Optimizing Solution Sets in Two-way Controlled Selection Problems

Sun-Woong Kim, Steven G. Heeringa, and Peter Solenberger, University of Michigan
Sun-Woong Kim, Institute for Social Research, 426 Thompson Street, Ann Arbor, Michigan 48106

Key Words: Controlled selection, optimal samples, linear programming

1. Introduction

Various controlled selection techniques have been developed since Goodman and Kish (1950) first suggested the method. Jessen (1970) proposed two methods which can be quite complicated to implement and sometimes fail to provide a solution. Jessen (1978) approached the controlled selection problem as probability lattice sampling.

Hess, Riedel and Fitzpatrick (1975) gave a detailed explanation of how to use controlled selection in order to select a sample of Michigan's hospitals. Groves and Hess (1975) suggested a formal computer algorithm for obtaining solutions to two- and three-dimensional controlled selection problems.

Causey, Cox and Ernst (1985) proposed an algorithm based on transportation theory to solve two-dimensional controlled selection problems, an approach originally suggested in a previous paper by Cox and Ernst (1982).

Following Rao and Nigam (1990, 1992), Sitter and Skinner (1994) used a linear programming approach to solve controlled selection problems. Tiwari and Nigam (1998) proposed using linear programming to reduce the selection probabilities of non-preferred combination of units.

Huang and Lin (1998) proposed a recursive algorithm using network flow to solve the two-dimensional controlled selection problem with row or column subtotals.

In this paper, we suggest a new linear programming approach. Our method adopts ordinary distance functions to minimize the overall distortion to cell sample size expectations in two-dimensional controlled selection problems.

2. Optimal Samples

Consider the two-way controlled selection problem that is denoted by the $R \times C$ tabular array A, which consists of cells that have real numbers, a_{ij}, $i = 1, \cdots, R$, $j = 1, \cdots, C$. Let B_k, $k = 1, \cdots, L$, denote possible samples where each sample is the replacement of the real numbers in A by the adjacent integers. Also, let b_{ijk} be each internal entry of B_k. Then b_{ijk} equals either $\lfloor a_{ij} \rfloor$ or $\lfloor a_{ij} \rfloor + 1$, where $\lfloor a_{ij} \rfloor$ is the integer part of a_{ij}.

The major restriction on finding a solution to this controlled selection problem is the fact that the selection probabilities of samples should depend only on the tabular array A. In other words, we have to consider a set of samples with selection probabilities that satisfy the constraints:

$$E\left(b_{ijk} \middle| i, j\right) = \sum_{ij \in B_k, B_k \in B} b_{ijk}\, p(B_k) = a_{ij} \quad (2.1)$$

and

$$\sum_{B_k \in B} p(B_k) = 1, \quad (2.2)$$

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

There may be a large number of sets of probability distributions $p(B_k)$ satisfying (2.1) and (2.2), although only one set of probabilities can be used to obtain a solution to the controlled selection problem. In this case, we may consider an algorithm to find the solution that reflects the closeness of each sample B_k to A, which is based on an appropriate and objective principle for measuring this "closeness."

For this purpose we consider several measures of closeness between A and B_k. The ordinary distance metric, which is often called the Euclidean metric, could be used.

$$d_1(A : B_k) = \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \left(a_{ij} - b_{ijk}\right)^2 \right]^{\frac{1}{2}}, \quad k = 1, \cdots, L \quad (2.3)$$

This function would be the most common measure to define the distance between arrays A and B_k.

The metric measure below could be also used to define another distance function for each integer m.

$$d_2(A : B_k) = \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \left(a_{ij} - b_{ijk}\right)^{2m} \right]^{\frac{1}{2m}}, \quad k = 1, \cdots, L,\ 1 < m < \infty \quad (2.4)$$

It is clear that we can define other distance functions using the following metric:

$$d_3(A : B_k) = \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \left|a_{ij} - b_{ijk}\right|^{p} \right]^{\frac{1}{p}},$$

$$k = 1, \cdots, L,\ 1 \le p < \infty \qquad (2.5)$$

Furthermore we can define the distance function below for $p = \infty$.

$$d_4(A : B_k) = \max\left\{\left|a_{ij} - b_{ijk}\right| : 1 \le i \le R, 1 \le j \le C\right\},$$
$$k = 1, \cdots, L \qquad (2.6)$$

which is motivated by the fact that:

$$d_4(A : B_k) = \lim_{p \to \infty} d_3(A : B_k) \qquad (2.7)$$

The different metrics for measuring the "closeness" of A and B_k, such as d_1, d_2, d_3, and d_4, give rise to a number of distinct metric spaces. Here we focus on d_1 or d_4, which are special cases of d_3 with $p = 2$ and $p = \infty$. These are chosen because d_1 represents the "overall distance" between A and B_k considering all $R \times C$ cells, whereas d_4 indicates the maximum deviation in a single cell.

