ISSN 1684-8403
Journal of Statistics
Volume 15, 2008, pp. 7-16

New Difference Estimator in Two-phase Sampling using Arbitrary Probabilities

Asifa Kamal[1] and Muhammad Qaiser Shahbaz[2]

Abstract

A new difference estimator has been constructed in two-phase sampling using two auxiliary variables w and x .The first phase sampling unit has been selected with probability proportional to measure of size and second phase sample is selected with equal probability without replacement. The proposed estimator has been found to be more efficient as compared to Raj (1965) in which single auxiliary characteristic has been used. The variance expression can be easily extended to p-auxiliary variates.

Keywords

Two-phase sampling, Auxiliary variable, Measure of size, Arbitrary probabilities

1. Introduction

Auxiliary information can be used to increase the precision of regression estimators. Often information on many aspects of study is either available or can be collected from sample survey. This information can be used to improve the efficiency of regression estimators. Furthermore, another technique of improving the precision is to use unequal probabilities for selection of sample.

An unbiased estimator has been developed by extending the number of auxiliary variables in regression type estimator whereas sample is selected with arbitrary probabilities. The unavailability of prior knowledge of auxiliary information leads to the two-phase sampling.

[1]Department of Statistics, Lahore College for Women University, Pakistan
Email: asifa_kamal@hotmail.com
[2]Department of Mathematics, COMSATS Institute of Information Technology, Lahore, Pakistan
Email: qshahbaz@gmail.com

If information on a measure of size z is available, the initial sample s' of size n' is drawn with probability proportional to measure of size z with replacement and information about w and x is collected from s'. The second phase sample of size n is drawn using simple random sampling without replacement from s' to collect information about w, x and y. Raj (1965) used the probability proportional to size technique and developed the difference estimator using single auxiliary characteristic. Raj (1965) difference estimator of population total is:

$$\tilde{y}_1 = \sum_{i=1}^{n} \frac{y_i}{np_i} + k\left(\sum_{i=1}^{n'} \frac{x_i}{n'p_i} - \sum_{i=1}^{n} \frac{x_i}{np_i} \right) \tag{1.1}$$

where $p_i = \frac{z_i}{Z}$ and $Z = \sum_{i=1}^{N} z_i$

The Variance of $\tilde{y}_1'$ is:

$$V\left(\tilde{y}_1\right) = \frac{V_p\left(y\right)}{n} + \left(\frac{1}{n} - \frac{1}{n'}\right)\left[k^2 V_p\left(x\right) - 2k\rho_{xy(p)}\sqrt{V_p\left(x\right)V_p\left(y\right)} \right] \tag{1.2}$$

An estimator of $V\left(\tilde{y}_1\right)$ is:

$$v\left(\tilde{y}_1'\right) = \frac{1}{n'\left(n'-1\right)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \sum_{i=1}^{n} \frac{y_i}{np_i} \right)^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \frac{1}{\left(n'-1\right)} \sum_{i=1}^{n} \left(\frac{d_i}{p_i} - \sum_{i=1}^{n} \frac{d_i}{np_i} \right)^2$$

where $d_i = y_i - kx_i$

and

$$\rho_{xy(p)} = \sum_{i=1}^{\mathrm{N}} \mathrm{P}_i \left(\frac{Y_i}{\mathrm{P}_i} - \mathrm{Y} \right) \frac{\left(\frac{X_i}{\mathrm{P}_i} - \mathrm{X} \right)}{\sqrt{V_p\left(x\right)V_p\left(y\right)}} = \frac{Cov_p\left(X,Y\right)}{\sqrt{V_p\left(x\right)V_p\left(y\right)}}$$

Srivenkataramana and Tracy (1989) developed a variance expression using the optimum value of k which is given as:

$$V\left(\tilde{y}_1\right)=\frac{1}{n}V_p\left(y\right)\left(1-\rho^2_{xy(p)}\right)+\frac{1}{n'}\rho^2_{xy(p)}\cdot V_p\left(y\right) \tag{1.3}$$

where

$$V_p\left(y\right)=\sum_{i=1}^{\mathrm{N}}\mathrm{P}i\left(\frac{Y_i}{\mathrm{P}_i}-Y\right)^2 \tag{1.4}$$

