Metrika (2005) 61: 1–7
DOI 10.1007/s001840400319

Metrika

A two-stage stratified Warner's randomized response model using optimal allocation

Jong-Min Kim[1] and M. E. Elam[2]

[1]Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, 56267, USA (E-mail: jongmink@mrs.umn.edu)
[2]Department of Industrial Engineering, The University of Alabama, Box 870288, Tuscaloosa, AL, 35487, USA

Abstract. This paper applies Kim and Warde's (2004) stratified Warner's randomized response model to Mangat and Singh's (1990) two-stage randomized response model. The proposed stratified randomized response model has an optimal allocation and a large gain in precision. Hence, the estimator based on the proposed method is more efficient than Kim and Warde's (2004) and Mangat and Singh's (1990) estimators under the conditions presented in both the case of completely truthful reporting and that of not completely truthful reporting by the respondents.

Key words: Randomized response technique, stratified random sampling, sensitive characteristics, dichotomous population, estimation of proportion

1. Introduction

Warner (1965) introduced the randomized response (RR) model as an alternative survey technique for socially undesirable or incriminating behavior questions in order to reduce response error, protect a respondent's privacy, and increase response rates. Warner's model draws respondents using simple random sampling with replacement from the population. It requires the interviewee to give a "Yes" or "No" answer either to the sensitive question or to its negative depending on the outcome of a randomizing device not reported to the interviewer.

Greenberg et al. (1969) derived results for Warner's model in the case of less than completely truthful reporting. Mangat and Singh (1990) proposed a two-stage RR model in which each interviewee (who is selected using simple random sampling with replacement) is provided with two randomization devices. The first one consists of two statements: 1) "I belong to the sensitive trait group" and 2) "Go to the second randomization device". The second

randomization device also consists of two statements, which are "I belong to the sensitive group" and "I do not belong to the sensitive group".Hong et al. (1994) suggested a stratified RR technique using a proportional allocation. A problem with the Hong et al. model is that it may cause a high cost because of the difficulty in obtaining a proportional sample from each stratum. To rectify this problem, Kim and Warde (2004) presented a stratified RR technique using an optimal allocation which is more efficient than that using a proportional allocation.

In this paper, we apply Kim and Warde's (2004) stratified Warner's RR model to Mangat and Singh's (1990) two-stage RR model. It is shown that the estimator resulting from the proposed model is more efficient than those for its component models under the conditions presented in the cases of completely and not completlely truthful reporting. It should be noted that Chaudhuri and Mukerjee (1988) and Singh and Mangat (1996) provide more comprehensive reviews of the RR literature.

2. Proposed model

In the proposed model, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement from each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. In the first stage of the survey interview, an individual respondent in the sample from stratum i is instructed to use the randomization device R_{1i} which consists of a sensitive question (S) card with probability M_i and a "Go to the randomization device R_{2i} in the second stage" direction card with probability $1 - M_i$. The respondents in the second stage of stratum i are instructed to use the randomization device R_{2i} which consists of a sensitive question (S) card with probability P_i and its negative question $(\bar{S})$ card with probability $1 - P_i$. The respondent should answer the question with a "Yes" or a "No" without reporting which question card he or she has in order to protect the respondent's privacy. Let n_i denote the number of units in the sample from stratum i and n denote the total number of units in the samples from all strata so that $n = \sum_{i=1}^{k} n_i$. Under the assumption that the "Yes" and "No" reports are made truthfully and M_i and P_i are set by the researcher, the probability of a "Yes" answer in stratum i for this procedure is:

$$Y_i = M_i\pi_{S_i} + (1 - M_i)[P_i\pi_{S_i} + (1 - P_i)(1 - \pi_{S_i})] \quad \text{for } i = 1, 2, \ldots, k \qquad (2.1)$$

where Y_i is the proportion of "Yes" responses and π_{S_i} is the proportion of respondents with the sensitive trait in the sample from stratum i. The maximum likelihood estimate of π_{S_i} is:

$$\hat{\pi}_{S_i} = \frac{\hat{Y}_i - -(1 - M_i)(1 - P_i)}{2P_i - 1 + 2M_i(1 - P_i)} \quad \text{for } i = 1, 2, \ldots, k \qquad (2.2)$$

where $\hat{Y}_i$ is the estimate of the proportion of "Yes" answers in the sample from stratum i. Since each $\hat{Y}_i$ has a binomial distribution $B(n_i, Y_i)$, $\hat{\pi}_{S_i}$ is an unbiased estimate for π_{S_i}. The variance of $\hat{\pi}_{S_i}$ is:

$$\operatorname{var}(\hat{\pi}_{S_i}) = \frac{\pi_{S_i}(1 - \pi_{S_i})}{n_i} + \frac{(1 - M_i)(1 - P_i)[1 - (1 - M_i)(1 - P_i))]}{n_i[2P_i - 1 + 2M_i(1 - P_i)]^2}. \qquad (2.3)$$

It should be noted that by removing the subscript i, (2.1)–(2.3) reduce to Mangat and Singh's (1990) equations (2.1)–(2.3) (with some differences in notation)).

