A Nonlinear Programming Problem Applying LINGO

[1] Tanveer Ahmad Tarray, [2] Muzafar Rasool Bhat
[1, 2] IUST, J&K India

Abstract – In this paper the problem of stratified random sampling where randomized response technique is used in presence of non-response. Misreporting and refusal to respond are two main causes of misleading results from direct or open surveys when we ask about sensitive issues directly. Moral support for child abuse, drug usage, racism, induced abortion and illegal activities are in those issues to which individual either misreport or refuse to respond. Generally, individuals do not want to unveil their true status and want to keep it confidential because of the stigma attached with the question asked. Warner (1965) introduced a randomized response model to estimate a population proportion for sensitive attribute .The problem is formulated as a Nonlinear Programming Problem (NLPP) and is solved using Branch and Bound method. Also the results are formulated through LINGO.

Index Terms – Randomized response technique, Stratified random sampling, Sensitive attribute, Branch and Bound method.

1. INTRODUCTION

The randomized response (RR) technique to procure trustworthy data for estimating the proportion of a population possessing a sensitive attribute "A" (say) was first introduced by Warner (1965). Warner's modal draws respondents using simple random sampling with replacement from the population. It requires the interviewee to give a "Yes" or "No" answers either to the sensitive question or to its negative depending on the outcome of a randomizing device not reported to the interviewer. This pioneering work of Warner's (1965) led to modifications and developments in various directions. Feeling that the cooperation of the respondent might be further enhanced if one of the two questions referred to a non – sensitive, innocuous attribute, say Y, unrelated to sensitive attribute A, Horvitz et al. (1967) proposed an unrelated question randomized response model (U-model) with known π_y the proportion of non – sensitive attribute Y. Theoretical details for this model were given by Greenberg et al. (1969). This technique has generated much interest in the statistical literature since the publication of Warner's randomized response model. Subsequently, several other workers have proposed different RR strategies for instance, see the review oriented references like Fox and Tracy (1986) and Tarray and Singh (2016). Some times in survey sampling certain amount of information is known about the elements of the population to be studied. For instance, information may be available on the geographical location of the area, e.g. if it is an inner city, a suburban or a rural area. Census information will provide a wealth of other information about the area, for instance, its population at the previous census, its rate of population change, the proportion of its population employed in manufacturing, or the proportion of its population with different origins. Supplementary information of this type can be used either at the design stage to improve the sample design, or at the analysis stage to improve the sample estimators, or both the essence of stratification is the classification of population in to sub-population or strata, based on some supplementary information and then the selection of separate samples from each of the strata. The benefits of stratification derive from the fact that the sample sizes in the strata are controlled by the sampler, rather than being randomly determined by the sampling process after the strata sample sizes are made proportional to the strata population sizes.

Hong et al. (1994) suggested a stratified randomized response technique that applied the same randomization device to every stratum. Stratified random sampling is usually achieved by dividing the population into non – overlapping groups called strata and selecting a simple random sample from each stratum. An RR technique using a stratified random sampling gives the group characteristics related to each stratum estimator. Also, stratified samples protect a researcher from the possibility of obtaining a poor sample. Under Hong et al.'s (1994) proportional sampling assumption, it may be easy to derive the variance of the proposed estimator, however, it may cause a high cost because of the difficulty in obtaining a proportional sample from some stratum. Kim and Warde (2004) presented a stratified RR technique using an optimal allocation which is more efficient than a stratified randomized response technique that using a proportional allocation.

A primary focus of this paper is the implementation of Stratified randomized response technique (RRT) usingTarray and Singh (2016) question randomized response strategy.