We define a few samples in the set of all possible samples, $\{B_k\}$, having the minimum distance value from d_1 or d_4 as 'optimal samples.' For identifying optimal samples, we would prefer d_4 rather than d_1 because the following relation always holds:

$$B_0 \subseteq B_1, \qquad (2.8)$$

where B_0 is the set of optimal samples under metric d_1 and B_1 is the set of optimal samples for metric d_4.

Empirically, there would be a very small number of samples to be added to our subset of optimal samples when we use d_4 compared to d_1.

On the other hand, if we define 'unfavorable samples' as samples which have the maximum distance value, d_1 yields a greater number of unfavorable samples than does d_4.

In the next section, we first review several controlled selection methods in the literature and then explain our new algorithm to obtain solutions of controlled selection problems. We use a simple linear programming approach which maximizes the selection probabilities of optimal samples under distance metric d_1 and d_4 and simultaneously minimizes the selection probabilities of unfavorable samples.

3. Optimal controlled selection

The algorithm developed by Causey et al. (1985) applies transportation theory to approximate nonlinear distance functions (2.5) and (2.6) by linear functions in order to obtain a solution for the controlled selection problem. To solve the problem specified by the array A, their algorithm requires the solution to a sequence of problems. Thus this algorithm is not only less direct, but also more purposive in making a decision on the selection probabilities of samples.

Sitter and Skinner (1994) showed how linear programming may be applied to controlled selection problems. Their key idea is to minimize "the expected lack of desirability" of samples with regard to margins of the tabular array A. Their method is primarily applicable to controlled selection problem A, with non-integer margins.

Huang and Lin (1998) adopted subgroup constraints raised by Goodman and Kish (1950) and solved a controlled selection problem as a network flow problem. Their method uses a recursive algorithm based on simple definitions of the selection probabilities of samples, which is similar to the method of Causey et al. (1985).

In this section, we present an algorithm to optimize the assignment of probability to each sample.

First, consider all possible samples for the controlled selection problem corresponding to rounding of tabular array A. As mentioned above, the set of possible samples is denoted by B. Second, establish the following linear programming problem:

$$\phi_1 = \sum_{B_k \in B} d_1(A : B_k)\, p(B_k) \qquad (3.1)$$

or

$$\phi_2 = \sum_{B_k \in B} d_2(A : B_k)\, p(B_k) \qquad (3.2)$$

subject to:

$$\sum_{ij \in B_k} p(B_k) = a_{ij}^*, \qquad (3.3)$$

$$p(B_k) \ge 0, \qquad (3.4)$$

$$\sum_{B_k \in B} p(B_k) = 1, \qquad (3.5)$$

where a_{ij}^* is the non-integer part of a_{ij}.

For distance measure d_2, this method can be employed for any integer, $m > 2$. Although d_3 could also be considered as the weight in objective functions such as (3.1) or (3.2), it is sufficient to explain the role of different distance functions in the linear programming approach using d_1 or d_4.

In particular, we can use the following objective function for $p = \infty$ under the same constraints (3.3), (3.4) and (3.5).

$$\phi_3 = \sum_{B_k \in B} d_4(A : B_k) P(B_k) \qquad (3.6)$$

It is evident that it would be more convenient to use ϕ_3 than ϕ_1 or ϕ_2 since d_4 has the simpler form and d_4 would cluster possible samples into several groups in which the number of groups is much

smaller than in using d_1 or d_2. It also makes the selection probability of each sample $p(B_k)$ easy to compute under the linear programming approach.

Third, selection probabilities of possible samples are obtained by minimizing each objective function subject to the constraints (3.3), (3.4) and (3.5).

Finally, we randomly select one sample from the sampling plan which is the solution, using the method of cumulative sums (probability proportionate to size).

Objective functions ϕ_1 ϕ_2 or ϕ_3 would reflect the "closeness" of each sample B_k to the original tabular array A, and this linear programming approach would maximize the selection probabilities of optimal samples under the given constraints.

