CHAPTER 24

Controlled Sampling

24.1 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 nonsampling errors involving nonresponse and investigator's bias increase because of inadequate supervision of fieldwork. Such samples, which are uneconomical and also create organizational and other difficulties, are termed nonpreferred 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 multicharacter 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 male 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 three 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 multistage sampling because 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

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00024-8

estimates than multiproportionate stratified sampling. The sample coordination 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 coordination, the expected overlap is maximized whereas in negative coordination, it is minimized.

Five different approaches of controlled sampling available in the literature include the following: (i) experimental design configurations, (ii) linear programming (LP), (iii) nearest proportional to size design, (iv) nonlinear programming, and (v) coordination of samples overtime. Combinatorial properties of experimental designs in controlled sampling designs were used by Chakrabarti (1963), Avadhani and Sukhatme (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 (1987a,b) used nearest proportional to size sampling design for selection of controlled sample. Coordination of samples overtime was considered by Keyfitz (1951), Fellegi (1963), Lanke (1974a,b), among others. A detailed review has been given by Arnab (2013). Some of the controlled sampling techniques have been described below.

24.2 PIONEERING METHOD

Consider the example of Goodman and Kish (1950) where a population is stratified into two strata. Strata 1 comprises 6 units *A, B, C, D, E*, and *F* whereas strata 2 comprises 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 coastal unit. The probability of selection assigned to each of the units is given below:

Stratum 1		**Stratum 2**	
Unit	**Probability**	**Unit**	**Probability**
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 (inland) = 0.80	
P (coastal) = 0.45		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:

$$\begin{aligned}\text{P (one inland, one coastal)} &= \text{P (inland from stratum 1)}\\ &\quad \times \text{P (coastal from stratum 2)}\\ &\quad + \text{P (coastal from stratum 1)}\\ &\quad \times \text{P (inland from stratum 2)}\\ &= 0.55 \times 0.20 + 0.45 \times 0.80 = 0.47\end{aligned}$$

$$\begin{aligned}\text{P (two inland)} &= \text{P (inland from stratum 1)} \times \text{P (inland from stratum 2)}\\ &= 0.55 \times 0.80 = 0.44\\ \text{P (two coastal)} &= \text{P (coastal from stratum 1)} \times \text{P (coastal from stratum 2)}\\ &= 0.45 \times 0.20 = 0.09\end{aligned}$$

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	**Probability**	**Unit**	**Probability**
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 coastal unit from stratum 2. In this method the probabilities of different combinations of selection of units are:

P (one inland one coastal 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 coastal unit) is made as large as possible within the limitation of probability sampling.

24.3 EXPERIMENTAL DESIGN CONFIGURATIONS

In an experimental design setup, one to one correspondence between a sampling design and a block design are 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 trial and error method while the remaining blocks are associated with the undesirable samples. One block (sample) is then selected at random with a preassigned probability so that the property of a probability sampling is maintained.

24.3.1 Equal Probability Sampling Design

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

$$\pi_i = \frac{n}{N} \text{ and } \pi_{ij} = \frac{n(n-1)}{N(N-1)} \tag{24.3.1}$$

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 nonpreferred 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 $\pi_{ij} = \lambda/b$. Furthermore, from the properties of a BIBD, *viz.*, (i) $bk = vr$ and (ii) $\lambda(v-1) = r(k-1)$ (Raghavrao, 1971), we find $\pi_i = \dfrac{n}{N}$ and $\pi_{ij} = \dfrac{n(n-1)}{N(N-1)}$. Thus the number of supports for a controlled sampling b is much less than the number of supports $\binom{N}{n}$ of an 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 24.3.1

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

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

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 nonpreferred sample.

For controlled sampling design, take the above seven blocks as possible samples, each of which has the same selection probability 1/7. In the above seven possible samples, only the Block 7: (7, 1, 3)* is an undesirable sample and the other six 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 solutions 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 nonpreferred 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 it does exist, it is difficult to construct. Furthermore, 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 $i = 1, 2, \ldots, k$. 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 $i = 1, 2, \ldots, k$. Then select a simple random subsample of the n'_i units from the N_i units of the ith group and do independently for $i = 1, 2, \ldots, k$.

