Statistics and Applications
Volume 11, Nos. 1&2, 2013 (New Series), pp. 127-146

Controlled Sampling: A Review

Raghunath Arnab
Department of Statistics, University of Botswana and Department of Statistics, University of South Africa

---

Abstract

Controlled sampling design was originated by Goodman and Kish (1950). The controlled sampling method reduces the probability of selection of undesirable samples while retaining properties associated with a probability sampling design. In this article we have reviewed some of the important controlled sampling designs in detail.

Key Words: BIBD, BSA(*m*), BSEC, Controlled sampling, Integer programming, IPPS sampling, Sampling design.

---

1 Introduction

In selecting a sample of size n from a finite population of N units by simple random sampling without replacement (SRSWOR) procedure, all the possible $\binom{N}{n}$ samples have equal probability of selection but all the samples are not equally advantageous for surveying purposes. Sampling units within a sample may be so widespread that the cost of data collection may be very expensive due to travel costs and at the same time non-sampling errors involving non-response and investigator's bias increases due to inadequate supervision of field work. Such samples, which are uneconomical and also create organizational and other difficulties, are termed as non-preferred or undesirable samples by Goodman and Kish (1950). The controlled sampling method, originated by Goodman and Kish (1950), reduces the probability of selection of undesirable samples

while retaining properties associated with a probability sampling design. The scope of control sampling may include appropriate distribution of sampling units over different subgroups of the population to obtain reliable estimates from each of the subgroups. Controlled sampling is also used for increasing efficiency of key estimates for a multi-character survey. Controlled selection can be easily achieved by stratification. For example if we choose a sample of 6 students from a group of 24 students comprising of 12 male and 12 female, there is a possibility of selecting all 6 male students or all 6 female students. If we wish to control the selection of male or female students, we may stratify the 24 students into two strata comprising of 12 males and 12 female students, and then select 3 students from each of the strata. We thereby control the selection of male and female students to a fixed number of 3 each. Goodman and Kish (1950) pointed out that the control selection cannot be obtained by applying stratification alone. The use of controlled sampling is not always safe for multi-stage sampling since unbiased variance estimation may not always be possible. Using data available from a survey of Scottish schools, Waterton (1983) showed that controlled sampling provides more efficient estimates than multi proportionate stratified sampling. The sample co-ordination problem is similar to controlled sampling where the overlaps of two or more samples drawn in different occasions are controlled. It is either positive or negative. In positive co-ordination, the expected overlap is maximized while in negative co-ordination it is minimized.

Five different approaches of controlled sampling available in the literature are: (i) experimental design configurations, (ii) linear programming, (iii) nonlinear programming, (iv) nearest proportional to size design and (v) co-ordination of samples over time. Combinatorial properties of experimental designs in controlled sampling designs was used by Chakrabarti (1963), Avadhani and Sukatame (1973), Foody and Hedayat (1977), Gupta et al. (1982, 2012), Nigam et al. (1984) etc. The application of linear and nonlinear programming in controlled sampling was used by Rao and Nigam (1990, 1992), Mandal et al. (2010, 2011) and Tiwari et al. (2007), while Gabler (1987) used nearest proportional to size sampling design for selection of controlled sample. Co-ordination of samples over time was considered by Keyfitz (1951), Fellegi (1963), Lanke (1974) among others. Some of the controlled sampling techniques are described in the sequel.

2 Pioneering method

Consider the example of Goodman and Kish (1950) where a population is stratified into two strata. Strata 1 comprises of 6 units *A, B, C, D, E* and *F* while the strata 2 comprises of 5 units *a, b, c, d* and *e*. The units *B, C, F* of the stratum 1 are identified as coastal units and the rest (*A, D* and *E*) are inland units. Similarly, for stratum 2 units *a, b, c* and *e* are

inland while the unit *d* is a costal unit. The probability of selection assigned to each of the units is given below:

Stratum 1		Stratum 2	
Unit	Unit	Unit	Unit
A	0.10	*a*	0.15
B	0.15	*b*	0.30
C	0.10	*c*	0.10
D	0.20	*d*	0.20
E	0.25	*e*	0.25
F	0.20		
P (inland) = 0 .55 P (coastal) = 0 .45		P (inland) = 0 .80 P (coastal) = 0 .20	

It is desirable to select one inland and one coastal unit. The selection of two coastal units is undesirable. Under stratified random sampling selecting one unit from each of the stratum, the probabilities of selection of different combinations of units are as follows:

P(one inland, one costal) = P (inland from stratum 1) $\times$ P (costal from stratum 2) + P (costal from stratum 1) $\times$ P (inland from stratum 2) = 0.55 $\times$ 0.20 + 0.45 $\times$ 0.80 = 0.47

P(two inland) = P (inland from stratum 1) $\times$ P (inland from stratum 2) = 0.55 $\times$ 0.80 = 0.44

P(two costal) = P (costal from stratum 1) $\times$ P (costal from stratum 2) = 0.45 $\times$ 0.20 = 0.09

Goodman and Kish (1950) in their proposed method rearranged units in the stratum 1 by listing *B, C, F* first, followed by *A, D,* and E. Then they rearranged the units in the stratum 2 by shifting *d* to the end, *i.e.*, by placing the unit *e* above *d*.

