Statistics and Applications {ISSN 2452-7395 (online)}
Volume 16 Nos. 1, 2018 (New Series), pp 145-169

Controlled Sampling – A Review

A K Nigam[1], Neeraj Tiwari[2] and B N Mandal[3]
[1]*Institute of Applied Statistics and Development Studies, Bengaluru*
[2]*Kumaun University, S.S.J. Campus, Almora*
[3]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

Received: November 25 2017; Revised: February 10 2018; Accepted: February 17 2018

Abstract

The purpose of the paper is to present developments that have taken place in the area of controlled sampling. The article first deals with one dimensional controlled sampling plans based on combinatorics of experimental designs and linear programming approaches. Some special classes of controlled sampling plans namely balanced sampling plans avoiding adjacent units and distance balanced sampling plans are discussed. An overview of two or more-dimensional controlled sampling plans is also given. Some statistical applications of controlled sampling are discussed in brief.

Key words: Controlled sampling, experimental designs, linear programming, quadratic programming, BSA plans, DBSP plans, IPPS plans, two or more-dimensional controlled sampling, sample coordination, controlled rounding

1. Introduction

Consider a finite population containing *N* distinct and identifiable units. The purpose in sample theory is the estimation of some population parameter e.g. population mean of a character of interest *Y by* observing a sample *s* of size *n* (< *N*) units from the population. Let *S* denote the set of all possible samples containing *n* distinct units. The set *S is*, called the sample space for selecting *n* units out of *N* units in the population and has cardinality $^{N}C_{n}$. In certain situations, the sample space *S* may be partitioned into two disjoint sets S_1 and S_2 such that the set S_1 contains preferred samples and the set S_2 contains non-preferred samples. The objective is to reduce the probability of selection of any sample from the set S_2 to as small as possible, preferably to zero and to increase the probability of selection from S_1 in such a way that the desirable features of uncontrolled sampling designs are retained. This partitioning of the sample space *S* may be due to several reasons. For example, administrative inconvenience because of geographical spread of the sampling units may lead to increase in travel cost or difficulty in accessing the sampling units. Here S_2 contains those samples containing far away sampling units. Sometimes the set of non-preferred samples S_2 may contain nearer units which gives almost identical information. A sampling plan containing minimum number of samples from S_2 is called a controlled sampling plan.

Goodman and Kish (1950) introduced controlled sampling as a sample selection method beyond stratification which reduces the probability of selecting non-preferred

Corresponding Author: A K Nigam
E-mail: dr_aknigam@yahoo.com

samples while retaining the desirable features of an un-controlled probability sampling design. It is a useful method for selecting first stage units in multi-stage sampling. While the method allows unbiased estimation of population total/mean, it does not provide unbiased variance estimation because sampling is no longer independent in the strata. For further work on Goodman-Kish method, one may refer to Hess et al. (1976), Waterton (1983), Causey et al. (1985) and Cox (1987).

A new dimension to controlled selection was added by Chakrabarty (1963) by establishing a connection between a balanced incomplete block (BIB) design and simple random sampling without replacement (SRSWOR). This approach was further exploited by Avadhani and Sukhatme (1973) who obtained controlled sampling designs with SRS properties. For further work on controlled sampling, one may refer to Wynn (1977), Foody and Hedayat (1977), Srivastava and Saleh (1985) and Mukhopadhyay and Vijayan (1996).

Combinatorial properties of experimental designs were further exploited by Nigam and Co-workers (Gupta et al., 1982; Nigam et al., 1984) for obtaining controlled sampling designs with inclusion probability proportional to size (IPPS). Hedayat et al. (1989) used the method of 'emptying boxes' to construct controlled IPPS sampling plans with the additional property $0 < \pi_{ij} \le \pi_i \pi_j, i < j = 1,2,..., N$ where π_i denotes the first order inclusion probability of the *i*th unit (i = 1,2, ..., N) and π_{ij} denotes the second order inclusion probability of the pair of units *i* and *j*, $i \ne j = 1,2,..., N$.

It is realized that the focus of these efforts was limited to reducing the support size. Rao and Nigam (1990) generalized these approaches by using linear programming to derive optimal controlled sampling designs with specified properties and minimum probability of selection of non-preferred samples. Rao and Nigam (1992) further generalized it to derive optimal controlled sampling plans which match the variance of a generalized linear unbiased estimator of a specified uncontrolled sampling plan. They covered the Narain-Horvitz-Thompson estimator under IPPS, Murthy’s estimator under sampling with probabilities proportional to sizes and without replacement and ratio estimator under a plan with probability of selection proportional to aggregate size of sample units.

A major drawback of all the above-mentioned approaches is that the methods select the whole sample of *n* units instead of sample selection through unit by unit selection. From the selected whole sample, the next step is to identify the selected units. Nigam and Gupta (1984) proposed a method which facilitates identification of units contained in the selected sample in case of SRSWOR. Nigam and Singh (1994) extended the procedure to SRS with replacement. However, both the procedures are difficult to implement for large *N*.

A new dimension was added by Hedayat et al. (1988) by developing controlled sampling plans excluding contiguous units. Further work on this came from a number of authors (Stufken, 1993; Colbourn and Ling, 1998, 1999; Stufken et al., 1999; Stufken and Wright, 2001, 2008; Mandal, 2007; Mandal et al., 2008, 2011; Tahir et al., 2010, 2012; and Kumar et al., 2016).

Tiwari and Co-Workers (Tiwari and Nigam, 1998; Tiwari et al., 2007; Tiwari and Nigam, 2010; Tiwari and Sud, 2011; Tiwari and Chilwal, 2013) developed controlled sampling plans facilitating control in two or more-dimensions. For other related literature, the reader may refer to Goodman and Kish (1950), Bryant et al., (1960), Bryant (1961), Hess and

Srikantan (1966), Moore et al., (1974), Jessen (1970, 1973, 1975), Gabler (1987), Sitter and Skinner (1994), and Lu and Sitter (2002). For an excellent review on controlled sampling designs till 2012, the readers may refer to Gupta et al. (2012).

2. Experimental design approach

There is a long history of using experimental designs in survey sampling. Some recent applications include controlled sampling, handling of sensitive questions and balanced subsamples for variance estimation, balanced bootstrap, among others. A good source of the applications of experimental designs in survey sampling is by Rao and Vijayan (2008). Some of the references cited in the last section relate to some of those exploiting the interplay between experimental designs and survey sampling.

Most work on controlled sampling exploits the combinatorial properties of various incomplete block designs to construct designs with minimum support size, that is, with minimum number of distinct blocks. Maximum possible numbers of distinct blocks are identified with the preferred samples and the remaining with the non-preferred samples. One block or sample is then selected at random or with pre-assigned probabilities from the totality of blocks, *b*, in the chosen design.

In a pioneering work, Chakrabarty (1963) showed that a balanced incomplete block (BIB) design can be used to obtain a controlled sampling plan with appealing properties. Considering the symbols of BIB design as sampling units, the blocks as samples and treatments in a block as the units in the sample, he showed that if from a BIB design with $v = N$ symbols in b $(\leq {}^{N}C_{n})$ blocks of size $k = n$ with number of replications r and number of concurrences of pairs of symbols λ, one block is selected with probability $1/b$, then this sampling design gives the same first and second order inclusion probabilities as in SRSWOR design. This can be easily seen that each unit in a BIB design belongs to r blocks and hence, with equal probability of selection $1/b$, the first order inclusion probability is $r/b = k/v = n/N$. Similarly, the second order inclusion probabilities for a pair of units is $\lambda/b = r(k-1)/b(v-1) = k(k-1)/v(v-1) = n(n-1)/N(N-1)$. The blocks of the BIB design are so selected that, as far as possible, they are from the set of preferred samples S_1. Clearly, with appropriate choice of N and n, the design has much reduced support size than an SRSWOR design.

The following example from Avadhani and Sukhatme (1973) illustrates the use of BIB designs in construction of a controlled sampling design.

Example 1: Suppose n = 3 villages are to be selected from N = 7 villages which are located as shown below.

*	2	*	1	*
7	*	5	*	4
*	6	*	3	*

The following 14 samples (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) and (4, 6,7) are considered as non-preferred because of inconvenience in field work. Avadhani and Sukhatme (1973) suggested to use the BIB design given below with parameters $v = N = 7$, $b = 7$, $k = n = 3$, $r = 3$ and $\lambda = 1$ as the controlled sampling design: Block 1: (1, 2, 4); Block 2: (2, 3, 5); Block 3: (3, 4, 6); Block 4: (4, 5, 7); Block 5: (5, 6, 1); Block 6: (6, 7, 2) and Block 7: (7, 1, 3)*.

