_1, x)\right)dx\right\} \tag{25}$$

Allowing t_2 to tend to b_2 in (24), we get

$$F_1(t_1) = \exp\left\{-\int_{t_1}^{b_1} \frac{1}{\bar{m}_1^w(x,b_2)}\left(x - \frac{\partial}{\partial x}\bar{m}_1^w(x,b_2)\right)dx\right\} \tag{26}$$

Now substitute (26) in (25). Hence the distribution of X is uniquely determined as

$$\begin{aligned} F(t_1,t_2) \;=\; & \exp\left\{-\int_{t_1}^{b_1} \frac{1}{\bar{m}_1^w(x,b_2)}\left(x - \frac{\partial}{\partial x}\bar{m}_1^w(x,b_2)\right)dx\right. \\ & \left. - \int_{t_2}^{b_2} \frac{1}{\bar{m}_2^w(t_1,x)}\left(x - \frac{\partial}{\partial x}\bar{m}_2^w(t_1,x)\right)dx\right\}. \end{aligned} \tag{27}$$

Or equivalently from (25) as

$$\begin{aligned} F(t_1,t_2) \;=\; & \exp\left\{-\int_{t_1}^{b_1} \frac{1}{\bar{m}_1^w(x,t_2)}\left(x - \frac{\partial}{\partial x}\bar{m}_1^w(x,t_2)\right)dx\right. \\ & \left. - \int_{t_2}^{b_2} \frac{1}{\bar{m}_2^w(b_1,x)}\left(x - \frac{\partial}{\partial x}\bar{m}_2^w(b_1,x)\right)dx\right\}. \end{aligned} \tag{28}$$

Hence from (23), we get $\bar{\mathscr{E}}_i^{*w}(X;t_1,t_2)$ determines $\bar{m}_{iX}^w(t_1,t_2)$. Again using the fact that the vector valued WEIT uniquely determines the bivariate distribution function, the proof is complete.

□

In the following theorem we characterize uniform distribution. The proof follows easily as X_1 and X_2 are independent.

Theorem 3. *Let $X = (X_1, X_2)$ be a bivariate random variable having joint distribution function F. Then X is said to follow bivariate uniform distribution with distribution function*

$$F(t_1,t_2) = \frac{t_1 t_2}{b\ d},\ 0 \le t_1 \le b,\ 0 \le t_2 \le d,$$

if and only if $\bar{\mathscr{E}}_i^{*w}(X;t_1,t_2) = \frac{t_i^2}{9},\ i = 1,2.$

In the following theorem we give another characterization result for uniform distribution with dependent components.

Theorem 4. *Let X be a non negative bivariate random vector with $\bar{\mathscr{E}}_i^{*w}(X;t_1,t_2) < \infty$ and WEIT $\bar{m}_i^w(t_1,t_2),\ i = 1,2$ for all $t_i \ge 0$. Then for $0 < t_1, t_2 < 1,\ \theta \le 0$,*

$$\bar{\mathscr{E}}_i^{*w}(X;t_1,t_2) = \left(\frac{1+\theta\log t_j}{3+\theta\log t_j}\right)\bar{m}_i^w(t_1,t_2),\ i \ne j = 1,2, \tag{29}$$

if and only if X is distributed as bivariate uniform with

$$F(t_1,t_2) = t_1^{1+\theta\log t_2}t_2,\ 0 < t_1,t_2 < 1,\ \theta \le 0. \tag{30}$$

Proof. The if part is straight forward if X follows the distribution (30), then

$\bar{m}_i^w(t_1,t_2) = \frac{t_i^2}{3+\theta\lg t_j}$ and $\bar{\mathscr{E}}_1^{*w}(X;t_1,t_2) = \frac{t_i^2(1+\theta\log t_j)}{(3+\theta\log t_j)^2},\ i \ne j = 1,2.$

To prove the converse, assume that (29) holds. Then differentiating (29) with respect to t_i and using (22), we get

$$\frac{\partial}{\partial t_i}\bar{m}_i^w(t_1,t_2) = \frac{2t_i}{3+\theta\log t_j}\ \ i \neq j = 1,2,$$

which on integration gives

$$\bar{m}_i^w(t_1,t_2) = \frac{t_i^2}{3+\theta\log t_j} + c_i(t_j)$$

where $c_i(t_j)$ is the constant of integration. Now, $c_i(t_j) = 0$ as $\bar{m}_i^w(t_1,t_2) \to 0$ for $t_i \to 0$, which in turn gives the bivariate WEIT. Hence the result follows on using the fact that bivariate WEIT determines the distribution function uniquely.

□

The following theorem gives characterization of the bivariate power distribution.

Theorem 5. *Let X be a non negative random vector in the support $(0,b_1)\times(0,b_2)$, $b_i < \infty$, $i = 1,2$ with $\bar{\mathcal{E}}_i^{*w}(X;t_1,t_2)$ finite. Then*

$$\bar{\mathcal{E}}_i^{*w}(X;t_1,t_2) = c_i(t_j)\bar{m}_i^w(t_1,t_2); i \neq j = 1,2, \tag{31}$$

if and only if X follows bivariate power distribution with distribution function

$$F(t_1,t_2) = \left(\frac{t_1}{b_1}\right)^{c_1}\left(\frac{t_2}{b_2}\right)^{c_2+\theta\log\left(\frac{t_1}{b_1}\right)},\ \ \theta \leq 0, \tag{32}$$

where $c_i = \frac{c_i(b_j)}{[1-c_i(b_j)]}$.

Proof. The if part of the theorem is straight forward. To prove the reverse part, if (29) holds, then differentiating both sides with respect to t_i, and on using (22), we get

$$\frac{\partial}{\partial t_i}m_i^w(t_1,t_2) = t_i[1-c_i(t_j)],\ \ i \neq j = 1,2,$$

which on integration gives

$$\bar{m}_i^w(t_1,t_2) = [1-c_i(t_j)]\frac{t_i^2}{2} + z_i(t_j),$$

where $z_i(t_j)$ is a constant of integration. Now $z_i(t_j) = 0$ as $\bar{m}_i^w(t_1,t_2) \to 0$ for $t_i \to 0$. Which in turn gives the bivariate weighted EIT. Hence the result follows on using the fact that bivariate weighted EIT determines the distribution function uniquely.

□

Now we define new classes of life distributions based on vector dynamic WCPE that are analogous to certain properties of the vector dynamic CRE defined in [13].

Definition 4. *A random variable X is said to be increasing (decreasing) bivariate reversed hazard rate if $\bar{h}_i(t_1,t_2)$ is increasing (decreasing) in t_i, $i = 1,2$.*

Definition 5. *A random variable X is said to be increasing (decreasing) weighted expected inactivity time if $\bar{m}_i^w(t_1,t_2)$ is increasing (decreasing) in t_i, $i = 1,2$.*

Definition 6. *For $t_1,t_2 > 0$, F is said to have bivariate increasing (decreasing) weighted uncertainty in past life BIWUPL (BDWUPL) if $\bar{\mathcal{E}}_i^{*w}(X;t_1,t_2)$ is increasing (decreasing) in t_i, $i = 1,2$.*

The following theorem gives an upper bound to bivariate WCPE in terms of weighted expected inactivity time as follows.

Theorem 6. *For $t_1, t_2 > 0$, if F is BIWUPL (BDWUPL)*

$$\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2) \leq (\geq) \bar{m}_i^w(t_1, t_2).$$

Proof. Differentiating $\bar{\mathscr{E}}_i^{*w}(X; t_1, t_2)$ with respect to t_i, we get (22). The proof follows by using the fact that $r_i^X(t_1, t_2)$ is non negative for all t_i.

□

Conditional Dynamic Weighted CPE for X_i given $X_j = t_j$

The determination of the joint distribution function of $X = (X_1, X_2)$, when conditional distributions of $(X_1|X_2 = t_2)$ and $(X_2|X_1 = t_1)$ are known, has been an important problem dealt with by many researchers in the past. This approach of identifying a bivariate density using the conditionals is called the conditional specification of the joint distribution. These conditional models are often useful in many two component reliability systems, when the operational status of one component is known. Let the distribution function of $\bar{Y}_i^* = (X_i|X_i < t_i, X_j = t_j)$, $i \neq j = 1, 2$ is defined as $F_i^*(t_i|t_j)$. Then, for an absolutely continuous nonnegative bivariate random vector X, the conditional dynamic DWCPE of $\bar{Y}_i^*$, $i \neq j = 1, 2$ is defined as

$$\bar{\zeta}_1^{*w}(X; t_1, t_2) = -\int_0^{t_1} x_1 \frac{F_1^*(x_1|t_2)}{F_1^*(t_1|t_2)} \log \frac{F_1^*(x_1|t_2)}{F_1^*(t_1|t_2)}\, dx_1 \tag{33}$$

and

$$\bar{\zeta}_2^{*w}(X; t_1, t_2) = -\int_0^{t_2} x_2 \frac{F_2^*(x_2|t_1)}{F_2^*(t_2|t_1)} \log \frac{F_2^*(x_2|t_1)}{F_2^*(t_2|t_1)}\, dx_2. \tag{34}$$

In particular if X_1 and X_2 are independent, then $\bar{\zeta}_i^{*w}(X; t_1, t_2) = \bar{\mathscr{E}}_{X_i}^w(t_i),\ i = 1, 2$.

In the sequel we give the definitions of bivariate reversed hazard rate and bivariate weighted EIT functions for the random variable $\bar{Y}_i^*$.

Definition 7. *For a random vector $X = (X_1, X_2)$ with distribution functions $F_i^*(t_i|t_j)$, $i \neq j = 1, 2$*

(i) the bivariate reversed hazard rate is defined as a vector,
$\bar{h}^*(t_1|t_2) = (\bar{h}_1^*(t_1|t_2), \bar{h}_2^*(t_1|t_2))$, *where*

$$\bar{h}_i^*(t_i|t_j) = \frac{\partial}{\partial t_i} \log F_i^*(t_i|t_j),\ i \neq j = 1, 2 \tag{35}$$

are the components of bivariate reversed hazard rate;

(ii) the bivariate WEIT is defined by the vector,
$\bar{m}^{*w}(t_1|t_2) = (\bar{m}_1^{*w}(t_1|t_2), \bar{m}_2^{*w}(t_1|t_2))$, *where*

$$\bar{m}_i^{*w}(t_i|t_j) = \frac{1}{F_i^*(t_i|t_j)} \int_0^{t_i} x_i F_i^*(x_i|t_j)\, dx_i,\ i \neq j = 1, 2. \tag{36}$$

which measures the the expected waiting time of X_i given that $X_i < t_i$ and $X_j = t_j$.

Differentiating (33) and (34) with respect to t_1 and t_2 respectively, we get in general

$$\frac{\partial}{\partial t_i}\bar{\zeta}_i^{*w}(X;t_1,t_2) = \bar{h}_i^*(t_1|t_2)[\bar{m}_i^{*w}(t_1|t_2) - \bar{\zeta}_i^{*w}(X;t_1,t_2)]. \quad (37)$$

Now we define new classes of life distributions based on vector dynamic WCPE that are analogous to certain properties of the vector dynamic CRE defined in [12]

Definition 8. *A random variable X is said to be increasing (decreasing) bivariate RHR if $\bar{h}_i^*(t_i|t_j)$ is increasing (decreasing) in $t_i,\ i=1,2$.*

Definition 9. *A random variable X is said to be increasing (decreasing) weighted expected inactivity time if $\bar{m}_i^{*w}(t_i|t_j)$ is increasing (decreasing) in $t_i,\ i=1,2$.*

The following theorem gives the necessary and sufficient condition for $\bar{\zeta}_i^{*w}(X;t_1,t_2)$ to be increasing (decreasing) vector dynamic WCPE.

Theorem 7. *For $t_1,t_2>0$, $\bar{\zeta}_i^{*w}(X;t_1,t_2)$ is increasing (decreasing) in $t_i,\ i=1,2$ if and only if*

$$\bar{\zeta}_i^{*w}(X;t_1,t_2) \le (\ge)\bar{m}_i^{*w}(t_i|t_j), i,j=1,2, i\ne j.$$

Proof. Differentiating $\bar{\zeta}_i^{*w}(X;t_1,t_2)$ with respect to t_i, we get (37). The proof follows by using the fact that $r_i^*(t_i|t_j)$ is non negative for all t_i. □

Analogous to Theorem 8, the following theorem establishes a fundamental relationship between dynamic WCPE and weighted EIT of X_i given $X_j=t_j, i\ne j=1,2$. The proof is omitted.

Theorem 8. *Let $X=(X_1,X_2)$ be an absolutely continuous non negative bivariate random vector with finite $\bar{\zeta}_i^{*w}(X;t_1,t_2)$ and bivariate weighted EIT components $\bar{m}_i^{*w}(t_i|t_j), i\ne j=1,2$. Then for all $t_1,t_2\ge 0$,*

$$\bar{\zeta}_1^{*w}(X;t_1,t_2) = \int_0^{t_1}\bar{m}_1^{*w}(x_1|t_2)\bar{f}_1^*(x_1|t_2)dx_1 \quad (38)$$

and

$$\bar{\zeta}_2^{*w}(X;t_1,t_2) = \int_0^{t_2}\bar{m}_2^{*w}(t_1|x_2)\bar{f}_2^*(t_1|x_2)dx_2, \quad (39)$$

where $\bar{f}_i^*(x_i;t_j)$ is the density function of $(X_i|X_i<t_i, X_j=t_j)$, $i\ne j=1,2$.

4 Estimation of Conditional Dynamic Weighted Cumulative Past Entropy

In this section, we focused on constructing non-parametric estimators for CDWCPE. Let $(X_{1i},X_{2i}); i=1,\ldots,n$ be n independent and identically distributed pairs of lifetimes with joint distribution function $F(x_1,x_2)$. Based on these observations and using the kernel density $k_i(.), i=1,2$, a non-parametric estimator for $\bar{F}(x_1,x_2)$ is defined as

$$\hat{F}(x_1,x_2) = \frac{1}{na_n^2}\sum_{j=1}^{n} K_1\left(\frac{x_1-X_{1j}}{a_n}\right)K_2\left(\frac{x_2-X_{2j}}{a_n}\right), \quad (40)$$

where

$$K_i(z) = a_n\int_0^z k_i(v)\,dv, i=1,2 \quad (41)$$

and $\{a_n\}$ is a non-increasing sequence of positive real numbers such that $a_n\to 0$ and $na_n\to\infty$, as $n\to\infty$. From (12) and (13), we propose a non-parametric kernel estimator for CDWCPE as follows.

Definition 10. *Let $X = (X_1, X_2)$ be a random sample drawn from a population having distribution function $F(x_1, x_2)$. From (12) and (13), we define the non-parametric kernel estimator for CDWCPE as follows.*

$$\bar{\mathscr{E}}_1^{*w}(X; t_1, t_2) = -\int_0^{t_1} x_1 \frac{\hat{F}(x_1, t_2)}{\hat{F}(t_1, t_2)} \log \frac{F(x_1, t_2)}{\hat{F}(t_1, t_2)}\, dx_1 \tag{42}$$

and

$$\bar{\mathscr{E}}_2^{*w}(X; t_1, t_2) = -\int_0^{t_2} x_2 \frac{\hat{F}(t_1, x_2)}{\hat{F}(t_1, t_2)} \log \frac{F(t_1, x_2)}{\hat{F}(t_1, t_2)}\, dx_2, \tag{43}$$

where

$$\hat{F}(t_1, t_2) = \frac{1}{n} \sum_{k=1}^{n} I(X_{1k} < t_1, X_{2k} < t_2) \quad \text{is the empirical distribution function}$$

and

$$I(X_{1k} < t_1, X_{2k} < t_2) = \begin{cases} 1, & X_{1k} < t_1, X_{2k} < t_2 \\ 0, & otherwise. \end{cases}$$

is the indicator function of the event $\{X_{1k} < t_1, X_{2k} < t_2\}$.

Numerical Illustration

In this section, we illustrate the usefulness of the proposed estimator given in (42) and (43) using a real life data set.

Example 3. *Consider the data-set reported by [8], which consists of the failure times of 20 sample units from a system consisting of three components. Here we consider the failure times of first two components.*

Here we use the bootstrapping procedure. At each value of (t_1, t_2), we calculate the bias and MSE of $\hat{\bar{\mathscr{E}}}_i^{*w}(X; t_1, t_2),\ i = 1, 2$ using 100 bootstrap samples of size 20. Table 1 presents the absolute values of the bias and MSE for $\hat{\bar{\mathscr{E}}}_i^{*w}(X; t_1, t_2), i = 1, 2$ with samples of size $n = 20$. From Table 1, we can see that for $t_1 > \ (<)\ t_2$, absolute values of bias and MSE of $\hat{\bar{\mathscr{E}}}_1^{*w}(X; t_1, t_2)$ is greater (less) than $\hat{\bar{\mathscr{E}}}_2^{*w}(X; t_1, t_2)$. When the entropy is lower, we know that there is a greater likelihood of an event happening. Table results indicate that there is a likelihood of an increase in the time to the second failure time as the time to the first failure time passes. If there was an immediate first failure, it would be expected that the second would follow shortly after.

(t_1, t_2)	Bias of $\hat{\bar{\mathscr{E}}}_1^{*w}(X; t_1, t_2)$	MSE of $\hat{\bar{\mathscr{E}}}_2^{*w}(X; t_1, t_2)$
(3.85, 3.88)	(-0.0871, -0.1673)	(0.0321, 0.2635)
(2.42, 3.86)	(0.2339, -0.5677)	(0.1044, 0.5018)
(2.70, 2.17)	(0.3994, 0.0234)	(0.2341, 0.0554)
(1.53, 2.08)	(-0.0015, -0.1939)	(0.0045, 0.0532)
(1.93, 2.08)	(-0.2183, -0.2249)	(0.0416, 0.0759)
(2.50, 2.08)	(-0.2363, -0.1207)	(0.0735, 0.0335)
(2.50, 3.88)	(-0.0639, 0.1871)	(0.0175, 0.4379)

Table 1: Bootstrap bias and MSE estimates of $\hat{\bar{\mathscr{E}}}_i^{*w}(X; t_1, t_2)$ for the failure times of the components

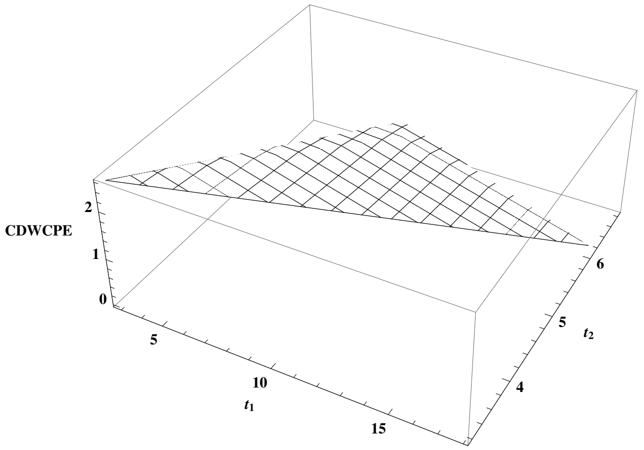

Figure 1: Graphs of $\hat{\hat{\mathscr{E}}}_i^{*w}(X;t_1,t_2),\ i=1,2$ for the failure times of two components.

Figure 1 shows the graph of $\hat{\hat{\mathscr{E}}}_i^{*w}(X;t_1,t_2), i = 1,2$ for the real life data set and we can see that the expected weighted uncertainty in the predictability about the failure times of two components are decreasing in t_1 and t_2. It implies that as time progresses, the uncertainty associated with predicting when each component will fail is decreasing. Lower uncertainty in failure predictions can lead to more effective maintenance strategies. If we can predict failures more accurately, we can optimize maintenance schedules, reduce downtime, and manage resources more efficiently.

To illustrate the effectiveness of our proposed estimator, we calculated the CDWCPE and univariate DWCPE for the real life data set. In particular, if $t_1 = 2.50$, the univariate DWCPE for the failure time of first component is

$$\hat{\hat{\mathscr{E}}}^w(X_1;t_1) = 2.369.$$

The univariate DWCPE for the failure time of second component is

$$\hat{\hat{\mathscr{E}}}^w(X;t_2) = 2.6913.$$

Using (42) and (43), the CDWCPE is

$$\left(\hat{\hat{\mathscr{E}}}_1^{w*}(X;t_1,t_2), \hat{\hat{\mathscr{E}}}_2^{w*}(X;t_1,t_2)\right) = (2.2189, 1.4519). \tag{44}$$

From (44) we can see that for all possible values of t_1 and t_2, when we consider the failure times of two components jointly, the expected weighted uncertainty about the predictability of failure time of components decreases and CDWCPE will give more information about the predictability of failure times of two components. This is an indication that in survival analysis, if we have the knowledge about the failure time of only one component, the expected weighted uncertainties are very high, where as if we have the knowledge about failure times of all the components, it decreases. So we suggest that in situations like this to consider the bivariate weighted CPE to study the failure times of multiple the components.

References

[1] Maliheh Abbasnejad. "Some characterization results based on dynamic survival and failure entropies." In: *Communications for Statistical Applications and Methods* 18 (2011), pp. 787–798.

[2] Jafar Ahmadi, Antonio Di Crescenzo, Maria Longobardi, et al. "On dynamic mutual information for bivariate lifetimes." In: *Advances in Applied Probability* 47 (2015), pp. 1157–1174.

[3] Mariana Belis and Silviu Guiasu. "A quantitative-qualitative measure of information in cybernetic systems (Corresp.)" In: *IEEE Transactions on Information Theory* 14 (1968), pp. 593–594.

[4] Suchismita Das. "On weighted generalized entropy." In: *Communications in Statistics-Theory and Methods* 46.12 (2017), pp. 5707–5727.

[5] Antonio Di Crescenzo and Maria Longobardi. "On cumulative entropies." In: *Journal of Statistical Planning and Inference* 139 (2009), pp. 4072–4087.

[6] Antonio Di Crescenzo and Maria Longobardi. "Stochastic comparisons of cumulative entropies." In: *Stochastic Orders in Reliability and Risk*. Springer, 2013, pp. 167–182.

[7] Suchandan Kayal. "On weighted generalized cumulative residual entropy of order n." In: *Methodology and Computing in Applied Probability* (2017), pp. 1–17.

[8] Hyoungtae Kim and Paul H Kvam. "Reliability estimation based on system data with an unknown load share rule." In: *Lifetime Data Analysis* 10 (2004), pp. 83–94.

[9] Amarjit Kundu and Chanchal Kundu. "Bivariate Extension of (Dynamic) Cumulative Past Entropy." In: *Communications in Statistics-Theory and Methods* (2016).

[10] MALIHEH Mirali, S Baratpour, and Vahid Fakoor. "On weighted cumulative residual entropy." In: *Communications in Statistics-Theory and Methods* 46.6 (2017), pp. 2857–2869.

[11] J Navarro, SM Sunoj, and MN Linu. "Characterizations of bivariate models using some dynamic conditional information divergence measures." In: *Communications in Statistics-Theory and Methods* 43 (2014), pp. 1939–1948.

[12] G Rajesh et al. "Bivariate extension of dynamic cumulative residual entropy." In: *Statistical Methodology* (2014), pp. 72–82.

[13] G Rajesh et al. "Bivariate generalized cumulative residual entropy." In: *Sankhya A* 76 (2014), pp. 101–122.

[14] G Rajesh et al. "The Conditional Dynamic Cumulative Residual Entropy." In: *Journal of the Japan Statistical Society* 2 (2016), pp. 99–119.

[15] C Radhakrishna Rao. "On discrete distributions arising out of methods of ascertainment." In: *Sankhyā: The Indian Journal of Statistics, Series A* (1965), pp. 311–324.

[16] Murali Rao et al. "Cumulative residual entropy: a new measure of information." In: *IEEE transactions on Information Theory* 50 (2004), pp. 1220–1228.

[17] CE Shannon. "A mathematical theory of communication, bell System technical Journal 27: 379-423 and 623–656." In: *Mathematical Reviews (MathSciNet): MR10, 133e* (1948).

[18] SM Sunoj and MN Linu. "Dynamic cumulative residual Renyi's entropy." In: *Statistics* 46 (2012), pp. 41–56.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Bayes' estimators for parameters of Gumbel's Type I bivariate exponential distribution with positive probability at instantaneous failure

Saju Verghese George[1] and Vaijayanti Ullas Dixit[2]

1. *Dept. of Statistics, St. Xavier's College, Mumbai, Maharashtra, Pin 400001, India*
2. *Dept. of Statistics, University of Mumbai, Mumbai, Maharashtra, Pin 400098, India*

Received: Dec 2023 / Accepted: July 2024

Abstract

The bivariate exponential distribution with exponential margins defined by Gumbel [1] Type-I model is a continuous lifetime distribution. In practice, however, there may be a possibility of instantaneous failure, that is, possibility of both variables taking value zero with positive probability. Thus we get a mixture distribution with positive probability at $X = 0$ and $Y = 0$ in bivariate exponential distribution, where the marginals too are exponential distribution with positive probability at instantaneous failure, that is, at $X = 0$ and at $Y = 0$. It is observed that the joint probability density function and the cumulative distribution function can be expressed in compact form. Here we present Bayes' method of estimation of the three parameters p_1, p_2 and θ of the bivariate exponential distribution with positive probability at instantaneous failure. We use squared error loss function. The bias and mean squared error (MSE) of these estimators are computed using simulation.

Keywords: Bayes estimator, squared error loss function.

1 Introduction

Exponential distribution has been studied extensively in the last centuries and has numerous applications, specially in the field of Reliability and Analysis of life-time data. Its cumulative distribution function (cdf) is given by

$$\begin{aligned} F(x, \lambda) &= 0 && ; x \leq 0 \\ &= 1 - exp(-\lambda x) && ; x \geq 0, \lambda > 0. \end{aligned} \tag{1.1}$$

Its probability density function (pdf) is

$$\begin{aligned} f(x, \lambda) &= \lambda \text{ exp}(-\lambda x) && ; x \geq 0, \lambda > 0, \\ &= 0 && ; \text{otherwise}, \end{aligned} \tag{1.2}$$

denoted as $X \sim Exp(\lambda)$, where λ is the scale parameter.

A mixture of degenerate distribution (degenerate at zero) and exponential distribution occurs frequently in many practical situations. For instance, the life of a component may have an exponential distribution but some of the components may fail instantaneously. Another possible application of the mixture distribution is in clinical trials where a drug may have no response with probability $1 - p$, $(0 < p \leq 1)$ but once there is a response, the length of response may follow exponential distribution. Hence, a mixed failure time distribution (MFTD) is proposed by Kleyle and Dahiya [2] as follows:

$$\begin{aligned} F(x) &= 0 \qquad\qquad ; x < 0 \\ &= 1 - pe^{-\lambda x} \qquad ; x \geq 0, 0 < p \leq 1, \lambda > 0, \end{aligned}$$

where, $P(X = 0) = 1 - p$. They have studied estimation of parameters of this distribution from Type-I censored data.

Jayade and Prasad [3] have studied estimation of parameters of this distribution based on modified sampling scheme. Dixit [4] studied statistical inference for AR (1) process when the error follows MFTD. Dixit [5] derived classical optimum tests of this MFTD. Shinde and Shanubhogue [6] also studied estimation of parameters and the mean life of MFTD. Dixit [7] further studied estimation of parameters of this distribution based on an extended modified sampling scheme.

A bivariate random variable (X, Y) is said to have a **bivariate exponential distribution** (BVED) if its marginals are exponential distributions. The concept of positive probability at instantaneous failure is introduced to bivariate exponential distribution model as described by Gumbel [1] (referred as Gumbel's Type I bivariate exponential distribution)

$$G(x, y) = 1 - e^{-x} - e^{-y} + e^{-x-y-\theta xy}; x \geq 0, y \geq 0, 0 \leq \theta \leq 1, \tag{1.3}$$

where θ is the parameter of association. When $\theta = 0$ in (1.3), it is obvious that the random variables are independent. Dixit and Karadkar [8] studied estimation of the parameters of the distribution function in (1.3). Khare [9] has also obtained Bayes' estimator of θ for squared error loss function. In Dixit et al. [10] Bayes estimator of θ is obtained for squared error loss function and its bias and MSE are computed for different values of sample size and θ.

If we consider a system of two components whose lifetimes follow Gumbel's Type I bivariate exponential distribution and there is a possibility of instantaneous failure in either of the components or instantaneous failure in both the components together, then we can have a model as follows:

Introducing parameters $p_1 = P(X > 0)$ and $p_2 = P(Y > 0)$, $0 < p_1 < 1, 0 < p_2 < 1$ in the appropriate place in the distribution function (1.3) we obtain the new mixture distribution function as

$$F(x, y) = 1 - p_1 e^{-x} - p_2 e^{-y} + p_1 p_2 e^{-x-y-\theta xy}; x \geq 0, y \geq 0, 0 < p_1, p_2 < 1, 0 \leq \theta \leq 1. \tag{1.4}$$

This is Gumbel's Type I bivariate exponential distribution with positive probability at instantaneous failure (BVEDP).

Saju & Dixit [11] have studied the properties of (1.4).The marginal distribution function of X and Y respectively are

$$\begin{aligned} F(x) &= 1 - p_1 e^{-x} \quad ; x \geq 0, 0 < p_1 < 1, \\ F(y) &= 1 - p_2 e^{-y} \quad ; y \geq 0, 0 < p_2 < 1, \end{aligned}$$

which are exponential with positive probability $(1 - p_1)$ at $X = 0$, and $(1 - p_2)$ at $Y = 0$ respectively and when $p_1 = p_2 = 1$ they are exponential distributions. It is obvious that when $p_1 = 1$ and $p_2 = 1$, (1.4) reduces to (1.3) and it satisfies all the conditions of bivariate distribution function.

The corresponding mixed joint probability density function (pdf) of (1.4) is given by

$$\begin{aligned} f(x, y) = &\Big[(1 - p_1)(1 - p_2)\Big]^{I(x)I(y)} \Big[(1 - p_1)p_2 e^{-y}\Big]^{I(x)[1-I(y)]} \\ &\Big[(1 - p_2)p_1 e^{-x}\Big]^{[1-I(x)]I(y)} \Big[p_1 p_2 e^{-x-y-\theta xy}\Big[(1 + \theta x)(1 + \theta y) - \theta\Big]\Big]^{[1-I(x)][1-I(y)]} \\ &; x \geq 0, y \geq 0, 0 < p_1, p_2 < 1, 0 \leq \theta \leq 1 \end{aligned} \tag{1.5}$$

which can also be written as

$$\begin{aligned} f(x,y) &= (1-p_1)(1-p_2) = P(X=0,Y=0) && ;x=0,y=0\\ &= (1-p_1)p_2e^{-y} && ;x=0,y>0\\ &= (1-p_2)p_1e^{-x} && ;x>0,y=0\\ &= p_1p_2e^{-x-y-\theta xy}\ [(1+\theta x)(1+\theta y)-\theta] && ;x>0,y>0,0<p_1,p_2<1,0\leq\theta\leq 1. \end{aligned}$$

$$\begin{aligned} \text{where } I(t) &= 1 \text{ if } t=0\\ &= 0 \text{ if } t>0. \end{aligned} \tag{1.6}$$

Further, Saju and Dixit [11] have obtained Maximum likelihood estimators (MLE) of p_1, p_2 and θ of (1.4). We have also studied finite sample and asymptotic properties of MLE, using simulation. In this paper we obtain Bayes' estimators of p_1, p_2 and θ of (1.4) for squared error loss function, assuming appropriate conjugate prior distribution for p_1, p_2 and θ. Further, we have obtained bias and mean squared error (MSE) of these Bayes estimators using simulation.

The bias and MSE of $\hat{p_1}$ and $\hat{p_2}$ are in closed form. Though Bayes estimator of θ appears in closed form its bias and MSE cannot be obtained in closed form. Hence as a computational illustration, we compute bias and MSE of the Bayes estimators of p_1, p_2 and θ using simulation which are presented in Table 1. The Table is prepared for Simulation size $N = 1000$, sample sizes $(n = 5, 10, 15), p_1 = 0.5, (p_2 = 0.5, 0.7$ and $0.9)$ and $(\theta = 0.3, 0.7, 0.9)$. We consider $a_1 = 2, b_1 = 3, a_2 = 2, b_2 = 3, c = 3, d = 2$.

Bayesian inference is studied by various authors namely, Pathak et al [12] for Poisson inverse exponential distribution (PIED). They have studied Bayesian and E-Bayesian estimators of parameters of PIED under squared error loss function (SELF), General Entropy Loss function (GELF) and Linear Exponential Loss function (LINEX) for progressive type-II censored data with binomial removals. Their risks are compared using simulation. These methods are applied to survival time of multiple myeloma patient data.

In addition, Pathak et al. [13] have studied Bayesian inference for Weibull Poisson model for censored data. They have considered estimation of parameters under progressive type-II censoring with binomial removals. The MLE's and Bayes estimators have been obtained under symmetric and asymmetric loss functions. Further, their performance is compared using simulation. These methods are illustrated through real bladder cancer data set.

2 Bayes' estimators for the parameters of Gumbel's Type I bivariate exponential distribution with positive probability at instantaneous failure

In this section we describe Bayes' estimators of the parameters p_1, p_2 and θ for the pdf in (1.5).

