USE OF MULTIVARIATE AUXILIARY INFORMATION IN SELECTION OF UNITS IN PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SAMPLING

By

S.K. Agarwal,

University of Jodhpur, Jodhpur

AND

M. Singh,

I.A.S.R.I., New Delhi

(Received : June, 1977)

1. Introduction

The use of auxiliary information in obtaining the sampling strategies, for population total/mean with increased efficiency may be made in the following basic ways (Tripathi, [10]).

(*i*) in stratification (construction of strata boundaries)

(*ii*) in selecting the sample (PPS Sampling)

(*iii*) in estimating the poplation total/mean (ratio, regression and product methods of estimation)

Whenever the information on more than one auxiliary variate is available, the efficiency of the estimator can be further increased. The use of multivariate information in constructing the estimators for the population total (mean) of a variate y has been considered, among others, by Olkin [6], Des Raj [2], Srivastava [9] and Tripathi [10]. Maiti and Tripathi [8] considered the use of two auxiliary characters in constructing the set of probabilities in case of PPSWR sampling and obtained the conditions under which the sampling strategy proposed by them may be better than sample mean in SYSWR and usual unbiased estimator in PPSWR sampling based on a single character. However, the information on more than one auxiliary character may be available which could be used in selecting the units with unequal probabilities. In this paper a criterion, using information on several auxiliary chatacters, has been suggested for selecting the units with unequal probability sampling and with replacement.

2. Formation of Selection Index

Suppose the information is available on p auxiliary variables for every unit of the population. It is desired to determine the selection probabilities in terms of the values of auxiliary variates by builping up an index based on them. This would be analogous to the case of a single auxiliary variate as a measure of size for determining the selection probabilities. Let

x_{ij} be the value of the j—th auxiliary variate for the i—th unit ($j=1,\ldots\ldots p$; $i=1,\ldots\ldots N$)

y_i be the corrresponding value for the character under study.

A linear function of the auxiliary variates may serve the purpose of an index. In single auxiliary variate case the auxiliary variates used for selection is taken to be the one highly correlated with the variable under study. Since the efficiency of the pps sampling, upto great extent depends upon the correlation between the main character and auxiliary character, in index which has the maximum correlation with character under study would be desirable as a measure of sizes.

Let the index (I) be

$$I=\underline{a}'\ \underline{X} \qquad \ldots(2.1)$$

where $\underline{a}'=(a_1, a_2,\ldots\ldots, a_p)$

$\underline{X}'=(x_1, x_2,\ldots\ldots,x_p)$

The theory of regression analysis reveals that the simple choice of $\underline{a}$ such that $\underline{a}'\ \underline{X}$ is the optimum (in the sense that $\underline{a}'\ \underline{X}$ has maximum correlation with y and the sum of squares of deviation of $\underline{a}'\ \underline{X}$ from y is minimum) is, given by

$$\underline{a}=\underset{\sim}{\Sigma}^{-1}\ \underline{\sigma}$$

where $\underset{\sim}{\Sigma}^{-1}$ is the inverse of a $p\times p$ matrix $\underset{\sim}{\Sigma}$ of variances and covariances for the auxilary variates and $\underline{\sigma}$ is a $p\times 1$ vector of covariances of auxiliary variates with the main character and are given by

$$\underset{\sim}{\Sigma}=\begin{Bmatrix} V_{11} & V_{12}\ldots\ldots V_{1p} \\ \vdots & \\ V_{p1} & V_{p2}\ldots\ldots V_{pp} \end{Bmatrix}$$

$$\underline{\sigma}=(V_{01}, V_{02},\ldots\ldots\ldots,V_{0p})'$$

$V_{ij}=\text{Cov}\ (x_i, x_j)$

$V_{oj}=\text{Cov}\ (y, x_j)$

$i, j=1, 2,\ldots\ldots\ldots p.$

The index (I) so proposed involves the knowledge of $\underset{\sim}{\Sigma}$ and $\underline{\sigma}$. Since the information on p-auxiliary variates is assumed to be known for every unit of the population, therefore, $\underset{\sim}{\Sigma}$ can be known in advance. However, the assumption regarding the knowledge of $\underline{\sigma}$ is very restrictive, which involves variance of y and correlation coefficients of p-auxiliary variates with character under study. Hence a new criterion of constructing the selection index (I^*) has been proposed which does not require the knowledge of $\underline{\sigma}$ as such.

