

**MODIFIED REGRESSION-TYPE ESTIMATOR IN TWO-PHASE
SAMPLING USING ARBITRARY PROBABILITIES**

Asifa Kamal¹, Muhammad Qaiser Shahbaz² and Muhammad Hanif³

¹ Department of Statistics, Lahore College for Women University,
Lahore, Pakistan. Email: asifa_kamal@hotmail.com

² Department of Mathematics, COMSATS Institute of Information
Technology, Lahore, Pakistan. Email: hafiz_shahbaz@yahoo.com

³ Department of Mathematics, Lahore University of Management
Sciences, Lahore, Pakistan. Email: hanif@lums.edu.pk

ABSTRACT

A modified regression-type estimator has been constructed for two-phase sampling using two auxiliary variables w, x and a measure of size z . The estimator has been developed by using probability proportional to a measure of size with replacement at first phase and equal probability sampling without replacement at the second phase. The mean square error of the modified estimator has been derived under arbitrary probabilities of selection. An empirical study has been made of the performance of new estimator with Roy's estimator (2003).

KEY WORDS

Two-phase sampling; auxiliary variable; measure of size; arbitrary probabilities; regression-type estimator.

1. INTRODUCTION

Use of equal probability for the purpose of sample selection is sometime difficult approach and gives unrepresentative sample when sampling units vary in size. If information about measure of size is available then unequal probability sampling is considered to be more appropriate. The rationale for choosing unequal probabilities of selection is to produce a more efficient estimator of the population total than by equal probabilities. The use of unequal probabilities of selection in two phase sampling has not been much widely used in literature. A notable reference where unequal probability sampling is used in connection with two phase sampling is of Raj (1965).

The use of unequal probabilities with two phase sampling can be adopted in many ways. The simplest way is to select a first phase sample of size n_1 with unequal probability sampling with replacement and the selection of second phase sample is done by using equal probabilities without replacement. In this paper we have adopted the same route to modify the Roy's (2003) estimator with unequal probabilities.

2. ESTIMATORS AVAILABLE IN LITERATURE

Unequal probabilities sample selection in two-phase sampling has been introduced by Raj (1965). Using probabilities proportional to size, Raj (1965) developed the following difference estimator:

$$t_1 = \sum_{i=1}^{n_2} \frac{y_i}{n_2 p_i} + k \left(\sum_{i=1}^{n_1} \frac{x_i}{n_1 p_i} - \sum_{i=1}^{n_2} \frac{x_i}{n_2 p_i} \right), \quad (2.1)$$

where $p_i = \frac{z_i}{Z}$, $Z = \sum_{i=1}^N z_i$, and k is the ratio between x and y .

The variance of t_1 is

$$\text{Var}(t_1) = \frac{V_p(y)}{n_2} + \left(\frac{1}{n_2} - \frac{1}{n_1} \right) \left[k^2 V_p(x) - 2k \rho_{xy(p)} \sqrt{V_p(x) V_p(y)} \right], \quad (2.2)$$

where

$$V_p(y) = \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i} - Y \right)^2,$$

and

$$\rho_{xy(p)} = \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i} - Y \right) \left(\frac{X_i}{p_i} - X \right) \frac{1}{\sqrt{V_p(x) V_p(y)}} = \frac{\text{Cov}_p(X, Y)}{\sqrt{V_p(x) V_p(y)}}.$$

Variance estimator of t_1 is

$$\text{var}(t_1) = \frac{1}{n_1(n_2-1)} \sum_{i=1}^{n_2} \left(\frac{y_i}{p_i} - \sum_{i=1}^{n_2} \frac{y_i}{n_2 p_i} \right)^2 + \left(\frac{1}{n_2} - \frac{1}{n_1} \right) \frac{1}{(n_2-1)} \sum_{i=1}^{n_2} \left(\frac{d_i}{p_i} - \sum_{i=1}^{n_2} \frac{d_i}{n_2 p_i} \right)^2, \quad (2.3)$$

where $d_i = y_i - kx_i$.

