METRON - *International Journal of Statistics*
2010, vol. LXVIII, n. 1, pp. 95-103

MARCIN KOZAK – HAI YING WANG

On stochastic optimization in sample allocation among strata

Summary - The usefulness of stochastic optimization for sample allocation in stratified sampling is studied. Three models of stochastic optimization are compared: E-Model, Modified E-model and V-model, recently presented by Díaz-García and Garay-Tápia (Comput. Statistics Data Anal., 3016–3026, 51, 2007), with the classical sample allocation, which distributes the costs among strata in such a way that the variance of an estimator is minimized. To make the comparison, a simulation study was conducted. None of the methods was the most efficient for all cases, but usually the classical allocation was the most efficient, followed by the E-model, quite similar to the former.

Key Words - Stratified sampling; Optimization; Cost allocation; Optimum allocation.

1. Introduction

Sample allocation in stratified sampling has been the subject of many studies (see, *e.g.*, Särndal *et al.* 1992 and citations therein), and it still is. Recently, Díaz-García and Garay-Tápia (2007) have proposed to use stochastic programming to allocate a sample among strata. They rightly noticed that the classical formulation of sample allocation suffers from an unrealistic assumption that an allocation variable is also a survey variable, quite an unlikely situation in practice. For this reason the stratum variances to be used in allocation are unknown and need to be estimated. In practice it is done based on the results of a previous survey or the knowledge of an auxiliary variable, which should be strongly correlated with the survey variable. This problem has also been indicated by many others, including Särndal *et al.* (1992).

In Díaz-García and Garay-Tápia's (2007) paper these unknown variances of the survey variable are replaced with their estimates from a sample, which are random variables. Taking this into account, the authors presented three stochastic programming models to find the optimum allocation, namely the E-

Received June, 2009 and revised March, 2010.

model, Modified E-model and V-model. All these models rely on the unknown stratum variances and, in the case of the last two models, stratum fourth moments. Thus the authors used the results of a previous survey to replace these quantities, as is done in the classical allocation. They showed that stochastic programming provides different allocation from the one the classical method does; they did not show, however, which of the approaches was better. By better allocation we understand allocation that provides more precise estimates of the parameter of interest.

Therefore, this research aims to choose the best allocation among the following: (i) classical allocation (hereafter also called the C-allocation), (2) allocation based on the E-model (the E-allocation), (2) allocation based on the Modified E-model (the ME-allocation), and (4) allocation based on the V-model (the V-allocation). In this paper we will focus only on the allocation that aims to minimize variance of the estimator of the population mean subject to fixed survey costs. We believe, however, that the results and conclusions will refer also to the dual problem, in which a survey cost is minimized subject to variance constraints.

2. Sample allocation

Consider a finite population U of size N subdivided into H non-overlapping strata U_h of sizes $\mathbf{N} = (N_1, ..., N_H)^T$; Y is a survey variable. A problem of interest is to allocate a survey cost among the strata so that the variance of the estimator $\hat{\theta}$ of some parameter θ studied is minimized.

Hereafter we will focus on estimation of the population mean, so $\theta = \bar{Y}$. In this case the variance to be minimized is

$$V\left(\hat{\theta}\right) = \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N}, \tag{1}$$

where $W_h = N_h/N$, $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(y_{ih} - \bar{Y}_h\right)^2$ is the population variance of the study character in the hth stratum, y_{ih} is the Y-value of the ith element in the hth stratum, and $\bar{Y}_h = \sum_{i=1}^{N_h} y_{ih}$.

In classical allocation the following problem is solved:

$$\min_{\mathbf{n}} \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N}$$

$$\text{s.t. } \mathbf{c}^T \mathbf{n} = C - C_0,$$

$$\mathbf{2} \leq \mathbf{n} \leq \mathbf{N},$$

where C is the total permissible survey cost, C_0 is the fixed survey cost, $\mathbf{c} = (c_1, ..., c_H)^T$ is the vector of costs of selecting one unit from the strata, and $\mathbf{2}$ is the H-vector of twos.

