Thailand Statistician
January 2008; 6(1) : 15-26
http://statassoc.or.th
Contributed paper

An Improved the Estimator in Inverse Adaptive Cluster Sampling

Nipaporn Pochai
Department of Mathematics, Faculty of Science, Mahasarakham University, 44150, Thailand.
Author for correspondence; e-mail: j3832024@yahoo.com

Received: 23 June 2007
Accepted: 19 September 2007.

Abstract

Christman and Lan [1] considered adaptive cluster sampling based on inverse sampling that use two stopping rule. We use the estimator in adaptive cluster sampling based on the estimator of Dryver and Thompson [3] for improving the estimator in inverse adaptive cluster sampling that use three stopping rule. Estimators are compared in the small simulation. The results indicate that the improved estimator in inverse adaptive cluster sampling have the smallest variance.

Keyword: adaptive cluster sampling, inverse sampling, stopping rule.

1. Introduction

Inverse sampling is the method of sampling which the sample of units is selected at random. This method is continued until certain pre-specified conditions have been fulfilled. Christman and Lan [1] applied inverse sampling design to the rare population. They considered inverse sampling design that stopping rules based on the number of rare units: that is, the initial sample of units size n_0 is selected by simple random sampling. If the number of units whose value of the variable of interest satisfies a pre-specified condition grater than or equal to a predetermined numbers, say k, we stop sampling; otherwise we keep sampling unit until k rare units are observed. They

also incorporated adaptive cluster sampling [7,8], which is an efficient method for sampling rare and hidden clustered populations, with the inverse sampling. In this paper, we will study the improved unbiased estimator [2,3] in inverse adaptive cluster sampling. Some comparisons are made using a simulation.

2. Background of Inverse Sampling Designs

As in the finite population sampling situation, the population consists of N units $P=\{u_1,u_2,\ldots,u_N\}$ index by their labels $S=\{1,2,\ldots,N\}$. With unit i is associated a variable of interest y_i. The population is divided into 2 subgroups according to whether the y-values satisfies a pre-specifies condition C, for example $\{i, y_i \geq C\}$. The 2 subgroups are denoted as $P_M=\{u: y_i \in C, i=1,2,\ldots,N\}$ and $P_{N-M}=\{u: y_i \notin C, i=1,2,\ldots,N\}$ where M is the unknown number of units. The subgroups to which a unit belongs are not known until the unit is sampled. A unit is selected at random from the population and the sampling is sequential, and then the y-value is obtained. Sampling is continued until a predetermined number of k unit $(1<k\leq M)$ from P_M are sampled.

The estimator of the population total $\tau_y=\sum_{i=1}^{N} y_i$ can be written as [1]

$$\hat{\tau}_y = M\bar{y}_M + (N-M)\bar{y}_{N-M} \quad , \tag{1}$$

where $\bar{y}_M=\frac{1}{k}\sum_{i\in S_M} y_i$; S_M is the index label that are member of P_M, $\bar{y}_{N-M}=\frac{1}{n_1-k}\sum_{i\in S_{N-M}} y_i$; S_{N-M} is the index label that are member of P_{N-M}, n_1 is the total sequential sample size.

In applications M will be not known. The unbiased estimator of M [1] is $\hat{M}=\frac{N(k-1)}{(n_1-1)}$; $k>1$. So an unbiased estimator of τ_y is

