The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 139-151

Comparison of Algorithms for Two-way Stratification Design[1)]

Sun-Woong Kim[2)]

Abstract

Kim *et al.* (2002) developed public use SAS-based software for two-way stratification design, which is called SOCSLP. We describe the details of a new approach implemented using SOCSLP and key differences between the approach and the sampling schemes of Sitter and Skinner (1994) and Winkler (2001). In addition, a numerical example is given to compare those methods with respect to the probabilities of selecting sample arrays.

Keywords : Linear Programming, Iterative Integer Linear Programming, Optimal Samples

1. Introduction

In sample surveys one may encounter populations for which there exist several effective stratifying variables and in such cases multi-way stratification design, which is often called controlled sampling design, can be used to determine the sizes of sample units in the cells formed by cross-classifying categories of those variables in order to guarantee gain in precision of the estimators. A problem with the approach is largely due to the fact that when selecting primary sampling units in household surveys, the expected number of sample units in each cell may be very small or the total expected number of sample units be less than the total number of cells.

A variety of controlled sampling techniques to overcome the problem have been suggested and those methods not only can yield sample units that are representative of the population but also can maintain the usual requirements for probability sampling.

Goodman and Kish (1950) first developed the method. Bryant *et al.* (1960) proposed a simple method for two-way stratification, but their method cannot be used if a specified cell size is zero. Jessen (1970) suggested several approaches for two-way and three-way stratification that are complicated to implement and a solution is not always available.

Hess *et al.* (1975) described how to employ controlled sampling to select a sample of Michigan's hospitals and patients. Causey *et al.* (1985) proposed an algorithm using

1) This work is supported by the Dongguk University research fund
2) Full-time Lecturer, Department of Statistics, Dongguk University, Seoul, 100-715, Korea
E-mail: sunwk@dongguk.edu

transportation theory to solve multi-way stratification problems.

Rao and Nigam (1990, 1992) showed how controlled sampling designs can be achieved by linear programing (LP). Sitter and Skinner (1994) applied their approach to controlled sampling method for multi-way stratification problems.

Winkler (2001) proposed an iterative integer LP procedure which is similar to that of Causey *et al.* (1985) but uses a different probabilistic structure for the design.

Kim *et al.* (2002) suggested a new LP approach to adopt ordinary distance functions to minimize the overall distortion to cell expectations and developed a software called SOCSLP (Software for Optimal Controlled Selection Linear Programming) in order to implement their approaches. A public version of the software can be downloaded from the URL: http://www.isr.umich.edu/src/smp/socs.

In this paper, we first describe the approach suggested by Kim *et al.* (2002) and a feature of SOCSLP, and then presents the details of key differences between the method of Kim *et al.* (2002) and those of Sitter and Skinner (1994) and Winkler (2001). In addition, we compare them through an example with respect to the selection probabilities of possible arrays.

2. Notation

Consider the two-way stratification problem that a population is partitioned by two stratifying criteria of classification, R rows and C columns, resulting in a $R\times C$ tabular array A. The tabular array of the RC cells having real numbers, a_{ij}, $i=1,\cdots,R$, $j=1,\cdots,C$ where each denotes the expected number of sample units in the ij th cell when a sample of a units is drawn from the population.

Letting S_k, $k=1,\cdots,L$ denote possible arrays where each array is the replacement of the real numbers in the tabular array A by the adjacent integer values, $P(S_k)$ represent the selection probability assigned to each S_k in a set S of possible arrays. We say that certain arrays where each is allocated the selection probability larger than zero are 'sample arrays' for A. Also, let $n_{ij}(S_k)$ be each adjacent value of S_k, which is the sample size in the ij th cell depending on S_k and equals either $[a_{ij}]$ or $[a_{ij}]+1$, where $[a_{ij}]$ is the integer part of a_{ij}.

