A CLASS OF DUAL FRAME SURVEY SAMPLING ESTIMATORS IN THE PRESENCE OF A COVARIATE: HOW AMY PREDICTS HER PRESIDENT

Sarjinder Singh and Stephen A. Sedory
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363, USA
E-mail: kuss2008@tamuk.edu

David Molina
Department of Statistics and Operational Research
University of Granada, Spain.
E-mail: dmmunoz@correo.ugr.es

ABSTRACT

In this paper, a fictitious story, "*How Amy Predicts Her President*", is introduced to motivate the research considered. In the course of the story we propose a new class of estimators in dual frame survey sampling that makes use of a power transformation. The estimator proposed by Hartley (1962, 1974) is shown to be a special case of the proposed class of estimators. The mean squared error of the proposed estimator is derived and compared to that of the Hartley estimator. A suggestion is given for improving the Fuller and Burmeister (1972) estimator along similar lines. Lastly, the work is extend to the case of multi-covariates. Note that we make no use of any known parameter of auxiliary information as in the ratio estimator due to Cochran (1940). In this regard the proposed class of estimators is different from the existing estimators in the literature of dual frame survey sampling. We show theoretically that the proposed class of estimators is always more efficient than the pioneer Hartley (1962, 1974) estimator. The results are also justified through extensive simulated numerical situations.

Key words: Dual frame survey, estimation of population total, power transformation.

1. INTRODUCTION

Let us motivate this contribution by a story. Every day, whenever *Amy* switches on her television, she finds a stream of very interesting news about politics in the United States. The news reader always seems to be talking about the latest prediction of who will be chosen in the coming election to be the next president. *Amy* finds that the main purpose of the news seems to be to discover whether "Democrats" or "Republicans" will win in the coming election.

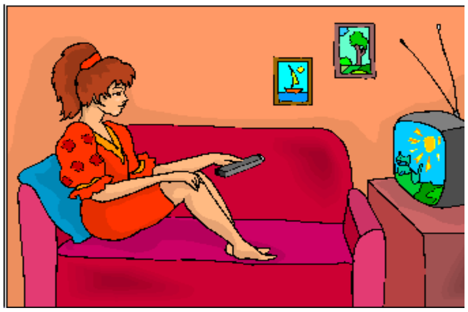
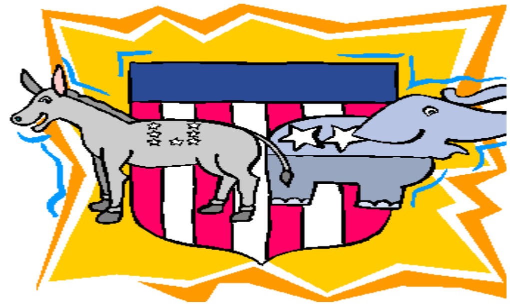

Fig. 1.1. Amy watching television   Political game of donkeys and elephants

Amy is a survey statistician. One day, *Amy* found a very interesting and challenging problem on the television in the program: "*Future of Politics*". In the episode, a government agency hires two private companies, with logos: "*Modern Analytics*" and "*Stat-Hawkers*". Both companies are assigned the job of taking samples in such a clever way that the general public would be made completely aware of the future president of the United States. The show makes the point that such predictions of the presidential winner are also helpful to the candidates who are competing in the election in preparing them for the almost certain outcome of their being the loser or winner. Otherwise a sudden shock of losing or gaining president position may cause heart problems to the candidates as well as to many who are deeply associated with the election. Thus predictions of future president of the United States (or of any other county) help people to stay calm during or after the time of final election.

The company "*Modern Analytics*" decides to use a frame A (say), which consists of all voters who have cell phones. The company "*Stat-Hawkers*" decides to use a frame B (say), which consists of voters who have land-line phones. *Amy* found that "*Modern Analytics*" selected a sample of n_A voters from the frame A and "*Stat-Hawkers*" selected a sample of n_B voters from the frame B. Both companies, "*Modern Analytics*" and "*Stat-Hawkers*", announce their results on the television. *Amy* became suspicious of the findings of both companies. *Amy* reaches both of the companies and is granted permission to look at the raw data collected by both companies. *Amy* noticed that one respondent, *Mr. Mobile* has only cell phone, another, *Miss Twinkle* has both cell phone and land-line phone, and still another, *Mr. Static* has only a land-line phone. *Amy* looks at the entire raw data sets collected by both of the companies "*Modern Analytics*" and "*Stat-Hawkers*". *Amy* found that out of the n_A voters selected by "*Modern Analytics*", n_a voters have only cell phones, and n_{ab} voters have both cell phones and land-line phones. Also *Amy* found that out of the n_B voters selected by "*Stat-Hawkers*", n_b voters have only land-line phones, and n_{ab} voters have both land-line and cell phones. Thus *Amy* wonders how this double counting from both frames in the sample can be utilized to draw better inferences about which candidate might be the future presidents from the target population consisting of the union of both frames. *Amy* feels that inferences based on samples collected only from frame *A*, or only from frame *B*, may provide misleading results. In addition, *Amy* finds that there is additional information about the voters selected in the sample from the presence of a co-variate. We take these observations by *Amy* as motivation for our movement in the following direction.

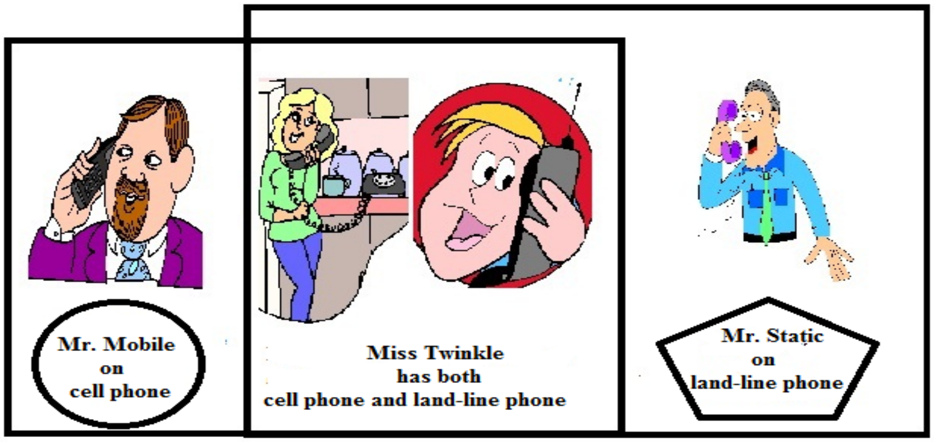

Fig. 1.2. Amy's motivation for dual frame survey sampling.

