Hacettepe Journal of Mathematics and Statistics
Volume 45 (5) (2016), 1519 – 1524

Moving towards an optimal sample using VNS algorithm

Priyaranjan Dash*† and Samrat Hore‡

Abstract

In almost all random sampling schemes, we adopt different sampling designs with an objective of obtaining a better representative sample (optimal sample) for the population. Application of different randomization techniques were adopted for providing a supportive basis for this. Now the question arises, whether the final sample selected, on which all our efforts are utilized, from the population is an optimal sample or not? No where we are checking about the *optimality of this sample, i.e.,* whether this sample is the best one or there exists any other sample which is more optimal than the selected one satisfying all the constraints. In all these procedures, we only assume but, nowhere we are establishing a guarantee about the achievement of such a representative sample. The present paper emphasizes on achieving an optimal sample by using *variable neighborhood search (VNS)* technique.

Keywords: Optimal Sample, Representative Sample, Variable Neighborhood Search.

2000 AMS Classification: 62D05

Received : 16.03.2015 *Accepted* : 10.09.2015 *Doi* : 10.15672/HJMS.20158712906

1. Introduction

In any sample survey, we first develop a frame, which emphasizes on specifying the sampled population identical with the target population lacking any kind of ambiguity there on. A sample plays a role of centripetal force in sampling theory literature. An optimum sample is always desirable and fetches attention at all phases because a poor sample ruins the entire effort of the survey whatever attention may be put to other aspects. We put our entire effort in sampling theory to develop methods of sample selection i.e. to get an optimum sample and to draw inferences on the principles of specified precision and minimum cost. In this connection, two rivalry methods of selection

*Department of Statistics, Tripura University, Email: prdashjsp@gmail.com
†Corresponding Author.
‡Department of Statistics, Tripura University, Email: sam.stat724@gmail.com

Table 1. All possible samples selected by SRSWOR scheme

Sample No.	Sample Units $(y_1,\ y_2)$	Sample Mean $\overline{y}$	$(\overline{y}-\overline{Y})^2$	
1	$y_1=5,\ y_2=3$	4	0	(Minimum)
2	$y_1=5,\ y_3=6$	5.5	2.25	
3	$y_1=5,\ y_4=2$	3.5	0.25	
4	$y_1=5,\ y_5=4$	4.5	0.25	
5	$y_2=3,\ y_3=6$	4.5	0.25	
6	$y_2=3,\ y_4=2$	2.5	2.25	
7	$y_2=3,\ y_5=4$	3.5	0.25	
8	$y_3=6,\ y_4=2$	4	0	(Minimum)
9	$y_3=6,\ y_5=4$	5	1	
10	$y_4=2,\ y_5=4$	3	1	

of a sample came into existence: (1) *random selection* and the other one is (2) *purposive (non-random) selection*. Jensen (1926) [10], Gini and Galvani (1929) [3], Neyman (1934) [12] advocated about these methods of selection. But, all of these based on the hope that the sample we get is a representative one. Since, our desire lies on getting an optimum sample (as a proper subset of the target population) whose characteristics $\hat{\Phi}_y = \widetilde{y}(y_1, y_2, \cdots, y_n)$ under study are almost similar with the population characteristic $\Phi_y = \widetilde{Y}(y_1, y_2, \cdots, y_N)$, when we have a sample of size n from the population of size N to infer about the variable y. Unfortunately, an optimum sample does not exist and even if it exists, it is very difficult, even not possible to identify it. In this regard Godambe(1955) [4], Hege (1965) [8], Hanurav (1966) [7] had given significant contributions. The following hypothetical example will clear this idea.

1.1. Example. Consider a finite population with $N = 5$ and $n = 2$. When the population values are known to us (say) $y_1 = 5, y_2 = 3, y_3 = 6, y_4 = 2, y_5 = 4$.So, we can have $\binom{5}{2} = 10$ different possible SRSWOR samples in total. (We are not emphasizing here regarding WR and WOR samples, as both the schemes are indifferent for large samples (Freedman, 1977 [2]). We have to estimate the population mean $\Phi_y = \overline{Y}$. We now calculate the sample means for these as follows. From the Table 1, it indicates that the sample number 1 and 8 are optimum samples and the sample numbers 2 and 6 are poor samples on the basis of the value of the expression $||\hat{\Phi}_y - \Phi_y|| = (\overline{y} - \overline{Y})^2$.
When we use equal probability scheme to select a sample, in that case all the samples are equally likely of being selected. Alternatively, if we use PPSWR sample, sample 2 is more likely to be selected than the others. So, in all the cases, we are not selecting an optimum sample. It emphasis the individual units to be present in the sample for an optimum sample.