Considering blocks as samples, each of the blocks are given equal probability of selection 1/7. Since only block 7 is a non-preferred sample, the probability of getting a non-preferred sample is only 1/7. Note here that the probability of getting a non-preferred sample using an uncontrolled SRSWOR design with n = 3 is = 14/${}^7C_3$ = 14/35 = 2/5.

For large N and n, a BIB design may not exist and if exists, it may be difficult to construct. For such situations, Avadhani and Sukhatme (1973) suggested the following approach.

i) Divide N units at random into g disjoint groups with ith group having N_i units so that $N_1 + N_2 + ... + N_g = N$.

ii) Let $n_i = nN_i/N$ be an integer, i = 1, 2, ..., g. Take an integer n_i' such that a BIB design with parameters ($v = n_i'$, b_i, r_i, n_i, λ_i) exists, where $n_i < n_i' < N_i$, i = 1, 2, ..., g. Then select a random sample of n_i' units from N_i units in the ith group and do it independently for each group.

iii) Find out the preferred combinations of n_i units from n_i' units and make a one-to-one correspondence between the blocks of the BIB design and the preferred combinations. Then select one block at random from the BIB design. This is done for each group independently. Thus, the g selected blocks from g BIB designs will constitute the controlled sample of size $n_1 + n_2 + ... + n_g = n$.

The idea of using block designs was further extended by Srivastava and Saleh (1985) and Mukhopadhyay and Vijayan (1996) who suggested the use of t-designs to construct controlled sampling plans. Wynn (1977) and Foody and Hedayat (1977) suggested use of BIB designs with repeated blocks for situations where a BIB design with $b < {}^vC_k$ does not exist.

Controlled plans with inclusion probabilities proportional to sizes of units (IPPS) were derived first by Gupta et al. (1982). They used BIB designs to obtain controlled plans satisfying IPPS property, that is, $\pi_i = np_i$, where $p_i = x_i/X$, x_i is a size measure attached to the i-th unit and X is the population total of the x_i's. These controlled plans are applicable only to those populations which satisfy $\sum_{i \in s} p_i > \frac{n-1}{N-1}$, for all $s \in S_c$. Here, s denotes the sample, S_c is the support of the plan and the y_i are the values of characteristic of interest. Hedayat et al. (1989) used the method of 'emptying boxes' to construct controlled IPPS sampling plans with the additional property $0 < \pi_{ij} \le \pi_i \pi_j, i < j = 1,2,...,N$. The latter property ensures the unbiasedness and non-negativity of the well-known Sen-Yates-Grundy (Sen, 1953; Yates and Grundy, 1953) variance estimator for the Narain-Horvitz Thompson (Narain, 1951; Horvitz and Thompson, 1952) estimator ($\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}, s \in S_c$) of the population total Y:

$$\hat{V}\left(\hat{Y}_{HT}\right) = \sum_{i<} \sum_{j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, s \in S_c . \tag{2.1}$$

However, there is no guarantee that the variance of $\hat{Y}_{HT}$ will always be smaller than the variance of the estimator under probability proportional to size (PPS) sampling with replacement. Nigam et al. (1984) used typical configurations of experimental designs, including BIB designs, to obtain controlled IPPS sampling plans with the additional property $c\pi_i\pi_j \le \pi_{ij} \le \pi_i\pi_j, i < j = 1,2,...,N$, where c (< 1) is a positive constant, say c = 1/2. The property $\pi_{ij} \ge c\pi_i\pi_j$ is expected to ensure that the variance estimator (2.1) remains stable. Wynn (1977) showed that for a given sampling plan $p_1(.)$, there always exists a sampling plan $p_2(.)$ with support size no greater than $N(N-1)/2$ and with the same π_{ij}'s (and hence π_i's) as $p_1(.)$. This result, however, does not show us how to construct such a controlled plan $p_2(.)$.

It should be clear from the above overview of controlled sampling that the focus of previous work has been on reducing the support size rather than minimizing the probability of selecting non-desired combinations to arrive at an optimal controlled plan. The criterion of minimum support size, in fact, is not necessarily relevant in constructing an optimal controlled plan (see example 1, Rao and Nigam, 1990). Further, the experimental design approach could often involve considerable trial and error and computations.

3. Linear programming approach

Experimental design approach is useful in obtaining controlled sampling designs. However, unless a design is chosen in such a way that it reduces the probability of selecting non-preferred samples, the chance of getting a non-preferred sample can be quite high. Such an exercise of selecting a suitable design involves a lot of trial and error. To this end, Rao and Nigam (1990, 1992) suggested linear programming approach which minimizes the probability of non-preferred samples and they termed such a controlled sampling design as optimal. They suggested the following formulation to obtain an optimal controlled sampling design with desired second order inclusion probabilities:

Minimize $\phi = \sum_{s \in S_2} p(s)$ subject to constraints

i) $\sum_{s \in S} p(s) = 1$

ii) $\sum_{s \ni i,j} p(s) = \pi_{ij} = \dfrac{n(n-1)}{N(N-1)} \forall i, j$ (3.1)

iii) $p(s) \ge 0 \;\; \forall s \in S$

where $p(s)$ denote the probability of selecting a sample s of size n.

They have shown that an optimal solution to the formulation readily gives an optimal controlled sampling plan. In fact, all the controlled sampling plans provided by earlier authors can be easily obtained by the proposed approach. For example, consider the problem of Avadhani and Sukhatme (1973) in Example 1. For instance, using the formulation (3.1) in Example 1 above, an optimal solution is obtained, which once again gives $\phi_{\min} = 1/7$, with an optimal controlled plan given in Table 1.

Table 1: Optimal controlled simple random sampling plan of Example 1

s	*p* (*s*)	*s*	*p* (*s*)
1 2 4	0.11429	2 3 5	0.14286
1 2 7	0.02857	2 4 6*	0.02857
1 3 4	0.02857	2 6 7	0.11429
1 3 7*	0.11429	3 4 6	0.11429
1 5 6	0.1428-6	3 6 7	0.02857
4 5 7	0.14286		

However, the optimal solution to the formulation (3.1) is not unique and there may be more than one solution to the same problem. For example, the solution of Avadhani and Sukhatme (1973) given in Example 1 is also an optimal solution to the formulation (3.1) with $\phi_{\min}$ = 1/7.

The constraints on *p*(*s*) in (3.1) were suitably changed by Rao and Nigam (1990) to obtain optimal controlled IPPS plans:

Minimize $\phi = \sum_{s\in S_2} p(s)$ subject to constraints

$$\text{i)} \quad \sum_{s\in S} p(s) = 1$$

$$\text{ii)} \ \sum_{s\ni i} p(s) = np_i \ \forall \ i = 1,2,\ldots,N \tag{3.2}$$

$$\text{iii)}\, c(np_i)(np_j) \le \sum_{s\ni i,j} p(s) \le (np_i)(np_j)$$

$$\text{iv)} \ \ p(s) \ge 0 \ \ \forall s \in S\,,$$

where *c* is a suitably chosen constant.

Using the formulation (3.2), Rao and Nigam (1990) also obtained the controlled IPPS plan with *N* = 7, *n* = 3 and p_i values 0.25, 0.19, 0.16, 0.15, 0.12, 0.08, 0.05 of Nigam et al. (1984). In this case, the value of $\phi_{\min}$ using (3.2) was 0.17 whereas the solution of Nigam et al. had ϕ = 0.32.

Moreover, the objective function may also be suitably modified by giving appropriate weights to the probability of selecting samples so that the expected cost of the survey is reduced. For details, please see Rao and Nigam (1990, 1992) and Gupta et al. (2012).

One drawback of the linear programming approach suggested by Rao and Nigam (1990, 1992) is that it becomes impractical for large *N* and *n*. To handle this problem, Lahiri and Mukerjee (2000) suggested a modified version to reduce the dimensionality of the problem. For this purpose, they divided the population of *N* units into *t* equivalence classes of cardinalities N_1, N_2, ..., N_t with $N_1 + N_2 + ... + N_t = N$. Within one equivalence class, units are associates of each other. Here two units are said to be associate if the set S_n remains unaltered if the roles of the units are interchanged. Then they defined a linear programming formulation which has fewer constraints and decision variables than the formulations (3.1) and (3.2).

4. Quadratic programming approach

Tiwari et al. (2007) used the idea of ‘nearest proportional to size sampling designs’, originated by Gabler (1987), to propose a one-dimensional optimal controlled IPPS sampling design that matches the original inclusion probabilities (π_i’s) of each unit in the population and ensures zero probability to non-preferred samples.

