21st Century Academic Forum Conference Proceedings
2014 Conference at Berkeley

A New Hybrid Meta-heuristic Approach for Stratified Sampling

Timur KESKİNTÜRK[1], Sultan KUZU[2], Bahadır F. YILDIRIM[3]
1 School of Business, İstanbul University, İstanbul, Turkey
2 School of Business, İstanbul University, İstanbul, Turkey
3 School of Business, İstanbul University, İstanbul, Turkey

Abstract

Stratified sampling is a methodology of dividing members of population into homogeneous subgroups before sampling. The aim of this paper is solving the combined problem of stratification and sample allocation with the hybrid meta-heuristic. Some numerical examples are given and the performance of hybrid meta-huristic is compared with the Kozak's random search (RSM) (2014) and Keskintürk and Er's GA (2007) methods. The results show that the new hybrid meta-heuristic for stratified sampling provides same or better results.

Keywords: Stratified sampling, Stratification, Sample allocation, Hybrid Meta-heuristic

Introduction

Stratified sampling is a methodology in which the elements of a heterogeneous population are classified into mutually exclusive and exhaustive subgroups (strata) based on one or more important characteristics (Cyert & Davidson, 1962). Stratified sampling involves taking a sample without replacement from each subgroup (Hedlin, 2003), and then combining those selected samples from each stratum. An extensive literature is available on the principles of stratified sampling, e.g. Cochran (1977).

One of the main objectives of stratified sampling is to reduce the variance of the estimator and to get more statistical precision than with the simple random sampling (Cochran, 1977). This objective is best achieved when the variability within each stratum is small and the stratum means are different from one another (Cyert & Davidson, 1962).

In this paper, in order to minimize the variance of the estimator, we first propose a hybrid meta-heuristic approach for the determination of stratum boundaries, using proportional, and Neyman allocation of sample elements among the strata. We then show how the hybrid meta-heuristic is used in both stratum boundary determination and sample size allocation. In the application of our proposed hybrid meta-heuristic approach, the total sample size and the number of strata are predetermined.

The paper is organized as follows: In the second section, stratified sampling is discussed. In the third section, we give a brief description of the hybrid meta-heuristic. The fourth section gives computational results, concluding remarks and future research.

Literature Review

STRATUM BOUNDARY DETERMINTION AND THE SAMPLE ALLOCATION METHODS

Khan et al (2009) compare the proposed method with the Dalenius and Hodges' cum$\sqrt{f}$ method (Dalenius & Hodges, 1959) and show how the proposed technique using dynamic programming is effective in determining the optimum strata boundaries. The advantage of the dynamic programming method is that it gives the global minimum of the objective function and it does not require any initial approximate solutions. A numerical example using a hospital population data is presented to illustrate the computational details of the solution procedure. From the experimental results they conclude that the proposed method within our frame work yields a gain in relative efficiency (Khan, Ahmad, & Kahn, 2009).

Brito et al. (2010) suggested that an iterative local search (ILS) meta-heuristic algorithm would obtain a good feasible solution. It is a search-based method that is intended to work for variables with any distribution. They implemented their algorithm on sixteen skewed populations; some real and some simulated, and showed that it produced better solutions than the random search algorithm of Kozak (2004) in most cases (Brito, Ochi, Montenegro, & Maculan, 2010).

Brito et al. (2010) suggested an algorithm based on using minimal path in a graph, and claimed that it guarantees optimum stratification boundaries. They tested the algorithm using real data from the Brazilian Central Statistics Office, and provided the CPU time for the algorithm's implementation; in some cases this was nearly three minutes (Brito, Maculan, Lila, & Montenegro, 2010).

Horgan (2011) suggest a modification, adding empirical rules for determining end points, outliers, take-none and take-all strata in order to improve the efficiency and ensure a feasible set of boundaries (Horgan, 2011).

Er (2012), examines the improvement in the efficiency ratios and stratum boundaries obtained with (Lavallée & Hidiroglou, 1988), Kozak (2004) and Keskintürk and Er's (2007) methods once the initial boundaries are obtained with geometric method (GA_GM). With the stratification of 16 heterogenous

populations that have different properties, higher variance of the estimates or infeasible solutions can be observed. As a result, researchers should be much more rigorous when using geometric method for the initial boundaries in algorithmic methods or else use the modified version of geometric method once the data has very extreme values (Er, 2012).

Kozak (2014), suggested that quite likely genetic algorithms have potential to be a means of very efficient stratification, especially in complex stratification problems.

