STRATIFICATION AND INTEGER ALLOCATION IN PRESENCE OF NONRESPONSE

Mushtaq A. Lone[1], Shakeel A. Mir , and Imran Khan
Division of Agri-Statistics SKUAST-K

ABSTRACT
In this article Warner (1965) randomized response model is used to determine the allocation of sample sizes in stratified sampling design. The problem is formulated as Nonlinear Programming Problem *(NLPP*) with non linear cost function. The formulated problem is solved using Branch and Bound method and the results are obtained through LINGO.

KEYWORDS.- Stratification, Optimum allocation, Randomized response technique, Nonresponse, Branch and Bound method, NLPP.

MSC. 62D05

RESUMEN
En su articulo Warner (1965) es usado el modelo de respuestas aleatorizadas para determinar la afijación de los tamaños de muestra en el diseño muestreo estratificado aleatorio. El problema es formulado como un Problema de Programación No-lineal *(NLPP*) con función de costo no-lineal. El formulado problema es resuelto mediante el método de Ramificación y Acotación y los resultdos son obtenidos usando LINGO.

PALABRAS CLAVE . Estratificación, Afijación Optima, Técnica de Respuestas Aleatorizadas, No-respuesta, Ramificación y Acotación, NLPP.

1. INTRODUCTION

In survey sampling, at early stages the attention was focused primarily on the methods of sampling and estimation of reduction of sampling errors subjected to budgetary and other practical constraints. In practical it was found that some other unavoidable sources of errors also dominate the scenario. Emphasis was then laid on studying the effects of such errors called nonsampling errors, which includes nonresponse and measurement errors or response errors. Nonresponses means failure to response and in actual surveys there are number of causes attributed to the incidence of nonresponse. (i) non-coverage (ii) not-at-home (iii) unable to answer (iv) refusal of the respondent to be interviewed. Now it is not enough to identifying the causes of nonresponse, one has to take remedial measures with additional funds and time. Generally it has been recommended to cover a part of the nonrespondent group at least a second attempt should be made. Thus the problem of non response is very serious and it has received attention of sampling practitioners in almost every field of actual surveys. The technique was first developed for the surveys in mailing the questionnaires and next to personal interview to a subsample of the non respondents. Hansen and Hurwitz (1946) presented the classical non response theory for eliciting responses from a subsample of the non respondents. Khare (1987) discussed the allocation of samples in the presence of nonresponse.The problem of optimum allocation in stratified random sampling for univeriate population is well known in sampling literature; see for example Cochran (1977) and Sukhatme *et al*.(1984): Lone *et al*.(2017) used Gradient Projection method for determination of optimal allocation of stratified sampling design. Khan *et al*.(2008) uses multivariate stratified sampling in presence of nonresponses to determine the problem of optimum allocation and the optimum sizes of subsamples to various strata which is formulated as a Nonlinear Programming Problem (NLPP): Lone *et al*.(2017) uses Branch and Bound Method to obtain integer solution in stratified sampling design. Shabbir and Gupta (2005) and Holbrook and Krosnicks (2010) used randomized response technique for conducting experiments and reported a procedure for reducing social desirability response bias by

[1] lonemushtaq11@gmail.com

allowing respondents to report secretly whether they voted. The significant contribution is by Greenberg *et al.* (1969), Moors (1971), Mangat and Singh (1990), Mangat (1994), Lone *et al*. (2018)and Singh *et al*. (2000): Hong *et al.* (1994): In this article Warner (1965) randomized response technique (RRT) is used for estimating the π_A proportion of respondents belonging to particular class.

2. **PROBLEM FORMULATION**

Consider the population of size N , divided into L strata of sizes N_h (h=1, 2, . . ., L), such that $\sum_{h=1}^{L} N_h = N$. Let. we assume that samples are drawn independently from each stratum and a simple random sample of size n_h be obtained from the h^{th} stratum to measure the response . Warner suggested method (RRT) for conducting such survey, which provides us reasonably accurate estimate of the quantity in which we are interested (e.g. the proportion of employees which are not satisfied): Suppose the respondents in a group A are sensitive revealing the same to an unknown interviewer. Instead of asking directly to the respondents "Do you belong to group A", the interviewer offers a pack of cards (which consists of two types, in type one it is written that i belong to group A and in other type it written that i do not belong group A) to the respondents. The respondent is instructed to choose a chard on random basis from the pack of cards without revealing to the interviewer what card they has drawn to answer "yes" or "no" truthfully to indicate whether they agrees to the statement written on the cards or not and $p_h \neq 0.5$ is set by the researcher. Now it is impossible for the interviewer to find out the group to which employee belongs and guaranteed about the privacy of the employee. Due to this enhancement of privacy it is expected that non-cooperation with the interviewer will decrease and the respondent will able to provide answers truthfully.

