Research Article

A Simplified Approach for Two-Dimensional Optimal Controlled Sampling Designs

Neeraj Tiwari and Akhil Chilwal

Department of Statistics, Kumaun University, S.S.J. Campus, Almora, Uttarakhand 263601, India

Correspondence should be addressed to Akhil Chilwal; akhil.stat@gmail.com

Received 9 February 2014; Accepted 8 May 2014; Published 25 May 2014

Academic Editor: Steve Su

Copyright © 2014 N. Tiwari and A. Chilwal. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Controlled sampling is a unique method of sample selection that minimizes the probability of selecting nondesirable combinations of units. Extending the concept of linear programming with an effective distance measure, we propose a simple method for two-dimensional optimal controlled selection that ensures zero probability to nondesired samples. Alternative estimators for population total and its variance have also been suggested. Some numerical examples have been considered to demonstrate the utility of the proposed procedure in comparison to the existing procedures.

1. Introduction

Goodman and Kish [1] introduced controlled sampling as a method of sample selection that increases the probability of desired samples. Controlled sampling may be described as a technique of sampling from finite universe, which allows multiple stratifications beyond what is possible by stratified random sampling. There often arises a situation where some combinations of units may be less beneficial or even undesirable to be included in the sample due to considerations such as distance, similarity of units, and cost. The samples having undesirable combinations of units are known as nonpreferred or undesirable samples. Using the technique of controlled selection, one can exclude the possibility of including undesirable combinations of units in the sample or assign them minimum probability of selection. This results in an increase in the selection probability of preferred samples.

The controlled sampling technique can be effectively used in two or more dimensions. Generally, researchers face multidimensional sampling problems in social research where various variables are involved in the population, requiring stratification in more than one dimension. The need of multidimensional stratification in various real life

situations was discussed by Bryant [2], Hess and Srikantan [3], Moore et al. [4], and Jessen [5]. Jessen [5] considered with 12 geographical areas and 12 income classes, resulting in a total of 144 strata cells, out of which only 24 cells were to be selected. In such situations, the researcher requires stratification techniques which could permit fewer cells to be selected than the total number of strata cells permitted under stratified sampling, without sacrificing the requirements of probability sampling. This is known as controls beyond stratification.

Goodman and Kish [1] were the first to address this problem under the name of two-dimensional controlled selection but did not provide any general method to solve such problems. Hess and Srikantan [3] and Groves and Hess [6] discussed the multidimensional controlled selection problem for hospital data in US and presented a formal algorithm for obtaining solutions to the two-dimensional and three-dimensional problems. However, there are simple examples, where their algorithm fails, even for two-dimensional problems. To select the set of feasible samples, Jessen [7] proposed two methods for two-way and three-way stratification but both the methods are quite complicated to implement, involve a lot of trails and errors and sometimes even fail to provide a solution. Ernst [8] was

the first to present a constructive solution for two-dimensional controlled selection problems, but his procedure is quite cumbersome.

Causey et al. [9] proposed an algorithm based on transportation theory to solve two-dimensional controlled selection problems, which is efficient but complex to implement. Inspired by the idea of Rao and Nigam [10, 11], Sitter and Skinner [12] proposed a linear programming approach to solve multidimensional controlled selection problems. Tiwari and Nigam [13] solved the two-dimensional optimal controlled selection problems with controls beyond stratification using simplex method in linear programming. The procedure of Tiwari and Nigam [13] is best suited to problems with integer marginals while the method of Sitter and Skinner [12] is best suited for noninteger marginals. Extending the idea of linear programming of Sitter and Skinner [12], Lu and Sitter [14] discussed some methods to reduce the amount of computation so that very large problems become feasible using the linear programming approach. Tiwari and Nigam [15] applied the idea of nearest proportional to size sampling design to two-dimensional optimal controlled sampling problems using quadratic programming, to introduce a sampling design which ensures zero probability to nonpreferred samples. The procedure of Tiwari and Nigam [15] is efficient but quite cumbersome in the sense that before applying the idea of nearest proportional to size design to obtain the desired controlled inclusion probability proportional to size (IPPS) design they have to first obtain an appropriate uncontrolled IPPS design and then define a non-IPPS design which totally avoids the nonpreferred samples to make their probabilities zero. In this paper, we introduce an effective distance measure as the objective function and a new constraint in linear programming problem to propose a very simple and effective method for two-dimensional controlled sampling which fully excludes the undesirable samples. The proposed procedure appears to perform better than the earlier two-dimensional controlled selection procedures, as it ensures zero probability to undesirable samples without complicating the implementation process.

Another problem that needs attention is of variance estimation in multidimensional controlled selection designs. For one-dimensional controlled selection problems, the Horvitz and Thompson [16] estimator is best suited as the stability, and nonnegativity conditions of the Yates-Grundy [17] form of the Horvitz-Thompson [16] variance estimator are satisfied for such designs. However, as observed by Tiwari and Nigam [15], the two-dimensional controlled selection problems do not satisfy these conditions, owing to the need for alternative estimators. To overcome this difficulty, Jessen [18], Tiwari and Nigam [13], and Tiwari and Nigam [15] suggested alternative variance estimation procedures using the “split sample,” “half sample,” and “random group” methods, respectively. In this paper, we propose a systematic method for estimation of population total and its variance for two-dimensional controlled selection problems. The proposed variance estimator appears to perform better than the existing estimators in terms of bias. We demonstrate its utility with the help of some examples.