(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 BIBD's will constitute the controlled sample of size n.

Here we can easily verify the following theorem.

Theorem 24.3.1

Let $\overline{y}_i$ be the sample mean based on n_i units selected from the ith group and $\overline{y}_w = \dfrac{1}{N}\sum_{i=1}^{k} N_i\overline{y}_i$. Then,

(i) $E(\overline{y}_w) = \overline{Y} =$ population mean and

(ii) $V(\overline{y}_w) = \left(\dfrac{1}{n} - \dfrac{1}{N}\right)S_y^2$, where S_y^2 is the population variance.

The theorem above indicates that the weighted controlled sample mean $\overline{y}_w$ 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. Furthermore, the controlled selection given above reduces the probability of selection of the nonpreferred sample. Wynn (1977) and Foody and Hedayat (1977) used BIBD with repeated blocks for controlled sampling when nontrivial BIBD do not exist.

24.3.2 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 ith 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 ($\nu = N$, b, r, $k \geq n$, λ) assuming it exists.

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

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

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

(i) If $k > n$, select a subsample of size n units from k units of the selected block s_j by SRSWOR method.

(ii) Associate the blocks of the BIBD with the maximum possible number with preferred samples and the rest with nonpreferred samples.

Inclusion probability for the ith unit in the above controlled sampling design is $\pi_i = np_i$. Hedayat and Stufken (1989) and Nigam et al. (1984) proposed alternative controlled IPPS sampling scheme, which possesses nonnegative Yates-Grundy (1953) variance estimators. Further details have been given in Section 5.4.2.5.

24.3.3 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 labeled by the numbers 1, 2,…, N; N is the total number of units in the population. In general, the units are labeled according to their physical positions. For examples, in household 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 (BSEC) units, 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. A sampling design with support BSEC constitutes the desired controlled sampling design.

Example 24.3.2

Hedayat et al. (1988) provided the following example of a BSEC sampling plan with population size $N = 9$ and sample size $n = 3$ where i and $(i + 1)$ mod 9 as contiguous units, i.e., units 9 and 1 are treated as contiguous units.

$$\{1,3,6\},\{1,4,8\},\{1,5,7\},\{2,4,7\},\{2,5,9\},\{2,6,8\},\{3,5,8\},$$
$$\{3,7,9\},\{4,6,9\}$$

If the samples are selected with equal probability, then the first- and second-order inclusion probabilities become

$$\pi_i = n/N = 1/3 \text{ and}$$
$$\pi_{ij} = \begin{cases} 1/9 & \text{if } i \text{ and } j \text{ are noncontiguous} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i \neq j.$$

Consider a population $U = \{1,\ldots, N\}$ of N units from which a sample s of size n is selected using a sampling design, which assign equal probability to each of the samples of a BSEC sampling plan where units i and $(i + 1)$ mod N are treated as contiguous units. Let $\overline{y}_{\text{BSEC}}$ be the sample mean of the variable under study y based on the selected sample s. Then we have the following theorem from Hedayat et al. (1988).

Theorem 24.3.2

(i) Inclusion probabilities for the *i*th, and *i*th and *j*th $(j \neq i)$ units are

$$\pi_i = n/N \text{ and } \pi_{ij} = \begin{cases} \dfrac{n(n-1)}{N(N-3)} & \text{if } (i,j) \text{ noncontiguous} \\ 0 & \text{if } (i,j) \text{ contiguous} \end{cases}$$

(ii) $\overline{y}_{\text{BSEC}}$ is an unbiased estimator for the population mean $\overline{Y}$

(iii) $V\left(\overline{y}_{\text{BSEC}}\right) = \dfrac{\sigma_y^2}{n}\left(1 - \dfrac{(1+2\rho_1)(n-1)}{N-3}\right)$

where $\sigma_y^2 = \dfrac{1}{N}\sum_{i \in U}\left(y_i - \overline{Y}\right)^2$ and $\rho_1 = \dfrac{\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)\left(y_{i+1} - \overline{Y}\right)}{N\sigma_y^2}$ is the serial correlation of first order.

