Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 19

11-1-2005

Kim And Warde's Mixed Randomized Response Technique For Complex Surveys

Amitava Saha
Directorate General of Mines Safety, India, saha_amitava@hotmail.com

DOI: 10.22237/jmasm/1130804280

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 538-544

1538 – 9472/05/$95.00

Kim And Warde's Mixed Randomized Response Technique For Complex Surveys

Amitava Saha
Directorate General of Mines Safety
India

The randomized response (RR) technique introduced by Warner (1965) was found to be an effective method for reducing answer bias and ensuring better respondent cooperation in estimating the proportion of people in a community bearing a sensitive attribute. Chaudhuri (2001a, 2001b, 2002, 2003) extended Warner's method and several other well-known RR devices to complex surveys adopting a varying probability sampling design. Kim and Warde (2004) proposed an RR model assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). Here, the method of estimation is presented when sample is chosen with varying selection probabilities and Kim and Warde's RR procedure is applied for estimating a sensitive proportion. Also illustrated is a numerical example that unequal probability sampling performs better than SRS.

Key words: Answer bias; randomized response; sensitive attribute; simple random sampling; varying probability sampling

Introduction

Warner (1965) proposed a method called randomized response (RR) to ensure better respondent cooperation and honest responses in surveys involving collection of information on certain sensitive attributes. It has been found that Warner's technique is capable of reducing answer bias and refusals considerably in surveys where a question of sensitive nature is involved. This method has been studied extensively and as a consequence, numerous modifications of it as well as several other methods have emerged in the literature of RR. Among many others, Horvitz et al. (1967), Greenberg et al. (1969), Kuk (1990), Christofides (2003), Mangat and Singh (1990) made notable contributions.

Most of the works cited here have been done assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). But in practice, in the socio-economic surveys, the respondents are usually selected with varying probability sampling. Thus, to meet the demand of the social surveys, Chaudhuri (2001a, 2001b, 2002, 2004) extended some of the RR procedures to complex survey situations.

Most of the works cited here have been done assuming that the sample is selected with simple random sampling (SRS) with replacement (SRSWR). But in practice, in the socio-economic surveys, the respondents are usually selected with varying probability sampling. Thus, to meet the demand of the social surveys, Chaudhuri (2001a, 2001b, 2002, 2004) extended some of the RR procedures to complex survey situations.

Kim and Warde (2005) proposed a mixed RR model in an attempt to improve Moors (1971) model after taking due consideration of the inherent privacy problem of Moors (1971) RR device. They have also discussed how their method may be applied when stratified sampling design is used. But the entire development of Kim and Warde (2005) is based on the assumption that the sample is selected with SRSWR. Since in large-scale sample surveys equal probability sampling is rarely used, necessary modifications need to be developed for adopting this method to complex sample surveys where varying probability sampling designs are often used. Here, Kim and

Contact information for Amitava Saha is Dhanbad, Jharkhand – 826001, India. E-Mail: saha_amitava@hotmail.com

Warde's (2005) procedure is presented when a varying probability sampling design is adopted rather than SRSWR. As well, a numerical illustration of the performance of the extended procedure under varying and equal probability sampling is presented.

Kim and Warde's (2005) Device in Complex Surveys

Kim and Warde's (2005) method for complex surveys is described in section 2. A numerical study for comparing the relative performances is reported in section 3.

Let $U=(1,\cdots\cdots,i,\cdots\cdots N)$ be a finite population of N individuals and y_i be the value of a variable of interest, say, y on the ith individual such that $y_i = 1$ if i bears a sensitive attribute A = 0 if i bears the complementary attribute A^C. The problem is to estimate the proportion of people in U bearing the character A, i.e., $\pi_A=\left(\sum_{i=1}^{N} y_i\right)\Big/N = Y/N$ where $Y=\sum_{i=1}^{N} y_i$ on choosing a sample, say, s of size n from U according to any arbitrary sampling design p.

It is also assumed that x_i be the value of a variable x on the ith individual in U such that $x_i = 1$ if i bears a non-sensitive attribute B = 0 if i bears B^C, the complement of B. Kim and Warde (2004) proposed a method for estimating π_A when a sample of size n is drawn from U by SRSWR. However, in this article it is assumed that instead of selecting the individuals by SRSWR only, they are chosen following any arbitrary sampling design p.