$$\hat{\tau}_I = \hat{M}\bar{y}_M + (N-\hat{M})\bar{y}_{N-M}. \tag{2}$$

Its variance is given by

$$\begin{aligned}\mathrm{Var}\left(\hat{\tau}_I\right) &= \mathrm{E}_{n_1}\left[\hat{M}^2\mathrm{Var}\left(\bar{y}_M\middle|n_1\right)+\left(N-\hat{M}\right)^2\mathrm{Var}\left(\bar{y}_{N-M}\middle|n_1\right)\right] \\ &\quad + \mathrm{Var}_{n_1}\left(\hat{M}\mu_M+\left(N-\hat{M}\right)\mu_{N-M}\right), \\ &= \frac{\sigma_M^2}{k}\mathrm{E}_{n_1}\left[\hat{M}^2\left(1-\frac{k}{M}\right)\right]+\sigma_{N-M}^2\mathrm{E}_{n_1}\left[\frac{\left(N-\hat{M}\right)^2}{n_1-k}\left(1-\frac{\left(n_1-k\right)}{N-M}\right)\right] \\ &\quad +\left(\mu_M-\mu_N-M\right)^2\mathrm{Var}_{n_1}\left(\hat{M}\right),\end{aligned} \tag{3}$$

where $\mu_M = M^{-1}\sum_{i\in U_M} y_i$, $\sigma_M^2 = M^{-1}\sum_{i\in U_M}\left(y_i-\mu_M\right)^2$,

$\mu_{N-M} = \left(N-M\right)^{-1}\sum_{i\in U_{N-M}} y_i$ and

$\sigma_{N-M}^2 = \left(N-M\right)^{-1}\sum_{i\in U_{N-M}}\left(y_i-\mu_{N-M}\right)^2$. $U_M\left(U_{N-M}\right)$ is the index set for the subpopulation $P_M\left(P_{N\text{-}M}\right)$. The variance of $\hat{M}$ is difficult to determine but has been shown to be bounded [4]:

$$\frac{M^2\left(1-M/N\right)}{k} \le \mathrm{Var}_{n_1}\left(\hat{M}\right) \le \frac{M^2\left(1-M/N\right)}{k+M/N-2}.$$

Christman and Lan [1], suppose a sample size n_0 is selected by simple random sampling. The stopping rule is to stop after the sample of size n_0 selections if any members of P_M are observed in the sample. Otherwise sampling until continues until k units from P_M are sampled. In this stopping rule the unbiased estimator is

$$\hat{\tau}_{mix} = \begin{cases} \dfrac{N}{n_0}\displaystyle\sum_{i=1}^{n_0} y_i, & \text{if } n_1 = n_0 \\ \hat{M}\bar{y}_M+\left(N-\hat{M}\right)\bar{y}_{N-M}, & \text{if } n_1 > n_0 \end{cases} \tag{4}$$

The variance of $\hat{\tau}_{mix}$ is given by

$$\mathrm{Var}\left(\hat{\tau}_{\mathrm{mix}}\right)=\begin{cases} N^2\left(1-\dfrac{n_0}{N}\right)\dfrac{S_0^2}{n_0} \ , & \text{if } n_1=n_0 \\ \mathrm{Var}\left(\hat{\tau}_{\mathrm{I}}\right) \ , & \text{if } n_1>n_0 \end{cases} \tag{5}$$

where $S_0^2=\left(N-1\right)^2\sum_{i=1}^{N}\left(y_i-\mu\right)^2$ and $\mu=\sum_{i=1}^{N}y_i\big/N$

Salehi and Seber [5] introduced the general inverse sampling design which is more practical sampling design. Beginning a sample size n_0 is selected by simple random sampling, we stop further sampling if at least k units from P_M are selected. Otherwise sampling until continues until k units from P_M are sampled but a limit is put on final sample size, that is $n_2=N$.

This estimator is

$$\hat{\tau}_{\mathrm{gI}}=\begin{cases} \dfrac{N}{n_0}\sum_{i=1}^{n_0}y_i & \text{if } [Q_0], \\ \hat{M}\bar{y}_M+\left(N-\hat{M}\right)\bar{y}_{N-M} & \text{if } [Q_1], \\ \dfrac{N}{n_2}\sum_{i=1}^{n_2}y_i & \text{if } [Q_2], \end{cases} \tag{6}$$

where $[Q_0]=\left\{n_1=n_0\right\}$

$[Q_1]=\left\{n_0<n_1<n_2 \ \text{ or } \ n_1=n_2 \ \&\ \left|S_M\right|=k\right\}$

$[Q_2]=\left\{ \ n_1=n_2 \ \&\ \left|S_M\right|<k\right\}$.

