J. Ind. Soc. Agril. Statist.
58(3), 2004 : 307-311

Unequal Probability Sampling with Nonzero Intercept

S.J. Amdekar
GB Pant University of Agriculture and Technology, Pantnagar – 263 145
(Received : August, 2003)

SUMMARY

Roychoudhury (1957) has given an estimator of population total based on unequal probability sampling which is particularly useful in situations where there is linear relationship between study and auxiliary variables with the intercept not necessarily near zero. In this paper we consider generalization of this method. From empirical study it is observed that the proposed estimator is even better than regression estimator, the efficiency increasing with increase in relative intercept.

Key words : Unequal probability sampling, Linear relationship with nonzero intercept, Roychoudhury method, Generalization.

1. Introduction

The estimator based on unequal probability sampling and the ratio estimator are generally more efficient than the simple estimator based on equal probability sampling when there is linear relationship between study and auxiliary variables with the intercept near zero. Roychoudhury (1957) gave an estimator based on unequal probability sampling which is efficient even when the intercept is away from origin. For easy reference, the Roychoudhury estimator is presented in this section. In the next section generalization is given and empirical study is made.

1.1 Roychoudhury Estimator

Let a population consisting of N units with x and y as auxiliary and study variables respectively, be divided into two groups containing N_1 and N_2 units respectively, first group containing those units for which $X_i < \overline{X}$ and second group containing remaining units (those units for which $X_i = \overline{X}$ are assigned to one group (any one) only). Let X_{1i}, Y_{1i} (i = 1, 2, ..., N_1) and X_{2j}, Y_{2j} (j = 1, 2, ..., N_2) denote values of the units in the two groups respectively. For these units

define $P_{1i} = \frac{(X_{1i} - \overline{X})}{\Sigma(X_{1i} - \overline{X})}$ and $P_{2j} = \frac{(X_{2i} - \overline{X})}{\Sigma(X_{2j} - \overline{X})}$. A population of $N_1 \times N_2$ pairs is formed by pairing each unit of first group with each unit of second.

For a typical pair, study variable is defined as $Z_{ij} = P_{1i}\, Y_{2j} + P_{2j}\, Y_{1i}$ and selection probability as $Q_{ij} = \frac{(P_{1i} + P_{2j})}{N}$. Note that $\Sigma\Sigma\, Z_{ij} = Y =$ total for original population and $\Sigma\Sigma Q_{ij} = 1$. From this population of pairs a sample of n pairs is selected with replacement with probability Q_{ij} and estimator of Y, as given by Roychoudhury is $\hat{Y}_R = \frac{1}{n}\Sigma\frac{Z_{ij}}{Q_{ij}}$. Roychoudhury has shown that $\hat{Y}_R$ reduces to Y when y and x are related as $y = \alpha + \beta x$ thereby showing its utility for the situations where regression line is away from origin. Now we propose generalization of this method.

2. Generalization

The population is divided into two groups as in case of Roychoudhury method. Let sample size n be written as $n = n_1 + n_2$ where n_1 and n_2 are sizes of the samples to be selected from two groups. Now $M_1 = {}^{N1}C_{n1}$ and $M_2 = {}^{N2}C_{n2}$ are number of possible samples in two groups respectively. Let P_{1i} and P_{2j} be probabilities associated with i-th and j-th samples of two groups ($i = 1, 2, ..., M_1$ and $j = 1, 2, ... M_2$). By associating each sample of first group with each of second, we get a sample space containing $M_1 \times M_2$ elements. One sample from this space is selected with probability $Q_{ij} = \frac{(P_{1i} + P_{2j})}{(M_1 + M_2)}$. Denoting by w_{1i} sum of y values in i-th sample of first group and w_{2j} that of second group, an unbiased estimator of population total is then defined as $\hat{Y}_G = \frac{Z_{ij}}{Q_{ij}}$ where

$$Z_{ij} = P_{1i}\,(w_{2j}/M_2^*) + P_{2j}(w_{1i}/M_1^*),\ M_1^* = {}^{N1-1}C_{n1-1} \text{ and } M_2^* = {}^{N2-1}C_{n2-1}$$

