QÜESTIIÓ, vol. 22, 2, p. 243-252, 1998

ON SOME STRATEGIES USING AUXILIARY INFORMATION FOR ESTIMATING FINITE POPULATION MEAN

L.N. SAHOO
Utkal University*

J. SAHOO
Orissa University**

M. RUÍZ ESPEJO
UNED*

This paper presents an empirical investigation of the performances of five strategies for estimating the finite population mean using parameters such as mean or variance or both of an auxiliary variable. The criteria used for the choices of these strategies are bias, efficiency and approach to normality (asymmetry).

Keywords: Auxiliary information, empirical study, finite population mean, sampling.

AMS Classification: 62 D 05

* Department of Statistics, Utkal University. Bhubaneswar 751004, India.
** Department of Statistics, Orissa University of Agriculture and Technology, Bhubaneswar 751003, India.
⋆ Departamento de Economía Aplicada Cuantitativa. Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid.
– Received January 1997.
– Accepted September 1997.

1. INTRODUCTION

Let y_i, x_i $(i = 1, 2, \ldots, N)$ be the values of a survey variable y and a positively correlated auxiliary variable x on the ith unit of a finite population of size N. Suppose that x_i's are known and an estimate is needed for the population mean $\bar{Y}$ of y on the basis of a sample s of fixed size n, selected by simple random sampling without replacement (SRSWOR). It is known that the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$$

provides an unbiased estimate of $\bar{Y}$. But, use of the auxiliary variable x very often provides an estimator with increased accuracy. With this objective when the population mean $\bar{X}$ of x is known, traditional ratio method estimates $\bar{Y}$ by

$$t_1 = \bar{y}\frac{\bar{X}}{\bar{x}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i\in s} x_i.$$

However, this can be rendered unbiased under Sen-Midzuno (1952) [SM, say] scheme in which the sample s is selected with probability

$$p(s) = \binom{N}{n}^{-1}\frac{\bar{x}}{\bar{X}}.$$

In many practical situations σ_x^2, the population variance of x or equivalently $S_x^2 = N\sigma_x^2/(N-1)$ is known, or can be calculated from the known x-values of the population units. Then it is quite reasonable to consider the ratio-type estimator of $\bar{Y}$

$$t_2 = \bar{y}\frac{S_x^2}{s_x^2}$$

where

$$s_x^2 = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2.$$

Singh and Srivastava (1980) forwarded a sampling scheme [SS, say] in which a pair of units, (i, j) say, is selected with probability proportional to $(x_i - x_j)^2$ and then an SRSWOR sample of $(n-2)$ units is drawn from the $(N-2)$ that remain. That is, the probability of selecting a sample is

$$p(s) = \binom{N}{n}^{-1}\frac{s_x^2}{S_x^2}.$$

Cleary, for this design, t_2 provides an unbiased estimator for $\bar{Y}$.

For many populations encountered in practice, both $\bar{X}$ and S_x^2 are known in advance. Then one may think of using these parameters simultaneously and lead to consider a ratio-in-ratio estimator (which is of course biased)

$$t_3 = \bar{y}\frac{\bar{X}}{\bar{x}}\frac{S_x^2}{s_x^2}.$$

One basic question, not yet definitely answered, is how to make a choice between the parameters $\bar{X}$ or S_x^2 or both simultaneously, providing a good estimate of $\bar{y}$. The literature to date offers little guidance in this choice. The main reason perhaps being that one cannot easily evaluate the relative performance of one estimator over another. Because, the estimators considered above do not involve the same parameters in the resulting expressions of their variances or mean square errors. Also, in most of the cases the results are only in asymptotical forms which do not usually provide any meaningful conclusions preferably for smallish samples. So, the present investigation on the comparative performance is made with the help of an empirical study carried over a wide variety of natural populations.

