A STUDY OF THE EFFECT OF VARIATE-TRANSFORMATIONS ON STRATEGIES OF SAMPLING FINITE POPULATINOS

By

ARIJIT CHAUDHURI AND ARUN KUMAR ADHIKARY[1]

Indian Statistical Institute, Calcutta

(Received : November, 1981)

SUMMARY

This is a study of the effect of variate-transformations on the 'small-sample' efficiencies of some standard strategies of sampling a finite population on postulating a 'super-population' regression model with a non-zero intercept and a gamma-distributed auxiliary variate. Exact efficiency of regression estimator being difficult to study in general a few competitors are considered; among them the one modifying the Midzuno strategy stands out as a very promising one in several situations.

INTRODUCTION

Srivenkataramana [10], following Mohanty and Das [8], recently considered a method (through variate-transformations) of improving on standard estimators (based on SRSWOR sampling scheme) for a population mean

$$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N} y_i$$

of a variate y assumed to have a linear regression (in a 'finite population' sense) on an auxiliary variate x whose values x_i's for all the units of the population $U=(1,\ldots\ldots,i,\ldots\ldots N)$ are positive and known. Here we extend this technique to the following strategies of sampling with varying probabilities adopting a 'super-population' model (detailed below) : (*i*) Midzuno [7] strategy involving ratio estimator based on his sampling scheme with selection probabilities of samples (size

1. Present Address : ORG, Baroda.

being supposed throughout for each sampling scheme we treat to be a fixed integer n) proportional to aggregates of size-measure (x_i, s), (*ii*) Horvitz-Thompson [6] estimator (HTE) based on any IPPS sampling scheme with inclusion probabilities π_i's proportional to x_i's (*i.e.* $\pi_i = np_i$ with $p_i = x_i / N\bar{X}$, where

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i = X/N),$$

(*iii*) Hansen-Hurwitz [5] strategy involving the usual estimator (HHE,) in brief) based on PPSWR sampling scheme with normed size-measures p_i's involving n draws, (*iv*) Rao-Hartely-Cochran [9] (RHC, in brief) strategy involving size measures x_i's (detailed description of this strategy is omitted to save space, but we mention that we assume that each group formed in adopting this scheme is of size N/n which is supposed to be an integer), and finally (*v*) the ratio estimator and (*vi*) the regression estimator both based on SRSWOR sampling scheme. We postulate super-population model M (say) so that we may write

$$Y_i = \alpha + \beta x_i + \gamma_i, \; i = 1, 2, \ldots N \qquad \ldots(1.1)$$

where $\alpha > 0$, $\beta \geqslant 0$ (both unknown otherwise), $\varepsilon(\gamma_i \mid x_i) = 0 \; \forall \; i$, ($\varepsilon \equiv$ operator for conditional expectation given x_i's in respect of a super-population from which the given finite population is supposed to be a random sample). Also we suppose $\varepsilon(\gamma_i^2 / x_i) = \delta x_i^g \; \forall i$, $\varepsilon(\gamma_i \gamma_j / x_i, x_j) = 0 \;\; i \neq j$ with $0 < \delta < \infty$ and $0 \leqslant g \leqslant 2$. Further, we assume the x_i's are positive values on random variables (also denoted as x_i's) each distributed independently and identically as a gamma variate with a single parameter namely the mean m (supposed to exceed 2) which we have [on the strength, if needed for *logicality*, of the law of large numbers which may be supposed to be applicable provided we are ready to assume N to be large, as we are, following Chakrabarti [3] as equal to $\bar{X}$ which is known for the given finite population (the corresponding expectation operator is denoted as ε_x for x standing for x_i's, $i = 1, \ldots N$). By E we mean generically the operator for expectation over sampling design for which $(p)s$ will generically mean selection probability of a samples (typical) according any of the sampling schemes we are studying here. The overall two and three-step expectation operators will be denoted as $\bar{\varepsilon} = \varepsilon_x \, \varepsilon$ and $\bar{e} = \varepsilon E = \varepsilon_x \, \varepsilon E$.