(iv) $V\left(\overline{y}_{\text{BSEC}}\right) \leq V\left(\overline{y}_{sr}\right)$ if $\rho_1 \geq -1/(N-1)$

where $\overline{y}_{sr}$ sample mean of SRSWOR sample of size *n*.

Proof

(i) From construction of the sampling design we have $\pi_i = \alpha$ (constant), $\pi_{ij} = \beta$ (constant) when the units (i, j) noncontiguous and $\pi_{ij} = 0$ when the units are contiguous. The set of contiguous units is $C = \{(1, 2), (2, 3), \ldots, (N-1, N), (N, 1)\}$ and the cardinality of C is N. Now using the consistency conditions of inclusion probabilities $\sum_{i=1}^{N}\pi_i = n$ and $\sum_{i\neq}^{N}\sum_{j=1}^{N}\pi_{ij} = n(n-1)$ given in Section 1.3.3 yield $N\lambda = n$ and $\{N(N-1) - 2N\}\beta = n(n-1)$, i.e., $\pi_i = \alpha = \dfrac{n}{N}$ and $\beta = \pi_{ij} = \dfrac{n(n-1)}{N(N-3)}$ for (i, j) noncontiguous.

$$\text{(ii)}\quad E\left(\overline{y}_{\text{BSEC}}\right) = \frac{1}{n}E\left(\sum_{i \in s} y_i\right)$$

$$= \frac{1}{n}\sum_{i \in U} y_i E(I_{si}) \text{ where } I_{si} = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if } i \notin s \end{cases}$$

$$= \frac{1}{n}\sum_{i \in U} y_i \pi_i$$

$$= \overline{Y}$$

(iii) $$V\left(\overline{y}_{\mathrm{BSEC}}\right)=\frac{1}{n^2}E\left(\sum_{i\in s}\left(y_i-\overline{Y}\right)\right)^2$$

$$=\frac{1}{n^2}E\left(\sum_{i\in U}\left(y_i-\overline{Y}\right)I_{si}\right)^2$$

$$=\frac{1}{n^2}\left[\sum_{i\in U}\left(y_i-\overline{Y}\right)^2\pi_i+\sum_{i\neq}\sum_{j\in U}\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\pi_{ij}\right]$$

$$=\frac{1}{n^2}\left[\frac{n}{N}\sum_{i\in U}\left(y_i-\overline{Y}\right)^2+\frac{n(n-1)}{N(N-3)}\sum_{i\neq}\sum_{j\in U|i,j\text{ noncontiguous}}\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\right]$$

$$=\frac{1}{n^2}\left[\frac{n}{N}\sum_{i\in U}\left(y_i-\overline{Y}\right)^2+\frac{n(n-1)}{N(N-3)}\sum_{i\neq}\sum_{j\in U}\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\right.$$

$$\left.-\frac{n(n-1)}{N(N-3)}\sum_{i\neq}\sum_{j\in U|i,j\text{ contiguous}}\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\right]$$

$$=\frac{1}{n^2}\left[\frac{n}{N}\left(1-\frac{n-1}{N-3}\right)\sum_{i\in U}\left(y_i-\overline{Y}\right)^2-\frac{2n(n-1)}{N(N-3)}\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)\left(y_{i+1}-\overline{Y}\right)\right]$$