2. The Basic Notations and Preliminaries

Let us consider a two-dimensional population array A of N units, consisting of cells that have real numbers, a_{ij} , ($i = 1, \dots, R$, $j = 1, \dots, C$). Suppose a sample of size n is to be obtained from this population. Let y be the characteristic under study, Y_{ij} the y -value for the ij th unit in the population ($i = 1, \dots, R$, $j = 1, \dots, C$), and y_l the y -value for the l th unit in the sample ($l = 1, \dots, n$). Let S_k , $k = 1, \dots, L$, denote the k th possible samples. Also let s_{ijk} be each internal entry of S_k . Then s_{ijk} equals either $[a_{ij}]$ or $[a_{ij}] + 1$, where $[a_{ij}]$ is the integer part of a_{ij} . We have to consider a set of samples with selection probabilities that satisfy the constraints:

$$\begin{aligned} E(s_{ijk} | i, j) &= \sum_{i, j \in S_k, S_k \in S} s_{ijk} p(S_k) = n a_{ij}, \\ \sum_{S_k \in S} p(S_k) &= 1, \end{aligned} \quad (1)$$

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 (1), although only one set of probabilities can be used to obtain a solution of the two-dimensional 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 we consider the following measures of closeness between A and S_k .

The first ordinary distance, which is often called the Euclidean distance, given as

$$\xi_1(A, S_k) = \left[\sum_{i=1}^R \sum_{j=1}^C (a_{ij} - s_{ijk})^2 \right]^{1/2}, \quad k = 1, \dots, L \quad (2)$$

is the most common measure to define the closeness between A and S_k , as it is easy to calculate.

Two other distance measures can also be used to define the distance between A and S_k . These are

(i) cosine distance function:

$$\xi_2(A, S_k) = 1 - \frac{\sum_{i=1}^R \sum_{j=1}^C a_{ij} s_{ijk}}{\|A\|_2 \|S_k\|_2}. \quad (3)$$

(ii) Bray-Curtis distance function:

$$\xi_3(A, S_k) = \frac{\sum_{i=1}^R \sum_{j=1}^C (a_{ij} - s_{ijk})}{\sum_{i=1}^R \sum_{j=1}^C (a_{ij} + s_{ijk})}. \quad (4)$$

Huang [19] and Khatri [20] compared all the above distance measures in their study and found that the cosine distance function works well in comparison to other distance functions. Different distance measures were evaluated empirically using seven data sets by Huang [19] and the results indicated that the cosine distance function performs

reasonably well. We have also applied these three distance functions (2), (3), and (4) to all the controlled sampling problems considered by us and found that the distance function ξ_2 given in (3) provides minimum bias, which supported the works of Huang [19] and Khatri [20]. In view of the above observations, we have decided to use ξ_2 as the distance measure in this paper. We have used OPTMODEL procedure in SAS 9.3 to solve linear objective programming and “pdist2” (pairwise distance between two sets of observations) method in MATLAB 10.0 to solve the three distance functions.

3. The Proposed Two-Dimensional Optimal Controlled Sampling Plan

Let S^* ($S^* \subset S$) denote the set of undesired samples, that is, the samples containing the undesired combinations of units. The required set of samples is obtained through the solution of the following linear programming problem.

Minimize the objective function ϕ , where

$$\phi = \sum_{S_k \in S} \xi_2(A, S_k) p(S_k). \quad (5)$$

Subject to the following constraints:

- (i) $p(S_k) \geq 0$,
- (ii) $\sum_{S_k \in S-S^*} p(S_k) = 1$,
- (iii) $\sum_{S_k \ni i,j} p(S_k) = na_{ij}$, $(i = 1, \dots, R, j = 1, \dots, C)$, (6)
- (iv) $\sum_{S_k \in S^*} p(S_k) = 0$.

The constraints (i) and (ii) in (6) are necessary for any sampling design and the constraint (iii) assures that the resultant design $p(S_k)$ is an IPPS design. The constraint (iv) ensures that the probabilities of undesired samples are equal to zero. We also tried to add one more constraint $\sum_{S_k \ni i,j} p(S_k) \leq \pi_{ijp}\pi_{ijq}$, $(i = 1, \dots, R, j = 1, \dots, C, p < q = 1, \dots, N)$ in (6), to ensure the nonnegativity of the Yates-Grundy form of Horvitz-Thompson variance estimator and applied it to all the two-dimensional controlled selection problems considered by us. However, in no case did it yield a solution. Consequently, we dropped the idea of adding this constraint and suggested an alternative procedure for variance estimation.

The solution of the linear programming problem, namely, minimization of (5) and subject to the constraints (6), using “pdist2” (pairwise distance between two sets of observations) method in MATLAB 10.0 and OPTMODEL procedure in SAS 9.3, provides us optimal controlled IPPS sampling plan that ensures zero probability of selection for the undesired samples. The proposed strategy also provides an opportunity to add more constraints to the controlled selection problem. The proposed plan performs better than the plans of Sitter and Skinner [12] and Tiwari and Nigam [13] in the sense that these plans only attempt to minimize the selection

TABLE 1: The expected sample cell counts (na_{ij}) for the 4×3 population.

