The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 151-159

Efficient Controlled Selection[1)]

Jea-Bok Ryu[2)] and Seung-Joo Lee[3)]

Abstract

In sample surveys, we expect preferred samples that reduce the survey cost and increase the precision of estimators will be selected. Goodman and Kish (1950) introduced controlled selection as a method of sample selection that increases the probability of drawing preferred samples, while decreases the probability of drawing nonpreferred samples. In this paper, we obtain the controlled plans using the maximum entropy principle, and when the order of nonpreferred samples is considered, we propose the algorithm to obtain a controlled plan.

1. Introduction

Among the samples selected from a finite population, a few units which are not important from the point of view of the characters under study will be contained. It may happen that the sampling units are too widespread or lie far from the interior. This may cause not only to increase the survey cost considerably, but also to bring difficulty on supervision and effective formation of fieldwork. In this situation the collected datas are seriously affected by non-response, and investigator bias, etc., and as a result nonsampling errors are increased. Such samples which would increase the cost and consequently reduce the precision of the estimate of the parameter are referred to as nonpreferred samples(Goodman and Kish, 1950). Particularly in multi-purpose surveys, it is desirable that a sample, if possible, should contain many characteristics of research being under study. Therefore, it may be possible to raise the precision of the estimate by increasing the probability of selecting preferred samples and decreasing the probability of selecting nonpreferred samples.

Goodman and Kish(1950) suggested controlled selection as a sampling procedure that increases the probability of selecting preferred samples beyond that which is possible with stratified random sampling. Hereafter, many researches have studied on controlled selection

1) This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1995.
2) Professor, Department of Applied Statistics, Chongju University, 36 Naedokdong Sangdanggu Chongju Chungbuk, 360-764, Korea.
2) Full-time Lecturer, Department of Applied Statistics, Chongju University, 36 Naedokdong Sangdanggu Chongju Chungbuk, 360-764, Korea.

and these researches can be divided into two parts. One part is related to the method of reducing the support size, while the other is concerned with minimizing the selection probabilities of nonpreferred samples.

Controlled selection is particularly useful in the selection of first-stage units in multi-stage sampling, and a controlled plan can be easily implemented using the method of cumulative sums or Lahiri's(1951) method.

In this paper, we use the maximum entropy principle to obtain a controlled plan that reduces the probability of drawing nonpreferred samples and support size. And when the order of nonpreferred samples is considered, a controlled plan, excepting the bad nonpreferred samples from the ordered nonpreferred samples, can not be obtained by previous methods. So we propose the algorithm to obtain a controlled plan under these circumstances.

Previous works and limitations of controlled selection are summarized in section 2. Section 3 deals with the general entropy principles. In section 4, we obtained a controlled plan using the maximum entropy principle and compared it with other methods through numerical examples. And when the order of nonpreferred samples is considered, we present the algorithm for a controlled plan in section 5.

2. Controlled Selection

Goodman and Kish gave the following statement in their paper published in 1950.

> "Earlier research workers , such as Neyman(1934), Bowley(1926), Jensen(1926, 1928), and Strand and Jessen(1943), had reported their analyses of stratified random sampling and purposive selection but their findings had been far from conclusive. Jensen, in 1928, described the purposive selection of a sample from records of the 1923 Danish Agricultural Census and showed that it represented the population well in respect to distributions of several farm variables. While others concluded that purposive selection is generally inferior to stratified random sampling. However, none of these investigators attempted to combine purposive selection with probability sampling."

Conceptually, controlled selection may be regarded as an extension of purposive selection that is a kind of nonprobability sampling. In order for controlled selection to be probability sampling, not just one but many purposive samples must be selected, until every unit in the population is included in one or more samples. The number of samples in which each unit appears must be exactly proportionate to its assigned probability of selection. Hence, controlled selection may be considered as a method of combining probability sampling with nonprobability sampling.

When the population size and the sample size are slightly large, it becomes complicated and

laborious to list all purposive samples. Therefore, it is very important to make purposive sample sets easily and conveniently.

Chakrabarti(1963) used balanced incomplete block(BIB) designs to construct sampling plans with reduced support size and Avadhani and Sukhatame(1973) proposed the use of the BIB design to obtain a controlled simple random sampling plan. Foody and Hedayat(1977), and Wynn(1977) used BIB designs with repeated blocks for situations where nontrivial BIB designs do not exist. Gupta, Nigam, and Kumar(1982) employed BIB designs in conjunction with the Horvitz-Thompson estimator.

