Applied Mathematical Sciences, Vol. 1, 2007, no. 37, 1803 – 1820

ESTIMATION OF SENSITIVE MULTI-CHARACTERS USING UNKNOWN VALUE OF UNRELATED QUESTION

S S Sidhu

Department of Mathematics Statistics and Physics
Punjab Agricultural University, Punjab, 141004, India
sidhusukhjinder@hotmail.com

M L Bansal

Department of Mathematics Statistics and Physics
Punjab Agricultural University, Punjab, 141004, India
mlb_stat@yahoo.co.in

Sarjinder Singh

Department of Statistics, St. Cloud State University
St. Cloud, MN 56301-4498, USA
sarjinder@yahoo.com

ABSTRACT

The confidence of the respondent to answer sensitive questions is more, if one of the two questions belongs to non-sensitive attribute unrelated to sensitive characteristics Greenberg *et al.*[2]. Seeing its simplicity and wide application when the value of the unrelated question may be unknown in advance we propose a set of alternative estimators for probability proportional to size with replacement (PPSWR) corresponding to multi-character survey that elicit simultaneous information on many sensitive study variables. The new estimators, which take into account the already known rough value of the correlation coefficient ρ between y (the characteristic under study) and p (the measure of size), are developed corresponding to usual estimators in PPSWR. The estimators proposed are all biased but it is expected that the extent of bias will be smaller, since the proposed estimators are suitable for situations in between those optimum for the usual estimators and the estimators based on multi-characters for no correlation. The Mean Square Error (MSE) expressions are derived for the proposed estimators. An empirical study has also been carried out to examine their performance.

Keywords*:* Total estimation, RRT, Sensitive Multi-characteristics, Mean square error, Super Population Model, Cost aspects and Empirical study

1. Introduction

Rao [6] has provided alternative estimators in multi-character survey when the study variable and size measure are unrelated and demonstrated that these alternative estimators are more efficient though biased. But Rao [6] model is not commonly encountered in practice since the correlation is not always zero. Kumar and Herzel [10] suggested estimator for the character in same form looking different from Rao [6]. Bansal and Singh [7] developed a transformed estimator of population total suitable for the characteristics covering entire range of positive correlation. Amahia *et al*. [4] and Grewal *et al* [5]. suggested simple alternatives to the transformations in Bansal and Singh [7]. The transformations of selection probabilities used are as follows:

$$P_{i0}^{*} = \frac{1}{N} \qquad \text{[Rao [6]]} \qquad (1.1)$$

$$P_{i1}^{*}=\left(1+\frac{1}{N}\right)^{(1-\rho)}(1+p_i)^{\rho}-1 \quad \text{[Bansal and Singh [7]]} \qquad (1.2)$$

$$P_{i2}^{*}=\frac{(1-\rho)}{N}+\rho p_i \quad \text{[Amahia \textit{et al} [4]]} \qquad (1.3)$$

$$P_{i3}^{*}=\left(\frac{1}{N}\right)^{(1-\rho)} p_i^{\rho} \quad [\quad ,, \quad] \qquad (1.4)$$

$$P_{i4}^{*}=\left[N(1-\rho)+\frac{\rho}{p_i}\right]^{-1} \quad [\quad ,, \quad] \qquad (1.5)$$

$$P_{i5}^{*}=\frac{\left(1-\rho^{\frac{1}{3}}\right)}{N}+\rho^{\frac{1}{3}}p_i \quad \text{[Grewal \textit{et al} [5]]} \qquad (1.6)$$

On the basis of these transformations, following types of estimators of population total Y under PPSWR sampling are available in the literature:

$$\left(\hat{Y}_{pps}\right)_h=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{P_{ih}^{*}} \quad \text{for} \quad h = 1,2,3,4,5. \qquad (1.7)$$

The transformations P_{ih}^{*} $(h=1,2,3,4,5)$ at (1.2) to (1.6) of the selection probabilities p_i are useful for positive correlation between x and y variables, whereas transformation (1.1), is useful under no correlation situation. If $\rho=0$ then P_{ih}^{*} $(h=1,2,3,4,5)$ reduce to P_{i0}^{*} at (1.1), and if $\rho=1$ then these transformations reduces to original selection probabilities, p_i .

