SOME IMPROVEMENTS ON AN ALGORITHM FOR CONTROLLED SELECTION

Ting-Kwong Lin*, Fred Hutchinson Cancer Research Center
1124 Columbia Street, MP702, Seattle, WA 98104

KEY WORDS: Pattern, contradiction, maximum, minimum

1. Introduction

The sampling technique called controlled selection was first described by Goodman and Kish (1950). It was found to be a very useful sampling technique among practicing survey samplers, especially in selecting first-stage units in multi-stage sampling. Hess, Riedel and Fitzpatrick (1961, 1975) have given a simple illustration on how it can be implemented in the sampling of hospitals in the state of Michigan.

Causey, Cox and Ernst (1985), using transportation theory, have shown for two-dimensional controlled selection problems complete solutions do always exist. They have given an algorithm on how the two-dimensional controlled selection problems can be solved. Computer programs are available to solve the transportation problems.

Groves and Hess (1975) gave a formal algorithm for obtaining solutions to the two-dimensional and the much more complex three-dimensional problems. A computer program written by Groves for the mainframe is available. However, this algorithm may not always yield a solution. There are simple examples that can be solved by hand but which the algorithm fails to solve, even in the two-dimensional situation.

In this paper, we show how the Groves-Hess algorithm can be improved and how further controls which are useful to survey practitioners can be built into the algorithm. Problems the old algorithm could not solve can now be solved by the new algorithm. The computer program written by Groves for the mainframe has been modified and now can run on a personal computer.

* Ting-Kwong Lin was a visiting scientist at Fred Hutchinson Cancer Research Center (1991-1992) and is currently Senior Lecturer, Department of Economics and Statistics, National University of Singapore, Kent Ridge, Singapore 0511. The author is grateful to Irene Hess, Sampling Section, Survey Research Center, Institute for Social Research, The University of Michigan, acting in her capacity as consultant, for her generosity in providing unsolved controlled selection problems, her hand calculation solutions and the computer program source code for the mainframe. The author also thanks Steven G. Heeringa of the same Survey Research Center for providing additional controlled selection problems.

2. THE OLD ALGORITHM

The algorithm as described by Groves and Hess (1975) consists mainly of three phases:

"Phase I:

Begin the pattern construction with the cell having minimum probability. That cell may be one of the individual cells, one of the marginals, or the total. The minimum probability may be the probability for the cell to take on its maximum value or the probability for its minimum value. This minimum probability is also the weight assigned to the pattern. For the first pattern, a minimum probability will be an initial probability from the input data. For the second and later patterns, a minimum probability is really a minimum remaining probability, since the initial probability of a cell value might have been diminished by that value's use in prior patterns.

"Phase II:

Search for cells whose values are implied by virtue of some other cell's value having been selected for use in the current pattern; that is we search for cells that can assume only one of their possible values given the selected cell's value.

"Phase III:

After all implied cells have been identified and the implications entered into the pattern, select a free cell (a cell that may take either its minimum or maximum value) and choose its value for the pattern being constructed. Then return to Phase II.

"After all free cells have been chosen at the end of Phase II, a pattern is complete. The algorithm then adjusts the remaining probabilities and returns to Phase I to begin the next pattern. Eventually all cells will become fixed after completing a pattern. At that point there is only one remaining pattern; it contains the fixed value of every cell and receives the remaining probability as a weight. Then the solution is complete".

3. NO SOLUTION

Two examples are available for which no solution

exists for three-dimensional controlled problems. The first is given by Causey, Cox and Ernst (1985), viz:

	I=1			I=2			I=T		
	K=1	K=2	K=T	K=1	K=2	K=T	K=1	K=2	K=T
J=1	0.5	0.0	0.5	0.0	0.5	0.5	0.5	0.5	1.0
J=2	0.0	0.5	0.5	0.5	0.0	0.5	0.5	0.5	1.0
J=T	0.5	0.5	1.0	0.5	0.5	1.0	1.0	1.0	2.0

(T stands for the marginal total.)

Causey et al. (1985) have given a proof that no solution exists for the above problem. We can also view the above problem in the following way. For I=T we can have the following two possible patterns (the number within the parentheses denotes the weight of the pattern):

	Pattern 1 I=T (0.5)			Pattern 2 I=T (0.5)		
	K=1	K=2	K=T	K=1	K=2	K=T
J=1	1	0	1	0	1	1
J=2	0	1	1	1	0	1
J=T	1	1	2	1	1	2

It can be easily shown that neither of the patterns can be satisfied simultaneously by I=1 and I=2. In fact we can start with any of the cells having minimum probability and end with a contradiction without ever the need to search for another free cell to imply.

For their algorithm, Groves and Hess (1975) have given two possibilities that may give rise to contradictions:

1. The marginal cell has not yet been chosen, and

a) Sum of the minimum values of cells not selected yet together with those selected values is greater than the marginal cell maximum; OR

b) Sum of the maximum values of cells not selected yet together with those selected values is less than the marginal cell minimum.