In this paper, we consider a new situation when a co-variate X is available for the units included in the sample taken from a dual frame survey, in addition to the main variate, Y, of interest. Let (Y_a, X_a), (Y_b, X_b), (Y_{ab}, X_{ab}), (Y_A, X_A) and $(Y_{B,} X_B)$ be the unknown population totals of the main variate Y and co-variate X, where the subscript *a* indicates the subpopulation of units only in frame A, *b* indicates units only in frame B, and *ab* indicates units found in both frames. Note that $Y_A = Y_a + Y_{ab}$, $X_A = X_a + X_{ab}$, $Y_B = Y_b + Y_{ab}$ and $X_B = X_b + X_{ab}$. A pictorial representation of such a dual frame survey structure is shown below:

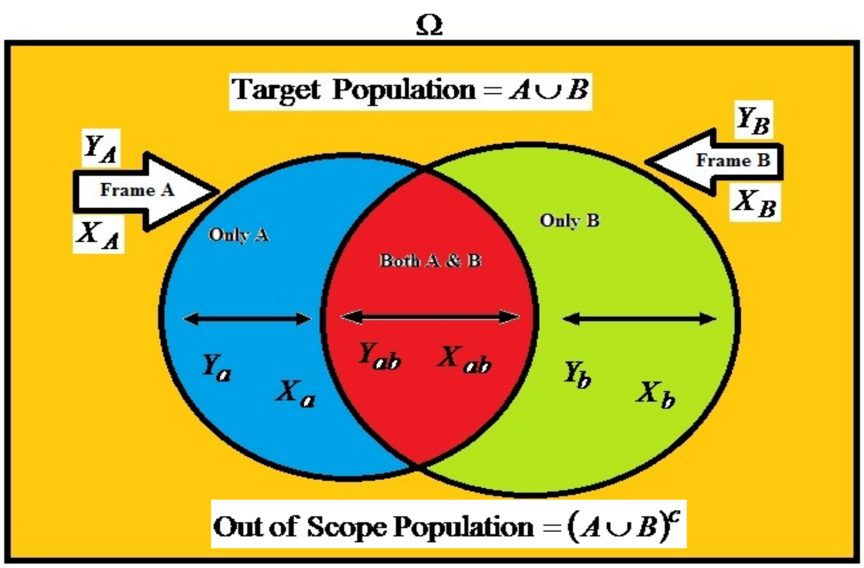

Fig. 1.3. A dual frame survey structure.

In the next section, we define a few notations which remain useful in this and future research in this area.

2. NOTATIONS

Assume s_A to be a sample of size n_A taken from the frame A and s_B to be an independent sample of size n_B taken from the frame B. Let $\pi_i^{(A)}$ be the probability of including ith unit in the sample s_A from the frame A and $\pi_i^{(B)}$ be the probability of including ith unit in the sample s_B from the frame B.

Following Horvitz and Thompson (1952), we have:

$\hat{Y}_A = \sum_{i \in s_A} \frac{y_i}{\pi_i^{(A)}}$ is an unbiased estimator of the population total Y_A,

and

$\hat{Y}_B = \sum_{i \in s_B} \frac{y_i}{\pi_i^{(B)}}$ is an unbiased estimator of the population total Y_B.

Let us define three indicator variables:

$$I_i^{(a)} = \begin{cases} 1, & \text{if } i \in a \\ 0, & \text{otherwise} \end{cases}, \quad I_i^{(b)} = \begin{cases} 1, & \text{if } i \in b \\ 0, & \text{otherwise} \end{cases} \text{ and } I_i^{(ab)} = \begin{cases} 1, & \text{if } i \in (ab) \\ 0, & \text{otherwise} \end{cases}.$$

By following Hartley (1962, 1974), we define:

$\hat{Y}_a = \sum_{i \in s_A} \frac{y_i}{\pi_i^{(A)}} I_i^{(a)}$, unbiased estimator of the domain population total Y_a

$\hat{Y}_b = \sum_{i \in s_B} \frac{y_i}{\pi_i^{(B)}} I_i^{(b)}$, unbiased estimator of the domain population total Y_b

$\hat{Y}_{ab} = \sum_{i \in s_A} \frac{y_i}{\pi_i^{(A)}} I_i^{(ab)}$, unbiased estimator of the domain population total Y_{ab} based on the sample from frame A, and

$\hat{Y}_{ba} = \sum_{i \in s_B} \frac{y_i}{\pi_i^{(B)}} I_i^{(ab)}$ as also an unbiased estimator of the domain population total Y_{ab} based on the sample from frame B.

In the same way, the unbiased estimators of X_A, X_B, X_a, X_b and X_{ab} are defined as $\hat{X}_A$, $\hat{X}_B$, $\hat{X}_a$, $\hat{X}_b$ and $\hat{X}_{ab}$ (or $\hat{X}_{ba}$) respectively.

Hartley (1962, 1974) proposed an estimator of the population total Y in a dual frame survey sampling as:

$$\hat{Y}_{\text{Hartley}} = \hat{Y}_a + \hat{Y}_b + \theta_H \hat{Y}_{ab} + (1-\theta_H)\hat{Y}_{ba}$$

The minimum variance of the estimator $\hat{Y}_{Hartley}$ with the optimum value of θ_H is given by:

$$Min.V(\hat{Y}_{\text{Hartley}}) = V(\hat{Y}_a) + V(\hat{Y}_b) + V(\hat{Y}_{ba}) + 2Cov(\hat{Y}_b, \hat{Y}_{ba}) - \frac{\{Cov(\hat{Y}_b, \hat{Y}_{ba}) + V(\hat{Y}_{ba}) - Cov(\hat{Y}_a, \hat{Y}_{ab})\}^2}{V(\hat{Y}_{ab}) + V(\hat{Y}_{ba})}.$$

