Journal of the Korean Statistical Society (2002), 31: 2, pp 143–152

An Empirical Comparison of Ratio and PPS Strategies

L. N. Sahoo[1] and M. Dalabehera[2]

ABSTRACT

In an effort to make a right choice among ratio estimation strategies and PPS sampling strategies, we conduct an empirical investigation of the relative performances of three ratio estimation strategies and four PPS estimation strategies using a set of 12 natural populations. The quality of a strategy is measured in the traditional way, namely with the consideration of efficiency, achieved coverage rate of the nominal 99% confidence interval and approach to normality (asymmetry).

Keywords. Auxiliary variable, confidence interval, efficiency, kurtosis, PPS strategy, ratio strategy, skewness.
AMS 2000 subject classifications. Primary 62D05.

1. Introduction

Consider a finite population of N units with observations $(y_i,\ x_i)$, $i = 1, 2, \cdots, N$ on the survey variable y and a positively correlated auxiliary variable x. We seek to estimate the population total Y of $y-$values from a random sample of n units, when the population total X of $x-$values is known. The $x-$values, when assumed to be known and non-negative, can be put to effective use either at the design stage in selecting units as in the various PPS sampling designs, or at the estimation stage in constructing a ratio or ratio–type estimator. Numerous attempts have been made in recent years to develop various practicable ratio estimation strategies as well as PPS sampling strategies. These strategies are expected to perform well when the relation between y and x is known to be linear passing through the origin *i.e.* $y = \beta x$, or approximately so.

Received August 1997; accepted September 1999.
[1]Department of Statistics, Utkal University, Bhubaneswar-751004, India
[2]Department of Statistics, Orissa University of Agriculture & Technology, Bhubaneswar-751003, India

A fundamental question that frequently arises in many survey situations concerning the effective use of x is how to make a unique choice between a PPS sampling strategy and a ratio estimation strategy, what ever be the survey variable y. Because, in the absence of the knowledge on $y-$values there is no scope of investigating the relationship $y = \beta x$ before hand. Further, more the relationship deviates from $y = \beta x$, less is the efficiency of a PPS strategy. On the other hand, there is no reason to believe that a ratio strategy would be more efficient than a PPS strategy when there is a high correlation between y and x. Thus, in practice we will probably desire a universal criterion for the choice of a suitable strategy. But, unfortunately, in the sample survey inference, there is no such guideline in this choice. In the face of this disadvantage, the only alternative is to study the relative comparison of various strategies. In this context, some model-motivated comparisons are available in Raj (1958), Foreman and Brewer (1971), Chaudhuri and Arnab (1979), Arnab (1979), Sahoo (1984), Chaudhuri and Adhikary (1989) among others.

In the present work, we tackle a problem of investigating some distributional characteristics of various strategies for estimating Y, based on ratio method of estimation under simple random sampling without replacement(SRSWOR) scheme and PPS sampling without replacement scheme using a variety of 12 natural populations. Attention is given to $n = 2$ only, because the calculations of inclusion probabilities become steadily more complex as n increases beyond 2. This is also a common situation which may exist in surveys with many small strata where separate estimators are used.

2. Strategies under Study and Their Performance Measures

Our aim is just to have an overall idea on the performances of ratio estimators and estimators based on PPS sampling schemes. So, we shall treat only some well known but potentially interesting strategies for this purpose. The strategies under consideration together with their variance estimators for $n = 2$ are described below. However, these strategies are clearly defined in every reliable text book on sampling theory (*cf.* Murthy, 1967; Cassel, Sarndal, Wretman; 1977).

1. The classical ratio estimation strategy. $H_1 = (\text{SRSWOR}, t_R)$ where $t_R = \frac{\bar{y}}{\bar{x}}X$, such that $\bar{y} = \frac{1}{2}(y_1 + y_2)$ and $\bar{x} = \frac{1}{2}(x_1 + x_2)$, 1 and 2 being the two units in the sample. There is no closed form for $\text{MSE}(t_R)$ or $\text{Var}(t_R)$. However both can be

estimated by an approximate variance estimator given by

$$v(H_1) = \frac{N(N-2)}{2} \sum_{i=1}^{2} \left(y_i - \frac{\bar{y}}{\bar{x}} x_i \right)^2$$

(Cochran, 1977, p. 155).