THE DETERMINATION OF THE STRATUM BOUNDARIES AND SAMPLE ALLOCATION WITH HYBRID META-HEURISTIC METHOD

Several algorithms are derived for constructing stratum boundaries in the literature. The cumulative root frequency method of Dalenius and Hodges (1959) is the most widely used. More recently Lavallée and Hidiroglou's (1988) method of minimizing the sample size for a given level of precision and Gunning and Horgan's (2004) method of equalizing the coefficients of variation have been derived specifically for skewed populations. In the present paper, we adopt the general strategy of minimizing the variance of the estimator V(ystrat)and introduce a GA approach for the determination of stratum boundaries. In order to explore the efficiency of GA approach, we compare its efficiency with the geometric and the cumulative root frequency method. For details of the application of the geometric approach see Gunning and Horgan (2004).

When the question is to allocate the sample size among strata, there are several alternative methods such as equal, proportional, and Neyman allocation (Neyman, 1934; Hess, Sethi, & Balakrishnan, 1966). The equal allocation method is the simplest method where each stratum sample size is the same. With the proportional allocation method, the sample sizes in each stratum are proportional to the size of that stratum. These two methods are efficient and suitable if the variances within the stratum are similar (Cyert, & Davidson, 1962). On the other hand, if the stratum variances differ substantially, as in for example highly skewed populations, the Neyman allocation method should be used. This method is based on the principle of sampling fewer elements from homogeneous strata and more elements from strata with high internal variability. In this study sampling costs are assumed to be equal for all strata.

The following notation will be used throughout the paper:

Y	:	stratification variable
N	:	population size
N	:	sample size
H	:	number of strata
Nh	:	number of elements in stratum h (h=1,...,H)
nh	:	sample size in stratum h
σ_{yh}^2	:	variance of the elements in stratum h
$\bar{Y}_h$	:	mean of elements in stratum h
$\bar{y}_{strat}$	:	estimated mean in stratified sampling

Estimated mean and the variance of the estimated mean $\bar{y}_{strat}$ is given by Cochran (1977)

$$\bar{y}_{strat} = \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{N} \qquad (1)$$

$$S^2_{\bar{y}_{strat}} = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{\sigma^2_{yh}}{n_h}\left(1-\frac{n_h}{N_h}\right) \quad (2)$$

where the variance of each stratum is assumed known and calculated as follows:

$$\sigma^2_{yh} = \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1) \quad (3)$$

where Yih is the i th element in the h th stratum.

It is assumed in Eq. (3) that Nh > 1; clearly when Nh =1, there will be no deviation. Sample sizes n1,...,nH are allocated with proportional, and Neyman allocation methods and these methods are briefly summarized below:

Proportional allocation method:

$$n_h = n\frac{N_h}{N}, \quad h = 1, 2, ..., H \quad (4)$$

Neyman allocation method:

$$n_h = n\frac{N_h\sigma_{yh}}{\sum_{i=1}^{H} N_i\sigma_{yi}}, \quad h = 1, 2, ..., H \quad (5)$$

In this paper we used genetic algorithm selection, mutation operators to determine the stratum boundaries and sample allocation. Our Hybrid Meta-heuristic uses GA operators and local search together. Keskinturk and Er (2007) used binary and real-valued chromosome structure in their work. In this paper instead of the binary and real-valued chromosome structure, we use real values to determine sampling and stratum boundaries. Operators are modified by these new structures. Thus, the chromosome length is reduced.

The principle of our genetic algorithm based hybrid meta-heuristic is given as follow:

Start	***:***	*Generate random initial generation*
Fitness Function	***:***	*Evaluate the fitness of each chromosome*
Local Search	***:***	*Local search for boundaries and sample size*
Selection	***:***	*Select the better individuals for the next generation*
Mutation	***:***	*With a mutation probability, mutate new offspring*
Loop	***:***	*If stopping criterion is not reached go to fitness function*
Stop	***:***	*Return the best solution in current generation*

In this paper, real-valued encoding is used for boundary determination sample allocation. Encoding for simple example of the combined stratification problem is illustrated in Figure 1.

Figure 1 Encoding for stratification

467	485	77	3	1

The number of gene is equal to the number of strata (H). The last gene represents the upper boundary of stratum which comes before the final stratum. Final stratum boundary equal to population size so it is not shown on any gene. The right side of the chromosome represent the sample size of each strata.

After constructing the initial generation, each chromosome is evaluated by an objective function, referred to as a fitness function, from which a fitness value is derived. In our algorithm the fitness value is the variance of estimator in stratified sampling denoted as Eq. (2) that must be minimized through the iteration process.

The next step after determination of the fitness values is the local search process. Local search process briefly summarized below:

Since there is a two sub-chromosome that first part represents stratum boundaries and the second one shows sample sizes of stratums, we have used two different operations for local search. For the boundaries we select one of them randomly and move it to the right or to the left. To avoid infeasible solutions, we use a control operator which checks overlap of stratums and the size of population that should be equal or greater than 2. For samples, we choose two of them randomly to change their sizes reciprocally.

