On the Relative Efficiencies of Single and Replicated Simple Random Samples

by
M.C. Agrawal*

Abstract

Subject to the same total expected cost (taken as proportional to effective sample size), a commonly used estimator based on k independent interpenetrating sub-samples of equal size selected according to SRSWOR method has been compared with the usual estimators based on (i) SRSWR and (ii) SRSWOR, and is found to be more efficient than the former but less than the latter. This estimator has also been compared for the same expected cost with an estimator based on 'dependent' sub-samples in interpenetrating sub-sampling.

Numerical results on the relative efficiencies of the above and some other estimators are presented and the effect of showing consideration to non-integer sample sizes has been studied.

1. Introduction

The usual estimator obtained by the technique of interpenetrating sub-samples (also known as replicated sampling) has been compared with the estimators based on a single sample drawn by employing (i) simple random sampling with replacement (SRSWR) and (ii) simple random sampling without replacement (SRSWOR). Singh and Bansal (1975) studied the relative efficiencies of estimators based on single sample drawn with SRSWOR and on independent replicated samples drawn with SRSWOR keeping the overall effective sample sizes equal for both the schemes. We consider here their relative efficiencies against the sample mean in SRSWR keeping the average effective sample size equal, implying thereby that the expected cost, which is taken proportional to effective sample size, is the same. Sections 4 and 6 are then devoted to a numerical comparison of some potentially competing estimators. In this connection we have also addressed ourselves to the question as to what is the effect of taking into account the possibility that the (average effective) sample size is not an integer.

Roy and Singh (1973) proposed an estimator (discussed in Section 3 of this paper) which is shown by them to be more efficient than the one under consideration in interpenetrating sub-sampling. Here, we shall compare them for the same expected cost taken proportional to effective sample size.

2. The relative efficiency

For a population of N units, let Y_j be the value of some Y-characteristics associated with the jth unit (j = 1, 2, ..., N).

An estimator of the population mean

$$\bar{Y}_N = \sum_{j=1}^{N} Y_j / N$$

generally used in applying the technique of independent interpenetrating sub-sampling briefly TIIS (with K sub-samples, each of size n^*/k, drawn with SRSWOR making up a sample of size n^*) is

*Visiting Professor, Statistical Center, University of the Philippines

$$y_T = \sum_{i=1}^{k} y_i / k$$

where y_i is the arithmetic mean of n^*/k observations in the i th sub-sample selected with SRSWOR.

2.1 TIIS Versus SRSWR

It is well-known that for a sample of size n in SRSWR the average effective sample size is

$$E(\upsilon) = N[1 - (1- 1/N)^n]. \qquad (2.1.1)$$

Also for k independent replicated samples taken as above, the average effective size is

$$E(\upsilon^*) = N[1 - (1 - n^*/kN)^k].$$

Equating $E(\upsilon)$ with $E(\upsilon^*)$, we determine n^* in terms of the other parameter. Then, the relative efficiency of the sample for SRSWR with respect to y_T works out as

$$\text{R.E.} = \frac{n(1 - 1/N)^{n/k}}{k(N-1)[1-(1-1/n)^{n/k}]} = \frac{(1 - \alpha)^{\beta-1}\alpha\beta}{1 - (1 - \alpha)\beta} = E1 \text{ (say)} \qquad (2.1.2)$$

where $\alpha = 1/N$ and $\beta = n/k$.

THEOREM 2.1.1 The relative efficiency expressed by (2.1.2) is always less than or equal to 1.

PROOF The relative efficiency will be less than or equal to 1 if

$$(1-\alpha)^{\beta-1}\alpha\beta \le 1 - (1-\alpha)^{\beta}$$

$$\text{or } (1-\alpha)^{-1}\alpha\beta \le (1-\alpha)^{-\beta} - 1$$

Since $0 < \alpha < 1$, we expand both sides and collect the coefficients of powers of α. Thus, we get

$$\beta \sum_{r=2}^{\infty} \alpha^r \frac{(\beta+1)(\beta+2)\ldots(\beta+r-1)}{r!} - 1 \ge 0,$$

which always holds as $\beta \ge 1$.