Stratum 1		Stratum 2	
Unit	Unit	Unit	Unit
B	0.15	*a*	0.15
C	0.10	*b*	0.30
F	0.20	*c*	0.10
A	0.10	*e*	0.25
D	0.20	*d*	0.20
E	0.25		

In this method the selection of units from both the strata was done by drawing a single random number from 1 to 100. If the selected random number is 45 or less, a coastal unit is selected from stratum 1 and an inland unit from stratum 2. If the selected number is between 46 and 80 an inland unit is selected from both the strata. If the selected number is greater than 80, an inland unit is selected from stratum 1 and a costal unit from stratum 2. In this method the probabilities of different combinations of selection of units are: P (one inland one costal unit) = 0.65 and P (two inland units) = 0.35. In this procedure the original assigned probabilities of all the units are rigorously maintained and the probability of selection of desirable samples (one inland and one costal unit) is made as large as possible within the limitation of probability sampling.

3 Experimental design configurations

In an experimental design setup, one to one correspondence between a sampling design and an block design is established. The *treatment* and *block* of a block design are termed as the *unit* and *sample* of a sampling design respectively. Thus, the total number of treatments v is equal to the total number of units N and the sample size n is equal to the block size k. The total number of blocks of an experimental design b will be treated as the total number of possible samples in a sampling design. In constructing a controlled sampling, properties of various incomplete block designs with minimum number of support (block) sizes are used. Preferred samples are assigned as many blocks as possible by the trial and error method while the remaining blocks are associated with the undesirable samples. One block (sample) is then selected at random with a pre-assigned probability so that the property of a probability sampling is maintained.

3.1 Equal probability sampling design

Here, one block (sample) is selected at random from b blocks so that that first and second order inclusion probabilities for the i^{th} unit and i and $j(i \neq j)$ units become respectively equal to

$$\pi_i = \frac{n}{N} \quad \text{and} \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)} \tag{3.1}$$

Avadhani and Sukhatme (1973) used the properties of the balanced incomplete block design (BIBD) in the construction of a controlled sampling design. In this method a BIBD is constructed with parameters (v, b, r, k, λ) assuming it exists, where $v = N =$ population size, $b =$ total number of blocks (samples), $r =$ replication of a treatment = total number of times a unit appears in b samples which is the same for all units, $k(= n)$ block (sample) size and $\lambda =$ number of times any two treatments (units) appear together in the same block (sample). Let us identify blocks of the BIBD with the preferred samples or maximum possible number of preferred samples and the rest with non-

preferred samples. One block is selected at random from the b blocks. The selected block constitutes the controlled sample. In this controlled sampling, the inclusion probabilities are $\pi_i = r/b$ and $\pi_{ij} = \lambda/b$. Further, from the properties of a BIBD, *viz*., (i) $bk = vr$ and (ii) $\lambda(v-1) = r(k-1)$ (Raghavarao, 1971), we find $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$. Thus the number of supports for a controlled sampling b is much less than the number of supports $\binom{N}{n}$ of a SRSWOR sampling design of size n. The number of supports becomes the minimum when $b = v = N$ *i.e.* if the BIBD is symmetric. Since the number of the preferred samples is identified as much as possible with the block, the probability of selection of preferred samples for controlled sampling is much higher than that of uncontrolled SRSWOR sampling using the same sample size. Where efficiency is concerned, the sample mean of the controlled sampling design is equally precise as the sample mean of SRSWOR sampling because both the designs possess the same first and second order inclusion probabilities.

Example 3.1: Consider the following example of Avadhani and Sukhatme (1973) which comprises of $N = 7$ and $n = 3$. Suppose that the units are located as follows:

$$\begin{array}{ccccc} & 2 & & 1 & \\ 7 & & 5 & & 4 \\ & 6 & & 3 & \end{array}$$

From the point of view of travel and the inconvenience of fieldwork, the following 14 samples are considered as undesirable samples:

$$(1,\ 2,\ 3),(1,\ 2,\ 6),(1,\ 3,\ 6),(1,\ 3,\ 7),(1,\ 4,\ 6),(1,\ 4,\ 7),(1,\ 6,\ 7),$$
$$(2,\ 3,\ 4),(2,\ 3,\ 6),(2,\ 3,\ 7),(2,\ 4,\ 6),(2,\ 4,\ 7),(3,\ 4,\ 7),(4,\ 6,7)$$

Consider the following BIBD with parameters $v = 7 = b, r = k = 3$ and $\lambda = 1$ with layout 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)*; here *denotes non-preferred sample.

For controlled sampling design, take the above 7 blocks as possible samples, each of which has the same selection probability 1/7. In the above 7 possible samples, only the Block 7: (7, 1, 3)*is an undesirable sample and the other 6 are desirable samples. The probability of selection of the undesirable sample under controlled sampling is 1/7 which is much less than probability of selection of an undesirable sample (14/35) from the uncontrolled SRSWOR sampling design of size 3. Since the solution of a BIBD is not unique, one can get different solution if another BIBD is chosen. For example, Rao and

Nigam (1990) have shown the following alternative layout of a BIBD with parameters $v = 7 = b, r = k = 3$ and $\lambda = 1$, whence the probability of a non-preferred samples is 3/7.
Block 1: (1, 3, 4); Block 2: (2, 4, 5); Block 3: (3, 5, 6); Block 4: (4, 6, 7)*; Block 5: (5, 7, 1); Block 6: (6, 1, 2)* and Block 7: (7, 2, 3)*.

