Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Optimal solutions in controlled selection problems with two-way stratification

by Sun Woong Kim, Steven G. Heeringa and Peter W. Solenberger

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 243-263
Statistics Canada, Catalogue No. 12-001-X

Optimal solutions in controlled selection problems with two-way stratification

Sun Woong Kim, Steven G. Heeringa and Peter W. Solenberger[1]

Abstract

When considering sample stratification by several variables, we often face the case where the expected number of sample units to be selected in each stratum is very small and the total number of units to be selected is smaller than the total number of strata. These stratified sample designs are specifically represented by the tabular arrays with real numbers, called controlled selection problems, and are beyond the reach of conventional methods of allocation. Many algorithms for solving these problems have been studied over about 60 years beginning with Goodman and Kish (1950). Those developed more recently are especially computer intensive and always find the solutions. However, there still remains the unanswered question: In what sense are the solutions to a controlled selection problem obtained from those algorithms optimal? We introduce the general concept of optimal solutions, and propose a new controlled selection algorithm based on typical distance functions to achieve solutions. This algorithm can be easily performed by a new SAS-based software. This study focuses on two-way stratification designs. The controlled selection solutions from the new algorithm are compared with those from existing algorithms using several examples. The new algorithm successfully obtains robust solutions to two-way controlled selection problems that meet the optimality criteria.

Key Words: Cell expectation; Probability sampling; Distance function; Optimum array; Linear programming problem; Simplex method.

1 Introduction

In the term, "Controlled Selection (or Controlled Sampling)", "control" has a broad meaning. The pioneering paper of Goodman and Kish (1950, page 351) defined controlled selection as "...any process of selection in which, while maintaining the assigned probability for each unit, the probabilities of selection for some or all preferred combinations of n out of N units are larger than in stratified random sampling".

The focus in this paper is upon **controls** required in deciding the number of units (e.g., primary sampling units (PSUs)) allocated to each stratum cell in a **two-way stratification design**, where the total number of units to be selected is smaller than the number of strata cells or the expected number of units to be selected from each stratum cell is very small. This assumes that given precision and cost constraints, simply reducing the number of strata cells or increasing the number of the sampled units is not appropriate for the design.

Here **controlled selection** refers to the following two-stage procedure. First, the **controlled selection problem** represented by a tabular array with real numbers formed by the two-way stratification design is solved according to a specified algorithm (or technique). The solution to the problem is a set of feasible arrays with nonnegative integer sample allocation to the cells of each array and probabilities of selection corresponding to each array. Second, a random selection of one of the solution arrays is made using the assigned probabilities. The integer number appearing in each cell of the selected solution array then serves as

1. Sun Woong Kim, Director, Survey & Health Policy Research Center, Professor, Department of Statistics, Dongguk University, 26, 3-Ga, Pil-Dong, Jung-Gu Seoul, South Korea 100-715. E-mail: sunwk@dongguk.edu; Steven G. Heeringa, Senior Research Scientist, Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106. E-mail: sheering@isr.umich.edu; Peter W. Solenberger, Applications Programmer Analyst Lead, Survey Research Center, Institute for Social Research, University of Michigan. E-mail: pws@isr.umich.edu.

the number of sample units to be allocated to that cell of the two-way stratification. The key to the controlled selection is the **algorithm** that defines a set of solution arrays that achieve the **controls** to solve the problem.

Many controlled selection techniques have been developed since Goodman and Kish (1950) first described the application of controlled selection to a specific problem of choosing 17 PSU's to represent the North Central States of the United States. Bryant, Hartley and Jessen (1960) proposed a simple method which was applicable in a limited number of sample situations. Raghunandanan and Bryant (1971) generalized their method and Chernick and Wright (1983) suggested an alternative. Jessen (1970) proposed two methods called "method 2" and "method 3", both quite complicated to implement and sometimes failing to provide a solution. Jessen (1978, chapter 11) introduced a simpler algorithm for solving controlled selection problems.

Hess, Riedel and Fitzpatrick (1975) gave a detailed explanation of how to use controlled selection in order to select a representative sample of Michigan's hospitals. Groves and Hess (1975) first suggested a formal computer algorithm for obtaining solutions to controlled selection problems with two-and three-way stratification. Heeringa and Hess (1983) reported the response to Roe Goodman's question: How does a computer solution of highly controlled selection compare with a manual solution? The answer was "For the same sample design, computer generated controlled selection often leads to slightly higher variances than does manual controlled selection; but since the differences in precision are small and manual controlled selection is laborious, computer generated controlled selection is preferred." Lin (1992) improved the algorithm of Groves and Hess (1975) and the software called "PCCONSEL" for their algorithm was presented by Heeringa (1998). Huang and Lin (1998) proposed a more efficient algorithm, which imposes additional constraints in the controlled selection problem with two-way stratification and uses any standard network flow computer package. Hess and Heeringa (2002) summarized investigations on controlled selection over 40 years that have been made at the Survey Research Center, University of Michigan.

Taking a different approach, Causey, Cox and Ernst (1985) proposed an algorithm that applied a transportation model to controlled selection problems with two-way stratification, based on the theory originally suggested in a previous paper of Cox and Ernst (1982). Winkler (2001) developed an integer programming algorithm quite similar to that of Causey et al. (1985). Deville and Tillé (2004) suggested an algorithm called the Cube method.

Following Rao and Nigam (1990, 1992), Sitter and Skinner (1994) applied a linear programming (LP) approach to solve controlled selection problems. Later, Tiwari and Nigam (1998) proposed an LP method that reduces the probabilities of selecting non-preferred samples.

In summary, many different algorithms for controlled selection have been investigated and described in the literature. Those most recently developed are especially **computer-intensive**, since they are highly dependent on available software and high speed computers. However, in spite of this evolution in the algorithms over about 60 years, a question still remains: In what sense are the solutions to a controlled selection problem obtained from those algorithms **optimal**?

In this paper, we define in Section 2 the two-way controlled selection problem and revisit several problems of this type that have appeared in the historical literature. In Section 3, we present the desirable constraints. In Section 4, we introduce our concept of **optimal solutions** to controlled selection problems. In Section 5, we describe the weaknesses in the previous algorithms. In Section 6, we suggest a new algorithm using the LP approach for achieving **optimal solutions** and a new publicly available **software** for implementing the new controlled selection algorithm is presented in Section 7. In Section 8, to show the

robustness of the new algorithm, it is applied to several example controlled selection problems and the results are compared to those obtained using existing algorithms. We conclude in Section 9.

