2002 Proceedings of the Autumn Conference, Korean Statistical Society

The Controlled Selection: Do Algorithms for Optimal Sampling Plan Exist?

Sun Woong Kim[1], Jae Bok Ryu[2], Joon Keun Yum[3]

Abstract

A number of controlled selection methods, which have some advantages for practical surveys in considering controls beyond stratification, have developed throughout the last half-century. With respect to the optimization of sampling plan, it is obvious that we may use optimal controlled selection in preference to satisfactory controlled selection. However, there are currently certain restrictions on the employment of optimal controlled selection. We present further research to improve an algorithm for optimal controlled selection and to develop standard software.

KEY WORDS: Multi-way Stratification, Satisfactory Controlled Selection, Optimal Controlled Selection

1. Introduction

A lot of survey researchers including sampling statisticians and sociologists have commonly insisted on stratified sampling method as a selection procedure to increase the precision of estimates of variables, which are closely related with the stratification variable. However, the method is neither practical nor efficient, not only when stratification is preferred concurrently by several variables, but also when controls beyond stratification are needed from the design point of view. The purpose of controlled selection, which is first suggested by Goodman and Kish (1950), is to achieve some advantages in sample selection that cannot be obtained more readily by other means such as stratified sampling. In general, there are several major advantages obtained from controlled selection. First, particularly when the sample size is less than the number of strata, controlled selection is applicable. Second, controls may be possible to achieve appropriate distribution geographically and to insure adequate sample size for subgroups of the population or domains of study. Third, it is anticipated to secure moderate reductions in the sampling

1) Full-time lecturer, Department of Statistics, Dongguk University. e-mail:sunwk@dongguk.edu
2) Professor, Department of Statistics, Chongju University. e-mail: jbryu@chongju.ac.kr
3) Professor, Department of Statistics, Dongguk University. e-mail: joonkeun@dongguk.edu

errors of a multiplicity of characters simultaneously. Fourth, it is noted that controlled selection does not increase the cost of data collection.

In the social research area, and increasingly in the research area for governmental statistics as well, various controlled selection techniques beyond simple stratification have been introduced for sample surveys throughout the last half-century. For the sake of convenience, we may classify a variety of controlled selection methods into two kinds: satisfactory controlled selection, optimal controlled selection.

In this paper, we give a brief overview with respect to the developments of algorithms for controlled selection, and at the same time discuss the limitations of those algorithms in the applications for practical surveys. In addition, we present further research to be useful in terms of an algorithm for optimal sampling plan and the software available.

2. Satisfactory Controlled Selection

Since Goodman and Kish, some researchers have suggested sampling schemes using quite complicated algorithms to solve two-way or three-way controlled selection problems.

Bryant, Hartley, and Jessen (1960) proposed a method for two-way stratification, but their method depends on the assumption that the rows and columns are independent, although it is not generally accepted. Jessen (1969, 1970) suggested several schemes as alternatives to such as quota sampling, deep stratification, and lattice sampling. But his approaches do not always lead to a solution. Hess, Ridel, and Fitzpatrick (1975) gave the details of how to select the Michigan sample of general hospitals and patients by using the algorithm of Groves and Hess (1975). The Groves-Hess algorithm, which was implemented in a FORTRAN IV program on the mainframe, was not able to obtain the solutions to certain problems. Lin (1992) improved their algorithm and was successful to solve those problems.

Causey, Cox, and Ernst (1985) suggested an algorithm using transportation theory, which is based on the previous paper by Cox and Ernst (1982), in order to solve two-way controlled selection problems. The algorithm has been used even to maximize the overlap between two different sampling schemes or surveys. However, their algorithm has a definite problem because different objective functions in the transportation problems would result in too many solutions, and we should decide which of them is preferable.

Huang and Lin (1998) developed a recursive algorithm using flow in network by adding row subtotals or column subtotals in a original controlled selection problem. Though it is desirable to divide the sampling procedure into two successive steps, the algorithm uses the simple definition of selection probability of each sample.

It is convinced that most methods above use the algorithms just to identify one of many solution sets that meet the constraints of the problem and to minimize the number of feasible samples required to form a set of solution. As a result, the implementation of the algorithms is most unlikely to optimize selection probability of each sample.

Thus, we may regard controlled selection methods using those algorithms as satisfactory controlled selection. In contrast, we may consider the sampling schemes mentioned in the next section as optimal controlled selection because they mainly focus on optimizing of the selection probability to each sample.

3. Optimal Controlled Selection

Rao and Nigam (1990, 1992) introduced lineal programming (LP) approach to solve one-way controlled selection problems under different sampling schemes. Their ideas encouraged Sitter and Skinner (1994) to devise the approach to solve multi-way controlled selection problems by LP. Their approach was comparatively simple to carry out and straightforward for computer programming. But when controlled selection problems have integer margins, it would not be efficient because the expected lack of desirability always becomes zero before finding the solutions.

Kim, Heeringa, and Solenberger (2002) suggested an algorithm to minimize the overall distortion to cells in two-way stratification problems. The approach identifies all possible samples and applies LP to optimize the assignment of the selection probability to each sample by solving the problem of the form:

$$\textit{minimize} \sum_{B_k \in B} d(A:B_k)\, p(B_k)\,,$$

where B denotes a set of possible samples and the weight $d(A:B_k)$ depends on the metric distance measures between the controlled selection problem A and each possible sample B_k. Also, $p(B_k)$ indicates the selection probability of each sample.

Also, the algorithm is easily implemented by using the program which is called SOCSLP (software for optimal controlled selection linear programming). This software now can run on a personal computer.

Although those algorithms are useful to obtain optimal sampling plan for surveys, there are two limitations in actual use. One is that there is no effective algorithm to find solutions of controlled selection problems with more than three-way. The other is that using those algorithms would be much more computationally intensive when we face large controlled selection problems.

4. Concluding Remarks

There are some potential advantages to be gained from the use of optimal controlled selection by LP for solutions of multi-way problems. One of them is to be available for optimal sampling plan. But it would be more accurate to say that the current algorithms for optimal controlled selection are not only inefficient but purely experimental for three-way problems.

Most approaches have been simply extended to three-way or others from two-way problems. In consideration of the complexity of three-way problems, we may differentiate them from two-way problems. And it would be considered that even three-way problems cannot be sometimes solved, although LP is used.

The construction of a modified algorithm for large controlled selection problems is also essential in further study. But it is unlikely to depend on high-speed computers to solve the problems.

On the other hand, the development of standard program to incorporate the algorithms will concentrate on some improvements on the software such as SOCSLP, which is SAS-based program. It may be computationally inexpensive compared to other software.

References

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

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, Rovert M., and Hess, Irene. (1975). An algorithm for controlled selection. Ch. VII in Hess, Irene, Ridel, Donald C., and Fitzpatrick, Thomas B. Probability Sampling of Hospitals and Patients, Second Edition, Ann Arbor: Health Administration Press.

Hess, Irene, Ridel, Donald C., and Fitzpatrick, Thomas B. (1975). Probability Sampling of Hospitals and Patients, Second 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, National University of Singapore.

Jessen, R.J. (1969). Some methods of probability non-replacement sampling. *Journal of the American Statistical Association*, 64, 175-193.

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

Kim Sun-Woong, Heeringa S., and Solenberger P. (2002). Optimizing solutions in two-way controlled selection problems. *Proceedings of the section on survey research methods*, Forthcoming.

Lin T.K. (1992). Some improvements on an algorithm for controlled selection. *Proceedings of the section on survey research methods*, 407-410.

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.

Sitter, R.R., and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20, 1, 65-73.