Díaz-García and Garay-Tápia (2007) introduced three models of stochastic programming for sample allocation. They can be presented by the following general model

$$\min_{\mathbf{n}} k_1 \left[\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h - 1} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N} \frac{n_h}{n_h - 1}\right]$$

$$+ k_2 \left[\sum_{h=1}^{H} \frac{W_h^4}{n_h (n_h - 1)^2} \left(C_{yh}^4 - S_h^4\right) - \sum_{h=1}^{H} \frac{W_h^2}{N^2} \frac{n_h}{(n_h - 1)^2} \left(C_{yh}^4 - S_h^4\right)\right]^{1/2}$$

$$\text{s.t. } \mathbf{c}^T \mathbf{n} = C - C_0,$$

$$\mathbf{2} \le \mathbf{n} \le \mathbf{N},$$

where $C_{yh}^4 = \frac{1}{N_h} \sum_{i=1}^{N_h} \left(y_{ih} - \bar{Y}_h\right)^4$, $h = 1, \ldots, N_h$, is the fourth moment of Y in the hth stratum, and k_1 and k_2 are nonnegative constants, the sum of which equals 1. For k_1=1 and k_2=0, the model becomes the E-model, for k_1=0 and k_2=1 the V-model, and for other choices the Modified E-model. As Díaz-García and Garay-Tápia (2007) did, in this paper we consider the Modified E-model with $k_1 = k_2 = 0.5$.

To arrive at these models, the asymptotic distributions of sample variances s_h^2 are used, for which both n_h and $N_h - n_h$ need to go to infinity. Thus the E-model is very similar to the C-model, for large populations. If the stratum variances are known, apparently the classical allocation will be the best one because it will reach the minimum variance of the estimator. Nonetheless, since the stratum variances are unknown and their estimates are used instead, we are not able to say which method is the best.

Díaz-García and Garay-Tápia (2007) considered Lagrangian-based allocation, in which the objective function is minimized subject to $\mathbf{c}^T \mathbf{n} = C - C_0$ via the Lagrange multipliers. The values obtained are rounded to integers, an operation that may provide non-optimum allocation. To overcome this difficulty the authors employed nonlinear integer programming. To allocate a sample, in this paper we will use the random search algorithm, recently proved very efficient for the optimum sample allocation (Kozak 2006). Note, however, that for the purpose of this paper it does not make difference which optimization procedure one chooses, provided that it yields optimality for the objective function subject to specified constraints. Our interest lies not in an allocation procedure itself, but in how an allocation problem is stated, namely via the classical approach or the stochastic programming.

3. Simulation study

To compare the four allocations, we conducted a simulation experiment designed as follows. Two populations (Y variables), one with N = 1000 and the other with N = 5000, were generated according to a gamma distribution with shape parameter 0.5. They were then sorted and stratified into H= 3 and 5 strata as given in Table 1, which also presents information about other parameters of the generated populations.

Table 1: *Description of the two artificial populations studied.*

h	N_h	S_h^2	C_h^4
		Population 1 (N = 1000), H = 3	
1	250	2.54×10^{-4}	1.27×10^{-7}
2	500	0.02653	1.49×10^{-3}
3	250	0.79064	6.3019
		Population 1 (N = 1000), H = 5	
1	50	2.33×10^{-7}	9.86×10^{-14}
2	300	7.32×10^{-4}	1.03×10^{-6}
3	300	8.62×10^{-3}	1.45×10^{-4}
4	300	0.1509	0.0582
5	50	1.0445	5.7016
		Population 2 (N = 5000), H = 3	
1	1000	8.72×10^{-5}	1.56×10^{-8}
2	3000	0.0468	5.45×10^{-3}
3	1000	0.6773	6.3181
		Population 2 (N = 5000), H = 5	
1	500	6.18×10^{-6}	7.79×10^{-11}
2	1500	1.34×10^{-3}	3.67×10^{-6}
3	1500	0.0126	3.14×10^{-4}
4	1000	0.0554	6.11×10^{-3}
5	500	0.7156	7.5046

To simulate the randomness of the variances used in the allocation procedures, for each such population and number of strata, 100,000 stratified samples were taken using simple random sampling without replacement. The overall sample size was $n = 0.1N$ and stratum sample sizes were proportionally distributed among strata. Based on each such sample, S_h^2 and C_h^4 were estimated and then used to allocate among strata the sample of size $n = 0.1N$ using the four allocations compared. Stratum costs of selecting one unit in each stratum were assumed to be the same and equal to 1. Then for each allocation the variance of the estimator was calculated using Eq. (1) based on known population values of S_h^2. These variances were compared, and between two allocations the one was more efficient for which the variance was smaller.