By t_i and t_i' $(i=1, \ldots, 5)$ we shall denote the standard estimators (for the first five situations mentioned above) based on the respective sampling schemes mentioned above (the respective strategies being denoted as D_i, D_i') and their modifications through variate-transformations (in fact, translations) which are respectively

$$t_1=t_1(\underline{y})=\frac{\bar{y}}{\bar{x}}, \bar{X}, \qquad t_1'=t_1(\underline{x})+\Theta$$

$$t_2=t_2(\underline{y})=\frac{1}{N}\sum_{i\in S}\frac{y_i}{\pi_i} \qquad t_2'=t_2(\underline{x})+\Theta$$

$$t_3=t_3(\underline{y})=\frac{1}{N}\cdot\frac{1}{n}\sum_{\gamma=1}^{n}\frac{y_r'}{p_r'},\quad t_3'=t_3(\underline{z})+\Theta$$

$$t_4=t_4(\underline{y})=\frac{1}{N}\sum Y_i\frac{p_i}{p_i'},\quad t_4'=t_4(\underline{z})+\Theta$$

$$t_5=t_5(\underline{y})=\frac{\bar{y}}{\bar{x}}\bar{X}, \qquad t_5'=t_5(\underline{z})+\Theta$$

writing $\bar{y}$, $\bar{x}$, $\bar{z}$ for sample means and $\underline{r}=(r_1,\ldots, r_i,\ldots, r_N)$ for $r=x, y, z$, $z_i=y_i-\Theta$, y_r' is the value of y_i for the unit chosen on the rth draw and p_r' the corresponding value of p_i. Here Θ is supposed to be a quantity of the *sampler's* choice. The regression estimator based on SRSWOR is

$$t_R=\bar{y}+b\,(m-\bar{x}), \text{ where } b=\frac{\sum_{1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{1}^{n}(x_s-\bar{x})^2}.$$

For varying probability sampling schemes, generalised regression estimators are available in the literature [vide Cassel, Särndal and Wretman [1] but we will not treat them here to avoid complicated formulae.

2. Study of Efficiencies of the Various Strategies

In order to study the relative efficiencies of the strategies we need compare the value of $\varepsilon\, E(e-\bar{y})^2$ for different e's, each standing for one of the estimators above.

For any design-unbiased homogeneous linear estimator $t=t(\underline{y})$ satisfying the condition $t(\underline{x})=\bar{x}$, it may be checked that for the above super-population model (in fact for a more general one with the common distribution of x_i's being of any arbitrary form) if we considered the competing estimator $t'(\underline{y})=t(\underline{z})+\ominus$, then it follows (as one may readily check) that

$$\bar{\varepsilon}\, E\,(t'-\bar{Y})^2<\bar{\varepsilon}\, E(t-\bar{Y})^2 \text{ if } 0<\ominus<2\alpha \qquad \ldots(1.2)$$

The estimators t and t' for the above noted strategies are of this form and hence this result applies to them, giving a rule to choose among t_i and t_i' $(i=1,\ldots 5)$ if we get, on calculations (details omitted to save space),

$$V_1=\frac{(\alpha-\ominus)^2}{(nm-1)}+\delta\left[m\ \frac{\overline{|m+g}}{\overline{|m}}\ \frac{1}{(nm+g-1)} - \frac{1}{N}\,\frac{\overline{|m+g}}{\overline{|m}}\right]$$

In calculating V_2 we neglect the term

$$(\alpha-\ominus)^2\sum_{i\neq j}\sum\left(\frac{\pi_{ij}}{\pi_i\,\pi_j}-1\right)$$

and thus get a conservative expression on assuming $\pi_{ij}<\pi_i\pi_j$ $\forall\ i, j$': otherwise no useful formula for V_2 is available.