Fuller and Burmeister (1972) suggested a modification in the Hartley's estimator by using an additional information about N_{ab} as:

$$\hat{Y}_{\text{FB}} = \hat{Y}_a + \hat{Y}_b + \theta_1 \hat{Y}_{ab} + (1-\theta_1)\hat{Y}_{ba} + \theta_2(\hat{N}_{ab} - \hat{N}_{ba})$$

Lohr and Rao (2000) have shown that the Fuller-Burmeister $\hat{Y}_{FB}$ estimator has the smallest asymptotic variance among the estimators considered by them. The estimator due to Fuller and Burmeister (1972) is internally inconsistent, see Lohr (2011) for detail about internal consistency. Later Skinner and Rao (1996) attempted to make it consistent by using pseudo-maximum likelihood (PML) estimator based on some simulation justifications, but no strong theoretical evidence is provided. Rao and Wu (2010) proposed a pseudo-empirical likelihood (PEL) estimator for a dual frame survey sampling estimator in the presence of known auxiliary information (Lohr, 2011, page 201). Again their constraints result in a different set of weights for each response variable leading to their proposed PEL being internally inconsistent. Rao and Wu (2010) also tried an alternative estimator in which the weight adjustment does not depend on the study variable, and in the absence of auxiliary variable their this approach leads back to the pioneer Hartley's estimator. The moral of the story of this review is that there is no clearly well defined estimator in the literature which, based on theoretical evidence, can be claimed to be more efficient than the pioneer Hartley's estimator in the absence of auxiliary information. For a review of such estimators, please refer to Lohr (2011).

In the next section, we propose a new class of estimator suitable for a dual frame survey in the presence of a covariate (note that no auxiliary information parameter is available). Then we show theoretically that it remains more efficient than the Hartley's estimator.

3. PROPOSED CLASS OF ESTIMATORS

We propose a new class of estimators of the population total Y in dual frame survey sampling as:

$$\hat{Y}_{\text{new}} = \left(\hat{Y}_a + \hat{Y}_b\right)\left[\frac{(\hat{X}_a + \hat{X}_{ba})(\hat{X}_b + \hat{X}_{ab})}{\hat{X}_A \hat{X}_B}\right]^{\alpha} + \gamma \hat{Y}_{ab} + (1-\gamma)\hat{Y}_{ba} \tag{3.1}$$

where α and γ are real known constants. If $\alpha = 0$ then $\hat{Y}_{\text{new}} = \hat{Y}_{\text{Hartley}}$, that is, the proposed class of estimators reduces to the Hartley's estimator. It will be worth mentioning that such a class of ratio type estimators in the presence of an auxiliary variable was initiated by Srivastava (1967), and today a huge body of literature making use of such power transformation estimators is available, many of which are quoted in Singh (2003). The present contribution has a similarly broad scope of extensions in the presence of a co-variate, which is a departure from the Srivastava (1967) class of estimators. There is also a huge body of literature in the field of survey sampling where optimum values of these types of constants α and γ are estimated from the given sample, and the resultant estimators are shown to maintain the same asymptotic mean squared errors, see Singh (2003) and Singh *et al.* (1995).

Using notations from the Appendix, and using binomial expansion the proposed class of estimators $\hat{Y}_{\text{new}}$, in terms of $\in_a, \in_b$, $\in_{ab}$, $\in_{ba}$, δ_a, δ_b, δ_{ab}, δ_{ba}, δ_A and δ_B, to the first order of approximation, can be expressed as:

$$\begin{aligned}\hat{Y}_{\text{new}} &= Y + Y_a \in_a + Y_b \in_b + \alpha(Y_a + Y_b)\psi + Y_{ab} \in_{ba} + \gamma Y_{ab}(\in_{ab} - \in_{ba}) \\ &\quad + \alpha Y_a \in_a \psi + \alpha Y_b \psi + \frac{\alpha(\alpha-1)}{2!}\psi^2 + .O(n^{-2})\end{aligned} \tag{3.2}$$

where

$$\psi = \frac{\begin{bmatrix} X_a X_b(\delta_a + \delta_b) + X_{ab} X_b(\delta_{ba} + \delta_b) + X_a X_{ab}(\delta_a + \delta_{ab}) \\ + X_{ab}^2(\delta_{ab} + \delta_{ba}) - X_A X_B(\delta_A + \delta_B) \end{bmatrix}}{X_A X_B} \tag{3.3}$$

Taking expected value on both sides of (3.2) and using results from the Appendix, we have the following theorem:

Theorem 3.1. The bias in the proposed class of estimators $\hat{Y}_{new}$ is given by:

$$\begin{aligned}B(\hat{Y}_{new}) &= \alpha\left(\frac{1}{X_B} - \frac{1}{X_A}\right)\left[Cov\left(\hat{Y}_a, \hat{X}_{ab}\right) - Cov\left(\hat{Y}_b, \hat{X}_{ba}\right)\right] \\ &+ \frac{\alpha(\alpha-1)(Y_a + Y_b)}{2!\, X_A^2 X_B^2}\begin{bmatrix}\left(X_a^2 + X_b^2\right)\left\{V(\hat{X}_{ab}) + V(\hat{X}_{ba})\right\} \\ + 2X_{ab}(X_b - X_a)Cov(\hat{X}_a, \hat{X}_{ab}) - 2X_b X_A V(\hat{X}_{ba})\end{bmatrix}\end{aligned} \tag{3.4}$$

In practice for most of the sampling designs, as the sample size increases:
$Cov(\hat{Y}_a, \hat{X}_{ab}) \to 0$, $Cov(\hat{Y}_b, \hat{X}_{ba}) \to 0$, $Cov(\hat{X}_a, \hat{X}_{ab}) \to 0$, $V(\hat{X}_{ba}) \to 0$, and $V(\hat{X}_{ab}) \to 0$

Thus from (3.4), it is clear that the bias in the proposed estimator is of first order of approximation and $B(\hat{Y}_{new}) \to 0$ as the sample sizes increases.