The above result encourages to design a sampling scheme, which will guide us at each step of selection of the units for moving towards optimality. However, we have to keep in view about the cost incurred for selecting the sample.

2. An Overview of VNS Algorithm

Mladenović and Hansen (1997) [11] used the variable neighborhood approach for solving the vehicle routing problems. Variable neighborhood search is the systematic change of neighborhood within a possibly randomized local search algorithm yields a simple and effective metaheuristic for combinatorial and global optimization (Hansen and Mladenović, (1999, 2001) [5, 6]). Contrary to the other metaheuristics based on local search

methods, VNS does not follow a trajectory but explores increasingly distant neighborhoods of the current solution, and jumps from this solution to a new one, if and only if an improvement has been made. In this way, favorable characteristics of the current solution (e.g., many variables are already at their optimal value), will often be kept and used to obtain promising neighboring solutions. Moreover, a local search routine is applied repeatedly to get from these neighboring solutions to local optima. This kind of VNS algorithm has recently been successfully applied in the field of design of experiments by finding optimum allocation of experimental units with known predictors into two treatment groups (Hore et al., 2014 [9]).

3. An Optimal Sample Using VNS Algorithm

Let x be the auxiliary variable closely related to the study variable y. The corresponding parametric function of interest for x is defined as

$$\Phi_x = \widetilde{X}(x_1, x_2, \cdots, x_N).$$

Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ to be a selected sample of size n over the design $\mathscr{P}$ from N units, gives the sample observation vector for auxiliary variable only. To obtain an optimal sample observation vector over the selected one, an iterative method through neighborhood search algorithm is propsed in this article. Let us denote the neighborhood of α by $\mathcal{N}(\alpha)$ and the neighborhood construction of the corresponding α is derived in Step 3. The proposed algorithm is framed by using the concept of VNS algorithm.
Another sample of size k ($\leq n$) also has been selected randomly from remaining $(N-n)$ population units gives the sample observation vector

$$\beta = \{\beta_1, \beta_2, \cdots, \beta_k\}$$

with the relation $\beta^{(j)} = \beta^{(j-1)} - \{\beta_j\}$; $\beta^{(0)} = \beta$, $j = 1, 2, \cdots, k$. We repeat the following steps until a *stopping condition* is met.

Step 1: Start with $j = 1$. Select the j^{th} unit from $\beta^{(j-1)}$.

Step 2: Choose the initial sample $\alpha^{(0)}$ and calculate the value of $\hat{\Phi}_x = \Phi_{\alpha^{(0)}}$ and the corresponding value of the objective function

$$V\left(\alpha^{(0)}\right) = ||\Phi_{\alpha^{(0)}} - \Phi_x|| \quad \text{(say)}.$$

Step 3: The neighborhood of $\alpha^{(0)}$ is constructed as

$$\mathcal{N}\left(\alpha^{(0)}\right) = \left\{\alpha^{(0)}_{(i)}, \beta_j : \ i = 1, 2, \cdots, n\right\},$$

where the $\alpha^{(0)}_{(i)}$ is the sample vector of size $(n-1)$ and it is constructed as,

$$\alpha^{(0)}_{(i)} = \alpha^{(0)} - \{\alpha_i\}, \quad i = 1, 2, \cdots, n.$$

Step 4: Consider all the allocations in $\mathcal{N}(\alpha^{(0)})$ and compute $V(\alpha')$, for all $\alpha' \in \mathcal{N}(\alpha^{(0)})$. Find the minimum objective function with respective sample $\alpha^{(0')}$, denoted as

$$\alpha^{(0')} = arg\ min\left\{V(\alpha'), \text{for all } \alpha' \in \mathcal{N}\left(\alpha^{(0)}\right)\right\}$$

Step 5: If $V(\alpha^{(0')}) < V(\alpha^{(0)})$, choose the next improved sample to be $\alpha^{(1)} = \alpha^{(0')}$. Otherwise select $\alpha^{(1)} = \alpha^{(0)}$.

Step 6: *(Moving towards optimality)* Replace $\alpha^{(0)}$ by $\alpha^{(1)}$, j by $j+1$ and start the algorithm again from **Step 1**.

