Statistical Papers 48, 215-233 (2007)

Statistical Papers
© Springer-Verlag 2007

A stratified unrelated question randomized response model

Jong-Min Kim[1], Matthew E. Elam[2]

[1] Division of Science and Mathematics,
Unversity of Minnesota, Morris, MN, 56267, USA
[2] Department of Industrial Engineering,
The University of Alabama, Tuscaloosa, AL, 35487, USA

Received: April 29, 2004; revised version: April 25, 2005

This paper presents a new randomized response model that combines Kim and Warde's (2004) stratified Warner's randomized response technique using optimal allocation with the unrelated question randomized response model. The empirical studies performed show that, for the prior information given, the new model is more efficient in terms of variance (in the case of completely truthful reporting) and mean square error (in the case of less than completely truthful reporting) than its component models.

Key words: Randomized response; Stratified random sampling; Sensitive characteristics; Dichotomous population; Estimation of proportion.

1. INTRODUCTION

In situations where potentially embarrassing or incrimi-nating responses are sought, the randomized response (RR) technique is effective in reducing non-sampling errors in sample surveys. Refusal to respond and lying in surveys are two main sources of such non-sampling errors, as the stigma attached to certain

practices (e.g., sexual behaviors and the use of illegal drugs) oftentimes leads to discrimination.

Warner (1965) did the pioneering work of an RR technique which minimizes underreporting of a data relative to socially undesirable or incriminating behavior questions. Hong et al. (1994) suggested a stratified RR technique under the proportional sampling assumption. Under Hong et al.'s (1994) proportional sampling assumption, it may be easy to derive the variance of the proposed estimator. However, it may come at a high cost in terms of time, effort, and money. For example, obtaining a fixed number of samples from a rural county in Minnesota through a proportional sampling method may be very difficult compared to the researcher's time, effort, and money.

To rectify this problem, Kim and Warde (2004) and Kim and Elam (2005) both presented stratified RR techniques using an optimal allocation which are more efficient than a stratified RR technique using a proportional allocation. The extension of the randomized response technique to stratified random sampling may be useful if the investigator is interested in estimating the proportion of HIV/AIDS positively affected persons at different levels such as by rural areas or urban areas, age groups, or income groups.

Here we develop a stratified RR model for Greenberg et al.'s (1969) unrelated question RR model. According to Greenberg et al. (1969), in the unrelated question approach, the respondent might be more truthful when presented with the opportunity to reply to one of two questions, one of which is completely innocuous and unrelated to the stigmatizing attribute. In Section 2 we review the relevant literature on randomized response techniques. In Section 3 we present our model in the cases where the proportion of respondents with the nonsensitive trait in a stratum is known and unknown. In Section 4 we show the results of two empirical studies, both in the case of completely truthful reporting. The first one shows that, for the prior information given, our new model is more efficient in terms of variance than Kim and Warde's (2004) stratified Warner's RR model. The second one shows that, for the prior information given, our new model is more efficient in terms of variance than Greenberg et al.'s (1969) unrelated question RR model. In Section 5 we present the less than completely truthful reporting counterparts to Sections 3 and 4. The empirical studies in

Section 5 indicate that, for the prior information given, our new model is more efficient in terms of mean square error than Kim and Warde's (2004) and Greenberg et al.'s (1969) models. In Section 6 we offer some concluding remarks.

2. A REVIEW OF RANDOMIZED RESPONSE TECHNIQUES

The Warner (1965) and Greenberg et al. (1969) models draw respondents using simple random sampling with replacement from the population. The Warner model required 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) proposed the unrelated question RR model that is a variation of Warner's (1965) RR model. Under the assumption of π_N (the proportion of people with the non-sensitive trait) unknown, Greenberg et al. (1969), in his equations (2.1) and (2.2), derived the probability Y_i of a respondent answering "Yes" to the sensitive statement in one of two independent, non-overlapping simple random samples of size n_1 and n_2 (see Greenberg et al. (1969), p.523) to be:

$$Y_i = P_i\pi_S + (1-P_i)\pi_N \quad \text{for } i=1,2 \tag{2.1}$$

where π_S is the proportion of people with the sensitive trait, P_i is the probability of selecting the sensitive question for each sample i, $1-P_i$ is the probability of selecting the nonsensitive question for each sample i, and $P_1 \neq P_2$ (see Greenberg et al. (1969), p.524).