The probabilities P_{1i} and P_{2j} used above are quite general. Now consider a particular case. Let $d_{1r} = \overline{X}_{1r} - \overline{X}, (r = 1, 2, ..., N_1)$ and $d_{2s} = X_{2s} - \overline{X}$ ($s = 1, 2, ..., N_2$) be the deviations of x in two groups with $D = \Sigma d_{1r} = \Sigma d_{2s}$. Let v_{1i} and v_{2j} be sum of deviations in i-th and j-th samples of two groups respectively. If P_{1i} and P_{2j} are taken proportional to v_{1i} and v_{2j} then

$P_{1i} = v_{1i}/M_1^*D$ and $P_{2j} = v_{2j}/M_2^*D$. Consequently, it can be shown that the estimator $\hat{Y}_G$ reduces to $\dfrac{(M_1 + M_2)(v_{2j}\, w_{1i} - v_{1i}w_{2j})}{(v_{2j}M_1^* - v_{1i}M_2^*)}$.

Further, it can easily be verified that with $n_1 = n_2$ and $N_1 = N_2$ this estimator will also reduce to Y when y and x are related as $y = \alpha + \beta x$. Thus, it will be advantageous to use this estimator when regression line is away from origin for populations symmetric with respect to x.

2.1 Variance Estimator

It is not possible to get a simplified expression for $V(\hat{Y}_G)$. Therefore one can not algebraically compare $\hat{Y}_G$ with other estimators. In a subsequent subsection nevertheless, it has been studied empirically. But variance estimator can be derived as explained below.

Noting that $V(\hat{Y}_G) = E(\hat{Y}_G^2) - Y^2$ we have $\hat{V}(\hat{Y}_G) = (\hat{Y}_G)^2 - (\hat{Y}^2)$. Thus we need estimator of Y^2. Let $Y = Y_1 + Y_2$ where $Y_1 = \Sigma Y_{1r}$ and $Y_2 = \Sigma Y_{2s}$. Thus $Y_1^2 = \Sigma\, Y_{1r}^2 + 2\,\Sigma\, Y_{1r}\, Y_{1r'}$ and $Y_2^2 = \Sigma Y_{2s}^2 + 2\Sigma Y_{2s} Y_{2s'}$. It is easy to see that $A = \dfrac{\left[P_{1i}\left(\dfrac{\Sigma Y_{2js}^2}{M_2^*}\right) + P_{2j}\left(\dfrac{\Sigma Y_{1ir}^2}{M_1^*}\right)\right]}{Q_{ij}}$ is an unbiased estimator of $\left(\Sigma Y_{2s}^2 + \Sigma Y_{1r}^2\right)$

Moreover $B = \dfrac{\left[P_{1i}\left(\dfrac{\Sigma Y_{2js}\, Y_{2js'}}{M_2^0} + P_{2j}\left(\dfrac{\Sigma Y_{1ir}\, Y_{1ir'}}{M_1^0}\right)\right)\right]}{Q_{ij}}$ is an unbiased estimator of $(\Sigma Y_{2s} Y_{2s'} + \Sigma Y_{1r} Y_{1r'})$ where $M_1^0 = {}^{N1-2}C_{n1-2}$ and $M_2^0 = {}^{N2-2}C_{n2-2}$. Lastly, $D = \dfrac{(\Sigma Y_{1ir}\, \Sigma Y_{2js})}{(M_1^* M_2^*\, Q_{ij})}$ is an unbiased estimator of $Y_1 \times Y_2$. Hence the unbiased variance estimator is given by $\hat{V}(\hat{Y}_G) = (\hat{Y}_G^2) - (A + 2B + 2D)$. Of course, for some samples this may take negative value.