Let the index I^* be

$$I^* = \underline{C}' \underline{X} \qquad \ldots(2.2)$$

where $\underline{C} = (c_1, c_2, \ldots\ldots, c_p)'$

Here C is to be determined so that I^* has maximum correlation with y. In order to avoide the knowledge of variance of y and also that of correlation coefficients between y and x's, it has been assumed that I^* has unit variance. Thus the problem of determining I^* reduces to that of determining vector $\underline{C}$ such that covariance between y and I^* is maximum subject to the condition that

$$\underline{C}' \underset{\sim}{\Sigma} \underline{C} = 1 \qquad \ldots(2.3)$$

The maximisation of covariance between y and I^*, that is Cov $(y, I^*) = \underline{C}' \underline{\sigma}$ subject to (2.3) using the Lagrangian multiplier leads to the following choice of vector $\underline{C}$.

$$\underline{C} = (\underset{\sim}{\Sigma}^{-1} \underline{\psi}) / (\underline{\psi}' \underset{\sim}{\Sigma}^{-1} \underline{\psi})^{\frac{1}{2}} \qquad \ldots(2.4)$$

Where $\underline{\psi} = (\psi_j)$; $p \times 1$

$$\psi_j = (\rho_{oj}/\rho_{op})(\sigma_j/\sigma_p), \quad j = 1, 2, \ldots\ldots p.$$

ρ_{oj} is the correlation coefficient between y and x_j and σ_j^2 is the variance of x_j, $j = 1, 2, \ldots p$.

From (2.4) it can be seen that in C neither the knowledge of variance of y nor the knowledge of correlation coefficients of y and x's as such are required. However the ratios of correlation coefficients of the form $\left(\frac{\rho_{oj}}{\rho_{op}};\ j = 1, 2, \ldots\ldots p-1\right)$ which can assume a fairly stable magnitude over time, are required.

The case of two auxiliary characters : In this case the Index I^* will assume the following form.

$$I^* = C_1 x_1 + C_2 x_2$$

where C_1 and C_2 are given by

$$C_1=\frac{\theta-\rho_{12}}{[(1-\rho_{12}^2)(\theta^2-2\theta\rho_{12}+1)V_{11}]^{\frac{1}{2}}}$$

and $$C_2=\frac{1-\theta\rho_{12}}{[(1-\rho_{12}^2)(\theta^2-2\theta\rho_{12}+1)V_{22}]^{\frac{1}{2}}}$$

where $\theta=\frac{\rho_{01}}{\rho_{02}}$; ρ_{12} is correlation coefficient between x_1 and x_2.

Now, the index I^* involves θ, the exact value of which will not usually be available. However, in the cases of many repeated surveys where the information is collected on same variates, it will be possible to predict or to have an accurate guess of θ. Similar problem regarding the constancy of several population parameters like θ have already been studied empirically by Reddy [7]. Therefore, for certain populations the prior knowledge of θ can be assumed.

3. Efficiency Comparison

In this section the comparison of efficiency of the estimator for population total, in which the selection probabilities are based on I^* with that of usual ppswr estimator where selection probabilities are based on single auxiliary variate has been made, Quite often the practice is made to take a super population model Des Raj [1] of the following type.

$$y_i=\beta I_i^* + e_i \qquad \ldots(3.1)$$

$$E(e_i/I_i^*)=0;\ \text{Var}(e_i/I_i^*)=\nu I_i^{*g}$$

$$E(e_i\, e_j' I_i^*, I_j^*)=0;\ \nu>0 \text{ and } g\geqslant 0$$

The model (3.1) can be regarded to be reasonable for the study variate y because

(i) The correlation coefficient between y and I^* is always greater than $\rho_{oj}\forall j=1,2,\ldots\ldots p$

