Jour. Ind. Soc. Ag. Statistics
44 *(1), 1992 : 45–54*

Supplementary Information Quantitative Randomized Response Model

Rajendra Singh* and O.P. Kathuria**
I.A.S.R.I., New Delhi - 110 012
(Received : October, 1989)

Summary

A supplementary randomized response design has been proposed. Estimator of population mean of sensitive variable has been developed and its variance derived. Rules for selection of design parameters have been obtained. It is shown that the proposed supplementary information model will never be less efficient than optimized model under any condition. The relative efficiency of the supplementary information quantitative randomized response model over the optimized model has been worked out for different values of the design parameters.

Key words : Supplementary Information, Quantitative, Un-related question, Randomized Response, Optimized model.

Introduction

Greenberg *et al* [2] extended the randomized response technique of reducing the response bias for answer to sensitive question for qualitative character to a situation where the response was quantitative. Singh [4] discussed in detail the optimization of unrelated question quantitative randomized response (UQQRR) model and concluded that the second sample should be solely employed to estimate the population mean of neutral variable as suggested by Moors [3] regarding unrelated question qualitative randomized response (UQQLRR) model.

Review of literature and above discussion reveal that the two-sample single alternate randomized response model proposed by Greenberg and co-workers for obtaining the data on continuous type sensitive random variable is more practicable and easy in handling than any other available model. The two-sample unrelated single alternate question quantitative randomized response model would

* I.V.R.I. Izatnagar, U.P.
** I.A.S.R.I., New Delhi-110012.

be optimal when P_2, the probability of 'yes' response to unrelated question in the second sample is equal to zero i.e. when the second sample is solely used to estimate population mean and variance of alternate variable. Henceforth this model would be referred to as optimized quantitative randomized response (OQRR) model.

To extract full benefit from OQRR model, the present article develops a randomized response design which uses the second sample more efficiently.

2. *Randomized Response Design*

The respondents in the second sample will be asked to answer openly two direct alternate questions Q_1 and Q_2 and through randomized device either sensitive question (A) or alternate question (Q_1) depending on its random selection in first sample.

The design can be described as

Technique used with respondent	Sample I	Sample II
Randomized Device	Question A Question Q_1	— —
Direct Question	— —	Question Q_1 Question Q_2

3. *Notations and Derivations*

Let X, Y and Y_s be the variables associated with sensitive question (A), alternate question (Q_1) and second alternate question (Q_2) with population means, μ_x, μ_y and μ_{ys} and variances σ_x^2, σ_y^2 and σ_{ys}^2 respectively. The second alternate question is selected in such a way that the variable y_s is correlated with variable y and let ρ be the correlation between y_s and y. In other words, one can say that response of second alternate question is used as supplementary information to improve the estimator of μ_y directly and estimator of μ_x indirectly.

Assume two independent samples of size n and m with replacement from the population, and let

p = Probability that sensitive question is selected by the first respondent in first sample.

$1-p =$ Probability that non-sensitive question (Q_1) is selected by the respondent in first sample,

$= q$

$Z_i =$ Observed response from individual i in first sample

$X_i =$ Response of individual i in case he selects sensitive question through randomized device in first sample,

$Y_{1i} =$ Response of individual i in case he selects alternate question (Q_1) through randomized device in first sample,

$Y_{2j} =$ Response of first alternate question (Q_1) directly asked from jth respondent in second sample,

$Y_{sj} =$ Response of second alternate question (Q_2) directly asked from jth respondent in second sample.

Now $$E(Z) = \mu_z = p\mu_x + (1-p)\,\mu_y \tag{3.1}$$

Estimator of μ_x from (3.1) takes the form

$$\hat{\mu}_{xo} = \frac{1}{p}\left[\,\hat{\mu}_Z - (1-p)\,\hat{\mu}_y\,\right] \tag{3.2}$$

where 'o' indicates estimator under the optimized version, i.e. taking $P_2 = 0$.