2. PROBLEM FORMULATION

In the proposed models, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement (SRSWR) in each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. Let N be the total number of units in the population. The population is partitioned in to k non –

overlapping groups such that $N = \sum_{h=1}^{k} N_h$, where N_h is number of units in the hth stratum *(h=1,2,...,k)*. In stratified population, $\pi_S = \sum_{h=1}^{k} w_h \pi_{Sh}$, where π_{Sh} is the proportion of respondents with the sensitive trait in the sample from stratum h and $w_h = N_h / N$. The proposed randomized response device consists of two – urns: Urn-I contains *Mh1* balls, out of which *rh1* balls bearing the statement, (a) "I belong to the group A", and the remaining *(Mh1-rh1)* balls are blank with no statement on them. Urn-II contains *Mh2* balls, out of which *rh2* balls bearing the statement, (b) "I do not belong to group A" and the remaining *(Mh2-rh2)* balls are blank with no statement on them. Each respondent selected in the sample from hth stratum is instructed as follows: if he /she belong to the group A, then he / she draws balls using without replacement sampling from the Urn-I until he/she gets *th1 (< rh1)* balls bar the statement (a), and report the total number of balls, say X_h , drawn by him / her. Thus X_i follows a negative hypergeometric distribution given by

$$P(X_h = x_h \mid M_{h1}, r_{h1}, t_{h1}) = \frac{\binom{x_h - 1}{t_{h1} - 1}\binom{M_{h1} - x_h}{r_{h1} - t_{h1}}}{\binom{M_{h1}}{r_{h1}}}$$

$$, x_h = t_{h1}, (t_{h1} + 1), ..., (M_{h1} - r_{h1} + t_{h1})$$

If he/she does not belong to the group A, then he / she draws balls without replacement sampling from the Urn – II until he /she gets *th2 (<rh2)* balls bearing the statement (b), and reports the total number of balls, say Y_h, drawn by him / her. Thus Y_h also follows a negative hypergeometric distribution, but with different parameters, given by

$$P(Y_h = y_h / M_{h2}, r_{h2}, t_{h2}) = \frac{\binom{y_h - 1}{t_{h2} - 1}\binom{M_{h2} - y_h}{r_{h2} - t_{h2}}}{\binom{M_{h2}}{r_{h2}}}$$

$$, y_h = t_{h2}, (t_{h2} + 1), ..., (M_{h2} - r_{h2} + t_{h2})$$

Let n_h denote the number of units in the sample from stratum *h* and *n* denote the total number of units in the samples from all strata so that $n = \sum_{h=1}^{k} n_h$ .Under the assumption that reports are made by the respondents truthfully, then the distribution of the observed response Z_{hi} is given by

$$Z_{hi} = \begin{cases} X_{hi} & if\ i \in A \\ Y_{hi} & if\ i \in A^c \end{cases} .$$

An unbiased estimator of the proportion π_{Sh} is given by

$$\hat{\pi}_{Sh} = \frac{\frac{(r_{h1}+1)(r_{h2}+1)}{n_h}\sum_{i=1}^{n_h} Z_{hi} - t_{h2}(r_{h1}+1)(M_{h2}+1)}{\{t_{h1}(M_{h1}+1)(r_{h2}+1) - t_{h2}(M_{h2}+1)(r_{h1}+1)\}} .$$

The variance of the estimator $\hat{\pi}_{Sh}$ is given by

$$V(\hat{\pi}_{Sh}) = \frac{\pi_{Sh}(1-\pi_{Sh})}{n_h} + \frac{\{\pi_{Sh} t_{h1} A_h + (1-\pi_{Sh}) t_{h2} B_h\}}{n_h (r_{h1}+2)(r_{h2}+2)\{t_{h1}(r_{h2}+1)(M_{h1}-1) - t_{h2}(r_{h1}+1)(M_{h2}+1)\}}$$

where

$$A_h = (M_{h1}+1)(M_{h1} - r_{h1})(r_{h1} + 1 - t_{h1})(r_{h2}+1)^2 (r_{h2}+2),$$

$$B_h = (M_{h2}+1)(M_{h2} - r_{h2})(r_{h2} + 1 - t_{h2})(r_{h1}+1)^2 (r_{h1}+2).$$