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

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

$$=-\frac{2(n-1)}{(N-1)(N-3)n}\{1+(N-1)\rho_1\}\sigma_y^2$$

$$\leq 0\text{ if }\rho_1\geq -1/(N-1)$$

From the theorem above, it follows that sample mean based on BSEC is more precise than the sample mean based on SRSWOR sample of the same size if the serial correlation $\rho_1 > -1/(N-1)$. 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 $V\left(\overline{y}_{BSEC}\right)$ cannot be estimated unbiasedly because the inclusion probabilities of the two contiguous units are zero. Several extensions of BSEC have been proposed. Stufken (1993) proposed a balanced sampling plans excluding adjacent units [BSA(m)] where all pairs of units whose distance are less than or 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, 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) whose first- and second-order inclusion probabilities are

$$\pi_i = n/N, \text{ for } i = 1, \dots, N \text{ and } \pi_{ij} = \frac{n(n-1)}{N} \frac{f_{ij}}{\sum\limits_{j(\neq i)=1}^{N} f_{ij}} \text{ for}$$

$$i \neq j = 1, \dots, N$$

where f_{ij} is a suitably defined nonnegative 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 have been given by Rao and Vijayan (2008), Mandal et al. (2010, 2011), Gupta et al. (2012), among others.

24.4 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 LP method. The proposed method not only minimizes the probability of selection of undesirable samples but also maintains conditions such as unbiasedness properties, controlling sampling variance to a certain level, and nonnegative variance estimation, which are desirable to sampling designs.

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. Our objective is to minimize $\sum\limits_{s \in S_1} p(s)$ subject to (i) IPPS condition: $\pi_i = np_i$, (ii) Nonnegative 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 prespecified constant such as $c = 1/2$. Thus we need to find a solution of $p(s)$ from the following LP problem:

Objective function:

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

Constraints:

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

(ii) $\sum\limits_{s \in S} p(s) = 1$

(iii) $\sum\limits_{s \supset i} p(s) = np_i$ for $i = 1, \ldots, N$

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

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

One can obviously choose a more general objective function $\phi = \sum\limits_{s \in S} c(s)p(s)$ 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, because 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 that 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 DBSPs.

24.5 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 $S_1(\subset S)$ is considered undesirable because of practical considerations. To eliminate the undesirable samples, we may consider the following sampling design p_0, which assigns zero probability of selection for each of the nonpreferred samples in S_1

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

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 $\widetilde{p}$ with support $\widetilde{S}(= S - S_1)$ 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 (1987a,b). In this method the design $\widetilde{p}$ is constructed in such a way that the distance between $\widetilde{p}$ and p_0 becomes as small as possible. Gabler (1987a,b) proposed the following distance measure analogous to the chi-square and used by Cassel and Särandal (1972).

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

Details on construction of such a design $\widetilde{p}$ have been given in Section 5.4.2.6.

24.6 APPLICATION OF NONLINEAR PROGRAMMING

Tiwari et al. (2007) obtained the solution of $\widetilde{p}$ by applying nonlinear programming method as follows:

$$\text{Minimize } \varphi = \sum_{s\in \widetilde{S}} \frac{\{\widetilde{p}(s) - p_0(s)\}^2}{p_0(s)}; \quad \widetilde{S} = S - S_1$$

subject to the following constraints:

(i) $\widetilde{p}(s) \geq 0 \quad \forall s \in \widetilde{S}$,

(ii) $\sum\limits_{s\in \widetilde{S}} \widetilde{p}(s) = 1$,

(iii) $\sum\limits_{s\supset i} \widetilde{p}(s) = \widetilde{\pi}_i \quad \forall i = 1, \ldots, N$,

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

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

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

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

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 LP 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 become 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 others.

24.7 COORDINATION OF SAMPLES OVERTIME

The sample coordination problem consists of managing the overlap of two or more samples drawn in different occasions. It is either positive or negative. In positive coordination, the expected overlap is maximized whereas in negative coordination it is minimized. This is important because the measure of size associated with the population unit changes overtime because of 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 coordination 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 evolved from 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 (1974a,b), Arthanari and Doge (1981), and Matei and Tillé (2005).