2. The marginal cell has been chosen, and

a) Sum of the minimum values of cells not selected yet together with those selected values is greater than the selected marginal cell value; OR

b) Sum of the maximum values of cells not selected yet together with those selected values is less than the selected marginal cell value.

The second "no solution" example is given by Hess (1975) as follows:

	I=1			I=2			I=T		
	K=1	K=2	K=T	K=1	K=2	K=T	K=1	K=2	K=T
J=1	0.2	0.2	0.4	0.2	0.4	0.6	0.4	0.6	1.0
J=2	0.4	0.0	0.4	0.6	0.6	1.2	1.0	0.6	1.6
J=3	0.6	0.4	1.0	0.4	0.0	0.4	1.0	0.4	1.4
J=4	1.2	0.6	1.8	1.2	1.0	2.2	2.4	1.6	4.0

She has also given a proof that no complete set of solutions exist. We can again view the problem as above. For I=T the following three patterns are possible with their respective weights given by the numbers in the parentheses:

	Pattern 1 I=T (0.4)			Pattern 2 I=T (0.2)			Pattern 3 I=T (0.4)		
	K=1	K=2	K=T	K=1	K=2	K=T	K=1	K=2	K=T
J=1	1	0	1	0	1	1	0	1	1
J=2	1	1	2	1	1	2	1	0	1
J=3	1	0	1	1	0	1	1	1	2
J=T	3	1	4	2	2	4	2	2	4

In this example, patterns 1 and 2 can be satisfied simultaneously by I=1 and I=2 but pattern 3 cannot. So, when the program is run for the above data, we will have a solution when either pattern 1 or pattern 2 is obtained but ultimately we receive a contradiction as pattern 3 cannot be constructed. We again observe that the contradiction occurred without searching and choosing another free cell.

Thus, one modification to the algorithm is this: if in the construction of a pattern a contradiction arises without even the need for searching for another free cell, the program is terminated with the output statement that most likely there is no complete solution to the problem.

4. SOME IMPROVEMENTS

In the old (Groves-Hess) algorithm, when a contradiction arises the program retracts all implications, setting them free once again; makes other necessary adjustments; and then sets the free choice to its complement and begins again to check for implications. If a second contradiction arises the algorithm stops. In doing this it assumes that both the minimum and maximum cell values of a chosen cell will lead to contradictions in implied values of related cells.

However, the following situation may occur. After setting the free choice to its complement and checking for all its implications, the program then searches for another free choice cell; then only does the contradiction occur. In this situation, it is no longer the same cell that leads to both contradictions. In stopping, the old algorithm removes the possibility that a solution may exist. To allow the algorithm to proceed, it is modified so that if the above situation arises the new free choice cell is set to its complement and the program continues to check for implications.

Even though both the minimum and maximum cell values of a chosen cell lead to contradictions in implied values of related cells, the following situations may arise:

1. At the beginning of the construction of a pattern, there may be more than one cell with minimum remaining probability.
2. There may be more than one free choice cell.

If either of the situations arises, there is a possibility that the selection of a different cell other than the one chosen may lead to a solution. Thus the algorithm is modified such that if either of the above situations arises, the program will begin construction of the pattern all over again. In order to do this we need two arrays to store the remaining probabilities and frequencies before the construction of a pattern begins. Then if either of the above situations occurs, the two arrays that are to contain the newly formed remaining probabilities and frequencies are set to begin with the values in the stored arrays.

5. IMPOSING OTHER CONDITIONS

The motivation for imposing further conditions is given by the following three-dimensional controlled selection problem. The old algorithm could not give a set of complete solutions to the problem. When it was solved by hand calculation, it was found that further conditions were needed in order to obtain the complete solution. Some restrictions have to be imposed on the marginal cells.

I=1

	K=1	K=2	K=3	K=4	K=T
J=1	0.000	0.244	1.005	0.649	1.898
J=2	0.000	0.119	0.500	0.167	0.786
J=3	0.000	0.096	0.372	0.213	0.681
J=4	0.000	0.031	0.362	0.115	0.508
J=5	0.000	0.000	0.000	0.000	0.000
J=6	0.000	0.000	0.000	0.000	0.000
J=T	0.000	0.490	2.239	1.144	3.873

I=2

	K=1	K=2	K=3	K=4	K=T
J=1	0.000	0.000	0.000	0.102	0.102
J=2	0.214	0.000	0.000	0.000	0.214
J=3	0.000	0.213	0.000	0.106	0.319
J=4	0.000	0.092	0.185	0.215	0.492
J=5	0.000	0.000	0.155	0.845	1.000
J=6	0.391	0.000	0.609	0.000	1.000
J=T	0.605	0.305	0.949	1.268	3.127

I=T

	K=1	K=2	K=3	K=4	K=T
J=1	0.000	0.244	1.005	0.751	2.000
J=2	0.214	0.119	0.500	0.167	1.000
J=3	0.000	0.309	0.372	0.319	1.000
J=4	0.000	0.123	0.547	0.330	1.000
J=5	0.000	0.000	0.155	0.845	1.000
J=6	0.391	0.000	0.609	0.000	1.000
J=T	0.605	0.795	3.188	2.412	7.000

In this illustration, we need to put controls over the K dimension. For I=T, J=T and K=1 and K=2, the sum of the initial probabilities of the maximum cell value for these two cells is 1.4. When the sum is equal to or greater than 1, the frequency in each cell should not be simultaneously the minimum value (in this case, zero) for any pattern. For I=T, J=T and K=3 and K=4, the sum of the initial probabilities of the maximum cell value for these two cells is 0.6, which is less than 1. In this case, the frequency in each cell should not be simultaneously the maximum value for any pattern.