Now we have following Lemmas:

Lemma 3.1. The expected value of ψ^2 is given by:

$$E(\psi^2) = \frac{(X_b^2 + X_a^2)(V(\hat{X}_{ab}) + V(\hat{X}_{ba})) + 2X_{ab}(X_b - X_a)Cov(\hat{X}_a, \hat{X}_{ab}) - 2X_b X_A V(\hat{X}_{ba})}{X_A^2 X_B^2} \tag{3.5}$$

Lemma 3.2. The expected value of $\in_a \psi$ is given by:

$$E(\in_a \psi) = \frac{1}{Y_a}\left(\frac{1}{X_B} - \frac{1}{X_A}\right)Cov(\hat{X}_{ab}, \hat{Y}_a) \tag{3.6}$$

Lemma 3.3. The expected value of $\in_b \psi$ is given by:

$$E(\in_b \psi) = \frac{1}{Y_b}\left(\frac{1}{X_A} - \frac{1}{X_B}\right)Cov(\hat{X}_{ba}, \hat{Y}_b) \tag{3.7}$$

Lemma 3.4. The expected value of $\in_{ab} \psi$ is given by:

$$E(\in_{ab} \psi) = \frac{1}{Y_{ab}}\left(\frac{1}{X_B} - \frac{1}{X_A}\right)Cov(\hat{X}_{ab}, \hat{Y}_{ab}) \tag{3.8}$$

Lemma 3.5. The expected value of $\in_{ba} \psi$ is given by:

$$E(\in_{ba} \psi) = \frac{1}{Y_{ab}}\left(\frac{1}{X_A} - \frac{1}{X_B}\right)Cov(\hat{X}_{ba}, \hat{Y}_{ba}) \tag{3.9}$$

Now using (3.2), to the first order of approximation, the mean squared error of the proposed class of estimators $\hat{Y}_{\text{new}}$ is given by:

$$\text{MSE}(\hat{Y}_{\text{new}}) = E\left[\hat{Y}_{\text{new}} - Y\right]^2$$

$$\approx E\left[Y_a \in_a + Y_b \in_b + \alpha(Y_a + Y_b)\psi + Y_{ab} \in_{ba} + \gamma Y_{ab}(\in_{ab} - \in_{ba})\right]^2$$

$$= V(\hat{Y}_a) + V(\hat{Y}_b) + V(\hat{Y}_{ba}) + 2Cov(\hat{Y}_b, \hat{Y}_{ba}) \quad + \gamma^2 \left\{V(\hat{Y}_{ab}) + V(\hat{Y}_{ba})\right\}$$

$$
+\alpha^2\left(\frac{Y_a+Y_b}{X_AX_B}\right)^2\left\{\begin{array}{l}(X_a^2+X_b^2)(V(\hat{X}_{ab})+V(\hat{X}_{ba}))\\ +2X_{ab}(X_b-X_a)Cov(\hat{X}_a,\hat{X}_{ab})-2X_bX_AV(\hat{X}_{ba})\end{array}\right\}
$$
$$
+2\gamma\left\{Cov(\hat{Y}_a,\hat{Y}_{ab})-Cov(\hat{Y}_b,\hat{Y}_{ba})-V(\hat{Y}_{ba})\right\}
$$
$$
+2\alpha(Y_a+Y_b)\left(\frac{1}{X_B}-\frac{1}{X_A}\right)\left\{Cov(\hat{Y}_a,\hat{X}_{ab})-Cov(\hat{X}_b,\hat{X}_{ba})-Cov(\hat{Y}_{ba},\hat{X}_{ba})\right\}
$$
$$
+2\alpha\gamma(Y_a+Y_b)\left(\frac{1}{X_B}-\frac{1}{X_A}\right)\left\{Cov(\hat{Y}_{ab},\hat{X}_{ab})+Cov(\hat{Y}_{ba},\hat{X}_{ba})\right\} \tag{3.10}
$$

To reduce the length of the expressions, let us consider:

$$
A_1=V(\hat{Y}_{ab})+V(\hat{Y}_{ba}) \tag{3.11}
$$

$$
A_2=\left(\frac{Y_a+Y_b}{X_AX_B}\right)^2\left\{\begin{array}{l}(X_a^2+X_b^2)(V(\hat{X}_{ab})+V(\hat{X}_{ba}))\\ +2X_{ab}(X_b-X_a)Cov(\hat{X}_a,\hat{X}_{ab})-2X_bX_AV(\hat{X}_{ba})\end{array}\right\} \tag{3.12}
$$
$$
A_3=Cov(\hat{Y}_b,\hat{Y}_{ba})+V(\hat{Y}_{ba})-Cov(\hat{Y}_a,\hat{Y}_{ab}) \tag{3.13}
$$
$$
A_4=(Y_a+Y_b)\left(\frac{1}{X_A}-\frac{1}{X_B}\right)\left\{Cov(\hat{Y}_a,\hat{X}_{ab})-Cov(\hat{X}_b,\hat{X}_{ba})-Cov(\hat{Y}_{ba},\hat{X}_{ba})\right\} \tag{3.14}
$$

and

$$
A_5=(Y_a+Y_b)\left(\frac{1}{X_B}-\frac{1}{X_A}\right)\left\{Cov(\hat{Y}_{ab},\hat{X}_{ab})+Cov(\hat{Y}_{ba},\hat{X}_{ba})\right\} \tag{3.15}
$$

The mean squared error of the proposed class of estimators $\hat{Y}_{\text{new}}$ can then be written as:

$$
\begin{aligned}MSE(\hat{Y}_{\text{new}})&=V(\hat{Y}_a)+V(\hat{Y}_b)+V(\hat{Y}_{ba})+2Cov(\hat{Y}_b,\hat{Y}_{ba})\\ &\quad+\gamma^2A_1+\alpha^2A_2-2\gamma A_3-2\alpha A_4+2\alpha\gamma A_5\end{aligned} \tag{3.16}
$$