$$V_2=(\alpha-\ominus)^2\left[\frac{m}{n(m-1)}-\frac{1}{N}\right]$$

$$+\delta\left[\frac{m}{n}\cdot\ \frac{\overline{|m+g-1}}{\overline{|m}}-\frac{1}{N}\cdot\ \frac{\overline{|m+g}}{\overline{|m}}\right]$$

$$V_3=\frac{(\alpha-\ominus)^2}{n(m-1)}+\delta\left[\frac{m}{N}\cdot\ \frac{\overline{|m\ \ g-1}}{\overline{|m}}-\frac{1}{n}\cdot\ \frac{1}{N}\,\frac{\overline{|m+g}}{\overline{|m}}\right]$$

$$V_4=\frac{N-n}{N-1}\cdot\ V_3$$

$$V_5=(\alpha-\ominus)^2\cdot\ \frac{(nm-2)}{(nm-1)\,(nm-2)}$$

$$+\delta\left[\frac{m^2n\ \overline{|m+g}}{\overline{|m}}\cdot\ \frac{1}{(nm+g-1)(nm+g-2)}\right.$$

$$\left.+\frac{1}{N}\,\frac{\overline{|m+g}}{\overline{|m}}-\frac{2mn}{N}\,\frac{\overline{|m+g}}{\overline{|m}}\cdot\ \frac{1}{(nm+g-1)}\right]$$

Unfortunately such an exact expression in a closed form is not available for

$$\bar{\varepsilon}\, E\,(t_R - \bar{Y})^2 = V_R \quad \text{(say)}$$

In fact, we have

$$V_R = \delta\left(\frac{1}{N} - \frac{1}{N}\right)\frac{\Gamma mg}{\Gamma m_i}$$

$$+\,\delta\, E\,\varepsilon_x\left[(m-\bar{x})^2 \cdot \frac{\sum_1^n x_i^g\,(x_i-\bar{x})^2}{\left\{\sum_1^n (x_i-\bar{x})^2\right\}^2}\right.$$

$$+2\left(\frac{1}{n} - \frac{1}{N}\right)(m-\bar{x})\,\frac{\sum_1^n x_i^g\,(x_i-\bar{x})}{\sum_1^n (x_i-\bar{x})^2} \text{ for } g \neq 0.$$

$$= \delta\, E\,\varepsilon_x\left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{(m-\bar{x})^2}{\sum_1^n (x_i-\bar{x})^2}\right] \text{for } g=0 \qquad \ldots(2.1)$$

Further simplifications are obviously difficult to achieve and hence it is not easy to study the exact efficiency of t_R under the present model. If, however, we use the usual asymptotic formula namely

$$E(t_R-\bar{Y})^2 \simeq \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2\,(1-\rho^2)$$

$$\left[\text{ with } \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2,\ \sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2,\right.$$

$$\left.\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})\,(x_i-\bar{X}) \text{ and } \rho = \sigma_{xy}/\sigma_x\,\sigma_y\right] \text{ available}$$

for n and N large then we get [on algebraic manipulations, with details omitted]

$$\xi E(t_R-\bar{Y})^2 \simeq \frac{N-2}{N-1}\;\frac{N-n}{Nn}\delta=\bar{V}_R \text{ (say), if } g=0$$

[general formula for $g \neq 0$ is too complicated for presentation].

On further calculations we may observe the following results in particular

(i) $V_5 > V_1 > \bar{V}_R > 0$, if $g=0$

(ii) $V_2 > V_1$, if $g < 1$,

(iii) $V_3 > V_1$ if $g < 1$,

(iv) $V_3 > V_4 \forall g$

(v) $V_5 > V_1$, if $1 < g < 2$

(vi) $V_4 > V_1$ if $nm < N(1-g)$ when $0 < g < 1$

(vii) $V_2 > V_3$ and $V_2 > V_4$; if terms $0(1/N)$ are neglected

(viii) $V_2 > V_5$, if terms $0(1/N)$ are neglected

(ix) $V_5 < V_3$, if $n \geqslant 5$ and if $g < g_0$ where g_0 is a root in $[0, 2]$ of $g^2-(n^2m-2nm+3)\,g+(n^2m-3nm+2)=0$

3. Numerical Values of Relative Efficiencies of Strategies

Defining efficiencies of the strategies as $E_i = 100\dfrac{V_2}{V_i}$ for $i=1, \ldots, 6$ and writing $V_6=\bar{V}_R$ we present below the values of relative efficiencies of these strategies for a few combinations of the parametric values under the model M considered above.