An individual respondent in the sample of any h^{th} stratum is selected to use the randomization device which consists of sensitive and its negative questions, with probability p_h and $1-p_h$ respectively.. If question (i) is selected then its probability is p_h and if question (ii) is selected then its probability will be $1-p_h$.

In the random sample size n, r denotes the number of 'yes' answers. The proportion ∅ is binomial estimate of 'yes' answers and is given by $\phi = \frac{r}{n}$

If the questions are answered truthfully, Cochran (1977) gave the relation between ϕ and π_A in the population is

$$\begin{aligned}\phi &= P_h\pi_A + (1-P_A)(1-\pi_A)\\ &= (2P_h-1)\pi_A + (1-\pi_A)\end{aligned} \tag{1}$$

Where π_A = proportion of respondents possess the attribute A.

The maximum likelihood estimate (MLE) of π_A is given by

$$\hat{\pi}_A = \frac{\phi-(1-P_h)}{(2P_h-1)} \quad ; \text{where } p_h \neq 0.5 \tag{2}$$

.

Let $W_h = \frac{N_h}{N}$ denote the stratum weight. The unbiased estimate of $\hat{\pi}_A$ is given by

$$\hat{\pi}_A = E\left[\hat{\pi}_{Ah}\right] = E\left[\sum_{h=1}^{L} W_h\hat{\pi}_{Ah}\right] = \sum_{h=1}^{L} W_h E(\hat{\pi}_{Ah}) = \sum_{h=1}^{L} W_h\pi_h = \pi_A \tag{3}$$

Thus we have

$$f_h = P_{Ah}p_{Ah} + (1- P_{Ah})(1- p_{Ah}) = (2P_{Ah} - 1)p_{Ah} + (1- p_{Ah})$$

$P_{Ah} \neq 0.5$,

Where,

P_{Ah} = probability that question (i) is selected from h^{th} stratum.

π_{Ah} = proportion of respondents that possess the attribute A from h[th] stratum.

ϕ_h = proportion of 'yes' answers from h[th] stratum and $\hat{\phi}_h$ is binomial estimate of ϕ_h.

The sampling variance of $\hat{\pi}_{Ah}$

$$V(\hat{p}_{Ah}) = \sum_{h=1}^{L} W_h^2 V(\hat{p}_{Ah}) = \sum_{i=1}^{L} W_h^2 \left[\frac{\hat{p}_{Ah}(1-\hat{p}_{Ah})}{n_h} + \frac{p_{Ah}(1-p_{Ah})}{n_h(2p_{Ah}-1)^2}\right] \quad P_{Ah} \neq 0.5, \qquad (4)$$

The problem of optimum allocation involves determining the sample sizes say $n_1, n_2, \ldots, n_h$ that minimize the total variance subjected to sampling cost. In RR model the interviewer have to approach the population units selected in the sample to get the answers from the each stratum. In each stratum the interviewer have to travel from unit to unit to contact them, this involves additional cost to the overhead cost. Here we consider the non linear cost function $C = C^o + \sum_{h=1}^{L} c_h n_h + \sum_{h=1}^{L} t_h \sqrt{n_h}$. It has been indicated by the mathematical studies that costs are better represented by the expression $\sum_{h=1}^{L} t_n \sqrt{n_h}$.