Under the assumption that the total number of "Yes" responses is known from the sample, the estimator $\hat{\pi}_M$ for π_s is given by Greenberg et al. (1969) (see his equation (2.4)) as follows:

$$\hat{\pi}_M = \frac{(1-P_2)\hat{Y}_1 - (1-P_1)\hat{Y}_2}{P_1 - P_2}, \tag{2.2}$$

which is an unbiased estimator because the observed proportions $\hat{Y}_1$ and $\hat{Y}_2$ are binomially distributed with parameters (n_1, Y_1) and (n_2, Y_2). Since $Var(\hat{Y}_i) = Y_i(1-Y_i)/n_i$ and $\hat{Y}_1$ and $\hat{Y}_2$ are independent, the variance of the unbiased estimator $\hat{\pi}_M$ is given by Greenberg et al. (1969) (see his equation (2.5)) as follows:

$$Var(\hat{\pi}_M) = \left(\frac{1}{P_1 - P_2}\right)^2 \left\{\frac{Y_1(1-Y_1)}{n_1}(1-P_2)^2 + \frac{Y_2(1-Y_2)}{n_2}(1-P_1)^2\right\}. \tag{2.3}$$

The mean square error of $\hat{\pi}'_M$, the unbiased estimator for π_S in the case of less than completely truthful "Yes" answers to the sensitive statement, is:

$$MSE(\hat{\pi}'_M) = \left(\frac{1}{P_1 - P_2}\right)^2 \times$$
$$\left\{\frac{Y_1'(1-Y_1')}{n_1}(1-P_2)^2 + \frac{Y_2'(1-Y_2')}{n_2}(1-P_1)^2\right\} + \{\pi_S(T_r - 1)\}^2 \tag{2.4}$$

where $Y_i' = P_i\pi_S + (1-P_i)\pi_N$ for $i = 1,2$, T_r is the probability that a respondent with the sensitive trait will report truthfully, and the right-hand-side of (2.4) is the squared bias (see Greenberg et al. (1969), p.530).

After the optimal allocation of n $(= n_1 + n_2)$ to n_1 and n_2, the minimum variance of an estimator $\hat{\pi}_M$ is:

$$Var(\hat{\pi}_M) = \frac{\left[(1-P_2)\sqrt{Y_1(1-Y_1)} + (1-P_1)\sqrt{Y_2(1-Y_2)}\right]^2}{n(P_1 - P_2)^2} \tag{2.5}$$

The mean square error of $\hat{\pi}'_M$ after the optimal allocation of n to n_1 and n_2 is:

$$MSE(\hat{\pi}'_M) = \frac{\left[(1-P_2)\sqrt{Y_1'(1-Y_1')} + (1-P_1)\sqrt{Y_2'(1-Y_2')}\right]^2}{n(P_1-P_2)^2} + \{\pi_S(T_r-1)\}^2. \quad (2.6)$$

Kim and Warde (2004) presented a stratified Warner's RR model using an optimal allocation which is more efficient than Hong et al.'s (1994) model. They derived the probability Z_i of a "Yes" answer in stratum i for this procedure as:

$$Z_i = P_i\pi_{S_i} + (1-P_i)(1-\pi_{S_i}) \quad \text{for } i = 1,2,\cdots,k \quad (2.7)$$

where π_{S_i} is the proportion of people with the sensitive trait in stratum i. Greenberg et al. (1969) (see Greenberg et al. (1969), p. 526) investigated how the variance of the estimator depends on the choice of P_i.

The maximum likelihood estimate $\hat{\pi}_{sw}$ of a sensitive proportion π_{sw} is given by:

$$\hat{\pi}_{sw} = \sum_{i=1}^{k} w_i\hat{\pi}_{S_i} = \sum_{i=1}^{k} w_i\left[\frac{\hat{Z}_i - (1-P_i)}{2P_i - 1}\right] \quad (2.8)$$

where $w_i = (N_i/N)$ for $i = 1,2,\cdots,k$ so that $w = \sum_{i=1}^{k} w_i = 1$ (N is the number of units in the whole population and N_i is the total number of units in stratum i) and $\hat{Z}_i$ is a point estimate of Z_i. The minimum variance of the estimator $\hat{\pi}_{sw}$ is given by:

$$Var(\hat{\pi}_{sw}) = \frac{1}{n}\left[\sum_{i=1}^{k} w_i\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2. \quad (2.9)$$