Thus the unbiased estimator of $\pi_S = \sum_{h=1}^{k} w_h \pi_{Sh}$ is given by

$$\hat{\pi}_S = \sum_{h=1}^{k} w_h \hat{\pi}_{Sh}$$

$$= \sum_{h=1}^{L} w_h \frac{\left\{\frac{(r_{h1}+1)(r_{h2}+1)}{n_h}\sum_{i=1}^{n_h} Z_{hi} - t_{h2}(r_{h1}+1)(M_{h2}+1)\right\}}{\{t_{h1}(M_{h1}+1)(r_{h2}+1) - t_{h2}(M_{h2}+1)(r_{h1}+1)\}} .$$

Since the selections in different strata are made independently, the variance of the unbiased estimator of $\hat{\pi}_S$ is given by

$$V(\hat{\pi}_S) = \sum_{h=1}^{k} w_h^2 V(\hat{\pi}_{Sh}) \quad (1)\ (1)$$

$$= \sum_{h=1}^{k} \frac{w_h^2}{n_h}\left[\pi_{Sh}(1-\pi_{Sh}) + \frac{\{\pi_{Sh} t_{h1} A_h + (1-\pi_{Sh}) t_{h2} B_h\}}{(r_{h1}+2)(r_{h2}+2)\{t_{h1}(r_{h2}+1)(M_{h1}+1) - t_{h2}(r_{h1}+1)(M_{h2}+1)\}}\right]$$

$$V(\hat{\pi}_S) = \sum_{h=1}^{k} \frac{w_h^2}{n_h} V_h \quad ,$$

where

$$V_h = \left[\pi_{Sh}(1-\pi_{Sh}) + \frac{\{\pi_{Sh} t_{h1} A_h + (1-\pi_{Sh}) t_{h2} B_h\}}{(r_{h1}+2)(r_{h2}+2)\{t_{h1}(r_{h2}+1)(M_{h1}+1) - t_{h2}(r_{h1}+1)(M_{h2}+1)\}} \right]$$

The problem of optimum allocation involves determining the sample size say n_1 , n_2 ,..., n_h that minimize the total variance V ($\hat{\pi}_S$) subject to sampling cost. The sampling cost function is of the form $\sum_{h=1}^{k} c_h n_h$, the cost is proportional to the size of the sample within any stratum. But when we move from stratum to stratum, the cost per unit i.e. c_h may vary. Under RRT 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 contract them, this involves additional cost to the overhead cost. Also, we define $c^0 = C - C^0$.

The linear cost function is $C = C^0 + \sum_{h=1}^{k} c_h n_h$,

where C^0 is the over head cost, c_h is the per unit cost of measurement in hth stratum, *C* is the available fixed budget for the survey

Equation (1) can be rewritten as

$$\mathrm{V}(\hat{\pi}_S) = \sum_{h=1}^{k} \frac{w_h^2}{n_h} V_h$$

where

$$V_h = \left[\pi_{Sh}(1-\pi_{Sh}) + \frac{\{\pi_{Sh} t_{h1} A_h + (1-\pi_{Sh}) t_{h2} B_h\}}{(r_{h1}+2)(r_{h2}+2)\{t_{h1}(r_{h2}+1)(M_{h1}+1) - t_{h2}(r_{h1}+1)(M_{h2}+1)\}} \right]$$

(2)

The problem of optimum allocation can be formulated as a non linear programming problem (NLPP) for fixed cost as

$$\left.\begin{aligned} &Minimize\ V(\hat{\pi}_S) = \sum_{h=1}^{k} \frac{w_h^2}{n_h} V_h \\ &subject\ to \sum_{h=1}^{k} c_h n_h \le c^0 \\ &\qquad 2 \le n_h \le N_h \quad and\ n_h\ integers, h = 1,2,...k \end{aligned}\right\} \tag{3}$$

The above NLPP can be solved using non linear integer programming technique. We can now apply Branch and Bound method to determine the optimal sample size in presence of non response. This method consists of two strategies, alternatively followed till the desired solution is obtained. One strategy consists in Branch a problem in to two sub problems and the other in solving each of the two sub problems to obtain the minimum or suitable lower bound of the objective function.