Mukherjee et al. (1987) have developed series of regression estimators. Following regression estimator was proposed with two auxiliary variables using equal probability without replacement at both phases:

$$\bar{y}_2=\bar{y}+\beta_{xy}\left(\bar{x}'-\bar{x}\right)+\beta_{wy}\left(\bar{w}'-\bar{w}\right) \tag{1.5}$$

where the optimum values of β_{xy} and β_{wy} are given as:

$$\beta_{xy}=\frac{S_y}{S_x}.\frac{\left(\rho_{xy}-\rho_{wx}\rho_{wy}\right)}{\left(1-\rho^2_{wx}\right)}$$

$$\beta_{wy}=\frac{S_y}{S_w}.\frac{\left(\rho_{wy}-\rho_{wx}\rho_{xy}\right)}{\left(1-\rho^2_{wx}\right)}$$

Since β_{xy} and β_{wy} are population quantities their consistent estimators b_{xy} and b_{wy} are computed and estimator (1.5) becomes

$$\bar{y}_2=\bar{y}+b_{xy}\left(\bar{x}'-\bar{x}\right)+b_{wy}\left(\bar{w}'-\bar{w}\right)$$

The expression of MSE given by Mukherjee et al. (1987) for (1.5) is

$$MSE\left(\bar{y}_2\right)=\frac{S^2_y}{n}\left(1-\rho^2_{y.wx}\right)+\frac{S^2_y}{n'}\rho^2_{y.wx}$$

where $\rho^2_{y.wx}=\dfrac{\rho^2_{xy}+\rho^2_{wy}-2\rho_{xy}\rho_{wx}\rho_{wy}}{\left(1-\rho^2_{wx}\right)}$ is multiple correlation coefficient.

2. The Proposed New Regression Estimator Using Arbitrary Probabilities

An unbiased estimator of population total Y using two auxiliary characteristics is proposed as:

$$\tilde{y}_3 = \sum_{i=1}^{n} \frac{y_i}{np_i} + \alpha\left(\sum_{i=1}^{n'} \frac{w_i}{n'p_i} - \sum_{i=1}^{n} \frac{w_i}{np_i}\right) + \beta\left(\sum_{i=1}^{n'} \frac{x_i}{n'p_i} - \sum_{i=1}^{n} \frac{x_i}{np_i}\right) \tag{2.1}$$

where α *and* β are constants.

The variance of $\tilde{y}_3$ is derived as:

$$V(\tilde{y}_3) = \left(\frac{1}{n}\right)V_p(y) + \left(\frac{1}{n} - \frac{1}{n'}\right)\left[\begin{array}{l}\alpha^2 V_p(w) + \beta^2 V_p(x) - 2\alpha\rho_{wy(p)}\sqrt{V_p(y)V_p(w)} \\ -2\beta\rho_{xy(p)}\sqrt{V_p(x)V_p(y)} + 2\alpha\beta\rho_{wx(p)}\sqrt{V_p(w)V_p(x)}\end{array}\right] \tag{2.2}$$

The optimum value of α *and* β is obtained by minimizing (2.2). Hence differentiating (2.2) with respect to α and β respectively and equating the resulting derivative equal to zero:

$$\frac{\partial}{\partial\alpha}V(\tilde{y}_3) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left[2\alpha V_p(w) - 2\rho_{wy(p)}\sqrt{V(y)V(w)} + 2\beta\rho_{wx(p)}\sqrt{V_p(w)V_p(x)}\right] = 0$$

$$\frac{\partial}{\partial\beta}V(\tilde{y}_3) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left[2\beta V_p(x) - 2\rho_{xy(p)}\sqrt{V(x)V(y)} + 2\alpha\rho_{wx(p)}\sqrt{V_p(w)V_p(x)}\right] = 0$$