Since the selections in different strata are made independently, the estimators for individual strata can be added together to obtain an estimator for the whole population. The maximum likelihood estimator of π_S, the proportion of respondents with the sensitive trait, is:

$$\hat{\pi}_S = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} = \sum_{i=1}^{k} w_I \left[\frac{\hat{Y}_i - (1 - M_i)(1 - P_i)}{2P_i - 1 + 2M_i(1 - P_i)} \right] \tag{2.4}$$

where N is the number of units in the whole population, N_i is the total number of units in stratum i, and $w_i = (N_i/N)$ for $i = 1, 2, \ldots, k$, so that $w = \sum_{i=1}^{k} w_i = 1$.

Theorem 2.1. *The proposed estimator $\hat{\pi}_S$ is an unbiased estimate for the population proportion.*

Proof. This follows from taking the expected value of (2.4). ■

Theorem 2.2. *The variance of the estimator $\hat{\pi}_S$ is:*

$$\operatorname{var}(\hat{\pi}_S) = \sum_{i=1}^{k} \frac{w_i^2}{n_i} \left\{ \pi_{S_i}(1 - \pi_{S_i}) + \frac{(1 - M_i)(1 - P_i)[1 - (1 - M_i)(1 - P_i)]}{[2P_i - 1 + 2M_i(1 - P_i)]^2} \right\}. \tag{2.5}$$

Proof. This follows from taking the variance of (2.4) and from corollary 1 in Sec. 5.9 of Cochran (1977).

Information on π_{S_i} is usually unavailable. But if prior information on π_{S_i} is available from past experience then we may derive the following optimal allocation formula. ■

Theorem 2.3. *The optimal allocation of n to $n_1, n_2, \cdots, n_{k-1}$ and n_k to derive the minimum variance of $\hat{\pi}_S$ subject to $n = \sum_{I=1}^{k} n_i$ is approximately given by:*

$$\frac{n_i}{n} = \frac{w_i \left\{ \pi_{S_i}(1 - \pi_{S_i}) + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{[2P_i-1+2M_i(1-P_i)]^2} \right\}^{1/2}}{\sum_{i=1}^{k} w_i \left\{ \pi_{S_i}(1 - \pi_{S_i}) + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{[2P_i-1+2M_i(1-P_i)]^2} \right\}^{1/2}}. \tag{2.6}$$

Proof. Follows from Sec. 5.9 of Cochran (1977).

The minimal variance of the estimator $\hat{\pi}_S$ is given by:

$$\operatorname{var}(\hat{\pi}_S) = \frac{1}{n} \left[\sum_{i=1}^{k} w_i \left\{ \pi_{S_i}(1-\pi_{S_i}) + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{[2P_i-1+2M_i(1-P_i)]^2} \right\}^{1/2} \right]^2. \tag{2.7}$$

By substituting $n_i - 1$ for n_i in (2.5), the unbiased minimal variance of the estimator $\hat{\pi}_S$ can be derived. ■

3. Efficiency comparison with variations of the Warner model

We will do an efficiency comparison of our stratified randomized response technique and the two-stage randomized response technique that was presented by Mangat and Singh (1990) by comparing variances.

Theorem 3.1. *Suppose that there are two strata in the population,* $n = n_1 + n_2$, $P = P_1 = P_2 \neq 0.5$ *(P is the probability of selecting the sensitive question in the second stage),* $M = M_1 = M_2$ *(M is the probability of selecting the sensitive question in the first stage), and* $\hat{\pi}_S = w_1\hat{\pi}_{S_1} + w_2\hat{\pi}_{S_2}$. *The proposed estimator* $\hat{\pi}_S$ *is more efficient than the Mangat and Singh (1990) estimator* $\hat{\pi}_{ms}$ *where* $\pi_{S_1} \neq \pi_{S_2}$.

The following theorem shows that our proposed estimator is more efficient than that of Kim and Warde (2004).