This algorithm may be a more direct approach than the method using transportation theory or flow in the network developed by Causey et al. (1985) and Huang and Lin (1998) respectively, since the controlled selection problem is specified directly as a linear programming problem. Also, it would reflect a more straightforward weight than the method proposed by Sitter and Skinner (1994), which uses marginal constraints in tabular array A.

We have developed public use SAS-based software for our linear programming approach to two-way controlled selection problems. Possible samples B_k, $k = 1, \cdots, L$ for original tabular array A are automatically produced. The objective functions such as ϕ_1, ϕ_2 and ϕ_3 can be used by simple options in the program.

In this software, a two-phase revised simplex method, implemented using SAS/OR LP Procedure, is employed to solve the controlled selection problem. A unique optimal solution set is obtained when the objective function is minimized under the given constraints (3.3), (3.4) and (3.5) through phase 1 and 2 of LP program.

In using the algorithm, there are no restrictions on the problem size, i.e., the number of all possible samples that can be considered for the solution. The problem size and the number of constrains that can be solved would depend on the memory capacity and the available disk space of the computer. A public use version of the SOCSLP (Software for Optimal Controlled Selection Linear Programming) software may be downloaded from the URL: http://www.isr.umich.edu/src/smp/socs.

4. Examples

We apply our linear programming method to solve three two-way controlled selection problems previously described in the literature. These problems are divided into two cases: integer margins and non-integer margins. The results from several methods are compared using as criteria the assigned selection probabilities of optimal samples and unfavorable samples.

Example 1: Jessen (1970)

We first examine the simple example given in the table of Jessen (1970), page 778. This example is a 3×3 controlled selection problem with integer margins and a total sample of size n = 6. There are six possible samples, B_k, that satisfy the marginal constraints imposed by the problem.

Table 1 presents the selection probability of each sample, resulting from Jessen's (1970) method 2 and 3 solutions, the Sitter and Skinner (1994) method, and our method using ϕ_1 and ϕ_3. For Sitter and Skinner's method, we used the SAS/OR LP Procedure to find the solution.

Table 1 shows that the solutions from all methods except Jessen's method 3 yield the same result for this simple controlled selection problem. In the common solutions, the optimal sample receives .5 probability of selection and the unfavorable sample has .2 chance of selection.

Example 2: Jessen (1978)

Consider another example that is a 4×4 controlled selection problem (Jessen 1978, p. 375) which has 30 possible samples. This problem also has integer margins. Jessen used a simple but effective probability lattice sampling method to find a solution set for this controlled selection problem. A comparison of the proposed method using ϕ_1 or ϕ_3 with those of Jessen (1978) and Sitter and Skinner (1994) is shown in Table 2. (Only samples that receive a non-zero probability for one or more methods are shown.)

Sitter and Skinner's method provides the lowest probability of 0.4 to optimal samples, whereas the suggested method using ϕ_1 or ϕ_3 allocates the highest probability of 0.8 to the samples. For unfavorable samples, the suggested methods give the probability of 0.2. Jessen's method and the Sitter and Skinner solution give zero probabilities to the unfavorable sample pattern.

Example 3: Causey et al. (1985)

Causey et al. (1985) used an 8×3 controlled selection problem to explain their transportation theory algorithm. Each row can be regarded as a stratum and each column considered a classification variable. This problem could produce 141 possible samples considerably larger than in the above two problems. Table 3 presents a comparison of the suggested method using ϕ_1 or ϕ_3 with Causey et al. (1985), Huang and Lin (1998), Sitter and Skinner (1994). (See the last page)

Table 1. Sampling Designs And Comparison For Example 1

Sample B_k	$p(B_k)$				
	JS2	**JS3**	**S - S**	**K- H- S(1)**	**K - H – S(2)**
0 1 1 1 0 1# 1 1 0	0.2	0.1	0.2	0.2	0.2
0 1 1 1 1 0 1 0 1	0.0	0.1	0.0	0.0	0.0
1 0 1 0 1 1 1 1 0	0.0	0.1	0.0	0.0	0.0
1 0 1 1 1 0 * 0 1 1	0.5	0.4	0.5	0.5	0.5
1 1 0 0 1 1 1 0 1	0.3	0.2	0.3	0.3	0.3
1 1 0 1 0 1 0 1 1	0.0	0.1	0.0	0.0	0.0
$\sum_1$	0.5	0.4	0.5	0.5	0.5
$\sum_2$	0.2	0.1	0.2	0.2	0.2