Consider a finite population $U = (1,\dots, N)$ of N identifiable units. Let, at a certain point of time, a sample $s(1) = \{i_1(1),\dots, i_n(1)\}$ of size n be selected using auxiliary information $\mathbf{x}(1) = (x_1(1),\dots, x_N(1))$. Later on, $\mathbf{x}(1)$ changes into $\mathbf{x}(2) = (x_1(2),\dots, x_N(2))$ and it is required to revise the sample accordingly, but on the other hand, one does not want to exchange units unnecessarily. Let $s(2) = \{i_1(2),\dots, i_n(2)\}$ be a sample of size n selected on the second occasion using $\mathbf{x}(2) = (x_1(2),\dots, x_N(2))$ as the auxiliary available. Our problem is to maximize $E\{|s(1) \cap s(2)|\}$, the expected number of units common between the samples $s(1)$ and $s(2)$. 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 method is given below.

24.7.1 Keyfitz Method

Let $p_i(1) = x_i(1)/X(1)$, $p_i(2) = x_i(2)/X(2)$, $X(1) = \sum_{i \in U} x_i(1)$, and $X(2) = \sum_{i \in U} x_i(2)$. Partition the population U into two groups $U(1) = \{i | p_i(1) > p_i(2)\}$ and $U(2) = \{i | p_i(1) \le p_i(2)\}$. Draw one unit i (say) using normed size measure $p_i(1)$ and take $s(1) = \{i_1(1) = i\}$. If the selected unit $i \in U(2)$, then retain i as the selected sample for $s(2)$, i.e., $s(2) = \{i_1(2) = i\}$. But if $i \in U(1)$, then perform a Bernoulli trial with success probability $p_i(2)/p_i(1)$. If the trial results in a success, retain the unit i for the sample $s(2)$, i.e., $\{i_1(2) = i\}$. On the other hand, if the trial results in failure, retain the jth unit from $U(2)$ with probability proportional to $p_j(2) - p_j(1)$.

Theorem 24.7.1

Keyfitz method selects samples $s(1) = \{i_1(1) = i\}$ and $s(2) = \{i_1(2) = i\}$ with probabilities $p_i(1)$ and $p_i(2)$, respectively.

Proof

The theorem is true for selection of $s(1)$ because in this method the unit is selected by PPSWR method using normed size measure $p_i(1)$ for the ith unit. For selection of sample $s(2)$, there are two scenarios:

If the unit $i \in U(1)$, then probability of selecting the unit i is

$$\text{Prob}\{i_1(2)=i\} = p_i(1) \times \frac{p_i(2)}{p_i(1)} = p_i(2) \tag{24.7.1}$$

If the unit $i \in U(2)$, then probability of selecting the unit i is

$$\begin{aligned}\text{Prob}\{i_1(2)=i\} &= \text{Prob}\{i_1(1)=i\} + \sum_{j\in U(1)} \text{Prob}\{i_1(1)=j, i_1(2)=i\} \\ &= p_i(1) + \sum_{j\in U(1)} p_j(1)\left(1-\frac{p_j(2)}{p_j(1)}\right)\frac{p_i(2)-p_i(1)}{\Delta(2)} \\ &\left(\text{where } \Delta(2) = \sum_{i\in U(2)}\{p_i(2)-p_i(1)\}\right)\end{aligned}$$

Now noting, $\Delta = \sum_{i\in U}\{p_i(2)-p_i(1)\} = \sum_{i\in U(1)}\{p_i(2)-p_i(1)\} + \sum_{i\in U(2)}\{p_i(2)-p_i(1)\} = 0$, we find

$$\text{Prob}\{i_1(2)=i\} = p_i(1) + \{p_i(2)-p_i(1)\} = p_i(2)$$

Theorem 24.7.2

Let $s(1)=\{i_1(1)\}$ and $s(2)=\{i_1(2)\}$ be PPSWR samples of size 1 each. Then the expected number of common units between $s(1)$ and $s(2)$ is

$$E(|s(1)\cap s(2)|) \le \sum_{i=1}^{N} \min\{p_i(1), p_i(2)\}$$

Proof

$$\begin{aligned}\text{E}(|s(1)\cap s(2)|) &= \sum_{i=1}^{N} \text{Prob}\{i_1(1)=i, i_1(2)=i\} \\ &\le \sum_{i=1}^{N} \min[\text{Prob}\{i_1(1)=i\}, \text{Prob}\{i_1(2)=i\}] \\ &= \sum_{i=1}^{N} \min\{p_i(1), p_i(2)\}\end{aligned}$$

