CSIRO PUBLISHING The South Pacific Journal of Natural and Applied Sciences, **29**, 40-42, 2011
www.publish.csiro.au/journals/spjns 10.1071/SP11008

Nonresponse in stratified sampling: A mathematical programming approach

Ummatul Fatima[1] and M. J. Ahsan[1]

[1]*Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India.*

Abstract

In sampling theory the term nonresponse is used for not being able to obtain from some units selected in the sample. Among other reasons nonresponse may be due to the refusal to answer or due to evasive answers in response to a sensitive question. Warner (1965) presented the Randomized Response (RR) technique to estimate the proportion of respondents to a sensitive question without the knowledge of the respondents' personal status. This paper addresses the problem of optimum allocation in stratified sampling under RR model as an All Integer Nonlinear Programming Problem (AINLPP) in the presence of nonresponse. The solution to the formulated problem is obtained using optimization software.

Keywords: *Stratified sampling, Optimum allocation, Nonresponse, Randomized response technique, All integer nonlinear programming.*

1 Introduction

In complex sample surveys where the measurements are practically difficult the surveyor may fail to measure some units selected in the sample for one reason or the other. This results in an incomplete and less informative sample. In the sampling literature, failure to measure some of the units selected in the sample is termed as 'nonresponse'. Cochran (1977) gave the following reasons for nonresponse: (i) noncoverage- due to the failure to locate or visit some units in the sample, (ii) not-at-home- due to the absence of the sampling unit from the given address, (iii) unable to answer- the respondent may not have the required information or may be unwilling to share it, (iv) the 'hard core'- persons who adamantly refuse to be interviewed.

In a questionnaire survey, if a question is highly sensitive or personal, the person may refuse to answer or may give evasive answer. This situation is covered in the 'hard core', that is, the (iv) reason stated above. To get response on such question the interviewer must encourage the truthful answers without revealing the identity of the person interviewed. Let π_A denotes the proportion of respondents belonging to a certain class 'A'. By using a randomizing device Warner (1965) showed that π_A can be estimated under the above situation. Warner's method is popularly known as Randomized Response (RR) technique in sampling literature. Many authors worked on RR methods (Greenberg *et al.,* 1969; Moors, 1971; Mangat, 1994; Mangat and Singh, 1990; Singh *et al.,* 2000). Kim and Warde (2004) used the Warner's model in stratified sampling design under optimum allocation. Shabbir and Gupta (2005) also used Warner's model in stratified sample surveys with nonresponse and obtained allocations under loglinear and nonlinear survey costs.

In this article the authors discussed the problem of optimum allocation in stratified random sampling when there are 'hard core' nonrespondents in the population. The problem is formulated as an All Integer Nonlinear Programming Problem (AINLPP). A numerical example is also presented and its solution is obtained by using optimization software *LINGO* (2004).

2 Formulation of the Problem

In Warner's RR method the two mutually exclusive and exhaustive questions asked are: (i) I am a member of class A, (ii) I am not a member of class A. Where A is certain attribute on the basis of which the population is to be classified. Each question require a 'yes' or 'no' response. By any randomizing device one of the questions is selected. The interviewer does not know which of the two questions is selected but does know the relative probabilities P and $(1-P)$ with which the two questions are selected. Let question (i) be selected with probability P. Obviously, the probability of selection of question (ii) will be $(1-P)$. Let in a random sample of size n the number of 'yes' answers be recorded as m. Then the binomial estimate of the proportion ϕ of the 'yes' answer is given by $\hat{\phi} = \frac{m}{n}$. Assuming correct answers from the respondents Cochran (1977) gave the relation between ϕ and π_A in the population as:

$$\phi = (2P-1)\pi_A + (1-P). \quad (1)$$

The estimated value of π_A is

$$\hat{\pi}_A = \left[\frac{\hat{\phi} - (1-P)}{(2P-1)}\right]; P \neq 0.5. \quad (2)$$

It can be seen that $\hat{\pi}_A$ is the maximum likelihood estimate (MLE) of π_A with a sampling variance

$$\mathrm{V}(\hat{\pi}_A) = \frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n} + \frac{p(1-P)}{n(2P-1)^2}; \; P \neq 0.5. \quad (3)$$

In a stratified population with L strata of sizes N_h; $h=1,2,\ldots,L$ let a simple random sample of size n_h be obtained from the h^{th} stratum. Under the RR model for h^{th} stratum define:

π_{Ah}: proportion of respondents who belong to certain class A.

ϕ_h: proportion of 'YES' answers.

$\hat{\phi}_h$: binomial estimate of ϕ_h.

P_{Ah} : probability of selection of question (i).