	Expected cell counts (na_{ij})			Total
	0.8	0.5	0.7	2.0
	0.8	0.9	0.3	2.0
	0.7	0.7	0.6	2.0
	0.7	0.9	0.4	2.0
Total	3.0	3.0	2.0	8.0

probabilities of the nonpreferred samples, whereas the proposed plan ensures zero probability to nonpreferred samples through constraint (iv) in (6). The exclusion of nonpreferred samples was also attempted by Tiwari and Nigam [15] for two-dimensional controlled selection problems, using the idea of nearest proportional to size design. However, their procedure is quite lengthy and tedious, as in their procedure first of all an uncontrolled IPPS design is to be manually constructed and then the required controlled IPPS design is achieved using the quadratic linear programming approach. The same advantage has been achieved in the proposed plan in a very simple manner by just adding one more constraint in the linear programming problem, ensuring zero probability to nonpreferred samples. The implementation of proposed design is very simple in comparison to the earlier designs. One limitation of proposed design is that it becomes impractical when the set of all possible samples (${}^N C_n$) is very large, as the process of enumerating of all possible samples and formation of the objective function and constraints becomes quit tedious. This limitation also holds for the optimum approach of Sitter and Skinner [12], Tiwari and Nigam [13], and Tiwari and Nigam [15]. However, with the help of faster computing techniques and modern statistical tools, there may not be much difficulty in using the proposed plan for moderately large populations. Nevertheless, the proposed procedure takes lesser computing time in comparison to the procedures of Sitter and Skinner [12], Tiwari and Nigam [13], and Tiwari and Nigam [15]. In what follows, we show the utility of the proposed procedure with the help of some numerical examples.

4. Empirical Evaluation

In this section, we will present some numerical examples to demonstrate the utility of the proposed procedure and compare it with the existing procedures of optimal controlled sampling designs.

Example 1. Let us consider a 4×3 hypothetical population borrowed from Tiwari and Nigam [15], given in Table 1. The desired sample size of $n = 8$ is less than the total number of cells, 12. The set of all possible samples consists of 12 samples, given in Table 2. Let the set of undesirable samples consists of those samples that do not contain all the three elements 1st, 5th, and 9th or 3rd, 5th, and 7th. Thus the sample numbers 6th and 9th are the nonpreferred samples.

TABLE 2: The set of all possible samples for 4×3 population.

Sample 1			Sample 2			Sample 3			Sample 4			Sample 5			Sample 6		
1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
1	1	0	1	1	0	0	1	1	0	1	1	1	0	1	1	0	1
0	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1
1	0	1	0	1	1	1	0	1	1	1	0	0	1	1	1	1	0
Sample 7			Sample 8			Sample 9			Sample 10			Sample 11			Sample 12		
0	1	1	0	1	1	0	1	1	1	0	1	1	0	1	1	0	1
1	1	0	1	1	0	1	0	1	1	1	0	1	1	0	0	1	1
1	1	0	1	0	1	1	1	0	1	1	0	0	1	1	1	1	0
1	0	1	1	1	0	1	1	0	0	1	1	1	1	0	1	1	0

TABLE 3: Probabilities of selection of samples.

Sample no.	TN-1	TN-2	$p(S_k)$
1	0.1	0.0462	0.1
2	0.2	0.0829	0.1
3	0.0	0.0121	0.0
4	0.0	0.0589	0.0
5	0.0	0.1	0.1
6 [#]	0.0	0.0	0.0
7	0.0	0.0417	0.0
8	0.1	0.1583	0.2
9 [#]	0.1	0.0	0.0
10	0.1	0.1171	0.1
11	0.2	0.2528	0.2
12	0.2	0.1290	0.2

[#]Undesirable sample.

Applying the Tiwari and Nigam [13] (to be denoted by TN-1), Tiwari and Nigam [15] (to be denoted by TN-2), and the proposed plan discussed in Section 3 to this population, we get the selection probabilities of the samples as shown in Table 3. For this example we find that the probability of nonpreferred samples for Tiwari and Nigam [13] plan is 0.1, whereas the proposed plan always assures zero probability to nonpreferred samples.

Example 2. Now let us consider another hypothetical example borrowed from Bryant et al. [21] given in Table 4. The desired sample of size 10 is less than the total number of units, 15. The integer parts of na_{ij} 's are known as "certainty proportion." For obtaining the set of feasible samples, we initially remove the certainty proportions and replace them at their original position after getting these samples. After removing the certainty proportions, we get a two-way array shown in Table 5.

After subtracting the certainty proportions, the problem is reduced to selecting 6 units from the array. The set S of all possible samples consists of $^{15}C_6$ samples, out of which 4989 samples do not satisfy the marginal constraints of the 5×3 population. Thus, the set of samples satisfying the marginal constraints have only 16 samples, given in Table 6.

TABLE 4: Expected sample cell counts (na_{ij}) under proportionate stratification with $n = 10$.

Type of community	Regions			Total
	Urban	Rural	Metropolitan	
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

TABLE 5: Expected sample cell counts (na_{ij}) for 5×3 population after removing certainty proportions.

Expected sample cell counts (na_{ij})			Total
0.0	0.5	0.5	1.0
0.2	0.3	0.5	1.0
0.2	0.6	0.2	1.0
0.6	0.8	0.6	2.0
0.0	0.8	0.2	1.0
Total	1.0	3.0	6.0

Now we suppose the situation of controls beyond stratification. Based on the considerations similar to those of Avadhani and Shukhatme [22], Tiwari and Nigam [13], and Tiwari and Nigam [15], we consider that if all three units 4th, 8th, and 12th or 6th, 8th, and 10th do not appear in a sample, then the sample is nonpreferred sample. Thus the set of all preferred samples consists of only 10 samples, that is, the sample numbers 1, 3, 5, 7, 9, 10, 11, 13, 14, and 16. Applying the proposed, Tiwari and Nigam [13] [TN-1] and Tiwari and Nigam [15] [TN-2] plans to the modified problem, we get the selection probabilities, shown in Table 7.