Hedayat, Lin, and Stufken(1989) used the method of "emptying boxes" to construct controlled inclusion probability proportional to size(IPPS) sampling plans with the following property :

$$0 < \pi_{ij} \leq \pi_i \pi_j, \quad i < j = 1, 2, \ldots, N,$$

where π_i and π_{ij} are respectively 1st-order and 2nd-order inclusion probabilities for simple random sampling.

Hess, Reidel, and Fitzpatrick(1975), Ernst(1981), and Lin(1992) proposed algorithms to construct the purposive sample sets for controlled selection efficiently and systematically.

Rao and Nigam(1990, 1992), and Sitter and Skinner(1994) obtained optimal controlled plans using linear programming to minimize the probability of selecting nonpreferred samples.

Let $p(s)$ be the probability of selecting the sample s ($\in S$), where S is the set of all possible samples and sample size n is greater than 2. If S_1 denotes the subset of nonpreferred samples, then the optimal controlled design, $p_c(s)$, is a solution to the linear programming problem that minimizes the objective function, $\phi = \sum_{s \in S_1} p(s)$, under the following linear constraints : (This is readily obtained by simplex method.)

$$\begin{aligned} &(i) \quad \sum_{i,j \in s} p(s) = \pi_{ij} \quad (i < j = 1, \ldots, N), \\ &(ii) \quad p(s) \geq 0, \quad \textit{all} \quad s \in S. \end{aligned} \tag{2-1}$$

Also, an optimal controlled plan, $p_c(s)$, may be obtained by a unified approach using standard linear programming, that is, by minimizing the objective function, $\phi = \sum_{s \in S_1} p(s)$, subject to :

$$\begin{aligned} &(i) \quad p(s) \geq 0, \quad s \in S, \\ &(ii) \quad \sum_{s \in S} p(s) = 1, \\ &(iii) \quad \textit{unbiasedness condition} : \sum_{i \in s} d_i(s)\, p(s) = 1, \quad i = 1, \ldots, N, \\ &(iv) \quad \textit{variance matching condition} : \\ &\qquad \sum_{i,j \in s} d_i(s)\, d_j(s)\, p(s) = d_{ij}^0 + 1, \quad i < j = 1, 2, \ldots, N, \end{aligned} \tag{2-2}$$

where $d_{ij} = E[d_i(s)d_j(s)] - 1$, $d_i(s)$ may depend on i or s or both, and the values of d_{ij} under the uncontrolled plans, $p(s)$, are denoted by d_{ij}^0. A more general objective function, $\phi = \sum_{s \in S} w(s)\, p(s)$, with specifed weights $w(s) \geq 0$, can also be used. This means that instead of specifying S_1, a cost function for all samples can be defined and expected cost is minimized. Since a solution obtained by solving linear programming problems is not always unique, an optimal solution, $p_c(s)$, should be chosen for our purpose.

Ryu(1996) reviewed the methods on controlled selection and considered that limitations and problems occurred by applying these methods to practical field survey. He also suggested the directions and the subjects for further studies on controlled selection.

3. Entropy Principle

Clausius(Rudolf Clausius, 1822-1888) was the first to use the term "entropy" to measure the loss quantities of available energy in his 1868 paper. Shannon(1948) defined the information entropy as a meaning of uncertainty.

Let $\boldsymbol{p} = (p_1, p_2, \ldots p_n)$ be a probability distribution such that $p_i \geq 0$ for all i and $\sum_{i=1}^{n} p_i = 1$. Shannon's measure of entropy, or uncertainty, for this distribution is given by

$$S(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \ln p_i \ . \tag{3-1}$$

Several methods for finding a probability distribution under the restricted conditions have been studied in many areas. Here we discussed two useful entropy principles, which are Janey's maximum entropy principle(MEP), and Kullback's minimum cross entropy principle (MCEP).

By Janey's MEP we can find a probability distribution that maximizes Shannon's measurement, $-\sum_{i=1}^{n} p_i \ln p_i$, subject to the following constraints :

$$\begin{aligned} &\sum_{i=1}^{n} p_i = 1 \ , \\ &\sum_{i=1}^{n} p_i g_{ri} = a_r \ , \quad r = 1, 2, \ldots, m \ , \\ &p_i \geq 0, \ i = 1, 2, \ldots, n \ , \end{aligned} \tag{3-2}$$

where g_{ri}, $r = 1, 2, \ldots, m$, is a arbitrary function whose expectation exists.