3. The Linear Programming Approach

3.1 The Algorithm and SOCSLP

Finding a solution to the two-way stratification problem is to determine $n_{ij}(S_k)$ for each cell according to a specified randomized procedure with associate probabilities $P(S_k)$, $k=1,\cdots,L$.

In order to maintain the requirements of probability sampling, we may restrict attention to sampling designs satisfying the following:

$$\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(S_k)=a \ , \tag{3.1}$$

$$E[n_{ij}(S_k) \mid i,j]=\sum_{S_k \ni i,j,\, S_k\in S} n_{ij}(S_k)P(S_k)=a_{ij} \ , \tag{3.2}$$

$$\sum_{S_k\in S} P(S_k)=1 \ . \tag{3.3}$$

There may be a large number of sets of probability distributions $P(S_k)$ satisfying (3.1), (3.2) and (3.3), although only one set of them should be chosen to obtain a solution to the two-way stratification problems. Moreover, the selection probabilities of possible arrays should depend only on the original tabular array A . In this case, we may consider an approach to guarantee and yield optimal solutions reflecting the closeness of each array S_k to A . The distance metric would be appropriate for measuring this "closeness."

For this purpose, we first consider the ordinary distance metric, which is often called the Euclidean metric, by defining

$$m_1(S_k:A)=\left[\sum_{i=1}^{R}\sum_{j=1}^{C}\{n_{ij}(S_k)-a_{ij}\}^2\right]^{\frac{1}{2}} \ , \quad k=1,\cdots,L \tag{3.4}$$

(3.4) would be the most common measure used to define the distance between A and S_k .

The following metric measure could be also offered to define another distance function for each integer p .

$$m_2(S_k:A)=\left[\sum_{i=1}^{R}\sum_{j=1}^{C}\{n_{ij}(S_k)-a_{ij}\}^{2p}\right]^{\frac{1}{2p}} \ , \quad k=1,\cdots,L \ , \quad 1<p<\infty \tag{3.5}$$

From (3.4) and (3.5) it is clear that other distance functions can be defined by letting

$$m_3(S_k:A)=\left[\sum_{i=1}^{R}\sum_{j=1}^{C}|n_{ij}(S_k)-a_{ij}|^{q}\right]^{\frac{1}{q}} \ , \quad k=1,\cdots,L \ , \quad 1\le q<\infty \tag{3.6}$$

Furthermore the simpler distance function can be specified by the formula

$$m_4(S_k:A) = \max\{|n_{ij}(S_k) - a_{ij}| : i=1,\cdots R, j=1,\cdots,C\}\ , \quad k=1,\cdots,L\ , \tag{3.7}$$

which is derived by

$$\lim_{q\to\infty} m_3(S_k:A)\ . \tag{3.8}$$

Since those distance functions that are different from each other, such as m_1 , m_2 , m_3 and m_4 , give rise to a number of distance metric spaces, we need to choose some of them. Here m_1 or m_4 , which are special cases of m_3 with $q=2$ and $q=\infty$ respectively, may be recommended because m_1 indicates the "overall deviation" between A and S_k considering all RC cells separately, whereas m_4 represents the "maximum deviation" by a single cell.

For the set S of possible arrays, we define a few arrays having the minimum distance value from m_1 or m_4 as 'optimal samples.' But we would prefer m_4 rather than m_1 due to the following relation

$$S_0 \subseteq S_1\ , \tag{3.9}$$

where S_0 is the set of optimal samples for the distance function m_1 and S_1 is the set of optimal samples under the function m_4 .

Also, we define some arrays having the maximum distance value as 'unfavorable samples.' For those samples, m_4 may be also preferable to m_1 , so as to compare under the same distance function as 'optimal samples.' Let S_2 be the set of unfavorable samples under the function m_4 . Note that m_4 can sometimes yield a greater number of unfavorable samples. In that case m_1 can be used instead.

Now we introduce the approach implemented by SOCSLP to optimize the assignment of the selection probability to each array.