After simplification, following expressions of α *and* β are obtained as:

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

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

Substituting the values of α *and* β in (2.2), the resulting expression is:

$$V(\tilde{y}_3)=\left(\frac{1}{n}\right)V_p(y)+\left(\frac{1}{n}-\frac{1}{n'}\right)\left[\begin{array}{l}\frac{V_p(y)}{V_p(w)}\cdot\frac{\left(\rho_{wy(p)}-\rho_{wx(p)}\rho_{xy(p)}\right)^2}{\left(1-\rho^2_{wx(p)}\right)^2}\cdot V_p(w)+\frac{V_p(y)}{V_p(x)}\cdot\frac{\left(\rho_{xy(p)}-\rho_{wx(p)}\rho_{wy(p)}\right)^2}{\left(1-\rho^2_{wx(p)}\right)^2}\cdot V_p(x)\\ -2\rho_{wy}\frac{\sqrt{V_p(y)}}{\sqrt{V_p(w)}}\cdot\frac{\left(\rho_{wy(p)}-\rho_{wx(p)}\rho_{xy(p)}\right)}{\left(1-\rho^2_{wx(p)}\right)}\cdot\sqrt{V_p(w)V_p(y)}\\ -2\rho_{xy}\frac{\sqrt{V_p(y)}}{\sqrt{V_p(x)}}\cdot\frac{\left(\rho_{xy(p)}-\rho_{wx(p)}\rho_{wy(p)}\right)}{\left(1-\rho^2_{wx(p)}\right)}\cdot\sqrt{V_p(x)V_p(y)}\\ +2\rho_{wx}\cdot\frac{\sqrt{V_p(y)}}{\sqrt{V_p(w)}}\cdot\frac{\left(\rho_{wy(p)}-\rho_{wx(p)}\rho_{xy(p)}\right)}{\left(1-\rho^2_{wx(p)}\right)}\cdot\frac{\sqrt{V_p(y)}}{\sqrt{V_p(x)}}\\ \cdot\frac{\left(\rho_{xy(p)}-\rho_{wx(p)}\rho_{wy(p)}\right)}{\left(1-\rho^2_{wx(p)}\right)}\sqrt{V_p(w)V_p(x)}\end{array}\right]$$

After simplification $V(\tilde{y}_3)$ is given as:

$$V(\tilde{y}_3)=\frac{V_p(y)}{n}\left(1-\rho^2_{y.wx(p)}\right)+\frac{V_p(y)}{n'}\rho^2_{y.wx(p)} \tag{2.3}$$

where

$$\rho^2_{y.wx(p)}=\frac{1}{\left(1-\rho^2_{wx(p)}\right)}\left[\rho^2_{xy(p)}+\rho^2_{wy(p)}-2\rho_{wx(p)}\rho_{wy(p)}\rho_{xy(p)}\right] \tag{2.4}$$

The result derived in (2.3) can easily be generalized to q auxiliary variables x_i where i=1,2……q.

$$V(\tilde{y}_4)=\frac{V_p(y)}{n}\left(1-\rho^2_{y.1,2,....q(p)}\right)+\frac{V_p(y)}{n'}\rho^2_{y.1,2,.....q(p)} \tag{2.5}$$

Comparison of new estimator has been made with variance expression for single auxiliary characteristic (1.3) derived by Srivenkataramana and Tracy (1989). Taking the difference of variance expression (2.3) and (1.3)

$$V(\tilde{y}_3)-V(\tilde{y}_1)=\left[\frac{V_p(y)}{n}\left(1-\rho^2_{y.wx(p)}\right)+\frac{V_p(y)}{n'}\rho^2_{y.wx(p)}-\frac{1}{n}V_p(y)\left(1-\rho^2_{xy(p)}\right)-\frac{1}{n'}\rho^2_{xy(p)}V_p(y)\right]<0$$

