KMITL Sci. J. Vol. 5 No. 1 Feb. 2005

Analysis of Adaptive Cluster Sampling Utilizing Standard Software Packages Without Complex Programming

Arthur L. Dryver* and Chang-Tai Chao

School of Applied Statistics, National Institute of Development Association,
Bangkok 10240 Thailand
email: dryver@gmail.com, fax: 662-377-7892
Department of Statistics, National Cheng-Kung University,
Tainan City, 701 Taiwan
email: ctchao@email.stat.ncku.edu.tw

Abstract

Simple random sampling with or without replacement is the easiest to analyze using standard software, such as SAS, SPSS, Minitab, etc. Better population estimates can be obtained through more complex sampling designs, but this introduces analysis issues. For example, regression is very straightforward with simple random sampling, but this is not always the case when more complicated sampling designs are used, such as adaptive cluster sampling. A serious concern with regression estimates introduced with many complicated designs is lack of independence, a necessary assumption. This paper covers an effective manner to analyze data collected from adaptive cluster sampling designs using standard software. Also, included is sample SAS code.

Key Words: Adaptive cluster sampling; Hansen-Hurwitz; SAS.

1 Introduction

First proposed by Thompson [1], Adaptive Cluster Sampling (ACS) has been widely used in different disciplines, such as Ecological science (e.g. [2], [3], [4]), Environmental science (e.g. [5], [6]), Geographical [7], and Social science [8], because of its advantageous, because of its advantageous practical flexibility as well as the ability to provide a more meaningful sample and more efficient estimate of the population parameter of interest. Associated with different initial sampling designs, various ACS designs have been proposed during the last 15 years (e.g. [9], [10], [11]). The advantages of ACS can be considerable especially for rare and clustered populations, with which the conventional sampling methods often fail to provide efficient inference.

Unlike the conventional sampling designs, in which the sampling selection is independent from the observed values of population variable of interest, ACS is an untraditional data-driven sampling selection procedure. Consequently, it requires extra care to establish the related statistical inference. For professionals from disciplines other than statistics, standard softwares such as SAS, SPSS might be the only option for them to analyze their survey data. Most, standard software packages, however, generally require the assumption of a simple random sample to perform the analysis without programming. Analyzing the data obtained from complex sampling designs can be complicated, but analyzing it in the same manner as a sample obtained from simple random sampling can lead to very biased estimates. The ultimate purpose of the user-friendly interface provided by the standard packages will then be totally nullified if the data is collected by some complex/unconventional sampling methods.

For example, when auxiliary information is available in the survey data, often one would like to use ratio or regression estimation to construct the related inference of the population quantity of interest. Though design-biased, the ratio or regression estimators are often found to be more favorable with smaller mean square error. They also can provide better estimation results in ACS ([12], [13]). With standard statistical softwares and a simple random sample, the ratio/regression estimator can be carried out straightforwardly since it is equivalent to the least square estimator in simple linear regression with or without intercept. When the data is collected under some ACS designs, however, the calculation of an associated ratio/regression estimators might cause certain difficulty for professionals who are familial to the standard statistical packages only. It might lead to unnecessary hesitance to make use of ACS in their surveys. In this paper, how to carry out the related estimates under ACS using SAS, which might be the most commonly used standard statistical software package, is described. It is expected to be helpful for professionals from other disciplines to better understand ACS, and motivate them to use it in their future research.

In Section 2, the sampling procedure of an ACS with a simple random sampling without replacement as the initial design is described. For understanding of associated SAS code to analyze the data obtained by ACS, certain fundamentals behind ACS will be useful. For example, how to relate ACS to the usual conventional equal probability sampling is explained in Section 3 using real data. In Section 4, sample SAS code which can appropriately calculate the associated estimators under ACS is introduced. The SAS code can be extended to ACS with other initial sampling designs with minor modifications. Some final comments are given in Section 5.