The unbiased minimum variance of the estimator $\hat{\pi}_{sw}$ follows by replacing n with $n-1$. Letting $\hat{\pi}'_{sw}$ be the estimator of the

sensitive proportion in the case of less than completely truthful "Yes" answers to the sensitive and nonsensitive statements, the mean square error of $\hat{\pi}'_{sw}$ is:

$$MSE(\hat{\pi}'_{sw}) = \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i} T_r (1 - \pi_{S_i} T_r) + \frac{P_i(1-P_i)}{(2P_i - 1)^2}\right\}^{1/2}\right]^2 + \left\{\sum_{i=1}^{k} w_i \pi_{S_i} (T_r - 1)\right\}^2 \quad (2.10)$$

3. ESTIMATION OF PARAMETERS FOR COMPLETELY TRUTHFUL REPORTING

In this paper we apply Kim and Warde's (2004) technique for Warner's RR model to Greenberg et al.'s (1969) unrelated question RR model, both for completely truthful reporting. We derive results below for π_{N_i} (the proportion with the non-sensitive trait in stratum i) both known and unknown. Since we are presenting a new RR model based on Greenberg et al.'s (1969) unrelated question RR model, we follow the rules of the choice of the probabilities P_1 and P_2 of selecting the sensitive question in samples 1 and 2, of the selection of the nonsensitive question, and the allocation of the sample numbers n_1 and n_2 with respect to the variance of the estimates (see Greenberg et al. (1969), p.526ff).

3.1. The proportion when the non-sensitive trait is known

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. An individual respondent in the sample from stratum i is instructed to use the randomization device R_i which consists of a sensitive question (S) card with probability P_i and a nonsensitive question ($\overline{S}$) card with probability $1 - P_i$. The respondent should answer the question by "Yes" or "No" without reporting which question card she or he has in order to protect the respondent's

privacy. So respondents belonging to samples in different strata will perform different randomization devices, each having different preassigned probabilities.

Let n_i denote the number of units in the sample from stratum i and n denote the total number of units in samples from all strata so that $n = \sum_{i=1}^{k} n_i$. Under the assumption that these "Yes" and "No" reports are made truthfully and P_i is set by the researcher, the probability Z_i of a "Yes" answer in stratum i for this procedure is:

$$Z_i = P_i \pi_{S_i} + (1 - P_i)\pi_{N_i} \quad \text{for } i = 1,2,\cdots,k. \tag{3.1}$$

Under the condition that π_{N_i} is known, the maximum likelihood estimate $\hat{\pi}_{S_i}$ of π_{S_i} is:

$$\hat{\pi}_{S_i} = \frac{\hat{Z}_i - (1 - P_i)\pi_{N_i}}{P_i} \quad \text{for } i = 1,2,\cdots,k \tag{3.2}$$

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

$$Var(\hat{\pi}_{S_i} \mid \pi_{N_i}) = \frac{Z_i(1 - Z_i)}{P_i^2 n_i}. \tag{3.3}$$

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 estimate $\hat{\pi}_S$ of π_S is easily shown to be:

$$\hat{\pi}_S = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} = \sum_{i=1}^{k} w_i \frac{\hat{Z}_i - (1 - P_i)\pi_{N_i}}{P_i}. \tag{3.4}$$

Theorem 3.1. The proposed estimator $\hat{\pi}_S$ is an unbiased estimate for the population proportion π_S.
Proof. This follows from taking the expected value of (3.4).

Theorem 3.2. The variance of the estimator $\hat{\pi}_S$ given π_N is:

$$Var(\hat{\pi}_S \mid \pi_{N_i}) = \sum_{i=1}^{k} \left(\frac{w_i}{P_i} \right)^2 \frac{Z_i(1-Z_i)}{n_i}. \tag{3.5}$$

Proof. Since the unbiased estimators $\hat{\pi}_{S_i}$ for all strata are independent, (3.5) is the variance of $\hat{\pi}_S$ using (3.3).

Information on π_{S_i} and π_{N_i} is usually unavailable. But if prior information on π_{S_i} and π_{N_i} is available from past experience then it helps to derive the following optimal allocation formula.

Theorem 3.3. 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:

$$\frac{n_i}{n} = \frac{\frac{w_i}{P_i}\sqrt{Z_i(1-Z_i)}}{\sum_{i=1}^{k} \frac{w_i}{P_i}\sqrt{Z_i(1-Z_i)}}. \tag{3.6}$$

Proof. Follows, for example, from Section 5.5 of Cochran (1977).