In their plan, using the given selection probabilities for *N* units of the population (p_i’s), Tiwari et al. (2007) first obtained an appropriate uncontrolled IPPS design p(s), such as Sampford (1967) or Midzuno-Sen (1952, 1953) design. After obtaining the initial IPPS design $p(s)$, the idea behind their plan is to get rid of the non-preferred samples S_2 by confining to the set $S_1 = S - S_2$ by introducing a new design $p_0(s)$ which assigns zero probability of selection to each of the non-preferred samples belonging to S_2, given by

$$p_0(s) = \begin{cases} \dfrac{p(s)}{1 - \sum\limits_{s \in S_2} p(s)}, & \text{for } s \in S_1 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

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), they obtained 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

$$D(p_0, p_1) = E_{p_0}\left[\frac{p_1}{p_0} - 1\right]^2 = \sum_s \frac{p_1^2(s)}{p_0(s)} - 1 \tag{4.2}$$

subject to the following constraints:

$$\begin{array}{ll} \text{(i)} & p_1(s) \geq 0 \\ \text{(ii)} & \sum\limits_{s \in S_1} p_1(s) = 1 \\ \text{(iii)} & \sum\limits_{s \ni i} p_1(s) = \pi_i \\ \text{(iv)} & \sum\limits_{s \ni i,j} p_1(s) > 0 \\ \text{(v)} & \sum\limits_{s \ni i,j} p_1(s) \leq \pi_i \pi_j \,. \end{array} \tag{4.3}$$

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 sampling design. Constraint (iii), which requires that the selection probabilities in the old and new schemes remain unchanged, is the requirement for IPPS design, which ensures that the resultant design will be IPPS. This constraint is strong and it affects the convergence properties of the proposed plan. Constraint (iv) is highly desirable because it ensures unbiased estimation of variance. Constraint (v) is desirable as it ensures the sufficient condition for non-negativity of the Sen-Yates-Grundy estimator of variance.

The solution to the above quadratic programming problem, viz., minimizing the objective function (4.2) subject to the constraints (4.3), 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 (4.1) 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. Due to the constraints (iv) and (v) in (4.3), the proposed plan also ensures the conditions $\pi_{ij} > 0$ and $\pi_{ij} \leq \pi_i\pi_j$ for Sen-Yates-Grundy estimator of the variance to be stable and non-negative.

5. Balanced sampling plans excluding adjacent units

As stated earlier, Hedayat et al. (1988) developed controlled sampling plans excluding contiguous units. These were termed as balanced sampling plans excluding contiguous units (BSEC plans). These plans are useful when population units are arranged in space or time and the contiguous units provide similar measurement and hence, it is desirable that contiguous units do not appear in a sample. The BSEC plans are sampling plans in which every pair of non-contiguous units has constant second order inclusion probabilities and every pair of contiguous units has zero second order inclusion probabilities. Stufken (1993) extended this idea to balanced sampling plans excluding adjacent units (BSA plans) where every pair of adjacent units has zero second order inclusion probabilities and every pair of non-adjacent units has constant second order probabilities. Here, two units are said to be adjacent whenever they are within a distance of *m* units with *m* suitably defined by the investigator. However, given *N*, *n* and *m*, BSA plans may not always exist. Existence conditions have been studied by Stufken (1993), Stufken et al. (1999) and Wright (2008). A special class of block designs called polygonal designs were introduced by Stufken et al. (1999) to obtain BSA plans. A polygonal design in *v* treatments and *b* blocks with each block of size *k* is an incomplete block design such that (i) no treatment appears more than once in a block, (ii) every treatment appears in *r* blocks in the design, and (iii) each pair of treatments which are at a distance of *m* units or less do not appear together in any block and each pair of treatments which are at distance of more than *m* units, appear together in λ blocks. The parameters *v*, *b*, *r*, *k*, λ and *m* satisfy the following necessary conditions (i) $vr = bk$ and (ii) $\lambda(v-2m-1) = r(k-1)$.

The design given in Table 2 is a polygonal design with $v = 9$, $b = 9$, $r = 3$, $k = 3, \lambda = 1, m = 1$.

Table 2: A polygonal design for v = 9, b = 9, r = 3, k = 3, $\lambda = 1$ and m = 1

1	3	6
2	4	7
3	5	8
4	6	9
5	7	1
6	8	2
7	9	3
8	1	4
9	2	5

A polygonal design has one to one correspondence with a BSA plan. With $N = v$ and $k = n$, consider the treatments as sampling units, the blocks as samples, the treatments in a block as the units in the sample and then if every block of a polygonal design is given probability of selection as $1/b$, then the polygonal design is equivalent to a BSA plan for population size N, sample size n and m. A result due to Mandal et al. (2008) and Stufken and Wright (2008) for constructing polygonal designs is given below.

Theorem 1. Let $B_1, B_2, ..., B_t$ denote t initial blocks with k distinct treatments from the set {1, 2, ..., v}. Let $\mathrm{B}_u = \{b_{u1}, b_{u2}, ..., b_{uk}\}$, u = 1, 2, ..., t. Then if in the $tk(k - 1)$ pair wise distances of the elements from the t blocks, distances 1, 2, ..., m do not appear and distances m + 1, m + 2, ...,[v/2] appear λ times then, a polygonal design is obtained by developing the t initial blocks modulo v. The parameters of the design are given by v, $b = tv$, $r = tk$, k, λ and m.

Theorem 1 was utilized to develop algorithms to obtain BSA plans by Stufken and Wright (2001) for m = 1 and by Mandal et al. (2008) and Stufken and Wright (2008) for $m \geq 1$. A catalogue of BSA plans is available in Mandal (2007) for $N \leq 40, n \leq 7, m \leq 4$. Theorem 1 always gives polygonal designs which are cyclic in nature. An integer linear programming formulation to identify the generator blocks of such cyclic polygonal designs was proposed by Mandal et al. (2011). They constructed all the polygonal designs for $v \leq 100, k = 3$ for all permissible m using the approach.

Mandal et al. (2008) proposed a linear programming approach to obtain BSA plans following the idea of Rao and Nigam (1990, 1992). They suggested the following linear programming formulation for obtaining a BSA plan:

Minimize $\phi = \sum_{s \in S_2} p(s)$

subject to constraints

i) $\sum_{s \ni i} p(s) = \dfrac{n}{N}$

ii) $\sum_{s \ni i, j} p(s) = 0$ if (i, j) are adjacent (5.1)

$= \dfrac{n(n-1)}{N(N-2m-1)}$ if (i, j) are non-adjacent

iii) $p(s) \geq 0 \; \forall s$

iv) $\sum_{s} p(s) = 1$

where S_2 denotes set of samples containing adjacent pair of units.

An optimum solution of the linear programming formulation, if exists, gives the full support of the plan along with the probability of selections. One of the limitations of the proposed linear programming approach for construction of a BSA plans is that for large N and n, the number of possible samples becomes very large and the linear programming formulation becomes impractical to adopt.

Several authors suggested alternative approaches to obtain BSA plans. Colbourn and Ling (1998) used partial triple system to solve the existence problem of BSA plans for k = 3 with m = 1. Colbourn and Ling (1999) completely solved the problem of existence of BSA plans for k = 4 and m = 1. Tahir et al. (2010, 2012) used cyclic shift method to construct polygonal designs for k = 3 for particular settings of λ and m.

The above researches assumed that the population units have a circular ordering. This assumption is unrealistic and hence, BSA plans have been obtained considering that units have linear ordering and units have two-dimensional ordering. Stufken and Wright (2008) presented the results on existence of linear BSA plans. Mandal et al. (2008) presented a linear programming approach to obtain linear BSA plans. Very recently, Kumar et al. (2016) obtained several new smaller BSA plans through a integer linear programming based algorithm and the algorithm can produce designs which may or may not be cyclic which is not the case in earlier works.

For the situations when the size measures of the units may vary greatly and the information on the size measures is available for all the *N* population units and nearer units provide similar information, inclusion probability proportional to size sampling plans excluding adjacent units (IPPSEA plans) have been introduced by Mandal et al. (2008). Under IPPSEA plans, in a sample, no two adjacent units appear together and $\pi_i = np_i$, $i = 1,2,...,N$ with $p_i = x_i / \sum x_i$, where x_i is the size measure of the *i*th sampling unit. IPPSEA plans may be obtained by trial and error methods using combinatorial properties of block designs. Mandal et al. (2008) also proposed linear programming approach to obtain IPPSEA plans for both circular and linear arrangement of the popular units.