2. STRATEGIES UNDER STUDY AND THEIR PERFORMANCE MEASURES

For the evaluation of the comparative performances, we now consider five competing strategies, viz $H_1 = (\text{SRSWOR}, t_1)$, $H_2 = (\text{SM}, t_1)$, $H_3 = (\text{SRSWOR}, t_2)$, $H_4 = (\text{SS}, t_2)$ and $H_5 = (\text{SRSWOR}, t_3)$. To further facilitate the comparison, especially with efficiency, we consider the conventional unbiased strategy $H_0 = (\text{SRSWOR}, \bar{y})$. The strategies are compared under the following performance measures:

(a) Relative Bias $(RB) = |Bias| / \bar{Y}$:

We leave aside the strategies H_2 and H_4 in this assessment since they are completely unbiased.

(b) Relative Efficiency (RE):

The relative efficiency of a strategy is calculated in comparison to the strategy H_0. For this purpose, mean square error or variance is taken as a measure of efficiency according as a strategy is biased or unbiased.

(c) Approach to Normality (Asymmetry):

The coefficients of skewness and kurtosis, i.e. β_1 and β_2 coefficients are considered as the indices for measuring the asymmetry of the sampling distribution of a strategy. We say asymmetricity is nullified when $\beta_1 = 0$ and $\beta_2 = 3$.

3. DESCRIPTION OF THE EMPIRICAL STUDY

Our empirical study considers 20 natural populations available in some traditional sampling texts and journals articles described in table 1. We draw all $\binom{N}{n}$ possible samples for $n = 3, 4$ and 5 from a given population and calculate the average behaviour of the estimators through different performance measures. To save space, the results for $n = 4$ are not given. However, the main findings are discussed in subsections 3.1 to 3.4.

3.1. Results Based on *RB*

The results in table 2 indicate that the *RB* of H_1 is the least in all the populations. It is also true for varying size of the sample. The strategy H_3 follows H_1 in almost all the cases except in population 11 for $n = 3$, and in populations 11 and 17 for $n = 5$. On the other hand, H_5 seems to be highly biased and has very poor performance in the sense of *RB*.

Table 1. *Description of populations*

Pop. n°	Source	Size (N)	Pop. n°	Source	Size (N)
1	Cochran (1977, p. 203)	10	11	Singh *et al.* (1986, p. 279)	20
2	Cochran (1977, p. 325)	12	12	Singh *et al.* (1986, p. 286)	16
3	Konijn (1973, p. 49)	16	13	Singh *et al.* (1986, p. 287)	12
4	Murthy (1967, p. 422)	24	14^{***}	Sukhatme *et al.* (1970, p. 185)	34
5	Singh *et al.* (1986, p. 144)	11	$15'$	Sukhatme *et al.* (1970, p. 185)	34
6	Singh *et al.* (1986, p. 155)	17	$16''$	Murthy (1967, p. 228)	32
7^{*}	Singh *et al.* (1986, p. 166)	16	$17'''$	Murthy (1967, p. 228)	32
8^{**}	Singh *et al.* (1986, p. 166)	16	18	Murthy (1967, p. 398)	32
9	Singh *et al.* (1986, p. 176)	13	19	Horvitz *et al.* (1952)	20
10	Singh *et al.* (1986, p. 176)	20	20	Sampford (1962, p. 61)	35

$^{*}x$ = area under wheat during 1978-79;
$^{**}x$ = total cultivated area during 1978-79;
$^{***}x$ = area under wheat in 1936;
$'x$ = total cultivated area in 1931;
$''x$ = n° of workers;
$'''x$ = fixed capital.