Suppose $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a random sample of size n from mixed density function given in (1.5). Let

$$\sum_{i=1}^{n} I(x_i)I(y_i) = n_1, \sum_{i=1}^{n} I(x_i)[1-I(y_i)] = n_2, \sum_{i=1}^{n}[1-I(x_i)]I(y_i) = n_3,$$
$$\sum_{i=1}^{n}[1-I(x_i)][1-I(y_i)] = n_4, \sum_{i=1}^{4} n_i = n$$

$I(.)$ is as defined in (1.6)

It can be noted that (N_1, N_2, N_3) follows multinomial distribution with pmf as follows:

$$P(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!(q_1q_2)^{n_1}(q_1p_2)^{n_2}(p_1q_2)^{n_3}(p_1p_2)^{n_4}}{n_1!\ n_2!\ n_3!\ n_4!} \tag{2.1}$$

$$\begin{aligned}\text{where, } & q_1 = 1 - p_1, q_2 = 1 - p_2, \\ & q_1q_2 + q_1p_2 + p_1q_2 + p_1p_2 = 1, \\ & n_1, n_2, n_3, n_4 = 0, 1, 2, ..., n \text{ and } \sum_{i=1}^{4} n_i = n.\end{aligned}$$

For the prior distribution we assume that $p_1 \sim$ Beta$(a_1, b_1), p_2 \sim$ Beta$(a_2, b_2), \theta \sim$ Beta(c, d). Further, we also assume that $(\vec{x}, \vec{y}), p_1, p_2$ and θ are independent. The joint distribution of $(\vec{x}, \vec{y}, p_1, p_2$ and $\theta)$ is given as

$$f(\vec{x}, \vec{y}, p_1, p_2, \theta) = \left[\prod_{i=1}^{n} f(x_i, y_i)\right].f_1(p_1).f_2(p_2).f_3(\theta)$$

where

$$\begin{aligned}\prod_{i=1}^{n} f(\vec{x}, \vec{y}, p_1, p_2, \theta) = {} & (1 - p_1)^{n_1+n_2}\ (1 - p_2)^{n_1+n_3}\ p_1^{n_3+n_4}\ p_2^{n_2+n_4} \\ & e^{-\sum_{i=1}^{n_3} x_i - \sum_{i=1}^{n_2} y_i - \sum_{i=1}^{n_4} x_i - \sum_{i=1}^{n_4} y_i} \\ & e^{-\theta \sum_{i=1}^{n_4} x_i y_i} \prod_{i=1}^{n_4} \Big[(1 + \theta x_i)(1 + \theta y_i) - \theta\Big] \text{ from (1.5)}\end{aligned}$$

$$\begin{aligned} f_1(p_1) &= \frac{p_1^{a_1-1}(1 - p_1)^{b_1-1}}{\beta(a_1, b_1)} && 0 < p_1 < 1, \\ f_2(p_2) &= \frac{p_2^{a_2-1}(1 - p_2)^{b_2-1}}{\beta(a_2, b_2)} && 0 < p_2 < 1, \text{ and} \\ f_3(\theta) &= \frac{\theta^{c-1}(1 - \theta)^{d-1}}{\beta(c, d)} && 0 < \theta < 1.\end{aligned}$$

Hence

$$\begin{aligned} f(\vec{x}, \vec{y}, p_1, p_2, \theta) = {} & (1 - p_1)^{n_1+n_2}\ (1 - p_2)^{n_1+n_3}\ p_1^{n_3+n_4}\ p_2^{n_2+n_4} \\ & e^{-\sum_{i=1}^{n_3} x_i - \sum_{i=1}^{n_2} y_i - \sum_{i=1}^{n_4} x_i - \sum_{i=1}^{n_4} y_i} \\ & e^{-\theta \sum_{i=1}^{n_4} x_i y_i} \prod_{i=1}^{n_4} \Big[(1 + \theta x_i)(1 + \theta y_i) - \theta\Big] \\ & \frac{p_1^{a_1-1}(1 - p_1)^{b_1-1}}{\beta(a_1, b_1)}\ \frac{p_2^{a_2-1}(1 - p_2)^{b_2-1}}{\beta(a_2, b_2)}\ \frac{\theta^{c-1}(1 - \theta)^{d-1}}{\beta(c, d)} \qquad ; n_4 > 0 \\ = {} & (1 - p_1)^{n_1+n_2}\ (1 - p_2)^{n_1+n_3}\ p_1^{n_3+n_4}\ p_2^{n_2+n_4} \\ & e^{-\sum_{i=1}^{n_3} x_i\ -\ \sum_{i=1}^{n_2} y_i}\ \frac{p_1^{a_1-1}(1 - p_1)^{b_1-1}}{\beta(a_1, b_1)}\ \frac{p_2^{a_2-1}(1 - p_2)^{b_2-1}}{\beta(a_2, b_2)}\ \frac{\theta^{c-1}(1 - \theta)^{d-1}}{\beta(c, d)} \qquad ; n_4 = 0 \\ & ; x, y \geq 0, 0 < p_1, p_2 < 1, 0 \leq \theta \leq 1, a_1, b_1, a_2, b_2, c, d > 0.\end{aligned} \tag{2.2}$$

$$\text{Let} \quad T_{3x} = \sum_{i=1}^{n_3} x_i\ ,\ T_{2y} = \sum_{i=1}^{n_2} y_i\ ,\ T_{4x} = \sum_{i=1}^{n_4} x_i\ ,\ T_{4y} = \sum_{i=1}^{n_4} y_i\ ,\ T_{4xy} = \sum_{i=1}^{n_4} x_i y_i \tag{2.3}$$

consider,

$$e^{-\theta\sum_{i=1}^{n_4} x_i y_i} = \sum_{i=0}^{\infty} \frac{(-1)^i\ (\theta\ T_{4xy})^i}{i!} = \sum_{i=0}^{\infty} \frac{(-T_{4xy})^i\ \theta^i}{i!} \tag{2.4}$$

Consider

$$\begin{aligned} R &= \prod_{i=1}^{n_4} \Big[(1+\theta x_i)\ (1+\theta y_i) - \theta\Big] \\ &= \prod_{i=1}^{n_4} \Big[1 + (x_i + y_i - 1)\ \theta + x_i y_i \theta^2\Big] \\ &= \prod_{i=1}^{n_4} \Big[1 + A_i\theta + B_i\theta^2\Big] \qquad \text{where } A_i = x_i + y_i - 1, B_i = x_i y_i \end{aligned}$$

for $n_4 = 1$

$$\begin{aligned} R &= 1 + A_1\theta + B_1\theta^2 \\ &= C_{10} + C_{11}\theta + C_{12}\theta^2 \\ &\text{where } C_{10} = 1, C_{11} = A_1, C_{12} = B_1 \end{aligned}$$

for $n_4 = 2$

$$\begin{aligned} R &= (1 + A_1\theta + B_1\theta^2)\ (1 + A_2\theta + B_2\theta^2) \\ &= 1 + (A_1 + A_2)\theta + (B_1 + A_1A_2 + B_2)\theta^2 + (B_1A_2 + A_1B_2)\theta^3 + (B_1B_2)\theta^4 \\ &= 1 + (C_{11} + A_2)\theta + (C_{12} + C_{11}A_2 + C_{10}B_2)\theta^2 \\ &\quad\ + (C_{13} + C_{12}A_2 + C_{11}B_2)\theta^3 + (C_{14} + C_{13}A_2 + C_{12}B_2)\theta^4 \\ &= C_{20} + C_{21}\theta + C_{22}\theta^2 + C_{23}\theta^3 + C_{24}\theta^4 \end{aligned}$$

for $n_4 = n_4$

$$R = C_{n_4,0} + C_{n_4,1}\theta + C_{n_4,2}\theta^2 + C_{n_4,3}\theta^3 + \ldots + C_{n_4,k}\theta^k + \ldots C_{n_4,2n_4}\theta^{2n_4}$$

where $C_{n_4,0} = 1$

$C_{n_4,1} = C_{n_4-1,1} + A_{n_4}$

$C_{n_4,k} = C_{n_4-1,k} + C_{n_4-1,k-1}A_{n_4} + C_{n_4-1,k-2}B_{n_4}; \quad (2 \le k \le 2_{n_4})$

and $C_{n_4-1,2n_4-1} = C_{n_4-1,2n_4} = 0$

$$\therefore R = \sum_{j=0}^{2n_4} C_{n_4,j}\theta^j \tag{2.5}$$

Substituting (2.3), (2.4) and (2.5) in (2.2), we get the joint pdf as

$$\begin{aligned}
f(\vec{x},\vec{y},p_1,p_2,\theta) &= \frac{p_1^{n_3+n_4+a_1-1}(1-p_1)^{n_1+n_2+b_1-1}}{\beta(a_1,b_1)}\frac{p_2^{n_2+n_4+a_2-1}\ (1-p_2)^{n_1+n_3+b_2-1}}{\beta(a_2,b_2)} \\
&\quad e^{-T_{3x}-T_{2y}-T_{4x}-T_{4y}}\sum_{j=0}^{2n_4}\frac{C_{n_4,j}\theta^j}{\beta(c,d)}\sum_{i=0}^{\infty}\frac{(-T_{4xy})^i\theta^i}{i!}\theta^{c-1}(1-\theta)^{d-1} \qquad ;n_4>0 \\
&= \frac{p_1^{n_3+n_4+a_1-1}\ (1-p_1)^{n_1+n_2+b_1-1}}{\beta(a_1,b_1)}\quad\frac{p_2^{n_2+n_4+a_2-1}(1-p_2)^{n_1+n_3+b_2-1}}{\beta(a_2,b_2)} \\
&\quad e^{-T_{3x}-T_{2y}}\frac{\theta^{c-1}(1-\theta)^{d-1}}{\beta(c,d)} \qquad ;n_4=0 \\
&\quad x,y\geq 0, 0<p_1,p_2<1, 0\leq\theta\leq 1, a_1,b_1,a_2,b_2,c,d>0.
\end{aligned}$$

To obtain Bayes' estimators of p_1, p_2 and θ we proceed as follows:
The marginal pdf of $\vec{x},\vec{y}$ is given by

$$\begin{aligned}
f(\vec{x},\vec{y}) &= \int_0^1\int_0^1\int_0^1 f(\vec{x},\vec{y},p_1,p_2,\theta)\ dp_1\ dp_2\ d\theta \\
&= \frac{e^{-T_{3x}-T_{2y}-T_{4x}-T_{4y}}}{\beta(a_1,b_1)\beta(a_2,b_2)\beta(c,d)}\int_0^1 p_1^{n_3+n_4+a_1-1}(1-p_1)^{n_1+n_2+b_1-1}\ dp_1 \\
&\qquad \int_0^1 p_2^{n_2+n_4+a_2-1}(1-p_2)^{n_1+n_3+b_2-1}\ dp_2 \\
&\qquad \sum_{j=0}^{2n_4}C_{n_4,j}\qquad\sum_{i=0}^{\infty}\frac{(-T_{4xy})^i}{i!}\int_0^1\theta^{i+j+c-1}(1-\theta)^{d-1}\ d\theta \\
&= \frac{e^{-T_{3x}-T_{2y}-T_{4x}-T_{4y}}}{\beta(a_1,b_1)\beta(a_2,b_2)\beta(c,d)}\ \beta(n_3+n_4+a_1,n_1+n_2+b_1) \\
&\qquad \beta(n_2+n_4+a_2,n_1+n_3+b_2)\ \sum_{i=0}^{\infty}\ \sum_{j=0}^{2n_4}C_{n_4,j}\ \beta(i+j+c,d)\frac{(-T_{4xy})^i}{i!} \\
&\qquad\qquad \text{if } n_4>0, x\geq 0, y\geq 0 \\
&\qquad\qquad 0<p_1,p_2<1, 0\leq\theta\leq 1, a_1,b_1,a_2,b_2,c,d>0. \\
&= e^{-T_{3x}-T_{2y}}\frac{\beta(n_3+a_1,n_1+n_2+b_1)\ \beta(n_2+a_2,n_1+n_3+b_2)}{\beta(a_1,b_1)\ \beta(a_2,b_2)} \\
&\quad \text{if } n_4=0, x\geq 0, y\geq 0. \\
&\quad a_1,b_1,a_2,b_2>0
\end{aligned}$$

The joint posterior distribution $h(p_1,p_2,\theta\mid\vec{x},\vec{y})$ is

$$h(p_1,p_2,\theta\mid\vec{x},\vec{y}) = \frac{f(\vec{x},\vec{y},p_1,p_2,\theta)}{f(x,y)}$$

$$= \left\{ \begin{array}{l} \dfrac{ \dfrac{p_1^{n_3+n_4+a_1-1}(1-p_1)^{n_1+n_2+b_1-1}}{\beta(a_1,b_1)} \quad \dfrac{p_2^{n_2+n_4+a_2-1}(1-p_2)^{n_1+n_3+b_2-1}}{\beta(a_2,b_2)} \\ e^{-T_{3x}-T_{2y}-T_{4x}-T_{4y}} \quad \displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \; \dfrac{\left(\dfrac{\sum_{i=0}^{\infty} (-1)^i \, {T_{4xy}}^i}{i!} \right) \theta^{i+j+c-1} \, (1-\theta)^{d-1}}{\beta(c,d)} }{ e^{-T_{3x}-T_{2y}-T_{4x}-T_{4y}} \; \dfrac{\beta(n_3+n_4+a_1, n_1+n_2+b_1)}{\beta(a_1,b_1)} \; \dfrac{\beta(n_2+n_4+a_2, n_1+n_3+b_2)}{\beta(a_2,b_2)} \\ \displaystyle\sum_{j=0}^{2n_4} \dfrac{C_{n_4,j}}{\beta(c,d)} \quad \sum_{c=0}^{\infty} \dfrac{(-1)^i {T_{4xy}}^i}{i!} \quad \beta(i+j+c,d) } \end{array} \right.$$

which is equal to

$$\left. \begin{array}{l} \dfrac{ p_1^{n_3+n_4+a_1-1} \, (1-p_1)^{n_1+n_2+b_1-1} \, p_2^{n_2+n_4+a_2-1} \, (1-p_2)^{n_1+n_3+b_2-1} \\ \displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \quad \sum_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \quad \theta^{i+j+c-1} \quad (1-\theta)^{d-1} }{ \beta(n_3+n_4+a_1, n_1+n_2+b_1) \quad \beta(n_2+n_4+a_2, n_1+n_3+b_2) \\ \displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \quad \sum_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \quad \beta(i+j+c,d) } \\ \qquad \text{if } n_4 > 0, x \geq 0, y \geq 0, 0 < p_1, p_2 < 1, 0 \leq \theta \leq 1 \\ \qquad a_1, b_1, a_2, b_2, c, d > 0 \end{array} \right\}$$

$$= \left\{ \begin{array}{l} \dfrac{p_1^{n_3+n_4+a_1-1} \quad (1-p_1)^{n_1+n_2+b_1-1}}{\beta(n_3+a_1, n_1+n_2+b_1)} \quad \dfrac{p_2^{n_2+n_4+a_2-1} \quad (1-p_2)^{n_1+n_3+b_2-1}}{\beta(n_2+a_2, n_1+n_3+b_2)} \quad \dfrac{\theta^{c-1}(1-\theta)^{d-1}}{\beta(c,d)} \\ \qquad \text{if } n_4 = 0, 0 < p_1, p_2 < 1, a_1, b_1, a_2, b_2 > 0. \end{array} \right.$$

The marginal posterior pdf of p_1, p_2 and θ given $(\vec{x}, \, \vec{y})$ are respectively given as follows:

$$h(p_1 \mid \vec{x}, \vec{y}) = \int_0^1 \int_0^1 h(p_1, p_2, \theta \mid \vec{x}, \vec{y}) \; dp_2 \; d\theta$$

which on simplification

$$= \frac{p_1^{n_3+n_4+a_1-1} \; (1-p_1)^{n_1+n_2+b_1-1}}{\beta(n_3+n_4+a_1, n_1+n_2+b_1)} \qquad ; n_4 \geq 0, 0 < p_1 < 1, a_1, b_1 > 0$$

similarly

$$h(p_2 \mid \vec{x}, \vec{y}) = \frac{p_2^{n_2+n_4+a_2-1} \; (1-p_2)^{n_1+n_3+b_2-1}}{\beta(n_2+n_4+a_2, n_1+n_3+b_2)} \qquad ; n_4 \geq 0, 0 < p_2 < 1, a_2, b_2 > 0$$

and similarly

$$
\begin{aligned}
h(\theta \mid \vec{x},\vec{y}) &= \int_0^1 \int_0^1 h(p_1,p_2,\theta \mid \vec{x},\vec{y})\; dp_1\; dp_2 \\
&= \frac{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \quad \theta^{i+j+c-1}\;(1-\theta)^{d-1}}{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \quad \beta(i+j+c,d)} \\
& \qquad\qquad\qquad\qquad\qquad\qquad ;n_4>0, x\geq 0, y\geq 0, 0\leq\theta\leq 1, c,d>0. \\
&= 1 \qquad\qquad\qquad\qquad\qquad\qquad ;n_4=0.
\end{aligned}
$$

It is well known that if the loss function is squared error, then the Bayes estimator is a mean of the posterior distribution.

Thus, Bayes' estimator for p_1 and p_2 are given as follows:

$$
\hat{p_1} = \int_0^1 p_1\; h(p_1 \mid \vec{x},\vec{y})\; dp_1 = \frac{n_3+n_4+a_1}{n+a_1+b_1}
$$

$$
\hat{p_2} = \int_0^1 p_2\; h(p_2 \mid \vec{x},\vec{y})\; dp_2 = \frac{n_2+n_4+a_2}{n+a_2+b_2}
$$

and Bayes' estimator for θ is given as follows:

$$
\begin{aligned}
\hat{\theta} &= E(\theta \mid \vec{x},\vec{y}) = \int_0^1 \theta\; h(\theta \mid \vec{x},\vec{y})\; d\theta \\
&= \frac{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-1)^i {T_{4xy}}^i}{i!} \displaystyle\int_0^1 \theta^{i+j+c+1-1}(1-\theta)^{d-1}\; d\theta}{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-1)^i {T_{4xy}}^i}{i!} \;\beta(i+j+c,d)} \\
&= \frac{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-1)^i {T_{4xy}}^i}{i!}\beta(i+j+c+1,d)}{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-1)^i {T_{4xy}}^i}{i!}\beta(i+j+c,d)} \\
&= \frac{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \dfrac{\Gamma(i+j+c+1)\Gamma(d)}{\Gamma(i+j+c+d+1)}}{\sum\limits_{j=0}^{2n_4} C_{n_4,j} \quad \sum\limits_{i=0}^{\infty} \dfrac{(-T_{4xy})^i}{i!} \dfrac{\Gamma(i+j+c)\Gamma(d)}{\Gamma(i+j+c+d)}}
\end{aligned}
$$

Multiplying and dividing numerator by $\dfrac{\Gamma(j+c+1)}{\Gamma(j+c+d+1)}$

and multiplying and dividing denominator by $\dfrac{\Gamma(j+c)}{\Gamma(j+c+d)}$we get,

$$= \frac{\displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \frac{\Gamma(j+c+1)}{\Gamma(j+c+d+1)} \quad \sum_{i=0}^{\infty} \frac{\Gamma(i+j+c+1)\Gamma(j+c+d+1)}{\Gamma(j+c+1)\Gamma(i+j+c+d+1)} \left(\frac{-T_{4xy}{}^{i}}{i!} \right)}{\displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \frac{\Gamma(j+c)}{\Gamma(j+c+d)} \quad \sum_{i=0}^{\infty} \frac{\Gamma(i+j+c)\Gamma(j+c+d)}{\Gamma(j+c)\Gamma(i+j+c+d)} \left(\frac{-T_{4xy}{}^{i}}{i} \right)}$$

$$= \frac{\displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \frac{\Gamma(j+c+1)}{\Gamma(j+c+d+1)} \quad {}_1F_1\Big(j+c+1; j+c+d+1; -T_{4xy}\Big)}{\displaystyle\sum_{j=0}^{2n_4} C_{n_4,j} \frac{\Gamma(j+c)}{\Gamma(j+c+d)} \quad {}_1F_1\Big(j+c; j+c+d; -T_{4xy}\Big)} \quad ; n_4 > 0$$

$$\hat{\theta} = \int_0^1 \frac{\theta \ \theta^{c-1}(1-\theta)^{d-1}}{\beta(c,d)} \ d\theta = \frac{c}{c+d} \quad ; \text{if } n_4 = 0$$

where, ${}_1F_1$ is called a confluent hypergeometric function and is computed in Abramouitz and Stegun [14] or can be computed using "hypergeometric 1F19a,b,c)" in R code.

The bias and MSE of the Bayes estimator $\hat{p_1}$ are as follows:

From (2.1) it is clear that $(N_3 + N_4) \sim$ Binomial $(n, p_1q_2 + p_1p_2 = p_1)$. Hence,

$$E(\hat{p_1}) = E\left(\frac{n_3 + n_4 + a_1}{n + a_1 + b_1}\right) = \frac{np_1 + a_1}{n + a_1 + b_1}$$

$$V(\hat{p_1}) = V\left(\frac{n_3 + n_4 + a_1}{n + a_1 + b_1}\right) = \frac{np_1q_1}{(n + a_1 + b_1)^2}$$

Thus, the bias of $\hat{p_1}$ is

$$Bias(\hat{p_1}) = E(\hat{p_1}) - p_1 = \frac{np_1 + a_1}{n + a_1 + b_1} - p_1 = \frac{a_1q_1 - b_1p_1}{n + a_1 + b_1} \tag{2.6}$$

$$\text{and } MSE(\hat{p_1}) = V(\hat{p_1}) + \Big[bias(\hat{p_1})\Big]^2 = \frac{np_1q_1 + (a_1q_1 - b_1p_1)^2}{(n + a_1 + b_1)^2} \tag{2.7}$$

Similarly,

$$Bias(\hat{p_2}) = \frac{a_2q_2 - b_2p_2}{n + a_2 + b_2} \tag{2.8}$$

$$\text{and } MSE(\hat{p_2}) = \frac{np_2q_2 + (a_2q_2 - b_2p_2)^2}{(n + a_2 + b_2)^2} \tag{2.9}$$

Thus, the bias and MSE of $\hat{p_1}$ and $\hat{p_2}$ are in closed form. Though Bayes estimator of θ appears in closed form its bias and MSE cannot be obtained in closed form. Hence as a computational illustration, we compute bias and MSE of the Bayes estimators of p_1, p_2 and θ using simulation which are presented in Table 1. The Table is prepared for Simulation size $N = 1000$, sample sizes $(n = 5, 10, 15), p_1 = 0.5, (p_2 = 0.5, 0.7$ and $0.9)$ and $(\theta = 0.3, 0.7, 0.9)$. We consider $a_1 = 2, b_1 = 3, a_2 = 2, b_2 = 3, c = 3, d = 2$.

From Table 1, it is seen that bias and MSE of $\hat{p_1}$ and $\hat{p_2}$ decreases as n increases for all values of θ. However, bias and MSE of $\hat{\theta}$ decreases as n increases for the values of θ near to zero (0.3) or near to $1(0.9)$ but for central values (0.7) of θ both bias and MSE are smaller with slight fluctuation.

3 Tables

Table 1: Bias and Mean squared error for the Bayes estimators of p_1, p_2 and θ of BVEDP $(p1, p2, 1, 1, \theta)$. Simulation size, $N = 1000$. $a_1 = 2, b_1 = 3, a_2 = 2, b_2 = 3, c = 3, d = 2$.

n	Bias(p_1)	Bias(p_2)	Bias(θ)	MSE(p_1)	MSE(p_2)	MSE(θ)
$p_1 = 0.5, p_2 = 0.5, \theta = 0.3$						
5	-0.0461	-0.0581	0.2943	0.0146	0.0158	0.0877
10	-0.0327	-0.0388	0.2889	0.0126	0.0127	0.0854
15	-0.0309	-0.0274	0.2822	0.0109	0.0099	0.0830
$p_1 = 0.5, p_2 = 0.5, \theta = 0.7$						
5	-0.0438	-0.0528	-0.0982	0.0145	0.0148	0.0103
10	-0.0350	-0.0318	-0.0972	0.0124	0.0121	0.0106
15	-0.0267	-0.0252	-0.0941	0.0097	0.0103	0.0106
$p_1 = 0.5, p_2 = 0.5, \theta = 0.9$						
5	-0.0494	-0.0473	-0.2943	0.0145	0.0146	0.0870
10	-0.0354	-0.0349	-0.2889	0.0128	0.0119	0.0844
15	-0.0226	-0.0258	-0.2865	0.0096	0.0103	0.0835
$p_1 = 0.5, p_2 = 0.7, \theta = 0.3$						
5	-0.0459	-0.1461	0.2911	0.0152	0.0314	0.0866
10	-0.0293	-0.1017	0.2858	0.0129	0.0204	0.0846
15	-0.0246	-0.0731	0.2761	0.0113	0.0128	0.0807
$p_1 = 0.5, p_2 = 0.7, \theta = 0.7$						
5	-0.0430	-0.1499	-0.0957	0.0140	0.0327	0.0100
10	-0.0349	-0.1041	-0.0940	0.0128	0.0198	0.0105
15	-0.0260	-0.0748	-0.0929	0.0106	0.0136	0.0107
$p_1 = 0.5, p_2 = 0.7, \theta = 0.9$						
5	-0.0517	-0.1470	-0.2941	0.0149	0.0317	0.0871
10	-0.0316	-0.1001	-0.2856	0.0121	0.0194	0.0829
15	-0.0272	-0.0752	-0.2800	0.0101	0.0142	0.0801
$p_1 = 0.5, p_2 = 0.9, \theta = 0.3$						
5	-0.0519	-0.2490	0.2898	0.0152	0.0663	0.0860
10	-0.0271	-0.1685	0.2781	0.0122	0.0326	0.0818
15	-0.0287	-0.1251	0.2702	0.0104	0.0191	0.0787
$p_1 = 0.5, p_2 = 0.9, \theta = 0.7$						
5	-0.0492	-0.2502	-0.0954	0.0146	0.0671	0.0100
10	-0.0323	-0.1689	-0.0896	0.0112	0.0326	0.0097
15	-0.0234	-0.1237	-0.0917	0.0102	0.0184	0.0115
$p_1 = 0.5, p_2 = 0.9, \theta = 0.9$						
5	-0.0463	-0.2494	-0.2920	0.0148	0.0667	0.0862
10	-0.0325	-0.1633	-0.2842	0.0122	0.0304	0.0823
15	-0.0237	-0.1235	-0.2750	0.0103	0.0187	0.0777

4 Conclusion

In this paper we have obtained Bayes estimators $\hat{p_1}, \hat{p_2}$ and $\hat{\theta}$ for squared error loss function for the parameters of Gumbel's (1960) type-I bivariate exponential distribution with positive probability at instantaneous failure given in (1.4). bias, MSE and Bayes risk of $\hat{p_1}$ and $\hat{p_2}$ are also obtained, which are in closed form. Though $\hat{\theta}$ is in closed form, its bias, MSE and Bayes risk are difficult to obtain, analytically. Hence a simulation study is carried out and tables are prepared for bias and MSE of $\hat{p_1}, \hat{p_2}$ and $\hat{\theta}$.

Acknowledgements

Authors are thankful to the Referees and Editors for their valuable suggestions and comments. Authors are also thankful to Dr. U. J. Dixit and Dr. R. N. Rattihalli for valuable discussion about the problem.

References

[1] E. J. Gumbel. "Bivariate exponential distributions." In: *Journal of the American Statistical Association* 55, No. 292. (1960), pp. 698–707.

[2] R.M. Kleyle and R.C. Dahiya. "Estimation of Parameters of Mixed Failure Time Distribution from Censored Data." In: *Communication in Statistics - Theory and Methods* 4(9) (1975), pp. 873–882.

[3] V. D. (Alias Dixit V. U.) Jayade and M.S. Prasad. "Estimation of parameters of mixed failure time distribution." In: *Communication in Statistics - Theory and Methods* 19(12) (1990), pp. 4667–4677.

[4] V. U. (Nee Jayade V. D.) Dixit. *Statistical Inference for AR(1) Process with mixed errors. Unpublished PhD. thesis.* Shivaji University, Kolhapur, 1993.

[5] Dixit V. *Tests of hypotheses about the parameters of mixed failure time distribution, Statistical Inference and Design of Experiments*. Edited by U. J. Dixit and M. R. Satam. Narosa Publishing House, 1999, pp. 20–27.

[6] R.L. Shinde and A. Shanubhogue. "Estimation of parameters and the mean life of a mixed failure time distribution." In: *Communication in Statistics - Theory and Methods* 29(11) (2000), pp. 2621–2642.

[7] Dixit V. U. "Estimation of parameters of mixed failure time distribution based on extended modified sampling scheme." In: *Communication in Statistics - Theory and Methods* 32 (10) (2003), pp. 1911–1923.

[8] U. J. Dixit and D. Karadkar (alias Shilpa Khare) Shilpa. *Estimation of the parameters of a Bivariate Exponential distribution; Statistical Inference and Design of Experiments*. Edited by U. J. Dixit and M. R. Satam. Narosa Publishing House, 1999, pp. 13–19.

[9] Shilpa N. Khare. *Estimation of the parameters of bivariate exponential distribution, unpublished PhD thesis.* University of Mumbai, Mumbai, 2002.

[10] Khare S. N. Dixit V. U. and Mooghabi Jabbari. "Bayes Estimation of parameters of Gumbel's Bivariate Exponential distribution." In: *Aligarh Journal of Statistics* 30 (2010), pp. 77–90.

[11] Saju V. G. and Dixit V. U. "Statistical Inference for Gumbel's Type I Bivariate Exponential distribution with positive probability at instantaneous failure." In: *Journal of the Indian Statistical Association* 61 No.1 and 2 (2023), pp. 35–53.

[12] Singh S. Pathak A. Kumar M. and Singh U. "Assessing the effect of E-Bayesian inference for Poisson inverse exponential distribution parameters under different loss functions and its application." In: *Communication in Statistics - Theory and Methods* 4(9) (2022), pp. 1–43.

[13] Singh S. Pathak A. Kumar M. and Singh U. "Bayesian inference: Weibull Poisson model for censored data using the expectation maximization algorithm and its application to bladder cancer data." In: *Journal of Applied Statistics* (2022), pp. 1–23.

[14] Abramowitz and Stegun I. *Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematical series*. 55th ed. US Government Printing Office., 1972.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

A Note on Cyclic and Partial Cyclic Solutions of Block Designs

Shyam Saurabh [1]

1. *Tata College, Kolhan University, Chaibasa*

Received: Dec 2023 / Accepted: July 2024

Abstract

A necessary and sufficient condition for a block design $D(v, b, r, k)$ to have a cyclic or partial cyclic solution is obtained. A 5–resolvable solution of regular group divisible design $R152a : v = 22$, $r = 10$, $k = 5$, $b = 44$, $\lambda_1 = 0$, $\lambda_2 = 2$, $m = 11$, $n = 2$ is also obtained using circulant matrices. This solution is not reported in [16].

Keywords: Cyclic and Partial cyclic solutions; Circulant matrices; Regular group divisible design; Resolvability

1 Introduction

Cyclic and partial cyclic solutions of two associate classes partially balanced incomplete block designs (PBIBDs) have been reported by Clatworthy [2] under the range of $2 \leq r \leq 10$ whenever possible. Later Hall [15] reported cyclic and partial cyclic solutions of balanced incomplete block designs (BIBDs) under the range of $3 \leq r \leq 20; k \leq v/2$ using automorphism. A recent survey on cyclic and partial cyclic solutions of block designs under the range of $r, k \leq 10$ may be found in [21]. They reported cyclic and partial cyclic solutions of block designs with higher efficiencies for the same v, b, r, k available in the literature. We go for partial solution of a block design $D(v, b, r, k)$ when cyclic solution is not available for the same v, b, r, k.

Partial cyclic solutions of block designs have been studied by Dey and Nigam [8], Mukerjee et al. [12], Dey et al. [7], Midha et al. [17], Saurabh [20], among others and for details on cyclic solutions, see [11], [12], [14], [15], [18], [19] and [24]. Generalized cyclic designs are further generalization of cyclic and partial cyclic designs, see [10].

2 Definitions

For the definitions of balanced incomplete block design, group divisible (GD) design and triangular design, see [6], [5], [22]. Some other relevant definitions in context of the paper are as follows:

Cyclic and partial cyclic designs

A block design $D(v, b, r, k)$ is cyclic if its solution may be obtained by adding the elements of a cyclic group $Z_v = \{0, 1, 2, \ldots, v\} mod v$ to the initial blocks of the design whereas a design is partial cyclic if its solution may be obtained by developing the initial blocks under a partial cycle: $1 \leftrightarrow q, q+1 \leftrightarrow 2q, \ldots, [q(p-1)+1] \leftrightarrow v = pq$ of length q where $(1 \leftrightarrow q) \leftrightarrow 1 \rightarrow 2, 2 \rightarrow 3, \ldots, (q-1) \rightarrow q, q \rightarrow 1$.

Example 1. *A cyclic solution of the BIBD: $v = b = 7$, $r = k = 3$, $\lambda = 1$ can be obtained by developing the initial block $(1, 2, 4)$ under addition modulo 7.*

Example 2. *A partial cyclic solution the GD design* $R80 : v = 14, r = 9, k = 3, b = 42, \lambda_1 = 6, \lambda_2 = 1, m = 7, n = 2$ *may be obtained by developing the initial blocks:* $(1,2,8);(1,8,9);(1,3,8);(1,8,10);(1,4,8);(1,8,11)$ *under a partial cycle* $1 \leftrightarrow 7, 8 \leftrightarrow 14$ *of length* 7 *[see [2]].*

Circulant matrix

An $n \times n$ matrix $C = [c_{ij}]_{(0 \leq i,j \leq n-1)}$ where $c_{ij} = c_{j-i(modn)}$ is called a circulant matrix of order n

i. e. $C = \begin{pmatrix} c_0 & c_1 & c_2 & \dots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \dots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \dots & c_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & \dots & c_0 \end{pmatrix} = circ(c_0, c_1, \dots, c_{n-1}).$

Clearly the elements of each row of C are identical to those of the previous row but are moved one position to the right and wrapped around anti-clockwise. It is easy to verify that a circulant matrix may also be generated using the first column by shifting the elements downword and wrapped around at the top. For details on circulant matrices, see [3].

Resolvable design

A block design $D(v,b,r,k)$ is α-resolvable if its incidence matrix N may be partitioned in to submatrices as: $N = (N_1|N_2|\dots|N_t)$ where each $N_i (1 \leq i \leq t)$ is a $v \times v\alpha/k$ matrix such that each row sum of N_i is α and for $\alpha = 1$, the design is resolvable.

Notations: I_n is the identity matrix of order n, $\alpha = circ(010\dots0)$ is a circulant matrix of order n such that $\alpha^n = I_n$ and $v|b$ denotes v divides b.

3 A necessary and sufficient condition

Theorem 1. *(a) If the incidence matrix of a block design* $D(v,b,r,k)$ *may be partitioned into circulant submatrices of order* q *where* $q|v, q|b$ *then the design has a partial cyclic solution.*

(b) A block design $D(v,b,r,k)$ *has a cyclic solution if and only if its incidence matrix may be partitioned into circulant submatrices of order* $v(v|b)$.

Proof. Suppose the incidence matrix N of a block design $D(v,b,r,k)$ may be partitioned as:

$$N = [N_{ij}]_{\substack{1 \leq i \leq s \\ 1 \leq j \leq t}} = \begin{pmatrix} N_{11} & N_{12} & \dots & N_{1t} \\ N_{21} & N_{22} & \dots & N_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ N_{s1} & N_{s2} & \dots & N_{st} \end{pmatrix} = (M_1|M_2|\dots|M_t),$$

where each N_{ij} is a circulant matrix of order q such that $q|v, q|b$ and $M_i = \begin{pmatrix} N_{1i} \\ N_{2i} \\ \vdots \\ N_{si} \end{pmatrix} (1 \leq i \leq t)$. Clearly each N_{ij} can be obtained using its first column by shifting the elements downward and wrapped around at the top. Since the ones in first column of $N_{11}, N_{21}, \dots, N_{s1}$ correspond to the number of treatments in the block B_1 (say), all the remaining $(q-1)$ blocks corresponding to M_1 can be developed under the partial cycle $1 \leftrightarrow q, (q+1) \leftrightarrow 2q, \dots, q(p-1)+1 \leftrightarrow v = pq$ of length q which is the size of each N_{ij}. The same argument applies to the block matrices $M_2, M_3, \dots, M_t$. Hence the design has a partial cyclic solution.

Further suppose that the incidence matrix N of a block design $D(v, b, r, k)$ may be partitioned as: $N = (M_1|M_2|\ldots|M_t); b = tv$ where each $M_i(1 \leq i \leq t)$ is a circulant matrix of order v, then the design has clearly a cyclic solution.