Note. * : optimal sample, # : unfavorable sample
JS2: Jessen's method 2, JS3: Jessen's method 3
S – S : Sitter and Skinner(1994)'s method
K – H – S(1): Proposed method using ϕ_1
K – H – S(2): Proposed method using ϕ_3
$\sum_1$: Sum of selection probabilities of optimal samples
$\sum_2$: Sum of selection probabilities of unfavorable samples

Table 2. Sampling Designs And Comparison For Example 2

Sample B_k	$p(B_k)$			
	JS	**S – S**	**K – H – S(1)**	**K – H –S(2)**
0 0 1 1 0 1 0 1 1 0 1 0 1 1 0 0	0.2	0.1	0.0	0.0
0 0 1 1 0 1 0 1# 1 1 0 0 1 0 1 0	0.0	0.0	0.2	0.2
0 0 1 1 1 0 0 1 0 1 1 0 1 1 0 0	0.0	0.1	0.0	0.0
0 0 1 1 1 1 0 0 * 0 0 1 1 1 1 0 0	0.2	0.1	0.2	0.2
0 0 1 1 1 1 0 0 0 1 0 1 1 0 1 0	0.0	0.1	0.0	0.0
0 1 1 0 0 0 1 1 1 0 0 1 1 1 0 0	0.0	0.1	0.0	0.0
0 1 1 0 1 0 0 1 * 0 0 1 1 1 1 0 0	0.0	0.0	0.2	0.2
0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0	0.2	0.0	0.0	0.0
0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0	0.0	0.1	0.0	0.0
0 1 1 0 1 0 1 0 * 1 0 0 1 0 1 0 1	0.4	0.3	0.4	0.4
0 1 1 0 1 1 0 0 0 0 1 1 1 0 0 1	0.0	0.1	0.0	0.0
$\sum_1$	0.6	0.4	0.8	0.8
$\sum_2$	0	0	0.2	0.2

Note. JS : Jessen's (1978) probability lattice sampling method. See notes for Table 1

We note that all these methods provide different solutions, although the four methods except Sitter and Skinner's method offer the same sum of selection probabilities of optimal samples. The proposed methods distribute the total probability of 0.4 to two optimal samples, whereas the first two methods just allocate the probability to only one optimal sample. Sitter and Skinner's method appears to be less effective for this problem. For this problem, we would prefer the method using ϕ_3 to one using ϕ_1 because the former gives the probability of 0.08 to the unfavorable samples, while the latter gives a higher probability of 0.2 to the undesired sample problem.

In conclusion, through the above examples, the effectiveness of the method using ϕ_1 or ϕ_3 is measured by the ability to maximize the selection probabilities of optimal samples. In particular, not only does the method using ϕ_3 maximize the possibility to be selected for optimal samples, it reduces the selection probabilities of unfavorable samples as the controlled selection problems have larger numbers of possible samples. Though we did not show the results of the comparisons between the suggested method using ϕ_2 and other methods, they are less effective than the method using ϕ_3 in reducing the selection probabilities of unfavorable samples.

5. Conclusion

In this paper, we propose using a linear programming approach with metric distance functions as a weight for each sample to find optimizing solution sets in two-way controlled selection problems. We have

implemented this procedure in a new SAS-based software.

As shown in the above examples, this method would offer solution sets not only to maximize the selection probabilities of optimal samples, but also to minimize the probability of choosing unfavorable samples in large controlled selection problems.

Based on the results for the two-way controlled selection problem, we expect that the suggested method would also contribute to controlled selection problems with three dimensions. We are currently working on an extension to those problems and to develop a more effective algorithm for large controlled selection problems.

References

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985), "Applications of Transportation Theory to Statistical Problems," *Journal of the American Statistical Association*, 80, 903-909.

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

Goodman, R., and Kish, L. (1950), "Controlled Selection–A Technique in Probability Sampling," *Journal of the American Statistical Association*, 45, 350-372.

Groves, R. M. and Hess, I. (1975), "An Algorithm for Controlled Selection," Ch. VII in Hess, I., Ridel, D. C., and Fitzpatrick, T. B. (eds.), *Probability Sampling of Hospitals and Patients*, 2nd edition, Ann Arbor: Health Administration Press.

Hess, I., Ridel, D. C., and Fitzpatrick, T. B. (1975), *Probability Sampling of Hospitals and Patients*, 2nd edition, Ann Arbor: Health Administration Press.