For this example, the probability of undesired samples is zero for the proposed plan and Tiwari and Nigam [15] plan. The proposed plan again ensures zero probability to undesirable samples, whereas the plan of Tiwari and Nigam [13] only attempts to minimize the probability of undesirable samples.

TABLE 6: The set of all preferred combinations for 5×3 population.

Sample 1			Sample 2			Sample 3			Sample 4			Sample 5			Sample 6		
1	1	0	1	1	0	1	0	1	1	1	0	1	1	0	1	0	1
1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0
0	1	1	0	0	2	0	1	1	1	0	1	1	0	1	1	0	1
0	2	1	0	2	1	0	2	1	0	2	1	0	2	1	0	2	1
1	0	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	0
Sample 7			Sample 8			Sample 9			Sample 10			Sample 11			Sample 12		
1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0
0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1
0	1	1	0	0	2	0	0	2	0	1	1	0	1	1	0	0	2
1	1	1	1	2	0	1	1	1	1	2	0	1	1	1	1	2	0
1	0	1	1	0	1	1	1	0	1	0	1	1	1	0	1	1	0
Sample 13			Sample 14			Sample 15			Sample 16								
1	0	1	1	0	1	1	0	1	1	0	1						
0	1	0	0	1	0	0	1	0	0	0	1						
0	1	1	0	1	1	0	0	2	0	1	1						
1	2	0	1	1	1	1	2	0	1	2	0						
1	0	1	1	1	0	1	1	0	1	1	0						

TABLE 7: Probabilities of selection of samples.

Sample no.	TN-1	TN-2	$p(S_k)$
1	0.1	0.028175	0.0
3	0.1	0.171825	0.2
5	0.2	0.2	0.2
7	0.0	0.0	0.0
9	0.2	0.2	0.2
10	0.0	0.071825	0.1
11	0.0	0.0	0.0
13	0.1	0.1	0.1
14	0.0	0.0	0.0
16	0.3	0.228175	0.2

Example 3. Let us consider a real life application, borrowed from Tiwari and Nigam [15], where two-dimensional stratification is required in plot sampling in field experiments. Consider the yield (in tons) of wheat given in Table 8 for an experiment involving blocks (B1, B2, B3, and B4) and 4 treatments (T1, T2, T3, and T4). The integer parts of na_{ij} 's are known as "certainty proportion." For obtaining the set of feasible samples, we initially remove the certainty proportions and replace them at their original position after getting these samples. After removing the certainty proportions, we get a two-way array shown in Table 9.

After subtracting the certainty proportions, the problem is reduced to selecting 8 units from the array. The set S of all possible samples consists of ${}^{16}C_8$ samples, out of which 12780 samples do not satisfy the marginal constraints of the 4×4 population. Thus, the set of samples satisfying the marginal constraints have only 90 samples. Now we suppose the situation of controls beyond stratification.

TABLE 8: Yield (in tons) for the 4×4 experiments.

	T1	T2	T3	T4	Total
B1	11.25	18.75	13.75	6.25	50.00
B2	16.25	3.75	6.25	23.75	50.00
B3	16.50	6.50	18.75	8.25	50.00
B4	6.00	21.00	11.25	11.75	50.00
Total	50.00	50.00	50.00	50.00	200.00

TABLE 9: Expected sample cell counts (na_{ij}) for the 4×4 experiments.

	Expected sample cell counts (na_{ij})				Total
	0.45	0.75	0.55	0.25	2.0
	0.65	0.15	0.25	0.95	2.0
	0.66	0.26	0.75	0.33	2.0
	0.24	0.84	0.45	0.47	2.0
Total	2.00	2.00	2.00	2.00	8.0

Based on the considerations similar to Tiwari and Nigam [15], we consider that if three or more diagonal units appear in a sample, then the sample is nonpreferred sample. Thus, the set of all preferred samples consists of only 33 samples, shown in Table 10. Applying the proposed, Tiwari and Nigam [15] [TN-2] plans to the modified problem, we get the selection probabilities, shown in Table 11. For this example, the probability of undesired samples is zero for the proposed and Tiwari and Nigam [15] plans.

Some other examples are also considered to analyse the performance of the proposed plan. Details of these examples are given in the Appendix. The probabilities of selecting the undesirable samples for the proposed plan, the plan of Tiwari

TABLE 10: The set of all preferred combinations for 4×4 population.