The surveys on human population had established the fact that the direct question about sensitive characters often result in either refusal to respond or falsification of the answer. This can bias the estimates. Warner [11] developed an interviewing procedure designed to reduce or eliminate the bias and called it Randomized Response Technique (RRT). It was felt that the confidence of the respondents in anonymity provided by RRT and hence reliability of their responses, might be further enhanced if one of the two question belong to non-sensitive, innocuous attribute unrelated to the sensitive characteristics. Following his suggestion, Horvitz *et al* [3] developed a procedure and called it 'unrelated question' UQ model for randomized responses. Greenberg *et al* [1] developed the theoretical framework for this model. He did not restrict the technique to nominal scale data, and thus modified his work to quantitative responses Greenberg *et al* [2]. It was found that the unrelated question technique was more efficient than the original Warner [11] model. Bansal *et al* [8] and Grewal *et al* [5] had discussed the multi-characteristics in RRT to estimate population total.

2. THE GREENBERG UQ MODEL

In the quantitative unrelated question random response model, using two questions, the overall distribution of responses is comprised of numerical answers to both questions, the answers being indistinguishable as to question. This distribution is a mixture of two pure distributions, which must be statistically separated to provide meaningful estimates of the parameters of interest, the population means of both the sensitive (Y) and unrelated non-sensitive (U) variables are μ_Y and μ_{U} and their respective variances σ_Y^2 and σ_U^2. When μ_U and σ_{U}^2 are unknown in the moderately sensitive surveys (earnings of head of household to estimate earned income), two independent, non-overlapping samples of sizes n_i $\mathrm{i} = 1,2$ are sampled by using simple random sampling with replacement (SRSWR). The respondent in the sample is provided with a randomization device consisting of sensitive and non sensitive statements about

(i.) how much money in dollars did the head of household, earn last year?

(ii) how much average money in dollars do you think the head of a household of your size earns in a year?

with probabilities T_k and (1-T_k) respectively in the sample k, (k=1,2). The respondent selects randomly one of the two statements, unobserved by the interviewer, and reports the answer. The expected response r_{ki} from individual i in sample k (k = 1,2) to these questions was a sum of money.

$$E(r_{ki}) = y_i T_k + (1 - T_k)u_i = \gamma_{ki} \qquad \text{(say)} \qquad (2.1)$$

The variance of the randomized response of i-th individual is

$$V(r_{ki}) = T_k(1 - T_k)(y_i - u_i)^2 \quad , \qquad k = 1,2 \qquad (2.2)$$

Keeping in view the importance of this model, we propose to extend this method to multi-character surveys. The behavior of the proposed estimators has been examined under the super-population model given below.

3. THE SUPER POPULATION MODEL

A general super population model for sensitive characteristic under study is

$$Y_i = \beta p_i + e_i \qquad (i = 1,2,\ldots.N) \qquad (3.1)$$

where e_i's are the error terms such that

$$E(e_i / p_i) = 0$$
$$E(e_i e_j / p_i p_j) = 0 \qquad (3.2)$$
$$\text{and } E(e_i^2 / p_i) = a p_i^g : a > 0 \quad g \geq 0$$

Here $E(e_i^2 / p_i)$ is the residual variances of Y for p = p_i . The expected value of this residual variance in the super population model is given by

$$E(a p_i^g) = aE(p_i^g)$$

and when the infinite super population is simulated by a finite large population of N units having the same characteristics it will be reduced to

$$E(a p_i^g) = \frac{a}{N}\sum_{i=1}^{N} p_i^g$$

Also the expected value of residual variance is known to be given by $\sigma_y^2(1-\rho^2)$.

Thus we have

$$\frac{a}{N}\sum_{i=1}^{N} p_i^g = \sigma_y^2(1-\rho^2) \quad \text{or} \quad \sigma_y^2 = \frac{\frac{a}{N}\sum_{i=1}^{N} p_i^g}{(1-\rho^2)}$$

Moreover

$$\beta^2 = \rho^2 \frac{\sigma_y^2}{\sigma_p^2} \;,\; = \frac{\rho^2}{(1-\rho^2)}\left[\frac{a}{N}\sum_{i=1}^{N} p_i^g / \sigma_p^2\right] \tag{3.3}$$

where $\sigma_p^2 = \frac{1}{N}\left[\sum_{i=1}^{N} p_i^2 - \frac{\left(\sum_{i=1}^{N} p_i\right)^2}{N}\right]$ (3.4)

The probability density function associated with sensitive and non-sensitive question may be different or same.

The super population model for unrelated non-sensitive question is

$$U_i = \beta^* p_i + e_i^* \qquad (\mathrm{i} = 1,2,\ldots.\mathrm{N}) \tag{3.5}$$

where e_i^*'s are the error terms satisfying all the conditions at (3.2)

It is assumed for simplicity that means of $\mathrm{Y_i}$ and $\mathrm{U_i}$ are different but the residual variances of U for p = $\mathrm{p_i}$ i.e. $E(e_i^{*2} / p_i)$ is same as of Y.