Table 2. *Features of the RB of strategies = |Bias|/Mean*

Pop.	$n = 3$			$n = 5$		
n°	H_1	H_3	H_5	H_1	H_3	H_5
1	0.0008	5.6959	5.6298	0.0003	0.3829	0.4005
2	0.0011	2.0859	2.1596	0.0005	0.3883	0.4099
3	0.0006	2.2992	2.3647	0.0003	0.2385	0.2755
4	0.0313	8.7504	20.358	0.0174	2.3678	2.8392
5	0.0131	1.3775	1.8390	0.0054	0.1909	0.2725
6	0.0006	3.1119	3.1378	0.0004	0.9403	0.9436
7	0.0069	4.6651	6.6123	0.0043	0.6373	1.1035
8	0.0158	4.3740	6.4912	0.0080	1.1520	1.9969
9	0.0242	7.8101	9.4719	0.0100	0.8412	0.9158
10	0.0460	12.975	16.398	0.0325	1.5978	3.0081
11	0.0006	7.2642	7.1822	0.0002	0.2522	0.1891
12	0.0002	4.8377	4.8564	0.0001	0.6161	0.6260
13	0.0159	1.4597	2.2027	0.0080	0.2632	0.4099
14	0.0089	9.9166	22.294	0.0023	1.0743	1.9404
15	0.0664	3.6774	4.0725	0.0360	0.3531	0.4757
16	0.0172	4.3482	6.7701	0.0059	1.0134	1.7833
17	0.0470	0.2536	0.6299	0.0287	0.1439	0.1182
18	0.0336	3.3065	4.5126	0.0193	0.4481	0.4912
19	0.0019	3.0279	3.1656	0.0009	0.6904	0.9251
20	0.0530	15.822	44.207	0.0205	7.7060	21.983

3.2. Results Based on *RE*

Table 3 reveals that the strategies H_3, H_4 and H_5 perform very badly in comparison to H_0 in view of efficiency. Both H_1 and H_2 fare well in relation to H_0 having noticeable performances in some populations like 1, 11 and 12 etc. H_2 is more efficient than H_1 in most of the cases except a very few, as for example in populations 11, 14 and 19

for $n = 3$, and in populations 6, 11, 14 and 18 for $n = 5$. However from $n = 3$ to $n = 5$, H_1 decreases weakly in *RE* for eight populations (3, 6, 7, 10, 14, 15, 18 and 19); also H_2 decreases weakly in *RE* from $n = 3$ to $n = 5$, for other seven populations (6, 7, 10, 13, 14, 15 and 18). But, for the rest of the twelve or thirteen natural populations, H_1 and H_2 (respectively) have bigger *RE* for $n = 5$ with respect $n = 3$.

Table 3. *Features of the RE of strategies w.r.t. H_0 (in %)*

Pop.	$n = 3$					$n = 5$				
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	1613	1639	0.0009	0.1487	0.0010	1636	1647	0.3194	1.0035	0.3050
2	155.00	157.23	0.0127	0.2755	0.0122	158.00	158.55	0.1569	0.6359	0.1435
3	653.15	653.20	0.0048	0.1607	0.0045	645.07	662.44	0.0942	0.5973	0.0795
4	779.82	1063	0.0057	4.4305	0.0008	1129	1186	0.5650	5.1541	0.4448
5	259.38	315.15	0.3843	5.9564	0.2618	302.65	330.68	7.0257	19.603	3.0228
6	131.58	132.21	0.0001	0.1969	0.0001	125.36	125.18	0.0006	0.0037	0.0006
7	1660	2050	0.0290	3.0787	0.0073	1608	1997	3.6131	11.828	1.2884
8	745.21	990.33	0.0576	3.5645	0.0310	773.64	995.88	1.0131	7.4796	0.3356
9	351.40	432.18	0.0024	1.2003	0.0019	386.78	433.88	1.5360	8.0791	1.3611
10	346.73	393.66	0.0041	1.1193	0.0030	237.43	330.28	2.2957	6.1539	0.1379
11	15324	15163	0.0002	0.1313	0.0003	19600	18195	1.1436	1.8824	1.3473
12	21366	21563	0.0017	0.2242	0.0018	22174	22284	0.1251	0.9129	0.1187
13	719.33	838.99	0.6454	13.439	0.3120	760.78	830.37	10.775	35.358	3.858
14	1618	1255	0.0012	2.8122	0.0016	986.89	920.75	1.1681	11.601	0.3320
15	571.11	590.67	0.0098	3.5286	0.0193	287.30	311.99	7.1565	22.604	4.5010
16	270.62	301.39	0.0012	0.8000	0.0005	621.97	734.83	0.0578	8.2344	0.0243
17	185.83	262.69	0.3355	2.9183	0.1267	535.12	891.82	26.591	13.875	19.117
18	132.03	152.14	0.0181	2.7558	0.0071	114.03	107.08	0.9121	6.5094	1.0867
19	446.11	426.88	0.0281	1.7332	0.0313	430.08	446.97	1.3172	4.9535	0.7936
20	252.31	330.51	0.0009	2.1083	0.0001	349.37	596.19	0.0275	4.4305	0.0033