It begins with the basic idea to use a sampling design $P(S_k)$ obtained by minimizing the problem:

$$\phi = \sum_{S_k\in S} W(S_k)P(S_k)\ , \tag{3.10}$$

under some linear constraints, where $W(S_k)$ is a weight for the array S_k .

The approach using (3.10) may assign a higher probability of selection to the sample arrays having the smaller weight. To specify this sampling strategy, we develop an algorithm consisting of the following steps:

***Step* 1.** Consider a set of all possible arrays, denoted by S , corresponding to rounding of the $R \times C$ tabular array A , which consists of the RC cells that have real numbers a_{ij} .

Step 2. Establish the following LP problems matching (3.10):

$$\phi_1 = \sum_{S_k \in S} m_1(S_k : A) P(S_k) = \sum_{S_k \in S} \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \{ n_{ij}(S_k) - a_{ij} \}^2 \right]^{\frac{1}{2}} P(S_k) \tag{3.11}$$

or

$$\phi_2 = \sum_{S_k \in S} m_2(S_k : A) P(S_k) = \sum_{S_k \in S} \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \{ n_{ij}(S_k) - a_{ij} \}^{2p} \right]^{\frac{1}{2p}} P(S_k) \tag{3.12}$$

or

$$\begin{aligned} \phi_3 &= \sum_{S_k \in S} m_4(S_k : A) P(S_k) \\ &= \sum_{S_k \in S} \max\{ | n_{ij}(S_k) - a_{ij} | : i = 1, \cdots R, j = 1, \cdots, C \} P(S_k) \end{aligned} \tag{3.13}$$

subject to

$$P(S_k) \geq 0 \ , \tag{3.14}$$

$$\sum_{S_k \in S} P(S_k) = 1 \ , \tag{3.15}$$

$$\sum_{S_k \ni i, j, S_k \in S} P(S_k) = a_{ij} \ . \tag{3.16}$$

Step 3. Minimize the objective function ϕ_1 or ϕ_2 or ϕ_3 with respect to the variables $\{ P(S_k), S \in S_k \}$, selection probabilities of possible arrays, subject to the constraints (3.14), (3.15) and (3.16).

Step 4. Randomly select a sample array S_k from the sampling design $P(S_k)$ which is the solution in the step 3, using the method of cumulative sums.

Since each a_{ij} , which is the expected number of sample units in the ij th cell, always includes the integer part, we can also follow the steps above after subtracting the part from a real number a_{ij} . It gives the same result because the integer number is considered as the fixed size and deducting the integer part does not influence the distance functions.

We do not consider the metric measure m_3 , although it can be used as the weight in the objective function. It would be sufficient to explore the role of different distance functions in the LP approach by using m_1 and m_2 instead of m_3 .

Objective functions ϕ_1 , ϕ_2 and ϕ_3 would reflect the “closeness” of each possible array S_k to the two-way stratification problem A , and this LP approach would maximize the selection probabilities of those samples that have the smaller distance values than others, as well as optimal samples under the given constraints.

In particular, it would be easier to use ϕ_3 than ϕ_1 or ϕ_2 since m_4 is the simpler form and m_4 would cluster possible arrays into several groups in which the number of groups is much smaller than in using m_1 and m_2 . Typically, a group means the collection of sample arrays having the same weight. Then ϕ_3 would be more efficient to achieve the desired goal.

In order to implement the algorithm, we can use a public use version of the SOCSLP which was upgraded in 2003. This software automatically produces all possible arrays S_k , $k=1,\cdots,L$ for the $R\times C$ tabular array A . The LP problems using the objective functions such as ϕ_1 , ϕ_2 and ϕ_3 are easily set up to take the step 2 and we can minimize those objective functions under the given constraints in the step 3, by some options in the program. In addition, the program automatically yields one random sample array from the sampling design $P(S_k)$ as the step 4.

To solve the two-way stratification problem, this software adopts a two-phase revised simplex method, implemented using SAS/OR LP procedure. A unique optimal solution set is provided when the chosen objective function is minimized through phase 1 and 2 of LP program.