In Kim and Warde's (2005) device every sampled person is requested to answer a direct question about his/her possession of a non-stigmatizing or innocuous character, say, B and on receiving a '*yes*' reply to this non-sensitive question the individual is instructed to use an RR device R_1 where a pack of cards marked A and B in proportions $p_1:(1-p_1), 0<p_1<1$ is kept. The respondent is then requested to draw a card at random from this pack, unnoticed by the interviewer and to report the true value of y or x according as A-marked or B-marked card is drawn. If a respondent answers '*no*' to the initial direct question, he/she is requested to go to another RR device, R_2, in which there is another pack of cards marked A and A^C in proportions $p_2:(1-p_2), 0<p_2<1, p_2\neq 1/2$. The respondent is then instructed to choose a card randomly from this pack and to report the true value of y, i.e., either '1' or '0', if there is a match (mismatch) between his/her true y character and the card type drawn. Here, it is assumed that the sensitive and the innocuous questions are unrelated and also that the RR devices R_1 and R_2 are independent.

Suppose that out of the n selected persons n_1 reply '*yes*' to the direct question and the remaining $n_2 = n - n_1$ persons provided a '*no*' answer to it. Now, the following are defined:

I_i = 1 if the ith selected individual bears the sensitive character and draws an A – marked card or if the individual bears the non-sensitive character and chooses a B – marked card

= 0 else on using R_1.

Then $P(I_i = y_i) = p_1$ and $P(I_i = x_i) = 1 - p_1$ and writing E_R, V_R as the expectation and variance operators with respect to any arbitrary RR device it is easy to check that,

$$E_R(I_i) = p_1 y_i + (1-p_1)x_i$$

$$= p_1 y_i + (1-p_1).$$

This is because a respondent using the device R_1 has already responded '*yes*' to the initial direct innocuous question. Thus, it follows that for

$$r_i = [I_i - (1-p_1)]/p_1, 0<p_1<1,\ E_R(r_i) = y_i,\ \forall i \in U$$

and

$$V_R(r_i) = \frac{V_R(I_i)}{p_1^{\,2}} = \frac{(1-p)(1-y_i)^2}{p_1} = V_{1i}.$$

It may be seen that r_i is an unbiased estimator for y_i and also an unbiased estimator for V_{1i} is given by $v_{1i} = \frac{(1-p)(1-r_i)^2}{p_1}$. Further,

let $J_i = 1$ if ith selected individual bears the sensitive attribute A and draws an A-marked card = 0 else, on applying R_2. Then,

$$P(J_i = y_i) = p_2 \text{ and } P(J_i = 1 - y_i) = 1 - p_2$$

and

$$E_R(J_i) = p_2 y_i + (1-p_2)(1-y_i) = (2p_2 - 1)y_i + (1-p_2),$$
$$V_R(J_i) = p_2(1-p_2).$$

For $u_i = [J_i - (1-p_2)]/(2p_2 - 1),\ p_2 \neq 1/2$, there is $E_R(u_i) = y_i,\ \forall\ i \in U$ and $V_R(u_i) = \dfrac{p_2(1-p_2)}{(2p_2-1)^2} = V_{2i}$, say. Thus, u_i is also unbiased for y_i and an unbiased estimator of V_{2i} is given by $v_{2i} = V_{2i}$.

Let s_1 and s_2 be respectively the sets of sampled individuals offering '*yes*' and '*no*' responses to the initial direct innocuous question such that $s_1 \cup s_2 = s$ and write E_p, V_p respectively to denote the operators for expectation and variance with respect to the probability design p. Suppose that $t_k = \sum_{i=1}^{N} b_{s_k i} I_{s_k i} y_i$ where $I_{s_k i} = 1\,(0)$, if $i \in s_k\ (\notin s_k), k = 1, 2$ and $b_{s_k i}$'s are constants free of $\underline{Y} = (y_1, \ldots\ldots\ldots\ldots, y_N)$ such that $E_p(b_{s_k i} I_{s_k i}) = 1,\ \forall\ i \in U$ be a homogeneous linear unbiased estimator for $Y = \sum_{i=1}^{N} y_i$. The following is written as:

