USE OF TRANSFORMATION IN SAMPLING

S. MOHANTY

Orissa University of Agriculture & Technology, Bhubaneswar-3

and

M. N. DAS

Institute of Agricultural Research Statistics, New Delhi-12

Introduction

In sample surveys, the precision in estimating the population parameters of a finite population may be increased by using an auxiliary variate, X which is correlated with the character under study. The auxiliary variate is usually used in two ways (i) in the selection stage of the sample, when probabilities of selection of units depend on the sizes of the corresponding auxiliary variate and (*ii*) in the estimation process, after the units are selected following any one of the specified schemes. The probabilities proportional sampling scheme belongs to the first type, whereas the ratio and regression method of estimation belongs to the second type. Among a large number of auxiliary variates available, corresponding to the character under study, one is selected for the purpose mentioned above.

In this paper we have considered the situation where the transformation of origin and scale of the measurement of the auxiliary variate is utilised to reduce the bias of the estimate and increase simultaneously the efficiency of the estimate, when ratio method of estimation is adopted. Actually the main object of this paper is to expose the idea of the use of transformation in theory of sampling.

2. Bias of Ratio Estimator $\bar{y}_R$

For commonly used selection procedure the ratio estimator

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}}\bar{X}, \qquad \text{.....(1)}$$

of the population mean $\bar{Y}$ is generally biased. To find the bias it is assumed that X's are positive and sample size is sufficiently large so that

$$|(\bar{x}-\bar{X})|/\bar{X} < 1 \qquad \text{.....(2)}$$

These assumptions are not unreasonable in actual sample survey, since X's are usually positive or can be made positive through transformations. In other words

equation (2) means that $\bar{x}$ lies between O and $2X$, which is likely when variation in $\bar{x}$ is not large. Under the assumption given above, bias of $\bar{y}_R$ as an estimate of $\bar{Y}$, if we assume the terms involving power two and higher be neglected is

$$B(\bar{y}_R)=E(\bar{y}-\bar{Y})=\bar{Y}\ \theta\ [C_{20}-C_{11}], \qquad \ldots\ldots(3)$$

where

$$\theta=\left(\frac{1}{n}-\frac{1}{N}\right),\ C_{ij}=\frac{K_{ij}}{\bar{X}^i\ \bar{Y}^j}$$

and K_{ij} is the (ij)th cumulants of X and Y. The bias is zero, when

$$C_{20}=C_{11}$$

or $$\bar{Y}=\bar{X}(S_{xy}/S^2_x)$$

or $$Y=\beta X, \qquad \ldots\ldots(4)$$

where β is the population regression coefficient of Y on X. Thus, when X is so chosen that the regression line of Y on X passes through the origin the ratio estimator becomes unbiased for the case where terms involving power two and higher are neglected as stated before.

3. Method for Reducing Bias

Let $Y=\alpha+\beta X$, be the regression line of Y on X in the population. We can write this line as

$$Y=\beta\ (X+\alpha/\beta)$$

or $$Y=\beta X',$$

where $X'=X+\alpha/\beta$. Thus, if instead of X, we use X' as auxiliary variate, which is obtained from X by change of origin to $-\alpha/\beta$, the bias of ratio estimator as stated in equation (3) becomes zero. It may be pointed out here that X' and X have same correlation with Y, as we have used a linear transformation for transforming X to X'.

There are a large number of auxiliary variates from which one is selected. Thus, for some character Y, two different auxiliary variates use in ratio estimate may give two completely different estimates of the population mean. Hence, finding an unbiased estimate depends largely on the selection of X as well as the scale and origin of its measurement. So when α and β are known, we can always change the origin and scale of X such that the bias becomes zero. By changing the scale and origin, the correlation coefficient between Y and the auxiliary variate remains unaffected. As α and β are usually unknown, approximate estimate of them may be obtained through some pilot enquiry. A quick estimate can also be obtained by plotting the observations of the pilot enquiry on a graph. From a straight line drawn to fit the points, the estimates of α and β are found, where α is estimated by the intercept of the line on the Y-axis, β is estimated by the tangent of the angle

made by the line with the X-axis. When all the points are not lying on the same line, a large number of lines can be drawn to fit these points. However, the best results from the graph can be obtained with better judgement and skill of the sampler drawing the graph.

Let the estimated value of α and β be a and b respectively. Hence, we can change the origin of the measurement of X from O to $-a/b$ and which in turn will reduce the bias $\bar{y}_R$ as the regression line will now pass closer to origin if not through the origin. This is evident from the following considerations.

