A NONLINEAR PROGRAMMING PROBLEM USING BRANCH AND BOUND METHOD

Tanveer A. Tarray and Muzafar R. Bhat
Islamic University of Science and Technology, Awantipora, Pulwama, Kashmir - India,192122.

ABSTRACT
Taking the clue from the pioneer work of Tarray and Singh (2015) we have suggested a new stratified randomized response model. In this paper the problem of optimal allocation in stratified random sampling where randomized response technique is used in presence of non response. 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.

KEYWORDS: Randomized response technique, Estimation of proportion, Respondents protection, Negative Binomial Distribution, Optimum allocation, Stratified random sampling, Dichotomous population, Sensitive attribute, Branch and Bound method.

MSC: 62D05.

RESUMEN
A partir de trabajo pionero de Tarray and Singh (2015) hemos sugerido un Nuevo modelo de respuestasaleatorizadas estratificado. En este trabajo se utiliza el método de las respuestas aleatorizadas para el problema de la afijación optimal ante la respuesta de no respuestas. El problema es formulado como uno de Programación No-lineal (NLPP) y se resuelve usando un método de "Branch and Bound". También son formulados los resultados a través del LINGO.

1. INTRODUCTION

Surveys are a means by which responses to questions concerning certain topics may be obtained from a sample of individuals selected in some manner from a population of interest. Results from surveys are affected by two main sources of error. The first is sampling error that results from taking a sample instead of enumerating the whole population. The second type of error is non-sampling error that cannot be attributed to sample-to-sample variability. Non-sampling error has two different errors which are random error and nonrandom error. Random error, which results from a reduction in the reliability of measurements, can be minimized over repeated measurements. However, nonrandom error, which is bias in the survey data, is difficult to cancel out over repeated measurements. In order to reduce non-response and response bias, a survey technique different from open or direct surveys was needed that made people comfortable and encouraged truthful answers. Warner (1965) developed such an alternative survey technique that is called randomized response (RR) technique. According to the method, for estimating the population proportion π possessing the sensitive character "G", a simple random with replacement sample of n persons is drawn from the population. Each interviewee in the sample is furnished an identical randomization device where the outcome "I possess character G" occurs with probability P while its complement "I do not possess character G" occurs with probability (1-P). The respondent answers "Yes" if the outcome of the randomization device tallies with his actual status otherwise he/she answers "No". Some modifications in the model have been suggested by Singh (2003), Zaizai et al. (2008) and Tarray and Singh (2015). Greenberg et al. (1969) provided theoretical framework for a modification to the Warner's model proposed by Horvitz et al. (1969). The proposed method consisted in modifying the randomization device where the second outcome "I do not possess the character G" was replaced by the outcome "I possess the character Y" where "Y" was unrelated to character "G". This modified model is now known as 'unrelated question model, or U- model'.
Hong et al. (1994) envisaged a stratified RR technique under the proportional sampling assumption. Under Hong et al.'s (1994) proportional sampling assumption, it may be easy to derive the variance of the proposed estimator. However, it may come at a high cost in terms of time, effort and money. For example, obtaining a fixed number of samples from a rural country in India through a proportional sampling method may be very difficult compared to the researcher's time, effort and money.
The study related with Kuk (1990), Singh and Grewal's (2013) and Hussain et al.'s (2014) randomized response models whose description are given in subsequent subsections.

2. RANDOMIZED MODELS

2.1 Kuk (1990) randomized response model

Kuk (1990) suggested a randomized response model in which respondents belonging to a sensitive group G are instructed to use a deck of cards having the proportion θ_1^* of cards with the statement, "I belong to group G" and if respondents belong to non – sensitive group $\overline{G}$ then they are instructed to use a different deck of cards having the proportion θ_2^* of cards with the statement, "I do not belong to group G". Let π_G be the true proportion of persons belonging to the sensitive group G. Then, the probability of a "Yes" answer in the Kuk's (1990) model is given by

$$\theta_k = \theta_1^* \pi_G + (1-\pi_G)\theta_2^*$$

Let a simple random sample with replacement (SRSWR) of n respondents be chosen from the population, and n_1 is the number of observed "Yes" answers. The number of people n_1 that answer "Yes" is binomially distributed with parameters $\theta_k = \theta_1^* \pi_G + (1-\pi_G)\theta_2^*$ and n. For the Kuk (1990) model, an unbiased estimator of the population proportion π_G is given by

$$\hat{\pi}_k = \frac{\hat{\theta} - \theta_2^*}{\theta_1^* - \theta_2^*}, \quad \theta_1^* \neq \theta_2^* \quad .$$