SOCSLP was originally developed for personal computers using the Microsoft Windows but is currently available for other systems such as Linux operating systems, Unix workstations using the Sun Solaris and IBM AIX operating systems.

In summary, SOCSLP provides the followings including a solution to a two-way stratification problem:

- The solution from the SAS Proc LP procedure
- The tables of all possible arrays
- The tables of the sample arrays selected by the LP procedure
- The final sample array

3.2 The differences between the three approaches

In this section we take a look at the major differences between the above-mentioned approach and those of Sitter and Skinner (1994) and Winkler (2001).

Sitter and Skinner (1994) extended ideas from Rao and Nigam (1990, 1992) and showed how LP may be applied to multi-way stratification problems. The purpose of their scheme is to minimize (3.10) using a loss function such as

$$W(S_k)=\sum_{i=1}^{R}\{n_{i.}(S_k)-a_{i.}\}^2+\sum_{j=1}^{C}\{n_{.j}(S_k)-a_{.j}\}^2 , \qquad (3.17)$$

where $n_{i.}(S_k)=\sum_{j=1}^{C} n_{ij}(S_k)$, $n_{.j}(S_k)=\sum_{i=1}^{R} n_{ij}(S_k)$, $a_{i.}=\sum_{j=1}^{C} a_{ij}$, $a_{.j}=\sum_{i=1}^{R} a_{ij}$.

They demonstrated their method on the two-way stratification problems in the literature to show the fact that there are the designs for which the minimal solution of (3.10) is zero. But their method does not generally yield an optimal solution in the meaning of reflecting the closeness of each array S_k to the two-way stratification problem since (3.17) indicates the weight for the margins of a_{ij} s. Note that m_1 , m_2 and m_4 used as different loss functions in (3.11), (3.12) and (3.13) respectively are for each a_{ij} . Thus the scheme of Sitter and Skinner would reflect a less straightforward weight than the developed approach .

On the other hand Winkler (2001) suggested an algorithm that is quite similar to one of Causey *et al.* (1985), which employs transportation theory. His method uses an iterative integer LP procedure to obtain a solution for the multi-way stratification problems, but it is unlikely to optimize the assignment of probability to each array because each of sample arrays is compulsively allocated the following selection probabilities:

$$P(M_1)=1-\max|(a_{ij}-M_{ij1})/(M_{ij1}-B_{ij1})| \ , \qquad (3.18)$$

where M_1 is the first array obtained by the iterative integer LP procedure, M_{ij1} is each internal entry of M_1 , and B_{ij1} is the smallest integer greater than a_{ij} or the largest integer smaller than a_{ij} ,

and

$$P(M_g)=(1-\sum_{h=1}^{g-1}P(M_h))(1-\max|(a_{ijg}-M_{ijg})/(M_{ijg}-B_{ijg})|) \ , \qquad (3.19)$$

where M_g , $g=2,\cdots,G$ are the following arrays obtained by the iterative integer LP procedure, a_{ijg} is the modified a_{ij} for each g , M_{ijg} is each internal entry of M_g , and B_{ijg} is the smallest integer greater than a_{ijg} or the largest integer smaller than a_{ijg} or zero.

Also, a key feature of Winkler scheme is that his algorithm used a small part of all possible arrays because all of those arrays should be enumerated prior to running Sitter and Skinner's LP algorithm. But since SOCSLP automatically provides those arrays as mentioned before, it would not be a big problem when employing the LP algorithm.

4. An Example

Winkler (2001) applies his algorithm to the two-way stratification problem given in Table 1, *which is the* 5×5 *tabular array with one integer entry, single zero and* 23 non-integers. Each entry indicates the expected number of sample units a_{ij} . A sample of $a=37$ units is totally drawn.

