Survey Methodology, June 2007
Vol. 33, No. 1, pp. 87-94
Statistics Canada, Catalogue No. 12-001

On an optimal controlled nearest proportional to size sampling scheme

Neeraj Tiwari, Arun Kumar Nigam and Ila Pant [1]

Abstract

The concept of 'nearest proportional to size sampling designs' originated by Gabler (1987) is used to obtain an optimal controlled sampling design, ensuring zero selection probabilities to non-preferred samples. Variance estimation for the proposed optimal controlled sampling design using the Yates-Grundy form of the Horvitz-Thompson estimator is discussed. The true sampling variance of the proposed procedure is compared with that of the existing optimal controlled and uncontrolled high entropy selection procedures. The utility of the proposed procedure is demonstrated with the help of examples.

Key Words: Controlled sampling; Non-preferred samples; Quadratic programming; High entropy variance.

1. Introduction

In many situations, some samples may be undesirable due to administrative inconvenience, long distance, similarity of units or cost considerations. Such samples are termed non-preferred samples and the technique for avoiding these samples is known as 'controlled selection' or 'controlled sampling'. This technique, originated by Goodman and Kish (1950) has received considerable attention in recent years due to its practical importance.

The technique of controlled sampling is most appropriate when financial or other considerations make it necessary to select a small number of large first stage units, such as hospitals, firms, schools *etc*., for inclusion in the study. The main purpose of controlled sampling is to increase the probability of selecting a preferred combination beyond that possible with stratified sampling, whilst simultaneously maintaining the initial selection probabilities for each unit of the population, thus preserving the property of a probability sample. This situation generally arises in field surveys where the practical considerations make selection of some units undesirable but it is necessary to follow probability sampling. Controls may be imposed to secure a proper distribution geographically or otherwise and to ensure adequate sample size for some subgroups of the population. Goodman and Kish (1950) considered the reduction of sampling variances of the key estimates as the principal objective of controlled selection, but they also cautioned that this might not always be attained. A real problem emphasizing the need for controls beyond stratification was also discussed by Goodman and Kish (1950, page 354) with the objective of selecting 21 primary sampling units to represent the North Central States. Hess and Srikantan (1966) used the data for the 1961 universe of nonfederal, short-term general medical hospitals in the United States to illustrate the applications of estimation and variance formulae for controlled selection. Waterton (1983) used the data available from a postal survey of Scottish school leavers carried out in 1977 to describe the advantages of controlled selection and compare the efficiency of controlled selection with multiple proportionate stratified random sampling (meaning the sampling scheme in which instead of one stratifying variable, many variables each of which is associated with the variable of interest y, are used by cross-classifying the population on the basis of these variables) and found the controlled selection to perform favourably.

Three different approaches have been advanced in the recent literature to implement controlled sampling. These are (i) using typical experimental design configurations, (ii) the method of emptying boxes and (iii) using linear programming approaches. While some researchers have used simple random sampling designs to construct the controlled sampling designs, one of the more popular strategies is the use of IPPS (inclusion probability proportional to size) sampling designs in conjunction with the Horvitz-Thompson (1952) estimator. To construct controlled simple random sampling designs, Chakrabarti (1963) and Avadhani and Sukhatme (1973) proposed the use of balanced incomplete block (BIB) designs with parameters $v = N$, $k = n$ and λ, where N is the population size and n is the sample size. Wynn (1977) and Foody and Hedayat (1977) used the BIB designs with repeated blocks for situations where non-trivial BIB designs do not exist. Gupta, Nigam, and Kumar (1982) studied controlled sampling designs with inclusion probabilities proportional to size and used BIB designs in conjunction with the Horvitz-Thompson estimator of the population total $Y (= \sum_{i=1}^{N} y_i$, where y_i is the value of the i^{th} unit of the population, U). Nigam, Kumar and Gupta (1984) used some configurations of different types of experimental designs, including BIB designs, to obtain controlled IPPS sampling plans with the

[1] Neeraj Tiwari, Ila Pant, Department of Statistics, Kumaon University, S.S.J. Campus, Almora-263601, India. E-mail: kumarn_amo@yahoo.com; Arun Kumar Nigam, Institute of Applied Statistics & Development Studies, Lucknow-226017, India. E-mail: dr_aknigam@yahoo.com.

additional property $c\pi_i\pi_j \leq \pi_{ij} \leq \pi_i\pi_j$ for all $i \neq j = 1, ..., N$ and some positive constant c such that $0 < c < 1$, where π_i and π_{ij} denote first and second order inclusion probabilities, respectively. Hedayat and Lin (1980) and Hedayat, Lin, and Stufken (1989) used the method of 'emptying boxes' to construct controlled IPPS sampling designs with the additional property $0 < \pi_{ij} \leq \pi_i\pi_j$, $i < j = 1, ..., N$. Srivastava and Saleh (1985) and Mukhopadhyay and Vijayan (1996) suggested the use of 't-designs' to replace simple random sampling without replacement (SRSWOR) designs to construct controlled sampling designs.

