

Estimation of Mean of Rare and Clustered Population using Edge Units in the Networks

Jayant K. Kshirsagar

Dept. of Statistics, NAC&S College, Ahmednagar, India

Sharad D. Gore

Prof. & Ex-Head, Dept. of Statistics, SPPU, Pune, India

Abstract- Adaptive cluster sampling (ACS) can provide efficient unbiased estimators of population mean and population total of rare and clustered population (see Thompson and Seber, 1996). These estimators are based on the networks that consider the unit not satisfying the predefined condition as the edge units. These edge units are identified during the adaptation stage, when the initial sample is expanded around sampling units that satisfy predefined condition. In this paper, we propose two new estimators of the population mean based on the networks that consider the units satisfying the predefined condition as the edge units. These edge units are identified during adaptation stage when the initial sample is expanded around sampling units that do not satisfy the predefined condition. The proposed estimators use these edge units in the estimation procedure. We have obtained unbiased estimators of the variances of these new estimators. The efficiencies of the new estimators are compared with those of Hansen- Hurwitz type estimator and Horvitz- Thompson type estimator using real life data.

Keywords: ACS, Network sampling, edge unit, inclusion probability, Hansen – Hurwitz type estimator, Horvitz - Thompson type estimator.

I. INTRODUCTION

The traditional sampling designs fail to detect and represent the aggregation pattern in a rare and clustered population. ACS can provide efficient unbiased estimators of the population mean and population total of such populations (Thompson & Seber, 1996). These estimators are based on the networks that consider the units not satisfying the predefined condition as the edge units. These edge units are identified during the adaptation stage, when the initial sample is expanded around sampling units that satisfy the predefined condition. Thus in ACS the population is divided in clusters that are of the convex type.

In this paper, we have considered use of ACS for the population divided into several clustered of non-convex type. For this situation, we propose two new unbiased estimators of the population mean based on the networks that consider the units satisfying the predefined condition as the edge units. These units are identified during the adaptation stage, when the initial sample is expanded around sampling units that do not satisfy the predefined condition. The proposed estimators use edge units in the estimation procedure. We have obtained unbiased estimators of the variances of these new estimators. The efficiencies of the new estimators are compared with those of Hansen-Hurwitz type (HH) estimator and Horvitz-Thompson type (HT) estimator using real life data.

II. SAMPLING DESIGNS

2.1 Adaptive Cluster Sampling (Without using edge units in estimation) :-

Consider a rare and clustered population that is divided into N rectangular grid points of equal sizes. Each grid point serves as a sampling unit. Let Y be the binary variable characteristic that takes value 1 if a particular grid point satisfies the predefined condition C and 0 otherwise. An ACS design used for the above type of population usually includes the following steps:

- (1) n initial units are selected by SRSWOR from the above population.
- (2) The neighbors (that is top, bottom, right and left units) of an initial sampling unit are sampled if it satisfies the predefined condition C . Further the neighbors of these units are also selected if they satisfy condition C . This is continued till all the desirable neighbors are sampled. The group of adjacent units which satisfy the condition C is called as a network. The adjacent units which are the outermost units in the network that do not satisfy the condition C are called as the edge units. They are not used in estimation procedure. Thompson & Seber have provided the following two unbiased estimators of population mean along with the unbiased estimators of variances of these two estimators.
 - I) Hansen – Hurwitz (HH) type estimator

$$\mu_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i} \sum_{j \in \Psi_i} y_j$$

Where

y_j : Density of unit j with respect to variable Y .

Ψ_i : i th network

m_i : Number of units in Ψ_i

n : Number of networks.