If we insert (3.6) in (3.5) the minimum variance of the estimator $\hat{\pi}_S$ given π_N is given by:

$$Var(\hat{\pi}_S \mid \pi_{N_i}) = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{w_i}{P_i}\sqrt{Z_i(1-Z_i)}\right]^2 . \qquad (3.7)$$

The unbiased minimum variance estimator of the estimator $\hat{\pi}_S$ given π_{N_i} is obtained upon replacing Z_i by $\hat{Z}_i$ and n_i by $n_i - 1$ in (3.5).

3.2. The proportion when the non-sensitive trait is unknown

In practice, π_{N_i} is rarely known and may be difficult to obtain. The following results are the π_{N_i} unknown counterparts to the preceding π_{N_i} known results. In the proposed model the population is partitioned into strata, and two independent non-overlapping simple random samples are drawn from each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known.

In this model, two sets of the randomization device in each stratum need to be employed. The first set is used for respondents in the first sample, and the second set is used for respondents in the second sample. An individual respondent in the first sample from stratum i is instructed to use the randomization device R_{i1} which consists of a sensitive question (S) card with probability P_{i1} and a nonsensitive question (N) card with probability $1-P_{i1}$. An individual respondent in the second sample from stratum i is instructed to use the randomization device R_{i2} which consists of a sensitive question (S) card with probability P_{i2} and a nonsensitive question (N) card with probability $1-P_{i2}$. The respondent should answer the question as "Yes" or "No" without reporting which question card she or he has in order to protect the respondent's privacy. So a respondent in different strata will perform different randomization devices, each having different preassigned probabilities.

Let n_{i1} denote the number of units in the first sample from stratum i, n_{i2} denote the number of units in the second sample from stratum i, and n_i denote the total number of units in two

samples from each stratum. So $n = \sum_{i=1}^{k} n_i$ is the total number of units in the samples from every strata. Under the assumption that these "Yes" and "No" reports are made truthfully, the probability of a "Yes" answer in stratum i for this procedure is:

$$\begin{aligned} Z_{i1} &= P_{i1}\pi_{S_i} + (1-P_{i1})\pi_{N_i} \\ Z_{i2} &= P_{i2}\pi_{S_i} + (1-P_{i2})\pi_{N_i} \quad \text{for } i = 1,2,\cdots,k \end{aligned} \tag{3.8}$$

where Z_{i1} and Z_{i2} are the proportions of "Yes" answers in the first and second samples, respectively, from stratum i.

From (3.8) we can derive the following:

$$\pi_{N_i} = \frac{Z_{i1} - P_{i1}\pi_{S_i}}{1-P_{i1}} = \frac{Z_{i2} - P_{i2}\pi_{S_i}}{1-P_{i2}}.$$

From the above equation, we can derive the following:

$$\pi_{S_i} = \frac{1}{P_{i1} - P_{i2}}\left[Z_{i1}(1-P_{i2}) - Z_{i2}(1-P_{i1})\right].$$

The maximum likelihood estimate $\hat{\pi}_{S_i}$ of π_{S_i} is shown to be:

$$\hat{\pi}_{S_i} = \frac{1}{P_{i1} - P_{i2}}\left[\hat{Z}_{i1}(1-P_{i2}) - \hat{Z}_{i2}(1-P_{i1})\right] \quad \text{for } i = 1,2,\cdots,k \tag{3.9}$$

where $\hat{Z}_{i1}$ and $\hat{Z}_{i2}$ are the observed proportion of "Yes" answers in the first and second samples, respectively, from stratum i.

Since $\hat{Z}_{i1}$ is a binomial distribution $B(n_{i1}, Z_{i1})$ and $\hat{Z}_{i2}$ is a binomial distribution $B(n_{i2}, Z_{i2})$, the estimator $\hat{\pi}_{S_i}$ is an unbiased estimate for π_{S_i} with:

$$Var(\hat{\pi}_{S_i}) = \left(\frac{1}{P_{i1} - P_{i2}}\right)^2 \left\{Var\left(\hat{Z}_{i1}\right)(1-P_{i2})^2 + Var\left(\hat{Z}_{i2}\right)(1-P_{i1})^2\right\}$$

$$= \left(\frac{1}{P_{i1} - P_{i2}}\right)^2 \left\{\frac{Z_{i1}(1-Z_{i1})}{n_{i1}}(1-P_{i2})^2 + \frac{Z_{i2}(1-Z_{i2})}{n_{i2}}(1-P_{i1})^2\right\}. \tag{3.10}$$