2. The strategy of Midzuno (1952) and Sen (1952) involving classical ratio estimator. $H_2 = (\text{pps}\Sigma x, t_R)$ where pps Σx is the design such that the first unit in the sample, i say, is drawn with PPS of x, that is, with probability $p_i = x_i/X$ and then without replacing this unit the second one is drawn with equal probability from the remaining $N-1$ units in the population. The unbiased variance estimator of H_2 is

$$v(H_2) = t_R^2 - N^2 \left\{ \bar{y}^2 - \frac{(N-2)}{2N} \sum_{i=1}^{2} (y_i - \bar{y})^2 \right\}.$$

3. The Hartley-Ross (1954) strategy. $H_3 = (\text{SRSWOR}, t_{HR})$ where $t_{HR} = \bar{r}X + 2(N-1)(\bar{y} - \bar{r}\bar{x})$ with $\bar{r} = \frac{1}{2}(\frac{y_1}{x_1} + \frac{y_2}{x_2})$.

Goodman and Hartley (1958) derived an unbiased variance estimator of H_3 under the assumption that N is large relative to n. But, as pointed out by the authors, this estimator is computationally cumbersome. However, an alternative unbiased estimator, based on the independent random group technique, suggested by them can also seriously underestimate $\text{Var}(H_3)$. So, for simplicity first we consider an approximate expression for $\text{Var}(H_3)$ as $\text{Var}(H_3) = 4(N-1)^2[\text{Var}(\bar{y}) - 2\bar{R}\text{Cov}(\bar{y}, \bar{x}) + \bar{R}^2\text{Var}(\bar{x})]$ obtained by replacing $\bar{r}$ by $\bar{R} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{x_i}$ in t_{HR}. Then, as a consistent estimator of $\text{Var}(H_3)$ we may consider the following approximate variance estimator for our purpose :

$$v(H_3) = \frac{2(N-1)^2(N-2)}{N} \sum_{i=1}^{2} [(y_i - \bar{y}) - \bar{r}(x_i - \bar{x})]^2.$$

4. Murthy (1957) strategy. $H_4 = (\text{ppsux}, t_{MR})$ where ppsux is the PPS without replacement design such that unit i in the sample is drawn with PPS of x for the remaining units and ignoring the order of the selection of the units,

$$t_{MR} = \frac{1}{2 - p_1 - p_2} \left\{ (1 - p_2)\frac{y_1}{p_1} + (1 - p_1)\frac{y_2}{p_2} \right\}.$$

An unbiased sample estimate of the variance of H_4 can be obtained as

$$v(H_4) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2.$$

5. Horvitz-Thompson (1952) strategy. $H_5 = (\text{ppsx}, t_{HT})$ where ppsx is the PPS of x without replacement design such that the inclusion probabilities $\pi_1 = p_1(1 + \frac{p_2}{1-p_2})$, $\pi_2 = p_2(1 + \frac{p_1}{1-p_1})$, $\pi_{12} = p_1p_2(\frac{1}{1-p_1} + \frac{1}{1-p_2})$ and $t_{HT} = (\frac{y_1}{\pi_1} + \frac{y_2}{\pi_2})$. An unbiased variance estimator (Yates-Grundy) of H_5 is provided by

$$v(H_5) = \frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}} \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2.$$

6. The strategy of Midzuno (1952) and Sen (1952) involving H-T estimator. $H_6 = (\text{pps}\Sigma x, t_{HT})$. An unbiased variance estimator of this strategy is $v(H_6) = v(H_5)$ with $\pi_i = \frac{N-2}{N-1}p_i + \frac{1}{N-1}$ $(i = 1, 2)$ and $\pi_{12} = \frac{1}{N-1}(p_1 + p_2)$.