The variance of $\hat{\pi}_k$ is given by

$$V(\hat{\pi}_k) = \frac{\theta_k(1-\theta_k)}{n(\theta_1^* - \theta_2^*)^2} \quad .$$

Singh and Grewal (2013) have suggested improvement in the Kuk (1990) model Geometric distribution as a randomization device. They have claimed that their method is more protective and efficient than the Kuk (1990) model while doing surveys in practice. The description of Singh and Grewal (2013) randomized response technique is given below.

2.2 Singh and Grewal (2013) randomized response model

In this RRT, an individual respondent in the sample is provided with two decks of cards in the same way as in Kuk (1990) model. In the first deck of cards, let θ_1^* be the proportion of cards with the statement "I belong to a sensitive group G" and $(1-\theta_1^*)$ be the proportion of cards with the statement, "I do not belong to a sensitive group G". In the second deck of cards, let θ_2^* be the proportion of cards with the statement, "I do not belong to group G" and $(1-\theta_1^*)$ be the proportion of cards with the statement, " I belong to a sensitive group G". Up to here, it is same as the Kuk (1990) randomized response model. If a respondent belongs to a sensitive group G, he/she is instructed to draw cards, one – by – one with replacement, from the first deck of cards until he / she gets the first card bearing the statement of his / her own status, and requested to report the total number of cards, say Xdrawn by him / her to obtain the first card of his/ her own status. If a respondent belongs to group $\overline{G}$, he / she is instructed to draw cards, one – by – one using with replacement, from the second deck of cards until he / she gets the first card bearing the statement of his/ her own status, and requested to report the total number of cards, say Y, drawn by him/ her to obtain the first card of his / her own status. Since cards are drawn using with replacement sampling, it is clear that X and Y follow geometric distribution with parameters θ_1^* and θ_2^*, respectively[see Singh and Grewal (2013,pp. 244-245)]. If Z_i denotes the number of cards reported by the i^{th} respondent then it can be expressed as

$$Z_i = \alpha_i X_i + (1-\alpha_i)Y_i \quad ,$$

where α_i is a Bernoulli random variable . An unbiased estimator of π_G due to Singh and Grewal (2013) is given by

$$\hat{\pi}_{G(SG)} = \frac{\theta_1^*\theta_2^*\overline{Z} - \theta_1^*}{\theta_2^* - \theta_1^*}, \quad \theta_1^* \neq \theta_2^*.$$

The variance of $\hat{\pi}_{G(SG)}$ is given by

$$V(\hat{\pi}_{G(SG)}) = \frac{\pi_G(1-\pi_G)}{n} + \frac{\left\{\theta_2^{*2}(1-\theta_1^*)\pi_G + \theta_1^{*2}(1-\theta_1^*)(1-\pi_G)\right\}}{n(\theta_2^* - \theta_1^*)^2}.$$

Hussain et al.'s (2014) have suggested improvement in the Singh and Grewal's (2013) model. They have claimed that their method is more protective and efficient than the Singh and Grewal (2013) model while doing surveys in practice. The description of Hussain et al.'s (2014) RRT is as follows:

2.3. Hussain et al.'s (2014) randomized response model

This model is same as that of Singh and Grewal's randomized response model (2013) RRT except that the respondent belonging to either using first deck or second deck of cards are instructed to report number of cards drawn to obtain r(>1) cards of his/her own status. Then X and Y follow Negative Binomial (NB) distribution with parameters (r, θ_1^*) and (r, θ_2^*), respectively [see Hussain et al.'s (2014)].)]. If R_i denotes the number of cards reported by the i^{th} respondent then it can be expressed as randomized response model.

$$R_i = \alpha_i X_i + (1-\alpha_i)Y_i,$$

where α_i is same as in Singh and Grewal (2013) randomized response model. An unbiased estimator of π_G proposed by Hussain et al.'s (2014) is given by

$$\hat{\pi}_{G(P)} = \frac{\theta_1^*\theta_2^*\overline{R} - r\theta_1^*}{r(\theta_2^* - \theta_1^*)}, \quad \theta_1^* \neq \theta_2^*, r > 1.$$

with variance given by

$$V(\hat{\pi}_{G(P)}) = \frac{\pi_G(1-\pi_G)}{n} + \frac{\left\{\theta_2^{*2}(1-\theta_1^*)\pi_G + \theta_1^{*2}(1-\theta_1^*)(1-\pi_G)\right\}}{nr(\theta_2^* - \theta_1^*)^2}.$$