$$V_p(t_k) = \sum_{i=1}^{N} y_i^2 c_{ki} + \sum_{i \neq j} y_i y_j c_{kij}$$

where

$$c_{ki} = E_p(b_{s_k i}^2 I_{s_k i}) - 1$$

and

$$c_{kij} = E_p(b_{s_k i} I_{s_k i} - 1)(b_{s_k j} I_{s_k j} - 1)$$

and an unbiased estimator of $V_p(t_k), k = 1, 2$ as

$$v_p(t_k) = \sum_{i=1}^{N} y_i^2 c_{s_k i} I_{s_k i} + \sum_{i \neq j} y_i y_j c_{s_k ij} I_{s_k ij}$$

where $I_{s_k ij} = I_{s_k i} I_{s_k j}$ and $c_{s_k i}$, $c_{s_k ij}$ are $\underline{Y}$-free constants satisfying $E_p(c_{s_k i} I_{s_k i}) = c_{ki}$ and

$$E_p(c_{s_k ij} I_{s_k ij}) = c_{kij},\ k = 1, 2.$$

Because y_i's are unascertainable, two unbiased estimators for Y based on s_1 and s_2 are obtained

$$e_1 = \sum_{i \in s_1} b_{s_1 i} I_{s_1 i} r_i$$

and

$$e_2 = \sum_{i \in s_1} b_{s_2 i} I_{s_2 i} u_i$$

and accordingly, two unbiased estimators for $\pi_A = Y/N$ are given by

$$\overline{e_1} = e_1 / N \quad \text{and} \quad \overline{e_2} = e_2 / N.$$

Now, following Raj (1968) and Rao (1975), two unbiased estimators for $V(e_1)$ and $V(e_2)$ are obtained as:

$$v_1(e_1) = v_p(t_1)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} b_{s_1 i} I_{s_1 i} v_{1i}$$

$$v_2(e_1) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} (b_{s_1 i}^2 - c_{s_1 i}) I_{s_1 i} v_{1i}$$

$$v_1(e_2) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} b_{s_2 i} I_{s_2 i} V_{2i}$$

$$v_2(e_2) = v_p(t_2)\big|_{\underline{Y} = \underline{R}} + \sum_{i=1}^{N} (b_{s_2 i}^2 - c_{s_2 i}) I_{s_2 i} V_{2i}.$$

Since both e_1 and e_2 are unbiased estimators for Y, an unbiased estimator of Y based on e_1 and e_2 is given by

$$e = \frac{n_1}{n} e_1 + \frac{n_2}{n} e_2$$

and

$$V(e) = \left(\frac{n_1}{n}\right)^2 V(e_1) + \left(\frac{n_2}{n}\right)^2 V(e_2)$$

$$= \left(\frac{n_1}{n}\right)^2 V(e_1) + \left(1 - \frac{n_1}{n}\right)^2 V(e_2).$$

Also, an unbiased estimator of π_A is given by $\hat{\pi}_A = \frac{n_1}{n}\bar{e}_1 + \frac{n_2}{n}\bar{e}_2$. Again, as the two RR devices are independent, unbiased variance estimators for $V(e)$ are derived as

$$v_1(e) = \left(\frac{n_1}{n}\right)^2 v_1(e_1) + \left(\frac{n_2}{n}\right)^2 v_1(e_2)$$

$$v_2(e) = \left(\frac{n_1}{n}\right)^2 v_2(e_1) + \left(\frac{n_2}{n}\right)^2 v_2(e_2)$$

and similarly, the unbiased estimators for $V(\hat{\pi}_A)$ are given by

$$v_1(\hat{\pi}_A) = \left(\frac{n_1}{n}\right)^2 v_1(\bar{e}_1) + \left(\frac{n_2}{n}\right)^2 v_1(\bar{e}_2)$$

$$v_2(\hat{\pi}_A) = \left(\frac{n_1}{n}\right)^2 v_2(\bar{e}_1) + \left(\frac{n_2}{n}\right)^2 v_2(\bar{e}_2).$$

A Numerical Example

Artificial data relating to a community of N = 129 individuals is considered. As well, the problem of estimating the proportion of individuals evading income tax during the last financial year in the said community on choosing a sample of n = 37 individuals is considered. The individuals from this population were selected according to three different sampling schemes, namely, simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR) and Rao-Hartley-Cochran (RHC, 1962) sampling scheme as a representative of varying probability sampling.