Similarly

$$\beta^{*2} = \rho^{*2}\frac{\sigma_U^2}{\sigma_p^2} \;, \qquad = \frac{\rho^{*2}}{(1-\rho^{*2})}\left[\frac{a}{N}\sum_{i=1}^{N} p_i^g / \sigma_p^2\right] \tag{3.6}$$

We first obtain the estimator of population total for PPSWR.

4. THE ESTIMATOR $(\hat{Y}_1)$

The estimator $(\hat{Y}_1)$ of population total Y for PPSWR is given by

$$\hat{Y}_1 = \frac{1}{T_1 - T_2}\left[\frac{(1-T_2)}{n_1}\sum_{i=1}^{n_1}\frac{r_{1i}}{p_i} - \frac{(1-T_1)}{n_2}\sum_{i=1}^{n_2}\frac{r_{2i}}{p_i}\right] \tag{4.1}$$

The estimator $(\hat{Y}_1)$ is unbiased for population total Y.

Theorem 4.1 The variance of the estimator $(\hat{Y}_1)$ is given by

$$V(\hat{Y}_1) = \frac{1}{(T_1 - T_2)^2}\left[\frac{(1-T_2)^2}{n_1}\left\{T_1(1-T_1)\sum_{i=1}^{N}\frac{(Y_i - U_i)^2}{p_i} + \left(\sum_{i=1}^{N}\frac{\gamma_{1i}^2}{p_i} - \left(\sum_{i=1}^{N}\gamma_{1i}\right)^2\right)\right\}\right.$$
$$\left. + \frac{(1-T_1)^2}{n_2}\left\{T_2(1-T_2)\sum_{i=1}^{N}\frac{(Y_i - U_i)^2}{p_i} + \left(\sum_{i=1}^{N}\frac{\gamma_{2i}^2}{p_i} - \left(\sum_{i=1}^{N}\gamma_{2i}\right)^2\right)\right\}\right] \tag{4.2}$$

where γ_{ki} is given in (2.1)

Proof: Let *E₁* & *E₂* denote the expected values with respect to sampling design and over randomization device respectively and let *V₁* & *V₂* be the corresponding variances.

Now

$$V(\hat{Y}_1) = E_1 V_2(\hat{Y}_1) + V_1 E_2(\hat{Y}_1) \tag{4.3}$$

$$E_1 V_2(\hat{\mathrm{Y}}_1) = E_1\left[\frac{1}{(T_1 - T_2)^2}\left\{\frac{(1-T_2)^2}{n_1^2}\sum_{i=1}^{n_1}\frac{T_1(1-T_1)(y_i - u_i)^2}{p_i^2} + \frac{(1-T_1)^2}{n_2^2}\sum_{i=1}^{n_2}\frac{T_2(1-T_2)(y_i - u_i)^2}{p_i^2}\right\}\right]$$

$$= \left[\frac{1}{(T_1 - T_2)^2}\left\{\frac{T_1(1-T_1)(1-T_2)^2}{n_1}\sum_{i=1}^{N}\frac{(Y_i - U_i)^2}{p_i} + \frac{T_2(1-T_2)(1-T_1)^2}{n_2}\sum_{i=1}^{N}\frac{(Y_i - U_i)^2}{p_i}\right\}\right]$$

and

$$V_1E_2(\hat{\mathrm{Y}}_1)=V_1\left[\frac{1}{T_1-T_2}\left\{\frac{(1-T_2)}{n_1}\sum_{i=1}^{n_1}\frac{\gamma_{1i}}{p_i}-\frac{(1-T_1)}{n_2}\sum_{i=1}^{n_2}\frac{\gamma_{2i}}{p_i}\right\}\right]\Bigg]$$

$$=\left[\frac{1}{(T_1-T_2)^2}\left\{\frac{(1-T_2)^2}{n_1}\left(\sum_{i=1}^{N}\frac{\gamma_{1i}^2}{p_i}-\left(\sum_{i=1}^{N}\gamma_{1i}\right)^2\right)-\frac{(1-T_1)^2}{n_2}\left(\sum_{i=1}^{N}\frac{\gamma_{2i}^2}{p_i}-\left(\sum_{i=1}^{N}\gamma_{2i}\right)^2\right)\right\}\right]$$

Thus we have from (4.3)

$$V(\hat{Y}_1)=\frac{1}{(T_1-T_2)^2}\left[\frac{(1-T_2)^2}{n_1}\left\{T_1(1-T_1)\sum_{i=1}^{N}\frac{(Y_i-U_i)^2}{p_i}+\left(\sum_{i=1}^{N}\frac{\gamma_{1i}^2}{p_i}-\left(\sum_{i=1}^{N}\gamma_{1i}\right)^2\right)\right\}\right.$$
$$\left.+\frac{(1-T_1)^2}{n_2}\left\{T_2(1-T_2)\sum_{i=1}^{N}\frac{(Y_i-U_i)^2}{p_i}+\left(\sum_{i=1}^{N}\frac{\gamma_{2i}^2}{p_i}-\left(\sum_{i=1}^{N}\gamma_{2i}\right)^2\right)\right\}\right]$$

We now extend the theory for the estimator obtained above to propose the estimators of population total in case of multi-character surveys.