Where, C^o = Overhead cost, c_h = per unit cost of measurement in h^{th} stratum, C = available fixed budget for the survey Equation (4) can be written as.

$$V(\hat{p}_{Ah}) = \sum_{h=1}^{L} W_h^2 V(\hat{p}_{Ah}) = \sum_{i=1}^{L} W_h^2 A_h \qquad (5)$$

Where,

$$A_h = \frac{\hat{\pi}_{Ah}(1-\hat{\pi}_{Ah})}{} + \frac{p_{Ah}(1-p_{Ah})}{(2p_{Ah}-1)^2}, \text{and } K_h = A_h W_h^2 \quad ; \quad h = 1, 2, \ldots, L.$$

The problem of optimum allocation can be formulated a Nonlinear Programming Problem (NLPP) as.

$$\left.\begin{aligned} &\textit{Minimize} \quad V(\hat{p}_{Ah}) = \sum_{i=1}^{L} \frac{W_h^2 A_h}{n_h} \\ &\textit{subject to} \\ &\sum_{h=1}^{L} c_h n_h + \sum_{h=1}^{L} t_h \sqrt{n_h} \leq c^o \\ &2 \leq n_h \leq N_h \qquad \textit{and } n_h \text{ integers}, h = 1, 2, \ldots, L \end{aligned}\right\} \qquad (6)$$

The above NLPP (6) can be solved using nonlinear integer programming technique. In this article, Branch and Bound method of Land and Doig (1960) has been used to determine the optimal sample size in presence of nonresponse.