Table 1. Two-way Stratification Problem

Variable 1 Strata	Variable 2 Strata 1	2	3	4	5	Total
1	2.000	2.483	1.052	0.103	0.362	6
2	2.182	1.061	1.101	1.046	0.610	6
3	0.000	1.614	1.914	2.200	1.272	7
4	0.860	0.377	0.930	2.840	2.993	8
5	0.958	0.465	2.003	1.811	4.763	10
Total	6	6	7	8	10	37

This problem has 159 possible arrays, which are provided by SOCSLP. Thus those arrays are dented by S_k , $k=1,\cdots,159$. As described in Section 3.1, we may use m_4 to decide optimal samples or unfavorable samples among possible arrays.

Appendix 1 presents a comparison of the methods using ϕ_1 and ϕ_3 implemented by SOCSLP with those of Sitter and Skinner (1994) and Winkler (2001). Note that all these methods provide different controlled sampling designs, although each design equally has 15 sample arrays. Specifically, Appendix 1 shows the sum of the assigned selection probabilities of optimal samples or unfavorable samples as the first criteria to compare the results of controlled sampling designs. For the selection probabilities of optimal samples, Winkler's method provides the lowest probability of 0.104, whereas other methods yield the same probability of 0.483 that is so higher than that of Winkler. For unfavorable samples, four methods have identical selection probabilities of 0.003, which is very low. In this case we would prefer the three methods allocating the higher probabilities for optimal samples rather than Winkler's method.

The next best comparison may be to examine the cumulative sums of the selection probabilities of sample arrays arranged in ascending order of the weight m_4 . It would be preferable that the lower weights that the sample arrays have, the higher assigned the cumulative sums of the selection probabilities are for those arrays. Appendix 2 presents 13 groups of sample arrays divided by the weight m_4 and the cumulative sums of the selection probabilities by those groups. It is clear from Appendix 2 that the method using ϕ_3

consistently assigns the higher cumulative sums of selection probabilities for the groups of sample arrays having the lower weights. But method using ϕ_1 and Sitter and Skinner's method, as well as Winkler's method are erratic in this regard.

It is also evident that when the cumulative sums arranged in descending order of the weight m_4 is considered, the method using ϕ_3 may assign regularly the lower cumulative sums of selection probabilities for sample groups having the higher weights.

In the result, the method using ϕ_3 systematically reflects the weight of each sample array and is useful not only to maximize the selection probabilities of optimal samples but also to minimize the probability of selecting unfavorable samples.

5. Concluding Remarks

The LP approach implemented using SOCSLP would be more useful than Winkler's method which repeatedly uses integer LP procedure. Also, it would be more efficient than Sitter and Skinner's method using a different LP approach.

As mentioned above concerning the enumeration of all possible arrays, one disadvantage of LP approach is that it may become computationally expensive as the number of those arrays in the two-way stratification problems increases. But since SOCSLP is available for Unix workstations using the Sun Solaris and IBM AIX operating systems, as well as personal computers, most stratification problems may be easily solved.

References

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

[2] Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of transportation theory to statistical problems, *The Journal of the American Statistical Association*, Vol. 80, 903-909.

[3] Goodman, R. and Kish, L. (1950). Controlled selection - a technique in probability sampling, *The Journal of the American Statistical Association*, Vol. 45, 350-372.

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

[5] Jessen, R. J. (1970). Probability sampling with marginal constraints, *The Journal of the American Statistical Association*, Vol. 65, 776-795.

[6] Kim, Sun-Woong, Heeringa S. G., and Solenberger P. W. (2002). Optimizing solution sets in two-way controlled selection problems, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1839-1844.

