The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 735 - 741

A New Mail Survey Method for Sensitive Character without Using Randomization Device[1)]

Ki Hak Hong[2)]

Abstract

In the present paper, we propose a new randomization device free mail survey method. The estimator based on proposed model is unbiased and more efficient than the estimator based on Singh, Mangat and Singh model(SMS-model)(1993) when $\pi < 1/2$, and more protective than SMS-model in view of the protection of privacy regardless of the values of π and π_Y only if we count the number of say 'Yes' from the respondents. However, If we consider the respondents that say 'No', the SMS-model is more protective than our model.

1. Introduction

The randomizing response (RR) procedure to procure trustworthy data for estimating the proportion π of the population belonging to a sensitive group was first introduced by Warner(1965). Since then, many people have developed the method. Especially Chaudhuri and Mukerjee(1988), Ryu et al.(1993) have systematically summarized various RR methods in their books.

These randomized response techniques need some randomization device such as a spinner (Warner,1965), a deck of cards with questions printed on them (Greenberg et al. 1969 ; Folsom et al. 1973), a pair of deck of cards with questions printed on them (Mangat and Singh;1990) or a sealed plastic box with colored beads(Liu and Chow ; 1976).

In mail surveys, the physical randomization device is not supplied to the respondent. Instead, he is instructed as how to provide the required information without disclosing his status with respect to the sensitive character under study, by using material at hand in the questionnaire.

Takahasi and Sakasegawa(1977) have suggested a RR technique without making use of any randomization device and suggested that such a technique would be applicable not only to

1) This research was supported by the Science Research Program of the Dongshin University, 1996.
2) Associate Professor, Dept. of Computer Science, Dongshin University, Daeho-Dong, Naju, Chonnam, 520-714, Korea.

face-to-face interview survey, but also to self-administrated, mail surveys. But their models seem to be complicated and not easily understandable by the respondent in the mail surveys.

Singh, Mangat and Singh(1993) have developed the Takahasi and Sakasegawa model (TS-model) to fit the mail survey and proposed a new RR model (SMS-model) using Hansen and Hurwitz(1946) technique of mail surveys. They have replaced p, the proportion in RR device representing the sensitive characteristic, by π_Y ,the proportion of population possessing nonstigmatized attribute (Y; preferring color or season) .

In this paper, we propose a new mail survey method and compare the efficiency with the SMS-model in view of the variance and the protection of privacy.

2. SMS-model

Singh, Mangat and Singh(1993) have proposed a mail survey model for the sensitive character without using randomization device that improved the TS-model. In their method, each interviewee in the simple random sampling without replacement sample of n respondents is instructed to say 'no' if he or she does not possess both the nonstigmatized attribute(Y) and stigmatized attribute (A). In all other cases he or she report 'yes'. The population consists of N respondents and all population can be divided into two groups. First group consists of N_1 respondents who will report at the first attempt i. e. from mail survey while the second group consists of those N_2 respondents who will not respond at the first attempt so that $N_1 + N_2 = N$.

Let $\pi_1 = \frac{N_{1A}}{N_1}$ and $\pi_2 = \frac{N_{2A}}{N_2}$ denote the proportion of the respondents possessing the attribute A in the first and second group of population, respectively. Then the proportion π of the respondents in the whole population possessing the attribute A is given by

$$\pi = N^{-1}\sum_{i=1}^{2} N_i\pi_i \ . \tag{2.1}$$

The probability of 'yes' answers for this procedure in the first and second attempt are respectively given by

$$\Theta_1 = \pi_1 + (1-\pi_1)\pi_Y \ , \tag{2.2}$$

$$\Theta_2 = \pi_2 + (1-\pi_2)\pi_Y \ . \tag{2.3}$$

The probability of 'yes' answers for the whole population is given by

$$\Theta = N^{-1}\sum_{i=1}^{2} N_i\Theta_i = \pi + (1-\pi)\pi_Y \ . \tag{2.4}$$

Suppose n respondents are selected using simple random sampling without replacement (SRSWOR) method. Let n_1 and n_2 denote the number of respondents who respond at the