Srivenkataramana and Tracy (1989) used the optimum value of k and derived the following variance

$$\text{Var}(t_1) = \frac{1}{n_2} V_p(y) (1 - \rho_{xy(p)}^2) + \frac{1}{n_1} \rho_{xy(p)}^2 \cdot V_p(y). \quad (2.4)$$

Mohanty (1967) has proposed a regression estimator by measuring additional auxiliary variable z and x from a simple random sampling of size n_1 . Then a sample of size n_2 is selected from amongst these n_1 units with probability proportional to p_i , where $p_i = z_i / n_1 \bar{Z}_{n_1}$. Because x and y are correlated, y_i is estimated by:

$$t_2 = y_i + b(\bar{x}_{n_1} - x_i). \quad (2.5)$$

The variance of t_2 is given by:

$$V(t_2) = \left(\frac{n_2 - 1}{n_2} \right) \left(\frac{1}{n_1} - \frac{1}{N} \right) S_y^2 + \frac{1}{n_2 n_1^2} \left[\frac{n_1}{N} \sum_{i=1}^N u_i'^2 + \frac{n_1(n_1 - 1)}{N(N-1)} \sum_{i \neq j} \frac{u_i'^2}{Z_i} \cdot Z_j \right] - \frac{1}{n_2} \bar{Y}_N^2, \quad (2.6)$$

where

$$u_i'' = y_i + b(x_{n_1} - x_i) \text{ and } u_i' = u_i''/n_1 p_i.$$

An estimator of (2.6) is

$$v(t_2) = \left(\frac{1}{n_1} - \frac{1}{N} \right) s_y^2 + \frac{1}{n_2(n_2 - 1)} \sum_{i=1}^{n_2} (u_i' - \bar{u}_n')^2. \quad (2.7)$$

Roy (2003) proposed the following unbiased regression-type estimator in two-phase sampling is,

$$t_3 = \bar{y}_2 + k_1 \left[\bar{x}_1 + k_2 (\bar{W} - \bar{w}_1) - \left\{ \bar{x}_2 + k_3 (\bar{W} - \bar{w}_2) \right\} \right], \quad (2.8)$$

where the constants are chosen so that the variance of the estimator is minimum. The optimum values of k_1, k_2 and k_3 that minimize the variance of t_3 are given as:

$$k_1 = \left(\frac{\rho_{yx} - \rho_{yw}\rho_{xw}}{1 - \rho_{xw}^2} \right) \sqrt{\frac{\text{Var}(y)}{\text{Var}(x)}}; \quad k_2 = \beta_{yx.w}; \quad k_3 = \beta_{xw} - \frac{\beta_{yw}}{\beta_{yx.w}}.$$

A consistent estimator of t_3 is

$$t_3 = \bar{y}_2 + b_{yx.w} (\bar{x}_1 - \bar{x}_2) + b_{yx.w} b_{xw} (\bar{w}_2 - \bar{w}_1) + b_{yw} (\bar{W} - \bar{w}_2), \quad (2.9)$$

with sample regression coefficient being replaced with population counterpart.

The mean square error (MSE) of $O(n^{-1})$; of above estimator is:

$$MSE(t_3) = \bar{Y}^2 C_y^2 \left[\theta_2 (1 - \rho_{yx.w}^2) + \theta_1 (1 - \rho_{yw}^2) \rho_{xy.w}^2 \right], \quad (2.10)$$

where

$$\rho_{yx.w}^2 = \frac{\rho_{yx}^2 + \rho_{yw}^2 - 2\rho_{yx}\rho_{yw}\rho_{xw}}{(1 - \rho_{xw}^2)} \text{ and } \rho_{xy.w}^2 = \frac{(\rho_{yx} - \rho_{yw}\rho_{xw})^2}{(1 - \rho_{yw}^2)(1 - \rho_{xw}^2)}.$$

Roy (2003) proved that t_3 outperforms other competing two-phase regression estimators. In the following section we have modified the Roy's (2003) estimator by using arbitrary probabilities of selection.