2 Controlled selection problems

In order to select a sample of n units, consider a two-way stratification design classifying a population of N units by two criteria with R and C categories, respectively. The controlled selection problem under two-way stratification is defined by the $R \times C$ tabular array A, which consists of RC cells that have nonnegative real numbers a_{ij}, called the **cell expectations**, representing the expected number of units to be drawn in each cell ij. The standard two-way controlled selection problem is described as in Table 2.1.

Table 2.1
$R \times C$ **Controlled selection problem**

a_{11}	a_{12}	. . . . .	a_{1C}	$a_{1.}$
a_{21}	a_{22}	. . . . .	a_{2C}	$a_{2.}$
.	.		.	.
.	.		.	.
.	.	. . a_{ij} . .	.	$a_{i.}$
.	.		.	.
.	.		.	.
a_{R1}	a_{R2}	. . . . .	a_{RC}	$a_{R.}$
$a_{.1}$	$a_{.2}$	. . $a_{.j}$ . .	$a_{.C}$	$a_{..}(=n)$

The **marginal expectations** $a_{i.}$ and $a_{.j}$ denote the sum of cell expectations in each row category i and each column category j. Hence $a_{..}$ denotes the sum of all cell expectations and equals the total sample size n.

Although Table 2.1 takes a simple two-way tabular form, it should be noted that typically $n < RC$, and furthermore a_{ij} can be very small (e.g., often less than 1). In this case deciding how to allocate n units to cells, that is, how to obtain an $R \times C$ array with cells rounded to a nonnegative integer for each a_{ij}, requires an algorithm to solve the problem.

A variety of controlled selection problems are used as examples in the literature. The first example of a controlled selection problem was the 17×4 array, described by Goodman and Kish (1950, page 356), for allocating 17 PSU's to 68 cells given by 17 strata and 4 groups of North Central States in the United States. The array may be formed as follows. Let N_{ij} denote the number of population elements in each cell ij and let $N_{i.}$ denote the total number of population elements in each stratum. Then $a_{ij} = N_{ij} / N_{i.}$, where some N_{ij} are zero and $0 \le a_{ij} < 1$. All $a_{i.}$ equal the integer 1, whereas $a_{.j}$ are nonintegers sums of the a_{ij} in column j. The problem is therefore one of selecting one PSU per sample stratum (i dimension) and simultaneously controlling the distribution to state groups (j dimension). A total of $n = 17$ PSUs will be selected.

The following paragraphs describe four additional problems found in the literature that will be used in the discussion and comparative evaluations presented in this paper.

Problem 2.1: Jessen (1970)

A 3×3 problem involving two stratifying variables is given by Jessen (1970, page 779). Each cell ij corresponds to one PSU and $N=9$. A sample of size $n=6$ is drawn. $a_{ij}=nX_{ij}/X$, where X_{ij} is a "measure of size" for the PSU in cell ij and $X=\sum_{i=1}^{R}\sum_{j=1}^{C}X_{ij}$. Note that in this problem, $0<a_{ij}<1$, and both $a_{i.}$ and $a_{.j}$ are equal to 2.

Problem 2.2: Jessen (1978)

An extended 4×4 version of Problem 2.1 comes from Jessen (1978, page 375). In this problem, $N=16$ and $n=8$. As in Problem 2.1, both $a_{i.}$ and $a_{.j}$ are equal to 2, but $0\le a_{ij}\le 1$.

Problem 2.3: Causey et al. (1985)

Causey et al. (1985, page 906) describe an 8×3 two-way stratification problem designed to select 10 PSU's, that is, $n=10$. Let $X_{ijq_{ij}}\left(q_{ij}=1,\ldots,r_{ij}\right)$ be some measure of size for the PSU q_{ij} in cell ij. Here $a_{ij}=n\,X_{ijq}/X_q$, where $X_{ijq}=\sum_{q_{ij}=1}^{r_{ij}}X_{ijq_{ij}}$ and $X_q=\sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{q_{ij}=1}^{r_{ij}}X_{ijq_{ij}}$. Note that in this problem, $0\le a_{ij}\le 2$, and most $a_{i.}$ and $a_{.j}$ are noninteger values.

Problem 2.4: Winkler (2001)

Winkler (2001) provides the 5×5 controlled selection problem with two stratifying variables shown in Table 2.2.

The objective in solving this problem is to select $n=37$ sample units from the population of $N=1,251$. The problem definition begins with a 5×5 array with cell population sizes N_{ij}, where some N_{ij} are quite small. The marginal row and column expectations, $a_{i.}$ and $a_{.j}$, are integer-valued and are predetermined using the prior information on precision (e.g., coefficients of variation).

Table 2.2
5×5 **Controlled selection problem**

2.000	2.483	1.052	0.103	0.362	6
2.182	1.061	1.101	1.046	0.610	6
0.000	1.614	1.914	2.200	1.272	7
0.860	0.377	0.930	2.840	2.993	8
0.958	0.465	2.003	1.811	4.763	10
6	**6**	**7**	**8**	**10**	37

Source: Table 4, Appendix, Winkler (2001). Reproduced with permission.

The cell expectations, a_{ij}, are obtained by applying the generalized iterative fitting procedure (GIFP) of Dykstra (1985a, 1985b) and Winkler (1990) to the initial array. The GIFP is used to ensure that $a_{ij} < N_{ij}$ for the cells with small N_{ij}, when $a_{i.}$ and $a_{.j}$ are given. Note that in the Table 2.2, the a_{ij} are given to 3 decimal places, and $0 \le a_{ij} < 5$.

The common characteristic shared by these controlled selection problems is that, as mentioned above, the total number of selected units is smaller than the number of cells (except for Problem 2.4, where $n = 37 > RC = 25$) and many a_{ij} are less than 1. The algorithms used to solve these problems must enforce some strict constraints described in next section. As described in Section 4, the solution to a controlled selection problem obtained by any algorithm is a set of some $R \times C$ arrays and probabilities of selection corresponding to each array.