It can, therefore, be stated that the TIIS estimator y_T is more efficient than the usual SRSWR estimator.

It is easy to see that the relative efficiency expressed by (2.1.2) does not change when n and k vary in such a manner that n/k is a fixed value for a given N. We have the following theorem regarding the behaviour of the relative efficiency.

THEOREM 2.1.2 The relative efficiency expressed by (2.1.2) is

(i) strictly increasing with k for fixed values of n and N

(ii) strictly decreasing with n for fixed values of k and N.

PROOF In order to prove the two parts of the theorem, we differentiate (2.1.2) with respect to β, and show that

$$\frac{\partial E_1}{\partial \beta} < 0.$$ Differentiation yields

$$\frac{\partial E_1}{\partial \beta} = \frac{\alpha(1-\alpha)^{\beta-1}[\log(1-\alpha)^{\beta} + 1 - (1-\alpha)^{\beta}]}{[1 - (1-\alpha)^{\beta}]^2}.$$

Now, let

$$1 - (1-\alpha)^{\beta} = y$$

so that

$$\log(1-\alpha)^{\beta} + 1 - (1-\alpha)^{\beta} = \log(1-y) + y.$$

Since for $0 < y < 1$

$$y + \log(1-y) < 0,$$

it follows immediately that

$$\frac{\partial E_1}{\partial \beta} < 0.$$

In order to enable the reader to get some idea about the relative efficiency E_1 of estimator and also to illustrate simultaneously the above theorem, we make simple numerical investigations.

Making use of (2.1.2), we have prepared the following table for N = 50.

n / k	2	3	4
6	.9790	.9899	.9949
10	.9600	.9766	.9849
15	.9356	.9600	.9724
20	.9116	.9432	.9600

[See also Section 4.]

2.2 TIIS versus SRSWOR

The relative efficiency of the sample mean based on n' draws in SRSWOR with respect to the TIIS estimator y_T, both based on the same average effective size is

$$R.E. = \frac{n'(1-n'/N)^{1/k}}{k(N-n')\,[1-(1-n'/N)^{1/k}]} = \frac{(1-\alpha')^{\beta'-1}\,\alpha'\beta'}{1-(1-\alpha')^{\beta'}} = E2 \text{ (say)} \qquad (2.2.1)$$

where $\alpha' = n'/N$ and $\beta' = 1/k$.

As has been shown by Singh and Bansal (1975), the relative efficiency expressed by (2.2.1) is greater than or equal to 1.

THEOREM 2.2.1 The relative efficiency expressed by (2.2.1) is strictly increasing with

(i) k for fixed values of n' and N

(ii) n' for fixed values of k and N.

PROOF Parts (i) and (ii) of the theorem easily follow by differentiating (2.2.1) with respect to β' and α' respectively. In view of the analogy between the expressions (2.1.2) and (2.2.1), we easily obtain as in the proof of the Theorem 2.1.2

$$\frac{\partial E_2}{\partial \beta'} < 0$$

and further

$$\frac{\partial E_2}{\partial \alpha'} = \frac{\beta'(1-a')^{\beta'-2}\,[1-\alpha'\beta'-(1-\alpha')^{\beta'}]}{[1-(1-\alpha')^{\beta'}]^2}$$

Since

$$1-\alpha'\beta'-(1-\alpha')^{\beta'} = 1-(n'/Nk)-(1-n'/N)^{1/k}$$
$$= ((k-1)/2k^2)\,(n'/N)^2 + ((k-1)(2k-1)/6k^3)\,(n'/N)^3 + \ldots,$$

it is then evident that

$$\frac{\partial E_2}{\partial \alpha'} < 0$$

as $k > 1$.

This completes the proof of the theorem.