Thus the choice of an appropriate BIBD requires trial and error practices.

For large N and n, a BIBD of the required type may not exist or even if does exist, it is difficult to construct. Further, the identification of the undesirable sample at the initial stage may not be possible because of a lack of adequate information. In this situation we may construct controlled sampling using the following method provided by Avadhani and Sukhatme (1973).

Mechanism for controlled sampling

(i) Divide the population of N units at random into k disjoint groups containing $N_1, N_2, ..., N_k$ units with $\sum_{i=1}^{K} N_i = N$.

(ii) Let $n_i = nN_i / N$ be an integer for $i = 1,2,..,k$. Choose an integer n_i' such that $n_i < n_i' < N_i$ and there exists a BIBD with parameters $(n_i', b_i, r_i, n_i, \lambda_i)$ for $i = 1,2,..,k$. Then select a simple random sub-sample of the n_i' units from the N_i units of the i^{th} group and do independently for $i = 1,2,..,k$.

(iii) Determine the preferred combination of n_i from n_i' units and establish a one to one correspondence between the blocks of BIBD's in (ii) and the preferred combinations. Select one block at random from the BIBD of each of the k groups independently. Then the collection of the selected blocks of the BIDS will constitute the controlled sample of size n.

Here we can easily verify the following theorem.

Theorem 3.1: Let $\bar{y}_i$ be the sample mean based on n_i units selected from the ith group and $\bar{y}_w = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i$. Then(i) $E(\bar{y}_w) = \bar{Y}$ = population mean and (ii) $V(\bar{y}_w) = (\frac{1}{n} - \frac{1}{N})S_y^2$, where S_y^2 is the population variance.

The above theorem indicates that the weighted controlled sample mean $\bar{y}_w$ is unbiased for the population mean and at the same time it is as efficient as the sample mean based on an uncontrolled SRSWOR sample of the same size n. Further the controlled selection given above reduces the probability of selection of the non-preferred sample. Wynn (1977) and Foody and Hedayat (1977) used BIBD with repeated blocks for controlled sampling when nontrivial BIBD do not exist.

3.2 Unequal probability sampling design

Let s be a sample of size n selected with probability $p(s)$ using a sampling design p. Let S be the support of p *i.e.* S is the collection of all possible samples such that $p(s) > 0$ with $\sum_{s \in S} p(s) = 1$. Let us further suppose that p be an IPPS or π ps (inclusion probability proportional to the measure of size) sampling design with the inclusion probability of the ith unit as $\pi_i = np_i$ where $p_i (> 0)$ is the normed size measure for the i^{th} unit with $\sum_{i=1}^{N} p_i = 1$.

Controlled IPPS sampling design

Gupta et al. (1982) proposed the following controlled sampling design:

(i) Select a BIBD with parameters $(v = N, b, r, k \geq n, \lambda)$ assuming it exists.

(ii) Select one block s_j from the BIBD mentioned above with probability

$$p(s_j) = \frac{v(r \sum_{i \in s_j} p_i - \lambda)}{b(r - \lambda)} \text{ for } j = 1, .., b \text{ (assuming } p(s_j) \geq 0)$$

If $k = n$, the selected block constitutes the required sample of size n.

(iii) If $k > n$, select a sub-sample of size n units from k units of the selected block s_j by SRSWOR method.

(iv) Associate the blocks of the BIBD with the maximum possible number with preferred samples and the rest with non-preferred samples.

Gupta et al. (1982) has shown the inclusion probability for the i^{th} unit in the above controlled sampling design is $\pi_i = np_i$

Remark 3.1: The controlled sampling scheme proposed by Gupta et al. (1982) is valid if $p(s_j) = \frac{v(r\sum_{i\in s_j} p_i - \lambda)}{b(r-\lambda)} > 0$ i.e. $\sum_{i\in s_j} p_i \geq (k-1)/(N-1)$ [noting $v = N$ and for a BIBD $\lambda(v-1) = r(k-1)$]. For $k = n$, the constraint $\sum_{i\in s_j} p_i \geq (k-1)/(N-1)$ reduces to $\sum_{i\in s_j} p_i \geq (n-1)/(N-1)$ which is less restrictive than the Midzuno-Sen (Sukhatme and Sukhatme, 1976) IPPS sampling scheme $p_i > (n-1)/\{n(N-1)\}$.

Remark 3.2: Instead of using a BIBD to construct the controlled sampling design, we could consider a doubly balanced incomplete block design (Hedayat and Kegeyama, 1980), where every triplet of treatments appears in α blocks. These designs have also been called 3-designs. For such designs, $\alpha(v-2) = \lambda(k-2)$. For sampling schemes based on doubly balanced incomplete block designs the inclusion probability for the i^{th} and j^{th} units $(i \neq j)$ is $\pi_{ij} = n(n-1)\{(p_i + p_j) - 1/(N-1)\}/(N-2)$

This expression of π_{ij} is the same as that for the Midzuno-Sen IPPS sampling scheme. Since for this sampling scheme $\pi_i\pi_j \geq \pi_{ij}$ for all $i \neq j$, the Yates and Grundy (1953) variance estimator is always nonnegative.

Hedayat and Stufken (1989) and Nigam et al. (1984) proposed alternative controlled IPPS sampling scheme which also possesses non-negative Yates-Grundy (1953) variance estimators.