[7] SAS/OR (2001). *User's guide: mathematical programming*, Version 8. Cary, NC:SAS

Institute Inc.
[8] Sitter, R. R. and Skinner, C. J. (1994). Multi-way stratification by linear programming, *Survey Methodology*, Vol. 20, No. 1, 65-73.
[9] Rao, J. N. K. and Nigam, A. K. (1990). Optimal controlled sampling design, *Biometrika*, Vol. 77, 807-814.
[10] Rao, J. N. K. and Nigam, A. K. (1992). Optimal controlled sampling: a unified approach, *International Statistical Review*, Vol. 60, 89-98.
[11] Winkler, E. W. (2001). Multi-way survey stratification and sampling, U.S. Bureau of the Census, Statistical Division Report RRS2001/01 (http://www.census.gov/srd/www/byyear.html).

[Received December 2003, Accepted March 2004]

Appendix 1. Controlled Sampling Designs and Comparison

Sample	S_k	$P(S_k)$	Sample	S_k	$P(S_k)$	Sample	S_k	$P(S_k)$	Sample	S_k	$P(S_k)$
1[a]	2 2 1 0 1	0.127[b]	10	2 2 1 0 1	0.000	19	2 2 1 1 0	0.000	28	2 2 1 1 0	0.000
	2 1 1 1 1	0.127[c]		2 1 2 1 0	0.000		2 1 1 1 1	0.000		3 1 1 1 0	0.000
	0 2 2 2 1			0 2 2 2 1			0 1 2 2 2			0 1 2 2 2	
	1 0 1 3 3	0.065[d]		1 1 0 3 3	0.000		1 1 1 2 3	0.055		0 1 1 3 3	0.035
	1 1 2 2 4	0.237[e]		1 0 2 2 5	0.032		1 1 2 2 4	0.000		1 1 2 1 5	0.000
2	2 2 1 0 1	0.007	11	2 2 1 0 1	0.009	20	2 2 1 1 0	0.000	29	2 2 1 1 0	0.039
	2 1 1 2 0	0.000		2 2 1 1 0	0.013		2 1 1 1 1	0.000		3 1 1 1 0	0.000
	0 1 2 2 2			0 1 2 2 2			0 1 2 2 2			0 1 2 2 2	
	1 1 1 3 2	0.000		1 0 1 3 3	0.061		1 1 1 3 2	0.007		1 1 1 2 3	0.000
	1 1 2 1 5	0.000		1 1 2 2 4	0.000		1 1 2 1 5	0.000		0 1 2 2 5	0.000
3	2 2 1 0 1	0.027	12	2 2 1 0 1	0.000	21	2 2 1 1 0	0.000	30	2 2 2 0 0	0.000
	2 1 1 2 0	0.000		2 2 1 1 0	0.000		2 1 1 1 1	0.000		2 1 1 1 1	0.000
	0 1 2 3 1			0 1 2 2 2			0 2 2 2 1			0 2 1 3 1	
	1 1 1 2 3	0.000		1 1 1 3 2	0.000		1 0 1 3 3	0.000		1 1 1 2 3	0.000
	1 1 2 1 5	0.000		1 0 2 2 5	0.007		1 1 2 1 5	0.103		1 0 2 2 5	0.017
4	2 2 1 0 1	0.000	13	2 2 1 0 1	0.048	22	2 2 1 1 0	0.012	31	2 2 2 0 0	0.000
	2 1 1 2 0	0.000		2 2 1 1 0	0.000		2 1 1 2 0	0.029		2 1 1 1 1	0.000
	0 2 2 2 1			0 1 2 3 1			0 1 2 2 2			0 2 2 2 1	
	1 0 1 3 3	0.000		1 1 1 2 3	0.000		1 1 1 2 3	0.000		1 1 0 3 3	0.000
	1 1 2 1 5	0.043		1 0 2 2 5	0.000		1 1 2 1 5	0.000		1 0 2 2 5	0.035
5	2 2 1 0 1	0.000	14	2 2 1 0 1	0.004	23	2 2 1 1 0	0.034	32	2 2 2 0 0	0.000
	2 1 1 2 0	0.000		2 2 1 1 0	0.000		2 1 2 1 0	0.004		2 1 1 2 0	0.017
	# 0 2 2 2 1			0 2 2 2 1			0 1 1 3 2			0 1 2 2 2	
	1 1 0 3 3	0.000		1 0 1 3 3	0.000		1 1 1 2 3	0.000		1 1 0 3 3	0.000
	1 0 3 1 5	0.003		1 0 2 2 5	0.000		1 1 2 1 5	0.000		1 1 2 1 5	0.000
6	2 2 1 0 1	0.000	15	2 2 1 0 1	0.000	24	2 2 1 1 0	0.018	33	2 2 2 0 0	0.049
	2 1 1 2 0	0.000		3 1 1 1 0	0.003		2 1 2 1 0	0.018		2 1 2 1 0	0.000
	0 2 2 2 1			# 0 1 1 3 2			0 1 2 2 2			0 1 1 3 2	
	1 1 1 2 3	0.046		1 1 1 3 2	0.000		1 1 0 3 3	0.000		1 1 0 3 3	0.010
	1 0 2 2 5	0.000		0 1 3 1 5	0.000		1 1 2 1 5	0.000		1 1 2 1 5	0.000
7	2 2 1 0 1	0.000	16	2 2 1 0 1	0.101	25	2 2 1 1 0	0.000	34	2 2 2 0 0	0.003
	2 1 2 1 0	0.075		3 1 1 1 0	0.022		2 1 2 1 0	0.000		3 1 1 1 0	0.000
	0 1 1 3 2			0 1 2 2 2			# 0 2 1 2 2			# 0 1 1 3 2	
	1 1 1 2 3	0.056		0 1 1 3 3	0.000		1 1 0 3 3	0.003		1 1 0 3 3	0.000
	1 1 2 2 4	0.000		1 1 2 2 4	0.000		1 0 3 1 5	0.000		0 1 3 1 5	0.000
8	2 2 1 0 1	0.000	17	2 2 1 0 1	0.000	26	2 2 1 1 0	0.000	35	2 2 2 0 0	0.000
	2 1 2 1 0	0.000		3 1 1 1 0	0.004		2 1 2 1 0	0.004		3 1 1 1 0	0.035
	0 1 2 3 1			0 1 2 2 2			0 2 1 2 2			0 1 2 2 2	
	1 1 0 3 3	0.015		1 1 1 3 2	0.000		1 1 1 2 3	0.003		1 1 0 3 3	0.042
	1 1 2 1 5	0.000		0 1 2 2 5	0.000		1 0 2 2 5	0.000		0 1 2 2 5	0.000
9	2 2 1 0 1	0.000	18	2 2 1 0 1	0.039	27	2 2 1 1 0	0.000	36	2 3 1 0 0	0.000
	2 1 2 1 0	0.000		3 1 1 1 0	0.118		2 2 1 1 0	0.048		2 1 1 1 1	0.000
	0 2 1 3 1			0 1 2 3 1			0 1 2 2 2			0 1 2 3 1	
	1 0 1 3 3	0.014		0 1 1 3 3	0.105		1 1 1 2 3	0.000		1 1 1 2 3	0.000
	1 1 2 1 5	0.040		1 1 2 1 5	0.000		1 0 2 2 5	0.000		1 0 2 2 5	0.114