6. Distance balanced sampling plans

One limitation of BSA plans is that they do not permit unbiased estimation of the variance of the Narain-Horvitz-Thompson estimator of the population mean. Variance approximation approaches were suggested by Wright and Stufken (2011) for BSA plans. Mandal et al. (2009) introduced distance balanced sampling plans (DBSP) where unbiased variance estimation is possible. In a DBSP, any two units which are at same distance from each other have constant second order inclusion probabilities and this inclusion probability is greater for a pair of units which are at a greater distance than a pair which are at a lesser distance. In other words, if the distance between two distinct units (*i*, *j*) is greater than the distance between two distinct units (*k*, *l*), then $\pi_{ij} \geq \pi_{kl}$ and π_{ij} = constant for all $i \neq j$ at same distance under DBSP.

To understand the properties of a DBSP, assume that the units are arranged in one-dimensional circular order unless otherwise specified. This assumption is needed for algebraic simplicity. Let the distance between two units *i* and *j* be denoted as δ(*i*, *j*), $i \neq j$ = 1,2,...,*N*. Under circular ordering of units, δ(*i*, *j*) can be written as δ(*i*, *j*) = *min*{|*i* – *j*|, *N* – |*i* – *j*|}. For example, if *N* = 8, then distance between units 4 and 6 is δ(4,6) = *min*{|4 – 6|,8 – |4 – 6|} = *min*{2,6} = 2. Under circular ordering, the possible values that δ(*i*, *j*) can take are 1,2,...,[*N*/2], where [*x*] denote the largest value contained in *x*.

Now, let $\pi_{ij} = \gamma\ f(i,j), i \neq j = 1,2,...,N$, where γ is a constant so that $0 \leq \pi_{ij} \leq 1$ and $f(i,j)$ is a non-decreasing function of δ(*i*, *j*). It is easy to see that $\pi_i = \sum_{j(\neq i)=1}^{N} \gamma \frac{f(i,j)}{n-1}$, $i = 1,2,...,N$. If we let all the first order inclusion probabilities to be equal then, $\pi_i = \frac{n}{N}$ and hence, it leads to

$$\pi_{ij} = \frac{n(n-1)\ f(i,j)}{N\ F_i}, \tag{6.1}$$

where, $F_i = \sum_{j(\neq i)=1}^{N} f(i, j)$. Note that here π_{ij}'s depend on the choice of $f(i, j)$. It may be seen that if for all i and j, $f(i, j) = f$, a constant, then DBSP reduces to SRSWOR, if $f(i, j) = 0$ for $\delta(i, j) \leq m$ and $f(i, j) = f$ for $\delta(i, j) > m$ then these reduce to BSA plans.

An example of a DBSP for N = 5 and n = 3 is given in Table 3. The plan has first order inclusion probabilities $\pi_i = 3/5$ and the second order inclusion probabilities $\pi_{12} = \pi_{23} = \pi_{34} = \pi_{45} = \pi_{15} = 1/5$ and $\pi_{13} = \pi_{24} = \pi_{35} = \pi_{14} = \pi_{25} = 2/5$.

Table 3: A distance balanced sampling plan for *N* = 5, *n* = 3

	s		***p*(*s*)**
1	2	4	1/5
2	3	5	1/5
3	4	1	1/5
4	5	2	1/5
5	1	3	1/5

Let us now consider the estimation of population mean Ȳ using the Narain-Horvitz-Thompson estimator which is given by

$$\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i}. \tag{6.2}$$

The Sen-Yates-Grundy form of variance of Narain-Horvitz-Thompson estimator for the DBSP, using (6.1) can be easily obtained as

$$V(\hat{\bar{Y}}_{HT})_{DBSP} = \sigma^2 - \frac{n-1}{Nn}\sum_{i=1}^{N}\sum_{i<j=1}^{N} d_{ij}(Y_i - Y_j)^2 \tag{6.3}$$

where $d_{ij} = \dfrac{f(i, j)}{\sum_{j(\neq i)=1}^{N} f(i, j)}$ and $\sigma^2 = \dfrac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

An unbiased estimator of variance of Narain-Horvitz-Thompson estimator of population mean under DBSP is given by

$$\hat{V}(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{i<j=1}^{n}\left(\frac{N}{n(n-1)d_{ij}} - \frac{N^2}{n^2}\right)(y_i - y_j)^2. \tag{6.4}$$

Clearly, with suitable choice of $f(i, j)$, an unbiased estimator of variance of Narain-Horvitz-Thompson estimator always exists. In simple words $f(i, j)$ should be chosen such that no p_{ij} is zero for unbiased variance estimation.

It is evident that a large number of DBSP may be obtained depending on the choice of the function $f(i, j)$. Mandal et al. (2009) considered two particular cases namely

1. two-point DBSP: $f(i, j) = f_1$ for $\delta(i, j) \leq m$ and $f(i, j) = f_2(\geq f_1)$ for $\delta(i, j) > m$ and

2. three-point DBSP: $f(i,j) = f_1$ for $\delta(i, j) \le m_1$ and $f(i,j) = f_2$ for $m_1 < \delta(i,j) \le m_2$ and $f(i,j) = f_3$ for $\delta(i,j) > m_2$ with $f_1 \le f_2 \le f_3$.

Later on, Mandal et al. (2010) considered *w*-point (*w* = 1,2,...,[*N*/2]), DBSP where $f(i,j) = f_{t+1}$ whenever $m_t < \delta(i,j) \le m_{t+1}$, *t* = 0,1,2,...,*w* – 1. Here, $m_0 = 0, m_w = [N/2]$, f_t's are non-negative integers such that $f_{t+1} \ge f_t$.

It has been shown by Mandal et al. (2009) that a DBSP always exist for *n* = 2. In that case, the plan is given by {*s*, *p*(*s*)} with $p(s) = \pi_{ij}$, where π_{ij} is obtained from (6.1). For example, let *N* = 8 and *f*(*i*, *j*) = δ(*i*, *j*), $i \ne j$ = 1, 2,..., *N*. Then the second order inclusion probabilities can be easily obtained as π_{ij} = 1/64 for δ(*i*, *j*) = 1, $\pi_{ij} = 2/64$ for δ(*i*, *j*) = 2, $\pi_{ij} = 3/64$ for δ(*i*, *j*) = 3 and $\pi_{ij} = 4/64$ for δ(*i*, *j*) = 4. Therefore, the DBSP plan is given as in Table 4.

Table 4: A DBSP for *N* = 8, *n* = 2

s		*p*(*s*)	*s*		*p*(*s*)	*s*		*p*(*s*)	*s*		*p*(*s*)
1	2	1/64	2	3	1/64	3	5	2/64	4	8	4/64
1	3	2/64	2	4	2/64	3	6	3/64	5	6	1/64
1	4	3/64	2	5	3/64	3	7	4/64	5	7	2/64
1	5	4/64	2	6	4/64	3	8	3/64	5	8	3/64
1	6	3/64	2	7	3/64	4	5	1/64	6	7	1/64
1	7	2/64	2	8	2/64	4	6	2/64	6	8	2/64
1	8	1/64	3	4	1/64	4	7	3/64	7	8	1/64

Establishing existence of the plans for *n* ≥ 3 is not trivial. Combinatorial properties of block designs may be used to construct such designs. Mandal et al. introduced a parallel block design structure called distance balanced incomplete block (DSBIB) designs, which are equivalent to DBSP.

Definition 6.1 A distance balanced incomplete block design is an incomplete block design in *v* treatments arranged in *b* blocks of size *k* each such that

1. each block contains *k* distinct treatments,
2. a pair of treatments (*i*, *j*) with distance $\delta(i, j)$ appear together in λ_{ij} blocks and
3. λ_{ij}'s are non-decreasing in $\delta(i, j)$.

In case of circularly arranged population, this implies that each treatment has same number of replications *r*. The parameters of the design are v, b, r, k, λ_{ij} and they satisfy following necessary conditions.

$$i)vr = bk \text{ and ii) } \sum_{j(\ne i)=1}^{v} \lambda_{ij} = r(k-1) \qquad (6.5)$$

The design given in Table 5 is an example of a DSBIB design for *v* = 7, *b* = 14, *r* = 6, *k* = 3 with columns representing blocks. It may be noted that the design has $\lambda_{ij} = 1$ if δ(*i*, *j*) = 1, $\lambda_{ij} = 2$ if δ(*i*, *j*) = 2 and $\lambda_{ij} = 3$ if δ(*i*, *j*) = 3. In other words, $\lambda_{12} = \lambda_{23} = \lambda_{34} = \lambda_{45} = \lambda_{56} = \lambda_{67} = \lambda_{71} = 1$, $\lambda_{13} = \lambda_{24} = \lambda_{35} = \lambda_{46} = \lambda_{57} = \lambda_{61} = \lambda_{72} = 2$ and $\lambda_{14} = \lambda_{25} = \lambda_{36} = \lambda_{47} = \lambda_{51} = \lambda_{62} = \lambda_{73} = 3$.