first and second attempt, respectively. Let h_2 be a sub-sample of n_2 selected by SRSWOR such that $n_2 = h_2 g$, ($g \geq 1$), and interview personally. If n_{1A} and h_{2A} are the number of 'yes' answers out of n_1 and h_2, respectively.
The estimators of π_1 , π_2 and π are respectively given by

$$\hat{\pi}_1 = (\frac{n_{1A}}{n_1} - \pi_Y)/(1-\pi_Y) , \tag{2.5}$$

$$\hat{\pi}'_2 = (\frac{h_{2A}}{h_2} - \pi_Y)/(1-\pi_Y) , \tag{2.6}$$

$$\hat{\pi} = (n_1 \hat{\pi}_1 + n_2 \hat{\pi}'_2)/n . \tag{2.7}$$

theorem 2.1: $\hat{\pi}$ is an unbiased estimator of population proportion π .

theorem 2.2: The variance of estimator $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \frac{N-n}{n(N-1)} [\pi(1-\pi) + \frac{\pi_Y(1-\pi_Y)}{(1-\pi_Y)}] + \frac{(g-1)N^2_2}{nN(N_2-1)} \cdot [\pi_2(1-\pi_2) + \frac{\pi_y(1-\pi_2)}{1-\pi_Y}]. \tag{2.8}$$

They showed that their model was always more efficient in view of the variance than the TS-model when $\pi_Y < 2/3$.

3. The proposed model

In this proposed mail survey model each respondent which is selected with SRSWOR is instructed to say 'no' if he or she does possess A but does not possess Y . In all other cases he or she is instructed to say 'yes'. Using the proportional model, the probabilities of 'yes' answer from the first and second groups of the population are respectively given by

$$\Phi_1 = \pi_1\pi_Y + (1-\pi_1), \tag{3.1}$$

$$\Phi_2 = \pi_2\pi_Y + (1-\pi_2). \tag{3.2}$$

The probability of 'yes' answers for the whole population is given by

$$\Phi = N^{-1}\sum_{i=1}^{2} N_i\Phi_i = \pi\pi_Y + (1-\pi). \tag{3.3}$$

theorem 3.1: The unbiased estimator of population proportion π is given by

$$\hat{\pi}^h = (n_1 \hat{\pi}_1^h + n_2 \hat{\pi}'^h_2)/n. \tag{3.4}$$

proof: Let E_2 , V_2 respectively denote the conditional expectation and variance when those samples n_1 and n_2 are fixed. Let E_1 , V_1 respectively denote the expectation and variance for all possible selections of those samples. Then we have

$$E(\hat{\pi}^h) = E_1E_2(\hat{\pi}^h) = E_1(\hat{\pi}^h{}_n) = \pi ,$$

where $\hat{\pi}_1{}^h = (1-\frac{n_{1A}}{n_1})/(1-\pi_Y)$,

$\hat{\pi}'_2{}^h = (1-\frac{h_{2A}}{h_2})/(1-\pi_Y)$,

$\hat{\pi}_2{}^h = (1-\frac{n_{2A}}{n_2})/(1-\pi_Y)$ and

$$\hat{\pi}^h{}_n = E_2(\hat{\pi}^h) = \frac{n_1\,\hat{\pi}_1{}^h + n_2E_2(\hat{\pi}'_2{}^h)}{n} = (n_1\,\hat{\pi}_1{}^h + n_2\,\hat{\pi}_2{}^h)/n$$

is the estimator of π based on a sample of size n in absence of any non-response. This proves the theorem .

theorem 3.2: The variance of estimator $\hat{\pi}^h$ is given by

$$V(\hat{\pi}^h) = \frac{N-n}{n(N-1)}[\pi(1-\pi)+\frac{\pi_Y\pi}{(1-\pi_Y)}]+\frac{(g-1)N^2_2}{nN(N_2-1)}\cdot [\pi_2(1-\pi_2) + \frac{\pi_y\pi_2}{1-\pi_Y}]. \tag{3.5}$$

proof: For every respondent i of n, let y_i stand for the indicator random variable defined as $y_i=1$ if ith respondent say 'Yes', $y_i=0$, otherwise. Clearly, we can see that $n_A = \sum_{i=1}^{n} y_i$, and $\frac{n_A}{n} = \bar{y}$. In the same way, we can define another two indicator variables y_{2i} , y'_{2i} for every respondent i of n_2 and h_2, and obtain the equation $\frac{n_{2A}}{n_2} = \bar{y}_2(=\hat{\Phi}_2)$, $\frac{h_{2A}}{h_2} = \bar{y'}_2$. The proof is obtained by the following procedure.

