Chiang Mai J. Sci. 2013; 40(4) : 736-742
http://it.science.cmu.ac.th/ejournal/
Contributed Paper

Unequal Probability Inverse Adaptive Cluster Sampling

Prayad Sangngam*
Department of Statistics, Faculty of Science, Silpakorn University, Nakorn Pathom, Thailand.
*Author for correspondence; e-mail: prayad@su.ac.th

Received: 25 August 2011
Accepted: 8 August 2013

ABSTRACT

This paper incorporates an unequal probability inverse sampling with adaptive cluster sampling. Unbiased estimators of the population total and its variance are given. The proposed sampling design is compared with the inverse adaptive cluster sampling design using simulation study. The results indicate that the proposed sampling design can be more efficient than inverse adaptive cluster sampling design. In particular, when an auxiliary variable is highly correlated to the study variable, the estimator under the proposed sampling design produces large efficiency.

Keywords: unequal probability sampling, adaptive sampling, inverse sampling, unbiased estimator

1. INTRODUCTION

A rare population is a population in which only a few units exhibit the characteristics of interest. A problem encountered from a fixed sample size sampling for the population is that the sample might yield a zero estimate for the population mean or total. In some surveys, population units can be partitioned into two classes. Assume that the class in which a unit belongs is not known until the unit is observed. Inverse sampling is an efficient sampling design for estimating the parameters of these populations.

In inverse sampling, units are drawn until the fixed number of units with characteristics of interest is obtained. Usually, the purpose of the sample survey is to estimate the proportion of units of interest or to estimate the parameters of the whole population. Haldane [1] considered an inverse sampling with equal probability with replacement. An unbiased estimator of the proportion of units of interest and its variance were derived. However, an unbiased estimator of the variance was not given. Finney [2] proposed an unbiased estimator of the variance. Chistman and Lan [3] considered inverse sampling with and without replacement when the selection of units is with equal probability in each draw. An unbiased estimator of the population total and its variance were provided but an unbiased estimator of the variance was not given. Salehi and Seber [4] proved that Murthy's estimator can be applied to sequential sampling design. Using this approach, they obtained an unbiased estimator of the variance for the estimator given by [3]. Greco and Naddeo [5] considered an unequal probability inverse sampling with replacement. They provided an unbiased estimator of the population total and the corresponding unbiased variance estimator.

One way for sampling a rare population is by adaptive cluster sampling. Thompson [6] proposed adaptive cluster sampling design which has been shown that to be useful application for rare and clustered population. However, an initial sample for adaptive cluster sampling is commonly drawn by a classical fixed sample size design so that a final sample may not consist of a unit of interest. Chistman and Lan [3] considered inverse adaptive cluster sampling when the initial sample is selected by inverse sampling with equal probability. Salehi and Seber [7] proposed general inverse adaptive cluster sampling.

Commonly, if the probability of selection units is highly correlated to the study variable, unequal probability sampling design can be higher efficient than equal probability sampling design. This paper combines the unequal probability inverse sampling with adaptive cluster sampling. The parameter to be estimated is the population total. An unbiased estimator of the parameter and an unbiased variance estimator are derived. A comparison of the proposed sampling design to the inverse adaptive cluster sampling design is performed using simulation study.

2. Proposed Sampling

A finite population consists of N distinct units labeled 1,2,...,N with their associated study values y_1, y_2,...,y_N. Let y_i be an initial selection probability of the i-th unit. The parameter of interest is the population total,

$$\tau = \sum_{i=1}^{N} y_i .$$

Assume that population units are divided into two classes according to whether the study values satisfy a condition. A common form of the condition is {y : y>c} where c is a given constant. The class of units in which study values satisfy the condition is defined to be the class C. The notation $\bar{C}$ is the class of the remaining units. Each unit in the population is defined to have a neighborhood, a set of other units associated with that unit.

In the proposed sampling, an initial sample is drawn by unequal probability inverse sampling with replacement. The units are selected with unequal probability (z_i) with replacement until the initial sample consists of m units from the class C. For the sample units in class C, their neighborhoods are added to be sampled and observed. The procedure continues until no more units in the class C are found. Since the sampling begins with unequal probability inverse sampling and incorporates to adaptive cluster sampling, this sampling scheme is called unequal probability inverse adaptive cluster sampling. The final sample consists of the initial sample and all adaptively sample units. The initial sample s can be partitioned into two parts: parts s_C and $s_{\bar{C}}$ are the set of sample units from the class C and $\bar{C}$, respectively.