Theorem 24.7.3

For the Keyfitz method

$$\text{Prob}\{i_1(1)=i_1(2)\}=\sum_{i=1}^{N}\min\{p_i(1),p_i(2)\}$$

Proof

$$\begin{aligned}\text{If } i\in U(2),\ \text{then Prob}\{i_1(1)=i_1(2)=i\}&=\text{Prob}\{i_1(1)=i\}\\&=p_i(1)(\le p_i(2))=\min(p_i(1),p_i(2))\end{aligned}\qquad(24.7.2)$$

$$\begin{aligned}\text{If } i\in U(1),\ \text{then Prob}\{i_1(1)=i_1(2)=i\}&=p_i(1)\times\frac{p_i(2)}{p_i(1)}=p_i(2)(\le p_i(1))\\&=\min(p_i(1),p_i(2))\end{aligned}\qquad(24.7.3)$$

The theorem follows from Eqs. (24.7.2) and (24.7.3).

24.7.2 Probability Proportional to Aggregate Size Sampling Scheme

Lanke (1974a,b) considered the problem of selection of two samples $s(1)$ and $s(2)$, each of size n so that probability of selection of $s(1)$ and $s(2)$ are proportional to the aggregate measure of size (PPAS) $x(1)$ and $x(2)$, respectively, at the same time the expected number of common units between $s(1)$ and $s(2)$ is maximized. The method is described as follows.

24.7.2.1 Lanke Method

Draw a pair of units $\{i_1(1), i_1(2)\}$ by the Keyfitz method using normed size measures $p_1(1),\ldots, p_N(1)$ and $p_1(2),\ldots, p_N(2)$, respectively. Then draw an SRSWOR sample $s_0(1)$ of size $n-1$ from the $U-\{i_1(1)\}$ and take

$$s(1)=\{i_1(1)\}\cup s_0(1)$$

$$s(2)=\begin{cases}s(1) & \text{if}\quad i_1(2)\in s(1)\\ \{i_1(2)\}\cup s_0(1) & \text{if}\quad i_1(2)\notin s(1)\end{cases}$$

The first unit $i_1(1)$ of $s(1)$ is selected with probability $p_i(1)$ and the remaining $n-1$ units are selected from $U-\{i_1(1)\}$. Hence, $s(1)$ is selected by the Lahiri–Mizuno–Sen (1951, 1952, 1953) sampling method and the probability of selection of $s(1)$ is $p\{s(1)\}=\left(\sum_{i\in s(1)}p_i(1)\right)\Big/M_1$ where

$M_1 = \binom{N-1}{n-1}$. To prove that $p\{s(2)\} = \left(\sum_{i\in s(2)} p_i(2)\right)/M_1$, we need to show that $s_0(2) = s(2) - i_1(2)$ is an SRSWOR sample from $U - \{i_1(2)\}$. If $i_1(1) = i_1(2)$, then $s_0(1) = s_0(2)$. If $i_1(1) \neq i_1(2)$, then

$$s_0(2) = \begin{cases} s_0(1) & \text{if} \quad i_1(2) \notin s_0(1) \\ \{i_1(1)\} \cup s_0(1) - \{i_1(2)\} & \text{if} \quad i_1(2) \in s_0(1) \end{cases}$$

Thus $s_0(2)$ is an SRSWOR sample from $U - \{i_1(1)\}$ where $i_1(2)$, whenever selected in the sample, is replaced by $i_1(1)$. Hence $s_0(2)$ is an SRSWOR sample selected from $U - \{i_1(2)\}$.

