American Journal of Mathematics and Statistics 2014, 4(3): 156-161
DOI: 10.5923/j.ajms.20140403.04

A Post-Stratified Randomized Response Model for Proportion

Sevil Bacanli[1,*], Tuğçe Tuncel[2]

[1]Hacettepe University, Faculty of Science, Department of Statistics, Beytepe, Ankara, Turkey
[2]Republic of Turkey Social Security Institution, Ankara, Turkey

Abstract In this study, the post-stratified randomized response (RR) models are proposed in order to estimate the proportion of persons bearing the sensitive characteristic. In addition to this, post-stratified RR models are compared according to their efficiencies. It is concluded that Kim-Warde's post-stratified RR model is more efficient than the Hong et al.'s post-stratified RR model.

Keywords Randomized response model, Post-stratified sampling, Stratified random sampling

1. Introduction

With the purpose of making people comfortable and encourage them to give truthful answers, a new survey technique was needed to eliminate non-response and response bias and this technique should be different from open and direct surveys.

The randomized response technique can be defined as a procedure of collecting the information about sensitive characteristics without revealing the identity of respondent.

The first study about randomized response technique was developed by Warner [1] as an alternative way of survey technique and is named as randomized response (RR) model. Warner's RR model is designed with different features such as estimating the proportion of people who bear a sensitive characteristic, reducing answer bias and keeping the respondents confidentiality. In oder to collect information about a sensitive characteristic, Warner [1] use a randomization device (R). This device could be a deck of cards in which each card has one of the two following two questions:

i) Do you have a sensitive characteristic (A) "selected with probability P"

ii) Do you have a non-sensitive characteristic (A^c) " selected with probability(1- P)".

In order to estimate the population proportion π belonging to the sensitive characteristic (A), a simple random sampling with replacement (SRSWR) of n respondents is drawn from the population, and the respondent is required to answer "Yes" or "No" according to her/his actual status and the statement chosen.

* Corresponding author:
sevil@hacettepe.edu.tr (Sevil Bacanli)
Published online at http://journal.sapub.org/ajms

Hong et. al. [2] proposed a stratified RR model of Warner [1] by using a proportional allocation that applied the same randomization device to each stratum. However, this model may have high expenses, because of the fact that it is difficult to acquire a proportional sample from each stratum. In order to solve this problem, Kim and Warde [3] expanded Hong et al.'s RR model with the optimal allocation and thus, each stratum sample provides different randomization devices. This demonstrated the fact that a stratified RR model with an optimal allocation is more efficient than the one with a proportional allocation. Afterwords, Son et al. [4] proposed the calibration procedure for the variance reduction of the stratified RR models, which are suggested by Hong et al. [2] and Kim and Warde [3], by using auxiliary information at the population level.

A part from the Warner's RR procedure, two-stage RR procedure is proposed by Mangat and Singh [5]. In this procedure, each interviewee with the SRSWR of n respondents is provided with two random devices. Also, Kim and Elam [6] are developed a stratified RR model by using the Mangat and Singh's RR model.

In recent years, various alternative RR models which use two decks of cards are developed. Odumade and Singh [7] proposed the use of two decks of cards in a RR model in which each of the decks included the two statements which are used in Warner's RR model. Furthermore, Abdelfatah et al. [8] [9] suggested a modified RR model of Odumade and Singh [7] by using Mangat and Singh's procedure instead of Warner's procedure in each stage. There after Hong et al. [10] developed an RR model by applying stratified sampling to Abdelfatah et al.'s RR model.

Therefore, in literature, RR models have been developed in simple and stratified sampling as the estimator of a sensitive proportion. In this study, the stratified RR models of Warner [1] are examined in post-stratified sampling by using a deck of cards.

Post stratified sampling is a very popular method among

the survey practitioners and is often used in sample surveys, when the identification of stratum cannot be achieved in advance. In post stratified sampling, first of all, a sample is selected by using simple random, and then, this selected sample is stratified into strata e.g. personal characteristics such as, age, gender, race, occupation, income, educational level and other factors. Therefore, this sampling method is particularly useful in multi-purpose surveys in which stratification factors are selected prior to sampling. [11], [12].

This article is organized as follows: In section 2, the stratified RR models of Warner [1] are briefly reviewed. In section 3, post-stratified sampling is described and how post-stratified sampling can be used in RR models are explained and RR models are compared according to their efficiencies. Conclusion is given in section 4.