After simplification

$$V(\tilde{y}_3)-V(\tilde{y}_1)=V_p(y)\left[\frac{1}{n'}\left(\rho^2_{y.wx(p)}-\rho^2_{xy(p)}\right)-\frac{1}{n}\left(\rho^2_{y.wx(p)}-\rho^2_{xy(p)}\right)\right]<0$$

$$V(\tilde{y}_3)-V(\tilde{y}_1)=V_p(y)\left[\frac{1}{n'}-\frac{1}{n}\right]<0$$

It is easy to conclude from the above expression that $(\tilde{y}_3)$ performs better than $(\tilde{y}_1)$ because first phase sample is always greater than second phase sample i.e. $(n'>n)$ which makes above expression negative. Hence increase in the auxiliary variables brings more precision in the regression estimator.

3. Special Result for Variance Derived for New Estimator

a. There is verification of variance expression. If $P_i=\frac{1}{N}$ is substituted in (2.1) and (2.3) then these two expressions become equivalent to $\tilde{y}_2$ and $MSE(\tilde{y}_2)$ respectively. Taking (2.1)

$$\tilde{y}_3=\sum_{i=1}^{n}\frac{y_i}{np_i}+\alpha\left(\sum_{i=1}^{n'}\frac{w_i}{n'p_i}-\sum_{i=1}^{n}\frac{w_i}{np_i}\right)+\beta\left(\sum_{i=1}^{n'}\frac{x_i}{n'p_i}-\sum_{i=1}^{n}\frac{x_i}{np_i}\right)$$

Subsituting $P_i=\frac{1}{N}$ in α *and* β respectively following results are obtained:

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

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

Hence the expression (2.1) becomes

$$\tilde{y}_3=N\sum_{i=1}^{n}\frac{y_i}{n}+\beta_{wy}\left(N\sum_{i=1}^{n'}\frac{w_i}{n'}-N\sum_{i=1}^{n}\frac{w_i}{n}\right)+\beta_{xy}\left(N\sum_{i=1}^{n'}\frac{x_i}{n'}-N\sum_{i=1}^{n}\frac{x_i}{n}\right)\qquad(3.1)$$

Similarly, substituting $P_i = \frac{1}{N}$ in (1.4) $V_{p(y)}$ and 2.4) $\rho^2_{y.wx(p)}$ then following results are obtained:

$$V_p(y) = N^2 S_y^2 \text{ and } \rho^2_{y.wx(p)} = \rho^2_{y.wx}$$

Hence (2.3) becomes equivalent to $MSE(\tilde{y}_2)$

$$V(\tilde{y}_3) = N^2\left[\frac{S_y^2}{n}\left(1-\rho^2_{y.wx}\right)+\frac{S_y^2}{n'}\rho^2_{y.wx}\right] \tag{3.2}$$
$$V(\tilde{y}_3) = MSE(\tilde{y}_2)$$

which is expression of estimator of population total of (5) and its MSE using simple random sampling with replacement at initial phase and simple random sampling without replacement at final phase.

b. These are the conditions under which variance of $\tilde{y}_3$ is more precise than competent estimator in simple random sampling. Comparing $V(\tilde{y}_3)$ in (2.3) and $MSE(\tilde{y}_2)$ in (3.2)

$$MSE(\tilde{y}_2) - V(\tilde{y}_3) = \frac{1}{n}\left(V(\tilde{y}_{srs}) - V_p(y)\right)$$
$$+\left(\frac{1}{n'}-\frac{1}{n}\right)\left[V(\tilde{y}_{srs})\rho^2_{y.wx} - V_p(y)\rho^2_{y.wx(p)}\right] \tag{3.3}$$

From (3.3) it is observed that (3.2) is greater as compared to (2.3) if following both inequalities hold that is

$$V(\tilde{y}_{srs}) - V_p(y) > 0 \tag{3.4}$$

and

$$\left[V(\tilde{y}_{srs})\rho^2_{y.wx} - V_p(y)\rho^2_{y.wx(p)}\right] > 0 \tag{3.5}$$

The inequality in (3.4) is greater than zero if $\sum_{i=1}^{N} \frac{Y_i^2}{Z_i}\left(Z_i - \bar{Z}\right) > 0$ that is estimated is closely related to measure of size. And also (3.5) holds true if both $\sum_{i=1}^{N} \frac{Y_i^2}{Z_i}\left(Z_i - \bar{Z}\right) > 0$ and $\rho_{y.wx}^2 > \rho_{y.wx(p)}^2$ hold.