In order to incoporate these conditions into the algorithm, we create another array, say, PT which contains the value of 1 if the sum of the initial probabilities of the maximum cell value of the pair of cells is equal to or greater than 1, and contains the value 0 otherwise. This is done before the construction of the first pattern. The first value in the array PT refers to the outcome of the sum of cells for which K=1 and K=2, the second value refers to the outcome of the sum of the cells for which K=3 and K=4.

Thus in the constuction of a pattern, whenever a cell having I=T and J=T is selected or implied, a check is made on whether K is odd or even. If it is odd, the index of the other cell in the pair is K+1. If it is even, the index of the other cell in the pair is K-1. The position in array PT to which the selected or implied cell refers to is determined by integer division of K+1 by 2. For example, for K having the value of 1 or 2 integer division of K+1 by 2 gives a value of 1. This refers to the first value in array PT. A check is made to determine whether the other cell in the pair is fixed or already selected. If it is not, then a check is made of the array PT to determine whether it contains the value of 0 or 1. If the implied or selected cell contains the minimum value and the value in PT is 1, then the other cell is set to its maximum value. If the implied or selected cell contains the maximum value and the value in PT is 0, then the other cell is set to its minimum value. Otherwise, nothing is done to the other cell.

The above control is built into the algorithm and the computer program is modified accordingly. When

it is run with the above controlled selection problem, a complete set of solutions is obtained without any difficulty.

The setting up of the control over the marginal having dimension K is arbitrary. One can easily set up controls over dimensions I or J. Also forming pairs of cells consecutively is only one of the ways of setting up the control.

Another useful way is this. For the cells having indices I=T and J=T, determine the positions of the cells (that is, the value of K) arranged in the order of descending initial probabilities of the maximum cell value of the cells. This information is stored in an array. Next, compute the sum of the first two largest of the initial probabilities of the maximum cell value, and then the next two and so on through all the K values. For each sum, a check is made to determine whether it is equal to or larger than 1. An indicator array keeps track of it.

The rest of the procedure follows in similar line as before. In the construction of a pattern, whenever a cell having I=T and J=T is selected or implied, a check is made on the position of ranking, whether it is odd or even in order to determine the index of the other cell in the pair. The position in the indicator array to which the selected or implied cell refers to is determined in the same way as before. A check is made to determine whether the other cell in the pair is fixed or has already been selected. If neither, a check is then made on whether the indicator array contains the value of 0 or 1. Again, if the implied or selected cell contains its minimum value and the value in the indicator array is 1 then the other cell is set to its maximum value. If the implied or selected cell contains its maximum value and the value in the indicator array is 0, then the other cell is set to its minimum value. Otherwise, nothing is done to the other cell.

The first procedure is usually preferable because survey samplers usually arrange in meaningful order the various classes within stratification variables.

6. CONCLUSION

Though the Groves-Hess algorithm has been improved, we cannot be certain that the new algorithm will solve every two-dimensional controlled selection problem. However, it has solved those that we know of that the old algorithm was not able to solve.

We have shown how to impose additional controls over the marginal cells. Different controlled selection problems may call for different sets of controls. For instance, instead of forming the sum of two cells, one can form the sum of three cells, or even more.

The original computer program, which was written by Groves and runs on the mainframe, has been modified and now can run on a personal computer. This allows greater access to the computer program. The restriction on the size of the dimensions of I, J, K, depends on the size of the memory of the personal computer. For a standard personal computer, the following configuration seems sufficient, viz: I=20, J=40 and K=12.

REFERENCES

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

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,* 2nd ed., Hess, I., Riedel, D. C., and Fitzpatrick, T. B., Ann Arbor, Michigan : Health Administration Press, 82-102.

Hess, I. (1975),"Controlled Selection Problem With No "Perfect" Solution," in *Probability Sampling of Hospitals and Patients,* 2nd ed., Hess, I., Riedel, D. C., and Fitzpatrick, T.B., Ann Arbor, Michigan: Health Administration Press, 159-160.

Hess, I., Riedel, D., and Fitzpatrick, T. B., (1961, 1975), *Probability Sampling of Hospitals and Patients,* 1st ed., 1961, 2nd ed., 1975, Ann Arbor, Michigan: Health Administration Press.

Acknowlegement:

The author wishes to thank Irene Hess, Head of Sampling, Emeritus, Survey Research Center, Institute for Social Research, The University of Michigan, for her careful reviewing of the paper.