Conversely if a cyclic design is obtained by developing the initial blocks: $(\theta_1^1, \theta_2^1, \ldots, \theta_k^1)$; $(\theta_1^2, \theta_2^2, \ldots, \theta_k^2)$;$\ldots, (\theta_1^t, \theta_2^t, \ldots, \theta_k^t)(modv)$ then its incidence matrix N may be expressed as: $N = (\alpha^{\theta_1^1} + \alpha^{\theta_2^1} + \cdots + \alpha^{\theta_k^1}|\alpha^{\theta_1^2} + \alpha^{\theta_2^2} + \cdots + \alpha^{\theta_k^2}|\ldots|\alpha^{\theta_1^t} + \alpha^{\theta_2^t} + \cdots + \alpha^{\theta_k^t})$, where $\alpha = circ(010\ldots0)$ is circulant matrix of order v such that $\alpha^v = I_v$. □

Example 3. *A cyclic solution of the GD design $R83 : v = 15, r = 9, k = 3, b = 45, \lambda_1 = 2, \lambda_2 = 1, m = 3, n = 5$ is obtained by developing the initial blocks: $[(1, 7, 13); (1, 4, 5); (1, 3, 8)](mod15)$ [see [2]]. Then its incidence matrix is given as: $N = (\alpha + \alpha^7 + \alpha^{13}|\alpha + \alpha^4 + \alpha^5|\alpha + \alpha^3 + \alpha^8)$, where $\alpha = circ(010\ldots0)$ is circulant matrix of order 15 such that $\alpha^{15} = I_{15}$ which may be easily verified.*

Example 4. *A cyclic solution of the BIBD with parameters: $v = 41, r = 10, k = 5, b = 82, \lambda = 1$ is obtained be developing the initial blocks $[(1, 10, 16, 18, 37); (5, 8, 9, 21, 39)](mod41)$, see [15]. Then its incidence matrix may be expressed as: $N = (\alpha + \alpha^{10} + \alpha^{16} + \alpha^{18} + \alpha^{37}|\alpha^5 + \alpha^8 + \alpha^9 + \alpha^{21} + \alpha^{39})$, where $\alpha = circ(010\ldots0)$ is circulant matrix of order 41 such that $\alpha^{41} = I_{41}$which may be easily verified.*

4 A 5-resolvable regular group divisible design

A 5-resolvable solution of $R152a$ [see [9]] using circulant matrices is presented below. This solution is not reported in [16]. $R152a : v = 22, r = 10, k = 5, b = 44, \lambda_1 = 0, \lambda_2 = 2, m = 11, n = 2$.

$$N = \left(\begin{array}{cc|cc} \alpha + \alpha^2 + \alpha^3 + \alpha^6 & \alpha^9 & \alpha + \alpha^3 + \alpha^8 & \alpha^2 + \alpha^{10} \\ \alpha^9 & \alpha + \alpha^2 + \alpha^3 + \alpha^6 & \alpha^2 + \alpha^{10} & \alpha + \alpha^3 + \alpha^8 \end{array}\right) = \left(S|T\right),$$

where $\alpha = circ(010\ldots0)$ is circulant matrix of order 11 such that $\alpha^{11} = I_{11}$. Since each row sum of partitioned block matrices S and T is 5, the design is $5-$ resolvable. The resolution classes are given below. The blocks can be obtained by developing the initial blocks under a partial cycle $1 \leftrightarrow 11$,$12 \leftrightarrow 22$ of length 11:

No.	Resolution Classes	Initial blocks
1	RI	(6, 9, 10, 11, 14); (3, 17, 20, 21, 22)
2	RII	(4, 9, 11, 13, 21); (2, 10, 15, 20, 22)

Table 1: Partial cyclic solution of R152a

The complete solution is as given below:

Resolution I	**Resolution** II
(6, 9, 10, 11, 14); (3, 17, 20, 21, 22)	(4, 9, 11, 13, 21); (2, 10, 15, 20, 22)
(7, 10, 11, 1, 15); (4, 18, 21, 22, 12)	(5, 10, 1, 14, 22); (3, 11, 16, 21, 12)
(8, 11, 1, 2, 16); (5, 19, 22, 12, 13)	(6, 11, 2, 15, 12); (4, 1, 17, 22, 13)
(9, 1, 2, 3, 17); (6, 20, 12, 13, 14)	(7, 1, 3, 16, 13); (5, 2, 18, 12, 14)
(10, 2, 3, 4, 18); (7, 21, 13, 14, 15)	(8, 2, 4, 17, 14); (6, 3, 19, 13, 15)
(11, 3, 4, 5, 19); (8, 22, 14, 15, 16)	(9, 3, 5, 18, 15); (7, 4, 20, 14, 16)
(1, 4, 5, 6, 20); (9, 12, 15, 16, 17)	(10, 4, 6, 19, 16); (8, 5, 21, 15, 17)
(2, 5, 6, 7, 21); (10, 13, 16, 17, 18)	(11, 5, 7, 20, 17); (9, 6, 22, 16, 18)
(3, 6, 7, 8, 22); (11, 14, 17, 18, 19)	(1, 6, 8, 21, 18); (10, 7, 12, 17, 19)
(4, 7, 8, 9, 12); (1, 15, 18, 19, 20)	(2, 7, 9, 22, 19); (11, 8, 13, 18, 20)
(5, 8, 9, 10, 13); (2, 16, 19, 20, 21)	(3, 8, 10, 12, 20); (1, 9, 14, 19, 21)

The 11×2 GD scheme is given as transpose of the array:

$$\begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 \end{array}$$

5 Conclusion

In this note a necessary and sufficient condition for a block design $D(v, b, r, k)$ to have a cyclic solution is obtained whereas only a sufficient condition for a block design $D(v, b, r, k)$ to have a partial cyclic solution is obtained using circulant matrices. The cyclic and partial cyclic solutions of designs are important from statistical point of view. Cyclic designs can be used for partial confounding in factorial experiments and in the constructions of LDPC codes [see [13], [4] and [25]].

The incidence matrices of all the cyclic group divisible designs given in [2], [9] and the cyclic BIBDs given in [15] may be easily written using circulant matrices. It would be interesting to generate the incidence matrix of a partial cyclic design using initial blocks and circulant matrices. A 5–resolvable solution of regular group divisible design $R152a : v = 22, r = 10, k = 5, b = 44, \lambda_1 = 0, \lambda_2 = 2, m = 11, n = 2$ is also obtained. This solution is not reported in [16].

Saurabh and Sinha [23] obtained some series of L_2-type designs using circulant matrices. Their solutions may also be put in partial cyclic form. Cyclic solutions of triangular and L_2-type designs are not available in the literature, see [2], [21] and elsewhere. It would be interesting to obtain some series of cyclic triangular and L_2-type designs however a generalized cyclic solution of the triangular design $T9 : v = b = 10, r = k = 3, \lambda_1 = 1, \lambda_2 = 0, n_1 = 6, n_2 = 3$ is reported in [1]. The solution is obtained by adding $0, 2, 4, 6, 8$ successively to the initial blocks $(0, 3, 7)$, $(2, 3, 4)$ where the corresponding triangular scheme is represented as:

$$\begin{array}{ccccc} * & 0 & 1 & 7 & 8 \\ 0 & * & 2 & 3 & 9 \\ 1 & 2 & * & 4 & 5 \\ 7 & 3 & 4 & * & 6 \\ 8 & 9 & 5 & 6 & * \end{array}.$$

References

[1] B Adhikary. “On a new class of two associate (para–cyclic) association schemes.” In: *Calcutta Statistical Association Bulletin* 18 (1969), pp. 43–49.

[2] W H Clatworthy. “Tables of Two–Associate–Class Partially Balanced Designs.” In: *National Bureau of Standards (U.S.), Applied Mathematics, Series* 63 (1973).

[3] P J Davis. *Circulant matrices*. 2nd ed. American Mathematical Society, 2012.

[4] A M Dean and S M Lewis. “A unified theory for generalized cyclic designs.” In: *Journal of Statistical Planning and Inference* 4 (1980), pp. 13–23.

[5] A Dey. *Incomplete Block Designs*. New Delhi: Hindustan Book Agency, 2010.

[6] A Dey. *Theory of Block Designs*. New Delhi: Wiley Eastern, 1986.

[7] A Dey and K Balasubramanian. “Construction of some families of group divisible designs.” In: *Utilitas Mathematica* 40 (1991), pp. 283–290.

[8] A Dey and A K Nigam. “Constructions of group divisible designs.” In: *Journal of the Indian Society of Agricultural Statistics* 37 (1985), pp. 163–166.

[9] G H Freeman. “A cyclic method of constructing regular group divisible incomplete block designs.” In: *Biometrika* 63 (1976), pp. 555–558.

[10] R G Jarrett and W B Hall. “Generalized cyclic incomplete block designs.” In: *Biometrika* 65 (1978), pp. 397–401.

[11] M Jimbo. “Recursive constructions for cyclic BIB designs and their generalizations.” In: *Discrete Mathematics* 116 (1993), pp. 79–95.

[12] R Mukerjee M Jimbo and S Kageyama. “On cyclic semi–regular group divisible designs.” In: *Osaka Journal of Mathematics* 24 (1987), 395–407.

[13] J A John. “Generalized cyclic designs in factorial experiments.” In: *Biometrika* 60 (1973), pp. 55–63.

[14] J A John and E R Williams. *Cyclic and Computer Generated Designs*. London: Chapman and Hall, 1995.

[15] M Hall Jr. *Combinatorial Theory*. New York: John Wiley, 1998.

[16] S Saurabh K Sinha and M K Singh. “A Survey on Resolvable Solutions of Partially Balanced designs.” In: *Communications in Statistics-Theory and Methods* 52 (2023), pp. 1946–1962.

[17] C K Midha and A Dey. “Cyclic group divisible designs.” In: *Calcutta Statistical Association Bulletin* 45 (1995), pp. 179–180.

[18] D Preece P Dobcsanyi and L H Soicher. “On balanced incomplete block designs with repeated blocks.” In: *European Journal of Combinatorics* 28 (2007), pp. 1950–1970.

[19] D Raghavarao. *Constructions and Combinatorial Problems in Design of Experiments*. New York: John Wiley, 1971.

[20] S Saurabh. “Partial cyclic solution of a $2-(p^2, p, 1)$design.” In: *Calcutta Statistical Association Bulletin* 74 (2022), pp. 59–61.

[21] S Saurabh and K Sinha. “A survey on cyclic solutions of block designs.” In: *Statistics and Applications* 20 (2022), pp. 123–133.

[22] S Saurabh and K Sinha. “Matrix approaches to constructions of group divisible designs.” In: *Bulletin of the ICA* 97 (2023), pp. 83–105.

[23] S Saurabh and K Sinha. “Some matrix constructions of L_2-type Latin square designs.” In: *Bulletin of the ICA* 95 (2022), pp. 93–104.

[24] A P Street and D J Street. *Combinatorics of Experimental Designs*. New York: Oxford University Press, 1987.

[25] B V Vasic and O Milenkovic. "Combinatorial constructions of low–density parity–check codes for iterative decoding." In: *IEEE Transactions on Information Theory* 50.6 (2004), pp. 1156–1176.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Compatibility of conditional densities: Some useful results

Indranil Ghosh[1]

1. *Department of Mathematics and Statistics, University of North Carolina Wilmington*

Received: Feb 2024 / Accepted: Sep 2024

Abstract

Compatibility of conditional distributions in the discrete case has been adequately discussed, see [1, 3], [8], [9] and the references cited therein. However, not much has been discussed on the issue of compatibility between two or more probability distributions which are absolutely continuous in nature, except in [5]Arnold and Gokhale (1999, 2000) and in [1]. In this paper, simple results for compatibility between two conditional densities are provided in the form of theorem(s) and proof(s). In higher dimension (say, dimension 4) and in more general, $m \geq 5$ dimension joint distribution compatibility is discussed. Several of these results have been independently observed in [11] but no formal proofs are provided. This plays a major motivation for this article.

Keywords: Compatible conditionals; conditional densities.

1 Introduction

Identification of joint distributions by means of conditional densities has received considerable attention in the last decade or so. Among several potential applications, prominent applications can be found in the area of model building in classical statistical settings and in the elicitation and construction of a multiparameter prior distributions in Bayesian scenarios. For example suppose that $\underline{Y} = (Y_1, Y_2, \cdots, Y_p)$ is a p-dimensional random vector taking on values in the finite range set $\underline{\mathcal{Y}}_1 \times \underline{\mathcal{Y}}_2 \times \cdots \times \underline{\mathcal{Y}}_p$ where $\underline{\mathcal{Y}}_i$ denote the possible values of $Y_i, i = 1, 2, \cdots, p$. Efforts to ascertain an appropriate distribution for $\underline{Y}$ frequently involve acceptance or rejection of a series of bets about the stochastic behavior of $\underline{Y}$. Suppose that in this situation we are facing a question of whether or not to accept with odds 7 to 1 a bet that Y_1 is equal to 1. Then if we accept the bet then it puts a bound on the probability that Y=1. Discussion regarding compatibility under the discrete set-up has been adequately discussed in the literature. For example, Arnold and his co-authors in a series of papers (see, [3, 2, 1], [6] have discussed the theory and methodology related to compatibility of two discrete conditional distributions and in higher dimension with potential limitations in studying in higher dimension. Several other researchers have also looked into this problem, and some useful discussions can be found in the works of [7], [9], [10] and the references cited therein. However, not much discussion has been made in this direction when the conditionals are in the continuous domain. To the best of the authors' knowledge, two most notable references would be the paper by [4]; the books by [1] and [11] in which specific chapters have been devoted to address this matter. This paper frequently cites these two books as the results that presented in the form of theorem and proof structure can be found as statements and/or open problems that are left to readers to explore (except for the theorem and proof when the dimension is 4). This plays a major motivation for this article. The remainder of this paper is organized as follows. In Section 2 we provide some useful preliminaries related to measures of divergence. In Section 3 we provide the main result in which details of the proofs related to the several theorems related to the notion of compatibility in the bivariate, fourth dimension, and in general m-variate ($m \geq 5$) are given. Some illustrative examples are provided in Section 4. Finally, some concluding remarks are presented in Section 5.

2 Useful preliminaries

Here we begin our discussion on the theory of compatibility of two conditional densities in the bivariate set-up. An extension to the higher dimension can certainly be envisioned albeit notational complexity which will be discussed in a separate article. Suppose that (X, Y) be an absolutely continuous bivariate random variable with respect to some product measure $\nu_1 \times \nu_2$, on $S(X) \times S(Y)$, where $S(X)$ (and $S(Y)$) denotes the support set of $X(Y)$. For the sake of our discussion, one may assume $S(X) = S(Y) \in \mathbb{R}$. Next, we denote with families of candidate conditional densities (with respect to the measure $\nu_1 \times \nu_2$) by

$$a\left(x, y\right) = f_{X|Y}\left(x|y\right), \quad x \in S(X), y \in S(Y), \tag{1}$$

and

$$b\left(x, y\right) = f_{Y|X}\left(y|x\right), \quad x \in S(X), y \in S(Y), \tag{2}$$

In addition, we defined the incidence sets by writing the following

$W_a = \{(x, y) : a\left(x, y\right) > 0,\ W_b = \{(x, y) : b\left(x, y\right) > 0$. Then, according to Arnold et al. (1999), Theorem 1.2 (pp8), a joint density $f\left(x, y\right)$, with $a\left(x, y\right)$ and $b\left(x, y\right)$ as its conditional densities, will exist if and only if:

- $W_a = W_b = W$, say, the common incidence set,
- there exist functions, say, u and v such that for all $(x, y) \in W$

  $$\frac{a\left(x, y\right)}{b\left(x, y\right)} = \frac{u(x)}{v(y)}, \tag{3}$$

  where $\int_{S(X)} u(x) d\nu_1(x) < \infty$.

Regarding the proof of this theorem, see [1]. We will utilize this theorem in Section 4 to discuss some examples in this regard. It is to be noted that there are several different forms of the given condition in Eq. (3) exist in the literature, see [11], [7] and the references cited therein.

3 Main Result

We begin this section with the following theorem.

Theorem 1. Let (X, Y) be a bivariate random variable with joint p.d.f. $f\left(x, y\right)$. Let $f\left(x|y\right)$ and $f\left(y|x\right)$ be the conditional densities $f\left(x|y\right)$ of X given $Y = y$, and Y given $X = x$, respectively. Then,

(a) If $f\left(x_0|y\right) > 0$, for all y, then

$$f\left(x, y\right) \propto \frac{f\left(x|y\right) f\left(y|x_0\right)}{f\left(x_0|y\right)}$$

this result holds under the assumption that the conditional densities $f\left(x|y\right)$ are given for all y, and $f\left(y|x_0\right)$ is given for a particular fixed value of $X = x_0$. Therefore, it can be said that for any two absolutely continuous random variables, the set of conditional densities given one variable, and the other conditional density uniquely determines the associated bivariate probability distribution. This assertion has been independently obtained by [11], [1] .

The next result on compatible conditional densities (in the bivariate case as well) holds under the assumption that the conditional densities of $f\left(.|y\right)$ are available for all y, and vice versa. Then we have the following result.

(b) A sufficient condition for $f\left(x|y\right)$ and $f\left(y|x\right)$ to be compatible with a joint density for (X,Y) is that

$$\frac{f\left(x_2|y\right)f\left(y|x_1\right)}{f\left(x_1|y\right)f\left(y|x_2\right)}$$

does not depend on y for any choice of (x_1,x_2) such that $x_1 \neq x_2$.

Proof. For part [a], we start from the right hand side

$$\begin{aligned}\frac{f\left(x|y\right)f\left(y|x_0\right)}{f\left(x_0|y\right)} &= \frac{f\left(x,y\right)}{g(y)}\times\frac{f\left(x_0,y\right)}{f_1(x_0)}\times\frac{g(y)}{f\left(x_0,y\right)}\\ &= \frac{f\left(x|y\right)}{f_1(x_0)}\\ &\propto f\left(x|y\right), \qquad (4)\end{aligned}$$

where $g(y)$ represents the marginal density of Y and $f_1(x)$ represents the marginal density of X. Hence, the proof.

For part [b], we also start from the right hand side of the given expression. On simplifying and doing some basic algebra, it can be written as

$$\frac{f\left(x_2|y\right)f\left(y|x_1\right)}{f\left(x_1|y\right)f\left(y|x_2\right)} = \frac{h(x_2)}{h(x_1)},$$

which doesn't depend on y for any choices of (x_1,x_2) such that $x_1 \neq x_2$. This completes the proof. □

As a natural extension, compatibility in the multivariate scenario can also be represented in the form of the following theorem (as necessary and sufficient conditions). At first we provide the sketch of a proof in case the dimension is four.

Theorem 2. Suppose, the conditional densities of $f_{1|234}\left(.|y,z,w\right)$ are defined for all y,z,w; $f_{2|134}\left(.|x_0,z,w\right)$ are defined for all y,w and fixed $X=x_0$; $f_{3|124}\left(.|x_0,y,w_0\right)$ are defined for all y and fixed $X=x_0$, $W=w_0$; $f_{4|123}\left(.|x_0,y_0,z_0\right)$ are defined for fixed $X=x_0$, $Y=y_0$, $Z=z_0$. Furthermore, assume that $f_{1|234}\left(x_0|y,z,w\right)>0$ for all y,z,w; $f_{2|134}\left(y_0|x_0,z,w\right)>0$ for all z,w; and $f_{3|124}\left(z_0|x_0,y_0,w\right)>0$ for all w.

(a) Then, the joint density of $X,Y,ZandW$ can be written as

$$f_{1|234}\left(x,y,z,w\right)\propto\frac{\left[f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_0,z,w\right)f_{3|124}\left(z|x_0,y,w_0\right)f_{4|123}\left(w|x_0,y_0,z_0\right)\right]}{\left[f_{1|234}\left(x_0|y,z,w\right)f_{2|134}\left(y_0|x_0,z,w\right)f_{3|124}\left(z_0|x_0,y_0,w\right)\right]}. \tag{5}$$

with the proportionality constant equal to $f_{123}\left(x_0,y_0,z_0\right).$

(b) If the conditional densities of $f_{1|234}\left(.|y,z,w\right)$ are defined for all y,z,w; $f_{2|134}\left(.|x,z,w\right)$ are defined for all x,z,w; $f_{3|124}\left(.|x,y,w\right)$ are defined for all x,y,w; and $f_{4|123}\left(.|x,y,z\right)$ are defined or all x,y,z; then a condition for compatibility is that

$$\begin{aligned}&\left[\frac{f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_1,z,w\right)f_{3|124}\left(z|x_1,y_1,w\right)f_{4|123}\left(w|x_1,y_1,z_1\right)}{f_{1|234}\left(x_1|y,z,w\right)f_{2|134}\left(y_1|x_1,z,w\right)f_{3|124}\left(z_1|x_1,y_1,w\right)f_{4|123}\left(w_1|x_1,y_1,z_1\right)}\right]\\ &\times\left[\frac{f_{1|234}\left(x_2|y,z,w\right)f_{2|134}\left(y_2|x_1,z,w\right)f_{3|124}\left(z_2|x_2,y_2,w\right)f_{4|123}\left(w_2|x_2,y_2,z_2\right)}{f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_2,z,w\right)f_{3|124}\left(z_1|x_2,y_2,w\right)f_{4|123}\left(w|x_2,y_2,z_2\right)}\right]\end{aligned} \tag{6}$$

does not depend on x,y,z,w for all choices of $\left(x_1,y_1,z_1,w_1\right)\neq\left(x_2,y_2,z_2,w_2\right).$

Proof. For part [a], consider the right hand side of the expression:

$$\begin{aligned}&\frac{\left[f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_0,z,w\right)f_{3|124}\left(z|x_0,y,w_0\right)f_{4|123}\left(w|x_0,y_0,z_0\right)\right]}{\left[f_{1|234}\left(x_0|y,z,w\right)f_{2|134}\left(y_0|x_0,z,w\right)f_{3|124}\left(z_0|x_0,y_0,w\right)\right]}\\ &=\left\{\frac{f_{1234}\left(x,y,z,w\right)}{g_1\left(y,z,w\right)}\right\}\times\left\{\frac{f_{1234}\left(x_0,y,z,w\right)}{g_2\left(x_0,z,w\right)}\right\}\times\left\{\frac{f_{1234}\left(x_0,y_0,z_0,w\right)}{g_3\left(x_0,y_0,w\right)}\right\}\times\left\{\frac{f_{1234}\left(x_0,y_0,z_0,w\right)}{g_4\left(x_0,y_0,z_0\right)}\right\}\\ &\times\left\{\frac{g_1\left(y,z,w\right)}{f_{1234}\left(x_0,y,z_0,w\right)}\right\}\times\left\{\frac{g_2\left(x_0,z,w\right)}{f_{1234}\left(x_0,y_0,z,w\right)}\right\}\times\left\{\frac{g_3\left(x_0,y_0,w\right)}{f_{1234}\left(x_0,y_0,z_0,w\right)}\right\}\\ &=\frac{f_{1234}\left(x,y,z,w\right)}{g_4\left(x_0,y_0,z_0\right)}\\ &\propto f_{1234}\left(x,y,z,w\right)\end{aligned}$$

Hence, the proof.

For part [b], again, let us consider the right hand side of the expression:

$$\begin{aligned}&\left[\frac{f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_1,z,w\right)f_{3|124}\left(z|x_1,y_1,w\right)f_{4|123}\left(w|x_1,y_1,z_1\right)}{f_{1|234}\left(x_1|y,z,w\right)f_{2|134}\left(y_1|x_1,z,w\right)f_{3|124}\left(z_1|x_1,y_1,w\right)f_{4|123}\left(w_1|x_1,y_1,z_1\right)}\right]\\ &\times\left[\frac{f_{1|234}\left(x_2|y,z,w\right)f_{2|134}\left(y_2|x_1,z,w\right)f_{3|124}\left(z_2|x_2,y_2,w\right)f_{4|123}\left(w_2|x_2,y_2,z_2\right)}{f_{1|234}\left(x|y,z,w\right)f_{2|134}\left(y|x_2,z,w\right)f_{3|124}\left(z_1|x_2,y_2,w\right)f_{4|123}\left(w|x_2,y_2,z_2\right)}\right]\\ &=\frac{f_{1234}\left(x_2,y_2,z_2,w_2\right)}{f_{1234}\left(x_1,y_1,z_1,w_1\right)},\end{aligned}$$

after some simple algebra.
which implies that the expression does not depend on x, y, z, w for all choices of $(x_1, y_1, z_1, w_1) \neq (x_2, y_2, z_2, w_2)$. Hence, the proof.

□

In the next, we provide a more general theorem which is an extension to the multivariate case (say, dimension m, where $m \geq 5$). The theorem has been originally mentioned in Joe (1997), however, the proof was not provided. We provide the proof here by stating the theorem in a slightly different manner.

Theorem 3. Let $(Y_1, Y_2, \cdots, Y_m)$ be a random variable with joint probability density function $f_{1,2,\cdots,m}(y_1, y_2, \cdots, y_m)$. Denote by $f_{i|r}(y_i|y_r)$ to be the conditional density of Y_i given the remaining variables Y_j, $j \neq i$. Then,

(a) It can be shown that

$$f(y_1, y_2, \cdots, y_m) \propto \frac{\prod_{i=1}^{m} f_{i|r}\left(y_i|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}{\prod_{i=1}^{m} f_{i|r}\left(y_i^0|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}.$$

for a given vector, say, $\underline{y}^0 = (y_1^0, y_2^0, \cdots, y_m^0)$ for which all the conditional densities in the above expression are positive.

(b) The conditional densities $f\left(y_i|y_{(r)}\right), \quad i = 1, 2, \cdots, m$ are compatible if the expression

$$\left[\frac{\prod_{i=1}^{m} f_{i|(r)}\left(y_i|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}{\prod_{i=1}^{m} f_{i|(r)}\left(y_i^0|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}\right] \times \left[\frac{\prod_{i=1}^{m} f_{i|(r)}\left(z_i^0|z_1^0, \cdots, z_{i-1}^0, y_{i+1}, \cdots, y_m\right)}{\prod_{i=1}^{m} f_{i|(r)}\left(y_i|z_1^0, \cdots, z_{i-1}^0, y_{i+1}, \cdots, y_m\right)}\right]$$

does not depend on $(y_1, y_2, \cdots, y_m)$ for $(y_1^0, y_2^0, \cdots, y_m^0) \neq (z_1^0, z_2^0, \cdots, z_m^0)$.

Proof. For part [a], it is simple and thus excluded. For part [b], let us consider the right hand side of the expression

$$\begin{aligned}
&\left[\frac{\prod_{i=1}^{m} f_{i|(r)}\left(y_i|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}{\prod_{i=1}^{m} f_{i|(r)}\left(y_i^0|y_1^0, \cdots, y_{i-1}^0, y_{i+1}, \cdots, y_m\right)}\right] \times \left[\frac{\prod_{i=1}^{m} f_{i|(r)}\left(z_i^0|z_1^0, \cdots, z_{i-1}^0, y_{i+1}, \cdots, y_m\right)}{\prod_{i=1}^{m} f_{i|(r)}\left(y_i|z_1^0, \cdots, z_{i-1}^0, y_{i+1}, \cdots, y_m\right)}\right] \\
= \; &\left[\frac{f_{1|(1)}\left(y_1|y_1, y_2, \cdots, y_m\right) \times f_{2|(2)}\left(y_2|y_1^0, y_3, y_4 \cdots, y_m\right) \times \cdots f_{m|(m)}\left(y_m|y_1^0, y_2^0, y_3^0 \cdots, y_{m-1}^0\right)}{f_{1|(1)}\left(y_1^0|y_2 \cdots, y_m\right) \times f_{2|(2)}\left(y_2^0|y_1^0, y_3, y_4 \cdots, y_m\right) \times \cdots f_{m|(m)}\left(y_m|y_1^0, y_2^0, y_3^0 \cdots, y_{m-1}^0\right)}\right] \\
&\times\left[\frac{f_{1|(1)}\left(z_1^0|y_2, y_3, \cdots, y_m\right) f_{2|(2)}\left(z_2^0|z_1^0, y_2, y_3, y_4 \cdots, y_m\right) \cdots f_{m|(m)}\left(z_m^0|z_1^0, z_2^0, z_3^0 \cdots, z_{m-1}^0\right)}{f_{1|(1)}\left(y_1|y_2, y_3 \cdots, y_m\right) f_{2|(2)}\left(y_2^0|y_1^0, y_3, y_4 \cdots, y_m\right) \cdots f_{m|(m)}\left(y_m|z_1^0, z_2^0, z_3^0 \cdots, z_{m-1}^0\right)}\right] \\
= \; &\delta\left(y_1^0, y_2^0, \cdots, y_m^0; z_1^0, z_2^0, \cdots, z_m^0\right),
\end{aligned}$$

say, where $\delta()$ is a function which depends on $(y_1^0, y_2^0, \cdots, y_m^0)$ and $(z_1^0, z_2^0, \cdots, z_m^0)$ and is equal to 1 if $y_i^0 = z_i^0, \quad i = 1, 2, \cdots, m$; and doesn't depend on $(y_1, y_2, \cdots, y_m)$ for $(y_1^0, y_2^0, \cdots, y_m^0) \neq (z_1^0, z_2^0, \cdots, z_m^0)$. Hence, the proof. □

4 Illustrative Examples

In this section, we discuss couple of examples (in the bivariate case) which are given as textbook problems (as a general discussion) in [1] without any solution. To the best of the knowledge of the author, no such solution and/or hint of solution exists in the literature. We list them as follows.

1. **Example 1.** Suppose that we have the following two putative conditional densities

$$\begin{aligned} f\left(x|y\right) &\propto y \exp\left(-yx\right), 0 < x < A; \\ f\left(y|x\right) &\propto x \exp\left(-xy\right), 0 < y < A; \end{aligned} \tag{7}$$

Then, we make the following conjecture. If the set A is finite, then the two conditionals are compatible whereas if the set A is is infinite, then the two conditionals are incompatible.

Proof. By direct application of the condition in Eq. (3), the ratio of the two given conditionals will be

$$\frac{f\left(x|y\right)}{f\left(y|x\right)} = \frac{y}{x},$$

with $u(x) = \frac{1}{x}$, and $v(y) = y$. Next, note that,

$$\int_{S(X)} u(x) d\nu_1(x) = \int_0^A \frac{1}{x} dx = \log(A), \tag{8}$$

provided $1 < A < \infty$ i.e., a finite number. Obviously, if A approaches to ∞ then the above integral becomes infinite/ doesn't exist. Hence, the proof. □

2. **Example 2.** Does there exist a joint bivariate density with the following conditionals of the form:

$$\begin{aligned} f\left(x|y\right) &= \lambda(y) \exp\left[-\lambda(y)x\right], x > 0, \lambda(y) > 0, \quad and \\ f\left(y|x\right) &= \frac{\alpha}{\sigma(x)}\left(1 + \frac{y}{\sigma(x)}\right)^{-(\alpha+1)}, y > 0, \sigma(x) > 0? \end{aligned}$$

Proof. By direct application of the condition in Eq. (3), the ratio of the two given conditionals will be (as before)

$$\frac{f\left(x|y\right)}{f\left(y|x\right)} = \frac{\lambda(y)\sigma(x)\exp\left[-\lambda(y)x\right]}{\alpha}\left(1 + \frac{y}{\sigma(x)}\right)^{(\alpha+1)}. \tag{9}$$

□

Next, consider the following cases

- **Case 1.** Suppose, $\lambda(y) = \theta_1 > 0$, and $\sigma(x) = \theta_2 > 0$. Then, from Eq. (9), one can compute (using Eq. (3))
  with $u(x) = \exp\left[-\theta_1 x\right]$, and $v(y) = y$. Next, note that,

$$\int_{S(X)} u(x) d\nu_1(x) = \int_0^\infty \exp\left[-\theta_1 x\right] dx < \infty, \tag{10}$$

  thereby the two given conditionals will be compatible. It appears that no other choices of $\lambda(y)$ and/or $\sigma(x)$ would make this two conditional densities compatible.

5 Concluding Remarks

Compatibility under the absolutely continuous set-up has not been discussed adequately due to possible hindrance from the computational aspects including, but not limited to the integrability in bivariate and in higher dimensions as well as visualizing the whole thing. In this paper, we look at some of the simple results/conditions with an extension to multivariate dimension regarding the compatibility of conditional densities. Compatibility of conditional densities in the presence of imprecise information and/or information on conditional moments/ percentiles will be the subject matter of a separate article.

Conflict of interest statement

The corresponding author states that there is no conflict of interest.

References

[1] Barry C Arnold, Enrique Castillo, and José María Sarabia. *Conditional specification of statistical models*. Springer, 1999.

[2] Barry C Arnold, Enrique Castillo, and José María Sarabia. "Exact and near compatibility of discrete conditional distributions." In: *Computational statistics & data analysis* 40.2 (2002), pp. 231–252.

[3] Barry C Arnold, Enrique Castillo, and José María Sarabia. "Quantification of incompatibility of conditional and marginal information." In: *Communications in Statistics-Theory and Methods* 30.3 (2001), pp. 381–395.

[4] Barry C Arnold and Dattaprabhakar V Gokhale. "Distributions most nearly compatible with given families of conditional distributions." In: *Test* 7 (1998), pp. 377–390.

[5] Barry C Arnold and DV Gokhale. *Remarks on incompatible conditional distributions*. Tech. rep. Technical report, 1998.

[6] Barry C Arnold and José María Sarabia. "Conditional specification of statistical models: classical models, new developments and challenges." In: *Journal of Multivariate Analysis* 188 (2022), p. 104801.

[7] Hua Yun Chen. "Compatibility of conditionally specified models." In: *Statistics & probability letters* 80.7-8 (2010), pp. 670–677.

[8] Indranil Ghosh. "On the issue of convergence of certain divergence measures related to finding most nearly compatible probability distribution under the discrete set-up." In: *Statistics & Probability Letters* 203 (2023), p. 109915.

[9] Indranil Ghosh and N Balakrishnan. “Study of incompatibility or near compatibility of bivariate discrete conditional probability distributions through divergence measures.” In: *Journal of Statistical Computation and Simulation* 85.1 (2015), pp. 117–130.

[10] Indranil Ghosh and Saralees Nadarajah. “On the construction of a joint distribution given two discrete conditionals.” In: *Studia Scientiarum Mathematicarum Hungarica* 54.2 (2017), pp. 178–204.

[11] Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC press, 1997.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

BIG DATA-FUNDAMENTALS

B.L.S. Prakasa Rao[1] [2]

[1] *This article is based on the presentation "Big Data - Fundamentals" delivered as the inaugural lecture at the International Conference on Statistics "New Age Statistics for Data Science and Artificial Intelligence" (ICS 2023) held at Nilkamal School of Mathematics, Applied Statistics and Analytics, SVKM's NMIMS Deemed to be University, Mumbai during the period March 20-22, 2023.*

[2] *CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India, Email: blsprao@gmail.com*

Received: Apr 2024 / Accepted: Sept 2024

Abstract

We present several aspects of what is now known as BIG DATA Analytics with its advantages as well as pitfalls in general.

Keywords: .

1 Introduction

We present several aspects of what is now known as the BIG DATA with its advantages as well as pitfalls in general. Far from being an exhaustive review of this emerging field, this is an overview from the point of view of a statistician and it is a compilation of ideas of many researchers, organizations and from online sources. Our thanks are due to all those authors whose contributions have been listed in the references and whose ideas have been presented here and also to all those authors whose work we have inadvertently missed in including in the list of references. Some of our ideas dealing with Big Data and Data Science are discussed in Prakasa Rao (2015, 2017, 2021, 2022) and in Pyne et al. (2016).

Without any doubt, the most discussed current trend in statistics is BIG DATA. Different people think of different things when they hear about Big Data. For statisticians, how to get usable information out of data bases that are so huge and complex that many of the traditional or classical methods cannot handle? For computer scientists, Big Data poses problems of data storage and management, communication and computation. For citizens, Big Data brings up questions of privacy and confidentiality.

What is Big Data? Big Data is *relentless*. It is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor-enabled equipment such as aerial sensing technologies (remote sensing), information-sensing mobile devices, wireless sensor networks etc. Big Data is relatable. It can be related, linked and integrated to provide highly detailed information. Such a detail makes it possible, for instance, for banks to introduce individually tailored services and for health care providers to offer personalized medicines. Big data is a class of data sets so large that it becomes difficult to process it using standard methods of data processing. The problems of such data include capture or collection, curation, storage, search, sharing, transfer, visualization and analysis. Big data is difficult to work with using most relational data base management systems, desktop statistics and visualization packages. Big Data usually includes data sets with size beyond the ability of commonly used software tools.

When do we say that a data is a Big Data? Is there a way of quantifying the size of the data? We will come back to this question later in this article. Advantage of studying Big Data is that additional information can be derived from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found. For instance, analysis of a large data in marketing a product will lead to information on business trend for that product. Big Data can make important contributions to international development. Analysis of Big Data leads to a cost-effective way to improve decision making in important areas such as health care, economic productivity, crime and security, natural disaster and resource management. Large data sets are encountered in meteorology, genomics, biological and environmental research. They are also present in other areas such as internet search, finance and business informatics. Data sets are big as they are gathered using sensor technologies. There are also examples of Big Data in areas which we can call Big Science and in Science for research (cf. Prakasa Rao (2022)).

For Government, Big Data is present for climate simulation and analysis in meteorology, for official statistics and for information connected with national security. For private sector companies such as Flipkart and Amazon, Big Data comes up from millions of back-end operations every day involving queries from customer transactions, from vendors and from others.

Big Data sizes are a constantly moving target. It involves increasing volume (amount of data), velocity (speed of data in and out) and variety (range of data types and sources). Big Data are of high volume, high velocity and/or high variety information assets. It requires new forms of processing to enable enhanced decision making, insight discovery and process optimization. During the last twenty years, several companies here and abroad are adopting to data-driven approach to conduct more targeted services to reduce risks in decision making and to improve performance. They are implementing specialized data analytics to collect, store, manage and analyze large data sets. For example, some of the available financial data sources include stock prices, currency and derivative trades, transaction records, high-frequency trades, unstructured news and texts, consumer confidence and business sentiments from social media and internet among others. Analyzing these massive data sets help measuring firms risks as well as systemic risks. Analysis of such data requires people who are familiar with sophisticated statistical techniques such as portfolio management, stock regulation, proprietary trading, financial consulting and risk management. Big Data are of various types and sizes. Massive amounts of data are hidden in social net works such as Google, Face book, Linked In , You tube and Twitter now known as X. These data reveal numerous individual characteristics and have been exploited.