3. MODIFICATION OF ROY'S (2003) REGRESSION-TYPE ESTIMATOR USING ARBITRARY PROBABILITIES

The modification of Roy's (2003) estimator with arbitrary probabilities is given as:

$$t_4 = \sum_{i=1}^{n_2} \frac{y_i}{n_2 p_i} + k_1 \left[\left\{ \sum_{i=1}^{n_1} \frac{x_i}{n_1 p_i} + k_2 \left(W - \sum_{i=1}^{n_1} \frac{w_i}{n_1 p_i} \right) \right\} - \left\{ \sum_{i=1}^{n_2} \frac{x_i}{n_2 p_i} + k_3 \left(W - \sum_{i=1}^{n_2} \frac{w_i}{n_2 p_i} \right) \right\} \right], \quad (3.1)$$

where W is the population total of w . The above estimator can be written as:

$$\begin{aligned} t_4 &= \sum_{i=1}^{n_2} \frac{y_i}{n_2 p_i} + k_1 \left(\sum_{i=1}^{n_1} \frac{x_i}{n_1 p_i} - \sum_{i=1}^{n_2} \frac{x_i}{n_2 p_i} \right) - k_1 k_2 \left(\sum_{i=1}^{n_1} \frac{w_i}{n_1 p_i} - W \right) + k_1 k_3 \left(\sum_{i=1}^{n_2} \frac{w_i}{n_2 p_i} - W \right) \\ &= \sum_{i=1}^{n_2} \frac{y_i}{n_2 p_i} + \alpha \left(\sum_{i=1}^{n_1} \frac{x_i}{n_1 p_i} - \sum_{i=1}^{n_2} \frac{x_i}{n_2 p_i} \right) - \beta \left(\sum_{i=1}^{n_1} \frac{w_i}{n_1 p_i} - W \right) + \gamma \left(\sum_{i=1}^{n_2} \frac{w_i}{n_2 p_i} - W \right), \end{aligned}$$

where $\alpha = k_1$, $\beta = k_1 k_2$ and $\gamma = k_1 k_3$. For unbiasedness we apply the second phase expectation.

$$\begin{aligned} E_2(t_4) &= \frac{n_2}{n_2} E_2 \left(\frac{y_i}{p_i} \right) + \alpha \left\{ \frac{n_1}{n_1} E_2 \left(\frac{x_i}{p_i} \right) - \frac{n_2}{n_2} E_2 \left(\frac{x_i}{p_i} \right) \right\} \\ &\quad - \beta \left\{ \frac{n_1}{n_1} E_2 \left(\frac{w_i}{p_i} \right) - W \right\} + \gamma \left\{ \frac{n_2}{n_2} E_2 \left(\frac{w_i}{p_i} \right) - W \right\} \\ &= \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{y_i}{p_i} - \beta \left\{ \sum_{i=1}^{n_1} \frac{w_i}{n_1 p_i} - W \right\} + \gamma \left\{ \sum_{i=1}^{n_1} \frac{w_i}{n_1 p_i} - W \right\}. \end{aligned}$$

Now applying the first phase expectation we have

$$\begin{aligned} E_1 \{ E_2(t_4) \} &= \frac{n_1}{n_1} E_1 \left(\frac{y_i}{p_i} \right) - \beta \left\{ \frac{n_1}{n_1} E_1 \left(\frac{w_i}{p_i} \right) - W \right\} + \gamma \left\{ \frac{n_1}{n_1} E_1 \left(\frac{w_i}{p_i} \right) - W \right\} \\ &= \sum_{i=1}^N \frac{y_i}{p_i} \cdot p_i - \beta \left\{ \sum_{i=1}^N \frac{w_i}{p_i} \cdot p_i - W \right\} + \gamma \left\{ \sum_{i=1}^N \frac{w_i}{p_i} \cdot p_i - W \right\} = Y. \end{aligned}$$

The variance of t_4 can be found by using

$$V(t_4) = E_1 \{ V_2(t_4) \} + V_1 \{ E_2(t_4) \}.$$

Now

$$V_1 \{ E_2(t_4) \} = \frac{1}{n_1} V_p(y) + (\beta - \gamma)^2 \frac{1}{n_1} V_p(w) - 2 \frac{1}{n_1} (\beta - \gamma) \sqrt{V_p(y) V_p(w)} \rho_{wy(p)}$$

$$V(t_4) = \frac{1}{n_1} V_p(y) + (\beta - \gamma)^2 \frac{1}{n_1} V_p(w) - 2 \frac{1}{n_1} (\beta - \gamma) \sqrt{V_p(w) V_p(y)} \rho_{wy(p)} + \theta_3 \left[\begin{aligned} &V_p(y) + \alpha^2 V_p(x) + \gamma^2 V_p(w) - 2\alpha \rho_{xy(p)} \sqrt{V_p(y) V_p(x)} \\ &+ 2\gamma \rho_{wy(p)} \sqrt{V_p(w) V_p(y)} - 2\alpha \gamma \rho_{wx(p)} \sqrt{V_p(w) V_p(x)} \end{aligned} \right], \quad (3.2)$$

where $\theta_3 = n_2^{-1} - n_1^{-1}$.