All the methods of controlled sampling discussed in the previous paragraph may be carried out manually with varying degrees of laboriousness, but none has exploited the advantage of modern computing. Using the simplex method in linear programming, Rao and Nigam (1990, 1992) proposed optimal controlled sampling designs that minimize the probability of selecting the non-preferred samples, while retaining certain properties of an associated uncontrolled plan. Utilizing the approach of Rao and Nigam (1990, 1992), Sitter and Skinner (1994) and Tiwari and Nigam (1998) used the simplex method in linear programming to solve multi-way stratification problems with 'controls beyond stratification'.

In the present article, we use quadratic programming to propose an optimal controlled sampling design which ensures that the probability of selecting non-preferred samples is exactly equal to zero, rather than minimizing it, without sacrificing the efficiency of the Horvitz-Thompson estimator based on an associated uncontrolled IPPS sampling plan. The idea of 'nearest proportional to size sampling designs', introduced by Gabler (1987), is used to construct the proposed design. The Microsoft Excel Solver of the Microsoft Office 2000 package is used to solve the quadratic programming problem. The applicability of the Horvitz-Thompson estimator to the proposed design is discussed. The true sampling variance of the estimate for the proposed design is empirically compared with the variances of the alternative optimal controlled designs of Rao and Nigam (1990, 1992) and uncontrolled high entropy selection procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In Section 3, some examples are considered to demonstrate the utility of the proposed procedure by comparing the probabilities of non-preferred samples and sampling variances of the estimates. Finally in Section 4, the findings of the paper are summarized.

2. The optimal controlled sampling design

In this section, we use the concept of 'nearest proportional to size sampling designs' to propose an optimal controlled IPPS sampling design that matches the original π_i values, satisfies the sufficient condition $\pi_{ij} \leq \pi_i\pi_j$ for non-negativity of the Yates-Grundy (1953) form of the Horvitz-Thompson (HT) (1952) estimator of the variance and also ensures that the probability of selecting non-preferred samples is exactly equal to zero. Before coming to the proposed plan, we briefly describe the Misdzuno-Sen and Sampford IPPS designs which will be used in the proposed plan for obtaining the initial IPPS design $p(s)$.

2.1 The Midzuno-Sen and Sampford IPPS designs

To introduce the concept of IPPS designs, we assume that a known positive quantity, x_i, is associated with the value of the i^{th} unit of the population and there is reason to believe that the y_i's are approximately proportional to x_i's. Here x_i is assumed to be known for all units of the population and y_i is to be collected for sampled units. In IPPS sampling designs, π_i, the probability of including the i^{th} unit in a sample of size n, is np_i, where p_i is the single draw probability of selecting the i^{th} unit in the population (also known as the normal size measure of unit i), given by

$$p_i = \frac{x_i}{\sum_{j=1}^{N} x_j}, \; i = 1, 2, ..., N.$$

We first describe the Midzuno-Sen IPPS scheme and then discuss Sampford's design.

The Midzuno-Sen (MS) (1952, 1953) scheme has a restriction that the probabilities of selecting the i^{th} unit in the population $(p_i$'s) must satisfy the condition

$$\frac{1}{n}.\frac{n-1}{N-1} \leq p_i \leq \frac{1}{n}, \qquad i = 1, 2, ..., N. \quad (1)$$

If (1) is satisfied for the p_i values of the population under consideration, we apply the MS scheme to get an IPPS plan with the revised probabilities of selection, p_i^*'s, [also known as revised normal size measures] given by

$$p_i^* = n\,p_i \,.\, \frac{N-1}{N-n} - \frac{n-1}{N-n}, \qquad i = 1, 2, ..., N. \quad (2)$$

Now, supposing that the s^{th} sample consists of units $i_1, i_2, ..., i_n$, the probability of including these units in the s^{th} sample under the MS scheme is given by

$$p(s) = \pi_{i_1, i_2, ..., i_n} = \frac{1}{\binom{N-1}{n-1}} \; (p_{i_1}^* + p_{i_2}^* + \; ... \; + p_{i_n}^*). \quad (3)$$

However, due to restriction (1), the MS plan limits the applicability of the method to units that are rather similar in size. Therefore, when the initial probabilities do not satisfy the condition of the MS plan, we suggest the use of Sampford's (1967) plan to obtain the initial IPPS design $p(s)$.