(ii) Under certain conditions, the absolute value of intercept (α_{I^*}) of y for $I^*=0$ is less than that of the absolute value of intercept (α_{x_t}) of y for

$$x_j=0\forall j=1,2,\ldots\ldots p$$

To derive the conditions under which

$|\alpha_{I^*}| \leqslant |\alpha_{xj}| \quad \forall j=1,2,\ldots\ldots p$, we assume,

without loss of generality, that α_{x_1} is the smallest of

α_{xj}'s $\forall j=1,2,\ldots\ldots p$.

The intercept on the y-axis made by the line of aggression of y on I^* is given by

$$\alpha_{I^*}=(\bar{Y}-V_{OI^*}\,\bar{I}^{\alpha}); \quad [\because V(I^*)=1]$$

where $V_{OI^*}=\mathrm{Cov}(y, I^*)$ and $\bar{I}^*$ is population mean for I^*

and $$\alpha_{x_1}=\bar{Y}-\frac{V_{01}}{V_{11}}\bar{X}_1$$

where $\bar{Y}$ and $\bar{X}_1$ are populations means for y and x_1 respectively.

Here we are concerned with the absolute magnitude of the intercept and will be interested in the smallest one. Hence according to the sign of α_{I^*} and α_{xj}, the following four cases can be discussed.

(i) both the intercepts positive

(ii) both the intercept negative.

(iii) α_{I^*} negative and α_{x_1} positive.

(iv) α_{I^*} positive and α_{x_1} negative.

(i) Both the intercepts positive

in this case,

$$|\alpha_{I^*}| \leqslant |\alpha_{x_1}|$$

or, $$(\underline{\sigma}' \underset{\sim}{\overset{-1}{\Sigma}} \underline{\sigma})^{\frac{1}{2}} I^* \geqslant \rho_{01}\sigma_{00}\bar{X}_1\sigma_{11}^{-1}$$

or, $$\underline{\sigma}' \overset{-1}{\Sigma}\, \underline{\bar{X}} \geqslant \rho_{01}\sigma_{00}\bar{X}_1/\sigma_{11}.$$

or, $$(\rho_{01}\rho_{12}-\rho_{02})\,(\rho_{12}-{}^{C}x_1/{}^{C}x_2)\geqslant 0$$

where ${}^{C}x_j$ is the coefficient of variation of x_j; $j=1,2$, and σ_{00}^2 being the variance of y. For the above condition to hold good

either $\rho_{12}>\max(\rho_{02}/\rho_{01},\ {}^{C}x_1/{}^{C}x_2)$

or $\quad <\min(\rho_{02}/\rho_{01},\ {}^{C}x_1/{}^{C}x_2)$

(ii) Both the intercepts negative

Proceedings on the lines of case (i) the conditions turn out to be

$$\frac{C_{X_1}}{C_{X_2}} > \rho_{12} > \frac{\rho_{02}}{\rho_{01}}$$

or $$\frac{\rho_{02}}{\rho_{01}} > \rho_{12} > \frac{C_{X_1}}{C_{X_2}}$$

Similarly the conditions can be obtained for other two cases.

For most of the natural population which are likely to be met in the practice, it has been found out empirically that

$$| \alpha_{j*} | \leqslant | \alpha_{xj} |$$

The variance of usual pps estimator

$$\hat{Y}_1 \left(= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_{i1}} \right)$$

for population total using

$$p_{i1} \left[= \frac{x_{i1}}{X_1} ; X_1 = \sum_{i=1}^{N} x_{i1} \right]$$

as the selection probabilities is given by

$$V_1(\hat{Y}_1) = \frac{1}{N} \left(\sum_{i=1}^{N} \frac{y_i^2}{p_{i1}} - Y^2 \right) \qquad \ldots(3.2)$$

whenever the selection probabilities are based on auxiliary variate x_{i2} ; variance of

$$\hat{Y}_2 \left(= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_{i2}} \right)$$