5 THE PROPOSED ESTIMATORS $(\hat{Y}_2)$

The estimator of population total $(\hat{Y}_2)_h$ for multi-characteristics is given by

$$(\hat{Y}_2)_h=\frac{1}{T_1-T_2}\left[\frac{(1-T_2)}{n_1}\sum_{i=1}^{n_1}\frac{r_{1i}}{p_{ih}^*}-\frac{(1-T_1)}{n_2}\sum_{i=1}^{n_2}\frac{r_{2i}}{p_{ih}^*}\right] \qquad (5.1)$$

where P_{ih}^* is defined in (1.2 to 1.7). The response r_{ki} is measured by randomized device described earlier. This estimator is biased. For this we have the following theorem.

Theorem 5.1 The bias of the estimator $(\hat{Y}_2)_h$ is given by

$$B(\hat{Y}_2)_h=\sum_{i=1}^{N}\left(\frac{p_i}{p_{ih}^*}-1\right)Y_i \qquad (5.2)$$

Theorem 5.2 The variance of the estimator $(\hat{Y}_2)_h$ is given by

$$V(\hat{Y}_2)_h = \frac{1}{(T_1 - T_2)^2}$$

$$\left[\frac{(1-T_2)^2}{n_1} \left\{ T_1(1-T_1) \sum_{i=1}^{N} \frac{(Y_i - U_i)^2}{p_{ih}^{*2}} p_i + \frac{(1-T_1)^2}{n_2} \left(\sum_{i=1}^{N} \frac{\gamma_{1i}^2}{p_{ih}^{*2}} p_i - \left(\sum_{i=1}^{N} \frac{\gamma_{1i}}{p_{ih}^{*}} p_i \right)^2 \right) \right\} \right.$$

$$\left. + \frac{(1-T_1)^2}{n_2} \left\{ T_2(1-T_2) \sum_{i=1}^{N} \frac{(Y_i - U_i)^2}{p_{ih}^{*2}} p_i + \left(\sum_{i=1}^{N} \frac{\gamma_{2i}^2}{p_{ih}^{*2}} p_i - \left(\sum_{i=1}^{N} \frac{\gamma_{2i}}{p_{ih}^{*}} p_i \right)^2 \right) \right\} \right]$$

where γ_{ki} is given in (2.1). (5.3)

Proof: Let E_1, E_2 and V_1, V_2 be as defined earlier.

$$V(\hat{Y}_2)_h = E_1 V_2 (\hat{Y}_2)_h + V_1 E_2 (\hat{Y}_2)_h \qquad (5.4)$$

Now

$$E_1 V_2 (\hat{Y}_2)_{\mathrm{h}} =$$

$$E_1 \left[\frac{1}{(T_1 - T_2)^2} \left\{ \frac{(1-T_2)^2}{n_1^2} \sum_{i=1}^{n_1} \frac{T_1(1-T_1)(Y_i - U_i)^2}{p_{ih}^{*2}} + \frac{(1-T_1)^2}{n_2^2} \sum_{i=1}^{n_2} \frac{T_2(1-T_2)(Y_i - U_i)^2}{p_{ih}^{*2}} \right\} \right]$$

$$= \left[\frac{1}{(T_1 - T_2)^2} \left\{ \frac{T_1(1-T_1)(1-T_2)^2}{n_1} \sum_{i=1}^{N} \frac{(Y_i - U_i)^2}{p_{ih}^{*2}} p_i + \frac{T_2(1-T_2)(1-T_1)^2}{n_2} \sum_{i=1}^{N} \frac{(Y_i - U_i)^2}{p_{ih}^{*2}} p_i \right\} \right]$$

and using (2.1)

$$V_1 E_2 (\hat{Y}_2)_h = V_1 \left[\frac{1}{T_1 - T_2} \left\{ \frac{(1-T_2)}{n_1} \sum_{i=1}^{n_1} \frac{\gamma_{1i}}{p_{ih}^{*}} - \frac{(1-T_1)}{n_2} \sum_{i=1}^{n_2} \frac{\gamma_{2i}}{p_{ih}^{*}} \right\} \right]$$

$$=\left[\frac{1}{(T_1-T_2)^2}\left\{\frac{(1-T_2)^2}{n_1}\left(\sum_{i=1}^{N}\frac{\gamma_{1i}^2}{p_{ih}^{*2}}p_i-\left(\sum_{i=1}^{N}\frac{\gamma_{1i}}{p_{ih}^{*}}p_i\right)^2\right)\right.\right.$$
$$\left.\left.+\frac{(1-T_1)^2}{n_2}\left(\sum_{i=1}^{N}\frac{\gamma_{2i}^2}{p_{ih}^{*2}}p_i-\left(\sum_{i=1}^{N}\frac{\gamma_{2i}}{p_{ih}^{*}}p_i\right)^2\right)\right\}\right]$$

