STATISTICA, anno LXVII, n. 1, 2007

A NOTE ON THE USE OF AN ORIGIN SHIFT IN SURVEY SAMPLING: A CONCEPTUAL PERSPECTIVE

J. Sahoo, L.N. Sahoo, B.C. Das

1. INTRODUCTION

Let (y_i, x_i) be the values of a study variable y and an auxiliary variable x for the ith unit, $i = 1,2,...,N$, of a finite population. Suppose that our aim is an estimation of the population mean $\bar{Y} = N^{-1}(y_1 + y_2 + ... + y_N)$ when the population mean $\bar{X} = N^{-1}(x_1 + x_2 + ... + x_N)$ of x is exactly known. One of the options then is to use a linear transformation of the auxiliary variable x to increase efficiency of a sampling scheme or of an estimation method (see *e.g.*, Mohanty and Das, 1971; Reddy and Rao, 1977; Srivenkataramana and Tracy, 1980, 1986; Stuart, 1986; Montanari, 1987; Sahoo *et al.*, 1994; Mohanty and Sahoo, 1995 and others). In this paper, we suggest a simple origin shifted auxiliary variable z by clearly justifying its basis and then consider this in unequal probability sampling under (Midzuno's, 1952) scheme as well as in ratio method of estimation.

2. THE ORIGIN SHIFTED AUXILIAY VARIABLE

Assume that a sample of n units is drawn from the population by simple random sampling without replacement (SRSWOR). Then, by defining $\bar{y} = n^{-1}(y_1 + y_2 + ... + y_n)$ and $\bar{x} = n^{-1}(x_1 + x_2 + ... + x_n)$ as the sample means of y and x respectively, we have the conventional ratio estimator $t_R = \bar{y}\frac{\bar{X}}{\bar{x}}$ for estimating $\bar{Y}$. It is known that t_R is more precise than the simple expansion estimator $\bar{y}$ when

$$\rho \frac{C_y}{C_x} > \frac{1}{2}, \tag{1}$$

where ρ is the correlation coefficient between (y,x); C_y, C_x are the coefficients of variation of y and x.

Note that the inequality (1) can be made to satisfy (if not) by reducing the value of C_x (without disturbing ρ and C_y) through the use of a suitable origin shift of $x-$ variable. Then, let us define

$$z_i = x_i + d, \quad i = 1,2,...,N, \tag{2}$$

where d is a positive constant.

It can be easily verified that the ratio estimator $t_{RZ} = \bar{y}\frac{\bar{Z}}{\bar{z}}$ ($\bar{z}$ and $\bar{Z}$ are respectively the sample mean and population mean of z) is better than $\bar{y}$ when

$$\left(1+\frac{d}{\bar{X}}\right)\rho\frac{C_y}{C_x} > \frac{1}{2}. \tag{3}$$

See that the factor $\left(1+\frac{d}{\bar{X}}\right)$ in the *l.h.s.* of (3) is an incremental factor that brings (1) to its validity. For instance, if $d = k\bar{X}$, then t_{RZ} is more efficient that $\bar{y}$ when $\rho\frac{C_y}{C_x} > [2(k+1)]^{-1}$, a condition which can be met very often. If we assume $y = \alpha + \beta x \ \ (\alpha, \beta > 0)$ to obtain zero variance, then we should have

$$d_{opt} = \frac{\alpha}{\beta} = \delta \text{ (say)}. \tag{4}$$

Remark 2.1: Linear transformation of the form $z = \frac{x}{c} + d$ is frequently used in practice. See that what ever may be the value of c, the variance can never be made zero unless we use d. Hence it is perhaps wise to consider only single parameter d instead of two in the process of obtaining z.

2.1. *A geometrical interpretation*

The regression $y = \alpha + \beta x$ is shown in figure 1. Here α is the intercept being measured by OB and β is the slope given by $\tan\theta$. See that in the $\Delta\, OAB$, $\angle A = \theta$ and $\tan\theta = \frac{OB}{OA}$. Since we want the line to pass through the origin O, it is required that the point A be shifted to right by an amount OA which is being given by

$$OA = \frac{OB}{\tan\theta} = \frac{\alpha}{\beta} = \delta \text{ [as in (4)].}$$

Since OA is negative, we consider d to be positive.