$$\begin{aligned} V(\hat{\pi}^h) &= V_1E_2(\hat{\pi}^h) + E_1V_2(\hat{\pi}^h), \\ &= V_1(\frac{1-n_A/n}{1-\pi_Y}) + E_1(\frac{n_2}{n})^2V_2(\hat{\pi}'_2{}^h), \\ &= \frac{1}{(1-\pi_Y)^2}V_1(\bar{y}) + E_1(\frac{n_2}{n})^2\frac{1}{(1-\pi_Y)^2}V_2(\bar{y'}_2), \\ &= \frac{(N-n)}{N-1}\frac{\Phi(1-\Phi)}{n(1-\pi_Y)^2} + [\frac{N_2}{N}(\frac{g-1}{n})\frac{1}{(1-\pi_Y)^2}\frac{N_2\Phi_2(1-\Phi_2)}{N_2-1}]. \end{aligned} \tag{3.6}$$

where

$$V_2(\hat{\pi}^h) = \left(\frac{n_2}{n}\right)^2 V_2(\hat{\pi}'^h_2) = \left(\frac{n_2}{n}\right)^2 \frac{1}{(1-\pi_Y)^2} V_2(\overline{y}'_2),$$

$$= \left(\frac{n_2}{n}\right)^2 \frac{1}{(1-\pi_Y)^2} \frac{\hat{\Phi}_2(1-\hat{\Phi}_2)}{(n_2-1)} \frac{n_2-h_2}{h_2},$$

$$= \left(\frac{n_2}{n}\right) \frac{1}{(1-\pi_Y)^2} \frac{g-1}{n} \frac{n_2\,\hat{\Phi}_2(1-\hat{\Phi}_2)}{(n_2-1)},$$

$$, \; g = n_2/h_2.$$

$$V_2(\overline{y}'_2) = \frac{\hat{S}_2^{\,2}}{h_2}(1-f_2),$$

$$= \frac{1}{h_2}\frac{1}{n_2-1} n_2\,\hat{\Phi}_2(1-\hat{\Phi}_2)\frac{n_2-h_2}{n_2},$$

and

$$V_1(\overline{y}) = \frac{S^2}{n}(1-f) = \frac{N-n}{N}\frac{1}{n}\frac{N}{N-1}\Phi(1-\Phi).$$

On putting the equations (3.1) and (3.3) respectively in the equation (3.6) we get the equation (3.5) after some simplifications. This proves the theorem.

4. Efficiency Comparison

Firstly, we compare the efficiency of our estimator based on the proposed model with the estimator based on the SMS model in view of the variance.

If $V(\hat{\pi}) > V(\hat{\pi}^h)$, then we can say that our estimator is more efficiency than that of SMS-model.

From the equation (2.8) and (3.5),

$$V(\hat{\pi}) - V(\hat{\pi}^h)$$

$$= \frac{N-n}{n(N-1)}[\pi(1-\pi)+\frac{\pi_Y(1-\pi)}{(1-\pi_Y)}]+\frac{(g-1)N^2_2}{nN(N_2-1)} \cdot$$

$$[\pi_2(1-\pi_2) + \frac{\pi_y(1-\pi_2)}{1-\pi_Y}]$$

$$- \frac{N-n}{n(N-1)}[\pi(1-\pi)+\frac{\pi_Y\pi}{(1-\pi_Y)}]+\frac{(g-1)N^2_2}{nN(N_2-1)} \cdot ,$$

$$[\pi_2(1-\pi_2) + \frac{\pi_y\pi_2}{1-\pi_Y}]$$

$$= \frac{N-n}{n(N-1)} \cdot \frac{\pi_Y}{(1-\pi_Y)}(1-2\pi) + \frac{(g-1)N_2^{\,2}}{nN(N_2-1)} \cdot \frac{\pi_Y}{(1-\pi_Y)}(1-2\pi_2),$$

$$> 0 . \qquad (4.1)$$

The inequality (4.1) will always hold no matter how the values of π_Y if $\pi < \frac{1}{2}$ and $\pi_2 < \frac{1}{2}$.