There are new types of data now. These data are not numbers but they come in the form of curves (functions), images, shapes or network. The data might be a “Functional Data" which may be a time series with measurements of the blood oxygenation taken at a particular point and at different moments in time. Here the observed function is a sample from an infinite dimensional space since it involves knowing the oxidation at infinitely many instants. The data from e-commerce is of functional type, for instance, results of auctioning of a commodity/item during a day by an auctioning company. Brain and neuroimaging data are typical examples of another type of functional data. These data is acquired to map the neuron activity of the human brain to find out how the human brain works. The next-generation functional data is not only a Big Data but a complex data.

Social media and internet contains massive amounts of information on the consumer preferences leading to information on the economic indicators, business cycles and political attitudes of the society. Analyzing large amount of economic and financial data is a difficult issue. One important tool for such analysis is the usual vector auto-regressive model involving generally at most ten variables and the number of parameters grows quadratically with the size of the model. Now a days econometricians need to analyze multivariate time series with hundreds of variables. Incorporating all these variables lead to over-fitting and bad prediction. One solution is to incorporate what is called the sparsity assumption. Another example, where a large number of variables might be present, is in portfolio optimization and risk management. Here the problem is estimating the covariance and inverse covariance matrices of the returns of the assets in the portfolio. If we have 1000 stocks to be managed, then there will be 500500 covariance parameters to be estimated. Even if we could estimate individual parameters, the total error in estimation can be large (Pourahmadi (2013)).

There are also concerns dealing with Big Data such as privacy and ethics. We will come back to this issue later in this article.

When is a data a BIG DATA? (cf. Fokoue (2015), Report of London Workshop (2014)) Big Data comes in various ways, types, shapes, forms and sizes. The dimensionality p of the input space (number of parameters) and the sample size n are usually the main ingredients in characterization of the bigness of the data. Large p and small n data sets will require different set of tools from those for the large n and small p sets of data. Here n is the data size and p is the number of unknown parameters/variables/covariates. There is no method which performs well on all types of data.

Let us consider a data set $\mathcal{D} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\}$ where $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of characteristics/covariates from the input space $\mathcal{X}$ and y_i is the corresponding response. The matrix $\mathbf{X}$ of order $n \times p$ given by

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2p} & y_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} & y_n \end{pmatrix}$$

is the data matrix. Five aspects of the data matrix are important:

(i) The dimension p representing the number of explanatory variables measured.

(ii) The sample size n representing the number of observations/sites at which the variables are measured or collected.

(iii) The relationship between p and n measured through the ratio of them.

(iv) The type of variables measured (categorical, interval, count, ordinal, real-valued, vector-valued, function-valued) and the indication of scales/units of measurement.

(v) The relationship among the columns of the data matrix to check multi-collinearity in the explanatory variables.

What is meant by "Massive or Big Data" as a function of p? Suppose we are dealing with a multiple linear regression problem with p covariates or explanatory variables under a Gaussian noise/error. For a model space search for variable selection, we have to find the best subset from among $2^p - 1$ models/sub-models. If $p = 20$, then $2^p - 1$ is about a million; if $p = 30$, then $2^p - 1$ is about a billion; and if p=40, then $2^p - 1$ is about a trillion. Hence any problem with more than $p = 50$ variables is a massive data problem. It involves searching a thousand trillion models which is a huge task even for modern computers. Hence any problem with more than 50 predictor variables can be called BIG DATA. If the number of predictor variables is more than 100, then it is called a MASSIVE DATA problem.

What is meant by "Massive or Big Data" as a function of n? We generally believe that the larger the sample from a population, the better is the inference, due to the law of large numbers. However the computational and statistical complexity in using methods of regression analysis involves inversion of $n \times n$ matrices which is computationally intensive when n is large. It takes $O(n^3)$ number of operations to invert an $n \times n$ matrix. Based on this observation, we might say that the data is observation-massive if $n > 1000$.

What is meant by "Massive or Big Data" as a function of n/p? Suppose that we are in a situation with a data where $p > 50$ or $n > 1000$. We have seen that the data can be considered massive in both cases. However the ratio n/p is even more important than n and p taken separately.Let us suppose that we have at least ten observations for each one of the p variables. Hence we have $n > 10p$. Let us also suppose that the information in the data is an increasing function of n. We have the following scenario (cf. Fokoue (2015)).

(A)$n/p < 1$	IP	$n << p, n > 1000$	Large p, Large n
(D)$n/p < 1$	IP	$n << p, n \leq 1000$	Large p, Small n
(B)$1 \leq n/p < 10$	IS	$n > 1000$	Small p, Large n
(E)$1 \leq n/p < 10$	IS	$n \leq 1000$	Small p, Smaller n
(C)$n/p \geq 10$	IA	$n >> p, n > 1000$	Smaller p, Large n
(F)$n/p \geq 10$	IA	$n >> p, n \leq 1000$	Smaller p, Small n

(IP=Information poverty; IS= Information scarcity; IA= Information abundance)

The BIG DATA problem is with the cases (A) and (D).

For statisticians, Big Data challenges some basic paradigms.The aim is to develop a model that describes how the response variable is related to p other variables or covariates and to determine which variables are important to characterize or explain the relationship. Fitting the model to data involves estimating the parameters from the data and assessing the evidence that they are different from zero indicating the importance of the variable. When $p >> n$, the number of parameters is huge relative to the information about them in the data. Thousands of irrelevant parameters will appear to be statistically significant if one uses small data statistics. Big Data has special features that are not present in the classical data sets. Big Data are characterized by massive sample size and high-dimensionality. Massive sample size allows one to discover hidden patterns associated with small sub-populations. Modeling the intrinsic heterogeneity of Big Data needs better statistical methods. The problems of high-dimensionality in data are noise accumulation, spurious correlation and incidental endogeny. Big Data is often a consequence of aggregation of many data sources corresponding to different sub-populations. Each sub-population might have a unique feature which is not shared by others. A large sample size enables one to better understand heterogeneity. A mixture model for the population may be appropriate for a Big data.

For example, a mixture probability density of the form

$$\lambda_1 \, p_1(y;\theta_1(\mathbf{x})) + \cdots + \lambda_m \, p_m(y;\theta_m(\mathbf{x}))$$

where $\lambda_j \geq 0$ represents the proportion of j-th subpopulation and $p_j(y;\theta_j(\mathbf{x}))$ is the probability density of the j-th sub-population given the covariate $\mathbf{x}$ with $\theta_j(\mathbf{x})$ as the parameter might be a good fit for a Big Data. In practice, λ_j is very small for some j. If the sample size n is small, then $n\lambda_j$ is also small and hence it might not be possible to infer about the parameter $\theta_j(\mathbf{x})$. Analyzing Big Data requires simultaneous estimation or testing of a large number of parameters. Errors made in inferring on these parameters accumulate when a decision on inference from the data depends on these parameters. Such a noise accumulation is severe in high-dimensional data and it may even dominate the true signal. This is handled by the sparsity assumption. High-dimensionality brings in spurious correlation due to the fact that many uncorrelated random variables may have high sample correlation coefficient in high dimensions.Spurious correlation leads to wrong inferences and hence false results. Unlike spurious correlation, incidental endogeny may be present in Big Data. It is the existence of correlation between variable "unintentionally" as well as due to "high-dimensionality". The former is analogous to finding two persons who look alike but have no genetic relationship where as the latter is similar to meeting an acquaintance by chance in a big city. Endogeny happens due to selection bias, measurement errors and omitted variables (cf. Fan et al.

(2013)). With the advantage of high-tech measurement techniques, it is now possible to collect as many features as possible.This increases the possibility that some of them might be correlated to the residual noise leading to incidental endogeny. Another reason for incidental endogeny is the following. Big Data are usually aggregated from multiple sources with possibly different data generating schemes. This increases the possibility of selection bias and measurement errors which also leads to possible incidental endogeny. Some statistical methods have been proposed to handle such issues such as penalized quasi-likelihood to handle noise accumulation issue.

Big Data are massive and very high-dimensional and involve large-scale optimization if one wants to use a likelihood or quasi-likelihood approach directly. Optimization with a large number of variables is not only expensive due to computational costs but also suffers from slow numerical rates of convergence and instability. It is also computationally infeasible to apply optimization methods on the raw data. To handle the data both from statistical and computational views, **dimension-reduction** techniques have to be adopted.

2 Some issues with the Big Data cf. Fokoue (2015), Buelens et al. (2014)

(i) **Batch data against incremental data production:** Big Data is delivered generally in a sequential and incremental manner leading to online learning methods. Online algorithms have the important advantage that the data does not have to be stored in memory. All that is required is in the storage of the built model at the given time in the sense that the stored model is akin to the underlying model. If the sample size n is very large, the data cannot fit into the computer memory and one can consider building a learning method that receives the data sequentially or incrementally rather than trying to load the complete data set into memory.This can be termed as *sequentialization*. Sequentialization is useful for streaming data and for massive data that is too large to be loaded into memory all at once.

(ii) **Missing values and imputation schemes:** In most of the cases of massive data, it is quite common to be faced with missing values. One should check at first whether they are missing systematically, that is in a pattern, or if they are missing at random and the rate at which they are missing. Three approaches are suggested to take care of this problem: (a) Deletion which consists of deleting all the rows in the Data matrix that contain any missing values ; (b) Central imputation which consists of filling the missing cells of the Data matrix with central tendencies like mean, mode or median;and (c) Model-based imputation methods such as EM-algorithm.

(iii) **Inherent lack of structure and importance of pre-processing:** Most of the Big Data is unstructured and needs pre-processing.

(iv) **Sparsity problem:** With the inherently unstructured data like text data, the pre-processing of data leads to data matrices whose entries are frequencies of terms in the case of text data, that contain too many zeroes leading to the sparsity problem. The sparsity problem in turn leads to modeling issues.

(v) **Homogeneity versus heterogeneity:** There are massive data sets which have input space homogeneous, that is, all the variables are of the same type. Examples of such data include audio processing, video processing and image processing. There are other types of Big Data where the input space consists of variables of different types. Such types of data arise in business, marketing and social sciences where the variables can be categorical, ordinal, interval, count and real-valued.

(vi) **Differences in units of measurement:** It is generally observed that the variables involved are measured on different scales leading to modeling problems. One way to take care of this problem is to perform transformations that project the variables onto the same scale. This is done either by standardization which leads all the variables to have mean zero and variance one or by unitization which consists in transform the variables so that the support for all of them is the unit interval [0,1].

(vii) **Selection bias and quality:** When Big Data are discussed in relation to official statistics, one point of criticism is that Big Data are collected by mechanisms unrelated to probability sampling and are therefore not suitable for production of official statistics for government. This is mainly because **Big Data sets are not representative of the population of interest. In other words, they are selective by nature and therefore yield biased results.** When a data set becomes available through some mechanism other than random sampling, there is no guarantee what so ever that the data is representative unless the coverage is full. **When considering the use of Big Data for official statistics, an assessment of selectivity has to be conducted.** How does one assess selectivity of Big Data?

(viii) **No clarity of target population:** Another problem of Big Data dealing with official statistics is that many Big data sources contain records of events not necessarily directly associated with statistical units such as household, persons or enterprises. Big Data is often a by-product of some process not primarily aimed at data collection. Analysis of Big Data is data-driven and not hypothesis-driven. **For Big Data, the coverage is large but incomplete and selective.** It may be unclear what the relevant target population is.

(ix) **Comparison of data sources:** Let us look at a comparison of different data sources for official statistics as compared to Big Data.

Comparison between Sample Survey and Big Data:

Data Source	Sample Survey	Big Data
Volume	Small	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Sample	Various Organizations
Coverage	Small fraction	Large/Incomplete

(Ref: Buelens et al. (2014))

Comparison between Census and Big Data:

Data Source	Census	Big Data
Volume	Large	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Administration	Various Organizations
Coverage	Large/Complete	Large/Incomplete

(Ref: Buelens et al. (2014))

(x) **Additional Remarks on the Use of Big Data for Official Statistics:** For Big Data dealing with the official statistics, there are no approaches developed till now to measure the errors or to check the quality. It is clear that bias due to selectivity has role to play in the accounting of Big Data. Big Data can be the single source of data for the production of some statistic about a population of interest. Assessing selectivity of the data is important. Correcting for selectivity can some times be achieved by choosing suitable method of model-based inference (cf. Breiman (2001)). These methods are aimed at predicting values for missing/unobserved units. The results will be biased if specific sub-populations are missing from the Big Data set. Big Data set can be used as auxiliary data set in a procedure mainly based on a sample survey. The possible gain of such an application for the sample survey is likely reduction in sample size and the associated cost. Using small area models, Big Data can be used as a predictor for survey based measurement. Big Data mechanism can be used as a data collection strategy for sample surveys. Big Data may be used irrespective of selectivity issues as a preliminary survey. Findings obtained from Big Data can be further checked and investigated through sample surveys (cf. Struijs et al. (2014)).

3 Methods of handling Big Data cf. Fokue (2015)

(i) **Dimension reduction:** Dimension reduction is an important method for handling Big Data as we mentioned earlier. Dimensionality reduction involves the determination of intrinsic dimensionality q of the input space where $q << p$. This can be done by orthogonalization techniques on the input space which reduces the problem to a lower dimensional orthogonal input space leading to variance reduction for the estimator. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are the methods for dimensionality reduction. However if $p >> n$, then most of these techniques cannot be used directly.

(ii) **Bagging:** As it was observed earlier, it is common in a massive data or a big data that a single model selected does not lead to optimal prediction. If there is a multi-collinearity between the variables which is bound to happen when p is very large, the estimators are unstable and of large variances. Bootstrap aggregation (also called bagging) reduces the variance of the estimators by aggregation of bootstrapped versions of the base estimators.

(iii) **Paralellization:** When the computational complexity for building the base estimator is high, the method of bagging becomes inefficient and not practical. One way to avoid this problem is to use parallel processing. Big Data analytics will need parallel processing or parallelization for speeding up computation or to handle massive data that cannot fit into a single computer memory. One way to make statistical procedures more efficient in analysis of Big Data is to parallelize them, that is, to write many algorithms that can run on many computers or many processors at the same time. The method of “Bootstrap" is a standard method for inferring the probability distribution from a sample. It is computationally intensive. However it is ideally suitable for parallelization because it involves generating numerous independent rounds of simulated data.

(iv) **Regularization:** With large p and small n, there exist a multiplicity of solutions for any optimization problem involving Big Data and hence the problem becomes ill-posed. Regularization methods are used to find a feasible optimal solution and one method of regularization is Lagrangian formulation of a constrained version of the problem. The method of LASSO (Tibshirani (1996)) is one such method in high-dimensional data analysis.

(v) **Assumption of sparsity:** As we noted earlier, thousands of irrelevant parameters will appear to be statistically significant if we use small data statistics for analysis of Big Data. In classical statistics, if the data implies occurrence of an event that has one-in-a million chance of occurring, then we are sure it is not by chance and hence consider it statistically significant. But if we are considering a Big Data with a large number of parameters, it is possible for the event to occur by chance and not due to significance of the relationship. Most data sets have only a few strong relationships between variables and everything else is noise. Thus most of the parameters do not matter. This leads to sparsity assumption which is to assume that all but a few parameters are negligible. This will allow a way of extracting information from a Big Data. One such method is L_1-minimization called LASSO mentioned earlier. This was used in the field of image processing to extract an image in sharp focus from blurry or noisy data.

(vi) **The problem of “Big n, Big p, Little t":** The speed at which one can process is an important element in analyzing Big Data. Classical statistics was always done in an off-line mode, the size was small and the the time for analysis was essentially unlimited. However, in the era of Big Data things are different. For a web company which is trying to predict user reaction and elicit user behavior such as clicking on an advertisement sponsored by a client, time is important. The web company might have only milliseconds to decide how to respond to a given user’s click. Furthermore the model constantly has to change to adopt to new users and new products. The objective of the person who is analyzing the data may not be to deliver a perfect answer but to deliver a good answer fast.

4 Privacy and Confidentiality for Big Data

How to keep privacy and confidentiality in the era of Big Data? Public concerns about privacy, confidentiality and misuse and abuse of individual data is a matter of concern in collection of Big Data. There are ways of masking Big Data. One way is to anonymize the records after they are collected by adding a random noise or to do matrix masking of the data matrix by a known mathematical operation so that individual information is difficult to retrieve. Cryptography is a discipline that applies mathematical transformations to data that are either irreversible or reversible only with a password or reversible only at a great expense that an opponent can ill afford to pay for it.

5 Computing issues for Big Data

(Fan et al. (2013)) As was mentioned earlier, the massive or very large sample size of Big data is a challenge for traditional computing infrastructure. Big Data is highly dynamic and not feasible or possible to store in a centralized data base. **The fundamental approach to store and process such data is to “divide and conquer".** The idea is to partition a large problem into more tractable and independent sub-problems. Each sub-problem is tackled in parallel by different processing units. Results from individual sub-problems are then combined to get the final result. "Hadoop" is an example of basic software and programming infrastructure for Big Data processing. "Map Reduce" is a programming model for processing large data sets in a parallel fashion. "Cloud Computing" is suitable for storing and processing of Big Data. We are not presenting the problems involved in storage and computation connected with Big Data in this article.

6 Why Big Data is in trouble?

Answer: They forgot about Applied Statistics (Jeff Leak, May 7, 2014 "Simply Statistics"). There were articles with titles "The parable of Google Flu: traps in Big Data analysis"."Big Data: are we making a big mistake ? ; "Google Flu trends: the limits of Big Data"; "Eight (No, Nine!) problems with Big Data".

All of the articles listed above and on-line point out the problems of Big Data such as sampling populations, multiple testing, selection bias and over-fitting besides others. Big Data is not a solution for all issues of data analysis. "There is a tendency for Big Data researcher and more traditional applied statistician to live in two different realms. Big Data offers enormous possibilities for understanding human interactions at a societal scale with rich spatial and temporal dynamics and for detecting complex interactions and nonlinearities among variables. However traditional “small data" often offer information that is not contained in Big Data" (Lazer et al. (2014)).

Acknowledgment: This review is based on several excellent articles mentioned in the list of references and other sources available on the web. Thanks are due to all the authors for their insights and explanations to understand what is Big Data. Work on this review is supported under the scheme "INSA Honorary Scientist" at the CR Rao Advanced Institute for Research in Mathematics, Statistics and Computer Science, Hyderabad 500046, India. Thanks are due to Professor Dilip Kumar Ghosh, Editor for inviting me to submit this article to the journal.

References

[1] L. Breiman (2013) "Statistical Modeling: The two cultures (with comments and a rejoinder by the author)"*Statistical Science*, **16**, 199-231.

[2] D. Buelens, P. Daas, J. Burger, and J. van den Brakel (2014) "Selectivity of Big Data", *Discussion paper, Statistics Netherlands.*

[3] E. Fakoue (2015) “A taxonomy of Big Data for optimal predictive machine learning and data mining", *arXiv.1501.0060v1 [Stat. ML] 3 Jan 2015.*

[4] J. Fan, F. Han, Fang and H. Liu (2013) "Challenges of Big Data analytics", *arXiv.1308.1479v1 [stat.ML] 7 Aug 2013*.

[5] D. Lazer, R. Kennedy, G. King and A. Vespignani (2014) “The parable of Google Flu traps in Big Data analysis", *Science*. **343**, 1203-1205.

[6] J. Leak (2014) "Why big data is in trouble; they forgot about applied statistics", *“Simply Statistics"*, May 7, 2014.

[7] M. Pourahmadi (2013) "Modern Methods in Covariance Estimation with High-dimensional Data"', Wiley, New York.

[8] B.L.S. Prakasa Rao (2015) "Brief Notes on Big Data: A Cursory Look", *Lecture Notes*, C R Rao Advanced Institute for Mathematics, Statistics and Computer Science, Hyderabad, India.

[9] B.L.S. Prakasa Rao (2017) "Brief Notes on Big Data", *Visleshana*, Computer Society of India, Special Interest Group - Big Data Analytics, **1** (3), April-June 2017, pp.9-12.

[10] B.L.S. Prakasa Rao (2021) "Big Data and Agriculture", In "Special Proceedings, 23rd Annual Conference of Society of Statistics, Computer and Applications", Hyderabad, pp.1-13.

[11] B.L.S. Prakasa Rao (2022) "What is Data Science and What is Big Data?-An Overview"', *Gujarat Journal of Statistics and Data Science"*, **1**, 1-6.

[12] Saumyadipta Pyne, B.L.S. Prakasa Rao and S.B. Rao (2016) *"Big Data Analytics"*, Edited by Saumyadipta Pyne, B.L.S. Prakasa Rao and S.B. Rao, Springer (India) Pvt. Ltd., New Delhi.

[13] Peter Struijs, Barteld Braaksma and Piet jH daas (2014) "Official statistics and Big data", *BIG DATA & Society*, DOI: 10.1177/2053951714538417, SAGE (2014).

[14] R. Tibshirani (1996) "Regression analysis and selection via the Lasso", *Journal of the Royal Statistical Society,Series B*, **58**, 267-288.

[15] "Current trends and future challenges in statistics: Big Data" *Statistics and Science: A Report of the London Workshop on future of the Statistical Sciences* (2014), pp. 20-25.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Dual To Separate Ratio-Cum-Product Type Exponential Estimator Of Finite Population Mean In Case Of Post-Stratification

Nitu Barod[1],
and Rajesh Tailor [2]

1. *School of Studies in Statistics, Vikram University, Ujjain-456010, M.P. , India.*
2. *School of Studies in Statistics, Vikram University, Ujjain-456010*

Received: May 2024 / Accepted: Sept 2024

Abstract

In this study, a new approach is recommended for calculating the finite population mean in situations where the sampling frame of the studied population is not accessible. This is crucial due to the fact that Stratified Random Sampling necessitates prior knowledge of both the sampling frame and strata weights. Post stratification is the superior approach compared to other sampling technique when sampling frame is not available. In this technique, from the population with heterogeneity, a sample of required size has been drawn using simple random sampling without replacement. Subsequently, the obtained sample is stratified into different strata based on some stratification factor.

In this paper a dual to separate ratio-cum-product type estimator is recommended in case of post stratification for estimating the finite population mean. For analyzing the efficiency of the estimator, the bias and the mean squared error of the developed estimator have been derived up to the first degree of approximation. The conditions for which the developed estimator performs better than other considered estimators have also been obtained up to the first degree of approximation. The bias comparison has been shown theoretically. Simulation study has been carried out to support the theoretical results.

From the simulation study, it is observed that the percent relative efficiency of the developed estimator increases with the increase in sample size and highest among all other estimators. This findings are visually represented with the help of bar graph.Thus, it can be concluded that, under specific conditions, the developed estimator provide more precise estimates compared to other considered estimators.

Keywords: : Bias, Mean squared error, Dual to separate ratio cum product type estimator, Simulation study.

1 INTRODUCTION

Stratified random sampling is often used to obtain a representative sample, which can produce precise estimates in the case of heterogeneous populations. In this approach, the heterogeneous population is divided into homogeneous subgroups or strata on the basis of some stratification factor, then from each stratum a small sample is drawn by using simple random sampling without replacement. Hansen, et al. (1946), Kadilar and Cingi (2003), Singh, et al. (2008) and many other researchers have used stratified random sampling for the estimation of the population mean. The limitation of stratified random sampling is that it requires prior knowledge of strata weights and sampling frames for each stratum, but there are some situations when strata weights and sampling frames are not available. In the absence of a sampling frame, the post-stratification technique is the better approach for the estimation of population mean in case of heterogeneous population. In this approach, a sample of the required is drawn using simple random sampling then the selected sample is stratified into different strata on the basis of some characteristics.

For the estimation of population mean in case of post stratification Ige and Tripathi (1989) developed ratio and product estimators. By using the exponential function Bahl and Tuteja (1991) introduced ratio and product type exponential estimator which were studied by Singh et al (2008) in stratified random sampling and the work was further extended by Tailor et al (2017) in post stratification. Motivated by Srivenkataraman (1980) and Bondyopadhyayh (1980), Lone and Tailor (2014) suggested dual to separate ratio type exponential estimator in post stratification. Tailor and Mehta (2019) proposed a ratio and ratio exponential estimator for finite population mean in case of post-stratification. Rather et al (2022) developed a new ratio type estimator for computation of population mean under post-stratification. Singh and Nigam (2022) proposed a two parameter ratio-product-ratio estimator in case of post stratification.

2 POST STRATIFICATION TECHNIQUE

Consider a population $U = (U_1, U_2, \ldots, U_N)$ of size N which is stratified into L strata of size $N_1, N_2, \ldots, N_L$ such that $\sum_{i=1}^{L} N_h = N$. Let x, y and z are the triplets indexed over the population U. The study variable y is positively correlated with the first auxiliary variable x and negatively correlated with the second auxiliary variable z. Let y_{hi}, x_{hi}, and z_{hi} be the observations on the ith unit of the hth stratum for the study variable y and auxiliary variables x and z respectively. The mean of the hth stratum of the study variable y is denoted by $\bar{Y}_h$, and the means of the hth stratum of x and z are denoted by $\bar{X}_h$ and $\bar{Y}_h$, respectively. $\bar{Y}$, $\bar{X}$, and $\bar{Z}$ are the population means of the study variable y and auxiliary variables x and z, respectively. By using simple random sampling without replacement, a sample of size n is drawn from a population U, then from the selected sample, it is recorded which units belong to the hth stratum. Let n_h be the size of the sample falling in the hth stratum such that $\sum_{h=1}^{L} n_h = n$. The possibility of n_h being zero is very small as it is assumed that n is so large.

where,

$$\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{h_i}, \quad \bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{h_i}, \quad \bar{Z}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} z_{h_i}, \quad \bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h,$$

$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$, $\bar{Z} = \sum_{h=1}^{L} W_h \bar{Z}_h$ and $W_h = \frac{N_h}{N}$ is the weight of the hth stratum.

3 SOME EXISTING ESTIMATORS

In post stratification approach, the usual unbiased estimator for population mean $\bar{Y}$ is defined as

$$\bar{y}_{PS} = \sum_{h=1}^{L} W_h \bar{Y}_h \tag{3.1}$$

where $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_i}$ is the sample mean of the hth stratum of size n_h.

To the fist degree of approximation the variance of $\bar{y}_{PS}$ is given as

$$V(\bar{y}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)\, S_{yh}^2 \tag{3.2}$$

For estimation of population mean of study variable y, Ige and Tripathi (1989) developed ratio estimator in post stratification which is defined as

$$\hat{\hat{Y}}_{RPS} = \bar{y}_{PS}\left(\frac{\bar{X}}{\bar{x}_{PS}}\right) \tag{3.3}$$

The separate version of Ige and Tripathi (1989) estimator is defined as

$$\hat{\bar{Y}}_{RPS}^{S} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\frac{\bar{X}_h}{\bar{x}_h}\right) \tag{3.4}$$

The bias and mean squared error of $\hat{\bar{Y}}_{RPS}^{S}$ up to the first degree of approximation are given as

$$B(\hat{\bar{Y}}_{RPS}^{S}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{S_{xh}^2 R_{1h} - S_{yxh}}{\bar{X}_h} \tag{3.5}$$

and

$$MSE(\hat{\bar{Y}}_{RPS}^{S}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h R_{1h}^2 S_{xh}^2 - 2\sum_{h=1}^{L} W_h R_{1h} S_{yxh}\right] \tag{3.6}$$

Using the dual approach the separate ratio estimator in post stratification is defined as

$$\hat{\bar{Y}}_{RPS}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\frac{\bar{x}_h^*}{\bar{X}_h}\right) \tag{3.7}$$

where, $\bar{x}_h^* = \frac{N_h \bar{X}_h - n_h \bar{x}_h}{N_h - n_h}$.
The bias and mean squared error of $\hat{\bar{Y}}_{RPS}^{*}$ up to the first degree of approximation are obtained as

$$B(\hat{\bar{Y}}_{RPS}^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{g'_h S_{yxh}}{\bar{X}_h} \tag{3.8}$$

and

$$MSE(\hat{\bar{Y}}_{RPS}^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h R_{1h}^2 g_h'^2 S_{xh}^2 - 2\sum_{h=1}^{L} W_h g'_h R_{1h} S_{yxh}\right] \tag{3.9}$$

Bahl and Tuteja (1991) used exponential function in ratio estimator for estimation of population mean. Motivated by Bahl and Tuteja (1991), Singh et al.(2008) developed the same estimator in stratified random sampling. Later in 2017, Tailor et al (2017) developed ratio type exponential estimator in post stratification for estimation of population mean as

$$\hat{\bar{Y}}_{RPS}^{Re} = \bar{y}_{PS} \exp\left(\frac{\bar{X} - \bar{x}_{PS}}{\bar{X} + \bar{x}_{PS}}\right) \tag{3.10}$$

The separate ratio type exponential estimator in case of post stratification can be defined as

$$\hat{\bar{Y}}_{PS}^{SRe} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{X}_h - \bar{x}_h}{\bar{X}_h + \bar{x}_h}\right) \tag{3.11}$$

The bias and mean squared error of separate ratio type exponential estimator $\hat{\bar{Y}}_{PS}^{S}Re$ up to the first degree of approximation are obtained as

$$B(\hat{\bar{Y}}_{PS}^{SRe}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{3S_{xh}^2 R_{1h} - 4S_{yxh}}{8\bar{X}_h} \tag{3.12}$$

and

$$MSE(\hat{\bar{Y}}_{PS}^{SRe}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{4}\sum_{h=1}^{L} W_h R_{1h}^2 S_{xh}^2 - \sum_{h=1}^{L} W_h R_{1h} S_{yxh}\right] \tag{3.13}$$

Dual to classical ratio estimator was proposed by Srivenkataramana (1980) and Bandhyopadhyayh (1980). Dual to Bahl and Tuteja (1991) estimator was developed by Tailor and Tailor (2012). Motivated by Srivenkataraman (1980) and Bandhyopadhyayh (1980), Lone and Tailor (2014) envisaged dual to separate ratio type exponential estimator for estimation of population mean $\bar{Y}$ post stratification technique as

$$\hat{\bar{Y}}_{PS}^{*Re} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{x}_h^* - \bar{X}_h}{\bar{x}_h^* + \bar{x}_h}\right) \tag{3.14}$$

Up to the first degree of approximation, the bias and mean squared error of $\hat{\bar{Y}}_{PS}^{*Re}$ are determined as

$$B(\hat{\bar{Y}}_{PS}^{*Re}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} \frac{-R_{1h} g_h^{,2} S_{xh}^2 - 4 g_h^{,} S_{yxh}}{8\bar{X}_h} \tag{3.15}$$

and

$$MSE(\hat{\bar{Y}}_{PS}^{*Re}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{4}\sum_{h=1}^{L} W_h g_h^{,2} R_{1h}^2 S_{xh}^2 - \sum_{h=1}^{L} W_h g_h^{,} R_{1h} S_{yxh}\right] \tag{3.16}$$

In case of post stratification, Ige and Tripathi (1989) developed product type estimator for estimation of population mean as

$$\hat{\bar{Y}}_{PPS} = \bar{y}_{PS}\left(\frac{\bar{x}_{PS}}{\bar{X}}\right) \tag{3.17}$$

Separate version of Ige and Tripathi (1989) product type estimator in can be written as

$$\hat{\bar{Y}}_{PPS}^{S} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\frac{\bar{x}_h}{\bar{X}_h}\right) \tag{3.18}$$

The bias and mean squared error of $\hat{\bar{Y}}_{PPS}^{S}$ up to the first degree of approximation are obtained as

$$B(\hat{\bar{Y}}_{PPS}^{S}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} \frac{S_{yxh}}{\bar{X}_h} \tag{3.19}$$

and

$$MSE(\hat{\bar{Y}}_{PPS}^{S}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h R_{1h}^2 S_{xh}^2 + 2\sum_{h=1}^{L} W_h R_{1h} S_{yxh}\right] \tag{3.20}$$

The dual to separate product type estimator for estimation of population mean in case of post stratification is defined as

$$\hat{\bar{Y}}_{PPS}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\frac{\bar{X}_h}{\bar{x}_h^*}\right) \tag{3.21}$$

The bias and mean squared error of $\hat{\bar{Y}}^{*}_{PPS}$ to the degree of approximation are

$$B(\hat{\bar{Y}}^{*}_{PPS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{g_h^{,2} S_{xh}^2 R_{1h} + g_h^{,} S_{yxh}}{\bar{X}_h} \tag{3.22}$$

and

$$MSE(\hat{\bar{Y}}^{*}_{PPS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h R_{1h}^2 g_h^{,2} S_{xh}^2 + 2\sum_{h=1}^{L} W_h g_h^{,} R_{1h} S_{yxh}\right] \tag{3.23}$$

For estimation of population mean, Bahl and Tuteja (1991) developed product type exponential estimator. Singh et al. (2008) modified the same estimator in stratified random sampling then motivated by Bahl and Tuteja (1991) and Singh et al. (2008), the product type exponential estimator was suggested by Tailor et al (2017) in case of post stratification as

$$\hat{\bar{Y}}^{Pe}_{PS} = \bar{y}_{PS} \exp\left(\frac{\bar{x}_{PS} - \bar{X}}{\bar{x}_{PS} + \bar{X}}\right) \tag{3.24}$$

In case of post stratification, the separate version of Tailor et al (2017) estimator can be defined as

$$\hat{\bar{Y}}^{SPe}_{PS} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{x}_h - \bar{X}_h}{\bar{x}_h + \bar{X}_h}\right) \tag{3.25}$$

The bias and mean squared error of separate product type exponential estimator in case of post stratification up to the first degree of approximation are obtained as

$$B(\hat{\bar{Y}}^{SPe}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{4S_{yxh} - S_{xh}^2 R_{1h}}{8\bar{X}_h} \tag{3.26}$$

and

$$MSE(\hat{\bar{Y}}^{SPe}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{4}\sum_{h=1}^{L} W_h R_{1h}^2 S_{xh}^2 + \sum_{h=1}^{L} W_h R_{1h} S_{yxh}\right] \tag{3.27}$$

For the estimation of population mean, motivated by Srivenkataraman (1980) and Bandyopadhyayh (1980) Lone and Tailor (2015) suggested the dual to separate product type exponential estimator in case of post stratification as

$$\hat{\bar{Y}}^{*Pe}_{PS} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\bar{X}_h - \bar{x}_h^*}{\bar{X}_h + \bar{x}_h^*}\right) \tag{3.28}$$

Up to the first degree of approximation, the bias and mean squared error of the estimator suggested by Lone and Tailor (2015) $\hat{\bar{Y}}^{*Pe}_{PS}$ are obtained as

$$B(\hat{\bar{Y}}^{*Pe}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \frac{3R_{1h} g_h^{,2} S_{xh}^2 + 4S_{yxh}}{8\bar{X}_h} \tag{3.29}$$

and

$$MSE(\hat{\bar{Y}}^{*Pe}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{4}\sum_{h=1}^{L} W_h g_h^{,2} R_{1h}^2 S_{xh}^2 + \sum_{h=1}^{L} W_h g_h^{,} R_{1h} S_{yxh}\right] \tag{3.30}$$

where, $S_{yh}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{L} \left(y_{hi} - \bar{Y}_h\right)^2$, $S_{xh}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{L} \left(x_{hi} - \bar{X}_h\right)^2$, $S_{yxh} = \frac{1}{N_h - 1} \sum_{h=1}^{L} \left(y_{hi} - \bar{Y}\right) \left(x_{hi} -\right.$
$g_h^{,} = \frac{n_h}{N_h - n_h}$, $R_{1h} = \frac{\bar{Y}}{\bar{X}}$.