The set of units that is adaptively sampled as a result of the unit i-th being sampled and that is also the member of class C is called the network to which the i-th unit belongs. The units that are adaptively sampled but are in the class $\bar{C}$ are called edge units. By this way, if any unit in the i-th network is selected in the initial sample, all units in the network are sampled. From definition of network, the population can be divided into K mutually exclusive networks.

3. Parameter Estimation

Let n_0 denote the initial sample size and n be the final sample size. So the initial sample consists of m units from the class C and $n_0 - m$ units from class $\bar{C}$. Let ψ_k denote the set of units in the k-th network and m_k denote a number of units in the network. The total value of the study variable in the network ψ_k is $y_k^* = \sum_{j \in \psi_k} y_j$ and the probability of selection of that network is $z_k^* = \sum_{j \in \psi_k} z_j$. The parameter to be estimated can be written as $\tau = \sum_{i=1}^{N} y_i = \sum_{k=1}^{K} y_k^*$. Since the probability of any edge unit is included in the final sample is not known, some estimators included edge units will be biased [6]. So the edge units in the final sample are excluded from the estimation stage. The proposed unbiased estimator for the population total uses the sample units in the class $\bar{C}$ only when they are drawn to be the initial sample. The estimator is formed by modifying the unbiased estimator given by Greco and Naddeo [5].

Theorem 1 Under the proposed sampling design, an unbiased estimator of the population total is

$$\hat{\tau} = \hat{P}\bar{y}_C + \left(1-\hat{P}\right)\bar{y}_{\bar{C}}, \qquad (1)$$

where $\hat{P} = \frac{m-1}{n_0 - 1}$, $\bar{y}_C = \frac{1}{m}\sum_{i \in s_C} \frac{y_i^*}{z_i^*}$ and

$$\bar{y}_{\bar{C}} = \frac{1}{n_0 - m}\sum_{i \in s_{\bar{C}}} \frac{y_i^*}{z_i^*}.$$

Proof: Let w_i represent the new value of a study variable of the i-th unit in the k-th network, given by $w_i = \frac{z_i y_k^*}{z_k^*}$. The population total is

$$\tau = \sum_{i=1}^{N} w_i = \sum_{k=1}^{K}\sum_{j \in \psi_k}\left(\frac{z_j y_k^*}{z_k^*}\right) = \sum_{i=1}^{N} y_i .$$

Note that

$$\begin{aligned}\hat{\tau} &= \hat{P}\bar{y}_C + \left(1-\hat{P}\right)\bar{y}_{\bar{C}} \\ &= \hat{P}\left[\frac{1}{m}\sum_{i \in s_C}\frac{y_i^*}{z_i^*}\right] + \left(1-\hat{P}\right)\left[\frac{1}{n_0 - m}\sum_{i \in s_{\bar{C}}}\frac{y_i^*}{z_i^*}\right], \\ &= \hat{P}\left[\frac{1}{m}\sum_{i \in s_C}\frac{w_i}{z_i}\right] + \left(1-\hat{P}\right)\left[\frac{1}{n_0 - m}\sum_{i \in s_{\bar{C}}}\frac{w_i}{z_i}\right].\end{aligned}$$

The expectation of $\hat{\tau}$ is

$$E\left(\hat{\tau}\right) = E_1E_2\left[\hat{P}\left(\frac{1}{m}\sum_{i \in s_C}\frac{w_i}{z_i}\right)\right] + E_1E_2\left[\left(1-\hat{P}\right)\left(\frac{1}{n_0 - m}\sum_{i \in s_{\bar{C}}}\frac{w_i}{z_i}\right)\right].$$