Substituting in (5.4), and re-arranging, we have

$$V(\hat{Y}_2)=\frac{1}{(T_1-T_2)^2}$$

$$\left[\frac{(1-T_2)^2}{n_1}\left\{T_1(1-T_1)\sum_{i=1}^{N}\frac{(Y_i-U_i)^2}{p_{ih}^{*2}}p_i+\frac{(1-T_1)^2}{n_2}\left(\sum_{i=1}^{N}\frac{\gamma_{1i}^2}{p_{ih}^{*2}}p_i-\left(\sum_{i=1}^{N}\frac{\gamma_{1i}}{p_{ih}^{*}}p_i\right)^2\right)\right\}\right.$$
$$\left.+\frac{(1-T_1)^2}{n_2}\left\{T_2(1-T_2)\sum_{i=1}^{N}\frac{(Y_i-U_i)^2}{p_{ih}^{*2}}p_i+\left(\sum_{i=1}^{N}\frac{\gamma_{2i}^2}{p_{ih}^{*2}}p_i-\left(\sum_{i=1}^{N}\frac{\gamma_{2i}}{p_{ih}^{*}}p_i\right)^2\right)\right\}\right]$$

To obtain the expected Mean Square Error (MSE) of proposed estimators $(\hat{Y}_2)_h$ under super population model we have the following theorem.

Theorem 5.3 The expected value of $MSE(\hat{Y}_2)_h$ under super population model is given by

$$E\{MSE(\hat{Y}_2)_h\}=\frac{1}{(T_1-T_2)^2}\left[\frac{D_1}{n_1}+\frac{D_2}{n_2}\right]+D_3 \qquad (5.5)$$

where

$$D_1=(1-T_2)^2\left[T_1(1-T_1)\left\{(\beta-\beta^*)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}+2a\sum_{i=1}^{N}\frac{p_i^{g+1}}{p_{ih}^{*2}}\right\}+\right.$$
$$a\left(T_1^2+(1-T_1)^2\right)\left\{\sum_{i=1}^{N}\frac{p_i^{g+1}}{p_{ih}^{*2}}-\sum_{i=1}^{N}\frac{p_i^{g+2}}{p_{ih}^{*2}}\right\}+$$
$$\left.\left(\beta T_1+(1-T_1)\beta^*\right)^2\left\{\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}-\left(\sum_{i=1}^{N}\frac{p_i^2}{p_{ih}^{*2}}\right)^2\right\}\right]$$

$$\begin{aligned}D_2 = (1-T_1)^2 & \left[T_2\left(1-T_2\right)\left\{(\beta-\beta^*)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}+2a\sum_{i=1}^{N}\frac{p_i^{g+1}}{p_{ih}^{*2}}\right\}+\right. \\ & a\left(T_2^2+(1-T_2)^2\right)\left\{\sum_{i=1}^{N}\frac{p_i^{g+1}}{p_{ih}^{*2}}-\sum_{i=1}^{N}\frac{p_i^{g+2}}{p_{ih}^{*2}}\right\}+ \\ & \left.\left(\beta T_2+(1-T_2)\beta^*\right)^2\left\{\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}-\left(\sum_{i=1}^{N}\frac{p_i^2}{p_{ih}^{*2}}\right)^2\right\}\right]\end{aligned}$$

and

$$D_3 = \beta^2\left(\sum_{i=1}^{N}\frac{p_i^2}{p_{ih}^{*}}-1\right)^2 + a\sum_{i=1}^{N}\left(\frac{p_i^{g+2}}{p_{ih}^{*2}}+p_i^g-2\frac{p_i^{g+1}}{p_{ih}^{*}}\right)$$

Proof:

We know that

$$E\{MSE(\hat{Y}_2)_h\} = E\left\{V(\hat{Y}_2)_h + \left(B(\hat{Y}_2)_h\right)^2\right\}$$