Table 5: A DSBIB design with *v* = 7, *b* = 14, *k* = 3

1	2	3	4	5	6	7	1	2	3	4	5	6	7
2	3	4	5	6	7	1	3	4	5	6	7	1	2
5	6	7	1	2	3	4	5	6	7	1	2	3	4

Note that DSBIB designs are a broad class of incomplete block designs and contain many sub-classes depending on how λ_{ij}'s are defined. If $\lambda_{ij} = \lambda$, a constant irrespective of distance between two treatments *i, j, $i \neq j$* then they reduce to BIB designs. When $\lambda_{ij} = \lambda$ whenever $\delta(i, j) > m$ and $\lambda_{ij} = 0$, whenever $\delta(i, j) \leq m$, then they reduce to polygonal designs introduced by Stufken et al. (1999). A *w*-point (*w* = 1, 2, ..., [*v*/2]) DSBIB designs may be defined analogous to a *w*-point DBSP. In a *w*-point DSBIB design, $\lambda_{ij} = \lambda_{t+1}$ whenever $m_t < \delta(i, j) \leq m_{t+1}$, *t* = 0, 1, ..., *w*–1 and $\lambda_{t+1} \geq \lambda_t$. In other words, in a *w*-point DSBIB design, λ_{ij}'s can take *w* distinct values and they are non-decreasing in distance. Therefore, in a DSBIB design, λ_{ij}'s can take at most [*v*/2] distinct values.

It is easy to see that with *N* = *v* and *k* = *n*, if every block of a DSBIB design is given probability of selection 1/*b*, then the design is equivalent to a DBSP. Thus, obtaining a DSBIB design is equivalent to obtaining a DBSP. Mandal et al. (2010) presented a integer linear programming formulation to obtain DBSPs. There is a lot of scope for development of algebraic methods of construction of DSBIB designs. For a detailed overview of DBSP plans, see Mandal et al. (2016).

7. Multi-dimensional controlled plans

For some populations, there may be two or more stratifying criteria that are desirable in the sampling design. Multi-way stratification allows increased precision of estimates of each of the variables whose precision is increased by typical univariate estimators corresponding to single criteria designs. The sample need not be allocated to every multi-way population cell induced by a set of single-criteria designs.

Multi-dimensional controlled selection is needed when number of strata cells exceeds the permissible sample size. In a survey on fish catch, Bryant (1961) used four different types of strata, viz., location, times of a day, season of summer and type of day. The study had five locations, two times of a day, four seasons, and two types of day. This made a total of 80 strata cells out of which only 46 cells were to be sampled due to financial considerations. Hess and Srikantan (1966) reported a hospital survey with hospital size, region of the state and size of community as three stratification variables. The study had 4 sizes, 4 regions and 3 community sizes leading to a total of 48 strata cells. Two different samples were investigated in this study with *n* = 50 and *n* = 100. Jessen (1970) reported a survey of households for the city of Santa Monica, California. The two stratification variables were area and home value. There were 12 geographical areas and 12 income classes, i.e., 144 strata cells. However, the funds were available to sample just 24 of the cells. Moore et al. (1974) described the National Assessment of Educational Progress (NAEP) study having regions of the country, states within country, states within regions, socio-economic levels and size of community levels as different strata. There were 4 regions, each region had 12 to 15 states, three socio-economic levels and 3 sizes of community levels. In a 12-state region, there was a potential for 12×3× 3 = 108 strata cells, however, the resources and other considerations permitted samples of 27 cells only.

It is easily realised that with two or more stratification criteria, there are not enough resources to provide each cell with an adequate sample size, particularly if estimates of variances are desired. Several procedures have been proposed to overcome these situations. Main contributions are from Goodman and Kish (1950) under the name of “controlled selection”, and from Bryant et al. (1960) and Jessen (1970, 1973, 1975) under the name of “lattice sampling”.

Equal probability methods

The equal probability methods are due to Bryant et al. (1960) and Jessen (1975), the latter being superior in many ways. We therefore restrict ourselves only to Jessen’s work.

When two variables are used for stratification and each has the same number of levels, say L, there are L^2 strata cells. The sample size is $n = r\times L$, where r cells are to be selected from each row and each column of a square lattice of order $L\times L$. Two methods are proposed by Jesssen. These are ‘random lattice’ and ‘Latin lattice’. Both these methods are, however, identical to choosing the cells corresponding to any r letters from a randomly selected Latin square of order L. Their analytical properties for estimating variance may differ in the sense that ‘random lattice’ is proposed to be analyzed by considering r split samples, each sample corresponding to a different letter of the Latin square. This results in $r - 1$ degrees of freedom. Similarly, the ‘Latin lattice’ is to be analyzed by taking all the $r\times r$ Latin lattices separately, there being p such lattices provided L is even and $L/p = r$.

Using analysis of variance arguments, Jessen showed that generally such a two-way selection procedure is more efficient than both a simple random sample of cells and a one-way stratification using either rows or columns.

When two variables used for stratification have unequal number of levels, the square lattice concept can be extended to rectangular lattices of say R rows and C columns. Here again, the sample selection can be made by choosing, say any r letters from a randomly selected incomplete Latin square with R rows and C columns. Similarly, in sampling from three dimensions, with each dimension having L levels, there will be L^3 strata cells and a sample of size $n = r^2\times L$ can be selected by using $r\times r\times r$ Latin cubes. The procedure will again be identical with that of choosing r letters from a randomly selected Graeco Latin square. An extension of this to multi-dimensional stratification should be obvious using hyper Graeco Latin squares.

Unequal probability Methods

Sometimes, the cells may contain an unequal number of units. It may then be desirable to sample the cells in such a way that this universe is taken care of. Such situations may arise when the sampling is done in two stages.

In this method, first a sample of cells is chosen and then a sample of units is drawn from the selected cells. Jessen (1970, 1973, 1975) discussed two methods of ‘probability lattices’ for the purpose. We discuss here some details from Jessen (1970).

Let A_{rc} be the relative measure of size (probability) of the rc^{th} element (strata cell), such that

$$\sum_{r=1}^{R}\sum_{c=1}^{C} A_{rc} = 1 \tag{7.1}$$

where the two-dimensional sampling frame consists of N elements arranged in R rows and C columns.

We wish to draw a sample of size n out of the N elements such that the inclusion probability of rc^{th} cell in a sample is

$$\pi_{rc} = n\,A_{rc} \;\leq 1, \tag{7.2}$$

with the constraints,

$$E\,\{n_{rc}\} = n\,A_{rc}, \tag{7.3}$$

$$|\, n_{rc} - n\,A_{rc}\,| < 1\,, \tag{7.4}$$

$$|\, n.c - n\,A.c\,| < \;1\,, \tag{7.5}$$

and

$$|\, nr. - n\,Ar.\,| \;<\; 1\,, \tag{7.6}$$

where $A.c$ and $Ar.$ denote the marginal column and row totals, respectively.

Linear and quadratic programming approach

Sitter and Skinner (1994) applied the ideas of Rao and Nigam (1990) to multi-way stratification. However, they did not consider controls beyond stratification. Tiwari and Nigam (1998) suggested a method for two-dimensional controlled selection using simplex method in linear programming. They also proposed an alternate variance estimator for controlled selection designs, as Narain-Horvitz-Thompson estimator could not be used to their plan due to non-fulfillment of the condition $\pi_{ij} \leq \pi_i\pi_j$. Lu and Sitter (2002) developed some methods to reduce the amount of computation so that very large problems became feasible using the linear programming approach.

Tiwari and Nigam (2010) extended the idea of Tiwari et al. (2007) to propose a two-dimensional optimal controlled IPPS sampling design that matches the original inclusion probabilities (π_i's) of each unit in the population and ensures zero probability to non-preferred samples. Proceeding in the similar manner as described in the case of a one-dimensional optimal controlled IPPS sampling designs described in Section 4, Tiwari and Nigam (2010) obtained the 'nearest proportional to size sampling design' $p_1(s)$ in the sense that $p_1(s)$ minimizes the directed distance $D(p_0,\, p_1)$ from the sampling design $p_0(s)$ to the sampling design $p_1(s)$, subject to the following constraints:

(i) $p_1(s) \geq 0$

(ii) $\sum\limits_{s \in S_1} p_1(s) = 1$ (7.7)

(iii) $\sum\limits_{S \ni i} p_1 = np_i$, for $p_1(s)$ to be an IPPS design,

where $D(p_0,\, p_1)$ and $p_0(s)$ are defined in (4.2) and (4.1), respectively.