From the regression equation of Y on X, we have

$$\bar{Y}=\beta X+\alpha$$

$$\text{or} \quad S_x^2/X = S_{xy}/\bar{Y} + \alpha S_x^2/(\bar{X}\bar{Y})$$

$$\text{or} \quad C_{20}=C_{11}+(\alpha/\bar{Y})\, C_{20}$$

$$\text{or} \quad C_{11}=C_{20}\,(1-\alpha/\bar{Y}).$$

Substituting the values of C_{11} given in the equation (3), we have,

$$B(\bar{y}_R) = \bar{Y}\ \theta\ (\alpha/\bar{Y})\ C_{20}=\theta\ \alpha\ C_{20}.$$

Similarly, when the origin is changed from O to $-a/b$, the bias of the ratio estimate using X' $\left(=X+\frac{a}{b}\right)$ as an auxiliary variate can be written as

$$B(\bar{y}_R)_T=\theta\alpha' C'_{20},$$

where $\alpha'=\alpha-\beta(a/b)$ and $C'_{20}=S_x'^2/(\bar{X}')^2=S_x^2/(\bar{X}')^2$ as the mean sum of square remains unaffected by the change of origin. Comparing above two equations we note that,

$$B(\bar{y}_R)_T < B(\bar{y}_R),$$

when $\quad \alpha' C'_{20}<\alpha C_{20}$

$$\text{or} \quad \left(\alpha-\beta\frac{a}{b}\right)\frac{S_x^2}{\bar{X}'^2}<\alpha\frac{S_x^2}{\bar{X}^2}$$

$$\text{or} \quad \alpha(\bar{X}^2-\bar{X}'^2) < \beta\frac{a}{b}\bar{X}^2 \qquad \ldots\ldots(5)$$

This equation will hold true only when the intercept of the regression line with the Y-axis is above the origin (*i.e.* $\alpha>0$), because,

(*i*) β is always positive as X and Y are positively correlated, when ratio method is used.

(*ii*) $(\bar{X}^2 - \bar{X}'^2)$ is negative as $\bar{X}' > \bar{X}$, when a and b are positive and

(*iii*) a and b being estimate of α and β are expected to be positive.

Thus, when regression line intercept the Y-axis above the origin, it is possible to reduce the bias of the estimate of the population mean, when ratio method of estimation is used.

The mean square found under the same assumption as that of $E(\bar{y}_R)$ is given by

$$M(\bar{y}_R) = \theta (S_y^2 + R^2 S_x^2 - 2RS_{xy}). \qquad \ldots\ldots(6)$$

When X is trans-formed to X′, we have

$$M(\bar{y}_R)_T = \theta (S_y^2 + R^2_T S_x^2 - 2R_T S_{xy}), \qquad \ldots\ldots(7)$$

where $R_T = \bar{Y}/\bar{X}'$ and mean sum of square and mean sum of product remain unaffected with the change of origin. When $Y = \beta X'$ is satisfied, R_T becomes equal to β and hence

$$M(\bar{y}_R)_T = (S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}), \qquad \ldots\ldots(8)$$

which is the mean square error of the estimator of population mean, when regression method of estimation is used. Thus, when the condition mentioned above is satisfied, the mean square errors of both ratio and regression estimators becoms equal. In this case the ratio estimator becomes unbiased whereas the regression estimator still remains biased. This is one of the reasons for the use of ratio estimator in preference to regression estimator. Now, comparing equations (6) and (7), we note that $M(\bar{y}_R) > M(\bar{y}_R)_T$

when

$$S_x^2 (R^2 - R^2_T) > 2 S_{xy} (R - R_T)$$

$$\text{or} \quad (R + R_T) > 2\beta, \qquad \ldots(9)$$

when $R > R_T$. This means that mean square error of the estimate will reduce along with the bias when,

$$R > R_T \geqslant \beta. \qquad \ldots(10)$$

In all other situations, the simultaneous reduction of bias and mean square are not possible.

4. Summary

It has been shown that the bias of the estimate in case of ratio method of estimation can be reduced when intercept of the regression line with the Y-axis is above the origin. It is also possible to reduce mean square error when R_T, the ratio using the transformed auxiliary variate is greater than the regression coefficient.

Acknowledgement

We are thankful to the referee for his valuable comments to improve the paper.

REFERENCES

Mohanty, S, (1970) Contribution to theory of sampling, Ph.D. thesis submitted to I.A.R.I., New Delhi.

Murthy, M. N. (1967) Sampling theory and methods, Statistical Publishing Society, Calcutta.