An unbiased variance estimator is given by

$$\mathrm{Var}\left(\hat{\tau}_{\mathrm{gI}}\right)=\begin{cases} N^2\left(1-\dfrac{n_0}{N}\right)\dfrac{S_0^2}{n_0} \ , & \text{if } [Q_0], \\ \mathrm{Var}\left(\hat{\tau}_{\mathrm{I}}\right) \ , & \text{if } [Q_1], \\ N^2\left(1-\dfrac{n_2}{N}\right)\dfrac{S_2^2}{n_2} \ , & \text{if } [Q_2], \end{cases} \tag{7}$$

3. Inverse Adaptive Cluster Sampling Designs

Adaptive cluster sampling, proposed by Thompson [7], is an efficient method for sampling rare and hidden clustered populations. In adaptive cluster sampling, an initial sample of units is selected by simple random sampling. If the value of the variable of interest from a sampled unit satisfies a pre-specified condition C, that is $\{i, y_i \geq c\}$, then the unit's neighborhood will also be added to the sample. If any other units that were "adaptively" added also satisfy the condition C, i.e. $y \in P_M$, then their neighborhoods are also added to the sample. This process is continued until no more units that satisfy the condition are found. The set of all units selected and all neighboring units that satisfy the condition is called a network. The adaptive sample units did not satisfy the condition called edge units. If the initial unit does not satisfy the condition C, i.e. $y \in P_{N-M}$, no further units added, and the network consists of just the initial unit. Let A_i is the index set of all units in the network to which the i^{th} unit belongs. Let m_i be the number of units in A_i. Let ν denote the final sample size. The variable of interest associated with A_i is $y_i^* = \sum_{j \in A_i} y_j$ and $\overline{y}_i^* = y_i^* / m_i$ is the mean of the y-values in A_i. The combination of inverse sampling and adaptive cluster sampling is called inverse adaptive cluster sampling. Upon replacing each y_i with $\overline{y}_i^*$, an unbiased estimator of the population total is

$$\hat{\tau}_{gI,A} = \begin{cases} \dfrac{N}{n_0} \displaystyle\sum_{i=1}^{n_0} \overline{y}_i^* & \text{if } [Q_0], \\ \hat{M}\overline{y}_M^* + \left(N - \hat{M}\right)\overline{y}_{N-M} & \text{if } [Q_1], \\ \dfrac{N}{n_2} \displaystyle\sum_{i=1}^{n_2} \overline{y}_i^* & \text{if } [Q_2], \end{cases} \tag{8}$$

where $\overline{y}_M^* = \dfrac{1}{k} \displaystyle\sum_{i \in S_M} \overline{y}_i^*$.

An unbiased variance of the estimator is given by

$$\mathrm{Var}\left(\hat{\tau}_{gI,A}\right)=\begin{cases} N^2\left(1-\dfrac{n_0}{N}\right)\dfrac{S_0^{*2}}{n_0} & \text{if } [Q_0], \\ \mathrm{Var}\left(\hat{\tau}_{gI,A}\right)_{Q_1} & \text{if } [Q_1], \\ N^2\left(1-\dfrac{n_2}{N}\right)\dfrac{S_2^{*2}}{n_2} & \text{if } [Q_2], \end{cases} \tag{9}$$

where

$$\mathrm{Var}\left(\hat{\tau}_{gI,A}\right)_{Q_1}=\frac{\sigma_M^{*2}}{k}E_{n_1}\left[\hat{M}^2\left(1-\frac{k}{M}\right)\right]+\sigma_{N-M}^{*2}E_{n_1}\left[\frac{\left(N-\hat{M}\right)^2}{n_1-k}\left(1-\frac{\left(n_1-k\right)}{N-M}\right)\right]$$
$$+\left(\mu_M-\mu_N-M\right)^2\mathrm{Var}_{n_1}\left(\hat{M}\right)$$