In many cases the simple "distribution free" moment estimator $\bar{z}$ of μ_x from the first sample and $\bar{y}_2$ of μ_x from the second sample will be appropriate, giving

$$\hat{\mu}_{xo} = \frac{1}{p}\left[\,\bar{z} - (1-p)\,\bar{y}_2\,\right] \tag{3.3}$$

where $$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i\,, \qquad \bar{y}_2 = \frac{1}{m}\sum_{j=1}^{m} y_{2j}$$

and $$\text{Var}\,(\hat{\mu}_{xo}) = \frac{1}{p^2}\left[\frac{\sigma_z^2}{n} + \frac{q^2\,\sigma_y^2}{m}\right] \tag{3.4}$$

where $\sigma_z^2 = p\sigma_x^2 + q\sigma_y^2 + pq\,(\mu_x - \mu_y)^2$

The optimal sub-division of the total sample of size N into n and m will be obtained by minimising Var($\hat{\mu}_{xo}$) with respect to n and m which gives

$$n = \frac{N\,\sigma_z}{\sigma_z + q\,\sigma_y}, \quad m = \frac{Nq\,\sigma_y}{\sigma_z + q\,\sigma_y} \tag{3.5}$$

Now, using Y_s as the supplementary variable, the usual linear regression estimator of μ_y in simple random sampling (Vide Cochran, [1]) is

$$\overline{Y}_{1r} = \bar{y}_2 + b\,(\mu_{ys} - \bar{y}_s) \tag{3.6}$$

where $\bar{y}_s = \frac{1}{m} \sum_{j=1}^{m} y_{sj}$

μ_{ys} is the population mean of variable y_s and is assumed to be known,

$$b = \frac{s_{yy_s}}{s^2_{y_s}}$$

$$s_{yy_s} = \frac{1}{m-1} \sum_{j=1}^{m} (y_{2j} - \bar{y}_2)\,(y_{sj} - \bar{y}_s)$$

$$s^2_{ys} = \frac{1}{m-1} \sum_{j=1}^{m} (y_{sj} - \bar{y}_s)^2$$

Substituting estimator $\overline{Y}_{lr}$ of $\bar{\mu}_y$ and $\bar{z}$ of μ_z in (3.2) another estimator of μ_x is

$$\hat{\mu}_{xb} = \frac{\bar{z} - (1-p)\,\overline{Y}_{lr}}{p} \tag{3.7}$$

with variance $\quad \text{Var}\,(\hat{\mu}_{xo}) = \frac{1}{p^2}\left[\frac{\sigma_z^2}{n} + \frac{q^2\,\sigma_y^2}{m}\,(1-\rho^2)\right]$

for $\frac{q^2}{mp^2}$ sufficiently small.

It is known that in randomized response model, a value of $p > 0.5$ provides optimal allocation of total sample into two samples of size n and m $(m < n)$ (Greenberg *et. al.* [2]). The assumption that $p > 0.5$ is not at all a restriction because for $p < 0.5$ the result will not change due to symmetry. The question arises, whether the sample size m is sufficiently large to use approximate variance formula of $\bar{y}_{lr}$? In other words, how large m should be for $\frac{q^2}{mp^2}$ to be negligible? In conventional regression estimator, if approximate variance formula can be used for N = 500 then in randomized response model it can be used for m equal to 91, 31 and 6 for p equal to 0.7, 0.8 and 0.9 respectively. This shows that even for moderately small value of m the approximate formula of Var $(\bar{y}_{lr})$ can be used in randomized response model

$$E(\hat{\mu}_{xb}) = \frac{1}{p} [\, E(\bar{z}) - qE(\bar{Y}_{lr}) \,]$$

It follows that to the first order of approximation

$$E(\hat{\mu}_{xb}) = \mu_x + \frac{q}{mp} \beta \left[\frac{\mu_{21}}{\sigma_{y_s y}} - \frac{\mu_{30}}{\sigma^2_{ys}} \right] \qquad (3.9)$$

where

$$\mu_{21} = E[\,(y_s - \mu_{ys})^2 (Y - \mu_y)]\,, \quad \mu_{30} = E(Y_s - \mu_{ys})^3\,,$$

$$\sigma_{y_s y} = E[\,(y_s - \mu_{ys})(Y - \mu_y)]\,, \quad \sigma^2_{y_s} = E(Y_s - \mu_{ys})^2$$

and β is the regression coefficient of alternate variable (Y) on supplementary variable (Y_s) in the population.

As

$$E(\hat{\mu}_{xb}) \neq \mu_x \qquad (3.10)$$

the estimator $\hat{\mu}_{xb}$ is a biased estimator of μ_x

It may be seen that for the bias to be negligible, the sample size required in randomized response model is smaller than that for the usual regression estimator for $p > 0.5$. For example, if bias of conventional regression estimator will be negligible for N = 500, the bias in randomized response case will be negligible for sample sizes 125 and 55 at $p = 0.8$ and $p = 0.9$ respectively.