The notation E_2 refers to the conditional expectation given the initial sample size n_0 and E_1 is the unconditional expectation taken under all possible initial samples. Under the initial sample, Greco and Naddeo [5] showed that for given the initial sample size n_0, the selection with the inverse sampling is exactly the same as the selection procedure with stratified sampling. The sample from two strata are independent and the selection probability for the i-th unit in the classes C and $\bar{C}$ are $z_i \big/ \sum_{j \in C} z_j$ and $z_i \big/ \sum_{j \in \bar{C}} z_j$, respectively. We have

$$E_2\left[\left(\frac{1}{m}\sum_{i \in s_C}\frac{w_i}{z_i}\right) | \, n_0\right] = \frac{\sum_{i \in C} w_i}{\sum_{i \in C} z_i} = \frac{\sum_{i \in C} y_i}{\sum_{i \in C} z_i} ,$$

$$E_2\left[\left(\frac{1}{n_0 - m}\sum_{i \in s_{\bar{C}}}\frac{w_i}{z_i}\right) | \, n_0\right] = \frac{\sum_{i \in \bar{C}} w_i}{\sum_{i \in \bar{C}} z_i} = \frac{\sum_{i \in \bar{C}} y_i}{1 - \sum_{i \in C} z_i}$$

and $E_1\left(\hat{P}\right) = \sum_{i \in C} z_i$. We obtain that

$$E\left(\hat{\tau}\right) = E_1\left[\hat{P}\, E_2\left(\frac{1}{m}\sum_{i \in s_C}\frac{w_i}{z_i}\right)\right]$$

$$+E_1\left[\left(1-\hat{P}\right)E_2\left(\frac{1}{n_0-m}\sum_{i\in s_{\bar{C}}}\frac{w_i}{z_i}\right)\right],$$

$$=\sum_{i\in C} z_i\left(\frac{\sum_{i\in C} y_i}{\sum_{i\in C} z_i}\right)+\left(1-\sum_{i\in C} z_i\right)\left(\frac{\sum_{i\in\bar{C}} y_i}{1-\sum_{i\in C} z_i}\right),$$

$$=\sum_{i\in C} y_i+\sum_{i\in\bar{C}} y_i=\tau\,.$$

Theorem 2 The variance of the estimator $\hat{\tau}$ for the parameter τ is

$$V\left(\hat{\tau}\right)=\left(\tau_C-\tau_{\bar{C}}\right)^2 V\left(\hat{P}\right)+\frac{\sigma_C^2}{m}E\left(\hat{P}^2\right)+\frac{\sigma_{\bar{C}}^2}{m-1}E\left[\hat{P}\left(1-\hat{P}\right)\right], \qquad (2)$$

where $\tau_C=\frac{\sum_{i\in C} y_i}{\sum_{i\in C} z_i}$, $\tau_{\bar{C}}=\frac{\sum_{i\in\bar{C}} y_i}{\sum_{i\in\bar{C}} z_i}$,

$\sigma_C^2=\frac{1}{z_C}\sum_{i\in C} z_i\left(\frac{y_i^*}{z_i^*}-\tau_C\right)^2$, $z_C=\sum_{i\in C} z_i$

$\sigma_{\bar{C}}^2=\frac{1}{z_{\bar{C}}}\sum_{i\in\bar{C}} z_i\left(\frac{y_i^*}{z_i^*}-\tau_{\bar{C}}\right)^2$, $z_{\bar{C}}=\sum_{i\in\bar{C}} z_i$ and $V(\hat{P})$ is the variance of $\hat{P}$ for the parameter z_C.

Proof: The variance of $\hat{\tau}$ is equal to

$$V\left(\hat{\tau}\right)=E_1V_2\left(\hat{\tau}\right)+V_1E_2\left(\hat{\tau}\right),$$

$$=E_1V_2\left[\hat{P}\bar{y}_C+\left(1-\hat{P}\right)\bar{y}_{\bar{C}}\right]$$

$$+V_1E_2\left[\hat{P}\bar{y}_C+\left(1-\hat{P}\right)\bar{y}_{\bar{C}}\right].$$

The symbols E_2 and V_2 denote the conditional expectation and variance given the initial sample size n_0, respectively. The notations E_1 and V_1 represent the unconditional expectation and variance taking over n_0, respectively. When the initial sample size n_0 is given, we have $E_2\left(\bar{y}_C\,|\,n_0\right)=\tau_C$, $E_2\left(\bar{y}_{\bar{C}}\,|\,n_0\right)=\tau_{\bar{C}}$,