The optimum values of α , β and γ which minimizes $V(t_3)$ are given as:

$$\alpha = \beta_{xy.w(p)} = \frac{\sqrt{V_p(y)}}{\sqrt{V_p(x)}} \cdot \left[\frac{\rho_{xy(p)} - \rho_{wy(p)} \rho_{wx(p)}}{(1 - \rho_{wx(p)}^2)} \right]$$

$$\beta = \left[\frac{\rho_{wy(p)} (1 - \rho_{wx(p)}^2) - (\rho_{wy(p)} - \rho_{xy(p)} \rho_{wx(p)})}{(1 - \rho_{wx(p)}^2)} \right] \cdot \frac{\sqrt{V_p(y)}}{\sqrt{V_p(w)}}$$

$$\gamma = - \left\{ \frac{\rho_{wy(p)} - \rho_{xy(p)} \rho_{wx(p)}}{(1 - \rho_{wx(p)}^2)} \cdot \frac{\sqrt{V_p(y)}}{\sqrt{V_p(w)}} \right\} = -\beta_{wy.x(p)}.$$

Since $\alpha = k_1$, $\beta = k_1 k_2$ and $\gamma = k_1 k_3$. So the values of k_1, k_2, k_3 are:

$$k_1 = \beta_{xy.w(p)} = \frac{\sqrt{V_p(y)}}{\sqrt{V_p(x)}} \cdot \left[\frac{\rho_{xy(p)} - \rho_{wy(p)} \rho_{wx(p)}}{(1 - \rho_{wx(p)}^2)} \right]$$

$$k_2 = \beta_{wx(p)} = \rho_{wx(p)} \frac{\sqrt{V_p(x)}}{\sqrt{V_p(w)}}; \quad k_3 = \beta_{xw(p)} - \frac{\beta_{yw(p)}}{\beta_{xy.w(p)}}.$$

Because $\beta_{xy.w(p)}$, $\beta_{wx(p)}$ and $\beta_{wy(p)}$ are population quantities usually unknown so their consistent estimators $b_{xy.w(p)}$, $b_{wx(p)}$ and $b_{wy(p)}$ are computed from the sample. Substituting the values of α, β and γ in (3.2), the minimum variance of t_4 is

$$V(t_4) = \left(\frac{1}{n_2} \right) V_p(y) (1 - \rho_{y.wx(p)}^2) + \frac{1}{n_1} V_p(y) (1 - \rho_{wy(p)}^2) \rho_{xy.w(p)}^2. \quad (3.3)$$

4. COMPARISON OF THE ESTIMATORS

In this section the comparison of new estimator has been done with existing estimators. For this consider the estimators given in (2.1) and (2.9). The mean square

errors are given in (2.4) and (2.10). The mean square error of new estimator is given in (3.3). For the comparison consider:

$$\begin{aligned} V(t_1) &= \frac{1}{n_2} V_p(y) (1 - \rho_{xy(p)}^2) + \frac{1}{n_1} V_p(y) \rho_{xy(p)}^2 = V_p(y) [\theta_2 (1 - \rho_{xy(p)}^2) + n_1^{-1}] \\ V(t_3) &= V(y) [\theta_2 (1 - \rho_{y.wx}^2) + n_1^{-1} (1 - \rho_{wy}^2)] \\ V(t_4) &= V_p(y) [\theta_2 (1 - \rho_{y.wx(p)}^2) + n_1^{-1} (1 - \rho_{wy(p)}^2)]. \end{aligned}$$

Now

$$\frac{V(t_1)}{V(t_4)} = \frac{\theta_2 (1 - \rho_{xy(p)}^2) + n_1^{-1}}{\theta_2 (1 - \rho_{y.wx(p)}^2) + n_1^{-1} (1 - \rho_{yw(p)}^2)}. \quad (4.1)$$

In (4.1) $\rho_{y.wx(p)}^2 > \rho_{xy(p)}^2$ and $(1 - \rho_{yw(p)}^2) < 1$, so the numerator always exceed the denominator and so $V(t_1)/V(t_4) > 1$.