Huang, H. C. and Lin, T. K. (1998), "On the Two-Dimensional Controlled Selection Problem," internal manuscript, Department of Statistics and Applied Probability, Singapore: National University of Singapore.

Jessen, R. J. (1970), "Probability Sampling With Marginal Constraints," *Journal of the American Statistical Association*, 65, 776-795.

Jessen, R. J. (1978), *Statistical Survey Techniques*, New York: John Wiley and Sons.

Rao, J. N. K., and Nigam, A. K. (1990), "Optimal Controlled Sampling Design," *Biometrika*, 77, 807-814.

Rao, J. N. K., and Nigam, A. K. (1992), "'Optimal' Controlled Sampling: A Unified Approach," *International Statistical Review*, 60, 89-98.

SAS/OR (2001), *User's Guide: Mathematical Programming*, Version 8, Cary, NC: SAS Institute Inc.

Sitter, R. R. and Skinner, C. J. (1994), "Multi-Way Stratification by Linear Programming," *Survey Methodology*, 20, 1, 65-73.

Tiwari, N., and Nigam, A. K. (1998), "On Two-Dimensional Optimal Controlled Selection," *Journal of Statistical Planning and Inference*, 69, 89-100.

Table 3. Sampling Designs And Comparison For Example 3

Sample B_k	$p(B_k)$	Sample B_k	$p(B_k)$	Sample B_k	$p(B_k)$
0 2 0 1 0 1 0 0 0 2 0 0 1 0 0 0 1 0 0 0 1 0 0 1	0.20 [a]	0 2 0 1 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0	0.00	0 2 0 2 0 1 0 0 0 1 1 0 1 0 0 0 0 0 0 1 0 0 0 1	0.00
	0.00 [b]		0.00		0.00
	0.00 [c]		0.05		0.00
	0.00 [d]		0.10		0.00
	0.00 [e]		0.04		0.02
0 2 0 1 0 1 0 0 0 2 0 0 1 0 1 0 0 1 0 1 0 0 0 0	0.00	0 2 0 1 0 1 1 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 1	0.00	0 2 0 2 0 1 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1	0.00
	0.00		0.00		0.00
	0.10		0.05		0.15
	0.00		0.00		0.00
	0.10		0.04		0.06
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 0 0 0 1 0 0 0 1	0.00	0 2 0 2 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 1	0.00	1 2 0 1 0 1 0 0 0 1 0 0* 1 1 0 0 1 0 0 0 1 0 0 0	0.00
	0.20		0.00		0.00
	0.00		0.00		0.10
	0.00		0.10		0.10
	0.00		0.00		0.16
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 0 1 0 0 1 0 0 0	0.00	0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0	0.00	1 2 0 1 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 0 1 0 0 0	0.00
	0.00		0.20		0.00
	0.00		0.00		0.05
	0.10		0.00		0.00
	0.00		0.00		0.00
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0.00	0 2 0 2 0 1 0 0 0 1 0 1# 1 0 1 0 0 0 0 1 0 0 0 0	0.00	1 2 0 1 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0	0.00
	0.00		0.00		0.00
	0.10		0.00		0.10
	0.10		0.10		0.00
	0.10		0.00		0.00
0 2 0 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 0 0	0.20	0 2 0 2 0 1 0 0 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 0	0.00	1 2 0 1 0 1 0 0 0 1 1 0* 1 1 0 0 0 1 0 0 0 0 0 0	0.40
	0.20		0.00		0.40
	0.00		0.05		0.15
	0.00		0.00		0.30
	0.00		0.06		0.24
0 2 0 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0.00	0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 0 1 0 0 0	0.00		
	0.00		0.00		
	0.10		0.00		
	0.00		0.00		
	0.04		0.06		
0 2 0 1 0 1 1 0 0 1 0 1# 1 0 0 0 0 0 0 1 0 0 0 1	0.00	0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0	0.20		
	0.00		0.00		
	0.00		0.00		
	0.10		0.00		
	0.08		0.00		
$\sum_1$	0.40 [a]		$\sum_2$	0.00 [a]	
	0.40 [b]			0.00 [b]	
	0.25 [c]			0.00 [c]	
	0.40 [d]			0.20 [d]	
	0.40 [e]			0.08 [e]	

Note a : Causey, Cox and Ernst (1985), b : Huang and Lin (1998)
c : Sitter and Skinner (1994) , d : Proposed method using ϕ_1
e: Proposed method using ϕ_3
See notes for Table 1.