By using the Cauchy-Schwarz inequality, we can derive the following:

$$\left(\frac{1}{P_{i1} - P_{i2}}\right)^2 \left\{\frac{Z_{i1}(1-Z_{i1})}{n_{i1}}(1-P_{i2})^2 + \frac{Z_{i2}(1-Z_{i2})}{n_{i2}}(1-P_{i1})^2\right\}\{n_{i1} + n_{i2}\}$$

$$\geq \left(\frac{1-P_{i2}}{P_{i1} - P_{i2}}\sqrt{Z_{i1}(1-Z_{i1})} + \frac{1-P_{i1}}{P_{i1} - P_{i2}}\sqrt{Z_{i2}(1-Z_{i2})}\right)^2.$$

By using the inequality, we can derive the minimum variance of the estimator $\hat{\pi}_{S_i}$ as follows:

$$Var(\hat{\pi}_{S_i}) = \frac{1}{n_i(P_{i1} - P_{i2})^2} \times$$
$$\left((1-P_{i2})\sqrt{Z_{i1}(1-Z_{i1})} + (1-P_{i1})\sqrt{Z_{i2}(1-Z_{i2})}\right)^2.$$

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 estimate of π_S is easily shown to be:

$$\hat{\pi}_S = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} = \sum_{i=1}^{k} \frac{w_i}{P_{i1} - P_{i2}}\left[\hat{Z}_{i1}(1-P_{i2}) - \hat{Z}_{i2}(1-P_{i1})\right]. \tag{3.11}$$

Theorem 3.4. The proposed estimator $\hat{\pi}_S$ is an unbiased estimate for the population proportion π_S.

Proof. This follows from taking the expected value of (3.11).

Theorem 3.5. The variance of the estimator $\hat{\pi}_S$ is:

$$Var(\hat{\pi}_S) = \sum_{i=1}^{k} \frac{w_i^2}{n_i (P_{i1} - P_{i2})^2} \times \left((1-P_{i2})\sqrt{Z_{i1}(1-Z_{i1})} + (1-P_{i1})\sqrt{Z_{i2}(1-Z_{i2})}\right)^2. \quad (3.12)$$

Information on π_{S_i} and π_{N_i} is usually unavailable. But if prior information on π_{S_i} and π_{N_i} is available from past experience then it helps to derive the following optimal allocation formula.

Theorem 3.6. 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:

$$\frac{n_i}{n} = \frac{\left[\frac{w_i\left((1-P_{i2})\sqrt{Z_{i1}(1-Z_{i1})} + (1-P_{i1})\sqrt{Z_{i2}(1-Z_{i2})}\right)}{(P_{i1} - P_{i2})}\right]}{\left[\sum_{i=1}^{k} \frac{w_i\left((1-P_{i2})\sqrt{Z_{i1}(1-Z_{i1})} + (1-P_{i1})\sqrt{Z_{i2}(1-Z_{i2})}\right)}{(P_{i1} - P_{i2})}\right]}. \quad (3.13)$$

Proof. Follows, for example, from Section 5.5 of Cochran (1977).

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

$$Var(\hat{\pi}_S) = \frac{1}{n} \times \left[\sum_{i=1}^{k} \frac{w_i\left((1-P_{i2})\sqrt{Z_{i1}(1-Z_{i1})} + (1-P_{i1})\sqrt{Z_{i2}(1-Z_{i2})}\right)}{(P_{i1} - P_{i2})}\right]^2. \quad (3.14)$$

The unbiased minimum variance estimator of the estimator $\hat{\pi}_S$ is obtained upon replacing Z_i by $\hat{Z}_i$ and n_i by $n_i - 1$ in (3.12).

4. EFFICIENCY COMPARISON IN THE CASE OF COMPLETELY TRUTHFUL REPORTING

4.1. Efficiency Comparison of a Stratified Warner's RR Model with the Stratified Unrelated Question RR Model (unknown π_{N_i} case).