3. NUMERICAL ILLUSTRATION

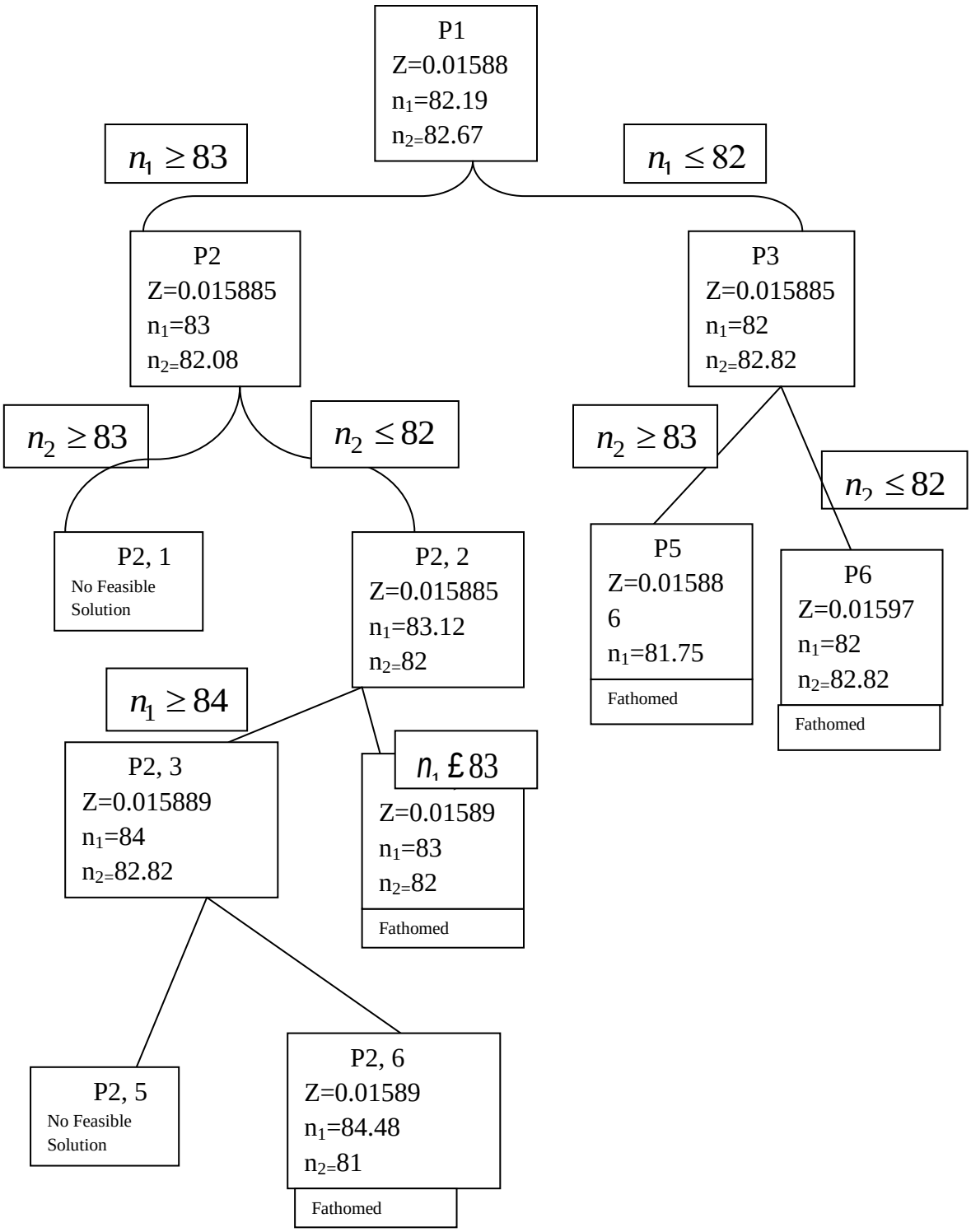

Fig. 1. Various nodes of *NLPP* (7)

Assume that C (available budget) =4500 units including C^O and C^o = 500 units (overhead cost):Therefore c^o=4500-500=4000 units. Also we assume that 400 and 700 are stratum sizes respectively as given in above table for h = 1, 2, N=400+700=1100 .The values of A_h and $A_h W_h^2$ are calculated as given in table below;

Table 1. stratified population with two strata Table 2. values of A_h and $A_h W_h^2$

Stratum(h)	N_h	W_h	π_{Ah}	P_{Ah}	C_h	t_h
1	400	0.3	0.2	0.6	15	11
2	800	0.7	0.4	0.7	20	16

Stratum(h)	A_h	$A_h W_h^2$
1	6.16	0.55
2	1.55	0.76

Substituting these values of the parameters into NLPP (6) we have

$$\begin{aligned} &\textit{Minimize} \quad V(\hat{p}_{Ah}) \text{ or } V(\hat{y}_{Ah}) = \frac{0.55}{n_1} + \frac{0.76}{n_2} \\ &\textit{subject to} \\ &15n_1 + 20n_2 + 11\sqrt{n_1} + 16\sqrt{n_2} \le 4000 \\ &2 \le n_1 \le 400 \\ &2 \le n_2 \le 700 \text{ and } n_1, n_2 \text{ integers}, h = 1,2. \end{aligned} \tag{7}$$

After solving equation (7), we get optimal solution as n_1=82.19 and n_2=82.67 and optimal value is 0.01588.Thus, instead of rounding the non integer solution to the nearest integer value. A Branch and Bound method of Land and Doig (1960) is used.
Various nodes for the *NLPP* (7) utilizing table1 and table2, are presented below in fig 1.

4. DISCUSSION

Since $n_1$and n_2 are required to be the integers, we branch problem P_1 into two sub problems P_2 and P_3 by introducing the constraints $n_1 \le 82$ and $n_1 \ge 83$ respectively indicated by the value n_2=82.19 which lies between 83 and 84. This process of replacing a problem by two sub problems is called branching. The solution of these two sub problems can be obtained using LINGO software as shown in Fig.1. Now again, we branch problem P_3 into two sub problems P_5 and P6 by introducing the constraints $n_2 \le 82$ and $n_2 \ge 83$ respectively indicated by the value n_2=82.82 which lies between 82 and 83. Since node problems P_5 and P_6 are is fathomed with integer value and we are left with P_2. we branch problem P_2 into two sub problems $P_{2,1}$ and $P_{2,2}$ by introducing the constraints $n_2 \le 82$ and $n_2 \ge 83$ respectively indicated by the value n_2=82.02 which lies between 83 and 84. since node $P_{2,1}$ has no solution, so we left with $P_{2,2}$ similarly, we branch problem $P_{2,2}$ into two sub problems $P_{2,3}$ and $P_{2,4}$ by introducing the constraints $n_1 \le 83$ and $n_1 \ge 84$ respectively indicated by the value n_1=83.12 which lies between 83 and 84. Thus the problem $P_{2,4}$ is fathomed. Now, we are left with problems $P_{2,3}$. Again, we branch problem $P_{2,3}$ into two sub problems $P_{2,5}$ and $P_{2,6}$ by introducing the constraints $n_2 \le 82$ and $n_2 \ge 83$ respectively indicated by the value n_1=82.82 which lies between 82 and 83. Similarly the same procedure is adopted until all notes becomes fathomed. Thus we stope at these sub

problems , because problem $P_{2,5}$ has no solution and $P_{2,6}$ has more objective value than previous and is fathomed.