3.3 Balanced Sampling plan without contiguous units

The first step of selection of a sample is to determine the sampling frame where the units of the populations are labelled by the numbers $1, 2, .., N$; N is the total number of units in the population. In general, the units are labelled according to their physical positions. For examples, in Households surveys adjacent enumeration areas receive contiguous numbers e.g. 101 and 102. Similarly, households within the enumeration areas are numbered serially according to their physical positions. In most situations the contiguous units possess similar information especially when ordering is done in time or space. In such situations samples containing contiguous units are treated as undesirable samples. Hedayat et al. (1988) proposed balanced sampling plan excluding contiguous units (BSEC) where each sample contains same number (n) of distinct units and no pair of contiguous units appear together in the same sample whereas all other pairs appear

equally often in the samples. If the samples are selected with equal probability, then the first and second order inclusion probabilities become

$\pi_i = n/N$ for $i = 1,..,N$ and

$$\pi_{ij} = \begin{cases} n(n-1)/[N(N-3)]; & (i,j) \quad \text{contiguous} \\ 0 & \text{if} \quad (i,j) \quad \text{non-contiguous} \end{cases}$$

Hedayat et al. (1988) assumed that the labelling of N units are done in a circular way with i and $(i+1) \bmod N$ as contiguous units i.e. N and 1 are contiguous units. In this case the serial correlation of first order is given by $\rho_1 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{Y})(y_{i+1} - \bar{Y})}{N\sigma_y^2}$

where $N\sigma_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2$.

For this BSEC the variance of the sample mean $\bar{y}_{BSEC}$ is

$$V(\bar{y}_{BSEC}) = \frac{\sigma_y^2}{n}\left(1 - \frac{(1+2\rho_1)(n-1)}{N-3}\right)$$

It can be checked that $V(\bar{y}_{BSEC})$ is smaller than $V(\bar{y}_s) = \dfrac{\sigma_y^2}{n}\dfrac{N-n}{N-1}$, the variance of the sample mean $\bar{y}_S$ based on SRSWOR if $\rho_1 > -1/(N-1)$. The condition $\rho_1 > -1/(N-1)$ is likely to be realized in practice especially when the contiguous units have high positive correlation. Hedayat et al. (1988) studied the existence and constructions of such BSEC. The main demerit of the BSEC is that $V(\bar{y}_{BSEC})$ cannot be estimated unbiasedly since the inclusion probabilities of the two contiguous units are zero. Several extensions of BSEC are proposed. Stufken (1993) proposed a balanced sampling plans excluding adjacent units [BSA(m)] where all pairs of units whose distance are less than equal to $m(\geq 1)$ are excluded. Clearly BSA(1) is equivalent to BSEC. Here also the unbiased estimator of the variance of the sample mean is not available. Stufken et al. (1999) introduced polygonal designs (PD) which is a generalization of BIBD to obtain a [BSA(m)]. Mandal et al. (2008) studied existence and constructions of such designs.

Mandal et al. (2009) proposed distance balanced sampling plan (DBSP) with the first and second order inclusion probabilities are respectively

$\pi_i = n/N$, for $i = 1,..,N$ and $\pi_{ij} = \dfrac{n(n-1)}{N}\dfrac{f_{ij}}{\sum_{j(\neq i)=1}^{N} f_{ij}}$ for $i \neq j = 1,..,N$

where f_{ij} is a suitably defined non-negative distance function between the units i and j. So for the proposed DBSP, the variance of the sample mean can be estimated unbiasedly.

More details are given by Rao and Vijayan (2008), Mandal et al. (2010, 2011), Gupta et al. (2012) among others.

4 Application of linear programming

The methods of controlled sampling focus on the reduction of support size through applications of experimental designs and increase of preferred samples by trial and error methods. The criterion of minimum support size is not even relevant for controlled sampling design (Rao and Nigam, 1992). Rao and Nigam (1990, 1992) constructed the optimum controlled sampling by applying linear programming method. The proposed method not only minimizes the probability of selection of undesirable samples but also maintains conditions desirable to sampling designs such as unbiasedness properties, controlling sampling variance to a certain level and nonnegative variance estimation.

Let S be the collection of all possible samples such that $p(s)>0$ for $s \in S$, $S_1(\subset S)$ is the collection of all undesirable samples and $S_0 = S - S_1$. Our objective is to minimize $\sum_{s \in S_1} p(s)$ subject to (i) IPPS condition: $\pi_i = np_i$, (ii) Non-negative Yates and Grundy's variance estimation: $\pi_i \pi_j \geq \pi_{ij}$ for $i \neq j$ and (iii) controlling the magnitude of variance: $c\pi_i \pi_j \leq \pi_{ij} \leq \pi_i \pi_j$ with $c(<1)$, a pre-specified constant such as $c = 1/2$. Thus we need to find a solution of $p(s)$ from the following linear programming (LP) problem:

Objective function:

$$\text{Minimize } \phi = \sum_{s \in S_1} p(s)$$

Constraints:

$$\text{(i) } p(s) \geq 0 \text{ for } s \in S$$

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

$$\text{(iii) } \sum_{s \supset i} p(s) = np_i \text{ for } i = 1,..,N$$

$$\text{(iv) } \sum_{s \supset i,j} p(s) \leq n^2 p_i p_j ; i \neq j = 1,..,N$$

$$\text{(v) } \sum_{s \supset i,j} p(s) \geq cn^2 p_i p_j ; i \neq j = 1,..,N$$

One can obviously choose a more general objective function $\phi = \sum_{s \in S} c(s)p(s)$ with suitable weight $c(s)$. The objective function ϕ reduces to the expected cost of the survey when $c(s)$ is the cost of selecting the sample s.