In this subsection we do an efficiency comparison of Kim and Warde's (2004) stratified Warner's RR model with our stratified unrelated question RR model (unknown π_{N_i} case) by way of variance comparison. We do an empirical study of the relative efficiency (RE) of $Var(\hat{\pi}_{sw})/Var(\hat{\pi}_S)$ (i.e., equations (2.9)/(3.14)) because it is difficult to derive mathematically $Var(\hat{\pi}_{sw})/Var(\hat{\pi}_S)$.

The results are in Table 1, which is available at http://tables.20m.com. We assumed $n = 1000$, two strata in the population, $\pi_N = \pi_{N_1} = \pi_{N_2}$, $P_1 = P_{11} = P_{21}$, $P_2 = P_{12} = P_{22}$, $P_1 + P_2 = 1$, and $P_1 \neq P_2$. Since all RE values in Table 1 are greater than one, our stratified unrelated question RR model is more efficient in terms of variance than Kim and Warde's (2004) stratified Warner's RR model under the assumptions given and the prior information used. Fig. 1 is a graphical representation of the Table 1 results. It illustrates that as π_S increases, the relative efficiency is decreasing. This is not a problem because the sensitive proportion is usually rare and between 0.1 and 0.3.

4.2. Efficiency Comparison of an Unrelated Question RR Model with the Stratified Unrelated Question RR Model (unknown π_{N_i} case).

In this subsection we do an efficiency comparison of Greenberg et al.'s (1969) unrelated question RR model with our stratified unrelated question RR model (unknown π_{N_i} case) by way of

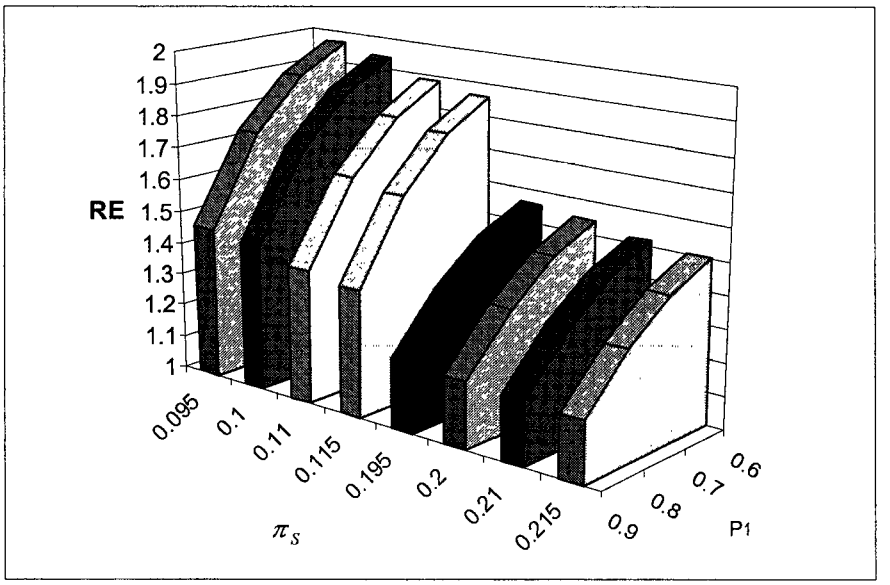

Fig. 1. Relative efficiency of $Var(\hat{\pi}_{sw})/Var(\hat{\pi}_S)$ when $\pi_N = 0.2$.

variance comparison. As in subsection 4.1, we do an empirical study of the RE of $Var(\hat{\pi}_M)/Var(\hat{\pi}_S)$ (i.e., equations (2.5)/(3.14)) because it is difficult to derive mathematically $Var(\hat{\pi}_M)/Var(\hat{\pi}_S)$.

The results are in Tables 2a and 2b, which have the same assumptions as Table 1 and are also available at http://tables.20m.com. Since all RE values in Tables 2a and 2b are, of course, greater than one, our stratified unrelated question RR model is more efficient in terms of variance than Greenberg et al.'s (1969) unrelated question RR model under the assumptions given and the prior information used. When comparing Tables 1 and 2a, we see that the gain in efficiency of our model is smaller when compared to the Greenberg et al. (1969) model that it was when compared to Kim and Warde's (2004) model. However, the gain in efficiency of our model increases when compared to the Greenberg et al. (1969) model for higher values of P_2 (see Table 2b).

Though the RE values in Tables 2a and 2b are greater than one, they are so by not much. In practice, one would have to determine if the small gain in efficiency from using our stratified unrelated question RR model over using Greenberg et al.'s (1969) unrelated question RR model is worth more than the extra cost in terms of time and effort to do the stratification.