We consider the following cases respectively denoted as $I-VI$ in Table 3.1 below and present the values of E_i for $\theta=0.1$, 0.5, 0.8 in the order from top to bottom in cases $I-IV$ and for $\theta=1.1$, 1.5, 1.8 for case V and $\theta=.5$, 1.5 and 2.0 in Case VI.

TABEL 3.1

Giving Relative Efficiencies of Strategies

	E_1	E_2	E_3	E_4	E_5	E_6
Case I						
$g=0.0$	159	100	91	115	137	190
	152	100	86	109	131	176
	156	100	89	112	134	184
$g=0.5$	126	100	86	109	112	
	121	100	83	105	108	
	124	100	84	107	110	
$g=1.0$	103	100	81	103	94	
	100	100	79	100	91	
	102	100	80	102	93	
$g=1.5$	86	100	77	97	81	
	85	100	75	95	79	
	86	100	76	96	80	
$g=2.0$	74	100	72	99	70	
	73	100	71	90	69	
	73	100	72	91	70	
$\alpha=0.5$ $\delta=2.0$ $m=3$ $n=5$ $N=20$						
Case II						
$g=0.0$	134	100	88	112	123	152
	127	100	13	106	116	141
	130	100	86	109	120	147
$g=0.5$	115	100	84	106	108	
	112	100	81	103	104	
	114	100	83	105	106	
$g=1.0$	102	100	80	102	96	
	100	100	79	100	95	
	101	100	80	101	96	
$g=1.5$	91	100	77	98	87	
	90	100	77	97	87	
	91	100	77	98	87	
$g=2.0$	82	100	75	95	80	
	82	100	74	94	80	
	82	100	75	95	80	
$\alpha=0.5$ $\delta=2.0$ $m=5$ $n=3$ $N=20$						

	E_1	E_2	E_3	E_4	E_5	E_8
Case III						
$g=0.0$	130	100	102	108	121	138
	122	100	95	101	114	128
	126	100	99	105	118	134
$g=0.5$	114	100	98	104	108	
	110	100	95	101	104	
	112	106	97	102	106	
$g=1.0$	102	100	96	102	98	
	100	100	94	100	96	
	101	100	95	101	97	
$g=1.5$	93	100	95	100	91	
	92	100	94	99	90	
	92	100	94	100	90	
$g=2.0$	85	100	94	99	85	
	85	100	93	99	85	
	85	100	93	99	85	

$\alpha=0.5$ $\delta=2.0$ $m=5$ $n=6$ $N=90$

	E_1	E_2	E_3	E_4	E_5	E_8
Case IV						
$g=0.0$	119	100	104	107	111	125
	111	100	97	100	104	116
	115	100	101	104	108	121
$g=0.5$	108	100	100	103	103	
	105	100	97	100	100	
	107	100	99	102	102	
$g=1.0$	101	100	99	101	98	
	100	100	97	100	97	
	101	100	98	101	97	
$g=1.5$	96	100	97	100	94	
	95	100	97	100	94	
	96	100	97	100	94	
$g=2.0$	91	100	97	100	94	
	91	100	97	100	94	
	91	100	97	100	91	