Using Sampford's scheme, the probability of including n units $i_1, i_2, ..., i_n$ in the s^{th} sample is given by

$$\begin{aligned} p(s) &= \pi_{i_1,\, i_2,\, ...,\, i_n} \\ &= n K_n \lambda_{i_1}, \lambda_{i_2}, ..., \lambda_{i_n} \left(1 - \sum_{u=1}^{n} p_{i_u}\right), \end{aligned} \qquad (4)$$

where $K_n = (\sum_{t=1}^{n} t L_{n-t} / n^t)^{-1}$, $\lambda_i = p_i/(1 - p_i)$ for a set $S(m)$ of $m \leq N$ different units, $i_1, i_2, ..., i_m$, and L_m is defined as

$$L_0 = 1,\ L_m = \sum_{S(m)} \lambda_{i_1}, \lambda_{i_2}, ..., \lambda_{i_m} \ \ (1 \leq m \leq N).$$

2.2 The proposed plan

Consider a population of N units. Suppose a sample of size n is to be selected from this population. The single draw selection probabilities of these N units of the population (p_i's) are known. Let S and S_1 denote respectively, the set of all possible samples and the set of non-preferred samples.

Given the selection probabilities for N units of the population, we first obtain an appropriate uncontrolled IPPS design $p(s)$, such as the Midzuno-Sen (1952, 1953) or Sampford (1967) design, as described in Section 2.1. After obtaining the initial IPPS design $p(s)$, the idea behind the proposed plan is to get rid of the non-preferred samples S_1 by confining ourselves to the set $S - S_1$ by introducing a new design $p_0(s)$ which assigns zero probability of selection to each of the non-preferred samples belonging to S_1, given by

$$p_0(s) = \begin{cases} \dfrac{p(s)}{1 - \sum_{s \in S_1} p(s)} & \text{for } s \in S - S_1 \\ \\ 0 & \text{otherwise,} \end{cases} \qquad (5)$$

where $p(s)$ is the initial uncontrolled IPPS sampling plan.

Consequently, $p_0(s)$ is no longer an IPPS design. So, applying the idea of Gabler (1987), we are interested in the 'nearest proportional to size sampling design' $p_1(s)$ in the sense that $p_1(s)$ minimizes the directed distance D from the sampling design $p_0(s)$ to the sampling design $p_1(s)$, defined as

$$D(p_0,\ p_1) = E_{p_0}\left[\frac{p_1}{p_0} - 1\right]^2 = \sum_{s \in S - S_1} \frac{p_1^2(s)}{p_0(s)} - 1 \qquad (6)$$

subject to the following constraints:

$$\begin{aligned} &\text{(i)} && p_1(s) \geq 0, \\ &\text{(ii)} && \sum_{s \in S - S_1} p_1(s) = 1, \\ &\text{(iii)} && \sum_{s \ni i} p_1(s) = \pi_i, \\ &\text{(iv)} && \sum_{s \ni i,\ j} p_1(s) > 0 \quad \text{and} \\ &\text{(v)} && \sum_{s \ni i,\ j} p_1(s) \leq \pi_i \pi_j. \end{aligned} \qquad (7)$$

The ordering of the above five constraints is carried out in accordance with their necessity and desirability. Constraints (i) and (ii) are necessary for any probability sampling design. Constraint (iii), which requires that the selection probabilities in the old and new schemes remain unchanged, which ensures that the resultant design will be IPPS. This constraint is a very strong constraint and it affects the convergence properties of the proposed plan to a great extent. Constraint (iv) is highly desirable because it ensures unbiased estimation of the variance. Constraint (v) is desirable as it ensures the sufficient condition for non-negativity of the Yates-Grundy estimator of the variance.

The solution to the above quadratic programming problem, *viz*., minimizing the objective function (6) subject to the constraints (7), provides us with the optimal controlled IPPS sampling plan that ensures zero probability of selection for the non-preferred samples. The proposed plan is as near as possible to the controlled design $p_0(s)$ defined in (5) and at the same time it achieves the same set of first order inclusion probabilities π_i, as for the original uncontrolled IPPS sampling plan $p(s)$. Due to the constraints (iv) and (v) in (7), the proposed plan also ensures the conditions $\pi_{ij} > 0$ and $\pi_{ij} \leq \pi_i \pi_j$ for the Yates-Grundy estimator of the variance to be stable and non-negative.