Step 7: *(Stopping Condition)* Continue repeating the above steps until all k units are examined one by one or $\beta^{(k)} = \phi$.

Table 2. Sizes of 15 Large United States Cities (in 1000's) in 1920 (x_i) and 1930 (y_i)

Sl No.	1	2	3	4	5	6	7	8
x_i	76	138	67	29	381	23	37	120
y_i	80	143	67	50	464	48	63	115
Sl No.	9	10	11	12	13	14	15	
x_i	61	387	93	172	78	66	60	
y_i	69	459	104	183	106	86	57	

4. Empirical Illustrations

Table 2 gives the number of inhabitants (in 1000's) of 15 cities of United States in the years 1920 and 1930. Cochran(2011) [1], p. 151-152.

In order to estimate the total number of inhabitants $Y = \sum_{i=1}^{15} y_i$ in these cities in the year 1930, we select an initial sample of 4 cities using SRSWOR scheme. Let the selected cities are 3, 6, 7 and 12. So, we have

$$\alpha^{(0)} = \{\alpha_1 = 67, \alpha_2 = 23, \alpha_3 = 37, \alpha_4 = 172\}.$$

Again we select another sample of size 3 from the remaining $15 - 4 = 11$ cities as 1, 5 and 8. Thus, $\beta = \{\beta_1 = 76, \beta_2 = 381, \beta_3 = 120\}$.

Step 1: Start with $j = 1$. Select the first unit from β as $\beta_1 = 76$.

Step 2: Choose the initial sample $\alpha^{(0)} = \{67, 23, 37, 172\}$ and calculate the value of

$$\begin{aligned} V\left(\alpha^{(0)}\right) &= ||\Phi_{\alpha^{(0)}} - \Phi_x|| = ||\hat{X} - X|| = (N\overline{x} - X)^2 \\ &= (1121.25 - 1788)^2 = 444555.6 \quad \text{(say)}. \end{aligned}$$

Step 3: To find the neighbors of $\alpha^{(0)}$, we consider 4 samples, each of size 3, as

$$\alpha^{(0)}_{(1)} = \{23, 37, 172\},\ \alpha^{(0)}_{(2)} = \{67, 37, 172\},\ \alpha^{(0)}_{(3)} = \{67, 23, 172\},\ \alpha^{(0)}_{(4)} = \{67, 23, 37\}.$$

The neighborhood of $\alpha^{(0)}$ is constructed as

$$\begin{aligned} \mathcal{N}\left(\alpha^{(0)}\right) &= \left\{\left\{\alpha^{(0)}_{(i)}, \beta_j\right\}, \quad i = 1, 2, \cdots, n\right\} \\ &= \{\{23, 37, 172, 76\}, \{67, 37, 172, 76\}, \\ &\qquad \{67, 23, 172, 76\}, \{67, 23, 37, 76\}\}. \end{aligned}$$

Step 4: Here,

$$\begin{aligned} \alpha^{(0')} &= argmin\left\{||V(\alpha')||,\ \forall \alpha^{(0')} \in \mathcal{N}\left(\alpha^{(0)}\right)\right\} \\ &= argmin\left\{(N\overline{x} - X)^2\right\} = \left\{\alpha^{(0)}_{(2)}, \beta_1\right\} = \{67, 37, 172, 76\} \end{aligned}$$

Step 5: Here $V\left(\alpha^{(0')}\right) = 219024 < V\left(\alpha^{(0)}\right)$.

So, the corresponding units $\{3rd,\ 7th,\ 12th,\ 1st\}$ gives a better representation of the population than the initial sample.

Step 6: *(Moving towards optimality)* We replace

$\alpha^{(0)}$ by $\alpha^{(1)} = \{67, 37, 172, 76\}$.

Table 3. Sample values for x and y

Sample units:	u_1	u_3	u_7	u_{12}
y values:	80	67	63	183
x values:	76	67	37	172

Table 4. Estimate of Variance of Different Estimators of Y and their Relative Gain in Efficiency (RGE).