We compute below E_2 from (2.2.1) for N = 50.

n' / k	2	3	4
6	1.0330	1.0452	1.0498
10	1.0590	1.0792	1.0894
15	1.0976	1.1316	1.1488
20	1.1455	1.1972	1.2235

The table points to the fact that TIIS can be considered fairly close to SRSWOR for suitably chosen number k and ratio n'/N. This encourages us to be inclined favourably towards TIIS which is additionally and intrinsically endowed with certain desirable properties. [See also Section 4.]

3. Interpenetrating sub-samples - with and without replacement

Under the same cost consideration as above, the usual estimator in TIIS will now be compared with an estimator based on sample obtained in such a manner that, in contradistinction to TIIS, sub-samples drawn according to SRSWOR are not replaced (and hence there is no common unit between the sub-samples). This sam-

pling scheme amounts, in practice, to drawing lk units by SRSWOR and then assigning the first l units to sub-sample 1, the second l units to sub-sample 2 and so on; or else lk units may be allotted to the k sub-samples in a prespecified manner. Roy and Singh (1973) have considered 'ordered' and 'unordered' estimators based on these k dependent sub-samples. Since the more efficient 'unordered' estimator, as shoen by them, is the same as the usual SRSWOR estimator, we need to consider, for the intended comparison with y_T, the 'ordered' estimator defined by

$$y_T' = 1/Nk\,(t_1 + \ldots + t_K)$$

where $t_1 = Ny_1$

$$t_r = l\,(y_1 + \ldots + y_{r-1}) + (N - (r-1)\,l)y_r \quad (r = 2,3,\ldots,k)$$

and y_i = mean of the i th sub-sample (i = 1, 2, . . ., k).

This estimator is unbiased and its variance is given by

$$V(y'_T) = \frac{1}{lkN^2}\left[N^2 - kNl + (l^2/3)\,(k^2 - 1)\right] S^2$$

where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - Y)^2$

Hence, the relative efficiency of y'_T with respect to y_T keeping the average effective size the same is obtained as

$$\text{R.E.} = \frac{l\,[1 - (lk/n) + (l^2/3N^2)(k^2 - 1)]^{-1}}{N\,[(1 - (lk/N)^{-1/k} - 1]} \qquad (3.1)$$

THEOREM 3.1 The relative efficiency expressed by (3.1) is greater than or equal to 1 if

$$N \geq \frac{5k + \sqrt{13k^2 + 12}}{6} \qquad (3.2)$$

which is always satisfied if $N \geq 3/2\ kl\ (k \geq 2)$.

PROOF Expanding the term in the denominator of (3.1), we get

$$\text{R.E.} = \frac{[1 - (lk/N) + (l^2/3N^2)\,(k^2 - 1)]^{-1}}{1 + ((k+1)/2)(l/N) + ((k+1)/2)((k+2)/3)(l^2/N^2) + \ldots}$$

Since

$$(rk + 1)/(r + 1) \leq k \quad (r \geq 1),$$

we have

$$\text{R.E.} \geq \frac{[1 - (lk/N) + (l^2/3N^2)\,(k^2 - 1)]^{-1}}{1 + ((k+1)/2)(l/N)[1 + (kl/N) + (k^2l^2/N^2 + \ldots]}$$

$$= \frac{(1 - (lk/N))\,[1 - (lk/N) + (l^2/3N^2)\,(k^2 - 1)]^{-1}}{\{1 - (k - 1)\,l/2N\}},$$

which is greater than or equal to 1 if (3.2) is satisfied, and hence it follows that, subject to the cost consideration discussed above, y'_T performs better than y_T if N exceeds 3/2 kl ($k \geq 2$).

It is clear from (3.1) that, for given l/N and k, the relative efficiency $V(y_T)/V(y'_T)$ is insensitive to changes in N. In order to give an idea about the relative efficiency expressed by (3.1), we have prepared a small table for N = 50.

	k = 2	k = 4
k = 8	1.0376	1.0588
k = 12	1.0536	1.0859
k = 16	1.0662	1.1101
k = 20	1.0739	1.1294

It may be noted from the table that the gains of y'_T over y_T are small unless the sampling ratio and the number of sub-samples are large.