In most of the practical situations, this biased estimator is an advantage in the case of randomized response design provided one selects supplementary variable suitably. In most of the surveys involving sensitive characters there is likelihood of under-estimation of μ_x due to false reporting. So, if one selects auxiliary variable in a manner such that

$$\frac{q}{n_2 p} \beta \left[\frac{\mu_{21}}{\sigma_{y_s y}} - \frac{\mu_{30}}{\sigma_y^2} \right] < |2 B_e| \qquad (3.11)$$

where B_e is the negative bias due to false reporting, then

$$|B_T(\hat{\mu}_{xb})| \le |B_T(\hat{\mu}_{xo})| \qquad (3.12)$$

where $B_T(.)$ denotes the total bias.

4. *Selection of Design Parameters*

The rule of selection of p and question Q_1 is same as given by Greenberg *et al* [2].

4.1 *Selection of Question*

Question Q_2 should be selected in such a way that correlation between first alternate variable (Y) and second alternate variable (Y_s) should be as close to 1 as possible, and bias in $\bar{Y}_{lr}$ should be positive or negative depending upon the possibility of under or over-estimation of μ_x respectively which could easily be identified before the start of survey on the basis of sensitive character considered for the study.

4.2 *Allocation of N into n′ and m′*

Let n′ and m′ be the sample sizes required for the supplementary information model. The optimal sub-division of total sample size N into n′ and m′ would be obtained by minimising $Var[\hat{\mu}_{xb}]$ with respect to n′ and m′. This gives

$$n' = \frac{N\sigma_z}{\sigma_z + q\sigma_y\sqrt{1-\rho^2}}, \quad m' = \frac{Nq\sigma_y\sqrt{1-\rho^2}}{\sigma_z + q\sigma_y\sqrt{1-\rho^2}}, \qquad (4.1)$$

Table 1. Values of n'/m' and n/m for different values of p and ρ and relations between μ_x, μ_y and σ_x, σ_y.

	Under Supplementary Information model									Under Optimized model with $P_2 = 0$	
	n'/m'					n'/m'				n/m	n/m
	$\lvert\mu_x - \mu_y\rvert = \sigma_x$					$\mu_x = \mu_y$				$\lvert\mu_x - \mu_y\rvert = \sigma_x$	$\mu_x = \mu_y$
	p	p=0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9		
$\sigma_x = 1.23\sigma_y$	0.6	3.252	3.583	4.344	1.117	2.988	3.291	3.991	6.839	3.102	2.850
	0.7	4.364	4.807	5.830	9.551	4.060	4.472	5.423	8.885	4.163	3.873
	0.8	6.546	7.211	8.745	14.327	6.202	6.831	8.284	13.572	6.245	5.916
	0.9	13.009	14.329	17.377	28.470	12.623	13.904	16.862	27.625	12.410	12.042
$\sigma_x = \sigma_y$	0.6	2.918	3.214	3.398	6.387	2.621	2.887	3.501	5.735	2.784	2.500
	0.7	3.844	4.234	5.134	8.412	3.494	3.849	4.668	7.647	3.667	3.333
	0.8	5.546	6.218	7.541	12.354	5.241	5.773	7.001	11.471	5.385	5.000
	0.9	10.944	12.055	14.619	23.952	10.483	11.547	14.002	22.942	10.440	10.000

Comparison of equations (4.1) and (3.5) reveals that $n'/m' > n/m$ for all $\rho \neq 0$

The values of n'/m' under supplementary information model and n/m under the optimized model for different values of design parameters are given in Table 1. It indicates that for $\sigma_y < \sigma_x$, a large proportion of units is allocated to first sample which decreases as σ_y approaches σ_x or becomes greater than σ_x. The values of n'/m' and n/m for $|\mu_x - \mu_y| = \sigma_x$ and $\mu_x = \mu_y$ differ marginally from each other indicating thereby that the sample allocation of n'/m' or n/m could be done under the assumption $\mu_x = \mu_y$ to make it more practical and easy to handle. The relation $|\mu_x - \mu_y| = \sigma_x$ has been used to examine the situation when μ_x and μ_y deviate from each other by σ_x in comparison to $\mu_x = \mu_y$.