is given by

$$V_2(\hat{Y}_2) = \frac{1}{n} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_{i2}} - Y^2 \right) \qquad \ldots3.3)$$

where $$p_{i2} = \frac{x_{i2}}{X_2} ; X_2 = \sum_{i=1}^{N} x_{i2}$$

The variance of the estimator

$$\dot{Y}_{I^*}\left(=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_{I_i^*}}\right)$$

of y for pps wr sampling using $p_{I_{i^*}}$ as the selection probabilities is given by

$$V_3\ (\hat{Y}_I{}^*)=\frac{1}{n}\left(\sum^{N}\frac{y_i^2}{p_{I_i^*}}-Y^2\right) \qquad \ldots (3.4)$$

The expected value of variances (3.5), (3.2), (3.3) and (3.4) under model (3.1) are

$EV_1\ (\hat{Y}_1)$

$$=V_1=\frac{1}{n}\left[\beta^2\left\{\sum_{i=1}^{N}\frac{I_i^*}{p_{I_i^*}}-I^{*2}\right\}+\nu\left(\sum_{i=I}^{N}\frac{I_i^{*g}}{p_{i1}}-\sum_{i=I}^{N}I^{*g}\right)\right] \qquad \ldots (3.5)$$

$EV_2(\hat{Y}_2)$

$$=V_2=\frac{1}{n}\left[\beta^2\left\{\sum_{i=1}^{N}\frac{I_i^*}{p_{I_i^*}}-I^{*2}\right\}+\nu\left(\sum_{i=1}^{N}\frac{I_i^{*g}}{p_{i2}}-\sum_{i=1}^{N}I^{*g}\right)\right] \qquad \ldots (3.6)$$

$$EV_3\ (\hat{Y}_{I^*})=V_3=\frac{\nu}{n}\left[\sum_{i=1}^{N}I_i^{*\,g-1}\ (I_i^*-I_1^*)\right] \qquad \ldots (3.7)$$

substracting (3.7) from (3.5) and from (3.6)

$$V_1-V_3=\frac{1}{n}\left[\beta^2\left\{\sum_{i=1}^{N}\frac{I_i^{*2}}{p_{i1}}-I^{*2}\right\}+\nu\left\{\left(\sum_{i=1}^{N}\frac{I_i^{*g}}{p_{i1}}-I^{*2}\right)\right\}\right] \ldots (3.8)$$

$$V_2-V_3=\frac{1}{n}\left[\beta^2\left\{\sum_{i=1}^{N}\frac{I_i^{*^2}}{p_{i2}}-I^{*2}\right\}+\nu\left\{\sum_{i=1}^{N}\frac{I_i^{*^2}}{p_{i2}}-I^{*2}\right\}\right] \ldots (3.9)$$

from (3.8) and (3.9) it is clear that the expected efficiency depends on the values of g, p_{i1} and p_{i2}. However, for $g=2$;

$$V_1 - V_3 = \frac{1}{n}(\beta^2+\nu) I^{*2} \left[\sum_{i=1}^{N} \frac{p_{I_i^*}}{p_{i1}} - 1 \right] > 0$$

$$V_2 - V_3 = \frac{1}{n}(\beta^2+\nu) I^{*2} \left[\sum_{i=1}^{N} \frac{p_{I_i^*}}{p_{i2}} - 1 \right] > 0$$

This shows that the selection scheme based on I^* leads to estimator less expected value of the variance under super-population model with $g=2$. It is worthwhile to point out that this comparison much depends on the model (3.1). If instead of I_i^* ; we consider x_{i1} or x_{i2} in the model, the conclusion will get reversed. However, the justification of the model has already been given. For $1 < g < 2$, nothing can be said. The comparison of

$$\hat{Y}_{I^*} \text{ with } \hat{Y}\left(= \frac{N}{n} \sum_{i=1}^{n} y_i \right)$$

under model (3.1) is well known hence omitted here.

4. Empirical Illustrations

To illustrate the idea presented above some natural populations are chosen. The characteristics of these populations are given in Table 1.