The optimum values of α and γ which minimizes the mean squared error in (3.16), are given by:

$$
\alpha=\frac{A_1A_4-A_3A_5}{A_1A_2-A_5^2} \tag{3.17}
$$

and

$$
\gamma=\frac{A_2A_3-A_4A_5}{A_1A_2-A_5^2} \tag{3.18}
$$

The resultant minimum mean squared error of the proposed class of estimators $\hat{Y}_{\text{new}}$ is given by

$$Min.MSE(\hat{Y}_{\text{new}}) = V(\hat{Y}_a) + V(\hat{Y}_b) + V(\hat{Y}_{ba}) + 2Cov(\hat{Y}_b, \hat{Y}_{ba}) - \frac{A_1 A_4^2 + A_2 A_3^2 - 2A_3 A_4 A_5}{A_1 A_2 - A_5^2} \tag{3.19}$$

or

$$Min.MSE(\hat{Y}_{\text{new}}) = Min.V(\hat{Y}_{\text{Hartley}}) + \frac{A_3^2}{A_1} - \frac{A_1 A_4^2 + A_2 A_3^2 - 2A_3 A_4 A_5}{A_1 A_2 - A_5^2} \tag{3.20}$$

where

$$Min.V(\hat{Y}_{\text{Hartley}}) = V(\hat{Y}_a) + V(\hat{Y}_b) + V(\hat{Y}_{ba}) + 2Cov(\hat{Y}_b, \hat{Y}_{ba}) - \frac{A_3^2}{A_1} \tag{3.21}$$

Thus we have:

$$Min.MSE(\hat{Y}_{\text{new}}) = Min.V(\hat{Y}_{\text{Hartley}}) + \frac{A_3^2 (A_1 A_2 - A_5^2) - A_1 (A_1 A_4^2 + A_2 A_3^2 - 2A_3 A_4 A_5)}{A_1 (A_1 A_2 - A_5^2)}$$

or

$$Min.MSE(\hat{Y}_{\text{new}}) = Min.V(\hat{Y}_{\text{Hartley}}) + \frac{A_3^2 A_1 A_2 - A_3^2 A_5^2 - A_1^2 A_4^2 - A_1 A_2 A_3^2 + 2A_1 A_3 A_4 A_5}{A_1 (A_1 A_2 - A_5^2)}$$

or

$$Min.MSE(\hat{Y}_{\text{new}}) = Min.V(\hat{Y}_{\text{Hartley}}) - \frac{(A_3 A_5 - A_1 A_4)^2}{A_1 (A_1 A_2 - A_5^2)} \tag{3.22}$$

Note that:

$$A_1 (A_1 A_2 - A_5^2) = A_1^2 A_2 \left(1 - \frac{A_5^2}{A_1 A_2}\right) = A_1^2 A_2 \left(1 - \rho^2\right) > 0 \tag{3.23}$$

where $\rho^2 = \dfrac{\{Cov\{(Y_a + Y_b)\psi, Y_{ab}(\in_{ab} - \in_{ba})\}\}^2}{V\big(Y_{ab}(\in_{ab} - \in_{ba})\big)V\big((Y_a + Y_b)\psi\big)}$ is a square of the usual correlation coefficient between two variables.

So from (3.22) and (3.23), the proposed class of estimators $\hat{Y}_{\text{new}}$ is always more efficient than the Hartley (1962, 1974) estimator. Hence no need of any simulation study or numerical results.

The reduction in variance $\dfrac{(A_3 A_5 - A_1 A_4)^2}{A_1 (A_1 A_2 - A_5^2)}$ could be small or large depending on the nature of population under study, see Srivastava and Jhajj (1980) where they used known parameters of auxiliary variables.

In order to see the magnitude of the percent relative efficiency of the new proposed class of estimators we consider hypothetical situation in the following section.

4. SIMULATION STUDY

Let:

$\rho_{Y_aY_{ab}} = \dfrac{Cov(\hat{Y}_a, \hat{Y}_{ab})}{\sqrt{V(\hat{Y}_a)V(\hat{Y}_{ab})}}$ be the correlation coefficient between $\hat{Y}_a$ and $\hat{Y}_{ab}$;

$\rho_{Y_bY_{ba}} = \dfrac{Cov(\hat{Y}_b, \hat{Y}_{ba})}{\sqrt{V(\hat{Y}_b)V(\hat{Y}_{ba})}}$ be the correlation coefficient between $\hat{Y}_b$ and $\hat{Y}_{ba}$;

$\rho_{X_aX_{ab}} = \dfrac{Cov(\hat{X}_a, \hat{X}_{ab})}{\sqrt{V(\hat{X}_a)V(\hat{X}_{ab})}}$ be the correlation coefficient between $\hat{X}_a$ and $\hat{X}_{ab}$;

$\rho_{Y_bX_{ba}} = \dfrac{Cov(\hat{Y}_b, \hat{X}_{ba})}{\sqrt{V(\hat{Y}_b)V(\hat{X}_{ba})}}$ be the correlation coefficient between $\hat{Y}_b$ and $\hat{X}_{ba}$;

$\rho_{Y_bX_{ab}} = \dfrac{Cov(\hat{Y}_b, \hat{X}_{ab})}{\sqrt{V(\hat{Y}_b)V(\hat{X}_{ab})}}$ be the correlation coefficient between $\hat{Y}_b$ and $\hat{X}_{ab}$;

$\rho_{X_bX_{ba}} = \dfrac{Cov(\hat{X}_b, \hat{X}_{ba})}{\sqrt{V(\hat{X}_b)V(\hat{X}_{ba})}}$ be the correlation coefficient between $\hat{X}_b$ and $\hat{X}_{ba}$;

$\rho_{Y_{ba}X_{ba}} = \dfrac{Cov(\hat{Y}_{ba}, \hat{X}_{ba})}{\sqrt{V(\hat{Y}_{ba})V(\hat{X}_{ba})}}$ be the correlation coefficient between $\hat{Y}_{ba}$ and $\hat{X}_{ba}$;

and

$\rho_{Y_{ab}X_{ab}} = \dfrac{Cov(\hat{Y}_{ab}, \hat{X}_{ab})}{\sqrt{V(\hat{Y}_{ab})V(\hat{X}_{ab})}}$ be the correlation coefficient between $\hat{Y}_{ab}$ and $\hat{X}_{ab}$;

It is likely that, for any sampling designs being used in the frames A and B, the values of the correlation coefficients $\rho_{Y_aY_{ab}}$, $\rho_{Y_bY_{ba}}$, $\rho_{X_aX_{ab}}$, $\rho_{Y_bX_{ba}}$, $\rho_{Y_bX_{ab}}$ and $\rho_{X_bX_{ba}}$ are negative. However the values of the correlation coefficients $\rho_{Y_{ba}X_{ba}}$ and $\rho_{Y_{ab}X_{ab}}$ could be positive or negative. By keeping these observations in mind, we simulated situations where the proposed class of estimator remains more efficient than the Hartley's estimator and the absolute value of the relative bias in the proposed estimator is negligible.