The constraints (i) and (ii) in (7.7) are necessary for any sampling design while the constraint (iii) ensures that the resultant design $p_1(s)$ is an IPPS design.

The greatest difficulty with the multi-dimensional controlled selection problems is that as the magnitude and complexity of the problem increases, the process of enumeration of all possible samples becomes quite tedious. The methodological modification in multi-dimensional approach over the one-dimensional approach is that only a sub-set of the ${}^{N}C_n$ combinations which satisfy the marginal constraints of the given multi-dimensional problem are considered as the set of all possible samples. With multi-dimensional controlled selection problems, the potential difficulty lies in the fact that the non-negativity condition of the Sen-Yates-Grundy form of the Narain-Horvitz-Thompson variance estimator is not satisfied. This leads to introduction of an alternative variance estimator for multi-dimensional controlled selection problems.

The distance measure $D(p_0, p_1)$ defined in (4.2) is like the χ^2 -statistic often employed in related problems and is also used by Cassel and Särndal (1972) and Gabler (1987). Some other distance measures are also discussed by Takeuchi et al. (1983). Two alternative distance measures may be defined as:

$$D(p_0,\, p_1) = \sum_s \left| p_0(s) - p_1(s) \right|$$

$$D(p_0, p_1) = \sum_s \frac{(p_0(s) - p_1(s))^2}{(p_0(s) + p_1(s))} \qquad (7.8)$$

These distance measures gave comparable results as obtained through the distance measure (4.2).

While all other two-dimensional optimal controlled selection plans discussed by earlier authors attempt to minimize the selection probabilities of the non-preferred samples, the proposed plan eliminates the non-preferred samples by assigning zero probabilities to them. The proposed plan is superior to the plans of Sitter and Skinner (1994) and Tiwari and Nigam (1998) in the sense that it ensures zero probability to non-preferred samples and is much nearer to the controlled design ($p_0(s)$), which we wanted to achieve due to practical considerations. Moreover, the proposed plan also incorporates the possibility of 'controls beyond stratification', which was not considered by Sitter and Skinner (1994).

Tiwari and Sud (2011) suggested minimum variance optimal controlled nearest proportional to size sampling scheme using multiple objective functions. Following the method suggested by Tiwari et al. (2007), they first obtained an appropriate uncontrolled inclusion probability proportional to size (IPPS) design $p(s)$. After obtaining the initial IPPS design $p(s)$, we get rid of the non-preferred samples (S_1) by restricting ourselves to the set $S - S_1$ by introducing the design $p_0(s)$, as proposed by Tiwari et al. (2007).

The first objective function (ϕ_1) in this plan is same as in the plan proposed by Tiwari et al. (2007). To minimize the true sampling variance of the Narain-Horvitz-Thompson estimator, Tiwari and Sud (2011) included one more objective function to the quadratic programming problem, given by

$$\phi_2 = Var\left(\hat{Y}_{HT}\right) = \frac{1}{N^2}\left\{np_i np_j - \sum_{s \ni i,j} p_1(s)\right\}\left(\frac{Y_i}{np_i} - \frac{Y_j}{np_j}\right)^2 \qquad (7.9)$$

Thus, the problem becomes

Minimize $(\phi_1 + \phi_2)$ subject to the constraints similar to those suggested by Tiwari et al. (2007).

Tiwari and Chilwal (2013) discussed a simplified approach for minimum variance optimal controlled selection. Let y be the characteristic under study, Y_i the y-value for the i^{th} unit in the population (i = 1, ... , N) and y_l the y-value for the lth unit in the sample (l = 1, ... , n). Let S_k, k = 1, ... , L, denote the k^{th} possible samples. Also let s_{ik} be each internal entry of S_k. Then s_{ik} equals either $[a_i]$ or $[a_i] + 1$, where $[a_i]$ is the integer part of a_i. We have to consider a set of samples with selection probabilities that satisfy the constraints

$$E(s_{ik} \mid i) = \sum_{i \in S_k, S_k \in S} s_{ik}\, p(S_k) = na_i \tag{7.10}$$

and

$$\sum_{S_k \in S} p(S_k) = 1, \tag{7.11}$$

where S is the set of all possible samples $\{S_k\}$, and $p(S_k)$ is the selection probability of each sample S_k.

There can be many sets of probability distributions $p(S_K)$ satisfying (7.10) and (7.11), although only one set of probabilities can be used to obtain a solution to the controlled selection problem. We may consider an algorithm based on an appropriate and objective principle to find the solution that reflects the closeness of each sample S_K to A. For this purpose, Tiwari and Chilwal (2013) considered the following measures of closeness between A and S_k.

(i) The ordinary distance, which is often called the Euclidean distance:

$$D_1(A, S_k) = \left[\sum_{i=1}^{N} (a_i - s_{ik})^2 \right]^{1/2}, k = 1, ..., L \tag{7.12}$$

(ii) The Cosine Distance Function:

$$D_2(A, S_k) = 1 - \sum_{i=1}^{N} \frac{a_i s_{ik}}{\|A\|_2 \|S_k\|_2} \tag{7.13}$$

(iii). The Bray-Curtis Distance Function:

$$D_3(A, S_k) = \frac{\sum_{i=1}^{N} (a_i - s_{ik})}{\sum_{i=1}^{N} (a_i + s_{ik})}. \tag{7.14}$$

The distance function D_2 provides minimum variance.

Multiple objective linear programming problem considered by Tiwari and Chilwal (2013) is described as below:

Minimize the objective function $\phi_1 + \phi_2$, where,

$$\phi_1 = \sum_{S_k \in S} D_2(A, S_k)\, p(S_k) \tag{7.15}$$

and

$$\phi_2 = Var\left(\hat{\hat{Y}}_{HT}\right) = \frac{1}{N^2} \left\{ na_i na_j - \sum_{S_k \ni i, j} p(S_k) \right\} \left[\frac{Y_i}{na_i} - \frac{Y_j}{na_j} \right]^2 \tag{7.16}$$

subject to the following constraints

$$
\begin{aligned}
&\text{(i)} && p(S_k) \geq 0 \\
&\text{(ii)} && \sum_{S_k \in S - S^*} p(S_k) = 1 \\
&\text{(iii)} && \sum_{S_k \ni i} p(S_k) = \pi_i \\
&\text{(iv)} && \sum_{S_k \ni i,j} p(S_k) > 0 \\
&\text{(v)} && \sum_{S_k \ni i,j} p(S_k) \leq \pi_i \pi_j \left(i < j = 1,\ldots,N\right). \\
&\text{(vi)} && \sum_{S_k \in S^*} p(S_k) = 0
\end{aligned}
\tag{7.17}
$$

where S^* denotes the set of non-preferred samples.

The solution of above linear programming problem, viz., minimization of sum of (7.15) and (7.16) subject to the constraints (7.17), provides optimal controlled IPPS sampling plan that ensures zero probability of selection for the non-preferred samples and also minimizes the true sampling variance of the HT estimator. This method also provides an opportunity to add more objective functions to the controlled selection problem.

The plan suggested by Tiwari and Chilwal (2013) is superior to the approach of Rao and Nigam (1990), Sitter and Skinner (1994) and Tiwari and Nigam (1998) in the sense that these plans only attempt to minimize the selection probabilities of the non-preferred samples. Whereas the proposed plan ensures zero probability to non-preferred samples through constraint (vi) in (7.17). The exclusion of non-preferred samples was also attempted by Tiwari et al. (2007) and Tiwari and Sud (2011), using the idea of nearest proportional to size design. However, their procedure is quite lengthy and tedious as an uncontrolled IPPS design is to be manually constructed and then the required controlled IPPS design is derived using the quadratic linear programming approach. The same advantage is achieved in the plan suggested by Tiwari and Chilwal (2013) in a very simple manner by just adding one more constraint in the linear programming problem that ensures zero probability to non-preferred samples. Their plan also minimizes the true sampling variance of HT estimator using the second objective function given in (7.16).

Two-dimensional BSA plans were first introduced by Bryant et al. (2002) when population units are arranged in two dimensions. To define adjacency between units in two dimensional populations, Wright (2008) introduced the concept of adjacency scheme. They also proposed one direct search algorithm for obtaining such plans. Gopinath et al. (2018) proposed a linear programming formulation to construct two dimensional BSA plans under different adjacency schemes given by Wright (2008).

8. Some statistical applications of controlled sampling

Now we discuss some important applications of controlled sampling to various statistical problems. These are sample coordination problem and statistical disclosure control. As the constraint of space prohibits us to discuss these applications in detail, in what follows is a brief description of these applications.