4 THE DEVELOPED ESTIMATOR

Srivenkataramana (1980) suggested dual to ratio estimator for estimation of population mean. Lone and Tailor (2015) discussed dual to separate ratio and separate product type estimator in case of post stratification. Then Lakhre (2015) envisaged ratio-cum-product type estimator in case of post stratification as

$$\hat{\bar{Y}} = \bar{y}_{PS} \exp \frac{\bar{X} - \bar{x}_{PS}}{\bar{X} + \bar{x}_{PS}} \exp \frac{\bar{Z} - \bar{z}_{PS}}{\bar{Z} + \bar{z}_{PS}} \tag{4.1}$$

The purpose of this paper is to develop the dual to separate version of Lakhre (2015) i.e. the dual to separate ratio-cum-product estimator in case of post stratification as

$$\hat{\bar{Y}}_{TN} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\exp \frac{\bar{x}_h^* - \bar{X}_h}{\bar{x}_h^* + \bar{X}_h} \right) \exp \left(\frac{\bar{Z}_h - \bar{z}_h^*}{\bar{Z}_h + \bar{z}_h^*} \right) \tag{4.2}$$

where, $\bar{x}_h^* = \frac{N_h \bar{X}_h - n_h \bar{x}_h}{N_h - n_h}$ and $\bar{z}_h^* = \frac{N_h \bar{Z}_h - n_h \bar{z}_h}{N_h - n_h}$.

To find the bias and mean squared error of the developed estimator $\hat{\bar{Y}}_{TN}$ it is assumed that
$\bar{y}_h = \bar{Y}_h(1 + e_{0h})$, $\bar{x}_h = \bar{X}_h(1 + e_{1h})$, $\bar{z}_h = \bar{Z}_h(1 + e_{2h})$.
Such that, $E(e_{0h}) = E(e_{1h}) = E(e_{2h}) = 0$
$E(e_{0h}^2) = (\frac{1}{nW_h} - \frac{1}{N_h})C_{yh}^2$,$E(e_{1h}^2) = (\frac{1}{nW_h} - \frac{1}{N_h})C_{xh}^2$,
$E(e_{2h}^2) = (\frac{1}{nW_h} - \frac{1}{N_h})C_{zh}^2$, $E(e_{0h}e_{1h}) = (\frac{1}{nW_h} - \frac{1}{N_h})\rho_{yxh}C_{yh}C_{xh}$,
$E(e_{0h}e_{2h}) = (\frac{1}{nW_h} - \frac{1}{N_h})\rho_{yzh}C_{yh}C_{zh}$, $E(e_{1h}e_{2h}) = (\frac{1}{nW_h} - \frac{1}{N_h})\rho_{xzh}C_{xh}C_{zh}$.

5 THE BIAS AND MEAN SQUARED ERROR OF THE DEVELOPED ESTIMATOR

The bias and mean squared error of the developed estimator $\hat{\bar{Y}}_{TN}$ up to the first degree of approximation are calculated as

$$B(\hat{\bar{Y}}_{TN}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \left[\frac{g_h^{,2}}{8} \left[\frac{3S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{2S_{xzh}R_{1h}}{\bar{Z}_h} - \frac{S_{xh}^2 R_{1h}}{\bar{X}_h} \right] + \frac{g_h^{,}}{2} \left[\frac{S_{yzh}}{\bar{Z}_h} - \frac{S_{yxh}}{\bar{X}_h} \right] \right] \tag{5.1}$$

and

$$MSE(\hat{\bar{Y}}_{TN}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{L} \left[S_{yh}^2 + \frac{g_h^{,2}}{4} \left[S_{zh}^2 R_{2h}^2 + S_{xh}^2 R_{1h}^2 - 2S_{xzh}R_{1h}R_{2h} \right] - g_h^{,} \left[S_{yxh}R_{1h} - S_{yzh}R_{2h} \right] \right] \tag{5.2}$$

where, $S_{zh}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{L} \left[z_{hi} - \bar{Z}_h \right]^2$, $S_{yzh} = \frac{1}{N_h - 1} \sum_{h=1}^{L} [y_{hi} - \bar{Y}_h][z_{hi} - \bar{Z}_h]$,
$R_{2h} = \frac{\bar{Y}_h}{\bar{Z}_h}$.

6 EFFICIENCY COMPARISON

(i) On comparing equation (3.2) and (5.2), it is found that the developed estimator $\hat{\bar{Y}}_{TN}$ is better than the usual unbiased estimator $\bar{y}_{PS}$ if,$MSE(\hat{\bar{Y}}_{TN}) \leq V(\hat{\bar{y}}_{PS})$ i.e.

$$\sum_{h=1}^{L} W_h g_h^{,2} S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h^{,2}}{2} S_{xzh} R_{1h} R_{2h} - \frac{g_h^{,2}}{4} S_{xh}^2 R_{1h}^2 + g_h^{,} SyxhR_{1h} - g_h^{,} SyzhR_{2h}\right] \tag{6.1}$$

(ii) When equation (3.6) and (5.2) are compared, it proved that the developed estimator $\hat{\hat{Y}}_{TN}$ is better than separate ratio type estimator $\hat{\hat{Y}}_{RPS}^{S}$if, $MSE(\hat{\hat{Y}}_{TN}) \le MSE(\hat{\hat{Y}}_{RPS}^{S})$ i.e.

$$\sum_{h=1}^{L} W_h g_h^{,2} S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h^{,2}}{2} S_{xzh} R_{1h} R_{2h} + \left(1 - \frac{g_h^{,2}}{4}\right) S_{xh}^2 R_{1h}^2 + SyxhR_{1h}\left(g_h^{,} - 2\right) - g_h^{,} SyzhR_{2h}\right] \tag{6.2}$$

(iii) The comparison of equation (3.9) and (5.2) proves that the developed estimator $\hat{\hat{Y}}_{TN}$ is superior to the dual to separate ratio type $\hat{\hat{Y}}_{RPS}^{*}$ if,$MSE(\hat{\hat{Y}}_{TN}) \le MSE(\hat{\hat{Y}}_{RPS}^{*})$ i.e.

$$\sum_{h=1}^{L} W_h g_h^{,2} S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h^{,2}}{2} S_{xzh} R_{1h} R_{2h} + \frac{3g_h^{,2}}{4} S_{xh}^2 R_{1h}^2 - SyxhR_{1h}g_h^{,} - g_h^{,} SyzhR_{2h}\right] \tag{6.3}$$

(iv) Comparison between equation (3.13) and (5.2) reveals that the developed estimator $\hat{\hat{Y}}_{TN}$ performs better than the separate ratio type exponential estimator,if $\mathrm{MSE}\left(\hat{\hat{Y}}_{TN}\right) \le \mathrm{MSE}\left(\hat{\hat{Y}}_{PS}^{SRe}\right)$i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} + \frac{S_{xh}^2 R_{1h}^2}{4}\left(1 - g_h'^2\right) + S_{yxh} R_{1h}\left(g_h' - 1\right) - g_h' S_{yzh} R_{2h}\right] \tag{6.4}$$

(v) When equation (3.16) and (5.2) are compared, the developed estimator proved $\hat{\hat{Y}}_{TN}$ to be better over dual to separate ratio type exponential estimator $\hat{\hat{Y}}_{PS}^{*Re}$ if,$MSE(\hat{\hat{Y}}_{TN}) \le MSE(\hat{\hat{Y}}_{PS}^{*Re})$ i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} - g_h' S_{yzh} R_{2h}\right] \tag{6.5}$$

(vi) On comparing equation (3.20) and (5.2), the developed estimator $\hat{\hat{Y}}_{TN}$ found to be better than separate version of Ige and Tripathi (1989) estimator $\hat{\hat{Y}}_{PPS}^{S}$ if, $MSE(\hat{\hat{Y}}_{TN}) \le MSE(\hat{\hat{Y}}_{PPS}^{S})$ i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} + S_{xh}^2 R_{1h}^2\left(1 - \frac{g_h'^2}{4}\right) + S_{yxh} R_{1h}\left(g_h' + 2\right) - g_h' S_{yzh} R_{2h}\right] \tag{6.6}$$

(vii) From equation (3.23) and (5.2), it is concluded that the developed estimator $\hat{\hat{Y}}_{TN}$ performs better than the dual to separate product type estimator $\hat{\hat{Y}}_{PPS}^{*}$ in case of post stratification if, $MSE(\hat{\hat{Y}}_{TN}) \le MSE(\hat{\hat{Y}}_{PPS}^{*})$ i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \le 4\sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} + \frac{3}{4} g_h'^2 S_{xh}^2 R_{1h}^2 + 3g_h' S_{yxh} R_{1h} - g_h' S_{yzh} R_{2h}\right] \tag{6.7}$$

(viii) From equation (3.27) and (5.2), the developed $\hat{\bar{Y}}_{TN}$ is more efficient than separate version of Chouhan (2012) product type exponential estimator $\hat{\bar{Y}}_{PS}^{SPe}$ in case of post stratification if, $MSE(\hat{\bar{Y}}_{TN}) \leq MSE(\hat{\bar{Y}}_{PS}^{SPe})$ i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \leq 4 \sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} + \frac{S_{xh}^2 R_{1h}^2}{4} \left(1 - g_h'^2\right) + S_{yxh} R_{1h} \left(g_h' + 1\right) - g_h' S_{yzh} R_{2h} \right] \tag{6.8}$$

(ix) From equation (3.30) and (5.2), it is obtained that the developed estimator $\hat{\bar{Y}}_{TN}$ performs better than the Lone and Tailor (2015) dual to separate product type exponential $\hat{\bar{Y}}_{PS}^{*Pe}$ in case of post stratification if, $MSE(\hat{\bar{Y}}_{TN}) \leq MSE(\hat{\bar{Y}}_{PS}^{*Pe})$ i.e.

$$\sum_{h=1}^{L} W_h g_h'^2 S_{zh}^2 R_{2h}^2 \leq 4 \sum_{h=1}^{L} W_h \left[\frac{g_h'^2}{2} S_{xzh} R_{1h} R_{2h} + 2 g_h' S_{yxh} R_{1h} - g_h' S_{yh} R_{2h} \right] \tag{6.9}$$

7 BIAS COMPARISON

This section shows the bias comparison of the developed estimator with other considered estimator. From equation (3.5), (3.8), (3.12), (3.15), (3.19), (3.22), (3.26), (3.29) and (5.1).

(i) $\left| B\left(\hat{\bar{Y}}_{TN}\right) \right| < \left| B\left(\hat{\bar{Y}}_{RPS}^{S}\right) \right|$ if,

$$\left[\begin{array}{c} \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{g_h'^2 S_{xzh} R_{1h}}{4\bar{Z}_h} - \frac{S_{xh}^2 R_{1h}}{\bar{X}_h} \left(\frac{g_h'^2}{8} + 1\right) + \frac{g_h' S_{yzh}}{2\bar{Z}_h} + \frac{S_{yxh}}{\bar{X}_h} \left(1 - \frac{g_h'}{2}\right) \right]\right) \right] \\ \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{g_h'^2 SxzhR_{1h}}{4\bar{Z}_h} + \frac{Sxh^2 R_{1h}}{\bar{X}_h} \left(1 - \frac{g_h'^2}{8}\right) + \frac{g_h' S_{yzh}}{2\bar{Z}_h} - \frac{S_{yxh}}{\bar{X}_h} \left(\frac{g_h'}{2} + 1\right) \right]\right) \end{array} \right] < 0 \tag{7.1}$$

(ii) $\left| B\left(\hat{\bar{Y}}_{TN}\right) \right| < \left| B\left(\hat{\bar{Y}}_{RPS}^{*}\right) \right|$ if,

$$\left[\begin{array}{c} \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{g_h'^2 S_{xzh} R_{1h}}{4\bar{Z}_h} - \frac{g_h'^2}{8} \frac{S_{xh}^2 R_{1h}}{\bar{X}_h} + \frac{g_h' S_{yzh}}{2\bar{Z}_h} \right]\right) \\ \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}h} - \frac{g_h'^2 S_{xzh} R_{1h}}{4\bar{Z}h} - \frac{g_h'^2}{8} \frac{S_{xh}^2 R_{1h}}{\bar{X}_h} + \frac{g_h' S_{yzh}}{2\bar{Z}_h} - \frac{2 g_h' S_{yxh}}{\bar{X}_h} \right]\right) \end{array} \right] < 0 \tag{7.2}$$

(iii) $\left| B\left(\hat{\bar{Y}}_{TN}\right) \right| < \left| B\left(\hat{\bar{Y}}_{PS}^{S\mathrm{Re}}\right) \right|$ if,

$$\left[\begin{array}{c} \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{g_h'^2}{4} \frac{S_{xzh} R_{1h}}{\bar{Z}_h} - \frac{S_{xh}^2 R_{1h}}{8} \left[g_h'^2 + 3\right] + \frac{g_h'}{2} \frac{S_{yzh}}{\bar{Z}_h} + \frac{S_{yxh}}{2\bar{X}_h} \left[1 - g_h'\right] \right]\right) \\ \left(\sum_{h=1}^{L} \left[\frac{3}{8} \frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h} - \frac{g_h'^2}{4} \frac{S_{xzh} R_{1h}}{\bar{Z}_h} + \frac{S_{xh}^2 R_{1h}}{8\bar{X}_h} \left[3 - g_h'^2\right] + \frac{g_h'}{2} \frac{S_{yzh}}{\bar{Z}_h} - \frac{S_{yxh}}{2\bar{X}_h} \left[1 + g_h'\right] \right]\right) \end{array} \right] < 0 \tag{7.3}$$

(iv) $\left| B\left(\hat{\bar{Y}}_{TN}\right) \right| < \left| B\left(\hat{\bar{Y}}_{PS}^{*\mathrm{Re}}\right) \right|$ if,

$$\left[\begin{array}{l} \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}\right.\right. \\ \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{S_{xh}^2 R_{1h} g_h'^2}{4\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}-\frac{S_{yxh}g_h'}{\bar{X}_h}\right]\right) \end{array}\right] < 0 \tag{7.4}$$

(v) $\left|B\left(\hat{\bar{Y}}_{TN}\right)\right| < \left|B\left(\hat{\bar{Y}}_{PPS}^{S}\right)\right|$ if,

$$\left[\begin{array}{l} \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{S_{xh}^2 R_{1h} g_h'^2}{8\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}-\frac{S_{yxh}}{\bar{X}_h}\left[1+\frac{g_h'}{2}\right]\right]\right) \\ \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{S_{xh}^2 R_{1h} g_h'^2}{8\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}+\frac{S_{yxh}}{\bar{X}_h}\left[1-\frac{g_h'}{2}\right]\right]\right) \end{array}\right] < 0 \tag{7.5}$$

(vi) $\left|B\left(\hat{\bar{Y}}_{TN}\right)\right| < \left|B\left(\hat{\bar{Y}}_{PPS}^{*}\right)\right|$ if,

$$\left[\begin{array}{l} \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{9S_{xh}^2 R_{1h} g_h'^2}{8\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}-\frac{3S_{yxh}g_h'}{2\bar{X}_h}\right]\right) \\ \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}+\frac{7S_{xh}^2 R_{1h} g_h'^2}{8\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}+\frac{S_{yxh}g_h'}{2\bar{X}_h}\right]\right) \end{array}\right] < 0 \tag{7.6}$$

(vii) $\left|B\left(\hat{\bar{Y}}_{TN}\right)\right| < \left|B\left(\hat{\bar{Y}}_{PS}^{SPe}\right)\right|$ if,

$$\left[\begin{array}{l} \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}+\frac{S_{xh}^2 R_{1h}}{8\bar{X}_h}\left[1-g_h'^2\right]+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}-\frac{S_{yxh}}{\bar{X}_h}\left[\frac{1}{2}+g_h'\right]\right]\right) \\ \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{S_{xh}^2 R_{1h}}{8\bar{X}_h}\left[1+g_h'^2\right]+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}+\frac{S_{yxh}}{\bar{X}_h}\left[\frac{1}{2}-g_h'\right]\right]\right) \end{array}\right] < 0 \tag{7.7}$$

(viii) $\left|B\left(\hat{\bar{Y}}_{TN}\right)\right| < \left|B\left(\hat{\bar{Y}}_{PS}^{*Pe}\right)\right|$ if,

$$\left[\begin{array}{l} \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}-\frac{S_{xh}^2 R_{1h} g_h'^2}{2\bar{X}_h}+\frac{g_h'}{2}\frac{S_{yzh}}{\bar{Z}_h}-\frac{S_{yxh}g_h'}{\bar{X}_h}\ \right]\right) \\ \left(\sum_{h=1}^{L}\left[\frac{3}{8}\frac{g_h'^2 S_{zh}^2 R_{2h}}{\bar{Z}_h}-\frac{g_h'^2}{4}\frac{S_{xzh}R_{1h}}{\bar{Z}_h}+\frac{S_{xh}^2 R_{1h} g_h'^2}{4\bar{X}_h}+\frac{g_h'^2}{2}\frac{S_{yzh}}{2\bar{Z}_h}\right]\right) \end{array}\right] < 0 \tag{7.8}$$

8 SIMULATION STUDY

Simulation study has been done using R-software for analyzing the performance of the developed estimator $\hat{\bar{Y}}_{TN}$ with respect to other considered estimator. The population of size N having four different strata of equal size has been generated. For the current study, four different sample sizes n = 700, 800, 900 and 1000 are taken from the population.

Step 1: Four different strata of equal size such that $N_1 = N_2 = N_3 = N_4 = 500$ have been generated from multivariate normal distribution which together constitute a population of size N=2000.

Step 2: From a population of size N, a sample of size n has been drawn using simple random sampling without replacement and then stratified.

Step 3: Used the sample data obtained for each stratum from Step 2 to find the mean squared error as well as percent relative efficiency of the developed estimator along with all other considered estimator.

Step 4: Repeated the Step 2 and Step 3, 10000 times to obtain 10000 values for calculating mean squared error and percent relative efficiency.

Step 5: Finally the mean squared error and percent relative efficiency of all the estimator have been calculated through following formulae

$MSE(E_i) = \frac{1}{10000}\sum_{i=1}^{10000}\left(E_i - \bar{Y}\right)^2$

$PRE(E_i) = \frac{MSE(\bar{y}_{PS})}{MSE(E)} * 100$

$\bar{Y}$is the population mean of study variable y.The results of mean squared error and percent relative efficiency for the developed as well as considered estimator are given in Table 1 and 2 respectively.

Estimator	700	800	900	1000
$\hat{\bar{y}}_{PS}$	0.01158	0.009464	0.007722	0.00629
$\hat{\bar{Y}}^{S}_{RPS}$	0.010649	0.008359	0.007013	0.005686
$\hat{\bar{Y}}^{*}_{RPS}$	0.007082	0.005844	0.005574	0.00573
$\hat{\bar{Y}}^{SRe}_{PS}$	0.007056	0.005634	0.004699	0.003825
$\hat{\bar{Y}}^{*Re}_{PS}$	0.008159	0.006203	0.004866	0.003844
$\hat{\bar{Y}}^{S}_{PPS}$	0.044906	0.036753	0.029745	0.024178
$\hat{\bar{Y}}^{*}_{PPS}$	0.025461	0.024703	0.024134	0.024189
$\hat{\bar{Y}}^{SPe}_{PS}$	0.024199	0.019838	0.016072	0.013075
$\hat{\bar{Y}}^{*Pe}_{PS}$	0.017347	0.01563	0.014144	0.01307
$\hat{\bar{Y}}_{TN}$	**0.004382**	**0.002879**	**0.001983**	**0.001575**

Table 1: Mean squared error of different estimators.

Estimator	700	800	900	1000
$\hat{\bar{y}}_{PS}$	100.00	100.00	100.00	100.00
$\hat{\bar{Y}}^{S}_{RPS}$	108.7446	113.2216	110.1234	110.6165
$\hat{\bar{Y}}^{*}_{RPS}$	163.5191	161.9509	138.5345	109.7763
$\hat{\bar{Y}}^{SRe}_{PS}$	164.1146	167.9821	164.3393	164.4506
$\hat{\bar{Y}}^{*Re}_{PS}$	141.9364	152.5795	158.6949	163.6408
$\hat{\bar{Y}}^{S}_{PPS}$	25.78753	25.75155	25.96188	26.01509
$\hat{\bar{Y}}^{*}_{PPS}$	45.48179	38.31326	31.99815	26.00291
$\hat{\bar{Y}}^{SPe}_{PS}$	47.85442	47.70897	48.04835	48.1066
$\hat{\bar{Y}}^{*Pe}_{PS}$	66.75625	60.55349	54.59974	48.12585
$\hat{\bar{Y}}_{TN}$	**264.2669**	**328.707**	**389.3731**	**399.4471**

Table 2: Percent relative efficiency of different estimators.

9 CONCLUSION

The objective of the present study is to develop the dual to separate ratio cum product type estimator and analyzing its performance through simulation study. Section 5 of this paper shows the bias and mean squared error of the developed estimator up to the first degree of approximation. In section 6, the conditions under which the developed estimator performs better than other considered estimator have been obtained up to the first degree of approximation. The theoretical bias comparison has also been shown in section 7. Section 8 shows the simulation study, where the performance of the developed estimator has been analyzed as compared to other considered estimators at different sample sizes and it is also observed that the mean squared error of the developed estimator is least at different sample sizes and decreases with the increase in sample size. The percent relative efficiency of the developed estimator increases with the increase in sample size and highest among all other estimators. Table 8.1 and 8.2 shows the mean squared error and percent relative efficiency of all the estimators with respect to the usual mean estimator in case of post stratification. Thus, the developed estimator performs better in terms of efficiency as compared to all other estimators considered in this study. Hence, the developed estimator has been recommended for the use in practice.

Conflict of interest

There is no conflict of interest.

References

Bahl, S. and Tuteja, R. K., (1991): Ratio and product type exponential estimators. *J. Info.& Optim. Sci.,12,1, 159-164.*

Bandyopadhyay (1980): Improved ratio and product estimator. *Sankhya*, 42, C, 45-49.

Hansen, M.H., Hurwitz, W.N. and Gurney, M. (1946): Problems and methods of the sample survey of business. *J. Am. Stat. Asso., 41, 173-189.*

Ige, A. F. and Tripathi, T. P. (1989): Estimation of population mean using post-stratification and auxiliary information. Abacus, 18, 2, 265-276.

K. Ul Islam Rather, M. Iqbal Jeelani, M. Younis Shah, S. E. H. Rizvi and M. Sharma (2022): A new ratio type estimator for computation of population mean under post-stratification. *J. Appl. Math., Stat. & Info.,*18 (1), 29-42.

Kadilar, C. and Cingi, H. (2003): Ratio estimators in stratified random sampling. *Biom J., 45(2), 218-225.*

Lakhre, A. (2015). Ratio-cum-product type estimator in case of post stratification. *Ph. D. Thesis, Vikram University, Ujjain, M. P., India.*

Lone, H. A. and Tailor, R. (2014): Dual to separate ratio type exponential estimator in post-stratificaiton. *J. Stat. Appl. & Pro.*, 3, 3, 425-432.

Lone, H. A. and Tailor, R. (2015): Dual to separate product type exponential estimator in sample surveys. *J. Stat. Appl. Pro. Lett.*, 2, 1, 89-96.

Singh, H. P., and Nigam, P. (2022): A two parameter Ratio-Product-Ratio Estimator in Post Stratification. *Pak.j.stat.oper.res. 18(1) 273-296.*

Singh, R., Kumar, M., Singh, R. D. and Chaudhary, M. K. (2008): Exponential ratio type estimators in stratified random sampling. *Presented in International Symposium on Optimization and Statistics (I.S.O.S) at A. M. U., Aligarh, India during 29-31 Dec 2008.*

Srivenkataramana, T. (1980): A dual to ratio estimator in sample surveys. *Biom J.*, 67, 1, 199-204.

Tailor R., Tailor R., and Chouhan S. (2017): Improved Ratio- and Product-Type Exponential Estimators for Population Mean in Case of Post-Stratification. *Communications in Statistics - Theory and Methods, Vol 46, No. 21, 10387-10393.*

Tailor, R. and Mehta, P. (2019). A Ratio and ratio exponential estimator for finite population mean in case of post-stratification. *J. Stat. Appl. & Pro.*, 8, 3, 241-246.

Tailor, R. and Tailor, R. (2012): Dual to ratio and product type exponential estimators for ratio of two population means in sample surveys. *Modl. Assist. Statist. Appls., 10.3233/MAS-140300.*

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

A Modified Genetic Algorithms Using Balanced Incomplete Block Designs And Balanced Ternary Designs

H. L. Sharma[1], Vijayshankar Shukla[2] and Varsha Shukla[3]

1. *Ex-Professor & Head, Department of Mathematics & Statistics , J.N. Agricultural University, Jabalpur (M.P.),India*
2. *Computer Science and Engineering, A.K.S. University, Satna (M.P.), India*
3. *Comptroller Office, J.N. Agricultural University, Jabalpur (M.P.),India*

Received: Jan 2024 / Accepted: Sept 2024

Abstract

This paper is concerned with a modified genetic algorithms using balanced incomplete block designs(BIBD) and balanced ternary designs (BTD). The different types of chromosomes of varying length and size have been considered with the help of BIBD and BTD which are found to be adequate for their offsprings in the next generations after applying the operator crossover and mutation. This has been illustrated with two examples, one is based on BIBD and other, on BTD which are added at the end.

Keywords: Genetic algorithm, Balanced incomplete block design (BIBD), Balanced ternary design (BTD), Crossover, Mutation, Offspring, Bits.

1 Introduction

Considerable evidences have been gathered by human biologists, geneticists ,demographers, social scientists, data scientist and data analyst, machine learning engineers and operation research workers during last few decades about the study of genetic algorithms in developing and developed countries utilizing the different techniques based on various sets of data. It has been initially developed by John Holland, his colleagues, his students at the University of Michigan with objective in two tires (i) to abstract and rigorously explain to adaptive processes of natural systems and (ii) to design artificial systems software that retains the important mechanisms of natural systems(Mitchell [8]; Lingarai [3]; Katoch et el.[6]; Goldberg[1]; Deb [5]; Dana-Bana-Hani [4]; Mullawaarchachi [7]; Jain [10]).

However, a genetic algorithm (GA) is a mathematical search algorithm based on the mechanics of natural selection and natural genetics combining into survival of the fittest among string structures including as a numerical optimization technique to be adequate for being applied to an extremely wide range of problems. In every generation, a new set of artificial creatures / strings is created using bits and pieces of the presence of genes to be a new part for next generation. They require the natural parameter set of the optimization problem to be coded as a finite length string over some finite alphabet. Here, we first code the switches as a finite length string. A simple code can be generated by considering a string of n consisting of 1's and 0's where each of the switches is represented by a 1's if the switch is on and a 0's if the switch is off. With this coding ,the string viz.,11110 codes the setting where the first four switches are on and the fifth switch is off. A simple GA that provides optimized results in many practical situations that is composed of three operators, (i) Reproduction (ii) Crossover, and (iii) Mutation.

Moreover, reproduction is a process in which individual strings are copied according to their objective function values, f (fitness function). It is to be noted that the function f is as some measure of profit, utility, or goodness that we want to maximize or minimize. In order to copy strings, one should know about the fitness function having high value that provide higher probability of contribution in one or more off springs in next generation. This operator, eventually, is an artificial version of natural selection, a Darvinian survival of the fittest among string creatures. Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. Offspring are created by exchanging the genes of parents among themselves until the crossover point is reached. Mutation occurs to maintain diversity within the population and prevent premature convergence. Thus, the algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation. The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance of surviving. This process keeps on iterating and at the end, a new generation with the fittest individuals will be achieved. The five phases are considered in a genetic algorithm. (i) Initial population (ii) Fitness function(iii) Selection (iv) Crossover (v) Mutation. The process begins with a set of individuals which is called a population. Each individual is a solution to the problem you want to solve. An individual is characterized by a set of parameters (variables) known as genes and are joined into a string to form a chromosome (solution). In a genetic algorithm, the set of genes of an individual is represented using a string, in terms of an alphabet. Usually, binary values are used (string of 1's and 0's). It applies evolution concepts such as reproduction and survival of the fittest to solve a variety of problems. In fact, it belongs to the larger class of evolutionary algorithms considering a chromosome (sometimes referred to as a string). Each part of a chromosome is called a "gene"while the value of each gene is called an "allele".

The objective of the present paper is to develop and initialize the chromosomes population in the form of 0's and 1's bits on the basis of BIBD and BTD (which is constructed with the help of BIBD) as 'presence'(1's) or 'absence'(0's). An expression in general for the chromosomes length n has also been considered about the phase of crossover and mutation along with the overall fitness in particular as compared to the original one.

2 MATERIAL AND METHODS

BALANCED INCOMPLETE BLOCK DESIGN (BIBD)

A BIB design is an arrangement of v treatments in b blocks each of size $k(< v)$ such that (i) Each treatment occurs at most once in a block (ii) Each treatment occurs in exactly r blocks (iii) Each pair of treatments occurs together in exactly λ blocks. Thus, the symbols v, b, r, k, λ are called the parameters of the design and satisfy the following relations.

$vr = bk, r(k-1) = \lambda(v-1)$ and $b \geq v$.

BALANCED TERNARY DESIGN (BTD)

A balanced ternary design is a collection of B blocks, each of cardinality K ($K \leq V$), chosen from a set of size V in such a way that each of the V treatments occurs R times altogether, each of the treatments occurring once in precisely in Q_1 blocks and twice in precisely Q_2 blocks, and with incidence matrix having inner product of any two rows Λ is denoted by BTD $(V, B, Q_1, Q_2, R, K, \Lambda)$. The following relations hold in the case of BTD $VR = BK$, $\Lambda(V-1) = R(K-1) - 2Q_2$ It is to be noted that $Q_1 + 2Q_2 = R$ (Gupta et el.,[2]; Sarvate and Seberry, [9]).

GENETIC ALGORITHM AND ITS CODED VALUES

In real-life problems, most decision variables are conditioned and bounded, mostly taking real numbers). For example, if you had an optimization problem that aimed to maximize the profit or minimize the cost for a certain multi-national company, it may desire to keep the maximum wages of labor hours providing other conditions. The company would tend to set a maximum wages and minimum wages per month, for example, it may say that the minimum and maximum wages of labor hours per week should not exceed 3000 to 8000, therefore, the lower bound of the wages of labor hours (x) would be 3000 and the upper bound would be 8000 $(3000 \leq x \leq 8000)$. It is to be noted that the decision variable wages of labor hours (x) between bounds, calculating a real number for the hours can be generated as:

1	1	0	1	0	0	0	1	0
2^8	2^7	2^6	2^5	2^4	2^3.......	2^2	2^1	2^0

Figure 1: Each gene represents with its corresponding 2^n image

For the chromosome in Figure 1, to be encoded with 0's and 1's each gene represents with 2^n (2 to the power of n, where n starts at 0,1 on the chromosome, i.e. $2^0, 2^1, 2^2, 2^3, \cdots, 2^n$). In order to decode GA chromosome into a real value, we use the following formula:

$$\sum_{i=0}^{n} 2^i \times \text{bit}\Big(\frac{\text{upper limit} - \text{lower limit}}{(2^n - 1)}\Big) + \text{lower limit}. \tag{1}$$

Thus, the chromosomes for our genetic algorithm will be sequences of 0's and 1's bits with a length of n , and have a range from $0(00000 \cdots, n)$ to $n(11111 \cdots, n)$.

The balanced ternary design (BTD) is constructed sometimes with the help of balanced incomplete block design (BIBD) consisting of the symbols 0,1 and 2. If we assume that the genes of each chromosome are represented by 0,1 and 2 in relation to the quality of genes such as recessive, heterozygote and dominant characters respectively. In order to maximize a function, these genes of a chromosome may be considered to prepare a model for genetic algorithms. The following table represents the initial population of chromosomes on the basis BTD which has been constructed on the basis of BIBD in the form of the symbols 0,1 and 2 respectively.

It is to be noted that the decision variable wages of labor hours (x) between bounds, calculating a real number for the hours can be generated in the range of (0,1,2).i.e. three genes are considered together as

1	2	0	2	0	1	2	0	2	1	2	0	1
3^{12}	3^{11}	3^{10}	3^9	3^8	3^7	3^6	3^5	3^4	3^3	3^2	3^1	3^0

Figure 2: Each gene represents with its corresponding 3^n image

For the chromosome in Figure 2, to be encoded with $0's$, $1's$ and $2's$ each gene represents with 3^n (3 to the power of n, where n starts at 0,1,2 on the chromosome, i.e.$3^0, 3^1, 3^2, 3^3, \cdots, 3^n$). In order to decode GA chromosome into a real value, we use the following formula:

$$\sum_{i=0}^{n} 3^i \times \text{bit}\Big(\frac{\text{upper limit} - \text{lower limit}}{(3^n - 1)}\Big) + \text{lower limit}. \tag{2}$$

Thus, the chromosomes for our genetic algorithm will be sequences of $0's$ and $1's$ bits with a length of n, and have a range from 0 $(00000 \cdots, n)$, $n(11111 \cdots, n)$ and $n(22222 \cdots, n)$.

3 ILLUSTRATIVE EXAMPLES

EXAMPLE 3.1

Let us take a function to show that whether GA can improve the solution from one generation to next generation for the following function to be maximized $f(x) = x^2$ subject to the condition that $1 \leq x \leq 16$ by considering the number and length of the chromosomes as 6 and 14 for the function to be maximized as $f(x) = \sqrt{x}$ subject to the condition that $1 \leq x \leq 25$ respectively on the basis of symmetrical BIBD .

Chromosomes/ string No.	Initial population	Decoded value(D)	X-Value $= D\big(\frac{16-1}{2^3-1}\big) + 1$	$f(x) = x^2$	Selection probability	Expected count
1	1 0 0	4	4 x (15/7)+1=9.57	91.58	0.178	1.07
2	0 1 0	2	2 x (15/7)+1= 5.28	27.88	0.054	0.32
3	0 0 1	1	1 x (15/7)+1=3.14	9.86	0.019	0.11
4	0 1 1	3	3 x (15/7)+1=7.43	55.2	0.107	0.65
5	1 0 1	5	5 x (15/7)+1=11.71	137.12	0.267	1.6
6	1 1 0	6	6 x (15/7)+1=13.86	192.1	0.374	2.24
				$\sum f_i = 513.84$ $\bar{f} = 85.64$		

Table 1: The initial population, decoded value of x, fitness value $f(x)$ along with selection probability of the chromosomes 7 of length 7 on the basis of BIBD.

Actual count Roulette wheel	Mating pool	Random Mating pair	Parents	Crossover Site	Offspring	Mutation	Decoded value X	X-Value $=$ $= D\big(\frac{16-1}{2^3-1}\big) + 1$	$f(x) = x^2$
1	1 0 0	1	1 0 0	1 | 0 0	1 0 1	1 0 1	5	5 x (15/7)+1=11.71	137.12
0	0 1 1	4	1 0 1	1 | 0 1	1 0 0	1 0 0	4	4 x (15/7)+1=9.57	91.58
0	1 0 1	2	0 1 1	0 1 | 1	0 1 0	1 1 0	6	6 x (15/7)+1=13.86	192.1
1	1 0 1	5	1 1 0	1 1 | 0	1 1 1	0 1 1	3	3 x (15/7)+1=7.43	55.2
2	1 1 0	3	1 0 1	1 | 0 1	1 1 0	1 1 0	6	6 x (15/7)+1=13.86	192.1
2	1 1 0	6	1 1 0	1 | 1 0	1 0 1	1 0 1	5	5 x (15/7)+1=11.71	137.12
								$\sum f_i = 805.22$ $\bar{f} = 134.20$ Maximum $f = 134.20$	

Table 2: The actual count Roulette wheel, mating pool, mating pair, parents, crossover, offspring and mutation of fourteen chromosomes given in Table 1 for value of $n = 6$

In order to generate the number and length of the chromosomes as 14 and 7 respectively, we consider the BIBD as 7,7,3,3,1 and its complement as 7,7,4,4,2 in the form of incidence matrix of symbols (0,1) and signify 1 as presence and 0 as the absence of the genes involved in the chromosomes. Seven number of chromosomes having length 7 on the basis of BIBD and put in the form of binary number as 0 and 1 and 13 number of chromosomes having length four on the basis of BTD are considered in Table 1 and Table 2 respectively for maximizing the function $f(x) = x^2$ subject to the condition that $1 \leq x \leq 16$.