The percent relative efficiency of the proposed class of estimator with respect the Hartley's estimator is defined as:

$$\text{RE} = \frac{Min.V(\hat{Y}_{\text{Hartley}})}{Min.MSE(\hat{Y}_{\text{new}})} \times 100\% \tag{4.1}$$

The percent relative bias in the proposed class of estimator is computed as:

$$\text{RB} = \frac{B(\hat{Y}_{new})}{Y_a + Y_b + Y_{ab}} \times 100\% \tag{4.2}$$

To search for situations where the proposed class of estimators has mean squared error smaller than the Hartley's estimator we wrote FORTRAN codes which can be had from the authors on a request. We consider hypothetical situations with the parameters:

$Y_a = 723$, $Y_b = 215$, $X_a = 523$, $X_b = 334$, $Y_{ab} = 312$, $X_{ab} = 212$, $V(\hat{Y}_a) = 75$, $V(\hat{Y}_b) = 80$, $V(\hat{Y}_{ab}) = 65$, $V(\hat{Y}_{ba}) = 70$, $V(\hat{X}_{ab}) = 75$, $V(\hat{X}_{ba}) = 75$, $V(\hat{X}_a) = 80$, and $V(\hat{X}_b) = 90$.

Realistically we also assumed that:

$$\rho_{X_b X_{ba}} = \rho_{X_a X_{ab}}, \ \rho_{Y_b Y_{ba}} = \rho_{Y_a Y_{ab}}, \ \rho_{Y_b X_{ba}} = \rho_{Y_a X_{ab}} \text{ and } \rho_{Y_{ba} X_{ba}} = \rho_{Y_{ab} X_{ab}}.$$

As said earlier, the percent relative efficiency depends on the situation being considered and it varies from 100% to 122.25% for the 18968 situations that were considered in the simulation study. As reported in Fig. 4.1, the percent relative bias (RB) can be seen to be close to zero in the range -0.0025% to +0.0025%. Table 4.1 with values below gives a summary of results obtained from the 18968 points for which $\rho_{Y_{ba} X_{ba}} = \rho_{Y_{ab} X_{ab}}$,with values between -0.91 to 0.89, with a step of 0.1.

Table 4.1 Descriptive summary of the RE

$\rho_{Y_{ab} X_{ab}}$	Freq	Min	Med	Max
-0.91	999	100.00	100.84	119.63
-0.81	999	100.00	100.81	118.50
-0.71	1000	100.00	100.78	117.57
-0.61	998	100.00	100.75	116.80
-0.51	998	100.00	100.74	116.16
-0.41	1000	100.00	100.71	115.64
-0.31	1000	100.00	100.70	115.21
-0.21	998	100.00	100.69	115.13
-0.11	997	100.00	100.69	115.17
-0.01	991	100.00	100.70	115.30
0.09	997	100.00	100.69	115.51
0.19	998	100.00	100.70	115.83
0.29	999	100.00	100.69	116.25
0.39	999	100.00	100.70	116.78
0.49	999	100.00	100.72	117.46
0.59	999	100.00	100.74	118.30
0.69	1000	100.00	100.76	119.34
0.79	999	100.00	100.78	120.63
0.89	998	100.00	100.82	122.25

A graphical representation of percent relative bias and percent relative efficiency is given in Fig. 4.1.

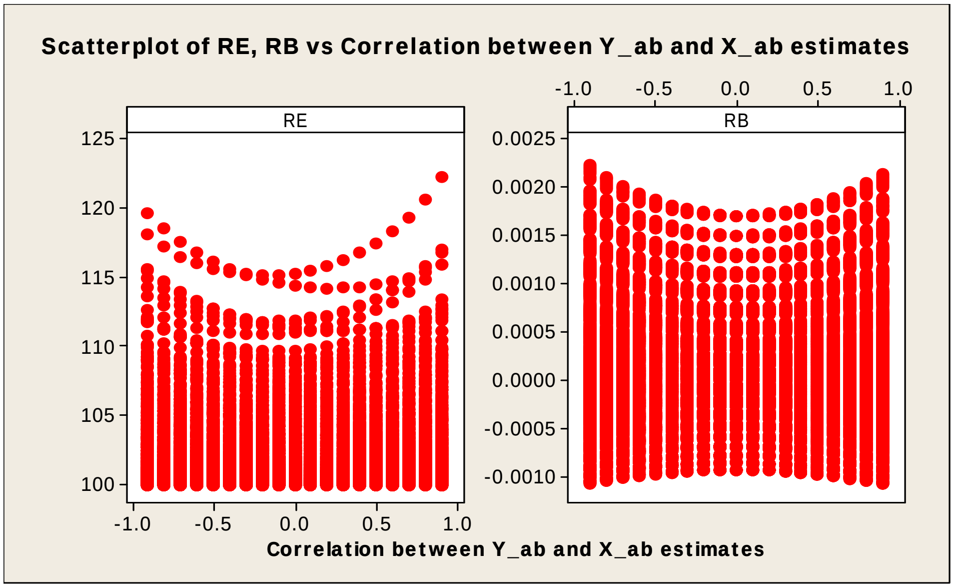

Fig. 4.1. Graphs of RE and RB obtained from 18968 data values.

Thus we conclude that there exists a choice of parameters in different populations where the proposed class of estimators can be efficiently used to estimate the population total when the data is collected from two frames no matter what the sampling designs have been used.

Remark: One obvious improvement of Fuller and Burmeister (1972) can be seen in a class room exercise:

$$\hat{Y}_{\text{FB(new)}} = (\hat{Y}_a + \hat{Y}_b)\left[\frac{(\hat{X}_a + \hat{X}_{ba})(\hat{X}_b + \hat{X}_{ab})}{\hat{X}_A \hat{X}_B}\right]^{\alpha_1} + \theta_1 \hat{Y}_{ab} + (1-\theta_1)\hat{Y}_{ba} + \theta_2(\hat{N}_{ab} - \hat{N}_{ba})$$

In the next section, we suggest a wider class of estimators making the use of multi-covariates.