Theorem 3.2. *Assume that there are two strata in the population,* $n = n_1 + n_2$, $P = P_1 = P_2 \neq 0.5$, *and* $M = M_1 = M_2$. *The proposed estimator* $\hat{\pi}_S$ *will be more efficient than the Kim and Warde (2004) estimator* $\hat{\pi}_{kw}$ *when* $\pi_{S_1} \neq \pi_{S_2}$ *under the following condition:*

$$M > (1 - 2P)/(1 - P). \tag{3.1}$$

If prior information on $\pi_{S_1}, \pi_{S_2}, w_1, w_2$, and n can be obtained and $M = M_1 = M_2$ and $P = P_1 = P_2 \neq 0.5$ are chosen by the researcher, then we can check the relative efficiency of $\text{var}(\hat{\pi}_{kw})/\text{var}(\hat{\pi}_S)$. We do this in Table 1 (which is available at http://tables.20m.com) for two strata in the population by setting seven different P's and three different M's which, in the appropriate combination, satisfy condition (3.1). The results show that the proposed estimator $\hat{\pi}_S$ is more efficient than the Kim and Warde (2004) estimator $\hat{\pi}_{kw}$. When $M = 0.6$, Fig. 1 shows that the relative efficiency of $\hat{\pi}_S$ with respect to $\hat{\pi}_{kw}$ increases as P increases, and that there is little reduction of the relative efficiency as π_S increases. If we set $M = 0$ in the proposed stratified RR model, then the model reduces to the Kim and Warde (2004) stratified RR model.

4. Less than completely truthful reporting

We denote T_1 to be the weighted probability $T_1 = \sum_{i=1}^{k} w_i T_{1_i}$, where T_{1_i} is the probability that a respondent with the sensitive trait will report truthfully at the first stage in a sample from stratum i. Additionally, we denote T_2 to be the weighted probability $T_2 = \sum_{i=1}^{k} w_i T_{2_i}$, where T_{2_i} is the probability that a respondent with the sensitive trait will report truthfully at the second stage in a sample from stratum i. We assume that the respondents with the non-sensitive trait will report truthfully.

The probability of a "Yes" answer in stratum i for this procedure is given by:

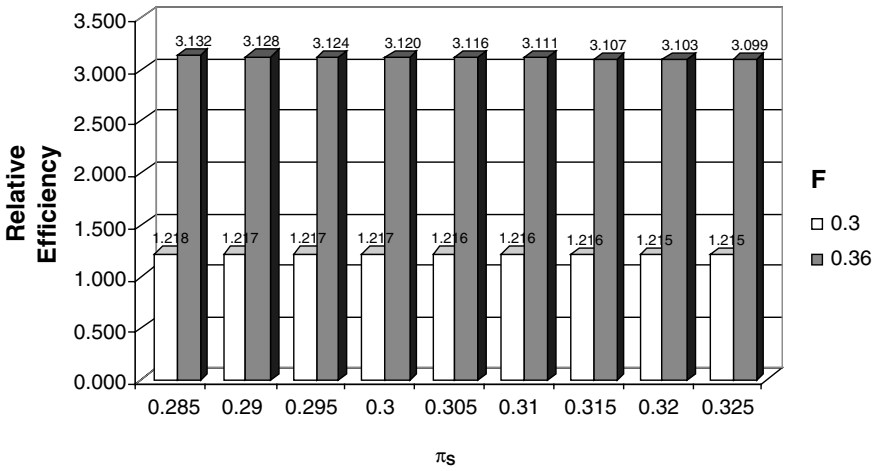

Fig. 1. The relative efficiency of $\mathrm{var}(\hat{\pi}_{kw})/\mathrm{var}(\hat{\pi}_S)$ when $M = 0.6$

$$Y_i^{'} = M_i\pi_{S_i}T_1 + (1-M_i)\{\pi_{S_i}P_iT_2 + \pi_{S_i}(1-P_i)(1-T_2) + (1-P_i)(1-\pi_{S_i})\} \tag{4.1}$$

where $i = 1, 2, \ldots, k$.

