ESTIMATION OF POPULATION MEAN OF A STIGMATIZED QUANTITATIVE VARIABLE USING DOUBLE SAMPLING

I.S. Grewal, M.L. Bansal, S. Singh

1. INTRODUCTION

The randomized response technique (RRT), an ingenious interviewing procedure for eliciting information on sensitive character, was introduced by Warner (1965). Since then, the recent developments in RRT are due to Franklin (1989), Kuk (1990), Mangat and Singh (1990) and Singh and Singh (1993). The technique was further extended by Greenberg *et al.* (1971), Eichhorn and Hayre (1983) and Chaudhuri and Adhikary (1990) for estimation of mean or total of quantitative variable which are sensitive in nature.

An estimator, for the population total of a sensitive quantitative variable, y suggested by Arnab (1990) is

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{p_i} \tag{1.1}$$

where r_i denote the response of the i[th] respondent in the sample on the sensitive character through RR strategy given by Chaudhuri and Adhikary (1990) and $p_i = X_i \Big/ \sum_{i=1}^{N} X_i$ denotes the probability of selecting i[th] respondent in the sample. The estimator given in (1.1) is unbiased and has the variance

$$V(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{N}\frac{V_i^2}{p_i} + \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{p_i} - Y^2\right) \tag{1.2}$$

where

$$V_i^2 = V(r_i) = \alpha Y_i^2 + \tau Y_i + \theta \qquad (\alpha > 0) \tag{1.3}$$

The symbols α, τ and θ in (1.3) are constants. For details the reader may refer to Arnab (1990). In this estimator, the auxiliary information

X_i $(i = 1,2,...,N)$ is assumed to be known. However, sometimes it may not be possible to collect information on X for all the units of the population. Such a situation necessitates the technique of double sampling.

For example, Menon (2001) published a news in the Indian Express shows that the Govt of India is worried about the growth of population in India. As mentioned in the News, the use of good quality and different sizes of prophylactics (e.g. condoms) may result in the control of population, but before making a more stronger statement, the concerned department is thinking of doing a survey about the quality of prophylactics. The null hypothesis is:

H_0: More and more condoms are getting torn the population is rising.

In India, the use of prophylactics is very sensitive issue due to social set-up. It may not be true that all married couples are using prophylactics. In the first-phase, select a list of shops/ agencies who sell or distribute prophylactics in different areas of a particular city, say New Delhi. Select a few areas in the second-phase based on the known total sale or distribution of prophylactics in different areas selected in the first-phase. Thus the areas having more distribution of prophylactics have more chances of selection in the second-phase sample according to the design considered in the present investigation. From all the families/ couples living in the selected areas of the second-phase, collect information about the quality of prophylactics they use. It seems a difficult task to ask every family that how many times they faced a problem with prophylactics during sex and how many prophylactics they used last month (say). Both questions can be asked with the technology developed in the present investigation. The ratio of the estimate of the average number of defective prophylactics to the estimate of average of the total number of prophylactics used in all areas may help in studying the quality of prophylactics. It is assumed that total number of defective prophylactics and used prophylactics have positive and high correlation with the total number of prophylactics sold or distributed in different areas. This may be the most powerful and cost-effective design that may be used to resolve the problem of estimation of quality of prophylactics in a country like India.

In this paper, we have proposed an estimator of population mean of a sensitive quantitative variable in double sampling.

2. THE USUAL ESTIMATOR

If the sample is selected by simple random sampling and without replacement (SRSWOR) then the usual estimator of population mean, $\overline{Y}$, can be obtained by replacing p_i by $1/N$ in Arnab (1990) estimator. Thus getting

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} r_i \tag{2.1}$$

The estimator (2.1) is unbiased and has the variance

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_y^2 + \frac{1}{nN}\sum_{i=1}^{N} V_i^2 \tag{2.2}$$

3. PROPOSED STRATEGY

In the first phase, we select a preliminary large sample of "m" units from the population of N units by using SRSWOR. The auxiliary information X is measured on these "m" units. Let the observations recorded be $X_1, X_2, ..., X_m$. In the second phase, a sub-sample of "n" units is drawn from the preliminary large sample of "m" units with PPSWR and then the scrambled responses r_i are measured through randomization device proposed by Chaudhuri and Adhikary (1990). The proposed estimator of population mean is given by

$$\bar{y}_p = \frac{1}{mn}\sum_{i=1}^{n} \frac{r_i}{p_i^*} \tag{3.1}$$

where $p_i^* = X_i \Big/ \sum_{i=1}^{m} X_i$ denote the probability of selecting ith unit from the given first phase sample.