On using (2.1), (5.2), (5.3) substituting Y_i and U_i from (3.1) & (3.5), we get on taking expectation, under the super population models

$$\begin{aligned}\mathbf{E\{}MSE(\hat{Y}_2)_h\mathbf{\}}= & \frac{(1-T_2)^2}{n_1(T_1-T_2)^2}\left[T_1\left(1-T_1\right)\left\{(\beta-\beta^*)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}+2a\sum_{i=1}^{N}\frac{p_i^{g+1}}{p_{ih}^{*2}}\right\}\right. \\ & +\left(T_1^2+(1-T_1)^2\right)\sum_{i=1}^{N}\frac{ap_i^{g+1}}{p_{ih}^{*2}}+\left(\beta T_1+(1-T_1)\beta^*\right)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}} \\ & \left.-a\left(T_1^2+(1-T_1)^2\right)\left(\sum_{i=1}^{N}\frac{p_i^{g+2}}{p_{ih}^{*2}}\right)-\left(\left(\beta T_1+(1-T_1)\beta^*\right)\sum_{i=1}^{N}\frac{p_i^2}{p_{ih}^{*}}\right)^2\right] \\ & +\frac{(1-T_1)^2}{n_2(T_1-T_2)^2}\left[T_2\left(1-T_2\right)\left\{(\beta-\beta^*)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}}+2a\sum_{i=1}^{N}p_i^{g+1}\right\}\right. \\ & +\left(T_2^2+(1-T_2)^2\right)\sum_{i=1}^{N}\frac{ap_i^{g+1}}{p_{ih}^{*2}}+\left(\beta T_2+(1-T_2)\beta^*\right)^2\sum_{i=1}^{N}\frac{p_i^3}{p_{ih}^{*2}} \\ & \left.-a\left(T_2^2+(1-T_2)^2\right)\left(\sum_{i=1}^{N}\frac{p_i^{g+2}}{p_{ih}^{*2}}\right)-\left(\left(\beta T_2+(1-T_2)\beta^*\right)\sum_{i=1}^{N}\frac{p_i^2}{p_{ih}^{*}}\right)^2\right]\end{aligned}$$

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

On re-arranging, we have

$$E\{MSE(\hat{Y}_2)_h\}=\frac{1}{(T_1-T_2)^2}\left[\frac{D_1}{n_1}+\frac{D_2}{n_2}\right]+D_3$$

where D_1, D_2, and D_3 are defined in (5.5).

A randomized response survey using quantitative questions the optimal design requires the appropriate choice of T_1 and T_2, the selection of a non-sensitive question U and efficient allocation of total sample size into n_1 and n_2. A good working rule is to select $T_1 + T_2 = 1$. We choose T_1 as far from 0.5 as is practicable without creating suspicion in the respondent that the randomization device is heavily weighted in favor of a particular question. In this way the randomization device is likely to affect both samples in an identical manner.

The selected innocuous question should be worded in such a way that the reply is in the same unit of measure as the sensitive question. Moreover, the classification is not done on the basis of a given individual reply but rather on the basis of groups using estimation procedures. It is clear from the (5.3), the variance of estimator $(\hat{Y}_2)_h$,that for a fixed value of (n_1, n_2) , T_1, T_2, and variances of expected value of randomized responses; the variance of the estimator $(\hat{Y}_2)_h$ increases as ($Y_i – U_i$) increases. There is no choice involved in mean and variance of character Y since these are fixed by the nature of the sensitive characteristic. For any fixed value of (n_1, n_2), the variances of the estimator decreases with decreasing variance of character Y and $Y_i – U_i$. Thus the important guideline in choosing a non-sensitive question is not how much it differs from or resembles the sensitive one in meaning but rather how uniform or nearly identical the replies will be. Obviously, a wise choice would be to chose a non-sensitive question such that mean of Y is close to U and with minimum variance of character Y. However, if variance of character Y is considerably less than variance of character U, there could be some loss in cooperation on the part of respondents. For this reason, we take the variance of character Y and U identical and that reliance are placed on n_1 and n_2 to reduce the variance. For this, we have the following section.

6. THE ALLOCATION OF SAMPLE SIZE

Here we find the optimum values of sample sizes when accuracy is more desirable to estimate Y. The minimization is done first with respect to overall sample size without considering the cost aspect, and then with cost consideration. For this, the expected MSE under super population model is used .

Theorem 6.1. Under the super population model, the optimum values of n_1 and n_2 that minimize $E\{MSE(\hat{Y}_2)_h\}$ are given by

$$(n_1)_{opt} = \frac{n\sqrt{D_1}}{\sqrt{D_1}+\sqrt{D_2}} \qquad (n_2)_{opt} = \frac{n\sqrt{D_2}}{\sqrt{D_1}+\sqrt{D_2}} \tag{6.1}$$

where D_1 and D_2 are defined in (5.5).