$V_2\left(\bar{y}_C\,|\,n_0\right)=\frac{\sigma_C^2}{m}$ and

$V_2\left(\bar{y}_{\bar{C}}\,|\,n_0\right)=\frac{\sigma_{\bar{C}}^2}{n_0-m}$. Therefore,

$$V\left(\hat{\tau}\right)=E_1\left[\hat{P}^2\frac{\sigma_C^2}{m}+\left(1-\hat{P}\right)^2\frac{\sigma_{\bar{C}}^2}{n_0-m}\right]$$

$$+V_1\left[\hat{P}\tau_C+\left(1-\hat{P}_h\right)\tau_{\bar{C}}\right],$$

$$=\frac{\sigma_C^2}{m}E\left(\hat{P}^2\right)+\sigma_{\bar{C}}^2E\left[\frac{\left(1-\hat{P}\right)^2}{n_0-m}\right]$$

$$+V\left[\hat{P}\tau_C+\left(1-\hat{P}_h\right)\tau_{\bar{C}}\right],$$

$$=\frac{\sigma_C^2}{m}E\left(\hat{P}^2\right)+\sigma_{\bar{C}}^2E\left[\frac{\left(1-\hat{P}\right)^2}{n_0-m}\right]$$

$$+\left(\tau_C-\tau_{\bar{C}}\right)^2V\left(\hat{P}\right).$$

In addition, $\frac{\left(1-\hat{P}\right)^2}{n_0-m}=\frac{\hat{P}\left(1-\hat{P}\right)}{m-1}$, so we obtain the result.

Theorem 3 An unbiased estimator of the variance $V(\hat{\tau})$ is

$$\hat{V}\left(\hat{\tau}\right)=\left(\bar{y}_C-\bar{y}_{\bar{C}}\right)^2\hat{V}\left(\hat{P}\right)+\frac{s_C^2}{m}\hat{P}^*+\frac{s_{\bar{C}}^2}{m-1}\left(\hat{P}-\frac{m-1}{m-2}\hat{P}^*\right), \qquad (3)$$

where $\hat{P}^*=\frac{(m-1)(m-2)}{(n_0-1)(n_0-2)}$,

$s_C^2=\frac{1}{m-1}\sum_{i\in s_C}\left(\frac{y_i^*}{z_i^*}-\bar{y}_C\right)^2$,

$s_{\bar{C}}^2=\frac{1}{n_0-m-1}\sum_{i\in s_{\bar{C}}}\left(\frac{y_i^*}{z_i^*}-\bar{y}_C\right)^2$ and $\hat{V}(\hat{P})=\frac{\hat{P}\left(1-\hat{P}\right)}{n_0-2}$.

Proof: Consider

$$E\left[\hat{V}\left(\hat{\tau}\right)\right]=E\left[\left(\bar{y}_C-\bar{y}_{\bar{C}}\right)^2\hat{V}\left(\hat{P}\right)+\frac{s_C^2}{m}\hat{P}^*\right],$$

$$+E\left[\frac{s_{\bar{C}}^2}{m-1}\left(\hat{P}-\frac{m-1}{m-2}\hat{P}^*\right)\right]$$

$$=E_1E_2\left[\left(\bar{y}_C-\bar{y}_{\bar{C}}\right)^2\hat{V}\left(\hat{P}\right)+\frac{s_C^2}{m}\hat{P}^*\,|\,n_0\right]$$

$$+E_1E_2\left[\frac{s_{\bar{C}}^2}{m-1}\left(\hat{P}-\frac{m-1}{m-2}\hat{P}^*\right)|\,n_0\right].$$

Conditioning on the initial sample size, we have

$E_2(\bar{y}_C|n_0)=\tau_C,\ E_2(\bar{y}_{\bar{C}}|n_0)=\tau_{\bar{C}},$
$E_2(\bar{y}_C\bar{y}_{\bar{C}}|n_0)=\tau_C\tau_{\bar{C}},$
$E_2(s_C^2|n_0)=\sigma_C^2,\ E_2(s_{\bar{C}}^2|n_0)=\sigma_{\bar{C}}^2,$
$E_2\left[\bar{y}_C^2|n_0\right]=\tau_C^2+\frac{\sigma_C^2}{m}$

and $E_2\left[\bar{y}_{\bar{C}}^2|n_0\right]=\tau_{\bar{C}}^2+\frac{\sigma_{\bar{C}}^2}{n_0-m}$.