The distance measure $D(p_0,\ p_1)$ defined in (6) is similar to the χ^2-statistic often employed in related problems and is also used by Cassel and S≅rndal (1972) and Gabler (1987). Other distance measures are also discussed by Takeuchi, Yanai and Mukherjee (1983). An alternative distance measure for the present discussion may be defined as

$$D(p_0,\ p_1) = \sum_{s} \frac{(p_0 - p_1)^2}{(p_0 + p_1)}. \qquad (8)$$

When applied on different numerical problems considered by us, we found that the use of (8) gave similar results to (6) in convergence and efficiency and so we will give results using (6) as the distance measure.

While all the other controlled sampling plans discussed by earlier authors attempt to minimize the selection probabilities of the non-preferred samples, the proposed plan completely excludes the possibility of selecting non-preferred samples by ensuring zero probability for them and at the same time it also ensures the non-negativity of the Yates-Grundy estimator of the variance. However, in some situations a feasible solution to the quadratic programming problem, satisfying all the constraints in (7), may not exist. Constraint (v) may then be relaxed. This may not guarantee the non-negativity of the Yates-Grundy form of the variance estimator. However, since the condition $\pi_{ij} \le \pi_i \pi_j$ is sufficient for non-negativity of the Yates-Grundy estimator of the variance but not necessary for $n > 2,$ as pointed out by Singh (1954), there will still be a possibility of obtaining a non-negative estimator of the variance. After relaxing the constraint (v) in (7), if the Yates-Grungy estimator of the variance comes out to be negative, an alternative variance estimator may be used. This has been demonstrated in Example 5 in Section 3. If even after relaxing constraint (v), a feasible solution of the quadratic programming problem is not found, constraint (iv) may also be relaxed and consequently an alternative variance estimator in place of the Yates-Grundy form of the HT variance estimator may be used. The effect of relaxing these constraints on efficiency of the proposed design is difficult to study, as after relaxing the non-negativity constraint (v) the Yates-Grundy estimator of the variance does not provide accurate results. Using the Yates-Grundy estimator of the variance, for some problems the variance estimate is smaller after relaxing constraint (v) [as in the case of Examples 2(a), 2(b) and 3(a) in Section 3] while for other problems it is larger [as in the case of Example 1(a), 1(b), 3(b), 4(a) and 4(b) in Section 3]. Relaxing a constraint leading to an increased variance estimate may be due to the inability of the Yates-Grundy form of the variance estimator to estimate the true sampling variance correctly, when the non-negativity condition is not satisfied.

The proposed method may also be considered superior to the earlier methods of optimal controlled selection in the sense that setting some samples to have zero selection probability is different from associating a cost with each sample and then trying to minimize the cost, the technique used in earlier approaches of controlled selection. The technique employed by the earlier authors for controlled selection was a crude approach giving some samples very high cost and others very low.

One limitation of the proposed plan is that it becomes impractical when $\binom{N}{n}$ is very large, as the process of enumeration of all possible samples and formation of the objective function and constraints becomes quite tedious. This limitation also holds for the optimum approach of Rao and Nigam (1990, 1992) and other controlled sampling approaches discussed in Section 1. However, with the advent of faster computing techniques and modern statistical packages, there may not be much difficulty in using the proposed procedure for moderately large populations. On the basis of the size of populations that we have considered in the empirical evaluation, we found that the proposed method can easily handle the controlled selection problems up to a population of 12 units and a sample of size 5. The proposed method may be used to select a small number of first-stage units from each of a large number of strata. This involves a solution of a series of quadratic programming problems, each of a reasonable size, provided the set of non-preferred samples is specified separately in each stratum.

As in the case of linear programming, there is no guarantee of convergence of a quadratic programming problem. Kuhn and Tucker (1951) have derived some necessary conditions for the optimum solution of a quadratic programming algorithm but no sufficient conditions exist for convergence. Therefore unless the Kuhn-Tucker conditions are satisfied in advance, there is no way of verifying whether a quadratic programming algorithm converges to an absolute (global) or relative (local) optimum. Also, there is no way to predict in advance that the solution of a quadratic programming problem exists or not.