4. Empirical Study

Empirical study has been conducted to illustrate the performance of the new estimator with the same estimator using single auxiliary characteristic and also with the competent estimator in simple random sampling in which two auxiliary characteristics have been used.

Four populations have been selected for this purpose from agriculture sector of Punjab (2005).
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)

Percent relative efficiency of new difference estimator (2.1) with Raj (1965) and Mukherjee et al.(1987) is calculated in Table 1 and Table 2, respectively, where $\left(f^* = \frac{n}{n'}\right)$ and $\left(f' = \frac{n'}{N}\right)$

Table 1: Percent Relative Efficiencies of New Difference Estimator $\tilde{y}_3$ in (2.1) with $\tilde{y}_1$ in (1.1)

Estimator	$f' = 0.2$			$f' = 0.4$			$f' = 0.5$		
	f^*			f^*			f^*		
	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4
Pop.I	100.3	100.1	100.1	100.3	100.2	100.1	100.3	100.2	100.1
Pop.II	100.3	100.2	100.1	100.3	100.2	100.1	100.3	100.3	100.2
Pop.III	105.2	103.9	102.8	105.1	104.3	102.6	105.4	104.2	102.7
Pop.IV	105.4	105.4	103.0	107.0	105.1	102.7	105.7	104.4	102.6

Table 2: Percent Relative Efficiencies of New Difference Estimator $\tilde{y}_3$ (2.1) with $\tilde{y}_2$ (3.1)

Estimator	$f' = 0.2$			$f' = 0.4$			$f' = 0.5$		
	f^*			f^*			f^*		
	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4
Pop.I	915.68	1116.35	1210.72	915.682	1038.06	1210.72	835.99	1048.42	1218.2
Pop.II	2425.76	3678.08	4466.68	2425.76	3132.17	4112.61	2081.40	2731.86	4032.94
Pop.III	2682.11	4669.27	6277.57	2851.44	3956.63	6606.11	2400.32	4212.8	6367.97
Pop.IV	1440.44	1440.44	2022.59	1038.23	1523.31	2114.94	1368.00	1691.60	2134.61

5. Conclusion

From empirical study it is concluded that variance for proposed estimators $\tilde{y}_3$ i.e. (2.1) is more precise than $\tilde{y}_2$ (3.1) and Raj (1965) regression estimator $\tilde{y}_1$ (1.1). Comparing the percent relative efficiencies it is observed that there is substantial increase in the percent relative efficiency of proposed estimator $\tilde{y}_3$ for all populations. From Table 2 it can be easily concluded that percent relative efficiencies always increase as $\left(f^* = \frac{n}{n'} \right)$ increases while in case of $\left(f' = \frac{n'}{N} \right)$ the increase is not always there.

It is therefore concluded that increase in auxiliary variable increases the efficiencies. Also using arbitrary probabilities improves the efficiency of estimator.

Acknowledgments

The authors acknowledge the referees and Chief Editor for their valuable comments and suggestions.

References

1. Mukherjee, R., Rao, T. J. and Vijayak, K. (1987). Regression type estimators using multiple auxiliary information. *Australian Journal of Statistics*, **29**, 244-254.
2. Punjab Development Statistics, Bureau of Statistics, Government of Punjab, Lahore, 2005.
3. Raj, D. (1965). On sampling over two occasions with probability proportionate to size. *Annals of Mathematical Statistics*, **36(1)**, 327-330.
4. Srivenkataramana, T. and Tracy, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample survey. *Biometrika*, **76(4)**, 818-821.