$\alpha=0.5$ $\delta=2.0$ $m=8$ $n=4$ $N=100$

	E_1	E_2	E_3	E_4	E_5	E_6
Case V						
$g=0.0$	119	100	100	108	112	128
	112	100	93	101	105	119
	116	100	97	105	109	124
$g=0.5$	108	100	95	103	103	—
	106	100	93	101	100	
	107	100	94	102	102	
$g=1.0$	101	100	93	101	97	
	100	100	92	100	97	
	101	100	93	101	97	
$g=1.5$	95	100	92	100	93	
	95	100	92	99	93	
	95	100	92	100	103	
$g=2.0$	91	100	91	99	90	
	91	100	91	99	90	
	91	160	91	99	90	

$\alpha=1.5$ $\delta=2.0$ $m=8$ $n=4$ $N=40$

	E_1	E_2	E_3	E_4	E_5	E_6
Case VI						
$g=0.0$	156	100	137	142	146	174
	111	100	97	100	104	116
	123	100	108	111	115	131
$g=0.5$	123	100	113	117	117	
	105	100	97	100	100	
	110	100	101	104	104	
$g=1.0$	106	100	103	100	103	
	100	100	97	100	97	
	102	100	98	102	98	
$g=1.5$	98	100	99	102	96	
	95	100	97	100	94	
	96	100	97	100	94	
$g=2.0$	92	100	97	100	92	
	91	100	97	100	91	
	91	100	97	100	91	

$\alpha=1.5$ $\delta=2.0$ $m=8$ $n=4$ $N=100$

Concluding remarks

The numerical values presented in Table 3 1 conform to the algebraic results derived in section 3. We find that for $g=0$, the regression estimator, of course, is the most efficient even for small n and N, but we considered only an asymptotic variance formula for this estimator but exact ones for the rest-(obviously not a fair approach). For $g \neq 0$, the regression estimator is not considered as its asymptotic or exact variance formula is not available and in this case the strategy D_1' fares best in case $g<1$ and D_2' fares best in case $g<1$ and D_1', D_2' and D_4' are equivalent if $g=1$ and $\theta=\alpha$, otherwise D_4' fares mid-way between D_1' and D_2' [consistently with Chaudhuri-Arnab (1979) in case when no variate-translation is made]. Interestingly, D_5' fares poorer compared to D_1' even if N is large compared to n.

ACKNOWLEDGEMENT

The authors are grateful to the referee for his recommendation to furnish results for an empirical study as given in the Appendix below and for his valuable remarks helpful in revising an earlier draft.

REFERENCES

[1] Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976) : Some results on generalised difference estimation and generalised regression estimation for finite populations. *Biometrika*, 63, 615-620.

[2] Chaudhuri, Arijit and Arnab, Raghunath (1979) : On the relative efficiencies of sampling strategies under a superpopulation model. *Sankhyā*, **C**, **41**, 40-43.

[3] Chakrabarti, R.P. (1973) : A note on the small sample theory of the ratio estimator in certain specified populations. *J. Ind. Soc. Agri. Statist.* **26**, 49-57.

[4] Cocharan, W.G. (1977) : Sampling Techniques. John Wiley and Sons, 325.

[5] Hansen, M.H. and Hurwitz, W.N. (1943) : On the theory of sampling from finite populations. *Ann. Math. Statist.* **14**, 333-362.

[6] Horvitz, D.G. and Thompson, D.J. (1952) : A generalisation of sampling without replacement from finite universes. *J. Amer. Static Assoc.* **47**, 663-685.

[7] Midzuno, H. (1952) : On the sampling system with probability proportional to sums of sizes. *Ann. Inst. Statist. Math.* **3**, 99-107.

[8] Mohanty, S. and Das, M.N. (1971) : Use of transformation in sampling. *J. Ind. Soc. Agri. Statist.* **23**, 83-87.

[9] Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). : On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc.* Ser. B. **24**, 482-491.

[10] Srivenketaramana, T, (1978). : Change of origin and scale in ratio and difference methods of estimation in sampling. *Canad. J. Statist.* **6**, 79-86.