Proof: Consider the function

$$L = \frac{1}{(T_1 - T_2)^2}\left[\frac{D_1}{n_1} + \frac{D_2}{n_2}\right] + D_3 + \phi(n_1 + n_2 - n) \tag{6.2}$$

Differentiating (6.2) partially w.r.t. n_1 and n_2 and taking $\frac{\partial L}{\partial n_1} = \frac{\partial L}{\partial n_2} = 0$, $n_1 + n_2 = n$ we have :

$$(n_1)_{opt} = \frac{n\sqrt{D_1}}{\sqrt{D_1}+\sqrt{D_2}} \qquad (n_2)_{opt} = \frac{n\sqrt{D_2}}{\sqrt{D_1}+\sqrt{D_2}}$$

Theorem 6.2. The minimum expected MSE for optimum sample size is given by

$$E\left\{MSE(\hat{Y}_2)\right\}_{\min} = \frac{\left(\sqrt{D_1} + \sqrt{D_2}\right)^2}{n(T_1 - T_2)^2} + D_3 \qquad (6.3)$$

where D_1, D_2, and D_3 are defined in (5.5).

Proof: The minimum expected MSE of $(\hat{Y}_2)_h$ is obtained by replacing n_1 and n_2 in (5.5) by $(n_1)_{opt}$ and $(\mathrm{n}_2)_{opt}$ given in (6.1).

The values $(n_1)_{opt}$ and $(\mathrm{n}_2)_{opt}$ used to get $E\left\{MSE(\hat{Y}_2)_h\right\}_{\min}$ are irrespective of cost consideration. For financial advantages the allocation of sample sizes is then determined with a view to minimize $E\left\{MSE(\hat{Y}_2)_h\right\}$ for a specified cost of conducting the sample survey.

Theorem 6.3 Under super population model the optimum values of n_1 and n_2 that minimize cost of conducting the sample survey are given by

$$(n_1)_{opt} = \frac{\sqrt{D_1}\,(C - C_0)}{\sqrt{C_1}\left(\sqrt{C_1 D_1} + \sqrt{C_2 D_2}\right)}, \quad (n_2)_{opt} = \frac{\sqrt{D_2}\,(C - C_0)}{\sqrt{C_2}\left(\sqrt{C_1 D_1} + \sqrt{C_2 D_2}\right)} \qquad (6.4)$$

where symbols have their usual meaning.

Proof: Let $C_1$and C_2 be the average cost of surveying one unit in the sample of sizes n_1 and n_2 respectively. Let C_0 be the fixed overhead cost. The cost function is given by

$C = C_0 + n_1C_1 + n_2C_2$

To determine the optimum value for n_i ,consider the function

$$\mathrm{L} = \frac{1}{(T_1 - T_2)^2}\left(\frac{D_1}{n_1} + \frac{D_2}{n_2}\right) + D_{11} + \psi\left(C_0 + n_1C_1 + n_2C_2\right) \qquad (6.5)$$

where ψ is some unknown constant.

Differentiating (6.5) partially w.r.t. n₁ and n₂ and taking $\frac{\partial L}{\partial n_1}=\frac{\partial L}{\partial n_2}=0$ we have

$$(n_1)_{opt}=\frac{\sqrt{D_1}\left(C-C_0\right)}{\sqrt{C_1}\left(\sqrt{C_1D_1}+\sqrt{C_2D_2}\right)},\ (n_2)_{opt}=\frac{\sqrt{D_2}\left(C-C_0\right)}{\sqrt{C_2}\left(\sqrt{C_1D_1}+\sqrt{C_2D_2}\right)} \tag{6.6}$$

Theorem 6.4 The minimum expected MSE for optimum sample size that optimizes cost of sample survey is given by

$$E\left\{MSE(\hat{Y}_2)_h\right\}_{\min}=\frac{\left(\sqrt{C_1D_1}+\sqrt{C_2D_2}\right)^2}{\left(C-C_0\right)\left(T_1-T_2\right)^2}+D_3 \tag{6.7}$$

where D₁, D₂, and D₃ are defined in (5.5).

Proof: The minimum expected MSE of $(\hat{Y}_2)_h$ is obtain by replacing n₁ and n₂ in (5.5) by $(n_1)_{opt}$ and $(\mathrm{n}_2)_{opt}$ given in (6.7).

To examine the relative efficiency of the proposed estimators $(\hat{Y}_2)_h$, (h=1,2,3,4,5.) with respect to $(\hat{Y}_2)_0$, when the value of the unrelated question is unknown, we resort to an empirical investigation. For this we have the following section.