The following notes need to be made to make this simulation study correct. Denote the optimal $(n_1, ..., n_H)^T$ by $\mathbf{n}$. The true variance of the estimator in the general case is

$$\begin{aligned} V\left(\hat{\theta}\right) &= V\left(\bar{y}\right) = E\left(V\left(\bar{y}|\mathbf{n}\right)\right) + V\left(E\left(\bar{y}|\mathbf{n}\right)\right) \\ &= E\left(\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N}\right) + V\left(\bar{Y}\right) \\ &= E\left(\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N}\right). \end{aligned}$$

Note, then, that the difference of this equation from equation (1) lies in the expectation only. Equation (1) is in fact $V(\bar{y}|\mathbf{n}) = V(\bar{y}|sample)$ (because the optimal $\mathbf{n}$ is determined by the sample drawn). So for a particular sample this can be seen as a Monte Carlo estimate of the true variance based on this one sample.

In our simulation, we have 100,000 samples, so we have 100,000 Monte Carlo values of $V(\bar{y}|\mathbf{n})$ for a particular allocation method. Then we can treat the mean of these 100,000 values of $V(\bar{y}|\mathbf{n})$ as the estimate of the true variance of the estimator, also for this particular allocation method. Note that these $V(\bar{y}|\mathbf{n})$ are i.i.d., so by the law of large numbers the mean will converge to the expectation, that is, the true variance of the estimator. Since we have 100,000 samples, this estimate of the mean is very close to the true value. Note also that the expressions of the true variances of the estimators are all the same. That is to say, under both E and V models, the variance is still

$$E\left(\sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h S_h^2}{N}\right).$$

Therefore, although we do not compare true variances of the estimators, we compare their Monte Carlo estimates that are obtained based on known true stratum variances S_h^2. For a particular iteration, a sample is drawn to estimate these variances S_h^2 in order to derive the optimum sample sizes for each allocation; based on these random sample sizes we estimate the true variances of the estimator for each allocation (this time we use known stratum variances S_h^2). As we discussed above, these are Monte Carlo estimates of the true variances, so they are comparable among the allocations. The allocation with the smallest variance estimated in that way will be the best for a particular sample. Hence finding which allocation is the best in most situations in these terms may be thought of as finding the most efficient allocation. In this simulation, it seems that we use the same values of survey variable in the first and second survey. But note that we only need the stratum variances in the second survey, and in real life the variances can be assumed more or less stable between a previous survey and the new survey, so it is reasonable to use the same populations.

Tables 2, 3 and 4 summarize the results of the experiment. Table 2 shows that the C-allocation and E-allocation were on average comparably efficient and the two most efficient allocations, although the former reached the optimum allocation most often. Moreover, from Tables 3 and 4 it follows that the C-allocation was in most cases more efficient than the E-allocation. Note, however, that there were cases in which these allocations were less efficient than the ME-allocation and V-allocation. Still these were the C-allocation and E-allocation that were the most efficient both on average and most often. Note also that the results of the ME-allocation and V-allocation were visibly less precise than those of the other two allocations. In addition, for many cases, especially when the populations were stratified into three strata, the ME-allocation and V-allocation provided the same sample sizes from strata and the same variances, showing a close similarity between the allocations.

Interestingly, for $N = 1000$ and $H = 5$, the variability in the variances obtained via all the allocations was very large, especially compared to the other situations studied (note the coefficients of variation and the maximum variances obtained). It resulted from the specificity of the population and its stratification; this variation came mainly from the very small (N_5=50) fifth stratum. According to the proportional allocation rule used to estimate the variance and fourth moment of the survey variable, we draw only five samples from this stratum, so the estimates were not stable. As a result, the number of units each method allocated to this stratum varied dramatically among the 100000 samples. Unfortunately, the variances of the estimator were strongly influenced by this stratum's variance. Having said that, it is worth noting that for this situation the same scenario for the performance of the allocation methods was obtained as for the other situations.