The next step after local search is the selection process. Selection determines whether chromosomes will survive in the next generation or not, according to their fitness values. Chromosomes with a better fitness value have more chance to survive. In this paper roulette wheel selection, one of the most popular, is used (Keskintürk & Er, 2007).

In this paper random exchange mutation, which is usually preferred on a permutation chromosome, is used. Random exchange mutation is applied so that two positions are selected at random along the chromosome and the genes contained in these positions are exchanged. The reason for using this mutation operator is to guarantee the number of strata be held fixed after mutation.

A Numerical Example

For the comparison we examined (iso487) dataset. Iso487 example consists of 487 Turkish manufacturing firms from the first 500 largest corporations belonging to the Istanbul Chamber of Industry (ISO) in year 2004. ISO collects data on net sales, gross value added, net profit, etc. of its members and publishes the data of the first 500 corporations annually. Same dataset used by Kozak (2014) and Er (2012) for the comprasion in their papers.

Iso487 example is divided into 2, 3, 4, 5, and 6 strata. For 2, 4, and 5 strata cases, the total sample size is 80. In order to allocate the total sample size into 3 and 6 strata evenly with the equal allocation method Keskintürk and Er (2007) increased the sample sizes to 81 and 84 for iso487 example. For comprasion we used same sample sizes into 3 and 6 strata. Er (2012) used iso487 dataset and divided into 2, 3, 4, 5, and 6 strata. All strata sample size is 100. For comprasion we used same sample size and run algoritm for 100 sample.

We implement our proposed hybrid meta-heuristic algorithm using Matlab programming language on a PC (Pentium 4 CPU 3.00 GHz, 512MB RAM). In order to compare the efficiency of the methods the

variance of the estimator given with Eq. (2) is calculated. Relative efficiencies (%) of hybrid meta-heuristic algorithm to other methods for the iso487 population used by Keskinturk and Er (2007), calculated as ratios of the variance of the estimator from the hybrid meta-heuristics to the variance of the estimator from the other methods. Table 1 and Table 2 presents relative efficiencies for the iso487 population.

Table 1 Efficiency of the estimators for stratification examples obtained with GA, RSM and hybrid meta-heuristic

	Keskintürk and Er (2007)			**Kozak (2014)**		**Hybrid MH**
H	***Proportional***	***Neyman***	***GA***	***Proportional***	***Neyman***	
2	0,11765	1,00000	1,00000	0,13263	1,00000	1,00000
3	0,07426	0,93029	1,00000	0,07426	1,00000	1,00000
4	0,04745	0,65572	1,00000	0,05285	1,00000	1,00000
5	0,04038	0,51796	0,92173	0,04216	0,99990	1,00000
6	0,02511	0,41054	0,98082	0,02659	0,99999	1,00000

Table 2 Efficiency of the estimators for stratification examples obtained with ga_gm and hybrid meta-heuristic

	Er (2012)	**Hybrid MH**
H	***GA_GM***	
2	-	-
3	1,00000	1,00000
4	1,00000	1,00000
5	1,00000	1,00000
6	0,99800	1,00000

GA, RSM, GA_GM and Hybrid Meta-heuristic results of stratum sizes, which correspond to stratum boundaries and sample sizes for all of the numerical examples, are reported in Table 3 and 4.

Table 3 Stratum boundaries for the iso487 example with GA, RSM and Hybrid Meta-heuristic

		Keskintürk and Er (2007)						**Kozak (2014)**				**Hybrid MH**	
		Neyman		***Proportional***		***GA***		***Neyman***		***Proportional***			
H		***N***	***n***	***N***	***N***	***N***	***n***	***N***	***n***	***N***	***n***	***N***	***n***
2	*1*	485	79	442	35	442	35	442	35	477	78	442	35
	2	2	1	45	45	45	45	45	45	10	2	45	45
3	*1*	467	77	357	25	351	26	351	26	467	77	351	26
	2	18	3	83	9	99	18	99	18	18	3	99	18
	3	2	1	47	47	37	37	37	37	2	1	37	37
4	*1*	460	75	141	1	251	13	251	13	431	70	251	13
	2	19	3	236	22	126	12	126	12	43	7	126	12
	3	6	1	61	8	68	13	68	13	11	2	68	13
	4	2	1	49	49	42	42	42	42	2	1	42	42
5	*1*	393	64	13	1	225	14	162	5	381	62	163	5
	2	74	12	217	14	130	13	129	7	81	13	129	7
	3	13	2	41	3	67	13	86	8	16	3	85	8
	4	5	1	35	4	34	9	62	12	7	1	62	12
	5	2	1	58	58	31	31	48	48	2	1	48	48
6	*1*	352	60	113	4	158	7	158	7	351	60	158	7
	2	87	15	89	1	113	7	107	6	90	15	108	6
	3	29	5	71	2	81	7	86	8	29	5	85	8
	4	12	2	120	10	49	8	43	5	9	2	43	5
	5	5	1	30	3	40	9	45	10	6	1	45	10
	6	2	1	64	64	46	46	48	48	2	1	48	48