2.2 Selection Procedure

Recall that in this method, from the sample space containing $M_1 \times M_2$ elements, one element is to be selected with probability $\dfrac{(P_{1i} + P_{2j})}{(M_1 + M_2)}$. This can be achieved using any of the methods described below.

First Method : Step (i) : Select one of the two groups with probability proportional to number of samples in that group. Step (ii) : From the group that is selected in step (i) select a sample with equal probability and from the other group select a sample with the probability associated with it. Selection of this sample can be made using Lahiri method (1951).

Second Method : Step (i) : Select one sample out of $(M_1 + M_2)$ samples with equal probability. Step (ii) : If the sample selected in step (i) belongs to first group then select one sample from second group with probability P_{2j} and vice versa. Selection of this sample can also be made using Lahiri method (1951).

2.3 Empirical Study

It has been stated earlier that the proposed estimator, $\hat{Y}_G$ reduces to parameter (the population total, Y) when there is perfect liner relationship with regression line having nonzero intercept for populations symmetric with respect to X. We therefore, make empirical comparison by considering a number of bivariate normal super populations with varying value of the intercept. Eight populations (designated as A1 to B4) of size N = 20 are drawn from bivariate normal distribution with parameters as given in Table 1 and then from each population a sample of n = 6 units is drawn.

Table 1. Values of the parameters of the super populations considered from empirical study

Parameter	A1	A2	A3	A4	B1	B2	B3	B4
μ_x	100	100	100	100	100	100	100	100
σ_x	20	20	20	20	20	20	20	20
μ_y	400	500	600	700	400	500	600	700
σ_y	50	50	50	50	80	80	80	80
ρ	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
α/μ_y	0.62	0.70	0.77	0.79	0.40	0.54	0.60	0.66

Observe that along with the usual five parameters we have also presented in the table value of α/μ_y, as a measure of intercept relative to mean of study variable. Here we have two sets (A and B) of super populations each consisting of four. μ_x, σ_x and ρ are same for both the sets. For the first set σ_y is same while value of μ_y is increasing (and hence that of α/μ_y). Similarly σ_y is same for second set with increasing value of μ_y (and hence that of α/μ_y).

In Table 2 variance of estimator of population mean and their relative efficiencies are presented. Note that in case of regression and ratio estimators exact mean square error (and not approximate variances) are worked out. We

observe that within each set with increase in the values of α/μ_y the efficiency of proposed, Roychoudhury and regression estimators also increases. A detailed study can be made considering more combinations of parameter values. But on the basis of present study it can be concluded that the proposed estimator is likely to perform better for populations with more variability in study variable and larger intercept.

Table 2. Variances (and efficiencies) of different estimators for two sets of super populations

Estimator	A1	A2	A3	A4	B1	B2	B3	B4
ppswr	407	773	1390	1091	630	1529	1941	2703
	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)
Ratio	307	552	973	780	311	756	952	1190
	(133)	(140)	(143)	(140)	(203)	(202)	(204)	(227)
Regression	230	266	178	114	135	76	152	85
	(200)	(291)	(781)	(957)	(467)	(2012)	(1277)	(3180)
$\hat{Y}_R$	248	317	199	142	191	114	208	121
	(164)	(244)	(698)	(768)	(330)	(1341)	(933)	(2234)
$\hat{Y}_G$	147	210	138	89	141	75	133	75
	(277)	(368)	(1007)	(1226)	(447)	(2039)	(1459)	(3604)

ACKNOWLEDGEMENT

Author is grateful to Prof. V.K. Sethi for his useful suggestions and encouragement and to the referee for constructive comments that led to improvement of the paper.

REFERENCES

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Instt.*, **33**, 133-140.

Roychoudhury, D.K. (1957). Unbiased sampling design using information provided by linear function of auxiliary variate. Thesis submitted for *Assoc. of Ind. Stat. Instt.*