Thus we have the following theorems:

Theorem 3.1. The proposed estimator $\bar{y}_p$ is an unbiased estimator of population mean, $\bar{Y}$.

Proof. We have

$$\mathrm{E}(\bar{y}_p) = \mathrm{E}_1\mathrm{E}_2\mathrm{E}_3\left(\frac{1}{mn}\sum_{i=1}^{n}\frac{r_i}{p_i^*}\right) = \mathrm{E}_1\mathrm{E}_2\left(\frac{1}{mn}\sum_{i=1}^{n}\frac{y_i}{p_i^*}\right) = \mathrm{E}_1\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) = \bar{Y}$$

Hence the theorem

Theorem 3.2. The variance of the proposed estimator $\bar{y}_p$ is given by

$$V(\bar{y}_p) = \frac{1}{n}\left[\alpha\left\{\bar{X}\,\bar{Z}_1 + \left(\frac{1}{m} - \frac{1}{N}\right)S_{XZ_1}\right\} + \tau\left\{\bar{X}\,\bar{Z}_2 + \left(\frac{1}{m} - \frac{1}{N}\right)S_{XZ_2}\right\} + \theta\left\{\bar{X}\,\bar{Z}_3 + \left(\frac{1}{m} - \frac{1}{N}\right)S_{XZ_3}\right\}\right]$$

$$+ \frac{\sigma_Z^2}{n} + \left(\frac{1}{m} - \frac{1}{N}\right)\sigma_Z^2 \tag{3.2}$$

where

$$\sigma_Z^2 = \sum_{i=1}^{N} \frac{Y_i^2}{p_i} - Y^2, \quad Z_{1i} = \frac{Y_i^2}{X_i}, \quad Z_{2i} = \frac{Y_i}{X_i} \text{ and } Z_{3i} = \frac{1}{X_i}.$$

Proof. We have

$$V(\bar{y}_p) = \mathrm{E}_1\mathrm{E}_2 V_3(\bar{y}_p) + \mathrm{E}_1 V_2 \mathrm{E}_3(\bar{y}_p) + V_1 \mathrm{E}_2\mathrm{E}_3(\bar{y}_p)$$

which on simplification reduces to (3.2). Hence the theorem.

4. SUPER POPULATION MODEL APPROACH

To have look into the relative efficiency of the proposed estimator with respect to the usual estimator, the expected variances of the proposed and the usual estimator are worked out under the super-population model suggested by Cochran (1963) as

$$y_i = \beta\, p_i + e_i \tag{4.1}$$

where

$$\mathrm{E}(e_i \mid p_i) = 0, \ \mathrm{E}(e_i e_j \mid p_i p_j) = 0 \ \ \forall i \neq j \text{ and } \mathrm{E}(e_i^2 \mid p_i) = a p_i^g, \text{ with } a > 0; g \geq 0.$$

We have the following theorems:

Theorem 4.1. The expected variance of the usual estimator $\bar{y}$ under model (4.1) is

$$\mathrm{E}_m[V(\bar{y})] = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{F_1}{N} + \frac{F_2}{nN} \tag{4.2}$$

where

$$F_1 = \beta^2 \sum_{i=1}^{N} p_i^2 + a \sum_{i=1}^{N} p_i^g - N^{-1}\left(\beta^2 + a \sum_{i=1}^{N} p_i^g\right)$$

and

$$F_2 = \alpha\beta^2 \sum_{i=1}^{N} p_i^2 + \alpha\, a \sum_{i=1}^{N} p_i^g + \tau\,\beta + N\theta.$$

Theorem 4.2. The expected variance of the proposed estimator $\bar{y}_p$ under model (4.1) is

$$\mathrm{E}_m(\overline{y}_p) = \frac{1}{n}\left(A_1 + \frac{A_4}{N}\right) + \frac{A_2}{m} + \frac{A_3}{mn} - \frac{A_5}{N} \tag{4.3}$$

where

$$A_1 = N^{-2}\left(\alpha\beta^2 + \alpha a\sum_{i=1}^{N} p_i^{g-1}\right) + N^{-1}\tau\beta + \theta N^{-2}\sum_{i=1}^{N}\frac{1}{p_i} + N^{-2}\left(\beta^2 + a\sum_{i=1}^{N} p_i^{g-1} - \beta^2 - \sum_{i=1}^{N} p_i^{g}\right)$$