Sample 1				Sample 2				Sample 3				Sample 4				Sample 5				Sample 6			
1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	1	1	0
0	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0	1	0	1	1	0	0
1	1	0	0	0	0	1	1	1	0	0	1	0	0	1	1	0	1	0	1	1	0	0	1
0	0	1	1	1	1	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1
Sample 7				Sample 8				Sample 9				Sample 10				Sample 11				Sample 12			
0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
0	1	1	0	0	0	1	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
1	0	0	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1	1	0	0	1
1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0	1	1	0
Sample 13				Sample 14				Sample 15				Sample 16				Sample 17				Sample 18			
0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1
1	0	0	1	1	0	0	1	1	0	1	0	0	1	0	1	1	1	0	0	1	1	0	0
1	0	1	0	0	1	0	1	1	0	0	1	1	0	0	1	1	1	0	0	0	0	1	1
0	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	0	1	1	1	1	0	0
Sample 19				Sample 20				Sample 21				Sample 22				Sample 23				Sample 24			
0	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0	0	1	1	0	1	0
1	1	0	0	1	0	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	1	1
1	0	0	1	1	1	0	0	1	0	1	0	1	0	0	1	0	1	1	0	1	1	0	0
0	1	1	0	0	1	1	0	1	1	0	0	0	1	1	0	0	1	1	0	0	1	0	1
Sample 25				Sample 26				Sample 27				Sample 28				Sample 29				Sample 30			
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1
1	0	0	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1	0	0	1	0	0	1
0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	1	0	0	1	1	1	0	1	0
0	1	1	0	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	0	1	1	0
Sample 31				Sample 32				Sample 33															
0	1	0	1	0	1	0	1	0	1	0	1												
1	0	1	0	1	0	1	0	0	1	0	1												
1	0	0	1	1	0	1	0	1	0	1	0												
0	1	1	0	0	1	0	1	1	0	1	0												

and Nigam [13] and Tiwari and Nigam [15], are given in Table 12. Table 12 again shows that while the plan of Tiwari and Nigam [13] only attempts to minimize the probability of undesirable samples, the proposed plan always ensures zero probability to undesirable samples. Tiwari and Nigam [15] plan also provides zero probability of nonpreferred samples, but as discussed earlier, it is quite difficult to implement.

5. Variance Estimation for the Proposed Procedure

Jessen [18] suggested split sample estimator as an alternative to Horvitz-Thompson (HT) estimator. This estimator also works in the situation where the nonnegativity condition of Yates-Grundy form of HT estimator is not satisfied. Jessen's split sample estimator is negatively biased and bias is found to be quite high. Using half sample method, Tiwari and Nigam

[13] introduced a method of variance estimation for two-dimensional controlled selection problems. Their variance estimator was found to be positively biased and the bias was low in comparison to Jessen's split sample estimator. An important limitation of both the estimators is that they require exactly two units from each row and column of the two-way array. The above two methods could not be applied if two units from each row and column are not available. Using the idea of random group, Tiwari and Nigam [15] introduced an alternative estimator for population total and its variance that can be used even when two units are not available from each row and column of the two-way array.

Using the procedure of systematic sampling for variance estimation, originally developed by W. G. Madow and L. H. Madow [23], we propose an alternative estimation procedure for the population total and its variance in two-dimensional controlled selection problems. The proposed estimator performs better than the split sample estimator of Jessen [18]

TABLE 11: Probabilities of selection of samples.

Sample (S_k)	TN-2	$p(S_k)$
1	0.023925	0.0
2	0.073749	0.0
3	0.051726	0.08
4	0.114597	0.25
5	0.000641	0.0
6	0.003188	0.0
7	0.000384	0.0
8	0.007685	0.0
9	0.03233	0.05
10	0.061156	0.09
11	0.0	0.0
12	0.0	0.0
13	0.257761	0.16
14	0.000339	0.0
15	0.0049	0.0
16	0.010486	0.0
17	0.000425	0.0
18	0.001588	0.0
19	0.001867	0.0
20	0.00327	0.0
21	0.057485	0.13
22	0.000456	0.0
23	0.07772	0.0
24	0.050686	0.12
25	0.007147	0.0
26	0.001349	0.0
27	0.046992	0.0
28	0.00096	0.0
29	0.000702	0.0
30	0.04646	0.05
31	0.001242	0.0
32	0.033256	0.05
33	0.025465	0.02

TABLE 12: The probabilities of selecting the undesirable samples for the proposed, Tiwari and Nigam [13] [TN-1] and Tiwari and Nigam [15] [TN-2] plans.

	TN-1	TN-2	Proposed plan
Example A.1 (4×3 population, $N = 12$, $n = 8$)	0.1	0	0
Example A.2 (3×3 population, $N = 9$, $n = 3$)	0.1	0	0
Example A.3 (8×3 population, $N = 24$, $n = 10$)	0	0	0

and the estimators proposed by Tiwari and Nigam [13] and Tiwari and Nigam [15] in terms of bias. The proposed procedure can also be used in the situations where exactly two units are not available from each row and column of the two-way array. The proposed approach is as follows.

To construct k ($k \geq 2$) systematic samples from a sample of size n drawn from a population of N units, we first arrange all n sample units in a list: they can be placed at random in the list or they can be placed in a particular sequence or they can be left in a sequence that they naturally occur.

Let y_i be the y -value for the i th unit in the sample ($i = 1, \dots, n$) and let x_i be the measure of size. Next, a cumulative measure of size, C_i , is calculated for each sample unit; that is, $C_i = \sum_{j=1}^i x_j$. To select a systematic sample of m units, a selection interval, say, B , is calculated as the total of all measures of size divided by m ; that is, $B = \sum_{i=1}^n x_i / m$. The selection interval B is not necessarily an integer but is typically rounded off to two or three decimal places. To initiate the sample selection process, a uniform random deviate, say, U , is chosen on the half open interval $(0, B]$. The m selection numbers for the sample are then $U, U + B, U + 2B, U + 3B, \dots, U + (m-1)B$. The sample unit identified for the systematic sample by each selection number is the first unit on the list for which the cumulative size, C_i , is greater than or equal to the selection number. With the help of this procedure the n units of the sample can be divided into k systematic samples. The various values of k will give various systematic samples and the proposed estimator will depend on the value of k . However, it has been found that the proposed estimator works satisfactorily in all the situations.