Recently, Tarray and Singh (2015) have suggested improvement regarding Hussain et al.'s (2014) model. They have claimed that their method is more protective and efficient than the Hussain et al.'s (2014) model while doing surveys in practice. The description of Tarray and Singh (2015) RRT is as follows:

2.4. Tarray and Singh (2015) randomized response model

The population is partitioned into L non – overlapping groups such that $N = \sum_{h=1}^{L} N_h$, where N_h is number of units in the h^{th}stratum (h=1,2,…,L). Let $w_h = N_h / N$ be the weight of the h^{th} stratum. An individual respondent in the sample of stratum h is provided with two decks of cards in the same way as in the Kuk (1990) RRT. In the first deck of cards θ_{1h}^* is the proportion of cards with the statement, "$I \in G$" and (1-θ_{1h}^*) is the proportion of cards with the statement, "$I \notin G$". In the second deck of cards θ_{2h}^* is the proportion of cards with the statement, "$I \notin G$" and (1- θ_{2h}^*) is the proportion of cards with the statement, "$I \notin G$" and (1-θ_{2h}^*) is the proportion of cards with the statement , "$I \in G$". Up to here, it is same as the Kuk (1990) RRT. If a respondent belongs to sensitive group G, he / she is instructed to draw cards, one by one using with replacement, from the first deck of cards. If a respondent belongs to non – sensitive group $\overline{G}$, he / she is instructed to draw cards, one by one using with replacement drawing, from the second deck of cards. Up to here, it is same as the Singh and Grewal (2013) RRT. The respondents belonging to either using first deck or second deck of cards are instructed to report number of cards drawn to obtain $r_h (>1)$ cards of his/her own status. Let X_h and Y_h be the number of cards drawn by the respondent from the first and second deck of cards respectively to obtain $r_h (>1)$ cards of his/her own

status. Then X_h and Y_h follow a Negative Binomial (NB) distribution with parameters(r_h, θ^*_{1h}) and (r_h, θ^*_{2h}), respectively.

Let R_{hi} be the number of cards reported by the i[th] respondent in the h[th] stratum, then it can be written as

$$R_{hi} = \alpha_{hi} X_{hi} + (1-\alpha_{hi}) Y_{hi},$$

where α_{hi} isa Bernoulli random variable with $E(\alpha_{hi}) = \pi_{Gh}$. In stratum h, the expected number of reported cards is given by

$$E(R_{hi}) = E(\alpha_{hi})E(X_{hi}) + E(1-\alpha_{hi})E(Y_{hi})$$

$$= \left[\frac{r_h \pi_{Gh}}{\theta^*_{1h}} + \frac{r_h (1-\pi_{Gh})}{\theta^*_{2h}} \right] = \left[\frac{r_h \pi_{Gh} (\theta^*_{2h} - \theta^*_{1h}) + r_h \theta^*_{1h}}{\theta^*_{1h} \theta^*_{2h}} \right].$$

Let $\overline{R}_h$ be the sample mean of reported response in stratum 'h'. Then an unbiased estimator of π_{Gh} is given by

$$\hat{\pi}_{Gh} = \left[\frac{(\theta^*_{1h} \theta^*_{2h} \overline{R}_h - r_h \theta^*_{1h})}{r_h (\theta^*_{2h} - \theta^*_{1h})} \right], \theta^*_{1h} \neq \theta^*_{2h}, r_h > 1.$$

Thus, an unbiased estimator of the population proportion $\pi_G = \sum_{h=1}^{L} w_h \pi_{Gh}$ is given by

$$\hat{\pi}_G = \sum_{h=1}^{L} w_h \hat{\pi}_{Gh} = \sum_{h=1}^{L} \frac{w_h (\theta^*_{1h} \theta^*_{2h} \overline{R}_h - r_h \theta^*_{1h})}{r_h (\theta^*_{2h} - \theta^*_{1h})}, \theta^*_{1h} \neq \theta^*_{2h}, r_h > 1.$$

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

$$V(\hat{\pi}_G) = \sum_{h=1}^{L} \frac{w_h^2}{n_h} V_h \qquad (1)$$

3. 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_i denote the number of units in the sample from stratum *i* and *n* denote the total number of units in sample from all strata so that $n = \sum_{i=1}^{k} n_i$.