$$A_2 = N^{-1}\left[\beta^2\sum_{i=1}^{N} p_i^2 + a\sum_{i=1}^{N} p_i^{g} - N^{-1}\left(\beta^2 + a\sum_{i=1}^{N} p_i^{g}\right)\right]$$

$$A_3 = \alpha\left\{\frac{\beta^2\sum_{i=1}^{N} p_i^2 + a\sum_{i=1}^{N} p_i^{g}}{N-1} - \frac{\beta^2 + a\sum_{i=1}^{N} p_i^{g-1}}{N(N-1)}\right\} + \frac{\theta}{(N-1)}\left\{N - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i}\right\}$$

$$A_4 = -\frac{1}{(N-1)}\left[\alpha\left\{\beta^2\sum_{i=1}^{N} p_i^2 + a\sum_{i=1}^{N} p_i^{g}\right\} - \frac{1}{N}\left\{\beta^2 + a\sum_{i=1}^{N} p_i^{g-1}\right\}\right] + \theta\left(N - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_i}\right)$$

and

$$A_5 = \frac{1}{N}\left[\beta^2\sum_{i=1}^{N} p_i^2 + a\sum_{i=1}^{N} p_i^{g} - \frac{1}{N}\left(\beta^2 + a\sum_{i=1}^{N} p_i^{g}\right)\right]$$

The proofs of these theorems are straightforward, hence are omitted to save the space.

5. COST ASPECT

For efficiency comparison, we shall choose the strategy which, for a fixed cost C_0 can estimate $\overline{Y}$ with maximum precision. For this consider the cost function as

$$C_0 = nC_1 + mC_2 \tag{5.1}$$

where C_1 and C_2 are the cost per unit of collecting information in the second and first phase, respectively.

Now an optimal double sampling strategy is one that minimises $\mathrm{E}_m[V(\overline{y}_p)]$ subject to the condition (5.1). For this consider the function

$$\phi = \mathrm{E}_m[V(\overline{y}_p)] + \lambda\{C_0 - nC_1 - mC_2\} \tag{5.2}$$

To have the optimal values of "m" and "n", we partially differentiate (5.2) with respect to "m" and "n" and after solving, we get the optimal values of "m" and "n" as

$$m_{opt} = \frac{D_4 \pm \sqrt{D_4^2 - 4D_3D_5}}{2D_3} \tag{5.3}$$

and

$$n_{opt} = \frac{C_0 - m_{opt}C_2}{C_1} \tag{5.4}$$

where

$$D_3 = \frac{NA_2C_{\tilde{x}}^2}{C_1} - C_1A_1N - C_2A_4, \quad D_4 = A_3C_3N + \frac{2NA_2C_0C_2}{C_1} + A_2C_2N \text{ and}$$

$$D_5 = \frac{NA_2C_0^2}{C_1} + A_3C_0N.$$

Substituting the values of m_{opt} and n_{opt} in (4.3), then the minimum expected variance of the proposed estimator, $\bar{y}_p$, for the fixed cost becomes

$$\mathrm{E}_m[V(\bar{y}_p)]_{opt} = \left(A_1 + \frac{A_4}{N}\right)\frac{1}{n_{opt}} + \left(A_2 + \frac{A_3}{n_{opt}}\right)\frac{1}{m_{opt}} - \frac{A_5}{N} \tag{5.5}$$

Now for the usual estimator, consider the cost C_0 of observing a sample of size n is

$$C_0 = nC_1 \tag{5.6}$$

To minimise $\mathrm{E}_m[V(\bar{y})]$ subject to the condition (5.6), consider the function

$$\phi = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{F_1}{N} + \frac{F_2}{nN} + \lambda(nC_1 - C_0) \tag{5.7}$$

To have the optimal value of "n", we partially differentiate (5.7) with respect to "n" and after solving, we get

$$n_{opt} = \frac{C_0}{C_1} \tag{5.8}$$

Substituting the value of optimal "n" from (5.8) in (4.2), we get

$$\mathrm{E}_m[V(\bar{y})_{opt}] = \left(\frac{C_1}{C_0} - \frac{1}{N}\right)\frac{F_1}{N} + \frac{C_1}{C_0 N}F_2 \tag{5.9}$$