7. The Rao-Hartley-Cochran (1962) strategy. $H_7 = (\text{ppsgx}, t_{RHC})$ where ppsgx denotes the design consisting in drawing by PPS of x, one unit from each of the two groups with N_1 and N_2 units, into which the population has been divided at random and

$$t_{RHC} = \frac{y_1}{p_1}P_1 + \frac{y_2}{p_2}P_2$$

where $P_i = \sum_{group\ i} p_i$ $(i = 1, 2)$. The unbiased variance estimator of H_7 is

$$v(H_7) = CP_1P_2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

where $C = \frac{N_1^2+N_2^2-N}{N^2-N_1^2-N_2^2}$. It is also assumed that $N_1 = N_2 = N/2$ when N is even and $N_1 = \frac{N-1}{2}$ and $N_2 = \frac{N+1}{2}$ when N is odd.

It may be noted here that the $p_i's$ of the strategies H_4, H_5 and H_7 are normed probabilities and $\Sigma_i p_i = 1$, while the $p_i's$ in H_6 are revised probabilities and $\Sigma_i\ p_i \neq 1$.

To study the relative performance of the competing strategies, the following performance measures were taken into consideration :

(a) Relative efficiency (RE)

The efficiency of biased strategy H_1 is measured in term of mean square error and the efficiencies of other strategies are measured in terms of their variances.

(b) Coverage rate (CR) based on the approximate $100(1 - \alpha)\%$ confidence interval

$$t \pm z_{1-\frac{\alpha}{2}}\sqrt{v(H)},$$

where the constant $z_{1-\frac{\alpha}{2}}$ is exceeded with probability $\frac{\alpha}{2}$ by the unit normal variate, $v(H)$ is the variance estimator of the strategy H and t is the corresponding estimator.

This performance measure gives us an idea about which percentage of the so constructed confidence intervals covers the true value of Y under repeated draws of samples by the given design.

(c) Approach to normality(asymmetry)

The coefficients of skewness and kurtosis, *i.e.* β_1 and β_2 coefficients are taken as the indices for the measurements of the symmetry of the sampling distribution of the strategy. It may be noted here that for a standard symmetrical(normal) distribution, $\beta_1 = 0$ and $\beta_2 = 3$.

3. Description of the Empirical Study

For the empirical comparison, 12 natural populations are used. The source, size(N), nature of y and x, correlation coefficient between y and $x(\rho)$ of these populations are provided in Table 3.1. The relevant numerical values computed from the populations considering all the $\binom{N}{n}$ possible samples, for $n = 2$, are shown in Tables 3.2 and 3.3. To save space, the numerical values on β_1 and β_2 are not given. But, only the main findings on these values are presented.

3.1. Results based on the RE

The percentage relative efficiencies of different strategies compared to $H_0 =$ (SRSWOR, $N\bar{y}$) are presented in Table 3.2. Among the ratio estimation strategies H_1,H_2 and H_3 we find that H_3 is clearly the worst performer whereas H_2 is the best performer except for populations 9 and 10 where the efficiency losses relative to H_1 are only marginal. Among the strategies H_5 and H_6 using $H-T$ estimator, H_5 is strongly preferred to H_6. Strategies H_4 and H_7 perform very similarly in the sense that each of them is the best performer in six cases among the seven strategies considered in this paper. The only consistent conclusion from Table 3.2 is that the strategy H_6 is the poorest among the competing strategies. However, other PPS sampling strategies *i.e.* H_4, H_5 and H_7 are greatly superior to the ratio estimation strategies except for populations 1 and 8 where H_7 appeared to be inferior to H_2.