2 Adaptive Cluster Sampling Design

In adaptive cluster sampling an initial sample of units can be selected by different types of conventional probability sampling. For this paper only adaptive cluster sampling in which an initial simple random sample is taken without replacement will be covered [1]. Whenever the variable of interest for a unit in the sample satisfies a pre-specified condition, linked units are added to the sample and this procedure continues until no more units are found that meet the pre-specified condition. The way in which units can be considered to be linked is very flexible. The main restriction is that if unit i is considered linked to unit j then unit j is considered linked to unit i. In adaptive cluster sampling a network is comprised of all linked units that meet the pre-specified condition. Units that are linked to a network but do not meet the condition are called edge units. A network together with its associated edge units form a cluster, thus the name adaptive cluster sampling. Due to the way in which the sample is obtained the final number of units and networks that will be in the sample is unknown prior to sampling and the standard sample mean, $\bar{y}$ is biased. A typical design unbiased estimator used to analyze data from an adaptive cluster sample has the form of the Hansen-Hurwitz estimator [14]. Edge units are only included in the estimate when using the standard Hansen-Hurwitz estimator if they were sampled in the initial sample. In addition, adaptive cluster can also be analyzed using a Horvitz-Thompson type estimator [14]. This paper focuses on ease of analysis using standard software and for this reason only the standard Hansen-Hurwitz type estimator will be covered.

2.1 Technical Notation

As in the typical finite population sampling situation, the population is considered to consist of N units labelled $1, 2, .., N$. Let y denote the primary variable of interest and x the auxiliary variable.

Let Ψ_i denote the network that includes unit i. Let m_i denote the number of units in that network. Let w_{xi} denote the average x-value in Ψ_i, that is $w_{xi} = \frac{1}{m_i}\sum_{j\in\Psi_i} x_j$. Let w_{yi} denote the average y-value in Ψ_i, that is $w_{yi} = \frac{1}{m_i}\sum_{j\in\Psi_i} y_j$. An unbiased estimator of the population mean, μ_y, and its variance for adaptive cluster sampling [14] are $\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} w_{yi}$ and $var(\hat{\mu}_y) = \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N}(w_{yi} - \mu)^2$ respectively. An unbiased estimator for the variance of $\hat{\mu}_y$ is $\widehat{var}(\hat{\mu}_y) = \frac{N-n}{Nn(n-1)}\sum_{i=1}^{n}(w_{yi} - \hat{\mu}_y)^2$.

3 Transformed Finite Population Approach

3.1 Univariate Case

The estimator $\hat{\mu}_y$ can be thought of as the sample mean via simple random sampling without replacement from a transformed population consisting of the average of y_i-values for each network as opposed to the y_i-values themselves. Then the expression for the variance and estimator of the variance follows from classical simple random sampling ([14], p304). Two populations of teal data will be used to illustrate the later concept. The condition to adaptively add units is if $x_i \geq 1$.

Table 1: Blue-winged teal data [5]

0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	10	103	0
0	0	0	0	2	3	2	0	13639	1
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

Table 2: Green-winged teal data x [5].

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	35	75	0
0	0	0	0	0	0	0	0	2255	13
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	24

Table 3: Transformed blue-winged teal data, w_y.

0	0	3	5	0	0	0	0	0	0
0	0	0	24	14	0	0	3438.25	3438.25	0
0	0	0	0	2	3	2	0	3438.25	3438.25
0	0	0	0	0	0	0	0	14	122
0	0	0	0	0	0	2	0	0	177

Table 4: Transformed green-winged teal data, w_x.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	594.5	594.5	0
0	0	0	0	0	0	0	0	594.5	594.5
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	24

An adaptive cluster sample with an initial simple random sample, or a specific type of network sample, which will be discussed in further detail later, taken from the population in Table 1 adaptively adding units according to the population in Table 2 can be thought of as a simple random sample taken from the transformed population in Table 3. Thus $\hat{\mu}_y$ from an original transformed population can be thought of as $\bar{w}_y = \frac{1}{n}\sum_{i=1}^{n} w_{yi}$ taken with a sample random sample. They are both unbiased estimators for the untransformed population in Table 1.