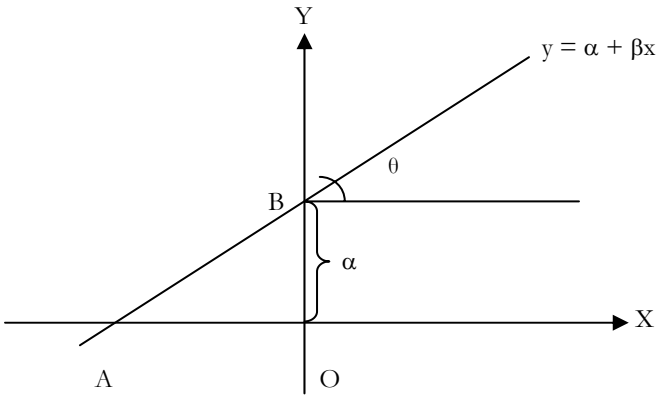

Figure 1 – Regression of y on x.

The above argument therefore encouraged us to consider z in its simpler form as in (2). In what follows, we shall use z in unequal probability sampling with (Midzuno's, 1952) scheme and suggest a retransformed variable u (obtainable from z) so that it is feasible to make this scheme a πpz scheme under revised probability of selection. The performance of Midzuno scheme is also compared with SRSWOR scheme under ratio method of estimation.

3. THE MIDZUNO SCHEME : HORVITZ-THOMPSON ESTIMATOR

We consider (Midzuno, 1952) scheme as it has the main advantages that it provides a no-negative Sen-Yates-Grundy variance estimator for the Horvitz-Thompson estimator, inclusion probabilities π_i and π_{ij} are simple to calculate, and it is possible to generate a set of revised probabilities to make the scheme a πps scheme. In the present context, the Midzuno scheme consists of drawing the first unit in the sample with probability proportional to z and the rest $(n-1)$ units by SRSWOR.

Thus, with $z-$variable the initial selection probability of unit i, inclusion probability of unit i and joint inclusion probability of units i and j, are

$$p_i = \frac{z_i}{N\bar{Z}} = \frac{x_i + d}{N(\bar{X} + d)}, \tag{5}$$

$$\pi_i = \frac{1}{N-1}\left[\frac{(N-n)x_i + N(n-1)\bar{X} + n(N-1)d}{N(\bar{X}+d)}\right], \tag{6}$$

and

$$\pi_{ij} = \frac{n-1}{(N-1)(N-2)}\left[\frac{(N-n)(x_i + x_j) + N(n-2)\bar{X} + n(N-2)d}{N(\bar{X}+d)}\right], \tag{7}$$

$i \neq j = 1,2,...,N$. Hence, after a considerable simplification we get

$$\pi_i\pi_j - \pi_{ij} = \frac{N-n}{N^2(N-1)^2(N-2)(\bar{X}+d)^2}\Delta, \tag{8}$$

where

$$\Delta = (N-n)(N-2)x\ x_j + (N-1)(N-2n)d(x_i + x_j)$$
$$+(n-1)N\bar{X}(N\bar{X} - x_i - x_j) + 2(n-1)N(N-1)d\bar{X} + n(N-1)(N-2)d^2.$$

The *r.h.s.* of (8) is always positive. This means that use of z does not handicap the Midzuno scheme of sampling. But, the main drawback is that it does not yield a πpz scheme. However, it is possible to create a set of revised probabilities so that the scheme becomes a πpz scheme. Thus we consider the revised probability p_i^* such that

$$\pi_i = \frac{N-n}{N-1}p_i^* + \frac{n-1}{N-1} = np_i, \tag{9}$$

where p_i is given by (5). This implies that

$$p_i^* = \frac{(N-1)np_i}{N-n} - \frac{n-1}{N-n}. \tag{10}$$

Since p_i^* should be always non-negative, we must have

$$p_i = \frac{x_i + d}{N(\bar{X}+d)} \geq \frac{n-1}{n(N-1)} = \eta \text{ (say).} \tag{11}$$

This is however true if

$$d \geq \frac{(n-1)N\bar{X} - n(N-1)x_{(1)}}{N-n}, \tag{12}$$

where $x_{(1)}$ is the smallest $x-$value. But since d is assumed positive, we need the *r.h.s.* of (12) to be positive. This means that we should have

$$\bar{X} > \frac{n}{n-1}\frac{N-1}{N}x_{(N)}, \tag{13}$$

where $x_{(N)}$ is the largest $x-$value.