To examine the relative efficiency of the proposed estimator with respect to usual estimator, we consider a practicable randomization device proposed by Chaudhuri and Adhikary (1990). According to this device, the i[th] respondent in the sample is required to choose independently at random two tickets numbered a_i and b_t out of boxes proposed by the investigator containing the tickets numbered (i) A_1, A_2, ..., A_m with known mean $\bar{A}$ and known variance σ_A^2 and (ii) B_1, B_2, ..., B_t with known mean $\bar{B}$ and variance σ_B^2. The respondent is required to report the response as $Z_i = a_j Y_i + b_t$. Thus $\mathrm{E}_R(Z_i) = \bar{A}Y_i + \bar{B}$, $R_i = \dfrac{(Z_i - \bar{B})}{\bar{A}}$ and $V_R(r_i) = V_i^2 = \dfrac{(\sigma_A^2 Y_i^2 + \sigma_B^2)}{\bar{A}^2}$, where E_R and V_R denote the expected value and variance corresponding to the randomization device. Thus on comparing (1.3) with the above randomization device, we get $\alpha = \sigma_A^2 / \bar{A}^2$, $\tau = 0$ and $\theta = \sigma_B^2 / \bar{A}^2$.

6. EMPIRICAL STUDY

To compare the proposed estimator with respect to usual estimator, we conducted an empirical study. The density function f(x) for the auxiliary character x are presented in Table 1. For simplicity of calculations, it is further assumed that number written on deck B of cards are same. Also let us define, the relative efficiency of the proposed estimator over the usual estimator as:

$$RE = \mathrm{E}_m[V(\bar{y})] \times 100 / \mathrm{E}_m[V(\bar{y}_p)] \tag{6.1}$$

For computing the relative efficiency, we consider different values of correlation coefficient between x and y namely ρ = 0.5, 0.6, 0.7, 0.8 and 0.9. Overall cost C_0 = \$1000.00, cost of selecting unit in the first phase C_2 = \$0.15, cost of selecting a unit in the second phase C_1 = \$10.00 and coefficient of variation of the scrambling device $\alpha = 20\%$. Table 2 gives the relative efficiency of the proposed estimator with respect to usual estimator. From this, one can conclude that the proposed estimator is far better than the usual estimator in most of the practical situations.

TABLE 1

Distributions used for generating the selection probabilities

Distribution	Density Function	Range
Right Triangular	$f(x) = 2(1-x)$	$0 \le x \le 1$
Exponential	$f(x) = e^{-x}$	$0 \le x < \infty$
Chi-Square at $v = 6$	$f(x) = \frac{1}{2^{v/2}\Gamma_{v/2}} e^{-x/2} x^{(v-2)/2}$	$0 \le x < \infty$
Gamma with p=2	$f(x) = \frac{1}{\Gamma_p} e^{-x} x^{p-1}$	$0 \le x < \infty$
Log-Normal	$f(x) = \frac{1}{x\sqrt{2\pi}} e^{-(\log(x))^2/2}$	$0 \le x < \infty$
Beta with p=3, q=2	$f(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$

TABLE 2

Plan of results for estimator of population mean in double sampling using RRT

ρ	m_{opt}	n_{opt}	RE	m_{opt}	n_{opt}	RE	m_{opt}	n_{opt}	RE
		g=0			g=1			g=2	
Right Triangular									
0.5	519	92	41.71	804	88	102.6	808	84	141.4
0.6	558	92	48.59	856	87	118.3	855	83	161.3
0.7	616	91	60.18	933	86	144.1	923	82	193.2
0.8	715	89	82.73	1057	84	191.9	1028	79	149.9
0.9	921	86	143.5	1291	81	308.5	1213	76	376.4
Exponential									
0.5	741	89	88.34	794	88	102.3	736	85	118.2
0.6	785	88	101.5	839	87	117.1	774	85	131.9
0.7	849	87	122.4	903	86	140.1	827	83	158.0
0.8	948	86	159.1	1002	85	179.7	906	82	197.8
0.9	1121	83	237.3	1168	82	261.1	1033	79	273.9
Chi-Square at 6 d.f.									
0.5	688	90	76.00	792	88	102.2	727	85	115.9
0.6	730	89	87.75	835	87	116.9	763	86	131.4
0.7	791	88	106.5	897	87	139.8	814	84	154.9
0.8	886	87	139.6	990	85	178.5	888	82	193.3
0.9	1053	84	211.5	1145	83	256.1	1006	80	265.2
Gamma (p=2)									
0.5	603	91	56.46	804	88	102.6	809	84	141.5
0.6	646	90	65.59	856	87	118.3	855	83	161.4
0.7	711	89	80.86	933	86	144.0	923	82	193.2
0.8	819	88	110.2	1057	84	191.9	1028	79	250.1
0.9	1040	84	187.4	1291	81	308.6	1214	76	367.7
Log-Normal									
0.5	580	91	57.58	808	88	102.8	918	82	185.6
0.6	623	91	59.94	864	87	118.9	973	81	212.4
0.7	689	90	74.14	947	86	146.1	1058	79	256.6
0.8	800	88	102.1	1085	84	198.8	1180	76	332.1
0.9	1036	84	180.4	1206	80	341.4	1262	72	543.4
Beta (p=3, q=2)									
0.5	718	89	86.47	775	88	101.5	687	86	106.6
0.6	751	89	98.58	607	88	114.6	712	86	119.4
0.7	797	88	116.5	850	87	133.5	747	85	137.6
0.8	863	87	114.4	910	86	161.9	794	84	164.2
0.9	964	86	192.6	998	85	207.9	860	83	205.4