2. Stratified Randomized Response Model

Let us consider a stratified random sampling with L strata. For each strata size (N_h), $h = 1,2,...,L$ and then, a sample (n_h) is selected with simple random sampling for each stratum. The number of units in each stratum is assumed to be known. Hong et al. [2] suggested a stratified RR model which applies the same randomization device to every stratum.

Each respondent in the sample stratum ($h = 1,2,...,L$) is provided with the randomization device R which includes a sensitive characteristic (A) card with probability P and its non-sensitive characteristic (A^c) card with probability $1-P$. The respondent is required to answer the question with "Yes" or "No" but s/he should not report the question card that she or he has. A respondent who is a part of the sample in different strata will perform the same randomization devices. Let n_h be the number of units in the sample from stratum h, and then, $n = \sum_{h=1}^{l} n_h$ is the total number of units in the sample from all strata. Assuming that the "Yes" or "No" reports are made truthfully and the researcher set the $P(0 < P < 1, P \neq 0.5)$, the proportion of a "Yes" answer in stratum h for this procedure is

$$P(X_i = 1)_h = Z_h = P\,\pi_h + (1-P)(1-\pi_h) \quad h = 1,...,l \tag{1}$$

where Z_h is the proportion of "yes" answer in stratum h, π_h is the proportion of respondents with sensitive characteristic in stratum h and P is the probability in a respondent's sensitive characteristic (A) card. The maximum likelihood estimator (MLE) of π_h is

$$\hat{\pi}_{h_H} = \frac{\hat{Z}_h - (1-P)}{2P-1} \quad P \neq 0.5 \tag{2}$$

where the proportion of "Yes" answer in a sample of the stratum h is $\hat{Z}_h$. The variance of $\hat{\pi}_{hH}$ can be given as

$$V(\hat{\pi}_{h_H}) = \frac{\pi_h - (1-\pi_h)}{n_h} + \frac{P(1-p)}{n_h(2p-1)^2} \tag{3}$$

Considering each $\hat{Z}_h$ is distributed with $B(n_h, Z_h)$ and selections in different strata are made independently, the estimate of is

$$\hat{\pi}_{st_H} = \sum_{h=1}^{l} W_h\ \hat{\pi}_{h_H} = \sum_{h=1}^{l} W_h \left[\frac{\hat{Z}_h - (1-P)}{2P-1} \right]. \tag{4}$$

The variance of $\hat{\pi}_{st_H}$ can be given as

$$V\left(\hat{\pi}_{st_H}\right) = \sum_{h=1}^{l} W_h^2 V\left(\hat{\pi}_{hH}\right) = \sum_{h=1}^{l} \frac{W_h^2}{n_h} \left(\pi_h (1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right) \tag{5}$$

[2]. Kim and Warde [3] introduced a different estimator from Hong et al's estimator for stratified random sampling. According to Kim and Warde [3], it is considered that each respondent in the sample stratum $h = 1,2,...,L$ is provided with the randomization device R_h and this device contains a sensitive characteristic (A) card with probability P_h and its negative question (A^c) card with probability $1-P_h$. The respondent is supposed to answer the question with "Yes" or "No" but s/he should not report the question card that he or she has. A respondent is a member of the sample in a different strata will perform different randomization devices and each of those devices has different preassigned probabilities. Assuming that the "Yes" or "No" reports are made truthfully and the researcher set the $P_h\ (\neq 0.5)$, the proportion of a "Yes" answer in stratum h for this procedure is:

$$Z_{h_{KW}} = P_h \pi_h + (1-P_h)(1-\pi_h) \quad h = 1,...,l \tag{6}$$

where $Z_{h_{KW}}$ is the proportion of "Yes" answer in stratum h, the proportion of respondents with a sensitive characteristic in stratum *h* is and, at the same time, is the probability of a sensitive question (A) card of a respondent in the sample stratum h. The maximum

likelihood estimator (MLE) of is

$$\hat{\pi}_{h_{KW}} = \frac{\hat{Z}_h - (1-P_h)}{2P_h - 1} \quad P_h \neq 0.5 \tag{7}$$

where Z_h is the proportion of "Yes" answer in a sample of the stratum h. The variance of $\hat{\pi}_{h_{KW}}$ is given by

$$V\left(\hat{\pi}_{h_{KW}}\right) = \frac{\pi_h(1-\pi_h)}{n_h} + \frac{P_h(1-P_h)}{n_h(2P_h-1)^2} \tag{8}$$