With the above definitions for, h^{th} stratum

$$\phi_h = (2P_{Ah}-1)\pi_{Ah} + (1-P_{Ah}); P_{Ah} \neq 0.5, \quad (4)$$

and

$$V(\hat{\pi}_{Ah}) = \frac{\pi_{Ah}(1-\pi_{Ah})}{n_h} + \frac{P_{Ah}(1-P_{Ah})}{n_h(2P_{Ah}-1)^2}; P_{Ah} \neq 0.5, \quad (5)$$

where $\hat{\pi}_{Ah}$ is the MLE of π_{Ah}.

If $W_h = N_h/N$ denote the proportion of population units falling in the h^{th} stratum then an unbiased estimate of π_A is given by

$$\hat{\pi}_A = \sum_{h=1}^{L} W_h \hat{\pi}_{Ah} \quad (6)$$

with a sampling variance

$$V(\hat{\pi}_A) = \sum_{h=1}^{L} W_h^2 V(\hat{\pi}_{Ah})$$

$$= \sum_{h=1}^{L} W_h^2 \left[\frac{\pi_{Ah}(1-\pi_{Ah})}{n_h} + \frac{P_{Ah}(1-P_{Ah})}{n_h(2P_{Ah}-1)^2} \right]; \; P_{Ah} \neq 0.5, \quad (7)$$

where $N = \sum_{h=1}^{L} N_h$ and $N_h; h = 1,2,\ldots,L$ is the strata sizes.

The interviewers have to approach the population units selected in the sample from each stratum to get the answer of the two proposed questions under the RR model. In each stratum the interviewers have to travel from unit to unit to contact them. This involves traveling cost in addition to the usual over head and measurement costs. Beardwood *et al.* (1959) showed that the traveling cost between *n* randomly scattered points may be given as $t\sqrt{n}$ where t is a constant. So that if the travel cost is substantial then instead of the usual linear cost function $C = c_0 + \sum_{h=1}^{L} c_h n_h$, it would be advisable to use

$$C = c_0 + \sum_{h=1}^{L} c_h n_h + \sum_{h=1}^{L} t_h \sqrt{n_h}, \quad (8)$$

as the cost function. Where c_0 = overhead cost, c_h = per unit cost of measurement in h^{th} stratum; $h = 1,2,\ldots,L$ and $t\sqrt{n}$ is the total cost involved in travelling within different strata between units selected in the sample.

Then, under Warner's RR method, the problem of allocation for fixed total cost may be expressed as the following AINLPP.

$$\left.\begin{aligned} &Minimize \; V(\hat{\pi}_A) \\ &Subject\ to \sum_{h=1}^{L} c_h n_h + \sum_{h=1}^{L} t_h \sqrt{n_h} \leq C_0 \\ &\qquad 2 \leq n_h \leq N_h \\ &\qquad n_h \; integers; \; h = 1,2,\ldots,L \end{aligned}\right\} \quad (9)$$

where $C_0 = C - c_0$ is the fixed budget available for the survey and $V(\hat{\pi}_A)$ is given by (7). The restrictions $2 \leq n_h$ and $n_h \geq N_h$ in AINLPP (9) are placed to have estimates of strata mean squares S_h^2 and to avoid oversampling, respectively. For practical implementation of the allocations the sampler needs their integer values, therefore, integer restrictions on n_h are imposed.

Substituting

$$A_h = \left[\pi_{Ah}(1-\pi_{Ah}) + \frac{P_{Ah}(1-P_{Ah})}{(2P_{Ah}-1)^2} \right]; \; h = 1,2,\ldots,L,$$ the

expression (7) for $V(\hat{\pi}_A)$ may be simplified as:

$$V(\hat{\pi}_A) = \sum_{h=1}^{L} \frac{A_h W_h^2}{n_h}, \text{ or } V(\hat{\pi}_A) = \sum_{h=1}^{L} \frac{K_h}{n_h},$$

where $K_h = A_h W_h^2; h = 1,2,\ldots,L.$

Thus, the AINLPP (9) can now be stated in a more simpler form as:

$$\left.\begin{aligned} &Minimize \; V(\hat{\pi}_A) = \sum_{h=1}^{L} \frac{K_h}{n_h} \\ &Subject to \sum_{h=1}^{L} c_h n_h + \sum_{h=1}^{L} t_h \sqrt{n_h} \leq C_0 \\ &\qquad 2 \leq n_h \leq N_h \\ &\qquad and \; n_h \; integers; \; h = 1,2,\ldots,L \end{aligned}\right\} \quad (10)$$

Using Lagrange Multipliers Technique (LMT) the AINLPP may be solved by taking equality in the constraint and ignoring the restrictions $2 \leq n_h \leq N_h$ and n_h integers; $h = 1,2,\ldots,L$. The noninteger solution may be rounded off to get integer allocations. If the rounded off values of n_h satisfy the restrictions $2 \leq n_h \leq N_h$; $h = 1,2,\ldots,L$ the AINLPP (10) is solved otherwise some integer nonlinear programming technique is to be used. For reasons given in Khan *et al.* (1997) 'rounding off' of the noninteger sample sizes is not always advisable because they may lead to infeasible or nonoptimum (or both) results. Therefore, the authors did not try the LMT. However, when the parameters K_h, c_h, t_h, C_0 and N_h of the AINLPP (10) are known, it can be solved by using an optimization software. The authors used *LINGO* (2004) which is a user's friendly software for constrained optimization of functions of several variables. For more details interested readers may visit the website www.lindo.com.