$S_j^{*2}=\left(N-1\right)^2\sum_{i=1}^{N}\left(\bar{y}_i^*-\mu\right)^2$, $j=0,1$, $\sigma_M^{*2}=M^{-1}\sum_{i\in U_M}\left(\bar{y}_i^*-\mu_M\right)^2$ and

$\sigma_{N-M}^{*2}=\left(N-M\right)^{-1}\sum_{i\in U_{N-M}}\left(\bar{y}_i^*-\mu_{N-M}\right)^2$.

The usual inverse adaptive cluster sampling can be improved by incorporating more of the information obtained in the final sample. In particular, the values of edge units are utilized in the estimator only for edge units that were picked in the initial sample [3].

The units selected in the initial sample are denoted by S_0 and the units in the final sample denoted by S can be partitioned into two parts: S_M is the set of all the distinct units in the sample for which satisfy the condition C and the remaining part S_{N-M} consists of all the distinct units in the sample for which does not satisfy the condition C. For unit i, let f_i be the number of times the network to which unit i belongs is intersected by the initial sample; that is f_i is the number of units in the initial sample that are in the network to which unit i belongs.

Let the statistic d^+ be defined as

$$d^+ = \left\{ (i, y_i, f_i) : i \in S_M, (j, y_j) : j \in S_{N-M} \right\}$$

Let D^+ denote a random variable that takes on the possible value of d^+. Also let D^+ denote the sample space for d^+.

For $i \in S$ define the indicator e_i as

$$e_i = \begin{cases} 1 \ ; \text{if } y_i < c \ \text{ and } \ i \ \text{is in the neighborhood of some } j \in S_M \\ 0 \ ; \text{otherwise} \end{cases}$$

Thus $e_i = 1$ if unit i is an edge unit and the network that makes it an edge unit is selected in the initial sample.

The number of sample edge units is

$$e_s = \sum_{i \in S} e_i .$$

The number of sample edge units picked in the initial sample s_0 is

$$e_{s_0} = \sum_{i=1}^{n_0} e_i = \sum_{i \in S_0} e_i .$$

The average y-value for the sample edge units in the final sample is

$$\overline{y}_e = \frac{\sum_{i \in s} e_i y_i}{e_s} .$$

For the i^{th} unit in the sample, define a new variable of interest $\overline{y}_i^{*'}$ by $\overline{y}_i^{*'} = \overline{y}_i^{*}(1 - e_i) + \overline{y}_e e_i$.

An unbiased estimator of the population total is

$$\hat{\tau}^+ = E\left[\hat{\tau}_{HH} \middle| D^+ = d^+ \right]$$

$$= N(n)^{-1} \sum_{i=1}^{n} \overline{y}_i^{*'} .$$

The variance of the $\hat{\tau}^{+}$ is

$$V\left(\hat{\tau}^{+}\right)=V\left(\hat{\tau}_{HH}\right)-E\left[\left(\hat{\tau}_{HH}-\hat{\tau}^{+}\right)^{2}\right].$$

By the Rao-Blackwell Theorem; $\hat{\tau}^{+}$ is an unbiased estimator of population total, since $\hat{\tau}_{HH}$ is an unbiased estimator, and d^{+} is the sufficient statistic then the variance of the $\hat{\tau}^{+}$ no more than the variance of the $\hat{\tau}_{HH}$.