7. THE EMPIRICAL STUDY

To investigate into the performance of the proposed estimators we resort to an empirical study under super population model given in section 3.0.For this the relative efficiency under unrelated question model (RE)ₕ of the proposed estimators $(\hat{Y}_2)_h$ for h=1,2,3,4,5 with respect to $(\hat{Y}_2)_0$ using the randomization device described in section 2.0. is given by

$$(RE_2)_h=\frac{E_m\left[MSE(\hat{Y}_2)_0\right]}{E_m\left[MSE(\hat{Y}_2)_h\right]}\times 100 \tag{7.1}$$

where symbols have their usual meanings. We assume that coefficient of variation for the randomized device is 20%. The probability associated with the statements in the device is 0.7 & 0.3 respectively. The density functions for the auxiliary character x are presented in Table1 below. For the sensitive character value of correlation coefficient between x & y is ρ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 but for unrelated question, correlation coefficient ρ^{*} = 0.15, 0.65, 0.95 is used. A PPSWR sample of size 20 was considered as drawn from a population consisting of 100 respondents.

Table 1: Density functions for various probability distributions.

Sr. No.	Distribution	Density function	Range
1	Right Triangular	$f(x) = 2(1-x)$	$0 \le x \le 1$
2	Exponential	$f(x) = e^{-x}$	$0 \le x < \infty$
3	Chi-square at $\nu = 6$	$f(x) = \frac{1}{2^{\nu/2}\Gamma_{\nu/2}} e^{-x/2} x^{(\nu-2)/2}$	$0 \le x < \infty$
4	Gamma, p=2	$f(x) = \frac{1}{\Gamma_p} e^{-x} x^{p-1}$	$0 \le x < \infty$
5	Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$	$-\infty < x < \infty$
6	Log Normal	$f(x) = \frac{1}{x\sqrt{2\pi}} e^{-\{log(x)\}^2/2}$	$0 < x < \infty$
7	Beta, p=3, q=2	$f(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1}$	$0 \le x \le 1$

The results obtained from these computations indicate that for every g = 0,1,2 and ρ^{*}=0.15, ρ^{*}=0.56, ρ^{*}=0.95 respectively the proposed estimator of population total of a sensitive character is more efficient than usual estimator for all the distributions discussed above in correlation range of 0.1—0.9. Therefore it is clear that the proposed estimators $(\hat{Y}_2)_h$ fares better than the estimator $(\hat{Y}_2)_0$ for all the p^*_{ih} (h= 1,2,3,4,5). The detailed computations results can be had from the authors on request.

Acknowledgement. The authors are thankful to the editor and a referee for the useful suggestions.

Bibliography

[1] B. G. Greenberg, A. L. Abul-Ela, W. R. Simmons and D.G.Horvitz, The unrelated question randomized response model theoretical framework, J Amer Statist Assoc, 64 (1969), 520-539.

[2] B. G. Greenberg, R. R. Kuelber, J. R. Abernathy and D. G. Horvitz, Application of the randomized response technique in obtaining quantitative data, J Amer Statist Assoc, 66 (1971), 243-250.

[3] D. G. Horvitz, B. V. Shah and W. R. Simmons, The unrelated question randomized response model Proc .Am Statist Assoc Social Statist Sect (1967), 65-72.

[4] G. N. Amahia, Y. P. Chaubey and T. J. Rao, Efficiency of a new estimator in PPS sampling for multiple characteristics, J Statist Plann Infer, 21 (1989), 72-84.

[5] I. S. Grewal, M. L. Bansal and S. Singh, An alternative estimator for multiple characteristics using randomized response technique in pps sampling. Aligarh J Statist , 19 (1997), 51-65.

[6] J. N. K. Rao, Alternative estimators in PPS sampling for multiple characteristics, Sankhyā Ser A 28 (1966), 47-60

[7] M. L. Bansal and R. Singh, An alternative estimator for multi characteristics in PPS sampling, J statist Plann Infer, 11 (1985), 313-320.

[8] M. L. Bansal, S. Singh and R. Singh, Multi-character survey using randomized response technique, Commun Statist - Theory Meth , 23 (1994), 1705-1715.

[9] N. S. Mangat, R. Singh and S. Singh, Sampling with varying probabilities without replacement, A review, Aligarh J Statist, 12 (1993), 75-75.

[10] P. Kumar and A. Herzel, Estimating population totals in surveys involving multi-characters, Metron 46 (1988), 33-46.

[11] S. L. Warner, Randomized response techniques for eliminating evasive answer bias, J Amer Statist Assoc., 60 (1965), 63-69.

[12] T. J. Rao, On certain alternative estimators for multiple characteristics in varying probability sampling, J Ind Soc Agril Statist, 45 (1993), 307-318.

Received: July 7, 2006