SYSTEMATIC SAMPLING USING VARYING PROBABILITIES

BY

RANJANA AGRAWAL[1], D. SINGH[2] AND PADAM SINGH[3]

(Received : March, 1982)

SUMMARY

A modified systematic sampling scheme is obtained by combining the concept af random interval with the use of unequal selection probabilities. The suggested scheme enjoys the simplicity of systematic sampling and is free from the drawback of non-estimability of variance. Three sets of probabilities for selecting sampling interval have been discussed. Empirical comparison with SRS and usual circular systematic sampling indicated that the suggested scheme can be used as an alternative to systematic sampling scheme.

INTRODUCTION

Systematic sampling is preferred over simple randomsampling due to simplicity of selection and efficiency of the estimator. But it has an obvious drawback of non-estimability of the variance of the estimator because inclusion probabilities for all pairs are not non-zero. Das [1] introduced the concept of random interval and thus removed the above drawback. But his scheme results in efficiency equal to that of SRS. In the present paper a modified scheme has been suggested which enjoys advantage of simplicity and efficiency over SRS and provides unbiased variance estimator. The scheme has been compared with SRS and usual systematic sampling empirically. The scheme is applicable to population sizes which are prime numbers.

Present address : 1. IASRI, New Delhi 12. B-4/126, Paschim Vihar, New Delhi 3. Planning Commision, New Delhi,

2. Sampling Scheme

Let there be $N(=2m+1)$ distinct and identifiable units in the population and a sample of size n is desired to be drawn from it. Let Y_i and y_i denote value of the character under study for i-th unit in the population and sample respectively.

The sampling scheme consists of the following steps

(*a*) Select a random number r from 1 to N with equal probabilitiy

(*b*) Select a random interval L from 1 to m with probability p_L

(*c*) Starting from r-th unit, select every L-th unit circularly till the sample of size n is selected.

It may be noted that as N is a prime number, there is no chance of repeatition for any unit in the sample.

3. Inclusion Probabilities

This scheme provides equal inclusion probability $\left(=\frac{n}{N}\right)$ to each unit of the population.

The inclusion probability for the pair of unit (i, j) can be obtained as

$$\pi_{ij} = \sum_{i<j}^{m} \sum_{L=1} p_L \, \pi_{ijL}$$

Here π_{ijL} denotes inclusion probability of the pair of units (i,j) given that random interval L has been selected. Again

$$\pi_{ijL} = \sum_{t=1}^{n-1} \pi_{ijL\,t}$$

where

$$\pi_{ijLt} = \frac{1}{N} \text{ if } j-i=Lt \text{ or } N-(j-i)=Lt$$
$$= 0 \text{ otherwise}$$

4. Estimation Procedure

Using inclusion probabilities and the values of the units in in the sample, Horvitz Thompson estimator of population total reduces to

$$\hat{Y}_{HT} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

The variance of $\hat{Y}_{HT}$ and variance estimator are respectively given by

$$V(\hat{Y}_{HT}) = \frac{1}{n^2} \sum_{i=1}^{N} \sum_{j>i}^{N} (n^2 - N^2\pi_{ij}) (Y_i - Y_j)^2$$

$$\hat{V}(\hat{Y}_{HT}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{n^2 - N^2\pi_{ij}}{\pi_{ij}} (y_i - y_j)^2$$

5 Probabilities for Selection of Interval

The probability for selecting sampling interval $L(p_L)$ ean be chosen in number of ways. In this paper three probability sets are considered which are discussed one by one.

5.1 *Set 1* : The probability of selecting the interval L can be taken in proportion of L

$$p_L \propto L, \; L = 1, 2, \ldots\ldots, m$$

$$= \frac{2L}{m(m+1)}$$

5.2 *Set 2.* : The probability of taking interval 'L' may be some value 'c' for $L=K$ and equally distributed for the remaining

$$p_L = c \text{ for } L = K$$

$$= \frac{1-c}{m-1} \text{ otherwise}$$

where K is the interval in the usual circular systematic sample for selecting a sample of size n from N and c is some arbitrarily chosen probability. The scheme reduces to SRS if $c = \frac{1}{m}$ and to systematic sampling if $c=1$. Obviously the larger the value assigned to c, the closer will be the efficiency of the proposed scheme to that of circular systematic sampling.