Unbiased estimator of the variance of this estimator is given as:

$$\tilde{V}(\mu_{HH}) = \frac{N-n}{N(n-1)} \sum_{i=1}^n (m_i - \mu_{HH})^2$$

Where $w_i = \frac{1}{m_i} \sum_{j \in \Psi_i} y_j$

II) Horvitz- Thompson (HT) type estimator.

$$\mu_{HT} = \frac{1}{N} \sum_{k=1}^K \frac{y_k z_k}{\alpha_k}$$

Where K : Total number of distinct networks in the population.

y_k : Sum of Y values for the k th network.

$$Z_k = \begin{cases} 1 & \text{if any unit of } k^{\text{th}} \text{ network is an initial sampling unit.} \\ 0 & \text{otherwise.} \end{cases}$$

α_k = Probability that a unit is included in the k^{th} network

$$= 1 - (N - x_k \binom{N-1}{n}) / (N \binom{N-1}{n})$$

x_k = Number of interior units in the k^{th} network.

Unbiased estimator of the variance of this estimator is given by

$$\tilde{V}(\mu_{HT}) = \frac{1}{N^2} \sum_{k=1}^K \sum_{h=1}^K \frac{y_k y_h z_k z_h (\alpha_{kh} - \alpha_k \alpha_h)}{\alpha_k \alpha_h}$$

Where

α_{kh} = The joint probability of two networks k and h being intersected in the initial sample.

$$= 1 - [(N - X_h \binom{N-1}{n}) + (N - X_k \binom{N-1}{n}) - (N - X_h - X_k \binom{N-1}{n})] / (N \binom{N-1}{n})$$

Note that $\alpha_{kk} = \alpha_k$

II.2. Adaptive Cluster Sampling (Using edge units in estimation):-

Consider the population used in II.1.

ACS design using edge units in estimation includes of the following steps:

- i) n initial units are selected by SRSWOR from the population.
- ii) The neighbors of initial sampling units which do not satisfy the predefined condition C are selected in the sample. This is continued till the desirable neighbors are sampled. The group of adjacent units which do not satisfy the condition C is called as a network. The adjacent units which are the outermost units in the network that satisfy condition C are now the edge units. They are used in estimation.

Based on the above method we propose the following new unbiased estimators of the population mean.

I) HH type estimator

$$\mu_1^* = \frac{1}{n} \sum_{i=1}^n w_i y_i$$

Where $w_i = \sum_{j \in \Psi_i} y_j / e_i$ $i = 1, 2, 3, \dots, n$.

Where y_j : Density of jth unit with respect to variable Y.

Ψ_i : i th network

e_i : Number of edge units in Ψ_i .

n : Number of networks.

Unbiased estimator of the variance of this estimator is given as:

$$\tilde{V}(\mu_1^*) = \frac{N-n}{N(n-1)} \sum_{i=1}^n (w_i - \mu_1^*)^2$$

II) HT Type estimator

Let K: Number of distinct networks in population.

Ψ_k : set of edge units in the k th network

x_k : Number of interior units that make up k th network.

e_k : number of units in Ψ_k

y_k^* : sum of y – values in the edge units of kth network.

$$= \sum_{j \in \Psi_k} y_j$$

α_k^* = the inclusion probability of network

$$= 1 - (N - x_k - e_k \binom{N-1}{n}) / (N \binom{N-1}{n})$$

Define $Z_k = 1$ if any unit of the k th network is an initial sampling unit.

$= 0$ otherwise

Then HT type estimator of population mean is given by :

$$\mu_2^* = \frac{1}{N} \sum_{k=1}^K \frac{Z_k y_k^*}{\alpha_k^*}$$

Unbiased estimator of the variance of this estimator is given by

$$\tilde{V}(\mu_2^*) = \frac{1}{N^2} \sum_{k=1}^K \sum_{h=1}^H \frac{y_{kh}^2 - \alpha_k \alpha_h}{\alpha_k \alpha_h}$$

Where $\alpha_{kh}^* = 1 - ((N - x_h - e_k \binom{N}{n}) + (N - x_h - e_h \binom{N}{n}) - (N - x_k - x_h - e_k - e_h \binom{N}{n})) / (N \binom{N}{n})$

Note that $\alpha_k^* = \alpha_k$


III. EXAMPLE

An area is divided into 100 rectangular quadrats of equal size for estimating the mean abundance of a particular species of animal.

	2	1	1			8	10	4	3
	5	*	3			2		*	2
	4		5			3			9
	1	2	4			*4	5	1	6
		4					9		
	1	*	2			2*			4
		3					3		

Fig. 1 The abundance of the species of animal in different quadrats.