2.3 Comparison of sampling variance of the estimate

To estimate the population mean $\bar{Y}(= N^{-1}\sum_{i=1}^{N} y_i)$ based on a sample s of size $n,$ we use the HT estimator of $\bar{Y}$ defined as

$$\hat{\bar{Y}}_{\text{HT}} = \sum_{i \in s} \frac{Y_i}{N\pi_i}. \tag{9}$$

Sen (1953) and Yates and Grundy (1953) showed independently that for fixed size sampling designs, $\hat{\bar{Y}}_{\text{HT}}$ has the variance

$$V(\hat{\bar{Y}}_{\text{HT}}) = \frac{1}{N^2} \sum_{i<j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2, \tag{10}$$

and an unbiased estimator of $V(\hat{\bar{Y}}_{\text{HT}})$ is given as

$$\hat{V}(\hat{\bar{Y}}_{\text{HT}}) = \frac{1}{N^2} \sum_{i<j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \tag{11}$$

Constraint (v), when used in the proposed plan, ensures the non-negativity of the variance estimator (11).

To demonstrate the utility of the proposed procedure, we use the empirical examples given in Section 3 to compare the true sampling variance of the HT estimator for the proposed procedure obtained through (10) with variances of the HT estimator using the optimal controlled plan of Rao and Nigam (1990, 1992) and those of two uncontrolled high entropy (meaning the absence of any detectable pattern or ordering in the selected sample units) procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In what follows, we reproduce the expressions for the variances of these two high entropy procedures.

The expression for variance of $\hat{Y}_{\text{HT}}$ correct to $O(N^{-2})$ using the procedure of Goodman and Kish (1950) is given as

$$V(\hat{Y}_{\text{HT}})_{\text{GK}} = \frac{1}{nN^2}\left[\sum_{i\in U} p_i A_i^2 - (n-1)\sum_{i\in U} p_i^2 A_i^2\right] - \frac{n-1}{nN^2} \times \left[2\sum_{i\in U} p_i^3 A_i^2 - \sum_{i\in U} p_i^2 \sum_{i\in U} p_i^2 A_i^2 - 2\left(\sum_{i\in U} p_i^2 A_i\right)^2\right], \quad (12)$$

where $A_i = Y_i / p_i - Y$, $Y = \sum_{i=1}^{N} Y_i$ and U denotes the finite population of N units.

Recently, Brewer and Donadio (2003) derived the π_{ij}-free formula for the high entropy variance of the HT estimator. They showed that the performance of this variance estimator, under conditions of high entropy, was reasonably good for all populations. Their expression for the variance of the HT estimator is given by

$$V(\hat{Y}_{\text{HT}})_{\text{BD}} = \frac{1}{N^2}\sum_{i\in U}\pi_i(1-c_i\pi_i)\left(\frac{Y_i}{\pi_i^{-1}} - \frac{Y}{n^{-1}}\right)^2, \quad (13)$$

where $c_i = (n-1)/\{n-(2n-1)(n-1)^{-1}\pi_i + (n-1)^{-1}\sum_{k\in U}\pi_k^2\}$ for all $i \in U$, which appears to perform better than the other values of c_i suggested by them.

3. Examples

In this section, we consider some empirical examples to demonstrate the utility of the proposed procedure and compare it with the existing procedures of optimal controlled sampling. In the present discussion, we begin with the Midzuno-Sen (1952, 1953) IPPS design to demonstrate our procedure, as it is relatively easy to compute the probability of drawing every potential sample under this scheme. However, if the conditions of the Midzuno-Sen scheme are not satisfied, we demonstrate that other IPPS sampling without replacement procedures, such as the Sampford (1967) procedure, may also be used to obtain the initial IPPS design $p(s)$. The true sampling variance of the HT estimator under the proposed plan is also compared with that of the existing procedures of optimal controlled selection and uncontrolled high entropy selection procedures given by (12) and (13).

Example 1: Let us consider a population consisting of six villages, borrowed from Hedayat and Lin (1980). The set S of all possible samples consists of 20 samples each of size $n = 3$. Due to the considerations of travel, organization of fieldwork and cost considerations, Rao and Nigam (1990) identified the following 7 samples as non-preferred samples:

123; 126; 136; 146; 234; 236; 246

(a). The Y_i and p_i values associated with the six villages of the population are:

Y_i:	12	15	17	24	17	19
p_i:	0.14	0.14	0.15	0.16	0.22	0.19

Since the p_i values satisfy the condition (1), we apply the MS scheme (3) to get an IPPS plan with the revised normal size measures (p_i^*'s) given by (2).

Applying the method discussed in Section 2 and solving the resulting quadratic programming problem with the Microsoft Excel Solver of Microsoft Office 2000 package, we obtain the controlled IPPS plan given in Table 1.