3 Desirable constraints

Each controlled selection problem of the form illustrated in Table 2.1 has many possible integer solutions. Let B_k denote one such solution, whose internal entries b_{ijk} are the replacement of the real numbers a_{ij} in the controlled selection problem A by the adjacent nonnegative integers. The entry, b_{ijk}, equals either $\lfloor a_{ij} \rfloor$ or $\lfloor a_{ij} \rfloor + 1$, where [ ] is the greatest integer function. If a_{ij} is a nonnegative integer, $b_{ijk} = a_{ij}$ for all k. The same rule is applied to the marginal expectations. As noted by Jessen (1970) and Causey et al. (1985), we primarily pay attention to B_k that simultaneously satisfy the following **constraints** for all i and j:

$$b_{ijk} \ge 0 \tag{3.1}$$

$$\left| b_{ijk} - a_{ij} \right| < 1 \tag{3.2}$$

$$\left| b_{i.k} - a_{i.} \right| < 1 \text{ and} \tag{3.3}$$

$$\left| b_{.jk} - a_{.j} \right| < 1, \tag{3.4}$$

where $b_{i.k} = \sum_{j=1}^{C} b_{ijk}$ equals either $\lfloor a_{i.} \rfloor$ or $\lfloor a_{i.} \rfloor + 1$, $b_{.jk} = \sum_{i=1}^{R} b_{ijk}$ equals either $\lfloor a_{.j} \rfloor$ or $\lfloor a_{.j} \rfloor + 1$, $\sum_{i=1}^{R} b_{i.k} = a_{..}$ and $\sum_{j=1}^{C} b_{.jk} = a_{..}$.

Consider the set of **all possible arrays**, $\mathfrak{B} = \{B_k, k = 1, \ldots, L\}$, satisfying (3.1) - (3.4). Since a_{ij} is the expectation of the sample allocation to each cell in A, the following **constraints** (3.5) and (3.6) on b_{ijk} in $B_k (\in \mathfrak{B})$ are especially important.

$$E\left(b_{ijk} \middle| i, j\right) = \sum_{B_k \in \mathfrak{B}} b_{ijk}\, p\left(B_k\right) = a_{ij}, \quad i = 1, \ldots, R, \text{ and } \quad j = 1, \ldots, C \tag{3.5}$$

and

$$\sum_{B_k \in \mathfrak{B}} p\left(B_k\right) = 1, \tag{3.6}$$

where $p\left(B_k\right)$, which depends on a specified algorithm for solving the controlled selection problem, is the selection probability of the array B_k and $p\left(B_k\right) \ge 0$.

Note that (3.5) and (3.6) will define a rigorous **probability sampling** method when randomly selecting any array in $\mathfrak{B}$. Also, note that since $\sum_{i=1}^{R}\sum_{j=1}^{C} E\left(b_{ijk}\middle|i,j\right) = a_{..}\sum_{B_k\in\mathfrak{B}} p\left(B_k\right) = a_{..}$, (3.5) implies (3.6) for any controlled selection problem such as those described in Problems 2.1 through 2.4. In addition, as an illustration, when applied to Problem 2.3, where $a_{ij} = n X_{ijq}/X_q$, (3.5) yields

$$E\left(b_{ijk}/X_{ijq}\middle|i,j\right) = a_{ij}/X_{ijq} = n/X_q, \tag{3.7}$$

which indicates the equal allocation for each cell.

4 Optimal solutions

Given the set of L possible arrays in $\mathfrak{B}$, consider the subset $\mathfrak{B}'\ (\subseteq \mathfrak{B})$ where

$$p\left(B_k\right) > 0.$$

A **solution set** to a controlled selection problem A denoted as

$$\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$$

is the set of the arrays that have the required positive selection probabilities $\left(p\left(B_k\right) > 0\right)$. This solution set, or simply a "solution" to the controlled selection problem, is usually obtained by an algorithm to control the constraints in (3.1) through (3.6). As described in the introduction, since Goodman and Kish (1950), many algorithms for obtaining solutions to controlled selection problems have been developed.

Until Groves and Hess (1975) suggested a computer algorithm, most solutions were manually obtained in a process that resembled solving a mathematical puzzle. Furthermore, for most problems it is possible that there is more than one solution set that meets the constraints. Since the 1980s, the computer-intensive controlled selection algorithms using transportation theory, network flow, integer programming, and LP have been developed. These algorithms may depend on highly specialized software or may be programmed to run in major software systems.

However, previous solutions ranging from manual to computer-intensive algorithms have rarely been compared empirically using a standard set of performance criteria. Therefore, we begin with the description of a concept called **optimal solution sets**, or more simply, **optimal solutions**.

The controlled selection problem A is only one array, but there may be many possible arrays in $\mathfrak{B}$. Also, only one array B_k from any solution to A is randomly chosen by $p\left(B_k\right)$ as the basis for choosing the stratified sample. In general then, we might define an optimal solution as that satisfying the following **requirements** (R1 and R2):

R1. The solution is obtained based on appropriate and objective measurements of the **closeness** between A and every single array B_k in $\mathfrak{B}$.

R2. The solution, as much as possible, maximizes the probabilities of selection over the **arrays nearest** to A under such measurements as referenced in R1.

The remainder of this section will address how to specify R1 and R2 for optimal solutions. First, in order to define **closeness** in R1, a real number $d(B_k : A)$ representing the distance between A and B_k, can be considered, where d is a distance function that satisfies the following axioms:

(i) $d(B_k, A) > 0$ if $B_k \neq A$; $d(A, A) = 0$;

(ii) $d(B_k, A) = d(A, B_k)$;

(iii) $d(B_k, A) \leq d(B_k, B_k') + d(B_k', A)$ for any $B_k' \in \mathfrak{B}$.

Axiom (iii) is termed the **triangle inequality axiom**. Distance functions satisfying (i), (ii), and (iii) can be defined by using the two-ordered RC-tuples $(a_{11}, a_{12}, \cdots, a_{RC})$ and $(b_{11k}, b_{12k}, \cdots, b_{RCk})$ for A and B_k. We first define the ordinary or **Euclidean distance** (**2-norm distance**):

$$d_2(B_k, A) = \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \left(b_{ijk} - a_{ij} \right)^2 \right]^{\frac{1}{2}}, \quad k = 1, \ldots, L. \tag{4.1}$$

This function is probably the most familiar measure to define the distance between B_k and A.

Also, we can define the function called the **Chebyshev distance** (**infinite norm distance**):

$$d_\infty(B_k, A) = \max \left\{ \left| b_{ijk} - a_{ij} \right| : i = 1, \ldots, R,\ j = 1, \ldots, C \right\}, \quad k = 1, \ldots, L. \tag{4.2}$$

These distance functions give rise to distinct distance spaces. Owing to (3.2), for any B_k, the following holds.

$$0 \leq d_2(B_k, A) < (RC)^{1/2} \tag{4.3}$$

and

$$0 \leq d_\infty(B_k, A) < 1. \tag{4.4}$$

For instance, for the 3×3 array in Problem 2.1 and the 8×3 array in Problem 2.3, $0 < d_2(B_k, A) < 3$ and $0 < d_2(B_k, A) < 4.9$, respectively.