[4]. Considering the fact that is a binomial distribution and selections in different strata are made independently, the MLE estimate $\pi_{st_{KW}}$ of a sensitive proportion can be given as

$$\hat{\pi}_{st_{KW}} = \sum_{h=1}^{l} W_h \ \hat{\pi}_{h_{KW}}. \tag{9}$$

The variance of $\hat{\pi}_{st\ KW}$ is

$$V\left(\hat{\pi}_{st_{KW}}\right) = \sum_{h=1}^{l} W_h^2 V\left(\hat{\pi}_{h_{KW}}\right) = \sum_{h=1}^{l} \frac{W_h^2}{n_h}\left[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\right] \tag{10}$$

[3], [4].

3. Post-Stratified Randomized Response Model

In a post-stratified sampling, firstly, a sample of n units is selected from the population of N units by using simple random sampling. The population is stratified into L strata on the basis of some known auxiliary information. In post stratified sampling, the values of N_h, where $h = 1, 2..., L$ and may or may not be known for each sample unit which is selected with the chosen design. After that, each sample unit which are post-stratified or placed in the h^{th} stratum based on the auxiliary information is associated as such $n = \sum_{h=1}^{L} n_h$. Thus the difference between stratified and post-stratified sampling schemes is that the sub-sample size n_h is a fixed or predefined number in stratified sampling, whereas it is a random variable in post-stratified sampling [12].

As for the proportion in the population, the appropriate estimate for post-stratified sampling is

$$\bar{y}_{pt} = \sum_{h}^{l} W_h P_h. \tag{11}$$

In post-stratified sampling, the variance of p_{pt} is

$$V(p_{pt}) = \sum_{h=1}^{l} W_h^2 \left(\frac{1-f_h}{n_h}\right) P_h Q_h = \sum_{h}^{l} W_h^2 P_h Q_h E\left(\frac{1}{n_h}\right) - \frac{1}{N}\sum_{h}^{l} W_h P_h Q_h \tag{12}$$

[13]. In this situation, a general expression for $V\left(p_{tb}\right)$ can be approximated by replacing $1/n_h$ with its expected value. It is difficult to find the expected value of the reciprocal of a random variable; thus, a good approximation can be given as

$$E\left(\frac{1}{n_h}\right) \cong \frac{1}{nW_h}\left\{1 + \frac{(1-W_h)}{nW_h}\right\} \cong \frac{1}{nW_h} + \frac{(1-W_h)}{n^2 W_h^2} \tag{13}$$

[14]. By replacing this with $1/n_h$ for variance of the mean estimator in equation, the variance of the post-stratified proportion estimator can be given as

$$\begin{aligned} V\left(p_{pt}\right) &= \frac{1}{n}\sum_{h=1}^{l} W_h P_h Q_h + \frac{1}{n^2}\sum_{h=1}^{l}(1-W_h) P_h Q_h - \frac{1}{N}\sum_{h=1}^{l} W_h P_h Q_h \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{l} W_h P_h Q_h + \frac{1}{n^2}\sum_{h=1}^{l}(1-W_h) P_h Q_h \\ &= \left(\frac{1-f}{n}\right)\sum_{h=1}^{l} W_h P_h Q_h + \frac{1}{n^2}\sum_{h=1}^{l}(1-W_h) P_h Q_h. \end{aligned} \tag{14}$$

If n_h were fixed, post-stratification of proportion estimator and variance would function as the proportion estimation in the stratified random sampling under proportional allocation [12].