The improved estimator of the population total is

$$\hat{\tau}_{gI,IA}=\begin{cases}\dfrac{N}{n_0}\displaystyle\sum_{i=1}^{n_0}\bar{y}_i^{*'} & \text{if } [Q_0],\\ \hat{M}\bar{y}_M^{*}+\dfrac{\left(N-\hat{M}\right)}{n_1-k}\displaystyle\sum_{i\in S_{N-M}}\left[y_i\left(1-e_i\right)+\bar{y}_e e_i\right] & \text{if } [Q_1],\\ \dfrac{N}{n_2}\displaystyle\sum_{i=1}^{n_2}\bar{y}_i^{*'} & \text{if } [Q_2],\end{cases} \tag{10}$$

where $[Q_0]=\left\{n_1=n_0\right\}$

$[Q_1]=\left\{n_0<n_1<n_2 \text{ or } n_1=n_2 \ \&\ \left|S_M\right|=k\right\}$

$[Q_2]=\left\{n_1=n_2 \ \&\ \left|S_M\right|<k\right\}$.

The variance of the estimator is given by

$$Var\left(\hat{\tau}_{gI,IA}\right)=\begin{cases}N^2\left(1-\dfrac{n_0}{N}\right)\dfrac{S_0^{*2}}{n_0}-N^2B_0 & \text{if } [Q_0],\\ Var\left(\hat{\tau}_{gI,A}\right)_{Q_1}-N^2B_1 & \text{if } [Q_1],\\ N^2\left(1-\dfrac{n_2}{N}\right)\dfrac{S_2^{*2}}{n_2}-N^2B_2 & \text{if } [Q_2].\end{cases} \tag{11}$$

Where

$$B_0 = E_{n_0}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\bar{y}_i^{*} - \frac{1}{n_0}\sum_{i=1}^{n_0}\bar{y}_i^{*'}\right]^2 \quad , \quad B_1 = E_{n_1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\bar{y}_i^{*} - \frac{1}{n_1}\sum_{i=1}^{n_1}\bar{y}_i^{*'}\right]^2 \text{ and}$$

$$B_2 = E_{n_2}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\bar{y}_i^{*} - \frac{1}{n_2}\sum_{i=1}^{n_2}\bar{y}_i^{*'}\right]^2 .$$

4. Simulation Study

In this section, the properties of the inverse sampling design is investigated in relation to k and n_0 and compare it with the inverse adaptive sampling design, the blue-winged teal data [6] were studied. The population region in an area of central Florida was partitioned into 200 quadrants and the numbers of blue-winged teal in each quadrant were records.

Table 1. The numbers of blue-winged teal into 200 quadrants.

0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	20	4	2	12	0	0	0	0	0	10	103	0	0	0
0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	150	7144	1	0
0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	6	6339	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	122	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	60
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	3	0

For each estimator 20,000 iterations were performed to obtain a precise estimate. The condition for added units in the sample is defined by $C = \{y : y \geq 5\}$. A varying of k = 2, 3, 4, 5, 6 and n_0 = 10, 20, 30, 40, 50 was used. The formula used to estimate the variance of the estimate total is

$$\mathrm{Var}(\hat{\tau}) = \frac{1}{20{,}000}\sum_{i=1}^{20{,}000}\left(\hat{\tau}_i - \bar{\tau}\right)^2 ,$$

where $\hat{\tau}_i$ is the value for the relevant estimator for sample i and $\bar{\tau}$ is the average of the $\hat{\tau}_i$. Let ν denote the final sample size.

The results of the inverse sampling, inverse adaptive cluster sampling and improved estimator of inverse adaptive cluster sampling are as follows:

Table 2. The variance of the estimators for the population total of the variable of interest.