Here, y_i = 1(0) is defined if the ith individual evades (does not evade) income tax during the last financial year and x_i = 1(0) if the ith individual prefers (does not prefer) football to basketball. The amount of expenditure incurred in a particular month in the household to which an individual belongs to is considered as the size-measure for selection of the individuals by RHC sampling strategy.

In the RHC scheme, first the population of N units is randomly divided into n random groups, the ith group having N_i units such that $\sum_n N_i = N$, where $\sum_n$ denotes the sum over the n random groups. Then, denoting $A_i = a_{i_1} + \cdots\cdots\cdots + a_{i_{N_i}}$ as the sum of the normed size-measures a_i's for the units belonging to the ith group, one unit is chosen from the ith group with a probability proportional to A_i divided by it's a-value. This process is repeated for all the n groups. Now, writing for simplicity (y_i, a_i) as the (y,a)-value for the unit selected from the ith group, an unbiased estimator for Y is given by

$$t = \sum_n (A_i / a_i)\, y_i$$

along with an unbiased variance estimator for $V(t)$ as

$$v(t) = B \sum_n A_i \left(\frac{y_i}{a_i} - t\right)^2$$

where

$$B = \left(\sum_n N_i^2 - N\right) \Big/ \left(N^2 - \sum_n N_i^2\right).$$

Here, y_i's are unknown and so are to be estimated. Suppose that w_i be an unbiased estimator for y_i and v_i be an unbiased estimator for $V_R(w_i)$. Then, one may employ the unbiased estimator

$$t = \sum_n (A_i / a_i)\, w_i$$

for estimating Y and an unbiased variance estimator of $V(e)$, following Chaudhuri, Adhikary and Dihidar (2000) is given by

$$v(e) = v(t)\big|_{\underline{Y}=\underline{W}} + \sum_{i=1}^{N} b_{si} I_{si} v_i$$

where $\underline{W} = (w_1, \cdots\cdots\cdots, w_N)$. Let e be any point estimator for the parameter θ and $v(e)$ be an unbiased estimator of $V(e)$. Then, assuming $\delta = (e - \theta) \Big/ \sqrt{v(e)}$ to be a standard normal deviate, the following two criteria are considered:

Table 1: Comparative performances of alternative procedures

p_1	p_2	RHC			SRSWOR			SRSWR		
		$\hat{\pi}_A$	CV	Length of CI	$\hat{\pi}_A$	CV	Length of CI	$\hat{\pi}_A$	CV	Length of CI
n_1 = 30										
0.98	0.47	0.65	11.4	0.366	0.40	16.9	0.264	0.59	18.5	0.265
0.92	0.48	0.74	15.0	0.397	0.37	17.4	0.281	0.46	18.9	0.313
0.93	0.76	0.68	14.9	0.475	0.32	17.3	0.276	0.40	18.1	0.315
0.81	0.84	0.85	17.9	0.466	0.34	21.6	0.319	0.34	24.9	0.362
0.89	0.68	0.65	16.4	0.491	0.32	19.4	0.290	0.42	22.1	0.327
n_1 = 25										
0.98	0.47	0.44	13.9	0.362	0.48	15.8	0.222	0.43	18.7	0.264
0.92	0.48	0.43	17.1	0.351	0.41	19.7	0.253	0.44	20.8	0.273
0.93	0.76	0.41	17.5	0.345	0.47	19.7	0.234	0.41	23.1	0.278
0.81	0.84	0.49	19.7	0.375	0.39	23.9	0.294	0.38	26.8	0.332
0.89	0.68	0.43	18.2	0.379	0.37	20.1	0.267	0.36	22.2	0.297
n_1 = 20										
0.98	0.47	0.33	15.1	0.282	0.35	18.9	0.217	0.32	20.3	0.242
0.92	0.48	0.39	18.6	0.229	0.39	21.2	0.210	0.32	23.7	0.258
0.93	0.76	0.32	19.4	0.260	0.31	22.6	0.235	0.30	24.6	0.260
0.81	0.84	0.29	21.7	0.206	0.24	24.1	0.275	0.24	27.6	0.297
0.89	0.68	0.27	21.6	0.257	0.36	24.2	0.230	0.30	26.8	0.267
n_1 = 15										
0.98	0.47	0.27	17.8	0.193	0.27	20.7	0.192	0.27	23.4	0.204
0.92	0.48	0.28	20.7	0.237	0.20	24.7	0.217	0.26	27.4	0.217
0.93	0.76	0.25	21.9	0.178	0.32	25.1	0.172	0.24	27.7	0.227
0.81	0.84	0.20	23.2	0.162	0.17	27.5	0.246	0.17	29.7	0.261
0.89	0.68	0.23	23.6	0.240	0.28	26.2	0.198	0.28	28.4	0.210