TABLE 2: *Summary statistics of the simulation results obtained for two populations studied via the four allocations compared. Mean, median and maximum values are given for variances divided by the optimum (minimum) variance.*

	Classic	E-model	M E-model	V-model
		Population 1, $H = 3$		
Mean	1.027	1.030	1.055	1.055
Median	1.014	1.014	1.032	1.032
Reached min[(1)]	6207	5520	2633	2633
Max	1.553	1.593	1.635	1.635
CV[(2)]	0.037	0.039	0.058	0.058
		Population 1, $H = 5$		
Mean	1.149	1.129	1.565	1.565
Median	1.052	1.046	1.098	1.098
Reached min	772	270	425	424
Max	4.502	4.510	5.211	5.198
CV	0.274	0.211	0.588	0.589
		Population 2, $H = 3$		
Mean	1.009	1.009	1.063	1.063
Median	1.004	1.005	1.029	1.029
Reached min	1752	2	107	107
Max	1.157	1.157	1.371	1.371
CV	0.011	0.011	0.076	0.076
		Population 2, $H = 5$		
Mean	1.022	1.022	1.089	1.089
Median	1.011	1.011	1.039	1.039
Reached min	28	31	23	24
Max	1.335	1.335	1.512	1.512
CV	0.026	0.026	0.104	0.104

[(1)] Number of times for the corresponding allocation for which the minimum variance of the estimator was reached.

[(2)] Coefficient of variation.

TABLE 3: *Results of the simulation study for* $N = 1000$ *and* $H = 3$ *(below diagonal) and* $H = 5$ *(above diagonal). Number of cases are given for which the column allocation was more efficient (i.e., had smaller variance) than the corresponding row allocation; in brackets, number of cases are given for which the column and row allocations provided the same results.*

	Classic	E-model	M E-model	V-model
Classic		46917 (3163)	27800 (1262)	27805 (1263)
E-model	48823 (30098)		27100 (1306)	27104 (1300)
M E-model	64718 (2139)	59450 (1343)		3085 (93773)
V-model	64718 (2139)	59450 (1343)	0 (100000)	

TABLE 4: *Results of the simulation study for* $N = 5000$ *and* $H = 3$ *(below diagonal) and* $H = 5$ *(above diagonal). Number of cases are given for which the column allocation was more efficient (i.e., had smaller variance) than the corresponding row allocation; in brackets, number of cases are given for which the column and row allocations provided the same results.*

	Classic	E-model	M E-model	V-model
Classic		38465 (10462)	30706 (3)	30711 (3)
E-model	88172 (1878)		31217 (1)	31192 (1)
M E-model	75477 (45)	74680 (23)		16042 (68122)
V-model	75477 (45)	74680 (23)	1 (99998)	

4. CONCLUSION

From the experiment it follows that among the four allocation methods compared, none could be considered the best one for every situation. Nonetheless, the C-allocation was most often the best followed by the E-allocation, a result that should not surprise given the similarity between them. The V-allocation and ME-allocation, the performance of which was very similar, were usually the worst.

The results do not mean that stochastic programming should be discounted in further research on sample allocation among strata. It is possible that for some other situations the allocation based on one of the models of stochastic programming (maybe a model not considered in this paper) might occur to be more efficient than the classical allocation. In other words, there may be some other equivalent deterministic problems that will work more efficiently than the classical method. Note that stochastic optimization offers the opportunity to find many different deterministic problems in sample allocation. Take the Modified E-model, for example. Intuitively, the E-model focuses on minimizing the variance of the estimator of the population mean, while the V-model on minimizing the variation of this variation. The Modified E-model attempts to strike a happy medium, minimizing a linear combination of the variance (the E-model's focus) and the variance's variance (the V-model's focus). Only the case $k_1 = k_2 = 0.5$ was considered in this paper, but better k_1 and k_2 may exist. With their optimal values, the Modified E-model may work more efficiently than both the E-model and V-model. Nonetheless, studying the way of searching for the optimal E-model or other efficient deterministic problems is beyond the scope of this work.

From our results it follows that at the moment we cannot claim that stochastic programming offers allocation that would perform better than the classical allocation method.

REFERENCES

Díaz-García, J. D. and Garay-Tápia, M. M. (2007) Optimum allocation in stratified surveys: Stochastic programming, *Computational Statistics & Data Analysis*, 51, 3016–3026.

Kozak, M. (2006) Multivariate sample allocation: application of random search method, *Statistics in Transition*, 7 (4), 889–900.

Särndal, C. E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag.

MARCIN KOZAK
Department of Experimental Design
and Bioinformatics
Faculty of Agriculture and Biology
Warsaw University of Life Sciences
Nowoursynowska 159
02-776 Warsaw, Poland
nyggus@gmail.com

HAI YING WANG
Academy of Mathematics
and Systems Sciences
Chinese Academy of Sciences
China