Under the assumption that these "Yes" or "No" reports are made truthfully and P_i is set by the researcher. The problem of optimum allocation involves determining the sample size say $n_1, n_2, \ldots, n_h$that minimize the total variance $V(\hat{\pi}_G)$ subject to sampling cost. The sampling cost function is of the form $\sum_{i=1}^{k} c_i n_i$, 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_i 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 $C = C^0 + \sum_{i=1}^{k} c_i n_i$,

where C^0 is the over head cost, c_i is the per unit cost of measurement in i[th] stratum, C is the available fixed budget for the survey.

Equation (1) can be rewritten as

$$V(\hat{\pi}_G) = \sum_{i=1}^{k} \frac{w_i^2}{n_i} A_i$$

where

$$A_h = \pi_{Gh}(1-\pi_{Gh}) + \frac{\left\{\theta_{2h}^{*2}(1-\theta_{1h}^{*})\pi_{Gh} + \theta_{1h}^{*2}(1-\theta_{1h}^{*})(1-\pi_{Gh})\right\}}{r_h(\theta_{2h}^{*}-\theta_{1h}^{*})^2}. \tag{2}$$

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

$$\left.\begin{aligned} &Minimize V(\hat{\pi}_G) = \sum_{i=1}^{k} \frac{w_i^2}{n_i} A_i \\ &sujbect\ to \sum_{i=1}^{k} c_i n_i \le c^0 \\ &\qquad 2 \le n_i \le N_i \quad and\ n_i\ integers, i = 1,2,..k \end{aligned}\right\} \tag{3}$$

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

$$\varphi = \sum_{i=1}^{k} \frac{w_i^2}{n_i} A_i + \lambda\left[\sum_{i=1}^{k} c_i n_i - c^0\right] \tag{4}$$

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

$$\frac{\bar{V}\varphi}{\bar{V}n_i} = 0 \Rightarrow n_i = \frac{w_i\sqrt{A_i}}{\sqrt{c_i}\sqrt{\lambda}} \tag{5}$$

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

$$\frac{\bar{V}\varphi}{\bar{V}\lambda} = 0 \Rightarrow c^0 = \sum_{i=1}^{k} c_i n_i \tag{6}$$

Solving (5) and (6), we have

$$\sqrt{\lambda} = \sum_{i=1}^{k} c_i \frac{w_i\sqrt{A_i}}{\sqrt{c_i}} \tag{7}$$

Substituting (7) in (5), we have

$$n_i = \frac{w_i\sqrt{A_i}}{\left[\sum_{i=1}^{k} c_i \frac{w_i\sqrt{A_i}}{c^0\sqrt{c_i}}\right]\sqrt{c_i}} \Rightarrow \frac{c^0 w_i \frac{\sqrt{A_i}}{\sqrt{c_i}}}{\left[\sum_{i=1}^{k} w_i\sqrt{A_i}\right]\sqrt{c_i}} \tag{8}$$

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

$$\varphi = \sum_{i \in 1\,r}^{k} \frac{w_i^2}{n_i} A_i + \lambda\left[\sum_{i \in 1\,r}^{k} c_i n_i - c^0\right] \tag{9}$$

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

$$n_i = \frac{w_i \sqrt{A_i}}{\sqrt{\lambda}\sqrt{c_i}} \tag{10}$$

At r[th] node,

$$\sum_{i \in 1\,r}^{k} c_i n_i = c^0 - \sum_{i \in 1\,r}^{k} c_i n_i$$

$$\Rightarrow \sqrt{\lambda} = \frac{c^0 - \sum_{i \in 1\,r}^{k} c_i n_i}{\sum_{i \in 1\,r}^{k} \sqrt{c_i}\, w_i \sqrt{A_i}} \quad . \tag{11}$$

After simplification, we get formula for r [th] node as

$$n_i = \frac{\left(c^0 - \sum_{i \in 1\,r}^{k} c_i n_i\right) \frac{\sqrt{A_i}\, w_i}{\sqrt{c_i}}}{\sum_{i \in 1\,r}^{k} \frac{\sqrt{A_i}\, w_i}{\sqrt{c_i}}} \tag{12}$$

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

4. NUMERICAL ILLUSTRATION

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

Table 1: The stratified population with two strata

Stratum (i)	N_i	w_i	π_{Gi}	R	θ_i^*	P_i	c_i
1	400	0.3	0.8	3.0	0.7	0.6	15
2	800	0.7	0.13	3.50	0.3	0.7	20

Assuming that C (available budget) = 4500 units including c^0 and c^0 =500 units (overhead cost). Therefore c^0 =4500-500=4000 units. Also we assume that 400 and 700 are stratum sizes respectively as given in above table for i = 1,2 , N = 400+700 = 1100. The values of A_i and $A_i w_i^2$ are calculated as given in table below.