With the help of this approach, an unbiased estimator of population total is given as

$$\hat{Y}_S = \sum_{t=1}^k \left(\frac{y_{t1}}{\pi_{t1}} + \frac{y_{t2}}{\pi_{t2}} \right), \quad (7)$$

where y_{t1} and y_{t2} are the observation from the t th systematic sample and π_{t1} and π_{t2} are their corresponding inclusion probabilities. An estimator of the variance of $\hat{Y}$ is given as

$$\widehat{V}(\hat{Y}_S) = \left(1 - n \sum_{i=1}^N p_i^2 \right) \left(\frac{1}{4} \right) \left[\sum_{t=1}^k \left(\frac{y_{t1}}{\pi_{t1}} - \frac{y_{t2}}{\pi_{t2}} \right)^2 \right], \quad (8)$$

where $(1 - n \sum_{i=1}^N p_i^2)$ is an approximate finite population correction factor.

The proposed procedure of variance estimation can be applied for square as well as for rectangular populations and works equally well even for the situation where the units selected from each row and column are not fixed and equal. When the nonnegativity condition of Yates-Grundy form of Horvitz-Thompson variance estimator is not satisfied, we can apply the variance estimator given in (8). The proposed variance estimator is always positive as it involves only the sum of squared quantities. We consider some examples to show the utility of proposed variance estimator and compare it with the Jessen's split sample estimator and the estimators suggested by Tiwari and Nigam [13, 15].

Example 4. Let us consider a 3×3 population borrowed from Jessen [7], shown in Table 13. Values of $\text{Var}(\hat{Y})$ obtained by Jessen's split sample estimator (to be denoted by S-S), the estimator given by Tiwari and Nigam [13] (to be denoted by TN-1), Tiwari and Nigam [15] (to be denoted by TN-2), and the proposed estimator are shown in Table 14.

TABLE 13: Basic data for the 3×3 population.

	na_{ij}			Total		Y_{ij}			Total
	0.8	0.5	0.7	2.0		5.6	3.0	6.3	14.9
	0.7	0.8	0.5	2.0		4.2	3.2	4.0	11.4
	0.5	0.7	0.8	2.0		1.5	3.5	5.6	10.6
Total	2.0	2.0	2.0	6.0	Total	11.3	9.7	15.9	36.9

TABLE 14: Comparison of various estimators of $\text{Var}(\hat{Y})$ for 3×3 population.

Sample	$p(s)$	$(\hat{Y})_s$	$\widehat{V}(\hat{Y})$			
			S-S	TN-1	TN-2	Proposed estimator
1	0.3	19/3	0.034445	0.03875	0.064583	0.0310145
2	0.1	35/6	0.008611	0.11625	0.075347	0.0452170
3	0.2	13/2	0.008611	0.025833	0.023680	0.0107650
4	0.2	6	0.137777	0.090417	0.051667	0.1162725
5	0.2	35/6	0.077500	0.099028	0.088264	0.0968930
$E[\widehat{\text{Var}}(\hat{Y})]$			0.0559722	0.0663056	0.0596319	0.0583512
Bias			-0.0021278	0.0082056	0.0015319	0.0002512
$V[\widehat{\text{Var}}(\hat{Y})]$			0.0022430	0.0011353	0.0004672	0.0016769
$(\text{Bias})^2$			0.0000045	0.0000673	0.0000023	0.0000006311
$\text{MSE}[\widehat{\text{Var}}(\hat{Y})]$			0.0022475	0.0012025	0.0004695	0.001676955
$(\text{Bias})^2/\text{MSE}[\widehat{\text{Var}}(\hat{Y})]$			0.0020022	0.0559667	0.0048988	0.00003757

The actual value of Y for this population is 123/20. From Table 13, we have

$$E(\hat{Y}_s) = \sum_{i=1}^5 p(s) (\hat{Y})_s = \frac{123}{20} = Y. \quad (9)$$

Thus $\hat{Y}_s$ is an unbiased estimate of Y . The expected value of $\widehat{V}(\hat{Y}_s)$ for proposed estimator is

$$E[\widehat{V}(\hat{Y}_s)] = \sum_{i=1}^5 p(s) \widehat{V}(\hat{Y})_s = 0.05835121. \quad (10)$$

The true value of $V(\hat{Y})$ for this population is 0.0581, which shows that the proposed estimator is positively biased. The bias of the proposed estimator is lowest among the four estimators, showing that the proposed estimator performs better than the previous estimators.

Example 5. To further evaluate the utility of the proposed variance estimator, we consider a 4×4 population borrowed from Jessen [24], shown in Table 15. A sample of size 8 is to be drawn from this population. The values of $\widehat{\text{Var}}(\hat{Y})$ for the four estimators and selection probabilities of all twenty possible samples are presented in Table 16.