Let us now determine the solution of problems (3) by ignoring upper and lower bounds and integer requirements. The Lagragian function may be

$$\varphi = \sum_{h=1}^{k} \frac{w_h^2}{n_h} V_h + \lambda \left[\sum_{h=1}^{k} c_h n_h - c^0 \right] \tag{4}$$

Differentiating (4) with respect to c_h and equate to zero, we get

$$\frac{\bar{V}\varphi}{\bar{V}n_i} = 0 \Rightarrow n_h = \frac{w_h \sqrt{V_h}}{\sqrt{c_h}\sqrt{\lambda}} \tag{5}$$

Again differentiating (4) with respect to λ in equation to zero, we get

$$\frac{\bar{V}\varphi}{\bar{V}\lambda} = 0 \Rightarrow c^0 = \sum_{h=1}^{k} c_h n_h \tag{6}$$

Solving (5) and (6), we have

$$\sqrt{\lambda} = \sum_{h=1}^{k} c_h \frac{w_h \sqrt{V_h}}{\sqrt{c_h}} \tag{7}$$

Substituting (7) in (5), we have

$$n_h = \frac{w_h\sqrt{V_h}}{\left[\sum_{h=1}^{k} c_h \frac{w_h\sqrt{V_h}}{c^0\sqrt{c_h}}\right]\sqrt{c_h}} \Rightarrow \frac{c^0 w_h \frac{\sqrt{V_h}}{\sqrt{c_h}}}{\left[\sum_{h=1}^{k} w_h\sqrt{V_h}\right]\sqrt{c_h}} \qquad (8)$$

The Branch and Bound method will require the solution of sub problems in which some of the n_i are fixed. Suppose that at rth node, the fixed values of n_h are for $h \in I\,r$. Then the required Lagrangian function is

$$\varphi = \sum_{h\in 1\,r}^{k} \frac{w_h^{\,2}}{n_h} V_h + \lambda\left[\sum_{h\in 1\,r}^{k} c_h n_h - c^0\right] \qquad (9)$$

Further, differentiating (9) with respect to n_h and equating to zero, we have

$$n_h = \frac{w_h\sqrt{V_h}}{\sqrt{\lambda}\sqrt{c_h}} \qquad (10)$$

At rth node,

$$\sum_{h\in 1\,r}^{k} c_h n_h = c^0 - \sum_{h\in 1\,r}^{k} c_h n_h$$

$$\Rightarrow \sqrt{\lambda} = \frac{c^0 - \sum_{h\in 1\,r}^{k} c_h n_h}{\sum_{h\in 1\,r}^{k} \sqrt{c_h} w_h \sqrt{V_h}} \qquad (11)$$

After simplification , we get formula for r th node as

$$n_h = \frac{\left(c^0 - \sum_{h\in 1\,r}^{k} c_h n_h\right)\frac{\sqrt{V_h} w_h}{\sqrt{c_h}}}{\sum_{h\in 1\,r}^{k} \frac{\sqrt{V_h} w_h}{\sqrt{c_h}}} \qquad (12)$$

where $1\,r$ is the set of indices which have been fixed at the rth node.

3. NUMERICAL ILLUSTRATION

Stratum (i)	N_i	T_i	w_i	r_i	π_{Si}	M_i	c_i
1	400	5	0.3	4.5	0.08	14	15
2	800	5	0.7	9	0.13	12	20

Table 1: The stratified population with two strata

To judge the performance of the proposed a numerical example is presented to illustrate the formulation of the problem.

Assuming that C (available budget) = 4500 units including c0 and c0 =500 units (overhead cost). Therefore c0 =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 *Vi* and $V_i w_i^2$ are calculated as given in table below.