Sample coordination is combining information from different sources (samples), say, about the income of the households from one sample and on literacy from another sample from the same population. It is usually desirable to select units that can be taken as a sample for both characteristics. It can be achieved by minimizing the number of distinct units in the union of the samples.

Causey et al. (1985) proposed an optimum linear programming procedure for maximizing the expected number of sampling units common to the two designs, when the two sets of sample units were chosen sequentially. Ernst and Ikeda (1995) also presented a linear programming procedure for overlap maximization under very general conditions. Other important references include Ernst (1996, 1998), Ernst and Paben (2002), Deville and Tillé (2000) and Matei and Tillé (2006).

Matei and Skinner (2009) constructed optimal sampling designs for given unit inclusion probabilities in order to realize maximum coordination. Tiwari and Sud (2012) proposed an improved method for sample coordination problem when sample units are selected simultaneously. Their method maximizes (or minimizes) the overlap of sampling units between the two designs, with identical stratifications, without putting any restriction on the number of sample units in a stratum. The procedure also facilitates variance estimation using Sen-Yates-Grundy) form of Narain-Horvitz-Thompson variance estimator.

Statistical disclosure control is the requirement of statistical offices to protect the confidentiality of data it collects. The procedure involves identification of sensitive cells and then protecting the confidential information contained in sensitive cells. Statistical disclosure control can be achieved through two methods namely, controlled rounding and cell suppression.

Controlled rounding is the problem of optimally rounding real valued entries in a tabular array to adjacent integer values in a manner that preserves the tabular structure of the array. Rounding methods are used for many purposes, like improving the readability of data values, to control statistical disclosure in tables, to solve the problem of iterative proportional fitting (or raking) in two-way tables and controlled selection.

Rounding techniques involve the replacement of the original data by multiples of a given rounding base. Controlled rounding problem is the problem of optimally rounding real valued entries in a tabular array to adjacent integer values in a manner that preserves the tabular structure of the array. Rounding methods are used for many purposes, such as for improving the readability of data values, to control statistical disclosure in tables, to solve the problem of iterative proportional fitting (or raking) in two-way tables and controlled selection. Statistical disclosure control is one of the area in which rounding methods are widely used. Fellegi (1975) proposed a technique for random rounding which unbiasedly rounds the cell values and also maintains the additivity of the rounded table. Cox and Ernst (1982) used the transportation theory in linear programming to obtain an optimal controlled rounding of a two way tabular array.

Causey et al. (1985) summarized the idea of Cox and Ernst (1982) and used the transportation theory to solve the controlled rounding problem. They discussed several statistical applications in which controlled rounding can be used and applied the concept of controlled rounding to solve the controlled selection problem. Cox (1987) presented a constructive algorithm for achieving unbiased controlled rounding which is simple to implement by hand. Tiwari and Nigam (1993) improved the method of Cox (1987) to terminate in fewer steps. Salazar (2005) proposed a technique, termed as cell perturbation, which allows reducing the data loss from controlled rounding. This method is closely related to the classical controlled rounding methods and has the advantage that it also ensures the protection of sensitive cells to a specified level, while minimizing the loss of information.

Another method widely used by different researchers for protecting sensitive cells in a table is the method of cell suppression in which sensitive cells are not published i.e. they are suppressed. These suppressed sensitive cells are called primary suppressions. To make sure that the primary suppressions cannot be derived by subtraction from published marginal totals, additional cells are selected for suppressions, which are known as complementary suppressions or secondary suppressions. Remaining cells in the table are published with their original values.

This problem has been widely discussed by Cox (1980, 1995), Sande (1984), Carvalho et al. (1994) and Fischetti and Salazar (2000). In cell suppression, a large amount of information is lost as in addition to suppression of sensitive cells, some non-sensitive cells are also suppressed. To reduce the loss of information, Fischetti and Salazar (2003) proposed an improved methodology, known as partial cell suppression, in which instead of wholly suppressing primary and complementary suppressed cells, some intervals obtained with the help of a mathematical model, are published for these cell entries. The loss of information in partial cell suppression is smaller in comparison to complete cell suppression. Tiwari (2012) used the idea of random rounding and quadratic programming to propose an improved methodology for disclosure control in an array that perturbs only the sensitive cells and adjusts some non-sensitive cells to preserve the marginal values of the array.

9. Open problems

A major drawback of all the above-mentioned approaches is that the methods select the whole sample of n units instead of sample selection through unit by unit selection. From the selected whole sample, the next step is to identify the selected units. This is a very cumbersome and time-consuming procedure even with moderately large N and n. Development of a controlled sampling plan using unit by unit selection procedure remains an open problem.

It is noted that so far, the controlled sampling is limited to only two disjoint sets, preferred and non-preferred. In general, we may have p disjoint sets, S_1, ..., S_p with p-order preferences. For instance, in example 1 of Avadhani and Sukhatme, samples 1, 3, 7 and 1, 4, 7 have larger distance than 1, 2, 3 and have higher degree of non-preference. There is a need to reframe the linear programming to address to the situations as above.

Very little work is available for BSA plans when population units are arranged in two dimensions. There is a need to look into the problem of existence and construction of BSA plans for such scenarios. DBSP plans are very attractive alternative to BSA plans. However, they are available only for populations with circular ordering of units and also for very limited number of sample sizes. Further work is required for existence and construction of DBSP for bigger sample sizes as well as for populations with linear and two-dimensional arrangement of units. Connection of DBSP with spatially balanced sampling may also be explored.

It remains to be investigated whether the two or more-dimensional controlled plans also lead to new row-column type designs with repeated blocks. In a row-column design there are two types of blocks, rows and columns. Latin square design is the most trivial row-column design.

Further attempts may be made to develop methods to reduce the amount of computation in single and multi dimensional controlled selection using linear and quadratic programming in the lines of Lahiri and Mukerjee (2000) and Lu and Sitter (2002), so that even large problems are addressed without much difficulty. Unbiased variance estimation in multi dimensional controlled selection problems, when the non-negativity condition of the Sen-Yates-Grundy form of Narain-Horvitz-Thompson variance estimator is not satisfied, is another open area where further work is needed.

Controlled selection using ranked set sampling is another area which may be explored extensively. One attempt in this direction was made by Al-Saleh and Zheng (2003) by using multistage ranked set sampling technique to obtain a controlled preferred sample. If the properties of ranked set sampling are exploited to obtain controlled selection designs, this may lead to significant changes in the method of sample selection and may also increase the efficiency of the estimates.

References

Al-Saleh, M.F. and Zheng, G. (2003). Controlled sampling using ranked set sampling. *Journal of Nonparametric Statistics*, **15(4-5)**, 505-516.

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

Bryant, D., Chang, Y., Rodger, C.A. and Wei, R. (2002). Two-dimensional balanced sampling plans excluding contiguous units. *Communications in Statistics - Theory and Methods*, **31(8)**, 1441-1455.

Bryant, E.C. (1961). *Sampling methods*, Seminar paper, Iowa State University.

Bryant, E.C., Hartley H.O. and Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of American Statistical Association*, **55**, 105-124.

Carvalho, F.D., Dellaert, N.P. and Osório, M.S. (1994). Statistical disclosure in two-dimensional tables: General tables. *Journal of American Statistical Association*, **89**, 1547-1557.

Cassel, C.M., and Sarndal, 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(2)**, 279-289.

Causey, B.D., Cox, L.H. and Ernst, L.R. (1985). Application of transportation theory to statistical problems. *Journal of American Statistical Association*, **80**, 903-909.

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.

Colbourn, C.J., and Ling, A.C. (1998). A class of partial triple systems with applications in survey sampling. *Communications in Statistics - Theory andMethods*, **27(4)**, 1009-1018.

Colbourn, C.J. and Ling, A.C.H. (1999). Balanced sampling plans with block size four excluding contiguous units. *Australasian Journal of Combinatorics*, **20**, 37-46.

Cox, L.H. (1980). Suppression methodology and statistical disclosure control. *Journal of American Statistical Association*, **75**, 377-385.

Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of American Statistical Association*, **82**, 420-424.

Cox, L.H. (1995). Network models for complementary cell suppression. *Journal of American Statistical Association*, **90**, 1453-1462.

Cox, L.H. and Ernst, L.R. (1982). Controlled rounding. *INFOR*, **20**, 423-432.

Deville, J.-C. and Tillé, Y. (2000). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, **86**, 215-227.

Ernst, L.R. (1996). Maximizing the overlap of sampling units for two designs with simultaneous selection. *Journal of Official Statistics*, **12**, 33-45.

Ernst, L.R. (1998). Maximizing and Minimizing overlap when selecting a large number of units per stratum simultaneously for two designs. *Journal of Official Statistics*, **14**, 297-314.