Theorem 24.7.4

Let $s(1)$ and $s(2)$ be PPAS samples each of size n, with normed size measures $p_1(1),\ldots, p_N(1)$ and $p_1(2),\ldots, p_N(2)$, respectively, then the expected number of common units between $s(1)$ and $s(2)$ satisfies

$$E\{|s(1) \cap s(2)|\} \leq \frac{N(n-1)}{N-1} + \frac{N-n}{N-1}\sum_{i=1}^{N} \min\{p_i(1), p_i(2)\}$$

Proof

Let $\gamma_i = \text{Prob}\{s(1) \cap s(2) \supset i\}$. Then

$$\begin{aligned} E\{s(1) \cap s(2)\} &= \sum_{i\in U} \gamma_i \\ &\leq \sum_{i\in U} \min[\text{Prob}\{s(1) \supset i\}, \text{Prob}\{s(2) \supset i\}] \qquad (24.7.4) \\ &= \sum_{i\in U} \min[\pi_i(1), \pi_i(2)] \end{aligned}$$

where $\pi_i(1)$ = inclusion probability of the ith unit for selection of sample $s(1)$ according to PPAS sampling design with normed size measure $p_1(1),\ldots, p_N(1)$.

$$= \frac{n-1}{N-1} + \frac{N-n}{N-1} p_i(1) \qquad (24.7.5)$$

Similarly $\pi_i(2)$ = inclusion probability of the ith unit for selection of sample $s(2)$ according to PPAS sampling design with normed size measure $p_1(2),\ldots, p_N(2)$

$$= \frac{n-1}{N-1} + \frac{N-n}{N-1} p_i(2) \qquad (24.7.6)$$

Substituting Eqs. (24.7.5) and (24.7.6) in Eq. (24.7.4), we get

$$E\{|s(1)\cap s(2)|\} \leq \frac{N(n-1)}{N-1} + \frac{N-n}{N-1}\sum_{i=1}^{N}\min\{p_i(1), p_i(2)\} \qquad (24.7.7)$$

Theorem 24.7.7

For Lanke sampling design the upper bound (24.7.7) is attained

$$\text{i.e., } E\{|s(1)\cap s(2)|\} = \frac{N(n-1)}{N-1} + \frac{N-n}{N-1}\sum_{i=1}^{N}\min\{p_i(1), p_i(2)\}$$

Proof

$$\begin{aligned}\mathrm{Prob}\{s(1)=s(2)\} &= \mathrm{Prob}\{i_1(1)=i_1(2)\} + \mathrm{Prob}\{i_1(1) \neq i_1(2)\}\\ &\quad\times \mathrm{Prob}\{s(1)=s(2)|i_1(1) \neq i_1(2)\}\\ &= \left(\sum_{i\in U}\min\{p_i(1), p_i(2)\}\right)\\ &\quad + \left(1-\sum_{i\in U}\min\{p_i(1), p_i(2)\}\right)\frac{n-1}{N-1}\\ &= \frac{n-1}{N-1} + \frac{N-n}{N-1}\sum_{i\in U}\min\{p_i(1), p_i(2)\}\end{aligned}$$

Since $|s(1) \cap s(2)|$ can take two values $n-1$ and n, we have

$$\begin{aligned}E\{|s(1)\cap s(2)|\} &= (n-1)\mathrm{Prob}\{s(1) \neq s(2)\} + n\mathrm{Prob}\{s(1)=s(2)\}\\ &= (n-1) + \mathrm{Prob}\{s(1)=s(2)\}\\ &= \frac{N(n-1)}{N-1} + \frac{N-n}{N-1}\sum_{i\in U}\min\{p_i(1), p_i(2)\}\end{aligned}$$

Remark 24.7.1

Lanke's scheme is not uniquely optimum, i.e., there exists at least one other method for which $E\{|s(1) \cap s(2)|\}$ attains the upper bound (24.7.7). Lanke also generalized this method for drawing $d(>2)$ PPAS samples, each of size n with different sets of measures of size.