Table 1 Optimal controlled IPPS plan corresponding to Midzuono-Sen (MS) and Sampford's (SAMP) schemes for Example 1

s	*p*₁(*s*) [MS]	*p*₁(*s*) [SAMP]	*s*	*p*₁(*s*) [MS]	*p*₁(*s*) [SAMP]
124	0.14	0.09	245	0.03	0.12
125	0.03	0.05	256	0.13	0.14
134	0.00	0.00	345	0.02	0.06
135	0.09	0.03	346	0.20	0.10
145	0.03	0.06	356	0.06	0.06
156	0.13	0.07	456	0.06	0.16
235	0.09	0.05			

This plan matches the original π_i values, satisfies the condition $\pi_{ij} \le \pi_i \pi_j$ and ensures that the probability of selecting non-preferred samples is exactly equal to zero. Obviously, due to the fulfillment of the condition $\pi_{ij} \le \pi_i \pi_j$, we can apply the Yates-Grundy form of the HT variance estimator for estimating the variance of the proposed plan.

We have also solved the above example, using plan (3) of Rao and Nigam (1990, page 809) with specified π_{ij}'s taken from the Sampford's plan [to be denoted by RN3] and their plan (4) [to be denoted by RN4]. Using the RN3 plan, the probability of non- preferred samples (ϕ) comes out to be 0.155253 and using the RN4 plan with $c = 0.005$, ϕ comes out to be zero, whereas the proposed plan always ensures zero probability to non-preferred samples.

The values of the true sampling variance of the HT estimator [$V(\hat{Y}_{\text{HT}})$] for the proposed plan, the RN3 plan, the RN4 plan, the Randomized Systematic IPPS sampling plan of Goodman and Kish (1950) [to be denoted by GK] and the uncontrolled high entropy sampling plan of Brewer and Donadio (2003) [to be denoted by BD] are produced in the first row of Table 2. It is clear from Table 2 that the proposed plan yields almost the same value of variance of the HT estimator as yielded by the RN4 plan. The value of $V(\hat{Y}_{\mathrm{HT}})$ for the proposed plan is slightly higher than those obtained from the RN3, GK and BD plans. This increase in variance may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 2 Values of the true sampling variance of the HT estimator [$V(\hat{Y}_{\mathrm{HT}})$] for the Proposed, RN3, RN4, GK and BD plans

$V(\hat{Y}_{\mathrm{HT}})$	RN3	RN4	GK	BD	PROPOSED PLAN
Ex1(a)					
$N = 6$, $n = 3$	2.93	4.02	3.03	2.92	4.06
Ex 1(b)					
$N = 6$, $n = 3$	4.76	5.07	4.89	4.15	4.78
Ex 2(a)					
$N = 7$, $n = 3$	4.48	5.01	4.61	4.45	3.56
Ex 2(b)					
$N = 7$, $n = 3$	11.97	14.52	12.25	11.44	9.49
Ex 3(a)					
$N = 8$, $n = 3$	4.85	4.29	4.96	4.86	3.90
Ex 3(b)					
$N = 8$, $n = 3$	7.29	8.43	7.74	7.37	8.17
Ex 4(a)					
$N = 8$, $n = 4$	3.19	3.46	3.23	3.15	3.75
Ex 4(b)					
$N = 8$, $n = 4$	2.41	2.53	2.54	2.38	2.25
Ex 5					
$N = 7$, $n = 4$	3.08	3.93	3.12	3.07	5.10

(b). Now suppose that the p_i values for the above population of 6 units are as follows:

p_i:	0.10	0.15	0.10	0.20	0.27	0.18

Since these values of p_i do not satisfy the condition (1) of the MS plan, we apply the Sampford (1967) plan to get the initial IPPS design $p(s)$ using (4).

Applying the method discussed in Section 2 and solving the resultant quadratic programming problem, we obtain the controlled IPPS plan given in Table 1. This plan again ensures zero probability to non-preferred samples and satisfies the non-negativity condition for the Yates-Grundy form of the HT variance estimator. This example was also solved by the RN3 and RN4 plans. The value of ϕ for the RN3 plan is 0.064135 and the value of ϕ for the RN4 plan with $c = 0.005$ is zero. The proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{Y}_{\mathrm{HT}})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are produced in the second row of Table 2. The proposed plan appears to perform better than the RN4 and GK plans and quite close to other plans considered by us.