Table 2: Calculated values of V_i and $V_i w_i^2$

Stratum (i)	V_i	$V_i w_i^2$
1	10.0059	0.9005
2	10.0454	4.9222

Substituting the above calculated values of the parameters into (3) non linear programming problem NLPP, we have

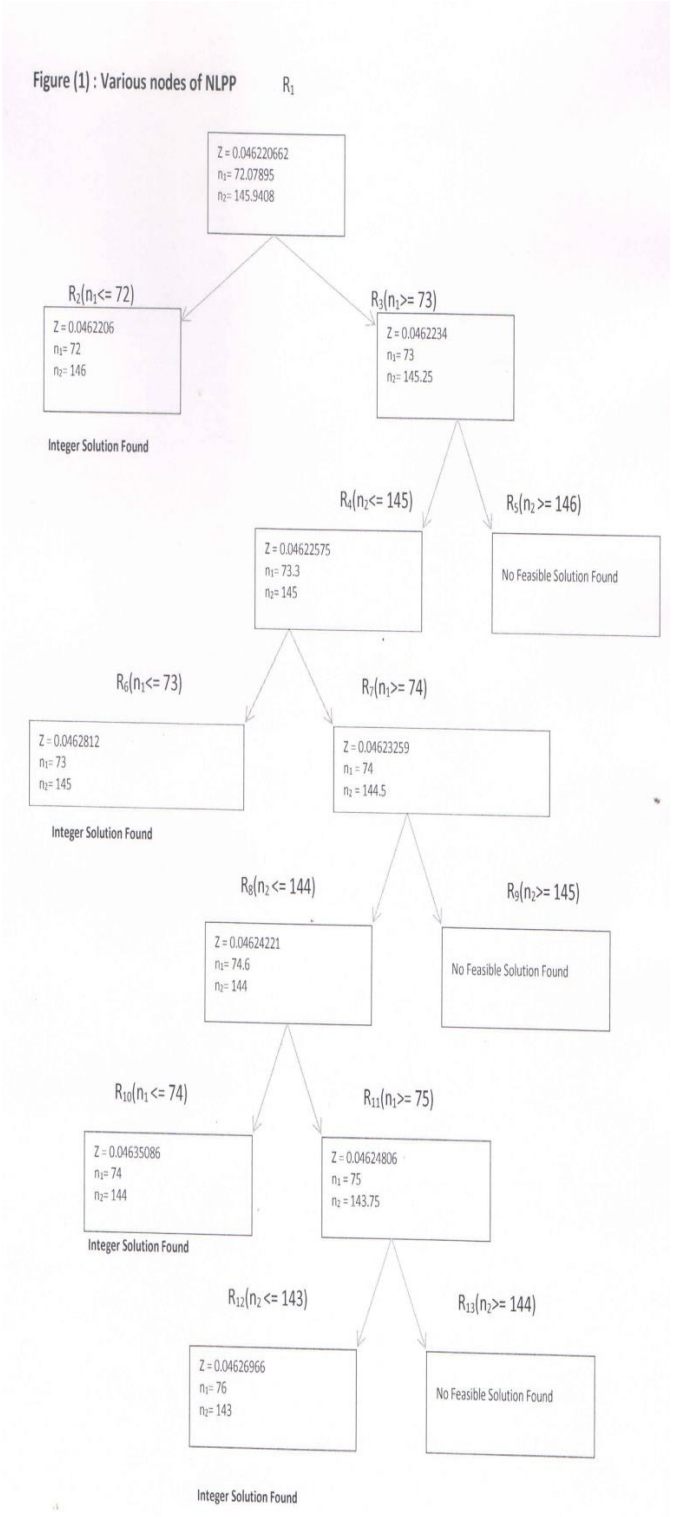
$$\left.\begin{aligned} &Minimize\ V(\ \hat{\pi}_S\ ) = \frac{0.9005}{n_1} + \frac{4.9222}{n_2} \\ &subject\ to\ \ 15n_1 + 20n_2 \le 4000 \\ &\qquad 2 \le n_1 \le 400 \\ &\qquad 2 \le n_2 \le 700 \quad and\ n_1, n_2\ integers, h = 1,2. \end{aligned}\right\}$$

Using the above minimization problem, we get optimal solution as n_1 = 72.07895, n_2 = 145.9408 and optimal value is Minimize $V(\ \hat{\pi}_S\ )$ = *0.04622062.*