Appendix 1. Controlled Sampling Designs and Comparison (Continued)

Sample S_k	$P(S_k)$	Sample S_k	$P(S_k)$	Sample S_k	$P(S_k)$
37 2 3 1 0 0	0.483	39 2 3 1 0 0	0.000	41 2 3 1 0 0	0.000
2 1 1 1 1	0.483	2 2 1 1 0	0.000	3 1 1 1 0	0.000
* 0 2 2 2 1	0.483	0 1 2 2 2	0.000	0 1 2 2 2	0.000
1 0 1 3 3		1 0 1 3 3		1 0 1 3 3	
1 0 2 2 5	0.104	1 0 2 2 5	0.054	0 1 2 2 5	0.042
38 2 3 1 0 0	0.000	40 2 3 1 0 0	0.000		
2 1 2 1 0	0.000	3 1 1 1 0	0.000		
0 1 1 3 2	0.000	0 1 2 2 2	0.000		
1 1 1 2 3		0 1 1 3 3			
1 0 2 2 5	0.029	1 0 2 2 5	0.140		
$\sum_{S_k \in S_1} P(S_k)$	0.483[b]			$\sum_{S_k \in S_2} P(S_k)$	0.003[b]
	0.483[c]				0.003[c]
	0.483[d]				0.003[d]
	0.104[e]				0.003[e]