(i) the coefficient of variation (CV) defined as $CV = \left(\sqrt{v(e)}/e\right)\times 100$; and

(ii) the length of the confidence intervals (CI's) $\left(e-1.96\sqrt{v(e)}, e+1.96\sqrt{v(e)}\right)$ given by $2\times 1.96\sqrt{v(e)}$

for comparing the relative performances of the alternative sampling procedures.

For the artificial population $\pi_A = 0.6202$. Table 1 outlines the performances of the alternative estimators for different choices of n_1, p_1 and p_2.

Conclusion

Irrespective of the values of n_1, SRSWOR performs better than SRSWR in terms of the two criteria for comparison considered here and the RHC scheme turns out to be the best sampling scheme in terms of the criterion CV. As the values of n_1, i.e. the number of individuals replying '*yes*' to the initial direct question increases, improvement in the efficiency level of the estimator is observed for all three sampling designs.

This implies that for producing efficient estimators by applying the method discussed above, one has to choose the direct innocuous question judiciously so that more numbers of interviewees answer '*yes*' to the initial direct question. Thus, the extended method of estimation as discussed here may be effectively used in complex sample surveys for collection of information on sensitive attributes.

References

Chaudhuri (2004). Christofides' randomized response technique in complex sample surveys. *Metrika,60*(3), 23-228.

Chaudhuri, A. (2002). Estimating sensitive proportions from randomized responses in unequal probability sampling. *Calcutta Statistical Association Bulletin, 52*, (205-208), 315-322.

Chaudhuri, A. (2001a). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. J*ournal of Statistical Planning & Inference, 94*, 37 - 42.

Chaudhuri, A. (2001b). Estimating sensitive proportions from unequal probability sample using randomized responses. *Pakistan Journal of Statistics, 17*(3), 259 - 270.

Chaudhuri, A., Adhikary, A.K. and Dihidar, S. (2000). Mean square error estimation in multi-stage sampling. *Metrika, 52*(2), 115-131.

Chaudhuri, A. & Mukerjee, R. (1988). *Randomized response: Theory and techniques*. Marcel Dekker Inc. N.Y.

Christofides, T. C. (2003). A generalized randomized response technique. *Metrika, 57*, 195 – 200.

Greenberg, B. G., Abul-Ela, Simmons, W. R. & Horvitz, D. G. (1969). The unrelated question randomized response model: theoritical framework. *Journal of the American Statistical Association, 64*, 520-539.

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section of the American Statistical Association.* 65-72.

Kim, Jong-Min & Warde, D. W. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference, 133*(1), 211-221.

Kim, Jong-Min & Warde, D. W. (2004). A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference 120*, 155-165.

Kuk, A.Y.C. (1990). Asking sensitive question indirectly. *Biometrika, 77*, 436-438.

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

Raj, D. (1968). *Sampling Theory*. Mc-graw Hill. N.Y.

Rao, J. N. K (1975). Unbiased variance estimation for multi-stage designs. Sankhya C, 37, 133-139.

Rao, J. N. K., Hartley, H. O., & Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. B, 24*, 482-491.

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