Chromosomes/ string No.	Initial population	Decoded value(D)	X-Value = $D\left(\frac{16-1}{2^3-1}\right)+1$	$f(x)=\sqrt{x}$	Selection probability	Expected Count
1	1 1 0 1 0 0 0	104	20.654	4.545	0.0928	1.3
2	0 1 1 0 1 0 0	52	10.827	3.29	0.0672	0.94
3	0 0 1 1 0 1 0	26	5.913	2.432	0.0497	0.7
4	0 0 0 1 1 0 1	13	3.457	1.859	0.038	0.53
5	1 0 0 0 1 1 0	70	14.228	3.772	0.077	1.08
6	0 1 0 0 0 1 1	35	7.614	2.759	0.0563	0.79
7	1 0 1 0 0 0 1	81	15.307	3.912	0.0799	1.12
8	0 0 1 0 1 1 1	23	5.346	2.312	0.0472	0.66
9	1 0 0 1 0 1 1	75	15.173	3.895	0.0795	1.11
10	1 1 0 0 1 0 1	101	21.220	4.607	0.0941	1.32
11	1 1 1 0 0 1 0	114	22.543	4.748	0.097	1.36
12	0 1 1 1 0 0 1	57	11.772	3.431	0.0701	0.98
13	1 0 1 1 1 0 0	92	18.386	4.288	0.0876	1.23
14	0 1 0 1 1 1 0	46	9.693	3.113	0.0636	0.89
				$\sum f_i$=48.963		
				$\bar{f}$=3.497		
				Maximum f=4.748		

Table 3: The initial population ,decoded value of x, fitness value $f(x)$ along with selection probability of the chromosomes 7 of length 7 on the basis of BIBD.

Actual count Roulette wheel	Mating pool	Random Mating pair	Parents	Crossover site	Offspring	Mutation
1	1 1 0 1 0 0 0	1	1 1 0 1 0 0 0	1 1 0 | 1 0 0 0	1 1 0 0 1 1 1	1 1 0 0 0 1 1
1	0 1 1 0 1 0 0	8	0 0 1 0 1 1 1	0 0 1 | 0 1 1 1	0 0 1 1 0 0 0	0 0 1 1 1 0 0
1	0 0 1 1 0 1 0	2	0 1 1 0 1 0 0	0 1 | 1 0 1 0 0	0 1 0 1 0 1 1	0 1 1 1 0 1 1
1	0 0 0 1 1 0 1	9	1 0 0 1 0 1 1	1 0 | 0 1 0 1 1	1 0 1 0 1 0 0	1 0 0 0 1 0 0
1	1 0 0 0 1 1 0	3	0 0 1 1 0 1 0	0 0 1 1 | 0 1 0	0 0 1 1 1 0 1	0 0 1 1 1 0 1
1	0 1 0 0 0 1 1	10	1 1 0 0 1 0 1	1 1 0 0 | 1 0 1	1 1 0 0 0 1 0	1 1 0 0 0 1 0
1	1 0 1 0 0 0 1	4	0 0 0 1 1 0 1	0 0 0 1 | 1 0 1	0 0 0 1 0 1 0	0 1 0 1 0 1 0
1	0 0 1 0 1 1 1	11	1 1 1 0 0 1 0	1 1 1 0 | 0 1 0	1 1 1 0 1 0 1	1 0 1 0 1 0 1
1	1 0 0 1 0 1 1	5	1 0 0 0 1 1 0	1 0 | 0 0 1 1 0	1 0 1 1 0 0 1	1 0 0 1 0 0 1
1	1 1 0 0 1 0 1	12	0 1 1 1 0 0 1	0 1 | 1 1 0 0 1	0 1 0 0 1 1 0	0 1 1 0 1 1 0
1	1 1 1 0 0 1 0	6	0 1 0 0 0 1 1	0 1 0 | 0 0 1 1	0 1 0 1 1 0 0	0 1 0 1 1 1 0
1	0 1 1 1 0 0 1	13	1 0 1 1 1 0 0	1 0 1 |1 1 0 0	1 0 1 0 0 1 1	1 0 1 0 0 0 1
1	1 0 1 1 1 0 0	7	1 0 1 0 0 0 1	1 0 1 0 | 0 0 1	1 0 1 0 1 1 0	0 0 1 0 1 1 0
1	0 1 0 1 1 1 0	14	0 1 0 1 1 1 0	0 1 0 1 | 1 1 0	0 1 0 1 0 0 1	1 1 0 1 0 0 1

Mutation	Decoded value X	X-Value = $D\left(\frac{16-1}{2^3-1}\right)+1$	$f(x)=\sqrt{x}$
1 1 0 0 0 1 1	99	99 x (24/127)+1=19.709	4.439
0 0 1 1 1 0 0	28	28 x (24/127)+1=6.291	2.508
0 1 1 1 0 1 1	59	59x (24/127)+1=12.150	3.486
1 0 0 0 1 0 0	68	68x (24/127)+1=13.850	3.722
0 0 1 1 1 0 1	28	28 x (24/127)+1=6.291	2.508
1 1 0 0 0 1 0	98	98 x (24/127)+1=19.520	4.418
0 1 0 1 0 1 0	42	42 x (24/127)+1=8.937	2.989
1 0 1 0 1 0 1	85	85 x (24/127)+1=17.063	4.131
1 0 0 1 0 0 1	73	73 x (24/127)+1=14.795	3.846
0 1 1 0 1 1 0	54	54 x (24/127)+1=11.205	3.347
0 1 0 1 1 1 0	46	46 x (24/127)+1=9.693	3.113
1 0 1 0 0 0 1	81	81 x (24/127)+1=16.307	4.038
0 0 1 0 1 1 0	22	22 x (24/127)+1=5.157	2.271
1 1 0 1 0 0 1	105	105 x (24/127)+1=20.843	4.565
			$\sum f_i$ =49.381
			$\bar{f}$ =3.527
			Maximum f=4.565

Table 4: The actual count Roulette wheel, mating pool, mating pair, parents, crossover, offspring and mutation of fourteen chromosomes given in Table 1 for value of $n=6$

EXAMPLE 3.2

Twelve number of chromosomes having length 4 on the basis of balanced ternary design (BTD) and arrange in the form of binary numbers as 0 , 1 and 2 for the function to be maximized $f(x) = \sqrt{x}$ subject to the condition that $1 \leq x \leq 25$ (Table 5).

With the existing BIB design, a self complementary BIB design can be generated (Mitra and Mandal,1998). In this case a self complementary BIB design which consists of the following parameters as $v = 4, b = 6, r = 3, k = 2$ and $\lambda = 1$ whose six blocks are given as (1,4) (2,4), (3,4),(2,3), (1,3) and (1,2).

Then balanced ternary design (BTD) are constructed by taking the combinations of above two blocks together at a time whose parameters are as follows:

$V = 4, B = 15, Q1 = 9, Q_2 = 3, R = 15, K = 4$ and $\Lambda = 13$.

(1,4,2,4), (1,4,3,4), (1,4,2,3), (1,4,1,3), (1,4,1,2), (2,4,3,4), (2,4,2,3), (2,4,1,3),

(2,4,1,2), (3,4,2,3), (3,4,1,3), (3,4,1,2), (2,3,1,3), (2,3,1,2), (1,3,1,2).

But the block numbers (iii), (viii) and (xii) (highlighted by red colour) are repeated blocks. Therefore, these three blocks are deleted. Then , we have the remaining twelve blocks which consists of BTD parameters as $V = 4, B = 12, Q_1 = 6, Q2 = 3, R = 12, K = 4$ and $\Lambda = 10$ which satisfy the parametric relations of BTD. After converting these 12 blocks in the form of incidence matrix whose blocks are given below:

(2 1 1 0), (1 1 2 0), (2 0 1 1), (1 0 2 1), (1 2 1 0), (2 1 0 1),

(1 2 0 1), (0 2 1 1), (0 1 2 1), (1 1 0 2) , (1 0 1 2), and (0 1 1 2).

These twelve blocks are used to represent 12 chromosomes of length 4 in genetic algorithms.

Chromosomes/ string No.	Initial population	Decoded value(D)	X-Value = $D\left(\frac{16-1}{3^4-1}\right)+1$	$f(x) = \sqrt{x}$	Selection probability	Expected count
1	2 1 1 0	66	66 x (24/80)+1=20.80	4.56	0.1076	1.3
2	1 1 2 0	42	42 x (24/80)+1= 13.60	3.69	0.0873	1.0
3	2 0 1 1	58	58 x (24/80)+1= 18.40	4.29	0.1012	1.2
4	1 0 2 1	34	34 x (24/80)+1= 11.20	3.35	0.0790	0.9
5	1 2 1 0	48	48 x (24/80)+1= 15.40	3.92	0.0925	1.1
6	2 1 0 1	64	64 x (24/80)+1= 20.2	4.49	0.1059	1.3
7	1 2 0 1	46	46 x (24/80)+1= 14.80	3.85	0.0908	1.1
8	0 2 1 1	22	22 x (24/80)+1= 7.60	2.76	0.0651	0.8
9	0 1 2 1	16	16 x (24/80)+1= 5.80	2.41	0.0569	0.7
10	1 1 0 2	38	38 x (24/80)+1= 12.40	3.52	0.0831	1.0
11	1 0 1 2	32	32 x (24/80)+1= 10.60	3.26	0.0769	0.9
12	0 1 1 2	14	14x (24/80)+1= 5.20	2.28	0.0538	0.6
			$\sum f_i =$	42.38		
			$\bar{f} =$	3.532	0.0833	
			Maximum $f =$	4.56		

Table 5: The initial population, decoded value of x, fitness value $f(x)$ along with selection probability of the chromosomes 12 of length 4 on the basis of BTD.

Actual count Roulette wheel	Mating pool	Random Mating pair	Parents	Crossover site	Offspring	Mutation
2	2 1 1 0	1	2 1 1 0	2 1 | 1 0	2 1 0 1	2 1 0 1
1	2 1 1 0	8	2 1 0 1	2 1 | 0 1	2 1 1 0	2 1 1 0
2	1 1 2 0	2	2 1 1 0	2 1 | 1 0	2 1 0 1	2 1 0 1
0	2 0 1 1	9	2 1 0 1	2 1 | 0 1	2 1 1 0	2 1 1 0
2	2 0 1 1	3	1 1 2 0	1 1 2 | 0	1 1 2 1	1 1 2 1
2	1 2 1 0	10	1 2 0 1	1 2 0 | 1	1 2 0 0	1 2 0 0
2	1 2 1 0	4	2 0 1 1	2 0 | 1 1	2 0 0 1	2 2 0 1
0	2 1 0 1	11	1 2 0 1	1 2 | 0 1	1 2 1 1	1 0 1 1
0	2 1 0 1	5	2 0 1 1	2 0 1 | 1	2 0 1 2	2 0 1 2
1	1 2 0 1	12	1 1 0 2	1 1 0 | 2	1 1 0 1	1 1 0 1
0	1 2 0 1	6	1 2 1 0	1 2 | 1 0	1 2 2 2	1 2 1 2
0	1 1 0 2	13	2 2 2 2	2 2 | 2 2	2 2 1 0	2 2 2 0

Mutation	Decoded value X	X-Value $= D\left(\frac{25-1}{3^4-1}\right)+1$	$f(x)=\sqrt{x}$
2 1 0 1	64	64 x (24/80)+1=20.20	4.49
1 2 1 0	48	48 x (24/80)+1=15.40	3.92
2 1 0 1	64	64 x (24/80)+1=20.20	4.49
2 1 1 0	66	66 x (24/80)+1=20.80	4.56
1 1 2 1	43	43 x (24/80)+1=13.90	3.73
2 1 0 0	63	63 x (24/80)+1=19.90	4.46
2 2 0 1	58	58 x (24/80)+1=18.40	4.29
2 0 1 1	46	46x (24/80)+1=14.80	3.85
1 2 0 1	58	58 x (24/80)+1=18.40	4.29
2 0 1 1	46	46 x (24/80)+1=14.80	3.85
1 2 0 1	50	50 x (24/80)+1=16.00	4.00
1 2 1 2	36	36 x (24/80)+1=11.80	3.44
			$\sum f_i =$ 49.37 $\overline{f} =$ 4.114 Maximum $f =$ 4.56

Table 6: The actual count Roulette wheel, mating pool, mating pair, parents, crossover, offspring and mutation of fourteen chromosomes given in Table 1 for value of $n = 6$.

4 Conclusion

The above final tables reveal that there is an improvement in the maximization of the function of the two functions namely $f(x) = x^2$ and $f(x) = \sqrt{x}$ in the range of $1 \leq x \leq 16$ and $1 \leq x \leq 25$ respectively. The improvement in genetic algorithms have been achieved using BIBD and BTD on the basis of incidence matrix having the various kinds of number of chromosomes and their length. Thus, genetic algorithm may tackle the problem of integer programming with more precision. It is started with a initial population randomly and out of these one solution may be found as global optimal solution. This paper can also be extended by choosing appropriate techniques consisting of the values 0 and 1 that should lie within a certain specified range.

Acknowledgement

The authors are grateful to the Editor in Chief, Managing Editor and specially Editorial Secretary Dr.Parag Shah for their critical comments and help that have eventually modified the shape of the paper. We are also indebted to Dr Raju, D. Chaudhari, Associate Professor for making the paper in LaTex format of this journal.

References

[1] D.E. Goldberg. *Genetic Algorithms in Search,Optimization and Machine Learning.* Addison Wesley publishing company, 1989.

[2] W.S. Gupta Sudhir; Lee and S. Kageyama. "Nested balanced n-ary designs." In: *Metrika* 42 (1995), pp. 411–419.

[3] Lingaraj Haldurai. "A Study on Genetic Algorithm and its Applications." In: *International Journal of Computer Sciences and Engineering* 10 (2016), pp. 139–143.

[4] Dana Bani Hani. *Genetic Algorithm (GA): A Simple and Intuitive Guide, year = 2020, url = Towards data science*.

[5] Deb Kalyanmoy. “An introduction to genetic algorithms.” In: *Sadhana* 24 (2016), pp. 293–315.

[6] S.S. Katoch Sourabh; Chauhan and Vijay Kumar. “A Review on Genetic algorithm,Past,Present and Future.” In: *Multimedia tools and applications* 80 (2020), pp. 8091–8126.

[7] Vijini Mallawaarachchi. *Introduction to Genetic Algorithms — Including Example Code, year = 2017, url = Towards data science*.

[8] Melanie Mitchell. *An Introduction to Genetic Algorithms.* MIT Press, 1998.

[9] D.G. Sarvate and J. Seberry. “Construction of balanced ternary designs based on generalized Bhaskar Rao designs.” In: *J. Statist. Plann. Inf.* 34 (1993), pp. 423–432.

[10] Jain Shubham. *Introduction to Genetic Algorithm & Their Application in Data Science*. 2020.

Gujarat Journal of Statistics and Data Science
GUJARAT STATISTICAL ASSOCIATION

Efficient Use of Two Auxiliary Variables in Estimating the Finite Population Mean Under Simple Random Sampling and Stratified Random Sampling.

Pragati Nigam[1] and Housila P. Singh[2]

1. *School of Studies in Agriculture Sciences, Vikram University, Ujjain-456010, M.P., India*
2. *School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India*

Received: Jan 2023 / Accepted: Nov 2023

Abstract

We have developed a class of estimators for estimating the population mean based on two auxiliary variables in simple random sampling without replacement (*SRSWOR*) scheme. Expressions of bias and mean squared error of the proposed class of estimators are obtained up to the first order of approximation. Optimum conditions are obtained at which the proposed class of estimators yields minimum mean squared error. We have compared the proposed class of estimators with some existing estimators $\hat{\bar{Y}}_{S(R)}$, $\hat{\bar{Y}}_{S(RExp)}$, $\hat{\bar{Y}}_{S(D_1)}$, $\hat{\bar{Y}}_{S(RP)}$, $\hat{\bar{Y}}_{S(RPExp)}$, $\hat{\bar{Y}}_{S(D_2)}$, $\hat{\bar{Y}}_{S(MSK)}$ and $\hat{\bar{Y}}_{S(SG)}$. The properties of the suggested class of estimators are also discussed in stratified random sampling. An empirical study is conducted in support of the present study.

Keywords: Study variable, Auxiliary variable, Bias, Mean squared error, Simple random sampling, Stratified random sampling.
Mathematics Subject Classification Code- 62D05

1 Introduction

Consider a finite population $U = \{U_1, U_2, ...U_N\}$ of N identifiable units with study variable Y and auxiliary variables (X,Z) defined on U taking the values y_i and (x_i, z_i) for unit U_i of U respectively. We define

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$: Population mean of the study variable Y,
$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$: Population mean of the auxiliary variable X,
$\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} z_i$: Population mean of the auxiliary variable Z,
$S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$: Population mean square of the study variable Y,
$S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$: Population mean square of the auxiliary variable X,
$S_z^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}\left(z_i - \bar{Z}\right)^2$: Population mean square of the auxiliary variable Z,
$C_y = \frac{S_y}{\bar{Y}}, C_x = \frac{S_x}{\bar{X}}$ and $C_z = \frac{S_z}{\bar{Z}}$ are the coefficient of variation of the study and auxiliary variable Y, X and Z respectively.
$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)$: Covariance between the study variable Y and auxiliary variable X.

$S_{yz} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(z_i - \bar{Z}\right)$: Covariance between the study variable *Y* and auxiliary variable *Z*.

$S_{xz} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(z_i - \bar{Z}\right)$: Covariance between the auxiliary variables *X* and *Z*.

$\rho_{yx} = \frac{S_{yx}}{S_y S_x}$, $\rho_{yz} = \frac{S_{yz}}{S_y S_z}$ and $\rho_{xz} = \frac{S_{xz}}{S_x S_z}$ are the correlation coefficient between the study and auxiliary variables (y, x) , (y, z) and (x, z) respectively.

A simple random sample of size *n* is drawn without replacement (*WOR*) method from the population *U* for estimating the population mean $\bar{Y}$. Let y_i, x_i and z_i be the values of the study and auxiliary variables on the i^{th} units (i = 1,2,. . . ,*N*). let

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: Sample mean of the study variable *Y*.
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$: Sample mean of the auxiliary variable *X*.
$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$: Sample mean of the auxiliary variable *Z*.

To obtain the bias and mean squared error (*MSE*) of the proposed class of estimators, we define the following error terms

$e_0 = \frac{\left(\bar{y}-\bar{Y}\right)}{\bar{Y}}$, $e_1 = \frac{\left(\bar{x}-\bar{X}\right)}{\bar{X}}$ and $e_2 = \frac{\left(\bar{z}-\bar{Z}\right)}{\bar{Z}}$ such that

$$E\left(e_0\right) = E\left(e_1\right) = E\left(e_2\right) = 0$$

and

$$E\left(e_0^2\right) = \left(\frac{1-f}{n}\right) C_y^2,\ E\left(e_1^2\right) = \left(\frac{1-f}{n}\right) C_x^2,\ E\left(e_2^2\right) = \left(\frac{1-f}{n}\right) C_z^2,$$

$E\left(e_0 e_1\right) = \left(\frac{1-f}{n}\right)\rho_{yx} C_y C_x$, $E\left(e_0 e_2\right) = \left(\frac{1-f}{n}\right)\rho_{yz} C_y C_z$ and $E\left(e_1 e_2\right) = \left(\frac{1-f}{n}\right)\rho_{xz} C_x C_z$.

where $f = \frac{n}{N}$ is the sampling fraction.

Sample mean estimator for population mean $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_{S(0)} = \bar{y} \tag{1}$$

The variance/*MSE* of $\bar{y}$ is given by

$$Var\left(\hat{\bar{Y}}_{S(0)}\right) = MSE\left(\hat{\bar{Y}}_{S(0)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2 C_y^2. \tag{2}$$

Ratio estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{S(R)} = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right), \tag{3}$$

The *MSE* of the ratio estimator $\hat{\bar{Y}}_{S(R)}$ to the first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}_{S(R)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x\right], \tag{4}$$

Ratio-type exponential estimator is given by

$$\hat{\bar{Y}}_{S(RExp)} = \bar{y}\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right), \tag{5}$$

The MSE of $\hat{\bar{Y}}_{S(RExp)}$ to the first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}_{S(RExp)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 + \frac{1}{4}C_x^2 - \rho_{yx} C_y C_x\right]. \tag{6}$$

Difference estimator for single auxiliary variable is defined by

$$\hat{\bar{Y}}_{S(D_1)} = \bar{y} + d_0 \left(\bar{X} - \bar{x} \right), \tag{7}$$

where d_0 is constant.

The minimum MSE of the difference estimator $\hat{\bar{Y}}_{S(D_1)}$ at optimum value of d_0 i.e., $d_{0(opt)} = \frac{\rho_{yx} C_y C_x}{C_x^2}$, is given by

$$MSE_{\min} \left(\hat{\bar{Y}}_{S(D_1)} \right) \cong \left(\frac{1-f}{n} \right) \bar{Y}^2 C_y^2 \left(1 - \rho_{yx}^2 \right). \tag{8}$$

Ratio cum product type estimator is defined by

$$\hat{\bar{Y}}_{S(RP)} = \bar{y} \left(\frac{\bar{X}}{\bar{x}} \right) \left(\frac{\bar{z}}{\bar{Z}} \right) \tag{9}$$

The *MSE* of $\hat{\bar{Y}}_{S(RP)}$ to first degree of approximation is given by

$$MSE \left(\hat{\bar{Y}}_{S(RP)} \right) = \bar{Y}^2 \left(\frac{1-f}{n} \right) \left[C_y^2 + C_x^2 + C_z^2 - 2\rho_{yx} C_y C_x + 2\rho_{yz} C_y C_z - 2\rho_{xz} C_x C_z \right] \tag{10}$$

Upadhyaya et. al. [11] proposed an exponential ratio cum product type estimator for two auxiliary variable as

$$\hat{\bar{Y}}_{S(RPExp)} = \bar{y} \exp \left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right) \exp \left(\frac{\bar{z} - \bar{Z}}{\bar{z} + \bar{Z}} \right) \tag{11}$$

The *MSE* of $\hat{\bar{Y}}_{S(RPExp)}$ to the first degree of approximation is given by

$$MSE \left(\hat{\bar{Y}}_{S(RPExp)} \right) = \bar{Y}^2 \left(\frac{1-f}{n} \right) \left[C_y^2 + \frac{1}{4} C_x^2 + \frac{1}{4} C_z^2 - \rho_{yx} C_y C_x + \rho_{yz} C_y C_z - \frac{1}{2} \rho_{xz} C_x C_z \right] \tag{12}$$

Traditional difference estimator for two auxiliary variables is defined by

$$\hat{\bar{Y}}_{S(D_2)} = \left\{ \bar{y} + d_1 \left(\bar{X} - \bar{x} \right) + d_2 \left(\bar{Z} - \bar{z} \right) \right\}, \tag{13}$$

where d_1 and d_2 are constants whose values are to be determined.

Minimum *MSE* of $\hat{\bar{Y}}_{S(D_2)}$ at optimum values of d_1 and d_2 i.e., $d_{1(opt)} = \frac{R_1 C_y (\rho_{yx} - \rho_{yz}\rho_{xz})}{C_x (1-\rho_{xz}^2)}$ and $d_{2(opt)} = \frac{R_2 C_y (\rho_{yz} - \rho_{yx}\rho_{xz})}{C_z (1-\rho_{xz}^2)}$, is given by

$$MSE_{\min} \left(\hat{\bar{Y}}_{S(D_2)} \right) \cong \left(\frac{1-f}{n} \right) \bar{Y}^2 C_y^2 \left(1 - R_{y.xz}^2 \right), \tag{14}$$

where $R_{y.xz}^2 = \frac{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}}{1-\rho_{xz}^2}$ is the multiple correlation coefficient, $R_1 = \frac{\bar{Y}}{\bar{X}}$ and $R_2 = \frac{\bar{Y}}{\bar{Z}}$.

Motivated by Gupta and Shabbir [4] and Singh and Singh [14], Muneer et. al. [8] proposed the following general class of estimators

$$\hat{\bar{Y}}_{S(MSK)} = \left[k_1 \bar{y} - k_2 \left(\bar{x} - \bar{X} \right) \right] \left[\alpha \left\{ 2 - \exp \left(\frac{\bar{z} - \bar{Z}}{\bar{z} + \bar{Z}} \right) + (1-\alpha) \exp \left(\frac{\bar{Z} - \bar{z}}{\bar{Z} + \bar{z}} \right) \right\} \right] \tag{15}$$

where $(k_i, i = 1, 2)$ are unknown constants whose values are to be determined.

The minimum *MSE* of $\hat{\hat{Y}}_{S(MSK)}$ at the optimum values of k_1 and k_2 to the first degree of approximation is given by

$$MSE_{\min}\left(\hat{\hat{Y}}_{S(MSK)}\right)=\bar{Y}^2\left[1-\left(\frac{1-f}{n}\right)\frac{\left(\rho_{xz}C_xC_z\right)^2}{4C_x^2}-\frac{U_1^2}{U_2}\right] \tag{16}$$

where

$$U_1=1+\left(\frac{3}{8}-\frac{\alpha}{4}\right)\left(\frac{1-f}{n}\right)C_z^2-\frac{1}{2}\left(\frac{1-f}{n}\right)\rho_{yz}C_yC_z-\left(\frac{1-f}{n}\right)\frac{\rho_{xz}C_xC_z\left(\rho_{xz}C_xC_z-\rho_{yx}C_yC_x\right)}{2C_x^2}$$

and

$$U_2=1+\left(\frac{1-f}{n}\right)C_y^2+\left(1-\frac{\alpha}{2}\right)\left(\frac{1-f}{n}\right)C_z^2-2\left(\frac{1-f}{n}\right)\rho_{yz}C_yC_z-\left(\frac{1-f}{n}\right)\frac{\left(\rho_{xz}C_xC_z-\rho_{yx}C_yC_x\right)^2}{C_x^2}.$$

Motivated by Gupta and Shabbir [4] and Grover and Kaur [3], Shabbir and Gupta [12] proposed a difference-cum-exponential ratio type estimator as

$$\hat{\hat{Y}}_{S(SG)}=\left\{d_1\bar{y}+d_2\left(\bar{X}-\bar{x}\right)+d_3\left(\bar{Z}-\bar{z}\right)\right\}\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right), \tag{17}$$

where $d_i\ (i=1,2,3)$ are constants.

Minimum mean squared error of Shabbir and Gupta [12] estimator is given by

$$MSE_{\min}\left(\hat{\hat{Y}}_{S(SG)}\right)=\bar{Y}^2\left\{1-\frac{\left(1+\frac{1}{64}\left(\frac{1-f}{n}\right)^2C_x^4\right)+\frac{1}{4}\left(\frac{1-f}{n}\right)^2C_y^2C_x^2\left(1-R_{y.xz}^2\right)}{1+\left(\frac{1-f}{n}\right)C_y^2\left(1-R_{y.xz}^2\right)}\right\} \tag{18}$$

In section 2 of this paper, we have proposed a class of estimators for population mean based on two auxiliary variables in simple random sampling and studied its properties up to the first order of approximation. Its stratified version has also been discussed in section 6. Theoretically and empirically we have shown that the proposed class of estimators is better than existing estimators in both simple random sampling and stratified random sampling.

2 Developed Estimator

Considering the same approach by Koyuncu [7], we have suggested the following class of estimators of population mean based on two auxiliary variables (*x*,*z*) as

$$T_S=\left\{w_0\bar{y}+w_1\left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha_1}+w_2\left(\frac{\bar{z}}{\bar{Z}}\right)^{\alpha_2}\right\}\exp\left\{\frac{\delta_1\left(\bar{X}-\bar{x}\right)}{\bar{X}+\bar{x}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z}-\bar{z}\right)}{\bar{Z}+\bar{z}}\right\}, \tag{19}$$

where (w_0,w_1,w_2) are suitably chosen weights and $(\alpha_1,\alpha_2,\delta_1,\delta_2)$ are design parameters.

Expanding (19) by using error terms, we have

$$T_S=\left[w_0\bar{Y}\left(1+e_0\right)+w_1\left(1+e_1\right)^{\alpha_1}+w_2\left(1+e_2\right)^{\alpha_2}\right]\exp\left\{\frac{-\delta_1e_1}{2+e_1}\right\}\exp\left\{\frac{-\delta_2e_2}{2+e_2}\right\}. \tag{20}$$

Expanding *RHS* of (20), multiplying and omitting the terms of *e*'s having power greater than two, we have

$$T_S = \left[\begin{array}{l} w_0\bar{Y}\left\{1+e_0-\left(\frac{\delta_1 e_1+\delta_2 e_2}{2}\right)-\left(\frac{\delta_1 e_0 e_1+\delta_2 e_0 e_2}{2}\right)+\frac{\delta_1(\delta_1+2)}{8}e_1^2+\frac{\delta_1\delta_2}{4}e_1e_2+\frac{\delta_2(\delta_2+2)}{8}e_2^2\right\} \\ +\frac{w_1}{R_1\bar{X}}\left\{1+\theta_1 e_1-\frac{\delta_2}{2}e_2+\frac{\theta_1(\theta_1-1)}{8}e_1^2-\frac{\delta_2\theta_1}{2}e_1e_2+\frac{\delta_2(\delta_2+2)}{8}e_2^2\right\} \\ +\frac{w_2}{R_2\bar{Z}}\left\{1+\theta_2 e_2-\frac{\delta_1}{2}e_1+\frac{\delta_1(\delta_1+2)}{8}e_1^2-\frac{\delta_1\theta_2}{2}e_1e_2+\frac{\theta_2(\theta_2-1)}{8}e_2^2\right\} \end{array} \right]$$

or

$$\left(T_S-\bar{Y}\right) = \left[\begin{array}{l} w_0\bar{Y}\left\{ \begin{array}{l} 1+e_0-\left(\frac{\delta_1 e_1+\delta_2 e_2}{2}\right)-\left(\frac{\delta_1 e_0 e_1+\delta_2 e_0 e_2}{2}\right)+\frac{\delta_1(\delta_1+2)}{8}e_1^2+\frac{\delta_1\delta_2}{4}e_1e_2 \\ +\frac{\delta_2(\delta_2+2)}{8}e_2^2 \end{array} \right\} \\ +\frac{w_1}{R_1\bar{X}}\left\{1+\theta_1 e_1-\frac{\delta_2}{2}e_2+\frac{\theta_1(\theta_1-1)}{2}e_1^2-\frac{\delta_2\theta_1}{2}e_1e_2+\frac{\delta_2(\delta_2+2)}{8}e_2^2\right\} \\ +\frac{w_2}{R_2\bar{Z}}\left\{1+\theta_2 e_2-\frac{\delta_1}{2}e_1+\frac{\delta_1(\delta_1+2)}{8}e_1^2-\frac{\delta_1\theta_2}{2}e_1e_2+\frac{\theta_2(\theta_2-1)}{2}e_2^2\right\}-\bar{Y} \end{array} \right] \tag{21}$$

where $\theta_1=\frac{(2\alpha_1-\delta_1)}{2}$ and $\theta_2=\frac{(2\alpha_2-\delta_2)}{2}$.

On taking expectation on both sides of (21), we get the bias of the proposed estimator T_S upto first order of approximation as

$$B\left(T_S\right)=\bar{Y}\left[w_0A_6+w_1A_7+w_2A_8-1\right], \tag{22}$$

where

$$\begin{aligned} A_6 &= \left[1+\frac{(1-f)}{n}\left\{\frac{\delta_1\left(\delta_1+2\right)}{8}C_x^2+\frac{\delta_1\delta_2}{4}\rho_{xz}C_xC_z+\frac{\delta_2\left(\delta_2+2\right)}{8}C_z^2-\frac{1}{2}\left(\delta_1\rho_{yx}C_yC_x+\delta_2\rho_{yz}C_yC_z\right)\right\}\right], \\ A_7 &= \frac{1}{R_1\bar{X}}\left[1+\frac{(1-f)}{n}\left\{\frac{\theta_1\left(\theta_1-1\right)}{2}C_x^2-\frac{\delta_2\theta_1}{2}\rho_{xz}C_xC_z+\frac{\delta_2\left(\delta_2+2\right)}{8}C_z^2\right\}\right], \\ A_8 &= \frac{1}{R_2\bar{Z}}\left[1+\frac{(1-f)}{n}\left\{\frac{\delta_1\left(\delta_1+2\right)}{8}C_x^2-\frac{\delta_1\theta_2}{2}\rho_{xz}C_xC_z+\frac{\theta_2\left(\theta_2-1\right)}{2}C_z^2\right\}\right]. \end{aligned}$$

On squaring both sides of (21), omitting terms of *e*'s having power greater than two and then taking expectation on both sides we get the *MSE* of the proposed class of estimators T_S to the first degree of approximation as

$$MSE\left(T_S\right)=\bar{Y}^2\left[\begin{array}{l} 1+w_0^2A_0+w_1^2A_1+w_2^2A_2+2w_0w_1A_3+2w_0w_2A_4+2w_1w_2A_5 \\ -2w_0A_6-2w_1A_7-2w_2A_8 \end{array} \right], \tag{23}$$

where

$$\begin{aligned} A_0 &= \left[1+\frac{(1-f)}{n}\left\{C_y^2+\frac{\delta_1\left(\delta_1+1\right)}{2}C_x^2+\delta_1\delta_2\rho_{xz}C_xC_z+\frac{\delta_2\left(\delta_2+1\right)}{2}C_z^2-2\left(\delta_1\rho_{yx}C_yC_x+\delta_2\rho_{yz}C_yC_z\right)\right\}\right], \\ A_1 &= \frac{1}{R_1^2\bar{X}^2}\left[1+\frac{(1-f)}{n}\left\{\theta_1\left(2\theta_1-1\right)C_x^2-2\theta_1\delta_2\rho_{xz}C_xC_z+\frac{\delta_2\left(\delta_2+1\right)}{2}C_z^2\right\}\right], \\ A_2 &= \frac{1}{R_2^2\bar{Z}^2}\left[1+\frac{(1-f)}{n}\left\{\frac{\delta_1\left(\delta_1+1\right)}{2}C_x^2-2\delta_1\theta_2\rho_{xz}C_xC_z+\theta_2\left(2\theta_2-1\right)C_z^2\right\}\right], \end{aligned}$$

$$\begin{aligned}
A_3 &= \frac{1}{R_1\bar{X}}\left[1+\frac{(1-f)}{n}\left\{\begin{array}{l}(\alpha_1-\delta_1)\,\rho_{yx}C_yC_x-\delta_2\rho_{yz}C_yC_z+\frac{(\alpha_1-\delta_1)(\alpha_1-\delta_1-1)}{2}C_x^2\\ -\delta_2\,(\alpha_1-\delta_1)\,\rho_{xz}C_xC_z+\frac{\delta_2(\delta_2+1)}{2}C_z^2\end{array}\right\}\right],\\
A_4 &= \frac{1}{R_2\bar{Z}}\left[1+\frac{(1-f)}{n}\left\{\begin{array}{l}(\alpha_2-\delta_2)\,\rho_{yz}C_yC_z+\frac{\delta_1(\delta_1+1)}{2}C_x^2-\delta_1\,(\alpha_2-\delta_2)\,\rho_{xz}C_xC_z\\ +\frac{(\alpha_2-\delta_2)(\alpha_2-\delta_2-1)}{2}C_z^2\end{array}\right\}\right],\\
A_5 &= \frac{1}{R_1R_2\bar{X}\bar{Z}}\left[1+\frac{(1-f)}{n}\left\{\begin{array}{l}\frac{(\alpha_1-\delta_1)(\alpha_1-\delta_1-1)}{2}C_x^2+\frac{(\alpha_2-\delta_2)(\alpha_2-\delta_2-1)}{2}C_z^2\\ +(\alpha_1-\delta_1)\,(\alpha_2-\delta_2)\,\rho_{xz}C_xC_z\end{array}\right\}\right]
\end{aligned}$$

A_6, A_7 and A_8 are same as defined earlier.