Second, as mentioned in R2, regarding the **arrays nearest** to A under such measurements described in R1, consider the set of arrays in $\mathfrak{B}$ having the minimum d_2 or d_∞ value from A. Let $\mathfrak{B}_2 (\subseteq \mathfrak{B}')$ be the set of the arrays having the minimum d_2 value from A and $\mathfrak{B}_\infty (\subseteq \mathfrak{B}')$ be the set of the arrays having the minimum d_∞ value from A.

Assuming that all possible arrays in $\mathfrak{B}$ are known, we define **optimum arrays** as follows.

Definition. The arrays in $\mathfrak{B}_2 \cup \mathfrak{B}_\infty$ are called the **optimum arrays**.

Note that in the new algorithm for controlled selection to be described in Section 6, d_2 or d_∞ are chosen based on preference. We avoid defining the intersection of $\mathfrak{B}_2$ and $\mathfrak{B}_\infty$ as the optimum arrays because this may exclude the other arrays not in $\mathfrak{B}_2 \cap \mathfrak{B}_\infty$ with the same minimum d_2 (d_∞) value. We illustrate below that there may exist a very small number of optimum arrays relative to the number of all possible arrays in $\mathfrak{B}$ for any A. The details of how to find all possible arrays will be described in Section 6 and Section 7.

Illustrations.

For Problem 2.1 through Problem 2.4, it is noted that $\mathfrak{B}_2 \subseteq \mathfrak{B}_\infty$. Thus, it is possible to use d_∞ only in illustrating the optimum arrays.

1. For Problem 2.1, there are six possible arrays satisfying (3.1), (3.2), (3.3), and (3.4). That is, $\mathfrak{B} = \{B_k, k = 1,\ldots,6\}$, as given in Table 4.1. There exists only one optimum array, B_2, with the minimum value of $d_\infty = 0.5$.

Table 4.1
3×3 **Controlled selection problem, optimum array with** $d_\infty = 0.5$ **and the other arrays**

A	B_1	B_2	B_3	B_4	B_5	B_6
0.8 0.5 0.7	0 1 1	1 0 1	1 1 0	0 1 1	1 0 1	1 1 0
0.7 0.8 0.5	1 0 1	1 1 0	0 1 1	1 1 0	0 1 1	1 0 1
0.5 0.7 0.8	1 1 0	0 1 1	1 0 1	1 0 1	1 1 0	0 1 1
d_∞	0.8	0.5	0.7	0.8	0.8	0.8

2. Problem 2.2 has 30 possible arrays, and there are three optimum arrays, shown in Table 4.2

Table 4.2
4×4 **Optimum arrays with** $d_\infty = 0.6$

0 0 1 1	0 1 1 0	0 1 1 0
1 1 0 0	1 0 0 1	1 0 1 0
0 0 1 1	0 0 1 1	1 0 0 1
1 1 0 0	1 1 0 0	0 1 0 1

3. Problem 2.3 has 141 possible arrays. There are six optimum arrays, where each array has the same $d_\infty = 0.6$. One of them is given in Table 4.3.

Table 4.3
One of six optimum arrays with $d_\infty = 0.6$

1	2	0
1	0	1
0	0	0
1	1	0
1	1	0
0	0	1
0	0	0
0	0	0

4. There are 159 possible arrays for Problem 2.4, and there is only one optimum array given in Table 4.4.

Table 4.4
5×5 **Optimum array with** $d_\infty = 0.517$

2	3	1	0	0
2	1	1	1	1
0	2	2	2	1
1	0	1	3	3
1	0	2	2	5

Accordingly, based on the definition of optimum arrays as well as d_2 and d_∞ satisfying the axioms (i), (ii), and (iii), we suggest the following **specifications** (S1 and S2) of R1 and R2 of optimal solutions:

S1. The solution is based on the values of the distance $d_2\left(d_\infty\right)$ between A and every single array B_k in $\mathfrak{B}$.

S2. The solution maximizes the probabilities of selection of optimum arrays.

S1 and S2 will be the rudiments of a new algorithm presented in Section 6, and in the next section we turn into the discussion on the previous algorithms from the viewpoint of optimal solutions.

5 Non-optimal properties of existing methods

As described in Section 4, the algorithms for controlled selection may be divided into two parts, manual algorithms before 1980s and computer-intensive algorithms since then. For large controlled selection problems with many cells, the latter class of algorithms may be preferred. But when the problem is small, the former can be easily used without the complexity of the latter. Therefore we would not say that the former is always inferior to the latter. More objective criteria for comparing them would be necessary, and the optimal solution may be adopted as one of the better criteria to compare their strengths or weaknesses.

As discussed by Jessen (1978, pages 375-376), the algorithms of Jessen (1970) aim to minimize the number of arrays in a solution set $\mathfrak{B}'$, and the algorithm of Jessen (1978) quite easily achieves that purpose relative to those of Jessen (1970). Thus his algorithms pursue "simplicity" in formulating a solution rather than an optimal solution.

The algorithm of Causey et al. (1985) may give a "partially" optimal solution. Other than the original problem, A, it sequentially creates a small number of new controlled selection problems, and then as a solution it finds only one array $B_k\,(\in\mathfrak{B})$ to be **nearest** to each problem, starting with A. Each problem is regarded as the transportation problem of Cox and Ernst (1982), which is formed by the objective function mimicking the behavior of

$$\sum_{i=1}^{R}\sum_{j=1}^{C}\left|b_{ijk}-a_{ij}\right|^p,\ k=1,\ldots,L,\ 1\le p<\infty. \tag{5.1}$$

Note that since function (5.1) violates the triangle inequality axiom (iii), it is not a distance function. It needs the inclusion of the p-th root to be a distance function. Also, each $p(B_k)$ is calculated by a simple formula. In view of the optimality requirements given by R1 and R2 the Causey et al. algorithm has the following weaknesses: 1) Since other controlled selection problems in addition to the original problem A are involved, it is difficult to obtain the solution consistently based on the **closeness** between the unique A and every individual B_k in $\mathfrak{B}$; 2) The maximization of the probabilities of selection for the **arrays nearest** to A is not guaranteed.