Observe that since $\frac{n}{n-1}\frac{N-1}{N} > 1$, (13) can never be satisfied. This indicates that we cannot have all p_i's in (11) are greater than η. It is also true that all p_i's can never be less than η. Then let us assume that

$$\eta > p_{(m)} = \frac{x_{(m)} + d}{N(\bar{X} + d)} \ , m > 1, \tag{14}$$

where $p_{(m)}$ is the selection probability of the unit corresponding to value $x_{(m)}$ or $z_{(m)}$. This means that we have inequalities of the form

$$p_{(N)} > p_{(N-1)} > ... > p_{(m+1)} > \eta > p_{(m)} > ... > p_{(1)}. \tag{15}$$

To overcome such a difficulty, let us retransform z to another variable u such that

$$u_i = \frac{z_i}{\bar{Z}} + \gamma \ , \ i = 1,2,...,N \ , \tag{16}$$

where γ is a positive scalar to be chosen in such a manner that setting of revised probabilities is feasible. Thus, we have

$$U = \sum_{i=1}^{N} u_i = N(1+\gamma). \tag{17}$$

Then, let us define a new set of initial probabilities $\{p_1', p_2', ..., p_N'\}$ such that

$$p_i' = \frac{u_i}{U} = \frac{x_i + \gamma\bar{X} + d(1+\gamma)}{N(\bar{X} + d)(1+\gamma)} \ , \ i = 1,2,...,N. \tag{18}$$

Accordingly, the first and second order inclusion probabilities are obtained as

$$\pi_i' = \frac{1}{N-1}\left[\frac{(N-n)x_i + (n-1)N\bar{X} + n(N-1)(\gamma\bar{X} + \gamma d + d)}{N(\bar{X} + d)(1+\gamma)}\right] \tag{19}$$

and

$$\pi'_{ij} = \frac{n-1}{(N-1)(N-2)} \times \left[\frac{(N-n)(x_i + x_j) + (n-2)N\bar{X} + n(N-2)(\gamma\bar{X} + \gamma d + d)}{N(\bar{X} + d)(1+\gamma)} \right]. \tag{20}$$

One can verify the inequality $\pi'_i \pi'_j > \pi'_{ij}$, $i \neq j = 1,2,...,N$, a condition needed for non-negativity of Yates-Grundy variance estimator of the Horvitz-Thompson estimator. Thus, it is possible to modify the Midzuno scheme to a $\pi p\chi$ scheme by generating a set of revised probabilities through the initial probabilities of selection p'_i.

4. DETERMINATION OF d AND γ

First we determine the value of γ such that

$$p'_{(1)} = \frac{x_{(1)} + \gamma\bar{X} + d(1+\gamma)}{N(\bar{X} + d)(1+\gamma)} > \eta. \tag{21}$$

This gives

$$\gamma = \gamma_1 + \gamma_2, \tag{22}$$

where $\gamma_1 = \eta_1(\eta - p_{(1)}), \eta_1 = \frac{n(N-1)N}{N-n}$ and γ_2 is a positive quantity. See that γ_1 is a known positive quantity since $\eta > p_{(1)}$ is true according to (14).

Note that an arbitrary γ_2 (positive) value can make the scheme operative. But, if we assume $y = \alpha + \beta x$ to achieve zero variance, then we have

$$\gamma_{2(opt)} = \frac{\delta - d}{\bar{X} + d} - \gamma_1, \tag{23}$$

with $\delta = \frac{\alpha}{\beta}$. Accordingly, this yields

$$\gamma_{(opt)} = \frac{\delta - d}{\bar{X} + d}. \tag{24}$$

Now since it is desired that $\gamma_{2(opt)} > 0$, we should have

$$0 < d < \frac{\delta - \gamma_1 \bar{X}}{1 + \gamma_1}, \tag{25}$$

with $\delta > \gamma_1 \bar{X}$.

Remark 4.1: When $d = \delta$, we have $\gamma_{(opt)} = 0$. This indicates that if an optimum value of d has already been used in (2), then obviously no further transformation as in (16) is needed since in this situation $p_i' = p_i$.

Remark 4.2: (Bedi and Rao, 1996) considered an origin shifted auxiliary variable $z = x + N\bar{X}$ and studied the efficiency of the transformed ratio estimator. This is a particular case of ours when $d = N\bar{X}$.