4. TIIS, SRSWR and SRSWOR

To have a simultaneous and collective view of the performance of the estimators in the three sampling schemes, viz., TIIS, SRSWR and SRSWOR, we may proceed with the computation of the relative sizes of the three variances conditioned by the same total expected cost. We take n draws in SRSWR and then bring about a parity in respect of the expected cost, i.e., in terms of the expected number of distinct units in TIIS and SRSWOR. In order to achieve this in the case of SRSWOR, we simply have to take the expected number of distinct units ν given by (2.1.1) as the sample size in SRSWOR and thus we shall get the following variance

$$V_{SRSWOR} = \frac{N-E(\nu)}{NE(\nu)} S^2 = \frac{(1-(1/N))^n}{N[1-(1-(1/N))^n]} S^2.$$

We may make use of the variance expressions of Section 2.1 for TIIS and SRSWR, i.e.,

$$V_{TIIS} = \frac{(1-(1/N))^{n/k}}{Nk[1-(1-(1/N))^{n/k}]} S^2. \qquad (4.1)$$

and $V_{SRSWR} = [(N-1)/(Nn)] S^2$.

It will not be out of place to consider the performance of an estimator which is the average of the distinct values obtained in an SRSWR sample of size n. The variance of this estimator is known to be

$$V_{SRSWR(D)} = (S^2/N) \sum_{j=1}^{N-1} (j/N)^{n-1}$$

Denoting the variances V_{SRSWOR}, $V_{SRSWR(D)}$, V_{TIIS} and V_{SRSWR} by V^*_1, V^*_2, V^*_3 and V^*_4 respectively, we prepare the following table:

n/N = 0.1

	V^*_1/V^*_4	V^*_2/V^*_4
N = 50	.9600	.9701
N = 200	.9531	.9556
N = 500	.9517	.9527
N = ∞	.9508	.9508

V^*_3/V^*_4

	k = 2	k = 3	k = 4
N = 50	.9848	.9931	.9975
N = 200	.9776	.9858	.9900
N = 500	.9762	.9844	.9885
N = ∞	.9752	.9835	.9876

V^*_1/V^*_3

	k = 2	k = 3	k = 4
N = 50	.9748	.9665	.9624
N = 200	.9749	.9668	.9627
N = 500	.9750	.9668	.9627
N = ∞	.9750	.9668	.9627

n/N = 0.2

	V^*_1/V^*_4	V^*_2/V^*_4
N = 50	.9116	.9214
N = 200	.9054	.9078
N = 500	.9041	.9051
N = ∞	.9033	.9033

V*3/V*4

	k = 2	k = 3	k = 4
N = 50	.9600	.9766	.9848
N = 200	.9531	.9694	.9776
N = 500	.9517	.9679	.9762
N = ∞	.9508	.9670	.9752

V^*_1/V^*_3

	k = 2	k = 3	k = 4
N = 50	.9495	.9335	.9256
N = 200	.9500	.9340	.9261
N = 500	.9500	.9341	.9262
N = ∞	.9500	.9341	.9263

If we think in terms of the same expected cost for the sampling schemes where cost is taken as proportional to the expected number of distinct units, then the following comments will be in order:

(1) In respect of the efficiency, the TIIS estimator considered in Section 3 falls between the SRSWOR estimator and the usual SRSWR estimator, and its performance depends on the factors n/N and k.

(2) For large N, the gain in efficiency attained by both the SRSWOR estimator and the estimator based on distinct units in SRSWR over the usual SRSWR estimator is hardly different, and it is no different as $N \to \infty$. This remark emerges analytically, as under the limit process

$$N \to \infty, n \to \infty \text{ and } n/N \to f_o,$$

we find that both

$$NS^{-2}V_{SRSWR(D)} = \sum_{j=1}^{N-1} (j/N)^{n-1}$$

and

$$NS^{-2}V_{SRSWOR} = \frac{[1 - (1/N)]^n}{N[1 - (1 - (1/N))^n]}$$

tend to $1/(e^{fo} - 1)$.