Instead of Shannon's measurement, the following Kullback and Leibler's measurement can be used :

$$D(\boldsymbol{p} : \boldsymbol{q}) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}, \tag{3-3}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots p_n)$ and $\boldsymbol{q} = (q_1, q_2, \ldots q_n)$ are probability distributions.

By Kullback's MCEP we can also find a probability distribution, $\boldsymbol{p} = (p_1, p_2, \ldots p_n)$, that minimizes Kullback and Leibler's measurement, $\sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$, subject to the prior distribution and constraints (3-2).

The solution has the form,

$$p_i = q_i \exp(-\lambda_0 - \lambda_1 g_{1i} - \lambda_2 g_{2i}, \cdots, \lambda_m g_{mi}),$$

where $\lambda_0, \lambda_1, \cdot, \cdot, \cdot \lambda_m$ may be determined by using the constraints.

Janey's MEP can be considered as a special case of Kullback's MCEP when a prior probability distribution, $\boldsymbol{q} = (q_1, q_2, \ldots q_n)$, is the uniform distribution..

4. Controlled Selection Using Maximum Entropy Principle

In this section maximum entropy principle is used to construct controlled plans. The controlled plans are the solution of the following nonlinear programming problem: maximize the objective function with respect to the variables $\{ p(s), s \in S \}$ subject to (2-1). The objective function of *MEP*1 is $\phi = -\sum_{s \in S} p(s) \ln p(s)$ which is the same as Shannon's measurement and that of *MEP*2 is $\phi = -\sum_{s \notin S_1} p(s) \ln p(s)$.

$$\begin{aligned} &MEP1 : \max -\sum_{s \in S} p(s) \ln p(s) \\ &MEP2 : \max -\sum_{s \notin S_1} p(s) \ln p(s) \end{aligned} \tag{4-1}$$

We used PROC NLP in SAS release 6.11 to solve the nonlinear programming problem in the following examples.

Example 1. Consider the following population from Rao and Nigam(1990) with $N = 6$, $n = 3$ and p_i values : $\frac{2}{17}, \frac{3}{17}, \frac{4}{17}, \frac{1}{17}, \frac{2}{17}, \frac{5}{17}$. Suppose that the following 7 samples are considered nonpreferred samples :

(1 2 3) (1 2 6) (1 3 6) (1 4 6) (2 3 4) (2 3 6) (2 4 6)

Using nonlinear programming, we obtained three controlled plans, given in table 1, that match π_{ij} computed by Sampford's(1967) method. ϕ for *MEP*1 is almost the same as ϕ

for Sampford's uncontrolled plan. However ϕ for *MEP*2 is less than Sampford's plan while greater than for Rao and Nigam.

<Table 1> Controlled Plan Matching Sampford's(1967) Method

s	Rao and Nigam	*MEP*1	*MEP*2
(1, 2, 3)*	0.020507	0.028953	0.020607
(1, 2, 4)	0.013375	0.003557	0.014833
(1, 2, 5)		0.008295	0.001091
(1, 2, 6)*	0.086172	0.079250	0.083523
(1, 3, 4)	0.018382	0.006923	0.016824
(1, 3, 5)	0.013119	0.015828	0.011927
(1, 3, 6)*	0.144649	0.144953	0.147298
(1, 4, 5)		0.001827	0.000100
(1, 4, 6)*		0.019450	
(1, 5, 6)	0.056737	0.043905	0.056737
(2, 3, 4)*		0.012832	
(2, 3, 5)	0.048403	0.028953	0.048403
(2, 3, 6)*	0.251017	0.249189	0.250917
(2, 4, 5)	0.003861	0.003557	0.001212
(2, 4, 6)*	0.038285	0.035576	0.039476
(2, 5, 6)	0.067791	0.079250	0.069349
(3, 4, 5)		0.006923	0.002649
(3, 4, 6)	0.074672	0.066376	0.073580
(3, 5, 6)	0.135134	0.144953	0.133676
(4, 5, 6)	0.027896	0.019450	0.027796
$\sum_{s \in S_1} p(s)$	0.540630	0.570200	0.541821

* nonpreferred sample

Example 2. We consider the following population from Avadhani and Sukhatme(1973) with $N=7$, $n=3$. Suppose that p_i values are 0.10, 0.12, 0.14, 0.15, 0.15, 0.16, 0.18.