5. MULTI-COVARIATES

Let $(Y_a, X_a^{(j)})$, $(Y_b, X_b^{(j)})$, $(Y_{ab}, X_{ab}^{(j)})$, $(Y_A, X_A^{(j)})$ and $(Y_B, X_B^{(j)})$, $j = 1,2,..,k$ be the unknown population totals of the main variate Y and k variables $X^{(j)}$. In such situation, we suggest a new wider class of estimators defined as:

$$\hat{Y}_{\text{new(w)}} = \left(\hat{Y}_a + \hat{Y}_b\right)\prod_{j=1}^{k}\left[\frac{(\hat{X}_a^{(j)} + \hat{X}_{ba}^{(j)})(\hat{X}_b^{(j)} + \hat{X}_{ab}^{(j)})}{\hat{X}_A^{(j)} \hat{X}_B^{(j)}}\right]^{\alpha_j} + \gamma \hat{Y}_{ab} + (1-\gamma)\hat{Y}_{ba} \tag{5.1}$$

where α_j, $j = 1,2,...,k$ are real constants to be determined such that the mean squared error of the proposed wider class of estimators is minimum. Such a determination seems could require a long class room exercise in extending the results, and can be solved if required.

ACKNOWLEDGEMENT

The use of Art Explosion 600,000 in producing the illustrations is duly acknowledged.

REFERENCES

Cochran, W. G. (1940). Some properties of estimators based on sampling scheme with varying probabilities. *Australian Journal of Statistics*, 17, 22-28.

Fuller, W.A. and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Hàjek, J. (1958). Some contribution to the theory of probability sampling. *Bull. Int. Statist. Inst.*, 36, 127-134.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya*, C, 36, 99-118.

Horvitz, D.G. and Thompson, D.J. (1952). A generalisation of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663--685.

Lohr, S.L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. Survey Methodology, 37(2), 197-213

Lohr, S.L. and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Rao, J.N.K. and Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Singh, S., Mangat, N.S. and Mahajan, P.K. (1995). General class of estimators. *Journal of the Indian Soc. Agricultural Statistics*, 47, 129-133.

Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael Selected Amy*. vol 1 & 2, Kluwer Academic Publihers, The Netherlands.

Srivastava, S.K. (1967). An estimator using auxiliary information in sample surveys. C*alcutta Statist. Assoc. Bull.*,16,121--132.

Srivastava, S.K. and Jhajj, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, 68, 341-343.

APPENDIX: NOTATIONS AND EXPECTED VALUES

Let

$$\in_a = \frac{\hat{Y}_a}{Y_a} - 1,\ \in_b = \frac{\hat{Y}_b}{Y_b} - 1,\ \in_{ab} = \frac{\hat{Y}_{ab}}{Y_{ab}} - 1,\ \in_{ba} = \frac{\hat{Y}_{ba}}{Y_{ab}} - 1,\ \in_A = \frac{\hat{Y}_A}{Y_A} - 1,$$

$$\in_B = \frac{\hat{Y}_B}{Y_B} - 1,\ \delta_a = \frac{\hat{X}_a}{X_a} - 1,\ \delta_b = \frac{\hat{X}_b}{X_b} - 1,\ \delta_{ab} = \frac{\hat{X}_{ab}}{X_{ab}} - 1,\ \delta_{ba} = \frac{\hat{X}_{ba}}{X_{ab}} - 1,$$

$$\delta_A = \frac{\hat{X}_A}{X_A} - 1 \text{ and } \delta_B = \frac{\hat{X}_B}{X_B} - 1$$

such that

$$E(\in_a) = E(\in_b) = E(\in_{ab}) = E(\in_{ba}) = E(\in_A) = E(\in_B) = 0$$

$$E(\delta_a) = E(\delta_b) = E(\delta_{ab}) = E(\delta_{ba}) = E(\delta_A) = E(\delta_B) = 0$$

$$E(\in_a^2) = \frac{V(\hat{Y}_a)}{Y_a^2},\ E(\in_b^2) = \frac{V(\hat{Y}_b)}{Y_b^2},\ E(\in_{ab}^2) = \frac{V(\hat{Y}_{ab})}{Y_{ab}^2},\ E(\in_{ba}^2) = \frac{V(\hat{Y}_{ba})}{Y_{ab}^2},$$

$$E(\in_A^2) = \frac{V(\hat{Y}_A)}{Y_A^2},\ E(\in_B^2) = \frac{V(\hat{Y}_B)}{Y_B^2},\ E(\in_a \in_b) = 0,\ E(\in_a \in_{ab}) = \frac{Cov(\hat{Y}_a, \hat{Y}_{ab})}{Y_a Y_{ab}},$$

$$E(\in_a \in_{ba}) = 0, E(\in_a \in_A) = \frac{V(\hat{Y}_a) + Cov(\hat{Y}_{ab}, \hat{Y}_a)}{Y_a Y_A}, E(\in_a \in_B) = 0, E(\in_b \in_{ab}) = 0,$$

$$E(\in_b \in_{ba}) = \frac{Cov(\hat{Y}_b, \hat{Y}_{ba})}{Y_b Y_{ab}}, E(\in_b \in_A) = 0,\ E(\in_b \in_B) = \frac{V(\hat{Y}_b) + Cov(\hat{Y}_{ba}, \hat{Y}_b)}{Y_b Y_B},$$

$$E(\in_{ab} \in_{ba}) = 0,\ E(\in_{ab} \in_A) = \frac{Cov(\hat{Y}_a, \hat{Y}_{ab}) + V(\hat{Y}_{ab})}{Y_{ab} Y_A},\ E(\in_{ab} \in_B) = 0,$$

$$E(\in_{ba} \in_A) = 0,\ E(\in_{ba} \in_B) = \frac{Cov(\hat{Y}_b, \hat{Y}_{ba}) + V(\hat{Y}_{ba})}{Y_{ab} Y_B},\ E(\in_A \in_B) = 0,$$

$$E(\delta_a^2) = \frac{V(\hat{X}_a)}{X_a^2},\ E(\delta_b^2) = \frac{V(\hat{X}_b)}{X_b^2}, E(\delta_{ab}^2) = \frac{V(\hat{X}_{ab})}{X_{ab}^2},\ E(\delta_{ba}^2) = \frac{V(\hat{X}_{ba})}{X_{ab}^2},$$