The solution of the LP, although not unique, can be obtained numerically by using the simplex method. The computer software package for LP is available. However, the proposed LP method becomes impractical if both N and n become large, since in that case the number of variables as well as the number of constraints increases very rapidly. Rao and Nigam (1992) suggested the use of stratified sampling if the undesirable samples can be identified separately in each of the stratum. Lahiri and Mukherjee (2000) suggested an alternative method which reduces the dimensionality of the problem and hence reduces the computing time to a great extent. Mandal et al. (2010, 2011) proposed linear integer programming approach for controlled sampling designs, particularly the balanced sampling plan and distance balanced sampling plans.

5 Application of Nonlinear Programming

Tiwari et al. (2007) obtained the solution of $\tilde{p}$ by applying non-linear programming method as follows:

Minimize

$$\varphi = \sum_{s \in \tilde{S}} \frac{\{\tilde{p}(s) - p_0(s)\}^2}{p_0(s)}$$

subject to the following constraints:

(i) $\tilde{p}(s) \geq 0 \ \forall s \in \tilde{S}$, (ii) $\sum_{s \in \tilde{S}} \tilde{p}(s) = 1$,

(iii) $\sum_{s \supset i} \tilde{p}(s) = \tilde{\pi}_i \ \forall i = 1,..,N$, (iv) $\sum_{s \supset i,j} \tilde{p}(s) = \tilde{\pi}_{ij} > 0 \ \forall i \neq j = 1,..,N$ and

(v) $\sum_{s \supset i,j} \tilde{p}(s) \leq \tilde{\pi}_i \tilde{\pi}_j \ \forall i \neq j = 1,..,N$.

Tiwari et al. (2007) also constructed a controlled sampling design by using the following alternative objective function suggested by Takeuchi et al. (1983)

$$\varphi^* = \sum_{s \in \tilde{S}} \frac{\{\tilde{p}(s) - p_0(s)\}^2}{\tilde{p}(s) + p_0(s)}$$

Tiwari et al. (2007) reported that both objective functions provide similar results on various numerical problems. They also observed that a feasible solution to the quadratic programming may not always exist. In this condition, one should try to get solutions by relaxing some of the constraints. One of the limitations of the proposed quadratic as well as linear programming methods is that the methods become impractical if $\binom{N}{n}$ is very large since enumeration of all possible samples and formulation of the objective functions with numerous constraints becomes highly tedious. Tiwari et al. (2007) studied different

controlled sampling procedures using numerical data. The empirical findings reveal that the quadratic programming method performs better than the rest.

6 Nearest proportional to size design

Suppose that a sampling design p with a support S is desirable due to theoretical considerations but the set of samples $S_1(\subset S)$ are considered undesirable due to practical considerations. In order to eliminate the undesirable sample, we may consider the following sampling design p_0 which assigns zero probability of selection for each of the non-preferred samples in S_1

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

where $p_0(s)$ and $p(s)$ denote respectively the selection probabilities of the sample s under p_0 and p.

The newly constructed sampling design p_0 may not have the desirable properties as the original sampling design p. So, we need to construct a sampling design $\tilde{p}$ with support $\tilde{S}(= S - S_1)$ consisting of preferred samples while at the same time retaining properties of the desirable design p. We can construct such a design using the method of Gabler (1987). In this method the design $\tilde{p}$ is constructed in such a way so that the distance between $\tilde{p}$ and p_0 becomes as small as possible. Gabler (1987) proposed the following distance measure analogous to the chi-square and used by Cassel and Särndal (1972).

$$D(p_0, \tilde{p}) = \sum_{s\in S-S_1} \frac{\{\tilde{p}(s) - p_0(s)\}^2}{p_0(s)}$$

Suppose it is required that the sampling design $\tilde{p}$ should be a fixed sample of size n design with a given inclusion probability $\tilde{\pi}_i$ for the ith unit for $i = 1,..,N$. Then our problem is to minimize

$D(p_0, \tilde{p})$ subject to the conditions (6.1)

$$\sum_{s\supset i} \tilde{p}(s) = \tilde{\pi}_i \text{ for } i = 1,..,N$$

where $\tilde{\pi}_1,...,\tilde{\pi}_N$ are known nonnegative numbers with $\sum_{i=1}^{N}\tilde{\pi}_i = n$ and N is the population size. If $\tilde{p}$ is an IPPS design, we have $\tilde{\pi}_i = np_i$ with $p_i(>0)$ as a normed size measure for the i^{th} unit.

We are given below the following theorem (without derivation) given by Gabler (1987).