Therefore, in accordance with this information, RR estimators of a sensitive proportion for post-stratified sampling which is given section 2 are demonstrated as below:

The Hong K. et al. RR estimator for post-stratified sampling can be given as

$$\hat{\pi}_{pt_H} = \sum_{h=1}^{l} W_h \hat{\pi}_{h_H} = \sum_{h=1}^{l} W_h \left[\frac{\hat{Z}_h - (1-P)}{2P-1} \right]. \tag{15}$$

Subsituting (13) into (5), variance of this estimator can be obtained as such;

$$\begin{aligned} V\left(\hat{\pi}_{pt_H}\right) &= V\left(\sum_{h=1}^{l} {W_k}^2 V\left(\hat{\pi}_{h_H}\right)\right) \\ &= \sum_{h=1}^{l} \frac{{W_h}^2}{n_h} \left[\pi_h (1-\pi_h) + \left(\frac{P(1-P)}{(2P-1)^2} \right) \right] \\ &= \sum_{h=1}^{l} {W_h}^2 E\left(\frac{1}{n_h}\right) \left[\pi_h (1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right] \\ &= \sum_{h=1}^{l} {W_h}^2 \left[\frac{1}{nW_h} + \frac{1-W_h}{n^2 {W_h}^2} \right] \left[\pi_h (1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right] \\ &= \sum_{h=1}^{l} \frac{W_h (n-1)+1}{n^2} \left[\pi_h (1-\pi_h) + \frac{P(1-P)}{(2P-1)^2} \right]. \end{aligned} \tag{16}$$

Similarly, Kim-Warde's RR estimator for post stratified sampling can be defined as

$$\hat{\pi}_{pt_{KW}} = \sum_{h=1}^{l} W_{h_{KW}} \hat{\pi}_{h_{KW}} = \sum_{h=1}^{l} W_h \left[\frac{\hat{Z}_h - (1-P_h)}{2P_h - 1} \right]. \tag{17}$$

In order to obtain the variance of this estimator, (13) is subsituted into (10) as follows

$$\begin{aligned} V\left(\hat{\pi}_{pt_{KW}}\right) &= V\left(\sum_{h=1}^{l} W^2 V\left(\hat{\pi}_h\right)\right) \\ &= \sum_{h=1}^{l} {W_h}^2 \left[\frac{1}{nW_h} + \frac{1-W_h}{n^2 {W_h}^2} \right] \left[\pi_h (1-\pi_h) + \frac{P(1-P_h)}{(2P_h-1)^2} \right] \\ &= \sum_{h=1}^{l} \frac{W_h (n-1)+1}{n^2} \left[\pi_h (1-\pi_h) + \frac{P(1-P_h)}{(2P_h-1)^2} \right]. \end{aligned} \tag{18}$$

Therefore, post-stratified RR- estimators are same as stratified RR-estimators. But the variance equations are different in post-stratified sampling.

3.1. Efficiency Comparison

In this section, a comparison of the relative efficiency (RE) is carried out in post-stratified RR models. Kim and Warde [3] resorted to an empirical study on RE for stratified random sampling and it is seen that Kim and Warde's stratified RR model is more efficient than Hong et al.'s- stratified RR model. Similar efficiency comparison is carried out for post-stratified sampling.

Let us assume that there are two strata in a population in which it is considered that $N_1 = 7000$, $N_2 = 3000$ and selection probabilities of sensitive question are $P_1 = 0.6$ to 0.9 by increments for stratum 1 and P_2 is different from $P_1 \left(P_2 > P_1\right)$. The RE of two variances in post-stratification is

$$RE = \frac{V\left(\hat{\pi}_{pt}\right)_H}{V\left(\hat{\pi}_{\wedge pt}\right)_{KW}} > 1. \tag{19}$$

Table 1. The relative efficiency of post-stratified $\hat{\pi}_{ptKW}$ with respect to $\hat{\pi}_{ptH}$

					P=P1							
					0,6		0,7		0,8		0,9	
π1	π2	W1	W2	πtb	P2=0,7	P2=0,8	P2=0,8	P2=0,9	P2=0,9	P2=0,95	P2=0,93	**P_2=0,95**
0,08	0,13	0,7	0,3	0,095	1,3010519	1,3778686	1,2293027	1,3365704	1,2080672	1,2799290	1,0751461	1,1222441
		0,3	0,7	0,115	2,1621342	2,7550482	1,7530509	2,3804648	1,6380526	1,9787377	1,1794724	1,3108626
0,28	0,33	0,7	0,3	0,295	1,2933997	1,3677119	1,2070876	1,3014151	1,1627926	1,2162189	1,0475573	1,0761394
		0,3	0,7	0,315	2,1170052	2,6691203	1,6597772	2,1581389	1,4753653	1,6924629	1,1153435	1,1921376
0,48	0,53	0,7	0,3	0,495	1,2908548	1,3643409	1,2004045	1,2909540	1,1514490	1,2005105	1,0422215	1,0673905
		0,3	0,7	0,515	2,1042144	2,6450942	1,6362689	2,1050514	1,4417657	1,6368943	1,1042336	1,1724870
0,68	0,73	0,7	0,3	0,695	1,2931563	1,3673894	1,2064345	1,3003904	1,1616439	1,2146236	1,0469938	1,0752130
		0,3	0,7	0,715	2,1215834	2,6777548	1,6684806	2,1780827	1,4884568	1,7144285	1,1198449	1,2001738
0,88	0,93	0,7	0,3	0,895	1,3005400	1,3771881	1,2277072	1,3340258	1,2043551	1,2746433	1,0724027	1,1175931
		0,3	0,7	0,915	2,1720896	2,7742507	1,7761176	2,4385201	1,6875187	2,0718994	1,2032265	1,3572745