It is usually hoped that for a population with a high positive value of ρ, the

TABLE 3.1 *Description of the populations*

Pop.no.	*Source*	*N*	*y*	*x*	*ρ*
1	Hajek (1981) p.10	37	yield	area	0.974
2	Murthy (1967) p.398	43	no. of absentees	no. of workers	0.661
3	Murthy (1967) p.399	34	area under wheat in 1964	area under wheat in 1961	0.980
4	Murthy (1967) p.399	34	area under wheat in 1964	area under wheat in 1963	0.988
5	Konijn (1973) p.49	16	food expenditure	total expenditure	0.954
6	Singh and Chaudhary (1986) p.166	16	area under wheat(1979-80)	area under wheat(1978-79)	0.978
7	Singh and Chaudhary (1986) p.176	20	no. of cows in milk enumerated	actual no. of cows in the previous year	0.889
8	Singh and Chaudhary (1986) p.287	12	bovine population in survey year	bovine population in census year	0.948
9	Sukhatme and Sukhatme (1970) p.51	25	area under rice	total cultivated area of the village	0.919
10	Sukhatme and Sukhatme (1970) p.166	20	no. of banana bunches	no.of banana pits	0.774
11	Sukhatme and Sukhatme (1970) p.185	34	area under wheat in 1937	area under wheat in 1936	0.930
12	Sukhatme and Sukhatme (1970) p.185	34	area under wheat in 1937	total cultivated area in 1936	0.900

strategies under consideration may yield dramatic efficiency gains compared to H_0. However, this phenomenon is not clearly observed in some cases especially in population 4, where $\rho = 0.988$ but RE's are low. The reason is that the population regression line of y on x intercepts the $y-$axis at some distance from the origin.

3.2. Results based on the CR

Using several strategies, the coverage rates of nominal 99% confidence intervals for Y are shown in Table 3.3. Results for the nominal level 95% are not shown, as they confirm more or less the tendencies found in the cases of 95%.

TABLE 3.2 *Features of the RE w.r.t.* H_0 *(in %)*

Pop.no.	H_1	H_2	H_3	H_4	H_5	H_6	H_7
1	1718	1835	1544	1904	1847	297	1755
2	169	183	165	188	186	164	209
3	2356	2859	1525	3103	3012	419	3061
4	464	525	361	577	567	265	557
5	565	655	642	662	609	208	721
6	1537	2025	1289	2131	2197	409	2260
7	452	499	309	559	553	299	781
8	674	847	552	911	823	290	813
9	728	718	505	887	867	312	1029
10	252	251	241	257	255	174	252
11	1377	1422	538	1548	1615	310	1847
12	462	524	355	578	570	269	555

For this level CRs of all strategies except H_4 are unpredictable and usually bear no resemblance to the nominal rates aimed at.

Table 3.3 gives an indication of better performance of H_3 than H_1 and H_2, and slightly better performance of H_5 than H_6 and H_7. The best performer ratio estimation strategy H_3 also appears inferior to PPS estimation strategies H_5 and H_7. However, the only consistent conclusion available to us from Table 3.3 is that the performance of Murthy's strategy H_4 is the most excellant among the seven strategies as it has the highest achieved CR.

3.3. Results based on the coefficient of skewness

Sampling distribution of none of the competing strategies shows a consistent behaviour in the approach to symmetry. Of course, in most of the populations, the distributions of the strategies are not for away from the symmetry. More specifically, the strategies H_2, H_4, H_5 and H_7 perform equally well in many cases. The remaining strategies have been superior in the sense of their approach to symmetry for some cases. The answer is not straight forward which strategy could be given preference over other under such a performance measure.

3.4. Results based on the coefficient of kurtosis

A similar trend as that in the case of skewness is also observed for the coefficient of kurtosis of the sampling distributions of the strategies. The distribution of H_2 seems to fluctuating more in comparison to others. The deviation from the

TABLE 3.3 *Features of the CR of nominal 99% confidence interval*

Pop.no.	H_1	H_2	H_3	H_4	H_5	H_6	H_7
1	76	61	78	95	81	78	79
2	71	66	73	93	74	73	75
3	63	69	71	94	74	72	73
4	69	72	73	93	79	67	76
5	82	58	82	95	88	76	83
6	75	67	81	94	78	82	88
7	81	69	83	94	81	78	91
8	67	70	80	92	74	77	75
9	64	64	71	93	79	80	79
10	77	58	77	94	79	80	83
11	72	74	80	95	82	71	81
12	69	72	73	93	79	67	76

normality is not remarkable for most of the strategies. The competing strategies seem to be all at par in their approach to normality.