* in a quadrat indicates its selection in the initial sample of ACS

 Indicates a sampling unit included in the network expanded around the initial sampling unit having abundance 2.

 Indicates a sampling unit included in the network of the initial sampling unit having abundance 4.

Using the sampling design II.1 for the above population we get 5 distinct networks such that: $m_1=12, m_2=1, m_3=1, m_4=1, m_5=1$

$$y_1=57, y_2=2, y_3=0, y_4=0, y_5=0$$

Hence

$$\alpha_1=0.48, \alpha_2=0.05: i=2, 3, 4, 5. \quad \alpha_{12}=0.02$$

I) $\hat{\mu}_{HH} = 1/5 (57/12+2/1+0/1+0/1+0/1) = 1.35$

$$W_1 = 4.75 \quad W_2=2 \quad W_3 = W_4 = W_5 = 0.$$

$$\hat{V}(\hat{\mu}_{HH}) = 0.57, \quad \hat{SE}(\hat{\mu}_{HH}) = 0.75$$


II) $\hat{\mu}_{HT} = 1/100 (57/0.48 + 2/0.05) = 1.59$

$$\hat{\mu}_{HT} = 0.79, \quad SE(\hat{\mu}_{HT}) = 0.89$$

If sampling design II.2 is used then layout of the networks gets changed.

	2	1	1			8	10	4	3
	5	*	3			2		*	2
	4		5			3			9
	1	2	4			*4	5	1	6
		4					9		
	1	*	2			2*			4
		3					3		

Fig.2 Layout if sampling design II.2 is used

 Indicates a sampling unit included in the network of the respective initial sampling units that do not satisfy the condition C.

In this case,

$$\begin{aligned} x_1=2, x_2=4, x_3=1, x_4=x_5=0, \\ e_1=6, e_2=8, e_3=4, e_4=e_5=1, \\ y_1=20, y_2=36, y_3=10, y_4=4, y_5=2 \\ w_1=3.67, w_2=4.5, w_3=4, w_4=2, w_5=2.5 \end{aligned}$$

Hence, $\mu_1^* = 3.33$

$$\hat{\mu}_1^* = 0.21, \quad SE(\hat{\mu}_1^*) = 0.46$$

Further $\alpha_1^* = 0.35, \alpha_2^* = 0.48, \alpha_3^* = 0.23, \alpha_4^* = \alpha_5^* = 0.05$

$$\mu_2^* = 2.53$$

$$\alpha_{12}^* = 0.15, \alpha_{13}^* = 0.07, \alpha_{14}^* = 0.02, \alpha_{15}^* = 0.02, \alpha_{23}^* = 0.10 \\ \alpha_{24}^* = 0.02, \alpha_{25}^* = 0.02, \alpha_{34}^* = 0.01, \alpha_{35}^* = 0.01, \alpha_{45}^* = 0.003$$

It gives

$$\hat{\mu}_2^* = 0.94, \quad SE(\hat{\mu}_2^*) = 0.97$$

IV. RESULTS AND DISCUSSION

Sampling Design	μ		$SE(\mu)$	
	HH	HT	HH	HT
II.1	1.35	1.59	0.75	0.89
II.2	3.33	2.53	0.46	0.97

In this case, sampling design II.1 gives more precise estimates of the population mean. But the ultimate sample size is 35. On the other hand if we use sample design II.2 the ultimate sample size is 27 which is less than that in design II.1. As the initial sample size increases this difference in the ultimate sample size is expected to grow much faster. If the cost of acquisition, collection, measurement is high then sampling design II.2 is preferable to II.1. But one has to take the risk of over estimation.

REFERENCES

- [1] "Adaptive cluster sampling" by Thompson S. K., JASA 85, 1050-1059, (1990)
- [2] "Sampling" by Thompson S. K., Johan Wiley & Sons Inc, New York, 339pp(1992):.
- [3] "Adaptive Sampling" Thompson S. K. and Sebar G.A.F., Johan Wiley & Sons Inc, New York, 265 pp(1996)