Among 35 possible samples, the following 14 samples are considered nonpreferred samples by Avadhani and Sukhatme(1973) for reasons of travel cost and field work.

(1 2 3) (1 2 6) (1 3 6) (1 3 7) (1 4 6) (1 4 7) (1 6 7)
(2 3 4) (2 3 6) (2 3 7) (2 4 6) (2 4 7) (3 4 7) (4 6 7)

<Table 2> Controlled Plan Matching Sampford's(1967) Method

s	Rao and Nigam	*MEP*1	*MEP*2
(1, 2, 3)*		0.012671	
(1, 2, 4)		0.014096	0.037179
(1, 2, 5)		0.014096	0.002660
(1, 2, 6)*		0.015653	
(1, 2, 7)	0.075785	0.019268	0.035946
(1, 3, 4)	0.086583	0.017572	0.048249
(1, 3, 5)		0.017572	0.013532
(1, 3, 6)*		0.019501	0.010734
(1, 3, 7)*	0.004708	0.023973	0.018777
(1, 4, 5)		0.019530	0.004091
(1, 4, 6)*		0.021666	
(1, 4, 7)*	0.012899	0.026617	0.009963
(1, 5, 6)	0.087441	0.021666	0.057580
(1, 5, 7)	0.012040	0.026617	0.021618
(1, 6, 7)*	0.020545	0.029501	0.039673
(2, 3, 4)*		0.022308	0.006016
(2, 3, 5)	0.109147	0.022308	0.057829
(2, 3, 6)*		0.024741	0.016122
(2, 3, 7)*	0.003261	0.030379	0.032442
(2, 4, 5)		0.024778	0.017485
(2, 4, 6)*	0.083232	0.027472	0.036033
(2, 4, 7)*	0.039131	0.033710	0.025652
(2, 5, 6)	0.013216	0.027472	0.032273
(2, 5, 7)		0.033710	0.012117
(2, 6, 7)	0.036228	0.037339	0.048249
(3, 4, 5)		0.030797	0.008690
(3, 4, 6)	0.039990	0.034124	0.055474
(3, 4, 7)*	0.020048	0.041819	0.028192
(3, 5, 6)		0.034124	0.019075
(3, 5, 7)	0.037473	0.041819	0.047494
(3, 6, 7)	0.118791	0.046291	0.057376
(4, 5, 6)	0.040391	0.037853	0.036980
(4, 5, 7)	0.118926	0.046359	0.092071
(4, 6, 7)*	0.008801	0.051300	0.043927
(5, 6, 7)	0.031365	0.051300	0.026505
$\sum_{s \in S_1} p(s)$	0.192624	0.381311	0.267531

* nonpreferred sample

Table 2 shows ϕ-values for three controlled plans. Also ϕ for Sampford's uncontrolled plan and ϕ for *MEP*1 are almost the same. However *MEP*2 has a larger ϕ value than

Rao and Nigam but smaller than $MEP1$ and Sampford's uncontrolled design. The results are similar to those of example 1.

5. Algorithm

Since the survey cost and the difficulty of field work are always considered as important variables in a practical survey, we want to construct a controlled plan excepting bad units which require a great deal of survey cost and field work. Therefore let us consider the order of nonpreferred samples ; $s_{I1} < s_{I2} < \cdots < s_{Ik}$, where s_{I1} denotes the worst sample and s_{I2} the next worse sample, $\cdots$, and s_{Ik} k-th worst sample in k nonpreferred sampling units. Among k nonpreferred samples, we are going to obtain a controlled plan excluding $k'(<k)$ bad samples; s_{I1}, $s_{I2}, \cdots, s_{Ik'}$. In this case, Rao and Nigam's method is not appropriate since these bad sampling units are often contained in a controlled plan. Therefore we propose the following algorithm to obtain a controlled plan excluding k' bad sampling units.

For example, suppose that we want to exclude three bad sampling units, (s_{I1}, s_{I2}, s_{I3}), in a controlled plan. Then we follow the next steps in order to obtain the efficient controlled plan satisfying the above conditions.