In the following article a numerical example is presented to illustrate the formulation of the problem for a given data. The solution of the formulated is obtained by the software *LINGO*.

3 A Numerical Example

Parts of the following data for a stratified population with two strata are from Shabbir and Gupta (2005).

Table 1. Data for a stratified population with two strata.

Stratum (h)	N_h	W_h	π_{Ah}	P_{Ah}	c_h	t_h
1	300	0.3	0.4	0.6	14	10
2	700	0.7	0.6	0.7	19	15

It is assumed that the available budget $C = 5500$ units including an overhead cost $c_0 = 500$ units. So that $C_0 = 5500 - 500 = 5000$ units. The strata sizes are assumed to be 300 and 700 respectively as given in Table 1. The computational values of $K_h, h = 1,2$ are presented in Table 2.

Table 2. Computation of $K_h, h = 1,2$.

h	$A_h = \left[\pi_{Ah}(1-\pi_{Ah}) + \frac{P_{Ah}(1-P_{Ah})}{(2P_{Ah}-1)^2} \right]$	$K_h = A_h W_h^2$
1	6.24	0.56
2	1.55	0.76

Substituting the values of the parameters from Table 1 and 2, the AINLPP (10) becomes:

$$\textit{Minimize } F(n_h) = \frac{0.56}{n_1} + \frac{0.76}{n_2}$$

$$\textit{Subject to } 14n_1 + 19n_2 + 10\sqrt{n_1} + 15\sqrt{n_2} \le 5000$$

$2 \le n_1 \le 300$ *and* $2 \le n_2 \le 700$; n_1, n_2 *integers.*

Using LINGO we obtain the required optimum allocation as:

$n_1 = 143$ and $n_2 = 142$.

The total cost under this allocation is 4998.33 < 5000 units.The optimal value of the objective function $V^*(\hat{\pi}_A) = 0.00926819$.

5 Conclusion

This paper gives a simple formulation of the problem of optimum allocation in stratified sampling in presence of nonresponse as an AINLPP. It has been suggested to use the optimization software *LINGO* to obtain the solution of the formulated AINLPP. A numerical example is also presented to illustrate the proposed formulation and the solution.

Acknowledgment

The authors are grateful to the referees and the editors for their valuable suggestions to improve the quality of this paper.

References

Beardwood, J. Halton, J.H. and Hammersley, J. 1959. The shortest path through many points. *Proceedings of Cambridge Philosophical Society* **55**, 299-327.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley and Sons.

Greenberg, B.G. Abul-Ela, Abdel-Latif. A., Simmons, W.R. and Horvitz, D.G. 1969. The unrelated question RR model theoretical frame-work. *Journal of American Statistical Asso*ciation **64**, 529-539.

Khan, M.G.M., Ahsan, M.J. and Jahan, N. 1997. Compromise allocation in multivariate stratified sampling: An integer solution. *Naval Research Logistics* **44**, 69-79.

Kim, J-M. and Warder, W.D. 2004. A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference* **120**, 155-165.

LINGO-User's Guide, 2004. *LINGO-User's Guide*. Publised by LINDO SYSTEM INC., 1415, North Dayton Street, Chicago, Illinois, 60622, USA.

Mangat, N.S. 1994. An improved randomized response strategy. *Journal of Royal Statistical Society*-B **56**, 93-95.

Mangat, N.S. and Singh, R. 1990. An alternative randomized response procedure. *Biometrika* **77**, 439-442.

Moors, J.J.A. 1971. Optimization of the unrelated question randomized response model. *Journal of American Statistical Asso*ciation **66**, 627-629.

Shabbir, J. and Gupta, S. 2005. Optimal allocation in stratified randomozed response model. *Pakistan Journal of Statistics and Operations Research* **1**, 15-22.

Singh, S., Singh, R. and Mangat, N. S. 2000. Some alternative strategies to Moor's model in randomized response model. *Journal of Statistical Planning and Inference* **83**, 243-255.

Warner, S.L. 1965. Randomize response: A Survey technique for eliminating evasive bias. *Journal of American Statistical Asso*ciation **60**, 63-69.

Correspondence to: Ummatul Fatima
E. mail: fatimau2011@yahoo.com