Table 4 Stratum boundaries for the iso487 example with GA_GM and hybrid meta-heuristic

		Er (2012) *GA_GM*		Hybrid MH	
H		***N***	***n***	***N***	***n***
3	*1*	312	23	312	23
	2	120	22	120	22
	3	55	55	55	55
4	*1*	229	14	229	14
	2	128	13	128	13
	3	74	17	74	17
	4	56	56	56	56
5	*1*	163	8	163	8
	2	129	11	129	11
	3	85	12	85	12
	4	54	13	54	13
	5	56	56	56	56
6	*1*	158	10	158	10
	2	108	9	108	9
	3	85	10	85	10
	4	42	7	42	7
	5	39	9	38	8
	6	55	55	56	56

This paper shows that our hybrid meta-heuristic algorithm improved the efficiency ratios of Keskintürk and Er (2007) , Er (2012) and Kozak's (2014) methods. We plan to apply our method to different problems and also to multivariate stratification.

References

Cyert, R. M., & Davidson, H. J. (1962). Statistical sampling for accounting information. Englewood Cliffs, N.J.: Prentice-Hall.

Cochran, W. G. (1977). Sampling techniques (3d ed.). New York: Wiley.

Hess, I., Sethi, V.K. & Balakrishnan, T.R., (1966). Stratification: a practical investigation. Journal of American Statistical. Association. 61 (313), 74–90.

Bretthauer, K.M., Ross, A., & Shetty, B., (1999). Nonlinear integer programming for optimal allocation in stratified sampling. European Journal of Operation Research, 116(3), 667–680.

Rao, P.S.R.S., (2000). Sampling Methodologies with Applications. Chapman & Hall/CRC Press,Washington DC.

Orhunbilge, N., (2000). Sampling Methods and Hypothesis Tests. (2nd ed.) Avcıol BasimYayin, Istanbul, Turkey. (In Turkish).

Hedlin, D., (2003). Minimum variance stratification of a finite population. Social Statistics Research Centre Methodology Working Paper, M03/07. URL (http://eprints.soton.ac.uk/7796/01/ssrc-workingpaper-m03-07.pdf).

Khan, M.G.M., Ahmad, N. & Kahn, S. (2009), Determining the optimum stratum boundaries using mathematical programming, Journal of Mathematical Modelling and Algorith, 8(4), 409–423 DOI 10.1007/s10852-009-9115-3.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition 6(5), 797–806.

Brito, J., Ochi, L., Montenegro, F., & Maculan, N.(2010), An Iterative Local Search Approach Applied To The Optimal Stratification Problem, International Transactions In Operational Research, 17(6), 753–764, DOI: 10.1111/j.1475-3995.2010.00773.x.

[11] Brito, J., Maculan, N. Lila, M. & Montenegro, F. (2010), An Exact Algorithm For The Stratification Problem With Proportional Allocation, Optim Lett 4, 185–195, DOI 10.1007/s11590-009-0157-2.

Horgan, J. M. (2011), Geometric Stratification Revisited, Int. Statistical Inst.: Proc. 58th World Statistical Congress, Dublin (Session STS058), 3319-3328.

Keskintürk, T. & Er, Ş¸ . (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. Computational Statistics and Data Analysis, 52(1), 53–67.

Er, Ş. (2012), Comparison of the efficiency of the various algorithms in stratified sampling when the initial solutions are determined with geometric method. International Journal of Statistics and Application, 2(1), 1-10, DOI: 10.5923/j.statistics.20120201.01.

Lavallée, P. & Hidiroglou, M. (1988). On the Stratication of Skewed Populations. Survey Methodology, 14(1), 33-43.

Dalenius, T. & Hodges Jr., J.L. (1959). Minimum variance stratification. Journal of American Statistical Associaton. 54(285), 88–101.

Gunning, P. & Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. Survey Methodology 30(2), 159–166.

Neyman, J., (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of Royal Statistical Society, 97(4), 558–625.

Kozak, M., (2014). Comparison of Random Search Method and Genetic Algorithm for Stratification. Communications in Statistics - Simulation and Computation 43(2), 249-253.