5. THE MIDZUNO SCHEME: RATIO METHOD OF ESTIMATION

Under the Midzuno scheme considered in section 3, the probability of selecting the sample s is given by

$$p(s) = ({}^{N}C_{n})^{-1} \frac{\bar{x} + d}{\bar{X} + d}. \tag{26}$$

Thus, the ratio estimator $t_{RZ} = \bar{y} \frac{\bar{X} + d}{\bar{x} + d}$ is found to be unbiased for $\bar{Y}$ under this scheme of sampling.

Let $H_1 = (SRSWOR, t_{RZ})$ and $H_2 = (MS, t_{RZ})$ respectively be the strategies under SRSWOR and Midzuno scheme with ratio estimator t_{RZ}. By denoting $M(H_1)$ and $V(H_2)$ as the mean square error of H_1 and variance of H_2 respectively, we have to a first order of approximation [see *e.g.*, (Kendall *et al.*, 1983)]

$$M(H_1) = V(H_2) = \theta \bar{Y}^2 (C_{20} - 2d_1 C_{11} + d_1^2 C_{02}), \tag{27}$$

where $\theta = n^{-1} - N^{-1}, d_1 = \frac{\bar{X}}{d + \bar{X}}$ and $C_{ij} = \frac{K_{ij}}{\bar{Y}^i \bar{X}^j}$; K_{ij} is the (i, j) th cumulant of (y, x). Note that (27) attains its minimum value when $d = \frac{\alpha}{\beta}$, as has been established in (4).

Since, to the first order approximation, the strategies H_1 and H_2 are equally efficient, a choice among them naturally depends on the comparison of $M(H_1)$

and $V(H_2)$ for higher order approximations. Therefore, following (Tin, 1965), (Singh, 1975) and (Sahoo, 1983) among others, we now obtain the following results by considering terms up to $O(n^{-2})$:

$$M(H_1) = V(H_2) + \bar{Y}^2\left[-\left(\theta_1 - \frac{3\theta}{N}\right)d_1(C_{21} - 2d_1C_{12} + d_1^2C_{03})\right]$$
$$+2\theta_1^2\bar{Y}^2d_1^2[3(d_1C_{02} - C_{11})^2 + C_{02}C_{20}(1-\rho^2)], \tag{28}$$

where $\rho = \dfrac{C_{11}}{\sqrt{C_{20}C_{02}}}$ and $\theta_1 = n^{-2} - N^{-2}$.

Observe that $M(H_1) > V(H_2)$ when

$$C_{21} - 2d_1C_{12} + d_1^2C_{03} \le 0,$$

i.e., $Cov[x,(y - Rd_1x)^2] \le 0 \ : \ R = \dfrac{\bar{Y}}{\bar{X}}$. (29)

Further, if the joint distribution of y and x is bivariate normal, then $M(H_1) > V(H_2)$ showing thereby that H_2 is more efficient than H_1.

5.1. *Efficiency comparison under a model*

Let us consider the following regression model (M_g):

$$y_i = \alpha + \beta x_i + e_i, \ i = 1,2,...,N, \tag{30}$$

where $\alpha \ge 0, \beta > 0$ and e_i's are uncorrelated random errors with conditional (given x_i) expectations $E(e_i / x_i) = 0$, $E(e_i^2 / x_i) = \xi x_i^g \ \ \forall i$ with $0 < \xi < \infty$, $0 \le g \le 2$; $E(e_ie_j / x_i, x_j) = 0 \ \ \forall \ \ i \ne j = 1,2,...,N$ and x_i's are assumed to be *i.i.d.* gamma variates with common single parameter $h > 0$ taken equal to the known value $\bar{X}$.

By the direct substitution under the model we get

$C_{20} = \dfrac{\beta^2 h + \xi H}{\bar{Y}^2}$, $C_{11} = \dfrac{\beta}{\bar{Y}}$, $C_{02} = \dfrac{1}{h}$, $C_{21} = \dfrac{2\beta^2 h + g\xi H}{\bar{Y}^2 h}$, $C_{12} = \dfrac{2\beta}{\bar{Y}h}$ and $C_{03} = \dfrac{2}{h^2}$, where $H = \dfrac{\Gamma(g+h-1)}{\Gamma(h-1)}$. Hence, under M_g, to $O(n^{-2})$, we obtain

$$M(H_1) = A[\theta h - 4Fd_1 + 9\theta^2 d_1^2] + B[\theta h - 2Fd_1 g + 3\theta^2 d_1^2] \tag{31}$$

$$V(H_2) = A[\theta h - 2Fd_1 + 3\theta^2 d_1^2] + B[\theta h - Fd_1 g + \theta^2 d_1^2], \tag{32}$$

where $A = \left(\frac{\alpha - \beta d}{d + h}\right)^2$, $B = \frac{\xi H}{h}$, $F = \theta_1 - \frac{3\theta}{N}$ and $d_1 = \frac{h}{d + h}$.