3.2 Multivariate Case: Regression Estimator

To perform something such as regression on data from an adaptive cluster sample can be very complicated, but to do regression on the transformed data, Table 3 and Table 4, and treating the data as if it were taken from a simple random sample, can be done by any standard software package with minimal additional work. An example is simple linear regression, where $w_{yi} = \beta_{0w} + \beta_{1w} w_{xi} + \epsilon_i$. An estimate for β_{0w} and β_{1w} can be derived in the typical manner for ordinary least squares (OLS) by minimizing the sum of squared error. Thus $\hat{\beta}_{0w} = \bar{w}_y - \hat{\beta}_{1w} \times \bar{w}_x$ and $\hat{\beta}_{1w} = \frac{\sum_{i=1}^{n}(w_{xi}-\bar{w}_x)(w_{yi}-\bar{w}_y)}{\sum_{i=1}^{n}(w_{xi}-\bar{w}_x)^2}$

4 Sample SAS Code

The following is sample SAS code to calculate the $\hat{\mu}_y$ and $\hat{\mu}_x$ and their associated standard deviations for an adaptive cluster sample when an initial sample is a simple random sample taken without replacement. In addition, is code to calculate regression estimates on the transformed populations of y and x. From this code it can be seen how to calculate estimates without complicated code. The data needed are unit labels (ulabel), network label (nlabel), x_i-value (xivalue); and y_i-value (yivalue). As you can see from the code the unit labels are not truly needed for calculations but it is highly recommended to have in your data always.

```
data acs1;                                    *1st enter the data;
input ulabel nlabel xivalue yivalue @@;
cards;
1 1 0 0
2 2 0 0
3 3 3 7
4 3 4 7

proc sort data=acs1;                          *2nd sort by network label;
by nlabel;
proc means data=acs1 noprint;                 *3rd solve for w_yi and w_xi;
by nlabel;
output out=acsn(drop=temp1 _type_ )
```

```
mean=templ wxivalue wyivalue;
proc means data=acsn;              *$4^{th}$ solve for $\hat{\mu}_y$ and $\hat{\mu}_x$;
var wyivalue wxivalue;             *and their standard deviations.;
proc reg data=acsn;                *$5^{th}$ and final step;
model wyivalue=wxivalue;           *perform regression;
run;                               *on the transformed population.;
```

5 Final Comments

In Section 3.1 methodology to estimate and understand a standard ACS univariate estimator of the population mean is proposed. A short coming of the approach recommended in Section 3.2 used for handling the multivariate case unlike in the univariate case is that the **findings relate directly to the transformed population and not the original populations.** Depending on your objectives though this can be greatly helpful. For example, since the population totals and population means are the same for the transformed and untransformed populations the results are directly applicable.

References

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

[2] Hanselman, D.H., Quinn, T.J., Lunsford, D, Heifetz, J and Clausen, D **2003** Applications in adaptive cluster sampling of Gulf of Alaska rockfish. *Fishery Bulletin 101*(3), 501-513.

[3] Conners, M.E. and Schwager, S.J. **2002** The use of adaptive cluster sampling for hydroacoustic surveys. *Ices Journal of Marine Science 59*(6), 1314-1325.

[4] Lo, N.CH, Griffith, D and Hunter, JR **1997** Using a restricted adaptive cluster sampling to estimate Pacific hake larval abundance , *California Cooperative Oceanic Fisheries Investigations Reports* 38: 103-113.

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

[6] Correll, R.L. **2001** The use of composite sampling in contaminated sites- a case study. *Environmental and Ecological Statistics 8*(3), 185-200.

[7] Boomer, K, Werner, C and Brantley S **2000** CO2 emissions related to the Yellowstone volcanic system 1. Developing a stratified adaptive cluster sampling plan , *Journal of Geophysical Research-Solid Earth*,105 (B5): 10817-10830.

[8] Thompson, S.K. and Collins, L.A. **2002**. Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence 68* S57-S67.

[9] Thompson, S.K., and Seber; G.A.F. **1996** *Adaptive Sampling.* New York: Wiley.

[10] Muttlak, H.A. and Khan, A. **2002** Adjusted two-stage adaptive cluster sampling. *Environmental and Ecological Statistics 9*(1), 111-120.

[11] Christman. M.C. **2003** Adaptive two-stage one-per-stratum sampling. *Environmental and Ecological Statistics 10*(1), 43-60.

[12] Chao, C.T. **2004** Ratio estimation on Adaptive Cluster Sampling. *Journal of Chinese Statistical Association 42*(3), 307-327

[13] Dryver, A.L. and Chao, C.T. **2005** Utilization of auxiliary information in adaptive cluster sampling, working manuscript.

[14] Thompson, S.K. **2002** *Sampling.* 2nd Ed. New York: Wiley.