Theorem 6.1: Let $\prod_{\mathbf{0}}$ be the inclusion probability matrices of the designs p_0 and let $\boldsymbol{\lambda} = (\lambda_1,..,\lambda_N)^t$ be solution of the linear equations

$$\prod_{\mathbf{0}} \boldsymbol{\lambda} = \tilde{\boldsymbol{\pi}} \tag{6.2}$$

where

$$\prod_{\mathbf{0}} = \begin{pmatrix} \pi^0_{11} & ... & \pi^0_{1i} & ... & \pi^0_{1N} \\ ... & ... & ... & ... & ... \\ \pi^0_{i1} & ... & \pi^0_{ii} & ... & \pi^0_{iN} \\ ... & ... & ... & ... & ... \\ \pi^0_{N1} & ... & \pi^0_{Ni} & ... & \pi^0_{NN} \end{pmatrix}, \quad \pi^0_{ii} = \pi^0_i = \sum_{s\supset i} p_0(s), \quad \pi^0_{ij} = \sum_{s\supset i,j} p_0(s) \quad \text{for} \quad i \neq j, \quad \tilde{\boldsymbol{\pi}} = \begin{pmatrix} \tilde{\pi}_1 \\ . \\ \tilde{\pi}_i \\ . \\ \tilde{\pi}_N \end{pmatrix} \text{ and}$$

$$\tilde{\pi}_i = \sum_{s\supset i} \tilde{p}(s)$$

If $\tilde{p}(s) = p_0(s)\sum_{i\in s}\lambda_i \geq 0 \ \forall s \in S - S_1$, then $\tilde{p}(s)$ is the solution of the minimization problem (6.1).

Gabler (1987) provided the following method of construction of $\tilde{p}$ from p_0.

Step 1: Select the i^{th} unit with probability $\lambda_i\pi^0_i$.

Step 2: Select a sample s with probability $p_0(s)$ and accept the sample s if $i \in s$. If $i \notin s$, repeat Step 2 until a sample is selected .

The probability of selecting a sample s^* under the above method given that the unit $i(\in s^*)$ at the first draw is

$$\lambda_i\pi^0_i[p_0(s^*) + \{\sum_s (1-I_{si})p_0(s)\}p_0(s^*) + \{\sum_s (1-I_{si})p_0(s)\}^2 p_0(s^*) + \cdots]$$

$$= \frac{\lambda_i\pi^0_i p_0(s^*)}{1-\sum_s (1-I_{si})p_0(s)} = \lambda_i p_0(s^*)$$

Hence the probability of selecting a sample $s*$ under the above method is $\tilde{p}(s*) = (\sum_{i \in s*} \lambda_i) p_0(s*)$.

7 Co-ordination of samples over time

The sample co-ordination problem consists in managing the overlap of two or more samples drawn in different occasions. It is either positive or negative. In positive co-ordination, the expected overlap is maximized while in negative co-ordination it is minimized. This is important because the measure of size associated with the population unit changes over time due to growth, birth, deaths, and mergers. Such changes in the auxiliary information should be incorporated to increase the efficiency of the estimates of the current occasion. Because of the high cost of obtaining information on the auxiliary and study variables, and of familiarizing new respondents with reporting procedures, it is often desirable to retain as many respondents as possible from the original sample (based on the outdated measure of size) for the new sample (based on the updated measure of size). So, the sample co-ordination problem reduces to controlled selection if the selected sample in the first occasion is treated as the desirable (or undesirable) sample in the second occasion. The pioneering work in this field was due to Patterson (1950) and Keyfitz (1951). Other contributors include Fellegi (1963, 1966), Kish (1963), Gray and Platek (1963), Raj (1968), Kish and Scott (1971), Brewer et. al. (1972), Lanke (1974), Arthanari and Doge (1981), and Matei and Tillé (2005).

Consider a finite population $U = (1, \cdots, N)$ of N identifiable units. Let at a certain point of time a sample $s' = \{i_1', \cdots, i_n'\}$ of size n be selected using auxiliary information $\mathbf{x}' = (x_1', \cdots, x_N')$. Later on, $\mathbf{x}'$ changes into $\mathbf{x}'' = (x_1'', \cdots, x_N'')$ and it is required to revise the sample accordingly, but on the other hand, one does not want to exchange units unnecessarily. Let $s'' = \{i_1'', \cdots, i_n''\}$ be a sample of size n selected on the second occasion using $\mathbf{x}'' = (x_1'', \cdots, x_N'')$ as the auxiliary available. Our problem is to maximize $E\{|s' \cap s''|\}$, the expected number of units common between the samples s' and s''. Keyfitz (1951) gave a procedure applicable for the probability proportional to size with replacement (ppswr) method for selection of one unit. His procedure is optimal in the sense that it maximizes the probability of the units drawn on the first occasion to be retained for the second occasion. Keyfitz's procedure can be easily extended to the general sample size n . Keyfitz's method is given below.

7.1 Keyfitz method

Let $p_i' = x_i' / X'$, $p_i'' = x_i'' / X''$, $X' = \sum_{i \in U} x_i'$ and $X'' = \sum_{i \in U} x_i''$. Partition the population U into two groups $U' = \{i \mid p_i' > p_i''\}$ and $U'' = \{i \mid p_i' \le p_i''\}$. Draw one unit i_1' (say) using normed size measure p_{i_1}' and take $s' = \{i_1'\}$. If the selected unit $i_1' \in U''$, then retain i_1' as the selected sample for s'' i.e. $s'' = \{i_1'' = i_1'\}$. But if $i_1' \in U'$, then perform a Bernoulli trial with success probability $p_{i_1'}' / p_{i_1'}''$. If the trial results in a success, retain the unit i_1' for the sample s'' i.e. $i_1'' = i_1'$. On the other hand, if the trial results in failure, one unit is selected from U'' with probability proportional to $p_i'' - p_i'$. The Keyfitz method yields the following theorems:

Theorem 7.1: The Keyfitz method selects samples $s' = \{i_1'\}$ and $s'' = \{i_1''\}$ with probabilities $p_{i_1'}'$ and $p_{i_1''}''$ respectively.