Winkler (2001) presented a modification of the method of Causey et al. (1985). Instead of using the transportation problem, he proposed integer linear programming, resulting in slight changes of the $p(B_k)$. Nevertheless, the Winkler (2001) algorithm is not free from the weaknesses of the Causey et al. (1985) method.

Adopting a network flow problem approach, the Huang and Lin (1998) algorithm imposes the additional subgroup constraints in A, raised by Goodman and Kish (1950). However, it does not attain objectives R1 and R2, just as in Causey et al. (1985) and Winkler (2001), since a new network, instead of a new controlled selection problem, is generated at every iteration, an arbitrary $B_k\,(\in \mathfrak{B})$ is obtained as a solution to the network, and $p(B_k)$ is calculated by a simple formula.

In contrast, the LP algorithms proposed by Sitter and Skinner (1994) and Tiwari and Nigam (1998) use all possible arrays in $\mathfrak{B}$. Note that finding all those arrays is an important issue, and that $p(B_k)$ for all possible arrays are simultaneously obtained by running the software for LP only once. The key idea underlying the algorithm of Sitter and Skinner (1994) is to use a "loss function" defined by

$$\sum_{i=1}^{R}\left(b_{i.k}-a_{i.}\right)^2+\sum_{j=1}^{C}\left(b_{.jk}-a_{.j}\right)^2. \tag{5.2}$$

In terms of R1 and R2, their algorithm has the following disadvantages: 1) The closeness between A and B_k is not well captured by loss function (5.2). This is because it is not a distance function that satisfies axiom (iii), as the marginal totals are used, instead of the cell entries; 2) Loss function (5.2) is irrelevant to the maximization of the probabilities of selection over the **arrays nearest** to A in Problems 2.1, 2.2, and 2.4, since it is always zero.

The LP method of Tiwari and Nigam (1998) can be used to reduce the selection probabilities of non-preferred arrays (e.g., arrays not containing the PSU corresponding to the cell $ij = 23$ in Problem 2.1), which are initially determined by the samplers. For controlled selection problems with integer margins and without considering the non-preferred arrays, their method will give the same solutions as that of Sitter and Skinner (1994).

The solutions from these previous methods will be compared with those from the proposed method in Section 6, on several examples in Section 8.

6 Suggested method

In this section, we present the details on an algorithm for achieving S1 and S2 of optimal solutions described in Section 4.

6.1 The algorithm

The algorithm has the following **characteristics**: 1) it finds a solution directly based on the values of the distance d_2 (d_∞) between the controlled selection problem A and each individual array B_k in $\mathfrak{B}$; 2) it is computer-intensive, but easily implemented by LP; 3) it is applicable to any type of controlled selection problem with two-way stratification.

The algorithm has five steps. They are as follows:

Step 1. Find the set of all possible arrays, $\mathfrak{B}$, satisfying (3.1) - (3.4) for a given controlled selection problem A. Specifically, if there are any noninteger marginal expectations in A, find all possible roundings of these marginal expectations by adjacent integers, which satisfy (3.3) and (3.4). Those rounded marginal integers will be $[a_{i.}]$ or $[a_{i.}]+1$ $\left([a_{.j}] \text{ or } [a_{.j}]+1\right)$, while the integer marginal expectations will remain, since $[a_{i.}] = a_{i.}$ $\left([a_{.j}] = a_{.j}\right)$. Next, find all possible arrays satisfying (3.1) and (3.2) under the rounded marginal integers and the other marginal integers.

Step 2. Choose either $d_2^*(B_k, A)$ or $d_\infty^*(B_k, A)$ (based on preference) and compute the chosen distance function for each $B_k\ (\in \mathfrak{B})$, where:

$$d_2^*(B_k, A) = d_2(B_k^*, A^*) = \left[\sum_{i=1}^{R}\sum_{j=1}^{C}\left(b_{ijk}^* - a_{ij}^*\right)^2\right]^{\frac{1}{2}} \tag{6.1}$$

$$d_\infty^*(B_k, A) = d_\infty(B_k^*, A^*) = \max\left\{\left|b_{ijk}^* - a_{ij}^*\right| : i = 1,\ldots,R,\ j = 1,\ldots,C\right\}. \tag{6.2}$$

Note that since each of the ij cells in the problem array, A, will receive a minimum allocation equal to $[a_{ij}]$ with certainty the distance functions need only consider the non-integer part of a_{ij}:

$$a_{ij}^* = a_{ij} - \left[a_{ij}\right], \tag{6.3}$$

and the integer difference (either 0 or 1) between the allocated sample size, b_{ijk}, for solution $k = 1,\ldots,L$ and the certainty count for the ij-th cell of A:

$$b_{ijk}^* = b_{ijk} - \left[a_{ij}\right]. \tag{6.4}$$

Step 3. According to the distance function chosen in Step 2, construct the following LP problem consisting of the minimization of the objective function (6.5) or (6.6), which is a linear form, with the linear constraints (6.7) and (6.8):

Minimize

$$OF_1 = \sum_{B_k \in \mathfrak{B}} d_2^*(B_k, A)\, p(B_k) \tag{6.5}$$

or

$$OF_2 = \sum_{B_k \in \mathfrak{B}} d_\infty^*\left(B_k, A\right) p\left(B_k\right) \tag{6.6}$$

subject to

$$\sum_{B_k \in \mathfrak{B}} b_{ijk}^*\, p\left(B_k\right) = a_{ij}^*,\ i = 1,\ldots,R,\ \ j = 1,\ldots,C, \tag{6.7}$$

and

$$p\left(B_k\right) \geq 0,\ k = 1,\ldots,L. \tag{6.8}$$

Step 4. By using an algorithm for LP, solve the LP problem established in Step 3 with respect to L unknown variables

$$\left\{p\left(B_k\right), B_k \in \mathfrak{B}\right\}. \tag{6.9}$$

Step 5. Obtain the solution set $\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$ to A consisting of arrays such that $p\left(B_k\right) > 0$ in the solution set to the LP problem obtained in Step 4.

Some remarks to be useful in implementing the algorithm are in order.