Again

$$\frac{V(t_3)}{V(t_4)} = \frac{V(y) [\theta_2 (1 - \rho_{y.wx}^2) + n_1^{-1} (1 - \rho_{yw}^2)]}{V_p(y) [\theta_2 (1 - \rho_{y.wx(p)}^2) + n_1^{-1} (1 - \rho_{yw(p)}^2)]}. \quad (4.2)$$

Now since $V(y) > V_p(y)$, $\rho_{y.xw}^2 < \rho_{y.xw(p)}^2$ and $\rho_{wy}^2 < \rho_{wy(p)}^2$ [variance of y for arbitrary probabilities is less as compared to equal probabilities but multiple correlation coefficient or simple correlation coefficient for arbitrary probabilities is greater than the multiple correlation coefficient or simple correlation coefficient for equal probabilities], the quantities in numerator exceed the denominator so $V(t_3)/V(t_4) > 1$ and so the new estimator always perform better than the estimator developed by Roy (2003).

5. EMPIRICAL STUDY

An empirical study has been conducted to illustrate the performance of the modified new regression-type estimator with those of Raj (1956) and Roy (2003). Four populations have been selected for this purpose from the agriculture sector of the Punjab.

z = Area Sown (thousand hectares)

y = Production 2004-05 (Thousand Metric Tons /Thousand Bales)

x = Production 2003-04 (Thousand Metric Tons /Thousand Bales)

w = Number of Tractors (Private and Government) 2004 census (March)

The source of the data is the Directorate of Agriculture, Crop Reporting Service, Punjab, Lahore.

Table 1
Percent Relative Efficiencies of proposed estimator t_4 with t_1

Estimator	$n_1 = 8$			$n_1 = 10$			$n_1 = 12$		
	n_2			n_2			n_2		
	2	3	4	2	4	6	2	4	6
Pop. I	105.59	106.17	106.52	105.23	106.26	106.72	104.91	106.01	106.52
Pop. II	101.12	101.28	101.39	101.04	101.04	101.46	100.97	101.23	101.39
Pop. III	105.68	105.23	104.84	105.88	105.14	104.58	106.02	105.37	104.84
Pop. IV	124.96	129.85	133.8	122.62	130.70	136.33	120.92	128.35	133.74

Table 2
Percent Relative Efficiencies of proposed estimator t_4 with t_3

Estimator	$n_1 = 8$			$n_1 = 10$			$n_1 = 12$		
	n_2			n_2			n_2		
	2	3	4	2	4	6	2	4	6
Pop. I	236.11	254.83	258.52	231.04	238.79	242.23	215.91	220.79	223.05
Pop. II	2166.27	2697.63	3082.92	1727.00	2534.69	3022.33	1391.8	2087.2	2527.72
Pop. III	2208.42	2980.24	3635.58	1714.25	2865.86	3755.26	1364.05	2295.31	3039.22
Pop. IV	1113.01	1383.15	1597.8	869.64	1241.62	1500.61	692.10	966.48	1165.71

Comparing the percent relative efficiencies it is observed that there is substantial increase in the percent relative efficiency of proposed estimator t_4 as compared to t_1 and t_3 .

REFERENCES

1. Mohanty, S. (1967). Combination of Regression and Ratio Type Estimator. *Jour. Ind. Statist. Assoc.* 5, 16-19.
2. Bureau of Statistics (2005). *Punjab Development Statistics*. Government of Punjab, Lahore.
3. Raj, D. (1965). On Sampling over Two Occasions with probability proportionate to size. *Ann. Math. Statist.*, 36, 327-330.
4. Roy, D.C. (2003). A regression-type estimator in two-phase sampling using two auxiliary variables. *Pak. J. Statist.*, 19(3), 281-290.
5. Srivenkataranana, T. and Tracy, D.S. (1989). Two-Phase Sampling for Selection with Probability Proportional to size in sample survey. *Biometika*, 76(4), 818-821.