Moreover, $E_1\left[\hat{V}(\hat{P})\right]=V(\hat{P})$ and

$E_1(\hat{P}^*)=\left(\sum_{i\in C} z_i\right)^2=z_C^2$.

We obtain that

$$\begin{aligned}
E\left[\hat{V}(\hat{\tau})\right]&=E_1\left[(\tau_C-\tau_{\bar{C}})^2\hat{V}(\hat{P})\right]\\
&+\frac{\sigma_C^2}{m}E_1\left[\hat{P}^*+\hat{V}(\hat{P})\right]\\
&+E_1\left[\frac{\sigma_{\bar{C}}^2}{m-1}\left(\hat{P}-\frac{m-1}{m-2}\hat{P}^*\right)\right]\\
&+E_1\left[\frac{\sigma_{\bar{C}}^2}{n_0-m}\hat{V}(\hat{P})\right],\\
&=(\tau_C-\tau_{\bar{C}})^2V(\hat{P})+\frac{\sigma_C^2}{m}\left(z_C^2+V(\hat{P})\right)\\
&+\frac{\sigma_{\bar{C}}^2}{m-1}\left(z_C-\frac{m-1}{m-2}z_C^2\right)+\frac{\sigma_{\bar{C}}^2}{n_0-m}V(\hat{P}),\\
&=(\tau_C-\tau_{\bar{C}})^2V(\hat{P})+\frac{\sigma_C^2}{m}E(\hat{P}^2)\\
&+\frac{\sigma_{\bar{C}}^2}{m-1}E\left[\hat{P}(1-\hat{P})\right].
\end{aligned}$$

Note that the variance of the estimator will be small when the value of y_i^*/z_i^* is closed to $\tau_{\bar{C}}$ for units in class C and value of y_i^*/z_i^* is closed to $\tau_{\bar{C}}$ for units in class $\bar{C}$. If the selection probabilities are equal for every unit, the proposed sampling design is the inverse adaptive cluster sampling given by Christman and Lan [3]. In addition, the unbiased estimator of the parameter is equivalent to the expression given by [3].

4. SIMULATION STUDY

The ring-necked ducks data given by Smith *et al* [8] was used as the study population. The population consists of N=200 units. The number of ring-necked ducks is used as the study variable (y). Auxiliary variables (x's) correlated to the study variable are created with the coefficients of correlation equal to 0.1, 0.2, 0.5, 0.6, 0.8 and 0.9. For unequal probability inverse adaptive cluster sampling, the population units are selected by probabilities proportional to the auxiliary variable. Simulations of sampling from the population were carried out to study the properties of the unequal probability inverse adaptive cluster sampling (UIACS) compared to the inverse adaptive cluster sampling (IACS) given by Christman and Lan [3]. We chose the condition {y:y>0} for dividing the units into class C or $\bar{C}$. The numbers of initial sample units satisfying the condition (m) in the initial sample are 2, 4, 6 and 8. The neighborhood of a unit is defined as the set of the four adjacent units. The simulation consists of 50,000 replications. The population total (τ) was estimated for each sample. In each sampling design, the values of the estimates ($\hat{\tau}$), and the final sample size (n) were averaged. The averages were interpreted as expected values, i.e.,

$$\tilde{E}(\hat{\tau})=\frac{1}{50,000}\sum_{i=1}^{50,000}\hat{\tau}_i$$

and

$$\tilde{E}(n)=\frac{1}{50,000}\sum_{i=1}^{50,000}n_i .$$

The estimate of the variance is

$$\tilde{V}(\hat{\tau})=\frac{1}{50,000-1}\sum_{i=1}^{50,000}(\hat{\tau}_i-\tilde{\tau}),$$

where $\tilde{\tau}=\tilde{E}(\hat{\tau})$. The estimate of standard error is equal to the squared root of the estimated variance.

In the Table 1, the averaged values of the estimates are close to the true value of the population total (τ=23,333). This result complies with Theory 1.

In Table 2, the results indicate that the averaged values of the final sample size under UIACS are smaller than the values under IACS and their values increase as m increase. Under UIACS with given the number m, when the coefficients of correlation increase, the averaged values of the final sample sizes decrease. Table 3 shows that the unequal probability inverse adaptive cluster sampling design outperforms the inverse adaptive cluster sampling design. With given the number m, the estimates of standard errors of the estimators under UIACS decrease when the coefficients of correlation increase.