5. ESTIMATION OF PARAMETERS FOR LESS THAN COMPLETELY TRUTHFUL REPORTING

In this section we apply Kim and Warde's (2004) stratified RR technique using optimal allocation to Greenberg et al.'s (1969) unrelated question RR model, both for less than completely truthful reporting. We derive results below for both π_{N_i} known and unknown. Since we use Greenberg et al. (1969) unrelated question RR model, we follow the same assumption as Greenberg et al. (1969) of less than completely truthful reporting; that is, the respondents confronted with a sensitive question may report untruthfully only when they have the sensitive trait but the respondents confronted with an unrelated question will report truthfully.

5.1. The proportion with the non-sensitive trait is known

For less than completely truthful reporting, we denote T_r to be the weighted probability $T_r = \sum_{i=1}^{k} w_i T_{r_i}$, where T_{r_i} is the probability that a respondent with the sensitive trait will report truthfully in a sample from stratum i. The probability $Z_i^{'}$ of a "Yes" answer in stratum i for this procedure is given by:

$$Z_i' = P_i(\pi_{S_i} T_{r_i}) + (1 - P_i)\pi_{N_i} \quad \text{where } i = 1,2,\cdots,k \tag{5.1}$$

Using results in Section 3.1, a point estimator $\hat{\pi}_S'$ of π_S in the population has the following bias and variance: for $\hat{\pi}_S' = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i}' = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} T_{r_i}$,

$$Bias(\hat{\pi}_S') = E(\hat{\pi}_S' - \hat{\pi}_S) = \sum_{i=1}^{k} w_i E(\hat{\pi}_{S_i}' - \hat{\pi}_{S_i}) = \sum_{i=1}^{k} w_i \pi_{S_i}(T_{r_i} - 1) \tag{5.2}$$

$$Var(\hat{\pi}_S') = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{w_i}{P_i}\left\{Z_i'(1 - Z_i')\right\}^{1/2}\right]^2. \tag{5.3}$$

So the mean square error of $\hat{\pi}'_S$ given π_{N_i} is given by:

$$MSE(\hat{\pi}'_S) = Var(\hat{\pi}'_S) + Bias(\hat{\pi}'_S)^2 = \frac{1}{n}\left[\sum_{i=1}^{k}\frac{w_i}{P_i}\left\{Z'_i\left(1-Z'_i\right)\right\}^{1/2}\right]^2 + \left\{\sum_{i=1}^{k} w_i\pi_{S_i}\left(T_{r_i}-1\right)\right\}^2. \tag{5.4}$$

5.2. The proportion with the non-sensitive trait is unknown

For less than completely truthful reporting, T_{r_i} is the probability that a respondent with the sensitive trait will report truthfully in two independent, overlapping samples 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:

$$\begin{aligned} Z'_{i1} &= P_{i1}(\pi_{S_i}T_{r_i} - \pi_{N_i}) + \pi_{N_i} \quad \text{and} \\ Z'_{i2} &= P_{i2}(\pi_{S_i}T_{r_i} - \pi_{N_i}) + \pi_{N_i} \quad \text{where } i = 1,2,\cdots,k \end{aligned} \tag{5.5}$$

Using results in Section 3.2, a point estimator $\hat{\pi}'_S$ of π_S in the population has the following bias and variance: for $\hat{\pi}'_S = \sum_{i=1}^{k} w_i\hat{\pi}'_{S_i} = \sum_{i=1}^{k} w_i\hat{\pi}_{S_i}T_{r_i}$,

$$Bias(\hat{\pi}'_S) = E\left(\hat{\pi}'_S - \hat{\pi}_S\right) = \sum_{i=1}^{k} w_i E\left(\hat{\pi}'_{S_i} - \hat{\pi}_{S_i}\right) = \sum_{i=1}^{k} w_i\pi_{S_i}\left(T_{r_i}-1\right) \tag{5.6}$$

$$Var(\hat{\pi}'_S) = \frac{1}{n}\times \left[\sum_{i=1}^{k}\frac{w_i\left((1-P_{i2})\sqrt{Z'_{i1}(1-Z'_{i1})} + (1-P_{i1})\sqrt{Z'_{i2}(1-Z'_{i2})}\right)}{(P_{i1}-P_{i2})}\right]^2 \tag{5.7}$$