$$E(\delta_A^2) = \frac{V(\hat{X}_A)}{X_A^2},\ E(\delta_B^2) = \frac{V(\hat{X}_B)}{X_B^2},\ E(\delta_a \delta_b) = 0,\ E(\delta_a \delta_{ab}) = \frac{Cov(\hat{X}_a, \hat{X}_{ab})}{X_a X_{ab}},$$

$$E(\delta_a \delta_b) = 0,\ E(\delta_a \delta_{ab}) = \frac{Cov(\hat{X}_a, \hat{X}_{ab})}{X_a X_{ab}},\ E(\delta_a \delta_{ba}) = 0,$$

$$E(\delta_a \delta_A) = \frac{V(\hat{X}_a) + Cov(\hat{X}_{ab}, \hat{X}_a)}{X_a X_A},\ E(\delta_a \delta_B) = 0,\ E(\delta_b \delta_{ab}) = 0,$$

$$E(\delta_b \delta_{ba}) = \frac{Cov(\hat{X}_b, \hat{X}_{ba})}{X_b X_{ab}},\ E(\delta_b \delta_A) = 0,\ E(\delta_b \delta_B) = \frac{V(\hat{X}_b) + Cov(\hat{X}_{ba}, \hat{X}_b)}{X_b X_B},$$

$$E(\delta_{ab}\delta_{ba})=0,\ E(\delta_{ab}\delta_{A})=\frac{Cov(\hat{X}_a,\hat{X}_{ab})+V(\hat{X}_{ab})}{X_{ab}X_A},\ E(\delta_{ab}\delta_{B})=0,$$

$$E(\delta_{ba}\delta_{A})=0,\ E(\delta_{ba}\delta_{B})=\frac{Cov(\hat{X}_b,\hat{X}_{ba})+V(\hat{X}_{ba})}{X_{ab}X_B},\ E(\delta_{A}\delta_{B})=0,$$

$$E(\in_a\ \delta_a)=\frac{Cov(\hat{Y}_a,\hat{X}_a)}{Y_aX_a},\quad E(\in_a\ \delta_b)=0,\quad E(\in_a\ \delta_{ab})=\frac{Cov(\hat{Y}_a,\hat{X}_{ab})}{Y_aX_{ab}},\quad E(\in_a\ \delta_{ba})=0,$$

$$E(\in_a\ \delta_A)=\frac{Cov(\hat{X}_a,\hat{Y}_a)+Cov(\hat{X}_{ab},\hat{Y}_a)}{X_AY_a},\ E(\in_a\ \delta_B)=0,$$

$$E(\in_b\ \delta_a)=0,\ E(\in_b\ \delta_b)=\frac{Cov(\hat{Y}_b,\hat{X}_b)}{X_bY_b},\ E(\in_{ab}\ \delta_b)=0,\ E(\in_b\ \delta_{ab})=0,$$

$$E(\in_b\ \delta_{ba})=\frac{Cov(\hat{Y}_b,\hat{X}_{ba})}{Y_bX_{ba}},\ E(\in_b\ \delta_A)=0,\ E(\in_b\ \delta_B)=\frac{Cov(\hat{X}_b,\hat{Y}_b)+Cov(\hat{X}_{ba},\hat{Y}_b)}{Y_bX_B},$$

$$E(\in_{ab}\ \delta_a)=\frac{Cov(\hat{X}_a,\hat{Y}_{ab})}{X_aY_{ab}},\ E(\in_{ab}\ \delta_{ab})=\frac{Cov(\hat{X}_{ab},\hat{Y}_{ab})}{X_{ab}Y_{ab}},$$

$$E(\in_{ab}\ \delta_A)=\frac{Cov(\hat{Y}_{ab},\hat{X}_{ab})+Cov(\hat{Y}_{ab},\hat{X}_a)}{Y_{ab}X_A},\ E(\in_{ab}\ \delta_B)=0,\ E(\in_{ba}\ \delta_a)=0,$$

$$E(\in_{ba}\ \delta_b)=\frac{Cov(\hat{X}_b,\hat{Y}_{ba})}{X_bY_{ab}},\ E(\in_{ba}\in_{ab})=0,\ E(\in_{ba}\ \delta_{ab})=0,$$

$$E(\in_{ba}\ \delta_{ba})=\frac{Cov(\hat{Y}_{ba},\hat{X}_{ba})}{Y_{ab}X_{ab}},\ E(\in_{ba}\ \delta_A)=0,$$

$$E(\in_{ba}\ \delta_B)=\frac{Cov(\hat{Y}_{ba},\hat{X}_b)+Cov(\hat{Y}_{ba},\hat{X}_{ba})}{Y_{ab}X_B},\ E(\in_{ab}\ \delta_{ba})=0,$$

$$E(\in_A\ \delta_a)=\frac{Cov(\hat{Y}_a,\hat{X}_a)+Cov(\hat{Y}_{ab},\hat{X}_a)}{Y_AX_a},\ E(\in_A\ \delta_b)=0,$$

$$E(\in_A\ \delta_{ab})=\frac{Cov(\hat{Y}_a,\hat{X}_{ab})+Cov(\hat{Y}_{ab},\hat{X}_{ab})}{Y_AX_{ab}},\ E(\in_A\ \delta_{ba})=0,$$

$$E(\in_A\ \delta_A)=\frac{Cov(\hat{Y}_A,\hat{X}_A)}{Y_AX_A},\ E(\in_A\ \delta_B)=0,\ E(\in_B\ \delta_a)=0,$$

$$E(\in_B\ \delta_b)=\frac{Cov(\hat{Y}_b,\hat{X}_b)+Cov(\hat{Y}_{ba},\hat{X}_b)}{Y_BX_b},\ E(\in_B\ \delta_{ab})=0,\ E(\in_B\ \delta_A)=0$$

$$E(\in_B\ \delta_{ba})=\frac{Cov(\hat{Y}_b,\hat{X}_{ba})+Cov(\hat{Y}_{ba},\hat{X}_{ba})}{Y_BX_{ab}},\ \text{and}\ E(\in_B\ \delta_B)=\frac{Cov(\hat{Y}_B,\hat{X}_B)}{Y_BX_B}.$$