Note. *: Optimal sample

#: Unfavorable sample

a: Sample array identification number (SAID)

b: Method using ϕ_1

c: Method using ϕ_2

d: Sitter and Skinner's method

e: Winkler's method

Appendix 2. Cumulative Sums of Selection Probabilities by Groups of Sample Arrays

Groups	a	Weights	b	c	d	e	Cumulative Sums b	Cumulative Sums c	Cumulative Sums d	Cumulative Sums e
1	* 37	0.517	0.483	0.483	0.483	0.104	0.483	0.483	0.483	0.104
2	1	0.763	0.127	0.127	0.065	0.237	0.610	0.610	0.548	0.341
3	36	0.840	0.000	0.000	0.000	0.114	0.610	0.610	0.548	0.455
4	16	0.860	0.101	0.022	0.000	0.000	0.750	0.750	0.653	0.595
	18		0.039	0.118	0.105	0.000				
	40		0.000	0.000	0.000	0.140				
5	19	0.897	0.000	0.000	0.055	0.000	0.750	0.750	0.743	0.698
	21		0.000	0.000	0.000	0.103				
	28		0.000	0.000	0.035	0.000				
6	7	0.914	0.000	0.075	0.056	0.000	0.784	0.833	0.816	0.767
	9		0.000	0.000	0.014	0.040				
	23		0.034	0.004	0.000	0.000				
	26		0.000	0.004	0.003	0.000				
	38		0.000	0.000	0.000	0.029				
7	8	0.930	0.000	0.000	0.015	0.000	0.802	0.851	0.831	0.799
	10		0.000	0.000	0.000	0.032				
	24		0.018	0.018	0.000	0.000				
8	11	0.939	0.009	0.013	0.061	0.000	0.863	0.912	0.892	0.853
	13		0.048	0.000	0.000	0.000				
	14		0.004	0.000	0.000	0.000				
	27		0.000	0.048	0.000	0.000				
	39		0.000	0.000	0.000	0.054				
9	30	0.948	0.000	0.000	0.000	0.017	0.912	0.912	0.902	0.905
	31		0.000	0.000	0.000	0.035				
	33		0.049	0.000	0.010	0.000				
10	3	0.954	0.027	0.000	0.000	0.000	0.951	0.958	0.948	0.948
	4		0.000	0.000	0.000	0.043				
	6		0.000	0.000	0.046	0.000				
	22		0.012	0.029	0.000	0.000				
	32		0.000	0.017	0.000	0.000				
11	29	0.958	0.039	0.000	0.000	0.000	0.990	0.993	0.990	0.990
	35		0.000	0.035	0.042	0.000				
	41		0.000	0.000	0.000	0.042				
12	2	0.993	0.007	0.000	0.000	0.000	0.997	0.997	0.997	0.997
	12		0.000	0.000	0.000	0.007				
	17		0.000	0.004	0.000	0.000				
	20		0.000	0.000	0.007	0.000				
13	# 5	0.997	0.000	0.000	0.000	0.003	1.000	1.000	1.000	1.000
	# 15		0.000	0.003	0.000	0.000				
	# 25		0.000	0.000	0.003	0.000				
	# 34		0.003	0.000	0.000	0.000				
Sums			1.000	1.000	1.000	1.000				

Note. See notes in Appendix 1 for *, #, a, b, c, d, and e.