(3) While sampling from a population of given size N, the relative efficiency V^*_3/V^*_4 can be matched with the relative efficiency V^*_1/V^*_4 if the ratio n/N (n = number of draws in SRSWR) on which the latter is based. This can be seen from the relevant variance expressions involved here. An indication of this is available from the column V^*_1/V^*_4 with n/N = 0.1 and the column V^*_3/V^*_4 with n/N = 0.2 and k = 2.

(4) Given n/N and k, the relative efficiency V^*_1/V^*_3 is practically unaffected by a change in the size of the population. A scrutiny of the relevant variance expression also points to this effect.

5. Effect of Consideration to non-integer sample sizes

We shall now undertake an investigation with a view to finding the effect of not taking into account the possibility of non-integer (average effective) sample sizes in our preceding discussion. In order to make efficiency comparisons for the same expected cost Ramakrishnan (1969) suggested an unbiased 'randomized' estimator in SRSWOR with due consideration to the fact that the expected number of distinct units in a with-replacement sample need not be an integer. This 'randomized' estimator defined by

$$y^*_{E(\upsilon)} = \begin{cases} y_{[E(\upsilon)]} & \text{with probability } P_1 \\ y_{[E(\upsilon)]+1} & \text{with probability } P_2 \end{cases}$$

has the variance given by

$$V(y^*_{E(\upsilon)}) = \{(2[E(\upsilon)] + 1 - E(\upsilon))/([E(\upsilon)]([E(\upsilon)]+1)) - (1/N)\} S^2$$

where $P_1 = 1 - E(\upsilon) + [E(\upsilon)]$, $P_2 = E(\upsilon) - [E(\upsilon)]$ and [z] denotes the integral part of z.

It may be pertinent to point out that, in the above context, the qualifier 'randomized' seems to be a misnomer, and hence we would instead like to call $y^*_{E(\upsilon)}$ an estimator based on a randomized sample size (rss).

Using an estimator with rss for each sub-sample of size n^*/k (n^* is chosen as explained in Section 2.1), we can strainghtaway write the new variance in TIIS as

$$V^*_3{}^* = 1/k \{((2[m^*] + 1 - m^*) / ([m^*]([m^*]+1)) - (1/N)\}S^2$$

where $m^* = n^*/k = N(1 - (1 - (1/N))^{n/k})$.

It can easily be verified that $V^*_3{}^* \geq V^*_3$, where V^*_3 has the same meaning as in the last section and is given by (4.1), i.e., V^*_3 stands for the variance obtained without regard to non-integer sample sizes.

To determine the effect of showing consideration to non-integer sample sizes, we shall examine the relative increase in V^*_3 given by

$$I_2 = (V^*_3{}^*/V^*_3) - 1 = \{m^* / (1 - (m^* / N))\} \{((2[m^*] + 1 - m^*) / ([m^*]([m^*]+1))) - 1\}.$$

Setting $m^* = x + a$ where x is the integral part of m^* and a is the fractional part lying between 0 and 1, we get after some simplification

$$I_2 = [(a(1-a)) / (x(x+1))][1 - ((x+a)/N)]^{-1}. \quad (5.1)$$

Obviously, I_2 will be zero if m^* is an integer. Assuming N to be large enough as compared to m^*, I_2 can be approximated by

$$I_2 \approx [a(1-a)] / [x(x+1)] \quad (5.2)$$

We shall now study the behaviour of I_2 expressed by (5.2) for non-integer values of m^*. For this purpose, we first of all notice that the supremum of I_2, given x, is at a = 1/2, while the unconditional supremum is at a = 1/2 and x = 1. Furthermore, the function I_2, for a given x, is symmetric about x + 1/2. It can also be easily seen that the function I_2, which is a measure of inflation in V^*_3 arising out of consideration to non-integer sample sizes, yields values all below or equal to 1% if m^* exceeds 5. Also, less than 1% inflation occurs even for non-integer $m^* < 5$ if m^* is close to an integer or it assumes values avoiding a certain range (depending on x) around x + 1/2 (x = 1, 2, 3 and 4). It may be mentioned that these observations apply equally well to the case of SRSWOR which occurs for k = 1.