Ernst, L.R. and Ikeda, M. (1995). A reduced size transportation algorithm for maximizing the overlap between surveys. *Survey Methodology*, **21**, 147-157.

Ernst, L.R. and Paben, S.P. (2002). Maximizing and minimizing the overlap when selecting any number of units per stratum simultaneously for two designs with different stratifications. *Journal of Official Statistics*, **18**, 185-202.

Fellegi, I.P. (1975). Controlled random rounding. *Survey Methodology*, **1**, 123-135.

Fischetti, M. and Salazar, J.J. (2000). Models and algorithms for optimizing cell suppression in tabular data with linear constraints. *Journal of American Statistical Association*, **95**, 916-928.

Fischetti, M. and Salazar, J.J. (2003). Partial cell suppression: A new methodology for statistical disclosure control. *Statistics and Computing*, **13**, 13-21.

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

Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics - Theory and 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.

Gopinath, P.P., Parsad, R. and Mandal, B.N. (2018). Two dimensional balanced sampling plans excluding adjacent units under sharing a border and island adjacency schemes, *Communications in Statistics - Simulation and Computation*, 47(3), 712-720. DOI: 10.1080/03610918.2017.1291959.

Gupta, V.K., Mandal, B.N. and Parsad, R. (2012). *Combinatorics in sample surveys vis-à-vis controlled selection*. Lambert Academic Publishing, Germany.

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.S., Rao, C.R., and Stufken, J. (1988). Sampling plans excluding contiguous units. *Journal of Statistical Planning and Inference*, **19(2)**, 159-170.

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 the probability sampling technique of controlled selection. *Health Services Research*, **1(1)**, 8-52.

Hess, I., Riedel, D.C. and Fitzpatrick, T.B. (1976). *Probability Sampling of Hospitals and Patients.* University of Michigan, Ann Arbor, second edition.

Horvitz, D. G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, **47**, 663–685.

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal of American Statistical Association*, **65**, 776-796.

Jessen, R.J. (1973). Some properties of probability lattice sampling. *Journal of American Statistical Association*, **68**, 26-28.

Jessen, R.J. (1975). Square and cubic lattice sampling. *Biometrics*, **31**, 449-471.

Kumar, R., Parsad, R. and Mandal, B.N. (2016). Smaller balanced sampling plans excluding adjacent units for one dimensional populations. *International Journal of Computational and Theoretical Statistics*, **3(2)**, 55-61.

Lahiri, P. and Mukerjee, R. (2000). On simplification of the linear programming approach to controlled sampling. *Statistica Sinica*, **10**, 1171-1178.

Lu, W. and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. *Survey Methodology*, **28(2)**, 199-207.

Mandal, B.N. (2007). *Combinatorics and its applications with special reference to sample surveys*. Unpublished Ph.D. Thesis, Indian Agricultural Research Institute, New Delhi.

Mandal, B.N., Parsad, R., and Gupta, V.K. (2008). IPPS sampling plans excluding adjacent units, *Communications in Statistics - Theory and Methods*, 37(16), 2532-2550.

Mandal, B.N., Gupta, V.K. and Parsad, R. (2011). Construction of polygonal designs using linear integer programming. *Communications in Statistics - Theory and Methods*, **40(10)**, 1787-1794.

Mandal, B.N., Gupta, V.K. and Parsad, R. (2016). *Distance balanced sampling plans - an overview*, in Statistical and Mathematical Sciences and their Applications, Eds. Neeraj Tiwari, Narosa Publishing House, India.

Mandal, B.N., Parsad, R. and Gupta, V.K. (2010). Linear integer programming approach to construct distance balanced sampling plans. *Journal of Indian. Society of Agricultural Statistics*, **64(2)**, 303–312.

Mandal, B.N., Parsad, R., and Gupta, V.K. (2008). Computer-aided constructions of balanced sampling plans excluding adjacent units. *Journal of Statistics and Applications*, **3**, 59-85.

Mandal, B.N., Parsad, R., Gupta, V.K. and Sud, U.C. (2009). A family of distance balanced sampling plans. *Journal of Statistical Planning and Inference*, **139(3)**, 860–874.

Matei, A. and Skinner, C. (2009). Optimal sample coordination using controlled selection. *Journal of Statistical Planning and Inference*, **139**, 3112-3121.

Matei, A. and Tillé, Y. (2006). Maximal and minimal sample co-ordination. *Sankhyā*: *An Indian Journal of Statistics*, **67**, 590-612.

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

Moore, R.P, Chromy, J.R. and Rogers, W.T. (1974). *The national assessment approach to sampling*. National Assessment of Educational Progress., Denver.

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

Narain, R.D. (1951). On sampling without replacement with varying probabilities, *Journal of Indian. Society of Agricultural Statistics*, **3**, 169-174.

Nigam, A.K. and Gupta, V.K. (1984). A method of sampling with equal or unequal probabilities without replacement. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), **33(2)**, 227-229.

Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society, Series B*, **46(3)**, 564-571.

Nigam, A.K. and Singh, R.K. (1994). A method of sampling with replacement. *Sankhya* B, **56(3)**, 369-373.

Rao, J.N.K. and Nigam, A.K. (1990). Optimum 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.

Rao, J.N.K. and Vijayan, K. (2008). Application of experimental designs in survey sampling. *Journal of Indian. Society of Agricultural Statistics*, **62**, 126-131.

Salazar, J.J. (2005). Controlled rounding and cell perturbation: Statistical disclosure limitation methods for tabular data. *Mathematical Programming* B, **105**, 583-603

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54(3-4)**, 499-513.

Sande, G. (1984). Automated cell suppression to preserve confidentiality of Business Statistics. *Statistical Journal of United Nations Economic Commission for Europe*, **2**, 33-41.

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

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.

Stufken, J. (1993). Combinatorial and statistical aspects of sampling plans to avoid the selection of adjacent units. *Journal of Combinatorics Information and System Science*, **18**, 81-92.

Stufken, J. and Wright, J.H. (2008). New balanced sampling plans excluding adjacent units. *Journal of Statistical Planning and Inference*, **138**, 3326 – 3335.

Stufken, J. and Wright, J.H. (2001). Polygonal designs with blocks of size $k \leq 10$. *Metrika*, **54**, 179-184.

Stufken, J., Song, S.Y., See, K. and Driessel, K.R. (1999). Polygonal Designs: some existence and non-existence results. *Journal of Statistical Planning and Inference*, **77**, 155-166.

Tahir, M.H., Iqbal, I., Akhtar, M. and Shabbir, J. (2010). Cyclic polygonal designs with block size 3 and $\lambda = 1$ for joint distance $\alpha = 6$ to 16. *Journal of Statistical Theory and Practice*, **4**(2), 203-220.

Tahir, M.H., Iqbal, I. and Shabbir J. (2012). Polygonal designs with block size 3 and single distance. *Hacettepe Journal of Mathematical Statistics*, **41(4)**, 587-604.

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

Tiwari, N. (2012). Statistical disclosure control using random rounding and quadratic programming. *Metodoloski zvezki* (Advances in Methodology of Statistics.), **9(1)**, 61-69.

Tiwari, N. and Chilwal, A. (2013). On minimum variance optimal controlled sampling: a simplified approach. *Journal of Statistical Theory and Practice*, **8(4)**, 692-706.

Tiwari, N. and Nigam, A.K. (1993). A note on constructive procedure for unbiased controlled rounding. *Statistics and Probability Letters*, **18**, 415-420.

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

Tiwari, N. and Nigam, A.K. (2010). Two dimensional optimal controlled nearest proportional to size sampling design using quadratic programming. *Statistical Methodology*, **7(6)**, 601-613.

Tiwari, N. and Sud, U.C. (2011). Minimum variance optimal controlled nearest proportional to size sampling scheme using multiple objective functions. *Journal of Indian Society of Agricultural Statistics*, **65(3)**, 297-304.

Tiwari, N. and Sud, U.C. (2012). An optimal procedure for sample coordination using multiple objective functions and nearest proportional to ipps size sampling designs. *Communications in Statistics-Theory and Methods*, **41**, 2014-2033.

Tiwari, N., Nigam, A.K and Pant, I. (2007). On an optimal controlled nearest proportional to size sampling scheme. *Survey Methodology*, **33(1)**, 87-94.

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

Wright, J.H. (2008). Two-dimensional balanced sampling plans excluding adjacent units. *Journal of Statistical Planning and Inference*, **138(1)**, 145-153.

Wright, J.H. and Stufken, J. (2011). Variance approximation under balanced sampling plans excluding adjacent units. *Journal of Statistical Theory and Practice*, **5(1)**, 147-160.

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

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society, Series B*, **15(2)**, 253–261.