From Table 15, we get

$$E(\hat{Y}_s) = \sum_{i=1}^{20} p(s) (\hat{Y})_s = 10. \quad (11)$$

Thus $\hat{Y}_s$ is an unbiased estimate of Y . The expected value of $\widehat{V}(\hat{Y}_s)$ for the proposed estimator is

$$E[\widehat{V}(\hat{Y}_s)] = \sum_{i=1}^{20} p(s) \widehat{V}(\hat{Y})_s = 0.246035. \quad (12)$$

The true value of $\text{Var}(\hat{Y})$ for this population is 0.24375, which shows that the proposed estimator is positively biased. The bias is lowest for the proposed estimator among the four estimators considered by us.

The outcomes of the above two examples show that the proposed variance estimator performs better than the estimators suggested by Jessen [18], Tiwari and Nigam [13], and Tiwari and Nigam [15]. The bias is minimum for the proposed estimator and it also performs favourably in the situations where the estimators of Jessen [18] and Tiwari and Nigam [13] cannot be applied.

6. Conclusion

In this paper, we have proposed a simple linear programming approach using distance measure as a weight for each sample to obtain an optimum solution in two-way controlled selection problems. In the proposed plan, we have introduced one more constraint in linear programming problem to ensure zero probability to nonpreferred samples. The proposed procedure is quite simple and flexible to implement. We have also proposed a new strategy for the estimation of variance in two-way controlled sampling designs. The proposed estimator appears to perform better than the earlier estimators for two-dimensional controlled sampling suggested by different

TABLE 15: Basic data for the 4×4 population.

na_{ij}					Total	Y_{ij}					Total
0.0	0.6	1.0	0.4	2.0		0.0	1.05	2.25	0.8	4.10	
0.8	0.4	0.4	0.4	2.0		0.6	0.70	0.65	0.4	2.35	
0.6	0.2	0.4	0.8	2.0		0.6	0.20	0.50	0.6	1.90	
0.6	0.8	0.2	0.4	2.0		0.3	0.80	0.25	0.3	1.65	
Total	2.0	2.0	2.0	2.0	8.0	Total	1.5	2.75	3.65	2.1	10.00

TABLE 16: Comparison of various values of $\widehat{V}(\widehat{Y})$ for the 4×4 population.

$\widehat{V}(\widehat{Y})$						
Sample	$p(s)$	$(\widehat{Y})_s$	S-S	TN-1	TN-2	Proposed estimator
1	0.1	10.250	0.578125	0.242813	0.329531	0.14453
2	0.1	9.875	0.052031	0.199453	0.342539	0.19799
3	0.0	9.375	0.005781	0.199453	0.221133	0.33097
4	0.0	9.750	0.208125	0.196563	0.167656	0.22546
5	0.0	9.750	0.023125	0.208125	0.052031	0.15609
6	0.3	9.250	0.023125	0.196563	0.248594	0.23346
7	0.0	9.250	0.023125	0.231250	0.040469	0.469727
8	0.0	10.375	0.005781	0.095391	0.383008	0.260156
9	0.1	10.250	0.023125	0.254375	0.398906	0.317960
10	0.0	9.870	0.052031	0.182109	0.319414	0.186445
11	0.0	10.250	0.208125	0.439375	0.375781	0.109844
12	0.0	10.750	0.578125	0.416250	0.387344	0.237031
13	0.0	10.250	0.023125	0.450938	0.271719	0.109844
14	0.0	10.750	0.578125	0.404688	0.352656	0.294844
15	0.1	10.375	0.283281	0.268828	0.359883	0.266881
16	0.0	9.750	0.023113	0.231250	0.387344	0.317969
17	0.2	10.875	0.144531	0.268828	0.296289	0.276831
18	0.0	10.875	0.052031	0.291953	0.307852	0.213906
19	0.1	10.750	0.578125	0.370000	0.375781	0.278953
20	0.0	10.750	0.023125	0.346875	0.260156	0.341094
$E[\widehat{\text{Var}}(\widehat{Y})]$			0.187202	0.246281	0.255927	0.24603500
Bias			-0.056548	0.002531	0.012177	0.00228500
$V[\widehat{\text{Var}}(\widehat{Y})]$			0.0444260	0.002652	0.016581	0.00216700
$(\text{Bias})^2$			0.0031977	0.0000064	0.000148	0.00000522
$\text{MSE}[\widehat{\text{Var}}(\widehat{Y})]$			0.0476237	0.0026584	0.016729	0.00217222
$(\text{Bias})^2/\text{MSE}[\widehat{\text{Var}}(\widehat{Y})]$			0.0671451	0.0024112	0.0088585	0.00240309

TABLE 17: The expected sample cell counts (na_{ij}) for the 4×3 population.

Expected cell counts (na_{ij})				Total
0.8	0.5	0.7	2.0	
0.8	0.9	0.3	2.0	
0.7	0.7	0.6	2.0	
0.7	0.9	0.4	2.0	
Total	3.0	3.0	2.0	8.0

TABLE 18: The expected sample cell counts (na_{ij}) for the 3×3 population.

Expected cell counts (na_{ij})				Total
0.2	0.3	0.5	1.0	
0.2	0.4	0.4	1.0	
0.6	0.3	0.1	1.0	
Total	1.0	1.0	1.0	3.0

researchers. The proposed procedure takes lesser computing time in comparison to the procedures of Tiwari and Nigam [13] and Tiwari and Nigam [15] and is found to be more advantageous than these plans.

Appendix

Example A.1. Consider a 4×3 hypothetical population, borrowed from Tiwari and Nigam [15] given in Table 17, with population size 12 and sample size is equal to 8. The samples

TABLE 19: The expected sample cell counts (na_{ij}) for the 8×3 population.