Further examples were constructed to analyze the performance of the proposed plan. The populations with Y_i and p_i values and the set of non-preferred samples for each population are summarized in the Appendix. The p_i values for Examples 2(a), 3(a) and 4(a) satisfy the condition (1) of Midzuno-Sen plan and hence for these examples the Midzuno-Sen IPPS plan is used to obtain the initial IPPS design $p(s)$. However, for Examples 2(b), 3(b) and 4(b) the p_i values do not satisfy this condition and therefore we apply the Sampford IPPS plan to obtain the initial IPPS design. The probabilities of non-preferred samples (ϕ) for these examples using the RN3 plan, the RN4 plan and the proposed method are produced in Table 3. Table 3 shows that while the RN3 and RN4 plans only attempt to minimize the probability of non-preferred samples, the proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{Y}_{\mathrm{HT}})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan for the population summarized in the Appendix are given in Table 2. From Table 2 we conclude that for all the empirical problems considered by us, the proposed plan appears to perform better than or quite close to the RN3, RN4, GK and BD plans. The increase in variance of the estimate for the proposed plan in some cases may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 3 The probabilities of non-preferred samples using RN3, RN4 and Proposed plans

Probability of non-preferred samples (ϕ)	RN3 PLAN	RN4 PLAN	Proposed Plan
Example 2(a) $N = 7$, $n = 3$	0.06	0 ($c = 0.5$)	0
Example 2(b) $N = 7$, $n = 3$	0.05	0 ($c = 0.5$)	0
Example 3(a) $N = 8$, $n = 3$	0.12	0 ($c = 0.005$)	0
Example 3(b) $N = 8$, $n = 3$	0.17	0 ($c = 0.005$)	0
Example 4(a) $N = 8$, $n = 4$	0.05	0 ($c = 0.005$)	0
Example 4(b) $N = 8$, $n = 4$	0.13	0 ($c = 0.005$)	0
Example 5 $N = 7$, $n = 4$	0.30	0.1008 ($c = 0.5$)	0

Example 5: We now consider one more example to demonstrate the situation where the proposed plan fails to provide a feasible solution satisfying all the constraints in (7). In such situations, we have to drop a constraint in (7) to obtain a feasible solution of the related quadratic programming problem.

Consider a population of seven villages. Suppose a sample of size $n = 4$ is to be drawn from this population. There are 35 possible samples, out of which the following 14 are considered as non-preferred:

1234; 1236; 1246; 1346; 1357; 1456; 1567;
2345; 2346; 2456; 2567; 3456; 3567; 4567.

Suppose that the following p_i values are associated with the seven villages:

p_i: 0.14 0.13 0.15 0.13 0.16 0.15 0.14 .

Since the p_i values satisfy condition (1), we apply the MS plan (3) to obtain the initial IPPS design $p(s)$ and solve the quadratic programming problem by the method discussed in Section 2. However, no feasible solution of the related quadratic programming problem exists in this case. Consequently, we drop constraint (v) in (7) for this particular problem to obtain a feasible solution of the quadratic programming problem. The probabilities of non-preferred samples using the RN3 plan, the RN4 plan and the Proposed plan for this empirical problem are given in the last row of Table 3. The proposed plan again matches the original π_i values and ensures the probability of selecting the non-preferred samples exactly equal to zero. However, due to non-fulfillment of the condition $\pi_{ij} \le \pi_i \pi_j$ for this example, the non-negativity of the Yates-Grundy estimator of the variance is not ensured. The values of the true variance, $V(\hat{\bar{Y}}_{\text{HT}})$, for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are produced in the last row of Table 2. The value of $V(\hat{\bar{Y}}_{\text{HT}})$ for this empirical example using the proposed plan does not appear to be satisfactory. For such problems where constraint (v) is not satisfied, we suggest the use of alternative variance estimators in place of the Yates-Grundy variance estimator.

We have also solved one more example with $N = 9$ and $n = 4$ using both the Midzuno-Sen and Sampford's methods for obtaining the initial IPPS design $p(s)$. The details of these solutions are omitted for brevity and can be obtained from the authors.

4. Conclusion

We have proposed a quadratic programming approach to solve the controlled sampling problems ensuring zero probability to non-preferred samples. The concept of 'nearest proportional to size sampling designs' of Gabler (1987) is used to obtain the proposed plan. The approach is simple in concept and is very flexible in allowing for a range of different objective functions as well as in permitting a variety of constraints. The only limitation of the procedure is that it cannot be applied to large populations, as the computational process becomes quite tedious for large populations. The utility of the proposed procedure is demonstrated with the help of examples and its true sampling variance is empirically compared with that of existing controlled sampling plans and uncontrolled high entropy sampling procedures. The proposed plan performs suitably.