Step 1 : If controlled plan using Rao and Nigam's method except for s_{I1} does not contain s_{I2} and s_{I3}, we choose this controlled plan. Otherwise we construct the controlled plan except s_{I2} and s_{I3} in order.

Step 2 : If three bad sampling units are not excepted in step 1, we begin by constructing the controlled plan except s_{I1} and s_{I2}. In this sampling plan, if three bad sampling units are not also excepted, we construct the controlled plan, except s_{I1} and s_{I3}, and the next, except s_{I2} and s_{I3}.

Step 3 : If three bad sampling units are not excepted in step 2, finially we construct a controlled plan, except for all three bad sampling units.

We can easily extend the above algorithm to a more general case in which $k'(<k)$ nonpreferred sampling units are deleted in the sampling plan. In particular, if we consider the order of preferred samples together, Rao and Nigam's method is not an appropriate method to obtain the efficient controlled plan. Therefore it is desired to use $MEP2$ to obtain the efficient controlled plan.

References

[1] Avadhani, M. S. and Sukhatme, B. V.(1973). Controlled Sampling with Equal Probabilities and without Replacement, *International Statistical Review*, Vol. 41, No. 2, 175-182.

[2] Chakrabarti, M. C.(1963). On the Use of Incidence Matrices of Designs in Sampling from Finite Populations, *Journal of the Indian Statistical Association*, Vol. 1, No. 1, 78-85.

[3] Ernst, L. R.(1981). A Constructive Solution for Two-Dimensional Controlled Selection Problems, *ASA 1981 Proceeding of the Section on Survey Research Methods*, 61-64.

[4] Food, W. and Hedayat, A.(1977). On Theory and Applications of BIB Designs with Repeated Blocks, *The Annals of Statistics*, Vol. 5, No. 5, 932-935.

[5] Goodman, R. and Kish, L.(1950). Controlled Selection - A Technique in Probability Sampling, *Journal of the American Statistical Association*, Vol. 45, 350-372.

[6] Gupta, V. K., Nigam, A. K., and Kumar, P.(1982). On a Family of Sampling Schemes with Inclusion Probability Proportional to Size, *Biometrika*, Vol. 69, No. 1, 191-196.

[7] Hedayat, A., Lin, Bing-ying, and Stufken, J.(1989). The Construction of PPS Sampling Designs through as Method of Emptying Boxes, *The Annals of Statistics*, Vol. 17, No. 4, 1886-1905.

[8] Hess, I., Riedel, D. C., and Fitzpatrick, T. B.(1975). *Probability Sampling of Hospitals and Patients*, 2nd ed. Ann Arbor, Michigan: Health Administration Press.

[9] Kapur, J. N. and Kesavan, H. K. (1992). *Entropy Optimization Principles with Applications*, Academic Press

[10] Lahiri, D. B.(1951). A Method of Sample Selection Providing Unbiased Ratio Estimates, *Bulletin of the International Statistical Institute*, Vol. 33, No. 2, 133-140.

[11] Lin, Ting-kwong(1992). Some Improvements on an Algorithm for Controlled Selection, *ASA 1992 Proceeding of the Section on Survey Research Methods*, 407-410.

[12] Nigam, A. K., Kumar, P., and Gupta, V. K.(1984). Some Methods of Inclusion Probability Proportional to Size Sampling, *Journal of the Royal Statistical Society*, Series B, Vol. 46, No. 3, 564-571.

[13] Rao, J. N. K. and Nigam, A. M.(1990). Optimal Controlled Sampling Design, *Biometrika*, Vol. 77, N0. 4, 807-814.

[14] Rao, J. N. K. and Nigam, A. M. (1992). Optimal Controlled Sampling: a Unified Approach, *International Statistical Review*, Vol. 60, No. 1, 89-98.

[15] Ryu, J. B. (1996). A study on the Controlled Selection, *Korean Communications in Statistics*, Vol. 3, No. 3, 135-144.

[16] Sitter, R. R. and Skinner, C. J.(1994). Multi-way Stratification by Linear Programing, *Survey Methodology*, Vol. 20, No. 1, 65-73.

[17] Wynn, H. P.(1977). Convex Sets of Finite Population Plans, *The Annals of Statistics*, Vol. 5, No. 2, 414-418.