Table 2 gives the characteristics of the populations under study for the Index I* constructed using the auxiliary characters x_1 and x_2 alongwith its efficiency when I* is used as size measures. It shows that I* has not only higher correlation with y but also smaller intercept. Therefore, I*, if used as auxiliary variate in place of x_1 or x_2, will always lead to higher efficiency. The performance of $\hat{Y}_2$ is poor than that of $N\bar{\bar{Y}}$ for the populations no. 2,3 and 4 due to either ρ_{02} is negative or very small or $|\alpha_{x_2}|$ is quite large as shown in Table 1.

TABLE 1

The characteristics of the populations chosen for empirical investigatious

S. No.	*Source*	*N*	*y*	x_1	x_2	ρ_{01}	ρ_{02}	ρ_{12}	α_{x_1}	α_{x_2}
1.	Goon-Gupta Das Gupta pp. 315	18	Yield of dry bark	Ht. in inches	Girth of a ht. of 6″	0.7679	0.7190	0.5205	9.32	6.12
2.	Snedecar P. 347	20	*	*	*	0.8096	−0.1355	−0·5364	8.00	23.94
3.	Census book Jodhpur	23	No. of H.H. in 23 villages	No. of Male cultivators	No. of female cultivators	0.8501	0.3291	0.1618	36.88	55.76
4.	Census Hand book Distt. Gujarat	54	No. of cultivators	No. of female cultivators	No. of Agr. Labour	0.8893	0.7381	0.4125	37.96	131.37

*Not given.

TABLE 2

The Characteristics of I* and percentage gain in efficiency of Y_{I*} over $\hat{Y}_1$, $\hat{Y}_2$ and $N\hat{\bar{Y}}$.

S. No.	ρ_{0I*}	$\lvert \alpha_{I*} \rvert$	*Percentage gain in efficiency of* $\hat{Y}_{I*}$ *over*		
			$\hat{Y}_1$	$\hat{Y}_2$	$N\hat{\bar{Y}}$
1.	0.8541	0.84	34.62	94.58	210.91
2.	0.8837	11.37	240.79	4800.70	288.96
3.	0.8720	5.12	95.74	310.78	289.78
4.	0.9783	22.96	449.68	2518.59	1845.95

SUMMARY

In this paper the use of multivariate auxiliary information in selecting units with unequal probability sampling and with replacement has been considered. The efficiency of the estimator of population total/mean based on a proposed criterion has been compared with the efficiency of usual PPSWR estimator using single auxiliary character. The gain in efficiency has been illustrated empirically for some well known populations.

ACKNOWLEDGEMENTS

The authors are thankful to the referees for their suggestions.

REFERENCES

[1] Des Raj (1958) : On the relative accuracy of some sampling techniques, JASA, 53, 98-101.

[2] Des Raj (1965) : On a method of using multi-auxiliary information in sample surveys, JASA, 60, 270-77.

[3] District Census Hand Book, Brauch (Gujarat).

[4] District Census Hank Book, Jodhpur (Rajasthan).

[5] Goon, A.M., Gupta, M.K., and Dasgupta, B. (1971) : *Fundamentals of Statistics*, Fourth revised Edition. World Press Private Ltd., Calcutta.

[6] Olkin, I. (1958) : Multivariate ratio estimators, *Biometrika* **45**, 154-65.

[7] Reddy, V.N. (1975) : A study on the use of prior knowledge on certain population parameters in estimation. Sankha, C, Sample Surveys—Theory.

[8] Maiti. P. and Tripathi, T.P. (1976) : The use of multivariate auxiliary information in selecting the sampling units. Proceedings of symposium on recent developments in survey methodology, held during March, 22-27, 1976 at I.S.I., Calcutta.

[9] Srivastava, S.K. (1965) : An estimate of the mean of finite population using several auxiliary variables, JISA, 3, 189-94.

[10] Tripathi, T.P. (1976) : On Double Sampling for multivariate ratio and difference methods of estimation, JISAS, Vol. 28, No. 1, pp. 33-54.