Therefore, a biased estimator $\hat{\pi}_S'$ of π_S in the population has the following bias and variance:

$$\begin{aligned} Bias(\hat{\pi}_S^{'} = E(\hat{\pi}_S^{'} - \hat{\pi}_S) &= \sum_{i=1}^{k} w_i E\left(\hat{\pi}_{S_i}^{'} - \hat{\pi}_{S_i}\right) \\ &= \sum_{i=1}^{k} w_i\pi_{S_i}\left(\frac{M_i(T_1-T_2)}{2P_i - 1 + 2M_i(1-P_i)} + T_2 - 1\right) \end{aligned} \tag{4.2}$$

Using equation (3.3) in Mangat and Singh (1990) and (2.7), we can derive the variance of $\hat{\pi}_S'$ as follows:

$$\begin{aligned} \mathrm{var}(\hat{\pi}_S^{'}) = \frac{1}{n}\Bigg[\sum_{i=1}^{k} w_i\Bigg\{ &\pi_{S_i}T_2(1-\pi_{S_i}T_2) + \frac{(1-M_i)(1-P_i)\{1-(1-M_i)(1-P_i)\}}{\{2P_i-1+2M_i(1-P_i)\}^2} \\ &+ \frac{\pi_{s_i}M_i(T_1-T_2)[1-2(1-M_i)(1-P_i)-\pi_{S_i}\{M_i(T_1-T_2)+4M_iT_2(1-P_i)+2T_2(2P_i-1)\}]}{\{2P_i-1+2M_i(1-P_i)\}^2}\Bigg\}^{1/2}\Bigg]^2. \end{aligned} \tag{4.3}$$

The mean square error of $\hat{\pi}_S'$ is given by:

$$\begin{aligned} MSE(\hat{\pi}_S') = \frac{1}{n}\Bigg[\sum_{i=1}^{k} w_i\Bigg\{ &\pi_{S_i}T_2(1-\pi_{S_i}T_2) + \frac{(1-M_i)(1-P_i)\{1-(1-M_i)(1-P_i)\}}{\{2P_i-1+2M_i(1-P_i)\}^2} \\ &+ \frac{\pi_{S_i}M_i(T_1-T_2)[1-2(1-M_i)(1-P_i)-\pi_{S_i}\{M_i(T_1-T_2)+4M_iT_2(1-P_i)+2T_2(2P_i-1)\}]}{\{2P_i-1+2M_i(1-P_i)\}^2}\Bigg\}^{1/2}\Bigg]^2 \\ &+ \left\{\sum_{i=1}^{2} w_i\pi_{S_i}\left(\frac{M_i(T_1-T_2)}{2P_i-1+2M_i(1-P_i)} + T_2 - 1\right)\right\}^2. \end{aligned} \tag{4.4}$$

Using equation (3.3) in Mangat and Singh (1990) and (4.4), the efficiency of the proposed estimator $\hat{\pi}'_S$ and the Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ in a situation of less than completely truthful reporting can be derived. Suppose there are two strata in the population and $P = P_1 = P_2$ and $M = M_1 = M_2 = 0$. $MSE(\hat{\pi}'_S)$ and $MSE(\hat{\pi}'_{ms})$ reduce to $MSE(\hat{\pi}'_{kw})$ and $MSE(\hat{\pi}'_w)$, respectfully. In the case where $M = M_1 = M_2 \neq 0$, we perform an empirical analysis. Using equation (3.3) in Mangat and Singh (1990), we get the mean square error of $\hat{\pi}'_{ms}$ from Mangat and Singh (1990):

$$MSE(\hat{\pi}'_{ms}) = \frac{\pi_S T_r(1-\pi_S T_r)}{n} + \frac{(1-M)(1-P)[1-(1-M)(1-P)]}{n[2P-1+2M(1-P)]^2} + [\pi_S(T_r-1)]^2$$
$$+\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$
$$-2(1-M)(1-P)-2\pi_s n\{2M(1-P)+2P-1\}]\left[n\{2P-1+2M(1-P)\}^2\right]^{-1}$$

where T and T_r are the probabilities that a respondent with the sensitive trait will report truthfully at the first and second stages, respectfully. Assume that there are two strata in the population, $T = T_1$, $T_r = T_2$, $M = M_1 = M_2 \neq 0$, and $P = P_1 = P_2 \neq 0.5$. If a researcher could obtain prior information on $\pi_{S_1}, \pi_{S_2}, w_1, w_2, n, M, P, T_1$, and T_2, then he or she can check the relative efficiency of $MSE(\hat{\pi}'_{ms})/MSE(\hat{\pi}'_S)$. We do this in Table 2 (which is available at http://tables.20m.com) for differing levels of n, P, T_1 and T_2. Table 2 shows that the values of the relative efficiency are more than one. Therefore, we can say that the proposed estimator $\hat{\pi}'_S$ is more efficient than the Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ in the case of two strata in terms of the relative efficiency $MSE(\hat{\pi}'_{ms})/MSE(\hat{\pi}'_S)$.