On comparison of (31) and (32), we find, ignoring *f.p.c.*, that $M(H_1) > M(H_2)$, when

$$d_1 > \frac{1}{3} \Rightarrow h > \frac{d}{2} \text{ and } g < \frac{2}{3}.$$

This leads to the conclusion that under the assumed model, the Midzuno scheme outperforms SRSWOR when used in ratio method of estimation with t_{RZ}, provided

$$h > \frac{d}{2} \text{ and } g < \frac{2}{3}.$$

6. CONCLUSIONS

The following conclusions are readily comprehensible from the present study:

(i) The origin shifted variable z, with an optimum value of d, makes the Midzuno scheme a πpz scheme.

(ii) If d is not optimum, then z can further be retransformed to another variable u so that the Midzuno scheme is feasible with revised probability of selection to yield a πpz scheme.

(iii) The Midzuno scheme with ratio method of estimation is better than SRSWOR scheme under design as well as model based comparisons when certain conditions, as given in the text, are satisfied.

Department of Statistics JAGATANANDA SAHOO
S.K.C.G. College, Parlakhemundi 761200
India

Department of Statistics LOKANATH SAHOO
Utkal University, Bhubaneswar 751004 BISHNUCHARAN DAS
India

ACKNOWLEDGEMENT

The authors are grateful to the referee whose constructive comments led to an improvement in the paper.

REFERENCES

P.K. BEDI, T.J. RAO (1996), *Efficient utilization of auxiliary information at estimation stage*, "Biometrical Journal", 38, pp. 973-976.

M.G. KENDAL, A. STUART, J.K. ORD (1983), *The advanced theory of Statistics*, volume III, Charles Griffin and Company Limited, London and High Wycombe, pp. 242.

H. MIDZUNO (1952), *On the sampling system with probability proportionate to sum of sizes*, "Annals of the Institute of Statistical Mathematics", 3, pp. 99-107.

S. MOHANTY, M.N. DAS (1971), *Use of transformation in sampling*, "Journal of the Indian Society of Agricultural Statistics", 23, pp. 83-87.

S. MOHANTY, J. SAHOO (1995), *A note on improving the ratio method of estimation through linear transformation using certain known population parameters*, "Sankhyā", series B, 57, pp. 93-102.

G.E. MONTANARI (1987), *Variance reduction through location shifts in unequal probability sampling*, "Metron", 45, pp. 213-234.

V.N. REDDY, T.J. RAO (1977), *Modified PPS method of estimation*, "Sankhyā", series C, 39, pp. 185-197.

L.N. SAHOO (1983), *On a method of bias reduction in ratio estimation*, "Journal of Statistical Research", 17, pp. 1-6.

J. SAHOO, L.N. SAHOO, S. MOHANTY (1994), *Unequal probability sampling using a transformed auxiliary variable*, "Metron", 52, pp. 71-83.

R. SINGH (1975), *A note on the efficiency of the ratio estimate with Midzuno scheme of sampling*, "Sankhyā", series C, 37, pp. 211-214.

T. SRIVENKATARAMANA, D.C. TRACY (1980), *An alternative to ratio method in sample survey*, "Annals of the Institute of Statistical Mathematics", 32, pp. 111-120.

T. SRIVENKATARAMANA, D.C. TRACY (1986), *Transformations after sampling*, "Statistics", 17, pp. 597-608.

A. STUART (1986), *Location shifts in sampling with unequal probabilities*, "Journal of the Royal Statistical Society", series A, 149, pp. 349-365.

M. TIN (1965), *Comparison of some ratio estimators*, "Journal of the American Statistical Association", 60, pp. 294-307.

SUMMARY

A note on the use of an origin shift in survey sampling: a conceptual perspective

In this paper, a simple origin shifted auxiliary variable z has been considered when the values of the main auxiliary variable x are completely known. The usability of z has been demonstrated both analytically and geometrically. The variable z has been used in (Midzuno's, 1952) scheme with unequal probability sampling and it is further retransformed to another variable u so that the Midzuno scheme is operative under revised probability of selection. The efficiencies of Midzuno scheme and SRSWOR scheme with ratio method of estimation have also been compared both under the design and an assumed model.