Table 4. *Features of the coefficient of skewness*

Pop.			$n=3$					$n=5$		
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	0.0001	0.0002	63.70	4790	63.95	0.0097	0.0113	12.94	12.51	10.39
2	0.0281	0.0316	25.66	190.7	24.87	0.0026	0.0035	16.11	22.59	13.76
3	0.0185	0.0157	47.59	445.6	46.67	0.0053	0.9431	72.49	73.81	74.53
4	0.2258	0.1513	1077	25914	136953	0.0311	0.0142	32.67	37.24	12.26
5	0.3009	0.2805	43.74	241.5	22.91	0.0793	0.0766	16.59	12.14	34.29
6	0.0578	0.0565	81.38	827.0	82.92	0.0166	0.1387	38.22	27.48	37.58
7	0.0001	0.0002	60.97	1926	267.6	0.0024	0.0392	3.8595	4.8007	10.56
8	0.1808	0.3463	46.85	777.7	35.75	0.1539	0.2099	8.8371	23.43	15.01
9	0.0186	0.0021	186.8	34483	119.3	0.0259	0.0325	38.92	37.01	21.22
10	0.0030	0.0041	206.6	5776	125.7	0.0417	0.8553	57.25	10.32	302.4
11	0.0579	0.0515	702.8	43523	677.9	0.0012	1.0053	1.9188	1.5136	1.9405
12	0.0137	0.0111	162.3	3409	160.9	0.0040	0.1524	45.30	62.45	44.32
13	0.1036	0.1666	37.56	209.9	32.40	0.0923	0.1243	13.30	10.96	22.59
14	1.2850	2.0191	488.4	268192	602.6	0.7797	0.8653	9.6147	36.24	11.35
15	0.0042	0.8857	1241	137580	311.4	0.0259	1.3711	3.7598	6.1541	3.9093
16	0.0341	0.2554	254.7	47038	255.7	0.0089	1.7938	645.4	7400	656.7
17	0.4526	0.4448	196.8	43.99	440.8	0.1674	3.5797	8.2159	0.7102	16.22
18	0.7658	0.0728	259.1	5513	384.5	0.5208	1.0834	46.00	77.16	41.46
19	0.5121	0.4977	136.1	1447	109.8	0.3882	0.2067	7.3445	9.1215	7.2631
20	0.0068	0.3711	572.2	109294	656.1	0.0020	1.8820	21.39	1230	23.99

3.3. Results Based on Skewness

The results in table 4 show that the strategies H_3, H_4 and H_5 have very much skewed distributions for $n = 3$. As the sample size increases to 5 through 4, there is a sudden fall in the asymmetricity of their distributions. Even for moderate size sample they have erratic behaviour and their performance under this measure is not at all

convincing excepting a very few. On the other hand, the strategies H_1 and H_2 have their distributions near the symmetry. H_2 dominates H_1 in 8 populations for $n = 3$ and in only 3 populations for $n = 5$ in the sense of approaching towards symmetry. This means that H_1 would have a decidedly better performance over H_2 when sample is of moderate size.