Expected cell counts (na_{ij})			Total
0.4	2.0	0.0	2.40
1.2	0.0	1.0	2.20
0.2	0.0	0.0	0.20
1.2	0.4	0.2	1.80
1.0	0.6	0.2	1.80
0.0	0.4	0.4	0.80
0.0	0.2	0.4	0.60
0.0	0.0	0.2	0.20
Total	4.0	3.6	2.4
			10.00

in which all three units 1st, 5th, and 9th or 3rd, 5th, and 7th do not appear, considered as undesired samples.

Example A.2. Consider a 3×3 hypothetical population with $N = 9$ and $n = 3$, borrowed from Tiwari and Nigam [15]. The proposed sample cell counts (na_{ij}) for this population are given in Table 18. The samples in which all the three elements 1st, 5th, and 9th appear together considered as undesired samples.

Example A.3. Let us consider an 8×3 population, borrowed from Causey et al. [9] consisting of 24 elements, and a sample of size 10 is to be drawn from it. The basic data for this population is shown in Table 19. The samples having two consecutive elements in a column are assumed as undesired samples.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] R. Goodman and L. Kish, "Controlled selection-a technique in probability sampling," *Journal of the American Statistical Association*, vol. 45, no. 251, pp. 350–372, 1950.
- [2] E. C. Bryant, "Sampling methods," Seminar Paper, Iowa State University, 1961.
- [3] I. Hess and K. S. Srikantan, "Some aspects of the probability sampling technique of controlled selection," *Health Services Research*, vol. 1, no. 1, pp. 8–52, 1966.
- [4] R. P. Moore, J. R. Chromy, and W. T. Rogers, *The National Assessment Approach to Sampling*, National Assessment of Educational Progress, Denver, Colo, USA, 1974.
- [5] R. J. Jensen, "Square and cubic lattice sampling," *Biometrics*, vol. 31, no. 2, pp. 449–471, 1975.
- [6] R. M. Groves and I. Hess, "An algorithm for controlled selection," in *Probability Sampling of Hospitals and Patients*, I. Hess, D. C. Ridel, and T. B. Fitzpatrick, Eds., chapter 7, Health Administration Press, Ann Arbor, Mich, USA, 2nd edition, 1975.
- [7] R. L. Jensen, "Probability sampling with marginal constraints," *Journal of the American Statistical Association*, vol. 65, no. 330, pp. 776–796, 1970.
- [8] L. R. Ernst, "A constructive solution for two-dimensional controlled selection problems," in *Proceeding of the Survey Research Methodology Section*, American Statistical Association, 1981.
- [9] B. D. Causey, L. H. Cox, and L. R. Ernst, "Application of transformation theory to statistical problem," *Journal of the American Statistical Association*, vol. 80, no. 392, pp. 903–909, 1985.
- [10] J. N. K. Rao and A. K. Nigam, "Optimal controlled sampling designs," *Biometrika*, vol. 77, no. 4, pp. 807–814, 1990.
- [11] J. N. K. Rao and A. K. Nigam, "Optimal controlled sampling: a unified approach," *International Statistical Review*, vol. 60, no. 1, pp. 89–98, 1992.
- [12] R. R. Sitter and C. J. Skinner, "Multi-way stratification by linear programming," *Survey Methodology*, vol. 20, no. 1, pp. 65–73, 1994.
- [13] N. Tiwari and A. K. Nigam, "On two-dimensional optimal controlled selection," *Journal of Statistical Planning and Inference*, vol. 69, no. 1, pp. 89–100, 1998.
- [14] W. Lu and R. R. Sitter, "Multi-way stratification by linear programming made practical," *Survey Methodology*, vol. 28, no. 2, pp. 199–207, 2002.
- [15] N. Tiwari and A. K. Nigam, "On two-dimensional optimal controlled nearest proportional to size sampling designs," *Statistical Methodology*, vol. 7, no. 6, pp. 601–613, 2010.
- [16] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.
- [17] F. Yates and P. M. Grundy, "Selection without replacement from within strata with probability proportional to size," *Journal of Royal Statistical Society*, vol. 15, no. 2, pp. 253–261, 1953.
- [18] R. L. Jensen, "Some properties of probability lattice sampling," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 26–28, 1973.
- [19] A. Huang, "Similarity measures for text document clustering," in *Proceeding of the New Zealand Computer Science Research Student Conference (NZCSRSC '08)*, pp. 49–56, Christchurch, New Zealand, April 2008.
- [20] M. Khatri, "Cosine similarity function for the temporal dynamic web data," *International Journal of Computer Science & Engineering Technology*, vol. 3, no. 8, pp. 315–318, 2012.
- [21] E. C. Bryant, H. O. Hartley, and R. J. Jensen, "Design and estimation in two-way stratification," *Journal of the American Statistical Association*, vol. 55, no. 289, pp. 105–124, 1960.
- [22] M. S. Avadhani and B. V. Shukhatme, "Controlled sampling with equal probabilities and without replacement," *International Statistical Review*, vol. 41, no. 2, pp. 175–182, 1973.
- [23] W. G. Madow and L. H. Madow, "On the theory of systematic sampling," *The Annals of Mathematical Statistics*, vol. 15, no. 1, pp. 1–24, 1944.
- [24] R. L. Jensen, *Statistical Theory Techniques*, Wiley, New York, NY, USA, 1978.