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

$$\left.\begin{aligned} &Minimize V(\hat{\pi}_G) = \frac{0.0239625}{n_1} + \frac{0.170394}{n_2} \\ &subject\ to\ \ 15n_1 + n_2 \le 4000 \\ &\qquad 2 \le n_1 \le 400 \\ &\qquad 2 \le n_2 \le 700 \quad and\ n_1, n_2\ integers, i = 1,2. \end{aligned}\right\}$$

Using the above minimization problem, we get optimal solution as n_1 = 65.37348, n_2 = 150.9699 and optimal value is Minimize $V(\hat{\pi}_G)$ = 0.00149521.

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 $n_1 \le 65$ and $n_1 \ge 66$ respectively indicated by the value n_1 = 65.37348 which lies between 65 and 66. 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 these two sub problems have optimal solutions in which the variables n_2 is non- integral, none of the sub problems has been fathomed. So both problems R_2 and R_3 are further branched into sub problems R_4 ;R_5 ;R_6 andR_7 with additional constraints as $n_2 \le 151$; $n_2 \ge 152$; $n_2 \le 150$ and $n_2 \ge 151$; respectively. Problems R_4 and R_5 stand fathomed as the optimal solution in each case is integral in n_1 and n_2. Problem R_7 has no feasible solution. Problem R_6 is not fathomed and is further branched into two sub problems, R_8

and R_9 by imposing the additional constraints $n_1 \leq 66$ and $n_1 \geq 67$ respectively, which are suggested by the non integral value $n_1 = 66.66$. Problem R_8 is fathomed with integer value. But the problem R_9 is not fathomed and is required to further branching into two sub problems R_{10} and R_{11} by imposing the additional constraints $n_2 \leq 149$ and $n_2 \geq 150$ respectively which are suggested by the non integral value $n_2 = 149.75$. Problem R_{11} has no feasible solution and problem R_{10} is fathomed with integer value. Now, all the terminal nodes are fathomed. The feasible fathomed node with the current best lower bound is node R_5. Hence the solution is treated as optimal. The optimal value is $n_1 = 64$ and $n_2 = 152$ and optimal solution is to Minimize V ($\widehat{\pi}_{1S}$)= 0.001495427 .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_i to the nearest integer. Let us suppose V ($\widehat{\pi}_G$) = Z, the various nodes for the NLPP (3) utilizing table1 and table2, are presented below in figure (1).