5.3. *Set 3.* The probability for $L=K$ may be taken as some value c and in proportion to L for the remaining

$$p_L = c \text{ for } L = K$$

$$= \frac{2L(1-c)}{m(m+1) - 2K} \text{ otherwise}$$

It is expected that with increase in c, the efficiency under the proposed scheme will be close to that of systematic sampling scheme.

6. COMPARISON

Variance of population total can also be expressed in the form

$$V(\hat{\bar{Y}}) = N \sum_{L=1}^{m} p_L \sum_{i=1}^{N} (\bar{Y}_{iL} - \bar{Y})^2$$

where $\bar{Y}_{iL}$ is the sample mean of i-th sample with interval L.

It can be further written as

$$V(\hat{\bar{Y}}) = N p_K \sum_{i=1}^{N} (\bar{Y}_{ik} - \bar{Y})^2 + N \sum_{L \neq K}^{m} p_L \sum_{i=1}^{N} (\bar{Y}_{iL} - \bar{Y})^2$$

$$= p_k V_{sys}(\hat{\bar{Y}}) + N \sum_{K \neq L}^{m} p_L \sum_{i=1}^{N} (\bar{Y}_{iL} - \bar{Y})^2$$

It can be seen from the above expression that the variance under proposed scheme will be closer to that under usual circular systematic sampling as p_k is increased and will be equal to that under systematic sampling for $p_k = 1$.

Above expression does not simplify further in general. But for Set 2, the above expression simplifies to

$$V(\hat{\bar{Y}}) = c V_{sys}(\hat{\bar{Y}}) + N \sum_{L \neq K}^{m} \frac{1-c}{m-1} \sum_{i=1}^{N} (\bar{Y}_{iL} - \bar{Y})^2$$

$$= \frac{cm-1}{m-1} V_{sys}(\hat{\bar{Y}}) + \frac{m(1-c)}{m-1} V_{SRS}(\hat{\bar{Y}})$$

It is evident from the above expression that the variance under Set 2 will be closer to that of circular systematic sampling for large value of c and to *SRS* for smaller value of c. For $c = 1$ and $\frac{1}{m}$, the variance will reduce to that of systematic and simple random sampling respectively.

EMPIRICAL COMPARISON FOR CERTAIN SPECIFIED POPULATIONS

It is not possible to compare the scheme under probability Set 1 and 3 with usual circular systematic sampling in general. However the scheme has been compared with SRS and usual circular systematic sampling empirically for certain artificially constructed and natural populations.

Following schemes have been considered for comparison.

1. Simple random sampling
2. Systematic Sampling
3. Suggested scheme with probability proportional to interval (set 1).
4. Suggested schme with probability set 2 (c = .80, .85, .90 & .95)
5. Suggested schcme with probability set 3 (c = .80, .85, .90 & .95)

The comparison for different populations is discussed in sections 7.1 to 7.4.

7.1. Population with Linear Trend

Let us consider a population of type

$$Y = \mu + \theta\, h$$

For simplicity let us take $\mu = 0$ and $h = 1$. For such populations, the variances for estimate of populattion total for above schemes denoted by V_1, V_2, V_3 for schemes 1, 2 & 3; V_4 to V_7 for scheme 4 and V_8 to V_{11} for scheme 5 are presented in Table 1 for different population and sample sizes.

Perusal of the table indicates that the performance of the suggested scheme with probability set 3 is at par with that of usual circular systematic sampling.

7.2 Populations with Periodic Trend

Three populations generated by Sine Curve

$$Y_i = m + a \text{ Sin } \frac{\pi i}{2}$$

have been considered. For these populations variances of estimate of population total under various schemes (denoted by V_1 to V_{11}) are given in Table 2.