To find the optimum values of $(w_0,\, w_1, w_2\,)$, minimization of $MSE\,(T_S)$ at (23) with respect to $(w_0,\, w_1, w_2\,)$, which yields

$$\begin{bmatrix} A_0 & A_3 & A_4\\ A_3 & A_1 & A_5\\ A_4 & A_5 & A_2\end{bmatrix}\begin{bmatrix} w_0\\ w_1\\ w_2\end{bmatrix}=\begin{bmatrix} A_6\\ A_7\\ A_8\end{bmatrix} \tag{24}$$

and after simplifying (24), the optimum values are

$$\left.\begin{array}{l} w_{0(opt)}=\frac{\Delta_0}{\Delta},\\ w_{1(opt)}=\frac{\Delta_1}{\Delta},\\ w_{2(opt)}=\frac{\Delta_2}{\Delta}.\end{array}\right\} \tag{25}$$

where

$$\Delta=\begin{vmatrix} A_0 & A_3 & A_4\\ A_3 & A_1 & A_5\\ A_4 & A_5 & A_2\end{vmatrix}=A_0\left(A_1A_2-A_5^2\right)-A_3\left(A_2A_3-A_4A_5\right)+A_4\left(A_3A_5-A_1A_4\right),$$

$$\Delta_0=\begin{vmatrix} A_6 & A_3 & A_4\\ A_7 & A_1 & A_5\\ A_8 & A_5 & A_2\end{vmatrix}=A_6\left(A_1A_2-A_5^2\right)-A_3\left(A_2A_7-A_5A_8\right)+A_4\left(A_5A_7-A_1A_8\right),$$

$$\Delta_1=\begin{vmatrix} A_0 & A_6 & A_4\\ A_3 & A_7 & A_5\\ A_4 & A_8 & A_2\end{vmatrix}=A_0\left(A_2A_7-A_5A_8\right)-A_6\left(A_2A_3-A_4A_5\right)+A_4\left(A_3A_8-A_4A_7\right),$$

$$\Delta_2=\begin{vmatrix} A_0 & A_3 & A_6\\ A_3 & A_1 & A_7\\ A_4 & A_5 & A_8\end{vmatrix}=A_0\left(A_1A_8-A_5A_7\right)-A_3\left(A_3A_8-A_4A_7\right)+A_6\left(A_3A_5-A_1A_4\right).$$

Thus the resulting minimum *MSE* of T_S is given by

$$MSE_{\min}\left(T_S\right)=\bar{Y}^2\left[1-\frac{\left(A_6\Delta_0+A_7\Delta_1+A_8\Delta_2\right)}{\Delta}\right]. \tag{26}$$

3 Special Cases

1. For $w_2=0$, the proposed class of estimators reduces to:

$$T_{S(1)} = \left[w_0\bar{y} + w_1\left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha_1}\right]\exp\left\{\frac{\delta_1\left(\bar{X}-\bar{x}\right)}{\bar{X}+\bar{x}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z}-\bar{z}\right)}{\bar{Z}+\bar{z}}\right\} \tag{27}$$

Bias and MSE of $T_{S(1)}$ are respectively given as

$$B\left(T_{S(1)}\right) = \bar{Y}\left[w_0A_6 + w_1A_7 - 1\right], \tag{28}$$

and

$$MSE\left(T_{S(1)}\right) = \bar{Y}^2\left[1 + w_0^2A_0 + w_1^2A_1 + 2w_0w_1A_3 - 2w_0A_6 - 2w_1A_7\right]. \tag{29}$$

Minimizing (29) with respect to $(w_0,\, w_1)$ yields,

$$\begin{bmatrix} A_0 & A_3 \\ A_3 & A_1 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} A_6 \\ A_7 \end{bmatrix} \tag{30}$$

After simplifying (30), we obtain the optimum values of $(w_0,\, w_1)$ as

$$\left.\begin{array}{l} w_0^* = \frac{\Delta_0^*}{\Delta^*}, \\ w_1^* = \frac{\Delta_1^*}{\Delta^*}. \end{array}\right\} \tag{31}$$

Thus the resulting minimum MSE of $T_{S(1)}$ is

$$MSE_{\min}\left(T_{S(1)}\right) = \bar{Y}^2\left[1 - \frac{\left(A_6\Delta_0^* + A_7\Delta_1^*\right)}{\Delta^*}\right]. \tag{32}$$

where

$$\Delta^* = \begin{vmatrix} A_0 & A_3 \\ A_3 & A_1 \end{vmatrix} = \left(A_0A_1 - A_3^2\right)$$

$$\Delta_0^* = \begin{vmatrix} A_6 & A_3 \\ A_7 & A_1 \end{vmatrix} = \left(A_1A_6 - A_3A_7\right)$$

$$\Delta_1^* = \begin{vmatrix} A_0 & A_6 \\ A_3 & A_7 \end{vmatrix} = \left(A_0A_7 - A_3A_6\right).$$

2. For $w_1 = 0$, the proposed class of estimators reduces to:

$$T_{S(2)} = \left[w_0\bar{y} + w_2\left(\frac{\bar{z}}{\bar{Z}}\right)^{\alpha_2}\right]\exp\left\{\frac{\delta_1\left(\bar{X}-\bar{x}\right)}{\bar{X}+\bar{x}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z}-\bar{z}\right)}{\bar{Z}+\bar{z}}\right\} \tag{33}$$

Bias and MSE of $T_{S(2)}$ are respectively given as

$$B\left(T_{S(2)}\right) = \bar{Y}\left[w_0A_6 + w_2A_8 - 1\right], \tag{34}$$

and

$$MSE\left(T_{S(2)}\right) = \bar{Y}^2\left[1 + w_0^2A_0 + w_2^2A_2 + 2w_0w_2A_4 - 2w_0A_6 - 2w_2A_8\right]. \tag{35}$$

Minimizing (35) with respect to $(w_0,\, w_2)$ yields,

$$\begin{bmatrix} A_0 & A_4 \\ A_4 & A_2 \end{bmatrix}\begin{bmatrix} w_0 \\ w_2 \end{bmatrix} = \begin{bmatrix} A_6 \\ A_8 \end{bmatrix} \tag{36}$$

After simplifying (36), we obtain the optimum values of $(w_0,\, w_2)$ as

$$\left.\begin{array}{l} w_0^{**} = \frac{\Delta_0^{**}}{\Delta^{**}}, \\ w_2^{**} = \frac{\Delta_2^{**}}{\Delta^{**}}. \end{array}\right\} \tag{37}$$

Thus the resulting minimum MSE of $T_{S(2)}$ is

$$MSE_{\min}\left(T_{S(2)}\right) = \bar{Y}^2\left[1 - \frac{\left(A_6\Delta_0^{**} + A_8\Delta_2^{**}\right)}{\Delta^{**}}\right]. \tag{38}$$

where

$$\Delta^{**} = \begin{vmatrix} A_0 & A_4 \\ A_4 & A_2 \end{vmatrix} = \left(A_0A_2 - A_4^2\right)$$

$$\Delta_0^{**} = \begin{vmatrix} A_6 & A_4 \\ A_8 & A_2 \end{vmatrix} = \left(A_6A_2 - A_4A_8\right)$$

$$\Delta_2^{**} = \begin{vmatrix} A_0 & A_6 \\ A_4 & A_8 \end{vmatrix} = \left(A_0A_8 - A_4A_6\right).$$

4 Efficiency Comparison

From (2), (4), (6) and (8) respectively, we have

$$MSE\left(\hat{\bar{Y}}_{S(0)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2C_y^2\rho_{yx}^2 \geq 0. \tag{39}$$

$$MSE\left(\hat{\bar{Y}}_{S(R)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2\rho_{yx}^2 - 2\rho_{yx}C_yC_x + C_x^2\right] \geq 0. \tag{40}$$

$$MSE\left(\hat{\bar{Y}}_{S(RExp)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2\rho_{yx}^2 - \rho_{yx}C_yC_x + \frac{1}{4}C_x^2\right] \geq 0. \tag{41}$$

It follows from (39), (40), and (41) that the difference estimator for single auxiliary variable is more efficient than $\hat{\bar{Y}}_{S(0)}$,$\hat{\bar{Y}}_{S(R)}$ and $\hat{\bar{Y}}_{S(RExp)}$.

From (10) and (14), we have

$$\begin{aligned} & MSE\left(\hat{\bar{Y}}_{S(RP)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) \\ = & \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2R_{y.xz}^2 + C_x^2 + C_z^2 - 2\rho_{yx}C_yC_x + 2\rho_{yz}C_yC_z - 2\rho_{xz}C_xC_z\right] \\ & \geq 0. \end{aligned} \tag{42}$$

Thus from (42), the difference estimator for two auxiliary variable is more efficient than $\hat{\bar{Y}}_{S(RP)}$.

From (12) and (14), we have

$$\begin{aligned} & MSE\left(\hat{\bar{Y}}_{S(RPExp)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) \\ = & \left(\frac{1-f}{n}\right)\bar{Y}^2\left[C_y^2 R_{y.xz}^2 + \frac{1}{4}C_x^2 + \frac{1}{4}C_z^2 - \rho_{yx}C_yC_x + \rho_{yz}C_yC_z - \frac{1}{2}\rho_{xz}C_xC_z\right] \\ & \geq 0. \end{aligned} \tag{43}$$

Thus from (43), the difference estimator for two auxiliary variable is more efficient than $\hat{\bar{Y}}_{S(RPExp)}$.

From (8) and (14), we have

$$MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) = \left(\frac{1-f}{n}\right)\bar{Y}^2 C_y^2\left(R_{y.xz}^2 - \rho_{yx}^2\right) \geq 0. \tag{44}$$

From (39), (40), (41) and (44), we have the following inequalities:

$$MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) \leq MSE\left(\hat{\bar{Y}}_{S(0)}\right) \tag{45}$$

$$MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) \leq MSE\left(\hat{\bar{Y}}_{S(R)}\right) \tag{46}$$

$$MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(D_1)}\right) \leq MSE\left(\hat{\bar{Y}}_{S(RExp)}\right) \tag{47}$$

Hence $\hat{\bar{Y}}_{S(D_2)}$ is more efficient than the sample mean, ratio-type, ratio-type exponential and difference estimator for single auxiliary variable.

From (14) and (26), we have that

$MSE_{\min}\left(T_S\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(D_2)}\right)$ if

$$\left[\left(\frac{1-f}{n}\right)C_y^2\left(1-R_{y.xz}^2\right)\right] \geq \left[1 - \frac{(A_6\Delta_0 + A_7\Delta_1 + A_8\Delta_2)}{\Delta}\right] \tag{48}$$

Thus the proposed estimator T_S is more efficient than the difference estimator for two auxiliary variables as long as the condition (48) holds.

Further from (42) to (47), we can see that our proposed class of estimators $T_S$is also more efficient than the sample mean, ratio-type, ratio-type exponential, difference estimator for single auxiliary variable, ratio cum product-type estimator and ratio cum product-type exponential estimators.

From (16) and (26), we have

$MSE_{\min}\left(T_S\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(MSK)}\right)$ if

$$\left[\frac{(A_6\Delta_0 + A_7\Delta_1 + A_8\Delta_2)}{\Delta}\right] \geq \left[\left(\frac{1-f}{n}\right)\frac{(\rho_{xz}C_xC_z)^2}{C_x^2} - \frac{U_1^2}{U_2}\right] \tag{49}$$

Thus the proposed estimator T_S is more efficient than Muneer et al (2016) estimator as long as the condition (49) is true.

From (18) and (26), we have

$MSE_{\min}\left(T_S\right) \leq MSE_{\min}\left(\hat{\bar{Y}}_{S(SG)}\right)$ if

$$\left[\frac{(A_6\Delta_0 + A_7\Delta_1 + A_8\Delta_2)}{\Delta}\right] \geq \left[\frac{1 + \frac{1}{64}\left(\frac{1-f}{n}\right)^2 C_x^4 + \frac{1}{4}\left(\frac{1-f}{n}\right)^2 C_y^2 C_x^2\left(1-R_{y.xz}^2\right)}{1 + \left(\frac{1-f}{n}\right)C_y^2\left(1-R_{y.xz}^2\right)}\right] \tag{50}$$

Thus the proposed estimator T_S is more efficient than Shabbir and Gupta [12] estimator as long as the condition (50) holds.

Now from (32) and (26), we have

$MSE_{\min}\left(T_S\right) \leq MSE_{\min}\left(T_{S(1)}\right)$ if

$$\left[\frac{A_6\Delta_0 + A_7\Delta_1 + A_8\Delta_2}{\Delta}\right] \geq \left[\frac{A_6\Delta_0^* + A_7\Delta_1^*}{\Delta^*}\right] \tag{51}$$

which is always true.

From (38) and (26), we have

$MSE_{\min}\left(T_S\right) \leq MSE_{\min}\left(T_{S(2)}\right)$, if

$$\left[\frac{A_6\Delta_0 + A_7\Delta_1 + A_8\Delta_2}{\Delta}\right] \geq \left[\frac{A_6\Delta_0^{**} + A_8\Delta_2^{**}}{\Delta^{**}}\right] \tag{52}$$

which always holds.

Thus the proposed class of estimators T_S is better than the estimators $T_{S(1)}$ and $T_{S(2)}$ at their optimum conditions.

5 Numerical Illustrations

For numerical comparisons we considered following four data sets

Data set 1 [Source: Hair [5]]

y: Preceived level of price charged by product suppliers.

x: Overall level of service necessary for maintaining a satisfactory relationship between suppliers and purchaser.

z: Overall image of manufacturer/suppliers.

$N = 100,\ n = 29,\ \bar{Y} = 2.3640,\ \bar{X} = 2.9250,\ \bar{Z} = 5.2390,$

$\rho_{yx} = 0.1602,\ \rho_{yz} = 0.0829,\ \rho_{xz} = 0.0846, C_y^2 = 2.5582,\ C_x^2 = 0.0661,\ C_z^2 = 0.0461.$

Data set 2 [Source: Singh [13], pp. 1119-1121]

y: Tobacco yield (metric tons) in specified countries during (1998).

x: Tobacco area (hectares) in specified countries during (1998).

z: Tobacco production (metric tons) in specified countries during (1998).

$N = 106,\ n = 31,\ \bar{Y} = 1.5507,\ \bar{X} = 34438.61,\ \bar{Z} = 52444.56,$

$\rho_{yx} = -0.0077,\ \rho_{yz} = 0.0304,\ \rho_{xz} = 0.9912, C_y^2 = 0.2629,\ C_x^2 = 18.8364,\ C_z^2 = 23.3405.$

Data set 3 [Source: MFA [2]]

y: District wise tomato production in tones of Pakistan for (2003).

x: District wise tomato production in tones of Pakistan for (2002).

z: District wise tomato production in tones of Pakistan for (2001).

$N = 97,\ n = 30,\ \bar{Y} = 3135.6186,\ \bar{X} = 3050.2784,\ \bar{Z} = 2743.9587,$

$\rho_{yx} = 0.8072,\ \rho_{yz} = 0.8501,\ \rho_{xz} = 0.6122, C_y^2 = 4.8674,\ C_x^2 = 5.4812,\ C_z^2 = 6.2422.$

Data set 4 [Source: Jhonston [6]]

y: Percentage of living affected by disease.

x: Mean January temperature.

z: Date of flowering of particular summer flowering species (no. of days from January 1).

$N = 10,\ n = 2,\ \bar{Y} = 52,\ \bar{X} = 42,\ \bar{Z} = 200,$

$\rho_{yx} = 0.7966,\ \rho_{yz} = -0.9364,\ \rho_{xz} = -0.7333, C_y^2 = 0.4997,\ C_x^2 = 0.2440,\ C_z^2 = 0.0021.$

Table 1 gives the *PRE*'s of different estimators considered in this paper with respect to $\bar{y}$.

Table 2 to 5 gives the *PRE*'s of the proposed class of estimator with respect to$\bar{y}$ at different values of $(\alpha_1, \alpha_2, \delta_1, \delta_2)$.

Estimator	Data 1	Data 2	Data 3	Data 4
$\hat{\bar{Y}}_{S(R)}$	102.63	1.37	242.17	266.67
$\hat{\bar{Y}}_{S(RExp)}$	101.97	5.27	235.33	176.86
$\hat{\bar{Y}}_{S(D_1)}$	102.63	100.01	287.00	273.65
$\hat{\bar{Y}}_{S(RP)}$	98.92	24.85	46.53	308.42
$\hat{\bar{Y}}_{S(RPExp)}$	100.45	51.76	74.60	191.43
$\hat{\bar{Y}}_{S(D_2)}$	103.15	109.00	685.89	1030.83
$\hat{\bar{Y}}_{S(MSK)}$(at $\alpha = 1$)	109.38	404.14	685.99	278.42
$\hat{\bar{Y}}_{S(MSK)}$(at $\alpha = 0$)	109.44	787.52	736.58	778.53
$\hat{\bar{Y}}_{S(SG)}$	109.45	297.93	731.30	1085.63

Table 1: *PRE*'s of $\hat{\bar{Y}}_{S(R)}, \hat{\bar{Y}}_{S(RExp)}, \hat{\bar{Y}}_{S(D_1)}, \hat{\bar{Y}}_{S(RP)}, \hat{\bar{Y}}_{S(RPExp)}, \hat{\bar{Y}}_{S(D_2)}$, $\hat{\bar{Y}}_{S(MSK)}$ and $\hat{\bar{Y}}_{S(SG)}$ with respect to $\bar{y}$.

α_1	α_2	δ_1	δ_2	*PRE*
1	1	-1	-1	2188.35
-1	-1	1	1	2359.94
1	1	0	0	8727.61
0.75	0.75	0	0	15563.07
0.25	0.25	1	1	15788.04
0.75	0.75	0.25	0.25	35012.66
0.5	0.5	1	1	35276.14
0.25	0.25	0.75	0.75	35489.64
1	-1	1	-1	111987.2
0.75	0.75	0.5	0.5	140060
0.75	0.75	1	1	140564.9
0.5	0.5	0.25	0.25	140748
0.5	0.5	0.75	0.75	141050.6
0.25	0.25	0	0	141679.5
0.25	0.25	0.5	0.5	141859.8
0.1	0.1	0	0	889758.5
0.09	0.09	0	0	1098841
0.05	0.05	0	0	3565267
0.1	0.1	0.1	0	3946147
0.05	0.05	0.01	0.01	5570472

Table 2: *PRE* of the proposed estimator T_S with respect to $\bar{y}$ for data set 1

From Tables 1 and 2 to 5 we found that our proposed class of estimators gives the highest *PRE* for all the data sets at different values of scalars $(\alpha_1, \alpha_2, \delta_1, \delta_2)$ (5570472, 23098.78, 160144.7 and 9842320 for data sets 1 to 4 respectively) which are higher than the estimators $\hat{\bar{Y}}_{S(R)}, \hat{\bar{Y}}_{S(RExp)}, \hat{\bar{Y}}_{S(D_1)}, \hat{\bar{Y}}_{S(RP)}, \hat{\bar{Y}}_{S(RPExp)}, \hat{\bar{Y}}_{S(D_2)}$, $\hat{\bar{Y}}_{S(MSK)}$ and $\hat{\bar{Y}}_{S(SG)}$.

6 Extension to Stratified Random Sampling

Let a population of size *N* is divided into *L* strata with the h^{th} stratum consisting of N_h units, such that $\sum_{h=1}^{L} N_h = N$, $(h = 1, 2, ...N)$. Let us drawn a sample of size n_h by *SRSWOR* from h^{th} stratum, such

α_1	α_2	δ_1	δ_2	*PRE*
0.06	0.06	0	0	795.81
0.05	0.05	0	0	1076.74
0.05	0.05	0.1	0.1	1081.47
0.2	0.2	0.05	0.1	1122.87
0.1	0.1	0.05	0.05	1511.94
0.1	0.1	0	0.1	1512.19
0.04	0.04	0	0	1588.34
0.05	0.05	0.01	0.01	1682.74
0.03	0.03	0	0	2681.58
0.04	0.04	0.01	0.01	2824.72
0.025	0.025	0	0	3772.71
0.02	0.02	0	0	5770.3
0.03	0.03	0.01	0.01	6036.85
0.1	0.05	0	0	6964.55
0.02	0.02	0.01	0.01	23098.78

Table 3: *PRE* of the proposed estimator T_S with respect to $\bar{y}$ for data set 2

α_1	α_2	δ_1	δ_2	*PRE*
-1	-1	1	1	759.22
0.1	0.1	1	1	828.79
0.25	0.25	1	1	877.46
0.75	0.75	1	1	2444.84
0.5	0.5	0.25	0.25	3837.87
0.5	0.5	0.75	0.75	4330.34
0.25	0.25	0	0	7851.37
0.25	0.25	0.5	0.5	8026.03
0.1	0	0.1	0.1	49382.6
0.1	0.1	0	0	74815.65
-1	-1	0.1	0.1	91094.72
0.09	0.09	0	0	94968.67
0.025	0.025	0.1	0.1	160144.7

Table 4: *PRE* of the proposed estimator T_S with respect to $\bar{y}$ for data set 3

that $\sum_{h=1}^{L} n_h = n$. For the i^{th} unit in the h^{th} stratum let y_{hi} and (x_{hi}, z_{hi}) be the values of the study and auxiliary variables respectively.

Let $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ and $\bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h$ be the sample means corresponding to the population means $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$, $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ and $\bar{Z} = \sum_{h=1}^{L} W_h \bar{Z}_h$ respectively, where $\bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}, \bar{x}_h = \sum_{i=1}^{n_h} \frac{x_{hi}}{n_h}$, $\bar{z}_h = \sum_{i=1}^{n_h} \frac{z_{hi}}{n_h}$, $\bar{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}$, $\bar{X}_h = \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}$, $\bar{Z}_h = \sum_{i=1}^{N_h} \frac{z_{hi}}{N_h}$, and $W_h = \frac{N_h}{N}$is the known stratum weight.

To obtain the bias and mean squared error (*MSE*) of the proposed estimator, we define the following error terms

$e_{0(st)} = \frac{(\bar{y}_{st} - \bar{Y})}{\bar{Y}}$, $e_{1(st)} = \frac{(\bar{x}_{st} - \bar{X})}{\bar{X}}$ and $e_{2(st)} = \frac{(\bar{z}_{st} - \bar{Z})}{\bar{Z}}$ such that

$$E\left(e_{0(st)}\right) = E\left(e_{1(st)}\right) = E\left(e_{2(st)}\right) = 0$$

and

$$E\left(e_{0(st)}^2\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{yh}^2}{\bar{Y}^2},\ E\left(e_{1(st)}^2\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{xh}^2}{\bar{X}^2},\ E\left(e_{2(st)}^2\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{zh}^2}{\bar{Z}^2},$$

$$E\left(e_{0(st)} e_{1(st)}\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{yxh}}{\bar{Y}\,\bar{X}},\ E\left(e_{0(st)} e_{2(st)}\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{yzh}}{\bar{Y}\,\bar{Z}} \text{and}$$

$$E\left(e_{1(st)} e_{2(st)}\right) = \frac{\sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) S_{xzh}}{\bar{X}\,\bar{Z}}.$$

α_1	α_2	δ_1	δ_2	*PRE*
-1	1	1	1	1230
1	1	2.5	2.5	109002.9
0.25	0.5	0.25	0.5	200124.8
1	1	0	0	217802.8
1	1	0.05	0.05	246186.3
0.5	0.5	1.25	1.25	371887.4
0.75	0.75	0	0	375909.5
0.25	0.25	1	1	536357.6
0.75	0.75	0.1	0.1	540596.2
1	-1	0	-1	694455.3
1	0	0	1	776344.7
0.5	0.5	0	0	821113.8
0.25	0.25	0.75	0.75	846732.7
0.5	0.5	1	1	959799.5
0.5	0.5	1	1	959799.5
0.25	0.25	1.25	1.25	1088420
0.5	0.5	0.1	0.1	1385435
0.25	0.25	0	0	3188881
0.25	0.25	0.5	0.5	3330473
0.25	0.25	0.1	0.1	9842320

Table 5: *PRE* of the proposed estimator T_S with respect to $\bar{y}$ for data set 4

where $S_{yxh}=\rho_{yxh}S_{yh}S_{xh},\ S_{yzh}=\rho_{yzh}S_{yh}S_{zh},\ S_{xzh}=\rho_{xzh}S_{xh}S_{zh}$ and $f_h=\frac{n_h}{N_h}$ is the sampling fraction.

Also $V_{200}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{yh}^2,\ V_{020}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xh}^2,\ V_{002}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{zh}^2,$

$$V_{110}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{yxh},\ V_{101}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{yzh},V_{011}=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{xzh},$$

$$C_{200}=\frac{V_{200}}{\bar{Y}^2},C_{020}=\frac{V_{020}}{\bar{X}^2},C_{002}=\frac{V_{002}}{\bar{Z}^2},C_{110}=\frac{V_{110}}{\bar{Y}\,\bar{X}},C_{101}=\frac{V_{101}}{\bar{Y}\,\bar{Z}},C_{011}=\frac{V_{011}}{\bar{X}\,\bar{Z}},$$

$R_1=\frac{\bar{Y}}{\bar{X}}$ and $R_2=\frac{\bar{Y}}{\bar{Z}}$.

Sample mean estimator for population mean $\bar{Y}$ in stratified random sampling is defined as

$$\hat{\bar{Y}}_{St(o)}=\bar{y}_{st} \tag{53}$$

The variance/*MSE* of $\bar{y}_{st}$is given by

$$Var\left(\hat{\bar{Y}}_{St(0)}\right)=MSE\left(\hat{\bar{Y}}_{St(0)}\right)=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)S_{yh}^2.=V_{200} \tag{54}$$

Combined ratio-type estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{St(R)}^{C}=\bar{y}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right), \tag{55}$$

The *MSE* of the ratio estimator $\hat{\bar{Y}}_{St(R)}^{C}$ to the first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}_{St(R)}^{C}\right)=V_{200}+R_1^2V_{020}-2R_1V_{110}, \tag{56}$$

Combined ratio-type exponential estimator is given by

$$\hat{\bar{Y}}_{St(RExp)}^{C}=\bar{y}_{st}\exp\left(\frac{\bar{X}-\bar{x}_{st}}{\bar{X}+\bar{x}_{st}}\right), \tag{57}$$

The MSE of $\hat{\bar{Y}}^{C}_{St(RExp)}$ to the first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}^{C}_{St(RExp)}\right) = V_{200} + \frac{1}{4}R_1^2 V_{020} - R_1 V_{110}. \tag{58}$$

Difference estimator for single auxiliary variable is defined by

$$\hat{\bar{Y}}^{C}_{St(D_1)} = \bar{y}_{st} + d_{0(st)}\left(\bar{X} - \bar{x}_{st}\right), \tag{59}$$

where$d_{0(st)}$ is constant.

The minimum MSE of the difference estimator$\hat{\bar{Y}}^{C}_{St(D_1)}$ at optimum value of $d_{0(st)}$i.e. $d_{0(st)(opt)} = \frac{V_{110}}{V_{020}}$, is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_1)}\right) \cong V_{200} - \frac{V_{110}^2}{V_{020}}. \tag{60}$$

Combined ratio cum product type estimator is defined by

$$\hat{\bar{Y}}^{C}_{St(RP)} = \bar{y}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right)\left(\frac{\bar{z}_{st}}{\bar{Z}}\right) \tag{61}$$

The *MSE* of $\hat{\bar{Y}}^{C}_{St(RP)}$to first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}^{C}_{St(RP)}\right) = \left[V_{200} + R_1^2 V_{020} + R_2^2 V_{002} - 2R_1 V_{110} + 2R_2 V_{101} - 2R_1 R_2 V_{011}\right] \tag{62}$$

Stratified version of Upadhyaya et. al. [11] exponential ratio cum product type estimator for two auxiliary variables is given by

$$\hat{\bar{Y}}^{C}_{St(RPExp)} = \bar{y}_{st}\exp\left(\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}}\right)\exp\left(\frac{\bar{z}_{st} - \bar{Z}}{\bar{z}_{st} + \bar{Z}}\right) \tag{63}$$

The *MSE* of $\hat{\bar{Y}}^{C}_{St(RPExp)}$to the first degree of approximation is given by

$$MSE\left(\hat{\bar{Y}}^{C}_{St(RPExp)}\right) = \left[V_{200} + \frac{R_1^2}{4}V_{020} + \frac{R_2^2}{4}V_{002} - R_1 V_{110} + R_2 V_{101} - \frac{R_1 R_2}{2}V_{011}\right] \tag{64}$$

Combined traditional difference estimator for two auxiliary variables is defined by

$$\hat{\bar{Y}}^{C}_{St(D_2)} = \left\{\bar{y}_{st} + d_{1(st)}\left(\bar{X} - \bar{x}_{st}\right) + d_{2(st)}\left(\bar{Z} - \bar{z}_{st}\right)\right\}, \tag{65}$$

where$d_{1(st)}$and $d_{2(st)}$ are constants whose values are to be determined.

Minimum *MSE* of $\hat{\bar{Y}}^{C}_{St(D_2)}$at optimum values of $d_{1(st)}$and $d_{2(st)}$ i.e $d_{1(st)(opt)} = \frac{(V_{002}V_{110} - V_{011}V_{101})}{\left(V_{020}V_{002} - V_{011}^2\right)}$ and $d_{2(st)(opt)} = \frac{(V_{020}V_{101} - V_{011}V_{110})}{\left(V_{020}V_{002} - V_{011}^2\right)}$, is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right) \cong \left[V_{200} - \frac{(V_{110}^2 V_{002} - 2V_{011}V_{101}V_{110} + V_{020}V_{101}^2)}{(V_{020}V_{002} - V_{011}^2)}\right], \tag{66}$$

Motivated by Gupta and Shabbir [4] and Singh and Singh [14], Muneer et. al. [8] proposed the following general class of estimators in stratified random sampling scheme as

$$\hat{\bar{Y}}^{C}_{St(MSK)} = \left[k_{3st}\bar{y}_{st} - k_{4st}\left(\bar{x}_{st} - \bar{X}\right)\right]\left[\alpha\left\{2 - \exp\left(\frac{\bar{z}_{st} - \bar{Z}}{\bar{z}_{st} + \bar{Z}}\right) + (1-\alpha)\exp\left(\frac{\bar{Z} - \bar{z}_{st}}{\bar{Z} + \bar{z}_{st}}\right)\right\}\right] \tag{67}$$

where $(k_i, i = 3, 4)$ are unknown constants whose values are to be determined.

The minimum *MSE* of $\hat{\bar{Y}}^C_{St(MSK)}$at the optimum values of k_3 and k_4 to the first degree of approximation is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^C_{St(MSK)}\right) = \bar{Y}^2\left[1 - \frac{C_{011}^2}{4C_{020}} - \frac{A_{st}^2}{B_{st}}\right] \tag{68}$$

where

$$A_{st} = 1 + \left(\frac{3}{8} - \frac{\alpha}{4}\right)C_{002} - \frac{1}{2}C_{101} - \frac{C_{011}\left(C_{011} - C_{110}\right)}{2C_{020}}$$

and

$$B_{st} = 1 + C_{200} + \left(1 - \frac{\alpha}{2}\right)C_{002} - 2C_{101} - \frac{\left(C_{011} - C_{110}\right)^2}{C_{020}}.$$

Motivated by Shabbir and Gupta [12], we define a class of combined difference estimator for population mean $\bar{Y}$ as

$$\hat{\bar{Y}}^C_{St(SG)} = \left\{d_1\bar{y}_{st} + d_2\left(\bar{X} - \bar{x}_{st}\right) + d_3\left(\bar{Z} - \bar{z}_{st}\right)\right\}\exp\left(\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}}\right), \tag{69}$$

where $d_i\,(i = 1, 2, 3)$ are suitably chosen constants.

To the first degree of approximation, the bias and mean squared error of $\hat{\bar{Y}}^C_{St(SG)}$ are respectively given by

$$B\left(\hat{\bar{Y}}^C_{St(SG)}\right) = \bar{Y}\left[a_7d_1 + a_8d_2 + a_9d_3 - 1\right], \tag{70}$$

and

$$MSE\left(\hat{\bar{Y}}^C_{St(SG)}\right) = \bar{Y}^2\left[\begin{array}{l} 1 + a_1d_1^2 + a_2d_2^2 + a_3d_3^2 + 2a_4d_1d_2 + 2a_5d_1d_3 + 2a_6d_2d_3 \\ -2a_7d_1 - 2a_8d_2 - 2a_9d_3 \end{array}\right]. \tag{71}$$

where

$$\begin{gathered}
a_1 = \left(1 + C_{200} + C_{020} - 2C_{110}\right)\,, \\
a_2 = \tfrac{1}{R_1^2}C_{020}\,,\ \ a_3 = \tfrac{1}{R_2^2}C_{002}\,, \\
a_4 = \tfrac{1}{R_1}\left(C_{020} - C_{110}\right)\,, \\
a_5 = \tfrac{1}{R_2}\left(C_{011} - C_{101}\right)\,, \\
a_6 = \tfrac{1}{R_1R_2}C_{011} \\
a_7 = \left(1 - \tfrac{1}{2}C_{110} + \tfrac{3}{8}C_{020}\right)\,, \\
a_8 = \frac{C_{020}}{2R_1}\,,\ \ a_9 = \frac{C_{011}}{2R_2}\,.
\end{gathered}$$

Minimum *MSE* of $\hat{\bar{Y}}^C_{St(SG)}$ at optimum values of $d_i\,(i = 1, 2, 3)$ is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^C_{St(SG)}\right) = \bar{Y}^2\left[1 - \left(\frac{a_7\Delta_{01(st)} + a_8\Delta_{02(st)} + a_9\Delta_{03(st)}}{\Delta_{00(st)}}\right)\right] \tag{72}$$

where

$$\begin{aligned}
\Delta_{00(st)} &= a_1\left[a_2a_3 - a_6^2\right] - a_4\left[a_3a_4 - a_5a_6\right] + a_5\left[a_4a_6 - a_2a_5\right]\ ,\\
\Delta_{01(st)} &= a_7\left[a_2a_3 - a_6^2\right] - a_4\left[a_3a_8 - a_6a_9\right] + a_5\left[a_6a_8 - a_2a_9\right]\ ,\\
\Delta_{02(st)} &= a_1\left[a_3a_9 - a_6a_9\right] - a_7\left[a_3a_4 - a_5a_6\right] + a_5\left[a_4a_9 - a_5a_9\right]\ ,\\
\Delta_{03(st)} &= a_1\left[a_3a_9 - a_6a_8\right] - a_4\left[a_4a_9 - a_5a_8\right] + a_7\left[a_4a_6 - a_2a_5\right]\ .
\end{aligned}$$

and $d_{1(opt)} = \frac{\Delta_{01(st)}}{\Delta_{00(st)}}, d_{2(opt)} = \frac{\Delta_{02(st)}}{\Delta_{00(st)}}, d_{3(opt)} = \frac{\Delta_{03(st)}}{\Delta_{00(st)}}$.

7 Suggested Class of Estimators in Stratified Random Sampling

The proposed class of estimators T_S at 2.1 in simple random sampling without replacement (*SRSWOR*) can be also studied in stratified random sampling. In practice the use of combined class of estimators and separate class of estimators for population mean $\bar{Y}$ can be made. However, Cochran [1] suggested that with only a small sample in each stratum, the combined estimate is to be recommended. Also for ease of computation, we use the combined estimator in our proposed setup. The suggested combined class of estimators in stratified random sampling is defined by

$$T_{St}^{C} = \left\{w_0\bar{y}_{st} + w_1\left(\frac{\bar{x}_{st}}{\bar{X}}\right)^{\alpha_1} + w_2\left(\frac{\bar{z}_{st}}{\bar{Z}}\right)^{\alpha_2}\right\}\exp\left\{\frac{\delta_1\left(\bar{X}-\bar{x}_{st}\right)}{\bar{X}+\bar{x}_{st}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z}-\bar{z}_{st}\right)}{\bar{Z}+\bar{z}_{st}}\right\}, \tag{73}$$

where (w_0, w_1, w_2) are suitably chosen weights and $(\alpha_1, \alpha_2, \delta_1, \delta_2)$ are design parameters.