Department of Mathematics, Statistics & Physics
Punjab Agricultural University, India
INDERJIT SINGH GREWAL
MOHAN LAL BANSAL

Department of Statistics
St. Cloud State University, USA
SARJINDER SINGH

ACKNOWLEDGEMENTS

The authors are heartily thankful to the Executive Editor Professor Stefania Mignani and a referee for their valuable comments and encouragement to bring the original manuscript in the present form. The opinion and results discussed in this paper are of authors' and not necessary of their institute(s).

REFERENCES

R. ARNAB (1990), *On commutative of design and model expectations in randomized response survey*. "Communications in Statistics - Theory and Methods", 19, 3751- 3757.

A. CHAUDHURI, A., A.K. ADHIKARY (1990), *Variance estimation with randomized response*. "Communications in Statistics - Theory and Methods", 19, 1119-1126.

W.G. COCHRAN (1963), *Sampling techniques*. Second Edition. John, Wiley and Sons, Inc., New York, London.

B.H. EICHHORN, L. S. HAYRE (1983), *Scrambled randomized response models for obtaining sensitive quantitative data*. "Journal of Statistical Planning and Inference", 7, 307-316.

L.A. FRANKLIN (1989), *Randomized response sampling from dichotomous populations with continuos randomization*. "Survey Methodology", 15, 225-235.

B.G. GREENBERG, R.R. KUEBLER, J.R ABERNATHY, D.G. HORVITZ, D.G. (1971), *Application of the randomized response technique in obtaining quantitative data*. "Journal of the American Statistical Association", 66, 243-250.

A.Y.C. KUK (1990), *Asking sensitive questions indirectly*. "Biometrika", 77, 436-438.

N. S. MANGAT, R. SINGH (1990), *An alternative randomized response procedure*. "Biometrika", 77, 439-442.

S. MENON (2001), *Health ministry to study the Indian penis, region by region; says this will shrink condom cock-ups, show the huge diversity of the Indian organ*. "Indian Express", dated Sept 30, 2001.

S. SINGH, R. SINGH (1993), *Generalized Franklin's model for randomized response sampling*. "Communications in Statistics - Theory and Methods", 22, 741-755.

S. L. WARNER (1965), *Randomized response: A survey technique for eliminating evasive answer bias*. "Journal of the American Statistical Association", 60, 63-69.

RIASSUNTO

Stima di una media di popolazione di una variabile quantitativa sensibile, ottenuta mediante un campionamento doppio

Il presente lavoro affronta il problema della stima della media di popolazione rispetto a una variabile sensibile. La tecnica delle risposte randomizzate (RRT), proposta da Chaudhuri e Adhikary (1990), è utilizzata per elicitare l'informazione. Viene effettuato anche uno studio empirico per mettere in evidenza l'efficienza relativa dello stimatore proposto rispetto allo stimatore solitamente usato nel modello di super-popolazione.

SUMMARY

Estimation of population mean of a stigmatized quantitative variable using double sampling

This paper considers the problem of estimation of population mean of a sensitive variable in double sampling. Randomized Response Technique (RRT) proposed by Chaudhuri and Adhikary (1990) is used to elicit the information on the sensitive character. An empirical study is included to show the relative efficiency of the proposed estimator over the usual estimator under the super-population model.