If need be, the multiplying factor $(1-((x + a)/N))^{-1}$ in (5.1) could be brought into consideration to adjust the earlier computations. Here, it may be noted that an increase in N, for a given m^*, is accompanied by a decrease in I_2.

6. TIIS with distinct units versus SRSWOR

This section is intended to bring out some features of a comparison between an estimator y_D in TIIS defined as an average of distinct units in k sub-samples of equal size l (= n^*/k) drawn independently according to SRSWOR and the SRSWOR estimator by observing the same cost consideration as before. In view of certain obvious difficulties in evaluating the combinatorial varinace expression for the former estimator, it is desirable to prepare a separate table to facilitate an exact comparison between these two estimators for appropriate values of n^* and k so that is an integer.

The variance of y_D is given by

$$V_{TIIS(D)} = \frac{1}{\binom{N}{l}^k} \sum_{r=1}^{N-l} \frac{\binom{N-r}{l}^k}{N-r} S^2.$$

[See Pathak (1964) and Agrawal] (1981).]

If n^* is the sample size in TIIS, then we know from Section 2.2 that the sample size n' in SRSWOR is determined by

$$n' = N[1 - (1 - (n^*/kN))^k].$$

The variance of the SRSWOR estimator of the population mean is then obtained as

$$V'_{SRSWOR} = [(N-n')/Nn'] S^2 = \{[1-(n^*/kN)]^k / N[1-(1-(n^*/kN))^k]\} S^2$$

An asymptotically interesting result follows if

$N \to \infty$, $k \to \infty$ and $[lk(=n^*)/N] \to f_o$.

Under these conditions

$$NS^{-2}V_{TIIS\,(D)} \to [\,1/(e^{fo}-1)\,]$$

and

$$NS^{-2}V'_{SRSWOR} \to [\,1/(e^{fo}-1)].$$

For fixed k and large N, the difference of the two variances will be obtained as

$$V_{TIIS(D)} - V'_{SRSWOR} = [1/2N]\,[\,(k-1)/k(kl-1)]\,S^2,$$

to terms of order N^{-1}.

To illustrate the above, we have computed the relative efficiency in the following table.

		$V'_{SRSWOR}/V_{TIIS(D)}$	
(n^*,N)	n^*/N	k = 2	k = 4
(8,80)	0.1	0.9963	0.9944
(20,000)	0.1	0.9986	0.9979
(40,000)	0.1	0.9993	0.9990
(16,80)	0.2	0.9963	0.9946
(40,200)	0.2	0.9986	0.9979
(80,400)	0.2	0.9993	0.9990
(40,80)	0.5	0.9958	0.9941
(100,200)	0.5	0.9984	0.9977
(200,400)	0.5	0.9992	0.9988

Subject to the cost aspect under consideration, the above table points to the fact that the estimator based on distinct units in TIIS is almost as efficient as the SRSWOR estimator. It is, further, clear from the table that the relative performce of the two estimators for a given population of size N is hardly affected by a change in the value of the n^*/N.

Acknowledgment

The author expresses his gratitude to Professor Gunnar Kulldorff for his help in writing this paper.

References

Agrawal, M.C. (1981). On averaging over distinct units in replicated samples. *Mathematische Operations forschung und Statistik*, Series Statistics.

Pathak, P.K. (1964). Sufficiency in sampling theory. *Ann. Math. Statist.*, 35, 797-808.

Ramakrishnan, M.K. (1969). Some results on the comparison of sampling with and without replacement. *Sankhya* Ser. A, 31, 333-342.

Roy, A.S. and Singh, M.P. (1973). Interpenetrating subsamples with and without replacement. *Metrika*, 20, 230-239.

Singh, R. and Bansal, M.L. (1975). On the efficiency of interpenetrating sub-samples in simple random sampling. *Sankhya* Ser. C, 37, 190-198.