The results reveal that the performance of suggested scheme with probability set 2 and 3 is at par and near to that of usual circular systematic sampling.

TABLE 1

Variance under different sampling schemes for populations with linear trend

N	n	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
5	3	8.33	5.56	7.41	6.67	6 38	6.11	5.83	6.67	6.38	6.11	5.83
7	3	43.56	21.78	36.30	28 31	26.68	25,64	23.41	26.13	25.04	23.96	22.87
7	4	24.50	12.25	20.42	15.93	15.01	14.09	13.17	14.70	14.09	13.47	12.86
11	3	322.67	134.44	250.96	181.50	169.74	157.97	146.21	166.22	158.28	150.33	142.39
11	4	211.75	75.62	162.34	109.66	101.15	92.64	84.13	97.30	91.88	86.46	81.04
11	5	145.20	48.40	112.26	72 60	66.55	60.56	54.45	63.14	59.46	55.77	52.08
11	6	100.83	33.61	77.98	50.42	46.22	42 01	37.81	43.85	41.29	38.73	36,17
13	3	657.22	262.89	500.74	357.53	333 87	310 21	286.55	321.65	306.96	292.27	277.58
13	5	315.47	135 20	232.42	178.46	167.65	156.83	146.02	157.88	152.21	146.54	140.87
13	7	168.99	48.28	126.13	77.26	70.01	62.77	55.52	65.49	61.19	56.89	52.59

TABLE 2

Variance under different sampling schemes for populations having periodic trend

N	m	a	n	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
17	2	1	3	44.33	16.00	49.58	22.48	20.86	19.24	17.62	24.05	22.04	20.03	18.01
17	2	1	5	22.80	11.92	24.84	14.41	13.78	13.16	12.54	14.74	14.03	13.32	12.62
17	2	1	7	13.57	2.12	13.40	4.74	4.08	3.43	2.78	4.51	3.91	3.31	2.72
17	5	2	3	177.33	64.00	198.32	89.90	83.43	76.95	70.48	96.24	88.18	80.12	72.06
17	5	2	5	91.20	47.68	99.36	57.63	55.14	52.65	50.17	58.96	5 .14	53.32	50.50
17	5	2	7	54.28	8.49	53.59	18.96	16.34	13.72	11.11	18.04	15.65	13.26	10.88
19	7	2	4	158.33	85.50	16.34	101.89	97.79	93.69	89.66	107.06	101.67	96.28	90.89
19	7	2	6	91.48	54.89	88.25	63.12	61.06	59.00	56.95	62.05	60.26	58.47	56.68
19	7	2	8	58.06	9.50	51.78	20.42	17.69	14.96	12.23	18.34	16.14	13.92	11.71

7.3 Auto Correlated Populations

Three populations having auto-correlation have been considered for this case. The three populations on American meat consumption from 1919 to 1941 (POP_1), Price of steers in Lbs at Cicago (POP_2) and milk price (POP_3) are given in Table 3.

TABLE 3

Populations with Auto-correlation

S. Na.	POP_1	POP_2	POP_3
1	171.5	99	99
2	167.0	98	97
3	164.5	97	94
4	169.3	96	91
5	179.4	95	89
6	179.2	94	88
7	172.6	93	88
8	170.5	93	89
9	168.6	92	90
10	164.7	92	91
11	163.0	91	92
12	162.1	90	92
13	160.2	89	90
14	161.2	88	88
15	165.8	87	85
16	163.5	86	83
17	146.7	85	81
18	160.2		
19	156.8		
20	156.8		
21	165.4		
22	174.7		
23	178.7		

The variances of estimate of population total for these populations under various schemes are given in Table 4.