Remark 6.1. In Step 2, note that $\left\lceil a_{ij} \right\rceil$ in (6.3) or (6.4) indicates the number of units to be selected with certainty in each cell. Also, note that

$$d_2^*\left(B_k, A\right) = d_2\left(B_k, A\right) \tag{6.10}$$

and

$$d_\infty^*\left(B_k, A\right) = d_\infty\left(B_k, A\right), \tag{6.11}$$

since $b_{ijk}^* - a_{ij}^* = b_{ijk} - a_{ij}$ due to (6.3) and (6.4).

Remark 6.2. In addition to the fact that d_2^* is the natural concept of distance and d_∞^* is the simplest and easiest to compute under the norm, there is sensible advice on the choice of d_2^* or d_∞^* in Step 2. Let D_2 and D_∞ be the sets of the distance values for all possible arrays calculated by d_2^* and d_∞^*, respectively. Let those arrays with the same distance value in D_2 (D_∞) be in the same group. Then logically, d_2^* would cluster possible arrays into many different groups, where the number of groups is larger than in d_∞^*, due to (4.3) and (4.4). Accordingly, when using d_2^* in LP problem, the number of arrays in $\mathfrak{B}$ such that $p(B_k) > 0$ would be larger than in using d_∞^*.

Remark 6.3. It is clear from (6.5) and (6.6) involving the distance values d_2^* or d_∞^* that the solution in Step 5 results in the safe achievement of S1. Furthermore, S2 is achieved efficiently using linear constraints (6.7) and (6.8).

Remark 6.4. In constructing the LP problem in Step 3, the constraints for the cells with $a_{ij}^{*}=0$ can be omitted in (6.7). For example, for the 5×5 controlled selection problem of Problem 2.4, the number of necessary constraints is 23, since two cells have $a_{ij}^{*}=0$. Also, the linear constraint (3.6) is not essential, because it is implied in (6.7).

6.2 Using the simplex method

The LP problem constructed in Step 3 with the system of constraints of RC equations in (6.7) for L nonnegative unknowns in (6.8) is in the "standard form" and no transformation is required.

Supposing that $RC<L$, the number of equations is smaller than the number of unknowns. Consequently it is an LP problem with a standard form, and it can always be solved by the simplex method by transforming with the system of RC constraints in canonical form. To change the system into canonical form, one could arbitrarily choose RC variables among L variables as **basic variables** and then, using a pivot operation, attempt to put the system into canonical form, where each basic variable has coefficient one in one equation and zero in the others, and each equation has exactly one basic variable with coefficient one.

Letting the other $L-RC$ variables except RC variables chosen as basic variables be 0 in the system in canonical form, the initial **basic feasible solution** is obtained. Next, by replacing exactly one basic variable, another basic feasible solution is obtained, and these steps are continued until the minimal value of the objective function is attained by any basic feasible solution. The set of these basic feasible solutions to the LP problem is convex. Many software packages for the simplex method are available for solving the LP problem. See Dantzig (1963) and Thie and Keough (2008, chapter 3) for the details on the simplex method.

6.3 The computational demands of the LP problem

It may be claimed that our algorithm is computationally expensive due to the following burdens:

a. Before solving the LP problem, all possible arrays to the controlled selection problem should be known.

b. The number of unknowns in the LP problem, L, is equal to the number of all possible arrays, which becomes large as RC, the number of cells in the controlled selection problem, increases. Hence, it is not unreasonable that L may be as large as the binomial coefficient

$$\binom{RC}{a_{..}^{*}}, \text{ where } a_{..}^{*}=a_{..}-\sum_{i=1}^{R}\sum_{j=1}^{C}\left[a_{ij}\right]. \tag{6.12}$$

c. If RC is large, it also yields a large number of constraints in (6.7).

Sitter and Skinner (1994), and Tiwari and Nigam (1998) also referred to these potential disadvantages in describing their LP algorithms. However, due to the following reasons, these computational burdens stated in a, b, and c may not be prohibitive in **actual operations**.

First, finding all possible arrays manually might be difficult for any controlled selection problem with a large number of cells, but this task is greatly simplified using an efficient algorithm and the power of modern computers. Using the software described in the next section, they can be easily obtained in seconds even in comparatively large problems such as Problems 2.3 and 2.4.

Second, applying (6.12) to Problems 2.1 through 2.4, respectively yields 84; 11,440; 10,626; and 4,457,400 arrays. However, the actual numbers for L are only 6, 30, 141 and 159, respectively. This is because marginal expectations of both rows and columns are simultaneously matched and some cell expectations are zero. The actual numbers can also be obtained from the software described in the next section.

Third, although the greater RC, the greater the number of constraints in the LP problem, the computational demands may depend on L as well as RC, and more specifically, on the number of basic feasible solutions, possibly denoted by

$$S = \binom{L}{RC}. \tag{6.13}$$

For example, if $L = 1{,}000$ and $RC = 100$, (6.13) gives 6.4E+139, which is an extremely large number. In this case, it is almost impossible to solve the LP problem, since each basic feasible solution should be investigated. But such cases would not happen in practice. According to Ross (2007, pages 221-224), when $RC < L$, the **number of necessary transitions**, say T, moving along the basic feasible solutions in solving the LP problem with standard form is approximately normally distributed with mean $E(T) = \log_e S$ and variance $Var(T) = \log_e S$, where

$$\log_e S \approx RC\left[1 + \log_e\left\{\left(L/RC\right) - 1\right\}\right]. \tag{6.14}$$

When applying this theory to the case of $L = 1{,}000$ and $RC = 100$, approximating both the mean and variance of T by (6.14) becomes 320, and the 95% confidence interval (CI) of T is (285, 355), which is smaller than the expected lower and upper limits.

Table 6.1
Comparison between S and T

	Problem 2.1	Problem 2.2	Problem 2.3	Problem 2.4
L	6	30	141	159
RC^*	9	14	13	23
S	NA	1.5E+8	7.9E+17	3.1E+27
$E(T)$	NA	16	43	64
95% CI of T	NA	(8, 24)	(30, 56)	(48, 80)

Note: NA - not available

Table 6.1 shows the results of the comparison between S and T for the four problems considered above. Note that due to Remark 6.4, RC in (6.13) and (6.14) is replaced by RC^*, that is, the number that results from subtracting the number of cells with $a_{ij}^* = 0$ from RC. The theory on T is not applied to Problem 2.1 because $RC^* > L$.

As shown in the table, the mean or confidence interval bounds of T are considerably smaller than S in each problem. In Section 8, T in Table 6.1 will be compared with the **actual number of transitions**, say t.