4. Conclusions

An examination of the results of this empirical study leads to the following tentative conclusions :

Horvitz-Thompson strategy H_5, is undoubtedly superior to the classical ratio estimation strategy H_1 in respect of the three performance measures. H_6 is less efficient than others showing that the use of Horvitz-Thompson estimator under Midzuno's scheme of sampling is definitely inferior to others. But, it compares favourably with H_1, H_2 and H_3 in respect of other two performance measures *i.e.*, achieved CR and approach to normality in most of the populations. Although, the strategies under consideration show very much erratic behaviour in respect of coefficients of skewness and kurtosis, we still observe that the distributions of PPS strategies have a tendency to be more symmetrical than the ratio estimation strategies in most of the populations. The coverage rates of H_7 have been poor in comparison to H_4 even though it is the most efficient in some cases. On the otherhand, H_5 has a very limited application when $n > 2$. Thus, we may conclude that Murthy's strategy H_4 is the most suitable strategy in respect of our performance measures.

From this empirical study we also conclude that, the overall performance of PPS sampling strategies compared to the ratio estimation strategies on the basis

of the three performance measures is very much satisfactory. Of course, these results are only indicative. Further investigations in this direction may be made for arriving at the conclusion.

Acknowledgement

The authors are grateful to the referees and Editor-in-Chief whose constructive comments led to an improvement in the paper.

References

Arnab, R. (1979). "On the relative efficiencies of some sampling strategies under a super-population model", *Journal of Indian Society of Agricultural Statistics*, **31**, 89–96.

Cassel, C. M., Sarndal, C. E. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*, John Wiley & Sons, New York.

Chaudhuri, A. and Adhikary, A. K. (1989). "On efficiency of the ratio estimator", *Metrika*, **36**, 55–59.

Chaudhuri, A. and Arnab, R. (1979). "On the relative efficiencies of sampling strategies under a super-population model", *Sankhyā*, **C41**, 40–53.

Cochran, W. G. (1977). *Sampling Techniques*, Third Edition, Wiley Eastern Limited.

Foreman, E. K. and Brewer, K. R. W. (1971). "The efficient use of supplementary information in standard sampling procedures", *Journal of the Royal Statistical Society*, **B33**, 391–400.

Goodman, L. A. and Hartley, H. O. (1958). "The precision of unbiased ratio-type estimators", *Journal of the American Statistical Association*, **53**, 491–508.

Hajek, J. (1981). *Sampling from a Finite Population*, Marcel Dekker Inc, New York.

Hartley, H. O. and Ross, A. (1954). "Unbiased ratio estimators", *Nature*, **174**, 270–271.

Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, **47**, 663–685.

Konijn, H. S. (1973). *Statistical Theory of Sample Survey Design and Analysis*, North Holland, Amsterdam.

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

Murthy, M. N. (1957). "Ordered and unordered estimators in sampling without replacement", *Sankhyā*, **18**, 379–390.

Murthy, M. N. (1967). *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

Raj, D. (1958). "On the relative accuracy of some sampling techniques", *Journal of the American Statistical Association*, **53**, 98–101.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). "A simple procedure of unequal probability sampling without replacement", *Journal of the Royal Statistical Society*, **B24**, 482–491.

Sahoo, L. N. (1984). "A comparison of ratio and PPS estimates under a superpopulation model", *Gujarat Stat. Rev.*, **11**, 69–72.

Sen, A. R. (1952). "Present status of probability sampling and its use in estimation of farm characteristics(abstract)", *Econometrica*, **20**, 130.

Singh, D. and Chaudhury, F. S. (1986). *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Limited, New Delhi.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*, Asia Publishing House, Calcutta.