Table 1. The averages of estimates with vary numbers of initial sample units in class C.

m	IACS	UIACS					
		ρ = 0.1	ρ = 0.2	ρ = 0.5	ρ = 0.6	ρ = 0.8	ρ = 0.9
2	23,396.31	23,444.59	23,516.31	23,612.57	23,661.67	23,354.22	23,427.58
4	23,491.37	23,239.50	23,222.15	23,092.35	23,206.13	23,220.56	23,271.66
6	23,449.78	23,280.25	23,404.85	23,333.90	23,255.43	23,176.89	23,225.47
8	23,299.32	23,194.47	23,243.25	23,213.96	23,242.99	23,290.70	23,334.18

Table 2. The averages of final sample sizes with vary numbers of initial sample units in class .

m	IACS	UIACS					
		ρ = 0.1	ρ = 0.2	ρ = 0.5	ρ = 0.6	ρ = 0.8	ρ = 0.9
2	35.47	35.27	34.98	34.00	33.62	32.17	30.90
4	70.88	70.52	70.01	68.17	67.40	64.67	62.02
6	106.44	105.51	104.73	101.81	100.70	96.72	92.58
8	142.03	140.86	139.82	136.06	134.62	129.09	123.62

Table 3. The estimates of standard error of the estimators with vary numbers of initial sample units in class C.

m	IACS	UIACS					
		ρ = 0.1	ρ = 0.2	ρ = 0.5	ρ = 0.6	ρ = 0.8	ρ = 0.9
2	72,379.44	66,081.23	61,370.73	52,641.63	50,491.04	42,286.85	37,434.59
4	38,304.30	35,458.89	33,619.88	28,277.31	26,957.17	23,037.24	20,876.93
6	29,412.98	26,789.38	25,788.74	21,787.25	20,549.86	17,653.88	15,669.58
8	24,374.67	22,484.54	21,414.35	18,247.97	17,274.90	14,871.70	13,355.46

5. Conclusion

An adaptive cluster sampling is an efficient sampling design for rare and clustered population. However, an initial sample in adaptive cluster sampling is commonly selected by fixed sample size design. This paper proposed an unequal probability inverse sampling to draw the initial sample from the population. The neighborhoods of sample units in the class of interest are added as in the adaptive cluster sampling. An unbiased estimator of the population total and an unbiased estimate of its variance are given. The simulation study showed that the efficiency of the proposed sampling design depends on the coefficient of correlation between the study variable and the auxiliary variable. When the auxiliary variable is highly correlated with the study variable, the unequal probability inverse adaptive cluster sampling design is more efficient than the inverse adaptive cluster sampling design. However, when the auxiliary variable is not appropriate for the study variable, an estimator of parameter by using the unequal probability inverse adaptive cluster sampling design may not always increase the efficiency.

Acknowledgements

This work was supported by the Faculty of Science, Silpakorn University, Nakorn Pathom, Thailand. I am grateful to referees and associate editors for their comments.

References

[1] Haldane J.B.S., On a method of estimating frequencies, *Biometrika*, 1945; **33:** 222–225.

[2] Finney D.J., On a method of estimating frequencies, *Biometrika*, 1949; **36(1)**: 223–234.

[3] Chistman M.C. and Lan F., Inverse adaptive cluster sampling, *Biometrics*, 2001; **57**: 1096-1105.

[4] Salehi M.M. and Seber G.A.F., A new proof of Murthy's estimator with applies to sequential sampling, *Aus. NZ. Stat.,* 2001; **43(3)**: 281-286.

[5] Greco L. and Naddeo N., Inverse sampling with unequal selection probabilities, *Communications in Statistics : Theory and Methods,* 2007; **36(5)**: 1039-1048.

[6] Thompson S.K., Adaptive cluster sampling, *J. Am. Stat. Assoc.,* 1990; **85**: 1050-1059.

[7] Salehi M.M. and Seber G.A.F., A general inverse sampling scheme and its application to adaptive cluster sampling, *Aus. NZ. J. Stat.*, 2004; **46(3)**: 483-494.

[8] Smith D.R., Conroy M.J. and Bralhage D.H., Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl, *Biometrics,* 1995; **51**: 777-788.