We now perform an empirical analysis to compare the MSE of the proposed estimator $\hat{\pi}'_S$ (see Equ. (4.4)) to that of the Kim and Warde (2004) estimator $\hat{\pi}'_{kw}$ in the case of two strata in the population, $T = T_1$, $T_r = T_2$, $T_1 \geq T_2$, $M = M_1 = M_2 \neq 0$, and $P = P_1 = P_2 > 0.5$. Under prior information on $\pi_{S_1}, \pi_{S_2}, w_1, w_2, M$, and differing levels of P, T_1, and T_2, Table 3 (which is also available at http://tables.20m.com) shows that the values of the relative efficiencies $MSE(\hat{\pi}'_{kw})/MSE(\hat{\pi}'_S)$ are more than one. Therefore, we can say that

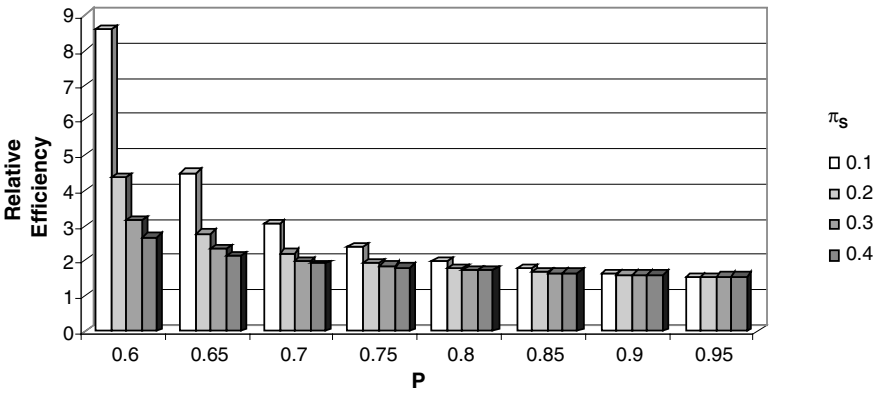

Fig. 2. The relative efficiency of $MSE(\hat{\pi}'_{kw})/MSE(\hat{\pi}'_S)$ when $T_1 = 0.8$, $T_2 = 0.7$, $M = 0.6$, and $n = 1000$

the proposed estimator $\hat{\pi}'_S$ is more efficient than the Kim and Warde (2004) estimator $\hat{\pi}'_{kw}$ in the case of two strata in the population. Fig. 2 shows results from Table 3 for $T_1 = 0.8$, $T_2 = 0.7$, $M = 0.6$, and $n = 1000$. The value of the relative efficiency is decreasing as P increases. Likewise, the relative efficiency decreases as π_S increases.

5. Discussion

This paper presented a new stratified randomized response model using the Mangat and Singh (1990) model. In the situations of completely truthful reporting and less than completely truthful reporting, we showed that the proposed randomized response model is more efficient than the Kim and Warde (2004) stratified randomized response model and the original Mangat and Singh (1990) model with the conditions presented. In addition to the gain in precision, the proposed method is more useful than the previous methods in that a stratified randomized response method helps to solve the limitation of randomized response that is the loss of the individual characteristics of the respondents. In future research, we will apply the stratified RR method to Mangat's (1994a) RR strategy and Mangat's (1994b) optional RR sampling technique. Researchers can apply the proposed method to medical- or criminal-related research topics with these advantages.

Acknowledgement. The authors are grateful to the referees for several valuable comments and suggestions.

References

Chaudhuri A, Mukerjee R (1988) Randomized Response: Theory and Techniques. Marcel Dekker, New York

Cochran WG (1977) Sampling Techniques, 3rd edn. John Wiley and Sons, New York

Greenberg BG, Abul-Ela AA, Simmons WR, Horvitz DG (1969) The unrelated question randomized response: theoretical framework. Journal of the American Statistical Association 64:529–539

Hong K, Yum J, Lee H (1994) A stratified randomized response technique. Korean Journal of Applied Statistics 7:141–147

Kim J-M, Warde WD (2004) A stratified Warner's randomized response model. Journal of Statistical Planning and Inference 120(1–2), pp 155–165

Mangat NS (1994a) An improved randomized response strategy. Journal of the Royal Statistical Society, B56(1):93–95

Mangat NS (1994b) An optional randomized response sampling technique. Journal of the Indian Statistical Association, 32(2):71–75

Mangat NS, Singh R (1990) An alternative randomized response procedure. Biometrika 77:439–442

Singh R, Mangat NS (1996) Elements of Survey Sampling. Kluwer Academic Publishers, The Netherlands

Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60:63–69