Since the sensitive proportion of population π that we wish to estimate has generally the very small value which is far less than 1/2, it is reasonable to assume that π and π_2 are less than $1/2$. we say that our proposed model is more efficiency than SMS-model.

The efficiency of SMS-model depends on the value of π_Y. So if we can't find the value of π_Y that satisfies the required condition, we can't use that in mail survey. But our model is only depends on the values of π and π_2 which usually have small values. Hence we can use our model regardless of the value of π_Y for most sensitive mail survey.

Secondary, we appreciate the proposed strategy by using the procedure described in Hong and Lee(1996). In SMS-model, the jeopardy function of 'yes' with respect to A and the jeopardy function of 'no' with respect to $\overline{A}$ are

$$g_{sms}(y) = g(Y, A) = \frac{P(Yes \mid A)}{P(Yes \mid \overline{A})} = \frac{1}{\pi_Y}, \tag{4.2}$$

$$g_{sms}(n) = g(N, A) = \frac{P(No \mid A)}{P(No \mid \overline{A})} = \frac{0}{1-\pi_Y} = 0, \tag{4.3}$$

The jeopardy functions of the proposed model in this paper are as follows

$$g_h(y) = g(Y, A) = \frac{P(Yes \mid A)}{P(Yes \mid \overline{A})} = \frac{\pi_Y}{1} = \pi_Y, \tag{4.4}$$

$$g_h(n) = g(N, A) = \frac{P(No \mid A)}{P(No \mid \overline{A})} = \frac{1-\pi_Y}{0} \approx \infty. \tag{4.5}$$

That the value of jeopardy function, g is greater than unity means increasing respondent's jeopardy for given answer and the attribute. In the case of 'yes' answer for attribute A, we can show that from the equations (4.2) and (4.4), our proposed model is always more protective than the SMS-model irrespective of the values of π_Y and π. In the case of 'no' answer for attribute $\overline{A}$, it is clear that from the equations (4.3), (4.5), the SMS-model is always more protective than our model irrespective of the values of π_Y and π. If only we count the number of say 'Yes' from the respondents, our model is more protective than the SMS-model. However, If we consider the respondents that say 'No', the SMS-model is more protective than our model.

References

[1] Chaudhuri,A. and Mukerjee, R. (1988). *Randomized response : theory and techniques*, Marcel Dekker, Inc., New York.

[2] Folsom,et al. (1973).The two alternative questions randomized response model for human surveys, *Journal of the American Statistical Association,* Vol. 68, 525-530.

[3] Greenberg, et al. (1969). The unrelated question randomized response model : model: theoretical framework, *Journal of the American Statistical Association,* June, 521-539.

[4] Hansen, M.H. and Hurwitz, W.N. (1946) The problems of non-response in sample surveys, *Journal of the American Statistical Association,* Vol. 41, 517-529.

[5] Hong, K.H. and Lee, K.S. (1996) The improvement of Mangat Strategy in view of the protection of privacy, *The Korean Communications in Statistics,* Vol. 3, No.2, 169-174.

[6] Leysieffer, F.W. and Warner,S.L. (1976). Respondent jeopardy and optimal designs in RR models, *Journal of the American Statistical Association,* Vol. 72, 649-656.

[7] Liu,P.T.and Chow, L.P. (1976).A new discrete quantitative randomized response model, *Journal of the American Statistical Association,* Vol. 71, 72-73.

[8] Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure, *Biometrika,* Vol. 77, 439-442.

[9] Sing,R. , Mangat, N.S. and Sing, S. (1993). A mail survey design for sensitive character without using randomization device, *Communications in statistics - Theory and methodology* -, Vol. 22, No. 9, 2661-2668.

[10] Ryu,J.B., Hong, K.H. and Lee, K.S. (1993). *Randomized response model,* Freedom Acadamy, Inc., Seoul .

[11] Takahasi,K. and Sakasegawa,H. (1977). A randomized response technique with-without making use of any randomizing device, *Annals Institute of Statistical Mathematics,* Vol. 29, Part A, 1-8.

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