24.8 DISCUSSIONS

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 because of 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 nonpreferred or undesirable samples. Controlled sampling procedure has been proposed to overcome such difficulties. Broadly, the methods are classified into five categories: (i) use of experimental designs for selection of sample, (ii) LP, (iii) nonlinear programming, (iv) nearest proportional to size design, and (v) coordination of samples overtime. However, none of the proposed methods is optimum in all the situations. The selection of samples using combinatorial properties of experimental designs, especially using balanced sampling, reduces the selection of nonpreferred samples drastically. 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 nonlinear programming methods cannot be used to study the properties of the estimators theoretically. In nearest proportional to size sampling design, one selects sample from a sampling design that is closed to the target sampling design. This method is quite advantageous because it is applicable for large sample 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 in some situations. 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.

24.9 EXERCISES

24.9.1 Agricultural farms are stratified into two strata. Stratum 1 comprises seven farms *A, B, C, D, E, F,* and *G* while stratum 2 comprises five farms *a, b, c, d,* and *e*. The farms *A, B, C, D* and *c, d, e* have irrigation facilities and the remaining farms have no such facilities. The normed size measures for the farms have been given in the following table. Select one farm from each of the stratum with the given normed size measure at the same time probability of selection of one irrigated and another nonirrigated farm is maximized.

Stratum 1		Stratum 2	
Units	Normed size measures	Units	Normed size measures
A	0.10	*a*	0.20
B	0.20	*b*	0.25
C	0.15	*c*	0.25
D	0.15	*d*	0.20
E	0.20	*e*	0.10
F	0.10		
G	0.10		

24.9.2 Consider the Example 24.3.1, where we need to select a sample of $n = 3$ units from a population of $N = 7$ units. Use appropriate LP problem to select 3 units so that inclusion probabilities of each of the 7 units becomes exactly 1/7 and at the same time the probability selection of the following sets of nonpreferred samples becomes a minimum.

$$\text{Nonpreferred units}: (1,2,3),(1,2,6),(1,3,6),(1,3,7),(1,4,6),(1,4,7),$$
$$(1,6,7),(2,3,4),(2,3,6),(2,3,7),(2,4,6),(2,4,7),$$
$$(3,4,7),(4,6,7)$$

24.9.3 The first- and second-order inclusion probability matrix of a sample of 3 units from a population of 6 units has been given in the following table. Using an LP method, select a sample of 3 units realizing the following inclusion probabilities (First-order inclusion probabilities have been given in the diagonal).

	Units					
Units	1	2	3	4	5	6
1	0.55	0.25	0.2	0.25	0.2	0.2
2	0.25	0.5	0.2	0.1	0.25	0.2
3	0.2	0.2	0.475	0.25	0.1	0.2
4	0.25	0.1	0.25	0.475	0.1	0.25
5	0.2	0.25	0.1	0.1	0.45	0.25
6	0.2	0.2	0.2	0.25	0.25	0.55

24.9.4 Expenditures on food and school fees of households are approximately proportional to the household size (hh) and number of school going children (z). It is decided to select two samples S_1 and S_2 each of size 3 from the same 10 households with replacement so that probability of selection of a unit in sample S_1 is proportional to the household size while probability of selection of a unit in S_2 is proportional to number of children. Use Keyfitz method, select such samples so that the number of common units in S_1 and S_2 is maximized. Find the expected number of common units for such sampling design.

House hold (hh)	1	2	3	4	5	6	7	8	9	10
hh size	4	3	6	5	5	4	7	6	6	4
Number of children	1	1	2	1	2	2	3	2	2	1

24.9.5 Consider the data of Example 24.9.4. Select two samples each of size 4 with probability proportional to the aggregate measure of sizes of households and number of children, respectively, and at the same time maximize the expected number of common units between them.

24.9.6 The following table gives the cost of travel between six villages and their measure of size.

Villages	**1**	**2**	**3**	**4**	**5**	**6**
1		112	115	124	130	120
2			120	130	115	118
3				135	130	20
4					140	250
5						240
Measure of size	10	15	20	30	25	15

Select a sample of three villages with inclusion probability proportional to their measure of size, and at the same time cost of travel for collecting data is minimized. Find also the minimum cost.