Expanding (73) by using error terms, we have

$$T_{St}^{C} = \left[w_0\bar{Y}\left(1+e_{0(st)}\right) + w_1\left(1+e_{1(st)}\right)^{\alpha_1} + w_2\left(1+e_{2(st)}\right)^{\alpha_2}\right]\exp\left\{\frac{-\delta_1e_{1(st)}}{2+e_{1(st)}}\right\}\exp\left\{\frac{-\delta_2e_{2(st)}}{2+e_{2(st)}}\right\} \tag{74}$$

Expanding *RHS* of (74), multiplying and omitting the terms of *e*'s having power greater than two, we have

$$T_{St}^{C} = \left[\begin{array}{l}
w_0\bar{Y}\left\{\begin{array}{l} 1+e_{0(st)} - \left(\frac{\delta_1e_{1(st)}+\delta_2e_{2(st)}}{2}\right) - \left(\frac{\delta_1e_{0(st)}e_{1(st)}+\delta_2e_{0(st)}e_{2(st)}}{2}\right) \\ +\frac{\delta_1(\delta_1+2)}{8}e_{1(st)}^2 + \frac{\delta_1\delta_2}{4}e_{1(st)}e_{2(st)} + \frac{\delta_2(\delta_2+2)}{8}e_{2(st)}^2 \end{array}\right\} \\
+w_1\left\{1+\theta_1e_{1(st)} - \frac{\delta_2}{2}e_{2(st)} + \frac{\theta_1(\theta_1-1)}{8}e_{1(st)}^2 - \frac{\delta_2\theta_1}{2}e_{1(st)}e_{2(st)} + \frac{\delta_2(\delta_2+2)}{8}e_{2(st)}^2\right\} \\
+w_2\left\{1+\theta_2e_{2(st)} - \frac{\delta_1}{2}e_{1(st)} + \frac{\delta_1(\delta_1+2)}{8}e_{1(st)}^2 - \frac{\delta_1\theta_2}{2}e_{1(st)}e_{2(st)} + \frac{\theta_2(\theta_2-1)}{8}e_{2(st)}^2\right\}
\end{array}\right]$$

or

$$\left(T_{St}^{C} - \bar{Y}\right) = \left[\begin{array}{l}
w_0\bar{Y}\left\{\begin{array}{l} 1+e_{0(st)} - \left(\frac{\delta_1e_{1(st)}+\delta_2e_{2(st)}}{2}\right) - \left(\frac{\delta_1e_{0(st)}e_{1(st)}+\delta_2e_{0(st)}e_{2(st)}}{2}\right) \\ +\frac{\delta_1(\delta_1+2)}{8}e_{1(st)}^2 + \frac{\delta_1\delta_2}{4}e_{1(st)}e_{2(st)} + \frac{\delta_2(\delta_2+2)}{8}e_{2(st)}^2 \end{array}\right\} \\
+w_1\left\{1+\theta_1e_{1(st)} - \frac{\delta_2}{2}e_{2(st)} + \frac{\theta_1(\theta_1-1)}{8}e_{1(st)}^2 - \frac{\delta_2\theta_1}{2}e_{1(st)}e_{2(st)} + \frac{\delta_2(\delta_2+2)}{8}e_{2(st)}^2\right\} \\
+w_2\left\{1+\theta_2e_{2(st)} - \frac{\delta_1}{2}e_{1(st)} + \frac{\delta_1(\delta_1+2)}{8}e_{1(st)}^2 - \frac{\delta_1\theta_2}{2}e_{1(st)}e_{2(st)} + \frac{\theta_2(\theta_2-1)}{8}e_{2(st)}^2\right\} \\
-\bar{Y}
\end{array}\right] \tag{75}$$

where $\theta_1 = \frac{(2\alpha_1-\delta_1)}{2}$ and $\theta_2 = \frac{(2\alpha_2-\delta_2)}{2}$.

On taking expectation on both sides of (75), we get the bias of the proposed estimator T_{St}^{C} upto first order of approximation as

$$B\left(T_{St}^{C}\right)=\bar{Y}\left[w_0A_{6(st)}+w_1A_{7(st)}+w_2A_{8(st)}-1\right], \tag{76}$$

where

$$A_{6(st)}=\left[1+\frac{\delta_1\left(\delta_1+2\right)}{8}C_{020}+\frac{\delta_1\delta_2}{4}C_{011}+\frac{\delta_2\left(\delta_2+2\right)}{8}C_{002}-\frac{1}{2}\left(\delta_1C_{110}+\delta_2C_{101}\right)\right],$$

$$A_{7(st)}=\frac{1}{R_1\bar{X}}\left[1+\frac{\theta_1\left(\theta_1-1\right)}{2}C_{020}-\frac{\delta_2\theta_1}{2}C_{011}+\frac{\delta_2\left(\delta_2+2\right)}{8}C_{002}\right],$$

$$A_{8(st)}=\frac{1}{R_2\bar{Z}}\left[1+\frac{\delta_1\left(\delta_1+2\right)}{8}C_{020}-\frac{\delta_1\theta_2}{2}C_{011}+\frac{\theta_2\left(\theta_2-1\right)}{2}C_{002}\right].$$

On squaring both sides of (75), omitting terms of e's having power greater than two and then taking expectation on both sides we get the *MSE* of the proposed estimator $T_{St}^{C}$to the first degree of approximation as

$$MSE\left(T_{St}^{C}\right)=\bar{Y}^2\left[\begin{array}{l}1+w_0^2A_{0(st)}+w_1^2A_{1(st)}+w_2^2A_{2(st)}+2w_0w_1A_{3(st)}+2w_0w_2A_{4(st)}+2w_1w_2A_{5(st)}\\-2w_0A_{6(st)}-2w_1A_{7(st)}-2w_2A_{8(st)}\end{array}\right] \tag{77}$$

where

$$\begin{aligned}
A_{0(st)}&=\left[1+C_{200}+\tfrac{\delta_1(\delta_1+1)}{2}C_{020}+\delta_1\delta_2C_{011}+\tfrac{\delta_2(\delta_2+1)}{2}C_{002}-2\left(\delta_1C_{110}+\delta_2C_{101}\right)\right],\\
A_{1(st)}&=\tfrac{1}{R_1^2\bar{X}^2}\left[1+\theta_1\left(2\theta_1-1\right)C_{020}-2\theta_1\delta_2C_{011}+\tfrac{\delta_2(\delta_2+1)}{2}C_{002}\right],\\
A_{2(st)}&=\tfrac{1}{R_2^2\bar{Z}^2}\left[1+\tfrac{\delta_1(\delta_1+1)}{2}C_{020}-2\delta_1\theta_2C_{011}+\theta_2\left(2\theta_2-1\right)C_{002}\right],\\
A_{3(st)}&=\tfrac{1}{R_1\bar{X}}\left[1+\left(\alpha_1-\delta_1\right)C_{110}-\delta_2C_{101}+\tfrac{(\alpha_1-\delta_1)(\alpha_1-\delta_1-1)}{2}C_{020}-\delta_2\left(\alpha_1-\delta_1\right)C_{011}+\tfrac{\delta_2(\delta_2+1)}{2}C_{002}\right],\\
A_{4(st)}&=\tfrac{1}{R_2\bar{Z}}\left[1+\left(\alpha_2-\delta_2\right)C_{101}+\tfrac{\delta_1(\delta_1+1)}{2}C_{020}-\delta_1\left(\alpha_2-\delta_2\right)C_{011}+\tfrac{(\alpha_2-\delta_2)(\alpha_2-\delta_2-1)}{2}C_{002}\right],\\
A_{5(st)}&=\tfrac{1}{R_1R_2\bar{X}\bar{Z}}\left[1+\tfrac{(\alpha_1-\delta_1)(\alpha_1-\delta_1-1)}{2}C_{020}+\tfrac{(\alpha_2-\delta_2)(\alpha_2-\delta_2-1)}{2}C_{002}+\left(\alpha_1-\delta_1\right)\left(\alpha_2-\delta_2\right)C_{011}\right],
\end{aligned}$$

$A_{6(st)}$, $A_{7(st)}$ and $A_{8(st)}$ are same as defined earlier.

To find the optimum values of$(w_0,\,w_1,w_2\,)$, minimizing $MSE\left(T_{St}^{C}\right)$at (77) with respect to$(w_0,\,w_1,w_2\,)$, which yields

$$\left[\begin{array}{ccc}A_{0(st)}&A_{3(st)}&A_{4(st)}\\A_{3(st)}&A_{1(st)}&A_{5(st)}\\A_{4\,(st)}&A_{5(st)}&A_{2(st)}\end{array}\right]\left[\begin{array}{c}w_0\\w_1\\w_2\end{array}\right]=\left[\begin{array}{c}A_{6(st)}\\A_{7(st)}\\A_{8(st)}\end{array}\right] \tag{78}$$

and after simplifying (78), the optimum values are

$$\left.\begin{aligned}w_{0(opt)}&=\tfrac{\Delta_{0(st)}}{\Delta_{(st)}},\\w_{1(opt)}&=\tfrac{\Delta_{1(st)}}{\Delta_{(st)}},\\w_{2(opt)}&=\tfrac{\Delta_{2(st)}}{\Delta_{(st)}}.\end{aligned}\right\} \tag{79}$$

where

$$\begin{aligned}
\Delta_{(st)} &= \begin{vmatrix} A_{0(st)} & A_{3(st)} & A_{4(st)} \\ A_{3(st)} & A_{1(st)} & A_{5(st)} \\ A_{4(st)} & A_{5(st)} & A_{2(st)} \end{vmatrix} \\
&= A_{0(st)}\left(A_{1(st)}A_{2(st)} - A^2_{5(st)}\right) - A_{3(st)}\left(A_{2(st)}A_{3(st)} - A_{4(st)}A_{5(st)}\right) \\
&\quad + A_{4(st)}\left(A_{3(st)}A_{5(st)} - A_{1(st)}A_{4(st)}\right)
\end{aligned}$$

$$\begin{aligned}
\Delta_{0(st)} &= \begin{vmatrix} A_{6(st)} & A_{3(st)} & A_{4(st)} \\ A_{7(st)} & A_{1(st)} & A_{5(st)} \\ A_{8(st)} & A_{5(st)} & A_{2(st)} \end{vmatrix} \\
&= A_{6(st)}\left(A_{1(st)}A_{2(st)} - A^2_{5(st)}\right) - A_{3(st)}\left(A_{2(st)}A_{7(st)} - A_{5(st)}A_{8(st)}\right) \\
&\quad + A_{4(st)}\left(A_{5(st)}A_{7(st)} - A_{1(st)}A_{8(st)}\right),
\end{aligned}$$

$$\begin{aligned}
\Delta_{1(st)} &= \begin{vmatrix} A_{0(st)} & A_{6(st)} & A_{4(st)} \\ A_{3(st)} & A_{7(st)} & A_{5(st)} \\ A_{4(st)} & A_{8(st)} & A_{2(st)} \end{vmatrix} \\
&= A_{0(st)}\left(A_{2(st)}A_{7(st)} - A_{5(st)}A_{8(st)}\right) - A_{6(st)}\left(A_{2(st)}A_{3(st)} - A_{4(st)}A_{5(st)}\right) \\
&\quad + A_{4(st)}\left(A_{3(st)}A_{8(st)} - A_{4(st)}A_{7(st)}\right),
\end{aligned}$$

$$\begin{aligned}
\Delta_{2(st)} &= \begin{vmatrix} A_{0(st)} & A_{3(st)} & A_{6(st)} \\ A_{3(st)} & A_{1(st)} & A_{7(st)} \\ A_{4(st)} & A_{5(st)} & A_{8(st)} \end{vmatrix} \\
&= A_{0(st)}\left(A_{1(st)}A_{8(st)} - A_{5(st)}A_{7(st)}\right) - A_{3(st)}\left(A_{3(st)}A_{8(st)} - A_{4(st)}A_{7(st)}\right) \\
&\quad + A_{6(st)}\left(A_{3(st)}A_{5(st)} - A_{1(st)}A_{4(st)}\right).
\end{aligned}$$

Thus the resulting minimum *MSE* of T^C_{St} is given by

$$MSE_{\min}\left(T^C_{St}\right) = \bar{Y}^2\left[1 - \frac{\left(A_{6(st)}\Delta_{0(st)} + A_{7(st)}\Delta_{1(st)} + A_{8(st)}\Delta_{2(st)}\right)}{\Delta_{(st)}}\right]. \tag{80}$$

8 Special Cases

For $w_2 = 0$, the proposed class of combined estimators reduces to:

$$T^C_{St(1)} = \left[w_0\bar{y}_{st} + w_1\left(\frac{\bar{x}_{st}}{\bar{X}}\right)^{\alpha_1}\right]\exp\left\{\frac{\delta_1\left(\bar{X} - \bar{x}_{st}\right)}{\bar{X} + \bar{x}_{st}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z} - \bar{z}_{st}\right)}{\bar{Z} + \bar{z}_{st}}\right\} \tag{81}$$

Bias and MSE of $T^C_{St(1)}$are respectively given as

$$B\left(T^C_{St(1)}\right) = \bar{Y}\left[w_0 A_{6(st)} + w_1 A_{7(st)} - 1\right], \tag{82}$$

and

$$MSE\left(T^C_{St(1)}\right) = \bar{Y}^2\left[1 + w_0^2 A_{0(st)} + w_1^2 A_{1(st)} + 2w_0 w_1 A_{3(st)} - 2w_0 A_{6(st)} - 2w_1 A_{7(st)}\right]. \tag{83}$$

Minimizing (83) with respect to $(w_0,\, w_1)$ yields,

$$\begin{bmatrix} A_{0(st)} & A_{3(st)} \\ A_{3(st)} & A_{1(st)} \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} A_{6(st)} \\ A_{7(st)} \end{bmatrix} \tag{84}$$

After simplifying (84), we obtain the optimum values of $(w_0,\, w_1)$ as

$$\left.\begin{array}{l} w_0^* = \frac{\Delta^*_{0(st)}}{\Delta^*_{(st)}}, \\ w_1^* = \frac{\Delta^*_{1(st)}}{\Delta^*_{(st)}}. \end{array}\right\} \tag{85}$$

Thus the resulting minimum MSE of $T^C_{St(1)}$ is

$$MSE_{\min}\left(T^C_{St(1)}\right) = \bar{Y}^2\left[1 - \frac{\left(A_{6(st)}\Delta^*_{0(st)} + A_{7(st)}\Delta^*_{1(st)}\right)}{\Delta^*_{(st)}}\right]. \tag{86}$$

where

$$\Delta^*_{(st)} = \begin{vmatrix} A_{0(st)} & A_{3(st)} \\ A_{3(st)} & A_{1(st)} \end{vmatrix} = \left(A_{0(st)}A_{1(st)} - A^2_{3(st)}\right)$$

$$\Delta^*_{0(st)} = \begin{vmatrix} A_{6(st)} & A_{3(st)} \\ A_{7(st)} & A_{1(st)} \end{vmatrix} = \left(A_{1(st)}A_{6(st)} - A_{3(st)}A_{7(st)}\right)$$

$$\Delta^*_{1(st)} = \begin{vmatrix} A_{0(st)} & A_{6(st)} \\ A_{3(st)} & A_{7(st)} \end{vmatrix} = \left(A_{0(st)}A_{7(st)} - A_{3(st)}A_{6(st)}\right).$$

For $w_1 = 0$, the proposed class of estimators reduces to:

$$T^C_{St(2)} = \left[w_0\bar{y}_{st} + w_2\left(\frac{\bar{z}_{st}}{\bar{Z}}\right)^{\alpha_2}\right]\exp\left\{\frac{\delta_1\left(\bar{X} - \bar{x}_{st}\right)}{\bar{X} + \bar{x}_{st}}\right\}\exp\left\{\frac{\delta_2\left(\bar{Z} - \bar{z}_{st}\right)}{\bar{Z} + \bar{z}_{st}}\right\} \tag{87}$$

Bias and MSE of $T^C_{St(2)}$are respectively given as

$$B\left(T^C_{St(2)}\right) = \bar{Y}\left[w_0 A_{6(st)} + w_2 A_{8(st)} - 1\right], \tag{88}$$

and

$$MSE\left(T^C_{St(2)}\right) = \bar{Y}^2\left[1 + w_0^2 A_{0(st)} + w_2^2 A_{2(st)} + 2w_0 w_2 A_{4(st)} - 2w_0 A_{6(st)} - 2w_2 A_{8(st)}\right]. \tag{89}$$

Minimizing (89) with respect to $(w_0,\, w_2)$ yields,

$$\begin{bmatrix} A_{0(st)} & A_{4(st)} \\ A_{4(st)} & A_{2(st)} \end{bmatrix}\begin{bmatrix} w_0 \\ w_2 \end{bmatrix} = \begin{bmatrix} A_{6(st)} \\ A_{8(st)} \end{bmatrix} \tag{90}$$

After simplifying (90), we obtain the optimum values of $(w_0,\, w_2)$ as

$$\left.\begin{aligned} w_0^{**} &= \frac{\Delta_{0(st)}^{**}}{\Delta_{(st)}^{**}}, \\ w_2^{**} &= \frac{\Delta_{2(st)}^{**}}{\Delta_{(st)}^{**}}. \end{aligned}\right\} \tag{91}$$

Thus the resulting minimum MSE of $T_{St(2)}^{C}$ is

$$MSE_{\min}\left(T_{St(2)}^{C}\right) = \bar{Y}^2\left[1 - \frac{\left(A_{6(st)}\Delta_{0(st)}^{**} + A_{8(st)}\Delta_{2(st)}^{**}\right)}{\Delta_{(st)}^{**}}\right]. \tag{92}$$

where

$$\Delta_{(st)}^{**} = \begin{vmatrix} A_{0(st)} & A_{4(st)} \\ A_{4(st)} & A_{2(st)} \end{vmatrix} = \left(A_{0(st)}A_{2(st)} - A_{4(st)}^2\right)\,,$$

$$\Delta_{0(st)}^{**} = \begin{vmatrix} A_{6(st)} & A_{4(st)} \\ A_{8(st)} & A_{2(st)} \end{vmatrix} = \left(A_{6(st)}A_{2(st)} - A_{4(st)}A_{8(st)}\right)\,,$$

$$\Delta_{2(st)}^{**} = \begin{vmatrix} A_{0(st)} & A_{6(st)} \\ A_{4(st)} & A_{8(st)} \end{vmatrix} = \left(A_{0(st)}A_{8(st)} - A_{4(st)}A_{6(st)}\right)\,.$$

9 Efficiency Comparison

From (54), (56), (58) and (60) respectively, we have

$$MSE\left(\hat{\bar{Y}}_{St(0)}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{St(D_1)}^{C}\right) = \frac{V_{110}^2}{V_{020}} \geq 0. \tag{93}$$

$$MSE\left(\hat{\bar{Y}}_{St(R)}^{C}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{St(D_1)}^{C}\right) = \left[R_1V_{020}\left(R_1V_{020} - 2V_{110}\right) + V_{110}^2\right] \geq 0. \tag{94}$$

$$MSE\left(\hat{\bar{Y}}_{St(RExp)}^{C}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{St(D_1)}^{C}\right) = \left[R_1V_{020}\left(\frac{1}{4}R_1V_{020} - V_{110}\right) + V_{110}^2\right] \geq 0. \tag{95}$$

It follows from (93), (94), and (95) that the combined difference estimator for single auxiliary variable is more efficient than$\hat{\bar{Y}}_{St(0)}$,$\hat{\bar{Y}}_{St(R)}^{C}$ and $\hat{\bar{Y}}_{St(R)}^{C}$.

From (62) and (66), we have

$$MSE\left(\hat{\bar{Y}}_{St(RP)}^{C}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{St(D_2)}^{C}\right) = \left[\begin{array}{l} \left(V_{020}V_{002} - V_{011}^2\right)\left(V_{020} + V_{002} - 2V_{110} + 2V_{101} - 2V_{011}\right) \\ +\left(V_{200}^2V_{002} - 2V_{011}V_{101}V_{110} + V_{020}V_{101}^2\right) \end{array}\right]$$

$$\geq 0. \tag{96}$$

Thus from (96), the difference estimator for two auxiliary variable is more efficient than $\hat{\bar{Y}}_{St(RP)}^{C}$.

From (64) and (66), we have

$$\begin{aligned} & MSE\left(\hat{\bar{Y}}_{St(RPExp)}^{C}\right) - MSE_{\min}\left(\hat{\bar{Y}}_{St(D_2)}^{C}\right) \\ = \;& \left(V_{020}V_{002} - V_{011}^2\right)\left(\frac{R_1^2}{4}V_{020} + \frac{R_2^2}{4}V_{002} - R_1V_{110} + R_2V_{101} - \frac{R_1R_2}{2}V_{011}\right) \\ & + \left(V_{200}^2V_{002} - 2V_{011}V_{101}V_{110} + V_{020}V_{101}^2\right) \\ \geq \;& 0. \end{aligned} \tag{97}$$

Thus from (97), the difference estimator for two auxiliary variable is more efficient than$\hat{\bar{Y}}^{C}_{St(RPExp)}$.
From (60) and (66), we have

$$\begin{aligned}
& MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_1)}\right)-MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right) \\
= & \left[V_{020}\left(V_{200}^{2}V_{002}-2V_{011}V_{101}V_{110}+V_{020}V_{101}^{2}\right)-V_{110}^{2}\left(V_{020}V_{002}-V_{011}^{2}\right)\right] \\
\geq & \ 0
\end{aligned} \tag{98}$$

From (93), (94), (95) and (98), we have the following inequalities:

$$MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_1)}\right)\leq MSE\left(\hat{\bar{Y}}_{St(0)}\right) \tag{99}$$

$$MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_1)}\right)\leq MSE\left(\hat{\bar{Y}}^{C}_{St(R)}\right) \tag{100}$$

$$MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_1)}\right)\leq MSE\left(\hat{\bar{Y}}^{C}_{St(RExp)}\right) \tag{101}$$

Hence $\hat{\bar{Y}}^{C}_{St(D_2)}$is more efficient than the sample mean, ratio-type, ratio-type exponential and difference estimator for single auxiliary variable.

From (66) and (80), we have that
$MSE_{\min}\left(T^{C}_{St}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(D_2)}\right)$, if

$$\begin{aligned}
& \left[\frac{\left(V_{020}V_{002}-V_{011}^{2}\right)}{\Delta_{(st)}}\left\{V_{200}\Delta_{(st)}+\bar{Y}^{2}\left(\begin{array}{l} A_{6(st)}\Delta_{0(st)}+A_{7(st)}\Delta_{1(st)} \\ +A_{8(st)}\Delta_{2(st)}\end{array}\right)\right\}\right] \\
\geq & \left[\begin{array}{l}\bar{Y}^{2}\left(V_{020}V_{002}-V_{011}^{2}\right) \\ +\left(V_{200}^{2}V_{002}-2V_{011}V_{101}V_{110}+V_{020}V_{101}^{2}\right)\end{array}\right]
\end{aligned} \tag{102}$$

Thus the proposed estimator $T^{C}_{St}$is more efficient than the difference estimator for two auxiliary variables as long as the condition (102) holds.

Further from (96) to (101), we can see that our proposed class of estimators $T^{C}_{St}$is also more efficient than the stratified(combined) sample mean, ratio-type, ratio-type exponential, difference estimator for single auxiliary variable, ratio cum product-type estimator and ratio cum product-type exponential estimators.

From (68) and (80), we have
$MSE_{\min}\left(T^{C}_{St}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(MSK)}\right)$, if

$$\left[\frac{\left(A_{6(st)}\Delta_{0(st)}+A_{7(st)}\Delta_{1(st)}+A_{8(st)}\Delta_{2(st)}\right)}{\Delta_{(st)}}\right]\geq\left[\bar{Y}^{2}\left(\frac{C_{011}^{2}}{4C_{020}}+\frac{A_{st}^{2}}{B_{st}}\right)\right] \tag{103}$$

Thus the proposed estimator $T^{C}_{St}$is more efficient than Muneer et al (2016) estimator as long as the condition (49) holds.

From (72) and (80), we have
$MSE_{\min}\left(T^{C}_{St}\right)\leq MSE_{\min}\left(\hat{\bar{Y}}^{C}_{St(SG)}\right)$, if

$$\left[\frac{\left(A_{6(st)}\Delta_{0(st)}+A_{7(st)}\Delta_{1(st)}+A_{8(st)}\Delta_{2(st)}\right)}{\Delta_{(st)}}\right] \geq \left[\frac{\left(a_{7}\Delta_{01(st)}+a_{8}\Delta_{02(st)}+a_{9}\Delta_{03(st)}\right)}{\Delta_{00(st)}}\right] \tag{104}$$

Thus the proposed estimator $T_{St}^{C}$is more efficient than Shabbir and Gupta [12] estimator as long as the condition (104) is true.

Now from (86) and (80), we have $MSE_{\min}\left(T_{St}^{C}\right) \leq MSE_{\min}\left(T_{St(1)}^{C}\right)$, if

$$\left[\frac{A_{6(st)}\Delta_{0(st)}+A_{7(st)}\Delta_{1(st)}+A_{8(st)}\Delta_{2(st)}}{\Delta_{(st)}}\right] \geq \left[\frac{A_{6(st)}\Delta_{0(st)}^{*}+A_{7(st)}\Delta_{1(st)}^{*}}{\Delta_{(st)}^{*}}\right] \tag{105}$$

which is always true.

From (92) and (80), we have
$MSE_{\min}\left(T_{St}^{C}\right) \leq MSE_{\min}\left(T_{St(2)}^{C}\right)$, if

$$\left[\frac{A_{6(st)}\Delta_{0(st)}+A_{7(st)}\Delta_{1(st)}+A_{8(st)}\Delta_{2(st)}}{\Delta_{(st)}}\right] \geq \left[\frac{A_{6(st)}\Delta_{0(st)}^{**}+A_{8(st)}\Delta_{2(st)}^{**}}{\Delta_{(st)}^{**}}\right] \tag{106}$$

which always holds.

Thus the proposed class of estimators $T_{St}^{C}$is better than the estimators $T_{St(\mathbf{??})}^{C}$and $T_{St(\mathbf{??})}^{C}$at their optimum conditions.

10 Numerical Illustration

For numerical illustration we use the following data sets
Data set 1 [Source: Tailor and Chouhan [10]]
y: Productivity (MT/hectare).
x: Production (000 tons).
z: Area (000 hectare).
$N = 20,\ n = 8,\ N_1 = 10,\ N_2 = 10,\ n_1 = 4,\ n_2 = 4$
$\bar{Y}_1 = 1.70,\ \bar{Y}_2 = 3.67, \bar{X}_1 = 10.40,\ \bar{X}_2 = 289.14, \bar{Z}_1 = 6.23,\ \bar{Z}_2 = 80.67,$
$S_{y1} = 0.50,\ S_{y2} = 1.41,\ S_{x1} = 3.53,\ S_{x2} = 111.61,\ S_{z1} = 1.18, S_{z2} = 10.81,$
$S_{yx1} = 1.60,\ S_{yx2} = 144.88,\ S_{xz1} = 1.38,\ S_{xz2} = -92.02,\ S_{yz1} = -0.05,\ S_{yz2} = -7.04.$
Data set 2 [Source: National Horticulture Board [9]]
y: Productivity (MT/hectare).
x: Production (000 tons).
z: Area (000 hectare).
$N = 10,\ n = 7,\ N_1 = 5,\ N_2 = 5,\ n_1 = 3,\ n_2 = 4,$
$\bar{Y}_1 = 1.70,\ \bar{Y}_2 = 3.67,\ \bar{X}_1 = 10.41,\ \bar{X}_2 = 309.14, \bar{Z}_1 = 6.20,\ \bar{Z}_2 = 80.67,$
$S_{y1}^2 = 0.2916,\ S_{y2}^2 = 1.9881,\ S_{x1}^2 = 1.4116,\ S_{x2}^2 = 3486.6916,$
$S_{z1}^2 = 1.4116, S_{z2}^2 = 116.8561,\ S_{yx1} = 1.6000,\ S_{yx2} = 83.47,$
$S_{xz1} = 1.7500,\ S_{xz2} = 64.9700,\ S_{yz1} = -0.2000,\ S_{yz2} = 5.5800.$

Table 6 gives the *PRE*'s of different estimators considered in this paper with respect to $\bar{y}_{st}$.

Table 7 gives the *PRE*'s of the proposed class of estimator with respect to $\bar{y}_{st}$ at different values of $(\alpha_1, \alpha_2, \delta_1, \delta_2)$.

Estimator	Data 1	Data 2
$\hat{\bar{Y}}^{C}_{St(R)}$	225.25	344.59
$\hat{\bar{Y}}^{C}_{St(RExp)}$	364.53	179.86
$\hat{\bar{Y}}^{C}_{St(D_1)}$	432.60	493.58
$\hat{\bar{Y}}^{C}_{St(RP)}$	290.76	158.69
$\hat{\bar{Y}}^{C}_{St(RPExp)}$	660.33	143.39
$\hat{\bar{Y}}^{C}_{St(D_2)}$	1072.78	589.44
$\hat{\bar{Y}}^{C}_{St(MSK)}$ (at $\alpha = 1$)	223.82	587.95
$\hat{\bar{Y}}^{C}_{St(MSK)}$ (at $\alpha = 0$)	223.91	588.01
$\hat{\bar{Y}}^{C}_{St(SG)}$	1048.10	587.94

Table 6: *PRE*'s of $\hat{\bar{Y}}^{C}_{St(R)}, \hat{\bar{Y}}^{C}_{St(RExp)}, \hat{\bar{Y}}^{C}_{St(D_1)}, \hat{\bar{Y}}^{C}_{St(RP)}, \hat{\bar{Y}}^{C}_{St(RPExp)}, \hat{\bar{Y}}^{C}_{St(D_2)}, \hat{\bar{Y}}^{C}_{St(MSK)}$and $\hat{\bar{Y}}^{C}_{St(SG)}$with respect to$\bar{y}$.

For Data Set 1						For Data Set 2				
α_1	α_2	δ_1	δ_2	*PRE*		α_1	α_2	δ_1	δ_2	*PRE*
1	1	0	0	1290.19		0.25	0.25	1	1	1407.53
1	1	0.25	0.25	2293.67		1	1	0	0	2111.01
0.75	0.75	0	0	2835.84		0.75	0.75	0	0	2733.85
1	0	1	1	2901.33		1	1	0.25	0.25	3759.05
0.25	0.25	1	1	4315.93		0.5	0.5	1	1	4389.42
1	1	0	1	5160.44		0.5	0.5	0	0	4392.65
1	1	1	0	5160.93		1	0	1	1	5009.18
1	1	0.5	0.5	5161.02		0	0.5	0.5	0.5	7872.07
0.5	0.5	0	0	7896.38		1	1	0	1	8444.21
0.5	0.5	1	1	7899.81		1	1	0.5	0.5	8518.06
0.5	0	0.5	0.5	11603.2		1	1	1	0	8548.12
1	1	0.75	0.75	20652.9		0.25	0.25	0.5	0.5	12667.3
0	0.5	0.5	0.5	24742.4		0.25	0.25	0	0	12673.3
0.75	0.75	1	1	25548.2		0.5	0.5	0	0.5	17567.2
0.5	0.5	0.25	0.25	31583.9		0.5	0.5	0.25	0.25	17608.1
0.5	0.5	0.5	0	31584.1		0.5	0.5	0.5	0	17620.5
0.5	0.5	0	0.5	31585.9		0.5	0	0.5	0.5	19792.4
0.25	0.25	0	0	38835		0.75	0.75	1	1	24945.4
0.25	0.25	0.5	0.5	38844.8		1	1	0.75	0.75	35688.9
0.25	0	0.25	0.25	46409.7		0.25	0.25	0	0.25	50687.3
0.25	0.25	0.25	0	155330		0.25	0.25	0.25	0	50728
0.25	0.25	0	0.25	155339		0.25	0	0.25	0.25	79119.9

Table 7: *PRE* of the proposed estimator T^{C}_{St} with respect to $\bar{y}$

From the Tables 6 and 7 we found that our proposed class of estimators gives the highest *PRE* for all the data sets at different values of scalars $(\alpha_1, \alpha_2, \delta_1, \delta_2)$(155339.2, 79119.91 for data sets 1 and 2 respectively) which are higher than the estimators $\hat{\hat{Y}}^C_{St(R)}$,$\hat{\hat{Y}}^C_{St(RExp)}$,$\hat{\hat{Y}}^C_{St(D_1)}$,$\hat{\hat{Y}}^C_{St(RP)}$,$\hat{\hat{Y}}^C_{St(RPExp)}$,$\hat{\hat{Y}}^C_{St(D_2)}$, $\hat{\hat{Y}}^C_{St(MSK)}$ and $\hat{\hat{Y}}^C_{St(SG)}$.

11 Conclusion

In this paper we have suggested a class of estimators for population mean based on two auxiliary variables in simple random sampling as well as in stratified random sampling. We studied their properties up to the first order of approximation and also find the optimum conditions in which our proposed class of estimators is better than other existing estimators in both the sampling schemes. For numerical illustration, we consider data sets (4 in simple random sampling and 2 in stratified random sampling) and found that our developed class of estimators has the highest *PRE*s as compared to other estimators. Hence we recommend our recommend their use in practice.

References

[1] W. G. Cochran. *Sampling Techniques*. Jhon Wiley and Sons, New York, 1977.

[2] *Crops area production.* Ministry of Food and Agriculture. Islamabad, Pakistan.

[3] L. K. Grover and P. K. Kaur. “An improved estimator of the finite population mean in simple random sampling.” In: *Model Assisted Statistics and Applications* 6(1) (2011).

[4] S. Gupta and J. Shabbir. “On improvement in estimating the population mean in simple random sampling.” In: *Journal of Applied Statistics* 35(5) (2008), pp. 559–566.

[5] A. T. Hair. *Analysing multivariate data.* Duxbury Press: Boston Massacusetts. USA., 2002.

[6] J. Jhonston. *Econometrics methods*. McGraw-Hill, New York., 1963.

[7] N. Koyuncu. “Efficient estimators of population mean using auxiliary attributes.” In: *Applied Mathematics and Computation* 218 (2004), pp. 10900–10905.

[8] J. Muneer S.; Shabbir and A. Khalil. “Estimation of finite population mean in simple random sampling and stratified random sampling using two auxiliary variables.” In: *Communications in Statistics-Theory and Methods* 465 (2016), pp. 2181–2192.

[9] *National Horticulture Board*. National Horticulture Board. URL: http://nhb.gov.in/statistics/area-productionstatistics.html.

[10] Tailor R. and Chouhan S. “Ratio-cum-product type exponential estimator of finite population mean in stratified random sampling.” In: *Communication in Statistics-Theory and Methods* 43 (2014), pp. 343–354.

[11] Upadhyaya L. N. ; Singh H. P.; Chatterjee S.; and Yadav R. “Improved ratio and product exponential type estimators.” In: *Journal of statistical Theory and Practice* 5(2) (2011), pp. 285–302.

[12] J. Shabbir and S. Gupta. “On estimation of finite population mean in simple and stratified random sampling using two auxiliary variables.” In: *Communications in Statistics-Theory and Methods* 46(20 (2016).

[13] S. Singh. *Advanced sampling theory with applications-How Michael ‘selected’ Amy*. Vol-II, Kluwer Academic Publishers, 2003.

[14] V. K. Singh and R. Singh. “Performance of an estimator for estimating population mean using simple and stratified random sampling.” In: *SOP Transactions on Statistics and Analysis* 1(1) (2014), pp. 1–8.

The Panel of the Reviewers in the current issue

Prof. Jyoti Divecha
Dept. of Statistics
Sardar Patel University, Vallabh
Vidyanagar
Gujarat INDIA

Dr. Indranil Ghosh
Dept. of Mathematics and Statistics
University of North Carolina, Wilmington
North Carolina, USA

Dr. Cinni Varghese
Principal Scientists
IASRI New Delhi
New Delhi-110012, INDIA

Prof. D. K. Ghosh
Dept. of Statistics
Saurashtra University, Rajkot
Gujarat, INDIA

Dr. Manoj Kumar
Associate Professor
Department of Statistics
University of Delhi
Delhi, INDIA

Prof. Bikas Kumar Sinha
Retired professor
ISI Kolkatta
WB INDIA

Prof. P. C. Gupta
Dept. of Statistics
University of Rajasthan
Jaipur, Rajasthan, India

Prof. Devang Garg
Dept. of Statistics
Punjabi University, Patiala
Punjab, INDIA

Prof. M. N. Patel
Retired Professor
Department of Statistics
Gujarat University, Ahmedabad
Gujarat, INDIA

Dr. K. Muralidharan
Dept. of Statistics
Faculty of Science
The Maharaja Sayajirao University of Baroda
Vadodara, Gujarat, India

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

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, 2024 Sd/- Dilip Kumar Ghosh

Signature of Publisher