Since the value of the RE is higher than one, Kim and Warde's post- stratified RR model is more efficient than the Hong et al.'s post- stratified RR model for all P. Table 1 shows that the values of the relative efficiency are higher than one for all parameter values tabled. Therefore, Kim and Warde's RR model is more efficient in both sampling methods; stratified and post-stratified.

In terms of relative efficiency, the results obtained from stratified sampling are similar to the results obtained from post-stratification sampling. However, usage advantages for post-stratification sampling are applied to RR models.

4. Conclusions

RR models are methods for eliminating response bias by keeping the respondent's confidentiality in surveys with a sensitive characteristic such as domestic violence, drug use, sexual behaviour, family income, tax evasion etc. In this study, Warner's RR models in stratified sampling are extended into post- stratified sampling by using a deck of cards. Warner's RR models are proposed for post-stratified sampling which is more advantageous in application and it is concluded that Kim & Warde's RR model is more efficient.

In forthcoming studies, we hope to examine the use of two decks of cards in a RR model for post-stratified sampling.

REFERENCES

[1] Warner, S.L., 1965, Randomized Response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association. 60,63-69.

[2] Hong, K., Yum, J., Lee, H., 1994, A stratified randomized response technique. The Korean Journal of Applied Statistics 7, 141-147.

[3] Kim, J., Warde, W.D., 2004, A stratified Warner's randomized response model. Journal of Statistical Planning and Inference. 120, 155-165.

[4] Son, C., Hong, K., Lee, G., Kim, J., 2008, The Calibration for Stratified Randomized Response Estimators. Communications of the Korean Statistical Society Vol.15, No. 4, pp.

[5] Mangat, N.S, Singh, R., 1990, An alternative randomized response procedure. Biometrika 77: 439-442

[6] Kim, J., Elam, M., 2005, A two-stage stratified Warner's randomized response model using optimal allocation. Metrika 61, 1-7.

[7] Odumade, O., Singh, S., 2009, Efficient Use of Two Decks of Cards in Randomized Response Sampling. Communication in Statistics- Theory and Methods, 38,439-446.

[8] Abdelfatah,S., Mazloum, R., Singh, S., 2011, An alternative Randomized Response model Using Two Decks of Cards. Statistica, 3,381-390.

[9] Abdelfatah,S., Mazloum, R., Singh, S., 2013, Efficient Use of a Two-Stage Randomized Response Procedure. Brazilian Journal of Probability and Statistics, 27,4, 608-617.

[10] Hong, K., Lee, G., Son, C., Kim, J., 2014, An Estimation of a Sensitive Attribute by Two Stage Stratified Randomized Response Model. Model Assisted Statistics and Applications 9, 25-35.

[11] Holt, D., Smith, T.M.F., 1979, Post-Stratification. Journal of the Royal Statistical Society A. 142, 33- 46.

[12] Singh, S., 2003, Advanced Sampling Theory with Applications: How Micheal Selected Amy, Volume II,

Kluwer Academic Publisher, ISBN Vol. 2:1-4020-1707-3The Netherlands, 889p.

[13] Cochran, W.G., 1977, Sampling techniques. 3rd edn New York: John Wiley and sons.

[14] Hansen, M.H., Hurwitz, W.N., Madow W.G., 1953, Sample Survey Methods and Theory Volume II-Theory. Canada; John Wiley & Sons, Incorporation.