Theorem 7.2: Let $s' = \{i_1'\}$ and $s'' = \{i_1''\}$ be PPSWR samples of size one each. Then the expected number of common units between s' and s' is $\mathrm{E}(|s' \cap s''|) \le \sum_{i=1}^{N} \mathrm{Min}(p_i', p_i'')$.

Theorem 7.3. For the Keyfitz Method $\mathrm{Prob}\{i_1' = i_1''\} = \sum_{i=1}^{N} \mathrm{Min}(p_i', p_i'')$

7.2 PPAS sampling scheme

Lanke (1974) considered the problem of selection two PPAS (probability proportional to aggregate measure of size) s' and s'' each of size n with different measure of size so that the expected number of common units between s' and s'' is maximized. The method is described as follows:

7.2.1 Lanke method

Draw a pair of units $\{i_1', i_1''\}$ by the Keyfitz method using normed size measures $p_1', \cdots, p_N'$ and $p_1'', \cdots, p_N''$ respectively. Then draw an SRSWOR sample s_0' of size $n-1$ from the $U - \{i_1'\}$ and take

$$s' = \{i_1'\} \cup s_0'$$

$$s'' = \begin{cases} s' & \text{if} \quad i_1'' \in s' \\ \{i_1''\} \cup s_0' & \text{if} \quad i_1'' \notin s' \end{cases}$$

The first unit i_1' of s' is selected with probability p_i' and the remaining $n-1$ units are selected from $U-\{i_1'\}$. Hence, s' is selected by the Lahiri-Midzuno-Sen (1951, 1952, 1953) sampling method and the probability of selection of s' is $p(s') = (\sum_{i \in s'} p_i') / M_2$ with $M_2 = \binom{N-1}{n-2}$. To prove that $p(s'') = (\sum_{i \in s''} p_i'') / M_2$, we need to show that $s_0'' = s'' - i_1''$ is an SRSWOR sample from $U-\{i_1''\}$. $i_1' = i_1''$ implies $s_0' = s_0''$. If $i_1' \neq i_1''$, then

$$s_0'' = \begin{cases} s_0' & \text{if} \quad i_1'' \notin s_0' \\ \{i_1''\} \cup s_0' - \{i_1''\} & \text{if} \quad i_1'' \in s_0' \end{cases}$$

Thus s_0'' is an SRSWOR sample from $U-\{i_1'\}$ where i_1'', whenever selected in the sample, it is replaced by i_1'. Hence s_0'' is an SRSWOR sample selected from $U-\{i_1''\}$.
We now state the following results without derivation.

Theorem 7.4: Let s' and s'' be PPAS samples each of size n, with normed size measures $p_1', \cdots, p_N'$ and $p_1'', \cdots, p_N''$ respectively, then the expected number of common units between s' and s'' satisfies

$$E(|s \cap s'|) \leq \frac{N(n-1)}{N-1} + \frac{N-n}{N-1} \sum_{i=1}^{N} \min(p_i', p_i'') \tag{7.1}$$

Theorem 7.5: For Lanke sampling design the upper bound (7.1) is attained
i.e. $E\{|s' \cap s''|\} = \frac{N(n-1)}{N-1} + \frac{N-n}{N-1} \sum_{i \in U} \text{Min}(p_i', p_i'')$

Remark 7.1: Lanke's scheme is not uniquely optimum *i.e.* there exist at least one other method for which $E\{|s' \cap s''|\}$ attains the upper bound of Theorem (7.1). Lanke also generalized this method for drawing $d(>2)$ PPAS samples, each of size n with different sets of measures of size.

Acknowledgements: The author wishes to thank the referee for his constructive suggestions which led to a considerable improvement of the earlier version of this manuscript.

8 Discussion/Future Direction

The main purpose of selection of a sample using an appropriate sampling design is to obtain efficient estimates of parameters of interest. But the selected units within the sample sometimes may be so widespread that the cost of data collection becomes very high due to travel costs and it may be inconvenient for administrative purposes also. Samples, which are uneconomical and also create organizational and other difficulties, are termed as non-preferred or undesirable samples. Controlled sampling procedure has been proposed to overcome such difficulties. Broadly, the methods are classified in five categories: (i) use of experimental designs for selection of sample, (ii) linear programing, (iii) non-linear programming, (iv) nearest proportional to size design and (v) co-ordination of samples over time. However, none of the proposed methods is optimum in all the situations. The selection of samples using combinatorial properties of experimental designs, especially using of balanced sampling reduces drastically selection of non-preferred samples. But, it does not always yield optimum solution. The method of linear and nonlinear programming can produce optimal solution numerically by using suitable objective function and constraints. Both methods have limited applications when the population and sample sizes are both very large. In general, linear and non-linear programming methods cannot be used to study the properties of the estimators theoretically. In nearest proportional to size sampling design one selects samples from a sampling design which is closed to the target sampling design. This method is quite advantageous because it is applicable for large sampling sizes. It can also be used for various varying probability sampling designs. The main demerit of this method is that it may fail to exist for some situation. Sampling coordination may be successfully achieved for some situations but has limited application especially for varying probability sampling designs. For practical purposes, the situation dictates the controlled sampling procedure to be used.