7 Software

To take the advantages of the power of modern computing, we have developed a public use SAS-based software called the SOCSLP (Software for Optimal Controlled Selection Linear Programming) for our algorithm to solve controlled selection problems with two-way stratification. The recent version may be downloaded from the URL: http://www.isr.umich.edu/src/smp/socslp.

In using the software, there are no restrictions on the number of all possible arrays that can be considered for the solution. The number of those arrays and the number of constraints that can be solved depend on the memory capacity and the available disk space of the computer.

The two-phase revised simplex method, implemented using SAS/OR LP Procedure, simply "PROC LP", is employed to solve the LP problem. A unique optimal solution to the LP problem is obtained when the objective function is minimized under the given constraints (6.7) through phase 1 and 2 of PROC LP, with the assumption that all unknown variables are nonnegative (6.8).

The software produces much information including the solution set to the controlled selection problem. Also, by choosing a simple option in the software, one array can be randomly selected from the solution set, completing the controlled selection. The SOCSLP is currently available for personal computers, and the details are provided through the User Guide on the website.

8 Comparisons of algorithms

Using the four controlled selection problems given in Section 2, we present some results from the **two methods** using d_2^* and d_∞^* in the new algorithm, and compare the solutions for these two methods to solutions generated under the algorithms previously described by Jessen (1970), Jessen (1978), Causey et al. (1985), Huang and Lin (1998), and Winkler (2001). The solutions from the two methods using d_2^* and d_∞^* were obtained by implementing the SOCSLP, running on the version 9.2 of SAS/OR (2008). Solutions for the algorithm of Sitter and Skinner (1994) using LP were also obtained using PROC LP of the version 9.2 of SAS/OR (2008). Solutions for the other methods are the results as they appeared in the original papers.

The answers to two questions help us compare the algorithms: 1) Are the solutions from the new methods different from those of the previous algorithms described in Section 5? 2) Do the solutions from the new methods give higher probabilities of selection for optimum arrays compared to those generated using the previous methods?

Prior to the comparison of the algorithms, we need to take a look at the results in Table 8.1 obtained from the two methods. In the table, the method using d_2^* and the one using d_∞^* are denoted by N_2 and N_∞, respectively. Since when calculated by d_2^* (d_∞^*), the arrays with the same distance value are in the same group, there would be different groups for all possible arrays (see Remark 6.2). Let G denote the number of the different groups. Also, let OF be the actual value of the objective function (6.5) or (6.6) and t the actual number of T, the number of transitions, introduced in Section 6.3. They are all obtained from the SOCSLP, and t especially indicates the number of iterations in phase 1 and 2 of the PROC LP in the software.

Table 8.1
Results with the new methods

	Problem 2.1		Problem 2.2		Problem 2.3		Problem 2.4	
	N_2	N_∞	N_2	N_∞	N_2	N_∞	N_2	N_∞
G	4	3	9	2	6	2	157	14
OF	1.336	0.620	1.689	0.640	1.582	0.720	1.661	0.701
t	2	2	8	6	18	15	43	41

As seen in the table, most values of G are much smaller than L, the number of all possible arrays given in Table 6.1, except for the case of the large value of "157" for Problem 2.4, which arises simply due to the fact that the a_{ij} are given to three decimal places. When using d_2^*, the values of OF range between 1 and 2, while they are always less than 1, when using d_∞^*. Most values of t do not reach the 95% CI of T shown at the bottom of Table 6.1. Thus, the actual computational demands are less than those expected in the theory.

The solutions from different algorithms for the first three problems are presented in order in Table 8.2 through Table 8.4. Results for Problem 2.4 are simply described below. (The table of solutions to this problem is available on request.) In Table 8.2, the method of Sitter and Skinner (1994), Jessen's (1970) method 2 and method 3 are denoted by SS, $J2$ and $J3$, respectively. The solutions for $J2$ and $J3$ in the table are from Jessen (1970, page 782). The table shows that all methods except Jessen's (1970) method 3 yield the same solution for the 3 x 3 array Problem 2.1. In the common solutions, the probability of selection for the optimum arrays, denoted by $\sum_{B_k \in \mathfrak{B}_\infty} p(B_k)$, is 0.5.

Table 8.2
Comparison of solutions to Problem 2.1

B_k	$p(B_k)$				
	N_2	N_∞	SS	$J2$	$J3$
0 1 1 1 0 1 1 1 0	0.2	0.2	0.2	0.2	0.1
1 0 1 * 1 1 0 0 1 1	0.5	0.5	0.5	0.5	0.4
1 1 0 0 1 1 1 0 1	0.3	0.3	0.3	0.3	0.2
0 1 1 1 1 0 1 0 1					0.1
1 0 1 0 1 1 1 1 0					0.1
1 1 0 1 0 1 0 1 1					0.1
Total	1.0	1.0	1.0	1.0	1.0
Total †	0.5	0.5	0.5	0.5	0.4

Note: * – Optimum array
† – The sum of probabilities of selection for optimum arrays

In Table 8.3, Jessen's (1978) method is denoted by JS. The solution for JS in the table is from Jessen (1978, pages 375-376). As shown in the table, the new methods using d_2^* and d_∞^* have the same solution for the Problem 2.2 4×4 array; however only one-half of the arrays in those solutions overlap with the arrays in the solutions from the methods of Sitter and Skinner (1994) and Jessen (1978). Also, the Sitter and Skinner and Jessen methods provide a lower probability of 0.6 to optimum arrays, whereas the new methods allocate the higher probability of 0.8 to the arrays.

Table 8.3
Comparison of solutions to Problem 2.2

B_k	$p(B_k)$			
	N_2	N_∞	SS	JS
0 0 1 1 0 1 0 1 1 1 0 0 1 0 1 0	0.2	0.2		
0 0 1 1 * 1 1 0 0 0 0 1 1 1 1 0 0	0.2	0.2	0.4	0.2
0 1 1 0 * 1 0 0 1 0 0 1 1 1 1 0 0	0.2	0.2		
0 1 1 0 * 1 0 1 0 1 0 0 1 0 1 0 1	0.4	0.4	0.2	0.4
0 1 1 0 0 0 1 1 1 1 0 0 1 0 0 1			0.2	
0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0			0.2	
0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0				0.2
0 0 1 1 0 1 0 1 1 0 1 0 1 1 0 0				0.2
Total	1.0	1.0	1.0	1.0
Total †	0.8	0.8	0.6	0.6

See note for Table 8.2.