TABLE 4

Variances under different sampling schemes for auto-correlated populations

Populations	n	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
P_1	4	6800.99	2488.85	4985.51	3437.51	3200.35	2963.19	2726.02	3038.11	2900.80	2763.48	2626.16
	6	4056.75	798.33	2952.75	1515.18	1335.96	1156.76	977.55	1257.01	1142.34	1027.67	913.00
	8	2684.60	538.16	1982.88	1010.37	892.32	774.27	656.21	840.86	765.19	689.51	613.84
P_2	3	1392.99	560.00	1097.90	750.40	702.80	655.20	607.60	689.10	656.82	624.55	592.27
	5	716.40	253.99	552.82	359.69	333.27	306.85	280.42	319.20	302.90	286.60	270.30
	7	426.43	229.02	339.06	274.14	262.86	251.58	240.30	252.32	246.50	240.67	234.85
P_3	3	1608.82	1228.23	1722.79	1315.21	1293.46	1271.73	1249.96	1346.92	1317.24	1287.57	1257.89
	5	827.40	384.72	970.84	485.90	460.61	435.31	410.02	512.60	480.63	448.66	416.69
	7	492.50	277.92	635.44	326.44	314.70	302.44	290.18	353.63	334.70	315.77	296.89

TABLE 5

Variances under different sampling schemes for natura l populations

Population	*n*	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
P_1	4	10.06	7.79	10.81	8.30	8.18	8.05	7.92	8.48	8.30	8.12	7.96
	6	5.81	4.61	6.34	4.88	4.82	4.75	4.68	4.98	4.89	4 80	4.71
	8	3.69	4.00	4.20	3.93	3.95	3.98	3.99	4.04	4.03	4.02	4.01
P_2	4	16.55	10.88	17.23	12.15	11.83	11.52	11.20	12.42	12.03	11.65	11.26
	6	9.56	8.30	10.48	8.58	8.52	8.44	8.37	8.77	8.65	8.53	8.42
	8	6.07	6.22	6.96	6.19	6.19	6.20	6.21	6.37	6.34	6.30	6.26
P_3	3	19.48	13.92	17.05	15.19	14.87	14.55	14.23	14.67	14.48	14.29	14.10
	5	10.01	6 91	8.34	7.62	7.44	7.26	7.09	7 22	7.14	7.07	6.99
	7	5.96	7.10	4.70	6.83	6.90	6.97	7.03	6.59	6.72	6.84	6.97
P_4	3	30.42	19.36	2o.65	21.89	21.26	23.63	19.99	21.11	20.67	20.24	19.80
	5	15.64	10.94	13.18	12.01	11.74	11.48	11.21	11.43	11.30	11.18	11.06
	7	9.31	8.99	7.41	9.06	9.04	9.03	9.01	8.66	8.74	8.82	8.91

The results indicate that the performance of the suggested scheme with probability set 3 is best and at par with usual systematic sampling. The scheme with probability set 2 is slightly inferior to that with probability set 3. The scheme with probability set 1, however, did not show much advantage. The results are similar to those obtained for populations with linear trend.

7.4. Natural Populations

Four human populations in various districts taken by Hanurav [2] have been considered. The first two populations each of size 19 give human population (in millions) in various districts of Punjab in 1951 and 1961. The third and fourth populations correspond to similar data for Gujarat (population of size 17). The variances of estimate of population total for these populations denoted by V_1 to V_{11} under various schemes are given in Table 5.

Perusal of the table indicates that the suggested scheme with probability set 2 and 3 performed at par and were near to the usual circular systematic sampling scheme. The scheme with probability set 1 proves to be advantageous only in few cases. For such population SRS vaniance was smaller than other variance when the sample size was large for three out of four populations.

Summing up, the suggested scheme with probability set 3 can be used in place of usual systematic sampling for certain types of population It provides efficiency at par with that of usual circular systematic sampling and is free from the darwback of non-estimability of variance estimator.

References

[1] Das, M.N. (1980) : Systematic sampling without any drawback. Indian Statistical Institute (Delhi Cantre), Technical Report No. 8206.

[2] Hanurav, T.V. (1967) : Optimum utilisation of auxiliary information, πPS sampling of two units from a stratum. *J.R. Statist. Soc.* 1329, 374-91.

[3] Horvitz, D.G. and Thompson, D.J. (1952) : A generalisation of sampling without replacement from a finite universe *J. Am. Statist. Assoc.* 47, 663-85.