n_0	k	Inverse sampling		Inverse adaptive cluster sampling		
		$E(\nu)$	$Var(\hat{\tau}_{gI})$	$E(\nu)$	$Var(\hat{\tau}_{gI,A})$	$Var(\hat{\tau}_{gI,IA})$
10	2	30.99	1,191,492,215	46.84	360,479,900	360,476,467
	3	44.41	506,975,610	62.65	215,799,053	215,785,886
	4	60.44	246,984,300	79.56	76,846,315	76,841,232
	5	74.41	170,593,762	92.71	46,125,905	46,117,090
	6	89.82	128,364,937	106.95	27,400,405	27,390,151
20	2	31.76	700,351,826	47.89	227,406,510	227,406,330
	3	45.65	399,684,128	64.19	143,950,190	143,939,415
	4	61.31	237,335,725	80.17	72,600,951	72,598,098
	5	74.23	181,104,339	92.58	53,375,238	53,369,182
	6	90.81	126,864,743	107.79	28,213,054	28,203,824
30	2	35.49	454,838,883	52.01	133,121,423	133,111,066
	3	47.26	316,653,324	65.89	95,215,515	95,203,577
	4	61.49	211,008,154	80.54	60,008,594	59,997,906
	5	75.57	169,243,161	93.93	45,270,897	45,260,074
	6	91.19	129,515,334	108.06	28,142,319	28,135,297
40	2	42.42	353,321,433	59.35	105,848,009	105,843,604
	3	49.93	312,763,634	68.29	89,925,155	89,918,008
	4	61.73	261,194,501	80.70	75,688,965	75,686,533
	5	74.59	177,403,084	92.94	47,137,138	47,127,496
	6	89.47	131,611,061	106.58	29,067,150	29,058,319
50	2	52.03	293,080,839	68.89	79,330,104	79,326,944
	3	56.10	258,011,038	74.58	74,385,698	74,379,356
	4	63.08	197,500,412	81.93	51,628,978	51,623,033
	5	75.52	168,874,018	93.91	40,240,311	40,232,428
	6	90.67	128,253,828	107.61	27,787,297	27,778,577

Table 2 lists the expected final sample sizes and variances for the inverse sampling and inverse adaptive cluster sampling for a selection of initial sample size. The variance of the $\hat{\tau}_{gI,IA}$ is less than the variance of the $\hat{\tau}_{gI,A}$ not more than 0.04% but it is less than the variance of the $\hat{\tau}_{gI}$ more than 50%.

5. Conclusions

Adaptive cluster sampling is an efficient method for sampling rare and hidden clustered populations. Inverse sampling design is applied in adaptive cluster sampling that use stopping rules based on the number of rare units observed. An estimator of a population total by inverse adaptive cluster sampling is improved for reducing the variance of the estimator. The numerical study shows that for the fixed initial sample size (n_0) the variance of the improved estimator decrease when a predetermined numbers (k) increase and for the fixed k the variance of the improved estimator decrease when n_0 increase. The variance of $\hat{\tau}_{gI,IA}$ is less than the variance of $\hat{\tau}_{gI,A}$ but The variance of $\hat{\tau}_{gI,IA}$ is much less than the variance of $\hat{\tau}_{gI}$.

6. Acknowledgements

I would like to profoundly thank Mr. Paveen Chutiman for his programming advice.

References

[1] Christman, M.C., and Lan, F., Inverse Adaptive Cluster Sampling: Biometrics. 2001; 57:1096-1105.

[2] Dryver, A.L., Adaptive Sampling Designs and Associated Estimators : Ph.D. Thesis, The Pennsylvania State University,1990.

[3] Dryver, A.L., and Thompson, S.K., Improved unbiased estimators in adaptive cluster sampling: Journal of the Royal Statistical Society, 2005; Ser.B, 67, 157-166.

[4] Mikulski, P.W. and Smith, P.J., A variance bound for unbiased estimation in inverse sampling: Biometrika. 1976; 63:216-217.

[5] Salehi, M.M, and Seber, G.A.F., A general Inverse Sampling Scheme and its Application to Adaptive Cluster Sampling: Australian & New Zealand Journal of statistics.2004; 46: 483-494.

[6] Smith, D.R.; Conroy, M.J. and Brakhage, D.H., Efficiency of Adaptive Cluster Sampling for Estimating Density Estimating Density Wintering Waterfowl, Biometrics,1995; 51: 777-788.

[7] Thompson, S.K., Adaptive Cluster Sampling: Journal of the American Statistical Association. 1990; 85: 1050 -1059.

[8] Thompson, S.K. Sampling**.** New York : Wiley,1992.