Problem 2.3, with 141 possible arrays, is considerably larger than the above two problems. The solutions to this problem under the five methods are compared in Table 8.4. In the table, the methods of Causey et al. (1985) and Huang and Lin (1998) are denoted by CA and HU, respectively. The solutions for CA and HU in the table are from Causey et al. (1985, page 906) and Huang and Lin (1998, Figure 3), respectively.

Table 8.4
Comparison of solutions to Problem 2.3

B_k	$p(B_k)$					B_k	$p(B_k)$					B_k	$p(B_k)$				
	N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0.2	0.2	0.2			0 2 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1			0.11			0 2 0 1 0 1 0 0 0 2 0 0 1 0 0 0 1 0 0 0 1 0 0 1				0.2	
0 2 0 1 0 1 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0	0.1	0.2	0.03			0 2 0 1 0 1 1 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0			0.03			0 2 0 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 0 0				0.2	0.2
0 2 0 1 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0	0.1					0 2 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1			0.03			0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0				0.2	
0 2 0 2 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1	0.1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0			0.09			0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 0 0 0 1 0 0 0 1					0.2
0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1	0.1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0			0.08			0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0					0.2
1 2 0* 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0	0.1		0.08			0 2 0 2 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0			0.03								
1 2 0* 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0	0.3	0.4	0.2	0.4	0.4	1 2 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 1			0.06								
0 2 0 2 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1		0.2				1 2 0 1 0 1 0 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0			0.06			Total	1.0	1.0	1.0	1.0	1.0
												Total †	0.4	0.4	0.28	0.4	0.4

See note for Table 8.2.

We note that all these methods provide different solutions, and about half of the arrays overlap between the new methods and the method of Sitter and Skinner (1994). Moreover, the solutions from the methods of Causey et al. (1985) and Huang and Lin (1998) are quite unlike the solution from the method using d_∞^*. The

method using d_2^* and Sitter and Skinner's method distribute the probabilities of selection to two optimum arrays, whereas the other three methods just allocate the probability to only one optimum array. Sitter and Skinner's method appears to be less effective in selecting optimum arrays since their method gives the probability of 0.28 to those, while the others give the higher probability of 0.4.

The solutions to Problem 2.4, which is the largest of the given problems, are compared under the four methods (N_2 , N_∞ , SS , and Winkler's (2001) method). Only two arrays, including one optimum, overlap in the solutions, and the two new methods give the same probabilities (0.127 and 0.483) to those arrays. Even when comparing the method using d_∞^* with the methods of Sitter and Skinner (1994) and Winkler (2001), their solutions are very different. Also, the new methods give the same probability of selection of 0.483 to the optimum array, whereas the other previous methods give the lower probabilities of 0.385 and 0.104, respectively.

In summary, it seems that the new methods successfully achieve S1 and S2 of optimal solutions. Note that the new methods consistently give higher probabilities of selection for optimum arrays and that the totals of those probabilities are always the same. The solutions from the new methods are very different from those obtained using previous methods, when the controlled selection problems are not small. This implies that the solutions from the previous methods may be far from optimal under criteria S1 and S2 (R1 and R2).

9 Concluding remarks

In this paper, we introduced the concept of optimal solutions to a controlled selection problem with two-way stratification, and proposed a new algorithm for finding such solutions. The algorithm has been easily and successfully implementedin the new SAS-based software (SOCSLP).

Since an optimal solution is a general idea, it may be adopted as one of the useful criteria for comparing the different algorithms. As shown in the above comparisons, the new algorithm results in solutions to large controlled selection problems that are very different from those derived using previously published methods. It is also likely to yield greater probabilities of selection for optimum arrays as compared to those obtained by the previous methods.

Based on the results for the two-way controlled selection problems, we expect that the suggested method would also contribute to improvements in the properties of solutions to controlled selection problems with three-way or more stratification dimensions.

Acknowledgements

This paper is in honor of I. Hess who dedicated her life to studying controlled selection. The authors wish to thank Jea-Bok Ryu in Chongju University for providing ideas and advice in the early stage of this study. We are also grateful to two anonymous referees, the Editor and the Associate Editor for their valuable comments and suggestions.

References

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

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.

Chernick, M.R. and Wright, T. (1983). Estimation of a population mean with two-way stratification using a systematic allocation scheme. *Journal of Statistical Planning and Inference*, 7, 219-231.

Cox, L.H. and Ernst, L.R. (1982). Controlled rounding. *INFOR: Information Systems and Operational Research*, 20, 423-432.

Dantzig, G.B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, New Jersey.

Deville, J-C and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91, 893-912.

Dykstra, R. (1985a). An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Annals of Probability*, 13, 975-984.

Dykstra, R. (1985b). Computational aspects of I-projections. *Journal of Statistical Computation and Simulation*, 21, 265-274.

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. In *Probability Sampling of Hospitals and Patients*, *Second Edition*, (Eds., I. Hess, D.C. Ridel and T.B. Fitzpatrick), Health Administration Press, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. (1998). PCCONSEL user guide. In *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess and S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. and Hess, I. (1983). More on controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 106-111.

Hess, I. and Heeringa, S.G. (2002). *Controlled Selection Continued.* Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Hess, I., Ridel, D.C. and Fitzpatrick, T.B. (1975). *Probability Sampling of Hospitals and Patients*, *Second Edition*. Health Administration Press, University of Michigan, Ann Arbor, USA.

Huang, H.C. and Lin, T.K. (1998). On the two-dimensional controlled selection problem. In *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess and S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

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

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

Lin, T.K. (1992). Some improvements on an algorithm for controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 407-410.

Raghunandanan, K. and Bryant, E.C. (1971). Variance in multi-way stratification. *Sankhyā, Series A*, 33, 221-226.

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

Rao, J.N.K. and Nigam, A.K. (1992). "Optimal" controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

Ross, S.M. (2007). *Introduction to Probability Models*. Burlington, MA: Academic Press.

SAS/OR (2008). *User's Guide: Mathematical Programming*. Version 9.2, 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.

Thie, P.R. and Keough, G.E. (2008). *An Introduction to Linear Programming and Game Theory, Third Edition*. Hoboken, New Jersey: John Wiley and Sons.

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

Winkler, W.E. (1990). On Dykstra's iterative fitting procedure. *Annals of Probability*, 18, 1410-1415.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. U.S. Census Bureau, *Statistical Research Division Report RRS 2001/01*. Available from: //www.census.gov/srd/www/byyear.html.