EMPIRICAL STUDY

In the above study we have compared the relative efficiencies of the strategies for a few combinations of the parametric values under the model M. Now to compare the efficiencies of the strategies for some actual population we consider the population considered by Cochran [4], (p. 325) which consists of the number of persons in a block, y, and number of rooms in a block, x, in 10 blocks. Writing $V_i' = E(t_i' - \bar{Y})^2$, $i=1, \ldots, 5$, for the variances of the estimators t_i' we present below in Table I the values of V_i' for a few selected values of θ viz. $\theta=12, 22, 32, 42$ and 52 to see how the transformed estimators behave over the corresponding original ones; we also consider the situation $\theta=0$ which corresponds to the case when no variate-transformation is made. In each case we take $n=2$ and for this n, the 45 possible samples were listed and the variances of the estimators t_1', t_5' and t_R (with $m=\bar{X}$) were computed from first principles, to avoid approximations. To calculate the variance of the estimator t_2' we have used the Midzuno [7], sampling scheme so modified as to give an *IPPS* sampling scheme because fortunately for the above population all the normed size measures satisfy the requirement for applying the modified Midzuno sampling scheme. We have also computed the variance of the regression estimator t_R by using the well-known asymptotic variance formula due to Cochran [4]. But this asymptotic formula underestimates the actual variance of the regression estimator substantially for small n which is clear from the last column of Table 1 in which the figure within the parenthesis gives the asymptotic variance of t_R which is much smaller than the actual variance (denoted by V_6') of t_R computed from first principles. Incidentally we note the variance of t_R remains invariant under variate-transformation.

Defining efficiencies of the strategies as $E_i' = 100\dfrac{V_2'}{V_i'}$ for $i=1,\ldots, 6$ we present below in table 2 the values of the relative efficiencies of the strategies for the selected values of θ.

TABLE 1

Variances of the sampling strategies for various value of θ

θ	V_1'	V_2'	V_3'	V_4'	V_5'	V_6'
0	61.32	63.74	71.16	63.25	63.06	2639.21 (54.92)
12	57.49	59.58	66.47	59.08	58.92	,,
22	55.89	57.63	64.31	57.17	57.09	,,
32	55.74	57.06	63.76	56.67	56.71	,,
42	57.04	57.88	64.79	57.59	57.80	,,
52	59.79	60.07	67.42	59.93	60.36	,,

TABLE 2

Relative efficiencies (E_i') of the sampling strategies for various values of θ

θ	E_1'	E_2'	E_3'	E_4'	E_5'	E_6'
0	103.95	100	89.57	100.77	101.08	2.42(116.06)
12	103,64	100	89.63	100.85	101.12	2.26(108.48)
22	103.11	100	89.61	100.80	100.96	2.18(104.93)
32	102.37	100	89.49	100.62	100.62	2.16(103.90)
42	101.47	100	89.33	100.50	100.13	2.19(105.39)
52	100.47	100	89.10	100.23	99.52	2.28(109.38)

This empirical study indicates that the strategy D_1' fares best and D_4' fares mid-way between D_1' and D_2' which is quite consistent with our theoretical findings under the model M in case $g<1$. Interestingly, here also D_5' fares poorer eompared to D^1 as we noted earlier.

To have an idea about the gain in efficiencies of the transformed strategies over the corresponding original ones we present below in Table 3 the values of the relative efficiencies of the transformed strategies defind as $E_i' = \frac{\text{Var }(t_i)}{\text{Var }(t_i')}$ 100, $i=1,\dots,5$, for various values of θ.

TABLE 3

Relative efficiencies (E_i^*) of the transformed strategies for various values of θ

θ	E_1^*	E_2^*	E_3^*	E_4^*	E_5^*
0	100	100	100	100	100
12	106.66	106.98	107.06	107.06	107.03
22	109.72	110.60	110.65	110.63	110.48
32	110.01	111.71	111.61	111.61	111.20
42	107.50	110.12	109.83	109.83	109.83
52	102.56	106.11	105.55	105.54	104.47

Among the values of θ considered above we find that the value $\theta=32$ leads to a higher gain in efficiency in each case compared to the other values of θ.