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

$$MSE(\hat{\pi}'_S) = \frac{1}{n} \times$$
$$\left[\sum_{i=1}^{k} \frac{w_i\left((1-P_{i2})\sqrt{Z'_{i1}(1-Z'_{i1})}+(1-P_{i1})\sqrt{Z'_{i2}(1-Z'_{i2})}\right)}{(P_{i1}-P_{i2})}\right]^2 + \left\{\sum_{i=1}^{k} w_i \pi_{S_i}\left(T_{r_i}-1\right)\right\}^2 \tag{5.8}$$

5.3. Efficiency Comparison of a Stratified Warner's RR Model with the Stratified Unrelated Question RR Model (unknown π_{N_i} case) in terms of MSE.

In this subsection we do an efficiency comparison of Kim and Warde's (2004) stratified Warner's RR model in the case of less than completely truthful reporting with our stratified unrelated question RR model (unknown π_{N_i} case) by way of MSE. We do an empirical study of $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ (i.e., equations (2.10)/(5.8)) because it is difficult to derive mathematically this ratio.

The results are in Table 3, which has the same assumptions as Table 1 and is also available at http://tables.20m.com. Since all values of $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ in Table 3 are greater than one, our stratified unrelated question RR model is more efficient in terms of MSE than Kim and Warde's (2004) stratified Warner's RR model under the assumptions given and the prior information used.

Further investigation of Table 3 reveals the following:

1. As P_1 (the probability that a respondent in the sample from stratum 1 has a sensitive question (S) card) increases (and consequently P_2 decreases), $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ decreases.
2. As π_{N_i} (the proportion of people with the non-sensitive trait) increases, $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ decreases.
3. As T_r (the probability of a respondent reporting truthfully) decreases, $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ decreases.

4. As π_S (the proportion of people with the sensitive trait) increases, $MSE(\hat{\pi}'_{sw})/MSE(\hat{\pi}'_S)$ for the most part decreases.

These results are not surprising, as higher values for P_1, π_N, and π_S, and lower values for T_r will cause the performance of our stratified unrelated question RR model to decline.

5.4. Efficiency Comparison of an Unrelated Question RR Model with the Stratified Unrelated Question RR Model (unknown π_{N_i} case) in terms of MSE.

In this subsection we do an efficiency comparison of Greenberg et al.'s (1969) unrelated question RR model with our stratified unrelated question RR model (unknown π_{N_i} case) by way of MSE. As in subsection 5.2, we do an empirical study of $MSE(\hat{\pi}'_M)/MSE(\hat{\pi}'_S)$ (i.e., equations (2.6)/(5.8)) because it is difficult to derive mathematically this ratio.

The results are in Table 4, which has the same assumptions as Table 1 and is also available at http://tables.20m.com. Since all values of $MSE(\hat{\pi}'_M)/MSE(\hat{\pi}'_S)$ in Table 4 are, of course, greater than one, our stratified unrelated question RR model is more efficient in terms of MSE than Greenberg et al.'s (1969) unrelated question RR model under the assumptions given and the prior information used. When comparing Tables 3 and 4, we see that in most cases the gain in efficiency of our model is smaller when compared to the Greenberg et al. (1969) model that it was when compared to Kim and Warde's (2004) model.

6. DISCUSSION

This paper applies stratified random sampling using optimal allocation to Greenberg et al.'s (1969) unrelated question RR model for both completely truthful reporting and less than completely truthful reporting. We showed that, for the prior information given, our new model is more efficient in terms of variance and MSE than Kim and Warde's (2004) stratified RR model and Greenberg et al.'s (1969) RR model. Two more advantages exist with stratified RR models using optimal allocation. The first is that they solve a limitation of RR which is

the loss of individual characteristics of the respondents. Also, using optimal allocation helps to overcome the high cost (in terms of time, effort, and money) incurred because of the difficulty in obtaining a proportional sample from a stratum. So this paper tries to extend the methodology of the RR techniques.

ACKNOWLEDGEMENT

The authors are grateful to Editor Dr. Sylvia Frühwirth-Schnatter and two referees for their valuable comments and suggestions.

REFERENCES

Cochran WG (1977) Sampling techniques, 3rd ed. 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**, 520-539

Hong K, Yum J, Lee H (1994) A stratified randomized response technique. The 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), 155-165

Kim J-M, Elam ME (2005) A two-stage stratified Warner's randomized response model using optimal allocation. Metrika **61**, 1-7

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