Optimizing Solution Sets in Two-way Controlled Selection Problems

Sun-Woong Kim
Steven G. Heeringa
Peter W. Solenberger

Survey Research Center
Institute for Social Research
University of Michigan

Overview

- Contributions to Controlled Selection
- Two-way Controlled Selection Problem
- Optimal Samples
- Optimal Controlled Selection
- New Public Use SAS-based Software
- Examples
- Further Studies

Contributions to Controlled Sampling

- **Goodman and Kish(1950)**

  Introduction of controlled sampling techniques
  for deep stratification and other highly constrained
  sample selection problems

- **Jessen(1970)**

  Method 1, Method 2, Method 3 in two-way
  and three-way stratification

- **Hess, Riedel and Fitzpatrick(1975)**

  Controlled selection for the Michigan sample of
  general hospitals and patients

- **Causey, Cox and Ernst(1985)**

  Algorithm using transportation theory for two-way
  stratification

- **Rao and Nigam(1990, 1992)**

  Linear programming approach under certain
  probability sampling schemes

- **Sitter and Skinner(1994)**

  Linear programming approach using marginal constraints in two-way or three-way stratification

- **Huang and Lin(1998)**

  Algorithm using network approach in two-way Stratification

Two-way Controlled Selection Problem

- **Two-way stratification desirable in the sample design**

(Hypothetical Example) Bryant et al. (1960)

Expected Sample Sizes($n = 10$)

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

Notation

Row Stratification Factor : R categories
Column Stratification Factor : C categories

$R \times C$ Tabular Array : A

Expected Sample Sizes : a_{ij} , $i = 1, \cdots, R$, $j = 1, \cdots, C$

Possible Samples : B_k, $k = 1, \cdots, L$

Internal Entry of B_k : $b_{ijk} = [a_{ij}]$ or $[a_{ij}] + 1$,
where $[a_{ij}]$ is the integer part of a_{ij}

Optimal Samples

- **Constraints for selection probabilities of samples**

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

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

where B : a set of possible samples,

$p(B_k)$: selection probability of B_k.

- **There may be a number of sets of probability distributions satisfying these constraints.**

- **Consider some measures of closeness between A and B_k.**

Metrics (or Distance Functions) :

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

: the overall distance between A and B_k for RC cells using the Euclidean metric

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

$$, \ 1 < m < \infty$$

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

$$, \; 1 \le p < \infty$$

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

$$= \max \left\{ \left| a_{ij} - b_{ijk} \right| : 1 \le i \le R, 1 \le j \le C \right\}$$

$$, k = 1, \cdots, L$$

: the simplest distance between A and B_k that is measured by only one cell among RC cells

Definition

Optimal Samples :

A few samples having the minimum distance values from d_4 **(or** d_1**)**

Unfavorable Samples :

A few samples having the maximum distance values from d_1 **(or** d_4**)**

Algorithm for Optimal Controlled Selection

PHASE 1. Find a set of possible samples satisfying the following internal and marginal constraints of the tabular array A:

$$|b_{ijk} - a_{ij}| < 1,$$

$$|b_{\cdot jk} - a_{\cdot j}| < 1,$$

$$|b_{i \cdot k} - a_{i \cdot}| < 1.$$

PHASE 2. Solve the following linear programming problem:

Determine $p(B_k)$ that minimize

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

or

$$\phi_2 = \sum_{B_k \in B} d_4(A : B_k)\ p(B_k)$$

subject to

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

$$p(B_k) \geq 0,$$

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

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

PHASE 3. Choose randomly one sample which is a final sampling plan by using the method of cumulative sums or Lahiri(1951)'s method.

Implementation of the Algorithm

- New public use SAS-based software
- Adapting two-phase revised simplex method to solve linear programming problem
- Obtained unique optimal solution set
- Maximizing the selection probabilities of optimal samples and at the same time minimizing the probabilities of unfavorable samples

Example

- Causey et al.(1985)

Further Studies

- Extension to controlled sampling problems more than three dimension

- Development of more effective algorithm for large controlled sampling problem