Now, all the terminal nodes are fathomed. The feasible fathomed node with the current best lower bound is node $P_{2,4}$. Hence the solution is treated as optimal solution. The optimal value is n_1=83 and n_2=82 and optimal solution $Minimize \quad V(\hat{\pi}_{Ah}) = 0.01589480.$ The total cost under this allocation is 3130.101<4000.

5. CONCLUSION

Thus we conclude that when a non integer solution of NLPP is obtained then instead of rounding the non inter solution to the nearest integer value. Branch and Bound method provides an integer value. Because some times in real situations rounding off non integer to the nearest integer value becomes impractical and the solution become infeasible.

RECEIVED: MARCH, 2019.
REVISED: FEBRUARY, 2020.

REFERENCES

[1] COCHRAN, W.G. (1977): **Sampling Techniques**. *3rd Edition*. John Wiley and Sons, New York.
[2] GREENBERG B. G, ABUL-ELA, ABDEL-LATIF.A, SIMMONS W. R, and HORVITZ D. G. (1969): The unrelated question RR model-theoretical frame-work. **J. Amer. Statist. Assoc** 64. 520-539.
[3] HANSEN, M.H. and HURWITZ, W.N. (1946): The problem of non response in sample surveys**. Journal of American Statistical Association,** 40. 517-529.
[4] HOLBROOK. A. and KROSNICKS, J. (2010): Measuring voter turnout by using the Randomized response technique Evidence calling into question the Method's validity. **Public Opinion Quarterly,** 74, 328-343.
[5] HONG H. *et al*.(1994): A stratified randomized response technique. **Korean Journal of Applied Statistics**,7.141-147.
[6] KHAN, M. G. M, KHAN, E. A. and AHSAN, M. J. (2008): Optimum Allocation in Multivariate Stratified Sampling in Presence of Non-Response. **Journal of the Indian Society of Agricultural Statistics,** 62, 42-48**.**
[7] KHARE, B. B. (1987): Allocation in stratified sampling in presence of non- response. **Metron** 45, 213-221.
[8] LAND, A.H and DOIG, A.G.(1960): An Automatic method for solving discrete Programming problems. **Econometrica**, 28, 479-520.
[9] LINGO 13.0, LINDO *inc.ltd.*
[10] LONE, M. A., MIR. S. A., MAQBOOL. S. and BHAT. M. A . (2015): An integer solution using Branch and Bound Method in Multi-objective stratified sampling design. **International Journal of Advanced Scientific and Technological Research**, 4, 172-181.
[11] LONE, M. A., MIR. S. A., KHAN, I. and WANI, M. S. (2017): Optimal allocation of stratified sampling design using Gradient Projection method. **Oriental Journal of Computer Science and Technology**, 10,11-17.
[12] LONE, M. A., MIR, S. A. and KHAN, I. (2018): Allocation problem in presence of nonresponse. a mathematical programming approach. **Int. J. Mathematics in Operational Research**, 12, 413-421.
[13] MANGAT, N. S. (1994): An improved randomized response strategy. **J. Roy.Statist. Soc**, *B* 56, 93-95.
[14] MANGAT, N. S and SINGH R. (1990): An alternative randomized response procedure. **Biometrika**, 77, 439-442.
[15] MOORS, J. J. A. (1971): Optimization of unrelated question randomized response model. **Journal of American Statistical Asssociation** 66, 627-629.
[16] SHABBIR, J. and GUPTA, S. (2005): Optimal allocation in stratified randomized response model. **Pakistan Journal of Statistics and Operations Research** 1, 15-22
[17] SINGH, S, SINGH R and MANGAT, N. S. (2000): Some alternative strategies to Moor's model in randomized response model, **J. Statist. Plann. Infer**.83, 243-255.
[18] SUKHATME, P.V, SUKHATME, B.V, SUKHATME, S. and ASOK, C. (1984): **Sampling Theory of Surveys with Applications**. Iowa State University Press, Ames and Indian Society of Agricultural Statistics, New Delhi
[19] WARNER, S. L., (1965): Randomized response. a survey technique for eliminating evasive bias. **J. Amer. Statist. Assoc**. 60, 63-69.