Since n_1 and n_2 are required to be the integers, we branch problem R_1 into two sub problems R_2 and R_3 by introducing the constraints $n_1 \le 72$ and $n_1 \ge 73$ respectively indicated by the value $n_1 = 72.07895$ which lies between *72* and *73*. 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 figure (1). Since only one sub problems have integer solutions. Problems R_2 stand fathomed as the optimal solution in each case is integral in n_1 and n_2 and problem R_3 is further branched into sub problems R_4 and R_5 with additional constraints as $n_2 \le 145$; $n_2 \ge 146$ respectively and R_5 has no feasible solution. Problem R_4 has been further branched into sub problems R_6 and R_7 with additional constraints as $n_1 \le 73$ and $n_1 \ge 74$; respectively. Problems R_6 stand fathomed as the optimal solution in each case is integral in n_1 and n_2 and problem R_7 is further branched into sub problems R_8 and R_9 with additional constraints as $n_2 \le 144$; $n_2 \ge 145$ respectively and R_9 has no feasible solution. Problem R_8 is not fathomed and is further branched into two

sub problems, R_{10} and R_{11} by imposing the additional constraints $n_2 \leq 74$ and $n_2 \geq 75$ respectively, which suggests that R_{10} is fathomed as the optimal solution in each case is integral in n_1 and n_2 but problem R_{11} is not fathomed and is required to further branching into two sub problems R_{12} and R_{13} by imposing the additional constraints $n_2 \leq 143$ and $n_2 \geq 144$ respectively, which suggests that R_{12} is fathomed as the optimal solution in each case is integral in n_1 and n_2 and R_{13} has no feasible solution.

Now, all the terminal nodes are fathomed. The feasible fathomed node with the current best lower bound is node R_2. Hence the solution is treated as optimal. The optimal value is $n_1 = 72$ and $n_2 = 146$ and optimal solution is to Minimize $V(\pi_S) = 0.0462206$. The total cost under this allocation is 4000 units. It may be noted that the optimal integer values are same as obtained by rounding the n_h to the nearest integer. Let us suppose $V(\pi_S) = Z$, the various nodes for the NLPP (3) utilizing table1 and table2, are presented below in figure (1).



4. DISCUSSION

A stratified randomized response method assists to solve the limitations of randomized response that is the loss of individual characteristics of the respondents. Formulating non linear programming problem (NLPP) of optimum allocation in stratified sampling with linear cost function in presence of non responses using Branch and Bound algorithm based on Tarray and Singh (2016) provides the optimum integer solution.

REFERENCES

[1] Greenberg B, Abul- Ela A, Simmons WR, Horvitz DG (1969): The unreleased question randomized response: Theoretical framework. Jour. Amer. Statist. Assoc., 64,529-539.
[2] Horvitz DG, Shah BV and Simmons WR (1967): The unrelated question randomized response model. Proc. of Social Statistics Section. Jour. Amer. Statist. Assoc., 65-72.
[3] Fox J. A. and Tracy P. E. (1986): *Randomized Response: A method of Sensitive Surveys*. Newbury Park, CA: SEGE Publications.
[4] Hong K, Yum J and Lee H (1994): A stratified randomized response technique. Korean Jour. Appl. Statist., 7, 141-147.
[5] Singh HP and Tarray TA (2014): An improvement over Kim and Elam stratified unrelated question randomized response model using Neyman allocation. Sankhya – B, 77(1) , DOI 10.1007/s13571-014-0088-5 .
[6] Tarray TA and Singh (2016): A stratified randomized response model for sensitive characteristics using the negative hypergeometric distribution. Comm. Statist. Theo. Metho., 45(4), 1014-1030, DOI: 10.1080/03610926.2013.853795.
[7] Singh R and Mangat NS (1996): Elements of Survey Sampling, Kluwer Academic Publishers, Dordrecht, The Netherlands.
[8] Warner SL (1965): Randomized response: A survey technique for eliminating evasive answer bias. Jour. Amer. Statist. Assoc., 60, 63-69.