Table 5. *Features of the coefficient of kurtosis*

Pop.	$n = 3$					$n = 5$				
n°	H_1	H_2	H_3	H_4	H_5	H_1	H_2	H_3	H_4	H_5
1	2.511	2.526	72.81	7135	73.17	2.576	2.583	25.45	30.88	21.75
2	2.583	2.599	32.01	339.6	31.26	2.566	2.571	29.29	47.85	25.15
3	2.598	2.594	59.28	869.2	57.70	2.608	2.630	101.4	208.3	103.7
4	3.730	4.188	1277	166953	1348	2.876	2.757	68.40	93.17	34.60
5	3.569	3.722	59.97	482.1	31.30	3.032	3.045	31.65	31.06	52.59
6	2.736	2.733	107.5	1867	109.7	2.646	2.691	84.73	77.95	83.77
7	3.132	3.288	67.99	3519	318.7	2.849	3.266	9.538	12.34	21.17
8	3.502	4.072	59.17	1550	45.16	3.255	3.845	16.62	46.85	29.93
9	3.125	3.040	207.1	58235	126.2	2.585	2.547	66.88	98.71	42.87
10	3.050	2.800	295.6	17439	174.1	2.825	3.041	130.4	37.40	421.7
11	2.617	2.586	792.8	128536	770.4	2.690	2.568	5.483	5.745	5.453
12	2.579	2.572	212.5	8364	211.0	2.657	2.599	65.60	156.3	63.59
13	2.793	2.898	45.34	416.1	42.02	2.805	2.905	27.08	27.36	34.10
14	5.480	4.976	561.2	533050	696.2	4.500	4.408	14.72	65.79	15.66
15	2.688	2.576	1404	272466	367.9	2.571	2.478	7.768	14.33	8.136
16	2.659	2.771	288.9	88816	288.9	2.565	1.938	768.8	26006	779.9
17	3.578	4.950	315.4	287.9	651.0	3.023	4.277	23.30	0.8021	40.96
18	3.675	3.285	315.1	13675	446.4	3.302	1.847	68.30	169.8	62.60
19	2.811	2.727	169.9	3582	138.5	2.733	2.729	16.73	21.13	14.97
20	2.465	3.344	658.8	375810	757.2	2.344	4.350	29.33	2089	32.82

3.4. Results Based on Kurtosis

The results in table 5 show that the strategies H_3, H_4 and H_5 have extremely leptokurtic type of distributions which lead to the conclusions that the strategies provide estimates highly concentrated around the parameter they are estimating. But with steady increase in the sample size, they considerably slow down their peakedness having distributions spectacularly above the normality. On the other hand, both H_1 and H_2 having their sampling distributions near the normality. They perform equally well in view of their approach to normality and this tendency is also nearly true even for moderate size of the sample. For $n = 5$, H_1 seems to be slightly better than H_2.

4. CONCLUSIONS

Since our study is realized for 20 particular populations, the conclusions following are justified for these concrete natural data. The present empirical study leads to the overall conclusions that the performances of the strategies H_3, H_4 and H_5 are decidedly very much unsatisfactory under different measures. H_1 has the least bias among the biased strategies. The unbiased strategy H_2 is more efficient than H_1 in most of the cases. But in view of asymmetry, H_1 seems to be better than H_2 although they are almost compatible when considered under their approach to normality. The study thus suggests a straightforward rejection of H_3, H_4 and H_5. H_2 is advisable to be preferred over H_1 from efficiency viewpoint. The only demerit that lies with H_2 is that it may assume negative variance estimators for certain samples. But, a sufficient condition that H_2 will provide non-negative variance estimator is given in a recent paper by Sahoo and Sahoo (1995). One can then choose a sample just fulfilling the requirements of such a condition.

5. REFERENCES

[1] **Cochran, W.G.** (1977). *Sampling Techniques.* Third Edition, Wiley Eastern Limited, New Delhi.

[2] **Horvitz, D.G.** and **Thompson, D.J.** (1952). «A generalization of sampling without replacement from a finite universe». *J. Amer. Statist. Assoc.,* **47**, 663-685.

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

[4] **Midzuno, H.** (1952). «On the sampling system with probability proportionate to sum of sizes». *Ann. Inst. Statist. Math.,* **3**, 99-107.

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

[6] **Sahoo, J.** and **Sahoo, L.N.** (1995). *On three unbiased strategies in sample surveys. Unpublished manuscript.*

[7] **Sampford, M.R.** (1962). *An Introduction to Sampling Theory.* Oliver and Boyd, Edinburg.

[8] **Sen, A.R.** (1952). «Present status of probability sampling and its use in estimation in farm characteristics (Abstract)». *Econometrica*, **20**, 130.

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

[10] **Singh, P.** and **Srivastava, A.K.** (1980). «Sampling schemes providing unbiased regression estimators». *Biometrika*, **67**, 205-209.

[11] **Sukhatme, P.V.** and **Sukhatme, B.V.** (1970). *Sampling Theory of Surveys with Applications.* Second Revised Edition, Iowa State University Press, Ames.