Sample Type	Different Estimators	Sample Type	Est. of Variance	Est. RGE of (a) to (b)
SRSWOR	$\hat{\Phi}_1 = N\overline{y}$	a	178470.4	39.91885
		b	249713.7	
Ratio Estimator	$\hat{\Phi}_2 = \frac{\overline{y}}{\overline{x}}X$	a	11669.76	142.9719
		b	28354.24	
Regression Estimator	$\hat{\Phi}_3 = N\left[\overline{y} + \hat{b}_{yx}\left(\overline{X} - \overline{x}\right)\right]$	a	10256.91	83.44258
		b	18815.54	
Ratio Est. in DS^*	$\hat{\Phi}_4 = \frac{\overline{y}}{\overline{x}}\overline{x}'$	a	80989.51	42.17238
		b	115144.7	
Regression Est. in DS^*	$\hat{\Phi}_5 = N\left[\overline{y} + \hat{b}_{yx}\left(\overline{x}' - \overline{x}\right)\right]$	a	80163.82	37.74642
		b	110422.8	

a. Optimum sample b. Traditional sample *DS: Double Sampling

Step 7: *(Stopping Condition)* Again, proceeding in the previous manner, after two such iterations, we can get $\beta^{(3)} = \phi$ and the corresponding sample units $\{3rd,\ 7th,\ 12th,\ 1st\}$ is the optimum sample as it has the smallest argument.

Here, we get the optimum sample as $s = \{u_3, u_7, u_{12}, u_1\}$. Now, we can only study these units for getting y values. The Table 3 gives the values of x and y for this optimum sample.

If an equivalent two phase sample is selected from this population with $n+k$ units to estimate the unknown population mean of auxiliary variable $\overline{X}$ and a second phase sample of size n units out of $n+k$ units, then in the present example (with $n = 4, k = 3$), observed sample values for x are 67, 23, 37, 172, 76, 381, 120. Table 4 gives the estimated standard errors of different estimators and relative gain in efficiency for estimating population total (Y) in adopting proposed optimal sample to the usual (initial) sample using different estimators under SRSWOR scheme.

5. Conclusion

In all traditional sample survey literature, we are emphasizing on improving the sampling design or the estimators there on by efficiently utilizing the auxiliary information but neglecting the representativeness of the selected sample. The present paper utilizes the readily available auxiliary information in order to get an *improved sample*, viewed by a *better representation of the population*, to estimate the parameters of interest. The proposed procedure provides, by sacrificing a little cost to study the auxiliary variable, a safeguard for arriving at a better representative sample employing *variable neighborhood search (VNS) technique*. It does not require any kind of abstract knowledge about the population values like population correlation coefficient (ρ) between y and x as in case of ratio and regression methods of estimation. The *optimality* of the final selected sample is established by the relative gain in efficiency to the traditional sample, on the basis of a

numerical study, shown in Table 4. Therefore, the proposed VNS algorithm for selecting an optimal sample strongly advocates about its better representativeness.

Acknowledgements

The authors are very much thankful to the referees for their valuable timely suggestions for the improvement of the paper.

References

[1] Cochran, W.G. (2011). Sampling Techniques. *Wiley India Pvt. Ltd.*, New Delhi, 3rd edition.

[2] Freedman, D. (1977). *A remark on the difference between sampling with and without replacement.* Journal of the American Statistical Association, **72** (359), 681.

[3] Gini, C. and Galvani, L. (1929). *Di una applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione* . Annali di Statistica, **6** (4), 1-107.

[4] Godambe, V. P.(1955). *A unified theory of sampling from finite populations.* Journal of the Royal Statistical Society. Series B (Methodological), **17** (2), 269-278.

[5] Hansen, P. and Mladenović, N. (1999). *An introduction to variable neighborhood search in: Metaheuristics, Advances and Trends in Local Search Paradigms for Optimization.* S. Voss et al., eds, Kluwer, Dordrecht.

[6] Hansen, P. and Mladenović, N. (2001). *Variable neighborhood search: Principles and applications.* European Journal of Operational Research, **130** (3), 449-467.

[7] Hanurav, T. V. (1966). *Some aspects of unified sampling theory.* Sankhya, **28**, 175-204.

[8] Hege, V.S. (1965). *Sampling designs which admit uniformly minimum variance unbiased estimators.* Calcutta Statistical Association Bulletin, **14**, 160-162.

[9] Hore, S., Dewanji, A. and Chatterjee, A. (2014) : *Design issues related to allocation of experimental units with known covariates into two treatment groups.* Journal of Statistical Planning and Inference, **155**, 117-126.

[10] Jensen, A. (1926). *Report on representative method in statistics.* Bulletin of the International Statistical Institute, **22** (1), 381-439.

[11] Mladenović, N. and Hansen, P. (1997). *Variable neighborhood search.* Computers & Operations Research, **24**, 1097-1100.

[12] 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 the Royal Statistical Society, **97** (4), 558-625.