5. *Efficiency*

To obtain the gain in efficiency due to adoption of supplementary information model in comparison to optimized model we have

$$E_f = \frac{\operatorname{Var}(\hat{\mu}_{xo})}{\operatorname{Var}(\hat{\mu}_{xb})}$$

$$= \left[\frac{\sigma_z^2}{n} + q^2 \frac{\sigma_y^2}{m}\right] \left[\frac{\sigma_z^2}{n'} + \frac{q^2}{m'} \sigma_y^2 (1-\rho^2)\right]^{-1}$$

Substituting from (4.1) in the above expression and after simplification, we get

$$E_f = \frac{\dfrac{\sigma_z^2}{n} + q^2 \dfrac{\sigma_y^2}{m}}{\dfrac{m(1-\rho^2)^{1/2}}{m'} \left[\dfrac{\sigma_z^2}{n} + \dfrac{q^2}{m}(1-\rho^2)^{1/2} \sigma_y^2\right]} \qquad (5.1)$$

For $\rho \neq 0$,

$$\frac{\sigma_z^2}{n} + \frac{q^2}{m}(1-\rho^2)^{1/2} \sigma_y^2 < \frac{\sigma_z^2}{n} + \frac{q^2}{m} \sigma_y^2$$

for any value of p, σ_z, n and m. So denominator inside the bracket is always smaller that numerator of (5.1).

Using (3.5) and (4.1) it can be seen that for

$\rho \neq 0$, $\frac{m(1-\rho^2)^{1/2}}{m'} \leq 1$. This proves that supplementary information model is always better than optimized model.

The relative efficiency E_f for different values of design parameters is shown in Table 2.

It follows from Table 2 that the relative efficiency increases as ρ increases or p decreases. The important point to be noted here is

Table 2. Relative efficiency E_f of supplementary information model in comparison to optimized model under different assumptions regarding design parameters

		$\|\mu_x - \mu_y\| = \sigma_x$				$\mu_x = \mu_y$			
	p	p=0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
$\sigma_x = 1.23\sigma_y$	0.6	1.02	1.06	1.14	1.25	1.02	1.07	1.15	1.27
	0.7	1.02	1.05	1.11	1.19	1.02	1.05	1.11	1.20
	0.8	1.01	1.04	1.07	1.13	1.01	1.04	1.08	1.13
	0.9	1.01	1.02	1.04	1.07	1.01	1.02	1.04	1.07
$\sigma_x = \sigma_y$	0.6	1.02	1.07	1.15	1.27	1.03	1.08	1.16	1.30
	0.7	1.02	1.06	1.12	1.21	1.02	1.06	1.13	1.23
	0.8	1.01	1.04	1.08	1.15	1.02	1.04	1.09	1.16
	0.9	1.01	1-02	1.05	1.08	1.01	1.02	1.05	1.08
$\sigma_x = 0.71\sigma_y$	0.6	1.03	1.08	1.17	1.31	1.03	1.09	1.19	1.36
	0.7	1.02	1.06	1.14	1.25	1.02	1.07	1.15	1.28
	0.8	1.02	1.05	1.10	1.18	1.02	1.05	1.11	1.20
	0.9	1.01	1.03	1.06	1.10	1.01	1.03	1.06	1.11

that when $\sigma_y > \sigma_x$ substantial increase in relative efficiency was observed than when $\sigma_y \leq \sigma_x$.

This study concludes that one should prefer supplementary information randomized model over optimized model when ρ is high, $p(p>0.5)$ is small, $\sigma_y > \sigma_x$ and $\mu_x = \mu_y$.

ACKNOWLEDGEMENT

The authors are grateful to the referee for valuable suggestions leading to improvement of the paper.

REFERENCES

[1] Cochran, W.G., 1963. Sampling Technique, 2nd Ed., John Wiley and Sons, New York.

[2] Greenberg, B.G., Kwbler, Roy, R., Jr. Abernathy, James R. and Horvitz, Daniel., 1971. Application of the randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.*, **66**, 243-250.

[3] Moors, J.J.A., 1971. Optimisation of the unrelated question randomized response model. *J. Amer. Statist. Assoc.*, **66**, 627-629.

[4] Singh, Rajendra, 1984. On randomized response technique for qualitative and quantitative characters. Unpublished Ph.D. Thesis. Indian Agricultural Statistics Research Institute, New Delhi.

[5] Warner, Stanley L., 1965. Randomized response : A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60**, 63-69.