References

Arthanari, T. and Doge, Y. (1981). *Mathematical programming in Statistics*. Wiley, New York.

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

Brewer, K., Early, L. and Joyce, S. (1972). Selecting several samples from a single population. *Austral. J. Statist.*, **14**, 231-239.

Cassel, C.M. and Särandal, C.E. (1972). A model for studying robustness of estimators in informativeness of labels in sampling with varying probabilities. *J. Roy. Statist. Soc*., **B 34**, 279-289.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *J. Ind. Statist. Assoc.*, **1**, 78-85.

Fellegi, I. (1963). Sampling with varying probabilities without replacement, rotation and non-rotating samples. *J. Amer. Statist. Assoc.*, **58**, 183-201.

Fellegi, I. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement. In *Proceedings of the Social Statistics Section, Amer. Statist. Assoc.*, Washington, 434-442.

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

Gabler, S. (1987). The nearest proportional to size sampling design. *Commn. Statist. Theory-Methods*., **5**, 932-945.

Goodman, R. and Kish, L. (1950). Controlled selection- a technique in probability sampling. *J. Amer. Statist. Assoc*., **45**, 350-372.

Gray, G. and Platek, R. (1963). Several methods of re-designing area samples utilizing probabilities proportion to size change significantly. *J. Amer. Statist. Assoc.*, **63**, 1280-1297.

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

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

Hedayat, A. and Kegeyama, S. (1980). The family of *t*-designs - Part I, *J. Statist. Plann. Infer*., **4**, 173-212.

Hedayat, A. and Stufken, J. (1989). The construction of IPPS sampling designs through a method of emptying boxes. *Ann. Statist.*, **17**, 1886-1905.

Hedayat, A.S., Rao, C.R. and Stufken J. (1988). Sampling plan excluding contiguous units. *J. Statist. Plan. Infer.*, **19**, 159-170.

Keyfitz, N. (1951). Sampling with probabilities proportional to size; adjustment for changes in probabilities. *J. Amer. Statist. Assoc*., **46**, 105-109.

Kish, L. (1963). Changing strata and selection probabilities. In *Proceedings of the Social Statistics Section, Amer. Statist. Assoc.*, Washington, 124-131.

Kish, L and Scott, A. (1971). Retaining units after changing strata and probabilities. *J. Amer. Statist. Assoc*., **66**, 461- 470.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimators. *Bull. Int. Statist. Inst.*, **33**, 33-140.

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

Lanke, J.(1974). *Some contribution to the theory of survey sampling*. Department of Mathematical Statistics, Lund.

Mandal, B.N., Prasad, Rajender and Gupta, V.K. (2008). Computer aided construction of balanced sampling plans excluding contiguous units. *J. Statist. Appl*., **3**, 59-85.

Mandal, B.N., Prasad, Rajender, Gupta, V.K. and Sud, U.C. (2009). A family of distance balanced sampling plans. *J. Statist. Plann. Infer.*, **139**, 860-874.

Mandal, B.N., Prasad, Rajender and Gupta, V.K. (2010). Linear programming approach to construct distance balanced sampling plan. *J. Ind. Soc. Agri. Statist*., **64**, 303-312

Mandal, B.N., Prasad, Rajender and Gupta, V.K. (2011). Construction of polygonal designs using linear integer programming. *Commn. Statist-Theory and Methods*, **40**, 1787-1794.

Matei, A. and Tillé, Y. (2005). Maximal and minimal sampling co-ordination, *Sankhyā*, **67**, 590-612.

Midzuno, H. (1952). On sampling system with probabilities proportionate to sum of sizes. *Ann. Inst. Statis. Math*., **3**, 99-107.

Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *J. Roy. Statist. Soc*., **B 46**, 546-571.

Patterson, H. (1950). Sampling on successive occasions with partial replacement of units, *J. Roy. Statist. Soc*., **B 12**, 241-256.

Raghavrao, D. (1971). *Constructions and Combinatorial problems in Design of Experiments*. New York: Wiley.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill, New York.

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

Rao, J.N.K. and Vijayan, K. (2008). Application of experimental designs in survey sampling. *J. Ind. Soc. Agri. Statist*., **62**, 126-131.

Rao, J.N.K. and Nigam, A.K. (1992). Optimal controlled sampling: A unified approach. *Int. Statist. Rev*., **60**, 89-98.

Sen, A.R. (1953). On the estimates of variance in sampling with varying probabilities. *J. Ind. Soc. Agri. Statist.*, **5**, 119-127.

Stufken, J. (1993). Combinatorial and statistical aspects of sampling plans to avoid the selection of adjustment units. *J. Comb. Info. Syst. Sci*., **18**, 81-92.

Stufken J., Song, S.Y., See, K. and Driessel, K.R. (1999). Polygonal design: Some existence and non-existence results. *J. Statist. Plann. Inf.*, **77**, 155-166.

Sukhatme, P.V. and Sukhatme, P.V. (1976). *Sampling theory and surveys with applications*. New Delhi: Indian Society of Agricultural Statistics, New Delhi.

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

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

Waterton, J.J. (1983). A exercise in controlled selections. *Appl. Statist.*, **32**, 150-164.

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

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc.*, **B 15**, 253-261.