Figure (1) : Various nodes of NLPP

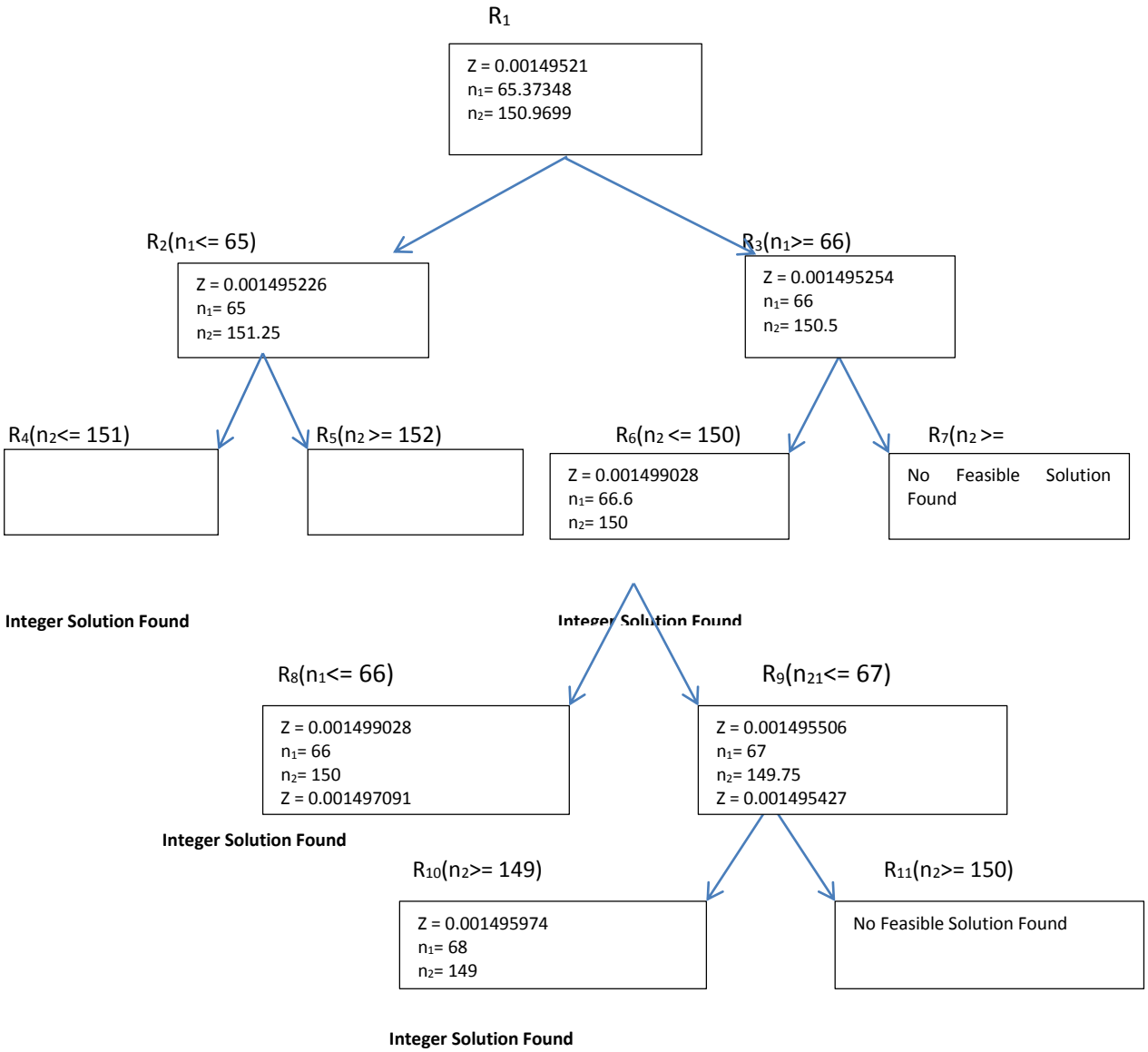

Table 2: Calculated values of A_i and $A_i w_i^2$

Stratum (i)	A_i	$A_i w_i^2$
1	0.26625	0.0239625
2	0.3477429	0.170394

5. CONCLUSION

This paper addresses the problem of estimating the population proportion of sensitive attribute π_G based on stratified sampling scheme. 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 (2015) provides the optimum integer solution.

Acknowledgements: The authors are thankful to the Editor – in- Chief, and to the anonymous learned referee for his valuable suggestions regarding improvement of the paper.

RECEIVED: APRIL, 2016
REVISED: OCTOBER, 2016

REFERENCES

[1] FOX J.A. and TRACY P.E. (1986): **Randomized Response: A method of Sensitive Surveys**. Newbury Park, CA: SEGE Publications.

[2] HONG K., YUM J. and LEE H. (1994): A stratified randomized response technique. **Korean Jour. Appl. Statist.**, 7, 141-147.

[3] KUK A.Y.C. (1990): Asking sensitive questions indirectly. **Biometrika** 77, 436-438.

[4] SINGH H.P. and MATHUR N (2005): Estimation of the population mean when the coefficient of variation is known using scrambled response technique. **Jour. Statist. Plann. Infere**. 131, 135-144.

[5] SINGH S. and SEDORY S.A. (2013): Geometric distribution as a randomization device: Implemented to the Kuk's model. **Int. Jour. Contemp. Math. Sci.** 8, 243-248.

[6] TARRAY T.A. (2016): **Statistical Sample Survey Methods and Theory.** Elite Publishers (onlinegatha.com), INDIA.

[7] TARRAY T.A. and SINGH H.P. (2015): A stratified randomized response model for sensitive characteristics using negative binomial distribution. **Revista Invest. Oper.** 36, 249-262.

[8] WARNER S.L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. **Jour. Amer. Statist. Assoc.**,60,63-69.