Acknowledgements

The authors are grateful to an Associate Editor and two referees for their valuable suggestions and constructive comments on an earlier version of this paper, which led to considerable improvement in presentation of this work.

Appendix

The populations for Example 2-4 with Y_i and p_i values and the set of non-preferred samples.

Example 2. N = 7, n = 3.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467.

Y_i:	12	15	17	24	17	19	25
(a). p_i:	0.12	0.12	0.13	0.14	0.20	0.15	0.14
(b). p_i:	0.08	0.08	0.16	0.11	0.24	0.20	0.13

Example 3. N = 8, n = 3.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467;
128; 178; 248; 458; 468; 478; 578.

Y_i:	12	15	17	24	17	19	25	18
(a). p_i:	0.10	0.10	0.11	0.12	0.18	0.13	0.12	0.14
(b). p_i:	0.05	0.09	0.20	0.15	0.10	0.11	0.12	0.18

Example 4. N = 8, n = 4.

Non-preferred samples: 1234; 1236; 1238; 1246; 1248; 1268; 1346;
1348; 1357; 1456; 1468; 1567; 1568; 1678;
2345; 2346; 2456; 2468; 2567; 2568; 2678;
3456; 3468; 3567; 3678; 4567; 4678; 5678.

Y_i:	12	15	17	24	17	19	25	18
(a). p_i:	0.11	0.11	0.12	0.13	0.17	0.12	0.11	0.13
(b). p_i:	0.09	0.09	0.18	0.11	0.12	0.14	0.17	0.10

References

Avadhani, M.S., and Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *International Statistical Review*, 41, 175-182.

Brewer, K.R.W., and Donadio, M.E. (2003). The high-entropy variance of the Hortivz-Thompson Estimator. *Survey Methodology*, 29, 189-196.

Cassel, C.M., and Särndal, C.-E. (1972). A model for studying robustness of estimators and informativeness of labels in sampling with varying probabilities. *Journal of Royal Statistical Society*, Series B, 34, 279-289.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *Journal of Indian Statistical Association*, 1, 78-85.

Foody, W., and Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *Annals of Statistics*, 5, 932-945.

Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics-Theory & Methods*, 16(4), 1117-1131.

Goodman, R., and Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of American Statistical Association*, 45, 350-372.

Gupta, V.K., Nigam, A.K. and Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, 69, 191-196.

Hedayat, A., and Lin, B.Y. (1980). Controlled probability proportional to size sampling designs. Technical Report, *University of Illinois at Chicago*.

Hedayat, A., Lin, B.Y. and Stufken, J. (1989). The construction of IPPS sampling designs through a method of emptying boxes. *Annals of Statistics*, 17, 1886-1905.

Hess, I., and Srikantan, K.S. (1966). Some aspects of probability sampling technique of controlled selection. *Health Serv. Res. Summer 1966*, 8-52.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *Journal of American Statistical Association*, 47, 663-85.

Kuhn, H.W., and Tucker A.W. (1951). Non-linear programming. *Proceedings of Second Berkely Symposium on Mathematical Statistics and Probability*, 481-492.

Midzuno, H. (1952). On the sampling system with probability proportional to sums of sizes. *Annals of Institute of Statistics & Mathematics*, 3, 99-107.

Mukhopadhyay, P., and Vijayan, K. (1996). On controlled sampling designs. *Journal of Statistical Planning & Inference*, 52, 375-378.

Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society*, B, 46, 564-571.

Rao, J.N.K., and Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.

Rao, J.N.K., and Nigam, A.K. (1992). 'Optimal' controlled sampling: A unified approach. *International Statistical Review*, 60, 89-98.

Sampford, M.R. (1967). On sampling with replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.

Singh, D. (1954). On efficiency of sampling with varying probabilities without replacement. *Journal of Indian Society of Agricultural Statistics*, 6, 48-57.

Sitter, R.R., and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20, 65-73.

Srivastava, J., and Saleh, F. (1985). Need of *t*-designs in sampling theory. *Utilitas Mathematica*, 28, 5-17.

Takeuchi, K., Yanai, H. and Mukherjee, B.N. (1983). The Foundations of Multivariate Analysis. 1st Ed. New Delhi: Wiley Eastern Ltd.

Tiwari, N., and Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning & Inference*, 69, 89-100.

Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Wynn, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics*, 5, 414-418.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society*, B, 15, 253-261.